frontiers | Research Topics

Crop improvement in the era of next-generation sequencing

Edited by
Manohar Chakrabarti, Umesh K. Reddy and Nabanita Chattopadhyay

Published in
Frontiers in Plant Science

ISSN 1664-8714
ISBN 978-2-8325-6056-3
DOI 10.3389/978-2-8325-6056-3

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

Crop improvement in the era of next-generation sequencing

Topic editors

Manohar Chakrabarti — The University of Texas Rio Grande Valley, United States

Umesh K. Reddy — West Virginia State University, United States

Nabanita Chattopadhyay — The University of Texas Rio Grande Valley, United States

Citation

Chakrabarti, M., Reddy, U. K., Chattopadhyay, N., eds. (2025). *Crop improvement in the era of next-generation sequencing*. Lausanne: Frontiers Media SA.

doi: 10.3389/978-2-8325-6056-3

Table of contents

frontiers | Frontiers in Plant Science

TYPE Review
PUBLISHED 02 June 2023
DOI 10.3389/fpls.2023.1204813

OPEN ACCESS

EDITED BY
Manohar Chakrabarti,
The University of Texas Rio Grande Valley, United States

REVIEWED BY
Yiping Qi,
University of Maryland, College Park, United States
Dhananjay K. Pandey,
Amity University, Jharkhand, India

*CORRESPONDENCE
Thomas Lübberstedt
thomasl@iastate.edu

RECEIVED 12 April 2023
ACCEPTED 17 May 2023
PUBLISHED 02 June 2023

CITATION
Cook TM, Isenegger D, Dutta S, Sahab S, Kay P, Aboobucker SI, Biswas E, Heerschap S, Nikolau BJ, Dong L and Lübberstedt T (2023) Overcoming roadblocks for *in vitro* nurseries in plants: induction of meiosis. *Front. Plant Sci.* 14:1204813. doi: 10.3389/fpls.2023.1204813

Overcoming roadblocks for *in vitro* nurseries in plants: induction of meiosis

Tanner M. Cook[1], Daniel Isenegger[2], Somak Dutta[3], Sareena Sahab[2], Pippa Kay[2], Siddique I. Aboobucker[1], Eva Biswas[3], Seth Heerschap[4], Basil J. Nikolau[5], Liang Dong[4] and Thomas Lübberstedt[1]*

[1]Iowa State University, Department of Agronomy, Ames, IA, United States, [2]Agriculture Victoria, Agribio, La Trobe University, Melbourne, VIC, Australia, [3]Iowa State University, Department of Statistics, Ames, IA, United States, [4]Iowa State University, Department of Electrical and Computer Engineering, Ames, IA, United States, [5]Iowa State University, Roy J. Carver Department of Biochemistry, Biophysics, and Molecular Biology, Ames, IA, United States

Efforts to increase genetic gains in breeding programs of flowering plants depend on making genetic crosses. Time to flowering, which can take months to decades depending on the species, can be a limiting factor in such breeding programs. It has been proposed that the rate of genetic gain can be increased by reducing the time between generations by circumventing flowering through the *in vitro* induction of meiosis. In this review, we assess technologies and approaches that may offer a path towards meiosis induction, the largest current bottleneck for *in vitro* plant breeding. Studies in non-plant, eukaryotic organisms indicate that the *in vitro* switch from mitotic cell division to meiosis is inefficient and occurs at very low rates. Yet, this has been achieved with mammalian cells by the manipulation of a limited number of genes. Therefore, to experimentally identify factors that switch mitosis to meiosis in plants, it is necessary to develop a high-throughput system to evaluate a large number of candidate genes and treatments, each using large numbers of cells, few of which may gain the ability to induce meiosis.

KEYWORDS

meiosis induction, *in vitro* biology, *in vitro* nurseries, high-throughput detection, plant breeding and biotechnology

1 Introduction to *in vitro* nurseries

Globally the number of undernourished people is expected to increase to 840 million by 2030 (FAO et al, 2020). Even though we need to produce more food in the future, the current levels of food production are at risk as climate change has the potential to disrupt food availability (Brummer et al., 2011; De La Fuente et al., 2013; Brown et al., 2015; US Embassy and Consulate in Italy, 2022). Innovative breeding techniques to improve food security to increase genetic gains are needed. Genetic gain is associated with selection

intensity, heritability, genetic variance, and the time needed for a breeding cycle (Figure 1; Li et al., 2018). Several breeding methods and technologies have been developed to increase genetic gain by reducing the time needed to complete a breeding cycle, and these include winter nurseries, doubled-haploids (Geiger, 2009; De La Fuente et al., 2013; Boerman et al., 2020), speed breeding (Watson et al., 2018; Jähne et al., 2020), gene editing or gene expression regulation (Gao et al., 2020; Pan et al., 2022; reviewed in Zhang et al., 2018), marker-assisted selection (Karunarathna et al., 2021; López-Malvar et al., 2021; Tibbs Cortes et al., 2021), genomic selection (Bhat et al., 2016; López-Malvar et al., 2021), phytohormonal induction of early flowering (Espinosa et al., 2017), and combination of doubled-haploids with other breeding strategies such as gene editing during haploid induction and genomic prediction (Kelliher et al., 2019; Wang et al., 2020). These technologies focus on reducing the number of generations needed to develop a line, or reducing the time to flowering and seed production, thereby allowing more generations per unit of time. Technologies such as speed breeding and hormone manipulation provide earlier flowering times but are only available for a limited number of crop species, (Iqbal et al., 2017; Watson et al., 2018; Jähne et al., 2020), and they still require that plants produce floral organs and gametes for sexual reproduction.

As a paradigm shift, Murray et al. (2013) and De La Fuente et al. (2013) suggested the concept of a cell-based *in vitro* breeding system (called *in vitro* nurseries; IVNs). In IVNs, breeding cycle time could be substantially reduced by enabling rapid cell-level breeding cycles, without the need for flowering. Somatic tissue from parental lines could be cultured and challenged to induce haploid cells after recombination without gametophyte development (will be referred to as artificial gametes throughout the text), these cells can then be fused artificially to develop sexual pairing *in vitro*. In addition to time, IVNs will significantly reduce required field space and avoid exposure to environmental risks in field settings. The benefits of IVNs are of particular interest for species with a long generation time. Some woody species do not produce flowers for more than 30 years (Hackett, 1985). For example, poplar trees provide many ecosystem services such as phytoremediation, substrate for biofuels, and other bioproducts (Zalesny et al., 2020), but can take 10 years to flower (Hsu et al., 2006). Coffee trees do not flower until the second year, and it is not until the third year that they reach maturity (Santos et al., 2019). Even in annual crops such as maize, where two generations per year are routinely completed using winter nurseries, more generations per year would increase the annual genetic gain significantly.

Successful implementation of IVNs will require systematically overcoming a variety of bottlenecks. We anticipate three distinct phases (Figure 2). Phase I addresses the main bottleneck: meiosis induction (or meiosis-like recombination followed by reductional division) from somatic vegetative tissue, which currently is unavailable in plants. We assume that there will be a need to evaluate a potentially large number of genes, external treatments, and their combinations before identifying a path for meiosis induction in plants. Thus, an assay for high-throughput and low-cost screening of

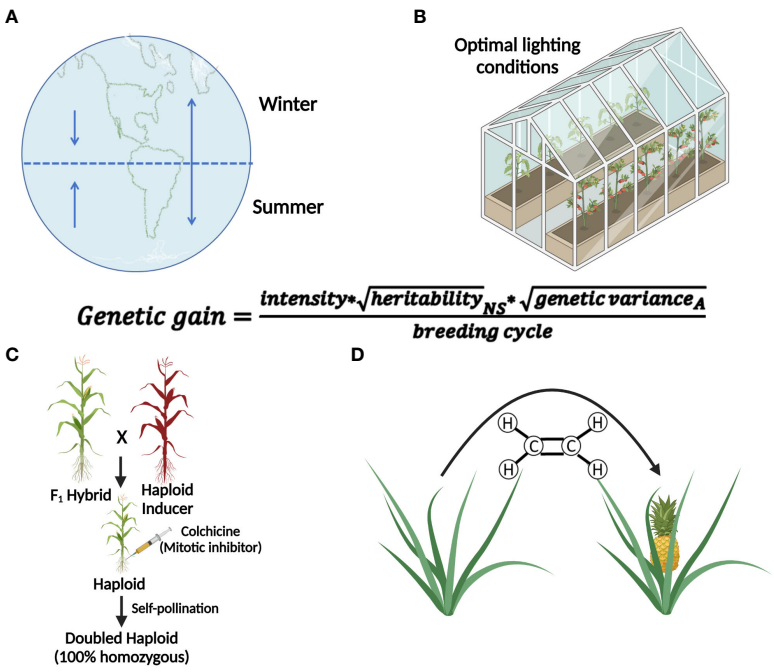


FIGURE 1
Methods that accelerate breeding cycles or generation time of plants. The genetic gain equation provides the basis for understanding how the use of each technique leads to genetic gain **(A)** The use of winter nurseries to capitalize on proximity to the equator or opposite hemispheric seasons, increase diurnal photoperiods, and warmer temperatures to increase the number of growing seasons per year. **(B)** Speed-breeding techniques that utilize optimal lighting conditions to induce early flowering to decrease breeding cycle times. **(C)** Doubled-haploid technology reduces the time needed to develop a homozygous line. **(D)** Use of chemicals (e.g. ethylene) to induce early flowering and fruit development, decreasing breeding cycle time. Created with BioRender.com.

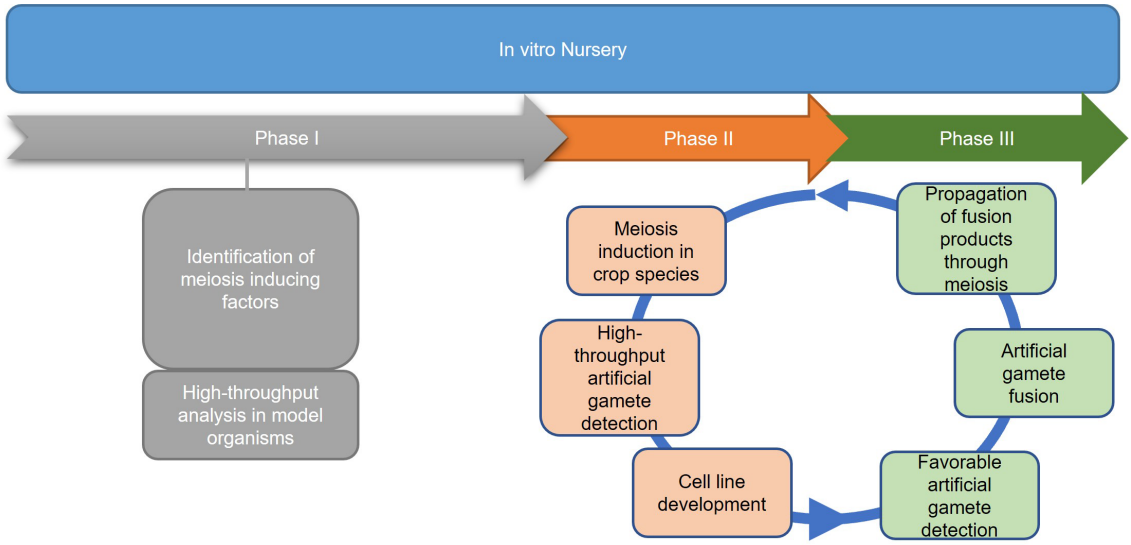


FIGURE 2
Three proposed phases of an *in vitro* nursery. Phase I emphasizes the identification of meiosis-inducing factors as a major obstacle in establishing IVNs. Identification of the meiosis-inducing factors needs a quantitative, high throughput assay to detect meiosis. Phase II involves the development of the nursery in the context of a specific crop species, using the meiosis-inducing factors identified in Phase I and developing haploid cell lines for the selection of desirable traits. Phase III identifies haploid cell lines that are expressing desirable traits by the use of markers and traditional breeding tools. After diploidization, the products are propagated by mitosis for further cycling through Phases II and III.

candidate genes or treatments for meiosis induction is needed. The other two phases are based on the successful development of protocols for meiosis induction. Phase II addresses artificial gamete formation and identification. Identifying and isolating these artificial gametes in a mixture consisting primarily of somatic cells will be critical for manipulation in the next phase. Phase III includes the assessment of induced artificial gametes that carry favorable alleles using genomic selection methods. This can only be done after artificial gametes have been isolated and developed into cell lines. Only then, can a sample of cells from each line be sacrificed for DNA isolation and genotyping for genomic selection. Further, this phase includes the fusion of selected artificial gametes to generate diploid cells, as a starting point for the next generation in IVNs. In this review, we will assess the concept of a cell-based *in vitro* breeding system, which circumvents the need for flowering (De la Fuente et al., 2013; Murray et al., 2013). The overall objective of this paper is to investigate the feasibility of Phase I through the (i) identification of bottlenecks and uncertainties (ii) while proposing possible solutions, and thus (iii) providing a starting point for the development of IVN technologies.

2 Eukaryotic meiosis

2.1 Meiosis in plants

Sex is a fundamental process shared among eukaryotes (Colnaghi et al., 2020), with meiosis being a key step to generating variation by recombining genomes. Meiosis consists of DNA replication followed by two divisions that reduce the genome size by half (Mercier et al., 2015). During meiosis, chromosomes recombine via crossovers (COs), a mechanism to reshuffle genes and respective physically linked alleles on a chromosome (Mercier et al., 2015). The major obstacle in establishing IVNs in plants is the inability to induce meiosis outside of the male or female reproductive cell structures of the flower. To be practical for IVNs, *in vitro* meiosis induction has to be based on a limited number of factors to enable a practical, routine application for artificial gamete formation.

In contrast to the predetermined germline of animals, the transition from vegetative to reproductive growth in plants occurs later in development where archesporial cells are generated from primordia, beginning the plant germline (Zhou et al., 2017). Differences between plant and human germline development have been outlined in Figure 3. In angiosperms, gametogenesis is a highly conserved process and occurs within specialized tissues of the anther and the ovule. The production of gametes proceeds in two steps: sporogenesis, followed by gametogenesis. In the anther, hypodermal archesporial cells divide to produce outer primary parietal cells which become somatic tissues, and inner primary sporogenous cells which will then divide mitotically to become microspore mother cells that undergo meiosis (Lora and Hormaza, 2021). Ovule initiation arises from the medial meristem tissue within the carpel. Immediately following ovule initiation three distinct regions arise: funiculus, chalaza, and nucellus. The nucellus gives rise to the megaspore mother cell which undergoes meiosis to generate four megaspores, three of which degrade leaving a single functional megaspore (Lora and Hormaza, 2021).

2.2 Controlling meiosis in other eukaryotes

Since the underlying evolutionary path of sexual reproduction is thought to have evolved only once in eukaryotes (Goodenough and Heitman, 2014), there is insight to gain from other non-plant,

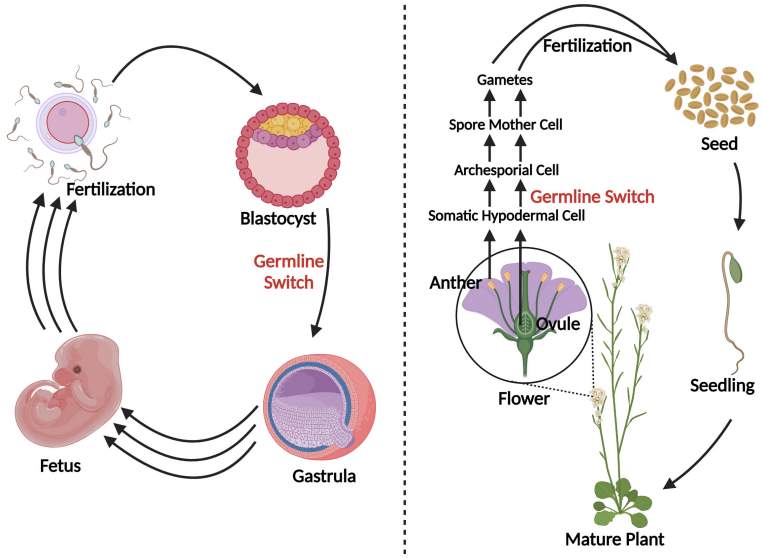


FIGURE 3
A schematic comparison of germline development in humans vs. plants. Blastocyst formation takes approximately five days (Popovic et al., 2021), after which germline in humans develops and can be detected as early as the beginning of gastrulation (Wen and Tang, 2019). Three sets of arrows indicate additional developmental processes that were not discussed. In plants, germline development is much later in a plant's life cycle, occurring during flowering. Somatic hypodermal cells divide to develop an archesporial cell, which is considered the origin of the germline (Zhou et al., 2017). Archesporial cells will then form spore mother cells which undergo meiosis to develop gametes. Created with BioRender.com.

eukaryotic species. For example, Medrano et al. (2016) found that human somatic cells can be converted into germline-like cells with the ectopic expression of six genes. These genes, *PRDM1*, *PRDM14*, *LIN28A*, *DAZL*, *VASA*, and *SYCP3*, have conserved regions in many plant species and the products have been shown to have important functions in such processes such as the repression of transposable elements, nucleic acid binding, and stem cell maintenance in human meiotic processes (Bateman, 2019; Howe et al., 2020). About 1% of these germline-like cells were able to complete meiosis (Medrano et al., 2016). Further, overexpression of human CD61 (integrin-β3) in canine adipose-derived mesenchymal stem cells led to the upregulation of markers for primordial germ-like cells (Fang et al., 2017). Vernet et al. (2020) suggested that the exogenous application of retinoic acid may force meiosis induction in mice. *In vitro* mouse studies of spermatogonia overexpressing telomerase catalytic component, mTERT, resulted in cells that could be induced to undergo meiosis *in vitro*, with the application of stem cell factor (Feng et al., 2002; Griswold, 2005; Riou et al., 2005). This outcome suggests that sex cells can be developed without structure-specific nurse cells, which is encouraging in the case of IVNs. In *S. cerevisiae*, antisense transcription was found to control meiotic cell entry by regulating IME4 (Initiator of Meiosis 4), an RNA methyltransferase (Hongay et al., 2006). This gene is also expressed in the testes and ovaries of *Drosophila* (Hongay and Orr-Weaver, 2011). In *Arabidopsis thaliana*, MTA (mRNA adenosine methylase), which is a homolog of IME4, was found to be essential for embryogenesis (Zhong et al., 2008). In addition to IME4, nutritional stress can also induce meiosis in yeast (Mata et al., 2002). Taken together, such studies in other eukaryotic species suggest that meiosis or meiotic precursors can be artificially induced and that this may also be achievable with plants.

2.3 Plant genetic factors involved in meiosis induction

In plants, a limited number of genetic factors have been identified to play a role in meiosis induction or early meiotic processes by studying aberrant phenotypes presented by mutant alleles. Table 1 summarizes the genes that have been found in previous studies. Maize AMEIOTIC 1 (AM1) is required for meiotic progression while it is also likely required for meiosis initiation as premeiotic cells with *am1* mutations led to mitosis instead of meiosis (Pawlowski et al., 2009). The SWI(SWITCH1)/DYAD protein is a putative homolog of AM1 in *Arabidopsis*, but its role appears to be more important in early meiosis instead of initiation. Consistent with this, SWI/DYAD maintains chromatid cohesion during meiosis as a WINGS APART-LIKE antagonist (Pawlowski et al., 2009; Yang et al., 2019). Evidence for the role of FEHLSTART (FST), a basic helix-loop-helix protein, in meiosis is shown with early meiotic entry in *Arabidopsis* mutants, and these mutants also show meiotic asynchrony (Li et al., 2015). KRP4, KRP6, and KRP 7 (KIP-RELATED PROTEIN 4,5,6) as well as RETINOBLASTOMA RELATED1 (RBR1), prevent the formation of supernumerary meiocytes from forming next to an already existing meiocyte while the repression of WUSCHEL (WUS) by RBR1 allows entry into meiosis (Zhao et al., 2017). An RNA-helicase (RH17) was found to play a role in reproduction as supernumerary reproductive cell lineages developed at a rate of over 20% in lines that were heterozygous for an *rh17* mutant allele (Stein et al., 2021). Evidence for potential clues in phase change induction is further supported by the *Mitosis instead of Meiosis* phenotype in rice and *Arabidopsis*, where mutations in only three genes prevent meiotic cell entry and instead meiocytes in the

TABLE 1 Genes involved in meiotic entry and regulation.

Gene/Nucleic Acid	Meiotic Role*	Species	Source
SWI1/DYAD/Am1	Cohesion, progression, and initiation	Arabidopsis/Rice/Maize	Pawlowski et al., 2009; Yang et al., 2019
MEL2	Regulates premeiotic G1/S-phase transition and synchrony	Rice	Nonomura et al., 2011
SPL/NZZ	Meiotic entry, meiotic fate acquisition, and ovule development	Rice/Arabidopsis	Wei et al., 2015; Ren et al., 2018
RBR1	WUS repression leading to meiotic entry	Arabidopsis	Zhao et al., 2017
FST	Meiotic entry and synchrony	Arabidopsis	Li et al., 2015
MIL1	Initiation and differentiation	Rice	Hong et al., 2012
AGO9/AGO104	Cell fate specification	Arabidopsis/Maize	Olmedo-Monfil et al., 2010; Singh et al., 2011
DTM1	Tapetum development and meiotic prophase 1 progression	Rice	Yi et al., 2012
MEI1	Meiotic-specific DNA repair	Arabidopsis	Mathilde et al., 2003
CDC45	Correct meiotic division progression	Arabidopsis	Stevens et al., 2004
XRI1	Meiotic DNA repair	Arabidopsis	Dean et al., 2009

*All homologous genes may not have all of the roles listed.
Many of these genes are reviewed in more detail by Mercier et al. (2015) and Wang et al. (2021).

gametophyte undergo mitosis (Mieulet et al., 2016). This phenotype can be developed with triple mutations in *REC8* and *OSD1* in combinations with either *SPO11* or *PRD1,2,3* mutants in *Arabidopsis* and with the combination of *REC8*, *PAIR1*, and *OSD1* mutations in rice (Mieulet et al., 2016). These previous studies lay a strong foundation on which we can build an understanding of meiosis induction in plants.

2.4 Plant hormonal and environmental factors

Hormonal cues from surrounding somatic tissue in the developing gametophyte also affect meiotic processes. Auxin signaling is likely to provide cues for the differentiation of egg cells vs. synergid cells in the egg apparatus (Sun et al., 2021). Auxin and brassinosteroids are important factors in meiocyte development as peak expression in biosynthesis and signaling is found in meiotic anthers (Dhaka et al., 2020). In addition, an auxin gradient appears to play a role in male germ cell development (Zheng et al., 2021). Cytokinin is shown to play a role in meiotic processes as well. For example, cytokinin histidine kinase receptors, AHK2, AHK3, and CRE1, are attributed with the ability to sense environmental cytokinin to create a kinase cascade, while triple knockouts of these three genes, result in cytokinin unresponsive plants (Inoue et al., 2001; Higuchi et al., 2004; Cheng et al, 2013). Loss of function with these cytokinin receptors results in female gametophytic lethality but can be recovered via TDNA complement insertions (Higuchi et al, 2004; Cheng et al, 2013). Environmental conditions such as hypoxia and oxidation-reduction have also been shown to induce meiotic fate (Kelliher and Walbot, 2012). Mutations in genes associated with redox reactions, like *MSCA1* (maize), *MIL1* (rice), *ROXY1*, and *ROXY2* (*Arabidopsis*), led to fertility disruptions (Xing and Zachgo, 2008; Hong et al., 2012; Kelliher & Walbot, 2012). Further, a switch from apomeiosis to meiosis occurs with increased oxidative stress treatment in *Boechera* premeiotic ovules (Mateo de Arias et al., 2020). In *Arabidopsis*, retinal was determined to be an endogenous metabolite that plays a role in root organogenesis and root clock functions (Dickinson et al., 2021). Interestingly, TEMPERATURE INDUCED LIPOCALIN (TIL) acts as a retinal binder in plants with protective functions in heat stress, light stress, and oxidative stress (Chi et al., 2009; Boca et al., 2014; Dickinson et al., 2021). Evidence for a stress-mediated switch between meiosis and apomeiosis has been demonstrated (Mateo de Arias et al., 2020), and since retinoic acid may force meiosis induction in mice (Vernet et al., 2020), the closely related retinal may function in stress response in plants, and possibly be of interest to explore for meiotic induction. Moreover, the number of candidate genes and factors for meiosis induction has grown substantially in the past years. However, an efficient test system is needed to determine the relevance of candidate factors in meiosis initiation.

3 Tools for meiotic factor testing

3.1 Cell-based system

Detailed analyses of specific genetic factors and growth hormones provide a great starting place to begin testing factors as meiotic induction candidates but low induction rates in mammals (Medrano et al., 2016) suggest that a high-throughput system is required to evaluate these candidates.

High-throughput, single-cell culture systems, such as protoplasts, may provide a robust approach to detecting the rare meiotic events induced by multiple factors. Protoplasts are spherical-shaped cells that are devoid of the cell wall, removed by enzymatic digestion, and can provide totipotent homogeneous populations of cells useful for plant genetic improvement studies in some species (Davey et al., 2005; Eeckhaut et al, 2013; Sahab et al., 2019). An important factor for the viability of protoplasts is maintaining osmotic stabilization to

prevent cell lysis after cell wall removal (Marx, 2016; Reed and Bargmann, 2021). Protoplast-based platforms can allow for the hybridization of different species via protoplast fusion and plant regeneration (Melchers et al., 1978). Moreover, protoplasts can enable the exploration of signal transduction and metabolic pathways (Sheen, 2001), cell type-specific functions (Petersson et al., 2015; Denyer et al., 2019), and determine the subcellular localization, transport, and interactions of intracellular proteins (Goodman et al., 2004; Zhang et al., 2011). Cellular division and subsequent regeneration from protoplasts have been reported in numerous species with varying levels of efficiency (Nagata and Takebe, 1971; Xu et al., 1982; Shillito et al., 1989; Kiełkowska and Adamus, 2012; Chupeau et al., 2013; Jeong et al., 2021). However, recalcitrance to protoplast regeneration has also been observed across many species and is particularly challenging in monocotyledonous species (Hahne et al., 1989; Xu et al., 2022). There is a range of factors that can influence establishing reliable protoplast transient assays and regeneration protocols (reviewed in Reed and Bargmann, 2021).

3.2 *In vitro* meiosis induction testing system using single cells

Using protoplasts allows many cells to be analyzed at one time while also providing the potential to be collected and used in downstream IVN experiments in addition to simple ploidy analysis. With these single cells, two options have been considered for high-throughput screening of meiosis induction. First, protoplasts can be isolated and then challenged to undergo cellular division. Division would then be followed by a meiotic induction treatment from which dividing cells can be reisolated for ploidy-state analysis. This option can be laborious but provides a means for the analysis of introduced genetic factors (Yoo et al., 2007). The second option can utilize dividing callus, which can be treated with meiotic induction factors followed by protoplast isolation for ploidy-state analysis. This system is potentially less laborious and enables efficient testing of exogenous factors, but the assessment of genetic elements would rely on an efficient transformation system. Both options, however, require callus formation as a result of cellular division of which cell lines could be maintained for analysis and subsequent selection. These protoplast-based approaches would also benefit from culture suspension as multiple factors could be tested while easily moving aliquots of cells for processing and systematic treatment application.

Efficient delivery of genetic elements and the induction and detection of meiosis may be difficult to establish in protoplasts, as cell survival, fitness, and division can be impacted by the product of transgenes and mutagenesis. To test genetic factors, DNA delivery into the cells is required. Conventionally, transgenic plants can be generated via the delivery of DNA-encoding gene constructs via microprojectile bombardment or *Agrobacterium*-mediated transformation (Cunningham et al, 2018). Other types of transformation methods can utilize nanoparticles (Cunningham et al, 2018; Mao et al., 2019), electroporation, microinjection, PEG-mediated direct delivery in protoplasts (Yoo et al, 2007), and viral-vectors (Catoni et al., 2018; reviewed by Abrahamian et al., 2020). Alternatively, DNA-free CRISPR/Cas genome editing systems can deploy ribonucleoproteins to cells using similar DNA delivery approaches for targeted mutagenesis of genes or regulatory regions to modulate the expression of genes (Woo et al, 2015; Liu et al., 2020; Ma et al., 2020; Zhang et al., 2021). In protoplasts, transcriptional regulation was also an effective means to control gene expression and could be multiplexed (Pan et al., 2021). Given the potential difficulties of genetic element testing in protoplasts, however, protoplasts derived from callus may be highly suited for evaluating chemical factors which can be simply applied in culture media. For instance, chemical factors have been applied to callus cultures to test cell cycle regulation and ploidy increase (Wan et al., 1989; Elmaghrabi et al., 2017), while hormones and stress factors added to callus culture media have been evaluated to increase metabolite production (Beygi et al., 2021). Interestingly, chemical agents can be assessed and deployed to reduce chromosome number, somewhat like haploids or meiosis-like reductions. For example, decades ago a chloramphenicol antibiotic treatment was shown to reduce chromosomes to a haploid state in root cells of barley seedlings (Yoshida and Yamaguchi, 1973). Caffeine treatments have been shown to induce somatic meiosis-like reductions in *Vicia* root tips (Chen et al., 2000). Meiosis-like reductions have also been observed in somatic embryogenic callus cultures of *Arabidopsis* (Yihua et al., 2001) and non-embryogenic carrot cell culture lines that were considered to be permanently expressed in a meiotic or sporogenous tissue state (Ronchi et al., 1992). These studies indicate the potential for screening and deploying chemical agents on explant tissue sources and in *in vitro* culture for meiosis induction in an IVN system. However, the reliable and efficient development of such approaches likely requires extensive work and validation that may vary across different species (Yan et al., 2017).

3.3 High-throughput fluorescence analysis

A potentially efficient method to detect and quantify meiosis induction with callus-derived protoplasts is through the use of a transgenic, bi-fluorescent system to track chromosomal segregation after meiotic cell division. By utilizing a dual marker system, an assay to detect the induction of meiosis can be achieved based on the presence or absence of fluorescence signals in cells through fluorescence-activated cell sorting (FACS) instruments (Bargmann and Birnbaum, 2010; Borges et al., 2012; Ortiz-Ramírez et al., 2018). DNA content analysis using FACS or flow cytometry can also be used to determine artificial gametes (haploid) and somatic cells (diploid), but traditional stains such as 4',6-diamidino-2-phenylindole and propidium iodide inefficiently pass through intact cell membranes (Wallberg et al., 2016), while there are other commercially available stains for DNA analysis of live cells, these will have to be controlled for and considered when testing meiotic candidates. These factors present difficulties for further downstream uses in IVN's and are the basis for the fluorescent system development suggestion. In the proposed fluorescent system, two different fluorescent single-copy reporter genes such as RFP or GFP can be integrated into the genome either in allelic or non-allelic positions (Figure 4) and detected without the need for DNA stains.

Generating a genotype with two different markers in allelic positions is more complex than for non-allelic markers. For example, one way to obtain different marker genes in allelic positions would be to establish a homozygous, fluorescent marker line, and use gene-editing techniques to replace this marker with another fluorescent marker in the allelic position. Low rates for homology-directed repair (HDR) in plants have prevented such targeted knock-ins from being efficiently accomplished. Recent developments, however, have provided more efficient approaches with HDR rates being reported at levels as high as 6.3% (Sun et al., 2023), 3.2% (Wang et al., 2023), 9.1% (Miki et al., 2018); and targeted T-DNA integration via *Agrobacterium*-mediated transformation in rice ranging from 4 to 5.3% (Lee et al., 2019). Moreover, constructs encoding a Cas9-VirD2 fusion have succeeded in improving HDR-mediated integration in rice transformation as well (Ali et al., 2020).

An informative marker tool is possible by using the resulting F1 progeny from a cross between parents carrying different fluorescent markers in allelic positions. Haploid cells would express only one fluorescent marker while diploid cells would express both (Figure 4). Repressor/activator systems such as the Q-system from *Neurospora crassa* may provide another option to track chromosomal segregation, as the presence of a repressor in a diploid containing an activator would prevent the expression of a marker, but in a haploid, the marker would be expressed due to the absence of the repressor. The opposite effect could also be obtained using only a transcriptional activator in the Q-system. Transcriptional activation also controls the Gal4/UAS system where the presence of Gal4 would lead to marker expression in a diploid while a haploid would be repressed. Both systems have been established as molecular tools in plant systems (Waki et al., 2013; Persad et al., 2020). These systems would still require allelic positioning to be relevant in IVNs, in addition to overcoming false identification with leaky signaling. These site-specific allelic placements of reporter genes could also be achieved using recombinases such as the Cre-lox and FLP-FRT gene-stacking system (Nandy et al., 2015).

Alternatively, a non-allelic bi-fluorescent reporter system may provide a readily available alternative to this process. The non-allelic, hybrid line would only require establishing, two single-locus fluorescent marker lines that would be crossed, and the resulting F_1, which would carry both markers, can be used for testing meiosis-inducing factors. However, this system has a decreased efficiency caused by a 50% reduction of "informative" artificial gametes, expressing only a single fluorophore compared to the allelic system. However, the speed of development of marker lines provides a relevant strategy for Phase I (Figure 4).

3.4 RNA sequencing and fusion technologies

Technological advances in RNA sequencing may also contribute to determining meiosis-induction factors in plants. Single-cell RNA sequencing (scRNA-seq) has proven to be an efficient and cost-effective approach to analyzing multiple tissue types in response to treatment (Shin et al., 2019; Shulse et al., 2019). This technology could be used as an assay to assess candidate meiotic induction factors and provide expression data. scRNA-seq with barcoding permits the sequencing of multiple samples through multiplexing, which allows for simultaneous evaluation of multiple treatments and factors (Macosko et al., 2015; Shin et al., 2019; Shulse et al., 2019). Nelms and Walbot (2019) demonstrated the use of this tool to determine gene expression profiles during different stages of meiosis development. Comparative studies of germline and somatic cells have provided insight into differential gene expression in the meiosis of plants (Dukowic-Schulze et al., 2014a; Nelms and Walbot, 2019; Barakate et al., 2021; Dukowic-Schulze et al., 2014b). The availability of reference genes for meiotic processes also provides an opportunity for factors to be tested using quantitative PCR (Ji et al., 2014; Nelms and Walbot, 2019; Garrido et al., 2020). Therefore, these molecular tools could be used to determine the onset of artificial meiosis induction. It has also been considered, that meiosis induction could be assessed by

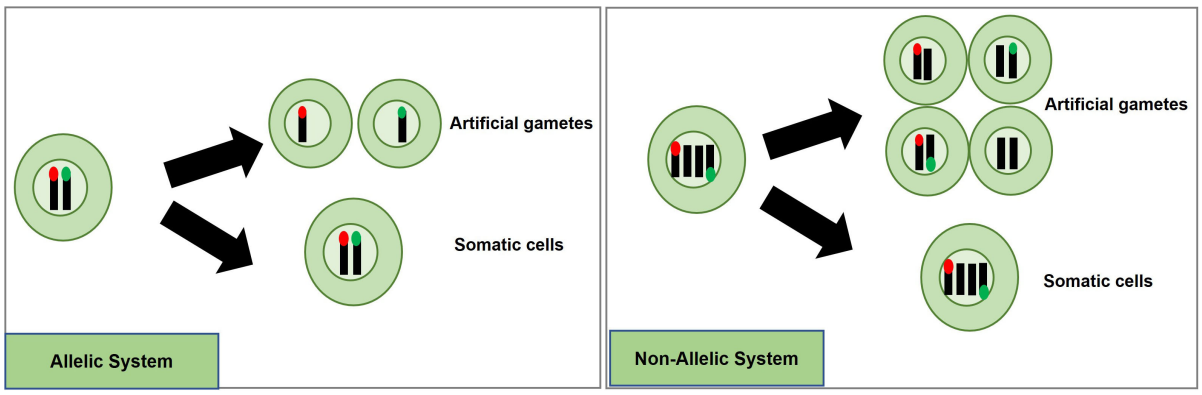


FIGURE 4
Proposed high-throughput meiosis induction detection tools using bi-fluorescent single cells to track chromosomal segregation. The left panel depicts an allelic system where chromosomal segregation can be identified 1:1. The right panel depicts a non-allelic system where chromosomal segregation can be detected with 50% less efficiency. Red and green dots represent different fluorescent markers on a chromosome, arrows indicate a treatment, and cells right of the arrows are the potential products after treatment, either artificial gametes (if meiosis was induced) or somatic cells (if meiosis was not induced).

fluorescent fusions with meiosis-specific genes such as the *PRO*$_{REC8}$*:REC8:GFP* line developed by Prusicki et al. (2019). These methods would provide evidence for specific steps in meiosis and could be scaled for high-throughput investigation. Given this, however, fluorescent-based markers may provide many benefits to tracking and assessing meiosis induction, especially with up-scaling and cost reduction using available commercial instruments, as laborious nucleic acid isolation would not be required.

3.5 Statistical approaches for the detection of rare events

The complexity of datasets and inherent variance expected among biological samples would require the optimization of robust statistical analysis methods to detect and discriminate artificial gametes at low meiotic induction rates. The method would rely on analyzing a multitude of data points and determining which factor(s), if any, play a role in meiosis induction. The limit of detection must be possible with induction rates as low as 1%, based on *in vitro* meiosis induction rates found in human cells (Medrano et al., 2016). Hence, large totipotent protoplast populations that can be analyzed are preferred.

The two bi-fluorescent systems outlined above can use flow cytometry or FACS to detect the different cell populations that show different fluorescence signals. Analysis of fluorescence values can be done using the popular method of "gating" (Adan et al., 2017). Gating is a technique where regions of fluorescence are manually selected to identify events, in our case artificial gametes and diploid cells. Figure 5 depicts a theoretical gating approach, where a balance between accuracy of cell identification and the number of cells identified must be reached. Cells containing both fluorescent markers (i.e., diploid cells), would show similar fluorescence signals for both markers (i.e., population near the middle of the plot), and cells containing only one of the fluorescent markers (i.e., haploid cells), would predominantly show fluorescence of one of the two markers (i.e., population near either of the two axes). Borges et al. (2012) used a similar FACS method to successfully sort nuclei tagged with either RFP in vegetative nuclei or GFP in sperm nuclei from intact bi-fluorescent pollen, obtaining purity rates as high as 99%. Gating is a potential solution to fluorescent cell discrimination, but the subjectivity in gating may induce unwarranted biases in the follow-up statistical analysis to determine the difference among multiple treatments. Alternative methods for gating could be support vector machines (SVM) or clustering. SVM (Cortes and Vapnik, 1995; Lee et al., 2012) is a supervised machine learning technique that learns from labeled training data and creates hyper-planes that separate the artificial gametes from the diploid cells. Clustering is an unsupervised learning (Lo et al., 2008) for automated gating of flow cytometry data that can estimate the cluster means, covariance matrices, and proportions of each cell type. All these data science tools have non-zero probabilities of misclassification, that is, classifying an artificial gamete as diploid and vice versa, and these misclassification probabilities affect the power of the statistical analyses. Thus, the number of cells needs to be adjusted to account for the loss in power due to misclassifications. Figure 6 shows a chi-squared test's power curves under possible misclassifications by SVM for a range of meiosis induction percentages. The number of cells required to detect, for example, a 1% meiosis induction rate with at least 80% power is around 13,000, which is well within the limit of flow cytometry. Further, these results suggest that pooling samples after meiosis induction treatment may provide an efficient approach to testing multiple factors to reduce flow cytometry costs. Pools of interest can then be analyzed more in-depth to identify the factor responsible for artificial gamete induction.

In summary, detection of the few induced gametic cells in a large population of predominantly diploid cells must be supported by a robust statistical framework to provide confidence in factors that result in artificial gametes at low rates, considerations such as this help to define testing procedures and technical limits.

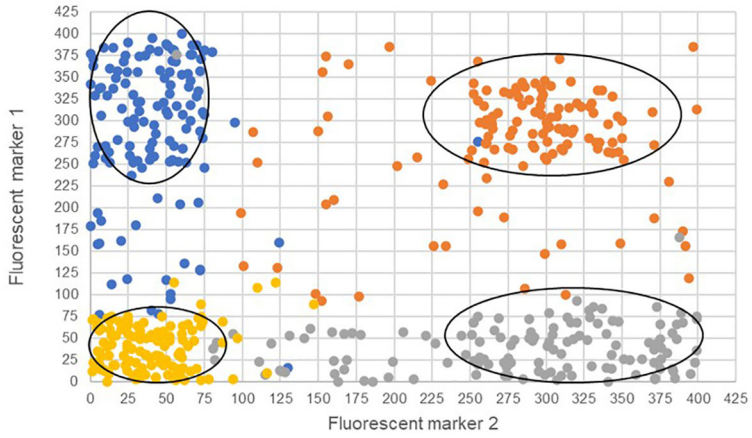


FIGURE 5
Four theoretical cell populations were produced to simulate mock flow cytometry analysis data with theoretical gating classifications (ovals) for either only fluorescent marker 1 (blue), only fluorescent marker 2 (gray), both fluorescent markers (orange), or the absence of fluorescent markers (gold). The ovals represent potential gating for individual populations. Misclassified cells are depicted as those that have fallen outside the gating ovals or those that have an incorrect fluorescence classification and are a different color than others in the same population.

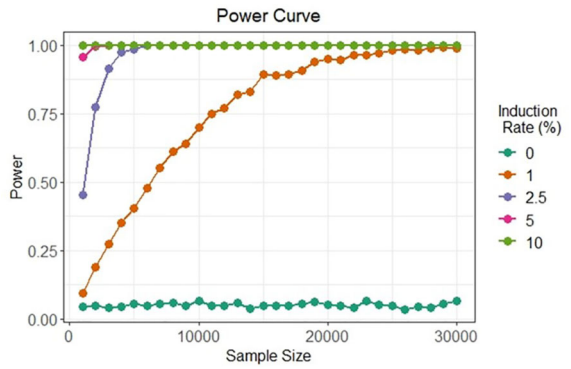


FIGURE 6
Simulated in silico power for statistical testing of fluorescent cell detection. Each curve denotes the power of the chi-squared test to detect artificial gametes at four different rates of meiosis induction upon the evaluation of different cell sample sizes under possible misclassification by an SVM classifier.

4 Discussion

The development of IVNs could greatly benefit plant breeding as a new tool to increase genetic gain. The conservation of meiotic processes in eukaryotes provides evidence of the potential to develop a universal system to induce meiosis for all plant species *in vitro* with only minimal changes to culture conditions. To make progress in IVNs, however, a cost-efficient, high-throughput detection tool must be developed for detecting artificial gametes, which is supported by a robust statistical framework. Such a tool would allow the evaluation of a large number of factors as potential inducers of meiosis. Additionally, as custom molecules for targeted biological processes such as the PROTAC system (reviewed in Békés et al., 2022) become more widely available, opportunities to target genetic factors may be tested more efficiently without needing genetic transformation.

There is much to learn from natural phenomena such as apomixis and parthenogenesis, which may provide insights into approaches that can be reversed in order to induce meiosis. Gene activation technologies paired with increased gene editing capabilities have promise in plant meiosis induction, especially in reversing the effect of a knock-out (Pan et al., 2021; Pan et al., 2022). Additionally, while environmental factors and hormone signaling show clear effects on reproduction, systematic testing of these factors will need to be well-thought out as these factors usually have global consequences on plants. By using liquid based culture systems, factors can be applied easier and in a more uniform fashion, which may further increase the scale of a meiosis induction screening system.

For detection we have proposed a protoplast system and while protoplasts can be isolated easily and in large numbers, which is amenable to the detection systems discussed in this article, protoplast regeneration can be species-dependent and recalcitrant (Hahne et al., 1989; Xu et al., 2022). There may be other technologies that provide different benefits to such a system and should also be explored. Additionally, as new cytometric and microfluidic technologies such as impedance flow cytometry (Heidmann et al., 2016) continue to improve, DNA stains and fluorescent markers may no longer be needed as cells could be detected, quantified, and sorted for downstream manipulation using label-free approaches.

5 Concluding remarks and future directions

IVN's have the potential to change cultivar development in big ways as they can increase genetic gain by decreasing breeding cycle time while also being kept in controlled laboratory conditions. In this review, we have assessed the bottleneck that we believe to be the most limiting at the current state, meiosis induction, but in order to implement and scale IVN's to efficient sizes, other bottlenecks will need to be overcome. These include the induction and detection of meiosis in crop species, artificial gamete selection, fusion, and subsequent propagation. These bottlenecks will be addressed in subsequent review articles as considerable research is needed. By fully understanding the gaps in knowledge in IVNs, solutions can be more efficiently explored and shared. Progress in single cell analyses, transformation, and sequencing technologies will continue to push IVN's from ideas to reality.

Author contributions

TC drafted and compiled manuscript. DI, SS, PK contributed to writing and editing manuscript. SA finalized figures and editing. SD and EB, contributed to writing and figure development. SH and LD contributed to draft writing. Additionally, BN and TL contributed to draft development and editing. All authors contributed to the article and approved the submitted version.

Funding

We want to thank the Iowa State University Plant Sciences Institute, RF Baker Center for Plant Breeding, and KJ Frey Chair in Agronomy for their generous support. This article is also a product of the Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa, Project No. IOW03717 (S.D.), and Project No. IOW04714(T.L.) which is supported by USDA/NIFA and State of Iowa funds. This work was also supported by the Foundation for Food & Agriculture Research under award number CA19-SS-0000000128 (T.L.). The content of this manuscript is solely the responsibility of the authors and does not necessarily represent the official views of the Foundation for Food & Agriculture Research or the U.S. Department of Agriculture.

Conflict of interest

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

Publisher's note

References

Abrahamian, P., Hammond, R. W., and Hammond, J. (2020). Plant virus-derived vectors: applications in agricultural and medical biotechnology. *Annu. Rev. Virol.* 7 (1), 513–535. doi: 10.1146/annurev-virology-010720-054958

Adan, A., Alizada, G., Kiraz, Y., Baran, Y., and Nalbant, A. (2017). Flow cytometry: basic principles and applications. *Crit. Rev. Biotechnol.* 37 (2), 163–176. doi: 10.3109/07388551.2015.1128876

Ali, Z., Shami, A., Sedeek, K., Kamel, R., Alhabsi, A., Tehseen, M., et al. (2020). Fusion of the Cas9 endonuclease and the VirD2 relaxase facilitates homology-directed repair for precise genome engineering in rice. *Commun. Biol.* 3 (1), 44. doi: 10.1038/s42003-020-0768-9

Barakate, A., Orr, J., Schreiber, M., Colas, I., Lewandowska, D., McCallum, N., et al. (2021). Barley anther and meiocyte transcriptome dynamics in meiotic prophase I. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.619404

Bargmann, B. O. R., and Birnbaum, K. D. (2010). Fluorescence activated cell sorting of plant protoplasts. *J. Visualized Experiments* 36, 1673. doi: 10.3791/1673

Bateman, A. (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 47 (D1), D506–D515. doi: 10.1093/nar/gky1049

Békés, M., Langley, D. R., and Crews, C. M. (2022). PROTAC targeted protein degraders: the past is prologue. *Nat. Rev. Drug Discovery* 21 (3), 181–200. doi: 10.1038/s41573-021-00371-6

Beygi, Z., Nezamzadeh, Z., Rabiei, M., and Mirakhorli, N. (2021). Enhanced accumulation of trigonelline by elicitation and osmotic stresses in fenugreek callus culture. *Plant Cell Tissue Organ Culture* 147 (1), 169–174. doi: 10.1007/s11240-021-02055-w

Bhat, J. A., Ali, S., Salgotra, R. K., Mir, Z. A., Dutta, S., Jadon, V., et al. (2016). Genomic selection in the era of next generation sequencing for complex traits in plant breeding. *Front. Genet.* 7. doi: 10.3389/fgene.2016.00221

Boca, S., Koestler, F., Ksas, B., Chevalier, A., Leymarie, J., Fekete, A., et al. (2014). Arabidopsis lipocalins AtCHL and AtTIL have distinct but overlapping functions essential for lipid protection and seed longevity. *Plant Cell Environ.* 37 (2), 368–381. doi: 10.1111/pce.12159

Boerman, N. A., Frei, U. K., and Lübberstedt, T. (2020). Impact of spontaneous haploid genome doubling in maize breeding. *Plants* 9 (3), 369. doi: 10.3390/plants9030369

Borges, F., Gardner, R., Lopes, T., Calarco, J. P., Boavida, L. C., Slotkin, R. K., et al. (2012). FACS-based purification of arabidopsis microspores, sperm cells and vegetative nuclei. *Plant Methods* 8 (1), 44. doi: 10.1186/1746-4811-8-44

Brown, M., Antle, J., Backlund, P., Carr, E., Easterling, W. E., Walsh, M. K., et al. (2015). *Climate change, global food security and the U.S. food system*. Available at: http://www.usda.gov/oce/climate_change/FoodSecurity2015Assessment/FullAssessment.pdf.

Brummer, C. E., Barber, W. T., Collier, S. M., Cox, T. S., Johnson, R., Murray, S. C., et al. (2011). Plant breeding for harmony between agriculture and the environment. *Front. Ecol. Environ.* 9 (10), 561–568. doi: 10.1890/100225

Catoni, M., Noris, E., Vaira, A. M., Jonesman, T., Matić, S., Soleimani, R., et al. (2018). Virus-mediated export of chromosomal DNA in plants. *Nat. Commun.* 9 (1), 5308. doi: 10.1038/s41467-018-07775-w

Chen, Y., Zhang, L., Zhou, Y., Geng, Y., and Chen, Z. (2000). Inducing somatic meiosis-like reduction at high frequency by caffeine in root-tip cells of vicia faba. *Mutat. Res. - Fundam. Mol. Mech. Mutagenesis* 452 (1), 67–72. doi: 10.1016/S0027-5107(00)00045-2

Cheng, C. Y., Mathews, D. E., Schaller, G. E., and Kieber, J. J. (2013). Cytokinin-dependent specification of the functional megaspore in the arabidopsis female gametophyte. *Plant J.* 73 (6), 929–940. doi: 10.1111/tpj.12084

Chi, W. T., Fung, R. W. M., Liu, H. C., Hsu, C. C., and Charng, Y. Y. (2009). Temperature-induced lipocalin is required for basal and acquired thermotolerance in arabidopsis. *Plant Cell Environ.* 32 (7), 917–927. doi: 10.1111/j.1365-3040.2009.01972.x

Chupeau, M. C., Granier, F., Pichon, O., Renou, J. P., Gaudin, V., and Chupeaua, Y. (2013). Characterization of the early events leading to totipotency in an arabidopsis protoplast liquid culture by temporal transcript profiling. *Plant Cell* 25 (7), 2444–2463. doi: 10.1105/tpc.113.109538

Colnaghi, M., Lane, N., and Pomiankowski, A. (2020). Genome expansion in early eukaryotes drove the transition from lateral gene transfer to meiotic sex. *ELife* 9, 1–16. doi: 10.7554/ELIFE.58873

Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.* 20 (3), 273–297. doi: 10.1007/bf00994018

Cunningham, F. J., Goh, N. S., Demirer, G. S., Matos, J. L., and Landry, M. P. (2018). Nanoparticle-mediated delivery towards advancing plant genetic engineering. *Trends Biotechnol.* 36 (9), 882–897. doi: 10.1016/j.tibtech.2018.03.009

Davey, M. R., Anthony, P., Power, J. B., and Lowe, K. C. (2005). Plant protoplasts: status and biotechnological perspectives. *Biotechnol. Adv.* 23 (2), 131–171. doi: 10.1016/j.biotechadv.2004.09.008

Dean, P. J., Siwiec, T., Waterworth, W. M., Schlögelhofer, P., Armstrong, S. J., and West, C. E. (2009). A novel ATM-dependent X-ray-inducible gene is essential for both plant meiosis and gametogenesis. *Plant J.* 58 (5), 791–802. doi: 10.1111/j.1365-313X.2009.03814.x

De La Fuente, G. N., Frei, U. K., and Lübberstedt, T. (2013). Accelerating plant breeding. *Trends Plant Sci.* 18 (12), 667–672. doi: 10.1016/j.tplants.2013.09.001

Denyer, T., Ma, X., Klesen, S., Scacchi, E., Nieselt, K., and Timmermans, M. C. P. (2019). Spatiotemporal developmental trajectories in the arabidopsis root revealed using high-throughput single-cell RNA sequencing. *Dev. Cell* 48 (6), 840–852.e5. doi: 10.1016/j.devcel.2019.02.022

Dhaka, N., Krishnan, K., Kandpal, M., Vashisht, I., Pal, M., Sharma, M. K., et al. (2020). Transcriptional trajectories of anther development provide candidates for engineering male fertility in sorghum. *Sci. Rep.* 10 (1), 897. doi: 10.1038/s41598-020-57717-0

Dickinson, A. J., Zhang, J., Luciano, M., Wachsman, G., Sandoval, E., Schnermann, M., et al. (2021). A plant lipocalin promotes retinal-mediated oscillatory lateral root initiation. *Science* 373 (6562), 1532–1536. doi: 10.1126/science.abf7461

Dukowic-Schulze, S., Harris, A., Li, J., Sundararajan, A., Mudge, J., Retzel, E. F., et al. (2014a). Comparative transcriptomics of early meiosis in arabidopsis and maize. *J. Genet. Genomics* 41 (3), 139–152. doi: 10.1016/j.jgg.2013.11.007

Dukowic-Schulze, S., Sundararajan, A., Mudge, J., Ramaraj, T., Farmer, A. D., Wang, M., et al. (2014b). The transcriptome landscape of early maize meiosis. *BMC Plant Biol.* 14 (1), 118. doi: 10.1186/1471-2229-14-118

Eeckhaut, T., Lakshmanan, P. S., Deryckere, D., Van Bockstaele, E., and Van Huylenbroeck, J. (2013). Progress in plant protoplast research. *Planta* 238 (6), 991–1003. doi: 10.1007/s00425-013-1936-7

Elmaghrabi, A. M., Rogers, H. J., Francis, D., and Ochatt, S. J. (2017). Peg induces high expression of the cell cycle checkpoint gene WEE1 in embryogenic callus of medicago truncatula: potential link between cell cycle checkpoint regulation and osmotic stress. *Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.01479

Espinosa, M.E.Á., Moreira, R. O., Lima, A. A., Ságio, S. A., Barreto, H. G., Luiz, S. L. P., et al. (2017). Early histological, hormonal, and molecular changes during pineapple (Ananas comosus (L.) Merrill) artificial flowering induction. *J. Plant Physiol.* 209, 11–19. doi: 10.1016/j.jplph.2016.11.009

Fang, J., Wei, Y., Lv, C., Peng, S., Zhao, S., and Hua, J. (2017). CD61 promotes the differentiation of canine ADMSCs into PGC-like cells through modulation of TGF-β signaling. *Sci. Rep.* 7 (1), 43851. doi: 10.1038/srep43851

FAO, IFAD, UNICEF, WFP and WHO. (2020). *The state of food security and nutrition in the world 2020*. (Rome: FAO). doi: 10.4060/ca9692en

Feng, L. X., Chen, Y., Dettin, L., Reijo Pera, R. A., Herr, J. C., Goldberg, E., et al. (2002). Generation and *in vitro* differentiation of a spermatogonial cell line. *Science* 297 (5580), 392–395. doi: 10.1126/science.1073162

Gao, H., Gadlage, M. J., Lafitte, H. R., Lenderts, B., Yang, M., Schroder, M., et al. (2020). Superior field performance of waxy corn engineered using CRISPR–Cas9. *Nat. Biotechnol.* 38 (5), 579–581. doi: 10.1038/s41587-020-0444-0

Garrido, J., Aguilar, M., and Prieto, P. (2020). Identification and validation of reference genes for RT-qPCR normalization in wheat meiosis. *Sci. Rep.* 10 (1), 2726. doi: 10.1038/s41598-020-59580-5

Geiger, H. H. (2009). Doubled Haploids. In: J. L. Bennetzen and S. Hake (eds) *Handbook of Maize*. Springer, New York, NY. doi: 10.1007/978-0-387-77863-1_32

Goodenough, U., and Heitman, J. (2014). Origins of eukaryotic sexual reproduction. *Cold Spring Harb Perspect. Biol.* 6 (3). doi: 10.1101/cshperspect.a016154

Goodman, C. D., Casati, P., and Walbot, V. (2004). A multidrug resistance-associated protein involved in anthocyanin transport in zea mays. *Plant Cell* 16 (7), 1812–1826. doi: 10.1105/tpc.022574

Griswold, M. D. (2005). Perspective on the Function of Sertoli Cells. *Sertoli Cell Biol.* 15–18. doi: 10.1016/B978-012647751-1/50003-9

Hackett, W. P. (2011). Juvenility, Maturation, and Rejuvenation in Woody Plants. In *Horticultural Reviews*, J. Janick (Ed.). doi: 10.1002/9781118060735.ch3

Hahne, B., Fleck, J., and Hahne, G. (1989). Colony formation from mesophyll protoplasts of a cereal, oat. *Proc. Natl. Acad. Sci.* 86 (16), 6157–6160. doi: 10.1073/pnas.86.16.6157

Heidmann, I., Schade-Kampmann, G., Lambalk, J., Ottiger, M., and Di Berardino, M. (2016). Impedance flow cytometry: a novel technique in pollen analysis. *PloS One* 11 (11), e0165531. doi: 10.1371/journal.pone.0165531

Higuchi, M., Pischke, M. S., Mähönen, A. P., Miyawaki, K., Hashimoto, Y., Seki, M., et al. (2004). In planta functions of the arabidopsis cytokinin receptor family. *Proc. Natl. Acad. Sci. United States America* 101 (23), 8821–8826. doi: 10.1073/pnas.0402887101

Hong, L., Tang, D., Zhu, K., Wang, K., Li, M., and Cheng, Z. (2012). Somatic and reproductive cell development in rice anther is regulated by a putative glutaredoxin. *Plant Cell* 24 (2), 577–588. doi: 10.1105/tpc.111.093740

Hongay, C. F., Grisafi, P. L., Galitski, T., and Fink, G. R. (2006). Antisense transcription controls cell fate in saccharomyces cerevisiae. *Cell* 127 (4), 735–745. doi: 10.1016/j.cell.2006.09.038

Hongay, C. F., and Orr-Weaver, T. L. (2011). Drosophila inducer of MEiosis 4 (IME4) is required for notch signaling during oogenesis. *Proc. Natl. Acad. Sci. United States America* 108 (36), 14855–14860. doi: 10.1073/pnas.1111577108

Howe, K. L., Contreras-Moreira, B., De Silva, N., Maslen, G., Akanni, W., Allen, J., et al. (2020). Ensembl genomes 2020-enabling non-vertebrate genomic research. *Nucleic Acids Res.* 48 (D1), D689–D695. doi: 10.1093/nar/gkz890

Hsu, C. Y., Liu, Y., Luthe, D. S., and Yuceer, C. (2006). Poplar FT2 shortens the juvenile phase and promotes seasonal flowering. *Plant Cell* 18 (8), 1856–1861. doi: 10.1105/tpc.106.041038

Inoue, T., Higuchi, M., Hashimoto, Y., Seki, M., Kobayashi, M., Kato, T., et al. (2001). Identification of CRE1 as a cytokinin receptor from arabidopsis. *Nature* 409 (6823), 1060–1063. doi: 10.1038/35059117

Iqbal, N., Khan, N. A., Ferrante, A., Trivellini, A., Francini, A., and Khan, M. I. R. (2017). Ethylene role in plant growth, development and senescence: interaction with other phytohormones. *Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.00475

Jähne, F., Hahn, V., Würschum, T., and Leiser, W. L. (2020). Speed breeding short-day crops by LED-controlled light schemes. *Theor. Appl. Genet.* 133 (8), 2335–2342. doi: 10.1007/s00122-020-03601-4

Jeong, Y. Y., Lee, H. Y., Kim, S. W., Noh, Y. S., and Seo, P. J. (2021). Optimization of protoplast regeneration in the model plant arabidopsis thaliana. *Plant Methods* 17 (1), 21. doi: 10.1186/s13007-021-00720-x

Ji, Y., Tu, P., Wang, K., Gao, F., Yang, W., Zhu, Y., et al. (2014). Defining reference genes for quantitative real-time PCR analysis of anther development in rice. *Acta Biochim. Biophys. Sin.* 46 (4), 305–312. doi: 10.1093/abbs/gmu002

Karunarathna, K. H. T., Mewan, K. M., Weerasena, O. V. D. S. J., Perera, S. A. C. N., and Edirisinghe, E. N. U. (2021). A functional molecular marker for detecting blister blight disease resistance in tea (Camellia sinensis l.). *Plant Cell Rep.* 40 (2), 351–359. doi: 10.1007/s00299-020-02637-6

Kelliher, T., Starr, D., Su, X., Tang, G., Chen, Z., Carter, J., et al. (2019). One-step genome editing of elite crop germplasm during haploid induction. *Nat. Biotechnol.* 37 (3), 287–292. doi: 10.1038/s41587-019-0038-x

Kelliher, T., and Walbot, V. (2012). Hypoxia triggers meiotic fate acquisition in maize. *Science* 337 (6092), 345–348. doi: 10.1126/science.1220080

Kiełkowska, A., and Adamus, A. (2012). An alginate-layer technique for culture of brassica oleracea l. protoplasts. *Vitro Cell. Dev. Biol. - Plant* 48 (2), 265–273. doi: 10.1007/s11627-012-9431-6

Lee, K., Eggenberger, A. L., Banakar, R., McCaw, M. E., Zhu, H., Main, M., et al. (2019). CRISPR/Cas9-mediated targeted T-DNA integration in rice. *Plant Mol. Biol.* 99 (4–5), 317–328. doi: 10.1007/s11103-018-00819-1

Lee, G., Stoolman, L., and Scott, C. (2012). *Transfer learning for auto-gating of flow cytometry data* Vol. 27. Eds. I. Guyon, G. Dror, V. Lemaire, G. Taylor and D. Silver (JMLR(workshop), 155–166. Available at: http://www.clopinet.com/isabelle/Projects/ICML2011/slides/lee11.pdf.

Li, J., Dukowic-Schulze, S., Lindquist, I. E., Farmer, A. D., Kelly, B., Li, T., et al. (2015). The plant-specific protein FEHLSTART controls male meiotic entry, initializing meiotic synchronization in arabidopsis. *Plant J.* 84 (4), 659–671. doi: 10.1111/tpj.13026

Li, H., Rasheed, A., Hickey, L. T., and He, Z. (2018). Fast-forwarding genetic gain. *Trends Plant Sci.* 23 (3), 184–186. doi: 10.1016/j.tplants.2018.01.007

Liu, W., Rudis, M. R., Cheplick, M. H., Millwood, R. J., Yang, J. P., Ondzighi-Assoume, C. A., et al. (2020). Lipofection-mediated genome editing using DNA-free delivery of the Cas9/gRNA ribonucleoprotein into plant cells. *Plant Cell Rep.* 39 (2), 245–257. doi: 10.1007/s00299-019-02488-w

Lo, K., Brinkman, R. R., and Gottardo, R. (2008). Automated gating of flow cytometry data via robust model-based clustering. *Cytometry Part A* 73 (4), 321–332. doi: 10.1002/cyto.a.20531

López-Malvar, A., Butron, A., Malvar, R. A., McQueen-Mason, S. J., Faas, L., Gómez, L. D., et al. (2021). Association mapping for maize stover yield and saccharification efficiency using a multiparent advanced generation intercross (MAGIC) population. *Sci. Rep.* 11 (1), 3425. doi: 10.1038/s41598-021-83107-1

Lora, J., and Hormaza, J. I. (2021). *Crosstalk between the sporophyte and the gametophyte during anther and ovule development in angiosperms BT - progress in botany*, Vol. 82. Eds. F. M. Cánovas, U. Lüttge, M.-C. Risueño and H. Pretzsch, (Cham: Springer) 113–129. doi: 10.1007/124_2020_50

Ma, X., Zhang, X., Liu, H., and Li, Z. (2020). Highly efficient DNA-free plant genome editing using virally delivered CRISPR–Cas9. *Nat. Plants* 6 (7), 773–779. doi: 10.1038/s41477-020-0704-5

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 161 (5), 1202–1214. doi: 10.1016/j.cell.2015.05.002

Mao, Y., Botella, J. R., Liu, Y., and Zhu, J. K. (2019). Gene editing in plants: progress and challenges. *Natl. Sci. Rev.* 6 (3), 421–437. doi: 10.1093/nsr/nwz005

Marx, V. (2016). Plants: a tool box of cell-based assays. *Nat. Methods* 13 (7), 551–554. doi: 10.1038/nmeth.3900

Mata, J., Lyne, R., Burns, G., and Bähler, J. (2002). The transcriptional program of meiosis and sporulation in fission yeast. *Nat. Genet.* 32 (1), 143–147. doi: 10.1038/ng951

Mathilde, G., Ghislaine, G., Daniel, V., and Georges, P. (2003). The arabidopsis MEI1 gene encodes a protein with five BRCT domains that is involved in meiosis-specific DNA repair events independent of SPO11-induced DSBs. *Plant J.* 35 (4), 465–475. doi: 10.1046/j.1365-313X.2003.01820.x

Mateo de Arias, M., Gao, L., Sherwood, D. A., Dwivedi, K. K., Price, B. J., Jamison, M., et al. (2020). Whether gametophytes are reduced or unreduced in angiosperms might be determined metabolically. *Genes* 11 (12), 1449. doi: 10.3390/genes11121449

Medrano, J. V., Martínez-Arroyo, A. M., Míguez, J. M., Moreno, I., Martínez, S., Quiñonero, A., et al. (2016). Human somatic cells subjected to genetic induction with six germ line-related factors display meiotic germ cell-like features. *Sci. Rep.* 6, 24956. doi: 10.1038/srep24956

Melchers, G., Sacristán, M. D., and Holder, A. A. (1978). Somatic hybrid plants of potato and tomato regenerated from fused protoplasts. *Carlsberg Res. Commun.* 43 (4), 203–218. doi: 10.1007/BF02906548

Mercier, R., Mézard, C., Jenczewski, E., Macaisne, N., and Grelon, M. (2015). The molecular biology of meiosis in plants. *Annu. Rev. Plant Biol.* 66 (1), 297–327. doi: 10.1146/annurev-arplant-050213-035923

Mieulet, D., Jolivet, S., Rivard, M., Cromer, L., Vernet, A., Mayonove, P., et al. (2016). Turning rice meiosis into mitosis. *Cell Res.* 26 (11), 1242–1254. doi: 10.1038/cr.2016.117

Miki, D., Zhang, W., Zeng, W., Feng, Z., and Zhu, J. K. (2018). CRISPR/Cas9-mediated gene targeting in arabidopsis using sequential transformation. *Nat. Commun.* 9 (1), 1967. doi: 10.1038/s41467-018-04416-0

Murray, S. C., Eckhoff, P., Wood, L., and Paterson, A. H. (2013). A proposal to use gamete cycling *in vitro* to improve crops and livestock. *Nat. Biotechnol.* 31 (10), 877–880. doi: 10.1038/nbt.2707

Nagata, T., and Takebe, I. (1971). Plating of isolated tobacco mesophyll protoplasts on agar medium. *Planta* 99 (1), 12–20. doi: 10.1007/BF00392116

Nandy, S., Zhao, S., Pathak, B. P., Manoharan, M., and Srivastava, V. (2015). Gene stacking in plant cell using recombinases for gene integration and nucleases for marker gene deletion. *BMC Biotechnol.* 15 (1), 93. doi: 10.1186/s12896-015-0212-2

Nelms, B., and Walbot, V. (2019). Defining the developmental program leading to meiosis in maize. *Science* 364 (6435), 52–56. doi: 10.1126/science.aav6428

Nonomura, K. I., Eiguchi, M., Nakano, M., Takashima, K., Komeda, N., Fukuchi, S., et al. (2011). A novel RNA-recognition-motif protein is required for premeiotic G1/s-phase transition in rice (Oryza sativa l.). *PloS Genet.* 7 (1), e1001265. doi: 10.1371/journal.pgen.1001265

Olmedo-Monfil, V., Durán-Figueroa, N., Arteaga-Vázquez, M., Demesa-Arévalo, E., Autran, D., Grimanelli, D., et al. (2010). Control of female gamete formation by a small RNA pathway in arabidopsis. *Nature* 464 (7288), 628–632. doi: 10.1038/nature08828

Ortiz-Ramírez, C., Arevalo, E. D., Xu, X., Jackson, D. P., and Birnbaum, K. D. (2018). An efficient cell sorting protocol for maize protoplasts. *Curr. Protoc. Plant Biol.* 3 (3), e20072. doi: 10.1002/cppb.20072

Pan, C., Li, G., Malzahn, A. A., Cheng, Y., Leyson, B., Sretenovic, S., et al. (2022). Boosting plant genome editing with a versatile CRISPR-combo system. *Nat. Plants* 8 (5), 513–525. doi: 10.1038/s41477-022-01151-9

Pan, C., Wu, X., Markel, K., Malzahn, A. A., Kundagrami, N., Sretenovic, S., et al. (2021). CRISPR–Act3.0 for highly efficient multiplexed gene activation in plants. *Nat. Plants* 7 (7), 942–953. doi: 10.1038/s41477-021-00953-7

Pawlowski, W. P., Wang, C. J. R., Golubovskaya, I. N., Szymaniak, J. M., Shi, L., Hamant, O., et al. (2009). Maize AMEIOTIC1 is essential for multiple early meiotic processes and likely required for the initiation of meiosis. *Proc. Natl. Acad. Sci. United States America* 106 (9), 3603–3608. doi: 10.1073/pnas.0810115106

Persad, R., Reuter, D. N., Dice, L. T., Nguyen, M. A., Rigoulot, S. B., Layton, J. S., et al. (2020). The q-system as a synthetic transcriptional regulator in plants. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.00245

Petersson, S. V., Lindén, P., Moritz, T., and Ljung, K. (2015). Cell-type specific metabolic profiling of arabidopsis thaliana protoplasts as a tool for plant systems biology. *Metabolomics* 11 (6), 1679–1689. doi: 10.1007/s11306-015-0814-7

Popovic, M., Azpiroz, F., and Chuva de Sousa Lopes, S. M. (2021). Engineered models of the human embryo. *Nat. Biotechnol.* 39 (8), 918–920. doi: 10.1038/s41587-021-01004-4

Prusicki, M. A., Keizer, E. M., Van Rosmalen, R. P., Komaki, S., Seifert, F., Müller, K., et al. (2019). Live cell imaging of meiosis in arabidopsis thaliana. *ELife* 8, e42834. doi: 10.7554/eLife.42834

Reed, K. M., and Bargmann, B. O. R. (2021). Protoplast regeneration and its use in new plant breeding technologies. *Front. Genome Editing* 3. doi: 10.3389/fgeed.2021.734951

Ren, L., Tang, D., Zhao, T., Zhang, F., Liu, C., Xue, Z., et al. (2018). OsSPL regulates meiotic fate acquisition in rice. *New Phytol.* 218 (2), 789–803. doi: 10.1111/nph.15017

Riou, L., Bastos, H., Lassalle, B., Coureuil, M., Testart, J., Boussin, F. D., et al. (2005). The telomerase activity of adult mouse testis resides in the spermatogonial α6-integrin-positive side population enriched in germinal stem cells. *Endocrinology* 146 (9), 3926–3932. doi: 10.1210/en.2005-0502

Ronchi, V. N., Giorgetti, L., Tonelli, M., and Martini, G. (1992). Ploidy reduction and genome segregation in cultured carrot cell lines. II. somatic meiosis. *Plant Cell Tissue Organ Culture* 30 (2), 115–119. doi: 10.1007/BF00034304

Sahab, S., Hayden, M. J., Mason, J., and Spangenberg, G. (2019). Mesophyll protoplasts and PEG-mediated transfections: Transient assays and generation of stable transgenic canola plants. In S. Kumar, P. Barone and M. Smith (Eds.). *Methods Molecular Biol.* 1864, 131–152. Springer New York. doi: 10.1007/978-1-4939-8778-8_10

Santos, T., Shuler, J., Guimarães, R., and Farah, A. (2019). CHAPTER 1. Introduction to Coffee Plant and Genetics: Production, Quality and Chemistry. (1–25). doi: 10.1039/9781782622437-00001

Sheen, J. (2001). Signal transduction in maize and arabidopsis mesophyll protoplasts. *Plant Physiol.* 127 (4), 1466–1475. doi: 10.1104/pp.010820

Shillito, R. D., Carswell, G. K., Johnson, C. M., Dimaio, J. J., and Harms, C. T. (1989). Regeneration of fertile plants from protoplasts of elite inbred maize. *Bio/Technology* 7 (6), 581–587. doi: 10.1038/nbt0689-581

Shin, D., Lee, W., Lee, J. H., and Bang, D. (2019). Multiplexed single-cell RNA-seq via transient barcoding for simultaneous expression profiling of various drug perturbations. *Sci. Adv.* 5 (5), eaav2249. doi: 10.1126/sciadv.aav2249

Shulse, C. N., Cole, B. J., Ciobanu, D., Lin, J., Yoshinaga, Y., Gouran, M., et al. (2019). High-throughput single-cell transcriptome profiling of plant cell types. *Cell Rep.* 27 (7), 2241–2247.e4. doi: 10.1016/j.celrep.2019.04.054

Singh, M., Goel, S., Meeley, R. B., Dantec, C., Parrinello, H., Michaud, C., et al. (2011). Production of viable gametes without meiosis in maize deficient for an ARGONAUTE protein. *Plant Cell* 23 (2), 443–458. doi: 10.1105/tpc.110.079020

Stein, R. E., Nauerth, B. H., Binmöller, L., Zühl, L., Loreth, A., Reinert, M., et al. (2021). RH17 restricts reproductive fate and represses autonomous seed coat development in sexual arabidopsis. *Dev. (Cambridge)* 148 (19), dev198739. doi: 10.1242/dev.198739

Stevens, R., Grelon, M., Vezon, D., Oh, J., Meyer, P., Perennes, C., et al. (2004). A CDC45 homolog in arabidopsis is essential for meiosis, as shown by RNA interference-induced gene silencing. *Plant Cell* 16 (1), 99–113. doi: 10.1105/tpc.016865

Sun, C., Lei, Y., Li, B., Gao, Q., Li, Y., Cao, W., et al. (2023). Precise integration of large DNA sequences in plant genomes using PrimeRoot editors. *Nat. Biotechnol.* doi: 10.1038/s41587-023-01769-w

Sun, Y., Wang, X., Pan, L., Xie, F., Dai, B., Sun, M., et al. (2021). Plant egg cell fate determination depends on its exact position in female gametophyte. *Proc. Natl. Acad. Sci. United States America* 118 (8), e2017488118. doi: 10.1073/pnas.2017488118

Tibbs Cortes, L., Zhang, Z., and Yu, J. (2021). Status and prospects of genome-wide association studies in plants. *Plant Genome* 14 (1), e20077. doi: 10.1002/tpg2.20077

US Embassy and Consulate in Italy (2022). "How climate change affects the food crisis," in *US Embassies and consulates in Italy*. Available at: https://it.usembassy.gov/how-climate-change-affects-the-food-crisis/.

Vernet, N., Condrea, D., Mayere, C., Féret, B., Klopfenstein, M., Magnant, W., et al. (2020). Meiosis occurs normally in the fetal ovary of mice lacking all retinoic acid receptors. *Sci. Adv.* 6 (21), eaaz1139. doi: 10.1126/sciadv.aaz1139

Waki, T., Miyashima, S., Nakanishi, M., Ikeda, Y., Hashimoto, T., and Nakajima, K. (2013). A GAL4-based targeted activation tagging system in arabidopsis thaliana. *Plant J.* 73 (3), 357–367. doi: 10.1111/tpj.12049

Wallberg, F., Tenev, T., and Meier, P. (2016). Analysis of Apoptosis and Necroptosis by Fluorescence-Activated Cell Sorting. *Cold Spring Harbor Protocols* 2016 (4), pdb.prot087387. doi: 10.1101/pdb.prot087387

Wan, Y., Petolino, J. F., and Widholm, J. M. (1989). Efficient production of doubled haploid plants through colchicine treatment of anther-derived maize callus. *Theor. And Appl. Genet.* 77 (6), 889–892. doi: 10.1007/BF00268344

Wang, N., Ryan, L., Sardesai, N., Wu, E., Lenderts, B., Lowe, K., et al. (2023). Leaf transformation for efficient random integration and targeted genome modification in maize and sorghum. *Nat. Plants* 9 (2), 255–270. doi: 10.1038/s41477-022-01338-0

Wang, Y., Van Rengs, W. M. J., Zaidan, M. W. A. M., and Underwood, C. J. (2021). Meiosis in crops: from genes to genomes. *J. Exp. Bot.* 72 (18), 6091–6109. doi: 10.1093/jxb/erab217

Wang, N., Wang, H., Zhang, A., Liu, Y., Yu, D., Hao, Z., et al. (2020). Genomic prediction across years in a maize doubled haploid breeding program to accelerate early-stage testcross testing. *Theor. Appl. Genet.* 133 (10), 2869–2879. doi: 10.1007/s00122-020-03638-5

Watson, A., Ghosh, S., Williams, M. J., Cuddy, W. S., Simmonds, J., Rey, M. D., et al. (2018). Speed breeding is a powerful tool to accelerate crop research and breeding. *Nat. Plants* 4 (1), 23–29. doi: 10.1038/s41477-017-0083-8

Wei, B., Zhang, J., Pang, C., Yu, H., Guo, D., Jiang, H., et al. (2015). The molecular mechanism of SPOROCYTELESS/NOZZLE in controlling arabidopsis ovule development. *Cell Res.* 25 (1), 121–134. doi: 10.1038/cr.2014.145

Wen, L., and Tang, F. (2019). Human germline cell development: from the perspective of single-cell sequencing. *Mol. Cell* 76 (2), 320–328. doi: 10.1016/j.molcel.2019.08.025

Woo, J. W., Kim, J., Kwon, S., Corvalán, C., Cho, S. W., Kim, H., et al. (2015). DNA-Free genome editing in plants with preassembled CRISPR-Cas9 ribonucleoproteins. *Nat. Biotechnol.* 33 (11), 1162–1164. doi: 10.1038/nbt.3389

Xing, S., and Zachgo, S. (2008). ROXY1 and ROXY2, two arabidopsis glutaredoxin genes, are required for anther development. *Plant J.* 53 (5), 790–801. doi: 10.1111/j.1365-313X.2007.03375.x

Xu, Z. H., Davey, M. R., and Cocking, E. C. (1982). Plant regeneration from root protoplasts of brassica. *Plant Sci. Lett.* 24 (1), 117–121. doi: 10.1016/0304-4211(82)90016-5

Xu, Y., Li, R., Luo, H., Wang, Z., Li, M. W., Lam, H. M., et al. (2022). Protoplasts: small cells with big roles in plant biology. *Trends Plant Sci.* 27 (8), 828–829. doi: 10.1016/j.tplants.2022.03.010

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

Yang, C., Hamamura, Y., Sofroni, K., Böwer, F., Stolze, S. C., Nakagami, H., et al. (2019). SWITCH 1/DYAD is a WINGS APART-LIKE antagonist that maintains sister chromatid cohesion in meiosis. *Nat. Commun.* 10 (1), 1755. doi: 10.1038/s41467-019-09759-w

Yi, J., Kim, S. R., Lee, D. Y., Moon, S., Lee, Y. S., Jung, K. H., et al. (2012). The rice gene DEFECTIVE TAPETUM and MEIOCYTES 1 (DTM1) is required for early tapetum development and meiosis. *Plant J.* 70 (2), 256–270. doi: 10.1111/j.1365-313X.2011.04864.x

Yihua, C., Lihua, Z., Yuxuan, G., and Zhenghua, C. (2001). Meiosis-like reduction during somatic embryogenesis of arabidopsis thaliana. *In Vitro Cell. Dev. Biol. Plant* 37 (5), 654–657. Available at: https://www.jstor.org/stable/4293529.

Yoo, S. D., Cho, Y. H., and Sheen, J. (2007). Arabidopsis mesophyll protoplasts: a versatile cell system for transient gene expression analysis. *Nat. Protoc.* 2 (7), 1565–1572. doi: 10.1038/nprot.2007.199

Yoshida, H., and Yamaguchi, H. (1973). Arrangement and association of somatic chromosomes induced by chloramphenicol in barley. *Chromosoma* 43 (4), 399–407. doi: 10.1007/BF00406746

Zalesny, R. S., Zhu, J. Y., Headlee, W. L., Gleisner, R., Pilipović, A., Van Acker, J., et al. (2020). Ecosystem services, physiology, and biofuels recalcitrance of poplars grown for landfill phytoremediation. *Plants* 9 (10), 1–26. doi: 10.3390/plants9101357

Zhang, Y., Iaffaldano, B., and Qi, Y. (2021). CRISPR ribonucleoprotein-mediated genetic engineering in plants. *Plant Commun.* 2 (2), 100168. doi: 10.1016/j.xplc.2021.100168

Zhang, Y., Massel, K., Godwin, I. D., and Gao, C. (2018). Applications and potential of genome editing in crop improvement. *Genome Biol.* 19 (1), 210. doi: 10.1186/s13059-018-1586-y

Zhang, Y., Su, J., Duan, S., Ao, Y., Dai, J., Liu, J., et al. (2011). A highly efficient rice green tissue protoplast system for transient gene expression and studying light/chloroplast-related processes. *Plant Methods* 7 (1), 30. doi: 10.1186/1746-4811-7-30

Zhao, X., Bramsiepe, J., Van Durme, M., Komaki, S., Prusicki, M. A., Maruyama, D., et al. (2017). RETINOBLASTOMA RELATED1 mediates germline entry in arabidopsis. *Science* 356 (6336), eaaf6532. doi: 10.1126/science.aaf6532

Zheng, Y., Wang, D., Ye, S., Chen, W., Li, G., Xu, Z., et al. (2021). Auxin guides germ-cell specification in arabidopsis anthers. *Proc. Natl. Acad. Sci. United States America* 118 (22), e2101492118. doi: 10.1073/pnas.2101492118

Zhong, S., Li, H., Bodi, Z., Button, J., Vespa, L., Herzog, M., et al. (2008). MTA is an arabidopsis messenger RNA adenosine methylase and interacts with a homolog of a sex-specific splicing factor. *Plant Cell* 20 (5), 1278–1288. doi: 10.1105/tpc.108.058883

Zhou, L. Z., Juranić, M., and Dresselhaus, T. (2017). Germline development and fertilization mechanisms in maize. *Mol. Plant* 10 (3), 389–401. doi: 10.1016/j.molp.2017.01.012

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 17 July 2023
DOI 10.3389/fpls.2023.1188149

OPEN ACCESS

EDITED BY
Umesh K. Reddy,
West Virginia State University, United States

REVIEWED BY
Mehboob-ur Rahman,
National Institute for Biotechnology and Genetic Engineering, Pakistan
Zhiqiang Wu,
Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, China
Wangsuo Liu,
Ningxia University, China

*CORRESPONDENCE
Kwang-Soo Cho
kscholvoe@korea.kr

†These authors have contributed equally to this work

RECEIVED 17 March 2023
ACCEPTED 26 June 2023
PUBLISHED 17 July 2023

CITATION
Choe M-E, Kim J-Y, Syed Nabi RB, Han S-I and Cho K-S (2023) Development of InDels markers for the identification of cytoplasmic male sterility in *Sorghum* by complete chloroplast genome sequences analysis.
Front. Plant Sci. 14:1188149.
doi: 10.3389/fpls.2023.1188149

Development of InDels markers for the identification of cytoplasmic male sterility in *Sorghum* by complete chloroplast genome sequences analysis

Myeong-Eun Choe†, Ji-Young Kim†, Rizwana Begum Syed Nabi, Sang-Ik Han and Kwang-Soo Cho*

Department of Southern Area Crop Science, National Institute of Crop Science, Rural Development Administration, Miryang, Republic of Korea

Cytoplasmic male sterility (CMS) is predominantly used for F1 hybrid breeding and seed production in *Sorghum*. DNA markers to distinguish between normal fertile (CMS-N) and sterile (CMS-S) male cytoplasm can facilitate F1 hybrid cultivar development in *Sorghum* breeding programs. In this study, the complete chloroplast (cp) genome sequences of CMS-S and Korean *Sorghum* cultivars were obtained using next-generation sequencing. The *de novo* assembled genome size of ATx623, the CMS-S line of the chloroplast, was 140,644bp. When compared to the CMS–S and CMS-N cp genomes, 19 single nucleotide polymorphisms (SNPs) and 142 insertions and deletions (InDels) were identified, which can be used for marker development for breeding, population genetics, and evolution studies. Two InDel markers with sizes greater than 20 bp were developed to distinguish cytotypes based on the copy number variation of lengths as 28 and 22 bp tandem repeats, respectively. Using the newly developed InDel markers with five pairs of CMS-S and their near isogenic maintainer line, we were able to easily identify their respective cytotypes. The InDel markers were further examined and applied to 1,104 plants from six Korean *Sorghum* cultivars to identify variant cytotypes. Additionally, the phylogenetic analysis of seven *Sorghum* species with complete cp genome sequences, including wild species, indicated that CMS-S and CMS-N contained *Milo* and *Kafir* cytotypes that might be hybridized from *S. propinquum* and *S. sudanese*, respectively. This study can facilitate F1 hybrid cultivar development by providing breeders with reliable tools for marker-assisted selection to breed desirable *Sorghum* varieties.

KEYWORDS

***Sorghum bicolor*, chloroplast genome, CMS, phylogenetic tree, InDel**

1 Introduction

Sorghum (*Sorghum bicolor* [L.] Moench) is the fifth most important major cereal cultivated worldwide and it is used not only for human nourishment, but also for animal fodder and feed, construction material, fencing, and brooms (Doggett, 1988; Mundia et al., 2019). *Sorghum* is a diploid C4 plant with outstanding tolerance to most types of abiotic stress (Tari et al., 2013). *Sorghum's* genome is significantly smaller genome than maize (around 800 vs. 2,500 Mb), and it has recently undergone high-quality diploid genome sequencing, making it an emerging model for highly productive C4 crops (Paterson et al., 2009; Mccormick et al., 2018). Heterosis, or hybrid vigor, is the ability of hybrids to outperform elite inbred line parents and is probably the most important strategy to increase grain yield in various crops, including *Sorghum* (Kim and Zhang, 2018). In several vegetable and cereal crops, commercial seed production is based on F1 hybrids produced using the cytoplasmic genetic male sterility (CMS) system. The CMS system relies on a set of male sterility-inducing cytoplasm that are complemented by alleles at genetic loci in the nuclear genome that either restore fertility or maintain sterility. In 1952, a milo CMS cytoplasm was identified in the offspring of a hybrid between two cultivars, *milo* and *kafir*, wherein *milo* as the female and *kafir* as the male. The A1 CMS line sourced from *milo* is the predominant CMS line used to produce hybrids in *Sorghum*. The presence of restorer genes enabling the production of fertile F1 hybrids using the CMS approach is essential for the cost-effective production of hybrid *Sorghum* seeds. To date, a total of nine distinct resources of CMS, namely, A1 (*milo*), A2, A3, A4, Indian A4 (A4M, A4VZM, and A4G), A5, A6, 9E, and KS cytoplasm, have been identified in *Sorghum* (Schertz, 1977; Li and Li, 1998). Although most grain *Sorghum* types currently utilized in agricultural production are hybrids in the world, hybrid cultivars are still not widely used in Korea, which explains the low average yields.

The chloroplast is the primary site of photosynthesis and carbon fixation in plants, is an essential organelle for land plants (Daniell et al., 2016), and is inherited maternally (Asaf et al., 2017). Some cp genome sequences have been used to distinguish between species and conduct phylogenetic studies, as the chloroplast (cp) genome is more conserved and shorter in length than the nuclear and mitochondrial (mt) genomes. The cp and mt genomes are often used to study plant evolution (Cho et al., 2015; Lu et al., 2016; Hong et al., 2017; Hong et al., 2019). Although the cp genome's specific structure and sequences remain conserved, the larger mt genome (200–2400 kb) has a significantly different structure and various isoforms, even within a single plant cell (Sugiyama et al., 2005; Cho et al., 2022). Therefore, polymorphic DNA originating from the cp genome is favorable for developing CMS markers, even though several CMS genes are mitochondrial as they are maternally inherited (Cho et al., 2006).

The mt and cp genomes of the normal fertile (CMS-N) of *S. bicolor*, BTx623, have previously been sequenced and are under the GenBank accession numbers NC_008360 and NC_008602, respectively. Recently, ergonomically important genes, such as those of the flowering time (Casto et al., 2019), dwarf (Hilley et al., 2017) and brown midrib (Bout and Vermerris, 2003) have been isolated using genetic mapping and comparative genomic studies between *Sorghum* and other crops. The mt genome rearrangement between CMS-S and CMS-N in *Sorghum* was analyzed, and it was discovered that the coding region of the *coxI* gene in CMS9E was found to be extended at the 3'- end by 303 nucleotides, resulting in an extension of 101 amino acids at the C-terminal of the protein. A novel chloroplast DNA deletion has been reported in most CMS lines of *Sorghum* and this deletion occurred in the middle of the gene *rpoC2*, coding for the *ß*"-subunit of RNA polymerase (Bailey-Serres et al., 1986; Chen et al., 1993). Chen et al. (1995) also reported that *rpoB*, *rbcL*, and *rpoC2* transcripts are low in inflorescence tissues and pollen of CMS. Molecular characterization of the cytoplasm using mitochondrial DNA probes revealed sufficient diversity to broaden the cytoplasmic base of *Sorghum* hybrids (Xu et al., 1995; Sivaramakrishnan et al., 1997). A previous study demonstrated the strict maternal inheritance of mt and cp DNA in the *Sorghum* cytoplasm (Pring et al., 1982). All genes encode proteins with a mitochondrial transit peptide and numerous penta-tatricopeptide repeats (Kante et al., 2018; Praveen et al., 2018). The cytoplasmic male sterile line (S *rfrf*) and its near-isogenic maintainer line (S or N *RfRf*) are essential for breeding F1 hybrids using CMS systems. The test cross is the most popular traditional method to identify the cytoplasmic type in *Sorghum*. DNA markers have been used for the indirect selection of major cultivation traits that distinguish the fertile and sterile individuals in several crops, such as onion, maize, wheat, cotton, and others (Bosacchi et al., 2015; Melonek et al., 2021).

This study aimed to: (i) Obtain complete chloroplast (cp) genome sequences of CMS-S and Korean *Sorghum* cultivars using next-generation sequencing, (ii) Identify single nucleotide polymorphisms (SNPs) and insertions and deletions (InDels) in the cp genomes that can serve as DNA markers for breeding, and phylogenetic studies, (iii) Develop InDel markers, including tandem repeats, to accurately distinguish between cytotypes (CMS-S and CMS-N) based on copy number variation, and validate their effectiveness in identifying cytotypes in Korean *Sorghum* cultivars.

2 Materials and methods

2.1 Plant materials and genome information

One male sterile line (ATx623) and four Korean cultivars of *S. bicolor* were used for the complete cp genome sequencing (Table 1). Five pairs of *S. bicolor* near-isogenic lines (male sterile and maintainer lines) and 1,104 individual plants from six Korean cultivars were used to identify cytoplasmic types with insertion and deletion (InDel) markers (Tables 2, 3). To conduct comparative genome analysis, the cp genome sequence information in *Sorghum* species was retrieved from the National Center for Biotechnology Information (Table 1). All plants were grown at the Department of Southern Area Crop Institute in Miryang, Korea.

TABLE 1 List of *Sorghum* species and GenBank accession numbers of complete chloroplast genome sequences.

Species	Cultivars	GenBank Accessions	Genome size (bp)	Remark
Sorghum bicolor	BTx623	EF115542	140,754	Saski et al. (2007),
	ATx623	MT459453	140,644	Complete chloroplast genome sequence was *de novo* assembled in this study
	Nampoongchal[z]	MT333847	140,753	
	Donganme[z]	MT333845	140,644	
	Sodamchal[z]	MT333848	140,644	
	Hwanggeumchal[z]	MT333846	140,753	
S. sudanense		MH926028	140,755	Song et al. (2019)
S. propinquum		MH926027	140,642	Song et al. (2019)
S. timorense		KF998272	140,629	Song et al. (2019)
S. halepense		LS398105	140,810	GenBank(NCBI)
S. arundinaceum		LS398103	140,821	GenBank(NCBI)
Hemisorghum mekongense		KY596136	140,765	Arthan et al. (2017)

[z]Korean Sorghum cultivars developed by line selection.

2.2 Extraction of DNA, sequencing and chloroplast genome assembly

DNA was extracted from approximately 100 mg of fresh leaf samples using the NucleoSpin Plant II Mini Kit (Macherey-Nagel, Germany), following the manufacturer's instructions. The quality and quantity of the genomic DNAs were examined using agarose gel electrophoresis and a NanoDrop 8000 spectrophotometer (Thermo Fisher, USA). Total DNA was sequenced using an Illumina HiSeq 2000 (Illumina, San Diego, USA), and raw reads ranged from 1.9 to 2.7 Gb (Supplementary Table 1). The cp genome sequences were determined from the *de novo* assembly of low-coverage whole-genome sequences according to previous reports (Hong et al., 2017; Hong et al., 2019). In particular, trimmed paired-end reads (Phred score > 20) were assembled using CLC Assembly Cell Packages (ver. 4.2.1, https://www.qiagenbioinformatics.com/products/clc-assembly-cell/) using default parameters. The cp genome sequence contigs were selected from the initial assembly through the Basic Local Alignment Search Tool using the *S. bicolor* cp genome sequence as a reference (GenBank accession number: EF115542). Gaps and ambiguous sequences were manually adjusted using Sanger sequencing. PCR amplification and Sanger sequencing were performed to verify the four junction regions between the inverted repeats (IRs) and large single copy (LSC)/small single copy (SSC). The cp genome annotation was conducted using GeSeq (Tillich et al., 2017) with the reference sequences of *S. bicolor* from

TABLE 2 Identification of cytoplasmic male sterile factors in *Sorghum bicolor* using chloroplast specific insertion and deletion (InDel) markers.

No.	Lines	Genotypes	Cytoplasmic male sterile factor	Amplicon sizes of InDel markers (bp)	
				cp_01	cp_02
1	ATx630	Male sterile line (S*rfrf*)	S	270	265
2	BTx630	Maintainer line(N*rfrf*)	N	242	243
3	ATx631	Male sterile line (S*rfrf*)	S	270	265
4	BTx631	Maintainer line(N*rfrf*)	N	242	243
5	ATx2928	Male sterile line (S*rfrf*)	S	270	265
6	BTx2928	Maintainer line(N*rfrf*)	N	242	243
7	A03017	Male sterile line (S*rfrf*)	S	270	265
8	B03017	Maintainer line(N*rfrf*)	N	242	243
9	A.arg-1	Male sterile line (S*rfrf*)	S	270	265
10	B.arg-1	Maintainer line(N*rfrf*)	N	242	243

TABLE 3 Application of cytoplasmic male sterile factors identification using the chloroplast genome specific markers (InDel cp_01) in the Korean *Sorghum bicolor* cultivars.

Cultivars	No. of individual plants			Remark		
	Total	S cytoplasm	N cytoplasm	Released Year	Bred by	Breeding methods
Nampoongchal	195	0	195	2015	RDA[z]	Pure line selection
Donganme	170	170	0	2015	RDA	Pure line selection
Sodamchal	236	236	0	2016	RDA	Pedigree
Hwanggeumchal	313	0	313	–	GARES[y]	Landraces
Bareme	95	95	0	2022	RDA	Pedigree
Noeulchal	95	0	95	2022	RDA	Pedigree
Total	1,104	501	603			

[z]Rural Development Administration.
[y]Gangwon-do Agricultural Research and Extension Services.

GenBank. The cp genome map was illustrated using the OGDraw software (Lohse et al., 2007).

2.3 Development of CMS specific markers and PCR amplification

Single nucleotide polymorphisms (SNPs) and InDels between the male sterile and maintainer lines were precisely identified using a variant calling process through MAFFT (Katoh and Standley, 2013). To amplify the InDel regions, 20 ng of genomic DNA was used in 20 μL PCR mixture comprising 2× TOP simple preMix-nTaq master mix (Enzynomics, Seoul, Korea) consisting of 0.2 U/μL of n-taq DNA polymerase, 3 mM of Mg^{2+}, and 0.4 mM of each deoxynucleotide triphosphate mixture with 10 pmol of each primer. The primer sequences used are listed in Table 4. PCR was conducted in a thermocycler (Veriti, Applied Biosystems, CA, USA) using the following cycling parameters: 95°C (5 min); 35 cycles at 95°C (20 s), 55°C (20 s), and 72°C (1 min); and the final extension was conducted at 72°C (5 min). The PCR products were analyzed by capillary electrophoresis (QIAxcel Advanced System, Qiagen, Germany) following the manufacturer's protocol. PCR products were purified with the Wizard SV Gel and PCR Clean-Up System (Promega, Madison, USA) and sequenced by direct sequencing in Bioneer Co. (Bioneer, Daejeon, South Korea). Sequences were aligned using ClustalW in MEGA 11.

2.4 Genetic distance and phylogenetic analyses

To investigate the phylogenetic position of *Sorghum* depending on the cytotype, we used eight complete cp genome sequences. Six complete cp genome sequences of the *Sorghum* species were retrieved from GenBank (Table 1). Phylogenetic analysis was conducted using the maximum composite likelihood model with 1,000 bootstrap replicates in MEGA 11 (Tamura et al., 2021). A phylogenetic tree was constructed using the neighbor-joining method (Tamura et al., 2004) with MEGA 11.

3 Results

3.1 Chloroplast genome assembly and characterization

We sequenced and assembled the complete cp genomes of one isogenic line (CMS-S, ATx623) and four Korean *Sorghum* varieties using the Illumina HiSeq 2000 system. The complete sequences of the five cp genomes were generated using *de novo* and reference-based assemblies. Sequencing with approximately 1,657–20,373X coverage generated 130.28 Gbp of paired-end reads (Supplemental Table 1). The complete size of CMS-N is 140,754 bp, as reported by Saski (Saski et al., 2007). We found that the complete cp genome sizes of *S. bicolor* ATx23 and BTx623 were 140,644 and 140,754 bp,

TABLE 4 The information of primers used in this study for the identification of cytoplasmic male sterile factors in *Sorghum bicolor* using the chloroplast genome sequences between the male sterile (ATx623) and maintainer (BTx623) lines.

Primers	Marker type	Sequence (5′-3′)	Length (bp)	Tm	Amplicon size (bp)		Location
					ATx623	BTx623	
cp_01	InDel	AGAGACCCCGTTTACCCCTA	20	59.8	270	242	*rpoC2 – rps2*
		TTGTTCCGATGGAACCTTCT	20	59.5			
cp_02	InDel	CGTGTTTGAAATTTTGGGTCT	21	58.9	265	243	*cemA - petA*
		CGAGTCTGTTGTCATTCTACTGC	23	59.1			

10.3389/fpls.2023.1188149

respectively and included a pair of IRs of 22,259 bp (CMS-N) and 22,782 bp (CMS-S) separated by SSC regions of 12,503 bp (CMS-N) and 12,506 bp (CMS-S) and LSC regions of 82,685 bp (CMS-N) and 82,574 bp (CMS-S) (Figure 1), respectively. The comparison of cp genomes of two inbred lines (ATx623 and BTx623) and four South Korean *Sorghum* cultivars (Nampoongchal, Donganme, Sodamchal, and Hwanggeumchal) showed no significant differences in their gene and gene order (Figure 1 and Supplementary Table 2). All six cp genome structures had a typical quadratic structure.

A total of 103 genes were identified in the *Sorghum* cp genome, including 40 photosynthesis-related genes, 29 transfer RNA (tRNA) genes, and 4 ribosomal RNA (rRNA) genes. Sixteen genes contained one, two, or three introns, and six of these were tRNAs (Supplemental Table 2). Notably, six protein-coding genes (*rps12, rps15, rps19, rps7, rpl2*, and *rpl23*), eight tRNA genes (*trnA-UGC, trnH-GUG, trnI-CAU, trnI-GAU, trnL-CAA, trnN-GUU, trnR-ACG*, and *trnV-GAC*), and all rRNA genes were duplicated in the IR regions, which is common in most *Poaceae* genomes. The *Sorghum* cp genome contained 16 intron-containing genes. Among them, ten protein-coding genes (*petB*, *petD*, *atpF*, *ndhB*, *ndhA*, *rpoC1*, *rps12*, *rps16*, *rpl16*, and *rpl2*) and six tRNA genes (*trnA-UGC, trnG-UCC, trnI-GAU, trnK-UUU, trnL-UAA*, and *trnV-UAC*) had a single intron, and two genes (*rps12*, and *ycf3*) contained two introns (Supplemental Table 2).

Comparing the cp genome sequences of four *Sorghum* cultivars revealed that Nampungchal and Hwanggeumchal were near, whereas Donganme and Sodamchal were identical. This might be attributed to cultivar development from the same cytoplasmic background genetic resources. Phenotypically, these cultivars showed diverse traits, such as seed or grain color, lodging tolerance, culm length, plant height, and waxy endosperm. Thus, these cultivars could be important materials for further genetic research and new cultivar development.

FIGURE 1
The complete chloroplast genome map of *Sorghum bicolor* in the male sterile (ATx623) and maintainer line (BTx623). The genes inside and outside the circle are transcribed in the clockwise and counterclockwise directions, respectively. Genes belonging to different functional groups are represented using different colors. Thick lines indicate the extent of the inverted repeats (IRa and IRb) that separate the genomes into small single-copy (SSC) and large single-copy (LSC) regions. The red-filled red triangles indicate the region of insertion and deletion (InDel) markers.

3.2 Development and validation of CMS specific markers

Although the content and genes order in the cp genomes of the two *Sorghum* inbred lines and four South Korean *Sorghum* cultivars were very similar, numerous polymorphic sites were found among them. In the cytoplasm of cp genomes, 19 single nucleotide polymorphisms (SNPs) and 142 InDels in the genic region that can be used for marker development for breeding, population genetics, and evolution studies (Supplementary Table 3). When complete cp genomes were aligned to identify polymorphisms that may distinguish between S- and N-cytoplasms, wherein the difference is more than 20bp, we found that differences of 28 and 22 bp in the intergenic regions of *rpoC2-rps2* and *cemA-petA*, respectively, were due to those in copy number variation with two major tandem repeats (Figure 2). The alignment showed that InDels with lengths of 28 and 22 bp were present at the same position in the CMS-S cytotype (Figure 2). Thus, two pairs of primers were designed based on the InDels with length of 28 and 22 bp to amplify the regions (Table 4). To evaluate the accuracy of this marker, we tested five pairs of near isogenic lines (NILs) of *Sorghum* and compared the results with those obtained with the CMS specific InDel markers. PCR amplification showed that *Sorghum* cytotypes could be clearly distinguished using gel electrophoresis (Figure 3). All male sterile lines known to contain S-cytoplasm had upper bands of 270 and 265 bp with the cp_01 and cp_02 InDel markers, respectively. In contrast, all maintainer lines with N-cytoplasm had lower bands of 242 and 243 bp (Table 2). We also identified the cytoplasmic male sterile factors, namely, S or N, by InDel markers with 1,104 plants from six South Korean *Sorghum* cultivars (Figure 3 and Table 3). The results of marker analysis revealed that individuals of each cultivar, Donganme, Sodamchalm, and Baremae, had 100% S-cytoplasm. In contrast, Hwanggeumchal, Nampungchal, and Noeulchal only contained N-cytoplasm types (Table 3). All cultivars containing S-cytoplasm were found to have a 28 bp insertion.

As *Sorghum* has a wide range of mating rates of more than 7-30% (Djè et al., 2000; Barnaud et al., 2008), we expected genetic variation in the cytoplasmic genome, but they were all identical. In *Sorghum* breeding, paper bags are used for repeated self-pollination and generation advancement to avoid outcrossing. Consequently, the developed varieties appear to have the same cytoplasmic type, implying that the breeding cultivars are genetically fixed, and the pure line is well maintained. Hwannggeumchal and Nampungchal

10.3389/fpls.2023.1188149

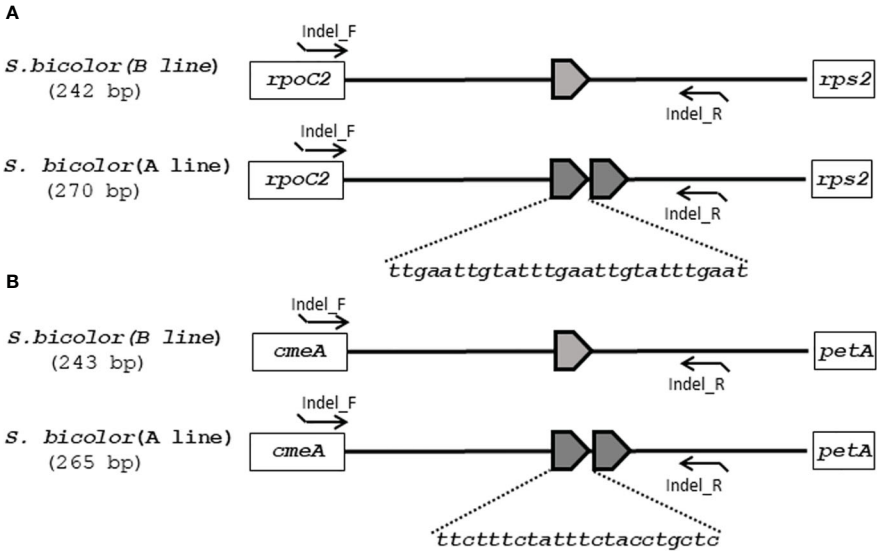


FIGURE 2
Schematic diagrams of **(A)** InDel cp_01 and **(B)** InDel cp_02 with cytoplasmic male sterile and maintainer lines of *Sorghum bicolor* in chloroplast genomes. **(A)** 28 and **(B)** 22 base pairs of tandem repeats (TRs) are represented as pentagons. The InDel cp_01 and InDel cp_02 marker primers are indicated as arrowed lines.

contain the S-cytoplasm type, whereas Donganme and Sodmchal contain the N-cytoplasm type (Table 1). In *Sorghum*, the first report described a 165 bp deletion in the middle of *rpoC2* in CMS lines containing A1, A2, A5, and A6 cytoplasm (Chen et al., 1995). Consistent with previous results, all male sterile lines used in this study were included in the A1 cytoplasm.

In Korea, *Sorghum* is usually bred by landrace selection and utilization of landraces in a breeding program. The cytoplasm types in Korean *Sorghum* varieties have not yet been identified. Through cp genome sequencing, the cytotypes of Korean varieties, Hwanggeumchal and Nampungchal contained 140,753 bp, whereas those of Donganme and Sodamchal cytotypes contained 140,644 bp (Table 1).

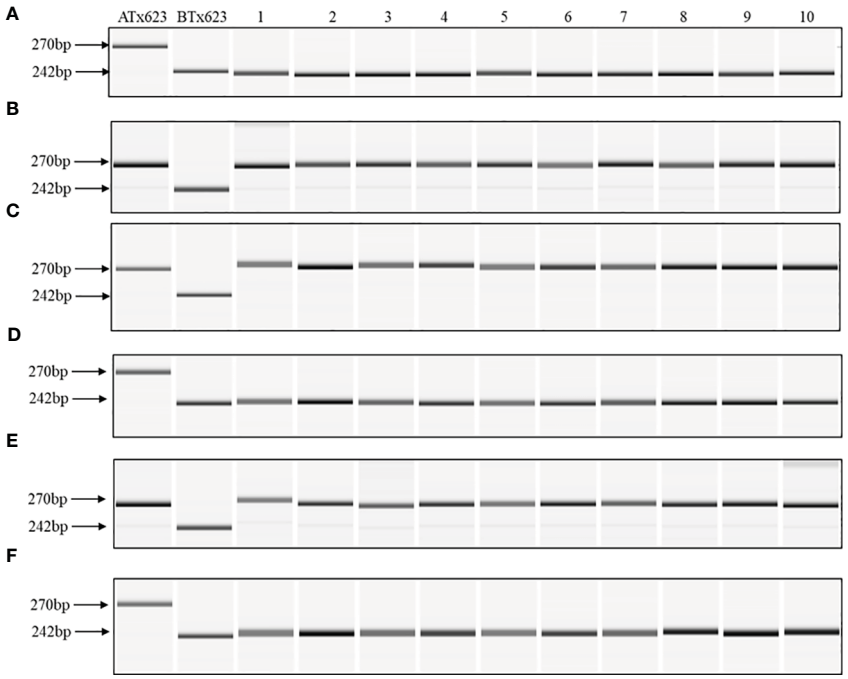


FIGURE 3
Capillary gel electrophoresis of Korean *Sorghum bicolor* cultivars with cytoplasmic male sterile factor-specific InDel cp_01 markers.
(A) Nampoongchal, **(B)** Donganme, **(C)** Sodamchal, **(D)** Hwanggeumchal, **(E)** Bareme, and **(F)** Noeulchal. Ten individual plants (1 to 10) of each cultivar and cytoplasmic male sterile (ATx623) and maintainer (BTx623) lines are depicted.

3.3 Comparative chloroplast genome analysis with congeneric species of *Sorghum*

Typically, IR regions have identical lengths; however, they can extend or contract inside the chloroplast. Therefore, we compared the cp genomes of LSC, SSC, and IR (IRa and IRb) among ATx623, *S. sudanense, S. propinquum*, and BTx623 (Table 5 and Figure 4). The total lengths of the cp genome of ATx623 and *S.propinquum* were nearly identical (140,644 and 140,642 bp, respectively), whereas those of *S.sudanens* and BTx623 were also identical (140,755 and 140,754 bp, respectively) (Table 5). However, the total cp genome lengths of *S. sudanense* and BTx623 were slightly longer (111-112 bp) than those of ATx623 and *S. propinquum* respectively. The guanine-cytosine content in the cp genome of *S. bicolor* (ATx623 and BTx623) and congeneric species (*S. sudanense* and *S. propinquum*) was 38.5%, and all the species contained 114 unique genes (Table 5). Additionally, we discovered that the size of intergenic spacers (IGSs) between the *rpl22-rps19* genes of *ATx623* in the junction between the LSC and IRb regions (LSC-IRb) were similar to those of *S. sudanense*, *S.propinquum* and BTx623 (Figure 4). Similarly, the IGSs between the *rps19- psbA* genes, located in the IRa-LSC junction, of ATx623, *S sudanense, S propinquum* and BTx623 were similar (Figure 4). The boundaries between the IRa regions were similar in size (1,182 bp) in all the compared species (Figure 4). Similarly, the *ndhH* gene spanned the IRb-SSC region, and the fragment located in the IRb region was equal in size (2,188 bp) among the compared species.

3.4 Phylogenetic analysis

Molecular phylogenetic analysis offers new perspectives on the evolutionary linkages between species. Thus, phylogenetic analysis was conducted using the complete cp genome sequences of the eight *Sorghum* species. The results of the maximum composite likelihood analysis are shown in the phylogenetic tree (Figure 5). The phylogenetic tree was monophyletic and formed two clades within these eight *Sorghum* species, wherein *S. timorense* was the outgroup. A strong bootstrap value (100%) was observed for three of the five nodes. Eusorghum species, such as *S. bicolor* (ATx623), *S. propinquum, S. halepense, S. sudanense, S. bicolor* (BTx623), and *S. arundinaceum* were grouped into one clade, whereas *hemisorghum mekongense* formed another group clade. In the Eusorghum species clade, *S. biclor* (ATx623) was the sister to *S.propinquum* in the same branch, whereas *S. sudanense* was the sister to *S. bicolor* (BTx623), with a short branch length, indicating a dispersed evolutionary history or that they are a closer ancestor. Genetic distance analysis revealed the considerably less genetic distance between the analyzed *Sorghum* species. The lowest genetic distance value of 0.0000 was observed between *S. propinquum* and *S. color* (ATx 623) followed by the second lowest value of 0.00002 between *S. bicolor* (BTx623) and *S. sudanese* (Supplementary Table 4).

4 Discussion

The cp genome is a useful tool for analyzing the evolutionary relationships among species. This is due to the fact that photosynthesis-related organelles such as chloroplasts, contain a circular genome that is comparatively stable and is passed along from the mother to offspring. Moreover, recent research has focused on the cp genome, as it provides essential genetic information to investigate the evolutionary links between related species.

The overall structural organization and introns, genes, and gene order of the analyzed cp genome of a CMS-S line and South Korean cultivars of *Sorghum* were conserved and showed no significant difference in the cp genome size. Similarly, the SSC, LSC, IR regions, and GC content of cp genomes (38.5%) (Table 5) were also found to be similar among the Eusorghum species. These results are consistent with previous studies on the cp genomes of *Sorghum* and other species from the *Poaceae* family (Lu et al., 2016; Song et al., 2017; Song et al., 2019). Recently, numerous taxonomists have focused on the cp genome to investigate the phylogenetic relationships of related species. For example, cp genomes can provide sufficient genetic information for species identification. In

TABLE 5 Summary of chloroplast genome characteristics for four *Sorghum* genera containing the male sterile (ATx623) and maintainer line (BTx623) of *Sorghum bicolor*.

Category	*Sorghum bicolor* (ATx623)	*Sorghum propinquum*	*Sorghum sudanense*	*Sorghum bicolor* (BTx623)
GenBank accession No.	MT459453	MH926027	MH926028	EF115542
Total length (bp)	140,644	140,642	140,755	140,754
LSC length (bp)	82,574	82,572	82,686	82,685
SSC length (bp)	12,506	12,506	12,503	12,503
IRa length (bp)	22,782	22,782	22,783	22,783
IRb length (bp)	22,782	22,782	22,783	22,783
Total GC content (%)	38.48	38.48	38.49	38.49
Total number of genes	114	114	114	114

LSC, Large Single Copy; SSC; IR, Inverted Repeat.

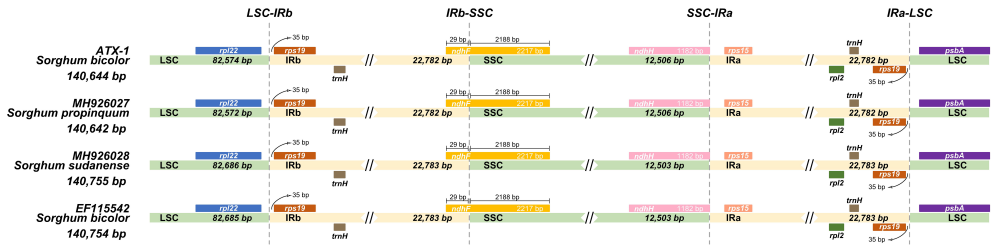


FIGURE 4
Comparison of the border position and size of LSC, SSC, and IR regions in the chloroplast genome of four Sorghum species. Gene names of each border are designated in boxes. LSC, Large Singe Copy; SSC, Small Single Copy; IRs, Inverted Repeats.

this study, we developed InDel markers based on sequence variation in the cp genome for the accurate cytoplasm identification of species and developed InDel markers for further cytoplasm evaluation of species.

In the chloroplasts, the *ndhD* gene is a component of the NADH dehydrogenase complexes. The specific mutations in the *ndhD* gene hinder the NADH dehydrogenase complex's ability to operate normally, which reduces the anther's ability to produce energy. In wheat, a mutation in the *ndhD* gene causes male sterility in wheat. The mutation is a single-nucleotide substitution that changes a cytosine to a thymine results in a frameshift that leads to the production of a truncated *ndhD* protein (Han et al., 2022). *PsaA* and *psaB* gene are the subunit of photosystem I (PSI) and PSI complex of proteins that uses light energy to drive the transfer of electrons from water to NADPH. These complex has been shown to play a role in the regulation of gene expression in plants (Azarin et al., 2020). Therefore, we analyzed the *ndhD, psaA, psaB* gene sequences (nucleotide and amino acids) with Clutal W and we found that there is no genetic variation such as SNP or InDel between male sterile line and maintainer line (data not shown).

Recently, cp genome sequence analysis has been successfully used to reconstruct phylogenetic relationships among plant lineages. Previous phylogenetic studies based on entire cp genomes have been used to resolve the difficult phylogenetic relationships among closely related species. In this study, the whole cp genomes of a *Sorghum* CMS-S line and South Korean cultivars were sequenced and assembled using next-generation sequencing. In a previous study, four *Sorghum* species were grouped into two groups. *S. sudanense, S. bicolor*, and *S. propinquum* formed groups. *S. sudanense, S. bicolor*, and *S. propinquum* belong to the subgenus *Sorghum* which contains 10 species (Song et al., 2019). Phylogenetic analysis using the complete cp genome of seven *Sorghum* species, including wild species, revealed that CMS-S and CMS-N of the *S. bicolor* cytoplasm were highly similar to *S. propinquum* and *S. sudanense*, respectively. These results were consistent with those of a previous study (Song et al., 2019; Ananda et al., 2021). *S. propinquum* is a wild perennial diploid rhizomatous species distributed across Southeast Asia and the Indian subcontinent. Various traits are potentially useful for the introgression of *S. propinquum* into *S. bicolor*. The primary gene pool of modern *Sorghum* cultivars contains the wild species, *S.*

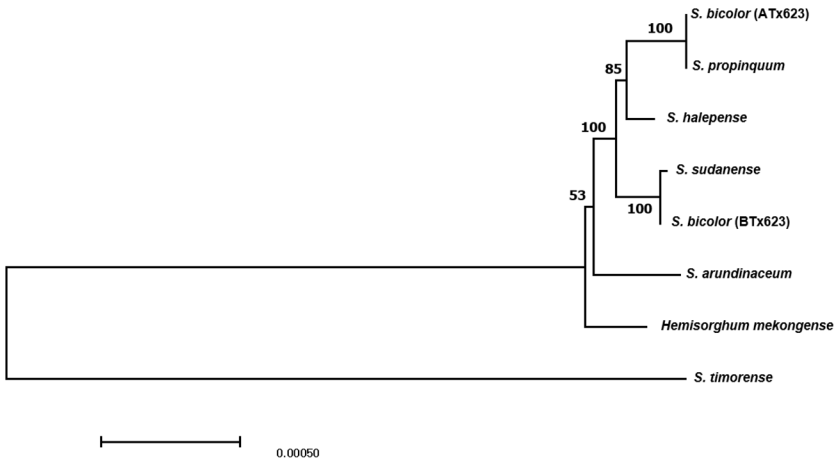


FIGURE 5
The phylogenic tree of six congeneric *Sorghum* species, including ATx623 (male sterile line) and BTx623 (maintainer line) of *Sorghum bicolor*, constructed using the maximum composite likelihood model. Bootstrap values are shown below each clade.

propinquum (De Wet and Harlan, 1971; De Wet, 1978). Previous results found that S. *propinquum* showed increased height, early maturity, and high yield (Ananda et al., 2020). This might explain the close relationship between *S. bicolor* (ATx623) and *S. propinquum*.

In this study, the S cytoplasm of *S. bicolor Milo* was found to have genetic exchange with *S. propinquum*. In contrast, *S. sudanense* is believed to be segregated from a natural hybrid of *S. bicolor* and *S.arundinaceum*. Our findings revealed that *S. sudanense* is closely related to *S. bicolor*, which represents CMS-N, including the maintenance and restoration lines; hence, these results are consistent with those of previous studies. In *S. bicolor*, the *milo* cytoplasm (A1) has been widely used in hybrid production, and *kafir* has been used as a maintainer line as it produces fully fertile hybrids when crossed with the milo parent. These results phylogenetically support the fact that the *milo* and *kafir* cytotypes originated from *S. propinquum* and *S. Sudanese*, respectively. Bayesian inference analysis indicated that the *Sorghum* genus diverged from Miscanthus about 19.5 million years ago (mya). Smaller spikelets are a distinctive feature of *S. propinquum*. This is consistent with the morphology of the small anther, and pollen depleted exone caused by a 165 bp deletion of the *rpoC2* region, such as in the A1, A2, A5, and A6 cytoplasms. Doggett (1988) proposed that durra (*milo*) originated in Ethiopian because it contains the entire set of wild-type bicolor-durra crosses.

In the CMS system, the breeding programs were divided into two groups. One group was devoted to the development of the female inbred line (A/B-line) and the second was devoted to the development of the male inbred line (R-line). Prior to the hybrid development program, testcrossing, or sterilization, a new line should be identified as maintainers or restorers by a testcross. If the lines with the S cytoplasm have a dominant allele present in the nuclear genome, the plant will be an R-line to restore male fertility. Unless the line lacks the dominant allele for fertility restoration, the plant will be male-sterile (Senthil et al., 1994). Maintainer lines have the N cytoplasm and lack a dominant *Rf* allele. It is easy to discover a new B-line or predict the male infertility gene type by simply identifying the cytoplasmic type using a marker prior to crossbreeding. When developing B-lines, resources with the N-cytoplasm are first selected as markers, and new lines can then be cultivated through crossbreeding between B-lines. N (*Rfrf*) can also be used if the combinatorial ability test is conducted on a lineage with the N-cytoplasm. In conclusion, the newly developed InDel markers based on the cp genome variation can facilitate a new F1 hybrid breeding system in Korea.

Data availability statement

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

Author contributions

M-EC conceived the design of the study, analyzed the data, and drafted the manuscript. J-YK and S-IH collected and grew *Sorghum* cultivars and lines in Miryang, Korea. RS conducted the bioinformatics work and was engaged in drafting the manuscript. K-SC was responsible for data analysis and writing of the manuscript. All authors read and approved the final manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was conducted with the support of the "Co-operative Research Program for Agriculture Science & Technology Development (Project No. PJ01505601)," Rural Development Administration, Republic of Korea. This project funded by Korean government not commercial affiliation.

Conflict of interest

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

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1188149/full#supplementary-material

SUPPLEMENTARY FIGURE 1
The comparison of the chloroplast genome structure of *Sorghum* genus using the mVISTA program.

SUPPLEMENTARY FIGURE 2
Multiple sequence alignment of InDel region in the chloroplast genome for the identification of cytoplasmic male sterile factors in *Sorghum bicolor*. (A) InDel cp_01 and (B) InDel cp_02. Each tandem repeat is shown as a black dotted line.

References

Ananda, G. K. S., Myrans, H., Norton, S. L., Gleadow, R., Furtado, A., and Henry, R. J. (2020). Wild sorghum as a promising resource for crop improvement. *Front. Plant Sci.* 11, 1108. doi: 10.3389/fpls.2020.01108

Ananda, G., Norton, S., Blomstedt, C., Furtado, A., Møller, B., Gleadow, R., et al. (2021). Phylogenetic relationships in the sorghum genus based on sequencing of the chloroplast and nuclear genes. *Plant Genome* 14 (3), e20123. doi: 10.1002/tpg2.20123

Arthan, W., Mckain, M. R., Traiperm, P., Welker, C. a. D., Teisher, J. K., and Kellogg, E. A. (2017). Phylogenomics of Andropogoneae (Panicoideae: Poaceae) of Mainland Southeast Asia. *Syst. Bot.* 42, 418–431.

Asaf, S., Khan, A. L., Khan, M. A., Waqas, M., Kang, S. M., Yun, B. W., et al. (2017). Chloroplast genomes of arabidopsis halleri ssp. gemmifera and arabidopsis lyrata ssp. petraea: structures and comparative analysis. *Sci. Rep.* 7, 7556. doi: 10.1038/s41598-017-07891-5

Azarin, K., Usatov, A., Makarenko, M., Kozel, N., Kovalevich, A., Dremuk, I., et al. (2020). A point mutation in the photosystem I P700 chlorophyll a apoprotein A1 gene confers variegation in helianthus annuus l. *Plant Mol. Biol.* 103, 373–389. doi: 10.1007/s11103-020-00997-x

Bailey-Serres, J., Hanson, D. K., Fox, T. D., and Leaver, C. J. (1986). Mitochondrial genome rearrangement leads to extension and relocation of the cytochrome c oxidase subunit I gene in sorghum. *Cell* 47, 567–576. doi: 10.1016/0092-8674(86)90621-5

Barnaud, A., Trigueros, G., Mckey, D., and Joly, H. (2008). High outcrossing rates in fields with mixed sorghum landraces: how are landraces maintained? *Heredity* 101, 445–452. doi: 10.1038/hdy.2008.77

Bosacchi, M., Gurdon, C., and Maliga, P. (2015). Plastid genotyping reveals the uniformity of cytoplasmic male sterile-T maize cytoplasms. *Plant Physiol.* 169, 2129–2137. doi: 10.1104/pp.15.01147

Bout, S., and Vermerris, W. (2003). A candidate-gene approach to clone the sorghum brown midrib gene encoding caffeic acid O-methyltransferase. *Mol. Genet. Genomics* 269, 205–214. doi: 10.1007/s00438-003-0824-4

Casto, A. L., Mattison, A. J., Olson, S. N., Thakran, M., Rooney, W. L., and Mullet, J. E. (2019). Maturity2, a novel regulator of flowering time in sorghum bicolor, increases expression of SbPRR37 and SbCO in long days delaying flowering. *PloS One* 14, e0212154. doi: 10.1371/journal.pone.0212154

Chen, Z., Muthukrishnan, S., Liang, G. H., Schertz, K. F., and Hart, G. E. (1993). A chloroplast DNA deletion located in RNA polymerase gene rpoC2 in CMS lines of sorghum. *Mol. Gen. Genet.* 236, 251–259. doi: 10.1007/BF00277120

Chen, Z., Schertz, K. F., Mullet, J. E., Dubell, A., and Hart, G. E. (1995). Characterization and expression of rpoC2 in CMS and fertile lines of sorghum. *Plant Mol. Biol.* 28, 799–809. doi: 10.1007/BF00042066

Cho, K.-S., Lee, H.-O., Lee, S.-C., Park, H.-J., Seo, J.-H., Cho, J.-H., et al. (2022). Mitochondrial genome recombination in somatic hybrids of solanum commersonii and s. tuberosum. *Sci. Rep.* 12 (1), 8659. doi: 10.1038/s41598-022-12661-z

Cho, K., Yang, T., Hong, S., Kwon, Y., Woo, J., and Park, H. (2006). Determination of cytoplasmic male sterile factors in onion plants (Allium cepa l.) using PCR-RFLP and SNP markers. *Molecules Cells* 21, 411.

Cho, K.-S., Yun, B.-K., Yoon, Y.-H., Hong, S.-Y., Mekapogu, M., Kim, K.-H., et al. (2015). Complete chloroplast genome sequence of tartary buckwheat (Fagopyrum tataricum) and comparative analysis with common buckwheat (F. esculentum). *PloS One* 10, e0125332. doi: 10.1371/journal.pone.0125332

Daniell, H., Lin, C. S., Yu, M., and Chang, W. J. (2016). Chloroplast genomes: diversity, evolution, and applications in genetic engineering. *Genome Biol.* 17, 134. doi: 10.1186/s13059-016-1004-2

De Wet, J. M. J. (1978). Systematics and evolution of sorghum sect. sorghum (Gramineae). *Am. J. Bot.* 65, 477–484. doi: 10.1002/j.1537-2197.1978.tb06096.x

De Wet, J. M. J., and Harlan, J. R. (1971). The origin and domestication of sorghum bicolor. *Economic Bot.* 25, 128–135. doi: 10.1007/BF02860074

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

Doggett, H. (1988). Sorghum 2nd edition tropical agriculture. *Ser. Longman Sci. Technical Essex England* 231 (4), 243–254.

Han, Y., Gao, Y., Li, Y., Zhai, X., Zhou, H., Ding, Q., et al. (2022). Chloroplast genes are involved in the Male-sterility of K-type CMS in wheat. *Genes (Basel)* 13 (2), 310. doi: 10.3390/genes13020310

Hilley, J. L., Weers, B. D., Truong, S. K., Mccormick, R. F., Mattison, A. J., Mckinley, B. A., et al. (2017). Sorghum Dw2 encodes a protein kinase regulator of stem internode length. *Sci. Rep.* 7, 4616. doi: 10.1038/s41598-017-04609-5

Hong, S. Y., Cheon, K. S., Yoo, K. O., Lee, H. O., Cho, K. S., Suh, J. T., et al. (2017). Complete chloroplast genome sequences and comparative analysis of chenopodium quinoa and c. album. *Front. Plant Sci.* 8, 1696. doi: 10.3389/fpls.2017.01696

Hong, S.-Y., Cheon, K.-S., Yoo, K.-O., Lee, H.-O., Mekapogu, M., and Cho, K.-S. (2019). Comparative analysis of the complete chloroplast genome sequences of three amaranthus species. *Plant Genet. Resour.* 17, 245–254. doi: 10.1017/S1479262118000485

Kante, M., Rattunde, H. F. W., Nébié, B., Weltzien, E., Haussmann, B. I. G., and Leiser, W. L. (2018). QTL mapping and validation of fertility restoration in West African sorghum A1 cytoplasm and identification of a potential causative mutation for Rf2. *Theor. Appl. Genet.* 131, 2397–2412. doi: 10.1007/s00122-018-3161-z

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

Kim, Y. J., and Zhang, D. (2018). Molecular control of Male fertility for crop hybrid breeding. *Trends Plant Sci.* 23, 53–65. doi: 10.1016/j.tplants.2017.10.001

Li, Y., and Li, C. (1998). Genetic contribution of Chinese landraces to the development of sorghum hybrids. *Euphytica* 102, 47–57. doi: 10.1023/A:1018374203792

Lohse, M., Drechsel, O., and Bock, R. (2007). OrganellarGenomeDRAW (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. *Curr. Genet.* 52, 267–274. doi: 10.1007/s00294-007-0161-y

Lu, D., Zhao, Y., Han, R., Wang, L., and Qin, P. (2016). The complete chloroplast genome sequence of the purple feathergrass stipa purpurea (Poales: poaceae). *Conserv. Genet. Resour.* 8, 101–104. doi: 10.1007/s12686-016-0519-x

Mccormick, R. F., Truong, S. K., Sreedasyam, A., Jenkins, J., Shu, S., Sims, D., et al. (2018). The *Sorghum bicolor* reference genome: improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. *Plant J.* 93, 338–354. doi: 10.1111/tpj.13781

Melonek, J., Duarte, J., Martin, J., Beuf, L., Murigneux, A., Varenne, P., et al. (2021). The genetic basis of cytoplasmic male sterility and fertility restoration in wheat. *Nat. Commun.* 12, 1036. doi: 10.1038/s41467-021-21225-0

Mundia, C. W., Secchi, S., Akamani, K., and Wang, G. (2019). A regional comparison of factors affecting global sorghum production: the case of north America, Asia and africa's sahel. *Sustainability* 11, 2135. doi: 10.3390/su11072135

Paterson, A. H., Bowers, J. E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., et al. (2009). The *Sorghum bicolor* genome and the diversification of grasses. *Nature* 457, 551–556. doi: 10.1038/nature07723

Praveen, M., Uttam, A. G., Tonapi, V. A., and Madhusudhana, R. (2018). Fine mapping of *Rf2*, a major locus controlling pollen fertility restoration in sorghum A1 cytoplasm, encodes a PPR gene and its validation through expression analysis. *Plant Breed.* 137, 148–161. doi: 10.1111/pbr.12569

Pring, D. R., Conde, M. F., Schertz, K. F., and Levings, C. S. (1982). Plasmid-like DNAs associated with mitochondria of cytoplasmic male-sterile *Sorghum*. *Mol. Gen. Genet. MGG* 186, 180–184. doi: 10.1007/BF00331848

Saski, C., Lee, S. B., Fjellheim, S., Guda, C., Jansen, R. K., Luo, H., et al. (2007). Complete chloroplast genome sequences of *Hordeum vulgare, sorghum bicolor* and *Agrostis stolonifera*, and comparative analyses with other grass genomes. *Theor. Appl. Genet.* 115, 571–590. doi: 10.1007/s00122-007-0567-4

Schertz, K. (1977). Registration of A2 Tx2753 and B Tx2753 sorgum germplasm 1 (Reg.No. GP 30 and 31). *Crop Sci.* 17 (6), 983–983. doi: 10.2135/cropsci1977.0011183X001700060056x

Senthil, N., Rangasamy, S. S., and Palanisamy, S. (1994). Male Sterility inducing cytoplasm in sorghum classification, genetics of sterility and fertility restoration studies. *Cereal Res. Commun.* 22 (3), 179–184.

Sivaramakrishnan, S., Seetha, K., and Reddy, B. V. (1997). Characterization of the a 4 cytoplasmic male-sterile lines of sorghum using RFLP of mtDNA. *Euphytica* 93, 301–305. doi: 10.1023/A:1002906606333

Song, Y., Chen, Y., Lv, J., Xu, J., Zhu, S., and Li, M. (2019). Comparative chloroplast genomes of *Sorghum* species: sequence divergence and phylogenetic relationships. *BioMed. Res. Int.* 2019, 5046958. doi: 10.1155/2019/5046958

Song, Y., Chen, Y., Lv, J., Xu, J., Zhu, S., Li, M., et al. (2017). Development of chloroplast genomic resources for *Oryza* species discrimination. *Front. Plant Sci.* 8, 1854. doi: 10.3389/fpls.2017.01854

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

Tamura, K., Nei, M., and Kumar, S. (2004). Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proc. Natl. Acad. Sci.* 101, 11030–11035. doi: 10.1073/pnas.0404206101

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

Tari, I., Laskay, G., Takács, Z., and Poór, P. (2013). Response of sorghum to abiotic stresses: a review. *J. Agron. Crop Sci.* 199, 264–274. doi: 10.1111/jac.12017

Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E. S., Fischer, A., Bock, R., et al. (2017). GeSeq–versatile and accurate annotation of organelle genomes. *Nucleic Acids Res.* 45, W6–W11. doi: 10.1093/nar/gkx391

Xu, G.-W., Cui, Y.-X., Schertz, K., and Hart, G. (1995). Isolation of mitochondrial DNA sequences that distinguish male-sterility-inducing cytoplasms in *Sorghum bicolor* (L.) moench. *Theor. Appl. Genet.* 90, 1180–1187. doi: 10.1007/BF00222941

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 14 August 2023
DOI 10.3389/fpls.2023.1199748

OPEN ACCESS

EDITED BY
Manohar Chakrabarti,
The University of Texas Rio Grande Valley, United States

REVIEWED BY
Dong-Liang Huang,
Guangxi Academy of Agricultural Sciences, China
Niranjan Baisakh,
Louisiana State University, United States

*CORRESPONDENCE
Robert Henry
robert.henry@uq.edu.au

RECEIVED 03 April 2023
ACCEPTED 17 July 2023
PUBLISHED 14 August 2023

CITATION
Thirugnanasambandam PP, Singode A, Thalambedu LP, Athiappan S, Krishnasamy M, Purakkal SV, Govind H, Furtado A and Henry R (2023) Long read transcriptome sequencing of a sugarcane hybrid and its progenitors, *Saccharum officinarum* and *S. spontaneum*. *Front. Plant Sci.* 14:1199748. doi: 10.3389/fpls.2023.1199748

Long read transcriptome sequencing of a sugarcane hybrid and its progenitors, *Saccharum officinarum* and *S. spontaneum*

Prathima Perumal Thirugnanasambandam[1], Avinash Singode[2], Lakshmi Pathy Thalambedu[1], Selvi Athiappan[1], Mohanraj Krishnasamy[1], Sobhakumari Valiya Purakkal[1], Hemaprabha Govind[1], Agnelo Furtado[3] and Robert Henry[3]*

[1]Crop Improvement Division, ICAR-Sugarcane Breeding Institute, Coimbatore, Tamil Nadu, India, [2]Indian Council of Agricultural Research (ICAR)-Indian Institute of Millets Research, Hyderabad, Telangana, India, [3]Queensland Alliance for Agriculture and Food Innovation, The University of Queensland, St. Lucia, Brisbane, QLD, Australia

Commercial sugarcane hybrids are derivatives from *Saccharum officinarum* and *Saccharum spontaneum* hybrids containing the full complement of *S. officinarum* and a few *S. spontaneum* chromosomes and recombinants with favorable agronomic characters from both the species. The combination of the two sub-genomes in varying proportions in addition to the recombinants presents a challenge in the study of gene expression and regulation in the hybrid. We now report the transcriptome analysis of the two progenitor species and a modern commercial sugarcane hybrid through long read sequencing technology. Transcripts were profiled in the two progenitor species *S. officinarum* (Black Cheribon), and *S. spontaneum* (Coimbatore accession) and a recent high yielding, high sugar variety Co 11015. The composition and contribution of the progenitors to a hybrid with respect to sugar, biomass, and disease resistance were established. Sugar related transcripts originated from *S. officinarum* while several stress and senescence related transcripts were from *S. spontaneum* in the hybrid. The hybrid had a higher number of transcripts related to sugar transporters, invertases, transcription factors, trehalose, UDP sugars, and cellulose than the two progenitor species. Both *S. officinarum* and the hybrid had an abundance of novel genes like sugar phosphate translocator, while *S. spontaneum* had just one. In general, the hybrid shared a larger number of transcripts with *S. officinarum* than with *S. spontaneum*, reflecting the genomic contribution, while the progenitors shared very few transcripts between them. The common isoforms among the three genotypes and unique isoforms specific to each genotype indicate that there is a high scope for improvement of the modern hybrids by utilizing novel gene isoforms from the progenitor species.

KEYWORDS

***Saccharum officinarum*, Black Cheribon, *Saccharum spontaneum*, Coimbatore accession, Co 11015, progenitors, long read transcriptome, sugar genes**

1 Introduction

Modern sugarcane hybrids are complex polyploids derived from polyploid progenitor species, *Saccharum officinarum* and *Saccharum spontaneum*. *S. officinarum*, originally found growing in the tropical Papua-New Guinea region, is rich in sugars, due to which it was called noble cane. *S. spontaneum* is a grassy wild species with extensive distribution from Africa to Southeast Asia and the Pacific islands and has a diverse gene pool for adaptability and resistance to biotic and abiotic stresses. *S. officinarum* ($2n = 8\times = 80$, $x = 10$) has high sugar content of about 18-20 degree Brix and is reported to have been domesticated around 8000 years ago from the wild species *S. robustum* (Pompidor et al., 2021). *S. spontaneum* has various cytotypes, many aneuploid forms ($2n = 5\times = 40$ to $16\times = 128$; $x = 8$), and has a sugar content of less than 10 degree Brix (Garsmeur et al., 2018).

The earliest sugarcane breeding and selection program in 1888 in Java, Indonesia, incorporated the disease resistance, hardiness, and tillering capacity of *S. spontaneum* into *S. officinarum* germplasm. The resultant hybrids were repeatedly backcrossed to *S. officinarum* as a recurrent female parent in a process called nobilization (Stevenson, 1965). An important phenomenon called "female restitution" occurs during the crossing with *S. officinarum*, wherein 2n+n and n+n transmission of chromosomes happens in the F1 hybrid and BC1 progeny respectively (Premachandran et al., 2011). The rapid recovery of high sugar commercial types from the interspecific hybridization of *S. officinarum* with *S. spontaneum* is attributed to the transmission of the diploid complement of the *S. officinarum* to the hybrid. The first interspecific hybrid, Co 205, a selection from a cross between *S. officinarum* cultivar Vellai and *S. spontaneum* Coimbatore was developed in India in 1912 while POJ2725 and POJ2878 were developed in Java in 1921(Jackson et al., 2014). These inter-specific hybrids served as the foundation for all the modern hybrids of sugarcane worldwide. Commercial sugarcane hybrids which are derivatives from such hybrids contain the full complement of *S. officinarum* and a few *S. spontaneum* chromosomes imparting the favorable agronomic characters from both the species. Such unequal contribution of each progenitor to the hybrid genome was revealed by genomic *in situ* hybridization (GISH) and fluorescent *in situ* hybridization (FISH) studies, demonstrating that the female parent *S. officinarum* contributed about 80% of the chromosomes to the genome of the hybrids, while the male parent *S. spontaneum* contributed only 10%–20% to the hybrid genome (D'Hont et al., 1996; Piperidis and D'Hont, 2001; Cuadrado et al., 2004; D'Hont, 2005). About 5%–17% of the chromosomes resulted from a recombination of chromosomes from the two parental species. Furthermore, each sugarcane hybrid cross most likely directly reflects the chromosome ratio originally from the two parental species, while phenotypically, the greater the contribution of the wild *S. spontaneum*, the greater the fiber content, hardiness, high tillering and vigor in the hybrid (Matsuoka et al., 2014). The resulting sugarcane hybrid genome is composed of a unique chromosome set (ranging from 100-130), containing up to 12-14 copies of each gene (Piperidis and D'Hont, 2001). The monoploid sugarcane genome is estimated to be 382 Mb in size (Garsmeur et al., 2018) while the polyploid sugarcane nuclear genome is about 10 Gb (D'Hont and Glaszmann, 2001; Hoarau et al., 2001; Le Cunff et al., 2008) The genomes of *S. officinarum* LA Purple and *S. spontaneum* SES208 were explored by earlier studies beginning from 1996 (D'Hont et al., 1996). Recently, genomes of *S. spontaneum*, *S. officinarum* and the hybrid genotype R570 were explored (Zhang et al., 2018; Wang et al., 2022; Zhang et al., 2022). However, the entire polyploid sugarcane genome is not sequenced yet due to the inherent genome complexity resulting from the varied contributions of two to three progenitor genomes (Pompidor et al., 2021), recombination, repetitive content, and alternative splicing (Thirugnanasambandam et al., 2018). The diversity existing in each species of the *Saccharum* complex is so high that sequencing a few genotypes may not truly represent sugarcane. The pan genome concept is very suitable for sugarcane as hybrids show differences in chromosome composition and number, and sequencing just one sugarcane hybrid as a representative might result in missing entire chromosome/chromosomes and their associated genomic information.

For this reason, transcriptomic resources remain a valuable means for unraveling this complex genome. Short read assemblies (Casu et al., 2004; Casu et al., 2007; Figueira et al., 2012; Cardoso-Silva et al., 2014; Park et al., 2015), sugarcane expressed sequence tags (SUCESTs) (Vettore et al., 2001), and *Saccharum officinarum* gene indices (SOGI) (Vettore et al., 2003; Hotta et al., 2010) have formed the basis for initial sugarcane transcriptome studies. However, the short read-based assemblies resulted in chimeric reads and artifacts that do not represent the real transcripts arising from the two different sub-genomes and their recombinant chromosomes. This necessitated the development of sugarcane transcriptome resources based on long read sequencing technology that can capture full length transcripts without the need for assembly. The first reported long read reference transcriptome for sugarcane with 107,598 transcripts was developed from stem, leaf, and root tissues from Australian sugarcane hybrid genotypes (Hoang et al., 2017). The benefits of such long read transcriptomes for sugarcane are enormous. There have been successful experiments in gene editing in sugarcane, leading to modified/altered sugar and biomass compositions (Zale et al., 2016; Kannan et al., 2018; Parajuli et al., 2020; Hussin et al., 2022). These studies were possible as a consequence of the sequencing and identification of the various copies and transcript variants of genes in sugarcane. Here, we show for the first time the sub-genomic origins of transcripts related to the most important traits, sugar and disease resistance, in a modern sugarcane hybrid, in comparison with the founding progenitor species. *S. officinarum* accession Black

Abbreviations: CDD, Conserved domain database; cDNA, Complementary DNA; GDP, Guanosine diphosphate; GTP, guanidine tri phosphate; mRNA, Messenger RNA; NADP, Nicotinamide Adenosine diphosphate; NCBI, National Center for Biotechnology Information; NGS, next generation sequencing; Nr, database, Non-redundant database; ORF, Open reading frame; RNA-seq, Ribonucleic acid sequencing; RPKM, Reads per kilobase per million mapped reads; SNP, single nucleotide polymorphism; SRA, Sugar Research Australia; SUGIT, sugarcane Iso-Seq transcriptome database; SuSy, sucrose synthase; UDP, uridine diphosphate; UTR, Untranslated region.

Cheribon, *S. spontaneum* accession Coimbatore, and the commercial hybrid Co 11015 were chosen for long read transcriptome sequencing using PacBio technology. These progenitors were selected as they occur in the pedigree of the commercial hybrid Co 11015 involving crosses for more than six generations.

2 Materials and methods

2.1 Plant material, RNA extraction and Iso seq sequencing

Three sugarcane genotypes, Co 11015 (commercial sugarcane hybrid), Black Cheribon (*S. officinarum*) and *S. spontaneum* (accession Coimbatore), were used in the study. The commercial hybrid Co 11015 was developed at the ICAR-Sugarcane Breeding Institute, Coimbatore (Hemaprabha et al., 2019). The pedigree of Co 11015 is presented in Figure 1 and the chromosome composition of all the three genotypes is given in the Figure 2. Co 11015 is one of the leading cultivars in Southern India and is considered early maturing (can be harvested from the 8th month of planting and has a high sugar content (24 °Brix). Standard crop management practices were followed to raise a healthy crop in the field. The progenitors were selected based on their occurrence in the breeding program of modern commercial hybrids, while Co 11015 was selected on the basis of performance in a field planting of 36 genotypes (data not shown here). Leaf and stem tissues were collected from Co 11015 and *S. spontaneum* planted in the research fields at ICAR-SBI, Coimbatore while *S. officinarum* Black Cheribon leaf and stem tissues were collected from ICAR-SBI Research station, Kannur, Kerala at 12 months after planting. The leaf sample was pooled from three biological replicates while stem samples were collected from top, middle and bottom internodes of three biological replicates and pooled together. The collected samples were immediately frozen in liquid nitrogen and total RNA extraction was performed using the RNeasy Plant Mini Kit (Qiagen, Hilden, Germany) separately for leaf and stem tissues. Total RNA of each sample was estimated by using a Nanodrop 2000 (Thermo Fisher Scientific, Massachusetts, USA) and a Qubit 3.0 fluorometer (Thermo Fisher Scientific, USA) using an RNA HS assay kit (Thermo Fisher #Q32851, Thermo Fisher Scientific, Massachusetts, USA). The integrity of RNA was evaluated on a 1% agarose gel and on an Agilent 2100 Bioanalyzer (Agilent Technologies, California, USA). The RNA was subjected to cDNA synthesis (pooled equimolar from leaf and stem RNA for three genotypes). The amplification of cDNA was done using the NEBNext® Single Cell/Low Input cDNA Synthesis and Amplification Module (New England Biolabs Inc., Massachusetts, USA) in conjunction with an Iso-Seq Express Oligo Kit (Pacific Biosciences, California, USA). Pronex beads (Promega, Wisconsin, USA) were used for the purification of the cDNA before amplification and later for size selection of the amplified product. The library was constructed using the SMRTbell Express template Preparation Kit 2.0 (Pacific Biosciences, California, USA) as per manufacturers' protocol. The library was purified using Pronex beads (Promega, Wisconsin, USA) and the library size was assessed using a Bioanalyzer (Agilent Technologies, California, USA). About 70 pM of the library was loaded onto one SMRTcell containing 8M ZMW and sequenced in a PacBio Sequel II system in CCS/HiFi mode at the sequencing facility of Nucleome Bioinformatics, Hyderabad, India.

2.1.1 Circular consensus sequence calling and demultiplexing

Calling the circular consensus sequence (CCS) is the very first step in processing the Iso-Seq data which was done using the SMRT tool 'ccs'. This combined multiple sub-reads of the same SMRT bell molecule using a statistical model to produce one highly accurate consensus sequence, also known as a HiFi read. We used Lima, the standard tool to identify barcode and primer sequences in PacBio single molecule sequencing data. The overall workflow for the bioinformatics analysis is given in Figure 3. Lima identifies and removes the 5' and 3' cDNA barcodes.

2.1.1.1 Refining and clustering

In this step, full-length non-chimeric reads (FLNC) were generated for each sample using the tool 'isoseq3 refine', which removes the poly (A) tail and concatemers from the reads. This tool filters for full-length (FL) reads that have a poly (A) tail with at least 20 base pairs and removes the identified tail. The trimmed FL reads are clustered at the isoform level and a consensus is called. Isoseq3 deems two reads to stem from the same transcript if they meet the following criteria: similar transcripts with <100bp 5'overhang,

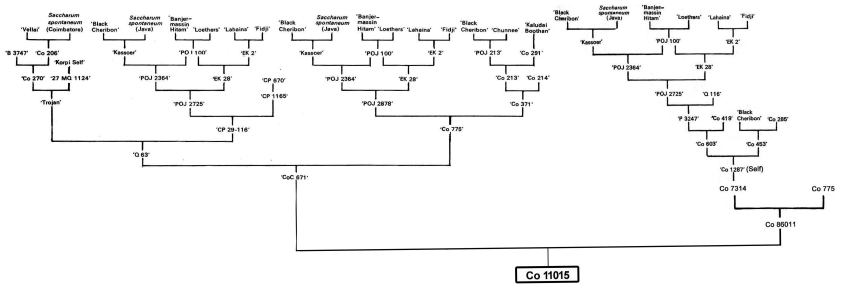


FIGURE 1
Pedigree map of the sugarcane hybrid cultivar Co 11015.

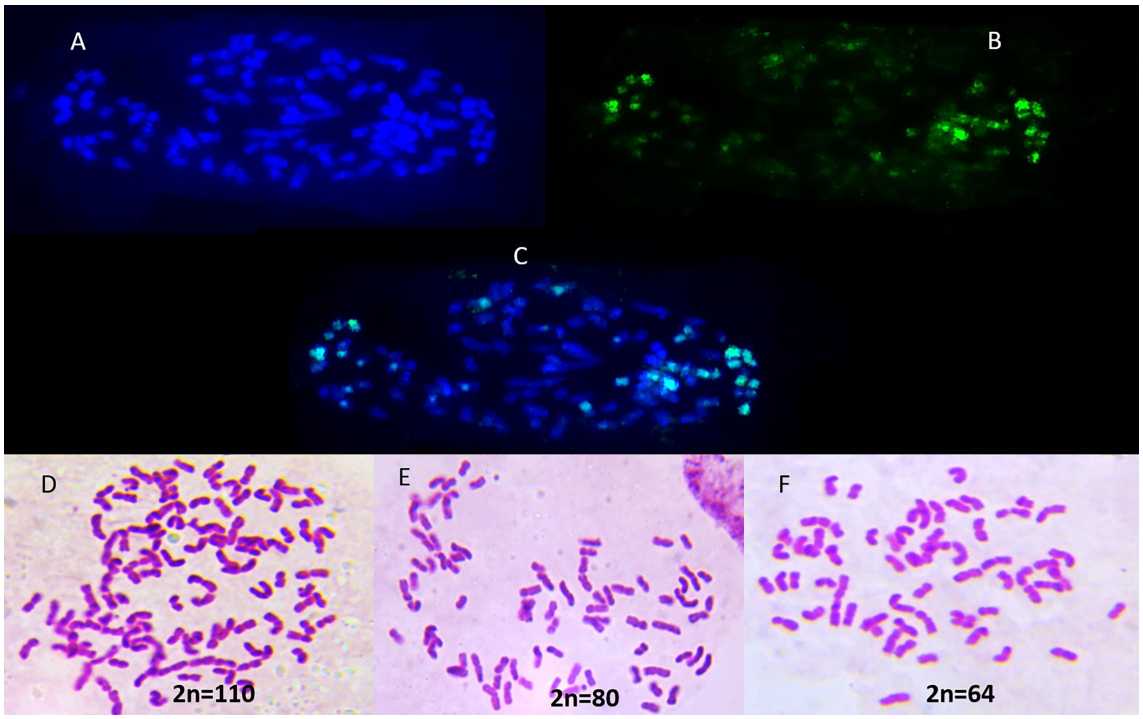


FIGURE 2
Chromosome composition of **(A)** Sugarcane hybrid Co 11015; **(B)** *S. spontaneum* Coimbatore accession as probe on Co 11015; **(C)** merged Co 11015 and *S. spontaneum* showing fluorescence of *S. spontaneum* (greenish blue); **(D–F)** Chromosome spread of Co 11015; *S. officinarum* Black Cheribon and *S. spontaneum* Coimbatore respectively.

<30bp 3' overhang, and <10bp gaps. The transcripts with predicted accuracy of ≥0.99 are considered high-quality reads and <0.99 are considered low-quality reads.

2.1.1.2 Reference mapping and collapsing

During library preparation, 5′ RNA degradation products can be formed and are subsequently sequenced. Collapsing is performed to remove the redundant transcript models and especially redundancy caused by reads originated from 5' degraded RNA. For collapsing the redundant isoforms, clustered high-quality reads were mapped to the reference genome of *Sorghum bicolor* (GCF 000003195.3; https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000003195.3/) using the pbmm2 tool. These mapped reads were then collapsed using the 'isoseq3 collapse' tool.

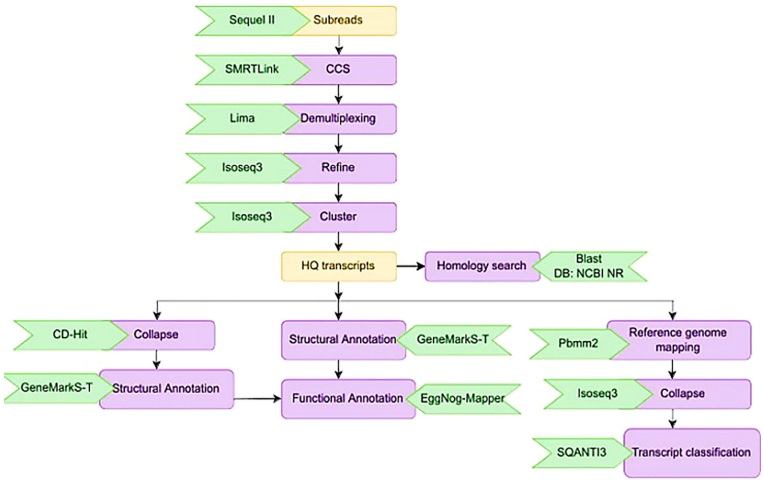


FIGURE 3
Schematic representation of the bioinformatics work flow for processing HiFi reads from three sugarcane genotypes.

2.1.1.3 Transcript classification

SQANTI3 (structural and quality annotation of novel transcript isoforms) was used for the classification of the long-read transcriptome. It classifies the transcripts according to their splice junctions and donor and acceptor sites. Transcripts matching a reference transcript at all the splice junctions are labelled as "full splice match" (FSM), and transcripts matching some consecutive, but not at all, are termed "incomplete splice match" (ISM). SQANTI further classifies the novel transcripts of known genes into two categories: "novel in catalogue" (NIC) and "novel not in catalogue" (NNC). Transcripts in novel genes are classified as "intergenic" if lying outside the boundaries of an annotated gene and as "genic intron" if lying entirely within the boundaries of an annotated intron. In addition, the "genic genomic" category encompasses transcripts with partial exon and intron/intergenic overlap in a known gene.

2.1.1.4 *De novo*-based analysis and annotation of high-quality transcripts

To remove the redundancy from the high-quality (HQ) transcripts, we used the cd-hit tool and collapsed the redundant transcripts and obtained the unique transcripts for each sample. The HQ transcripts of each sample were annotated using NCBI Blastn against the NCBI nr database. From the HQ transcripts, protein-coding regions were predicted using the GeneMarkST tool. This uses the heuristic method of initialization of the hidden semi-Markov model and the viterbi algorithm for finding the maximum likelihood parse of the transcript sequence into coding and non-coding regions. Also, it does iterative self-training on sequences. These predicted cds sequences were used for functional annotation with the help of the eggNOG-mapper v2.1 tool.

2.2 Comparative genomics

The transcriptomes (HQ transcripts) of *S. spontaneum*, *S. officinarum*, and the sugarcane hybrid Co 11015 were aligned to each other using the alignment and mapping tool in CLC Genomics Workbench v22 (CLC-GWB; Qiagen, Aarhus, Denmark). The percentage of mapping was used to determine the total amount of mapped sequence and the percent identity among the three transcriptomes. The probable sub-genome constitution of the sugarcane hybrid was checked by mapping to both the progenitors. The Co 11015 assembly was compared to the assemblies of *S. officinarum* and *S. spontaneum* by mapping the Co 11015 contigs separately using the two progenitor assemblies as references. The mapping was performed to each reference sequence with a 0.8 length fraction and 0.8 similarity fraction indicating an alignment of two reads with 80% length and similarity coverage. The settings were also varied to 0.9 and 0.9 and 1.0 and 1.0 length and similarity fractions, respectively, to capture the exact transcripts, if any, matching in the reference. The transcriptomes from the three genotypes were aligned to other published genomes of sugarcane and sorghum including *Sorghum* genome v5.1 (https://phytozome-next.jgi.doe.gov/) and sugarcane R570 genome (https://www.ncbi.nlm.nih.gov/genome/10780?genome_assembly_id=386616) with 80% identity and 80% coverage threshold for comparison.

2.3 Analyses of sugar and disease resistance genes in the transcriptomes

2.3.1 Sugar genes

The annotated transcriptomes from three genotypes were used for searching for sugar related genes. For this, search terms such as "sugar" and "sucrose" and specific gene lists of the sucrose pathway such as sucrose phosphate synthase (SPS), sucrose phosphate phosphatase, sucrose synthase, and invertases were used for filtering the transcripts.

2.3.1.1 Analysis of sucrose phosphate synthase genes from the three transcriptomes

The SPS genes were filtered from the three transcriptomes and the length distribution of the transcripts was analysed. The transcripts were translated to protein sequences using the Expasy tool (https://web.expasy.org/translate). The protein coding full-length sequences were used for multiple sequence alignment and phylogeny (CLC WB, V22). The tree file was exported in the Newick format and viewed using NCBI Treeviewer. Motif distribution among the SPS transcripts from all three genotypes was found using MEME (https://meme-suite.org/meme/). In addition, sucrose phosphate phosphatase (SPP) and sucrose phosphate translocator (SPT) were also profiled using a similar approach.

2.3.1.2 Disease resistance genes

Similar to the search for sugar genes, for disease resistance, terms such as disease, senescence, -responsive, pathogen, and resistance were used in addition to a list of genes including chitinase, glucanase, and ethylene. The transcripts were filtered accordingly for all three samples, and further analyses were performed.

2.4 RNA seq analyses for expression profiling

For expression profiling, RNA Seq reads from sugarcane hybrid genotypes (Mason et al., 2022; Bioproject PRJNA317338) and PRJNA317338 were retrieved from NCBI. The module 'Expression Analysis using RNA-seq' in CLC-Genome Work Bench (CLC-GWB) version 22 was used. The abundance of each isoform (contributed by SP, BC and Co 11015) was estimated by alignment of the Illumina RNA-seq data of each sample to the three transcriptomes individually using the RNA-seq analysis function to have an understanding of the sub-genomic origin of transcripts related to sugar and disease resistance. The reads were aligned to the transcript reads using "one reference sequence per transcript" in the CLC-GWB's RNA-seq package. Normalised expression values were obtained as Reads Per Kilobase of transcript per Million mapped

reads (RPKM) and Transcripts Per Million reads (TPM) for further analyses. To identify differentially expressed transcripts, the differential expression (DE) for the RNA-seq data analysis function in CLC-GWB was used. The DE analysis in CLC-GWB uses multi-factorial statistics based on a negative binomial model (generalised linear model) that considers the various sequencing depths of each sample, facilitating the identification of differentially expressed genes (Mason et al., 2022). A differential gene expression (DGE) table containing the fold changes between samples based on Bonferroni and false discovery rate (FDR) corrected p-values were used for filtering the expression data.

3 Results

3.1 Iso seq sequencing of sugarcane progenitors and hybrid

Iso seq sequencing of the samples of *S. spontaneum* (SP), *S. officinarum* (BC), and Co 11015(11) was subjected to initial processing which included generating HiFi reads using the SMRT *ccs* tool. The number of genes and isoforms from each sample and other read statistics are given in the Table 1. The majority of the HiFi reads were 2kb to 4kb long. The quality of the 900,000 reads obtained was above Q50 on the Phred scale. The number of HiFi reads after demultiplexing were 679606 in SP, 1076156 in BC and 1268630 in 11. Using 'Isoseq3 refine', the high-quality full-length non-chimeric (FLNC) reads generated were found to represent > 95% of total HiFi reads (Table 2). These reads were clustered using 'Isoseq3 cluster' to create high-quality isoforms with a prediction accuracy of ≥ 0.99. Clusters with ≤ 0.99 prediction accuracy were considered low-quality and were excluded from analyses. Among the three genotypes used in the study, the highest number of clusters was found in BC followed by Co 11015. The high-quality FLNC reads from SP, BC, and 11 were 49908, 119662, and 92500 respectively. The number of splice sites identified in BC transcripts was more than in SP and 11.

3.2 Comparative genomics of sugarcane progenitor species and hybrid

The comparative analyses of transcriptomes revealed shared ancestry between *S. spontaneum*, *S. officinarum*, and the sugarcane hybrid Co 11015. The nobilization of sugarcane has harnessed the desirable agronomic and quality traits from both the genomes. Though the sugarcane hybrid Co 11015 and other commercial cultivars are derivatives of BC and SP, significant variation in phenotype and transcript diversity is observed. In our study, the transcriptome of sugarcane hybrid Co 11015 mapped up to 68.7% with *S. spontaneum* and 75% with *S. officinarum*. However, 79% of the Co 11015 transcriptome was mapped on the combined transcriptome of *S. spontaneum* and with *S. officinarum*. A total of 36,287 unique transcripts were found in the combined transcriptome of SP, BC, and 11. The greatest number of transcripts were represented in BC (73.6%) followed by 11 (57.3%) and SP (40.0%). There were 8541 common unique elements in BC, 11, and SP. As expected, the number of common unique elements in SP and 11 were less than the common unique elements found between 11 and BC (Figure 4). Comparative transcriptome analysis between the three

TABLE 1 Transcript category statistics based on reference genome.

Category	SP-Leaf-stem		11-Leaf-stem		BC-Leaf-Stem	
FSM	3841		5918		5927	
ISM	2539		3296		11262	
NIC	430		976		1362	
NNC	13268		23589		26692	
Genic Genomic	510		853		517	
Antisense	29		31		58	
Fusion	106		234		307	
Intergenic	95		91		121	
Genic Intron	0		0		1	
Splice Junction Classification						
	SP-Leaf-stem		11-Leaf-stem		BC-Leaf-Stem	
	SJs count	%	SJs count	%	SJs count	%
Known canonical	41960	77.14	49701	72.10	53848	73.74
Known Non-canonical	23	0.04	28	0.04	27	0.04
Novel canonical	1523	2.80	3170	4.60	4494	6.15
Novel Non-canonical	10892	20.02	16030	23.26	14651	20.00

TABLE 2 Details of HiFi reads and clusters obtained for the three genotypes.

Sample	HiFi reads	FLNC reads	% of FLNCs	Clusters of HQ FLNC reads	Unique Transcripts
SP	679,606	663,562	97.6	49,908	40584
11	1,268,630	1,214,852	95.8	92,500	72391
BC	1,076,156	1,069,379	99.4	119,662	99841

genotypes and the reference genome of *Saccharum* hybrid cvr SP80-3280 showed a similarity of 36.4%, 40.4%, and 34% with 11, BC, and SP, respectively. Mapping with the reference *S. spontaneum* genome assembly (accession number GCA_022457205.1_ASM2245720v1) was 39.8% for 11, 40.1% for BC, and 43.4% for SP. When *Saccharum* the hybrid R570 assembly MTP (accession number GCA_900465005.1) was used for mapping, a mapping percentage of 30.8% for 11, 31.5% for BC, and 29.5% for SP was recorded. With the *S. officinarum* LA Purple reference genome (GCA_020631735.1_ASM2063173v1), 51.9% of BC, 47.2% of 11, and 29.0% of SP mapped. The results of reference genomes and their mapping percentage with SP, BC, and 11 are given in the Table 3.

3.3 Transcript diversity in the hybrid and the progenitor genomes

The final Iso seq transcripts were annotated to assign gene function. The total number of transcripts in BC was more (119,662) than in 11 (92,500) and the least was in SP (49,908). The sequences were blasted in the NCBI database. Most of the transcripts matched *Miscanthus* spp. which is a genus related to *Saccharum*. The other transcripts found hits in 15 other genera, among them the most frequent hits were to *Sorghum* sequences in the database. Most of the hits came from C4 members of the Panicoideae sub-family (Supplementary Figure 1A–C). The annotated transcripts in SP, BC, and 11 were filtered using functional keywords such as "sucrose", "sugar", and "transporters". Likewise, 37 groups were identified using the filtering keywords that were related to important functions in the plant system. Variation in the number of transcripts related to a function were filtered and counted in the three genotypes. Overall, the most abundant transcripts were related to sugar, transporters, and pyruvate (Figure 5). Transcripts related to pyruvate and carboxylases were greater in SP than in BC and 11. In BC, the transcripts related to sucrose pathways in general, i.e., SPP, sucrose synthase, sugar SWEETs, and retrotransposons, were high. Co 11015 showed a higher number of transcripts related to uridine diphosphate (UDP); transcription

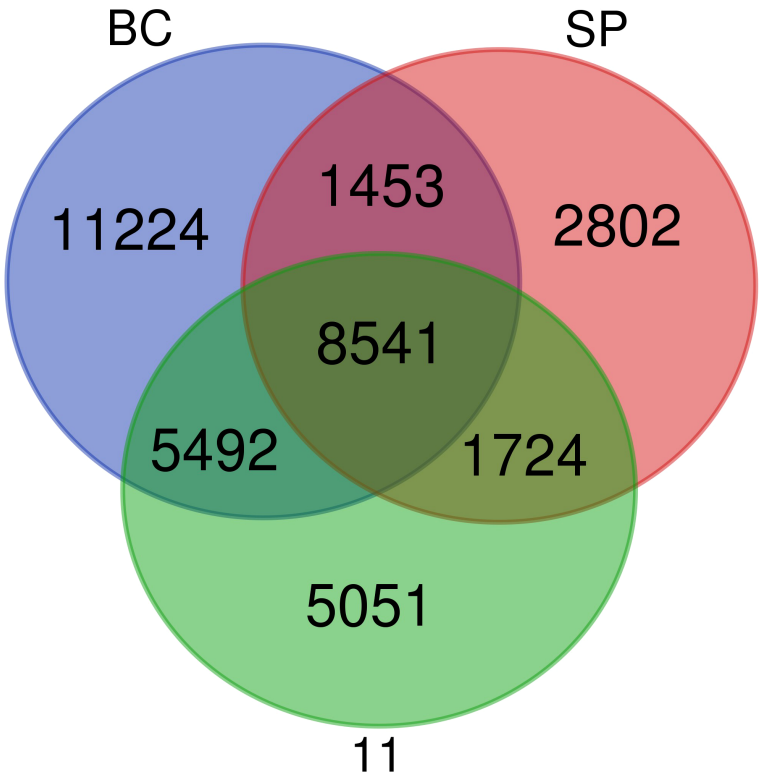


FIGURE 4
Comparative transcriptome analysis of the sugarcane progenitor species and the hybrid.

TABLE 3 Mapping results of SP, BC and 11 with different reference genomes.

S.No	Organism	Reference	Percentage mapping(%)		
			BC	SP	11
1.	*Sorghum bicolor* (cvr: BTx623)	GCA_000003195.3	26.32	28.96	26.61
2.	*S. spontaneum* (Isolate Np-X)	GCA_022457205.1	40.10	43.39	39.81
3.	*S. spontaneum* (AP85-441)	GCA_003544955.1	40.64	44.01	40.63
4.	*S. officinarum* (LA purple)	GCA_020631735.1	51.90	28.96	47.22
5.	Saccharum hybrid (cvr R570)	GCA_900465005.1	31.48	29.35	30.82
6.	Saccharum hybrid (cvr SP80-3280)	GCA_008692665.1	47.17	41.59	43.81
7.	Saccharum hybrid (cvr SP80-3280)	GCA_002018215.1	40.42	33.99	36.41
8.	Saccharum hybrid (cvr SP80-3280)	GCA_009173535.1	2.26	2.03	2.25
9.	Saccharum hybrid (cvr CC_01_1940)	GCA_020102875.1	43.71	40.62	42.21
10.	Sugarcane SUGIT transcriptome	GFH_J01000000	62.11	62.25	57.21

factors; invertases; other sugars such as xylose, trehalose, and galactose; and stress responsive genes related to DREB, heat, senescence, and dehydration. Lignin related transcripts were absent in BC while "mannose" related transcripts were more abundant in BC. Among the transcripts related to photosynthesis, 11 showed more abundance than SP and BC. Another interesting observation was that 11 had the greatest number of invertases along with the concomitant expression of invertase inhibitors transcripts whereas transcripts for invertase inhibitors were not found in BC or SP.

3.4 Analyses of sugar and disease resistance genes

There were 231, 1792, and 482 genes related to sugar and sucrose in SP, BC, and 11 respectively. Transcripts for other sugars such as trehalose, mannose, and xylose were also checked (Figure 6). The number of transcripts related to each component in the three transcriptomes is presented in Supplementary Tables 1A–C. The sugar genes were found to be higher in BC than in 11 and SP. Most of the sugar genes were related to transporters. The transcripts related to each of the genes/enzymes of the pathway are presented in Supplementary Table 1C.

3.4.1 Analysis of sucrose phosphate synthase genes from the three transcriptomes

The SPS genes were filtered from the three transcriptomes and the length distribution of the transcripts was visualized (Figure 7). The translated full length protein sequences were used for multiple sequence alignment, phylogeny, and motif distribution analysis (Figure 8; Supplementary Figures 2A–F). In total, there were 67, 113, and 69 SPS transcripts from 11, BC, and SP respectively. The total transcripts were further classified into four categories: A, B, II,

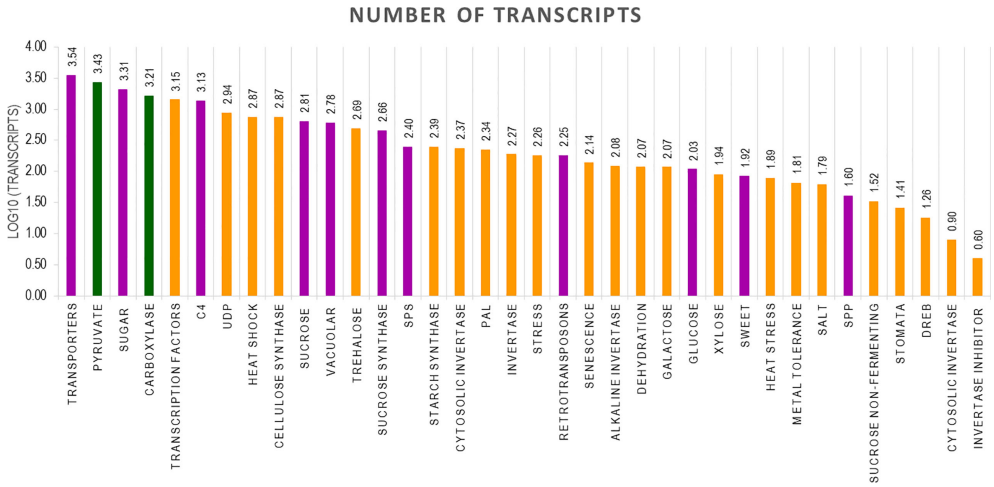


FIGURE 5
Number of transcripts in 37 groups. The highest number of transcripts in three genotypes are depicted in different colours. Green bar indicates SP, Purple bar indicates BC and Yellow bar 11.

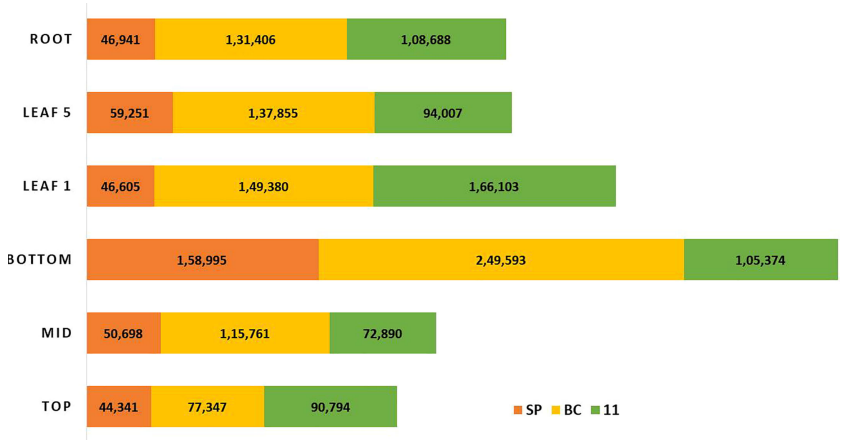


FIGURE 6
The overall expression pattern of transcripts originating from the three genotypes; red-*S. spontaneum*; yellow-*S. officinarum*; green-*Saccharum* hybrid in the sugarcane tissues, leaf 1 and 5, and the top, middle and bottom tissues of the culm and root.

and SPS. The composition of SPS transcripts from each genotype is shown in Figure 9. SP had a higher number of SPSB, whereas 11 had a higher number of SPSA while BC had all four categories almost equal in proportion. In the RNA seq expression analysis with SP and BC, SPSA was expressed in the middle tissues of the culm while with 11, there was no expression of SPSA transcripts. SPSB was expressed at high levels in the leaf tissues of all three transcriptomes (Supplementary Figures 3A–D). The transcript details for SPT and their expression profiling are shown in Supplementary Figure 3E.

3.5 RNA seq analysis using the hybrid and progenitor transcriptomes

RNA Seq reads derived from leaf 1, leaf 5, root, and culm samples from the top, middle, and bottom of sugarcane hybrid genotypes were used for profiling spatial expression bias from sub-genomes and the hybrid. The proportion of each tissue-expressed transcripts originating from the three transcriptomes and the results are presented in Figures 10 and 11. Similarly, RNA seq reads from stressed and control samples for water stress revealed that the higher expressions of stress related transcripts in the stressed samples were from SP and 11 compared to BC (Supplementary Figures 4A–E).

4 Discussion

Modern sugarcane hybrids are complex polyploids derived from inter-specific hybridization involving two progenitor species, *S. officinarum* and *S. spontaneum*. Sugar content and disease resistance were the characteristic traits for which the modern hybrids were selected over generations of breeding programs. Knowledge of the share of the progenitors in manifesting higher genetic gains in terms of these two traits would help in widening the gene pool further in developing future-ready cultivars. The long read transcriptomes of a modern hybrid, Co 11015, and its progenitors, *S. officinarum* Black Cheribon and *S. spontaneum* Coimbatore, were developed and dissected for sugar and disease resistance. *S. officinarum* Black Cheribon served as a common parent in almost all the sugarcane breeding programs and would probably be the most common ancestor for all the sugarcane cultivars being grown around the world. The proportional genome content of *S. officinarum* in the progenies seems to determine the sucrose synthesis and accumulation potential of the genotype. *S. spontaneum* clones are wild weedy plants, which are mainly non-cane forming types, and are therefore not used as immediate parents for commercial breeding purposes. Co 11015 is an early maturing variety developed at ICAR-Sugarcane Breeding Institute, Coimbatore. This genotype was a selection from the cross between high sucrose clones, CoC 671 and Co 86011, for which POJ 2725 is a common parent (Figure 1).

Iso seq sequencing of *S. spontaneum* (SP), *S. officinarum* (BC), and Co 11015 (11) resulted in 679606, 1076156, and 1268630 HiFi reads in SP, BC, and 11 respectively, mostly in the range of 2kb to 4kb in length. The FLNC reads represented more than 95% of the HiFi reads out of which only the high-quality isoforms were used for further analyses. There were 49908, 92500, and 119662 clustered high-quality (HQ) reads in SP, 11, and BC, respectively. Among the three genotypes used in the study, the highest number of clusters as well as splice junctions were found in BC, followed by 11, suggesting that BC has a more complex genome than the hybrid. The comparative analyses of transcriptomes revealed that sugarcane hybrid Co 11015 mapped up to 68.7% with *S. spontaneum* and 75% with *S. officinarum*. When the three transcriptomes were combined, 36,287 transcripts were found to be unique. BC had a major share (73.6%) of the unique transcripts followed by 11 (57.3%) and SP (40%), further suggesting its complexity. The unique transcripts in BC might have undergone the process of unconscious negative selection for those genes during the course of selection of the intermediate parents that ultimately gave rise to Co 11015. This might be one of the reasons for the low mapping percentages of the progenitor transcripts with the hybrid. The unique transcripts in the hybrid might have partly originated

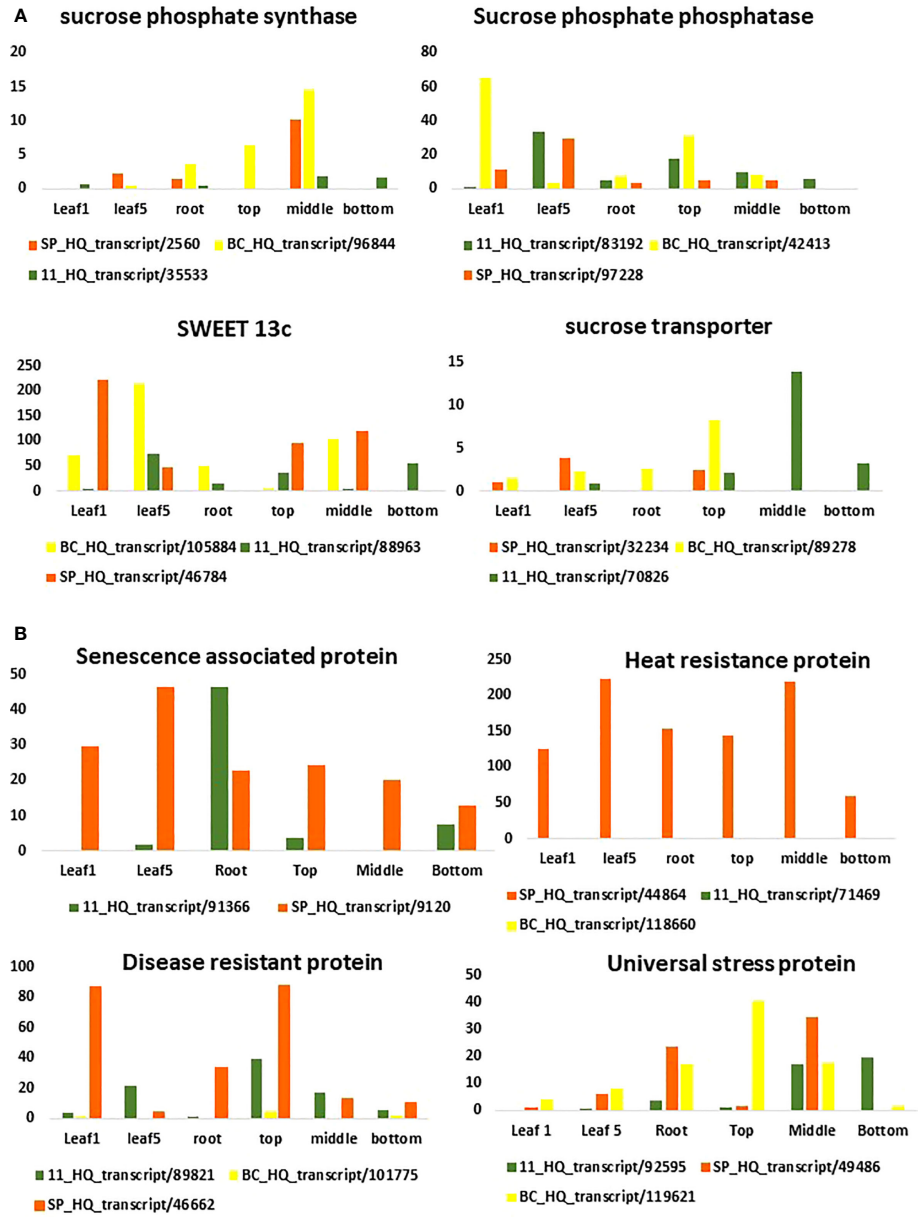


FIGURE 7
Expression pattern of sugar **(A)** and disease related genes **(B)** in different tissues using the three transcriptomes as references for RNA seq analysis.

from the other *S. officinarum* and *S. spontaneum* genotypes which were part of its ontogeny such as Loethers, Banjermassin Hitham, chunnee, and *S. spontaneum* Java. Comparative transcriptome analysis of the three transcriptomes with published reference genomes indicated a shared common ancestry with cultivars such as R570 and SP80-3280.

The hybrid's and the progenitor's transcriptomes were analyzed for overall gene expression in general and genes for the important traits sugar content and disease resistance, in particular. The transporters category was the largest in all the three genotypes, with BC having higher number of transcripts compared to 11 and SP. Second was the transcription factor (TF) category with 11 having the highest number of transcripts than BC and SP. Transcripts for trehalose, UDP, phenyl ammonia lysase, cellulose, heat, stress, senescence, starch, pyruvate, metal and salt tolerance, drought, invertases, and invertase inhibitor were higher in 11. In fact, the transcripts for trehalose are higher than the transcripts for sucrose in 11 (Supplementary Tables 1A–C). The large number of UDP and photosynthesis related transcripts in 11 suggests the availability of substrate for many cellular processes translates into higher sugar and biomass in the hybrids. Another interesting observation was that the retrotransposons were lower in 11 than BC, suggesting that the hybrid is less complex than the parental genome of BC which is also corroborated by the lower number of splice junctions and total number of unique transcripts observed in 11 compared to BC (Table 2). SP has a very low number of retrotransposons compared to BC and 11.

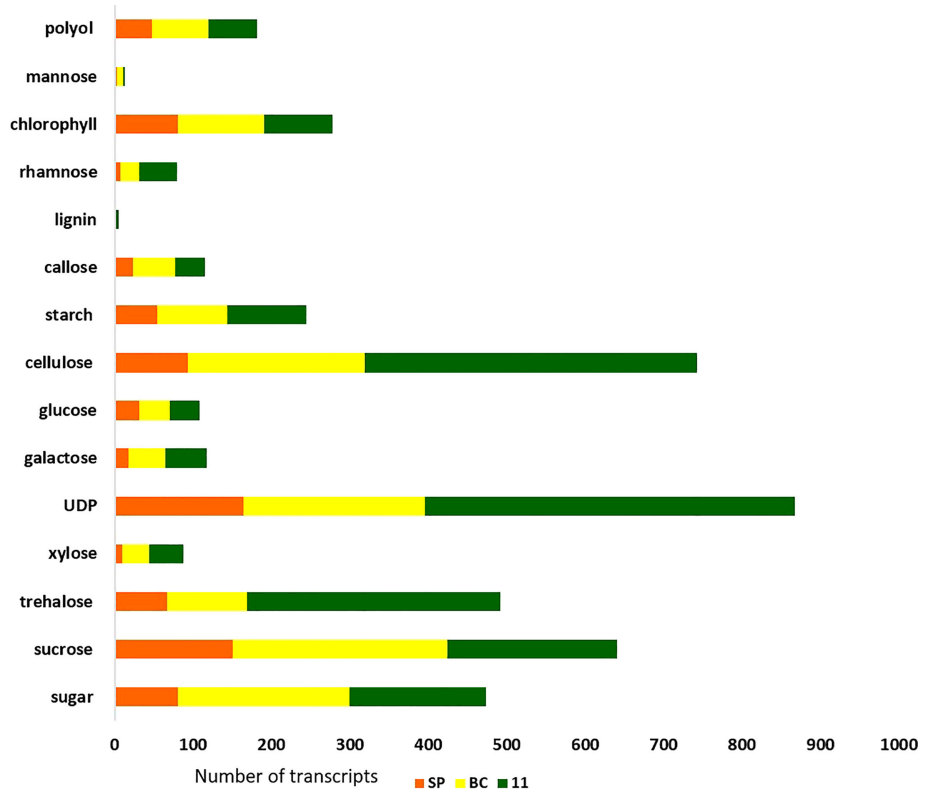


FIGURE 8
Sugar and other metabolite related transcript expression in the three genotypes: 11, SP, and BC.

For studying the gene expression pattern for sucrose, sucrose phosphate synthase (SPS) was selected for a detailed analysis. It is a key regulatory enzyme involved in sucrose biosynthesis. This enzyme catalyzes the transfer of a hexosyl group from UDP glucose to D-fructose 6-phosphate to form UDP and D-sucrose-6-phosphate. SPS is critical in the accumulation of sucrose because the reaction is irreversible. Sucrose synthase, on the other hand, involves a reversible reaction that allows sucrose to engage in a variety of metabolic activities, including tissue formation, material storage, and plant cell metabolism (Huber, 1983). SPS genes were categorized into three distinct families (A, B, and C) with different evolutionary histories in dicots (A family) and monocots (B family)

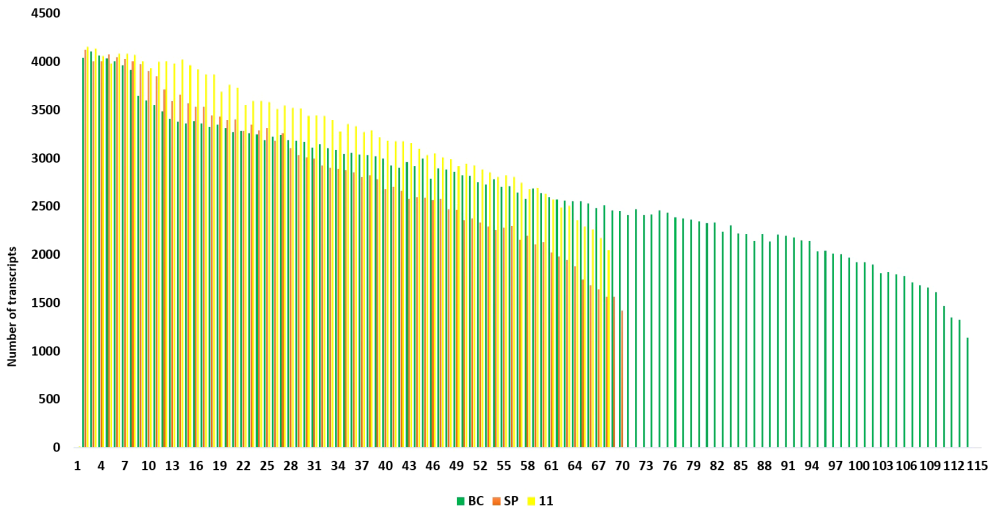


FIGURE 9
Transcript length distribution for sucrose phosphate synthase genes in *S. spontaneum* (red); *Saccharum* hybrid Co 11015, and *S. officinarum* (green).

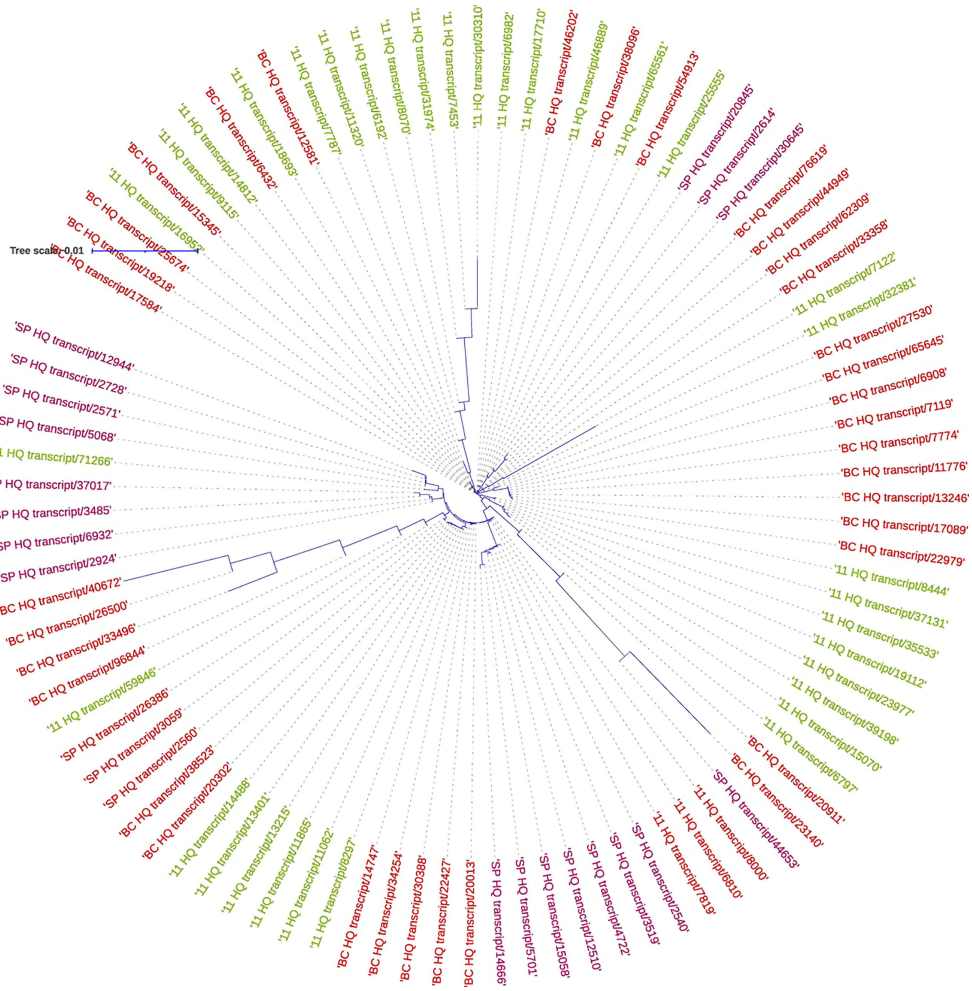


FIGURE 10
Phylogenetic analysis of sucrose phosphate synthase genes in *S. spontaneum* SP (purple), *Saccharum* hybrid Co 11015 (green), and *S. officinarum* BC (red).

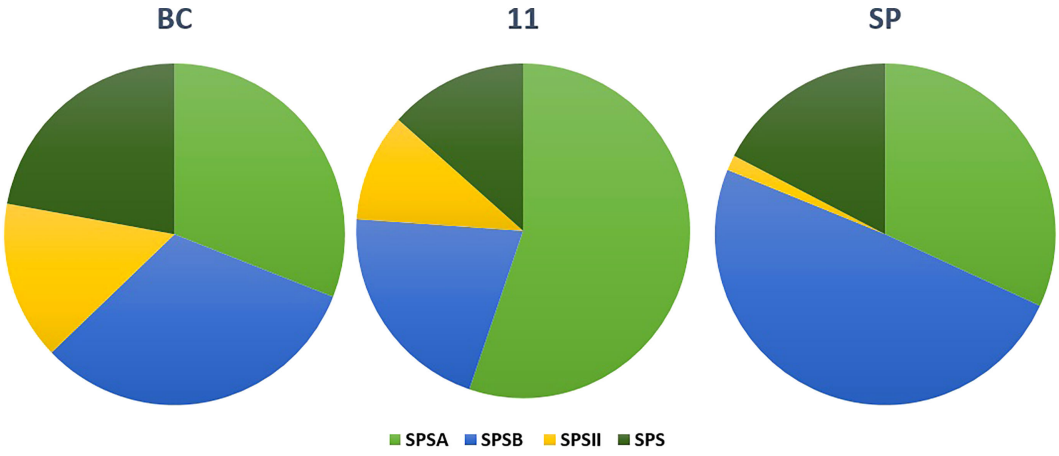


FIGURE 11
Composition of sucrose phosphate synthase (SPS) transcripts in the three genotypes: SP, BC, and 11.

(Langenkamper et al., 2002). Expression studies on *S. officinarum* and *S. spontaneum* have shown that *S. officinarum* had a higher expression of SPS A and SPS B than *S. spontaneum* (Ma et al., 2020). We observed a large number of SPS B in SP and SPS A in the sugarcane hybrid while BC had all the categories in equal proportions. The physiological relevance of these categories is not yet known and needs further study. However, it can be observed from the tissue specific expression study that SPS B is expressed in leaf tissues while SPS A shows higher expression levels in the culm tissues (Supplementary Figures 3A, B), suggesting SPS A could have an important role in stem sugar accumulation. In the case of disease resistance, the expression profiling of dehydrins, heat shock proteins, abscisic stress-ripening, aquaporin, senescence associated protein, etc. clearly show a higher contribution from SP than BC and 11 of hybrid genotypes under stress (Supplementary Figures 4A–E). A similar trend was observed in the transcriptomes with a higher number of transcripts in the stress, heat, and senescence categories in SP and 11 than BC. It must be noted that although the total number of transcripts in SP was only 49908 compared to 119662 (BC) and 92500(11), it showed an equal or higher number of transcripts for stress, senescence, heat, etc. as that of 11 and BC. (Supplementary Table 1A).

An interesting aspect of the inter-specific hybrids explored by previous studies is that *S. spontaneum* is a potential source of genes for sugar content. Studies on sugar composition from the *S. spontaneum* genotypes in the world collection in Miami, Florida, revealed positive alleles for sugar content (Tai and Miller, 2001). *S. spontaneum*-specific polymorphic markers for sugar content were identified and used for tagging positive *S. spontaneum* alleles for introgression into commercial sugarcane genotypes (da Silva et al., 2007). Although the sucrose accumulating potential of *S. spontaneum* accessions can hardly be estimated based on their performance *per se*, progeny performance can be taken as an indirect measure of the breeding value of the parent. This principle formed the basis for developing linkage maps and identifying the genomic regions governing sucrose content in *S. spontaneum* (Ming et al., 2001; Aitken et al., 2005; da Silva and Bressiani, 2005). However, these studies were limited by the number of molecular markers and the genotypes that were used in the experiments. In this study, SP was found to have a large number of transcripts for pyruvate carboxylase, sucrose transporters, trehalose phosphate phosphatase, acid invertase, and sucrose non-fermenting kinases, and, for some genes, there was similarity in the number of transcripts with 11 (Supplementary Figures 5A, B). *S. spontaneum* could also be speculated to be a source of sugar genes due to the presence of SPS B and a large number of transcripts related to pyruvate and trehalose, but this needs further study.

The transcriptomes from the progenitors described above and from the hybrid very clearly indicate the potential of such resources in understanding the gene regulation for important traits at the molecular level. It also suggests that the hybrid transcriptome of 11 has evolved its own genetic makeup apart from the mixture of genomes from the progenitors (Figure 8). From the expression profiling experiments, it is clearly evident that the genes for vigor and adaptation to various biotic and abiotic stresses in the hybrid might have been contributed by *S. spontaneum* while the genes for sugar and transporters were from *S. officinarum*. Studies on each and every gene set would be exhaustive, highly informational, and would provide us clues of the inheritance pattern of traits from the crosses. As observed from the transcripts, BC probably has a much more complicated genome structure than the hybrids, however, the hybrid seems to have a higher degree of sophistication in terms of transcription factors, biomass related genes, invertases, and the transcripts related to several sugars other than sucrose.

Unlike several other polyploid crops such as Brassica or wheat, sugarcane hybrids did not involve diploid progenitors. Diploid progenitors are often used as model systems for studying the gene expression bias/dominance and genomic changes in the formation of hybrids resulting from polyploidization. Extensive alterations occur during the merger of diverged polyploid genomes at each level of crossing, resulting in the formation of novel transcriptome networks. The extent of homolog expression bias changes over generations, from the initial sub-genome merger through to the incorporation of new genomes until the hybrid is selected. Novel transcripts, gene networks, and regulatory elements can emerge in the hybrids, however, the progenitors, which themselves are polyploids, may still hold important genes and perform similarly or equivalently to hybrids. There are no studies on the expression level dominance of sub-genomes in sugarcane hybrids due to limitations in the currently available genomics tools. However, to a certain extent, the direction of expression level dominance could be observed from the transcriptome of sugarcane hybrid Co 11015 in comparison with the transcriptomes of the progenitors.

5 Conclusion

The gene pool of sugarcane hybrids needs to be widened with more input from the valuable germplasm available from around the world to meet the fresh demands of agriculture today. We have developed transcriptome resources from progenitor species and a hybrid for a comparative study to look deeper into the parental materials for new perspectives in terms of their contribution to the hybrid and to improve the traits of interest in a more precise manner. The long reads offer a great advantage compared to short reads, particularly, as there is no assembly involved. This facilitates the identification of the exact isoform of a gene that may help in modifying a trait, for example, to increase the sugar content, and, in the future, the exact time of its expression. Further studies on more genotypes and their sequence information will provide a comprehensive understanding of the sugarcane genome complexity and gene regulation.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/ and https://doi.org/10.6084/m9.figshare.21974702. NCBI Sequence BioProject ID PRJNA479814, study Accession Number SRP152893 and PRJNA317338, study accession number SRP075950.

Ethics statement

Sugarcane commercial genotypes and germplasm collection were collected from the field planting at ICAR-Sugarcane Breeding Institute, Coimbatore, Tamil Nadu, India. No ethics approval was required for the conduct of experiments in this study.

Author contributions

PT conceived and designed the experiments. PT, MK, and LT collected the samples. PT, AS, SA, and SP conducted analyses. PT prepared the first draft. PT, RH, AF, AS, and HG critically revised the manuscript. All authors contributed to the article and approved the submitted version.

Funding

We gratefully acknowledge the financial support for the project on sequencing whole transcriptomes from *Saccharum* progenitor species and hybrid from ICAR, Government of India.

Acknowledgments

We gratefully acknowledge the help rendered by students Saranga and Sivasakthi Technical officer Rabisha during sample collection and Dr. Chandran, Head, ICAR-Sugarcane Breeding Institute, Research Centre, Kannur for the Black Cheribon samples.

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.2023.1199748/full#supplementary-material

References

Aitken, K. S., Jackson, P. A., and McIntyre, C. L. (2005). A combination of AFLP and SSR markers provides extensive map coverage and identification of homo (eo) logous linkage groups in a sugarcane cultivar. *Theor. Appl. Genet.* 110, 789–801. doi: 10.1007/s00122-004-1813-7

Cardoso-Silva, C. B., Costa, E. A., Mancini, M. C., Balsalobre, T. W. A., Canesin, L. E. C., Pinto, L. R., et al. (2014). *De novo* assembly and transcriptome analysis of contrasting sugarcane varieties. *PLoS One* 9, e88462. doi: 10.1371/journal.pone.0088462

Casu, R. E., Dimmock, C. M., Chapman, S. C., Grof, C. P. L., McIntyre, C. L., Bonnett, G. D., et al. (2004). Identification of differentially expressed transcripts from maturing stem of sugarcane by in silico analysis of stem expressed sequence tags and gene expression profiling. *Plant Mol. Biol.* 54, 503–517. doi: 10.1023/B:PLAN.0000038255.96128.41

Casu, R. E., Jarmey, J. M., Bonnett, G. D., and Manners, J. M. (2007). Identification of transcripts associated with cell wall metabolism and development in the stem of sugarcane by Affymetrix GeneChip Sugarcane Genome Array expression profiling. *Funct. Integr. Genomics* 7, 153–167. doi: 10.1007/s10142-006-0038-z

Cuadrado, A., Acevedo, R., Moreno Díaz de la Espina, S., Jouve, N., and de la Torre, C. (2004). Genome remodelling in three modern S. officinarum× S. spontaneum sugarcane cultivars. *J. Exp. Bot.* 55 (398), 847–854. doi: 10.1093/jxb/erh093

da Silva, J. A., and Bressiani, J. A. (2005). Sucrose synthase EST-derived RFLP marker associated to sugar content in elite sugarcane progeny. *Genet. Mol. Biol.* 28, 294–298. doi: 10.1590/S1415-47572005000200020

da Silva, J. A., Veremis, J., and Solis-Gracia, N. (2007). Saccharum spontaneum gene tagging by markers developed from sugarcane expressed sequence tags. *Subtropical Plant Sci.* 58, 6–14.

D'Hont, A. (2005). Unraveling the genome structure of polyploids using FISH and GISH; examples of sugarcane and banana. *Cytogenet. Genome Res.* 109, 27–33. doi: 10.1159/000082378

D'Hont, A., and Glaszmann, J. (2001). Sugarcane genome analysis with molecular markers, a first decade of research. *Proc. Int. Soc. Sugar Cane Technol.* 24, 556–559.

D'Hont, A., Grivet, L., Feldmann, P., Glaszmann, J. C., Rao, S., and Berding, N. (1996). Characterisation of the double genome structure of modern sugarcane cultivars (Saccharum spp.) by molecular cytogenetics. *Mol. Gen. Genet.* 250, 405–413. doi: 10.1007/BF02174028

Figueira, T. R. E. S., Okura, V., Rodrigues da Silva, F., Jose da Silva, M., Kudrna, D., Ammiraju, J. S., et al. (2012). A BAC library of the SP80-3280 sugarcane variety (saccharum sp.) and its inferred microsynteny with the sorghum genome. *BMC Res. Notes* 5, 1–11. doi: 10.1186/1756-0500-5-185

Garsmeur, O., Droc, G., Antonise, R., Grimwood, J., Potier, B., Aitken, K., et al. (2018). A mosaic monoploid reference sequence for the highly complex genome of sugarcane. *Nat. Commun.* 9 (1), 2638. doi: 10.1038/s41467-018-05051-5

Hemaprabha, G., Appunu, C., Mohanraj, K., Durai, A. A., Alarmelu, S., Sreenivasa, V., et al. (2019). Co 11015 (Atulya): a recently notified sugarcane variety for Tamil Nadu. *J. Sugarcane Res.* 9 (2), 193–195. doi: 10.37580/JSR.2019.2.9.193-195

Hoang, N. V., Furtado, A., Mason, P. J., Marquardt, A., Kasirajan, L., Thirugnanasambandam, P. P., et al. (2017). A survey of the complex transcriptome from the highly polyploid sugarcane genome using full-length isoform sequencing and *de novo* assembly from short read sequencing. *BMC Genomics* 18 (1), 1–22. doi: 10.1186/S12864-017-3757-8

Hoarau, J. Y., Offmann, B., D'Hont, A., Risterucci, A. M., Roques, D., Glaszmann, J. C., et al. (2001). Genetic dissection of a modern sugarcane cultivar (Saccharum spp.). I. Genome mapping with AFLP markers. *Theor. Appl. Genet.* 103 (1), 84–97. doi: 10.1007/s001220000390

Hotta, C. T., Lembke, C. G., Domingues, D. S., Ochoa, E. A., Cruz, G. M. Q., Melotto-Passarin, D. M., et al. (2010). The biotechnology roadmap for sugarcane improvement. *Trop. Plant Biol.* 3 (2), 75–87. doi: 10.1007/S12042-010-9050-5

Huber, S. C. (1983). Role of sucrose-phosphate synthase in partitioning of carbon in leaves. *Plant Physiol.* 71 (4), 818–821. doi: 10.1104/pp.71.4.818

Hussin, S. H., Liu, X., Li, C., Diaby, M., Jatoi, G. H., Ahmed, R., et al. (2022). An Updated Overview on Insights into Sugarcane Genome Editing *via* CRISPR/Cas9 for Sustainable Production. *Sustainability* 14 (19), 12285. doi: 10.3390/su141912285

Jackson, P., Hale, A., Bonnett, G., and Lakshmanan, P. (2014). Sugarcane. In: Pratap, A., and Kumar, J. (eds) *Alien Gene Transfer in Crop Plants*. New York, NY: Springer. 2, 317–345. doi: 10.1007/978-1-4614-9572-7_14

Kannan, B., Jung, J. H., Moxley, G. W., Lee, S. M., and Altpeter, F. (2018). TALEN-mediated targeted mutagenesis of more than 100 COMT copies/alleles in highly polyploid sugarcane improves saccharification efficiency without compromising biomass yield. *Plant Biotechnol. J.* 16 (4), 856–866. doi: 10.1111/pbi.12833

Langenkamper, G., Fung, R. W., Newcomb, R. D., Atkinson, R. G., Gardner, R. C., and MacRae, E. A. (2002). Sucrose phosphate synthase genes in plants belong to three different families. *J. Mol. Evol.* 54 (3), 322–332. doi: 10.1007/s00239-001-0047-4

Le Cunff, L., Garsmeur, O., Raboin, L. M., Pauquet, J., Telismart, H., Selvi, A., et al. (2008). Diploid/polyploid syntenic shuttle mapping and haplotype-specific chromosome walking toward a rust resistance gene (Bru1) in highly polyploid sugarcane (2n ∼ 12x ∼ 115). *Genetics* 180 (1), 649–660. doi: 10.1534/genetics.108.091355

Ma, P., Zhang, X., Chen, L., Zhao, Q., Zhang, Q., Hua, X., et al. (2020). Comparative analysis of sucrose phosphate synthase (SPS) gene family between Saccharum officinarum and Saccharum spontaneum. *BMC Plant Biol.* 20 (1), 422. doi: 10.1186/s12870-020-02599-7

Mason, P. J., Hoang, N. V., Botha, F. C., Furtado, A., Marquardt, A., and Henry, R. J. (2022). Comparison of the root, leaf and internode transcriptomes in sugarcane (Saccharum spp. hybrids). *Curr. Res. Biotechnol.* 4, 167–178. doi: 10.1016/j.crbiot.2022.02.005

Matsuoka, S., Kennedy, A. J., Santos, E. G. D., Tomazela, A. L., and Rubio, L. C. S. (2014). Energy cane: its concept, development, characteristics, and prospects. *Adv. Bot.* 2014, 1–13. doi: 10.1155/2014/597275

Ming, R., Liu, S.-C., Moore, P. H., Irvine, J. E., and Paterson, A. H. (2001). QTL analysis in a complex autopolyploid: genetic control of sugar content in sugarcane. *Genome Res.* 11, 2075–2084. doi: 10.1101/gr.198801

Parajuli, S., Kannan, B., Karan, R., Sanahuja, G., Liu, H., Garcia-Ruiz, E., et al. (2020). Towards oilcane: Engineering hyperaccumulation of triacylglycerol into sugarcane stems. *GCB Bioenergy* 12, 476–490. doi: 10.1111/gcbb.12684

Park, J.-W., Benatti, T. R., Marconi, T., Yu, Q., Solis-Gracia, N., Mora, V., et al. (2015). Cold responsive gene expression profiling of sugarcane and Saccharum spontaneum with functional analysis of a cold inducible Saccharum homolog of NOD26-like intrinsic protein to salt and water stress. *PLoS One* 10, e0125810. doi: 10.1371/journal.pone.0125810

Piperidis, G., and D'Hont, A. (2001). Chromosome composition analysis of various Saccharum interspecific hybrids by genomic *in situ* hybridisation (GISH). *Int. Soc. Sugar Cane Technol. Congress* 11, 565–566.

Pompidor, N., Charron, C., Hervouet, C., Bocs, S., Droc, G., Rivallan, R., et al. (2021). Three founding ancestral genomes involved in the origin of sugarcane. *Ann. Bot.* 127 (6), 827–840. doi: 10.1093/aob/mcab008

Premachandran, M. N., Prathima, P. T., and Lekshmi, M. (2011). Sugarcane and polyploidy: a review. *J. Sugarcane Res.* 1 (2), 1–15.

Stevenson, G. C. (1965). *Genetics and Breeding of Sugar Cane* (London, UK: Longman).

Tai, P. Y. P., and Miller, J. D. (2001). A core collection for Saccharum spontaneum L. from the world collection of sugarcane. *Crop Sci.* 41 (3), 879–885. doi: 10.2135/cropsci2001.413879x

Thirugnanasambandam, P. P., Hoang, N. V., and Henry, R. J. (2018). The challenge of analyzing the sugarcane genome. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.00616

Vettore, A. L., da Silva, F. R., Kemper, E. L., Souza, G. M., da Silva, A. M., Ferro, M. I. T., et al. (2003). Analysis and functional annotation of an expressed sequence tag collection for tropical crop sugarcane. *Genome Res.* 13, 2725–2735. doi: 10.1101/GR.1532103

Vettore, A. L., Silva, F. R., Kemper, E. L., and Arruda, P. (2001). The libraries that made SUCEST. *Genet. Mol. Biol.* 24, 1–7. doi: 10.1590/S1415-47572001000100002

Wang, K., Cheng, H., Han, J., Esh, A., Liu, J., Zhang, Y., et al. (2022). A comprehensive molecular cytogenetic analysis of the genome architecture in modern sugarcane cultivars. *Chromosome Res.* 30, 29–41. doi: 10.1007/s10577-021-09680-3

Zale, J., Jung, J. H., Kim, J. Y., Pathak, B., Karan, R., Liu, H., et al. (2016). Metabolic engineering of sugarcane to accumulate energy-dense triacylglycerols in vegetative biomass. *Plant Biotechnol. J.* 14 (2), 661–669. doi: 10.1111/pbi.12411

Zhang, Q., Qi, Y., Pan, H., Tang, H., Wang, G., Hua, X., et al. (2022). Genomic insights into the recent chromosome reduction of autopolyploid sugarcane Saccharum spontaneum. *Nat. Genet.* 54, 88. doi: 10.1038/s41588-022-01084-1

Zhang, J., Zhang, X., Tang, H., Zhang, Q., Hua, X., Ma, X., et al. (2018). Allele-defined genome of the autopolyploid sugarcane Saccharum spontaneum L. *Nat. Genet.* 50 (11), 1565–1573. doi: 10.1038/s41588-018-0237-2

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 06 September 2023
DOI 10.3389/fpls.2023.1233996

OPEN ACCESS

EDITED BY
Umesh K. Reddy,
West Virginia State University, United States

REVIEWED BY
Javed Akhatar,
Punjab Agricultural University, India
Sareena Sahab,
Department of Economic Development Jobs Transport and Resources, Australia

*CORRESPONDENCE
Harsh Raman
harsh.raman@dpi.nsw.gov.au

†These authors have contributed equally to this work

RECEIVED 03 June 2023
ACCEPTED 31 July 2023
PUBLISHED 06 September 2023

CITATION
Raman H, Raman R, Sharma N, Cui X, McVittie B, Qiu Y, Zhang Y, Hu Q, Liu S and Gororo N (2023) Novel quantitative trait loci from an interspecific *Brassica rapa* derivative improve pod shatter resistance in *Brassica napus*.
Front. Plant Sci. 14:1233996.
doi: 10.3389/fpls.2023.1233996

Novel quantitative trait loci from an interspecific *Brassica rapa* derivative improve pod shatter resistance in *Brassica napus*

Harsh Raman[1*], Rosy Raman[1], Niharika Sharma[2†], Xiaobo Cui[3†], Brett McVittie[1†], Yu Qiu[1†], Yuanyuan Zhang[3], Qiong Hu[3], Shengyi Liu[3] and Nelson Gororo[4]

[1]New South Wales (NSW) Department of Primary Industries, Wagga Wagga Agricultural Institute, Wagga Wagga, NSW, Australia, [2]New South Wales (NSW) Department of Primary Industries, Orange Agricultural Institute, Orange, NSW, Australia, [3]Oil Crops Research Institute, Chinese Academy of Agricultural Sciences, Wuhan, Hubei, China, [4]Nuseed Pty Ltd, Horsham, VIC, Australia

Pod shatter is a trait of agricultural relevance that ensures plants dehisce seeds in their native environment and has been subjected to domestication and selection for non-shattering types in several broadacre crops. However, pod shattering causes a significant yield reduction in canola (*Brassica napus* L.) crops. An interspecific breeding line BC95042 derived from a *B. rapa/B. napus* cross showed improved pod shatter resistance (up to 12-fold than a shatter-prone *B. napus* variety). To uncover the genetic basis and improve pod shatter resistance in new varieties, we analysed F_2 and $F_{2:3}$ derived populations from the cross between BC95042 and an advanced breeding line, BC95041, and genotyped with 15,498 DArTseq markers. Through genome scan, interval and inclusive composite interval mapping analyses, we identified seven quantitative trait loci (QTLs) associated with pod rupture energy, a measure for pod shatter resistance or pod strength, and they locate on A02, A03, A05, A09 and C01 chromosomes. Both parental lines contributed alleles for pod shatter resistance. We identified five pairs of significant epistatic QTLs for additive x additive, additive dominance and dominance x dominance interactions between A01/C01, A03/A07, A07/C03, A03/C03, and C01/C02 chromosomes for rupture energy. QTL effects on A03/A07 and A01/C01 were in the repulsion phase. Comparative mapping identified several candidate genes (*AG*, *ABI3*, *ARF3*, *BP1*, *CEL6*, *FIL*, *FUL*, *GA2OX2*, *IND*, *LATE*, *LEUNIG*, *MAGL15*, *RPL*, *QRT2*, *RGA*, *SPT* and *TCP10*) underlying main QTL and epistatic QTL interactions for pod shatter resistance. Three QTLs detected on A02, A03, and A09 were near the *FUL (FRUITFULL)* homologues *BnaA03g39820D* and *BnaA09g05500D*. Focusing on the *FUL*, we investigated putative motifs, sequence variants and the evolutionary rate of its homologues in 373 resequenced *B. napus* accessions of interest. *BnaA09g05500D* is subjected

to purifying selection as it had a low Ka/Ks ratio compared to other *FUL* homologues in *B. napus*. This study provides a valuable resource for genetic improvement for yield through an understanding of the genetic mechanism controlling pod shatter resistance in *Brassica* species.

KEYWORDS

pod shattering, domestication, genetic mapping, canola, genetic analysis, sequence variation

1 Introduction

Plants have evolved vivid mechanisms for survival and fitness across various ecological niches. In the wild, plants dehisce their fruits and disperse seeds to ensure the multiplication and adaptation of their progenies and confront challenges posed by climatic and ecological vagaries. Seeds of the *Brassicaceae* family members are enclosed in a silique (pod), which consists of two congenitally fused carpels (valves); each is separated with a thin layer called a pseudo-septum or replum (Figure 1) (Bowman et al., 1999). Both valves and replum are differentiated with valve margins where pod dehiscence and seed abscission occur via pod drop and seed shattering, possibly by similar molecular mechanisms (Balanzà et al., 2016). Pod drop – a phenomenon where a whole fruit (silique) drops on the ground, is a common problem in some canola production regions, particularly Canada. As the pod matures physiologically, valves detach from the replum, resulting in pod dehiscence (Figure S1A) and the seeds attached to the replum with a funiculus fall to the ground (Figure S1C). Pod dehiscence occurs via the dehiscence zone formation at the valve margins by two layers: a lignification layer of 1-2 thick and rigid cells and the separation (also called abscission) layer of iso-diametrically shaped cells, separating the valve from the replum (Spence et al., 1996; Rajani and Sundaresan, 2001; Dinneny and Yanofsky, 2005). At maturity, cells in the separation layer degrade by polygalacturonase, cellulase, and mannanase enzymes (Ogawa et al., 2009). Shattering occurs when the abscission force becomes more significant than the binding force of the pod valve (Lee et al., 2017). External influences such as wind velocity, machinery, and high temperatures further escalate pod shattering in brassicas.

Molecular mechanisms underlying pod dehiscence are well-dissected in a model plant, *Arabidopsis thaliana* - a distant relative of *Brassica napus* L. At least thirteen genes that are responsible for pod dehiscence in Arabidopsis have been identified, such as MADS-box genes: *SHATTERPROOF1* (*SHP1), SHATTERPROOF2 (SHP2*) and *FRUITFULL* (*FUL*); Basic-loop-helix genes: *INDEHISCENT* (*IND*), *ALCATRAZ* (*ALC*) and *SPATULA (SPT); REPLUMLESS* (*RPL*) and *APETALA2* (*AP2*), *ARABIDOPSIS DEHISCENCE ZONE POLYGALACTUROSE1* (*ADPG1*), ADPG2, a C2H2 zinc finger transcription factors *JAGGED* (*JAG*) and BnLATE FLOWERING (*BnLATE)*; NAC SECONDARY WALL THICKENING PROMOTING FACTOR1 (*NST1*), ENDO-BETA-*MANNANASE7* (*MAN7*), and *CELLULASE6* (Ferrandiz et al., 2000; Liiljegren et al., 2000; Rajani and Sundaresan, 2001; Roeder et al., 2003; Sorefan et al., 2009; He et al., 2018; Li et al., 2021). Different genes involved in auxin, gibberellin and cytokinin biosynthesis also regulate pod development and dehiscence (Sorefan et al., 2009; Arnaud et al., 2010; Marsch-Martínez et al., 2012).

Canola, the second most crucial oilseed crop after soybean, contributes about 13-16% of global vegetable oil production. The allotetraploid canola genome ($2n$ = 4× = 38, genome AACC) originated about 7,500 years ago via ancient hybridisation events between two diploid progenitors *Brassica* species, *B. rapa* ($2n$ = 2× = 20, AA genome) and *B. oleracea* ($2n$ = 2× = 18, CC genome) (Chalhoub et al., 2014; Lu et al., 2019). However, seed shattering (commonly referred to as pod-shattering) is a universal constraint in canola production, and in the literature, none of the domesticated accessions of *B. napus* is reported to be 'completely' resistant to pod shattering. Generally, canola pods are highly sensitive to pre-mature shattering, significantly reducing yield. The seed loss varies from 8 to 70% across environments depending on genotypic attributes (canopy architecture, resistance to lodging and diseases), method of harvesting (windrow/direct heading), and time of harvesting (early, optimal time vs late) and environmental conditions at the time of harvest (MacLeod, 1981; Price et al., 1996; Child et al., 1998; Vera et al., 2007; de la Pasture, 2018). Shattered seeds grow in the field at a much higher rate (60x) than those sowed initially (Figure S1) and become a weed in the next crop; hence must be controlled (Wang et al., 2007).

To overcome pod-shattering, the majority of broadacre canola varieties are harvested by windrowing/swathing - a practice of cutting plants at physiological maturity (50 to 60% seed colour change from green to dark brown, red or black) and leaving them in the field before threshing with a combine harvester. This practice can also lead to significant losses from seed shattering, mainly when not accomplished at the 'right' time. The window for windrowing is often small and subjected to labour and combined harvester availability and congenial weather conditions. High temperatures, high-velocity winds, rainfall, and hailstorm events significantly impact canola seed yield and oil content. High yield is essential for meeting global demands for healthy vegetable oil, protein for animal feed, and canola growers for return on their investment.

Understanding the genetic determinants and novel alleles underlying this domestication trait would provide an improved

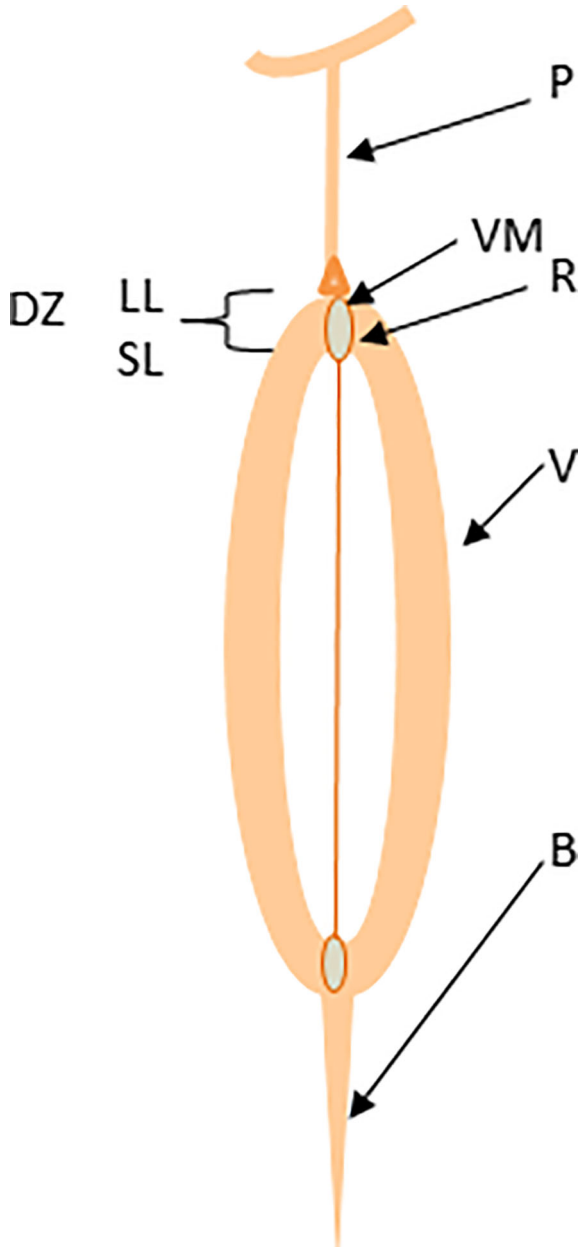


FIGURE 1
Schematic representation of canola pod structure. The valve (V), dehiscence zone (DZ), separation layer (SL), lignification layer (LL), valve margin (VM), replum (R), and beak (B).

genetics-based solution to reduce yield loss in *B. napus*. The functionality of some of Arabidopsis pod dehiscence genes has also been demonstrated in *Brassica* species via overexpression, RNAi, gene editing, and induced mutation studies (Ostergaard et al., 2006; Kord et al., 2015; Lawrenson et al., 2015; Braatz et al., 2018a; Braatz et al., 2018b; Stephenson et al., 2019; Li et al., 2021). Recently, it has also been shown that miR319-targeted *TEOSINTE BRANCHED 1*, *CYCLOIDEA*, and *PROFEERATIN CELL NUCLEAR ANTIGEN BINDING FACTOR* (*TCPs*) inhibit pod elongation and dehiscence via regulation of *FUL* expression in *A. thaliana* and *B. napus* (Cao et al., 2022). Although the network of pod dehiscence genes has been investigated in Arabidopsis, their expression level has not been fine-tuned in commercial canola varieties with genetic modification approaches, except in POD GURAD varieties where TILLING has been deployed only in the BASF canola breeding program (Laga et al., 2008). In fact, ectopic (over-) expression of *FUL* and *SHP* genes led to indehiscent pods due to the non-lignification of cells between the valve and replum and the absence of dehiscence zone formation (Ferrandiz et al., 2000; Liiljegren et al., 2000; Ostergaard et al., 2006).

Previous research has shown a limited range of genetic variation for pod shatter resistance in *B. napus* (Morgan et al., 2007; Raman et al., 2014). However, a wide range of genetic variation for pod shattering is observed in diploid and amphidiploid species of *Brassica*, such as *B. rapa, B. juncea* ($2n = 4\times = 36$, AABB), and *B. carinata* ($2n = 4\times = 34$, BBCC) (Kadkol et al., 1984; Kadkol et al., 1985; Raman et al., 2017). In a previous study, Raman et al. (2014) reported that pod shatter resistance could improve up to 12-fold in a shatter-prone variety of *B. napus* via the introgression of resistant alleles from *B. rapa*. To uncover the genetic basis underlying seed shattering in this interspecific source, we investigated an F_2 mapping population and its $F_{2:3}$ progenies derived from a cross between *B. napus* (BC95041) and *B. rapa*/*B. napus* (BC95042). We further identified epistatic quantitative trait loci (QTLs) for additive × additive, additive dominance, and dominance × dominance interactions. Candidate genes and their sequence variants in parental lines underlying QTL regions for pod shatter resistance were identified, which could regulate variation in pod shatter resistance.

2 Materials and methods

2.1 Construction of mapping population

An interspecific line derived *B. rapa*/*B. napus* with the highest pod rupture energy (RE), BC95042 (shatter resistant with high RE (Raman et al., 2014)) was crossed with the advanced breeding lines of *B. napus*, BLN3303 (BC95041, maternal parent, shatter prone with low RE). This study utilised an F_2 population comprising 203 individuals generated from the self-pollination of a single F_1 cross from BC95041/BC94042. Each F_2 line was selfed to generate an $F_{2:3}$ population for confirming phenotypes.

2.2 Evaluation for pod shatter resistance

The two parental lines and their F_2 population of 203 plants were grown in 2021 in white plastic pots (Garden City Plastics, NSW, Australia)) under birdcage conditions at the Wagga Wagga Agricultural Institute, New South Wales, Australia. The cultivation of canola plants followed standard management practices. Plants were watered thrice per week, fertilised weekly using in-line liquid fertilisers, and protected from blackleg and sclerotinia diseases by applications of Prosaro® 420 SC and Aviator fungicides (Bayer Crop Sciences, Australia) and aphids using chemicals recommended in Australia. Day to flowering was recorded daily for each F_2 plant. To avoid outcrossing and get pure F_3 progenies, all

F_2 plants were bagged with perforated pollination bags before flower initiation, leaving the primary stem out for the natural pod development for shatter testing. Ten pods were collected from each line at maturity (BBCH scale 95) in the 50 mL plastic tubes containing a silica sachet, as detailed in our previous study (Raman et al., 2014). Pods were desiccated in a dehydrator (G. T. D. Pty. Ltd., Australia) at 40°C for 48 hours to reduce variation due to moisture content and further tested for variation in pod rupture energy. For validation, 40 $F_{2:3}$ families (20 high rupture energy and 20 low rupture energy) and parents were grown in pots in 2016 under birdcage conditions and tested with a pendulum test described earlier (Raman et al., 2014). The phenotypic means for each genotype were used for further genetic analysis. A pair-wise correlation between rupture energy and pod length in F_2 and $F_{2:3}$ populations was calculated. The rupture energy of five pods of each F_2 plant was averaged and used for QTL analysis.

2.3 DNA isolation and genotyping

Young leaf tissue of the field-grown plants was collected from each line in a 96-well format. The tissue was frozen immediately and kept at - 80°C until used for DNA isolation. Tissue was ground in liquid nitrogen and extracted for DNA using a method described by Raman et al. (2005). DNA concentration was determined by a Qubit fluorometer and Qubit dsDNA broad-range assay kit according to the manufacturer's recommendation. DNA quality was checked on the Tris-Acetate-EDTA buffered 0.8% agarose gel. The F_2 population and parental lines were genotyped with the genotyping-by-sequencing-based DArTseq marker approach (Raman et al., 2014) using the HiSeq 2500 system (Illumina, USA) at the DArT P/L, University of Canberra, Bruce, Australia. We considered only high-quality DArTseq markers, which included SNPs (single nucleotide polymorphism) and *in-sillco* presence-absence markers, having BLAST alignments (E-value: $5e^{-5}$) and minimum sequence identity of 90% with the reference *B. napus* cv. Darmor-*bzh* v 4.1.

2.4 Map construction and QTL identification for pod shatter resistance

The linkage map of the F_2 population was constructed using DArT P/L's OCD MAPPING program (Petroli et al., 2012), as described previously (Raman et al., 2017). The association between markers and rupture energy was tested using linear marker regression, Fisher's exact test, and the X^2 test. We applied the additive, dominant and recessive models and full scan permutation with 1000 iterations for the genome scan. Haplotype blocks (HB) were detected using 0.98 upper confidence and 0.7 lower bound recombination value at threshold 0.01, Expectation maximization algorithm (EM) iteration 1,000 and EM convergence tolerance value of 0.00010 (Gabriel et al., 2002). *P* values for haplotyping association test were determined using 10,000 iterated permutations in the SVS package (Golden Helix, Bozeman, USA). We used binary data of contrasting 141 F_2 phenotypes for resistance or sensitivity to shattering (Table S5a) for haplotype analysis. Manhattan plots were generated in the SVS package (Golden Helix, Bozeman, USA).

QTL mapping was performed by single interval mapping (IM), inclusive composite interval mapping (ICIM-ADD) of additive and dominant QTL, and inclusive composite interval mapping of epistatic QTL (ICIM-EPI) functions implemented in the QTL IciMapping v4.1 (www.isbreeding.net). The threshold logarithm of odds (LOD) value was determined by a permutation test involving 1,000 runs at a significance level of $P = 0.05$. Threshold *P* values for ICIM and IM for rupture energy were 3.07 and 3.25, respectively. While for pod length, threshold *P* values for ICIM and IM are 2.66 and 1.78, respectively. QTLs having LOD values more than the estimated threshold were declared as significant. LOD score greater than 2.5 but less than estimated threshold *P* values were termed suggestive QTL. The phenotypic variance explained (% PVE) and the additive effects of QTLs were directly derived from the QTL analysis outputs files. For digenic epistatic QTL interactions, LOD threshold values for each trait were estimated after 1,000 permutations using a type I error = 0.05. Epistatic effect QTLs were analysed using ICIM-EPI at the threshold LOD 4.87. Favorable parental alleles that enhance the trait expression were identified using an additive effect's direction (+ and -ve).

2.5 Alignment of markers with the Brassica reference genomes

The physical map positions of significant markers associated with pod shatter resistance were obtained using the reference *B. napus cv* Darmor-*bzh* genome by BlastN (Altschul et al., 1990) searches, as detailed in Raman et al. (2014). We also used the BnaOmics platform (https://bnaomics.ocri-genomics.net/) that integrates pan-genome and multi-omics data of *B. napus* (Cui et al., 2023) to search candidate genes. The only single top hit with the cutt-of E value of $1E^{-5}$ was considered for identifying syntenic region underlying candidate genes. *B. napus* annotated genes which were mapped within the marker intervals with ICIM/ICIM-EPI, were assumed candidate genes. The candidates that map within 500 kb from the significant markers identified with genome scan approaches were also identified. Genes involved in the pod shatter trait of *Arabidopsis* (Table S14) were used to search the corresponding copies in *B. napus*, with an e-value of $1e^{-10}$.

2.6 Identifying *FUL* homologues in *B. napus* based on homology to *ATFUL* (AT5G60910)

Arabidopsis thaliana genic and protein sequences of AT5G60910 from the Arabidopsis Information Resource (TAIR) were used to search the homologues in *B. napus* using TBLASTN and BLASTP (*B. napus* cv. Darmor-*bzh* genome, versions 4.1; http://www.genoscope.cns.fr, and the pan-genome) (Cui et al., 2023).

2.7 Phylogenetic relationship and Ka/Ks ratios

We used the Geneious tree builder pipeline to generate a Neighbour-Joining phylogenetic tree of DNA sequences from *B. rapa*, *B. oleracea* and *B. napus* for *FUL* (Figure 2) and *FUL-Like* genes (Figure S5). Sequences were aligned with global alignment with free end gaps, Blosum62 cost matrix, and Jukes-Cantor genetic distance model, implemented in the Geneious prime package (https://www.geneious.com). *A. thaliana FUL* gene was used as an outgroup to verify functional divergence. The synonymous substitution rate (Ks), non-synonymous substitution rate (Ka), and Ka/Ks ratio were calculated with SNPGenie (https://github.com/chasewnelson/SNPGenie).

2.8 Gene structure and motif conserved domains and *cis*-acting elements identification of *FUL* homologues

The intron-exon distribution of *FUL* genes was obtained from genome annotation files from the online resources described above and confirmed using sequence analysis with *AtFUL*. Multiple sequence alignment of protein sequences was performed with ClustalX 2.0 (http://www.custal.org/clustal2/) and implemented in the BioEdit package to visualise functional variation in the *FUL* genes. Conserved domains in the FUL were predicted using the NCBI Conserved Domain Database (http://www.ncbi.nih.gov/cdd) at E-value <0.001. Analysis of 5Kb upstream sequences of five *FUL* homologues for locating known motifs in the cis-acting regulatory elements was conducted using SIGNALSCAN program in Plant cis-Regulatory DNA Elements (PLACE, https://www.dna.affrc.go.jp/PLACE/?action=newplace). The number of motifs identified for each type were counted, and their roles were described (https://www.dna.affrc.go.jp/PLACE/place_seq.shtml). Also, the same dataset (5Kb upstream sequences of *FUL* homologues) was investigated for the presence of any novel motifs (sequence pattern that repeatedly occurs in a group of related protein or DNA sequences) using MEME (Multiple EM for Motif Elicitation, https://meme-suite.org/meme/tools/meme).

2.9 Microscopic analysis of pod anatomy

Anatomical features of valve margins from pods of parental lines were collected 35 to 40 days after anthesis. Hand sections were prepared from the middle of the pod, where the replum was narrow. Fresh sections were observed for autofluorescence using a fluorescence microscope. Photographs were taken using a Zeiss Axiphot microscope fitted with a Sony Cyber-shot digital camera.

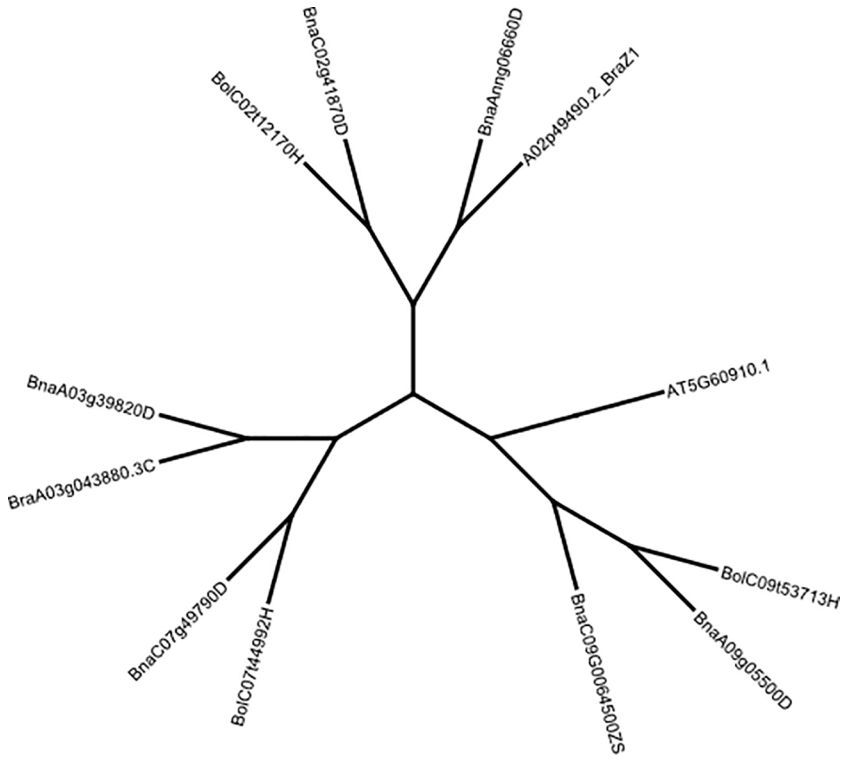


FIGURE 2
Neighbour-joining tree showing the grouping of *B. rapa*, *B oleracea* and *B. napus FUL* copies using Jukes-Cantor distance and Blosum62 cost matrices implemented in Geneious Prime. The *FUL* gene of *A. thaliana* (AT5G60910, TAIR) was used as an outgroup. Multiple sequence alignments were also carried-out for the *FUL* homologues. FUL protein sequences were retrieved from the BRAD database (www.brassicadb.cn, Accessed 18 April 2023).

3 Results

3.1 Inheritance of pod shatter resistance

We evaluated 203 F_2 lines derived from a cross between the *B. napus* line BLN3343-C00402 (maternal parent, NBGIP accession BC95041, shattering type) and interspecific line BC95042 (paternal parent derived from *B. rapa/B. napus*, resistant to pod shattering, Raman et al., 2014) using the pendulum test to investigate the genetic inheritance and genetic determinants underlying pod-shattering resistance. Herein, we implemented the pendulum test to detect genetic variation in rupture energy - a measure of pod strength/resistance to shattering (Kadkol et al., 1984; Kadkol et al., 1986; Liu et al., 1994; Raman et al., 2014). The interspecific line, BC95042, required a higher level of force to break up the pod and release seed; therefore, it had a higher value for rupture energy than the maternal line BC95041.

The F_2 population derived from a single F_1 plant showed a continuous distribution of rupture energy scores, ranging from 2.32 mJ to 17.76 mJ) (Figure 3A). We observe that both pod valves separate length-wise (vertically) under field conditions (Figure S1A). This shattering pattern differs from pod drop, which often occurs in related species of *Brassica*, such as *Raphanus raphanistrum subsp. sativus* (L.) (Figure S1B). Microscopic analysis revealed that the dehiscence zone is well-differentiated in shatter-prone parental lines of the mapping population BC95041 compared to pod shatter-resistant parental lines (BC95042). Interspecific line BC95042 required high energy to rupture the pod (threshing) than the shatter-prone line BC95041 (Figure 3A). In the resistant parental line, there was less lignification of cells near the dehiscence zone and a less conspicuous distinction between lignified and separation layer from the replum compared to shatter-prone lines (Figures 3B, C). These observations suggest that the pod shatter resistance genes play an essential role in the dehiscence zone differentiating and subsequent seed dispersal (Liiljegren et al., 2000). To verify the rupture energy scores of the F_2 lines, we raised a subset of 40 $F_{2:3}$ progenies representing extreme phenotypes (the top 20 and bottom 20 F_2 lines based on their pod energy scores) under natural field conditions. A positive correlation (r = 0.7) between the rupture energy scores of F_2 plants and their $F_{2:3}$ progenies (Figure 3D) indicates that rupture energy scores are reliable and suitable for genetic analysis.

3.2 Multiple loci associated with resistance to pod shatter

Using the DArTseq technology (Raman et al., 2014), a total of 26,002 high-quality SNPs (single nucleotide polymorphism) and *in-sillco* presence-absence markers, which showed (i) polymorphism

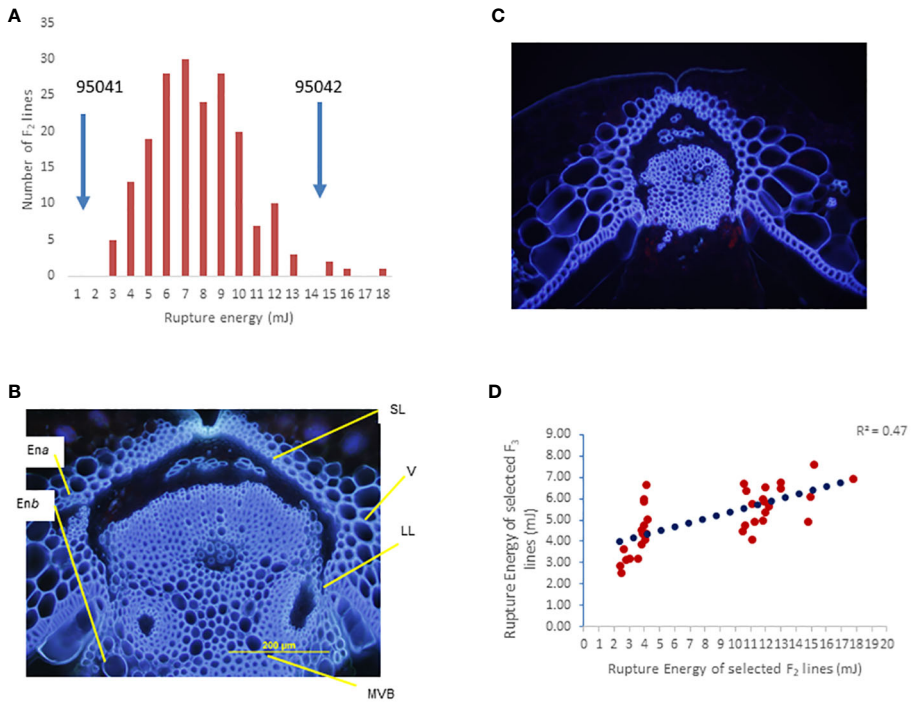


FIGURE 3
Genetic analysis of the BC95041 (shatter prone)/BC95042 (shatter resistant) F_2 population for pod shatter resistance. **(A)** Frequency distribution of rupture energy (RE) scores in a segregation population containing 179 individuals. Solid arrows indicate the average RE scores of the maternal line BC95041 and the paternal interspecific line 95042. **(B)** Cross section of developing pods showing well-developed dehiscence zone/abscission layer (DZ) in BC95041, whereas BC95042 shows limited DZ differentiation. The dehiscence zone: DZ, valve: V, the main vascular bundle of replum: MVB, the two endocarp layers, endocarp a: En*a*, and endocarp b: En*b* are shown. **(C)** Arrowheads indicate a lack of complete cell separation in the pod shatter-prone line. **(D)** Relationship `of pod rupture energy scores between F_2 and F_3 individuals.

between the parents and (ii) segregation in a mapping population, were used. We constructed a genetic linkage map spanning a total length of 2117.53 cM, with an average interval of 7.32 cM. The length of the chromosomes (linkage groups) ranged from 22.25 (C02) to 179.81 cM (A09). The marker density of the linkage groups ranged from 3.61 (A02) to 10.15 (A10). On average, 80.51% of markers were anchored to the 19 linkage groups, representing the A^n and C^n subgenomes of the reference *B. napus* cv. Darmor-*bzh* genome (Table 1). Using a genetic framework map based on 15,498 DArTseq markers (Table S1), we identified and located the significant QTLs conferring resistance to pod shatter on the *B. napus* genome. Different algorithms were used to identify robust associations for breeding use. Linear regression analysis using an additive model revealed that the top 99 markers mapped on chromosomes A01, A05, A09, C03 and C04 have a significant association (LOD ≥3.00) with resistance to pod shatter (Figure 4A, Table S2A). Of them, the top 16 markers were localised on A09 within 4.59 to 21.47 cM, and *in-silico* DArTseq marker 3101411 showed the most significant association ($-log_{10}P$ = 5.16) with resistance to pod shatter (Supplementary Table S2B). This marker showed a complete linkage with 15 other markers (Table S2C). Haplotype-based association test was conducted to detect the association between observed variations of pod shatter and marker haplotypes rather than single SNPs using the SVS package. We detected 677 haplotype blocks (HB, Supplementary Table S3A) following parameters described by Gabriel et al. (2002). Two markers in HB 303 on A09 detected the most significant association for pod shatter resistance with logistic regression (Table S3B). Haplotype trend regression revealed that HB308 (delimited with 3105829|F|0-8:C>G-8:C>G, 5121480|F|0-11:T>C-11:T>C, 3074795|F|0-19:G>T-19:G>T, 5050199|F|0-8:T>C-8:T>C markers, followed by HB309 with 3146480|F|0-46:A>G-46:A>G was the most significantly associated with pod rupture energy in the BC95041/BC95042 population (Table S4).

We further detected QTLs associated with rupture energy and pod length using the simple interval mapping (IM) and composite

TABLE 1 Linkage map showing genetic distance, distribution and distance (cM) of DArTseq markers in the F_2 population from BC95041/BC95042.

Chromosome	Mapped markers (No)	Total length (cM)	Average marker density	Markers mapped on AC genome	Markers mapped on the physical *B. napus* cv Darmor-*bzh* genome (%)
A01	1060	136.94	7.74	224	78.87
A02	246	68.23	3.61	45	81.71
A03	1050	165.86	6.33	234	77.71
A04	757	89.89	8.42	158	79.13
A05	1020	118.09	8.64	226	77.84
A06	1465	149.78	9.78	268	81.71
A07	892	121.99	7.31	166	81.39
A08	618	64.91	9.52	113	81.72
A09	1481	179.81	8.24	284	80.82
A10	902	88.83	10.15	170	81.15
Total A subgenome	9491	1184.34	8.01	1888	80.11
C1	492	106.23	4.63	106	78.46
C2	83	22.25	3.73	6	92.77
C3	1214	174.60	6.95	226	81.38
C4	984	137.40	7.16	230	76.63
C5	427	88.19	4.84	65	84.78
C6	524	96.52	5.43	101	80.73
C7	1012	148.07	6.83	171	83.10
C8	626	78.63	7.96	104	83.39
C9	645	81.30	7.93	123	80.93
Total C subgenome	6007	933.19	6.44	1132	81.16
Total A and C genomes	15498	2117.53	7.32	3020	80.51

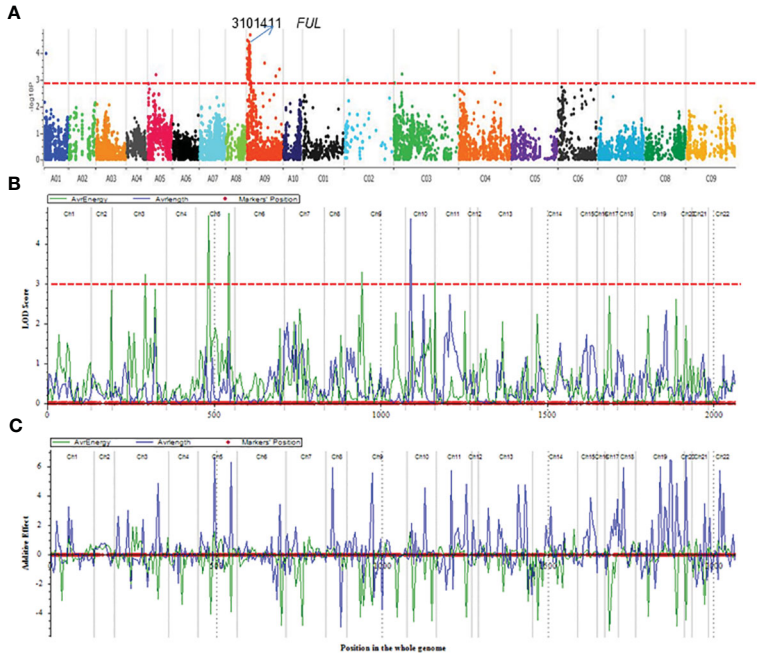


FIGURE 4
QTL mapping for pod shatter resistance measured as rupture energy (mj) by the pendulum test and pod length in the $F_{2:3}$ population derived from BC95041/BC95042. **(A)** Manhattan plots showing genomic regions associated with resistance to pod shatter: significant regions are labelled. **(B)** Gene scan showing a single QTL on chromosome A09 for pod shatter resistance in an F_2 population derived from a cross between BC95041 and BC95042. Significant QTL having a LOD score of 4 are shown by the dashed line (in blue colour). Pod shatter resistance was evaluated under birdcage conditions at Wagga Wagga, Australia and tested for rupture energy using a pendulum. **(C)** Allelic effects estimated by CIM approach. Linkage groups: Ch1-Ch10 relate to chromosomes A1-A10, Ch11-Ch14 to C01-C04; Ch15-16 to C05, Ch17-18 to C06, Ch19 to C07, Ch20-21 to C08 and Ch22 to C09.

interval mapping (CIM) approaches using the ICIM package. Five to seven significant QTLs for rupture energy were detected on chromosomes A03, A05 and A09 and C01 with IM and CIM (Table 2, Figure 4B). Three consistent QTLs were localised to the same genomic regions on chromosomes A02 and A05 across the analytical methods (Table 2). LOD scores of QTLs ranged from 2.8 to 4.77 and accounted proportion of variance explained (PVE) from 6.29% to 20.80% (Table 2). QTLs displayed both additive and dominant effects. Both parental lines contributed alleles for pod shatter resistance (Figure 4C). However, the interspecific paternal line BC95042 showed higher allelic effects (more than 2 folds) than the maternal *B. napus* line BC95041.

To investigate the major genetic determinants controlling rupture energy, we binned pod shatter variation scores into two discrete categories, resistant (1, rupture energy: 2.32 to 6.94 mJ) and susceptible (0, rupture energy: 7.0 to 17.76 mJ) phenotypes, in conjunction with the seven highly significant markers (Supplementary Table S5A) and performed haplotype analysis to determine trait-marker association. The chi-squared analysis supported the presence of a single shatter resistance gene in BnF_2 ($\chi^2_{3:1}$ = 0.17, with 1 degree of freedom, Two-tailed *P* value = 0.90). The HB 309 (defined by 15 SNPs: 3146480|F|0-46:A>G-46:A>G, 3096696|F|0-28:T>C-28:T>C, 3101752|F|0-29:C>T-29:C>T, 3159673|F|0-15:T>G-15:T>G, 5818650|F|0-5:C>T-5:C>T, 7250077|F|0-9:G>A-9:G>A, 3076890|F|0-52:A>T-52:A>T, 3079266|F|0-41:T>A-41:T>A, 3113543|F|0-40:A>C-40:A>C, 5121412|F|0-9:A>G-9:A>G, 5120748|F|0-29:G>C-29:G>C, 7249512|F|0-32:T>C-32:T>C, 3076528|F|0-55:T>C-55:T>C, 3077272|F|0-18:C>T-18:C>T, 3081487|F|0-26:C>T-26:C>T) revealed the most significant marker association with pod shatter resistance (χ^2 -log_{10}P: 9.99, Supplementary Table S5B) on chromosome A09. No other significant association was detected on *B. napus* chromosomes. Significantly associated markers detected on A09 showed collinearity between genetic and physical maps (Figure S2A). Different analytic methods revealed at least one significant locus on chromosome A09 that conditions variation in pod shatter resistance in the BC95041/BC95042 population. Mendelisation of quantitative variation revealed the limitation of identifying significant QTLs for trait variation (Tables 2, 3).

3.3 Pod length QTLs are not related to pod shattering

Previous studies showed pod shatter resistance, measured as a random impact test, correlates with pod length (Cui, 2013). To determine whether pod length variation relates to pod shattering tested with pendulum test in the F_2 population from BC95041/BC95042, we mapped QTLs associated with pod length on A02,

TABLE 2 Quantitative Trait Loci (QTLs) associated with pod shatter resistance measured as average rupture energy with the pendulum test.

Mapping approach	Chromosomal location	DArTseq Marker	Physical position on Darmor-bzh v4.1	DArTseq Marker	Physical position on Darmor-bzh v4.1	LOD	PVE (%)	Additive effect	Dominant effect
Composite interval mapping of additive QTL									
	A02	****3129258|F|0-32:G>A-32:G>A***	***23443447***	***4335059|F|0-41:T>C-41:T>C***	***24434057***	***2.84***	***9.42***	***0.07***	***2.05***
	A03	*3095606|F|0-36:A>T-36:A>T*	*14823303*	**3100670|F|0-31:A>G-31:A>G*	*12171871 on chrAnn_random*	*3.24*	*20.80*	*-1.57*	*-1.50*
	A03	*5048176|F|0-11:C>T-11:C>T*	*19780019*	**3100404|F|0-57:G>T-57:G>T*	*21580461*	*2.87*	*19.25*	*-3.02*	*-3.14*
	A05	***3089648|F|0-11:G>A-11:G>A***	***5420258***	****3089864|F|0-22:T>C-22:T>C***	***5947676***	***4.71***	***13.06***	***-4.04***	***-5.12***
	A05	***4116883|F|0-10:C>T-10:C>T***	***19860330***	****3101784|F|0-53:A>G-53:A>G***	***20067798***	***4.77***	***16.30***	***-3.89***	***-3.79***
	A09	*3082931|F|0-57:C>T-57:C>T*	*6081612*	*4167404|F|0-5:A>G-5:A>G*	*8328617*	*3.29*	*15.72*	*-3.51*	*-3.67*
	C01	*3101048|F|0-47:C>T-47:C>T*	*1404201*	*4110108|F|0-53:C>T-53:C>T*	*1469395*	*3.03*	*6.29*	*-1.05*	*0.04*
Single Interval mapping of additive QTL									
	A02	****3129258|F|0-32:G>A-32:G>A***	***23443447***	***4335059|F|0-41:T>C-41:T>C***	***24434057***	***2.93***	***12.04***	***-0.17***	***2.30***
	A05	***3089648|F|0-11:G>A-11:G>A***	***5420258***	****3089864|F|0-22:T>C-22:T>C***	***5947676***	***3.69***	***14.07***	***-4.28***	***-5.32***
	A05	***4116883|F|0-10:C>T-10:C>T***	***19860330***	****3101784|F|0-53:A>G-53:A>G***	***20067798***	***3.83***	***18.51***	***-4.15***	***-4.36***
	A09	*5050053|F|0-9:T>G-9:T>G*	*1798316*	*5121480|F|0-11:T>C-11:T>C*	*4340953*	*2.91*	*9.22*	*1.17*	*0.49*
	A09	*5049291|F|0-34:G>A-34:G>A*	*2530510*	*3140648|F|0-36:T>C-36:T>C*	*2767343*	*2.80*	*17.72*	*1.78*	*-0.05*

DArTseq markers were binned, and DArTseq SNPs were used for QTL analysis. The logarithm of the odds (LOD) scores, additive effects, and the proportion of phenotypic variance (PVE) were estimated using the ICIM package. Permutation Loci detected across Composite Interval (ICIM) and simple interval mapping (IM) were in bold. *Distance, based on cosegregating loci as linked marker did not return a significant hit.

A05, A07, A08, A10, C02 and C05 (Supplementary Table S6A). Simple interval mapping identified three significant QTLs on chromosomes A05, A07, A10, and C02, whereas composite interval mapping identified two QTLs on A10 and C01 (Supplementary Table S7A). None of the QTLs associated with pod length was collocated with QTLs for rupture energy, suggesting that pod length is genetically not associated with rupture energy (Table 2, Supplementary Table S7A). This was further substantiated by the lack of phenotypic correlation between pod length and shatter resistance scores (r = 0.01, Figure S2B).

3.4 Epistatic QTL interactions modulate variation in pod shatter resistance

Using a threshold estimated by permutation test at P = 0.05, 1,000 iteration (4.87), five pairs of significant epistatic QTLs for rupture energy were detected on A01/C01, A03/A07, A03/C03, A07/C03, and C01/C02 and revealed effects for additive × additive, additive × dominance and dominance × dominance interactions (Figure 5, Table 3). These EPI-QTLs accounted for 16.61% to 28.44% of PVE. Both parental alleles contributed to the epistasis in the intercross population. Additive

10.3389/fpls.2023.1233996

TABLE 3 Epistatic Quantitative Trait Loci (QTL) associated with pod shatter resistance measured as average rupture energy with the pendulum test.

Chromosome	LeftMarker1	Physical position on Darmor-bzh	RightMarker1	Physical position on Darmor-*bzh*	Chromosome	LeftMarker2	Physical position on Darmor-bzh	RightMarker2	Physical position on Darmor-bzh	LOD	PVE (%)	Add1	Add2	Dom1	Dom2	AddbyDom1	AddbyDom2	DombyAdd1	DombyAdd2
A01	4110587|F|0-9:C>G-9:C>G	2276310	3132222|F|0-58:T>C-58:T>C (A1 random)	155773	C01	**3101048|F|0-47:C>T-47:C>T**	**1713593**	4110108|F|0-53:C>T-53:C>T	**1644880**	5.44	17.45	2.39	-3.34	-3.98	-2.48	-2.55	-2.51	3.23	3.82
A03	5148873|F|0-19:G>A-19:G>A	5666135	4118427|F|0-10:A>G-10:A>G	6024022	C03	4121078|F|0-63:C>T-63:C>T	12585496	3141033|F|0-28:T>C-28:T>C	13581980	4.87	28.44	-0.46	-0.36	0.46	1.59	0.48	-2.75	0.50	-2.36
A03	***3100404|F|0-57:G>T-57:G>T**	**14324688**	**5048176|F|0-11:C>T-11:C>T**	**21730375**	**A07**	**5029215|F|0-26:C>T-26:C>T**	**2562779**	***3078953|F|0-62:C>T-62:C>T**	**2646233**	5.27	17.00	0.88	-0.77	-1.17	-0.84	-1.27	-1.43	0.99	0.51
A07	**5029215|F|0-26:C>T-26:C>T**	**2562779**	***3078953|F|0-62:C>T-62:C>T**	**2646233**	C03	4116381|F|0-24:G>C-24:G>C	18167918	4338040|F|0-47:G>T-47:G>T	20237766	5.03	23.12	-1.11	-2.40	-1.29	-1.93	2.31	0.76	2.32	0.47
C01	**3101048|F|0-47:C>T-47:C>T**	**1713593**	**4110108|F|0-53:C>T-53:C>T**	**1644880**	C02	4166149|F|0-37:C>G-37:C>G	2621703	3145176|F|0-14:T>A-14:T>A	2632328	5.31	16.61	-3.27	-2.60	-2.32	-2.47	2.78	2.41	3.14	2.46

The logarithm of the odds (LOD) scores, additive effects (Add1 with marker1 and Add2 with marker 2 interval), Dominant (Dom 1 with marker1 and Dom2 with marker 2 intervals), Additive x dominance (Add by Dom1 and Add by Dom2 with marker1 and 2 intervals), dominance x additive (Dom By Add1 and Dom by Add2 with marker 1 and 2 intervals) effects and the proportion of phenotypic variance were estimated using the EPI-CIM-ADD algorithm implemented in the ICIM package. Loci detected across digenic interaction were bold (see Table 2). *Distance, based on cosegregating loci as linked markers did not return a significant hit. DArTseq markers were binned, and DArTseq SNPs were used to identify digenic epistatic interactions.

marker effects between A03 and A07 chromosomes and A01 and C01 were in the repulsion phase. Epistatic QTLs for pod length were identified on chromosomes; A03/C07, A03/A05, A05/A08 and A05/A09, A05/C01, A05/C03, A09/C08, A09/A10, A10/C03 and A10/C08 at threshold 5 (Table S7B). However, using the threshold permutation test value estimated using 1,000 iterations, we did not identify any significant epistatic QTL for pod length.

3.5 Prioritized candidate genes underlying QTLs for pod shatter resistance

We searched for the physical location of significant markers flanking QTLs for main effects and epistatic interactions (Tables 2, 3) using the *B. napus cv.* Darmor-*bzh* reference genome v4.1 (Supplementary Tables S8A, B). Annotated genes mapped with the QTLs marker intervals and the homologues of *priori* genes involved in pod shattering of *A. thaliana.* were inspected. Annotated genes in the reference assemblies located within QTL intervals in reference assemblies were prioritized as candidates for pod shatter resistance. The highly significant marker 3101411 associated with pod shatter resistance on A09 was mapped to the reference sequence of C08, and other cosegregating markers with 3101411 that were located at the same locus on the genetic map (16.45 cM) were mapped to the 2,177,920 to 2,443,302 bp of the Darmor-*bzh v4.1* reference sequence (Supplementary Table S2C). Comparative analysis identified several candidate genes, including *AP2*, *ABI3*, *ARF*, *BP1*, *CEL6*, CESA3, *FIL*, *FUL*, *GA2OX2*, IAA31, *IND*, *LAC4*, *LEUNIG*, *KNOTTED*, *MAGL15*, *PG1, RPL*, *QRT2*, *RGA*, *SPL* and *TCP10*) underlying main QTL and epistatic QTL interactions for pod shatter resistance. Three copies of the *FUL* gene underlie the QTLs for pod shatter resistance on chromosomes A02, A03 and A09 (Table 2, Supplementary Table S10). Marker 3129258|F|0-32:G>A-32:G>A was located 63.6 kb from BnaAnng06660D homologue of *FUL* on A02 (Supplementary Table S9A). The A03 QTL delimited with 5048176|F|0-11:C>T-11:C>T was mapped ~116kb apart from the *FUL* homolog (BnaA03g39820D), accounting for 19.25% % PVE. QTL on chromosome A09 delimited with 5121480|F|0-11:T>C-11:T>C marker (19.25% of the total PVE) was located near the *FUL* gene (~248Kb, BnaA09g05500D, Table 2). Therefore, *FUL* may contribute to genetic variation in pod shatter resistance in the population used herein. To check whether there are candidate genes that could not be retrieved based on a single reference (Darmor-*bzh* versions v4.1/10) genome assembly, we utilised the BnaOmics platform that integrates pan-genome of 26 *B. napus* reference genomes and re-sequencing data of 2,885 accessions (Cui et al., 2023). At least two *FUL* copies of A02 and A03 were located in the pan-genome (Table S10).

3.6 Sequence divergence of FRUITFULL in 373 *B. napus* varieties

FUL is a MADS-box transcription factor that is shown to be a part of a complex regulatory network that controls floral meristem identity, shoot maturation, floral transition, cell proliferation in pod valves and cell differentiation by limiting the dehiscence zone formation in *A.*

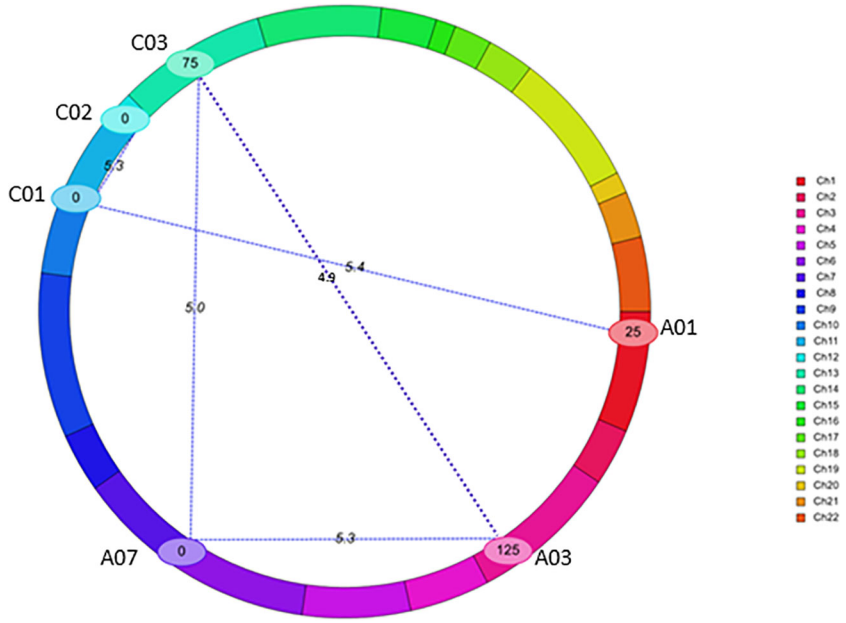


FIGURE 5
QTL interactions that showed epistatic effects for pod shatter resistance in the F2 population from BC95041/BC95042. The epistatic interaction was identified using the CIM-EPI approach in the ICIM package. Different linkage groups relating to *B. napus* chromosomes are shown (Ch1-Ch10 relate to chromosomes A1-A10, Ch11-Ch14 to C01-C04; Ch15-16 to C05, Ch17-18 to C06, Ch19 to C07, Ch20-21 to C08 and Ch22 to C09) in different colours. Interactions are shown with blue lines. Chromosomes that showed significant interactions are labelled.

thaliana, *B. napus* and *B. juncea* (Gu et al., 1998; Ferrandiz et al., 2000; Rajani and Sundaresan, 2001; Ostergaard et al., 2006). TBLASTN and reciprocal BLASTP searches against Arabidopsis proteins confirmed that the *FUL* (AGL8, AT5G60910) clade includes five homologues in *B. napus* on chromosomes Ann_random (BnaAnng06660D, A02 in the pan-genome), A03 (BnaA03g39820D), A09 (BnaA09g05500D), C02 (BnaC02g41870D) and C07 (BnaC07g49790D) detected in both reference genome assemblies v4.1 and 10 (Figures S3, S5). However, seven homologues were annotated in the *B. napus* pan-genome gene assembly on A02, A03, A09, C02, C07 and C09 chromosomes and validated for the presence of MADS-box domain-containing protein with a K-box coil and the MEF2 DNA-binding/dimerisation regions (Table S10B). *FUL* homologues of *B. napus*: BnaA03g39820D, BnaA09g0550D and BnaAnn06660D were clustered into distinct clades with *B. rapa* and BnaC02g41870D and BnaC07g49790D with *B. oleracea* clade, as expected (Figure 2). *FUL* homologue of *B. oleracea* (LOC10631378) showed grouping with BnaA09g0550D. Since we identified several QTLs that map near to MADS-box transcription factors such as *AGAMOUS* (*AG*), *APETALA* and *AG-LIKE* transcription factors could also regulate *FUL* expression throughout vegetative and reproductive phases during the plant development; we performed phylogenetic analysis using the Bayesian clustering method. This analysis differentiated *AG*, *FUL (AGL8)*, *SHP1* (*AGL1*), *SHP2 (AGL5)*, and *AGL3/SEPALLATA4* (*SEP4*) clades (Figure S5).

To date, BnaA09g05500D is the only *FUL* orthologue of *A. thaliana* and its closely related MADS-box gene in *Sinapis alba: MADSB*, which is shown to be involved in pod dehiscence via gene expression studies (Ferrandiz et al., 2000; Liiljegren et al., 2000; Chandler et al., 2005). Therefore, we further investigated its gene structure, evolution rate, and sequence variants using a dataset of 373 resequenced *B. napus* accessions utilised in the Australian National Brassica germplasm improvement program for gene discovery projects (Table S10). To determine the gene structure of the *FUL homologues* in *B. napus*, we used *AtFUL* (AT5G60910, TAIR). Sequence analysis of BnaA09g0550D) revealed that it encodes a 726 bp transcript with a 242 amino acid protein and comprises 8 exons and 7 introns (http://www.genoscopegen.cns.fr/brassicanapus/cgi-bin/geneView?src=colza;name=BnaA09g55330D) (Figure S3). The size of the first intron (intron 1) varied from 861 (A02) to 2462 (C02) bp, in contrast to some plant species, such as tomato and the wild D-genome progenitor of bread wheat, *Aegilops tauschii* (Takumi et al., 2009; Maheepala et al., 2019). The parental lines of the mapping population from BC95041 and BC95042 revealed 364 polymorphic SNPs and deletions NCBI, Banklt accession ID 2735083, (Table S11), and two non-synonymous variants were identified in exon 1 (c.A25G:p.K9E) and 7 (c.G616T:p.A206S). There were five non-synonymous SNV in exon 1 (c.G166A:p.E56K, c.G155A:p.G52D, c.G139A:p.V47I and c.A25G:p.K9E) and exon 7 (c.G616T:p.A206S) of BnaA09g05500D. Among all 373 accessions, up to 578 variants were detected in *FUL* homologues in *B. napus*; the majority (~50%) occurred in the intergenic region, followed by intronic regions (Table S11). Sequence variants were detected in the exonic and upstream sequence of *FUL* homologues, ranging from 19 to 36 and 11-99, respectively. We also identified splice variants for BnaA09g05500D (1 variant) and BnaAnng06660D gene (2 variants).

We performed selection pressure analysis to determine the evolution rate as the ratio of Ka/Ks of *FUL* copies. Our results show that BnaA09g05500D copy on chromosome A09 had purifying selection (<0.1) followed by copies on C02, suggesting conserved

function compared to BnaA03g39820D and BnaC07g49790D on A03 and C07, respectively (Supplementary Table S10C).

Analysis of 5 kb upstream regions of five *FUL* homologues with the SIGNALSCAN program within the PLACE database (https://www.dna.affrc.go.jp/PLACE/?action=newplace) revealed several motifs found in plant cis-acting regulatory DNA elements. The search identified 183 motifs, ranging from 127 in BnaC02g41870D to 145 in BnaA03g39820D). Of these 183 motifs, 91 common motifs were present in all five homologues, while 25 were unique to one of them. The duplication frequency of these common motifs in all five genes is depicted in Figure 6A, and the numbers are given in Table S12. Among the common motifs, DOFCOREZM is the most abundant one, with duplication frequency of 66 to 98 in the 5 Kb upstream region of *FUL* homologues, followed by CACTFTPPCA1, GT1CONSENSUS GATABOX and CAATBOX1. The *FUL* gene is shown to bind to a specific CArG box, with the consensus sequence CC(A/T)6GG (de Folter and GC, 2006). In *B. napus*, 2 to 20 CArG motifs (CARGCW8GAT and CARGATCONSENSUS) were found in the upstream sequence of *FUL* homologs. We identified CArG consensus sequence (CCWWWWWWGG) in BnaAnng06660D and BnaC07g49790D only, whereas a variant of CArG motif with a more extended A/T-rich core (CWWWWWWWWG) is found in upstream sequences of all five *FUL* homologues (Figure 6B). There were 14 motifs (ABRELATERD1, ACGTATERD1, ACGTABREMOTIFA2OSEM, CBFHV, DRECRTCOREAT, LTRECOREATCOR15, MYB1AT, MYB2AT, MYBATRD22, MYBCORE, MYB2CONSENSUSAT, MYCCONSENSUSAT, MYCATERD1 and MYCATRD22) detected in the dataset which are associated with water stress or dehydration. Consistent with previous studies, we also found auxin response elements (GGTCCCATGMSAUR, AUXREPSIAA4, AUXRETGA1GMGH3, ARFAT, SURECOREATSULTR11 and CATATGGMSAUR) in our upstream sequences dataset. Among these motifs, SURECOREATSULTR11 and CATATGGMSAUR were found in the upstream sequences of all five genes, whereas GGTCCCATGMSAUR and AUXREPSIAA4 were unique to the upstream sequence of BnaA03g39820D (Table S12). Furthermore, seven motifs (WRKY71OS, PYRIMIDINEBOXOSRAMY1A, PYRIMIDINEBOXHVEPB1, GAREAT, MYBGAHV, GADOWNAT and GARE2OSREP1) were associated with gibberellin signalling pathway. The chromosome A09 *FUL* copy also had the maximum number (14) of SAUR (Small Auxin-Up RNA, CATATGGMSAUR) motifs, implicated in auxin responsiveness (Xu and Guilfoyle, 1997). Copy number variation and distribution of motifs in the upstream regulatory region of *FUL* may account for natural variation in gene expression and regulation of valve growth by interacting with other genes involved in valve margin differentiation, such as *SHP1*, *SHP2*, *IND* and *ALC*. *IND* also forms auxin minimum by coordinating auxin efflux in separation layer cells (Sorefan et al., 2009). We also found the GTGANTG10 motif (with duplication frequency 28-43), which shows homology to pectate lyase (Rogers et al., 2001).

We also discovered three unknown motifs in the 5 Kb upstream sequences of all five *FUL* homologues. The first motif KYKTGWG YCTMCMSTKWSGCWWRCGTKKKWWCMGTRMCGTAM GKGATKT (GCGTGTGCCTCCCCTGTCGCAAGCGTGGGAAC CGTGCCGTACGGGATGT) is potentially located within first 500bp upstream, whereas the second motif KATRTKTWK GBCHYHTYARVDCHMAAVTBTGKHYCWTTTBTTC (GATG CGTTGGCCCCCTCAGCGCCCAACTGTGGCCCATTTCTTC) and the third motif TWYGKGMRATATAMYATAT GMKKTMTTGWSAWGTTSWCWTA (TACGGGCGATA

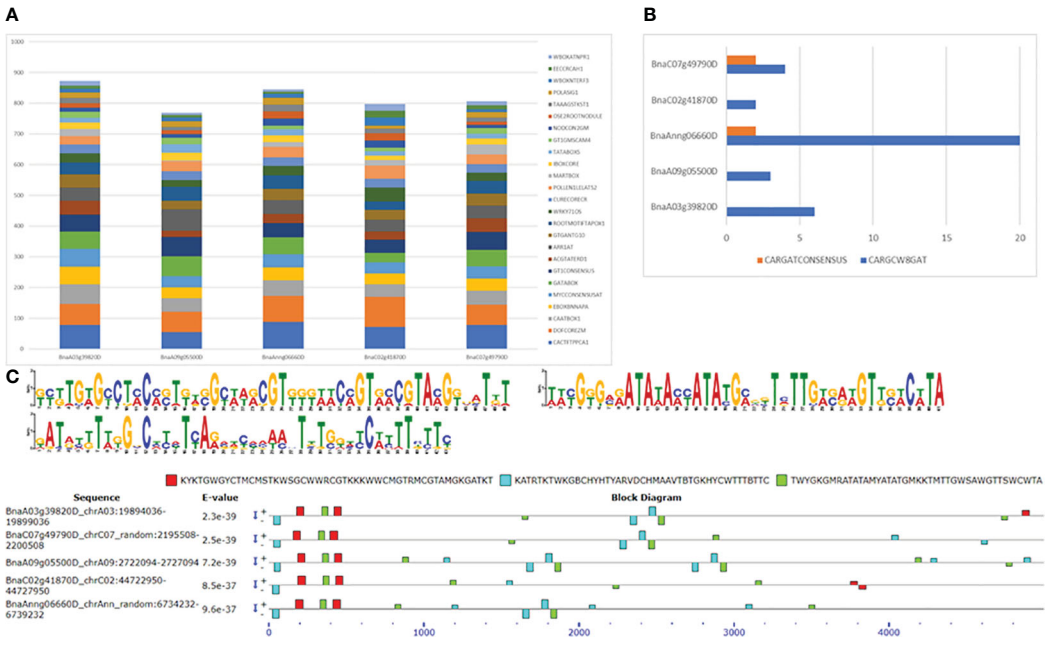


FIGURE 6
Motif identification in 5 Kb upstream sequences of five *FUL* homologues in *B. napus*. **(A)** Duplication frequency of 25 most abundant motifs. **(B)** CArG motifs and their frequency of occurrence. **(C)** Three novel motifs found using MEME and their occurrence in the sequence. + and – indicates the motif occurrence on sense and antisense strands.

TACCATATGCGGTCTTGACAAGTTCACATA) are randomly dispersed with no particular pattern detected in their occurrence with respect to positions (Figure 6C). Also, the first motif is mainly detected on the sense strand, whereas the second and third motifs are comparatively present on both sense and antisense strands.

4 Discussion

Seed shattering is a massive issue in commercial canola production worldwide, underpinning growers' profitability. Pod shatter-resistant varieties suitable for direct harvesting with combines are essential to reduce (i) reliance on windrowing, (ii) yield losses, (iii) inputs cost (labour and fuel for windrowing and controlling rogues in subsequent crops), (iv) carbon emissions occurred while windrowing followed by threshing with combine harvesters, and to improve (v) gross margins of farmers (return on the investment).

Herein, we investigated the genetic basis of pod shatter resistance in an interspecific derivative of *B. rapa*/*B. napus*. In this study, we used the pendulum test to describe the genetic variation for pod shatter resistance in a quantitative manner and understand its underlying genetic and anatomical bases. Previously, several methods, such as the number of seeds lost from pods, the number of seedlings germinated, the random impact test, and the pendulum test, have been used to determine genetic variation for pod shatter resistance in *Brassica* species (Morgan et al., 1998). There were 6.23-fold differences in pod shatter resistance between parental lines, suggesting that the interspecific source, BC95042, could be used to improve resistance to pod shatter.

Genetic analysis showed that pod shatter resistance is due to seven QTLs located on A02, A03, A05, A09 and C01 chromosomes in an F_2 population derived from a cross between BC95041 and BC95042 (Table 2). With linear marker regression, HTR, IM, and CIM algorithms, we repeatedly detected four QTLs for pod shatter resistance on A02, A05 and A09, suggesting these QTLs are reliable for research and development activities such as introducing appropriate favourable alleles into canola varieties. Using different mapping algorithms with robust statistical power ensured the identification of significant marker-trait associations by reducing false positives to make genetic gains in canola breeding programs. Previous genetic mapping studies identified QTLs for pod shatter resistance in *B. rapa* (Mongkolporn et al., 2003; Bagheri et al., 2012), *B. juncea* (Kaur et al., 2020) and *B. napus* (Hu et al., 2012; Wen et al., 2013; Raman et al., 2014; Liu et al., 2016). Some of the QTLs were located in similar genetic positions on *B. napus* genome, which were detected in earlier studies (Table S9). However, there were no overlapping QTL regions across populations of Chinese origin. For example, Liu et al. (2016)) reported six significant QTLs for pod shatter resistance in a *B. napus* GWAS panel and two structured biparental populations on A01, A06, A07, A09, C02, and C05 chromosomes. Two QTLs on A06 and A09 were repeatedly detected across environments and mapping panels. QTL on A09 delimited with an Illumina SNP marker, Bn-A09-p30171993, was mapped near the *SHP1* gene (A09_random chromosome on the 4.1 Darmor-*bzh* assembly). However, *SHP1* and Bn-A09-p30171993 were located at the distal end of the A09 chromosome (Darmor-*bzh* version 10). However, this study identified three QTLs on chromosomes A02, A03 and A09 that significantly contributed to pod shatter resistance, accounting for 9.42% and 19.25% of the total PVE, respectively, and map near the *FUL* homologues (BnaAnng06660D, BnaA03g39820D and BnaA09g05500D, Table 1). These QTLs were not detected in other *B. napus* populations (Wen et al., 2013; Liu et al., 2016). We could not compare the map position of 13 QTLs for pod shatter resistance, measured by improved random impact method on A01, A04, A07, A08, C05, and C08 (Wen et al., 2013) as they were not mapped on any physical map of *B. napus*. Our study did not detect any QTL on A06 for pod shatter resistance located near the *GIBBERELLEIN-3-OXIDASE1* gene in *B. napus* populations of Chinese origin (Liu et al., 2016). Most QTLs on A01, C02, and C05 were not closely mapped. These observations hint that selection for pod-shattering may have occurred at several independent loci and shaped the genomic architecture of pod-shatter resistance during cultivation and selective breeding in *B. napus*. This hypothesis is supported by independent seed-shattering QTLs (on A03, A09, this study) and the absence of the *SHP1* and *TCP8* genes, as shown in earlier studies (Liu et al., 2016; Liu et al., 2020; Chu et al., 2022). During domestication, Brassica species may have acquired several shattering resistance mechanisms to reach the desirable level of shattering resistance, suitable for manual harvesting, probably under humid climates, e.g., Europe and Wuhan. However, the resistance level is insufficient for hot and dry climates, e.g., Australia.

The PVE (6.29 to 20.80%) and additive effects from both parental lines (-4.28 to 1.78) that we identified in this study were consistent with most of the published *B. napus* studies revealing a small to moderate proportion of genotypic variation (4.01 to 28.9%) in pod shatter resistance (Wen et al., 2013; Raman et al., 2014; Liu et al., 2016; Liu et al., 2020). A recent study shows a major gene (i.e. *TCP8* on C09) effect on pod shatter resistance via a lignified-layer bridge in a *B. napus* population (Chu et al., 2022). Our digenic interaction analysis showed five epistatic QTL interactions between chromosomes (A01-C01, A03-A07, A07-C03, A03-C03, and C01-C02). The positive epistatic effect of additive x additive suggested that the two epistatic loci (e.g. A03/C03, A07/C03, and C01/C02) with homozygous/heterozygous alleles from the same parent could increase the pod shatter resistance. However, the positive additive × dominance epistatic effect indicated that BC95042 could increase the pod shatter resistance. Breeding programs must consider additive and additive x additive epistatic interactions to improve resistance to pod shatter.

Based on the physical location of linked markers associated with pod shatter resistance, we prioritized *AG*, *ABI3*, *ARF3*, *BP1*, *CEL6*, *FIL*, *FUL*, *GA2OX2*, *IND*, *LATE*, *LEUNIG*, *MAGL15*, *RPL*, *QRT2*, *RGA*, *SPT* and *TCP10*, as candidate genes for pod shatter resistance (Table S9). The mechanisms and genetic factors involved in pod dehiscence have been investigated in *A. thaliana* and its closely related Brassica species. MADX-box transcription factors encoding *FUL*, *SHP1*, and *SHP2* are the major players that control fruit patterning, lignin deposition, and pod dehiscence in Arabidopsis (Gu et al., 1998; Liiljegren et al., 2000). *FUL* negatively regulates *SHP* and *IND* expression in the valve margin and *APETALA 1* in the outer whorl of the flower (Ferrandiz et al., 2000; Kaufmann et al., 2010). *FUL* and BEL-subfamily homeodomain gene *RPL* also negatively regulate *SHP* expression in the valve margin (Roeder et al., 2003). The floral homeotic gene *AP2* also negatively regulates

the expression of *SHP*, *RPL*, and *IND* genes and the expansion of replum and lignified layers (Ripoll et al., 2011). *SHP1* and *SHP2*, which act redundantly, regulate the expression of basic helix-loop-helix (bHLH) genes: *ALC*, *IND*, and *SPATULA* (*SPT*). *SHP1/2* and *IND* cause pod dehiscence by promoting cell proliferation and are involved in the differentiation of the lignification and separation layers in the stripes of the valve margin, whereas *ALC* and *SPT* are involved in forming the separation layer (Rajani and Sundaresan, 2001; Liljegren et al., 2004; Lewis et al., 2006; Groszmann et al., 2011). *IND* activates the expression of *ALC* and *SPT* but also promotes its own heterodimerisation with them through DELLA protein degradation (Girin et al., 2010; Girin et al., 2011). Finally, *ALC* and *SPT* are able to repress *IND* expression (Lenser and Theissen, 2013). *IND* regulates gibberellin levels through the *GA3 Oxidase 1/GA4* gene (Arnaud et al., 2010; Kay et al., 2013). *FIL*, *YABBY* and *JAG* can control the expression patterns of *FUL* and *SHP* in the valve and valve margins ((Dinneny and Yanofsky, 2005; Mühlhausen et al., 2013). We also identified downstream genes such as *BETA-1-4 GLUCANASE* (*CELLULASE6*), ENDO-POLYGALACTURONASE (*RDPG1*, *QRT2*), *MAN7*, *NST1/3* and other MADS family transcription factors like *SEPALLATA3*, *AGL15*, *SEP4*, associated with pod shatter resistance in the mapping population. These genes are implicated in pod dehiscence in *A. thaliana* and *B. napus* (Jiang et al., 2016; Li et al., 2021). Di Marzo et al. (2022), found that the expression of α-XYLOSIDASE1 (*XYL1*) is directly regulated in developing seeds and fruit by the MADS-box transcription factor *SEEDSTICK* (*STK*). They demonstrated that *XYL1* complement the *stk* smaller seed phenotype, confirming the importance of cell wall modulation in shaping organs. Some *priori* genes for pod shatter resistance were localised more than 1Mb from significant QTL regions. Small populations with low-density markers cannot resolve

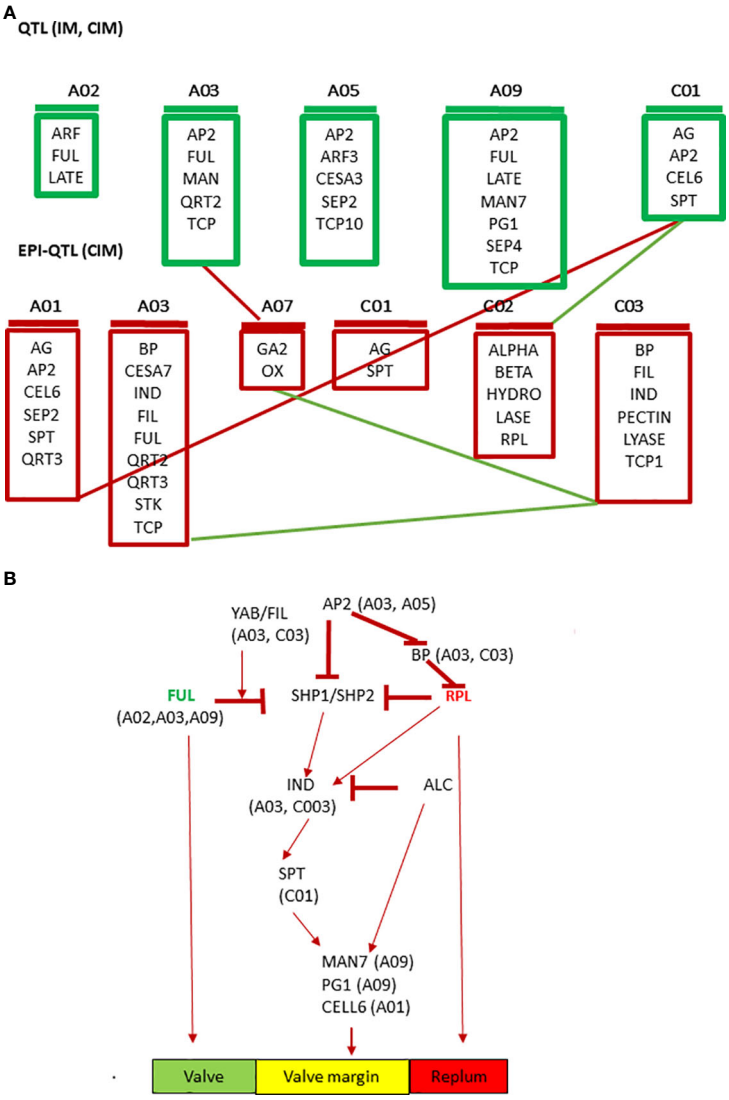


FIGURE 7
Prioritized candidate genes underlying QTL for pod shatter resistance using the simple interval, composite interval mapping, and epistatic-composite interval mapping algorithms implemented in the ICIM package. **(A)** Cartoon showing QTL with main effects (IM, CIM) and epistatic interactions (epi-QTL) along with their chromosomal location and **(B)** Extrapolated QTL-based candidate genes involved in pod shatter resistance network in *B. napus*. The green colour indicates valve, the yellow colour valve margin identity-related genes, and the orange colour indicates the replum.

recombination between markers and candidate genes (Raman et al., 2016). However, the homologs of pod shatter resistance genes that map further apart from significantly associated markers on other chromosomes could regulate genetic variation in pod shatter resistance. Further research is required to substantiate this hypothesis. We identified sequence variants between the parental lines of the mapping population and other elite lines of *B. napus*. Further studies are required to establish the role of sequence variants in pod shatter resistance genes and their functional role via gene expression and gene editing approaches. Overall, our data on genetic mapping and putative candidate/priority genes suggest the complex network involved in pod shatter resistance in *B. napus* germplasm, broadly consistent with *A. thaliana* (Figure 7), as reiterated earlier (Stephenson et al., 2019). This observation is consistent with the high syntenic relationships between *B. napus* and *A. thaliana* (Parkin et al., 2005).

In summary, we constructed the genetic framework map and identified seven genomic regions associated with pod rupture energy on A02, A03, A05, A09, and C01 chromosomes in an F_2 population derived from the BC95041/BC95042 line developed from *B. rapa*/*B. napus*. In addition, five pairs of significant epistatic QTL interactions for rupture energy between A01/C01, A03/A07, A07/C03, A03/C03, and C01/C02 chromosomes. Overall, our results showed that independent QTLs (on A02, A03, A05, A09 and C01 chromosomes) and interactive QTLs (on A01/C01, A03/A07, A07/C03, A03/C03, and C01/C02) contribute to genetic variation in pod shatter resistance. Epistatic QTL interactions possibly reflect the regulatory network (repressor and activators) involved in pod dehiscence in *A. thaliana*. Several QTL regions were mapped near the candidate genes (*AG, ABI3, ARF3, BP1, CEL6, FIL, FUL, GA2OX2, IND, LATE, LEUNIG, MAGL15, RPL, QRT2, RGA, SPT*, and *TCP10*) which are involved in pod dehiscence, primarily in Arabidopsis. We described putative *cis*-acting motifs and sequence variants in genic and promoter regions of *FUL* homologues in 373 *B. napus* accessions. This study provides a valuable resource for gene discovery, the molecular mechanism underlying pod shatter resistance and yield improvement in *Brassica* species. DNA markers could accelerate the use of QTL in the Brassica breeding programs for marker-assisted selection, backcross, and genomic selection pipelines.

5 Conclusions

This study found that the interspecific line, BC94052 has superior alleles for resistance to pod shatter. Our genetic mapping suggests pod shatter resistance is due to multiple loci; three QTLs map to the A02, A03 and A09 chromosomes near *FUL* homologues. Our research provides a valuable genetic resource for improving pod shatter resistance in canola and for future studies on understanding molecular mechanisms underlying pod shatter resistance.

Data availability statement

The datasets presented in this study can be found in online repositories. The Illumima sequence data of FULL genes can be found in NCBI Banklt accesssion 2735083.

Ethics statement

The authors declare that the experiments comply with the current laws of the country in which they were performed and comply with ethical standards.

Author contributions

HR and RR designed the research and analyzed the data. RR developed the mapping population and conducted the experiments. YQ and BM assisted in phenotyping and performed pod anatomy and DNA extractions. NS, XC, YZ, QH, HR, and SL aligned DArTseq data with the reference genomes and analysed the dataset. NG provided the seeds of an interspecific line. HR wrote the first draft, NS contributed to the sections and all authors approved the final draft of the manuscript.

Funding

We thank the GRDC and NSW DPI for supporting this research under the DAN00208 and the Key Research Project of Hubei province (No.2021EHB026).

Acknowledgments

We thank Ms. Hannah Roe and Mr. John Bromfield for the pendulum testing of F_2 and $F_{2:3}$ families and Dr Gururaj Kadkol and Greg Buzza for the discussion. HR thanks Ms. Charmaine Carlisle, Charles Sturt University, Wagga Wagga, for making the microscope available.

Conflict of interest

NG works for Nuseed Pty Ltd.

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

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1233996/full#supplementary-material

References

Altschul, S., Gish, W., Miller, W., Myers, E., and Lipman, D. (1990). Basic local alignment search tool. *J. Mole Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

Arnaud, N., Girin, T., Sorefan, K., Fuentes, S., Wood, T. A., Lawrenson, T., et al. (2010). Gibberellins control fruit patterning in *Arabidopsis thaliana*. *Genes Dev.* 24 (19), 2127–2132. doi: 10.1101/gad.593410

Bagheri, H., El-Soda, M., van Oorschot, I., Hanhart, C., Bonnema, G., Jansen - van den Bosch, T., et al. (2012). Genetic analysis of morphological traits in a new, versatile, rapid-cycling *Brassica rapa* recombinant inbred line population. *Front. Plant Sci.* 3 (183). doi: 10.3389/fpls.2012.00183

Balanzà, V., Roig-Villanova, I., Di Marzo, M., Masiero, S., and Colombo, L. (2016). Seed abscission and fruit dehiscence required for seed dispersal rely on similar genetic networks. *Development* 143 (18), 3372–3381. doi: 10.1242/dev.135202

Bowman, J. L., Baum, S. F., Eshed, Y., Putterill, J., and Alvarez, J. (1999). Molecular genetics of gynoecium development in Arabidopsis. *Curr. Top. Dev. Biol.* 45, 155–205. doi: 10.1016/S0070-2153(08)60316-6

Braatz, J., Harloff, H. J., Emrani, N., Elisha, C., Heepe, L., Gorb, S. N., et al. (2018a). The effect of INDEHISCENT point mutations on silique shatter resistance in oilseed rape (*Brassica napus*). *Theor. Appl. Genet.* 131 (4), 959–971. doi: 10.1007/s00122-018-3051-4

Braatz, J., Harloff, H.-J., and Jung, C. (2018b). EMS-induced point mutations in ALCATRAZ homoeologs increase silique shatter resistance of oilseed rape (*Brassica napus*). *Euphytica* 214 (2), 29. doi: 10.1007/s10681-018-2113-7

Cao, B., Wang, H., Bai, J., Wang, X., Li, X., Zhang, Y., et al. (2022). miR319-regulated TCP3 modulates silique development associated with seed shattering in brassicaceae. *Cells* 11 (19). doi: 10.3390/cells11193096

Chalhoub, B., Denoeud, F., Liu, S., Parkin, I. A., Tang, H., Wang, X., et al. (2014). Early allopolyploid evolution in the post-Neolithic Brassica napus oilseed genome. *Science* 345 (6199), 950–953. doi: 10.1126/science.1253435

Chandler, J., Corbesier, L., Spielmann, P., Dettendorfer, J., Stahl, D., Apel, K., et al. (2005). Modulating flowering time and prevention of pod shatter in oilseed rape. *Mol. Breed.* 15 (1), 87–94. doi: 10.1007/s11032-004-2735-4

Child, R., Chauvaux, N., John, K., Ulvskov, P., and Onckelen, H. (1998). Ethylene biosynthesis in oilseed rape pods in relation to pod shatter. *J. Exp. Bot.* 49, 829–838. doi: 10.1093/jxb/49.322.829

Chu, W., Liu, J., Cheng, H., Li, C., Fu, L., Wang, W., et al. (2022). A lignified-layer bridge controlled by a single recessive gene is associated with high pod-shatter resistance in. *Brassica napus L. Crop J.* 10 (3), 638–646. doi: 10.1016/j.cj.2021.09.005

Cui, J., Mei, D., Li, Y., Liu, J., Fu, L., Peng, P., et al (2013). Genetic contribution of silique related traits to silique shatter resistance of *Brassica napus* L. *Chin. J. Oil Crop Sci.* 35 (5), 461. doi: 10.7505/j.issn.1007-9084.2013.05.001

Cui, X., Hu, M., Yao, S., Zhang, Y., Tang, M., Liu, L., et al. (2023). BnaOmics: a comprehensive platform combining pan-genome and multi-omics data of Brassica napus. *Plant Commun.*, 100609. doi: 10.1016/j.xplc.2023.100609

de Folter, S., and GC, A. (2006). trans meets cis in MADS science. *Trends Plant Sci.* 11, 224–231. doi: 10.1016/j.tplants.2006.03.008

de la Pasture, L. (2018). *Pod shatter—shattering implications for pod pop. Crop Production Magazine*. Available at: https://www.cpm-magazine.co.uk/technical/pod-shatter-shattering-implications-pod-pop/ (Accessed 25 March 2023).

Di Marzo, M., Viana, V. E., Banfi, C., Cassina, V., Corti, R., Herrera-Ubaldo, H., et al. (2022). Cell wall modifications by α-XYLOSIDASE1 are required for control of seed and fruit size in Arabidopsis. *J. Exp. Bot.* 73 (5), 1499–1515. doi: 10.1093/jxb/erab514

Dinneny, J. R., and Yanofsky, M. F. (2005). Drawing lines and borders: how the dehiscent fruit of Arabidopsis is patterned. *BioEssays* 27 (1), 42–49. doi: 10.1002/bies.20165

Ferrandiz, C., Liljegren, S. J., and Yanofsky, M. F. (2000). Negative regulation of the SHATTERPROOF genes by *FRUITFULL* during Arabidopsis fruit development. *Science* 289 (5478), 436–438.

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., et al. (2002). The structure of haplotype blocks in the human genome. *Science* 296 (5576), 2225–2229. doi: 10.1126/science.1069424

Girin, T., Paicu, T., Stephenson, P., Fuentes, S., Körner, E., O'Brien, M., et al. (2011). INDEHISCENT and SPATULA interact to specify carpel and valve margin tissue and thus promote seed dispersal in arabidopsis. *Plant Cell* 23 (10), 3641–3653. doi: 10.1105/tpc.111.090944

Girin, T., Stephenson, P., Goldsack, C. M. P., Kempin, S. A., Perez, A., Pires, N., et al. (2010). Brassicaceae INDEHISCENT genes specify valve margin cell fate and repress replum formation. *Plant J.* 63 (2), 329–338. doi: 10.1111/j.1365-313X.2010.04244.x

Groszmann, M., Paicu, T., Alvarez, J. P., Swain, S. M., and Smyth, D. R. (2011). SPATULA and ALCATRAZ, are partially redundant, functionally diverging bHLH genes required for Arabidopsis gynoecium and fruit development. *Plant J.* 68 (5), 816–829. doi: 10.1111/j.1365-313X.2011.04732.x

Gu, Q., Ferrándiz, C., Yanofsky, M. F., and Martienssen, R. (1998). The FRUITFULL MADS-box gene mediates cell differentiation during Arabidopsis fruit development. *Development* 125 (8), 1509–1517. doi: 10.1242/dev.125.8.1509

He, H., Bai, M., Tong, P., Hu, Y., Yang, M., and Wu, H. (2018). CELLULASE6 and MANNANASE7 affect cell differentiation and silique dehiscence. *Plant Physiol.* 176 (3), 2186–2201. doi: 10.1104/pp.17.01494

Hu, Z., Hua, W., Huang, S., Yang, H., Zhan, G., Wang, X., et al. (2012). Discovery of pod shatter-resistant associated SNPs by deep sequencing of a representative library followed by bulk segregant analysis in rapeseed. *PloS One* 7 (4), e34253. doi: 10.1371/journal.pone.0034253

Jiang, X., He, H., Wang, T., Wang, X., and Wu, H. (2016). Gene expression profile analysis indicate SEPALLATA3 and AGL15 potentially involved in arabidopsis silique dehiscence by regulating glycosyl hydrolase. *J. Plant Biol.* 59 (2), 133–142. doi: 10.1007/s12374-016-0567-5

Kadkol, G., Halloran, G., and Macmillan, R. (1985). Evaluation of *Brassica* genotypes for resistance to shatter. II. Variation in siliqua strength within and between accessions. *Euphytica* 34, 915–924. doi: 10.1007/BF00035431

Kadkol, G., Halloran, G., and MacMillan, R. (1986). Inheritance of siliqua strength in *Brassica campestris* L. I. Studies of F_2 and backcross populations. *Can. J. Genetical Cytology* 28, 365–373. doi: 10.1139/g86-054

Kadkol, G. P., Macmillan, R. H., Burrow, R. P., and Halloran, G. M. (1984). Evaluation of Brassica genotypes for resistance to shatter. I. Development of a laboratory test. *Euphytica* 33 (1), 63–73. doi: 10.1007/BF00022751

Kaufmann, K., Wellmer, F., Muiño, J. M., Ferrier, T., Wuest, S. E., Kumar, V., et al. (2010). Orchestration of floral initiation by APETALA1. *Science* 328 (5974), 85–89. doi: 10.1126/science.118524

Kaur, J., Akhatar, J., Goyal, A., Kaur, N., Kaur, S., Mittal, M., et al. (2020). Genome wide association mapping and candidate gene analysis for pod shatter resistance in Brassica juncea and its progenitor species. *Mol. Biol. Rep.* 47 (4), 2963–2974. doi: 10.1007/s11033-020-05384-9

Kay, P., Groszmann, M., Ross, J. J., Parish, R. W., and Swain, S. M. (2013). Modifications of a conserved regulatory network involving INDEHISCENT controls multiple aspects of reproductive tissue development in Arabidopsis. *New Phytol.* 197 (1), 73–87. doi: 10.1111/j.1469-8137.2012.04373.x

Kord, H., Shakib, A. M., Daneshvar, M. H., Azadi, P., Bayat, V., Mashayekhi, M., et al. (2015). RNAi-mediated down-regulation of SHATTERPROOF gene in transgenic oilseed rape. *3 Biotech.* 5 (3), 271–277. doi: 10.1007/s13205-014-0226-9

Laga, B., den Boer, B., and Lambert, B. (2008). *Brassica plant comprising a mutant indehiscent allele* (Patent US8809635B2, Bayer CropScience NV). Available at: patents.google.com/patent/US9475849B2/en.

Lawrenson, T., Shorinola, O., Stacey, N., Li, C., Ostergaard, L., Patron, N., et al. (2015). Induction of targeted, heritable mutations in barley and *Brassica oleracea* using RNA-guided Cas9 nuclease. *Genome Biol.* 16, 258. doi: 10.1186/s13059-015-0826-7

Lee, J. S., Kim, K. R., Ha, B.-K., and Kang, S. (2017). Identification of SNPs tightly linked to the QTL for pod shattering in soybean. *Mol. Breed.* 37 (4), 54. doi: 10.1007/s11032-017-0656-2

Lenser, T., and Theissen, G. (2013). Conservation of fruit dehiscence pathways between *Lepidium campestre* and *Arabidopsis thaliana* sheds light on the regulation of INDEHISCENT. *Plant J.* 76, 545–556. doi: 10.1111/tpj.12321

Lewis, M. W., Leslie, M. E., and Liljegren, S. J. (2006). Plant separation: 50 ways to leave your mother. *Curr. Opin. Plant Biol.* 9 (1), 59–65. doi: 10.1016/j.pbi.2005.11.009

Li, Y. L., Yu, Y. K., Zhu, K. M., Ding, L. N., Wang, Z., Yang, Y. H., et al. (2021). Down-regulation of MANNANASE7 gene in *B. napus* L. enhances silique dehiscence-resistance. *Plant Cell Rep.* 40 (2), 361–374. doi: 10.1007/s00299-020-02638-5

Liiljegren, S. J., Ditta, G. S., Eshed, Y., Savidge, B., Bowman, J. L., and Yanofsky, M. F. (2000). *SHATTERPROOF* MADS-box genes control seed dispersal in *Arabidopsis*. *Nature* 404, 766–770. doi: 10.1038/35008089

Liljegren, S. J., Roeder, A. H. K., Kempin, S. A., Gremski, K., Østergaard, L., Guimil, S., et al. (2004). Control of fruit patterning in arabidopsis by INDEHISCENT. *Cell* 116 (6), 843–853. doi: 10.1016/S0092-8674(04)00217-X

Liu, X. Y., Macmillan, R. H., Burrow, R. P., Kadkol, G. P., and Halloran, G. M. (1994). Pendulum test for evaluation of rupture strength of seed pods. *J. Texture Stud.* 25, 179–189. doi: 10.1111/j.1745-4603.1994.tb01325.x

Liu, J., Wang, J., Wang, H., Wang, W., Zhou, R., Mei, D., et al. (2016). Multigenic control of pod shattering resistance in chinese rapeseed germplasm revealed by genome-wide association and linkage analyses. *Front. Plant Sci.* 7 (1058). doi: 10.3389/fpls.2016.01058

Liu, J., Zhou, R., Wang, W., Wang, H., Qiu, Y., Raman, R., et al. (2020). A copia like-retrotransposon insertion in the upstream region of SHATTERPROOF 1 gene, BnSHP1.A9 is associated with quantitative variation in pod shattering resistance in oilseed rape. *J. Exp. Bot.* 71 (18), 5402–5413. doi: 10.1093/jxb/eraa281

Lu, K., Wei, L. J., Li, X. L., Wang, Y. T., Wu, J., Liu, M., et al. (2019). Whole-genome resequencing reveals Brassica napus origin and genetic loci involved in its improvement. *Nat. Commun.* 10. doi: 10.1038/s41467-019-09134-9

MacLeod, J. (1981). *Harvesting in oilseed rape* (Cambridge: Cambridge Agricultural Publishing), 107–120.

Maheepala, D. C., Emerling, C. A., Rajewski, A., Macon, J., Strahl, M., Pabón-Mora, N., et al. (2019). Evolution and diversification of FRUITFULL genes in solanaceae. *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.00043

Marsch-Martínez, N., Ramos-Cruz, D., Irepan Reyes-Olalde, J., Lozano-Sotomayor, P., Zúñiga-Mayo, V. M., and de Folter, S. (2012). The role of cytokinin during Arabidopsis gynoecia and fruit morphogenesis and patterning. *Plant J.* 72 (2), 222–234. doi: 10.1111/j.1365-313X.2012.05062.x

Mongkolporn, O., Kadkol, G. P., Pang, C. K., and Taylor, P. W. J. (2003). Identification of RAPD markers linked to recessive genes conferring siliqua shatter resistance in *Brassica rapa*. *Plant Breed.* 122, 479–484. doi: 10.1046/j.0179-9541.2003.00910.x

Morgan, C., Bavage, A., Bancroft, I., Bruce, D., Child, R., Chinoy, C., et al. (2007). Using novel variation in Brassica species to reduce agricultural inputs and improve agronomy of oilseed rape—a case study in pod shatter resistance. *Plant Genet. Resour.* 1 (1), 59–65. doi: 10.1079/PGR200311

Morgan, C. L., Bruce, D. M., Child, R., Ladbrooke, Z. L., and Arthur, A. E. (1998). Genetic variation for pod shatter resistance among lines of oilseed rape developed from synthetic *B. napus*. *Field Crops Res.* 58, 153–165. doi: 10.1016/S0378-4290(98)00099-9

Mühlhausen, A., Lenser, T., Mummenhoff, K., and Theißen, G. (2013). Evidence that an evolutionary transition from dehiscent to indehiscent fruits in Lepidium (Brassicaceae) was caused by a change in the control of valve margin identity genes. *Plant J.* 73 (5), 824–835. doi: 10.1111/tpj.12079

Ogawa, M., Kay, P., Wilson, S., and Swain, S. M. (2009). Arabidopsis dehiscence zone polygalacturonase1 (*ADPG1*), ADPG2, and *QUARTET2* are polygalacturonases required for cell separation during reproductive development in Arabidopsis. *Plant Cell* 21 (1), 216–233. doi: 10.1105/tpc.108.063768

Ostergaard, L., Kempin, S. A., Bies, D., Klee, H. J., and Yanofsky, M. F. (2006). Pod shatter-resistant Brassica fruit produced by ectopic expression of the FRUITFULL gene. *Plant Biotechnol. J.* 4 (1), 45–51. doi: 10.1111/j.1467-7652.2005.00156.x

Parkin, I. A. P., Gulden, S. M., Sharpe, A. G., Lukens, L., Trick, M., Osborn, T. C., et al. (2005). Segmental structure of the *Brassica napus* genome based on comparative analysis with *Arabidopsis thaliana*. *Genetics* 171 (2), 765–781. doi: 10.1534/genetics.105.042093

Petroli, C. D., Sansaloni, C. P., Carling, J., Steane, D. A., Vaillancourt, R. E., Myburg, A. A., et al. (2012). Genomic characterisation of DArT markers based on high-density linkage analysis and physical mapping to the eucalyptus genome. *PloS One* 7 (9), e44684. doi: 10.1371/journal.pone.0044684

Price, J. S., Hobson, R. N., Neale, M. A., and Bruce, D. M. (1996). Seed losses in commercial harvesting of oilseed rape. *J. Agric. Eng. Res.* 65 (3), 183–191. doi: 10.1006/jaer.1996.0091

Rajani, S., and Sundaresan, V. (2001). The Arabidopsis myc/bHLH gene ALCATRAZ enables cell separation in fruit dehiscence. *Curr. Biol.* 11 (24), 1914–1922. doi: 10.1016/S0960-9822(01)00593-0

Raman, R., Qiu, Y., Coombes, N., Song, J., Kilian, A., and Raman, H. (2017). Molecular diversity analysis and genetic mapping of pod shatter resistance loci in *Brassica carinata* L. *Front. Plant Science.* 8 (1765). doi: 10.3389/fpls.2017.01765

Raman, H., Raman, R., Coombes, N., Song, J., Prangnell, R., Bandaranayake, C., et al. (2016). Genome-wide association analyses reveal complex genetic architecture underlying natural variation for flowering time in canola. *Plant Cell Environ.* 39 (6), 1228–1239. doi: 10.1111/pce.12644

Raman, R., Raman, H., Johnstone, K., Lisle, C., Smith, A., Martin, P., et al. (2005). Genetic and in *silico* comparative mapping of the polyphenol oxidase gene in bread wheat (*Triticum aestivum* L.). *Funct. Integrated Genomics* 5, 185–200. doi: 10.1007/s10142-005-0144-3

Raman, H., Raman, R., Kilian, A., Detering, F., Carling, J., Coombes, N., et al. (2014). Genome-wide delineation of natural variation for pod shatter resistance in *B. napus*. *PloS One* 9 (7), e101673. doi: 10.1371/journal.pone.0101673

Ripoll, J. J., Roeder, A. H., Ditta, G. S., and Yanofsky, M. F. (2011). A novel role for the floral homeotic gene APETALA2 during Arabidopsis fruit development. *Development* 138 (23), 5167–5176. doi: 10.1242/dev.073031

Roeder, A. H. K., Ferrándiz, C., and Yanofsky, M. F. (2003). The role of the REPLUMLESS homeodomain protein in patterning the arabidopsis fruit. *Curr. Biol.* 13 (18), 1630–1635. doi: 10.1016/j.cub.2003.08.027

Rogers, H. J., Bate, N., Combe, J., Sullivan, J., Sweetman, J., Swan, C., et al. (2001). Functional analysis of cis-regulatory elements within the promoter of the tobacco late pollen gene g10. *Plant Mol. Biol.* 45 (5), 577–585. doi: 10.1023/A:1010695226241

Sorefan, K., Girin, T., Liljegren, S. J., Ljung, K., Robles, P., Galvan-Ampudia, C. S., et al. (2009). A regulated auxin minimum is required for seed dispersal in Arabidopsis. *Nature* 459, 583–586. doi: 10.1038/nature07875

Spence, J., Vercher, Y., Gates, P., and Harris, N. (1996). 'Pod shatter' in Arabidopsis thaliana, *Brassica napus* and *B. juncea*. *J. Microscopy* 181 (2), 195–203. doi: 10.1046/j.1365-2818.1996.111391.x

Stephenson, P., Stacey, N., Brüser, M., Pullen, N., Ilyas, M., O'Neill, C., et al. (2019). The power of model-to-crop translation illustrated by reducing seed loss from pod shatter in oilseed rape. *Plant Reprod.* 32 (4), 331–340. doi: 10.1007/s00497-019-00374-9

Takumi, S., Nishioka, E., Morihiro, H., Kawahara, T., and Matuoka, Y. (2009). Natural variation of morphological traits in wild wheat progenitor *Aegilops tauschii* Coss. *Breed. Sci.* 59, 579–588. doi: 10.1270/jsbbs.59.579

Vera, C. L., Downey, R. K., Woods, S. M., Raney, J. P., McGregor, D. I., Elliott, R. H., et al. (2007). Yield and quality of canola seed as affected by stage of maturity at swathing. *Can. J. Plant Sci.* 87 (1), 13–26. doi: 10.4141/P05-077

Wang, R., Ripley, V. L., and Rakow, G. (2007). Pod shatter resistance evaluation in cultivars and breeding lines of *Brassica napus, B. juncea* and *Sinapis alba*. *Plant Breed.* 126 (6), 588–595.

Wen, Y. C., Zhang, S. F., Yi, B., Wen, J., Wang, J. P., Zhu, J. C., et al. (2013). Identification of QTLs involved in pod-shatter resistance in *Brassica napus* L. *Crop Pasture Sci.* 63 (12), 1082–1089.

Xu N, H. G., and Guilfoyle, T. (1997). Multiple auxin response modules in the soybean SAUR 15A promoter. *Plant Sci.* 126, 193–201. doi: 10.1016/S0168-9452(97)00110-6

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 12 September 2023
DOI 10.3389/fpls.2023.1223778

OPEN ACCESS

EDITED BY
Umesh K. Reddy,
West Virginia State University, United States

REVIEWED BY
Sareena Sahab,
Department of Economic Development Jobs Transport and Resources, Australia
Manohar Chakrabarti,
The University of Texas Rio Grande Valley, United States

*CORRESPONDENCE
Fatima Chigri
fchigri@uni-bonn.de

†These authors share senior authorship

RECEIVED 16 May 2023
ACCEPTED 23 August 2023
PUBLISHED 12 September 2023

CITATION
Bhattacharyya S, Giridhar M, Meier B, Peiter E, Vothknecht UC and Chigri F (2023) Global transcriptome profiling reveals root- and leaf-specific responses of barley (*Hordeum vulgare* L.) to H_2O_2. *Front. Plant Sci.* 14:1223778. doi: 10.3389/fpls.2023.1223778

Global transcriptome profiling reveals root- and leaf-specific responses of barley (*Hordeum vulgare* L.) to H_2O_2

Sabarna Bhattacharyya[1], Maya Giridhar[1,2], Bastian Meier[3], Edgar Peiter[3], Ute C. Vothknecht[1†] and Fatima Chigri[1*†]

[1]Institute for Cellular and Molecular Botany, University of Bonn, Bonn, Germany, [2]Leibniz Institute for Food Systems Biology at the Technical University of Munich, Freising, Germany, [3]Institute of Agricultural and Nutritional Sciences, Faculty of Natural Sciences III, Martin Luther University Halle-Wittenberg, Halle, Germany

In cereal crops, such as barley (*Hordeum vulgare* L.), the ability to appropriately respond to environmental cues is an important factor for yield stability and thus for agricultural production. Reactive oxygen species (ROS), such as hydrogen peroxide (H_2O_2), are key components of signal transduction cascades involved in plant adaptation to changing environmental conditions. H_2O_2-mediated stress responses include the modulation of expression of stress-responsive genes required to cope with different abiotic and biotic stresses. Despite its importance, knowledge of the effects of H_2O_2 on the barley transcriptome is still scarce. In this study, we identified global transcriptomic changes induced after application of 10 mM H_2O_2 to five-day-old barley plants. In total, 1883 and 1001 differentially expressed genes (DEGs) were identified in roots and leaves, respectively. Most of these DEGs were organ-specific, with only 209 DEGs commonly regulated and 37 counter-regulated between both plant parts. A GO term analysis further confirmed that different processes were affected in roots and leaves. It revealed that DEGs in leaves mostly comprised genes associated with hormone signaling, response to H_2O_2 and abiotic stresses. This includes many transcriptions factors and small heat shock proteins. DEGs in roots mostly comprised genes linked to crucial aspects of H_2O_2 catabolism and oxidant detoxification, glutathione metabolism, as well as cell wall modulation. These categories include many peroxidases and glutathione transferases. As with leaves, the H_2O_2 response category in roots contains small heat shock proteins, however, mostly different members of this family were affected and they were all regulated in the opposite direction in the two plant parts. Validation of the expression of the selected commonly regulated DEGs by qRT-PCR was consistent with the RNA-seq data. The data obtained in this study provide an insight into the molecular mechanisms of oxidative stress responses in barley, which might also play a role upon other stresses that induce oxidative bursts.

KEYWORDS

barley, H_2O_2, oxidative stress, RNA-sequencing, reactive oxygen species (ROS), transcriptome profiling, stress response

1 Introduction

In aerobic organisms, reactive oxygen species (ROS) are generated as by-products of certain metabolic pathways in plant organelles such as chloroplasts, mitochondria, and peroxisomes (Huang et al., 2019; Smirnoff and Arnaud, 2019). Because of their high reactivity with cellular components, aerobic organisms have developed systems for enzymatic ROS removal based on the activity of ascorbate peroxidase (APX), superoxide dismutase (SOD), and catalase (CAT) as well as non-enzymatic antioxidative systems such as ascorbic acid, proline, and glutathione (GSH) (Foyer and Noctor, 2003; Ahmad et al., 2010). Plants also actively produce ROS as part of signaling cascades that coordinate the appropriate responses to environmental stimuli and contribute to stress tolerance (Pei et al., 2000; Zhu, 2016; Mohanta et al., 2018). It is proposed that systemic communication *via* redox systems is very fundamental to all photosynthetic organisms.

The ROS species hydrogen peroxide (H_2O_2) has been shown to play a role in various processes such as cell differentiation, senescence, and cell wall formation (Kärkönen and Kuchitsu, 2015; Ribeiro et al., 2017; Zeng et al., 2017). It is generated from superoxide in various cellular compartments as well as the apoplast as a result of a highly conserved superoxide dismutation reaction (Smirnoff and Arnaud, 2019). H_2O_2 is also known to be transported across the cell membrane by specific aquaporins (Bienert et al., 2007) and to participate in long distance cell signaling (Mittler et al., 2011). Exogenous treatment with H_2O_2 has been shown to increase the tolerance of plants to abiotic stress by regulating multiple stress-responsive pathways and expression of genes including heat shock proteins and genes involved in abscisic acid (ABA) biosynthesis (Wahid et al., 2007; Terzi et al., 2014). An activation of ROS-dependent signaling by H_2O_2 causes the accumulation of defense proteins such as ROS-scavenging enzymes, transcription factors (TFs), and other response factors (Hossain et al., 2015), and it thus increases the tolerance of plants to abiotic stress. For example, certain HEAT SHOCK TRANSCRIPTION FACTORS (HSFs) have been suggested to serve as sensors that perceive H_2O_2 and regulate the expression of oxidative stress response genes (Miller and Mittler, 2006).

An early transcriptomic approach pursued to elucidate the effect of H_2O_2 was performed in *Arabidopsis thaliana* cell suspension cultures and showed that various TFs, hormone-associated pathways, and genes associated with other vital metabolic pathways like photosynthesis and fatty acid biosynthesis were affected (Desikan et al., 2001). Other studies revealed the role of H_2O_2 as a signaling molecule in a variety of plant species and under various conditions. For instance, H_2O_2 is involved in the response of plants to a variety of environmental cues, such as salt stress in tomato (Li et al., 2019), heat stress in rice (Wang et al., 2014), chilling stress in mung beans and manila grass (Yu et al., 2003; Wang et al., 2010), copper stress in maize and mung bean (Guzel and Terzi, 2013; Fariduddin et al., 2014), and many more (Khan et al., 2018).

Barley is one of the oldest cultivated cereal crops and has a high tolerance to stresses like salt, drought, and heat (Munns et al., 2006; Rollins et al., 2013; Gürel et al., 2016). Whereas changes in the barley transcriptome upon those stresses have been analyzed (Janiak et al., 2018; Osthoff et al., 2019; Nefissi Ouertani et al., 2021), a global transcriptome analysis in response to H_2O_2 has not been performed yet.

In the present study, we used RNA sequencing (RNA-Seq) to analyze changes in the transcriptome of barley roots and leaves upon application of H_2O_2. This analysis identified a total of 1001 and 1883 differentially expressed genes (DEGs) in response to H_2O_2 in leaves and roots, respectively. Comparative and quantitative analyses of gene expression patterns revealed commonly regulated key genes related to H_2O_2 stress between both tissues, nine of which were further confirmed by qRT-PCR analysis. The data obtained in this study contribute to the understanding of molecular mechanisms of oxidative stress response in barley, which might also play a role upon other stresses that induce oxidative bursts.

2 Materials and methods

2.1 Plant material and growth conditions

Barley plants *(Hordeum vulgare* cultivar Golden Promise) were grown in pots filled with water-soaked vermiculite in a climate-controlled growth chamber under long-day conditions with 16 h light at 20°C and a light intensity of 120 µmol photons m^{-2} s^{-1} (Philips TLD 18W of alternating 830/840 light color temperature) and 8 h darkness at 18°C for five days.

2.2 H_2O_2 application and RNA isolation

Five-day-old seedlings were harvested and washed carefully to remove any remaining vermiculite prior to submersion in 10 mM H_2O_2 (Carl Roth, Germany) or ddH_2O (control) for three hours. The duration of H_2O_2 treatment was selected based on previous studies, which showed that at this time point H_2O_2 induced the strongest changes in the expression of most of the H_2O_2-responsive genes (Desikan et al., 2001; Stanley Kim et al., 2005; Hieno et al., 2019). Subsequently, seedlings were carefully rinsed with ddH_2O and dissected into roots and leaves. Samples were shock-frozen in liquid nitrogen and homogenized using a sterile, ice-cold mortar and pestle. Total RNA was extracted using the Quick-RNA miniprep Kit (Zymo Research, USA) according to the manufacturer's instructions. The yield and purity of extracted RNA was determined with a NABI Nanodrop UV/Vis Spectrophotometer (MicroDigital, South Korea). The integrity of the extracted RNA was verified by separation of the 28S and 18S rRNA bands on a 1% agarose gel.

2.3 RNA-sequencing and data analyses

RNA sequencing was performed on three biological replicates for each treatment. Each replicate furthermore consisted of pooled material from three plants. Library preparation and transcriptome sequencing (3' mRNA sequencing) were carried out at the NGS

Core Facility (Medical Faculty at the University of Bonn, Germany) using a NOVASEQ 6000 (Illumina, USA) with a read length of 1x100 bases and an average sequencing depth of >10 million raw reads per sample (Table 1). 3' end sequencing libraries were prepared using the QuantSeq protocol (Moll et al., 2014). Briefly, oligo dT priming were followed by synthesis of the complementary first strand without any prior removal of ribosomal RNA. After successful introduction of Illumina specific adapter sequences, the resulting cDNA was further purified with magnetic beads. The unpaired reads were processed for quality control using fastQC and cutAdapt (Martin, 2011) in order to trim any remaining adapter sequences. They were then aligned using Tophat2 software (Trapnell et al., 2012) against a *H. vulgare* IBSC v2 reference genome obtained from Ensembl (http://plants.ensembl.org/info/data/ftp/index.html) using a Bowtie index (Langmead and Salzberg, 2012) created with the help of the reference genome (in FASTA format; the individual FASTA files of the chromosomes were concatenated using the "cat" command in UNIX shell). The alignment with Tophat2 was performed on an Ubuntu 18.04 LTS operating system, in a UNIX shell environment. Every step after alignment was performed using R 4.0.0 (R Core Team, 2020). Gene counts from the aligned BAM files were generated using featureCounts function in RStudio (Liao et al., 2014). Differential gene expression analyses was carried out using DESeq2 (Love et al., 2014). The p-values were corrected using the False Discovery Rate (FDR) method (Benjamini and Hochberg, 1995) and subsequently the FDR and the $\log_2$FC cutoffs were set to 0.01 and 1, respectively. Principal Component Analyses (PCA) plots were prepared with the raw gene counts for all samples and replicates using the tidyverse and ggplot2 packages. The volcano plots and heatmaps were generated using the EnhancedVolcano and Pheatmap packages, respectively. In addition, transcript per million (TPM) values of each gene were calculated using a separate function designed in the R environment (Supplementary Table S1). With common regulated DEGs, a clustering was performed with four predefined clusters based on FDR and $\log_2$FC cutoffs of 0.01 and 0.5, respectively. The first and second cluster consisted of commonly down- and up-regulated genes, respectively, while the third and fourth cluster contained counter-regulated genes between leaves and roots of barley. The clusters were then represented as heatmaps using the pheatmap package and line plots using the ggpubr package.

Gene ontology (GO) and enrichment analyses were carried out using shinyGO (Ge et al., 2020). Categories were chosen as significant if the FDR was less than 0.05 (Benjamini and Hochberg, 1995). Homology searches against the *A. thaliana* genome were carried out using the BaRT (Barley Reference Transcript) tool available on www.ics.hutton.ac.uk (Mascher et al., 2017) based on a E-value cutoff of $1e^{-30}$.

2.4 Quantification of transcript levels by qRT-PCR

qRT-PCR was performed with three replicates for each sample. Each replicate consisted of the pooled RNA material from three different plants. Synthesis of first strand cDNA for qRT-PCR was carried out from at least 1 µg of total RNA using the RevertAid first strand cDNA synthesis kit (Thermo Fisher Scientific, USA) with oligo-dT_{18} primers following the manufacturer's instructions. The quality of cDNA was assessed using a NABI UV/Vis Nanodrop Spectrophotometer. Gene expression was quantified in 48-well plates using a BioRad CFX 96 real-time PCR detection system (BioRad, Germany) and a SYBR Green PCR master mix (Thermo Fisher Scientific, USA). All forward and reverse primers used for qRT-PCR are listed in Supplementary Table S2. Data were quantified using the BioRad CFX Maestro software, and the

TABLE 1 Summary of total reads and aligned reads in the RNA-seq samples from barley roots and leaves obtained under H_2O_2 treatment and control conditions.

Sample	Replicate	Total Reads	Aligned Reads	% Aligned Reads
root control	RC1	15222810	12333400	81.02
	RC2	13555021	10223311	75.42
	RC3	12544002	9988003	79.62
leaf control	LC1	12392862	9242908	74.58
	LC2	14067426	10125991	71.98
	LC3	12314839	9224084	74.90
root + H_2O_2	RT1	12123370	8559783	70.61
	RT2	13079745	9303393	71.13
	RT3	12698432	10154310	79.97
leaf + H_2O_2	LT1	13222658	11555866	87.39
	LT2	14555200	12333012	84.73
	LT3	12220331	10214419	83.59

For each treatment three biological replicates were performed, each containing the combined RNA from three plants. LC-Leaf control, LT-Leaf H_2O_2 treated, RC-Root control, and RT-Root H_2O_2 treated.

expression was estimated using the $2^{-\delta\delta Ct}$ method (Livak and Schmittgen, 2001) after normalization against the two reference genes *HvACTIN* and *HvGAPDH*, as the C_q values of both genes were unchanged upon H_2O_2 treatment. Data were analyzed statistically with one-way analysis of variance (ANOVA) and Tukey' Post-Hoc HSD test using the agricolae and tidyverse packages, respectively. Graphs were prepared using the ggpubr package.

2.5 H_2O_2 staining and microscopic analyses

Staining of hydrogen peroxide in barley leaves and roots was performed with 2',7'-dichlorodihydrofluorescein diacetate (H_2-DCFDA; Thermo Fisher Scientific, USA) based on a modified protocol (Kaur et al., 2016). Briefly, five-day-old barley seedlings were treated with either 10 mM H_2O_2 or ddH_2O (control) for 3 hours. Afterwards, the seedlings were briefly rinsed and treated with 10 µM H_2-DCFDA prepared from a 4 mM stock dissolved in DMSO for 1 hour in the dark. After staining, seedlings were washed, and roots and leaves were mounted separately on a microscopy slide. 2',7'-Dichlorfluorescein (DCF) fluorescence was analyzed using a Leica SP8 Lightning confocal laser scanning microscope (Leica Microsystems, Germany). For excitation, an argon laser with a wavelength of 488 nm was used, and emission of 517-527 nm was detected using a HyD Detector. Fluorescence intensity was quantified in regions of interest (ROI) using the integrated LASX software.

3 Results

3.1 Differential gene expression in leaves and roots of barley in response to application of H_2O_2

To investigate the transcriptomic modulation in barley (*Hordeum vulgare* cv. Golden Promise) in response to oxidative stress, five-day-old plants were exposed for three hours to 10 mM H_2O_2 or to ddH_2O as control (Figure 1A). H_2-DCFDA staining confirmed that H_2O_2 penetrated both roots and leaves (Figures 1B, C and Supplementary Figure 1). RNA was then extracted separately from roots and leaves, and RNA-seq analysis was carried out on three biological replicates per tissue and treatment, each comprising the pooled RNA from three different plants (Supplementary Table S1). On average approximately 13 million total reads were obtained per sample. About 75-85% of these reads could be aligned to the

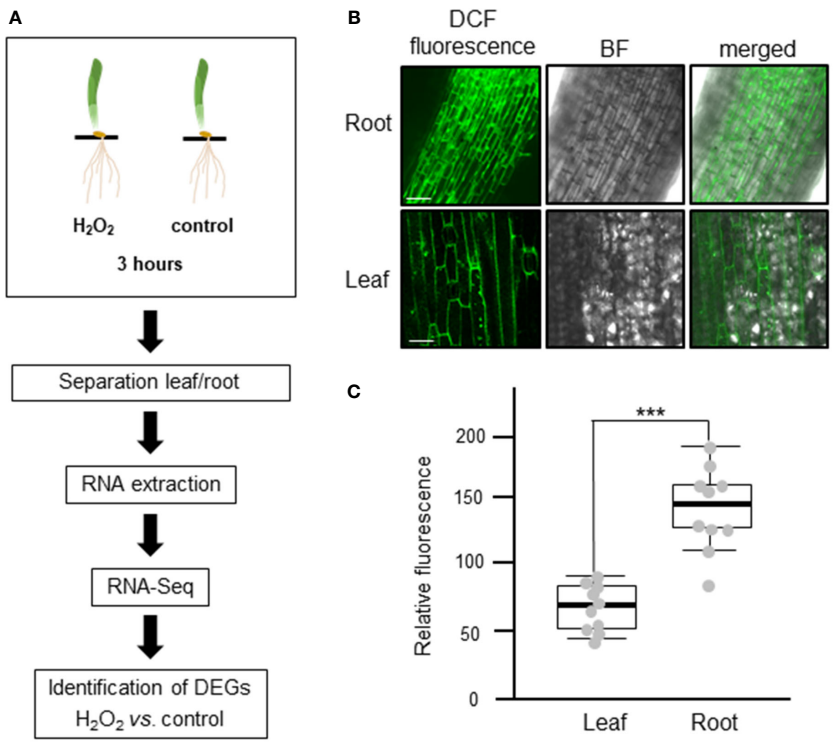


FIGURE 1
Experimental design to analyze the transcriptional changes of barley plants to oxidative stress. **(A)** Schematic representation of the study design. Five-day-old barley plants were treated with either 10 mM H_2O_2 or water (control) for three hours. After the treatment, leaves and roots were separated, RNA was extracted, and three independent biological replicates, each containing the pooled RNA from three plants, were submitted to RNA-Seq analyses. The raw reads obtained were subjected to quality control and aligned against the barley reference genome. Based on raw gene counts, a differential expression analysis was carried out using DESeq2. **(B)** Uptake of H_2O_2 in roots (upper panel) and leaves (lower panel) visualized by H_2-DCFDA. Green fluorescence of the 2',7'-Dichlorfluorescein (DCF) was observed using a Leica SP8 lightning confocal laser scanning microscope. BF: bright field; bar: 100 µm. **(C)** Quantification of fluorescence intensity of H_2-DCFDA relative to untreated control tissues. Each dot represents the average of five regions of interests (ROIs). ROIs were taken from two independent images from three biological replicates (n=6). Statistical analysis was carried out using the two-tailed t-test (*** = P<0.001).

barley reference genome (Table 1). To assess the main variances within the dataset, a principal component analysis (PCA) was performed. The result showed that PC1 (X-axis), which separates the samples by tissue, represents the largest variation in our dataset compared to PC2 (Y-axis), which separates the samples by treatment (Figure 2A). Consequently, the differential gene expression analysis was separately performed for the leaf and root samples.

Differentially expressed genes (DEGs) between H_2O_2-treated and control samples were identified based on fold change (FC) $|Log_2FC \geq 1|$ and FDR < 0.01 (Supplementary Table S3). A total number of 2884 DEGs were detected across both tissues. H_2O_2 application clearly resulted in stronger transcriptional changes in roots compared to leaves (Figure 2B). Of the 1883 DEGs detected in roots, 701 were up- and 1182 were down-regulated, while in leaves 1001 DEGs were identified with 546 up- and 455 down-regulated (Figure 2C). Among all DEGs only 75 and 134 were commonly up- and down-regulated, respectively, in both tissues, while 37 were counter-regulated.

3.2 Gene ontology analyses

GO classification was used to identify the 20 most significant biological process categories within the DEGs. The results show that not only the number of genes, but also the biological processes affected by H_2O_2 were clearly different between leaves and roots (Figure 3). In leaves, GO terms associated with genes that showed the highest fold change were related to protein complex oligomerization, response to H_2O_2 and jasmonate. Further categories with lower fold change but often higher number of genes comprised quite global stress effects associated with different, mostly abiotic stimuli, but also wounding (Figure 3A). In roots, many of the enriched GOs were associated with response to oxygenic stress including H_2O_2 catabolism, glutathione and ROS metabolism, or cellular oxidant detoxification as well as with cell wall modulation (Figure 3B).

3.2.1 Differentially expressed genes in barley leaves in response to H_2O_2

In barley leaves, the most highly enriched GO term category upon exposure to H_2O_2 was the response to H_2O_2 and protein complex oligomerization (Figure 3A). Both categories consist of the same SMALL HEAT SHOCK PROTEINS (SHSP domain-containing proteins) (Table 2). SHSPs are ubiquitous in prokaryotic and eukaryotic organisms and function as chaperone proteins involved in the response to many abiotic stresses (Basha et al., 2012; Waters, 2013). Their expression levels were shown in different plant species to increase upon stress and to enhance stress tolerance. Here, barley leaves exposed to H_2O_2 showed an increased

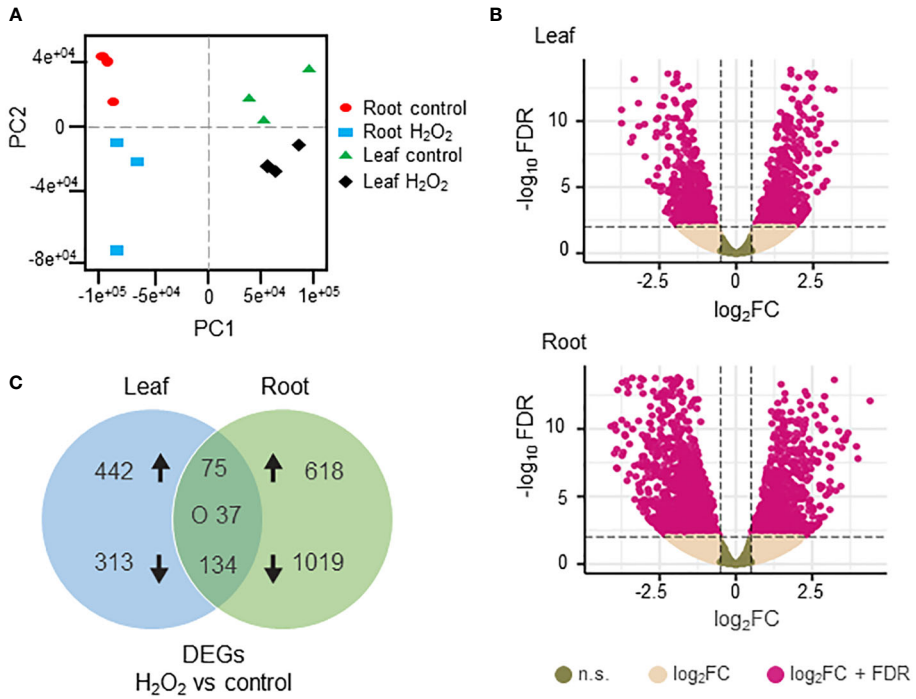


FIGURE 2
Differentially expressed genes (DEGs) in H_2O_2-treated and untreated barley plants. **(A)** Principal component analysis of the RNAseq data showing the homogeneity of the different samples. PC1 (X axis) separates the samples by tissue while PC2 (Y axis) separates the samples by treatment. **(B)** Volcano plots of the DEGs in leaves (upper panel) and roots (lower panel). The X axis represents the fold change (Log_2FC) of the DEGs (H_2O_2 *vs.* control), whereas the Y axis represents the statistical significance ($log_{10}FDR$). Pink dots indicate genes that fit the DESeq criteria of FDRand $|Log_2FCin|$, while beige dots represent DEGs that fit only Log_2FC. N.S.: not significant **(C)** Venn diagram representing DEGs (DESeq, adjusted to FDR<0.01 and $|Log_2FC| \geq 1$) between H_2O_2-treated and untreated samples in leaves and roots. Arrows indicate up- and down-regulation. 'O' indicates counter-regulated genes.

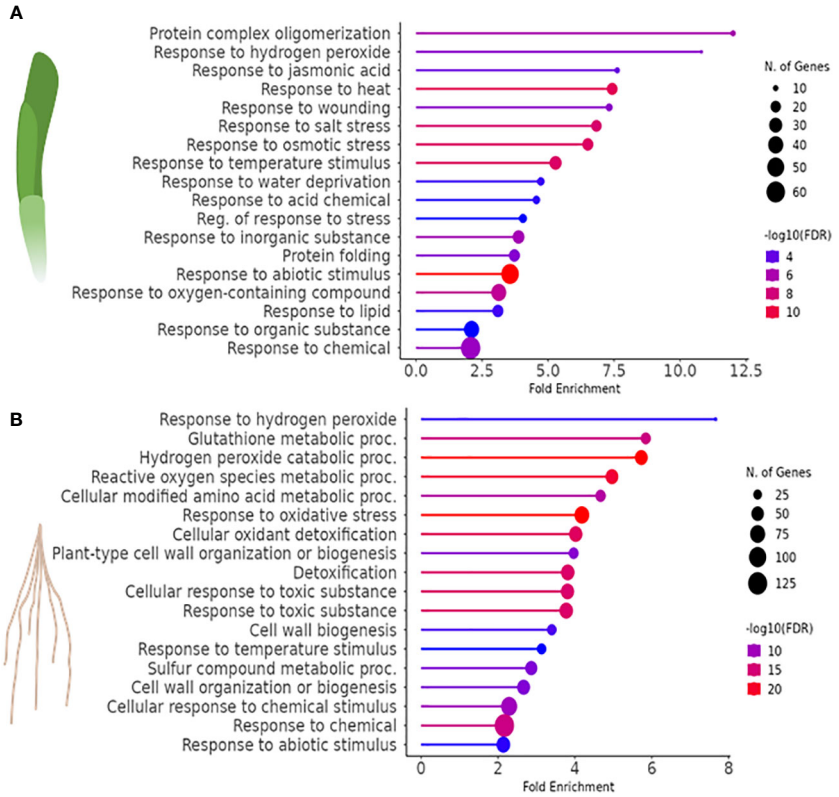


FIGURE 3
Gene ontology (GO) enrichment analysis to identify biological processes associated with the DEGs (FDR<0.01) of H_2O_2-treated *vs.* control samples in **(A)** leaves and **(B)** roots of barley.

expression of SHSPs, except for the 18.8 kDa class V heat shock protein (HORVU2Hr1G046370), which was down-regulated. All of the differentially regulated SHSPs have close orthologs in Arabidopsis (Li and Liu, 2019) with the majority being orthologous to *AtHSP17.6II* (At5g12020).

An enrichment was also found for genes involved in hormone biosynthesis and signaling, especially jasmonate, auxin, and abscisic acid (ABA). Jasmonate-related DEGs were represented by the specific GO-term category 'response to jasmonic acid'. This category comprised two up-regulated TIFY domain-containing proteins with no direct homologs in Arabidopsis (Table 2). The TIFY domain is also known as ZIM domain which is present in members of the transcriptional repressor JASMONATE ZIM-domain (JAZ) family, key elements in the jasmonate signaling pathway (Chung and Howe, 2009; Pauwels and Goossens, 2011). This category also includes genes that encode for enzymes involved in jasmonate biosynthesis (Schaller and Stintzi, 2009; Bittner et al., 2022) such as ALLENE OXIDE CYCLASE (*AOC*), and OXOPHYTODIENOATE-REDUCTASE (*OPR*) as well as ALLENE OXIDE SYNTHASE (*AOS*) but with a FC less than 2 (FC 1.69, $Log_2FC=0.76$). By contrast, genes related to other hormone signaling pathways were found redundantly interspersed in the two GO terms 'response to abiotic stimulus' and 'response to salt stress' (Figure 3A). With regard to auxin, a number of orthologs to auxin-responsive genes from Arabidopsis, especially IAA-type TFs, were found. Similar to the jasmonate signaling pathway, H_2O_2 seems to affect the auxin pathway differentially since both, up- and down-regulated DEGs, were identified. All components related to the phytohormone ABA were up-regulated and those related to APETALA2/ETHYLENE RESPONSIVE FACTOR (*AP2/ERF*) domain-containing proteins, known to be involved in abiotic stress responses and associated with various hormones, were down-regulated. Similar to the GO term categories related to auxin, both sets comprise mostly orthologs to TFs or co-regulators known in Arabidopsis (Table 2).

In leaves, genes associated with photosynthesis light harvesting in photosystem I, were also affected, however, the category did not appear in the top GOs since for several of the genes the FC was less than 2 but mostly higher than 1.5 (Table 2; Log_2FC between 0.5 and 1). This category contained mostly down-regulated DEGs, including several orthologs of Arabidopsis LHCII trimer components, i.e., genes encoding for LHCb1 and LHCb3, and the LHCa1 protein. It furthermore comprised orthologs to the photosystem I subunits PSAF and PSAL but also the oxygen evolving complex subunit PSBP-1 and the large subunit of RIBULOSE-1,4-BISPHOSPHATE-CARBOXYLASE/OXYGENASE (Rubisco) (Table 2).

3.2.2 Differentially expressed genes in barley roots in response to H_2O_2

In barley roots, the most enriched GO terms are associated with response to oxidative stress and detoxification (Figure 3B). This is

TABLE 2 Selected DEGs associated with top GO terms in leaves of barley in response to H_2O_2.

Category	*Gene ID*	*log_2FC*	*Functional protein*	*Predicted ortholog in A. thaliana*
Response to H_2O_2/ protein complex oligomerization				
	HORVU2Hr1G046370	-3.74	SHSP domain-containing protein	AT4G21870 (AtHSP15.4)
	HORVU3Hr1G020500	2.24	SHSP domain-containing protein	AT5G12020/AT5g12030 (AtHSP17.6)
	HORVU3Hr1G020490	3.03	SHSP domain-containing protein	AT5G12020/AT5g12030 (AtHSP17.6)
	HORVU3Hr1G020390	1.267	SHSP domain-containing protein	AT5G12020/AT5g12030 (AtHSP17.6)
	HORVU0Hr1G020420	1.54	SHSP domain-containing protein	AT5G37670 (AtHSP15.7)
	HORVU3Hr1G020520	1.84	SHSP domain-containing protein	AT5G12020/AT5g12030 (AtHSP17.6)
	HORVU6Hr1G082360	2.98	SHSP domain-containing protein	At1G54050 (AtHSP17.4)
Response to jasmonic acid				
	HORVU5Hr1G062290	2.34	TIFY domain-containing protein	AT1G74950 (AtJAZ12)
	HORVU4Hr1G076850	1.80	TIFY domain-containing protein	no homolog
	HORVU5Hr1G098090	1.21	Uncharacterized protein	AT1G13280 (AtAOC4)
	HORVU7Hr1G118010	-1.44	Oxidored FMN domain-containing	AT1G76680 (AtOPR1)
	HORVU2Hr1G004230	-1.55	Oxidored FMN domain- containing	AT1G76690 (AtOPR2)
	HORVU6Hr1G081000	0.76	Allene oxide synthase	AT5G42650 (AtCYP74A/AtAOS)
Response to abiotic stimulus/ osmotic stress/ hormones				
Auxin	HORVU7Hr1G084940	1.81	Auxin responsive protein	AT4G14550 (AtIAA14/AtSLR)
	HORVU5Hr1G087880	1.48	Auxin responsive protein	AT5G65980 (AtPILS7)
	HORVU7Hr1G033820	1.22	Auxin responsive protein	AT1G19220 (AtARF19)
	HORVU1Hr1G086070	1.00	Auxin responsive protein	no homolog
	HORVU1Hr1G086070	1.00	Auxin responsive protein	no homolog
	HORVU6Hr1G058890	-1.52	Auxin response factor	AT4G30080 (AtARF16)
	HORVU7Hr1G077110	-1.62	Auxin responsive protein	no homolog
	HORVU5Hr1G093580	-2.40	Auxin responsive protein	AT3G04730 (AtIAA16)
Abcisic acid	HORVU7Hr1G085130	2.34	Multiple protein bridging factor	AT3G24500 (AtMBF1c)
	HORVU7Hr1G035500	1.58	bZIP domain-containing protein	AT4G34000 (AtABF3/AtbZIP37)
	HORVU3Hr1G069590	1.37	HSF_domain-containing protein	AT3G24520 (AtHsfC1)

(Continued)

TABLE 2 Continued

Category	Gene ID	log_2FC	Functional protein	Predicted ortholog in A. thaliana
	HORVU6Hr1G028790	1.30	WRKY domain-containing protein	AT4G31800 (AtWRKY18)
	HORVU5Hr1G115100	1.03	GRAM domain-containing protein	At5G13200 (AtGEML5/AtGER5/AtGRE5)
other	HORVU5Hr1G097560	1.62	HTH MYB domain-containing protein	AT2G38090
	HORVU3Hr1G085180	1.26	MYB domain-containing protein	no homolog
	HORVU6Hr1G091700	-1.13	Ethylene receptor domain-containing protein	AT3G04580 (AtEIN4)
	HORVU4Hr1G077310	-1.31	AP2/ERF domain-containing protein	no homolog
	HORVU4Hr1G000700	-1.92	AP2/ERF domain-containing protein	AT3G23240 (AtERF092/AtERF1b)
	HORVU3Hr1G010190	-3.31	AP2/ERF domain-containing protein	AT1G68840 (AtEDF2/AtRAV2/AtTEM2)
Photosynthesis				
	HORVU6Hr1G091660	-1.67	Chlorophyll a-b binding protein	AT2G34420 (AtLHCb1.5)
	HORVU1Hr1G088920	-1.37	Chlorophyll a-b binding protein	AT2G34420 (AtLHCb1.5)
	HORVU7Hr1G040370	-1.16	Chlorophyll a-b binding protein	AT2G34420 (AtLHCb1.5)
	HORVU6Hr1G047870	-1.11	Ribulose bisphosphate carboxylase LSU	ATCG00490 (RubisCo LSU)
	HORVU5Hr1G109250	-1.07	Chlorophyll a-b binding protein	AT1G29930 (AtLHCb1.3)
	HORVU5Hr1G109260	-0.93	Chlorophyll a-b binding protein	AT2G34420 (AtLHCb1.5)
	HORVU2Hr1G040780	-0.92	Chlorophyll a-b binding protein	AT5G54270 (AtLHCb3)
	HORVU1Hr1G078380	-0.91	Chlorophyll a-b binding protein	AT2G34420 (AtLHCb1.5)
	HORVU2Hr1G060880	-0.87	PsbP domain-containing protein	AT1G06680 (AtPsP1)
	HORVU5Hr1G100140	-0.81	PSI-F	AT1G31330 (AtPsaF)
	HORVU7Hr1G046320	-0.72	Chlorophyll a-b binding protein	AT3G54890 (AtLHCa1)
	HORVU3Hr1G009210	-0.71	PSI subunit V	AT4G12800 (AtPsaL)
	HORVU1Hr1G088870	-0.68	Chlorophyll a-b binding protein	AT2G34430 (AtLHCb1.4)

also evident by the fact that many DEGs within those GO terms are class-III peroxidases, catalases, or genes related to glutathione metabolism, which were grouped together as a category named 'Detoxification of H_2O_2' (Table 3). In plants, class-III peroxidases have been described in association with a wide variety of biotic and abiotic stresses along with plant defense mechanisms (Almagro et al., 2009; Shigeto and Tsutsumi, 2016). While most peroxidases were up-regulated, some were down-regulated along with a number of glutathione transferases, an ascorbate peroxidase (*APX*), and *CATALASE 1*. We also found strong up-regulation of the genes for two putative detoxification efflux carriers/multidrug and toxic compound extrusion (*DTX/MATE*) transporters. These metabolite transporters have been described to be associated with plant stress responses and overexpression of a gene encoding a cotton DXT

TABLE 3 Selected DEGs associated with top GO terms in roots of barley in response to H_2O_2.

Category	*Gene ID*	*log_2FC*	*Functional annotation*	*Predicted ortholog in* A. thaliana
Response to H_2O_2				
	HORVU0Hr1G020420	-1.21	SHSP domain containing protein	AT5G37670 (AtHSP15.7)
	HORVU2Hr1G077710	-1.59	SHSP domain containing protein	AT4G10250 (AtHSP22)
	HORVU3Hr1G006940	-2.24	SHSP domain containing protein	No ortholog
	HORVU3Hr1G020390	-1.92	SHSP domain containing protein	AT5G12020 (AtHSP17.6II)
	HORVU3Hr1G020490	-2.79	SHSP domain containing protein	AT5G12020 (AtHSP17.6II)
	HORVU3Hr1G020520	-2.96	SHSP domain containing protein	AT5G12020 (AtHSP17.6II)
	HORVU4Hr1G015170	-3.2	SHSP domain containing protein	AT4G10250 (AtHSP22)
	HORVU4Hr1G060720	-1.34	SHSP domain containing protein	AT3G46230 (AtHSP17.4)
	HORVU4Hr1G060760	-2.88	SHSP domain containing protein	AT1G53540 (AtHSP17.6C)
	HORVU6Hr1G008640	-2.55	Catalase	AT1G20630 (AtCAT1)
	HORVU7Hr1G014870	-1.54	ABC transporter domain containing protein	AT1G31770 (AtABCG14)
Detoxification of H_2O_2				
H_2O_2 catabolism	HORVU7Hr1G039550	3.97	Peroxidase	AT1G05260 (AtPRX3)
	HORVU2Hr1G026640	3.65	Peroxidase	AT1G05260 (AtPRX3)
	HORVU7Hr1G010280	3.598	Peroxidase	AT4G11290 (AtPRX39)
	HORVU1Hr1G016730	2.96	Peroxidase	AT2G18140 (AtPRX14)
	HORVU2Hr1G018550	2.91	Peroxidase	AT5G05340 (AtPRX52)
	HORVU7Hr1G039590	2.74	Peroxidase	AT1G05260 (AtPRX3)
	HORVU2Hr1G018530	2.60	Peroxidase	AT5G05340 (AtPRX52)
	HORVU7Hr1G039570	2.21	Peroxidase	AT1G05260 (AtPRX3)
	HORVU0Hr1G002840	2.17	Peroxidase	AT4G11290 (AtPRX39)
	HORVU2Hr1G100610	2.07	Peroxidase	AT5G17820 (AtPRX57/AtPRXR10)
	HORVU1Hr1G016770	2.01	Peroxidase	AT4G11290 (AtPRX39)
	HORVU2Hr1G026590	1.93	Peroxidase	AT4G11290 (AtPRX39)
	HORVU2Hr1G026520	1.84	Peroxidase	AT4G11290 (AtPRX39)

(Continued)

TABLE 3 Continued

Category	Gene ID	log_2FC	Functional annotation	Predicted ortholog in A. thaliana
	HORVU2Hr1G026540	1.83	Peroxidase	AT4G11290 (AtPRX39)
	HORVU6Hr1G026600	1.67	Peroxidase	AT5G05340 (AtPRX52)
	HORVU7Hr1G039560	1.52	Peroxidase	AT1G05260 (AtPRX3)
	HORVU1Hr1G016870	-1.84	Peroxidase	AT5G66390 (AtPRX72/AtPRXR8)
	HORVU2Hr1G124930	-1.99	Peroxidase	AT1G71695 (AtPRX12/AtPRXR6)
	HORVU4Hr1G022280	-2.15	Peroxidase	AT5G05340 (AtPRX52)
Glutathione metabolism	HORVU6Hr1G063830	-1.47	Glutathione peroxidase	AT4G11600 (AtGPX6/AtGPXL6)
	HORVU5Hr1G006330	-1.17	Glutathione transferase	no homolog
	HORVU1Hr1G049230	-1.28	Glutathione transferase	AT2G29470 (AtGSTU3)
	HORVU1Hr1G021140	-1.36	Glutathione transferase	AT3G62760 (AtGSTF13)
	HORVU6Hr1G011120	-2.16	GST_C terminal domain-containing protein	AT4G19880
	HORVU5Hr1G006330	-1.17	Glutathione transferase	no homolog
	HORVU1Hr1G049070	-2.86	GST_N terminal domain-containing protein	AT1G10370 (AtGSTU17)
Response to ROS / Detoxification	HORVU4Hr1G057170	-1.31	APX domain-containing protein	AT1G07890 (AtAPX1/AtC3H)
	HORVU6Hr1G008640	-2.55	Catalase	AT1G20630 (AtCAT1)
	HORVU4Hr1G011690	2.26	DTX/MATE metabolite transporter	AT3G26590 (AtDTX29)
	HORVU0Hr1G022350	-4.09	DTX/MATE metabolite transporter	AT5G52450 (AtDTX16)
Cell wall				
	HORVU4Hr1G028720	2.70	Xyloglucan endotransglucosylase/ hydrolase	AT5G13870 (AtXTH5/AtXTR12)
	HORVU2Hr1G010800	2.37	ExpansinA11	AT1G20190 (AtEXPA11)
	HORVU3Hr1G116470	2.07	Pectin acetylesterase	no homolog
	HORVU3Hr1G016820	2.04	Xyloglucan endotransglucosylase/ hydrolase	AT5G57550 (AtXTH25)
	HORVU2Hr1G120100	1.47	Endoglucanase	AT1G48930 (AtGH9C1/AtCEL6)
	HORVU3Hr1G016800	1.44	Xyloglucan endotransglucosylase/ hydrolase	AT5G57550 (AtXTH25)
	HORVU5Hr1G118270	1.43	Cellulose synthase	AT5G64740 (AtCESA6/AtIRX2)
	HORVU7Hr1G093680	1.27	Expansin	AT4G38210 (AtEXPA20)

(Continued)

TABLE 3 Continued

Category	Gene ID	log_2FC	Functional annotation	Predicted ortholog in A. thaliana
	HORVU7Hr1G098370	1.55	Xyloglucan endotransglycosylase	AT4G25810 (AtXTH23/AtXTR6)
	HORVU3Hr1G091360	257	Pectin esterase	AT5G09760 (AtPME51)

protein in Arabidopsis reduced stress-induced levels of H_2O_2 (Lu et al., 2019).

As in leaves, the most highly enriched GO term category in roots upon exposure to H_2O_2 was the response to H_2O_2, albeit with very few genes (Figure 3B). Similar to leaves, this category includes several SHSP domain-containing proteins, but in contrast to leaves, they were down-regulated (Table 3). All of the differentially regulated SHSPs have close orthologs in Arabidopsis, with several of them being orthologous to AtHSP17.6. This category contains also down-regulated catalase and ABC transporter containing domain proteins.

H_2O_2 treatment also induced up-regulation of components of cell wall biogenesis and modulation, such as xyloglucan endotransglucosylase/hydrolase, expansin, endo-1,4-beta glucanase, pectin acetyl esterase, and cellulose synthase (Table 3) that were found interspersed in several GO term categories. Indeed, H_2O_2 and peroxidases were shown to be involved in cell wall remodeling upon environmental stress (Tenhaken, 2015).

3.3 Common DEGs of leaves and roots in response to H_2O_2

As described above, we identified a total of 246 common DEGs between leaves and roots of barley when using a $|\log_2FC \geq 1|$ cutoff (Supplementary Table S3, Figure 2C). For several genes, we noticed that they were differentially regulated in both tissues, however, in one tissue they showed an expression with a FC>2 ($|\log_2FC \geq 1|$) while in the other tissue a FC less than 2 but higher as 1.5 Thus, for ($|\log_2FC$ between 1 and 0.5$|$) was detected. determination of commonly regulated genes in leaves and roots we used a cutoff of $Log_2FC \geq 0.5$ and listed these genes separately in Supplementary Table S3. Using this cut-off, a total 349 common DEGs were identified between roots and leaves of barley (Supplementary Figure S2; Supplementary Table S3). Of these, 116 and 176 genes were up- and down-regulated, respectively, while 58 genes showed counter-regulation. These common DEGs were organized in four clearly distinguishable clusters (Figure 4A), with either commonly down- (cluster 1) and up-regulated (cluster 2) genes or genes up-regulated in leaves but down-regulated in roots (cluster 3) and *vice versa* (cluster 4). Heat maps and line plots were constructed to visualize the changes in gene expression pattern for each cluster (Figures 4A, B).

3.3.1 Commonly up- and down-regulated genes

Cluster 1 contains DEGs commonly down-regulated in leaves and roots upon H_2O_2 treatment (Supplementary Table S3), among them members of important transcription factors such as *AP2/ERF*, *WRKY*, *CBF1*, *NAC*, and *HD-ZIP HOMEOBOX* (Supplementary Table S4, Figure 5A). Cluster 1 also comprises orthologs to the Arabidopsis sugar transporters *SWEET10* and *SWEET5*. Other transporters were orthologs to the phosphate transporter *PHT1;7* and the aquaporin *TIP4;1*. TIP aquaporins in plants had been shown to not only transport water molecules but also other molecules like H_2O_2 (Kurowska et al., 2020). In addition to components of oxidative stress, detoxification or cell wall biogenesis and modification that were already discussed in chapter 3.2.2, cluster 1 also contained several kinases including orthologs to the *CYSTEINE-RICH RECEPTOR-LIKE PROTEIN KINASES* (CRKs), *CRK29* and *CRK25*. *CRKs* are presented in Arabidopsis by a large gene family with over 40 members and have been associated with various abiotic and biotic stresses (Bourdais et al., 2015).

Cluster 2 contains DEGs commonly up-regulated in leaves and roots (Supplementary Table S3). Interestingly, it contains TFs of similar families as cluster 1, like *WRKY* and *AP2/ERF* but also orthologs of the LOB DOMAIN CONTAINING PROTEIN 41 (*LBD41*) from Arabidopsis (Supplementary Table S4; Figure 5B). DEGs associated with primary metabolism like amino acid and nucleic acid metabolism were also found in cluster 2. Genes associated with primary metabolism were also shown to be up-regulated in other transcriptome studies associated with abiotic stress (Hirai et al., 2004; Wang et al., 2014) and DEGs found in cluster 2 do not seem to be related to any specific metabolic pathway. Two MITOGEN-ACTIVATED PROTEIN KINASEs (MAPKs) identified in cluster 2 are orthologs to *AtMAPKKK16* and *AtMAPKKK17*, both of which were shown to be regulated by ABA (Wang et al., 2011).

3.3.2 Counter-regulated genes

Cluster 3 consists of 42 DEGs up-regulated in leaves and down-regulated in roots of barley upon H_2O_2 treatment (Supplementary Table S3). Nine of these DEGs are orthologs to different small heat shock proteins from Arabidopsis (Supplementary Table S4; Figure 6). The cluster furthermore comprises an assorted set of genes whose orthologs in Arabidopsis are connected with various metabolic pathways and hormone signaling.

Cluster 4 consists of only 15 genes and no common functional categories were found (Supplementary Table S4). However, they include genes, whose Arabidopsis orthologs have been associated with hormones, or cell wall modification, i.e. the *COPPER-CONTAINING AMINE OXIDASE 3 (CUAO3)* that was suggested to be involved in stress response since it was up-regulated upon treatment with several hormones or flagellin (Planas-Portell et al., 2013).

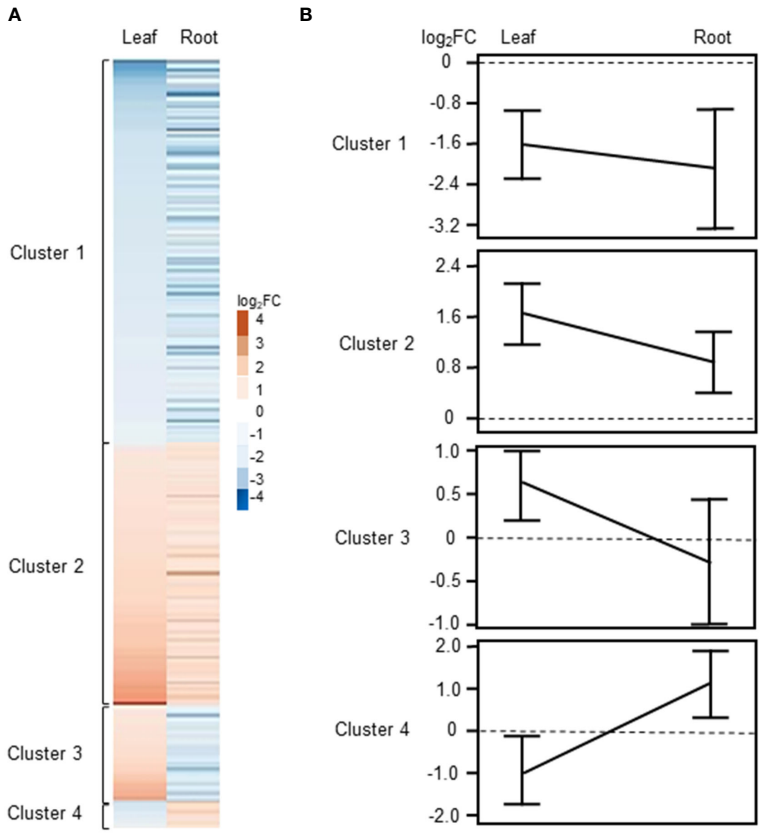


FIGURE 4
Clustering of DEGs commonly regulated or counter-regulated in leaves and roots of barley upon H_2O_2 treatment ($|Log_2FC| \geq 0.5$ and FDR<0.01). **(A)** Heat map showing the Log_2FC associated with each gene in leaves and roots. **(B)** Line plot showing the mean ± SE of the $|Log_2FC|$ associated with each cluster in leaves and roots.

Overall, clusters 3 and 4 show very few genes previously described to be associated with oxidative stress.

3.4 qRT-PCR confirmation of selected DEGs

In order to confirm the results obtained from RNA-seq analyses, we performed quantitative RT-PCRs (qRT-PCR) on some of the identified DEGs. For these, we selected several DEGs that showed common regulation in leaves and roots in our dataset and which, based on their functional annotation, could be related to oxidative stress (Supplementary Table S5). Orthologs to some of them had already been shown to play an important role in H_2O_2 and ROS-related signaling not only in Arabidopsis but also in important crops like wheat, maize, and rice (Polidoros et al., 2005; Mylona et al., 2007; Steffens, 2014; Dudziak et al., 2019). They also represent different levels of regulation, some being among the most highly up- or down-regulated genes and other showing a much more subtle response. These genes represent different gene ontologies, and encode for a catalase, a peroxidase, a glutathione S-transferase, several TFs, a MAPKKK, and a xyloglucan endotransglucosyalase, a protein involved in cell wall modification. As shown in Figure 7 and in Supplementary Table S5, the log_2FC changes observed with the different techniques were often quite close and, in all cases, the results of the qRT-PCR matched the trend observed in the RNA-seq data.

4 Discussion

In plants, H_2O_2 is a crucial ROS which plays a dual role as a harmful by-product of cell metabolism and as a secondary messenger that affects development and growth. Complex cross-talk between H_2O_2 and other signaling molecules, such as Ca^{2+} ions and hormones, plays a key role in regulating different biological processes that contribute to the response to various biotic and abiotic stresses (Peiter, 2016; Saxena et al., 2016). Despite its importance, very little is known about H_2O_2-induced changes of the transcriptome in barley. In this study, an analysis of the barley transcriptome in response to H_2O_2 was performed using next generation sequencing. First, a suitable concentration of H_2O_2 that was shown to initiate a stress response in barley was selected on basis of previously performed experiments (Dodd et al., 2010; Giridhar et al., 2022). An increase in cytosolic Ca^{2+} ($[Ca^{2+}]_{cyt}$) is one of the first responses of plants to most biotic and abiotic stresses (Dodd et al., 2010) that in turn leads to downstream stimulus-specific cellular responses. H_2O_2 was shown to induce such

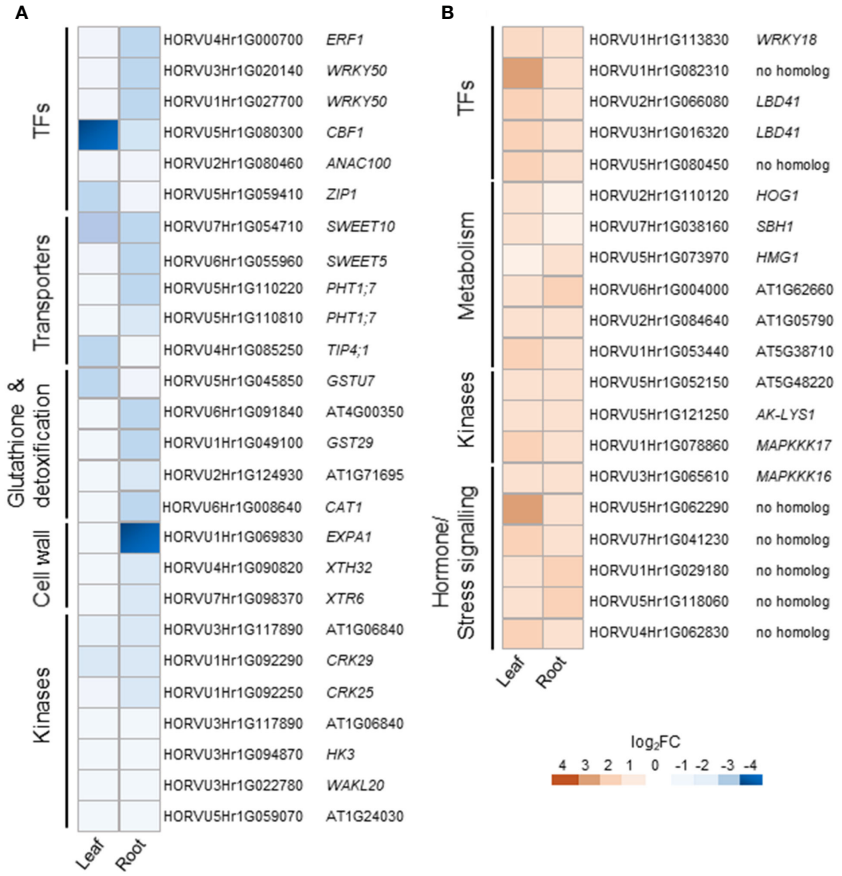


FIGURE 5
Selected DEGs commonly regulated in leaves and roots of barley upon H_2O_2 treatment. Down-regulated **(A)** and up-regulated **(B)** genes are grouped by functional category and presented with their Arabidopsis orthologs. TFs, transcription factors.

transient changes of $[Ca^{2+}]_{cyt}$ with 10 mM eliciting the highest response in barley roots and leaves (Giridhar et al., 2022). Staining of intact plants with the ROS indicator H_2-DCFDA confirmed that the exogenously applied H_2O_2 penetrated into both organs (Figures 1B, C, Supplementary Figure 1). To exclude natural degradation of RNA and changes of the transcriptome driven by processes such as senescence or tillering, five-day-old barley plants were used. Growth of monocotyledonous leaves is initiated from the base and the leaf blade shows developmental gradients, i.e., disappearance of poly (A+) RNA levels along the developing blade (Hellmann et al., 1995). Moreover, plant senescence is a natural process known to be initiated by ROS that in turn activates transcription factors interacting with senescence associated genes (Bieker et al., 2012; Shimakawa et al., 2020). Thus, the growth conditions and plant age used in the analysis ensure as much as possible a solely treatment-dependent change of the transcriptome.

Overall, the RNA-seq analysis showed that under the chosen conditions H_2O_2 caused more transcriptional changes in roots compared to leaves (Figure 2). Most of the identified DEGs were found exclusively in one of the two plant parts, further confirming organ-specific responses. While this difference may be in part due to a difference in H_2O_2 penetration into roots and leaves, it is more likely caused by differential response of the two tissues to H_2O_2 signals and/or oxidative stress. Only about 10% of the DEGs were found to be up- and down-regulated in leaves as well as in roots, some of which showed counter-regulation. This difference in response is also mirrored by the GO terms associated with the identified DEGs that only showed a minor overlap (Figure 3).

4.1 Leaf-specific transcriptomic changes in response to H_2O_2

Our data showed that several genes encoding for small heat shock proteins (SHSPs) were up-regulated by H_2O_2 in barley leaves (Table 2). In barley, the roles of several HSPs in response to a diverse range of abiotic stimuli have been characterized (Hlaváčková et al., 2013; Chaudhary et al., 2019; Landi et al., 2019). HSPs have also been shown to play crucial roles during abiotic stresses such as cold and heat in other important crop genera, like rice, maize, and wheat (ul Haq et al., 2019). SHSPs are a subgroup of HSPs defined by their size and a conserved α-crystalline C-terminal domain. They are known to form oligomeric complexes and prevent denatured proteins from aggregation until they can be refolded by other HSPs. They have been speculated to interact with transcription factors of the HEAT SHOCK FACTOR (HSF) family to create the HSP-HSF complex, alteration of which can drive essential reactions in response to ROS

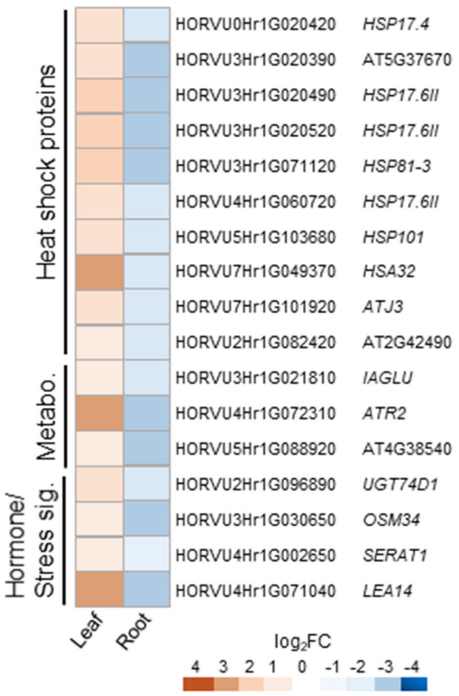


FIGURE 6
Selected counter-regulated DEGs in leaves and roots upon H_2O_2 treatment. Genes up-regulated in leaves and down-regulated in roots are grouped by functional category and presented with their Arabidopsis orthologs. Metabo., metabolism; sig., signaling.

(Driedonks et al., 2015). The SHSPs in our data set belong to subfamilies with close orthologs in Arabidopsis, i.e. HSP17.6, 15.4, 15.7, and 17.4 (Li and Liu, 2019). HSP17.6 and HSP15.7 have been shown to be localized in the peroxisomes in Arabidopsis (Ma et al., 2006; Li et al., 2017). Peroxisomes are one of the main subcellular compartments in which ROS are produced by processes such as ß-oxidation and photorespiration, and which are crucial for antioxidant defense (Sandalio et al., 2013; del Río and López-Huertas, 2016). Additionally, *HSP17.4* and *17.6* have been shown to exhibit increased transcript levels during periods of abiotic stress in Arabidopsis (Swindell et al., 2007). Thus, the induction of these HSPs points to a potential role of these proteins in increasing the tolerance to oxidative stress also in barley leaves. The single down-regulated SHSP is an ortholog to *AtHSP15.4*, for which this contrary behavior upon stress was already described (Siddique et al., 2008).

Not surprising, considering the well-established juxtaposition between ROS production and photosynthesis, the application of H_2O_2 negatively affected several photosynthetic components (Table 2). The most affected group represents chlorophyll a/b binding proteins orthologous to various light-harvesting complex proteins of the LHCb-type and to a component of the light-harvesting complex I, LHCa1, of Arabidopsis. Down-regulation of LHCb-type proteins upon oxidative stress has been previously described (Staneloni et al., 2008). It is likely part of an established photoprotection mechanism to alleviate increased ROS levels generated when the photosynthesis reaction becomes unbalanced, e.g., under high light conditions.

The role of phytohormones like ABA and jasmonate in response to several biotic and abiotic stimuli has been extensively studied in plants (Verma et al., 2016). In our data, several genes related to jasmonate signaling were found to be down-regulated (Table 2), including an ortholog of Arabidopsis *12-OXOPHYTODIENOATE REDUCTASE* (*OPR*). The OPR3 protein of Arabidopsis has been denoted as one of the most crucial enzymes in jasmonate synthesis, which converts 12-oxophytodieonic acid (*cis*-OPDA) to OPC8:0 in peroxisomes (Bittner et al., 2022). However, recent studies highlighted the role of an OPR3-independent pathway for jasmonic acid (JA) biosynthesis, involving an OPR2-mediated alternative bypass *via* dinor-OPDA (dnOPDA) and 4,5-didehydro-JA, which is then converted to JA (Chini et al., 2018). Interestingly, we found a down-regulation of the barley ortholog of *OPR2* in leaves, the consequence of which remains speculative due to the unclear role of the OPR3-independent bypass pathway. By contrast, genes coding for *ALLENE OXIDE CYCLASE* (*AOC*) and *ALLENE OXIDE SYNTHASE* (*AOS*) were up-regulated in leaves. These enzymes catalyze the generation of both *cis*-OPDA and dnOPDA, which in turn would increase OPDA production for both pathways. This is interesting, because OPDA is believed to have an independent regulatory function both on transcription (similar to JA-Ile), but also on protein activity by OPDadylation. Moreover, OPDA-mediated signaling seems closely associated with thiol metabolism and redox-mediated processes (Böttcher and Weiler, 2007; Ohkama-Ohtsu et al., 2011; Bittner et al., 2022). Also related to jasmonate signaling are two TIFY domain-containing proteins that were induced in response to H_2O_2 (Table 2). The TIFY domain is found in members of the JASMONATE ZIM DOMAIN (JAZ)-type transcriptional repressors involved in jasmonate signaling (Chung and Howe, 2009; Pauwels and Goossens, 2011). However, no regulation of TFs associated with jasmonate signaling was detected in our data set.

By contrast, many of the genes associated with other phytohormones, e.g. auxins and ABA, encode TFs or other proteins involved in transcription regulation (Table 2). Several of these genes belong to the large family of AP2/ERF-type TFs, members of which have been associated with environmental stresses including hypoxia and oxidative stress. While mostly associated with ethylene, AP2/ERF function is also connected to ABA, gibberellic acid, cytokinin, and brassinosteroids (Xie et al., 2019). The largest group of genes associated with hormones relates to auxin (Table 2), the role of which is mostly associated with development and growth. However, experimental evidence linked auxin also to oxidative stress, especially auxin-mediated stress-dependent cell proliferation including the RSL-type TF ROOT HAIR DEFECTIVE SIX-LIKE4 (RSL4) that targets NADPH oxidases also known as respiratory burst oxidase homologs (RBOHs) and secreted plant-specific type III peroxidases that impact apoplastic ROS homeostasis and in turn stimulate root hair cell elongation (Pasternak et al., 2005; Iglesias et al., 2010; Mangano et al., 2017).

4.2 Root-specific transcriptomic changes in response to H_2O_2

In roots, many DEGs were found to be associated with the detoxification of H_2O_2 (Table 3), especially peroxidases and genes related to glutathione metabolism. *GLUTATHIONE TRANSFERASES (GSTs)* and *GLUTATHIONE PEROXIDASES*

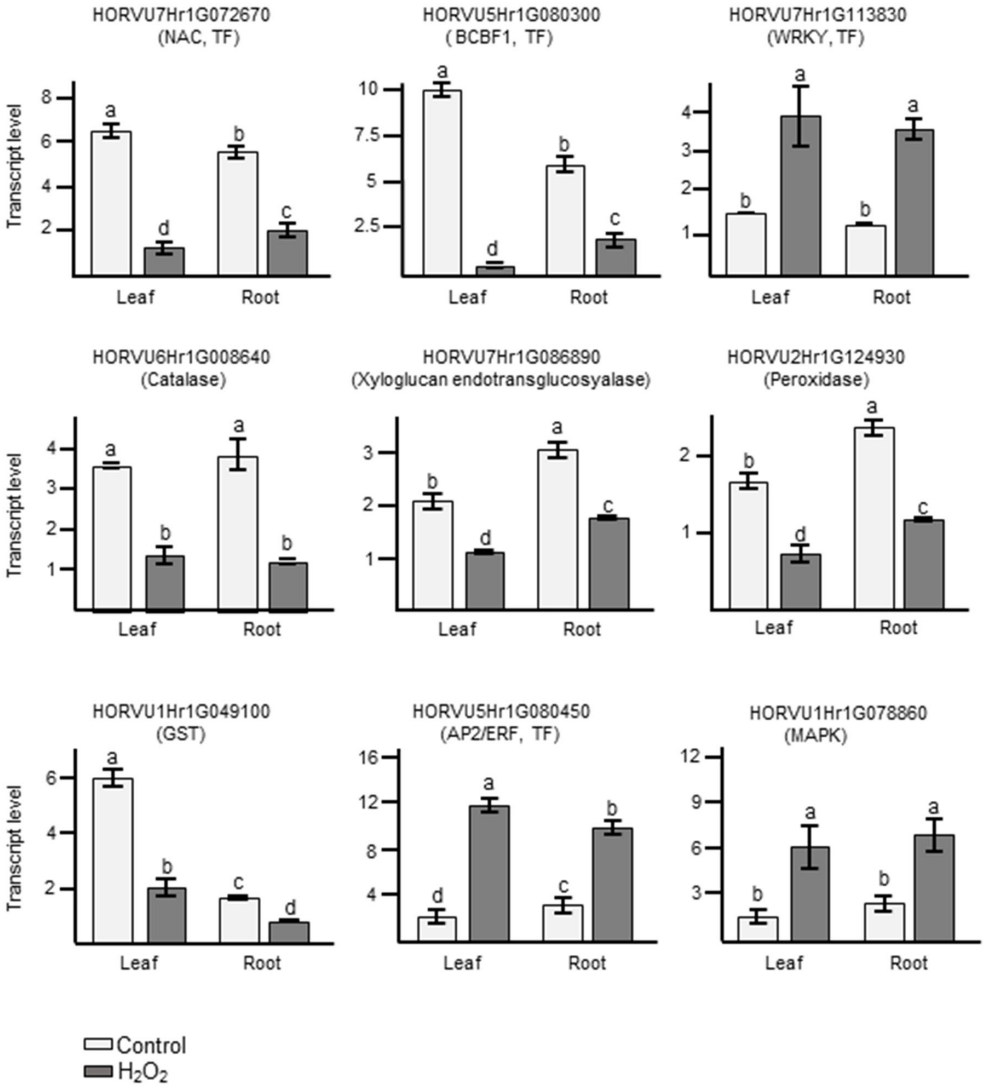


FIGURE 7
Analyses of transcript levels for selected candidate genes by qRT-PCR. Data represent means ± SE of three biological replicates (n=3), each having two technical repeats. Transcript levels were normalized to *HvACTIN* and *HvGAPDH*. Letters represent significant differences estimated using one-way ANOVA and Tukey's Post-Hoc HSD test (P<0.05). Potential functions of the genes inferred from orthologous genes in Arabidopsis are indicated in brackets.

(GTPs) have both been shown to be involved in plant stress responses (Bela et al., 2015; Nianiou-Obeidat et al., 2017). However, somewhat surprisingly, our data showed clear down-regulation of several *GSTs* and *GTPs* along with other key players associated with H_2O_2 detoxification such as orthologs of Arabidopsis *ASCORBATE PEROXIDASE* 1 (*APX1*) and *CATALASE 1(CAT1)*. Moreover, two putative *DETOXIFICATION EFFLUX CARRIERS/MULTIDRUG AND TOXIC COMPOUND EXTRUSION* (DXT/MATE) proteins were strongly up-regulated in roots. The MATE family proteins facilitate the efflux of various compounds including substances, such as hormones or flavonoids, that improve adaptation to stress (Ku et al., 2022).

The largest set of genes whose expression was affected in response to H_2O_2 belongs to class III plant type peroxidases (Table 3), whose role in plant defense mechanisms in response to a wide variety of biotic and abiotic stresses is well established. They play an important role in the cellular redox homeostasis upon stress. In addition, they also catalyze the oxidation of a variety of substrates and have been linked to processes involved in cell wall stability, including lignin and suberin polymerization in response to stress (Kidwai et al., 2020). Thus, the up-regulation of these peroxidases in roots upon H_2O_2 treatment is in line with the up-regulation of genes involved in cell wall metabolism observed in this study. Some components of the cell wall architecture, particularly the xyloglucans, have been shown to play an important role in imparting abiotic stress tolerance by coordinating with hormonal and other signaling cascades. For example, a xyloglucan galactosyl transferase from Arabidopsis, SHORT ROOT IN SALT MEDIUM 3 (RSA3), was shown to play a crucial role under salt stress by assembling actin microfilaments and thus preventing ROS

accumulation induced by disruption of actin microfilaments (Cho et al., 2006; Li et al., 2013). Also the role of xyloglucan modifying enzymes along with expansins in loosening and expanding the cell wall network upon abiotic stresses has already been described (Tenhaken, 2015).

4.3 Commonly and counter-regulated DEGs in responses to H_2O_2

Overall, leaves and roots showed very unique transcriptional responses upon H_2O_2 treatment. Not only the number of DEGs was much higher in roots compared to leaves, the change in transcription also affected a quite different set of genes (Figures 2, 3). Nevertheless, there are DEGs that were found in both plant parts (Figure 4). These 349 DEGs were further divided into four clusters, depending on their expression pattern. Looking at the two larger clusters, the commonly up- or down-regulated DEGs (Figure 5, Supplementary Table S3 and S4), certain patterns in the functional categories can be observed. Both clusters include TFs from different families. This is not unexpected and highlights their versatility in differentially regulating genes as an important part of all stress responses (Javed et al., 2020). However, of the TFs identified in this study, only few have previously been associated with oxidative stress, such as an Arabidopsis ortholog to *HORVU2Hr1G066080* and *HORVU3Hr1G016320*, the *LOB DOMAIN CONTAINING PROTEIN 41 (LBD41)*, that was previously identified in relation with low-oxygen endurance or high-light-induced increase in H_2O_2 (Mustroph et al., 2009; Vanderauwera et al., 2011). However, some were found associated with stresses, such as herbivory, that include ROS-mediated signaling or mutations that cause increased levels of ROS (Paudel et al., 2013; Garcia et al., 2016).

Several transporters were found commonly down-regulated (Supplementary Table S4 and Figure 5A). The aquaporin encoded by HORVU4Hr1G085250 is orthologous to the *TONOPLAST INTRINSIC PROTEIN 4;1 (TIP4;1)* of Arabidopsis and rice. Aquaporins not only transport water but also other molecules including H_2O_2. *TIP4;1* from barley was shown to be up-regulated by ABA in roots and gibberellic acid in shoots (Ligaba et al., 2011). Moreover, its expression was also up-regulated upon drought (Kurowska et al., 2019). Also sugar transporters of the SWEET-type and PHT1.7 phosphate transporters have been demonstrated to play a role in abiotic stress tolerance and showed variable expression patterns under stress conditions (Cao et al., 2020; Gautam et al., 2022).

We also found common down-regulation of orthologs to *RECEPTOR-LIKE PROTEIN KINASES(RLKs)* from different subfamilies, i.e., WAK, LLR, CRK and RLCK (Supplementary Table S4 and Figure 5A). Experimental evidence suggests that RKLs are a vital part of the growth-defense trade-off, i.e. by facilitating the cross-talk between different phytohormones (Zhu et al., 2023). However, of the specific *RLKs* found commonly down-regulated in barley leaves and roots, only the pepper ortholog of *WAKL20* was described in relation to stress (Zhu et al., 2023). DEGs connected to various facets of primary metabolism were found commonly up-regulated (Supplementary Table S4 and Figure 5B). While several of them are involved in pathways that play a role in stress responses, an obvious connection between these specific DEGs is lacking. Overall, even if no clear connection to oxidative stress exists, many of the commonly regulated DEGs have been described or postulated previously to be involved in stress tolerance mechanisms.

A very small number of DEGs was found counter-regulated upon treatment with H_2O_2 (Supplementary Table S4 and Figure 6), the majority of which showing up-regulation in leaves and down-regulation in roots. Several of those genes are connected to aspects of metabolism and hormone signaling, and some orthologous genes of other plant species, such as *SERAT1*, *OSM34*, and *UGT74D1* of tomato, grapevine and Arabidopsis have been previously connected to stress, ABA signaling, or auxin (Tavares et al., 2015; Jin et al., 2021; Park and Kim, 2021; Liu et al., 2022). Remarkably, this cluster also includes a group of nine *HSPs*, and this different expression in leaves and roots raises questions about their specific role in stress response in the different tissues.

5 Conclusions

Plant adaptation to changing environmental cues requires acclimation, enabling them to fulfil their lifecycle. This adaptation is based to a large extent on substantial changes on transcriptional level. Our data reveal that H_2O_2 modulates the expression of a wide range of genes within the barley genome. The results provide first insights into the significant role of H_2O_2 in altering cellular activities in this important crop species. However, in which manner all these genes are coordinated within the cell to provide an appropriate response during stress-induced H_2O_2 increase is an important question that needs to be addressed in further research. Many of them have previously been associated to stress responses in barley or more often *via* their orthologs in Arabidopsis or other crops. This reveals a high degree of similarity in the responses of these plants to situations where cellular H_2O_2 levels increase either as a toxic by-product of stress or as a dedicated signaling molecule. Other genes identified in this screen have so far not been associated with stress. As important redox molecules participating in plant cell signaling, developmental processes stress responses, as well as causing oxidative damage, uncovering the effect of ROS generally and H_2O_2 specifically on gene expression provides good insights into the molecular mechanisms of oxidative stress responses in barley. Such understanding might increase our ability to improve stress resistance in barley and other crops to optimize crop performance and productivity in present and future environmental climate challenges. Particularly, the highest up- or down-regulated genes in our dataset in both tissues were mostly uncharacterized and information on the exact nature of the genes is missing. These data can be used to guide future studies aimed to functionally characterize novel stress-related genes using state-of-the-art experimental designs including generation of mutants and ectopic expression lines. This will enable us to better understand H_2O_2 mediated regulation of adaptive processes not only in barley but also in other crops and might thus support targeted breeding of more resilient crops.

Data availability statement

The datasets presented in this study can be found in online repositories (https://www.ncbi.nlm.nih.gov/sra/PRJNA973626).

Author contributions

SB contributed to conceptualization, investigation (responsible for most experimental work), formal analysis (responsible for all bioinformatic analysis), validation, visualization, and writing - original draft as well as review & editing. MG contributed to investigation. BM contributed to validation (qRT-PCR) and writing - review & editing. EP contributed to supervision and writing - review and editing. UV contributed to conceptualization, validation, funding acquisition, project administration, supervision, and writing - review & editing. FC contributed to conceptualization, formal analysis, validation, visualization, supervision, and writing - original draft as well as review & editing. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, INST 217/939-1 FUGG to UV and GRK 2064 to MG and UV).

Acknowledgments

We would like to thank the NSG Core Facility of the Medical Faculty at the University of Bonn for providing support. We would also like to thank Elena Ulland Rodriguez for technical 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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1223778/full#supplementary-material

References

Ahmad, P., Jaleel, C. A., Salem, M. A., Nabi, G., and Sharma, S. (2010). Roles of enzymatic and nonenzymatic antioxidants in plants during abiotic stress. *Crit. Rev. Biotechnol.* 30, 161–175. doi: 10.3109/07388550903524243

Almagro, L., Gómez Ros, L. V., Belchi-Navarro, S., Bru, R., Ros Barceló, A., and Pedreño, M. A. (2009). Class III peroxidases in plant defence reactions. *J. Exp. Bot.* 60, 377–390. doi: 10.1093/jxb/ern277

Basha, E., O'Neill, H., and Vierling, E. (2012). Small heat shock proteins and α-crystallins: dynamic proteins with flexible functions. *Trends Biochem. Sci.* 37, 106–117. doi: 10.1016/j.tibs.2011.11.005

Bela, K., Horváth, E., Gallé, Á., Szabados, L., Tari, I., and Csiszár, J. (2015). Plant glutathione peroxidases: Emerging role of the antioxidant enzymes in plant development and stress responses. *J. Plant Physiol.* 176, 192–201. doi: 10.1016/j.jplph.2014.12.014

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc Ser. B Methodol.* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Bieker, S., Riester, L., Stahl, M., Franzaring, J., and Zentgraf, U. (2012). Senescence-specific Alteration of Hydrogen Peroxide Levels in Arabidopsis thaliana and Oilseed Rape Spring Variety Brassica napus L. cv. MozartF. *J. Integr. Plant Biol.* 54, 540–554. doi: 10.1111/j.1744-7909.2012.01147.x

Bienert, G. P., Møller, A. L. B., Kristiansen, K. A., Schulz, A., Møller, I. M., Schjoerring, J. K., et al. (2007). Specific aquaporins facilitate the diffusion of hydrogen peroxide across membranes*. *J. Biol. Chem.* 282, 1183–1192. doi: 10.1074/jbc.M603761200

Bittner, A., Cieśla, A., Gruden, K., Lukan, T., Mahmud, S., Teige, M., et al. (2022). Organelles and phytohormones: a network of interactions in plant stress responses. *J. Exp. Bot.* 73, 7165–7181. doi: 10.1093/jxb/erac384

Böttcher, C., and Weiler, E. W. (2007). cyclo-Oxylipin-galactolipids in plants: occurrence and dynamics. *Planta* 226, 629–637. doi: 10.1007/s00425-007-0511-5

Bourdais, G., Burdiak, P., Gauthier, A., Nitsch, L., Salojärvi, J., Rayapuram, C., et al. (2015). Large-scale phenomics identifies primary and fine-tuning roles for CRKs in responses related to oxidative stress. *PloS Genet.* 11, e1005373. doi: 10.1371/journal.pgen.1005373

Cao, M., Liu, H., Zhang, C., Wang, D., Liu, X., and Chen, Q. (2020). Functional analysis of stPHT1;7, a solanum tuberosum L. Phosphate transporter gene, in growth and drought tolerance. *Plants* 9, 1384. doi: 10.3390/plants9101384

Chaudhary, R., Baranwal, V. K., Kumar, R., Sircar, D., and Chauhan, H. (2019). Genome-wide identification and expression analysis of Hsp70, Hsp90, and Hsp100 heat shock protein genes in barley under stress conditions and reproductive development. *Funct. Integr. Genomics* 19, 1007–1022. doi: 10.1007/s10142-019-00695-y

Chini, A., Monte, I., Zamarreño, A. M., Hamberg, M., Lassueur, S., Reymond, P., et al. (2018). An OPR3-independent pathway uses 4,5-didehydrojasmonate for jasmonate synthesis. *Nat. Chem. Biol.* 14, 171–178. doi: 10.1038/nchembio.2540

Cho, S. K., Kim, J. E., Park, J.-A., Eom, T. J., and Kim, W. T. (2006). Constitutive expression of abiotic stress-inducible hot pepper CaXTH3, which encodes a xyloglucan endotransglucosylase/hydrolase homolog, improves drought and salt tolerance in transgenic Arabidopsis plants. *FEBS Lett.* 580, 3136–3144. doi: 10.1016/j.febslet.2006.04.062

Chung, H. S., and Howe, G. A. (2009). A critical role for the TIFY motif in repression of jasmonate signaling by a stabilized splice variant of the JASMONATE ZIM-domain protein JAZ10 in arabidopsis. *Plant Cell* 21, 131–145. doi: 10.1105/tpc.108.064097

del Río, L. A., and López-Huertas, E. (2016). ROS generation in peroxisomes and its role in cell signaling. *Plant Cell Physiol.* 57, 1364–1376. doi: 10.1093/pcp/pcw076

Desikan, R., A.-H.-Mackerness, S., Hancock, J. T., and Neill, S. J. (2001). Regulation of the arabidopsis transcriptome by oxidative stress. *Plant Physiol.* 127, 159–172. doi: 10.1104/pp.127.1.159

Dodd, A. N., Kudla, J., and Sanders, D. (2010). The language of calcium signaling. *Annu. Rev. Plant Biol.* 61, 593–620. doi: 10.1146/annurev-arplant-070109-104628

Driedonks, N., Xu, J., Peters, J. L., Park, S., and Rieu, I. (2015). Multi-level interactions between heat shock factors, heat shock proteins, and the redox system regulate acclimation to heat. *Front. Plant Sci.* 6. doi: 10.3389/fpls.2015.00999

Dudziak, K., Zapalska, M., Börner, A., Szczerba, H., Kowalczyk, K., and Nowak, M. (2019). Analysis of wheat gene expression related to the oxidative stress response and signal transduction under short-term osmotic stress. *Sci. Rep.* 92743. doi: 10.1038/s41598-019-39154-w

Fariduddin, Q., Khan, T. A., and Yusuf, M. (2014). Hydrogen peroxide mediated tolerance to copper stress in the presence of 28-homobrassinolide in Vigna radiata. *Acta Physiol. Plant* 36, 2767–2778. doi: 10.1007/s11738-014-1647-0

Foyer, C. H., and Noctor, G. (2003). Redox sensing and signalling associated with reactive oxygen in chloroplasts, peroxisomes and mitochondria. *Physiol. Plant* 119, 355–364. doi: 10.1034/j.1399-3054.2003.00223.x

Garcia, L., Welchen, E., Gey, U., Arce, A. L., Steinebrunner, I., and Gonzalez, D. H. (2016). The cytochrome c oxidase biogenesis factor AtCOX17 modulates stress responses in Arabidopsis. *Plant Cell Environ.* 39, 628–644. doi: 10.1111/pce.12647

Gautam, T., Dutta, M., Jaiswal, V., Zinta, G., Gahlaut, V., and Kumar, S. (2022). Emerging roles of SWEET sugar transporters in plant development and abiotic stress responses. *Cells* 11, 1303. doi: 10.3390/cells11081303

Ge, S. X., Jung, D., and Yao, R. (2020). ShinyGO: a graphical gene-set enrichment tool for animals and plants. *Bioinformatics* 36, 2628–2629. doi: 10.1093/bioinformatics/btz931

Giridhar, M., Meier, B., Imani, J., Kogel, K.-H., Peiter, E., Vothknecht, U. C., et al. (2022). Comparative analysis of stress-induced calcium signals in the crop species barley and the model plant Arabidopsis thaliana. *BMC Plant Biol.* 22, 447. doi: 10.1186/s12870-022-03820-5

Gürel, F., Öztürk, Z. N., Uçarlı, C., and Rosellini, D. (2016). Barley genes as tools to confer abiotic stress tolerance in crops. *Front. Plant Sci.* 7. doi: 10.3389/fpls.2016.01137

Guzel, S., and Terzi, R. (2013). Exogenous hydrogen peroxide increases dry matter production, mineral content and level of osmotic solutes in young maize leaves and alleviates deleterious effects of copper stress. *Bot. Stud.* 54, 26. doi: 10.1186/1999-3110-54-26

Hellmann, A., Meyer, C. U., and Wernicke, W. (1995). Tubulin gene expression during growth and maturation of leaves with different developmental patterns. *Cell Motil.* 30, 67–72. doi: 10.1002/cm.970300108

Hieno, A., Naznin, H. A., Inaba-Hasegawa, K., Yokogawa, T., Hayami, N., Nomoto, M., et al. (2019). Transcriptome analysis and identification of a transcriptional regulatory network in the response to H2O2. *Plant Physiol.* 180, 1629–1646. doi: 10.1104/pp.18.01426

Hirai, M. Y., Yano, M., Goodenowe, D. B., Kanaya, S., Kimura, T., Awazuhara, M., et al. (2004). Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in Arabidopsis thaliana. *Proc. Natl. Acad. Sci.* 101, 10205–10210. doi: 10.1073/pnas.0403218101

Hlaváčková, I., Vítámvás, P., Šantrůček, J., Kosová, K., Zelenková, S., Prášil, I. T., et al. (2013). Proteins Involved in Distinct Phases of Cold Hardening Process in Frost Resistant Winter Barley (Hordeum vulgare L.) cv Luxor. *Int. J. Mol. Sci.* 14, 8000–8024. doi: 10.3390/ijms14048000

Hossain, M. A., Bhattacharjee, S., Armin, S.-M., Qian, P., Xin, W., Li, H.-Y., et al. (2015). Hydrogen peroxide priming modulates abiotic oxidative stress tolerance: insights from ROS detoxification and scavenging. *Front. Plant Sci.* 6. doi: 10.3389/fpls.2015.00420

Huang, H., Ullah, F., Zhou, D.-X., Yi, M., and Zhao, Y. (2019). Mechanisms of ROS regulation of plant development and stress responses. *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.00800

Iglesias, M. J., Terrile, M. C., Bartoli, C. G., D'Ippólito, S., and Casalongué, C. A. (2010). Auxin signaling participates in the adaptative response against oxidative stress and salinity by interacting with redox metabolism in Arabidopsis. *Plant Mol. Biol.* 74, 215–222. doi: 10.1007/s11103-010-9667-7

Janiak, A., Kwasniewski, M., Sowa, M., Gajek, K., Żmuda, K., Kościelniak, J., et al. (2018). No time to waste: transcriptome study reveals that drought tolerance in barley may be attributed to stressed-like expression patterns that exist before the occurrence of stress. *Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.02212

Javed, T., Shabbir, R., Ali, A., Afzal, I., Zaheer, U., and Gao, S.-J. (2020). Transcription factors in plant stress responses: challenges and potential for sugarcane improvement. *Plants* 9, 491. doi: 10.3390/plants9040491

Jin, S., Hou, B., and Zhang, G. (2021). The ectopic expression of Arabidopsis glucosyltransferase UGT74D1 affects leaf positioning through modulating indole-3-acetic acid homeostasis. *Sci. Rep.* 11, 1154. doi: 10.1038/s41598-021-81016-x

Kärkönen, A., and Kuchitsu, K. (2015). Reactive oxygen species in cell wall metabolism and development in plants. *Mem. G Paul Bolwell Plant Cell Wall Dyn.* 112, 22–32. doi: 10.1016/j.phytochem.2014.09.016

Kaur, N., Dhawan, M., Sharma, I., and Pati, P. K. (2016). Interdependency of Reactive Oxygen Species generating and scavenging system in salt sensitive and salt tolerant cultivars of rice. *BMC Plant Biol.* 16, 131. doi: 10.1186/s12870-016-0824-2

Khan, T. A., Yusuf, M., and Fariduddin, Q. (2018). Hydrogen peroxide in regulation of plant metabolism: Signalling and its effect under abiotic stress. *Photosynthetica* 56, 1237–1248. doi: 10.1007/s11099-018-0830-8

Kidwai, M., Ahmad, I. Z., and Chakrabarty, D. (2020). Class III peroxidase: an indispensable enzyme for biotic/abiotic stress tolerance and a potent candidate for crop improvement. *Plant Cell Rep.* 39, 1381–1393. doi: 10.1007/s00299-020-02588-y

Ku, Y.-S., Cheng, S.-S., Cheung, M.-Y., and Lam, H.-M. (2022). The roles of multidrug and toxic compound extrusion (MATE) transporters in regulating agronomic traits. *Agronomy* 12, 878. doi: 10.3390/agronomy12040878

Kurowska, M. Małgorzata, Fahad, S., Saud, S., Chen, Y., Wu, C., and Wang, D. (2020). "TIP aquaporins in plants: role in abiotic stress tolerance," in *Abiotic stress in plants*(RijekaIntechOpen). doi: 10.5772/intechopen.94165

Kurowska, M. M., Wiecha, K., Gajek, K., and Szarejko, I. (2019). Drought stress and re-watering affect the abundance of TIP aquaporin transcripts in barley. *PloS One* 14, e0226423. doi: 10.1371/journal.pone.0226423

Landi, S., Capasso, G., Ben Azaiez, F. E., Jallouli, S., Ayadi, S., Trifa, Y., et al. (2019). Different Roles of Heat Shock Proteins (70 kDa) During Abiotic Stresses in Barley (Hordeum vulgare) Genotypes. *Plants* 8, 248. doi: 10.3390/plants8080248

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

Li, W., Guan, Q., Wang, Z.-Y., Wang, Y., and Zhu, J. (2013). A bi-functional xyloglucan galactosyltransferase is an indispensable salt stress tolerance determinant in arabidopsis. *Mol. Plant* 6, 1344–1354. doi: 10.1093/mp/sst062

Li, X., Li, Y., Ahammed, G. J., Zhang, X.-N., Ying, L., Zhang, L., et al. (2019). RBOH1-dependent apoplastic H2O2 mediates epigallocatechin-3-gallate-induced abiotic stress tolerance in Solanum lycopersicum L. *Revisiting Role ROS RNS Plants Change Environ.* 161, 357–366. doi: 10.1016/j.envexpbot.2018.11.013

Li, G., Li, J., Hao, R., and Guo, Y. (2017). Activation of catalase activity by a peroxisome-localized small heat shock protein Hsp17.6CII. *J. Genet. Genomics* 44, 395–404. doi: 10.1016/j.jgg.2017.03.009

Li, J., and Liu, X. (2019). Genome-wide identification and expression profile analysis of the Hsp20 gene family in Barley (Hordeum vulgare L.). *PeerJ* 7, e6832. doi: 10.7717/peerj.6832

Liao, Y., Smyth, G. K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30, 923–930. doi: 10.1093/bioinformatics/btt656

Ligaba, A., Katsuhara, M., Shibasaka, M., and Djira, G. (2011). Abiotic stresses modulate expression of major intrinsic proteins in barley (Hordeum vulgare). *C. R. Biol.* 334, 127–139. doi: 10.1016/j.crvi.2010.11.005

Liu, D., Li, M., Guo, T., Lu, J., Xie, Y., Hao, Y., et al. (2022). Functional characterization of the Serine acetyltransferase family genes uncovers the diversification and conservation of cysteine biosynthesis in tomato. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.913856

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

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15, 550. doi: 10.1186/s13059-014-0550-8

Lu, P., Magwanga, R. O., Kirungu, J. N., Hu, Y., Dong, Q., Cai, X., et al. (2019). Overexpression of cotton a DTX/MATE gene enhances drought, salt, and cold stress tolerance in transgenic arabidopsis. *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.00299

Ma, C., Haslbeck, M., Babujee, L., Jahn, O., and Reumann, S. (2006). Identification and characterization of a stress-inducible and a constitutive small heat-shock protein targeted to the matrix of plant peroxisomes. *Plant Physiol.* 141, 47–60. doi: 10.1104/pp.105.073841

Mangano, S., Denita-Juarez, S. P., Choi, H.-S., Marzol, E., Hwang, Y., Ranocha, P., et al. (2017). Molecular link between auxin and ROS-mediated polar growth. *Proc. Natl. Acad. Sci.* 114, 5289–5294. doi: 10.1073/pnas.1701536114

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetjournal* 17 (1), 10–12. doi: 10.14806/ej.17.1.200

Mascher, M., Gundlach, H., Himmelbach, A., Beier, S., Twardziok, S. O., Wicker, T., et al. (2017). A chromosome conformation capture ordered sequence of the barley genome. *Nature* 544, 427–433. doi: 10.1038/nature22043

Miller, G., and Mittler, R. (2006). Could heat shock transcription factors function as hydrogen peroxide sensors in plants? *Ann. Bot.* 98, 279–288. doi: 10.1093/aob/mcl107

Mittler, R., Vanderauwera, S., Suzuki, N., Miller, G., Tognetti, V. B., Vandepoele, K., et al. (2011). ROS signaling: the new wave? *Trends Plant Sci.* 16, 300–309. doi: 10.1016/j.tplants.2011.03.007

Mohanta, T. K., Bashir, T., Hashem, A., Abd_Allah, E. F., Khan, A. L., and Al-Harrasi, A. S. (2018). Early events in plant abiotic stress signaling: interplay between calcium, reactive oxygen species and phytohormones. *J. Plant Growth Regul.* 37, 1033–1049. doi: 10.1007/s00344-018-9833-8

Moll, P., Ante, M., Seitz, A., and Reda, T. (2014). QuantSeq 3′ mRNA sequencing for RNA quantification. *Nat. Methods* 11, i–iii. doi: 10.1038/nmeth.f.376

Munns, R., James, R. A., and Läuchli, A. (2006). Approaches to increasing the salt tolerance of wheat and other cereals. *J. Exp. Bot.* 57, 1025–1043. doi: 10.1093/jxb/erj100

Mustroph, A., Zanetti, M. E., Jang, C. J. H., Holtan, H. E., Repetti, P. P., Galbraith, D. W., et al. (2009). Profiling translatomes of discrete cell populations resolves altered cellular priorities during hypoxia in Arabidopsis. *Proc. Natl. Acad. Sci.* 106, 18843–18848. doi: 10.1073/pnas.0906131106

Mylona, P. V., Polidoros, A. N., and Scandalios, J. G. (2007). Antioxidant gene responses to ROS-generating xenobiotics in developing and germinated scutella of maize. *J. Exp. Bot.* 58, 1301–1312. doi: 10.1093/jxb/erl292

Nefissi Ouertani, R., Arasappan, D., Abid, G., Ben Chikha, M., Jardak, R., Mahmoudi, H., et al. (2021). Transcriptomic analysis of salt-stress-responsive genes in barley roots and leaves. *Int. J. Mol. Sci.* 22, 8155. doi: 10.3390/ijms22158155

Nianiou-Obeidat, I., Madesis, P., Kissoudis, C., Voulgari, G., Chronopoulou, E., Tsaftaris, A., et al. (2017). Plant glutathione transferase-mediated stress tolerance: functions and biotechnological applications. *Plant Cell Rep.* 36, 791–805. doi: 10.1007/s00299-017-2139-7

Ohkama-Ohtsu, N., Sasaki-Sekimoto, Y., Oikawa, A., Jikumaru, Y., Shinoda, S., Inoue, E., et al. (2011). 12-oxo-phytodienoic acid–glutathione conjugate is transported into the vacuole in arabidopsis. *Plant Cell Physiol.* 52, 205–209. doi: 10.1093/pcp/pcq181

Osthoff, A., Donà dalle Rose, P., Baldauf, J. A., Piepho, H.-P., and Hochholdinger, F. (2019). Transcriptomic reprogramming of barley seminal roots by combined water deficit and salt stress. *BMC Genomics* 20, 325. doi: 10.1186/s12864-019-5634-0

Park, E.-J., and Kim, T.-H. (2021). Arabidopsis OSMOTIN 34 functions in the ABA signaling pathway and is regulated by proteolysis. *Int. J. Mol. Sci.* 22, 7915. doi: 10.3390/ijms22157915

Pasternak, T., Potters, G., Caubergs, R., and Jansen, M. A. K. (2005). Complementary interactions between oxidative stress and auxins control plant growth responses at plant, organ, and cellular level. *J. Exp. Bot.* 56, 1991–2001. doi: 10.1093/jxb/eri196

Paudel, J., Copley, T., Amirizian, A., Prado, A., and Bede, J. (2013). Arabidopsis redox status in response to caterpillar herbivory. *Front. Plant Sci.* 4. doi: 10.3389/fpls.2013.00113

Pauwels, L., and Goossens, A. (2011). The JAZ proteins: A crucial interface in the jasmonate signaling cascade. *Plant Cell* 23, 3089–3100. doi: 10.1105/tpc.111.089300

Pei, Z.-M., Murata, Y., Benning, G., Thomine, S., Klüsener, B., Allen, G. J., et al. (2000). Calcium channels activated by hydrogen peroxide mediate abscisic acid signalling in guard cells. *Nature* 406, 731–734. doi: 10.1038/35021067

Peiter, E. (2016). The ever-closer union of signals: propagating waves of calcium and ROS are inextricably linked. *Plant Physiol.* 172, 3–4. doi: 10.1104/pp.16.01037

Planas-Portell, J., Gallart, M., Tiburcio, A. F., and Altabella, T. (2013). Copper-containing amine oxidases contribute to terminal polyamine oxidation in peroxisomes and apoplast of Arabidopsis thaliana. *BMC Plant Biol.* 13, 109. doi: 10.1186/1471-2229-13-109

Polidoros, A. N., Mylona, P. V., Pasentsis, K., Scandalios, J. G., and Tsaftaris, A. S. (2005). The maize alternative oxidase 1a (Aox1a) gene is regulated by signals related to oxidative stress. *Redox Rep.* 10, 71–78. doi: 10.1179/135100005X21688

R Core Team. (2020). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for statistical computing. Available at: https://www.R-project.org.

Ribeiro, C. W., Korbes, A. P., Garighan, J. A., Jardim-Messeder, D., Carvalho, F. E. L., Sousa, R. H. V., et al. (2017). Rice peroxisomal ascorbate peroxidase knockdown affects ROS signaling and triggers early leaf senescence. *Plant Sci.* 263, 55–65. doi: 10.1016/j.plantsci.2017.07.009

Rollins, J. A., Habte, E., Templer, S. E., Colby, T., Schmidt, J., and von Korff, M. (2013). Leaf proteome alterations in the context of physiological and morphological responses to drought and heat stress in barley (Hordeum vulgare L.). *J. Exp. Bot.* 64, 3201–3212. doi: 10.1093/jxb/ert158

Sandalio, L. M., Rodríguez-Serrano, M., Romero-Puertas, M. C., and del Río, L. A. (2013). "Role of peroxisomes as a source of reactive oxygen species (ROS) signaling molecules," in *Peroxisomes and their key role in cellular signaling and metabolism*. Ed. L.A.d. Río (Dordrecht: Springer Netherlands), 231–255. doi: 10.1007/978-94-007-6889-5_13

Saxena, I., Srikanth, S., and Chen, Z. (2016). Cross Talk between H2O2 and Interacting Signal Molecules under Plant Stress Response. *Front. Plant Sci.* 7. doi: 10.3389/fpls.2016.00570

Schaller, A., and Stintzi, A. (2009). Enzymes in jasmonate biosynthesis – Structure, function, regulation. *Jasmonates Stress Responses Dev.* 70, 1532–1538. doi: 10.1016/j.phytochem.2009.07.032

Shigeto, J., and Tsutsumi, Y. (2016). Diverse functions and reactions of class III peroxidases. *New Phytol.* 209, 1395–1402. doi: 10.1111/nph.13738

Shimakawa, G., Roach, T., and Krieger-Liszkay, A. (2020). Changes in photosynthetic electron transport during leaf senescence in two barley varieties grown in contrasting growth regimes. *Plant Cell Physiol.* 61, 1986–1994. doi: 10.1093/pcp/pcaa114

Siddique, M., Gernhard, S., von Koskull-Döring, P., Vierling, E., and Scharf, K.-D. (2008). The plant sHSP superfamily: five new members in Arabidopsis thaliana with unexpected properties. *Cell Stress Chaperones* 13, 183–197. doi: 10.1007/s12192-008-0032-6

Smirnoff, N., and Arnaud, D. (2019). Hydrogen peroxide metabolism and functions in plants. *New Phytol.* 221, 1197–1214. doi: 10.1111/nph.15488

Staneloni, R. J., Rodriguez-Batiller, M. J., and Casal, J. J. (2008). Abscisic acid, high-light, and oxidative stress down-regulate a photosynthetic gene *via* a promoter motif not involved in phytochrome-mediated transcriptional regulation. *Mol. Plant* 1, 75–83. doi: 10.1093/mp/ssm007

Stanley Kim, H., Yu, Y., Snesrud, E. C., Moy, L. P., Linford, L. D., Haas, B. J., et al. (2005). Transcriptional divergence of the duplicated oxidative stress-responsive genes in the Arabidopsis genome. *Plant J.* 41, 212–220. doi: 10.1111/j.1365-313X.2004.02295.x

Steffens, B. (2014). The role of ethylene and ROS in salinity, heavy metal, and flooding responses in rice. *Front. Plant Sci.* 5. doi: 10.3389/fpls.2014.00685

Swindell, W. R., Huebner, M., and Weber, A. P. (2007). Transcriptional profiling of Arabidopsis heat shock proteins and transcription factors reveals extensive overlap between heat and non-heat stress response pathways. *BMC Genomics* 8, 125. doi: 10.1186/1471-2164-8-125

Tavares, S., Wirtz, M., Beier, M. P., Bogs, J., Hell, R., and Amâncio, S. (2015). Characterization of the serine acetyltransferase gene family of Vitis vinifera uncovers differences in regulation of OAS synthesis in woody plants. *Front. Plant Sci.* 6. doi: 10.3389/fpls.2015.00074

Tenhaken, R. (2015). Cell wall remodeling under abiotic stress. *Front. Plant Sci.* 5. doi: 10.3389/fpls.2014.00771

Terzi, R., Kadioglu, A., Kalaycioglu, E., and Saglam, A. (2014). Hydrogen peroxide pretreatment induces osmotic stress tolerance by influencing osmolyte and abscisic acid levels in maize leaves. *J. Plant Interact.* 9, 559–565. doi: 10.1080/17429145.2013.871077

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, 562–578. doi: 10.1038/nprot.2012.016

ul Haq, S., Khan, A., Ali, M., Khattak, A. M., Gai, W.-X., Zhang, H.-X., et al. (2019). Heat shock proteins: dynamic biomolecules to counter plant biotic and abiotic stresses. *Int. J. Mol. Sci.* 20, 5321. doi: 10.3390/ijms20215321

Vanderauwera, S., Suzuki, N., Miller, G., van de Cotte, B., Morsa, S., Ravanat, J.-L., et al. (2011). Extranuclear protection of chromosomal DNA from oxidative stress. *Proc. Natl. Acad. Sci.* 108, 1711–1716. doi: 10.1073/pnas.1018359108

Verma, V., Ravindran, P., and Kumar, P. P. (2016). Plant hormone-mediated regulation of stress responses. *BMC Plant Biol.* 16, 86. doi: 10.1186/s12870-016-0771-y

Wahid, A., Gelani, S., Ashraf, M., and Foolad, M. R. (2007). Heat tolerance in plants: An overview. *Environ. Exp. Bot.* 61, 199–223. doi: 10.1016/j.envexpbot.2007.05.011

Wang, Y., Li, J., Wang, J., and Li, Z. (2010). Exogenous H2O2 improves the chilling tolerance of manilagrass and mascarenegrass by activating the antioxidative system. *Plant Growth Regul.* 61, 195–204. doi: 10.1007/s10725-010-9470-0

Wang, R., Liu, S., Zhou, F., and Ding, C. (2014). Exogenous ascorbic acid and glutathione alleviate oxidative stress induced by salt stress in the chloroplasts of oryza sativa L. *Z. Naturforsch C J Biosci* 69, 226–236. doi: 10.5560/znc.2013-0117

Wang, R.-S., Pandey, S., Li, S., Gookin, T. E., Zhao, Z., Albert, R., et al. (2011). Common and unique elements of the ABA-regulated transcriptome of Arabidopsis guard cells. *BMC Genomics* 12, 216. doi: 10.1186/1471-2164-12-216

Waters, E. R. (2013). The evolution, function, structure, and expression of the plant sHSPs. *J. Exp. Bot.* 64, 391–403. doi: 10.1093/jxb/ers355

Xie, Z., Nolan, T. M., Jiang, H., and Yin, Y. (2019). AP2/ERF transcription factor regulatory networks in hormone and abiotic stress responses in arabidopsis. *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.00228

Yu, C.-W., Murphy, T. M., and Lin, C.-H. (2003). Hydrogen peroxide-induced chilling tolerance in mung beans mediated through ABA-independent glutathione accumulation. *Funct. Plant Biol.* 30, 955–963. doi: 10.1071/FP03091

Zeng, J., Dong, Z., Wu, H., Tian, Z., and Zhao, Z. (2017). Redox regulation of plant stem cell fate. *EMBO J.* 36, 2844–2855. doi: 10.15252/embj.201695955

Zhu, J.-K. (2016). Abiotic stress signaling and responses in plants. *Cell* 167, 313–324. doi: 10.1016/j.cell.2016.08.029

Zhu, Q., Feng, Y., Xue, J., Chen, P., Zhang, A., and Yu, Y. (2023). Advances in receptor-like protein kinases in balancing plant growth and stress responses. *Plants* 12, 427. doi: 10.3390/plants12030427

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 20 September 2023
DOI 10.3389/fpls.2023.1237426

OPEN ACCESS

EDITED BY
Manohar Chakrabarti,
The University of Texas Rio Grande Valley, United States

REVIEWED BY
Niharika Sharma,
NSW Government, Australia
Xiujun Zhang,
Chinese Academy of Sciences (CAS), China

*CORRESPONDENCE
Achraf El Allali
achraf.elallali@um6p.ma
Morad M. Mokhtar
morad.mokhtar@ageri.sci.eg

RECEIVED 09 June 2023
ACCEPTED 21 August 2023
PUBLISHED 20 September 2023

CITATION
Mokhtar MM and El Allali A (2023) MegaLTR: a web server and standalone pipeline for detecting and annotating LTR-retrotransposons in plant genomes. *Front. Plant Sci.* 14:1237426. doi: 10.3389/fpls.2023.1237426

MegaLTR: a web server and standalone pipeline for detecting and annotating LTR-retrotransposons in plant genomes

Morad M. Mokhtar* and Achraf El Allali*

African Genome Center, Mohammed VI Polytechnic University, Benguerir, Morocco

LTR-retrotransposons (LTR-RTs) are a class of RNA-replicating transposon elements (TEs) that can alter genome structure and function by moving positions, repositioning genes, shifting exons, and causing chromosomal rearrangements. LTR-RTs are widespread in many plant genomes and constitute a significant portion of the genome. Their movement and activity in eukaryotic genomes can provide insight into genome evolution and gene function, especially when LTR-RTs are located near or within genes. Building the redundant and non-redundant LTR-RTs libraries and their annotations for species lacking this resource requires extensive bioinformatics pipelines and expensive computing power to analyze large amounts of genomic data. This increases the need for online services that provide computational resources with minimal overhead and maximum efficiency. Here, we present MegaLTR as a web server and standalone pipeline that detects intact LTR-RTs at the whole-genome level and integrates multiple tools for structure-based, homologybased, and *de novo* identification, classification, annotation, insertion time determination, and LTR-RT gene chimera analysis. MegaLTR also provides statistical analysis and visualization with multiple tools and can be used to accelerate plant species discovery and assist breeding programs in their efforts to improve genomic resources. We hope that the development of online services such as MegaLTR, which can analyze large amounts of genomic data, will become increasingly important for the automated detection and annotation of LTR-RT elements.

KEYWORDS

LTR-retrotransposons, plant genomes, webserver, insertion age, LTR-RT gene chimeras, non-redundant LTR-RTs library

1 Introduction

Long Terminal Repeat (LTR) Retrotransposons (LTR-RTs) are a class of transposon elements (TEs) belonging to the repetitive DNA sequences that have played a crucial role in shaping the structure and function of eukaryotic genomes (Vitte and Panaud, 2005). LTR-RTs are characterized by their ability to move within genomes *via* a "copy-and-paste" mechanism that involves transcription into RNA, reverse transcription into DNA, and subsequent insertion into new genomic locations (Lopes et al., 2013). These elements have been found in various organisms, including plants, where they contribute significantly to genome size and complexity. LTR-RTs are of great interest in the field of genomics because of their importance in genome evolution, gene regulation, and understanding plant biology (Bennetzen and Wang, 2014). Plant genomes are often characterized by a high proportion of TEs, with LTR-RTs being one of the major contributors to these elements. TEs can make up a substantial portion of the plant genome, as in maize, where TEs account for 85% of the genome, of which LTR-RTs account for 75% (Schnable et al., 2009). This wide distribution highlights their importance in shaping genome architecture and dynamics (Schnable et al., 2009). LTR-RTs are known to play a role in creating genetic diversity, promoting chromosomal rearrangements and influencing gene expression through their insertion sites and regulatory sequences (Bennetzen and Wang, 2014). Therefore, the study of LTR-RTs is crucial to unravel the complexity of plant genomes and understand their functional implications (Xia et al., 2020). The study of LTR-RTs provides insights into various aspects of plant genome biology. For example, studying their structural diversity, insertion patterns, and distribution in plant taxa can provide insight into evolutionary history and interspecies relationships (Grandbastien et al., 2005). In addition, understanding the regulation of LTR-RTs activity and its interplay with host factors can provide insight into the mechanisms of genome stability (Vitte et al., 2014). Because LTR-RTs can influence nearby gene expression through epigenetic modifications and transcriptional interference, studying these elements contributes to our understanding of gene regulatory networks in plants (Zhao et al., 2016; Mokhtar et al., 2021).

The movement of LTR-RTs within genomes contributes to genome evolution by generating genetic variation and driving genome expansion (Vitte et al., 2014). These elements can facilitate chromosomal rearrangements through unequal homologous recombination between LTRs or ectopic recombination between non-homologous LTRs. Such events can lead to gene duplications, deletions, and chromosomal rearrangements that contribute to plant genome diversification (Ma et al., 2004). LTR-RTs may also serve as targets for silencing by small RNAs, which could affect their transposition rates and influence the evolutionary development of plant species (Franco-Zorrilla et al., 2007). While some LTR-RTs are likely to be transcriptionally inactive, accumulating evidence suggests that many elements have been co-opted for useful functions in plant genomes. For example, some LTR-RTs have been domesticated to provide regulatory sequences such as promoters and enhancers for nearby genes (Jung et al., 2019). In addition, they have been associated with stress responses, chromatin remodeling, and even symbiotic interactions (Ito et al., 2016; Pereira, 2016). Understanding the functional significance of LTR-RTs in plant genomes provides insights into the intricate interplay between repetitive DNA elements and the evolution of novel traits.

LTR-RTs consist of several different structural elements that play different roles in the movement and regulation of the element within the genome. Common elements include target site duplication (TSD), two semi-identical LTRs, polypurine tract (PPT), primer binding site (PBS), *GAG* and *Pol* genes (Kumar, 1998). LTRs are long stretches of DNA located at both ends of the element and are typically several hundred base pairs long. LTRs contain regulatory elements (promoters, enhancers) and are thought to be important for the integration and stability of the element in the genome (Kumar, 1998). *GAG* and *Pol* genes are genes that encode proteins involved in the movement and replication of the element (Eickbush and Jamburuthugoda, 2008). The *GAG* gene encodes a structural protein involved in the assembly of the element, while the *Pol* gene consists of several different functional domains, including protease (PROT), reverse transcriptase (RT), RNase H (RH), and integrase (INT) (Ustyantsev et al., 2015).The RT domain is responsible for synthesizing a DNA copy of the RNA template of the element, while the INT domain is responsible for integrating the element into the genome (Zhao et al., 2016). The PROT domain is responsible for cleavage of the *Pol* protein into its functional domains; the RH domain is involved in degradation of the RNA template during reverse transcription; and other domains that are involved in various aspects of movement and regulation of the element (Gao et al., 2003; Ustyantsev et al., 2015). LTR-RTs are divided into two main categories based on their mode of movement: autonomous and non-autonomous. Autonomous LTR-RTs are capable of moving by themselves, whereas non-autonomous LTR-RTs require the assistance of an autonomous element to move (Wicker et al., 2007). In addition, LTR-RTs are classified into superfamilies *Copia* and *Gypsy* based on internal domain arrangements (Wicker et al., 2007). Other LTR-RTs groups include *LARD* (LArge Retrotransposon Derivatives), *BARE-2* (Barley RetroElement-2), *TR-GAG* (Terminal Repeat Retrotransposons with *GAG* domain), and *TRIM* (Terminal Repeats In Miniature)((Witte et al., 2001; Kalendar et al., 2004; Tanskanen et al., 2007; Chaparro et al., 2015), respectively).

Despite their widespread use and importance, LTR-RTs remain difficult to identify and annotate in most non-model organisms (Ou et al., 2019). One reason is that they are often difficult to identify and track in the genome. They are also difficult to study because they have complex and variable structures and can interact in complex ways with other DNA sequences (Ou et al., 2019). However, research on LTR-RTs has increased in recent years, thanks to advances in sequencing technology and bioinformatics that have improved our understanding of the role of LTR-RT in genomes. Several tools, pipelines, and databases exist to identify LTR-RTs and support current and future functional genomics research. These tools include Tandem Repeats Finder [TRF, (Benson, 1999)], LTR_STRUC (McCarthy and McDonald, 2003), LTR_FINDER (Xu and Wang, 2007), LTRdigest (Steinbiss et al., 2009), LTRharvest (Ellinghaus et al., 2008), RepeatMasker (Smit et al., 2015), MGEScan3 (Lee et al., 2016), LTR_retriever (Ou and Jiang, 2017), LtrDetector (Valencia and Girgis, 2019), DARTS

(Biryukov and Ustyantsev, 2021), and TEsorter (Zhang et al., 2022). Once LTR-RTs are identified, they can be annotated using various databases and resources. Some examples of databases and resources developed for this purpose are TREP (Wicker et al., 2002), RepBase (Jurka et al., 2005), REXdb (Neumann et al., 2019), PlantRep (Amselem et al., 2019), and PlantLTRdb (Mokhtar et al., 2023b). These tools and databases have been used to create automatized pipelines for LTR-RT analysis, including REPCLASS (Feschotte et al., 2009), EDTA (Ou et al., 2019), and Inpactor2 (Orozco-Arias et al., 2022).

EDTA is a pipeline that integrates structural-, homology-based, and *de novo* identification methods to create TEs libraries. EDTA combines LTRharvest, LTR_FINDER, and LTR_retriever to analyze LTR-RTs. In addition, Generic Repeat Finder (Shi and Liang, 2019), TIR-Learner (Su et al., 2019), HelitronScanner (Xiong et al., 2014), and RepeatModeler (Smit et al., 2015) are used for other TEs. For LTR-RTs, EDTA performs identification, superfamily-level classification (*Copia* and *Gypsy*), and insertion age estimation with highly efficient tools. Another available pipeline is Inpactor2. It integrates the process of identification and classification of LTR-RTs at the lineage level and runs in a reasonable time. While EDTA and Inpactor2 are comprehensive pipelines for creating LTR-RTs libraries, it lacks some features, such as putative autonomous and non-autonomous classification, identification of LTR-RT gene chimeras, detection of LTR-RTs near genes, statistical analysis and visualization of LTR-RTs, and adjustable parameters for each analysis step. It is also not available as a web server and requires some level of technical computer skills. Like any machine learning-based algorithm, Inpactor2 is dependent on the quality of its training dataset (Orozco-Arias et al., 2022), a fact that users should consider when using this algorithm.

Here we introduce MegaLTR as a web server and standalone pipeline that detects intact LTR-RTs at the whole genome level. MegaLTR integrates multiple tools for structure-based, homology-based, and *de novo* identification, classification, and annotation. MegaLTR performs classification into putative autonomous and non-autonomous, superfamilial and lineage levels. It also identifies LTR-RT gene chimeras, detects LTR-RTs near genes, statistical analysis and visualization of LTR-RT. MegaLTR is easy to use and allows customization of parameters for each analysis step in both its web server and standalone versions.

2 Materials and methods

2.1 Genomic data

The complete genome sequences and annotations of 26 plant species were downloaded from the NCBI database (Wheeler et al., 2007). These genomes were selected based on some criteria, such as annotation and LTR assembly index (LAI) score (Ou et al., 2018), genome size, number of pseudomolecules/scaffolds, and the fact that they were model and non-model plants. The LAI score has been widely used in recent years to assess the quality of genome assemblies. It has been shown to be useful in determining the quality of assemblies, as a higher LAI score is associated with a higher quality assembly (Ou et al., 2018). The LAI score of each species was taken from the PlantLAI database (Mokhtar et al., 2023a). The plant name, NCBI taxonomy ID, GenBank accession number, assembly level, LAI score, genome size, evolutionary rate, and number of pseudomolecules/scaffolds of the studied species are listed in Table S1.

2.2 MegaLTR design and workflow

MegaLTR's workflow includes multiple programs interconnected by data adapters to ensure that data is routed from the server to a high-performance computer (HPC) and back to the server and processed as an end-to-end pipeline. The implementation of MegaLTR was summarized in Data Sheet 1. The MegaLTR workflow is shown schematically in Figure 1. MegaLTR is designed to accept FASTA sequences and their GFF annotation as input. It is capable of processing whole genome sequences in any form, including chromesomes, pseudomolecules, scaffolds, contigs, and fragments, which is useful in draft genome analysis. Analysis with MegaLTR consists of eight main steps: 1) LTR-RTs identification with LTR_FINDER (Xu and Wang, 2007; Ou and Jiang, 2019) and LTRharvest (Ellinghaus et al., 2008); 2) filtering LTR-RTs with LTR_retriever (Ou and Jiang, 2017); 3) annotation of internal domains and clades with TEsorter (Zhang et al., 2022); 4) PBS and PPT annotation with LTRdigest (Steinbiss et al., 2009) and PltRNAdb (Mokhtar and El Allali, 2022); 5) insertion age estimation with REANNOTATE (Pereira, 2008) and ClustalW (Thompson et al., 2003); 6) LTR-RTs classification with Python scripts and create a non-redundant LTRRTs library using USEARCH v11.0 (Edgar, 2010); 7) LTR-RTs detection within and near genes with Perl scripts; 8) statistical analysis and visualization with Python, R scripts and RIdeograms (Hao et al., 2020). The user can set the parameters for each analysis step.

For identification of LTR-RT candidates, LTR_FINDER and LTRharvest are used because they are very effective in identifying LTR-RTs and outperform all other programs in sensitivity (Ou and Jiang, 2017). However, these programs tend to produce a number of false-positive predictions (Lerat, 2010). To effectively remove false-positive predictions made by the original softwares, the results were combined into one file and used as input to LTR_retriever. The LTR_retriever tool uses a combination of several programs, including HMMER (Wheeler and Eddy, 2013), CD-HIT (Li and Godzik, 2006), BLAST+ (Camacho et al., 2009), RepeatMasker (Smit et al., 2015), and TRF (Benson, 1999) to identify and filter out all false candidates for LTR-RTs. MegaLTR only considers intact LTR-RT candidates that pass these filtering steps in the post analysis. The intact LTR-RT, defined as candidates, contain two identical/semi-identical LTRs and a target site duplication at both ends. The LTRs contain conserved sequences such as the TG-CA, which may play a role in regulating retrotransposon expression and/or retrotransposition. To accurately identify features within a potential LTR-RT, MegaLTR uses LTRdigest to detect PPT, and PBS and TEsorter to analyze internal protein domains. The PBS is generally located near the 5'LTR, while PPT is relatively close to the 3'LTR. To identify the PBS, a tRNA sequence library is used to

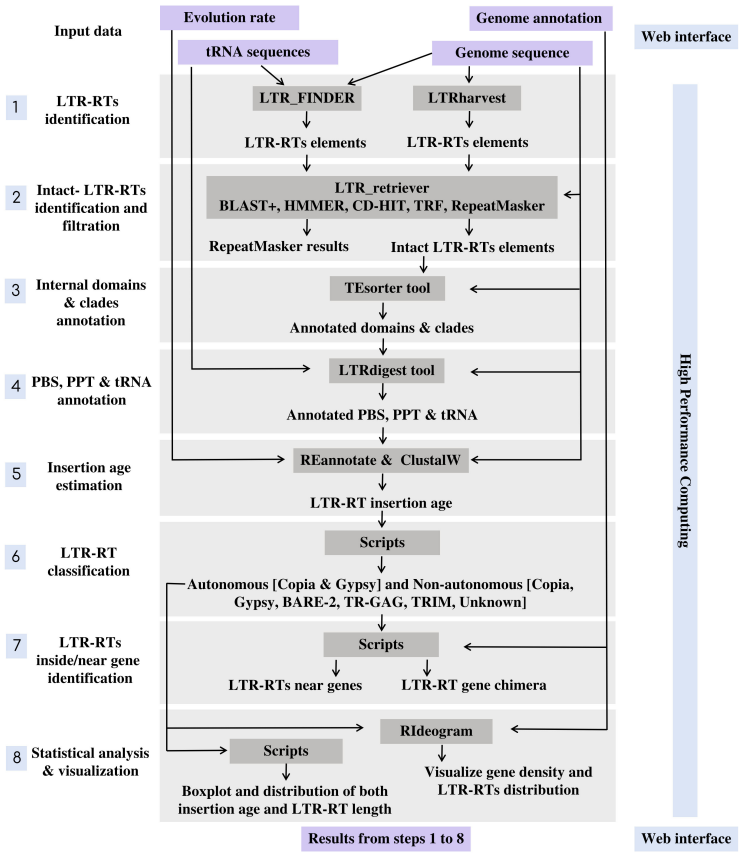


FIGURE 1
An Overview of MegaLTR Workflow and procedure.

search for regions in the LTR-RT candidate that are complementary to the tRNA. The tRNA sequences for this analysis are from the plant tRNA database [PltRNAdb, (Mokhtar and El Allali, 2022)]. This procedure allows reliable identification of PBS and PPT within a LTR-RT candidate. To annotate protein domains, TEsorter searched one of the databases REXdb (Neumann et al., 2019) and GyDB (http://gydb.org) using HMMScan (Eddy, 1998) to identify putative domains such as capsid protein, protease, reverse transcriptase, RNase H, and integrase.

The next step is to classify LTR-RTs in clades. Previous studies have proposed different clade-level classifications for LTR-RTs. Neumann et al. (2019) divided *Copia* to the clades *Ale, Alesia, SIRE, Bianca, Lyco, Ikeros, Gymco I-IV, Bryco, Osser, TAR, Angela, Ivana,* and *Tork*. They also divided *Gypsy* into the clades *Chlamyvir, CRM, Tcn1, Reina, Galadriel, Tekay, Tat-I-III, Athila, Ogre, Phygy, Selgy*, and *Retand*. This classification is based on the protein domain databases for clade-level classification of LTR-RT. The TEsorter tool uses these databases as well as REXdb and GyDB to classify LTR-RTs into superfamilies and further classify them into clades. To estimate the insertion age of LTR-RT, MegaLTR uses the tools REANNOTATE and ClustalW in combination to estimate the insertion age of LTR-RT elements based on a comparative analysis of their 5' and 3' LTRs. To calculate the insertion age, the Kimura-2 parameter model (Kimura, 1980) is used to calculate the substitutions per site rate (K) between LTRs. The age is then estimated as T= K/2r (Kimura, 1980), where (r) is the evolution rate. In MegaLTR, the evolution rate is usually set by the user. It is important to note that evolution rates can vary significantly between species.

LTR-RT can be divided into two main categories based on their structure: autonomous and nonautonomous. According to Wicker et al. (2007), the structure of autonomous *Gypsy* and *Copia* is based on domains arranged within the element LTR-RT. The structure of *Gypsy* is TSD-LTR-PBS-GAG-PROT-RTRH-INT-PPT-LTR-TSD, while *Copia* is TSD-LTR-PBS-GAG-PROT-INT-RT-RH-PPT-LTR-TSD. *Copia* and *Gypsy* elements that no longer have any of the previous structures are classified as non-autonomous Copia and non-autonomous *Gypsy*. Non-autonomous LTR-RT can be further subdivided based on their specific structure and the presence or absence of certain domains. Examples of non-autonomous elements include *LARD, TRIM, TR-GAG*, and *BARE-2* (Figure 2). The specific criteria for classifying LTR-RT elements into these categories have been described in several research studies, including Kalendar et al. (2004); Witte et al. (2001); Chaparro et al. (2015); Tanskanen et al. (2007). LTR-RT elements that do not fit into any of these categories and are not classified as autonomous or non-autonomous *Copia* or *Gypsy* elements are classified as "unknown".

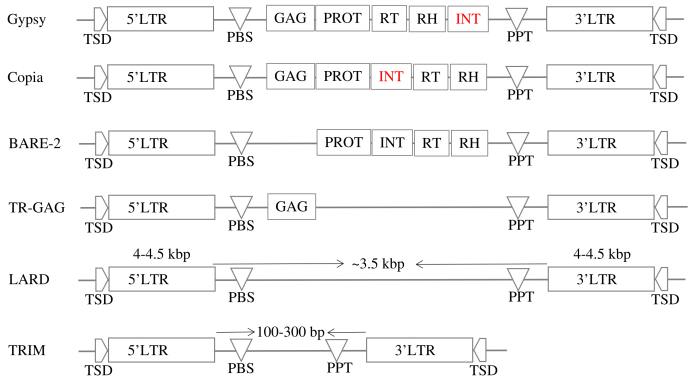


FIGURE 2
The structures of autonomous (*Gypsy* and *Copia*) and non-autonomous LTR-RTs (*LARD, TRIM, TR-GAG*, and *BARE-2*).

Because LTR-RTs sometimes insert themselves into or near genes and can affect gene function. MegaLTR identifies LTR-RTs that are inside or near genes using Perl scripts. To classify LTR-RT elements based on their genomic location, the start and end coordinates of the gene and the start and end coordinates of the LTR-RT element within the genome can be compared. If the LTR-RT element is located within the coordinates of the gene, it is considered a gene chimera. If the LTR-RT element is located near a gene, the distance upstream and downstream of the LTR-RT element can be determined in base pairs. This distance is usually determined by the user and may vary depending on the specific research question and desired sensitivity for detecting LTR-RT elements near genes. In the final step, MegaLTR performs two statistical analyses using boxplot. One for LTR-RT length by bps and the other for LTR-RT insertion age. Boxplots are useful for quickly conveying information about the variability and skewness of a data set. The next step is a visualization of the distribution of the identified LTR-RT and gene density in each pseudomolecules/scaffolds using RIdeograms (Hao et al., 2020).

2.3 Standalone version

The standalone version of MegaLTR is also available (https://github.com/MoradMMokhtar/MegaLTR). It has been thoroughly tested on Ubuntu 18.04 and 20.04. Installation is effortless *via* a Conda environment with the command: conda env create -f MegaLTR.yml. This command not only installs MegaLTR, but also takes care of installing the associated dependencies. Using MegaLTR standalone, the user can define all parameters using the following flags: -A (the analysis type), -F (fasta file), -G (GFF file), -T (species name for tRNA database), -P (prefix for outfiles), -l (minimum length of 5' & 3'LTR, -L (maximum length of 5' & 3'LTR), -d (minimum distance between 5' & 3'LTR), -D (maximum distance between 5' & 3'LTR), -S (similarity threshold), -M (minimum length of exact match pair), -B (name of TE database that TEsorter will use "gydb, rexdb, rexdb-plant, rexdb-metazoa"), -C (minimum coverage for protein domains in HMMScan), -V (maximum E value for protein domains in HMMScan), -Q (classification rule [identity - coverage - length]), -E (hmm database), -R (mutation rate of neutral species), -U (distance upstream LTR-RTs to determine nearby genes), -X (distance downstream LTR-RTs), -W (gene density window size), -N (number of chromosomes), -t (number of CPUs to run MegaLTR).

3 Results and discussion

3.1 Validation and comparison

To test the performance and validate the quality of the intact LTR-RTs identified by MegaLTR, a manual curation of LTR-RTs library from *Oryza sativa* was used to compare the non-redundant library generated by MegaLTR. The curated *Oryza sativa* library included 897 LTR-RT elements and was previously established by Ou and Jiang (2017). RepeatMasker v4.0.7 with the parameters "-e ncbi -pa 56 -no_is -q -norna -div 40 -nolow -lib [LTR -library] -cutoff 225 genome.fa" was applied to the MegaLTR library and the curated library to compute the performance metrics. We used six metrics proposed by Ou and Jiang (2017) to characterize the annotation performance of the non-redundant LTR-RT library generated by MegaLTR. These metrics include sensitivity (the ability to annotate target sequences correctly), specificity (the ability to exclude non-target sequences correctly), accuracy (true discrimination rate between target and non-target sequences), precision (true detection rate), FDR (false detection rate), and F1 measure (harmonic mean of precision and sensitivity). The True-positives (TP), false-positives (FP), false-negatives (FN), and true-negatives (TN) rates were computed using the EDTA toolkit. The performance metrics are defined as:

$$Sensitivity = \frac{TP}{TP+FN} \quad Specificity = \frac{TN}{FP+TN} \quad Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

$$Precision = \frac{TP}{TP+FP} \quad F1 = \frac{2*TP}{2*TP+FP+FN} \quad FDR = 1 - \frac{TP}{TP+FP}$$

MegaLTR results show consistently high specificity (96.59%), accuracy (94.98%), precision (89.38%), sensitivity (89.92%), and F1 measure (89.65%). The relatively low FDR (10.61%) confirms the accuracy and reliability of the LTR-RTs identified by MegaLTR. For comparison purposes, the EDTA pipeline was used to analyze the whole genome of *Oryza sativa* using the same parameters used in

MegaLTR (-D 15000 -d 1000 -L 7000 -l 100 -p 20 -M 0.85). The EDTA-generated LTR-RTs library was compared with the curated *Oryza sativa* LTR library. Similar to the evaluation of MegaLTR, RepeatMasker and the script "lib-test.pl" were used to calculate the evaluation metrics. The results of the EDTA metrics were: specificity (96.23%), accuracy (94.61%), precision (88.34%), sensitivity (89.52%), F1 measure (88.93%), and FDR (11.65%). As shown in Table 1, MegaLTR has relatively higher specificity, accuracy, precision and sensitivity with low FDR compared to EDTA.

Overall, the comparison of MegaLTR with both the manually curated *Oryza sativa* library and EDTA demonstrates the robustness and effectiveness of MegaLTR in identifying intact LTR-RTs and provides valuable insights for future studies on retrotransposons in plant genomes. Table 2 shows a comparison of various features between the MegaLTR and EDTA. The features compared include the class of TEs identified, the level of classification (autonomous, non-autonomous, superfamily, lineage level), the identification of LTR-RT near and within genes, and the form of availability.

To validate MegaLTR, 26 whole-genome sequences with a total volume of 15.33 Gbp representing 58,392 scaffolds/pseudomolecules were used. Plant species were selected based on their LAI score, genome size, and number of scaffolds/pseudomolecules. As suggested by Ou et al., (2018), the LAI score for draft genomes is below 10, while reference genomes have a LAI score between 10 and 20. Gold genomes have LAI scores greater than 20. The LAI scores of the selected genomes were retrieved from the PlantLAI database (Mokhtar et al., 2023a) and ranged from 8.7 (*Citrus unshiu*) to 29.45 (*Zea mays*), covering the different qualities of genome sequences (draft, reference, and gold quality). Genome sizes also varied, ranging from 119.6 Mbp for *Arabidopsis thaliana* to 2182.79 Mbp for *Zea mays*. In addition, the number of scaffolds/pseudomolecules varied from 7 to 20,876 for *Arabidopsis thaliana* and *Citrus unshiu* (Table S1).

Table 3 shows a comparison between MegaLTR and EDTA based on runtime and number of identified LTR-RTs in each classified superfamily using the same parameters mentioned above. Since EDTA performs the analysis of all TEs (LTR, TIR, and Helitron), we used the LTR [–type ltr] option to analyze only the LTR-RTs candidates. For MegaLTR, the total number of identified autonomous (*Gypsy* and *Copia*) and nonautonomous LTR-RTs (*Gypsy, Copia, BARE-2, TR-GAG*, unknown) was reported for the genomes examined. The *LARD* and *TRIM* structures were not detected in these genomes. However, EDTA classified LTR-RTs into *Gypsy, Copia* and unknown elements. As can be seen in Table 3, MegaLTR reported a small number of unknown elements compared to EDTA, as MegaLTR performed further analyses to annotate and classify the identified LTR-RTs.

The EDTA runtime given in Table 3 refers to LTR-RTs identification and classification as given by EDTA. However, the runtime given by MegaLTR refers to all analyses, including identification, annotation, classification of LTR-RTs, identification of LTR-RT gene chimeras, detection of LTR-RTs near genes, statistical analysis, and visualization of the density of LTR-RTs. The run times of each step reported by EDTA and MegaLTR can be found in Data Sheet 3 and Data Sheet 4, respectively. For MegaLTR, analysis times range from 9 minutes for *Arabidopsis lyrata* (206.8 Mb), *Arabidopsis thaliana* (119.6 Mb), and *Cucumis sativus* (226.6 Mb) to 10 hours and 16 minutes for *Zea mays* (2182.7 Mb). For EDTA, analysis times range from 6 minutes for *Cucumis sativus* to 27 hours and 39 minutes for *Zea mays*. For large genomes such as *Zea mays* (2182.7 Mb) and *Mikania micrantha* (1790.6 Mb), MegaLTR is more than 2x faster than EDTA. Figure S1 shows

TABLE 1 Comparison of six metrics between MegaLTR and EDTA using the genome of *Oryza sativa*.

Pipeline name	Sensitivity	Specificity	Accuracy	Precision	FDR	F1
MegaLTR	89.92%	96.59%	94.98%	89.38%	10.61%	89.65%
EDTA	89.52%	96.23%	94.61%	88.34%	11.65%	88.93%

TABLE 2 Comparison of some features between MegaLTR and EDTA.

	Identified TEs		Classification level			Identify LTR-RT		Availability	
	DNA TEs	LTR-RTs	Autonomous and non-autonomous	Superfamily	Lineage level	Gene chimeras	Near genes	Web server	Standalone
MegaLTR	X	✓	✓	✓	✓	✓	✓	✓	✓
EDTA	✓	✓	X	✓	X	X	X	X	✓
LTR-RTs sub-classification level									
	Copia	***Gypsy***	**Unknown**	***LARD***	***TRIM***	***TR-GAG***	***BARE-2***		
MegaLTR	✓	✓	✓	✓	✓	✓	✓		
EDTA	✓	✓	✓	X	X	X	X		

"✓" refer to the feature is found, and "X" refers to the feature is missing.

TABLE 3 Analysis runtime in hours and minutes (h:m), the total number of identified LTR-RTs in each classified superfamily for the 26 plant species using MegaLTR and EDTA.

Species name	Run time		MegaLTR							EDTA		
	MegaLTR	EDTA	Autonomous				Nonautonomous					
			Gypsy	*Copia*	*Gypsy*	*Copia*	*BARE-2*	*TR-GAG*	unknown	*Gypsy*	*Copia*	unknown
Arabidopsis thaliana	0:09	0:14	2	–	118	80	–	1	2	105	75	27
Brassica rapa	0:48	0:38	189	65	1138	1238	34	18	228	1196	1074	588
Citrus clementina	0:15	0:24	60	23	771	846	1	20	14	820	765	62
Citrus unshiu	0:17	0:20	26	4	340	178	–	2	2	343	161	32
Cucumis sativus	0:09	0:06	31	7	135	219	2	–	22	159	196	73
Glycine max	0:54	1:13	227	69	2046	2335	22	322	145	2433	1388	694
Medicago truncatula	0:26	0:31	25	21	743	1385	2	–	130	710	1329	335
Mikania micrantha	4:51	14:43	470	145	6930	15356	9	58	705	7009	14100	2671
Oryza sativa Japonica	0:28	0:25	35	9	488	1399	4	4	168	502	1504	226
Panicum hallii	1:03	1:23	46	46	841	3539	1	3	293	866	3512	641
Phoenix dactylifera	1:21	2:29	498	226	6233	3016	2	69	148	6625	2176	503
Physcomitrella patens	0:21	0:27	–	–	184	3225	–	–	13	140	3069	122
Populus trichocarpa	0:18	0:26	59	19	501	474	–	1	59	523	448	170
Prunus persica	0:16	0:30	43	21	632	326	5	1	141	637	319	548
Rosa chinensis	0:44	1:43	113	20	3806	1614	38	21	745	3498	1426	1884
Salvia splendens	2:16	3:18	198	296	4183	5687	23	70	459	3898	5462	2009
Selaginella moellendorffii	0:11	0:10	–	102	26	557	1	7	337	34	627	355
Sesamum indicum	0:21	0:28	5	20	258	176	1	6	6	240	185	38
Setaria viridis	0:25	0:36	3	39	829	1071	1	5	10	802	1091	105
Solanum lycopersicum	0:32	0:35	93	15	945	774	–	5	30	899	719	196
Solanum pennellii	0:37	0:42	143	1	1714	443	1	6	58	1622	361	310
Sorghum bicolor	1:28	2:35	58	109	1096	7779	3	7	893	1167	6927	2120
Trifolium pratense	0:38	1:00	83	95	2692	1566	5	34	943	2470	1469	2074
Vitis vinifera	0:22	0:35	247	10	1090	613	25	1	64	1275	583	247
Zea mays	10:16	27:39	3687	191	16301	28080	70	496	3163	19836	26376	6221

that MegaLTR is faster than EDTA for large genomes. The total number of LTR-RTs identified is also similar and only slightly different between MegaLTR and EDTA (Figure S1).

3.2 Case study: *Arabidopsis thaliana* genome

Arabidopsis thaliana was selected as a case study for MegaLTR results and serves as a comparison between the output of MegaLTR and EDTA in terms of classification. *Arabidopsis thaliana* is a model organism with a well-structured genome arranged in chromosomes and a high LAI score of 16.91. EDTA identified a total of 207 intact LTR-RTs elements, including 105 *Gypsy*, 75 *Copia*, and 27 unknown. For MegaLTR, a total of 203 intact LTR-RT elements were identified, classified, and annotated. Of the 203 intact LTR-RTs elements, 2 elements were classified as autonomous *Gypsy* and 201 as non-autonomous LTR-RTs. Non-autonomous elements included 118 *Gypsy*, 80 *Copia*, 1 *TR-GAG*, and 2 unknown (Table 3). Based on the position of the identified elements in the genome sequence, the LTR-RT results of EDTA and MegaLTR were compared. Of the 207 LTR-RTs identified by EDTA, 193 elements matched MegaLTR and 14 did not match. These 14 LTR-RTs included one element that did not pass the LTR_retriever filter and 11 elements that did not pass the TEsorter filter in MegaLTR. EDTA assigned these 11 elements to the NA class, consistent with their exclusion by MegaLTR. The remaining 2 elements were not found in the MegaLTR data. On the other hand, MegaLTR identified 10 LTR-RTs not found by EDTA (Figure S2A), including 7 *Gypsy* and 3 *Copia*. These elements are assigned to 7

clades, including 3 *Athila*, 2 *Retand*, 1 *Ale*, 1 *Ivana*, 1 *Reina*, 1 *SIRE*, and 1 *Tekay*. As for the internal domains, 6 elements contain all the domains necessary for transposition (*GAG*, PROT, INT, RT, RH) for Copia and (*GAG*, PROT, RT, RH, INT) for *Gypsy*. The remaining 4 elements include one element containing the domains *GAG* and PROT, 2 elements containing the domain PROT, and one containing the domain *GAG*. The annotation of MegaLTR's unique results suggests that MegaLTR is able to identify more intact LTR-RTs with a high degree of filtering and annotation. In contrast, EDTA reported a number of elements that do not belong to LTR (Data Sheet 5).

MegaLTR is able to classify intact LTR-RT elements into autonomous (*Gypsy*) and non-autonomous (*Copia, Gypsy*, and *TR-GAG*) based on their structure. In addition, MegaLTR is able to classify unknown elements into superfamilies. EDTA reported 27 unknown elements, while MegaLTR reported only 2 unknown elements. As shown in Figure (S2B), MegaLTR and EDTA have 2 unknown elements in common, while EDTA has 25 unique unknown elements. The 25 unknown elements include 12 elements that did not pass MegaLTR filtering steps and 13 elements that were annotated and classified as nonautonomous (*Copia* and *Gypsy*). Data Sheet 5 lists the common LTR-RTs, the unique LTR-RTs in EDTA, the unique LTR-RTs in MegaLTR, the unknown elements in EDTA, and the unknown elements in MegaLTR.

3.3 Runtime vs. number of CPUs

In MegaLTR, multithreading was implemented to reduce the execution time. By splitting the genome sequence into scaffold/chromosome without splitting the individual sequences, MegaLTR can analyze multiple sequences simultaneously using multiple CPU cores (threads). In the standalone version, the user can specify the number of threads to use, while the MegaLTR web server currently uses 56 CPU cores for parallel processing. We tested the effect of the number of threads on runtime using the *Brassica rapa* genome. This genome was selected based on its medium genome size (352.9 Mbp) and the number of pseudomolecules/scaffolds (1100). Figure 3 and Data Sheet 2 show the runtime for different CPU numbers from 1 to 30. To analyse the Brassica rapa genome using a single thread, MegaLTR required 1382 minutes, while using 2 threads reduced the runtime to 707 minutes and using 30 threads reduced the runtime to 102 minutes, demonstrating the gain achieved through parallel processing in MegaLTR.

3.4 Generated output

The web server and standalone version of MegaLTR automatically generate a series of tables, FASTA files, and images, some of which are listed in Table S2. These files contain tables with the position of the identified LTR-RT within the sequence, the start and end of all identified features, classification into autonomous, non-autonomous, superfamily and lineage levels, estimated insertion age, LTR-RT-gene chimeras and LTR-RTs-near genes. It also generates redundant and non-redundant LTR-RTs libraries in FASTA format. The full list of generated results can be found in the MegaLTR online documentation (https://github.com/MoradMMokhtar/MegaLTR). Using *Arabidopsis thaliana* genome, Figure 4 shows an example of statistical analysis of the length of LTR-RT, the age of insertion of LTR-RT, and visualization of the density of genes and LTR-RTs on chromosomes.

4 Conclusion and future directions

With the increasing availability of plant genome projects, researchers need accurate, robust, and easy-to-use pipelines for processing large amounts of data to study the effects of LTR-RTs on plant genome evolution and functionality. These pipelines, in the form of a web server, would be valuable for efforts to integrate LTR-RTs as a possible element for studying the gene regulatory system. MegaLTR is a web server and stand-alone pipeline that detects intact LTR-RTs at the whole-genome level and integrates multiple tools for homology-, structure-, and *de novo*-based identification, classification, and annotation of intact LTR-RT. In addition, a comprehensive pipeline is also needed to create a

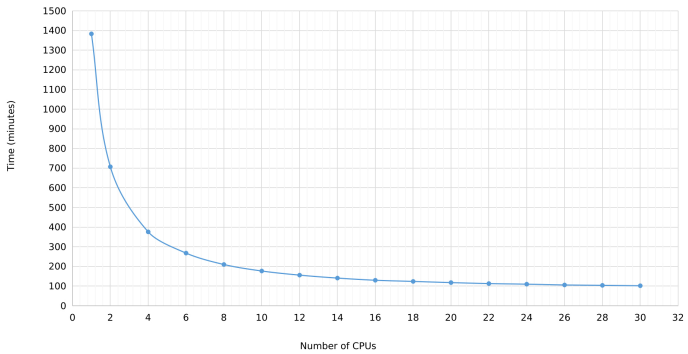


FIGURE 3
MegaLTR run time of *Brassica rapa* genome using different number of threads ranging from 1 to 30.

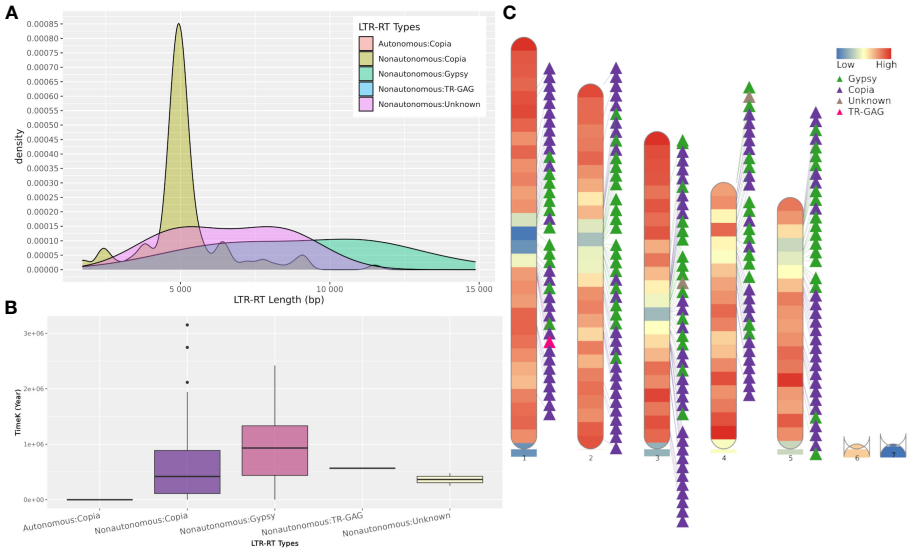


FIGURE 4
Example of MegaLTR generated results of *Arabidopsis thaliana*. **(A)** LTR-RTs length distribution. **(B)** Boxplot of insertion age, **(C)** Visualization of the density of genes and LTR-RTs on chromosomes.

non-redundant library of LTR-RTs for species that lack this resource for annotating whole genome LTR-RTs. MegaLTR is able to classify intact LTR-RT elements into putative autonomous (*Copia* and *Gypsy*) and non-autonomous (*Copia, Gypsy, LARD, TRIM, TR-GAG* and *BARE-2*), superfamily and lineage levels. It also identifies LTR-RT gene chimeras, detects LTR-RTs near genes, and provides statistical analysis and visualization of LTR-RT. For detection of LTR-RTs, MegaLTR shows high specificity, accuracy, precision, sensitivity and low FDR. The development of an online server such as MegaLTR, which provides computational resources for analyzing large amounts of genomic data, is becoming increasingly important for the automated analysis of LTR-RT elements. The current version of MegaLTR focuses on genome-level analysis LTR-RT, with work currently underway to integrate tools optimized for studying LTR-RTs at the transcriptomic level. MegaLTR web server is freely accessible at: https://bioinformatics.um6p.ma/MegaLTR and the standalone version at https://github.com/MoradMMokhtar/MegaLTR.

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. MegaLTR web server is freely available at: https://bioinformatics.um6p.ma/MegaLTR and the standalone version at https://github.com/MoradMMokhtar/MegaLTR.

Author contributions

Conceptualization MM and AA; Methodology MM and AA; Scripting MM and AA; Data curation MM; Writing–original draft MM and AA. All authors reviewed the manuscript. All authors contributed to the article and approved the submitted version.

Acknowledgments

The authors acknowledge the African Supercomputing Center at Mohammed VI Polytechnic University for supercomputing resources (https://ascc.um6p.ma/) made available for conducting the research reported in this paper.

Conflict of interest

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

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1237426/full#supplementary-material

References

Amselem, J., Cornut, G., Choisne, N., Alaux, M., Alfama-Depauw, F., Jamilloux, V., et al. (2019). RepetDB: a unified resource for transposable element references. *Mobile DNA* 10, 1–8. doi: 10.1186/s13100-019-0150-y

Bennetzen, J. L., and Wang, H. (2014). The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Annu. Rev. Plant Biol.* 65, 505–530. doi: 10.1146/annurev-arplant-050213-035811

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

Biryukov, M., and Ustyantsev, K. (2021). Darts: an algorithm for domain-associated retrotransposon search in genome assemblies. *Genes* 13, 9. doi: 10.3390/genes13010009

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). Blast+: architecture and applications. *BMC Bioinf.* 10, 1–9. doi: 10.1186/1471-2105-10-421

Chaparro, C., Gayraud, T., de Souza, R. F., Domingues, D. S., Akaffou, S., Laforga Vanzela, A. L., et al. (2015). Terminal-repeat retrotransposons with gag domain in plant genomes: a new testimony on the complex world of transposable elements. *Genome Biol. Evol.* 7, 493–504. doi: 10.1093/gbe/evv001

Eddy, S. R. (1998). Profile hidden markov models. *Bioinf. (Oxford England)* 14, 755–763. doi: 10.1093/bioinformatics/14.9.755

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than blast. *Bioinformatics* 26, 2460–2461. doi: 10.1093/bioinformatics/btq461

Eickbush, T. H., and Jamburuthugoda, V. K. (2008). The diversity of retrotransposons and the properties of their reverse transcriptases. *Virus Res.* 134, 221–234. doi: 10.1016/j.virusres.2007.12.010

Ellinghaus, D., Kurtz, S., and Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. *BMC Bioinf.* 9, 1–14. doi: 10.1186/1471-2105-9-18

Feschotte, C., Keswani, U., Ranganathan, N., Guibotsy, M. L., and Levine, D. (2009). Exploring repetitive dna landscapes using repclass, a tool that automates the classification of transposable elements in eukaryotic genomes. *Genome Biol. Evol.* 1, 205–220. doi: 10.1093/gbe/evp023

Franco-Zorrilla, J. M., Valli, A., Todesco, M., Mateos, I., Puga, M. I., Rubio-Somoza, I., et al. (2007). Target mimicry provides a new mechanism for regulation of microrna activity. *Nat. Genet.* 39, 1033–1037. doi: 10.1038/ng2079

Gao, X., Havecker, E. R., Baranov, P. V., Atkins, J. F., and Voytas, D. F. (2003). Translational recoding signals between gag and pol in diverse LTR retrotransposons. *RNA* 9, 1422–1430. doi: 10.1261/rna.5105503

Grandbastien, M.-A., Audeon, C., Bonnivard, E., Casacuberta, J. M., Chalhoub, B., Costa, A.-P., et al. (2005). Stress activation and genomic impact of Tnt1 retrotransposons in Solanaceae. *Cytogenetic Genome Res.* 110, 229–241. doi: 10.1159/000084957

Hao, Z., Lv, D., Ge, Y., Shi, J., Weijers, D., Yu, G., et al. (2020). Rideogram: drawing svg graphics to visualize and map genome-wide data on the idiograms. *PeerJ Comput. Sci.* 6, e251. doi: 10.7717/peerj-cs.251

Ito, H., Kim, J. M., Matsunaga, W., Saze, H., Matsui, A., Endo, T. A., et al. (2016). A stress-activated transposon in arabidopsis induces transgenerational abscisic acid insensitivity. *Sci. Rep.* 6, 23181. doi: 10.1038/srep23181

Jung, S., Venkatesh, J., Kang, M.-Y., Kwon, J.-K., and Kang, B.-C. (2019). A non-ltr retrotransposon activates anthocyanin biosynthesis by regulating a myb transcription factor in capsicum annuum. *Plant Sci.* 287, 110181. doi: 10.1016/j.plantsci.2019.110181

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase update, a database of eukaryotic repetitive elements. *Cytogenetic Genome Res.* 110, 462–467. doi: 10.1159/000084979

Kalendar, R., Vicient, C. M., Peleg, O., Anamthawat-Jonsson, K., Bolshoy, A., and Schulman, A. H. (2004). Large retrotransposon derivatives: abundant, conserved but nonautonomous retroelements of barley and related genomes. *Genetics* 166, 1437–1450. doi: 10.1534/genetics.166.3.1437

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

Kumar, A. (1998). The evolution of plant retroviruses: moving to green pastures. *Trends Plant Sci.* 3, 371–374. doi: 10.1016/S1360-1385(98)01304-1

Lee, H., Lee, M., Mohammed Ismail, W., Rho, M., Fox, G. C., Oh, S., et al. (2016). Mgescan: a galaxy-based system for identifying retrotransposons in genomes. *Bioinformatics* 32, 2502–2504. doi: 10.1093/bioinformatics/btw157

Lerat, E. (2010). Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. *Heredity* 104, 520–533. doi: 10.1038/hdy.2009.165

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659. doi: 10.1093/bioinformatics/btl158

Lopes, F. R., Jjingo, D., Silva, C. R. D., Andrade, A. C., Marraccini, P., Teixeira, J. B., et al. (2013). Transcriptional activity, chromosomal distribution and expression effects of transposable elements in coffea genomes. *PloS One* 8, e78931. doi: 10.1371/journal.pone.0078931

Ma, J., Devos, K. M., and Bennetzen, J. L. (2004). Analyses of ltr-retrotransposon structures reveal recent and rapid genomic dna loss in rice. *Genome Res.* 14, 860–869. doi: 10.1101/gr.1466204

McCarthy, E. M., and McDonald, J. F. (2003). LTR_STRUC: a novel search and identification program for ltr retrotransposons. *Bioinformatics* 19, 362–367. doi: 10.1093/bioinformatics/btf878

Mokhtar, M. M., Abd-Elhalim, H. M., and El Allali, A. (2023a). A Large-scale assessment of the quality of plant genome assemblies using the LTR assembly index. *AoB Plants* 15 (3). doi: 10.1093/aobpla/plad015

Mokhtar, M. M., Alsamman, A. M., Abd-Elhalim, H. M., and El Allali, A. (2021). Cicersptedb: A web-based database for high-resolution genome-wide identification of transposable elements in cicer species. *PloS One* 16, e0259540. doi: 10.1371/journal.pone.0259540

Mokhtar, M. M., Alsamman, A. M., and El Allali, A. (2023b). Plantltrdb: An interactive database for 195 plant species ltr-retrotransposons. *Front. Plant Sci.* 14, 1134627. doi: 10.3389/fpls.2023.1134627

Mokhtar, M. M., and El Allali, A. (2022). Pltrnadb: Plant transfer rna database. *PloS One* 17, 1–12. doi: 10.1371/journal.pone.0268904

Neumann, P., Novák, P., Hoštáková, N., and Macas, J. (2019). Systematic survey of plant LTRretrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. *Mobile DNA* 10, 1. doi: 10.1186/s13100-018-0144-1

Orozco-Arias, S., Humberto Lopez-Murillo, L., Candamil-Cortes, ´, M. S., Arias, M., Jaimes, P. A., Rossi Paschoal, A., et al. (2022). Inpactor2: a software based on deep learning to identify and classify ltr-retrotransposons in plant genomes. *Briefings Bioinf.* 24:bbac511. doi: 10.1093/bib/bbac511

Ou, S., Chen, J., and Jiang, N. (2018). Assessing genome assembly quality using the ltr assembly index (lai). *Nucleic Acids Res.* 46, e126–e126. doi: 10.1093/nar/gky730

Ou, S., and Jiang, N. (2017). 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, 1–3. doi: 10.1186/s13100-019-0193-0

Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J. R., Hellinga, A. J., et al. (2019). Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* 20, 1–18. doi: 10.1186/s13059-019-1905-y

Pereira, V. (2008). Automated paleontology of repetitive DNA with REANNOTATE. *BMC Genomics* 9, 614. doi: 10.1186/1471-2164-9-614

Pereira, A. (2016). Plant abiotic stress challenges from the changing environment. *Front. Plant Sci.* 7. doi: 10.3389/fpls.2016.01123

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

Shi, J., and Liang, C. (2019). Generic repeat finder: a high-sensitivity tool for genome-wide *de novo* repeat detection. *Plant Physiol.* 180, 1803–1815. doi: 10.1104/pp.19.00386

Smit, A., Hubley, R., and Green, P. (2015) *Repeatmodeler open-1.0. 2008–2015* (Seattle, USA: Institute for Systems Biology) (Accessed May 1, 2018). httpwww.repeatmasker.org.

Steinbiss, S., Willhoeft, U., Gremme, G., and Kurtz, S. (2009). Fine-grained annotation and classification of *de novo* predicted ltr retrotransposons. *Nucleic Acids Res.* 37, 7002–7013. doi: 10.1093/nar/gkp759

Su, W., Gu, X., and Peterson, T. (2019). Tir-learner, a new ensemble method for tir transposable element annotation, provides evidence for abundant new transposable elements in the maize genome. *Mol. Plant* 12, 447–460. doi: 10.1016/j.molp.2019.02.008

Tanskanen, J. A., Sabot, F., Vicient, C., and Schulman, A. H. (2007). Life without gag: The bare-2 retrotransposon as a parasite's parasite. *Gene* 390, 166–174. doi: 10.1016/j.gene.2006.09.009

Thompson, J. D., Gibson, T. J., and Higgins, D. G. (2003). Multiple sequence alignment using clustalw and clustalx. *Curr. Protoc. Bioinf.* 00, 2.3.1–2.3.22. doi: 10.1002/0471250953.bi0203s00

Ustyantsev, K., Novikova, O., Blinov, A., and Smyshlyaev, G. (2015). Convergent evolution of ribonuclease h in ltr retrotransposons and retroviruses. *Mol. Biol. Evol.* 32, 1197–1207. doi: 10.1093/molbev/msv008

Valencia, J. D., and Girgis, H. Z. (2019). Ltrdetector: A tool-suite for detecting long terminal repeat retrotransposons *de-novo*. *BMC Genomics* 20, 1–14. doi: 10.1186/s12864-019-5796-9

Vitte, C., Fustier, M.-A., Alix, K., and Tenaillon, M. I. (2014). The bright side of transposons in crop evolution. *Briefings Funct. Genomics* 13, 276–295. doi: 10.1093/bfgp/elu002

Vitte, C., and Panaud, O. (2005). Ltr retrotransposons and flowering plant genome size: emergence of the increase/decrease model. *Cytogenetic Genome Res.* 110, 91–107. doi: 10.1159/000084941

Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., et al. (2007). Database resources of the national center for biotechnology information. *Nucleic Acids Res.* 36, D13–D21. doi: 10.1093/nar/gkm1000

Wheeler, T. J., and Eddy, S. R. (2013). nhmmer: Dna homology search with profile hmms. *Bioinformatics* 29, 2487–2489. doi: 10.1093/bioinformatics/btt403

Wicker, T., Matthews, D. E., and Keller, B. (2002). TREP: a database for Triticeae repetitive elements, Dataset. *Trends Plant Sci.* 7, 561–562. doi: 10.1016/S1360-1385(02)02372-5

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., et al. (2007). A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* 8, 973–982. doi: 10.1038/nrg2165

Witte, C.-P., Le, Q. H., Bureau, T., and Kumar, A. (2001). Terminal-repeat retrotransposons in miniature (trim) are involved in restructuring plant genomes. *Proc. Natl. Acad. Sci.* 98, 13778–13783. doi: 10.1073/pnas.241341898

Xia, E., Tong, W., Hou, Y., An, Y., Chen, L., Wu, Q., et al. (2020). The reference genome of tea plant and resequencing of 81 diverse accessions provide insights into its genome evolution and adaptation. *Mol. Plant* 13, 1013–1026. doi: 10.1016/j.molp.2020.04.010

Xiong, W., He, L., Lai, J., Dooner, H. K., and Du, C. (2014). Helitronscanner uncovers a large overlooked cache of helitron transposons in many plant genomes. *Proc. Natl. Acad. Sci.* 111, 10263–10268. doi: 10.1073/pnas.1410068111

Xu, Z., and Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* 35, W265–W268. doi: 10.1093/nar/gkm286

Zhang, R.-G., Li, G.-Y., Wang, X.-L., Dainat, J., Wang, Z.-X., Ou, S., et al. (2022). Tesorter: an accurate and fast method to classify ltr-retrotransposons in plant genomes. *Hortic. Res.* 9, uhac017. doi: 10.1093/hr/uhac017

Zhao, D., Ferguson, A. A., and Jiang, N. (2016). What makes up plant genomes: The vanishing line between transposable elements and genes. *Biochim. Biophys. Acta (BBA)-Gene Regul. Mech.* 1859, 366–380. doi: 10.1016/j.bbagrm.2015.12.005

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 09 November 2023
DOI 10.3389/fpls.2023.1302337

OPEN ACCESS

EDITED BY
Manohar Chakrabarti,
The University of Texas Rio Grande Valley, United States

REVIEWED BY
Sanjay Joshi,
University of Kentucky, United States
Sareena Sahab,
Department of Economic Development Jobs Transport and Resources, Australia

*CORRESPONDENCE
Guglielmo Puccio
gugpuccio@gmail.com
Francesco Sunseri
francesco.sunseri@unirc.it
Francesco Mercati
francesco.mercati@ibbr.cnr.it

RECEIVED 26 September 2023
ACCEPTED 25 October 2023
PUBLISHED 09 November 2023

CITATION
Puccio G, Ingraffia R, Giambalvo D, Frenda AS, Harkess A, Sunseri F and Mercati F (2023) Exploring the genetic landscape of nitrogen uptake in durum wheat: genome-wide characterization and expression profiling of NPF and NRT2 gene families.
Front. Plant Sci. 14:1302337.
doi: 10.3389/fpls.2023.1302337

Exploring the genetic landscape of nitrogen uptake in durum wheat: genome-wide characterization and expression profiling of NPF and NRT2 gene families

Guglielmo Puccio[1,2*], Rosolino Ingraffia[1], Dario Giambalvo[1], Alfonso S. Frenda[1], Alex Harkess[3], Francesco Sunseri[2,4*] and Francesco Mercati[2*]

[1]Department of Agricultural, Food and Forestry Sciences, University of Palermo, Palermo, Italy, [2]Institute of Biosciences and BioResources (IBBR), National Research Council, Palermo, Italy, [3]HudsonAlpha Institute for Biotechnology, Huntsville, AL, United States, [4]Department Agraria , University Mediterranea of Reggio Calabria, Reggio Calabria, Italy

Nitrate uptake by plants primarily relies on two gene families: Nitrate transporter 1/peptide transporter (NPF) and Nitrate transporter 2 (NRT2). Here, we extensively characterized the NPF and NRT2 families in the durum wheat genome, revealing 211 NPF and 20 NRT2 genes. The two families share many Cis Regulatory Elements (CREs) and Transcription Factor binding sites, highlighting a partially overlapping regulatory system and suggesting a coordinated response for nitrate transport and utilization. Analyzing RNA-seq data from 9 tissues and 20 cultivars, we explored expression profiles and co-expression relationships of both gene families. We observed a strong correlation between nucleotide variation and gene expression within the NRT2 gene family, implicating a shared selection mechanism operating on both coding and regulatory regions. Furthermore, NPF genes showed highly tissue-specific expression profiles, while NRT2s were mainly divided in two co-expression modules, one expressed in roots (NAR2/NRT3 dependent) and the other induced in anthers and/ovaries during maturation. Our evidences confirmed that the majority of these genes were retained after small-scale duplication events, suggesting a neo- or sub-functionalization of many NPFs and NRT2s. Altogether, these findings indicate that the expansion of these gene families in durum wheat could provide valuable genetic variability useful to identify NUE-related and candidate genes for future breeding programs in the context of low-impact and sustainable agriculture.

KEYWORDS

durum wheat, nitrogen, N uptake, nitrate transporters, NPF and NRT2 family, Nitrogen Use Efficiency (NUE), N uptake, Weighted Gene Co-expression Network Analysis (WGCNA)

1 Introduction

Nitrogen (N) is a crucial nutrient for plant growth and development. Suboptimal nitrogen utilization can lead to diminished yields and significant environmental repercussions. Excessive or misapplied nitrogen fertilizers often lead to an increased risk of nitrogen escaping into the environment through processes like denitrification, leaching, or volatilization. This contributes to higher levels of nitrate in both surface and groundwater, as well as the release of N_2O and NH_3 into the atmosphere. Therefore, improving the efficiency of nitrogen utilization is crucial to address issues such as environmental degradation, climate change, and food security (Javed et al., 2022). Despite valuable research efforts in this field and the development of various technologies (i.e., slow-release fertilizers, inhibitors for nitrification and urease, fertigation, and advanced precision agriculture techniques) nitrogen efficiency remains relatively low for many crops, particularly for cereals where it typically ranges between 25% and 50% of the applied nitrogen (Giambalvo et al., 2018; Javed et al., 2022). This can be attributed to the complexity of Nitrogen Use Efficiency (NUE), which involves a multitude of factors related to agronomy, physiology, and molecular biology. Nitrate (NO_3^-) is one of the major N-forms taken up by plants from the soil. NO_3^- availability in the soil is highly variable and its uptake is governed by at least two transport systems, depending on soil NO_3^- concentrations: the low affinity NO_3^- transport (LATS) and the high affinity NO_3^- transport (HATS) systems. LATS is mediated by the NO_3^- transporter 1/peptide transporter (NRT1/NPF) family, which comprises a diverse array of membrane transport proteins found within multiple cell types and tissues, whereas HATS is facilitated by the NRT2 family, and is specific for NO_3^-. These two transport systems are responsible for the uptake of NO_3^- at different range of concentrations from millimolar to micromolar. N uptake is an important component of NUE, defined as the total biomass (or yield) per unit of N supplied (Moll et al., 1982), it is a complex trait influenced by interacting environmental factors and controlled by gene networks involved in N uptake, assimilation, and remobilization. NUE is divided in two main components, the Nitrogen Uptake Efficiency (NUpE), referred to the ability of the plant to take up N from the soil, and the Nitrogen Utilization Efficiency (NUtE), which encompasses the ability of the plant to assimilate, transfer, and utilize N to the harvestable part of the crop (Good et al., 2004; Xu et al., 2012).

The NPF and NRT2 families differ in both their structure and copy number across angiosperms. The NPF family harbors a conserved structural arrangement consisting of twelve transmembrane domains (TM) connected by short peptides and a central hydrophilic loop of about 90 amino acids between the sixth and the seventh TM domains (Wang et al., 2018b). They were previously known as NRT1s (NO_3^- transporters) and/or PTRs (peptide transporters) depending on their first discovered substrates. Based on a wide multi-species phylogenetic analysis, Léran et al. (2014) proposed a unified nomenclature for the NO_3^- transporter/Peptide transporter family (NPF), defining eight subfamilies (NPF1-8). The first NPF gene member isolated in plants and one of the most studied is the *Arabidopsis thaliana* NPF6.3 (*AtNPF6.3*), previously known as *CHL1/AtNRT1.1*. It is considered a dual-affinity NO_3^- transporter contributing to root NO_3^- uptake at both low (LATS) and high (HATS) NO_3^- availability, acting also as an NO_3^- sensor or 'transceptor' (Liu et al., 1999; Gojon et al., 2011; Xuan et al., 2017). *AtNPF6.3* can also act as a chlorate transporter (per the old name *CHL1* was awarded) when NO_3^- is less available and as an auxin transporter, a process negatively regulated by NO_3^- (Mounier et al., 2014; Maghiaoui et al., 2020; Meier et al., 2020). The interaction between auxin and NO_3^- is associated to NO_3^- sensing and it is involved in the regulation of N-dependent root development (Bouguyon et al., 2015). NPF proteins can transport a high number of different substrates other than NO_3^-, including phytohormones such as ABA and auxin, but also peptides, potassium, and secondary metabolites (Chiba et al., 2015; Tal et al., 2016; Kanstrup and Nour-Eldin, 2022). Although the NPF family is often involved in the LATS, many members also show high affinity transport in many species such as *ZmNPF6.6* and *MtNPF6.8* in maize and *Medicago truncatula*, respectively (Bagchi et al., 2012; Wen et al., 2017).

Land plant genomes typically contain a higher number of NPF/PTR genes compared to bacteria, animals, and algae, with 20 members in the moss *Physcomitrella patens*, 52 members in *Arabidopsis thaliana* and even more members in polyploid species such as *Brassica napus* (199) and *Triticum aestivum* (331) (Bajgain et al., 2018; Longo et al., 2018). In *Brassica napus*, allopolyploidy greatly contributed to the gene family expansion of the NPF family (Wen et al., 2020). A recent characterization of the NPF and NRT2 families in bread wheat also showed an expansion of these families (331 and 46, respectively) mainly due to tandem and segmental duplication (Bajgain et al., 2018; Li et al., 2021). The retention of multiple gene copies, after duplication, can be associated with the acquisition of new beneficial functions or the reduction of their full capacity, compared to that of the single-copy ancestral gene (Lynch and Conery, 2000). The high number of NPF genes in allopolyploid species suggested that the transporters encoded by these genes may have evolved for new unknown roles in plants (Corratg'e-Faillie and Lacombe, 2017; Longo et al., 2018). Thus, exploration for novel functions within these large gene families in polyploid crops is necessary. The NRT2 genes are primarily involved in HATS, and mainly active in roots, although some members are expressed in other tissues such as seeds or leaves to allow NO_3^- remobilization and storage (Chopin et al., 2007; Miller et al., 2007). Seven members were characterized in *Arabidopsis thaliana*, while five were detected in *Oryza sativa*. Similarly to the NPF family, a higher number of NRT2 members were discovered in allopolyploid species such as *Triticum aestivum* and *Brassica napus* with 47 and 17 genes, respectively (Tong et al., 2020; Li et al., 2021). This family was deeply studied in *Arabidopsis thaliana*, *AtNRT2.1* resulted the most studied member due to its main role in high affinity NO_3^- uptake in roots (Li et al., 2007). The NRT2 genes are usually identified based on the sequence homology to known NO_3^- transporters, then their functions are predicted through gene expression analysis and heterologous expression in *Xenopus* oocytes. Nonetheless, several studies on monocot species such as wheat and rice have highlighted high sequence divergence

with dicot species, making it hard to directly infer gene functions relying only on sequence identity (Plett et al., 2010; Pellizzaro et al., 2015; Wang et al., 2019a). Therefore, the utilization of multi-tissue and -condition expression data become mandatory to fully characterize these genes in monocot crops, mainly in the allopolyploids.

The hexaploid bread wheat (*Triticum aestivum* L.; genome AABBDD) is among the most important global crop species, shaped heavily by polyploidy and hybridization between the tetraploid durum wheat (*Triticum turgidum* L.; genome AABB) and *Aegilops tauschii* (genome DD). The NPF and NRT2 gene families have been investigated mainly in bread wheat, exploring their expression levels under different abiotic stresses, such as drought, salt and N deficiency, in response to Arbuscular Mycorrhizal Fungi (AMF), and in several tissues and development stages (Buchner and Hawkesford, 2014; Duan et al., 2016; Tian et al., 2017; Bajgain et al., 2018). Recently, the increase of grain NO_3^- uptake through the *TaNRT2.5* overexpression, localized in the grain cell tonoplast, was reported (Li et al., 2020). Many other studies highlighted improved crop yield, shoot biomass, and N uptake when NPF or NRT2 genes were overexpressed (Hu et al., 2015; Fan et al., 2016; Sol et al., 2019; Wang et al., 2021). Furthermore, the nucleotide variability in protein-coding regions of the NPF genes seems to affect NUE related traits such as yield and shoot N content (Li et al., 2021). These findings suggested that further efforts in the detection and functional characterization of both gene families may greatly aid the selection of N-use efficient wheat cultivars. The high wheat genetic variability, the high number of duplicated genes, and its economic relevance make this plant a key species for the screening of potentially beneficial genes.

Compared to bread wheat, much less is known about the phylogenetic diversity, evolution, and expression of the NPF and NRT2 gene families in tetraploid durum wheat, which is an important crop in the Mediterranean basin (Hawkesford, 2017; Hawkesford and Griffiths, 2019; Lupini et al., 2021). The detection of key genes involved in NO_3^- transport is a primary goal for NUE improvement, and a gene family comparison between durum and bread wheat can elucidate the impact of polyploidy on NUE components. Hence, it is crucial to undertake a thorough characterization and annotation of nitrate transporters in the durum wheat genome. In this study, we have identified and analyzed both NPF and NRT2 gene families, exploring their phylogenetic relationships, gene and protein structures, regulatory elements, and expression profiles within the durum wheat genome.

2 Materials and methods

2.1 NPF and NRT2 identification in durum wheat genome

To identify NPF and NRT2 genes in the durum wheat genome, the protein sequences of NPF and NRT2 genes of *Arabidopsis thaliana*, barley (*Hordeum vulgare*), maize (*Zea mays*), rice (*Oryza sativa*), and bread wheat (*Triticum aestivum*) were downloaded from Ensembl plants (http://plants.ensembl.org/). These sequences were used for a BLASTP search against the entire durum wheat proteome, also downloaded from Ensembl plants, using an e-value threshold of $1e^{-10}$ and a minimum sequence identity of 50%. The durum wheat BLASTP best hits were then used as input for HMMER3 (Mistry et al., 2013), using the hmmscan command and the 'Proton-dependent oligopeptide transporter family' (IPR000109) HMM profile with an e-value cut-off of $1e^{-}05$ for the NPF genes. Furthermore, Pfam (Bateman et al., 2004) and NCBI protein sequence analysis tools were used to check that all the NPF protein sequences belonged to the PTR2 family (PF00854) and that all the NRT2 protein sequences contained the NCBI conserved domain PLN00028. The final set of genes was then used to identify homologous groups. These were defined through a reciprocal BLASTN using nucleotide sequence identity >95%.

2.2 Motif discovery, TF binding site, CREs prediction, and gene structure analysis

Gene structure of both TdNPF and TdNRT2 family members using Webscipio2 was analyzed (Hatje et al., 2011). The Multiple Em for Motif Elicitation (MEME) suite (Bailey et al., 2015) was used to identify conserved motifs with the following parameters: classic mode algorithm, 6 and 100 for minimum and maximum motif width, and a maximum number of 30 motifs per sequence. Conserved motifs were further analyzed through the NCBI protein domain search tool (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) and the Conserved Domain Database (CDD), using an e-value threshold of 0.01. Transmembrane helices and protein localization prediction was performed using the TMHMM2.0 tool (Krogh et al., 2001) and both WoLF PSORT (Horton et al., 2007) and PProwler1.2 (Hawkins and Bodén, 2006), respectively. Chromosome location was extracted from the durum wheat genome annotation v1.0 and then displayed using the MG2C online tool (Chao et al., 2021). Significantly enriched chromosomal locations for both NPF and NRT2 were detected with ShinyGO (Ge et al., 2020) using a sliding window size of 6Mb and an FDR cutoff of 1e-05. The same tool was used to perform a GO enrichment analysis of both TdNPF and TdNRT2 genes.

Transcription Factor (TF) binding site prediction was performed on the promoter region using the binding site prediction tool of the Plant Transcription Factor Database (http://plantregmap.gao-lab.org/binding_site_prediction.php) with a p-value threshold of $1e^{-}06$ and the *Triticum aestivum* orthologs. The UniProtKB database (www.uniprot.org) was then used to extract protein domain information and annotation of the predicted TFs. Cis-regulatory Elements (CREs) in upstream promoter regions (– 2000 bp) of *TdNPFs* and *TdNRT2s* were predicted using PlantCARE (Lescot et al., 2002).

2.3 Collinearity and gene duplications analysis

The intraspecific collinearity was analyzed using both TdNPF and TdNRT2 gene sets. A reciprocal BLASTP was performed using an e-value threshold of $1e^{-10}$. MCScanX was used to evaluate collinearity and duplication events using an e-value threshold of $1e^{-05}$ and a match score of 50. MCScanX was also used to display the collinear blocks among five *Poaceae* species selected on the knowledge about the genesis of both durum and bread wheat (Kimber and Feldman, 1987; Matsuoka, 2011) (*Aegilops speltoides* Tausch: closer to B genome, *Triticum urartu*: A genome, *Triticum durum*: A and B genomes, *Triticum aestivum*: A, B and D genomes, and *Aegilops tauschii*: D genome). Collinear blocks between species were used for the evaluation of non-synonymous (Ka) and synonymous (Ks) values using TBtools (Chen et al., 2020). Tandem and collinear gene pairs inside the durum wheat genome were further used to evaluate both Ka and Ks using TBtools.

2.4 Phylogenetic analyses

Phylogenetic trees including *Arabidopsis thaliana* and *Oryza sativa* NPF and NRT2 genes and those here identified on durum wheat were constructed. The final dataset included 357 and 31 protein sequences for NPF and NRT2 families, respectively. Alignment was performed with the online tool CLUSTALW (Sievers et al., 2011) with default parameters. The unrooted phylogenetic tree was generated through the IQ-TREE software v. 2.2 (Nguyen et al., 2015) with the maximum likelihood method, 1000 bootstrap replicates, and the JTT + G4 model for both NPF and NRT2 trees, selected by the IQ-TREE best-fit model selection. Gene trees were visualized and analyzed through FigTree v. 1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/).

2.5 Expression profiles of TdNPF and TdNRT2 genes and co-expression analysis

A total of 195 wheat RNA-Seq datasets were downloaded from the Sequence Read Archive (SRA). These included 13 durum cultivars, 9 tissues, and 25 phenological stages (Zadoks scale: from Z12 to Z90) (Table S1). Raw reads were trimmed with the Trimmomatic tool (Bolger et al., 2014) using the options: LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:50. Clean reads were then quantified using Salmon (Patro et al., 2017) with default parameters and normalized through DESeq2 (Love et al., 2014). Reads were further filtered using the SVA package (Leek et al., 2012) to remove any batch effect or unwanted sources of variation using 10 surrogate variables. A co-expression network analysis was carried out by using the Weighted Gene Co-Expression Analysis (WGCNA) method (Langfelder and Horvath, 2008) with the following parameters: soft threshold=12, minimum module size=100, mergeCutHeight=0.3. Co-expression networks for each module were analyzed using Cytoscape (Shannon, 2003) and the hub genes for each network were selected using the CytoHubba pluging (Chin et al., 2014). Furthermore, module-trait (conditions) relationship was evaluated as correlation between the eigengenes of each module and a binary matrix representing each condition. Heatmaps were generated using the Pheatmap R package (Kolde and Kolde, 2018) using log-transformed normalized counts.

2.6 Data retrieval

The sequences and annotation files of all genomes were downloaded from the Ensembl plants database (http://plants.ensembl.org) (Bolser et al., 2016). The *Aegilops speltoides Tausch*. genome was obtained from the e!DAL - Plant Genomics & Phenomics Research Data Repository (https://doi.org/10.5447/ipk/2022/0) (Avni et al., 2022). The RNA-Seq datasets used for the expression profile and the co-expression analyses were obtained from the SRA archive (Leinonen et al., 2010) (Table S1).

3 Results

3.1 Durum wheat NPF and NRT2 genes identification

To identify NRT2 and NPF genes in durum wheat, a BLASTP search against all predicted protein sequences of the genome using the full-length amino acid sequences from five different plant species was carried out. The output of the BLAST search was further scanned using the HMMER3 tool with the 'Proton-dependent oligopeptide transporter family' profile (IPR000109) and the PLN00028 NCBI domain, and finally 211 and 20 NPF and NRT2 genes, respectively in the durum wheat genome were identified. *NPFs* and *NRT2s* showed 103 and 6 homologous groups, respectively, between A and B genomes.

The *TdNPFs* showed high variability in both gene length and amino acids content. The nucleotide sequences of the 211 genes showed a 3400 bp average gene length and encoded proteins ranging from 71 to 943 amino acids, with an average length of 583 amino acids, and molecular weights ranging from 7 to 105 kDa.Eighty percent of the TdNPF proteins showed 12 predicted transmembrane domains, while almost 95% of these proteins were localized in the plasma membrane (Figure S1). Like in durum wheat, NRT2 is a smaller gene family with a lower variability compared to the NPF family. The twenty TdNRT2 members showed a 1600 bp average gene length and encoded proteins ranging from 113 to 573 amino acids, with a mean length of 509 amino acids. Their molecular weights ranged from 12 to 62 kDa. Seventy-five percent of the TdNRT2 proteins showed 12 predicted transmembrane domains, while 90% were predicted to be localized in the plasma membrane (Table S2).

10.3389/fpls.2023.1302337

3.2 *TdNPFs* and *TdNRT2s* phylogenetic analysis

To explore the molecular evolution and the TdNPF gene family organization, we performed a phylogenetic analysis including protein sequences from *Arabidopsis thaliana* (53 *AtNPFs)*, *Oryza sativa* (93 *OsNPFs)*, and the 211 *TdNPFs* here identified in *Triticum durum* for a total of 357 NPF sequences. The Multiple Sequence Alignment (MSA) performed by CLUSTALW was used as input to IQ-TREE for both the model selection and the maximum-likelihood tree estimation. The phylogenetic tree showed a distinct clustering among the eight known NPF sub-families (Figure 1A). All the key nodes between sub-families are well supported with bootstrap values > 98 and all the genes from *Arabidopsis* and rice belonging to the same sub-family clustered together (Figure 1B). These results ensured the accuracy and reliability of the tree construction, suggesting a higher sequence variability between sub-families compared to the interspecific variability of each sub-family. The *TdNPFs* were assigned to the eight sub-families, namely from TdNPF1 to TdNPF8, following the tree topology and the previous classifications from other species. The sub-families TdNPF5 and TdNPF8 included the highest numbers of members (63 and 52, respectively), while TdNPF1 and TdNPF3 were the smaller sub-families with four and 8 genes, respectively (Table S2). TdNPF4, TdNPF5 and TdNPF6 were the only monophyletic groups, while the sub-families TdNPF1, TdNPF2, TdNPF3 and TdNPF7, TdNPF8 formed two distinct clusters with TdNPF1 clustering inside the TdNPF2 branch.

The NRT2 gene family was analyzed by using a similar approach; the maximum-likelihood phylogenetic tree was constructed based on the 31 NRT2 protein sequences, of which 7, 4 and 20 from *Arabidopsis*, rice and durum wheat, respectively. The phylogenetic tree revealed distinct evolutionary relationships among the NRT2 proteins of durum wheat and the other two species. Specifically, almost all the durum wheat NRT2 proteins formed a separate cluster, with only one protein (TdNRT2.2) closely grouped with the NRT2 proteins of *Oryza sativa* (OsNRT2.1 and 2.2) (Figure 1B). More interestingly, two other proteins (TdNRT2.19 and TdNRT2.20) showed a more ancient evolutionary divergence compared to all other TdNRT2 proteins, forming adistinct basal cluster, while all the AtNRT2 proteins clustered in three sub-clusters.

3.3 Chromosome location

The *TdNPFs* were evenly distributed along chromosomes in the A and B genomes (Figure S2A). The 2B chromosomal region (R2B) showed the highest gene density while the central chromosomal regions showed a lower gene density on average. Interestingly, four *NPFs*-enriched regions in chromosomes 2B, 3A, 3B, and 4B, also located in the R2B, were found, ranging from 5 to 7 genes per window (6Mb).

By contrast, the *TdNRT2s* were unevenly distributed along the genome, with chromosome 6 in both genomes (A and B) significantly enriched with 9 and 8 genes in 6B and 6A chromosomes, respectively (Figure S2B). Interestingly, all the *TdNRT2s* on chromosome 6 were located in the R1 in a

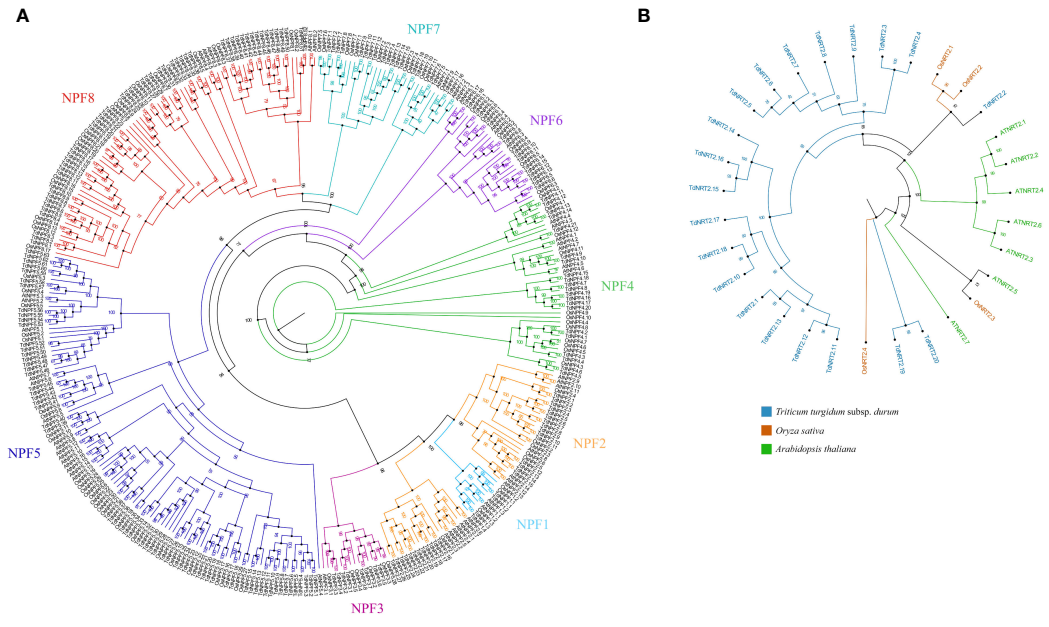


FIGURE 1
Phylogenetic analysis of the two gene families from *Arabidopsis thaliana*, *Oryza sativa* and *Triticum durum*. **(A)** Maximum likelihood tree of the NPFs full length protein sequences. The eight NPF sub-families are highlighted using colors: 1:Turquoise – 2:Orange – 3:Pink – 4:Green – 5:Blue – 6: Violet – 7:Cyan – 8:Red. **(B)** Maximum likelihood tree of the NRT2s full length protein sequences. The three species were highlighted using colors as indicated in the figure legend. Species specific branches were further highlighted using the same colors. Both trees were constructed using iqtree2.2, visualized, and modified using Figtree. Bootstrap values (1000 replicates) are shown for each node.

significantly enriched window in both genomes (A and B) while only two gene members were found in the R2B (Figure S2A).

3.4 Gene structure and conserved motifs prediction

The gene structure of both gene families showed a significant difference in the number of exons and transcript isoforms (Figure S3). Most of the 211 *TdNPFs* exhibited more than two exons, with 85% of genes ranging from 3 to 6. The distribution of the transcript isoforms number was significantly different compared to the genome, with an average number of 3 transcript isoforms per gene. The *TdNRT2* gene family showed a lower number of exons (9 out of 20 genes with one exon) and a lower number of transcripts per gene with more than 50% of genes showing only one transcript isoform.

To highlight conserved motifs and analyze their distribution among sub-families, protein motif analysis was carried out using the MEME tool. All the TdNPF proteins showed highly conserved motifs patterns (Figure S4). Despite that, the spatial organization and distance between conserved motifs were highly variable. The intra-motif variability was very high, with only few positions conserved in almost all the protein sequences (Figure S5).

Functional characterization of these motifs was performed using the NCBI protein domain search using the most represented sequence for each of the 25 conserved motifs. 14 motifs were assigned to the Major Facilitator Superfamily (MFS) while the remaining (11) were not assigned to any known protein domain. The motif#1 was identified in all the 211 NPF proteins, while the less conserved motif#19 and motif#21 were found only in 98 and 99 proteins, respectively. Furthermore, motif#18 (FILGN**EFFER** LAYYG), shared by 147 TdNPF proteins, contains the highly conserved ExxER/K peptide, suggesting its involvement in proton binding and transport. Among these sequences, both glutamic acids (E) were conserved in 80% of sequences, while the arginine residue (R) is less conserved. Rare motif variants such as ExxDR and ExxEE were also detected.

15 conserved motifs were identified in the TdNRT2 gene family (Figure S6). Nine out of 15 were assigned to the NO_3^- transmembrane transporter superfamily (PLN00028). Moreover, the distribution and position of the motifs created regular patterns and showed lower sequence variability compared to TdNPF family. Almost all the NRT2 genes shared many of the conserved motifs, except four highly variable genes (TRITD7Av1G231010, TRITD7Bv1G180680, TRITD2Av1G017380, TRITD6Bv1G008700).

3.5 Transcription factor binding sites, and CREs prediction

Transcription Factors (TFs) are essential for modulating gene transcription levels and many TFs directly regulate the expression of NPF and NRT2 genes (Marchive et al., 2013; Liu et al., 2017). We predicted the TF binding sites in promoter regions (3,000 bp upstream of transcription start site) of TdNPFs and TdNRT2 using the Binding Site Prediction tool of the PlantTFDB, and more than four thousand (4,072) binding sites for 163 TFs were identified in the promoter regions of 197 *TdNPFs*. The most abundant families of TFs were MYB, AP2, and NAC (Figure S7). One hundred twenty (120) binding sites for 53 TFs were detected in the promoter region of 19 *TdNRT2s*, of which the AP2 family resulted the most abundant. Interestingly, almost 96% (51 out of 53) of the TFs families were shared between NPF and NRT2 genes promoter region.

Cis-regulatory elements (CRE) are non-coding DNA regions also involved in the transcription regulation of neighboring genes (Bai et al., 2013). Here, we predicted CREs in the promoter regions of both *TdNPFs* and *TdNRT2s* using PlantCARE. Five thousand one hundred and twenty-one (5,121) CREs of 27 different types in the 211 promoter regions of TdNPF were found (Figure S7). The most abundant sites were the ABA responsive element (ABRE), DRE and MYB binding sites, activation sequence-1 (as-1), and the stress response element STRE, accounting for 70% of all the CREs. Other less abundant CREs were involved in light-response (G-box), biotic and abiotic stress response (MYC), and the common TATA-box and CAAT-box. One thousand five hundred and eighteen (1,518) CREs were predicted in the promoter regions of *TdNRT2s*. They were highly enriched in MYB and MYC binding sites, with many genes showing more than 5 sites in their upstream sequence accounting for almost 40% of CREs, in agreement with the previously described TF binding site prediction.

3.6 Expression profiles and co-expression analysis

The expression profiles of 211 TdNPF and 20 TdNRT2 genes were detected using publicly available RNA-seq datasets from the Sequence Reads Archive (SRA) covering 9 tissues at different growth stages from 13 different cultivars (Figure 2A). The hierarchical clustering based on TdNPF genes showed a clear tissue-specific signal, with almost all the samples from the same tissue clustering together (Figure 2B). The 211 NPF genes were further divided roughly into 9 clusters based on their expression patterns. These clusters ranged from 7 to 79 genes, with an average of 18 genes. Almost all the clusters showed the highest expression in roots, stem, and leaf, with five clusters being the most expressed (Cluster1-5). Interestingly, the cluster with the higher number of genes (Cluster 6) showed very constant low gene expression levels in almost all samples, except two small groups of genes induced in anthers, endosperm and roots.

TdNRT2 showed an opposite trend compared to TdNPF, with limited correlation between gene expression and tissue, except in the roots and seedlings (Figure S8). As might be expected, the higher gene expression was detected in roots, with 7 NRT2 members, highly similar to *AtNRT2.1* and *AtNRT2.4*, that showed higher expression (TdNRT2.3,.4,.5,.6,.7,.8,.9) (Figure S8).

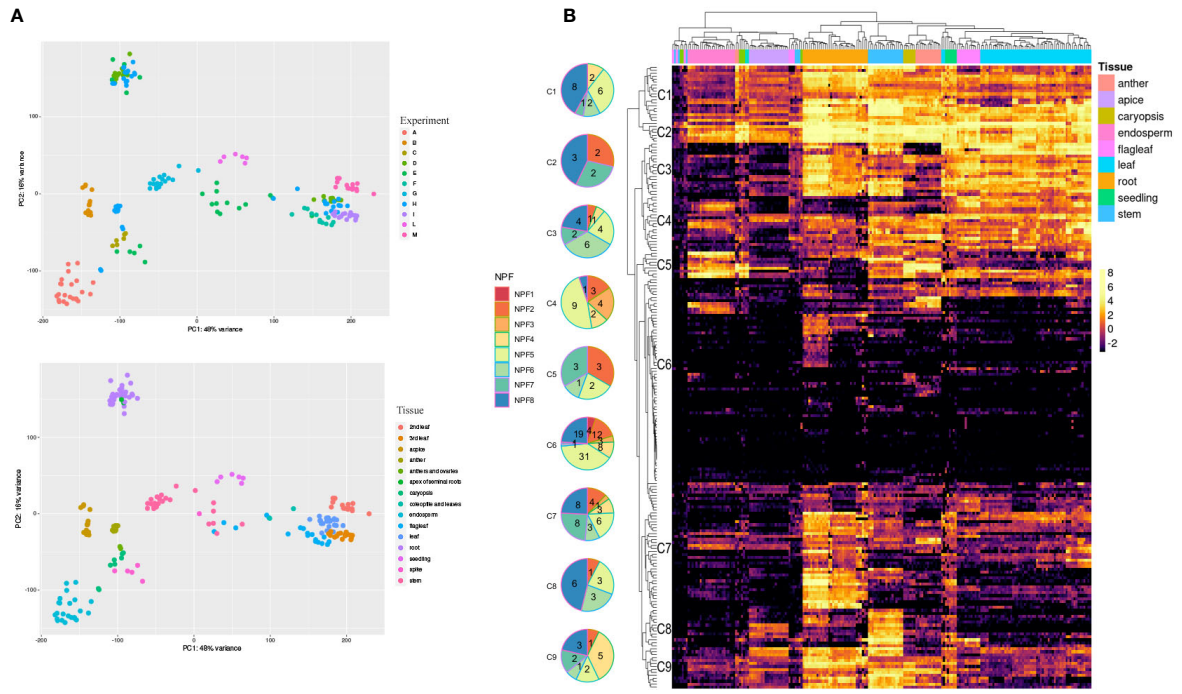


FIGURE 2
RNA-Seq of 195 samples obtained from the Short Reads Archive (SRA). **(A)** Principal Component Analysis performed with DESeq2 after the SVA correction. We highlighted SRA experiments (Top) and tissues (bottom). **(B)** Expression profiles of TdNPF genes in 9 tissues. Hierarchical clustering was performed both on rows and columns. Rows were roughly divided into 9 groups according to the similar expression levels. Pie charts were used to highlight the abundance of each NPF sub-family in each cluster. Tissues were highlighted as colored bars at the top of the heatmap.

The same dataset was used to evaluate the co-expression of both NRT2 and NPF gene families adopting the Weighted Gene Co-expression Network Analysis (WGCNA) method. We detected fourteen co-expression modules which showed highly variable expression profiles. 87.2% of durum wheat genes were assigned to co-expression modules, with four modules, colored salmon, red, blue and brown, showing a significantly higher number of genes ranging from roughly thirteen thousand to eight thousand. Module-tissue relationships were evaluated to highlight each module expression profile (Figure 3A). TdNRT2s were assigned to modules brown (8), green (8), red (2) and turquoise (2) which were highly induced in roots, anthers, endosperm-apex and leaves-flag leaves, respectively (Table S7). These expression profiles closely correlated to the phylogenetic tree distribution of the NRT2 genes, with almost all *TdNRT2* in the same co-expression module clustering together (Figure 3B). Furthermore, all six NAR2/NRT3 genes in the durum wheat genome were assigned to the brown module, highlighting their combined action mainly in roots.

TdNPFs showed a wider range of expression patterns, in agreement with the hierarchical clustering. They were assigned to ten of the fourteen modules detected. The majority of NPF genes belonged to the brown (52) and salmon (48) modules which were induced in root and flag leaf, respectively (Figure 3C). Interestingly 35 TdNPFs were assigned to three modules, red (23), yellow (9) and pink (3), highly upregulated in the caryopsis, especially in the endosperm, with three slightly different expression profiles (Figure S9). Furthermore, the network analysis allowed us to detect the hub-genes in each module. Among these we detected three NPF genes, *TdNPF6.12*, *TdNPF6.8* and *TdNPF5.61*, belonging to brown, red and salmon modules, respectively. We further used co-expression modules to detect expression patterns in homologous genes among the two genomes (A and B). The half (50.4%) of the TdNPFs homologous belonged to different co-expression modules while only two NRT2 genes did not cluster in the same module.

3.7 NPF gene sequence divergence and collinearity in five species of the Triticeae tribe

Evolutionary constraints of durum wheat NPF and NRT2 genes was evaluated through pairwise comparisons of Ka/Ks values from five species belonging to the Triticeae tribe (Figure 4). In detail, the durum wheat *TdNPFs* were compared to their orthologs in *T. urartu*, *Ae. speltoides*, *Ae. tauschii* and *T. aestivum* genomes. Furthermore, the Ka/Ks values of each duplicated TdNPF gene pairs were also evaluated. Ka/Ks was evaluated for 170, 91, 83, 74 orthologs in the *durum*/*aestivum*, *durum*/*speltoides*, *durum*/*urartu* and *aestivum*/*tauschii* comparisons, respectively. Interestingly, both the *durum*/speltoides and the *aestivum*/*tauschii* comparison showed very low Ka/Ks values with an average of 0.19 and 0.23, respectively, by contrast, the highest values were detected in the *durum*/*aestivum* comparison with an average of 0.49. Five genes (TdNPF3.5, TdNPF4.7, TdNPF5.42, TdNPF7.12, TdNPF8.38) exhibited Ka/Ks values greater than 1.5, indicating a substantial

10.3389/fpls.2023.1302337

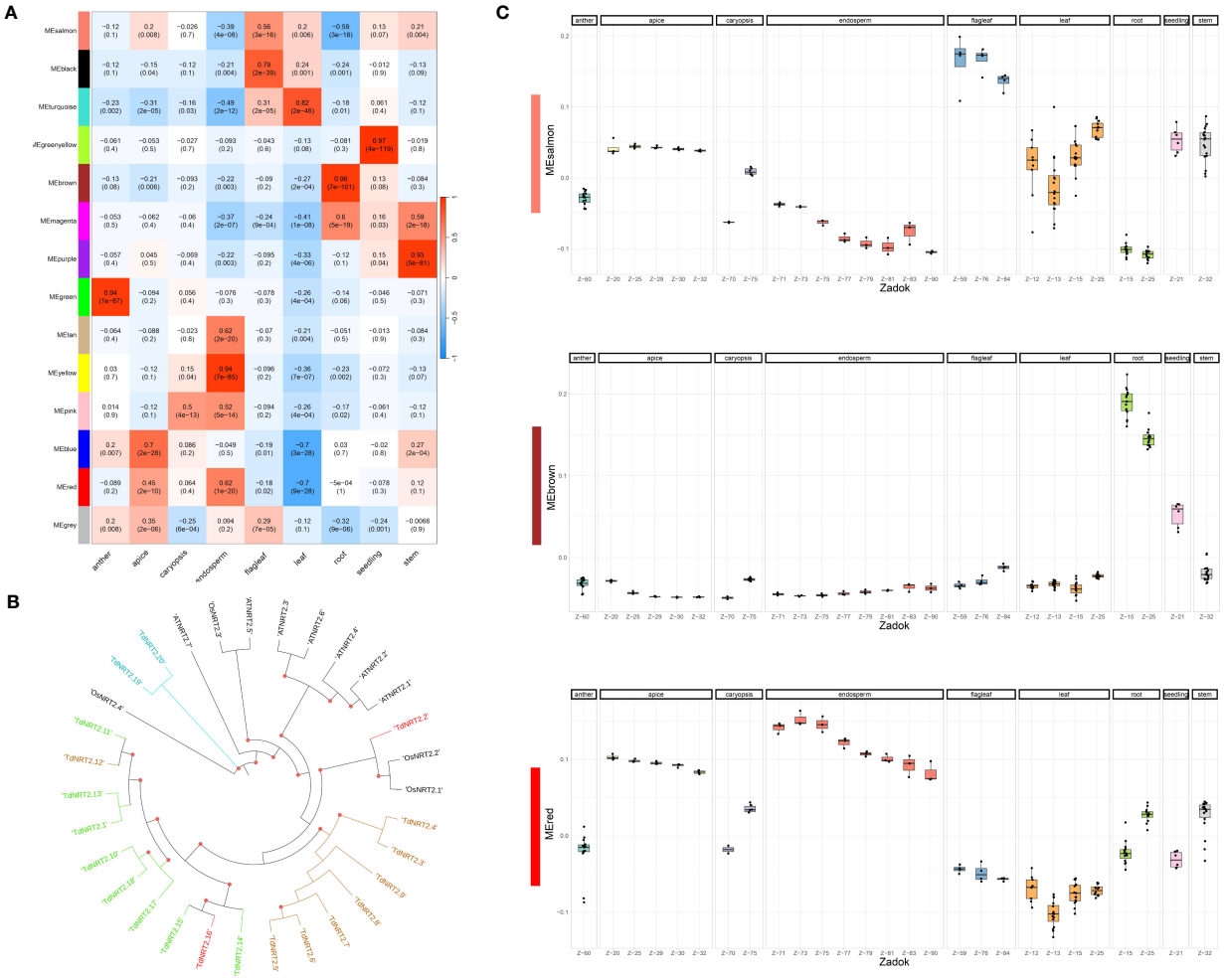


FIGURE 3
Weighted Gene Co-expression Network Analysis (WGCNA). **(A)** Heatmap of the Correlation coefficients between Module Eigengenes (WGCNA) and tissues. Pearson Correlation coefficient was evaluated using a binary matrix representing each tissue (1) against all the others (0). **(B)** Phylogenetic tree of TdNRT2 genes. Genes within each co-expression module form distinct clusters on the tree, visually distinguished by the use of module-specific colors (turquoise, red, brown and green). **(C)** Expression profile of the three most abundant co-expression modules visualized using the Boxplots of Module Eigengenes. Each tissue was highlighted in facets, while the developmental stage (Zadok) was highlighted on the X-axis.

positive selection acting on these genes. On the other hand, ten genes displayed a Ka/Ks value close to 1, suggesting a relatively neutral selection. Among the five genes with Ka/Ks > 1.5, three were associated with the brown co-expression module, indicating their upregulation specifically in roots. Based on these results, the significant difference between *durum*/*aestivum* and the other three comparisons was confirmed by Tukey's test (Figure S10).

Ka/Ks was evaluated also on durum wheat NPF and NRT2 tandem duplicated genes and NPF collinear genes between the two sub-genomes. All the gene pairs comparisons showed Ka/Ks values lower than 1, rarely higher than 0.5, suggesting strong purifying selection acting on duplicated genes, regardless the duplication event type. In particular, tandem duplications showed a slightly higher Ka/Ks among the NPF genes with an average of 0.31 compared to collinear duplicated NPF genes between sub-genome A and B (average 0.26). Finally, NRT2 showed a drastically lower Ka/Ks value ranging from 0.1 to 0.01.

Furthermore, using collinearity analysis through MCScanX we were able to characterize the relationships and the duplication events of both gene families inside the durum wheat genome and between these five species (Figure 5). In the durum wheat genome, almost 45% of *TdNPFs* were included in collinear pairs detected between A and B genomes. In detail, 77 segmental and 94 tandem duplications, as well as fewer dispersed (30) and proximal duplication (10) were detected; 42% of *TdNPFs* formed tandem blocks, with 11 blocks including three or more genes.

TdNRT2s are mainly located in two enriched regions on chromosomes 6A and 6B, as previously highlighted. These formed 5 tandem blocks, 3 and 2 located on 6B and 6A chromosomes, respectively. Furthermore, 14 tandem and no segmental duplications were detected.

Interspecific analysis of NPF genes revealed 23 and 25 collinear blocks in durum-speltoides and durum-urartu comparisons, respectively, 48 and 75 pairs in aestivum-aegilops and durum-aestivum comparisons, respectively. Almost all the blocks were detected between homologous chromosomes among the five genomes, significant differences in the number of blocks between A, B or D sub-genomes were not detected.

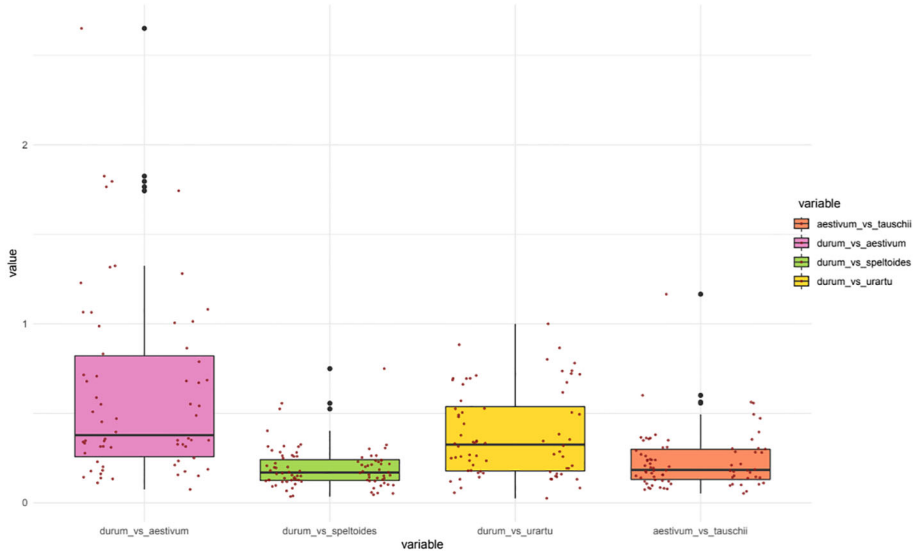


FIGURE 4
Boxplots of the Ka/Ks values for the orthologous NPF genes between five poaceae species: Durum wheat (*Triticum turgidum* subsp. *durum*), Bread wheat (*Triticum aestivum*), *Aegilops speltoides* Tausch, *Triticum urartu* and *Aegilops tauschii*. Two scatter plots (left: TdNPF on the A sub-genome; right: TdNPF on the B sub-genome) highlights the single Ka/Ks for each gene pairs.

4 Discussion

In plants, both the NPF and the NRT2 gene families are involved in nitrate/nitrite uptake, translocation and remobilization. *NPFs* are also involved in the transport of many other substrates such as hormones, secondary metabolites, peptides, chloride and potassium. A deeper characterization of these gene families is crucial to understand plant nitrate and metabolite transport.

In the present study, 211 *TdNPFs* and 20 *TdNRT2s* were identified in the *Triticum turgidum* L. subsp. *durum* (Desf.) Husn. genome. These numbers were comparable to other allopolyploid species such as *Brassica napus* (199 NPFs, 17 NRT2s), *Saccharum spontaneum* (178, 20), and *Triticum aestivum* (331, 46) and significantly higher than many diploid monocots and dicots such as *Arabidopsis thaliana* (53 and 7), *Oryza sativa* (93, 4) and *Zea mays* (79, 1). The NPF gene family expansion in plants seems to have arisen from neo- and sub-functionalization, as suggested by many reports (Lynch and Conery, 2000; O'Brien et al., 2016; Jørgensen et al., 2017; Wang et al., 2019a). In wheat, the large number of members in both gene families could be involved in highly differentiated responses to the availability of various substrates. Indeed, the high number of TdNPF and TdNRT2 genes, deriving from recent polyploidization and duplication events, may provide a higher modularity in terms of substrate affinity, condition or tissue specific gene expression induction and new protein-protein interactions. Similar effects were reported in many allopolyploid species such as rice, soybean, cotton and, in the MIKC-type MADS-box gene group, in bread wheat (Flagel et al.,

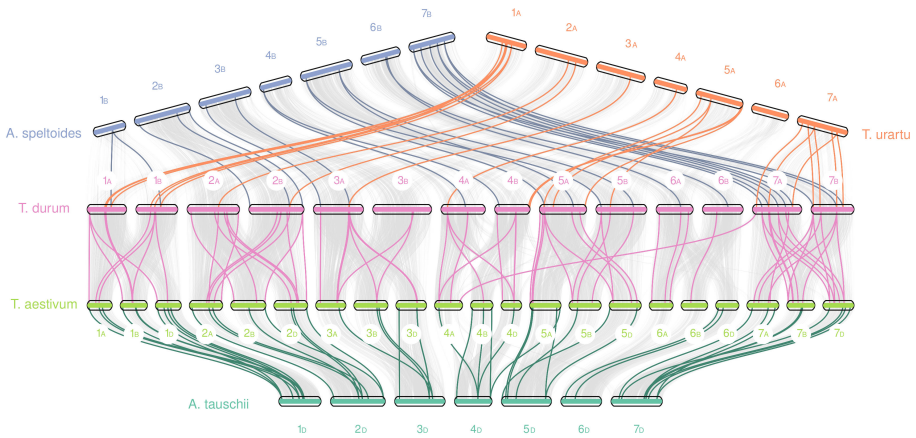


FIGURE 5
Collinearity analysis of NPF genes between five species from the Tritiitceae tribe. Grey lines indicate collinear blocks between genomes, while the colored lines indicate NPF genes detected inside collinear blocks. Sub-genomes were highlighted using the A, B or D letters.

2008; Schilling et al., 2020; Lee and Szymanski, 2021). In *Triticum aestivum*, the diversification of both NPF and NRT2 gene families was likely due to drift, leading to significant differences in the N-use efficiency of subpopulations clustered based on Single Nucleotide Polymorphisms (SNPs) within the NPF and NRT2 genes (Li et al., 2021).

Our phylogenetic analysis highlighted a high divergence among NRT2 genes from durum wheat, rice and *Arabidopsis* with genes from each species clustering together in distinguished groups. The best hit (BLASTP) of many *TdNRT2s* against the *Arabidopsis* NRT2 protein sequences is often *AtNRT2.1*, making it hard to functionally characterize these transporters solely based on their sequence homology, thus the association of previous functional annotations from model species to the newly identified genes in durum wheat become challenging. Here, we focused on sequence, gene expression and protein domains characterization, but further studies will be needed to fully describe *TdNRT2s* at a functional level.

The phylogenetic analysis on the NPF genes yielded more informative results, with all the orthologous genes belonging to the same sub-family clustering together, allowing us a much more reliable annotation of the novel *TdNPFs*. These results support the hypothesis of NPF family divergence before the separation of monocots and dicots, as suggested by Wang et al. (2019b). Furthermore, the NPF1 branch clustered inside the NPF2 branch, breaking it in two sub-classes. This feature was observed in other species such as *Brassica napus* and often led to the definition of more than eight sub-families, with NPF2 split into NPF2-1 and NPF2-2(Wen et al., 2020). Interestingly, this split was not detected in *Triticum aestivum* (Li et al., 2021; Kumar et al., 2022) potentially due to slightly different plant species utilized or to slightly different clustering method.

The chromosome location of both transporter families showed a non-random distribution inside the genome. Interestingly, *TdNRT2s* are highly concentrated on chromosomes 6 from both genomes (A and B), probably due to the multiple tandem duplication events, as shown by the results of collinearity analysis. In *Arabidopsis*, the *AtNRT2.1* and *AtNRT2.2* genes are adjacent, and end to end on chromosome 1, and this apparent duplication has been seen for orthologues in other species. AtNRT2.5 is also located on chromosome 1. Three other NRT2s are located on chromosome 5, with only AtNR2.6 located on chromosome 3. A similar enrichment on chromosome 6 of all three genomes (A, B and D) was detected in bread wheat, also deriving from tandem duplication that was suggested to have arisen from unequal crossing-over events (Li et al., 2021). Although similar, the number of NRT2 genes in these genomic regions is higher in bread wheat, suggesting that some of these duplication events should have occurred after or during the hybridization of durum wheat with *Aegilops tauschii* (genome D) as supposed by the International Wheat Genome Sequencing Consortium (IWGSC) (2014). Further studies on the intraspecific variability of these gene families among the main wheat species could help to deeply understand how their expansions occurred and what type of mechanisms underlie their preservation after duplication.

In *Arabidopsis*, several putative NO_3^- response cis-regulatory elements (CREs) have been detected in many promoters of N-related genes, while limited information is available for other plant species (Konishi and Yanagisawa, 2010; Wang et al., 2010; Rolly and Yun, 2021). Here, a high number of CREs related to ABA signaling and binding of Drought Responsive Element (DRE) and MYB TFs were detected in the *TdNPFs* promoter regions. Interestingly, the *TdNRT2s* upstream region was also highly enriched with MYB binding sites, as shown by both CREs and TF binding site prediction. MYB TFs are often involved in abiotic and biotic stress responses as well as in plant development, root and flower development (Kaur et al., 2017), although their role in NO_3^-related regulation has also been reported (Todd et al., 2004; Wang et al., 2018a; Zhang et al., 2021; Puccio et al., 2022). Interestingly, both gene family promoters showed multiple putative MYB binding sites for many genes. The presence of multiple binding sites for the same TF on the promoter of one or more genes has been often associated with a higher sensitivity to specific TFs (Howard and Davidson, 2004; Yáñez-Cuna et al., 2013; Brendolise et al., 2017).

The presence of many CREs and the partial overlap of their functions between the two gene families suggested that a complex regulatory network may be involved in modulating and fine-tuning their expression, with some TFs putatively involved in the regulation of members of both families. These could be involved in the spatiotemporal- or tissue-specific activation of transporter genes or may take part in the signaling cascade in response to the fluctuations of specific substrates concentration into the soil. Interestingly, the same analysis performed on the NPFs from *Brassica napus* yielded similar results on NPF genes (Wen et al., 2020), suggesting that the regulation of this gene family may involve the same TFs classes and could be evolutionarily conserved.

Fourteen motifs were assigned to the Major Facilitator Superfamily (MFS), the remaining 11 were not assigned to any known protein domain, suggesting a highly specific function for these peptides (putatively species-specific). Interestingly, 11 TdNPF7 proteins showed the monocot-specific variant ExxES, which is associated to non-proton dependent nitrate uptake and specific to the NPF7 sub-family (Longo et al., 2018). These genes were defined as NPF7a in rice, involved in the low-affinity nitrate transport system with some being tonoplast located (Hu et al., 2016).

Overall, *NPFs* showed a more variable gene structure and sequence variability in their conserved motifs compared to *NRT2s*. Interestingly, the *NRT2s* showed a simpler gene structure, with only one or two exons and highly clustered genomic locations. Both the gene structure and the chromosome locations of *NRT2s* seemed highly conserved among many monocots such as *Brachypodium distachyon*, *Saccharum spontaneum*, and bread wheat, mainly distributed on two chromosomes and harboring mainly one or two exons (Wang et al., 2019a; Li et al., 2021).

The high percentage of *NPFs* deriving from segmental and tandem duplication (37% and 44%, respectively) found in durum wheat supported the role of these genomic events in the expansion of NPF genes already reported in bread wheat (Li et al., 2021). Furthermore, the collinear analysis between durum and bread wheat and the putative A, B and D genomes was not able to detect a significantly higher number of collinear blocks between each putative sub-genome donor and the respective sub-genomes in durum or bread wheat. This observation does not directly support the idea that the expansion of these families derives from ancient duplication events in diploid wheat species, which should have

occurred before hybridization into allopolyploid species (Salse et al., 2008). Instead, our results indirectly support the idea that the substantial increase in gene members in both these families is mainly due to tandem and segmental duplications in the tetraploid or hexaploid ancestral genomes, and not in the diploid ancestral genomes. These duplications seemed favored by polyploidization events, with bread wheat showing a higher number of duplication events (Buchner and Hawkesford, 2014; Kumar et al., 2022). Furthermore, tandem duplicated NPF genes in durum wheat genomes showed strong purifying selection, suggesting preserved function after duplication, in agreement with many studies on other gene families (Hu et al., 2018; Hajiahmadi et al., 2020; Zhu et al., 2020).

In allopolyploid species, gene expression patterns can be significantly altered and this is one of the main sources of phenotypic variation (Jackson and Chen, 2010). Here, by using 195 RNA-seq durum wheat datasets the expression profiles highlighted different trends in both gene families. The *TdNPF*s expression patterns resulted highly tissue-specific, with most samples from specific tissue forming distinct clusters. By contrast, NRT2 genes were predominantly expressed in roots and anthers, being assigned to brown and green modules. The distinctiveness of these two groups of *TdNRT2*s becomes even more evident, as we observed that all the NAR2/NRT3 genes are present in the brown module. This finding implies that most of the NRT2 genes in wheat are either engaged in root uptake, facilitated by NAR2/NRT3, or have undergone evolutionary adaptations for translocation or accumulation in anthers/seeds. Additionally, four members of *TdNRT2* showed a more complex expression profile, with *TdNRT2.2* and *TdNRT2.16* being highly induced in apex, grain and endosperm during maturation, while *TdNRT2.19* and *TdNRT2.20* being expressed in leaves and flag-leaves. The detection of NRT2 genes responsible for seed N-accumulation, such as *TdNRT2.2* and *TdNRT2.16*, could be crucial to increase yield and higher N content, as already demonstrated by their overexpression in bread wheat (Li et al., 2020). *TdNRT2.19* and *TdNRT2.20* were the most basal genes in our phylogenetic analysis together with *OsNRT2.4* and *AtNRT2.7* in agreement with recent reports (Li et al., 2021; Deng et al., 2023; Kumar et al., 2023). Interestingly, both *OsNRT2.4* and *AtNRT2.7* are mainly expressed in the tonoplast of maturing seeds and roots, which seems to suggest differentiated functions of these basal genes in the vacuole (Chopin et al., 2007; Wei et al., 2018) and contrasting with most other family members located in the plasma membrane.

TdNRT2.2 was closely related to both *OsNRT2.1* and *OsNRT2.2*, which are usually expressed in root and germinating seeds (Feng et al., 2011). Furthermore, both phylogenetic and co-expression clustering yielded mostly the same results, with almost all the *TdNRT2s* from the same phylogenetic branch belonging to the same co-expression module. These results highlighted a close relationship between nucleotide variation and gene expression in this family, suggesting a shared selection mechanism between coding and regulatory regions. Similar coordinated evolution has been already observed in many gene families in mammals and plants (Necsulea and Kaessmann, 2014; Wang et al., 2020; Winkelmüller et al., 2021). Furthermore, duplication events may induce expression shifts favored by gene neo-functionalization as suggested by Fukushima and Pollock (2020), and this hypothesis could enhance the co-evolution of genome and transcriptome in the *NRT2* gene family in durum wheat.

The expression profiles of homologous genes showed significant variation, mainly within the *NPF* family. Indeed, about half of the NPF homologues exhibited dissimilar expression patterns. Although these differences may have already been present in ancestral genomes, the maintenance or development of highly similar genes with different expression patterns may provide a greater degree of modularity for regulation.

5 Conclusions

Our approach led to a comprehensive characterization of the NPF and NRT2 gene families in the durum wheat genome. Manual annotation of these transporters is crucial for understanding NO_3^- and N dynamics and their impact on NUE in durum wheat. This study identified 211 TdNPF and 20 TdNRT2 genes for the first time, providing detailed insights into their protein sequences and conserved domains and on their regulatory elements. By extensively analyzing nearly all publicly available RNA-seq datasets, we achieved the most comprehensive characterization of both gene expression profiles and co-expression relationships. This investigation confirmed that a considerable number of these genes underwent neo- or sub-functionalization following small-scale duplication events. These findings indicate that the expansion of these gene families in wheat holds promising potential as a valuable resource for identifying NUE-related genes and as potential candidates for molecular markers and the development of transgenic plants. By understanding the key players involved in durum wheat production and incorporating these findings into future research, we can take significant steps towards more eco-friendly and sustainable durum wheat fertilization management, addressing a critical challenge in modern agriculture.

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

GP: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. RI: Conceptualization, Supervision, Validation, Visualization, Writing – review & editing. DG: Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing – review & editing. AF: Funding acquisition, Project administration, Resources, Supervision, Writing – review & editing. AH: Conceptualization, Formal Analysis, Investigation, Methodology, Supervision, Writing – review & editing.

FS: Conceptualization, Formal Analysis, Investigation, Methodology, Resources, Supervision, Validation, Writing – review & editing. FM: Conceptualization, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Validation, Visualization, 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 Agritech National Research Center funded within the European Union Next-Generation EU program (the National Recovery and Resilience Plan, mission 4, component 2, investment 1.4 – D.D. 1032 del 17/06/2022, project CN00000022). This manuscript reflects only the authors' views and opinions, neither the European Union nor European Commission can be considered responsible for them. This work was also partially supported by the ANCIENT project (n. 942500454), granted by the Assessorato Regionale dell'agricoltura, dello sviluppo rurale e della pesca mediterranea, Regione Sicilia (P.S.R. Sicilia. 2014/2020, Sottomisura 16.1, DRS n. 3390/2022 del 11.08.2022). This research was also partially funded by BIAS - Innovative biofertilizers for sustainable agriculture to protect human health and the environment (project number 082015000275), funded by Sicily Region through the European Regional Development Fund (PO-FESR Sicilia 2014-2020).

Conflict of interest

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

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

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1302337/full#supplementary-material

References

Avni, R., Lux, T., Minz-Dub, A., Millet, E., Sela, H., Distelfeld, A., et al. (2022). Genome sequences of three Aegilops species of the section Sitopsis reveal phylogenetic relationships and provide resources for wheat improvement. *Plant J.* 110, 179–192. doi: 10.1111/tpj.15664

Bagchi, R., Salehin, M., Adeyemo, O. S., Salazar, C., Shulaev, V., Sherrier, D. J., et al. (2012). Functional assessment of the Medicago truncatula NIP/LATD protein demonstrates that it is a high-affinity nitrate transporter. *Plant Physiol.* 160, 906–916. doi: 10.1104/pp.112.196444

Bai, H., Euring, D., Volmer, K., Janz, D., and Polle, A. (2013). The nitrate transporter (NRT) gene family in poplar. *PloS One* 8, e72126. doi: 10.1371/journal.pone.0072126

Bailey, T. L., Johnson, J., Grant, C. E., and Noble, W. S. (2015). The MEME suite. *Nucleic Acids Res.* 43, W39–W49. doi: 10.1093/nar/gkv416

Bajgain, P., Russell, B., and Mohammadi, M. (2018). Phylogenetic analyses and in-seedling expression of ammonium and nitrate transporters in wheat. *Sci. Rep.* 8, 1–13. doi: 10.1038/s41598-018-25430-8

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., et al. (2004). The Pfam protein families database. *Nucleic Acids Res.* 32, D138–D141. doi: 10.1093/nar/gkh121

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

Bolser, D., Staines, D. M., Pritchard, E., and Kersey, P. (2016). Ensembl plants: integrating tools for visualizing, mining, and analyzing plant genomics data. *Plant Bioinformatics: Methods Protoc.* 1374, 115–140. doi: 10.1007/978-1-4939-3167-5_6

Bouguyon, E., Brun, F., Meynard, D., Kubeš, M., Pervent, M., Leran, S., et al. (2015). Multiple mechanisms of nitrate sensing by Arabidopsis nitrate transceptor NRT1.1. *Nat. Plants* 1, 15015. doi: 10.1038/nplants.2015.15

Brendolise, C., Montes, E., Hummel, I., Gouesnard, B., Glöckler, J., Martin, O., et al. (2017). Multiple copies in maize of Male Transmission Ratio Distortion (MTRD) region corresponding to an insertion in the reference genome. *J. Exp. Bot.* 68, 2075–2086. doi: 10.1093/jxb/erx090

Buchner, P., and Hawkesford, M. J. (2014). Complex phylogeny and gene expression patterns of members of the NITRATE TRANSPORTER 1/PEPTIDE TRANSPORTER family (NPF) in wheat. *J. Exp. Bot.* 65, 5697–5710. doi: 10.1093/jxb/eru231

Chao, H., He, J., Cai, Q., Zhao, W., Fu, H., Hua, Y., et al. (2021). The expression characteristics of NPF genes and their response to vernalization and nitrogen deficiency in rapeseed. *Int. J. Mol. Sci.* 22, 4944. doi: 10.3390/ijms22094944

Chen, C., Chen, H., Zhang, Y., Thomas, H. R., Frank, M. H., He, Y., et al. (2020). TBtools: an integrative toolkit developed for interactive analyses of big biological data. *Mol. Plant* 13, 1194–1202. doi: 10.1016/j.molp.2020.06.009

Chiba, Y., Shimizu, T., Miyakawa, S., Kanno, Y., Koshiba, T., Kamiya, Y., et al. (2015). Identification of Arabidopsis thaliana NRT1/PTR FAMILY (NPF) proteins capable of transporting plant hormones. *J. Plant Res.* 128, 679–686. doi: 10.1007/s10265-015-0710-2

Chin, C. H., Chen, S. H., Wu, H. H., Ho, C. W., Ko, M. T., and Lin, C. Y. (2014). cytoHubba: identifying hub objects and sub-networks from complex interactome. *BMC Syst. Biol.* 8, 1–7. doi: 10.1186/1752-0509-8-S4-S11

Chopin, F., Orsel, M., Dorbe, M. F., Chardon, F., Truong, H. N., Miller, A. J., et al. (2007). The Arabidopsis ATNRT2. 7 nitrate transporter controls nitrate content in seeds. *Plant Cell* 19, 1590–1602. doi: 10.1105/tpc.107.050542

Corratg'e-Faillie, C., and Lacombe, B. (2017). Substrate (un)specificity of arabidopsis NRT1/PTR FAMILY (NPF) proteins. *J. Exp. Bot.* 68, 3107–3113. doi: 10.1093/jxb/erw499

Deng, Q.-Y., Luo, J.-T., Zheng, J.-M., Tan, W.-F., Pu, Z.-J., and Wang, F. (2023). Genome-wide systematic characterization of the NRT2 gene family and its expression profile in wheat (Triticum aestivum L.) during plant growth and in response to nitrate deficiency. *BMC Plant Biol.* 23, 353. doi: 10.1186/s12870-023-04333-5

Duan, J., Tian, H., and Gao, Y. (2016). Expression of nitrogen transporter genes in roots of winter wheat (Triticum aestivum L.) in response to soil drought with contrasting nitrogen supplies. *Crop Pasture Sci.* 67, 128–136. doi: 10.1071/CP15152

Fan, X., Tang, Z., Tan, Y., Zhang, Y., Luo, B., Yang, M., et al. (2016). Overexpression of a pH-sensitive nitrate transporter in rice increases crop yields. *Proc. Natl. Acad. Sci.* 113, 7118–7123. doi: 10.1073/pnas.1525184113

Feng, H., Yan, M., Fan, X.-L., Li, B.-J., Shen, Q., Miller, A. J., et al. (2011). Spatial expression and regulation of rice high-affinity nitrate transporters by nitrogen and carbon status. *J. Exp. Bot.* 62, 2319–2332. doi: 10.1093/jxb/erq403

Flagel, L. E., Udall, J. A., Nettleton, D., and Wendel, J. F. (2008). Duplicate gene expression in allopolyploid Gossypium reveals two temporally distinct phases of expression evolution. *BMC Biol.* 6, 16. doi: 10.1186/1741-7007-6-16

Fukushima, K., and Pollock, D. D. (2020). Amalgamated cross-species transcriptomes reveal organ-specific propensity in gene expression evolution. *Nat. Commun.* 11, 4459. doi: 10.1038/s41467-020-18090-8

Ge, S.-X., Jung, D., and Yao, R. (2020). ShinyGO: a graphical gene-set enrichment tool for animals and plants. *Bioinformatics* 36, 2628–2629. doi: 10.1093/bioinformatics/btz931

Giambalvo, D., Amato, G., Badagliacca, G., Ingraffia, R., Di Miceli, G., Frenda, A. S., et al. (2018). Switching from conventional tillage to no-tillage: Soil N availability, N uptake, 15N fertilizer recovery, and grain yield of durum wheat. *Field Crops Res.* 218, 171–181. doi: 10.1016/j.fcr.2018.01.018

Gojon, A., Krouk, G., Perrine-Walker, F., and Laugier, E. (2011). Nitrate transceptor (s) in plants. *J. Exp. Bot.* 62, 2299–2308. doi: 10.1093/jxb/erq419

Good, A. G., Shrawat, A. K., and Muench, D. G. (2004). Can less yield more? Is reducing nutrient input into the environment compatible with maintaining crop production? *Trends Plant Sci.* 9, 597–605. doi: 10.1016/j.tplants.2004.10.008

Hajiahmadi, Z., Abedi, A., Wei, H., Sun, W., Ruan, H., Zhuge, Q., et al. (2020). Identification, evolution, expression, and docking studies of fatty acid desaturase genes in wheat (Triticum aestivum L.). *BMC Genomics* 21, 778. doi: 10.1186/s12864-020-07199-1

Hatje, K., Keller, O., Hammesfahr, B., Pillmann, H., Waack, S., and Kollmar, M. (2011). Cross-species protein sequence and gene structure prediction with fine-tuned Webscipio 2.0 and Scipio. *BMC Res. Notes* 4, 1–20. doi: 10.1186/1756-0500-4-265

Hawkesford, M. J. (2017). Genetic variation in traits for nitrogen use efficiency in wheat. *J. Exp. Bot.* 68, 2627–2632. doi: 10.1093/jxb/erx079

Hawkesford, M. J., and Griffiths, S. (2019). Exploiting genetic variation in nitrogen use efficiency for cereal crop improvement. *Curr. Opin. Plant Biol.* 49, 35–42. doi: 10.1016/j.pbi.2019.05.003

Hawkins, J., and Bodén, M. (2006). Detecting and sorting targeting peptides with neural networks and support vector machines. *J. Bioinf. Comput. Biol.* 4, 1–18. doi: 10.1142/S0219720006001771

Horton, P., Park, K.-J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C. J., et al. (2007). WoLF PSORT: protein localization predictor. *Nucleic Acids Res.* 35, W585–W587. doi: 10.1093/nar/gkm259

Howard, M. L., and Davidson, E. H. (2004). cis-Regulatory control circuits in development. *Dev. Biol.* 271, 109–118. doi: 10.1016/j.ydbio.2004.03.031

Hu, R., Qiu, D., Chen, Y., Miller, A. J., Fan, X., Pan, X., et al. (2016). Knock-down of a tonoplast localized low-affinity nitrate transporter osNPF7.2 affects rice growth under high nitrate supply. *Front. Plant Sci.* 7. doi: 10.3389/fpls.2016.01529

Hu, B., Wang, W., Ou, S., Tang, J., Li, H., Che, R., et al. (2015). Variation in NRT1. 1B contributes to nitrate-use divergence between rice subspecies. *Nat. Genet.* 47, 834–838. doi: 10.1038/ng.3337

Hu, R., Xiao, J., Gu, T., Yu, X., Zhang, Y., Chang, J., et al. (2018). Genome-wide identification and analysis of WD40 proteins in wheat (Triticum aestivum L.). *BMC Genomics* 19, 803. doi: 10.1186/s12864-018-5157-0

International Wheat Genome Sequencing Consortium (IWGSC), Mayer, K. F., Rogers, J., Doležel, J., Pozniak, C., Eversole, K., et al. (2014). A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. *Science* 345, 1251788. doi: 10.1126/science.1251788

Jackson, S., and Chen, Z. J. (2010). Genomic and expression plasticity of polyploidy. *Curr. Opin. Plant Biol.* 13, 153–159. doi: 10.1016/j.pbi.2009.11.004

Javed, T., Singhal, R., Shabbir, R., Kumar, P., Shah, A. N., Jinger, D., et al. (2022). Recent advances in agronomic and physio-molecular approaches for improving nitrogen use efficiency in crop plants. *Front. Plant Sci.* 13, 877544. doi: 10.3389/fpls.2022.877544

Jørgensen, M. E., Xu, D., Crocoll, C., Ernst, H. A., Ramírez, D., Motawia, M. S., et al. (2017). Origin and evolution of transporter substrate specificity within the NPF family. *eLife* 6, e19466. doi: 10.7554/eLife.19466

Kanstrup, C., and Nour-Eldin, H. H. (2022). The emerging role of the nitrate and peptide transporter family: NPF in plant specialized metabolism. *Curr. Opin. Plant Biol.* 68, 102243. doi: 10.1016/j.pbi.2022.102243

Kaur, A., Pati, P. K., Pati, A. M., and Nagpal, A. K. (2017). In-silico analysis of cis-acting regulatory elements of pathogenesis-related proteins of Arabidopsis thaliana and Oryza sativa. *PloS One* 12, e0184523. doi: 10.1371/journal.pone.0184523

Kimber, G., and Feldman, M. (1987). *Wild wheat. An introduction* (Columbia, Mo, U.S.A.: College of Agriculture University of Missouri). 353, Wild wheat. An introduction.

Kolde, R., and Kolde, M. R. (2015). *Package 'pheatmap'* (R Package) 1, 790.

Konishi, M., and Yanagisawa, S. (2010). Identification of a nitrate-responsive cis-element in the Arabidopsis NIR1 promoter defines the presence of multiple cis-regulatory elements for nitrogen response. *Plant J.* 63, 269–282. doi: 10.1111/j.1365-313X.2010.04239.x

Krogh, A., Larsson, B., Von Heijne, G., and Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* 305, 567–580. doi: 10.1006/jmbi.2000.4315

Kumar, A., Jayaswal, P. K., Mahato, A. K., Arya, A., Mandal, P. K., Singh, N. K., et al. (2023). Growth stage and nitrate limiting response of NRT2 and NAR2 gene families of bread wheat, and complementation and retrieval of nitrate uptake of atnrt2. 1 mutant by a wheat NRT2 gene. *Environ. Exp. Bot.* 207, 105205. doi: 10.1016/j.envexpbot.2022.105205

Kumar, A., Sandhu, N., Kumar, P., Pruthi, G., Singh, J., Kaur, S., et al. (2022). Genome-wide identification and in silico analysis of NPF, NRT2, CLC and SLAC1/SLAH nitrate transporters in hexaploid wheat (Triticum aestivum). *Sci. Rep.* 12, 11227. doi: 10.1038/s41598-022-15202-w

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinf.* 9, 559. doi: 10.1186/1471-2105-9-559

Lee, Y., and Szymanski, D. B. (2021). Multimerization variants as potential drivers of neofunctionalization. *Sci. Adv.* 7, eabf0984. doi: 10.1126/sciadv.abf0984

Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., and Storey, J. D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* 28, 882–883. doi: 10.1093/bioinformatics/bts034

Leinonen, R., Sugawara, H., Shumway, M., and Collaboration, I. S. D. (2010). The sequence read archive. *Nucleic Acids Res.* 39, D19–D21. doi: 10.1093/nar/gkq1019

Léran, S., Varala, K., Boyer, J.-C., Chiurazzi, M., Crawford, N., Daniel-Vedele, F., et al. (2014). A unified nomenclature of NITRATE TRANSPORTER 1/PEPTIDE TRANSPORTER family members in plants. *Trends Plant Sci.* 19, 5–9. doi: 10.1016/j.tplants.2013.08.008

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, W., He, X., Chen, Y., Jing, Y., Shen, C., Yang, J., et al. (2020). A wheat transcription factor positively sets seed vigour by regulating the grain nitrate signal. *New Phytol.* 225, 1667–1680. doi: 10.1111/nph.16234

Li, M., Tian, H., and Gao, Y. (2021). A genome-wide analysis of NPF and NRT2 transporter gene families in bread wheat provides new insights into the distribution, function, regulation and evolution of nitrate transporters. *Plant Soil* 465, 47–63. doi: 10.1007/s11104-021-04927-8

Li, W., Wang, Y., Okamoto, M., Crawford, N. M., Siddiqi, M. Y., and Glass, A. D. M. (2007). Dissection of the AtNRT2.1–AtNRT2.2 inducible high-affinity nitrate transporter gene cluster. *Plant Physiol.* 143, 425–433. doi: 10.1104/pp.106.091223

Liu, K.-H., Huang, C.-Y., and Tsay, Y.-F. (1999). CHL1 is a dual-affinity nitrate transporter of arabidopsis involved in multiple phases of nitrate uptake. *Plant Cell* 11, 865–874. doi: 10.1105/tpc.11.5.865

Liu, K. H., Niu, Y., Konishi, M., Wu, Y., Du, H., and Chung, H. S. (2017). Discovery of nitrate–CPK–NLP signalling in central nutrient–growth networks. *Nature* 545, 311–316. doi: 10.1038/nature22077

Longo, A., Miles, N. W., and Dickstein, R. (2018). Genome mining of plant NPFs reveals varying conservation of signature motifs associated with the mechanism of transport. *Front. Plant Sci.* 9, 1668. doi: 10.3389/fpls.2018.01668

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15, 550. doi: 10.1186/s13059-014-0550-8

Lupini, A., Preiti, G., Badagliacca, G., Abenavoli, M. R., Sunseri, F., Monti, M., et al. (2021). Nitrogen use efficiency in durum wheat under different nitrogen and water regimes in the mediterranean basin. *Front. Plant Sci.* 11, 607226. doi: 10.3389/fpls.2020.607226

Lynch, M., and Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science* 290, 1151–1155. doi: 10.1126/science.290.5494.1151

Maghiaoui, A., Bouguyon, E., Cuesta, C., Perrine-Walker, F., Alcon, C., Krouk, G., et al. (2020). The Arabidopsis NRT1.1 transceptor coordinately controls auxin biosynthesis and transport to regulate root branching in response to nitrate. (A Murphy, Ed.). *J. Exp. Bot.* 71, 4480–4494. doi: 10.1093/jxb/eraa242

Marchive, C., Roudier, F., Castaings, L., Brehaut, L., Blondet, E., and Colot, V. (2013). Nuclear retention of the transcription factor NLP7 orchestrates the early response to nitrate in plants. *Nat. Commun.* 4, 1713. doi: 10.1038/ncomms2650

Matsuoka, Y. (2011). Evolution of polyploid Triticum wheats under cultivation: the role of domestication, natural hybridization and allopolyploid speciation in their diversification. *Plant Cell Physiol.* 52, 750–764. doi: 10.1093/pcp/pcr018

Meier, M., Liu, Y., Lay-Pruitt, K. S., Takahashi, H., and Von Wirén, N. (2020). Auxin-mediated root branching is determined by the form of available nitrogen. *Nat. Plants* 6, 1136–1145. doi: 10.1038/s41477-020-00756-2

Miller, A. J., Fan, X., Shen, Q., and Smith, S. J. (2007). Expression and functional analysis of rice NRT2 nitrate transporters. *Comp. Biochem. Physiol. Part A: Mol. Integr. Physiol.* 146, S241. doi: 10.1016/j.cbpa.2007.01.618

Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A., and Punta, M. (2013). Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* 41, e121–e121. doi: 10.1093/nar/gkt263

Moll, R. H., Kamprath, E. J., and Jackson, W. A. (1982). Analysis and interpretation of factors which contribute to efficiency of nitrogen utilization1. *Agron. J.* 74, 562–564. doi: 10.2134/agronj1982.00021962007400030037x

Mounier, E., Pervent, M., Ljung, K., Gojon, A., and Nacry, P. (2014). Auxin-mediated nitrate signalling by NRT1.1 participates in the adaptive response of *Arabidopsis* root architecture to the spatial heterogeneity of nitrate availability: Nitrate signalling by NRT1.1. *Plant Cell Environ.* 37, 162–174. doi: 10.1111/pce.12143

Necsulea, A., and Kaessmann, H. (2014). Evolutionary dynamics of coding and non-coding transcriptomes. *Nat. Rev. Genet.* 15, 734–748. doi: 10.1038/nrg3802

Nguyen, L. T., Schmidt, H. A., Von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi: 10.1093/molbev/msu300

O'Brien, J. A., Vega, A., Bouguyon, E., Krouk, G., Gojon, A., Coruzzi, G., et al. (2016). Nitrate transport, sensing, and responses in plants. *Mol. Plant* 9, 837–856. doi: 10.1016/j.molp.2016.05.004

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* 14, 417–419. doi: 10.1038/nmeth.4197

Pellizzaro, A., Clochard, T., Planchet, E., Limami, A. M., and Mor`ere-Le Paven, M. (2015). Identification and molecular characterization of Medicago truncatula NRT2 and NAR2 families. *Physiologia Plantarum* 154, 256–269. doi: 10.1111/ppl.12314

Plett, D., Toubia, J., Garnett, T., Tester, M., Kaiser, B. N., and Baumann, U. (2010). Dichotomy in the NRT gene families of dicots and grass species. *PloS One* 5, e15289. doi: 10.1371/journal.pone.0015289

Puccio, G., Crucitti, A., Tiberini, A., Mauceri, A., Taglienti, A., Palumbo Piccionello, A., et al. (2022). WRKY gene family drives dormancy release in onion bulbs. *Cells* 11, 1100. doi: 10.3390/cells11071100

Rolly, N. K., and Yun, B. W. (2021). Regulation of nitrate (NO3) transporters and glutamate synthase-Encoding genes under drought stress in arabidopsis: the regulatory role of atbZIP62 transcription factor. *Plants* 10, 2149. doi: 10.3390/plants10102149

Salse, J., Bolot, S., Throude, M., Jouffe, V., Piegu, B., Quraish, U. M., et al (2008). Identifica-tion and characterization of shared duplications between rice and wheat provide new in-sight into grass genome evolution. *Plant Cell* 20, 11–24. doi: 10.1105/tpc.107.056309

Schilling, S., Kennedy, A., Pan, S., Jermiin, L. S., and Melzer, R. (2020). Genome-wide analysis of MIKC-type MADS-box genes in wheat: pervasive duplications, functional conservation and putative neofunctionalization. *New Phytol.* 225, 511–529. doi: 10.1111/nph.16122

Shannon, P. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* 7, 539. doi: 10.1038/msb.2011.75

Sol, S., Valkov, V. T., Rogato, A., Noguero, M., Gargiulo, L., Mele, G., et al. (2019). Disruption of the Lotus japonicus transporter LjNPF2.9 increases shoot biomass and nitrate content without affecting symbiotic performances. *BMC Plant Biol.* 19, 1–14. doi: 10.1186/s12870-019-1978-5

Tal, I., Zhang, Y., Jørgensen, M. E., Pisanty, O., Barbosa, I. C., Zourelidou, M., et al. (2016). The Arabidopsis NPF3 protein is a GA transporter. *Nat. Commun.* 7, 1–11. doi: 10.1038/ncomms11486

Tian, H., Yuan, X., Duan, J., Li, W., Zhai, B., Gao, Y., et al. (2017). Influence of nutrient signals and carbon allocation on the expression of phosphate and nitrogen transporter genes in winter wheat (Triticum aestivum L.) roots colonized by arbuscular mycorrhizal fungi. *PloS One* 12, e0172154. doi: 10.1371/journal.pone.0172154

Todd, C. D., Zeng, P., Huete, A. M. R., Hoyos, M. E., and Polacco, J. C. (2004). Transcripts of MYB-like genes respond to phosphorous and nitrogen deprivation in Arabidopsis. *Planta* 219, 1003–1009. doi: 10.1007/s00425-004-1305-7

Tong, J., Walk, T. C., Han, P., Chen, L., Shen, X., Li, Y., et al. (2020). Genome-wide identification and analysis of high-affinity nitrate transporter 2 (NRT2) family genes in rapeseed (Brassica napus L.) and their responses to various stresses. *BMC Plant Biol.* 20, 464. doi: 10.1186/s12870-020-02648-1

Wang, X. F., An, J. P., Liu, X., Su, L., You, C. X., and Hao, Y. J. (2018a). The nitrate-responsive protein MdBT2 regulates anthocyanin biosynthesis by interacting with the MdMYB1 transcription factor. *Plant Physiol.* 178, 890–906. doi: 10.1104/pp.18.00244

Wang, X., Cai, X., Xu, C., and Wang, Q. (2021). Identification and characterization of the NPF, NRT2 and NRT3 in spinach. *Plant Physiol. Biochem.* 158, 297–307. doi: 10.1016/j.plaphy.2020.11.017

Wang, R., Guan, P., Chen, M., Xing, X., Zhang, Y., and Crawford, N. M. (2010). Multiple regulatory elements in the Arabidopsis NIA1 promoter act synergistically to form a nitrate enhancer. *Plant Physiol.* 154, 423–432. doi: 10.1104/pp.110.162586

Wang, J., H"uner, N., and Tian, L. (2019a). Identification and molecular characterization of the Brachypodium distachyon NRT2 family, with a major role of BdNRT2.1. *Physiologia plantarum* 165, 498–510. doi: 10.1111/ppl.12716

Wang, W., Hu, B., Yuan, D., Liu, Y., Che, R., Hu, Y., et al. (2018b). Expression of the nitrate transporter gene OsNRT1.1A/OsNPF6.3 confers high yield and early maturation in rice. *Plant Cell* 30, 638–651. doi: 10.1105/tpc.17.00809

Wang, Z., Leushkin, E., Liechti, A., Ovchinnikova, S., Mößinger, K., Brüning, T., et al. (2020). Transcriptome and translatome co-evolution in mammals. *Nature* 588, 642–647. doi: 10.1038/s41586-020-2899-z

Wang, J., Li, Y., Zhu, F., Ming, R., and Chen, L.-Q. (2019b). Genome-wide analysis of nitrate transporter (NRT/NPF) family in sugarcane saccharum spontaneum L. *Trop. Plant Biol.* 12, 133–149. doi: 10.1007/s12042-019-09220-8

Wei, J., Zheng, Y., Feng, H., Qu, H., Fan, X., Yamaji, N., et al. (2018). OsNRT2. 4 encodes a dual-affinity nitrate transporter and functions in nitrate-regulated root growth and nitrate distribution in rice. *J. Exp. Bot.* 69, 1095–1107. doi: 10.1093/jxb/erx486

Wen, J., Li, P., Ran, F., Guo, P., Zhu, J., Yang, J., et al. (2020). Genome-wide characterization, expression analyses, and functional prediction of the NPF family in Brassica napus. *BMC Genomics* 21, 1–17. doi: 10.1186/s12864-020-07274-7

Wen, Z.-Y., Tyerman, S. D., Dechorgnat, J., Ovchinnikova, E., Dhugga, K. S., and Kaiser, B. N. (2017). Maize NPF6 proteins are homologs of Arabidopsis CHL1 that are selective for both nitrate and chloride. *Plant Cell* 29, 2581–2596. doi: 10.1105/tpc.16.00724

Winkelmüller, T. M., Entila, F., Anver, S., Piasecka, A., Song, B., Dahms, E., et al. (2021). Gene expression evolution in pattern-triggered immunity within Arabidopsis thaliana and across Brassicaceae species. *Plant Cell* 33, 1863–1887. doi: 10.1093/plcell/koab073

Xu, G., Fan, X., and Miller, A. J. (2012). Plant nitrogen assimilation and use efficiency. *Annu. Rev. Plant Biol.* 63, 153–182. doi: 10.1146/annurev-arplant-042811-105532

Xuan, W., Beeckman, T., and Xu, G. (2017). Plant nitrogen nutrition: sensing and signaling. *Curr. Opin. Plant Biol.* 39, 57–65. doi: 10.1016/j.pbi.2017.05.010

Yáñez-Cuna, J. O., Kvon, E. Z., and Stark, A. (2013). Deciphering the transcriptional cis-regulatory code. *Trends Genet.* 29, 11–22. doi: 10.1016/j.tig.2012.09.007

Zhang, Z., Li, Z., Wang, W., Jiang, Z., Guo, L., Wang, X., et al. (2021). Modulation of nitrate-induced phosphate response by the MYB transcription factor RLI1/HINGE1 in the nucleus. *Mol. Plant* 14, 517–529. doi: 10.1016/j.molp.2020.12.005

Zhu, T., Liu, Y., Ma, L., Wang, X., Zhang, D., Han, Y., et al. (2020). Genome-wide identification, phylogeny and expression analysis of the SPL gene family in wheat. *BMC Plant Biol.* 20, 420. doi: 10.1186/s12870-020-02576-0

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 16 November 2023
DOI 10.3389/fpls.2023.1271625

OPEN ACCESS

EDITED BY
Manohar Chakrabarti,
The University of Texas Rio Grande Valley, United States

REVIEWED BY
Mir Asif Iquebal,
Indian Council of Agricultural Research, India
Munevver Dogramaci,
Agricultural Research Service (USDA), United States

*CORRESPONDENCE
Jiming Jiang
jiangjm@msu.edu
C. Robin Buell
robin.buell@uga.edu

†These authors have contributed equally to this work

RECEIVED 02 August 2023
ACCEPTED 26 October 2023
PUBLISHED 16 November 2023

CITATION
Fang C, Hamilton JP, Vaillancourt B, Wang Y-W, Wood JC, Deans NC, Scroggs T, Carlton L, Mailloux K, Douches DS, Nadakuduti SS, Jiang J and Buell CR (2023) Cold stress induces differential gene expression of retained homeologs in *Camelina sativa* cv Suneson. *Front. Plant Sci.* 14:1271625. doi: 10.3389/fpls.2023.1271625

Cold stress induces differential gene expression of retained homeologs in *Camelina sativa* cv Suneson

Chao Fang[1†], John P. Hamilton[2,3†], Brieanne Vaillancourt[2], Yi-Wen Wang[2], Joshua C. Wood[2], Natalie C. Deans[2], Taylor Scroggs[4], Lemor Carlton[2], Kathrine Mailloux[2], David S. Douches[5], Satya Swathi Nadakuduti[6,7], Jiming Jiang[1,8*] and C. Robin Buell[2,3,9*]

[1]Department of Plant Biology, Michigan State University, East Lansing, MI, United States, [2]Center for Applied Genetic Technologies, University of Georgia, Athens, GA, United States, [3]Department of Crop & Soil Sciences, University of Georgia, Athens, GA, United States, [4]Department of Genetics, University of Georgia, Athens, GA, United States, [5]Department of Plant, Soil & Microbial Sciences, Michigan State University, East Lansing, MI, United States, [6]Department of Environmental Horticulture, University of Florida, Gainesville, FL, United States, [7]Plant Molecular and Cellular Biology Program, University of Florida, Gainesville, FL, United States, [8]Department of Horticulture, Michigan State University, East Lansing, MI, United States, [9]Institute of Plant Breeding, Genetics & Genomics, University of Georgia, Athens, GA, United States

Camelina sativa (L.) Crantz, a member of the Brassicaceae, has potential as a biofuel feedstock which is attributable to the production of fatty acids in its seeds, its fast growth cycle, and low input requirements. While a genome assembly is available for camelina, it was generated from short sequence reads and is thus highly fragmented in nature. Using long read sequences, we generated a chromosome-scale, highly contiguous genome assembly (644,491,969 bp) for the spring biotype cultivar 'Suneson' with an N50 contig length of 12,031,512 bp and a scaffold N50 length of 32,184,682 bp. Annotation of protein-coding genes revealed 91,877 genes that encode 133,355 gene models. We identified a total of 4,467 genes that were significantly up-regulated under cold stress which were enriched in gene ontology terms associated with "response to cold" and "response to abiotic stress". Coexpression analyses revealed multiple coexpression modules that were enriched in genes differentially expressed following cold stress that had putative functions involved in stress adaptation, specifically within the plastid. With access to a highly contiguous genome assembly, comparative analyses with *Arabidopsis thaliana* revealed 23,625 *A. thaliana* genes syntenic with 45,453 Suneson genes. Of these, 24,960 Suneson genes were syntenic to 8,320 *A. thaliana* genes reflecting a 3 camelina homeolog to 1 Arabidopsis gene relationship and retention of all three homeologs. Some of the retained triplicated homeologs showed conserved gene expression patterns under control and cold-stressed conditions whereas other triplicated homeologs

displayed diverged expression patterns revealing sub- and neo-functionalization of the homeologs at the transcription level. Access to the chromosome-scale assembly of Suneson will enable both basic and applied research efforts in the improvement of camelina as a sustainable biofuel feedstock.

KEYWORDS

camelina, cold stress, genome assembly, homeolog, lipid

1 Introduction

Camelina sativa (L.) Crantz, also known as false flax or gold-of-pleasure, is a low-cost renewable crop with multiple uses in food, feed, and bio-based applications. It has a broad environmental adaptability with a short life cycle of 85 to 100 days and can be grown in marginalized conditions with minimal agricultural inputs (Vollmann and Eynck, 2015; Malik et al., 2018; Zanetti et al., 2021). Camelina is a member of the Brassicaceae family and produces seeds with up to 40% oil by weight (Rodríguez-Rodríguez et al., 2013; Berti et al., 2016). If blended with conventional jet A fuel in equal proportions, camelina-based biofuel has been shown to reduce particle number and mass emissions by 50–70% (Moore et al., 2017). Key to its use as a sustainable biofuel is the development of camelina cultivars that are adapted to different climates and have favorable seed oil yield and fatty acid profiles. At the biochemical level, the deep knowledge of lipid metabolism in *Arabidopsis thaliana* (hereafter Arabidopsis) has been leveraged to camelina resulting in editing of fatty acid desaturase genes to alter the fatty acid profile in seed oil (Jiang et al., 2017; Morineau et al., 2017; Lee et al., 2021). However, in addition to serving as a storage molecule in seeds, fatty acids are integral components of membranes in which the composition of fatty acids (saturated vs unsaturated) impacts membrane fluidity.

To date, field studies on the impact of climate on camelina have shown that temperature, moisture, and soil type can impact seed yield and fatty acid profiles (Obour et al., 2017; Raziei et al., 2018). Furthermore, a controlled growth chamber experiment revealed that altered temperature resulted in significant changes in seed oil fatty acid profiles (Brock et al., 2020). In acclimation experiments in which spring and winter biotypes were exposed to low and then freezing temperatures, both physiological and gene expression changes were apparent between the biotypes reflecting differential responses to temperature (Anderson et al., 2022; Soorni et al., 2022). Gene expression differences were also observed between a spring and winter biotype following an 8-week cold acclimation period (Wang et al., 2022). In a limited study of genes involved in lipid metabolism, a cold stress treatment (4°C) induced expression of *CsPDAT1-A* and *CsPDAT1-C* that encode phospholipid: diacylglycerol acyltransferases which catalyze the final acylation step in triacylglycerol (TAG) biosynthesis (Yuan et al., 2017). Obtaining a better understanding of the impacts of climate on not only fatty acid and lipid profiles across organs but also other key agronomic traits is critical to developing camelina as a biofuel crop with resilience to climate variation.

The genome sequence of the doubled haploid DH55 *C. sativa* accession (641 Mb assembly) was published in 2014 and encodes ~89,000 genes (Kagale et al., 2014). Comparative genome analyses are consistent with a recent whole genome triplication in camelina that resulted in a highly undifferentiated hexaploid genome structure (Kagale et al., 2014). This is supported by recent chromosome painting, genome *in situ* hybridization, and phylogenetic analyses, which suggests that *C. sativa* is derived from an auto-allotetraploid *C. neglecta*-like species and the diploid species *C. hispida* (Mandáková et al., 2019). As a polyploid, genome fractionation has occurred in camelina along with subgenome dominance (Kagale et al., 2014). Conserved as well as sub- and neo-functionalization of gene expression among the homeologs has been reported (Kagale et al., 2014; Heydarian et al., 2018; Gomez-Cano et al., 2022). In addition to genomic and a suite of transcriptomic resources (Kagale et al., 2014; Abdullah et al., 2016; Abdullah et al., 2018; Gomez-Cano et al., 2020; Gomez-Cano et al., 2022), population genetics studies have been performed that associate agronomic traits with genomic loci (King et al., 2019; Luo et al., 2019; Chaudhary et al., 2020; Li et al., 2021).

While the DH55 reference genome has been highly useful to the community and a new version (v2) of the reference genome has been released (http://cruciferseq.ca), both genome assemblies are highly fragmented due to exclusive use of short read sequences in the assembly process. The use of long read sequencing platforms, coupled with significantly improved algorithms for genome assembly, permit the construction of a chromosome-scale, high quality camelina genome assembly that can enable an improved understanding gene function including regulation of homeologs. In this study, we generated a chromosome-scale, high quality reference genome sequence and annotation for the spring biotype cultivar 'Suneson' which has been widely used by the research community (Na et al., 2018; Ozseyhan et al., 2018; King et al., 2019; Na et al., 2019; Lhamo et al., 2020; Gomez-Cano et al., 2022; Bengtsson et al., 2023). To further our understanding of the impact of gene regulation, we examined gene expression in camelina leaves exposed to cold stress revealing conserved as well as differential gene expression among retained homeologs. The Suneson genome resource will be of value to researchers interested in engineering camelina as a biofuel crop and in understanding genome evolution in polyploids.

2 Materials and methods

2.1 Plant material

For generation of high molecular weight DNA, *Camelina sativa* cv Suneson seeds were subjected to one round of single seed descent, planted in soil, and grown at 25.5°C day/18°C night for 19 days under a 15 hr photoperiod with 500 $\mu E \cdot m^{-2} \cdot s^{-1}$ of light. Plants were stored in the dark for 24 hours, leaves were harvested, and flash frozen in liquid nitrogen. For the Illumina whole genome shotgun (WGS) library, immature leaves were harvested from plants grown for 22 days in soil at 22°C day/18°C night under 400 $\mu E \cdot m^{-2} \cdot s^{-1}$ of light with a 15 hr photoperiod, and then dark treated for 24 hours prior to DNA isolation. For generation of transcript data to support genome annotation, we harvested mature leaf, immature seed stage 1, immature seed stage 2, stem, open flower, and root without the dark treatment and flash froze tissues in liquid nitrogen prior to RNA isolation.

To examine the response of Suneson to cold treatment, bulk seeds were sterilized three times with 75% ethyl alcohol for 5 minutes and then plated on Murashige and Skoog plates for 4 days prior to transfer to soil. Potted seedlings were grown in a growth chamber at 22°C day/18°C night under a 16 hr photoperiod. After three weeks, plants were exposed to cold temperature (10°C/day, 6°C/night) for 48 hrs; plants not exposed to cold stress were used as a control.

2.2 Nucleic acid isolation, library construction, and sequencing

High molecular weight genomic DNA was isolated from dark-treated immature leaves using the Takara Bio Nucleobond HMW DNA Kit (Takara Bio USA, San Jose CA); short fragments were eliminated using the Short Read Eliminator kit (Pacific Biosciences, Menlo Park, CA). Oxford Nanopore Technologies (ONT) libraries were constructed using the Ligation Sequencing Kit (SQK-LSK114, Q20+ chemistry) and sequenced on FLO-MIN114 flow cells as described previously (Li et al., 2023b) (Supplementary Table 1). Bases were called using Guppy v6.3.7 in super high accuracy mode (https://nanoporetech.com/community). DNA for error correction was isolated from dark-treated immature leaves using the Qiagen genomic tip method (Vaillancourt and Buell, 2019) and WGS libraries were constructed using the PerkinElmer NEXTFLEX Rapid XP DNA-Seq Kit HT (Perkin Elmer, Waltham, MA) (Supplementary Table 2). Libraries (five in total) were sequenced on an Illumina NovaSeq 6000 in paired-end mode generating 150 nt reads.

2.3 Genome assembly and chromosome scaffolding

Jellyfish (v2.2.10) (Marçais and Kingsford, 2011) was used to count k-mers (k = 21) in the WGS reads which were analyzed with GenomeScope (v2.0) (Vurture et al., 2017) to determine the extent of heterozygosity and shared k-mers among the homeologs. In addition, a smudgeplot was generated using Smudgeplot (v0.2.5) (Ranallo-Benavidez et al., 2020) with k-mers (k = 21) counted with KMC (v3.1.1) (Kokot et al., 2017). ONT gDNA reads were filtered using seqtk (v1.3) (https://github.com/lh3/seqtk) to remove reads less than 15 kb. The genome was assembled using Flye (v2.9.1) (Kolmogorov et al., 2019) with the genome size set to .785g, zero polishing iterations, and asm-coverage 60. The initial assembly was error-corrected through two rounds of Medaka (v1.7.2; https://github.com/nanoporetech/medaka) with all of the ONT genomic DNA reads using the model r1041_e82_400bps_sup_g615. This was followed by two rounds of Pilon (v1.24) (Walker et al., 2014) using the alignments from Cudadapt-trimmed reads (v4.1) (Martin, 2011) that were aligned to the assembly using bwa-mem2 (v2.2.1) (Li, 2013). Contigs less than 50 kb were filtered out using seqkit (v2.3.0) (Shen et al., 2016). Two rounds of RagTag (v2.1.0) (Alonge et al., 2022) was used to generate a chromosome-scale assembly using the reference genome DH55 v2.0 (http://cruciferseq.ca). Kraken 2 (v2.1.2) (Wood et al., 2019) was used to check all reads and the final assembly for contamination. KAT (v 2.4.1) (Mapleson et al., 2017) was used to determine the representation of k-mers in the final assembly. Genome completeness was assessed using Benchmarking Universal Single Copy Orthologs (BUSCO, v5.4.3) (Waterhouse et al., 2018) with the embryophyta_odb10 database.

2.4 Preparation of RNA-seq and full-length cDNA libraries

To support high quality gene annotation, RNA was isolated from a diverse set of tissues using either the hot borate method (Wan and Wilkins, 1994) (mature leaf and stem) or Purelink RNA isolation kit (Thermo Fisher Scientific, Waltham MA) (immature seed stage 1, immature seed stage 2, open flower, and root). Total RNA was treated with Turbo DNAse (Thermo Fisher Scientific, Waltham MA) following the manufacturer's directions. ONT cDNA libraries were constructed using the SQK-PCB109 library preparation kit (Oxford Nanopore Technologies, Oxford UK) and sequenced on FLO-MIN106 flow cells (Supplementary Table 1). Bases were called using Guppy v6.3.7 (https://nanoporetech.com/community) in the super high accuracy mode with barcode trimming disabled.

For cold stress experiments, two biological replicates of leaf tissue were collected from control and cold-treated plants and ground into a fine powder in liquid nitrogen. Total RNA was extracted using the RNeasy Plant Mini Kit (Qiagen, Germantown MD) and RNA-seq libraries were constructed using the KAPA mRNA HyperPrep Kit protocol (KAPA Biosystems, Wilmington, MA). RNA-seq libraries were sequenced in paired-end mode generating 150 nt reads on an Illumina NovaSeq 6000 (Supplementary Table 1). ONT cDNA libraries were constructed, sequenced, and bases called as described above.

2.5 Genome annotation

The Suneson genome was annotated for protein-coding genes as described previously (Pham et al., 2020). In brief, repetitive

sequences were identified using RepeatModeler (v2.0.3) (Flynn et al., 2020) from which protein-coding genes were removed using Protex (v1.2)(Campbell et al., 2014). These filtered repeat sequences were added to the Repbase Viriplantae repeat dataset (v20150807) to construct a final repeat library. Prior to annotation, RepeatMasker (v4.1.2-p1) (Chen, 2004) was used to mask the genome using the parameters -s -nolow -no_is -gff. RNA-seq reads were cleaned of low quality sequences and adapters using Cutadapt (v2.10) (Martin, 2011) with a quality cutoff of 10 and a minimum length of 100nt (Supplementary Table 2). Cleaned reads were aligned to the Suneson genome using HISAT2 (v2.1.0) (Kim et al., 2019) with a maximum intron length of 5000 and genome-guided transcript assemblies were generated using Stringtie 2 (v2.2.1) (Kovaka et al., 2019). The BRAKER2 pipeline (v2.1.6) (Hoff et al., 2019) was used to predict gene models using the RNA-Seq alignments as hints. Gene models were refined through two rounds of PASA (v2.5.2) (Haas et al., 2003; Campbell et al., 2006) using the RNA-seq and ONT cDNA reads resulting in a set of 145,971 working gene models. To identify high confidence gene models, gene expression data were generated using Kallisto (v0.46.2) (Bray et al., 2016) with the mRNAseq reads and Stringtie (v2.2.1) (Kovaka et al., 2019) with the ONT cDNA reads. Predicted proteins were searched against the Arabidopsis v11 predicted proteome (Araport.org) using Diamond (v0.9.36) (Buchfink et al., 2015) and Pfam domains were identified using the PFAM database (v32.0) (El-Gebali et al., 2019) with HMMER (v3.3) (Mistry et al., 2013). High confidence gene models were determined based on gene expression (TPM > 0) and/or protein match to Arabidopsis and/or presence of a Pfam domain. Functional annotation was assigned to the gene models using matches to Arabidopsis, the presence of Pfam domains, and expression evidence. Transcription factors were predicted using iTAK v1.7 (Zheng et al., 2016) with the high-confidence representative peptide sequences.

2.6 Gene expression abundances, differential gene expression, and gene coexpression analyses

Gene expression abundance estimations were calculated for cold-stressed and control leaves (this study) along with publicly available data that was downloaded from the National Center for Biotechnology Information Sequence Read Archive. First, reads were cleaned using Cutadapt (v4.1) (Martin, 2011) with a minimum read length of 40, 3' end quality cutoff of 30, flanking N base removal, and 3' adapter sequence trimming. The Kallisto quant algorithm (v0.48.0) (Bray et al., 2016) was used to quantify expression with a k-mer size of 21; libraries that were sequenced in single end mode were run with two additional parameters, a fragment length of 200 and standard deviation of 20. Libraries that were sequenced in paired end mode were run with the –rf-stranded parameter. Gene coexpression networks were constructed using Simple Tidy GeneCoEx in R (Li et al., 2023a) with genes that had a TPM > 1.

To detect differential gene expression, RNA-seq reads were mapped to Suneson genome using HISAT2 (version 2.0.0-beta) (Kim et al., 2019) and expression abundances were calculated by StringTie (v1.3.3b) (Pertea et al., 2016) to determine gene expression abundances. Significantly differentially expressed homeologous genes among a triplet ($p < 0.01$) or between cold and control samples ($|\log2FC| > 1$; $p < 0.01$) were identified using EdgeR (Robinson et al., 2010).

2.7 Homeologous gene identification

Genome annotation for *Arabidopsis lyrata* and *A. thaliana* (Araport11) was downloaded from Phytozome (v13) (Goodstein et al., 2012). The GENESPACE pipeline (Lovell et al., 2022) was run with the Suneson genome and *A. thaliana* using GENESPACE v0.9.3 to identify triplicated homeologs within Suneson relative to *A. thaliana*. For *A. lyrata*, GENESPACE v1.1.4 was used with the representative gene model annotations to identify syntelogs between Suneson and *A. lyrata*. The default pipeline options were used except for the ploidy option which was set to 1,3. The syntelogs were exported from the pan-genome databases using the query_pangenes GENESPACE function.

2.8 Gene expression of triplicated homeologous genes

We first classified variation in expression across triplicated homeologs under control conditions by ranking the three homeologous genes based on their average FPKM value. If the highest expressed gene in a triplet of homeologs showed a significantly higher expression level ($p < 0.01$ and fold change of FPKM > 2) than the other two copies, this homeolog was classified as a Class 1 homeolog. If two genes in the triplet showed a significantly higher expression level ($p < 0.01$ and fold change of FPKM > 2) than the third copy, this homeolog was classified as a Class 2 homeolog. If all of the copies of a triplet showed similar expression levels (fold change of FPKM between every two copies < 1.5), this homeolog was classified as a Class 3 homeolog. Gene ontology analyses of the three classes of homeologs were performed and displayed using TBtools (Chen et al., 2020).

In the response to cold stress, if a triplicated homeolog had a significantly higher expression level following cold stress relative to the control sample, this gene was termed a cold-induced gene. A cold-induced homeolog was classified as a Type 1 triplet if all three homeologs were cold-induced; a homeolog was classified as a Type 2 homeolog if two of the three copies were cold-induced; and a homeolog was classified as a Type 3 homeolog if one of the three copies was cold-induced. For every homeolog in a Type 1 triplet, we calculated the fold change of its FPKM value between cold-treated and control samples and used the fold change to represent the cold response level of this homeolog. We ranked the three homeologous genes based on their cold inducibility with the copy exhibiting the highest level ranked first and the copy with the lowest level ranked

third. The cold inducibility of the first and third copies were compared to detect the divergence of their response to cold stress.

3 Results and discussion

3.1 Genome assembly of *C. sativa* cv Suneson

As the camelina genome is a hexaploid, we assessed the number of unique k-mers in the Suneson genome using Illumina WGS reads. The k-mer distribution plot (Supplementary Figure 1) is consistent with a diploidized genome in which the majority of k-mers (k = 21) were present in single copy with a subset present in two copies and an even smaller subset in three copies. We also examined the pattern of near-identical k-mers using SmudgePlot (Supplementary Figure 2) in which 49% of the k-mer pairs were present as AAB, 46% present as AB, and 5% as AAAB, consistent with the hypothesized origin of hexaploid camelina being derived from an auto-allotetraploid *C. neglecta*-like species and the diploid species *C. hispida* (Mandáková et al., 2019). Using ~42× coverage ONT genomic reads greater than 15 kb and the Flye assembler software, we assembled 647,473,868 bp of the Suneson genome into 551 contigs with an N50 contig length of 12,024,690 bp (Supplementary Table 3). Two rounds of error correction with Medaka followed by two rounds of Pilon were performed. The assembly was filtered to remove contigs less than 50 kb, yielding a 644,482,469 bp assembly contained in 157 contigs with an N50 length of 12,031,512 bp (Supplementary Table 3). To assemble to the 20 camelina chromosomes, Ragtag was used with the DH55 reference assembly resulting in 98.3% of the Suneson assembly anchored to the chromosomes (Supplementary Table 4). To validate the assembly, we used the KAT program to determine the representation of WGS-derived k-mers in the final assembly. As shown in Supplementary Figure 3, the majority of k-mers were present in single copy within the assembly with limited numbers present at two copies and a small set of k-mers present in three copies. To assess the representation of genic sequences in the genome assembly, we ran BUSCO with the embryophyta_odb10 database. A total of 1,606 of the 1,614 (99.5%) BUSCO orthologs were complete in the Suneson assembly, of which, 1,581 (98.0%) are duplicated as expected due to the hexaploid nature of the camelina genome; a mere 0.2% and 0.3% were fragmented or missing, respectively (Supplementary Table 5).

3.2 Genome annotation

Repetitive sequences in the Suneson genome were identified using a combination of *de novo* repeat identification and sequence similarity to existing Viridiplantae repetitive sequences. In total, 44.2% of the genome was annotated as repetitive (Supplementary Table 6), which is substantially higher than the percentage (25%) identified in the DH55 assembly which is attributable to the short-read-derived DH55 genome sequence. Retroelements (25%) dominated the annotated repetitive sequences relative to DNA transposons (2.85%). Annotation of protein-coding genes using five mRNA-seq libraries and full-length cDNA sequences derived from eight different tissues (Supplementary Table 7) resulted in 145,971 working gene models from 103,435 loci (Supplementary Table 8). Of these working models, 133,355 were high confidence models derived from 91,877 loci. While the number of total genes was similar between the Suneson and DH55 v2 annotation, access to substantial full-length cDNA sequence data permitted annotation of more gene models in the Suneson assembly compared to the DH55 assembly. In the Suneson annotation, the average number of gene models per locus in the working and high confidence gene sets to 1.41 and 1.45, respectively, which is higher compared to the average of 1.06 gene models per locus in the DH55 annotation (Supplementary Table 8). To assess the quality of the genome annotation, we examined the representation of BUSCO orthologs. Within the representative high confidence gene model set, 98.4% of the BUSCO orthologs were present with 92.8% present as duplicated, consistent with the hexaploid nature of the camelina genome (Supplementary Table 5).

Kagale et al. (2014) reported a high degree of synteny between *C. sativa* and *A. lyrata*. In the Suneson genome, 59,645 genes were syntenic to 20,934 *A. lyrata* genes as shown in the riparian plot of synteny between these two species (Figure 1A). To understand the relationship of the subgenomes within Suneson, we identified syntelogs between *A. thaliana* and Suneson using GENESPACE resulting in 23,625 *A. thaliana* genes syntenic with 45,453 Suneson genes (Supplementary Table 9). The riparian plot of the three subgenomes relative to the five chromosomes of *A. thaliana* (Figure 1B) highlights the significant degree of conservation between *A. thaliana* and the three subgenomes in Suneson. Of these, 24,960 Suneson genes were syntenic to 8,320 *A. thaliana* genes reflecting a 3 camelina homeolog to 1 A. *thaliana* relationship and retention of all three homeologs (Supplementary Table 10). In addition to fully retained homeologs, 2,829 *A. thaliana* genes were syntenic to 5,658 *C. sativa* genes (1:2 ratio) while 7,253 *A. thaliana* genes were syntenic to 7,253 *C. sativa* genes (1:1 ratio).

3.3 Response of camelina leaves to cold stress

We performed RNA-seq using leaf tissue collected from cold-treated plants to gain insight on the impact of short-term cold stress on leaf tissue. A total of 4,467 genes showed a significantly higher expression level in cold-treated samples compared to their expression under control temperature; 4,851 were down-regulated under cold stress (Supplementary Tables 11, 12). Gene ontology terms related to stress response were enriched in cold-inducible genes including "response to cold", "response to temperature stimulus", and "cold acclimation" (Figure 2A; Supplementary Table 13).

As temperature impacts fatty acid and lipid composition, we examined the expression of genes involved in lipid and fatty acid metabolism under cold stress in Suneson leaves. Using 552 A. *thaliana* genes previously associated with lipid and fatty acid

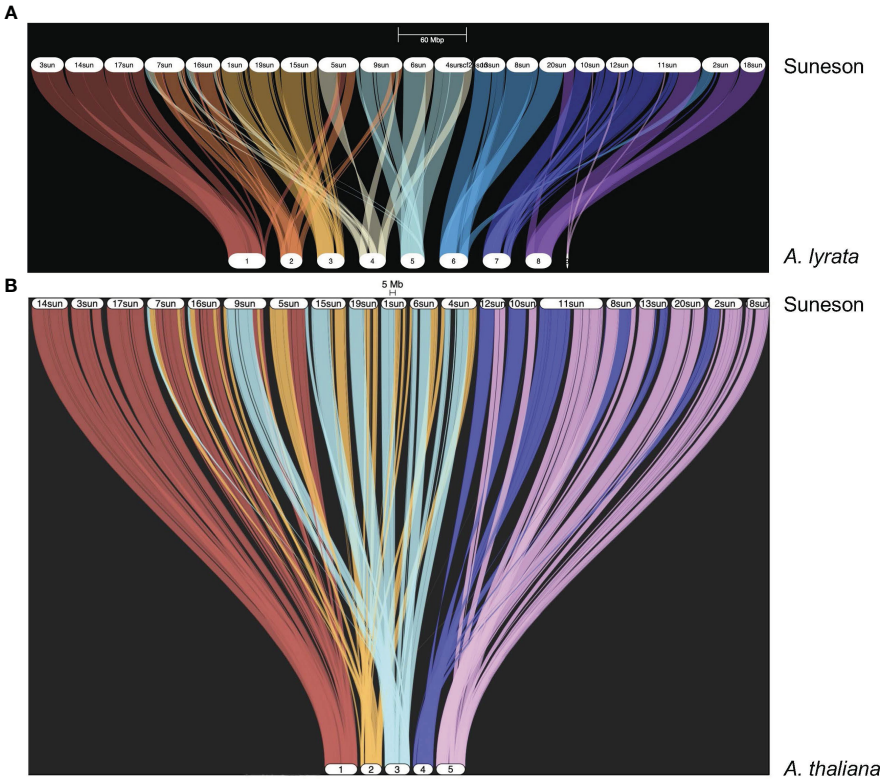


FIGURE 1
Syntelogs between *Arabidopsis* species and *Camelina sativa* cv Suneson. GENESPACE was used to identify syntelogs **(A)** between *Arabidopsis lyrata* and *Camelina sativa* cv Suneson and **(B)** between *Arabidopsis thaliana* and *Camelina sativa* cv Suneson.

metabolism (Nguyen et al., 2013), we identified 1,474 genes in the Suneson assembly involved in lipid and fatty acid metabolism (Supplementary Table 14). Expression of these genes were strikingly different between leaf and developing seeds as shown in Figure 2B consistent with the diverged function of lipid metabolism in these tissues. Of these, 60 were differentially up-regulated in cold stressed leaves while 111 were down-regulated. Multiple genes involved in remodeling membrane lipids were upregulated (Supplementary Table 11) including lipid transfer proteins functioning in phospholipid transfer between cell membranes. Also up-regulated were genes encoding phosphoinositide-specific phospholipase C which is associated with hormone signaling, abiotic stresses, and pathogen responses (Rupwate and Rajasekharan, 2012). Notably, a *DIACYLGLYCEROL KINASE (DGK)* gene was upregulated when exposed to cold temperatures (Supplementary Table 11). As plants balance the levels of phosphatidic acid, diacylglycerol, and triacylglycerol during cold stress, DGKs plays a major role in remodeling cold-responsive lipids (Tan et al., 2018). We did not observe up-regulation of phospholipid:diacylglycerol acyltransferases which have been reported to be induced by cold stress (Yuan et al., 2017) and shown to enhance fitness under cold stress in *A. thaliana* (Demski et al., 2020). This may be attributable to the warmer and shorter cold stress conditions employed in this study which may not have been sufficient to induce gene expression. In contrast, different classes of lipoxygenases (LOX gene family) involved in lipid catabolism and the formation of oxylipins including the defense-related hormone jasmonic acid were downregulated under cold stress (Supplementary Table 12) consistent with a previous study (Zhu et al., 2018). Oxylipins have been reported to play a role in cold stress through jasmonic acid-mediated regulation of Inducer of CBF (ICE) – C-Repeat Binding Factor (CBF)/DRE Binding Factor 1 through the alleviation of oxidative damage in cells (Hu et al., 2013).

To identify genes that are co-regulated under cold stress, we constructed gene coexpression networks. After filtering for the top 20% variable genes and an $r > 0.7$, 14,765 genes with 13,803,988 edges were used to construct coexpression modules (Figure 2C; Supplementary Table 15). With respect to gene expression following cold stress, four modules were of interest (Modules 5, 23, 124, 167). Module 5 contained 233 genes, of which, 62 were up-regulated and 22 were down-regulated following cold stress. Genes in Module 5 were enriched in GO terms associated with response to light, abiotic stress, and environmental stimuli as well as regulation of genes involved in photosynthesis (Supplementary Table 16). With respect to cellular compartment, Module 5 genes were associated with the plastid and the peroxisome. Module 23 (93 genes) had 35 and 4 genes up-regulated and down-regulated, respectively, in response to cold stress; GO terms associated with Module 23 were associated with response to abiotic stress, cold, and temperature stimulus and were associated with plastic cellular compartment (Supplementary Table 16). Similar to Module 5,

10.3389/fpls.2023.1271625

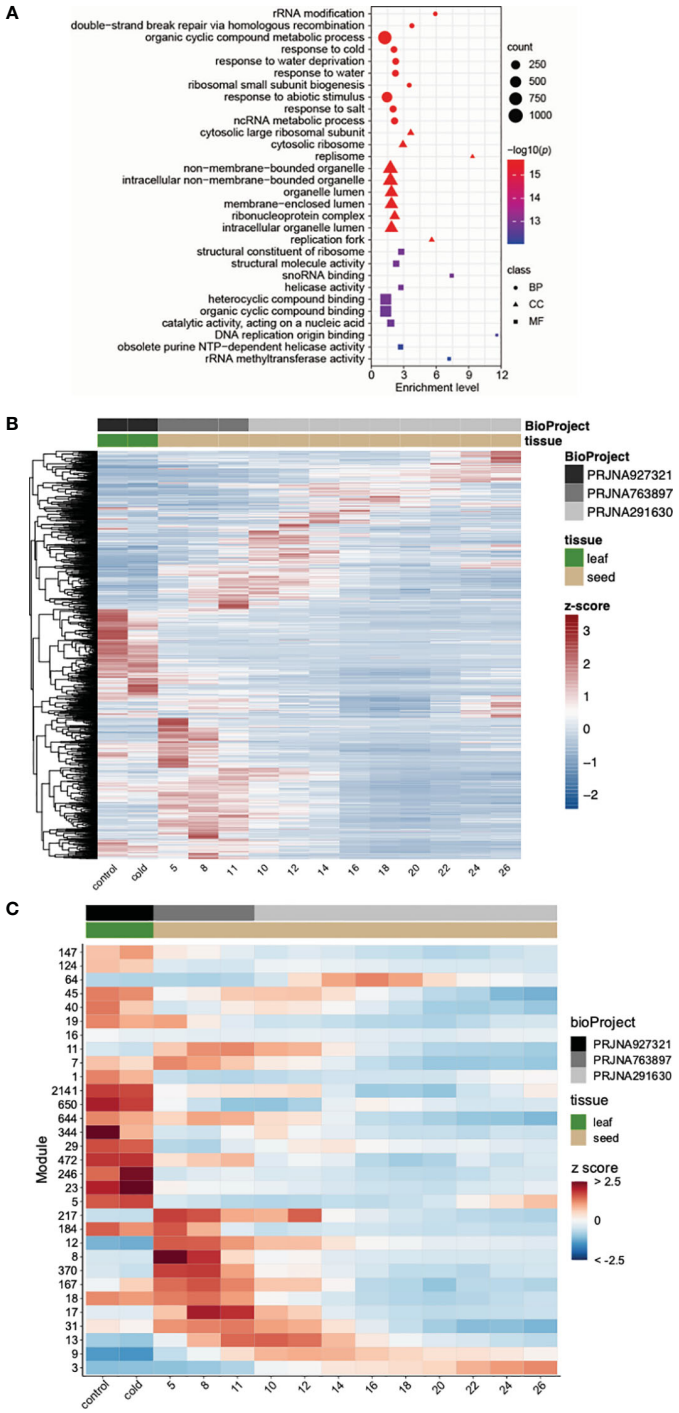


FIGURE 2
Gene expression in *Camelina sativa* cv Suneson following cold stress. **(A)** Gene ontology (GO) terms enriched in cold induced genes. Only the top 10 enriched GO terms are shown here. A complete list of enriched GO terms is included in Supplementary Table 13. Sizes of symbols reflect numbers of genes, color reflects $\log_{10}$ *p*-value and symbols reflect GO categories (BP: Biological process; CC: Cellular compartment; MF: Molecular function). **(B)** Gene expression in *C. sativa* cv Suneson leaf and developing seeds. Gene expression abundances for 1,452 genes involved in fatty acid and lipid metabolism were calculated using Kallisto for control and cold stressed leaves (this study) and two sets of seed development series obtained from the National Center for Biotechnology Information Sequence Read Archive. Numbers for seed samples reflect days post anthesis. **(C)** Modules of coexpressed genes using gene expression abundances from control and cold-stressed leaves (this study) and two seed development studies. Numbers for seed samples reflect days post anthesis.

Module 124 included 99 genes, of which, 24 were up-regulated and 29 were down-regulated and associated with GO terms involved in response to light and abiotic stimuli with localization within the plastid compartment (Supplementary Table 16). While only containing 38 genes, Module 167 had 23 genes up-regulated and a single gene down-regulated following cold-stress; GO associations suggest this module was associated with DNA repair (Supplementary Table 16).

3.4 Expression patterns of retained homeologous genes in *Camelina sativa*

We analyzed the transcriptional divergence of triplicated homeologous genes in the Suneson genome. We identified 8,320 sets of triplicated homeologous genes (see Methods) and analyzed their expression in leaf tissue. We first investigated the expression patterns of these homeologs under control conditions. Among these homeologs, over 76% (6,323/8,320) exhibited expression of all three homeologous genes, while 6.0% (499/8,320) and 6.4% (529/8,320) of the triplets displayed expression of only one and two of the three homeologous copies, respectively. All three copies of the remaining 969 (11.6%) triplets were not expressed in control leaf tissue.

The triplicated homeologs were then cataloged into Class 1, 2, and 3 based on the expression levels of the three homeologs. We identified 587 triplets (Class 1) in which the expression level of one copy is significantly higher ($p < 0.01$) than both of the two other copies (Figure 3A). We performed GO analysis on this group of genes. Interestingly, genes responsive to stimuli such as chemicals, stress, and endogenous stimuli were highly enriched in this group (Figure 3B; Supplementary Table 17). These data suggest neofunctionalization at the expression level among these homeologs. A similar result was reported in *A. thaliana* in which one copy of duplicated genes tends to retain their ancestral stress responses following gene duplication (Zou et al., 2009). We identified 1,044 triplicated homeologs (Class 3) in which the three copies showed a similar level of expression (Figure 3A); specifically, the fold change in expression between any two of the three copies was less than 1.5. Gene ontology analysis of this group of genes revealed enrichment of genes related to fundamental processes including biosynthetic processes, metabolic processes, cellular processes, developmental processes, and rhythmic processes (Figure 3C; Supplementary Table 18). This result indicates that it is favorable to retain expression of all three copies of genes related to fundamental biological processes. For comparison, we identified a set of 502 triplicated homeologs (Class 2) in which the expression levels of two copies were significantly higher ($p < 0.01$) than the third copy (Figure 3A); GO analysis revealed that signaling

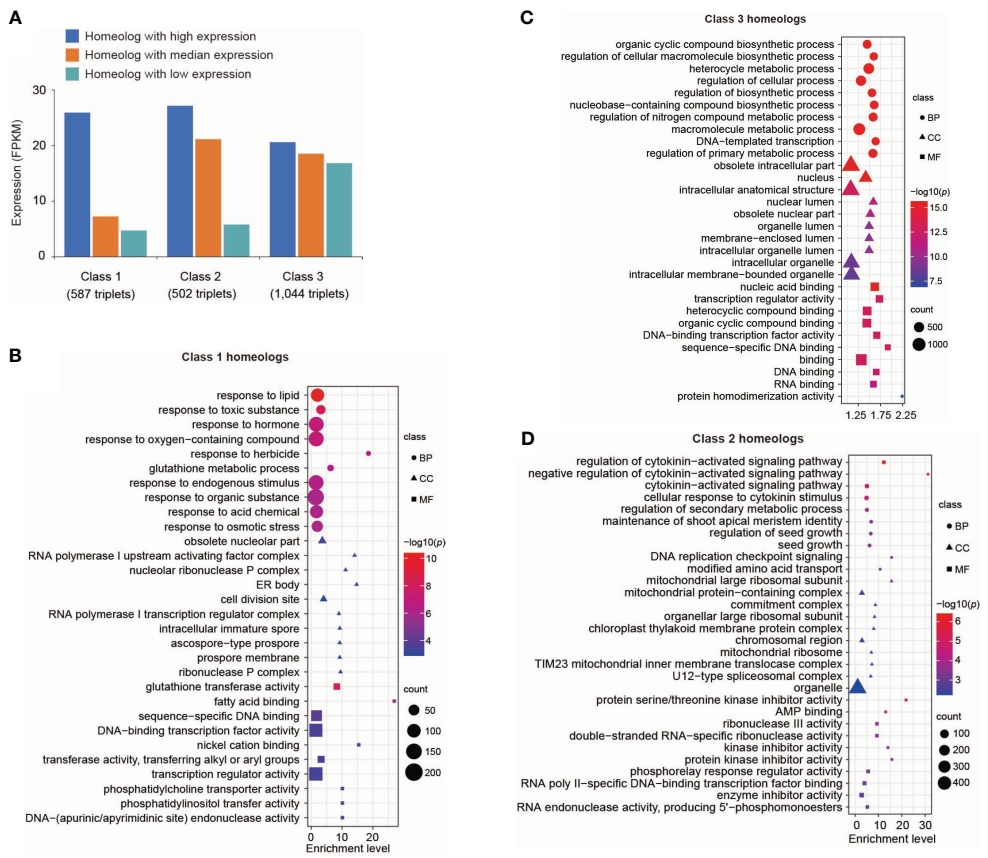


FIGURE 3
Transcriptional divergence of triplicated homeologous genes in control leaves. **(A)** Classification of three classes of triplicated homeologous genes based on gene expression. Class 1 represents the homeolog in which the highest expressed gene showed a significantly higher expression level ($p < 0.01$ and fold change of FPKM > 2) than the other two copies. Class 2 represents the homeolog in which two genes showed a significantly higher expression level ($p < 0.01$ and fold change of FPKM > 2) than the third copy. Class 3 represents the homeologs in which three copies showed similar expression levels (fold change of FPKM between every two copies < 1.5). **(B)** Gene ontology (GO) terms enriched in Class 1 genes. Only the top 10 enriched GO terms are shown. A complete list of enriched GO terms is included in Supplementary Table 17. **(C)** GO terms enriched in Class 3 genes. Only the top 10 enriched GO terms are shown here. A complete list of enriched GO terms is included in Supplementary Table 18. **(D)** GO terms enriched in Class 2 genes. Only the top 10 enriched GO terms are shown here. A complete list of enriched GO terms is included in Supplementary Table 19. Sizes of symbols reflect numbers of genes, color reflects $\log_{10}$ p-value and symbols reflect GO categories (BP, Biological process; CC, Cellular compartment; MF, Molecular function).

pathways were highly enriched in this class of genes (Figure 3D; Supplementary Table 19).

In allopolyploids, genes from one subgenome were often preferentially retained or achieved a higher level of expression than those from other subgenomes, which is known as subgenome dominance and has been documented in an increasing number of plant species (Thomas et al., 2006; Schnable et al., 2011; Alger and Edger, 2020). If one of the parental progenitors of an allopolyploid is highly adapted to the environment where the polyploid species originated, then genes responsible for environmental adaptation from this progenitor may be preferentially retained. For example, disease resistance genes were found to be preferentially retained and associated with subgenome dominance in strawberry (Barbey et al., 2019; Edger et al., 2019) and *Brassica napus* (de Jong and Adams, 2023). As noted above, in *A. thaliana* one copy of duplicated genes tends to retain their ancestral stress responses following gene duplication (Zou et al., 2009). Here, we show in *C. sativa* that a specific homeolog of genes responsive to stimuli tends to gain dominance in transcription in comparison to other homeologs. These results suggest that stress responsive genes have a distinct evolutionary trajectory in the evolution of allopolyploid species.

3.5 Diverged responses to cold stress among retained homeologous genes

We investigated how the expression of retained triplicated homeologous genes evolved in their response to an environmental cue, that of cold stress. Of the 4,467 genes up-regulated in response to cold stress, 36.5% (1,632/4,467) belong to 935 triplicated homeologs. We classified the 935 homeologs into three types. Type 1: all three homeologous copies were cold-inducible; Type 2: two of the three copies were cold-inducible; Type 3: only one copy was cold inducible. Our analysis revealed that 75% of the triplicated homeologous genes displayed diverged cold responses as at least one homoeologous copy was not induced after cold treatment (Type 2 or 3) (Figure 4A). Examples of Type 1 triplicated homeologs with retained expression are Camsa.SUN.04G044570, Camsa.SUN.04G044580, Camsa.SUN.05G006530, Camsa.SUN.05G006520, Camsa.SUN.06G040820, and Camsa.SUN.06G040830 which are syntelogs with the *A. thaliana* cold-regulated (*COR*) genes AT2G42530 (*COR15b*) and AT2G42540 (*COR15a*) present in tandem on *A. thaliana* chromosome 2 (Figure 4B). Arabidopsis COR15a and COR15b are small chloroplast-targeted polypeptides induced under cold stress, localized in the chloroplast stroma which function in freezing tolerance (Lin and Thomashow, 1992a; Lin and Thomashow, 1992b; Wilhelm and Thomashow, 1993; Artus et al., 1996; Thomashow, 1998; Thalhammer and Hincha, 2013). While all six Suneson genes are up-regulated in response to cold stress, the triplicated homeologs differ in basal gene expression levels and in the extent of up-regulation (Figure 4C). Syntelogs of AT2G42530 had lower basal expression but higher log2 fold-change relative to the AT2G42540 syntelogs which had a higher basal expression but lower log2 fold-change (Figure 4C). In addition, the extent of cold-induction within each set of the triplicated homeologs differed. For example, the log2 fold-change of Camsa.SUN.04G044570 is lower than Camsa.SUN.05G006530 and Camsa.SUN.06G040820 (Figure 4C). Similar cold-specific expression of the *Wcor15* homeolog has been documented in allopolyploid wheat and suggested to play an important role in cold hardiness in wheat and barley (Takumi et al., 2003).

To further compare the "cold inducibility" of each gene within all of the 233 Type 1 triplicated homeologs, we calculated the fold change in expression levels between control and cold treatment sample. We then ranked the three homeologous copies of each triplicated homeolog based on their cold inducibility, with the copy exhibiting the highest expression level ranked first and the copy with the lowest level ranked third (Figure 4A). Our analysis revealed that >10% of the Type 1 homeologs exhibited a two-fold or greater difference in cold inducibility between the first and third-ranked copies, suggesting divergence in the degree of cold inducibility among the homeologous genes. Such homeolog expression bias, where one homeolog is preferentially expressed relative to the other, has been reported in multiple other allopolyploid species including *Gossypium* (Hovav et al., 2008; Flagel and Wendel, 2010), *Triticum* (Bottley et al., 2006; Wei et al., 2019), *Brassica* (Auger et al., 2009; Wu et al., 2018; Lee and Adams, 2020), and other species (Grover et al., 2012). Abiotic stress conditions, especially cold stress, considerably impacts expression bias of homeologs involved in physiological responses. Homeologs with differential gene expression are involved in the CBF-COR signaling pathway, fatty acid metabolism which impacts plasma membrane fluidity and stabilization, scavenging reactive oxygen species, sucrose metabolism, and accumulation of secondary metabolites, all which contribute to cold tolerance (Combes et al., 2013; Lee and Adams, 2020; Park and Jang, 2020; Wu et al., 2022). Therefore, studying homeolog gene expression and their sub-functionalization provides foundational knowledge that can be utilized in engineering cold tolerant camelina.

4 Conclusions

Access to chromosome-scale genome assemblies and high quality annotation have been foundational resources for genomics-enabled improvement of crop plants. These data have facilitated the understanding of genetic diversity, population structure, structural variation, and quantitative genetics across many crop species. For camelina to be an adaptable biofuel feedstock, improvements in agronomic performance and optimization of seed oil composition and yield will be required. This will entail both conventional breeding and biotechnological approaches that will be enabled by tapping into genetic diversity (Luo et al., 2019; Li et al., 2021) and facile transformation via floral dip (Liu et al., 2012). Access to a chromosome-scale, highly contiguous genome assembly for the widely used spring biotype Suneson was generated in this study and will enable not only basic research on molecular, physiological, and biochemical traits but also breeding cultivars with improved agronomic and biofuel traits. In

10.3389/fpls.2023.1271625

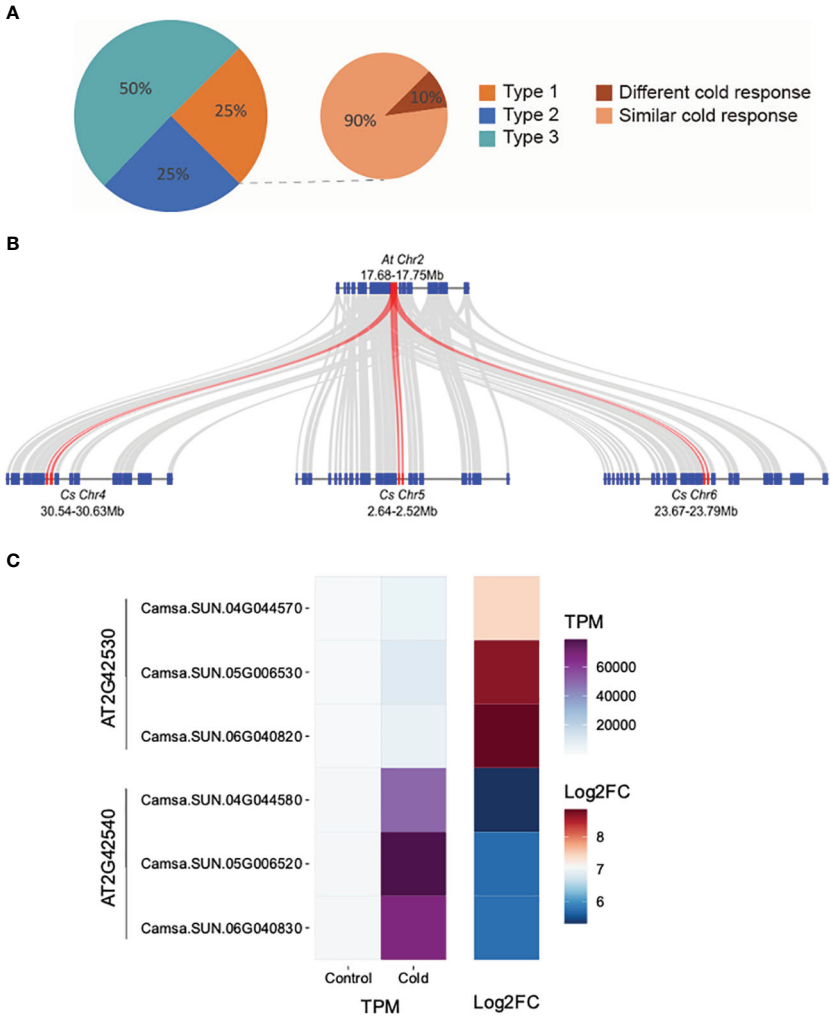


FIGURE 4
Transcriptional divergence of triplicated homeologous genes in *C. sativa* leaves following cold stress. **(A)** Classification of the three types of cold-response triplicated homeologs. The three homeologous genes of Type 1 triplets were all cold-induced; Two of the three homeologs from Type 2 triplets were cold-induced; Only one of the three homeologs from Type 3 triplets was cold-induced. "Different cold response" indicates the Type 1 triplets in which one homeolog showed at least two-times higher cold inducibility than the two other homeologs. The remaining Type 1 triplets are termed as "similar cold response". **(B)** McScan was used to display the systemic relationship of *C. sativa* homeologs of *Arabidopsis thaliana COR15* genes (red) and flanking genes. **(C)** Gene expression of *COR15* homeologs in control and cold-treated leaves.

addition to generation of a chromosome-scale assembly of Suneson and classification of syntelogs with two *Arabidopsis* species, we documented the transcriptional response to cold stress in vegetative leaves including identification of differentially expressed genes, generation of coexpression modules, and characterization of conserved/diverged expression of homeologous genes. These datasets provide a foundation for more detailed interrogation of gene function and regulation in camelina as well as how these diverged from the model species, *A. thaliana*.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: Raw sequence reads for all generated data are available through the National Center for Biotechnology Information Sequence Read Archive under BioProject ID PRJNA927321. The genome assembly, annotation, gene expression abundances and GENESPACE results are available on Figshare (https://figshare.com/s/6f95ce23f7c4eded54d6).

Author contributions

CF: Formal Analysis, Methodology, Writing – original draft, Writing – review & editing, Data curation, Investigation, Software, Visualization. JH: Data curation, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. BV: Data curation, Formal Analysis,

Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. Y-WW: Data curation, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing – review & editing. JW: Investigation, Methodology, Writing – review & editing, Supervision. ND: Investigation, Writing – review & editing, Formal Analysis, Software, Visualization. TS: Writing – review & editing, Methodology. LC: Methodology, Writing – review & editing. KM: Methodology, Writing – review & editing, Supervision. DD: Supervision, Writing – review & editing, Conceptualization, Funding acquisition. SN: Conceptualization, Funding acquisition, Writing – review & editing, Formal Analysis, Investigation, Resources, Writing – original draft. JJ: Conceptualization, Funding acquisition, Investigation, Writing – review & editing, Supervision. CRB: Conceptualization, Funding acquisition, Supervision, Writing – review & editing, Formal Analysis, Methodology, Project administration, Resources, Writing – original draft.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by funds from the University of Georgia, Georgia Seed Development, and the Georgia Research Alliance to CRB, the USDA National Institute of Food and Agriculture Biotechnology Risk Assessment Grant Program (2018-33522-28736) to CRB, SN and DD, and startup funds from Michigan State University to JJ.

Acknowledgments

We acknowledge the expertise of the RTSF Genomics Core at Michigan State University and the Texas A&M AgriLife Research: Genomics and Bioinformatics Service in providing sequencing services.

Conflict of interest

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

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1271625/full#supplementary-material

References

Abdullah, H. M., Akbari, P., Paulose, B., Schnell, D., Qi, W., Park, Y., et al. (2016). Transcriptome profiling of *Camelina sativa* to identify genes involved in triacylglycerol biosynthesis and accumulation in the developing seeds. *Biotechnol. Biofuels* 9, 136. doi: 10.1186/s13068-016-0555-5

Abdullah, H. M., Chhikara, S., Akbari, P., Schnell, D. J., Pareek, A., and Dhankher, O. P. (2018). Comparative transcriptome and metabolome analysis suggests bottlenecks that limit seed and oil yields in transgenic *Camelina sativa* expressing diacylglycerol acyltransferase 1 and glycerol-3-phosphate dehydrogenase. *Biotechnol. Biofuels* 11, 335. doi: 10.1186/s13068-018-1326-2

Alger, E. I., and Edger, P. P. (2020). One subgenome to rule them all: underlying mechanisms of subgenome dominance. *Curr. Opin. Plant Biol.* 54, 108–113. doi: 10.1016/j.pbi.2020.03.004

Alonge, M., Lebeigle, L., Kirsche, M., Jenike, K., Ou, S., Aganezov, S., et al. (2022). Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biol.* 23, 258. doi: 10.1186/s13059-022-02823-7

Anderson, J. V., Neubauer, M., Horvath, D. P., Chao, W. S., and Berti, M. T. (2022). Analysis of *Camelina sativa* transcriptomes identified specific transcription factors and processes associated with freezing tolerance in a winter biotype. *Ind. Crops Prod.* 177, 114414. doi: 10.1016/j.indcrop.2021.114414

Artus, N. N., Uemura, M., Steponkus, P. L., Gilmour, S. J., Lin, C., and Thomashow, M. F. (1996). Constitutive expression of the cold-regulated *Arabidopsis thaliana* COR15a gene affects both chloroplast and protoplast freezing tolerance. *Proc. Natl. Acad. Sci. U. S. A.* 93, 13404–13409. doi: 10.1073/pnas.93.23.13404

Auger, B., Baron, C., Lucas, M.-O., Vautrin, S., Bergès, H., Chalhoub, B., et al. (2009). Brassica orthologs from BANYULS belong to a small multigene family, which is involved in procyanidin accumulation in the seed. *Planta* 230, 1167–1183. doi: 10.1007/s00425-009-1017-0

Barbey, C. R., Lee, S., Verma, S., Bird, K. A., Yocca, A. E., Edger, P. P., et al. (2019). Disease resistance genetics and genomics in octoploid strawberry. *G3* 9, 3315–3332. doi: 10.1534/g3.119.400597

Bengtsson, J. D., Wallis, J. G., Bai, S., and Browse, J. (2023). The coexpression of two desaturases provides an optimized reduction of saturates in camelina oil. *Plant Biotechnol. J.* 21, 497–505. doi: 10.1111/pbi.13966

Berti, M., Gesch, R., Eynck, C., Anderson, J., and Cermak, S. (2016). Camelina uses, genetics, genomics, production, and management. *Ind. Crops Prod.* 94, 690–710. doi: 10.1016/j.indcrop.2016.09.034

Bottley, A., Xia, G. M., and Koebner, R. M. D. (2006). Homoeologous gene silencing in hexaploid wheat. *Plant J.* 47, 897–906. doi: 10.1111/j.1365-313X.2006.02841.x

Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* 34, 525–527. doi: 10.1038/nbt.3519

Brock, J. R., Scott, T., Lee, A. Y., Mosyakin, S. L., and Olsen, K. M. (2020). Interactions between genetics and environment shape Camelina seed oil composition. *BMC Plant Biol.* 20, 423. doi: 10.1186/s12870-020-02641-8

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12, 59–60. doi: 10.1038/nmeth.3176

Campbell, M. A., Haas, B. J., Hamilton, J. P., Mount, S. M., and Buell, C. R. (2006). Comprehensive analysis of alternative splicing in rice and comparative analyses with Arabidopsis. *BMC Genomics* 7, 327. doi: 10.1186/1471-2164-7-327

Campbell, M. S., Law, M., Holt, C., Stein, J. C., Moghe, G. D., Hufnagel, D. E., et al. (2014). MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Physiol.* 164, 513–524. doi: 10.1104/pp.113.230144

Chaudhary, R., Koh, C. S., Kagale, S., Tang, L., Wu, S. W., Lv, Z., et al. (2020). Assessing Diversity in the Camelina Genus Provides Insights into the Genome Structure of Camelina sativa. *G3* 10, 1297–1308. doi: 10.1534/g3.119.400957

Chen, C., Chen, H., Zhang, Y., Thomas, H. R., Frank, M. H., He, Y., et al. (2020). TBtools: an integrative toolkit developed for interactive analyses of big biological data. *Mol. Plant* 13, 1194–1202. doi: 10.1016/j.molp.2020.06.009

Chen, N. (2004). Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinf.* 5, 4.10 11–14.10. 14.. doi: 10.1002/0471250953.bi0410s05

Combes, M.-C., Dereeper, A., Severac, D., Bertrand, B., and Lashermes, P. (2013). Contribution of subgenomes to the transcriptome and their intertwined regulation in the allopolyploid *Coffea arabica* grown at contrasted temperatures. *New Phytol.* 200, 251–260. doi: 10.1111/nph.12371

de Jong, G. W., and Adams, K. L. (2023). Subgenome-dominant expression and alternative splicing in response to Sclerotinia infection in polyploid *Brassica napus* and progenitors. *Plant J.* 114, 142–158. doi: 10.1111/tpj.16127

Demski, K., Łosiewska, A., Jasieniecka-Gazarkiewicz, K., Klińska, S., and Banaś, A. (2020). Phospholipid:Diacylglycerol acyltransferase1 overexpression delays senescence and enhances post-heat and cold exposure fitness. *Front. Plant Sci.* 11, 611897. doi: 10.3389/fpls.2020.611897

Edger, P. P., Poorten, T. J., VanBuren, R., Hardigan, M. A., Colle, M., McKain, M. R., et al. (2019). Origin and evolution of the octoploid strawberry genome. *Nat. Genet.* 51, 541–547. doi: 10.1038/s41588-019-0356-4

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, D427–D432. doi: 10.1093/nar/gky995

Flagel, L. E., and Wendel, J. F. (2010). Evolutionary rate variation, genomic dominance and duplicate gene expression evolution during allotetraploid cotton speciation. *New Phytol.* 186, 184–193. doi: 10.1111/j.1469-8137.2009.03107.x

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, 9451–9457. doi: 10.1073/pnas.1921046117

Gomez-Cano, F., Carey, L., Lucas, K., García Navarrete, T., Mukundi, E., Lundback, S., et al. (2020). CamRegBase: a gene regulation database for the biofuel crop, *Camelina sativa*. *Database* 2020, baaa075. doi: 10.1093/database/baaa075

Gomez-Cano, F., Chu, Y.-H., Cruz-Gomez, M., Abdullah, H. M., Lee, Y. S., Schnell, D. J., et al. (2022). Exploring *Camelina sativa* lipid metabolism regulation by combining gene co-expression and DNA affinity purification analyses. *Plant J.* 110, 589–606. doi: 10.1111/tpj.15682

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., et al. (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* 40, D1178–D1186. doi: 10.1093/nar/gkr944

Grover, C. E., Gallagher, J. P., Szadkowski, E. P., Yoo, M. J., Flagel, L. E., and Wendel, J. F. (2012). Homoeolog expression bias and expression level dominance in allopolyploids. *New Phytol.* 196, 966–971. doi: 10.1111/j.1469-8137.2012.04365.x

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

Heydarian, Z., Yu, M., Gruber, M., Coutu, C., Robinson, S. J., and Hegedus, D. D. (2018). Changes in gene expression in Camelina sativa roots and vegetative tissues in response to salinity stress. *Sci. Rep.* 8, 9804. doi: 10.1038/s41598-018-28204-4

Hoff, K. J., Lomsadze, A., Borodovsky, M., and Stanke, M. (2019). "Whole-genome annotation with BRAKER," in *Gene Prediction: Methods and Protocols*. Ed. M. Kollmar (New York, NY: Springer New York), 65–95.

Hovav, R., Udall, J. A., Chaudhary, B., Rapp, R., Flagel, L., and Wendel, J. F. (2008). Partitioned expression of duplicated genes during development and evolution of a single cell in a polyploid plant. *Proc. Natl. Acad. Sci. U. S. A.* 105, 6191–6195. doi: 10.1073/pnas.0711569105

Hu, Y., Jiang, L., Wang, F., and Yu, D. (2013). Jasmonate regulates the INDUCER OF CBF EXPRESSION–C-REPEAT BINDING FACTOR/DRE BINDING FACTOR1 cascade and freezing tolerance in arabidopsis. *Plant Cell* 25, 2907–2924. doi: 10.1105/tpc.113.112631

Jiang, W. Z., Henry, I. M., Lynagh, P. G., Comai, L., Cahoon, E. B., and Weeks, D. P. (2017). Significant enhancement of fatty acid composition in seeds of the allohexaploid, *Camelina sativa*, using CRISPR/Cas9 gene editing. *Plant Biotechnol. J.* 15, 648–657. doi: 10.1111/pbi.12663

Kagale, S., Koh, C., Nixon, J., Bollina, V., Clarke, W. E., Tuteja, R., et al. (2014). The emerging biofuel crop *Camelina sativa* retains a highly undifferentiated hexaploid genome structure. *Nat. Commun.* 5, 3706. doi: 10.1038/ncomms4706

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

King, K., Li, H., Kang, J., and Lu, C. (2019). Mapping quantitative trait loci for seed traits in Camelina sativa. *Theor. Appl. Genet.* 132, 2567–2577. doi: 10.1007/s00122-019-03371-8

Kokot, M., Dlugosz, M., and Deorowicz, S. (2017). KMC 3: counting and manipulating k-mer statistics. *Bioinformatics* 33, 2759–2761. doi: 10.1093/bioinformatics/btx304

Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* 37, 540–546. doi: 10.1038/s41587-019-0072-8

Kovaka, S., Zimin, A. V., Pertea, G. M., Razaghi, R., Salzberg, S. L., and Pertea, M. (2019). Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* 20, 278. doi: 10.1186/s13059-019-1910-1

Lee, J. S., and Adams, K. L. (2020). Global insights into duplicated gene expression and alternative splicing in polyploid *Brassica napus* under heat, cold, and drought stress. *Plant Genome* 13, e20057. doi: 10.1002/tpg2.20057

Lee, K. R., Jeon, I., Yu, H., Kim, S. G., Kim, H. S., Ahn, S. J., et al. (2021). Increasing monounsaturated fatty acid contents in hexaploid *Camelina sativa* seed oil by FAD2 gene knockout using CRISPR-cas9. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.702930

Lhamo, D., Shao, Q., Tang, R., and Luan, S. (2020). Genome-wide analysis of the five phosphate transporter families in camelina sativa and their expressions in response to low-P. *Int. J. Mol. Sci.* 21, 8365. doi: 10.3390/ijms21218365

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*. doi: 10.48550/arXiv.1303.3997

Li, C., Deans, N. C., and Buell, C. R. (2023a). "Simple Tidy GeneCoEx": A gene co-expression analysis workflow powered by tidyverse and graph-based clustering in R. *Plant Genome* 16, e20323. doi: 10.1002/tpg2.20323

Li, H., Hu, X., Lovell, J. T., Grabowski, P. P., Mamidi, S., Chen, C., et al. (2021). Genetic dissection of natural variation in oilseed traits of camelina by whole-genome resequencing and QTL mapping. *Plant Genome* 14, e20110. doi: 10.1002/tpg2.20110

Li, C., Wood, J. C., Vu, A. H., Hamilton, J. P., Rodriguez Lopez, C. E., Payne, R. M. E., et al. (2023b). Single-cell multi-omics enabled discovery of alkaloid biosynthetic pathway genes in the medical plant *Catharanthus roseus*. *Nat. Chem. Biol.* 19, 1031–1041. doi: 10.1038/s41589-023-01327-0

Lin, C., and Thomashow, M. F. (1992a). A cold-regulated Arabidopsis gene encodes a polypeptide having potent cryoprotective activity. *Biochem. Biophys. Res. Commun.* 183, 1103–1108. doi: 10.1016/S0006-291X(05)80304-3

Lin, C., and Thomashow, M. F. (1992b). DNA sequence analysis of a complementary DNA for cold-regulated arabidopsis gene cor15 and characterization of the COR 15 polypeptide. *Plant Physiol.* 99, 519–525. doi: 10.1104/pp.99.2.519

Liu, X., Brost, J., Hutcheon, C., Guilfoil, R., Wilson, A., Leung, S., et al. (2012). Transformation of the oilseed crop *Camelina sativa* by *Agrobacterium*-mediated floral dip and simple large-scale screening of transformants. *In Vitro Cell. Dev. Biol. - Plant* 48, 462–468. doi: 10.1007/s11627-012-9459-7

Lovell, J. T., Sreedasyam, A., Schranz, M. E., Wilson, M., Carlson, J. W., Harkess, A., et al. (2022). GENESPACE tracks regions of interest and gene copy number variation across multiple genomes. *Elife* 11, e78526. doi: 10.7554/eLife.78526

Luo, Z., Tomasi, P., Fahlgren, N., and Abdel-Haleem, H. (2019). Genome-wide association study (GWAS) of leaf cuticular wax components in *Camelina sativa* identifies genetic loci related to intracellular wax transport. *BMC Plant Biol.* 19, 187. doi: 10.1186/s12870-019-1776-0

Malik, M. R., Tang, J., Sharma, N., Burkitt, C., Ji, Y., Mykytyshyn, M., et al. (2018). *Camelina sativa*, an oilseed at the nexus between model system and commercial crop. *Plant Cell Rep.* 37, 1367–1381. doi: 10.1007/s00299-018-2308-3

Mandáková, T., Pouch, M., Brock, J. R., Al-Shehbaz, I. A., and Lysak, M. A. (2019). Origin and evolution of diploid and allopolyploid *Camelina* genomes were accompanied by chromosome shattering. *Plant Cell* 31 (11), 2596–2612. doi: 10.1105/tpc.19.00366

Mapleson, D., Garcia Accinelli, G., Kettleborough, G., Wright, J., and Clavijo, B. J. (2017). KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics* 33, 574–576. doi: 10.1093/bioinformatics/btw663

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

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 17, 10–12. doi: 10.14806/ej.17.1.200

Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A., and Punta, M. (2013). Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* 41, e121. doi: 10.1093/nar/gkt263

Moore, R. H., Thornhill, K. L., Weinzierl, B., Sauer, D., D'Ascoli, E., Kim, J., et al. (2017). Biofuel blending reduces particle emissions from aircraft engines at cruise conditions. *Nature* 543, 411–415. doi: 10.1038/nature21420

Morineau, C., Bellec, Y., Tellier, F., Gissot, L., Kelemen, Z., Nogue, F., et al. (2017). Selective gene dosage by CRISPR-Cas9 genome editing in hexaploid *Camelina sativa*. *Plant Biotechnol. J.* 15, 729–739. doi: 10.1111/pbi.12671

Na, G., Aryal, N., Fatihi, A., Kang, J., and Lu, C. (2018). Seed-specific suppression of ADP-glucose pyrophosphorylase in *Camelina sativa* increases seed size and weight. *Biotechnol. Biofuels* 11, 330. doi: 10.1186/s13068-018-1334-2

Na, G., Mu, X., Grabowski, P., Schmutz, J., and Lu, C. (2019). Enhancing microRNA167A expression in seed decreases the α-linolenic acid content and increases seed size in *Camelina sativa*. *Plant J.* 98, 346–358. doi: 10.1111/tpj.14223

Nguyen, H. T., Silva, J. E., Podicheti, R., Macrander, J., Yang, W., Nazarenus, T. J., et al. (2013). Camelina seed transcriptome: a tool for meal and oil improvement and translational research. *Plant Biotechnol. J.* 11, 759–769. doi: 10.1111/pbi.12068

Obour, A. K., Obeng, E., Mohammed, Y. A., Ciampitti, I. A., Durrett, T. P., Aznar-Moreno, J. A., et al. (2017). Camelina seed yield and fatty acids as influenced by genotype and environment. *Agron. J.* 109, 947–956. doi: 10.2134/agronj2016.05.0256

Ozseyhan, M. E., Kang, J., Mu, X., and Lu, C. (2018). Mutagenesis of the FAE1 genes significantly changes fatty acid composition in seeds of *Camelina sativa*. *Plant Physiol. Biochem.* 123, 1–7. doi: 10.1016/j.plaphy.2017.11.021

Park, Y. C., and Jang, C. S. (2020). Molecular dissection of two homoeologous wheat genes encoding RING H2-type E3 ligases: TaSIRFP-3A and TaSIRFP-3B. *Planta* 252, 26. doi: 10.1007/s00425-020-03431-0

Pertea, M., Kim, D., Pertea, G. M., Leek, J. T., and Salzberg, S. L. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc.* 11, 1650–1667. doi: 10.1038/nprot.2016.095

Pham, G. M., Hamilton, J. P., Wood, J. C., Burke, J. T., Zhao, H., Vaillancourt, B., et al. (2020). Construction of a chromosome-scale long-read reference genome assembly for potato. *Gigascience* 9, giaa100. doi: 10.1093/gigascience/giaa100

Ranallo-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

Raziei, Z., Kahrizi, D., and Rostami-Ahmadvandi, H. (2018). Effects of climate on fatty acid profile in *Camelina sativa*. *Cell. Mol. Biol.* 64, 91–96. doi: 10.14715/cmb/2018.64.5.15

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616

Rodríguez-Rodríguez, M. F., Sánchez-García, A., Salas, J. J., Garcés, R., and Martínez-Force, E. (2013). Characterization of the morphological changes and fatty acid profile of developing *Camelina sativa* seeds. *Ind. Crops Prod.* 50, 673–679. doi: 10.1016/j.indcrop.2013.07.042

Rupwate, S. D., and Rajasekharan, R. (2012). Plant phosphoinositide-specific phospholipase C: an insight. *Plant Signal. Behav.* 7, 1281–1283. doi: 10.4161/psb.21436

Schnable, J. C., Springer, N. M., and Freeling, M. (2011). Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc. Natl. Acad. Sci. U. S. A.* 108, 4069–4074. doi: 10.1073/pnas.1101368108

Shen, W., Le, S., Li, Y., and Hu, F. (2016). SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PloS One* 11, e0163962. doi: 10.1371/journal.pone.0163962

Soorni, J., Kazemitabar, S. K., Kahrizi, D., Dehestani, A., Bagheri, N., Kiss, A., et al. (2022). Biochemical and transcriptional responses in cold-acclimated and non-acclimated contrasting camelina biotypes under freezing stress. *Plants* 11, 3178. doi: 10.3390/plants11223178

Takumi, S., Koike, A., Nakata, M., Kume, S., Ohno, R., and Nakamura, C. (2003). Cold-specific and light-stimulated expression of a wheat (*Triticum aestivum* L.) Cor gene Wcor15 encoding a chloroplast-targeted protein. *J. Exp. Bot.* 54, 2265–2274. doi: 10.1093/jxb/erg247

Tan, W.-J., Yang, Y.-C., Zhou, Y., Huang, L.-P., Xu, L., Chen, Q.-F., et al. (2018). DIACYLGLYCEROL ACYLTRANSFERASE and DIACYLGLYCEROL KINASE modulate triacylglycerol and phosphatidic acid production in the plant response to freezing stress. *Plant Physiol.* 177, 1303–1318. doi: 10.1104/pp.18.00402

Thalhammer, A., and Hincha, D. K. (2013). "The function and evolution of closely related COR/LEA (Cold-regulated/late embryogenesis abundant) proteins in Arabidopsis thaliana," in *Plant and Microbe Adaptations to Cold in a Changing World* (New York: Springer), 89–105.

Thomas, B. C., Pedersen, B., and Freeling, M. (2006). Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Res.* 16, 934–946. doi: 10.1101/gr.4708406

Thomashow, M. F. (1998). Role of cold-responsive genes in plant freezing tolerance. *Plant Physiol.* 118, 1–8. doi: 10.1104/pp.118.1.1

Vaillancourt, B., and Buell, C. R. (2019). High molecular weight DNA isolation method from diverse plant species for use with Oxford Nanopore sequencing. *BioRxiv*. doi: 10.1101/783159

Vollmann, J., and Eynck, C. (2015). Camelina as a sustainable oilseed crop: Contributions of plant breeding and genetic engineering. *Biotechnol. J.* 10, 525–535. doi: 10.1002/biot.201400200

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., et al. (2017). GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* 33, 2202–2204. doi: 10.1093/bioinformatics/btx153

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

Wan, C. Y., and Wilkins, T. A. (1994). A modified hot borate method significantly enhances the yield of high-quality RNA from cotton (*Gossypium hirsutum* L.). *Anal. Biochem.* 223, 7–12. doi: 10.1006/abio.1994.1538

Wang, H., Doğramacı, M., Anderson, J. V., Horvath, D. P., and Chao, W. S. (2022). Transcript profiles differentiate cold acclimation-induced processes in a summer and winter biotype of Camelina. *Plant Mol. Biol. Rep.* 40, 359–375. doi: 10.1007/s11105-021-01324-4

Waterhouse, R. M., Seppey, M., Simão, F. A., Manni, M., Ioannidis, P., Klioutchnikov, G., et al. (2018). BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol.* 35, 543–548. doi: 10.1093/molbev/msx319

Wei, J., Cao, H., Liu, J.-D., Zuo, J.-H., Fang, Y., Lin, C.-T., et al. (2019). Insights into transcriptional characteristics and homoeolog expression bias of embryo and de-embryonated kernels in developing grain through RNA-Seq and Iso-Seq. *Funct. Integr. Genomics* 19, 919–932. doi: 10.1007/s10142-019-00693-0

Wilhelm, K. S., and Thomashow, M. F. (1993). Arabidopsis thaliana cor15b, an apparent homologue of cor15a, is strongly responsive to cold and ABA, but not drought. *Plant Mol. Biol.* 23, 1073–1077. doi: 10.1007/BF00021822

Wood, D. E., Lu, J., and Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biol.* 20, 257. doi: 10.1186/s13059-019-1891-0

Wu, W., Guo, W., Ni, G., Wang, L., Zhang, H., and Ng, W. L. (2022). Expression Level Dominance and Homeolog Expression Bias Upon Cold Stress in the F1 Hybrid Between the Invasive *Sphagneticola trilobata* and the Native *S. calendulacea* in South China, and Implications for Its Invasiveness. *Front. Genet.* 13. doi: 10.3389/fgene.2022.833406

Wu, J., Lin, L., Xu, M., Chen, P., Liu, D., Sun, Q., et al. (2018). Homoeolog expression bias and expression level dominance in resynthesized allopolyploid *Brassica napus*. *BMC Genomics* 19, 586. doi: 10.1186/s12864-018-4966-5

Yuan, L., Mao, X., Zhao, K., Ji, X., Ji, C., Xue, J., et al. (2017). Characterisation of phospholipid: diacylglycerol acyltransferases (PDATs) from *Camelina sativa* and their roles in stress responses. *Biol. Open* 6, 1024–1034. doi: 10.1242/bio.026534

Zanetti, F., Alberghini, B., Marjanović Jeromela, A., Grahovac, N., Rajković, D., Kiprovski, B., et al. (2021). Camelina, an ancient oilseed crop actively contributing to the rural renaissance in Europe. A review. *Agron. Sustain. Dev.* 41, 2. doi: 10.1007/s13593-020-00663-y

Zhu, J., Wang, X., Guo, L., Xu, Q., Zhao, S., Li, F., et al. (2018). Characterization and alternative splicing profiles of the lipoxygenase gene family in tea plant (*Camellia sinensis*). *Plant Cell Physiol.* 59, 1765–1781. doi: 10.1093/pcp/pcy091

Zou, C., Lehti-Shiu, M. D., Thomashow, M., and Shiu, S. H. (2009). Evolution of stress-regulated gene expression in duplicate genes of *Arabidopsis thaliana*. *PloS Genet.* 5, e1000581. doi: 10.1371/journal.pgen.1000581

Zheng, Y, Jiao, C, Sun, H, Rosli, HG, Pombo, MA, and Zhang, P (2016). iTAK: A program for genome-wide prediction and classification of plant transcription factors, transcriptional regulators, and protein kinases *Arabidopsis thaliana*. *Mol Plant.* 9 (12), 1667–1670. doi: 10.1016/j.molp.2016.09.014

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 04 January 2024
DOI 10.3389/fpls.2023.1337736

OPEN ACCESS

EDITED BY
Manohar Chakrabarti,
The University of Texas Rio Grande Valley, United States

REVIEWED BY
Prathima Perumal Thirugnanasambandam, Indian Council of Agricultural Research, Coimbatore, India
Dong-Liang Huang, Guangxi Academy of Agricultural Sciences, China
Avinash Singode, Indian Council of Agricultural Research (ICAR), India

*CORRESPONDENCE
Gul Shad Ali
Gul.Ali@usda.gov
Dapeng Zhang
dapeng.zhang@usda.gov

RECEIVED 13 November 2023
ACCEPTED 15 December 2023
PUBLISHED 04 January 2024

CITATION
Park S, Zhang D and Ali GS (2024) Assessing the genetic integrity of sugarcane germplasm in the USDA-ARS National Plant Germplasm System collection using single-dose SNP markers. *Front. Plant Sci.* 14:1337736. doi: 10.3389/fpls.2023.1337736

Assessing the genetic integrity of sugarcane germplasm in the USDA-ARS National Plant Germplasm System collection using single-dose SNP markers

Sunchung Park[1], Dapeng Zhang[1]* and Gul Shad Ali[2]*

[1]Sustainable Perennial Crops Laboratory, United States Department of Agriculture, Agriculture Research Service, Beltsville, MD, United States, [2]Subtropical Horticulture Research Station, United States Department of Agriculture, Agriculture Research Service, Miami, FL, United States

The World Collection of Sugarcane and Related Grasses, maintained at the USDA-ARS in Miami, FL, is one of the largest sugarcane germplasm repositories in the world. However, the genetic integrity of the *Saccharum* spp. germplasm in this collection has not been fully analyzed. In this study, we employed a single-dose SNP panel to genotype 901 sugarcane accessions, representing six *Saccharum* species and various hybrids. Our analysis uncovered a high rate of clone mislabeling in the collection. Specifically, we identified 86 groups of duplicates, characterized by identical SNP genotypes, which encompassed 211 accessions (23% of the total clones), while 135 groups, constituting 471 clones (52% of the total), exhibited near-identical genotypes. In addition, twenty-seven homonymous groups were detected, which shared the same clone name but differed in SNP genotypes. Hierarchical analysis of population structure partitioned the *Saccharum* germplasm into five clusters, corresponding to *S. barberi, S. sinense, S. officinarum, S. spontaneum* and *S. robustum/S. edule*. An assignment test, based on the five *Saccharum* species, enabled correcting 141 instances of mislabeled species memberships and inaccuracies. Moreover, we clarified the species membership and parentage of 298 clones that had ambiguous passport records (e.g., '*Saccharum spp*', 'unknown', and 'hybrid'). Population structure and genetic diversity in these five species were further supported by Principal Coordinate Analysis and neighbor-joining clustering analysis. Analysis of Molecular Variance revealed that within-species genetic variations accounted for 85% of the total molecular variance, with the remaining 15% attributed to among-species genetic variations. The single-dose SNP markers developed in this study offer a robust tool for characterizing sugarcane germplasm worldwide. These findings have important implications for sugarcane genebank management, germplasm exchange, and crop genetic improvement.

KEYWORDS

Saccharum, germplasm, sugarcane, genetic diversity, population structure

1 Introduction

Sugarcane (*Saccharum* spp.) is a prolific energy crop that serves as a substantial source of sugar, biofuel, and other industrial chemicals (Formann et al., 2020). It is cultivated worldwide in over 95 countries across 26.4 million hectares, with a total production of 1.86 billion metric tons and a gross production value of $96.5 billion dollars (FAO, 2023). Sugarcane plays a crucial role in the economies of many tropical and subtropical countries, meeting nearly 80% of global sugar demands for food consumption and accounting for approximately 40% of the world's bioethanol needs (Lam et al., 2009). Furthermore, lignocellulosic biomass derived from sugarcane and energy cane is recognized as a promising feedstock for biofuel production. As the demand for sugar and biofuel continues to rise, a challenge for sugarcane breeding programs is to develop improved varieties with higher yield, sucrose content, disease resistance, improved ratooning ability and adaptability to environmental stresses. Genetic diversity is crucial in developing such varieties to unlock the full potential of sugarcane as a feedstock for both sugar and fiber production. (Lam et al., 2009; Hoang et al., 2015).Sugarcane species belong to the grass family *Poacea*, genus *Saccharum*, and share genetic similarities with *Sorghum* and other grasses (Spangler et al., 1999). Within the *Saccharum* genus, six main species are recognized: two wild species, *S. spontaneum* ($2n$ = 40 -128, x = 8) and *S. robustum* ($2n$ = 60 – 80), and four cultivated species, *S. officinarum* ($2n$ = 80, x =10), *S. barberi* ($2n$ = 81 -124), *S. sinense* ($2n$ = 111 – 120), and *S. edule* ($2n$ = 60, 70, 80) (Moore et al., 2013). Genetic studies suggest that *S. officinarum* and *S. edule* originated from *S. robustum* (Grivet et al., 2004; Grivet et al., 2006) and that *S. sinense* and *S. barberi* are interspecific hybrids resulting from a cross between *S. officinarum* an *S. spontaneum*, with 32 -39% of their genomes derived from *S. spontaneum* (D'Hont et al., 1996; Piperidis et al., 2010). Most modern sugarcane cultivars are complex polyploids ($2n$ = $4x$ to $12x$, totaling 100 – 128 chromosomes), resulting from interspecific crosses between sugar-rich *S. officinarum* and *S. spontaneum* with disease resistance, vigor and other agronomic traits (Piperidis et al., 2010).

Since the early 1970s, sugarcane productivity has steadily increased, largely attributed to improved varieties and agronomic practices (Moore et al., 2013; Hale et al., 2022). However, the sugarcane industry faces challenges posed by diseases, pests, adaptability to different soil types, water availability and environmental stresses, underscoring the need for the development of new and resilient sugarcane varieties with high sucrose content. The current repertoire of modern sugarcane cultivars is the result of crosses made in the early 1900s, involving fewer than 20 *S. officinarum* and *S. spontaneum* clones as parents (Deren, 1995; Raboin et al., 2006), essentially resulting in a monoculture of a few dominant sugarcane varieties grown across large geographic areas, and making them vulnerable to disease and pest outbreaks. To enhance the resilience of sugarcane varieties, it is important to increase the genetic diversity of sugarcane germplasm. Recognizing the importance of landraces and wild relatives as sources of novel genetic traits, efforts for enhancing genetic diversity should be focused on introgression of genes from landraces, wild species within *Saccharum* species complex, and crossable wild relatives such as *Miscanthus* and *Tripidium*. The World Collection of Sugarcanes and Related Grasses (WCSRG), maintained by the United States Department of Agriculture (USDA), Agriculture Research Service (ARS), Subtropical Horticulture Research Station (SHRS), serves as a repository for one of the world largest collections of sugarcane and its wild relatives, originating from various geographical regions. This collection has been used in breeding programs and biological studies to improve sugarcane varieties (Nayak et al., 2014; Zhang et al., 2018; You et al., 2019; Fickett et al., 2020; Hale et al., 2022; Wang et al., 2022). Moreover, as a USDA National Plant Germplasm collection, the clones are freely distributed to international sugarcane community. From 2010 to 2021, a total of 9439 cuttings of various *Saccharum* spp., and 298 cuttings of 4 *Tripidium* spp., were distributed to researchers and breeders in the USA (44%) and internationally (56%) (Hemaprabha et al., 2022).

The WCSRG currently houses approximately 1164 accessions, the majority of which belong to *Saccharum* spp, including 307 *S. spontaneum* accessions, 158 *S. officinarum* accessions, 127 *S. hybrid* accessions, 81 *S. robustum* accessions, 48 *S. sinense* accessions, and 33 *S. barberi* accessions. These accessions originated from various geoclimatic regions and likely harbor genes for adaptation to different climatic stresses, pests, and diseases. Ample information on genetic diversity and population structure within this collection has been generated using molecular markers (Brown et al., 2007; Nayak et al., 2014; Fickett et al., 2020; Xiong et al., 2022) and candidate genes (Parco et al., 2017). Based on simple sequence repeat (SSR) genotyping results, a core collection including 300 accessions was proposed (Nayak et al., 2014). Furthermore, target enrichment sequencing was performed on 307 germplasm accessions from this collection, leading to the identification of ancestor of ancient and modern hybrids in *Saccharum* spp. (Yang et al., 2019). Based on the sequencing data, a genome-wide association study was performed on this diversity panel and candidate genes for agronomic traits and disease resistance were identified (Yang et al., 2019; Yang et al., 2020).

Despite the progress achieved in the molecular characterization of the WCSRG, genetic integrity of the sugarcane germplasm maintained in this collection has not been systematically analyzed. This is mainly because accurate identification of sugarcane germplasm has been technically challenging, due to the high polyploidy (and aneuploidy) nature of this crop (Brown et al., 2007; Song et al., 2016). For any given locus in sugarcane, there can be 8 to 12 alleles in different configurations, which ambiguates genotype identification. Therefore, single-dose molecular markers are needed to distinguish among sugarcane genotypes with complex allele configurations (Song et al., 2016). Recently, You et al. (2019) reported the target enrichment sequencing of 300 sugarcane accessions selected from the world collection and developed an Affymetrix Axiom 100K SNP array. This array, which comprises 31,449 single-dose SNPs and 68,648 low-dosage SNPs, provides a powerful tool for using single-dose SNPs in sugarcane germplasm identification.

In this study, we selected 2000 single-dose SNP markers from the Affymetrix Axiom SNP array (You et al., 2019). After validating

the selected candidate SNPs in a pilot study, we selected the final genotyping panel and used it to genotype all the *Saccharum* germplasm, including six *Saccharum* species and hybrids. Through comprehensive genotyping and population structure analyses, we assessed genetic integrity, population structure, and genetic diversity in *Saccharum* germplasm. We identified a high rate of clone mislabeling and redundancy within the sugarcane germplasm collection, characterized by clone duplicate errors, homonymous off-types, and mistakes in species memberships. Moreover, our analyses of population structure and genetic diversity revealed novel insights into the classification and inter-relationships of *Saccharum* species. Overall, these findings provide valuable information for sugarcane research community to improve the accuracy and efficiency of managing and utilizing the sugarcane genetic resources in the WCSRG.

2 Materials and methods

2.1 Plant materials

The sugarcane accessions reported in this study are part of the WCSRG, which is curated by the USDA-ARS, SHRS, in Miami, FL. Figure 1 provides a summary of the geographical distribution of the germplasm accessions. A detailed list of all accessions is provided in Table S1. The S. spontaneum accessions are maintained in 7-gallon pots on a concrete pad and not allowed to flower as they are considered invasive. The rest of the accessions are planted in the field and rotated to new field plots every 4 years. The mature plants are cut to the ground every year in the early spring until replanting. The species name of each accession in the WCSRG was defined based on the curator's naming system.

From each sugarcane plant, one fully expanded young leaf was collected into labeled paper envelopes. A total of eight leaf disks were collected using the BioArk Leaf kit provided by LGC, Biosearch Technologies (https://www.biosearchtech.com/). The prepared BioArk Leaf kits were then submitted to LGC Genomics (Middleton, WI) for DNA extraction and subsequent genotyping.

2.2 SNP markers, Genotyping and SNP calling

The single-dose markers were initially selected from the Axiom Sugarcane 100K SNP array, which was developed based on five *Saccharum* species (*S. sinense*, *S. barberi*, *S. robustum*, *S. officinarum*, and *S. spontaneum*) and 37 sugarcane hybrids (You et al., 2019). A set of two thousand bi-allelic and single-dose SNPs were randomly selected from the Axiom Sugarcane 100K SNP array. Probes targeting these SNPs were designed by Biosearch Technologies (https://www.biosearchtech.com) and used to amplify sugarcane genome DNA libraries, which were then sequenced using the 1x 75 bp Illumina sequencing platform. Reads were trimmed by removing the first 40 bases and quality-checked with a Q value >20. The 2000 candidate SNPs were first evaluated in a pilot study using 196 sugarcane germplasm accessions (Table S1). Based on call rate, Minor Allele Frequency (MAF), and Linkage Disequilibrium (LD), we then selected a low-density genotyping panel including 400 SNPs and used it to genotype all the *Saccharum* clones including six *Saccharum* species and inter-specific hybrids.

The genotyping was performed using a targeted genotyping-by-sequencing approach called SeqSNP, which has been successfully used in several crops (Zhang et al., 2020; Jo et al., 2021; Ziarsolo et al., 2021). Sequence reads were aligned to the sequences (300bp)

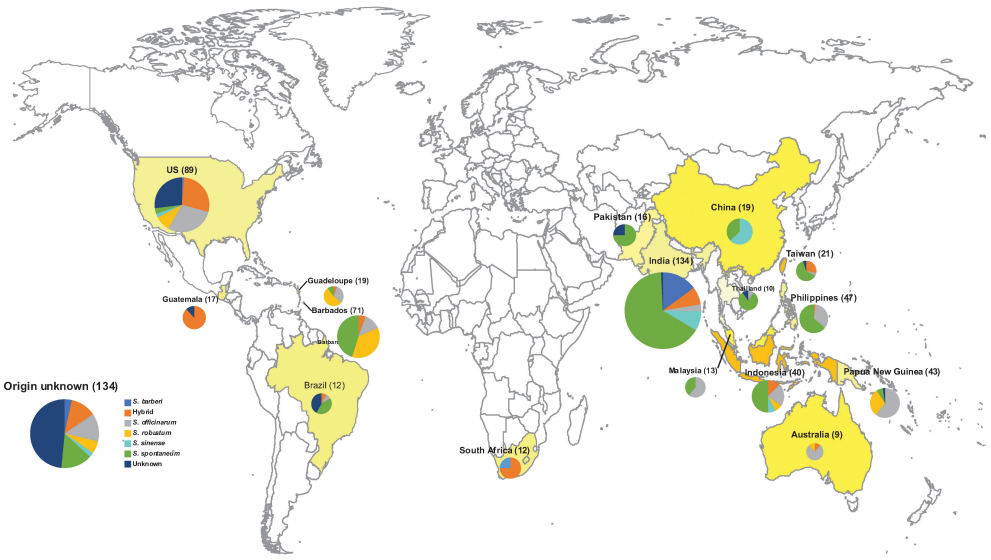


FIGURE 1
The Geographical origin of 901 *Saccharum* germplasm accessions analyzed in the present study, including *S. spontaneum* (283), *S. robustum* (76), *S. officinarum* (158), *S. barberi* (32), *S. sinense* (46), *S. edule* (3), *Saccharum* spp. (175), Unknown (3), hybrid (123). A detailed list with passport information is presented in Table S1.

flanking the SNP markers, using Bowtie2 (version 2.4.5) (Langmead and Salzberg, 2012). The subsequent alignments (BAM files) were used to call SNP variants by the freebayes program (version 1.1.0) (Garrison and Marth, 2012) with the following parameters: min-quality 20, –min-supporting-allele-qsum 10; –read-mismatch-limit 4; –mismatch-base-quality-threshold 10; –exclude-unobserved-genotypes; –no-mnps; –no-complex; –ploidy 4; –min-alternate-fraction.08333; –legacy-gls. The called SNP variants (VCF files) were further filtered using the following criteria: 1) SNPs must be biallelic; 2) SNPs must be supported by at least 20 reads, otherwise, they were marked as missing genotypes; 3) both alternative and reference alleles must each be supported by at least 2 reads. After the initial filtering, informative SNPs were selected by excluding SNPs with a missing rate of 10% and a minor allele frequency of <5%. Additionally, samples with 10% or more missing SNP genotyping were also excluded. These filtering processes resulted in a dataset of 357 SNPs and 901 samples for downstream analyses. To assess the degree of variation among SNP markers, genetic parameters such as minor allele frequency (MAF), expected heterozygosity (H_{exp}), and observed heterozygosity (H_{obs}) for each SNP marker were measured using the R-package snpReady (version 0.9.6) (Granato et al., 2018).

2.3 Clone mislabeling and genetic redundancy

For this study, we defined three types of problems related to genetic integrity of sugarcane germplasm. The first type was synonymous mislabeling or "duplicate error," meaning that sugarcane clones had different names but shared the same SNP genotype. The second was homonymous mislabeling, meaning that individual clones had the same name in this collection, but they had different SNP genotypes. The third type was mistakes, inaccuracies, or a lack of information in species classification, where the species membership of a given clone was wrongfully recorded (Brown et al., 2007; Yang et al., 2019).

To identify synonymous mislabeling among sugarcane clones, the allele difference between each pair of individuals was computed using the R-package poppr (Kamvar et al., 2014). Individuals with zero allele difference at all loci were considered duplicates. The groups of duplicates were visually inspected by constructing a network graph using the R-packages of network and ggplot2. Since genotyping errors are not uncommon, a pair of clones that differ by one or two loci could be the same clone (Kalinowski et al., 2006; Zhang et al., 2006). To assess potential genotyping error, we included 12 sugarcane samples in genotyping. These 12 samples were propagated from a single clone "P-Mag-84-2" (Table S1) and served as an internal control. Mismatched SNP loci among these 12 samples were calculated and used as a baseline to determine the "mismatch threshold" for clone identification. Any pair of samples that had mismatched loci below the threshold (near-identical genotypes) were considered as putative duplicates (Zhang et al., 2006; Akpertey et al., 2021).

The statistical rigor of duplicate identification was assessed using the probability of identity that two individuals may share the same multilocus genotype by chance (Waits et al., 2001). The computer program GenAlEx 6.5 (Peakall and Smouse, 2006; Peakall and Smouse, 2012) was used to calculate the probability of identity among siblings (PIDsib). PIDsib is defined as the probability that two sibling individuals drawn at random from a population have the same mutilocus genotype (Waits et al., 2001).

To identify homonymous mislabeling, SNP genotypes of the clones with the same name were manually grouped and compared using multi-locus matching. If the clones differ by more than two loci, then these clones were considered to have different SNP genotypes thus were claimed as homonymous mislabeling.

To identify clones with mistake in species membership, assignment test based on Bayesian clustering analysis of population structure (Pritchard et al., 2000) was used. Clones with wrongfully assigned species membership were detected and corrected based on the assignment result (see the next section).

2.4 Population structure and genetic diversity

To assess population structure in the collection, we only used accessions with explicit species names recorded in the passport data. Clones recorded as 'Hybrid' or 'Unknown' were excluded in this stage, which led to the retention of 591clones for the population structure analysis. The computer program STRUCTURE ver. 2.3.4 (Pritchard et al., 2000) was used. The program was run at 10 independent repetitions for population numbers ranging from K = 1 to K = 10, with a burn-in period of 50,000 and 100,000 Markov chain Monte Carlo (MCMC). The optimal number of model components (K) was determined based on delta K (Evanno et al., 2005). Based on the result, iterative runs were performed on each partitioned cluster to explore the sub-structures within each cluster, as recommended by Evanno et al. (2005). Ancestry and admixture proportions were visualized using computer program CLUMPAK (Kopelman et al., 2015).

Based on the result of the STRUCTURE analysis, clones with the assignment coefficient above 0.80 (Q value >0.80) were considered core members of each cluster and retained for subsequent genetic diversity analysis, including computation of F statistics, Analysis of Molecular Variance (AMOVA), Principal Coordinate Analysis and Neighbor-Joining Clustering Analysis.

AMOVA was performed using the program GenAlex 6.5 (Peakall and Smouse, 2006; Peakall and Smouse, 2012). The significance of fixation index (F_{ST}) was tested using 999 random permutations. In addition, the F_{ST} for each pair of core germplasm groups was calculated and the statistical significance was tested using permutations with the program GenAlEx 6.5.

Key summary statistics including gene diversity (H_{exp}) and observed heterozygosity (H_{obs}) were calculated for each species using the program GenAlEx 6.5 (Peakall and Smouse, 2006; Peakall and Smouse, 2012). To illustrate genetic relationships among the species, a distance-based multivariate analysis was performed. Pairwise genetic distances were computed using the Distance option, and Principal Coordinates Analysis (PCoA) within the GenAlEx 6.5 program. The PCoA results are presented as two-axis PCO plots, and both plots axis 1 vs 2 and axis 1 vs 3 are presented separately.

To further examine the genetic relationship among different species, a neighbor-joining (NJ) clustering analysis was performed. The NJ tree was constructed based on the SNP genotype data using the R-package poppr (Kamvar et al., 2014). Pairwise genetic distances between the sugarcane clones were estimated and the neighbor-joining method was used to construct the tree. The final tree was visualized using FigTree version1.4.4 (http://tree.bio.ed.ac.uk/software/figtree).

To assess species membership and parentage for clones recorded as 'hybrid' or 'unknown', we used the core members of the five *Saccharum* species as references and increased their samples size to 500 for each species, using the Simulation procedure implemented in the computer program ONCOR (https://www.montana.edu/kalinowski/software/oncor.html). The simulated populations were then analyzed together with the 298 clones with unclarified species membership (clones labeled as 'unknown' or 'hybrid') using STRUCTURE 2.3.4. An admixed model was selected, and the number of clusters (K value) was set to five, corresponding to the five *Saccharum* species. Ten independent runs were conducted at K = 5, each consisting of 100,000 iterations after a burn-in period of 50,000 iterations. From the 10 independent runs, the mean membership score was presented as the inferred species/parentage or species membership.

3 Results

3.1 Genotyping with SNP markers

After applying an initial filtering process to exclude SNPs and samples with a missing rate of 10% or greater, a total of 751 SNP markers and 901 clones were obtained. Among the *Saccharum* clones, there were 286 *S. spontaneum*, 175 unknown, 158 *S. officinarum*, 123 hybrids, 76 *S. robustum*, 46 *S. sinense*, 32 *S. barberi*, 3 *S. edule*, 1 *S. narenga*, 1 *S. arundinaceum*, as recorded in passport data (Table S1). Of the SNPs, 217 (29%) were found to be monomorphic. To obtain informative SNP markers, we further filtered out SNPs with a minor allele frequency of less than 5%, resulting in a final set of 357 markers. The final genotype data for these markers showed an average missing rate of 0.15%, ranging from 0 to 7.6%. The sugarcane clones, on the other hand, showed an average missing rate of 0.15%, ranging from 0 to 6.2%. Among the 357 SNP markers, the expected heterozygosity ranged from 0.09 to 0.5, with an average of 0.23. The observed heterozygosity ranged from 0.09 to 0.99 with an average of 0.3. Additionally, the minor allele frequency ranged from 0.05 to 0.5 with an average of 0.153 (Table S2).

3.2 Clone mislabeling and genetic redundancy in the collection

Clonal propagation and field maintenance of sugarcane germplasm plants can often lead to mislabeling and name loss. The WCSRG, which houses collections from diverse locations worldwide, often encounters duplicated accessions with different regional names but identical clones. To estimate mislabeling and clonal redundancy, we measured the allele difference distance between each pair of sugarcane accessions at all SNP loci. Synonymous groups (duplicates) were determined when clones shared the same alleles at all SNP loci. Our analysis revealed 86 groups of duplicates comprising 211 accessions (23% of the examined *Saccharum* clones), demonstrating a high rate of synonymous mislabeling and genetic redundancy within the collection (Figure 2). The number of duplicated clones within each group ranged from two to nine clones, with 67 groups consisting of two clones (Table 1; Table S3).

To estimate genotyping error, twelve clonal samples propagated from a single clone of 'P-Mag-84-2' were included as an internal control. Of the 12 samples, however, only eight samples (Group1 in the Table S3) were identified as duplicates with zero allele difference, while three samples showed one allele difference, and one sample showed two allele differences. These differences were attributed to likely genotyping errors at four loci. Assuming no mutation in the clonal plants, these results indicated an error rate of 0.093% in our genotyping, as four loci were called wrongly out of a total of 4,284 loci in the 12 clonal samples. Based on the observed error rate, we relaxed the threshold of detecting duplicates. Any pair of samples that had up to two allele differences were considered putative duplicates. Based on this relaxed threshold, we identified 135 groups consisting of 471 clones (52% of the total clones) as putative duplicates. The number of duplicated clones within each group ranged from 2 to 38 (Table S4).

The result of duplicate identification was supported by the probability of identity among siblings (PID-sib). The cumulative PID-sib of the first 48 SNPs ranged from 3.85E-04 (*S. robustum*) to 6.73E-07 (*S. officinarum*), which demonstrated that a high level of statistical power can be achieved in sugarcane germplasm analysis using only a small fraction of the 357 SNP markers (Table S5; Figure S1). When all 357 SNP loci were included, the cumulative PID-sib

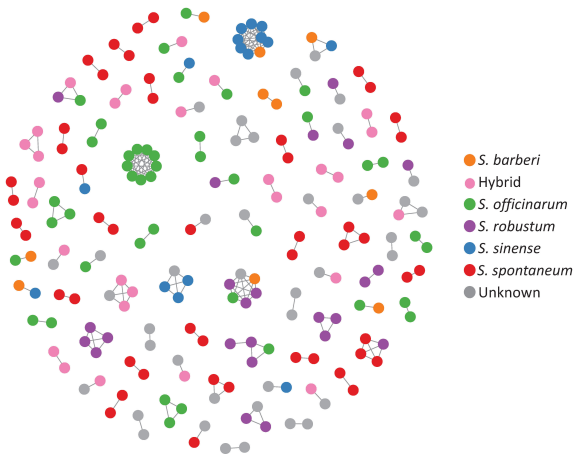


FIGURE 2
A network representing the genetic relationships among sugarcane clones based on allele difference at the SNP loci. The network was constructed with clones with zero allele difference (identical genotype) where nodes represent clones, and the connections indicate identical genotype between the clones.

TABLE 1 Examples of the identified synonymous mislabeling groups (duplicates) in the *Saccharum* germplasm maintained in the WCSRG.

Group	Clone code	Species	Clone Name	Source of introduction
2	SAC0472	*S. sinense*	UBA NAQUIN	US
	SAC0513	*S. sinense*	NEPAL 3	Nepal
	SAC0518	*S. sinense*	CHINA	South Africa
	SAC0519	*S. sinense*	AGAUL	South Africa
	SAC0531	*S. sinense*	TANZHOU	china
	SAC0616	*S. sinense*	MCILKRUM	US
	SAC0670	*S. barberi*	Kinar	India
	SAC0711	*S. sinense*	Uba Striped	Unknown
	SAC0760	*S. sinense*	Cayana 10	Unknown
3	SAC0321	*S. robustum*	IN 84-045	Barbados
	SAC0425	*S. barberi*	MESANGEN	Guyana
	SAC0426	*S. robustum*	IN 84-045	Barbados
	SAC0458	*S. officinarum*	HORNE	Barbados
	SAC0461	*S. robustum*	NG 28-251	Guadeloupe
	SAC0476	unknown	UNKNOWN	Unknown
4	SAC0323	*S. officinarum*	NG 28-014 (SS 58-08)	Papua New Guinea
	SAC0436	*S. robustum*	NG 57-208	US
	SAC0447	*S. robustum*	NG 57-208	US
	SAC1234	*S. robustum*	NG 57-208	US

The full list of identified duplicates and near-identical genotypes (clones differing by one or two alleles) was listed in Table S3 and S4.

ranged from 2.4E-20 (*S. robustum*) to 7.6E-42 (*S. officinarum*), which indicates that there is almost a null probability of finding two individual clones with the same genotype within any of the five *Saccharum* species.

To identify homonymous mislabeling, clones with the same name were compared for their SNP genotypes using multi-locus matching. A total of 27 homonymous mislabeling groups were detected in all the studied *Saccharum* species, except *S. edule* (Table 2). Most of the identified homonymous groups were collected from the same country and geographical region, indicating that mislabeling occurred before the germplasm were introduced into WCSRG.

3.3 Population structure and genetic diversity

3.3.1 Bayesian clustering analysis

The results of population structure analysis on 591 clones (with explicit passport records of species names) are presented in Figure 3. According to the delta K method (Evanno et al., 2005), the most probable number of genetically distinct groups (K) was estimated to be two (Figure 3A). At K = 2, the *S. spontaneum* clones were clearly classified as a distinct cluster, while the other five species were assigned to the second cluster (shown in orange in Figure 3B). It's noticeable that majority of the *S. sinense* clones had full population membership of the second cluster (in orange), whereas the majority of the *S. barberi* clones showed admixed genotypes between the first cluster (in blue) and the second cluster (in orange).

To further explore the substructure in the two clusters (*S. spontaneum* vs. the rest species), we repeated STRUCTURE analysis on each cluster using the same procedure and parameters. Through hierarchical analysis, the most probable number of genetically distinct groups (K) was two in the *S. spontaneum* cluster and four in the rest of the species (Figures 4A, C), based on Evanno's Delta K method. Therefore, the hierarchical partitioning classified the *Saccharum* germplasm into six sub-clusters: the *S. spontaneum* clones were classified into two populations, of which the first population was mainly originated from India and nearby countries in South Asia, whereas the second population were mainly originated from Southeast Asia and Barbados. Hereinafter, we used *S. spontaneum* (India) and *S. spontaneum* (SE Asia) to represent these two populations in subsequent analyses. The other four clusters correspond to four distinguishable species including 1) *S. barberi*, 2) *S. officinarum*, 3) *S. robustum*/*S. edule*, and 4) *S. sinense* (Figure 4D; Table S6).

This result is highly compatible with the current taxonomy framework of *Saccharum* (*sensu stricto*), which proposes six *Saccharum* species, with *S. edule* considered as a mutant of *S.*

10.3389/fpls.2023.1337736

TABLE 2 Identified homonymous mislabeling groups in the *Saccharum* germplasm maintained in the WCSRG.

Homonymous mislabeling group	Name	Code	Species	Origin
1	AGOULE	SAC0451	*S. sinense*	India Tamil Nadu
2	BA 11569	SAC0638	*S. officinarum*	Barbados
3	Chino	SAC0577	*S. officinarum*	Hawaii
4	CO 312	SAC0636	Hybrid	South Africa
5	CO 313	SAC0635	Hybrid	South Africa
6	CP 01-1372	SAC0552	Unknown	Florida
7	CP 91-555	SAC0653	Unknown	Louisiana
8	F 154	SAC0605	Hybrid	Taiwan
9	F31-762	SAC0608	Unknown	Hawaii
10	F36-819	SAC0299	*S. officinarum*	Hawaii
11	HC 71	SAC0511	*S. officinarum*	Hawaii
12	IJ 76-414	SAC0460	*S. robustum*	Barbados
13	IJ 76-478	SAC0659	*S. officinarum*	Indonesia
14	IJ 76-480	SAC0602	*S. robustum*	Barbados
15	IJ 76-547	SAC0322	*S. robustum*	Guadeloupe
16	IN 81-014	SAC0328	*S. robustum*	Barbados
17	Kerah	SAC0515	*S. sinense*	Indonesia Java
18	Longchuan (Yunan)	SAC1228	*S. spontaneum*	China
19	MESANGEN	SAC0514	*S. barberi*	Guyana
20	MOL 6077	SAC0481	*S. robustum*	Hawaii
21	MOL 6427	SAC0569	Unknown	Hawaii
22	N 26-14	SAC0522	Unknown	Unknown
23	NG 57-208	SAC1234	*S. robustum*	Hawaii
24	NG 57-238	SAC0598	*S. robustum*	Barbados
25	NG 77-094	SAC0396	*S. robustum*	Papua New Guinea
26	SES 519	SAC218	*S. spontaneum*	India
27	Tanzhou	SAC0646	*S. sinense*	China Guangxi

Each of the 27 accessions have at least one homonymous accession with different SNP genotype.

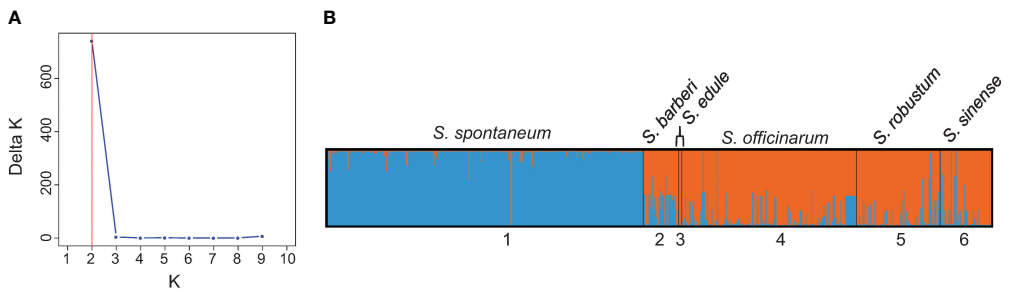


FIGURE 3
(A) Number of clusters based on the Evanno's Delta K value. **(B)** Population structure of the 591 *Saccharum* germplasm accessions with explicit passport record of species memberships partitioned using Structure v2.3.4. Black vertical lines indicate the separation of the *Saccharum* species. Multiple colors within the genetic group imply admixed individuals under the scenario of K=2.

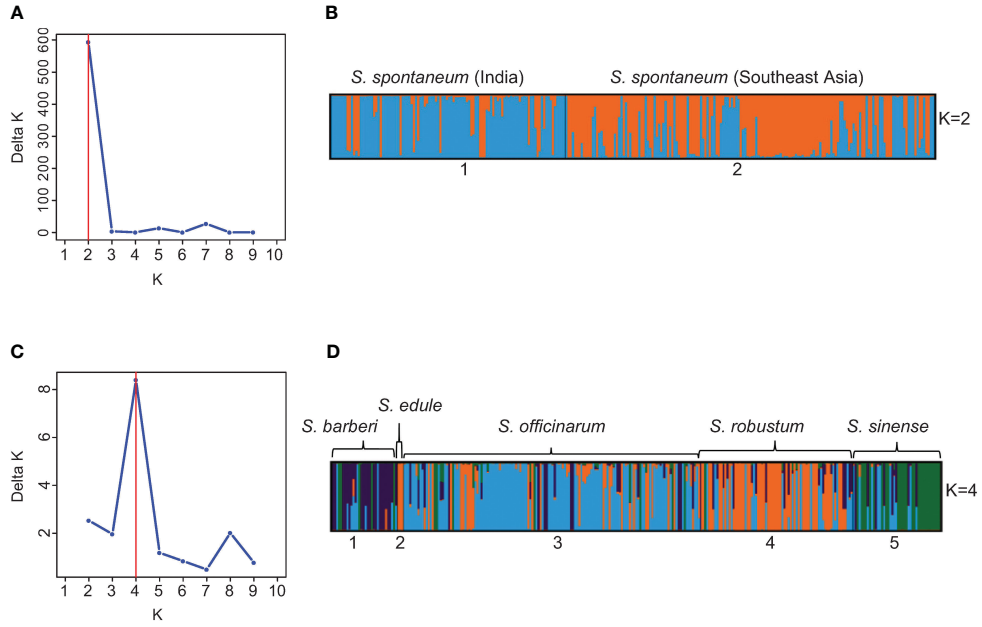


FIGURE 4
(A) Number of clusters in the 273 *S. spontaneum* clones based on the Evanno's Delta K value. **(B)** Partitioned result of the 273 *S. spontaneum* clones at K = 2 using Structure v2.3.4. **(C)** Number of clusters in 299 clones of *S. barberi, S. officinarum, S. robustum/S. edule*, and *S. sinense* based on the Evanno's Delta K value. **(D)** Population structure of the 299 clones of *S. barberi, S. officinarum, S. robustum/S. edule*, and *S. sinense* obtained using Structure v2.3.4. Multiple colors within the genetic group imply admixed individuals.

robustum (Daniels and Roach, 1987). Based on this result, we also observed that many clones had mislabeled species membership. In total, we detected 141 cases of mistakes or inaccuracies in species membership. The largest group was found in *S. officinarum* (65), followed by *S. robustum* (32), *S. sinense* (16) and *S. barberi* (12). In contrast, only one clone of *S. spontaneum* was found to have mislabeled species membership ("IN 84-072" from Indonesia), in addition to 10 clones of hybrids derived from *S. spontaneum* (Figures 4B, D; Table S6).

To further understand the genetic relationships among the five *Saccharum* species, we selected the core members of each species, based on the membership coefficient (Q-value) generated by the STRUCTURE analysis, with the threshold ≥ 0.80 (Table S6). This stringent cutoff enabled the inclusion of clones with minimal admixture. A total of 412 core members with unique genotypes were retained and each clone had a unique SNP genotype. These 412 core members were used in subsequent analysis of genetic diversity, including PCoA, Neighbor-Joining clustering analyses and AMOVA.

3.3.2 Principal Coordinates Analysis

The PCoA based on the Euclidean distance provided additional information on the relationships among the *Saccharum* germplasm clones (Figure 5). The first three principal coordinates accounted for 42.7% of the total variation, with the first, second and third coordinates explaining 21.7%, 13.0%, and 8.0%, respectively. Consistent with the findings from the STRUCTURE analysis, the core members of the five species were clearly separated from each other in both Figures 5A, B.

3.3.3 Neighbor-Joining clustering analysis

The Neighbor-Joining tree (Figure 6) revealed consistent results with STRUCTURE (Figure 4C) and PCoA analyses (Figure 5). There were two main clusters in the core members of the five species. Cluster 1 consisted of *S. officinarum*, *S. robustum*, *S. barberi*, and *S. sinense*, while Cluster 2 included the two populations of *S. spontaneum* from India and Southeast Asia. Within Cluster 1, *S. officinarum* and *S. robustum* were grouped together, showing their close relationship (Figure 6).

3.3.4 F statistics and Analysis of Molecular variance

The pattern of genetic differentiation between the five sugarcane species was also reflected by pairwise F_{ST} values, where higher F_{ST} values indicate greater genetic differentiation (Weir and Hill, 2002). The lowest pairwise F_{ST} values (0.073) was found between the two *S. spontaneum* populations (India vs. Southeast Asia), indicating a low level of differentiation. Among the five species, the pairwise F_{ST} ranged from 0.103 (*S. officinarum vs S. barberi*) to 0.323 (*S. robustum* vs *S. sinense*) (Table 3). The pairwise F_{ST} values generally align with the results from the PCoA plot (Figure 5) and the phylogenetic tree (Figure 6). All F_{ST} were highly significant ($P < 0.001$) based on permutation test. However, it's noteworthy that *S. sinense* was found to have the highest mean F_{ST} value (0.239), followed by *S. robustum* (0.224), *S. barberi* (0.178), *S. spontaneum* (0.176) and *S. officinarum* (0.123).

AMOVA was employed to assess the distribution of the observed genetic variance among the five sugarcane species, excluding hybrids and unknown species. The results showed that

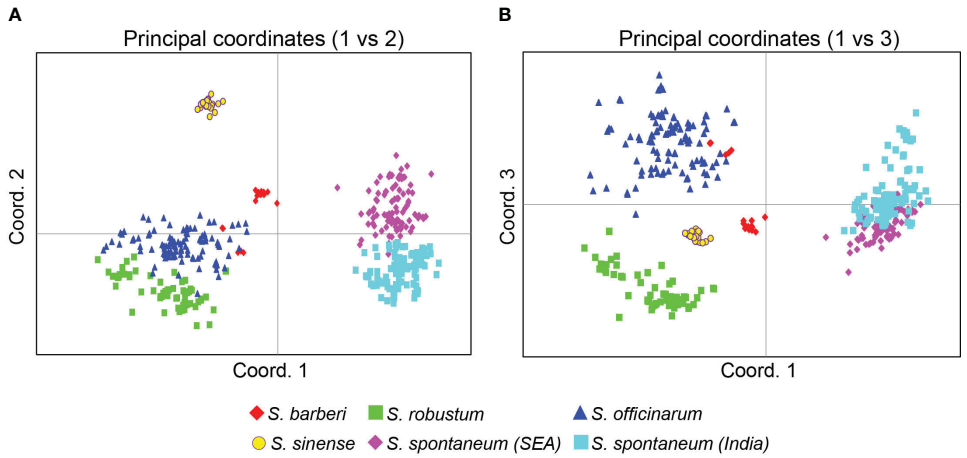


FIGURE 5
PCoA plots of the five *Saccharum* species represented by 412 core members with unique genotypes. **(A)** coordinates 1 vs 2 and **(B)** coordinates 2 vs 3. The analysis includes clones classified by STRUCTURE analysis with Q value >0.80. Each point represents an individual clone. The first three main axes accounted for the following percentages of the total variation: first axis = 21.7%, second axis = 13.0% and third axis = 8.0%.

a substantial proportion of the total genetic variation (85%) was attributed to within-species variation, while the remaining 15% of the total genetic variance was found in variation between species (Table 4). This shows that the observed genetic variations primarily arise from variation among individual clones within species rather than between different species. Out of the five species, *S. officinarum* has the highest intra-specific molecular variance (46.1), indicating its status as a cultivated hybrid species with extensive gene introgressions. In contrast, *S. robustum* had the lowest intra-specific molecular variance (21.8), suggesting its status as an ancient species with limited inter-specific gene flow. Intra-specific molecular variance in *S. spontaneum*, *S. barberi* and *S. sinense* are comparable. It's noticeable that within *S. spontaneum*, the population from Southeast Asia had higher molecular variance (35.0) than the population from India (30.8).

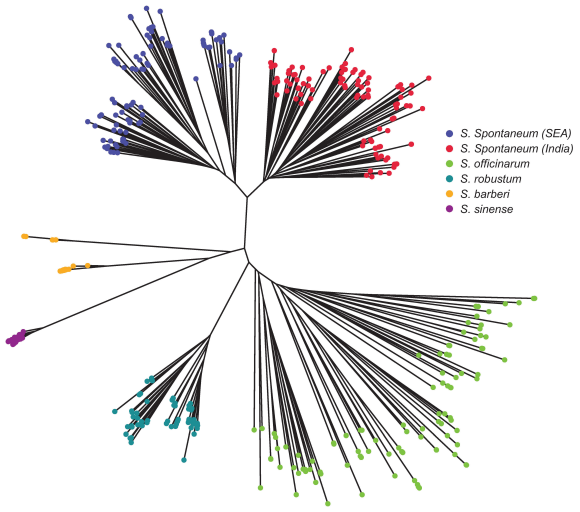


FIGURE 6
Neighbor-Joining tree depicting the relationships among the five *Saccharum* species represented by 412 core members with unique genotypes. The species *S. spontaneum* included two populations from India and Southeast Asia.

3.3.5 Observed heterozygosity and genetic diversity

S. spontaneum and *S. robustum* are considered as wild sugarcane species (Dinesh Babu et al., 2022). It is interesting to find that these two species exhibited relatively lower gene diversity and higher homogeneity compared to other cultivated species (Table 5). Notably, Zhang et al. (2018) also reported low gene diversity across 64 *S. spontaneum* accessions, based on genome sequencing data. They showed that the nucleotide diversity was much lower than that of other clonally propagated crops such as potato, cassava, grape, and citrus. Modern sugarcane cultivars have been extensively developed through interspecific crosses between *S. spontaneum* and *S. officinarum*. This disparity between wild and cultivated accessions suggests that the lower gene diversity observed in *S. spontaneum* is a characteristic of natural populations without human intervention. In contrast, the extensive intercrossing between species has likely contributed to the increased gene diversity and heterozygosity in hybrid cultivars.

3.4 Inferred species membership and parentage for clones with missing passport information

Of the 901 *Saccharum* (*sensu stricto*) accessions maintained in this sugarcane germplasm collection, there were 175 clones that do not have clear passport data for their species membership. These clones were recorded as 'unknown' in the collection. In addition, there were 123 clones recorded as 'Hybrid', but their parentage information was lacking. Using the selected core members of the five *Saccharum* species as references, we were able to assign the

TABLE 3 Pairwise population F_{ST} analysis among the five *Saccharum* species, represented by 412 core members with unique genotypes.

Species	*S. barberi*	*S. robustum*	*S. officinarum*	*S. sinense*	*S. spontaneum* (SE Asia)	*S. spontaneum* (India)
S. barberi	0.000	0.248	0.103	0.204	0.161	0.176
S. robustum	0.248	0.000	0.107	0.323	0.218	0.224
S. officinarum	0.103	0.107	0.000	0.168	0.119	0.120
S. sinense	0.204	0.323	0.168	0.000	0.216	0.286
S. spontaneum (SE Asia)	0.161	0.218	0.119	0.216	0.000	0.073
S. spontaneum (India)	0.176	0.224	0.120	0.286	0.073	0.000
	0.178	0.224	0.123	0.239	0.157	0.176

The probability, P (rand >= data) based on 999 permutations is shown above the diagonal.

species membership and parentage for all the clones that have ambiguous passport records. Our result showed that most of these clones have a species membership of *S. officinarum* (Table S8). Among the 175 clones recorded as 'unknown', two were assigned to *S. barberi*, 12 to *S. sinense*, 12 to *S. spontaneum*, 10 to *S. robustum*, 91 to *S. officinarum*, and 40 to inter-specific hybrids. Similarly, among the 123 clones that were recorded as 'Hybrid', two were assigned to *S. spontaneum*, one to *S. robustum*, 96 to *S. officinarum*, and 24 to inter-specific hybrid (Table S9).

4 Discussion

Genetic integrity is crucial for efficient conservation and use of plant germplasm for genetic improvement. This is particularly the case for many tropical/subtropical crops such as sugarcane, which is typically maintained in the field and propagated clonally, a process that can often result in mislabeling or loss of identifiers . Moreover, complex hybridization among sugarcane species further complicates the germplasm identification and highlights the need for comprehensive molecular and morphological characterization. However, accurate identification of sugarcane germplasm has been technically challenging, due to the high polyploidy (and aneuploidy) nature of this crop. In this study, we selected single-dose SNP markers and employed a targeted genotyping-by-sequencing (Seq-SNP) method to assess clone identity, population structure and genetic diversity in the *Saccharum* germplasm maintained in the WCSRG. This approach was chosen for its compatibility with sugarcane's diverse ploidy levels and large genome. The single-dose SNP markers offers advantages for polyploid plants such as sugarcane by enabling the identification and differentiation of alleles present across multiple sets of chromosomes without the need of ploidy determination (Sorrells, 1992; Aitken, 2022). By employing this high-throughput genotyping technique, we significantly improved our understanding of the genetic integrity and species relationship of sugarcane germplasm. The result provided valuable information to ensure the accuracy and efficiency in managing the sugarcane collection and facilitate its utilization in breeding programs.

4.1 Clone mislabeling and genetic redundancy

Using the selected SNP markers, we genotyped 901 *Saccharum* clones from six species including hybrids. Based on the SNP

TABLE 4 AMOVA and partitioning of total molecular variance within and among the five *Saccharum* species represented by 412 core members with unique genotypes.

Source	df	SS	MS	Est. Var.	%	P value
Among Pops	5	4310.9	862.2	6.4	15%	<0.001
Within Pops	818	28509.8	34.9	34.9	85%	
S. barberi	37	1257.5	34.0			
S. robustum	123	2682.6	21.8			
S. officinarum	207	9543.0	46.1			
S. sinense	55	2000.1	36.4			
S. spontaneum (S.E. Asia)	197	6898.5	35.0			
S. spontaneum (India)	199	6128.1	30.8			
Total	823	32820.7		41.2	100%	

*Probability, P (rand >= data), for F_{ST} is based on standard permutation across the full data set.

TABLE 5 Sample size (N), Observed heterozygosity (H_{obs}) and Gene diversity (H_{exp}) in the five *Saccharum* species represented by 412 core members.

Species	N	H_{obs}		H_{exp}	
		Mean	SE	Mean	SE
S. barberi	19	0.302	0.020	0.183	0.011
S. robustum	62	0.162	0.014	0.114	0.009
S. officinarum	104	0.354	0.014	0.256	0.008
S. sinense	28	0.389	0.025	0.199	0.013
S. spontaneum (SEA)	99	0.285	0.017	0.193	0.010
S. spontaneum (India)	100	0.238	0.015	0.169	0.009

genotypes, we discovered a high rate of clone mislabeling and genetic redundancy within the studied sugarcane collection. Because genotyping errors frequently occur, a pair of clones with a small number of mismatched loci could be duplicates as well (Zhang et al., 2006). Therefore, a threshold of mismatches to determine duplicates needs to be established. To evaluate the genotyping error rate, we included 12 samples propagated from the same clone as an internal control. Among the 12 samples, eight samples exhibited zero allele difference, while three samples showed one allele difference, and one sample showed two allele differences, suggesting that a two-allele difference could be used as the threshold for duplicate identification. Based on this threshold, we found that that half of the clones had at least one other clone with up to a two-allele difference (Table S4). We assessed how many SNP markers are needed to provide sufficient statistical power for sugarcane duplicate identification. Based on the cumulative PID-sib values for each species, we demonstrated that when utilizing 48 SNPs, the probability that two sibling individuals may share the same multilocus genotype by chance (Waits et al., 2001) was smaller than 0.001 (PIDsib <0.001) (Table S8; Figure S1). Therefore, the panel of 357 SNPs is far more than sufficient to identify synonymously mislabeled clones in each species (Table S5; Figure S1).

In addition to the detection of synonymous groups, the single-dose SNP genotyping enabled the identification of 27 homonymous mislabeling groups (Table 2), where clones with the same name had different SNP genotypes. These homonymous mislabeling groups were detected in all five *Saccharum* species, often in germplasm accessions collected from the same geographical region. For example, the three clones of "AGOULE" were all collected from Tamil Nadu, India. However, they exhibited two different SNP genotypes. In another case, two clones were labeled as "Chino", both from Hawaii, but their SNP genotypes were different. A more noteworthy example is the two "Uba" clones from India. Although they were both classified correctly as *S. sinense*, they had different SNP genotypes. Since "Uba" has been widely used as an important source of disease resistance in sugarcane breeding, the identified homonymous mislabeling has significant implications on sugarcane genetic studies and new variety development.

Furthermore, a high rate of mislabeling and inaccuracies was also detected in recorded species membership. Most of the mislabeling and inaccuracies occurred in species pairs that shared morphological similarities, such as *S. officinarum* vs *S. robustum* and *S. barberi* vs *S. sinense*. In contrast, there was almost no mislabeling of species membership between *S. spontaneum* and the rest of species. The high rate of mislabeling revealed in the present study was likely due to sugarcane's feature as a clonally propagated crop, which has allowed the exchange of sugarcane germplasm as clones among regions, countries, and continents. However, passport data, such as records and labels of the germplasm have not always followed the same naming conventions, leading to limited information about their correct identity. In fact, the majority of mislabeling and redundancy were observed in cultivated species, indicating a more intensive exchange of cultivated germplasm than wild species. Additional efforts of characterization are needed to fully resolve the mislabeling problem. SNP profiles for reference sugarcane clones need to be established through international collaboration. The putative mislabeled clones need to be compared with established references to correct the mislabeling errors. For the putative duplicate groups with near-identical genotypes, SNP genotyping will need to be repeated to confirm their clone identity. Moreover, since somaclonal mutation can occur in sugarcane, phenotypic examination remains essential to complement the result of molecular characterization.

4.2 Population structure and relationships among *saccharum* species

The sugarcane research community usually regarded *Saccharum* (*sensu stricto*) as containing six species, including two wild species (*S. spontaneum* and *S. robustum*), and four cultivated species - *S. officinarum*, *S. edule*, *S. barberi*, and *S. sinense* (Purseglove, 1972; Daniels and Roach, 1987; Grivet et al., 2006; Hemaprabha et al., 2022). The present study, using both model-based clustering and multivariant analysis based on 357 single dose SNP makers, supported the current classification of *Saccharum* germplasm. The only exception is *S. edule*, which could not be differentiated from *S. robustum*. *S. edule* is cultivated in New Guinea and nearby islands for its aborted edible inflorescences. Our result is consistent with the hypothesis that *S. edule* is a small group of sterile mutants that originated from *S. robustum* (Purseglove, 1972; Grivet et al., 2004; Grivet et al., 2006).

The present study also showed that *S. spontaneum* is a well-differentiated wild species, as shown by the analytical results of STRUCTURE (Figure 3), PCoA (Figure 5) and the NJ-tree (Figure 6). This observation is consistent with previous reports based on SSR markers (Brown et al., 2007), genome re-sequencing data (Yang et al., 2019; Fickett et al., 2020), and plastid genome sequences (Evans and Joshi, 2016). In the present study, we found very few mislabeling between *S. spontaneum* and the other four species. Nonetheless, the result of the hierarchical STRUCTURE analysis on *S. spontaneum* showed that there were two sub-groups within this species (Figure 4A, B). The first sub-group was mainly from India and south Asia, whereas the second sub-group was dominantly from Southeast Asia countries (Table S6). This result agrees with the recent report of Pompidor et al. (2021) and further indicates the importance of maintaining differentiated populations based on broader geographical regions.

A close genetic relationship was observed between *S. robustum* and *S. officinarum*. This observation is consistent with the recent finding of Pompidor et al. (2021), which suggested that both *S. officinarum* and *S. robustum* were derived from the same two ancestral genomes (A and B genomes), indicating a common origin of both species. Nonetheless, our result showed that the two species could be clearly differentiated at the molecular level, as demonstrated by the results of STRUCTURE (Figure 4B), PCoA (Figure 5) and NJ tree (Figure 6).

S. barberi and *S. sinense* are two cultivated species that are closely related (Purseglove, 1972; Lu et al., 1994; Hemaprabha et al., 2022). However, the taxonomy status, as well as the relationship between these two species, has been a subject of debate. It was proposed that *S. officinarum* hybridized with *S. spontaneum* in Asia continental and the hybrid progenies developed into *S. barberi* in India and *S. sinense* in China (Brandes, 1956; Daniels and Roach, 1987; D'Hont et al., 2002; Li et al., 2022). The geographical barrier (Southern China for *C. sinensis* vs. Northern India for *C. barberi*) could played significant role in the genetic differentiation of these two species. Using target enrichment sequencing of 307 germplasm accessions from WCSRG, Yang et al. (2019) showed that *S. sinense* and *S. barberi* were different in terms of genome compositions and potential ancestor accessions. Our result showed that *S. barberi* had the closest relationship with *S. officinarum*, which supported the proposal that *S. barberi* is a hybrid of *S. officinarum* and *S. spontaneum*. However, the present result also showed that relative to *S. barberi*, *S. sinense* had a larger genetic differentiation from *S. officinarum* and *S. spontaneum* (Table 3). Nonetheless, the number of samples of *S. barberi* and *S. sinense* used in the present study is relatively small. A systematic collection of samples representing the full geographical range of these two species is needed for a comprehensive analysis of population structure and genetic diversity in these two species.

To elucidate the underlying patterns of genetic variation in the sugarcane population, we conducted AMOVA to partition the observed variation among sugarcane clones. According to the AMOVA results, a significant proportion of the observed genetic diversity was attributed to variation among individual clones, accounting for 85% of the total variation, while the remaining 15% of the variation was attributed to variation between species (Table 4). These results are consistent with previous studies (Nayak et al., 2014; Manechini et al., 2018; Fickett et al., 2020; Singh et al., 2020), indicating that the primary source of genetic diversity resides within species rather than between different species. The relatively low percentage of genetic variation between species suggests that there may be a considerable level of introgression and systematic crossing occurring between different sugarcane species, reflecting intercrossing nature among sugarcane species (Moore et al., 2013). In all species, the observed heterozygosity was higher than the expected heterozygosity, which suggests more intercrosses between isolated populations than the founders. The inter-specific gene flow in sugarcane is well-supported by historical accounts that, throughout the seventeenth, eighteenth, and nineteen centuries, there was an extensive exchange of varieties among the sugarcane planters worldwide (Warner, 1962). The development of improved cultivars involved frequent intercrossing between the species, which also likely contributed to the low genetic variation between species. This is supported by the observed heterozygosity exceeding Hardy-Weinberg expectations, suggesting a greater number of intercrosses than expected from the founders alone.

In conclusion, accurate germplasm identity is critical for efficient management of sugarcane germplasm. Using single-dose SNP markers, we genotyped all the *Saccharum* clones in WCSRG, maintained at USDA-ARS. The single-dose SNP genotypes enabled us to detect a high rate of mislabeling and genetic redundancy in this collection. In addition, an analysis of population structure using both ordination and model-based clustering, revealed five genetic groups in the *Saccharum* germplasm, corresponding to *S. barberi, S. robustum, S. officinarum, S. sinense, and S. spontaneum*. The pattern of genetic structure in the *Saccharum* gene pool suggested a high level of gene flow among sugarcane species and across different geographical regions, likely facilitated by human intervention, as evident from the lower genetic variation observed between species. This extensive germplasm exchange, predominantly as clonal material, may contribute to the mislabeling and redundancy observed within the sugarcane collections. Through comprehensive analyses of genetic identity, we were able to detect genetic redundancy in the collection. Furthermore, we assessed the ancestral species/populations among the *Saccharum* germplasm clones and ascertained the presence of core members in each species. Using these core members as references, we were able to correct mistakes and/or inaccuracy in species membership, as well as clarify the parentage for hybrid clones. The corrected mislabeling in species membership needs to be validated based on phenotypic characteristics. Our results ensured that the preserved clones in the WCSRG have distinct genetic makeup. This is the first time that a large germplasm collection of sugarcane—a complicated polyploidy crop - was systematically characterized in terms of clone integrity and genetic redundancy. The single-dose SNP markers developed by this study offer a powerful tool for characterizing sugarcane germplasm worldwide. These markers can also be potentially used for identifying chromosomes. The reported findings have important implications for sugarcane genebank management, germplasm exchange, and crop genetic improvement.

Data availability statement

The genomic DNA sequencing data are accessible in the Sequence Read Archive (SRA) under bioproject ID PRJNA1051683 (http://www.ncbi.nlm.nih.gov/bioproject).

Author contributions

SP: Formal analysis, Investigation, Methodology, Software, Validation, Writing – original draft, Writing – review & editing. DZ: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Resources, Software, Supervision, Writing – original draft, Writing – review & editing. GA: Conceptualization, Data curation, Investigation, Methodology, Project administration, Resources, Supervision, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was funded by appropriated funds from the United States Department of Agriculture, Agricultural Research Service (USDA-ARS) (Project Number: 6038-21000-024-00D and 8042-21000-303-000-D.

Acknowledgments

We are grateful to Dr. Aliya Momotaz and Dr. Md. Sariful Islam, United States Department of Agriculture, Agriculture Research Service, Sugarcane Field Station, Canal Point, Florida, for critically reading the manuscript. Help of Barbara Freeman, Douglas DeStefano and Brandon Rodriguez with collecting leaf samples is appreciated. Mention of a trade name, proprietary product, or vendor does not constitute an endorsement, guarantee, or warranty by the USDA and does not imply its approval to the exclusion of other products or vendors that may be suitable.

Conflict of interest

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

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1337736/full#supplementary-material

References

Aitken, K. S. (2022). History and development of molecular markers for sugarcane breeding. *Sugar Tech* 24, 341–353. doi: 10.1007/s12355-021-01000-7

Akpertey, A., Padi, F. K., Meinhardt, L., and Zhang, D. (2021). Effectiveness of single nucleotide polymorphism markers in genotyping germplasm collections of coffea canephora using KASP assay. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.612593

Brandes, E. (1956). Origin, dispersal and use in breeding of the Melanesian garden sugarcane and their derivatives, Saccharum officinarum L. *Proc. Int. Soc. Sugar Cane Technol.* 9, 709–750.

Brown, J. S., Schnell, R. J., Power, E. J., Douglas, S. L., and Kuhn, D. N. (2007). Analysis of clonal germplasm from five Saccharum species: S. barberi, S. robustum, S. officinarum, S. sinense and S. spontaneum. A study of inter- and intra species relationships using microsatellite markers. *Genet. Resour. Crop Evol.* 54, 627–648. doi: 10.1007/s10722-006-0035-z

Daniels, J., and Roach, B. T. (1987). "Chapter 2 - taxonomy and evolution," in *Developments in crop science sugarcane improvement through breeding*. Ed. D. J. Heinz (New York: Elsevier), 7–84. doi: 10.1016/B978-0-444-42769-4.50007-2

Deren, C. W. (1995). Genetic base of U.S. Mainland sugarcane. *Crop Sci.* 35, cropsci1995.0011183X003500040047x. doi: 10.2135/cropsci1995.0011183X003500040047x

D'Hont, A., Grivet, L., Feldmann, P., Rao, S., Berding, N., and Glaszmann, J. C. (1996). Characterisation of the double genome structure of modern sugarcane cultivars (Saccharum spp.) by molecular cytogenetics. *Mol. Gen. Genet. MGG* 250, 405–413. doi: 10.1007/BF02174028

D'Hont, A., Paulet, F., and Glaszmann, J. C. (2002). Oligoclonal interspecific origin of 'North Indian' and 'Chinese' sugarcanes. *Chromosome Res.* 10, 253–262. doi: 10.1023/A:1015204424287

Dinesh Babu, K. S., Janakiraman, V., Palaniswamy, H., Kasirajan, L., Gomathi, R., and Ramkumar, T. R. (2022). A short review on sugarcane: its domestication, molecular manipulations and future perspectives. *Genet. Resour. Crop Evol.* 69, 2623–2643. doi: 10.1007/s10722-022-01430-6

Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol.* 14, 2611–2620. doi: 10.1111/j.1365-294X.2005.02553.x

Evans, D. L., and Joshi, S. V. (2016). Complete chloroplast genomes of Saccharum spontaneum, Saccharum officinarum and Miscanthus floridulus (Panicoideae: Andropogoneae) reveal the plastid view on sugarcane origins. *Syst. Biodivers.* 14, 548–571. doi: 10.1080/14772000.2016.1197336

FAO (2023). *Food and Agricultural data*. Available at: https://www.fao.org/faostat/en/#home (Accessed July 13, 2023).

Fickett, N. D., Ebrahimi, L., Parco, A. P., Gutierrez, A. V., Hale, A. L., Pontif, M. J., et al. (2020). An enriched sugarcane diversity panel for utilization in genetic improvement of sugarcane. *Sci. Rep.* 10, 13390. doi: 10.1038/s41598-020-70292-8

Formann, S., Hahn, A., Janke, L., Stinner, W., Sträuber, H., Logroño, W., et al. (2020). Beyond sugar and ethanol production: value generation opportunities through sugarcane residues. *Front. Energy Res.* 8. doi: 10.3389/fenrg.2020.579577

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

Granato, I. S. C., Galli, G., de Oliveira Couto, E. G., e Souza, M. B., Mendonça, L. F., and Fritsche-Neto, R. (2018). snpReady: a tool to assist breeders in genomic analysis. *Mol. Breed.* 38, 102. doi: 10.1007/s11032-018-0844-8

Grivet, L., Daniels, C., Glaszmann, J. C., and D'Hont, A. (2004). A review of recent molecular genetics evidence for sugarcane evolution and domestication. *Ethnobot. Res. Appl.* 2, 009. doi: 10.17348/era.2.0.9-17

Grivet, L., Glaszmann, J.-C., and D'Hont, A. (2006). "3. Molecular evidence of sugarcane evolution and domestication," in *3. Molecular evidence of sugarcane evolution and domestication* (New York: Columbia University Press), 49–66. doi: 10.7312/motl13316-004

Hale, A. L., Todd, J. R., Gravois, K. A., Mollov, D., Malapi-Wight, M., Momotaz, A., et al. (2022). Sugarcane breeding programs in the USA. *Sugar Tech* 24, 97–111. doi: 10.1007/s12355-021-01018-x

Hemaprabha, G., Pathy, T. L., Mohanraj, K., Alarmelu, S., and Ram, B. (2022). Population structure of coimbatore canes developed in a century of sugarcane breeding in India. *Sugar Tech* 24, 1449–1460. doi: 10.1007/s12355-021-01093-0

Hoang, N. V., Furtado, A., Botha, F. C., Simmons, B. A., and Henry, R. J. (2015). Potential for genetic improvement of sugarcane as a source of biomass for biofuels. *Front. Bioeng. Biotechnol.* 3. doi: 10.3389/fbioe.2015.00182

Jo, J., Kim, Y., Kim, G. W., Kwon, J.-K., and Kang, B.-C. (2021). Development of a panel of genotyping-in-thousands by sequencing in capsicum. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.769473

Kalinowski, S. T., Wagner, A. P., and Taper, M. L. (2006). ml-relate: a computer program for maximum likelihood estimation of relatedness and relationship. *Mol. Ecol. Notes* 6, 576–579. doi: 10.1111/j.1471-8286.2006.01256.x

Kamvar, Z. N., Tabima, J. F., and Grünwald, N. J. (2014). Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ* 2, e281. doi: 10.7717/peerj.281

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

Lam, E., Shine, J. Jr., Da Silva, J., Lawton, M., Bonos, S., Calvino, M., et al. (2009). Improving sugarcane for biofuel: engineering for an even better feedstock. *GCB Bioenergy* 1, 251–255. doi: 10.1111/j.1757-1707.2009.01016.x

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

Li, S., Duan, W., Zhao, J., Jing, Y., Feng, M., Kuang, B., et al. (2022). Comparative analysis of chloroplast genome in saccharum spp. and related members of 'Saccharum complex.'. *Int. J. Mol. Sci.* 23, 7661. doi: 10.3390/ijms23147661

Lu, Y. H., D'Hont, A., Walker, D. I. T., Rao, P. S., Feldmann, P., and Glaszmann, J. C. (1994). Relationships among ancestral species of sugarcane revealed with RFLP using single copy maize nuclear probes. *Euphytica* 78, 7–18. doi: 10.1007/BF00021393

Manechini, J. R. V., Costa, J. B., Pereira, B. T., Carlini-Garcia, L. A., Xavier, M. A., de A., M. G., et al. (2018). Unraveling the genetic structure of Brazilian commercial sugarcane cultivars through microsatellite markers. *PloS One* 13, e0195623. doi: 10.1371/journal.pone.0195623

Moore, P. H., Paterson, A. H., and Tew, T. (2013). "Sugarcane: the crop, the plant, and domestication," in *Sugarcane: physiology, biochemistry, and functional biology* (Hoboken, NJ: John Wiley & Sons, Ltd), 1–17. doi: 10.1002/9781118771280.ch1

Nayak, S. N., Song, J., Villa, A., Pathak, B., Ayala-Silva, T., Yang, X., et al. (2014). Promoting Utilization of Saccharum spp. Genetic Resources through Genetic Diversity Analysis and Core Collection Construction. *PloS One* 9, e110856. doi: 10.1371/journal.pone.0110856

Parco, A. S., Hale, A. L., Avellaneda, M. C., Hoy, J. W., Kimbeng, C. A., Pontif, M. J., et al. (2017). Distribution and frequency of Bru1, a major brown rust resistance gene, in the sugarcane world collection. *Plant Breed.* 136, 637–651. doi: 10.1111/pbr.12508

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

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

Piperidis, G., Piperidis, N., and D'Hont, A. (2010). Molecular cytogenetic investigation of chromosome composition and transmission in sugarcane. *Mol. Genet. Genomics* 284, 65–73. doi: 10.1007/s00438-010-0546-3

Pompidor, N., Charron, C., Hervouet, C., Bocs, S., Droc, G., Rivallan, R., et al. (2021). Three founding ancestral genomes involved in the origin of sugarcane. *Ann. Bot.* 127, 827–840. doi: 10.1093/aob/mcab008

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959. doi: 10.1093/genetics/155.2.945

Purseglove, J. (1972). *Tropical crops: monocotyledons* (London: Longman Group).

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

Singh, R. B., Mahenderakar, M. D., Jugran, A. K., Singh, R. K., and Srivastava, R. K. (2020). Assessing genetic diversity and population structure of sugarcane cultivars, progenitor species and genera using microsatellite (SSR) markers. *Gene* 753, 144800. doi: 10.1016/j.gene.2020.144800

Song, J., Yang, X., Resende, M. F. R.Jr., Neves, L. G., Todd, J., Zhang, J., et al. (2016). Natural allelic variations in highly polyploidy saccharum complex. *Front. Plant Sci.* 7. doi: 10.3389/fpls.2016.00804

Sorrells, M. E. (1992). Development and application of RFLPs in polyploids. *Crop Sci.* 32, cropsci1992.0011183X003200050003x. doi: 10.2135/cropsci1992.0011183X003200050003x

Spangler, R., Zaitchik, B., Russo, E., and Kellogg, E. (1999). Andropogoneae evolution and generic limits in sorghum (Poaceae) using ndhF sequences. *Syst. Bot.* 24, 267–281. doi: 10.2307/2419552

Waits, L. P., Luikart, G., and Taberlet, P. (2001). Estimating the probability of identity among genotypes in natural populations: cautions and guidelines. *Mol. Ecol.* 10, 249–256. doi: 10.1046/j.1365-294x.2001.01185.x

Wang, T., Fang, J., and Zhang, J. (2022). Advances in sugarcane genomics and genetics. *Sugar Tech* 24, 354–368. doi: 10.1007/s12355-021-01065-4

Warner, J. N. (1962). Sugar cane: an indigenous papuan cultigen. *Ethnology* 1, 405–411. doi: 10.2307/3772848

Weir, B. S., and Hill, W. G. (2002). Estimating F-statistics. *Annu. Rev. Genet.* 36, 721–750. doi: 10.1146/annurev.genet.36.050802.093940

Xiong, H., Chen, Y., Gao, S.-J., Pan, Y.-B., and Shi, A. (2022). Population Structure and Genetic Diversity Analysis in Sugarcane (Saccharum spp. hybrids) and Six Related Saccharum Species. *Agronomy* 12, 412. doi: 10.3390/agronomy12020412

Yang, X., Luo, Z., Todd, J., Sood, S., and Wang, J. (2020). Genome-wide association study of multiple yield traits in a diversity panel of polyploid sugarcane (Saccharum spp.). *Plant Genome* 13, e20006. doi: 10.1002/tpg2.20006

Yang, X., Song, J., Todd, J., Peng, Z., Paudel, D., Luo, Z., et al. (2019). Target enrichment sequencing of 307 germplasm accessions identified ancestry of ancient and modern hybrids and signatures of adaptation and selection in sugarcane (Saccharum spp.), a 'sweet' crop with 'bitter' genomes. *Plant Biotechnol. J.* 17, 488–498. doi: 10.1111/pbi.12992

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

Zhang, D., Mischke, S., Goenaga, R., Hemeida, A. A., and Saunders, J. A. (2006). Accuracy and reliability of high-throughput microsatellite genotyping for cacao clone identification. *Crop Sci.* 46, 2084–2092. doi: 10.2135/cropsci2006.01.0004

Zhang, D., Vega, F. E., Infante, F., Solano, W., Johnson, E. S., and Meinhardt, L. W. (2020). Accurate differentiation of green beans of arabica and robusta coffee using nanofluidic array of single nucleotide polymorphism (SNP) markers. *J. AOAC Int.* 103, 315–324. doi: 10.1093/jaocint/qsz002

Zhang, J., Zhang, X., Tang, H., Zhang, Q., Hua, X., Ma, X., et al. (2018). Allele-defined genome of the autopolyploid sugarcane Saccharum spontaneum L. *Nat. Genet.* 50, 1565–1573. doi: 10.1038/s41588-018-0237-2

Ziarsolo, P., Hasing, T., Hilario, R., Garcia-Carpintero, V., Blanca, J., Bombarely, A., et al. (2021). K-seq, an affordable, reliable, and open Klenow NGS-based genotyping technology. *Plant Methods* 17, 30. doi: 10.1186/s13007-021-00733-6

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 22 January 2024
DOI 10.3389/fpls.2023.1303398

OPEN ACCESS

EDITED BY
Nabanita Chattopadhyay,
The University of Texas Rio Grande Valley, United States

REVIEWED BY
Congting Ye,
Xiamen University, China
Xiaohui Wu,
Soochow University, Taiwan

*CORRESPONDENCE
Arthur G. Hunt
aghunt00@uky.edu

†PRESENT ADDRESS
Kai Li,
Department of Veterinary Science, University of Kentucky, Lexington, KY, United States

RECEIVED 27 September 2023
ACCEPTED 22 December 2023
PUBLISHED 22 January 2024

CITATION
Zhou L, Li K and Hunt AG (2024) Natural variation in the plant polyadenylation complex. *Front. Plant Sci.* 14:1303398. doi: 10.3389/fpls.2023.1303398

Natural variation in the plant polyadenylation complex

Lichun Zhou, Kai Li† and Arthur G. Hunt*

Department of Plant and Soil Sciences, University of Kentucky, Lexington, KY, United States

Messenger RNA polyadenylation, the process wherein the primary RNA polymerase II transcript is cleaved and a poly(A) tract added, is a key step in the expression of genes in plants. Moreover, it is a point at which gene expression may be regulated by determining the functionality of the mature mRNA. Polyadenylation is mediated by a complex (the polyadenylation complex, or PAC) that consists of between 15 and 20 subunits. While the general functioning of these subunits may be inferred by extending paradigms established in well-developed eukaryotic models, much remains to be learned about the roles of individual subunits in the regulation of polyadenylation in plants. To gain further insight into this, we conducted a survey of variability in the plant PAC. For this, we drew upon a database of naturally-occurring variation in numerous geographic isolates of *Arabidopsis thaliana*. For a subset of genes encoding PAC subunits, the patterns of variability included the occurrence of premature stop codons in some *Arabidopsis* accessions. These and other observations lead us to conclude that some genes purported to encode PAC subunits in *Arabidopsis* are actually pseudogenes, and that others may encode proteins with dispensable functions in the plant. Many subunits of the PAC showed patterns of variability that were consistent with their roles as essential proteins in the cell. Several other PAC subunits exhibit patterns of variability consistent with selection for new or altered function. We propose that these latter subunits participate in regulatory interactions important for differential usage of poly(A) sites.

KEYWORDS

alternative polyadenyaltion, natural variation, positive selection, *Arabidopsis* 1001 genomes, pseudogenes

1 Introduction

Messenger RNA polyadenylation process is an essential step for gene expression and regulation in eukaryotes (Edmonds, 2002). This process is mediated by a complex (the polyadenylation complex, or PAC) that consists of between 15 and 20 subunits (Boreikaitė and Passmore, 2023). These subunits function in the context of a set of subcomplexes, scaffolds, and enzymes - Cleavage and Polyadenylation Specificity Factor (CPSF), Cleavage stimulatory Factor (CstF), Cleavage Factor I (CFIm) and II (CFIIm), FIP1, symplekin, RBBP6, Poly(A) Polymerase (PAP), and Poly(A) Binding Protein-Nuclear (PABN). In

mammals, CPSF recognizes the conserved AAUAAA hexamer cis-element and cleaves the pre-mRNA at the polyadenylation site (Zhao et al., 1999; Chan et al., 2014; Schonemann et al., 2014; Clerici et al., 2018; Zhang et al., 2020). CstF recognizes sequences downstream of the mammalian poly(A) site and plays important roles in regulating alternative polyadenylation (APA) and in mediating between transcription and DNA repair (Pérez Cañadillas and Varani, 2003; Mirkin et al., 2008; Grozdanov et al., 2018; Yang et al., 2018; Zhang et al., 2020). CFIm recognizes the UGUA upstream cis-element and influences alternative poly(A) site selection, mRNA export and mRNA splicing (Martin et al., 2012; Zhu et al., 2018; Ghosh et al., 2022). Cleavage factor II (CFIIm) contributes to the recognition of cleavage/polyadenylation substrates through interaction with G-rich far-downstream sequence elements (Schäfer et al., 2018). CFIIm also plays roles in transcription termination (Sadowski et al., 2003; Zhang et al., 2005; West and Proudfoot, 2008; Kamieniarz-Gdula et al., 2019). FIP1 is a scaffold that links PAP with other parts of the PAC (Helmling et al., 2001; Kaufmann et al., 2004; Zhang et al., 2020; Muckenfuss et al., 2022). Symplekin and RBBP6 are additional scaffolds that coordinate the subcomplexes and enzymes in the course of the reaction (Ghazy et al., 2009; Kennedy et al., 2009; Xiang et al., 2010; Ruepp et al., 2011; Di Giammartino et al., 2014; Zhang et al., 2020; Rodríguez-Molina et al., 2022; Schmidt et al., 2022). PAP is the nucleotidyltransferase that adds the poly(A) tract to the 3' end of the cleaved pre-mRNA (Bard et al., 2000). In mammals, PABN regulates the length of poly (A) tail (Kerwitz et al., 2003; Kuhn and Wahle, 2004; Kühn et al., 2009).

With the possible exception of RBBP6 (discussed below), higher plants possess orthologs for the suite of core subunits of the mammalian and yeast PACs (Hunt et al., 2012). In plants, various PAC subunits have been implicated in important aspects of plant growth and development. CPSF30 is important in linking environmental signals and poly(A) regulation (Bruggeman et al., 2014; Hunt, 2014; Chakrabarti and Hunt, 2015). Both CPSF30 and FIP1 proteins participate in nitrate signaling and regulation (Li et al., 2017; Tellez-Robledo et al., 2019; Hou et al., 2021). In addition, FIP1 is important for plant response to stress and root development (Tellez-Robledo et al., 2019) and for seed dormancy (Li et al., 2023). CstF77 and CstF64 have been linked with the control of flowering time (Liu et al., 2010) and with responses to auxin (Zeng et al., 2019). One of the two *Arabidopsis* CFIm25 orthologs is important for maintaining the 3' UTR length in *Arabidopsis*, and mutation of this ortholog causes abnormal phenotypes (Zhang et al., 2022). CPSF73 plays roles in reproductive development in *Arabidopsis* (Xu et al., 2006). CPSF100 has functions in embryogenesis, seed production and root development (Lin et al., 2017). One Pcf11 ortholog, PCFS4, plays roles in the control of flowering time (Xing et al., 2008b). CLPS3 functions in embryo development (Xing et al., 2008a). Different PAP orthologs have been linked with the control of flowering time, defense responses, and aspects of gamete development and function (Vi et al., 2013; Trost et al., 2014; Kappel et al., 2015; Czesnick and Lenhard, 2016; Zhang et al., 2019; Ramming et al., 2023).

In plants, APA has been linked to numerous biological processes. For example, the choice of proximal and distal poly(A) site choice of transcripts encoded by the FCA gene, controlled by the core PAC subunit FY, determines the expression of FCA, a regulator of flowering time (Simpson et al., 2003). FY and FCA moreover cooperate to determine the usage of distal or proximal poly(A) sites associated with antisense transcripts that in turn regulate expression of FLC, a central regulator of flowering time in *Arabidopsis* (Whittaker and Dean, 2017). Usage of the poly(A) sites associated with antisense FLC transcripts is also linked with CstF77 and CstF64 (Simpson et al., 2003; Henderson et al., 2005; Liu et al., 2010; Whittaker and Dean, 2017). On a more global basis, poly(A) site choice varies genome-wide at different developmental stages in rice and *Arabidopsis* (Shen et al., 2011; Fu et al., 2016; Zhou et al., 2019). A large number of genes undergo APA in response to abiotic and biotic stress in sorghum (Chakrabarti et al., 2020), rice (Fu et al., 2016; Ye et al., 2019), *Populus trichocarpa* (Yan et al., 2021) and *Arabidopsis* (Hunt, 2014; de Lorenzo et al., 2017; Ma et al., 2022). Several plant PAC subunits have been implicated in the regulation of APA, including CstF77 (Zeng et al., 2019; Kim et al., 2023), CPSF30 (Liu et al., 2014), FIP1 (Tellez-Robledo et al., 2019), and FY (Yu et al., 2019).

While the impact of APA in plants is clear, much remains to be learned regarding the mechanisms that connect the PAC with environmental and developmental cues. Chief among the outstanding questions is that regarding the interactions of different PAC subunits with the larger gene regulatory network. One approach towards a better understanding of enzymes, complexes, and processes involves the assessment of naturally-occurring variability in the respective proteins (Alonso-Blanco et al., 2016; Hamm et al., 2019; Kadirjan-Kalbach et al., 2019; Zan and Carlborg, 2019). In this study, we compile and assess naturally-occurring variants in the subunits of the *Arabidopsis* PAC. Our results reinforce other studies that indicate essential roles for many core PAC subunits. In addition, they suggest that a subset of PAC subunits may be subject to diversifying selection, possibly indicative of functional specialization and roles in regulatory processes. Our results identify several genes as probable pseudogenes, thus tightening the focus of PAC subunits in *Arabidopsis* and answering questions about their absence in most other plants. Finally, we find that two evolutionarily-conserved PAC subunits, CstF50 and PAPS3, may not be essential in *Arabidopsis*, raising questions about their widespread conservation and possibilities about their roles in the PAC and in APA.

2 Methods

2.1 Plant growth and characterization

Four *Arabidopsis* strains (CS76822, CS76769, CS77397, CS7884) were ordered from the ABRC Stock center. Seeds were sown in soil, and grown in a temperature-controlled growth room at 22°C with a 16/8 hr light/dark cycle. After 20 days growth, leaves were collected and DNA was isolated using Plant DNAzol (Life Technologies) following the manufacturer's instructions. The

respective regions of interest were amplified by PCR using the primers listed in Supplementary File 7. PCR reactions consisted of: 0.25 ul Phire Hot Start II DNA Polymerase, 5 ul 5X phire reaction buffer, 2.5 ul 2.5 mM dNTP, 1 ul 10 uM forward primer, 1 ul 10 uM reverse primer, 1 ul of extracted DNA (concentration range 200-400 ng/ul), and 14.25 ul water. The cycle temperatures and durations were 95°C for 15 seconds, 55°C for 15 seconds, and 72°C for 30 seconds. PCRs were run for 25 cycles. PCR products were gel-purified using QIAquick Gel Extraction Kit as described in the user's manual. PCR products were sequenced by Eurofins Genomics; primers for sequencing reactions are indicated in Supplementary File 7. Sequencing results were aligned to the Col-0 reference sequences and displayed to confirm the homozygous nature of mutations; bioinformatics was conducted using various tools in the CLC Genomics Workbench package. After sampling for DNA, plants were grown until flowering, and then photographed.

2.2 Data collection and analyses

SNPs and variants that affect the protein coding regions (and not non-coding parts of genes such as promoters, untranslated regions, and introns) for the 31 genes that encode probable PAC subunit orthologs were downloaded from the *Arabidopsis* 1001 Genomes website using Polymorph 1001 tools; the list of genes is given in Table 1. The PCFS2 and SYM annotations in the *Arabidopsis* 1001 database were from an outdated annotation and were accordingly updated prior to data downloading. Specifically, PCFS2 was "formed" by merging the AT2G36485 and AT2G36480 annotations, and SYM by merging AT1G27590 and AT1G27595. *Arabidopsis* orthologs of the human RBBP6 were identified using BLASTP with the human RBBP6 as a query; this yielded two possible orthologs (denoted Mpe1 and PQT3 in Table 1 and elsewhere in this study).

The missense, silent mutations, nonsense mutations, and indels for each gene were tabulated and assembled into Supplementary File 1. These data were used to evaluate various features as described in the text. R studio software (data.table, dplyr, ggplot2 packages) was used to calculated the frequency for each PAC and draw Figures 1 and 2.

Analyses of synonymous and nonsynonymous substitutions were conducted using the Visualizing Variation (ViVa) analysis package (Hamm et al., 2019) run in R. This package extracts and compiles sequence variation information from the *Arabidopsis* 1001 Genomes database; included in the compilation are calculations of ratios of collective non-synonymous to synonymous diversity (π_N/π_S) for each protein-coding region. Details of the use of this package and of the π_N/π_S calculations may be found in Hamm et al. (Hamm et al., 2019).

To determine the expression of PAPS3-like genes in different plant species, PAPS-like genes in a select set of plant species were identified by BLASTP using the *Arabidopsis* PAPS3 protein sequence as a query; for this, plant species were chosen based on their presence in the EVOREPRO database (https://evorepro.sbs.ntu.edu.sg). These genes were used as queries to extract expression information in different tissues. Expression results were displayed after normalizing each sample such that the lowest expression level was equal to 1.

TABLE 1 A list of *Arabidopsis* polyadenylation complex subunits.

transcript id	gene name
AT1G13190.1	CFIm68-1
AT5G55670.1	CFIm68-2
AT4G29820.1	CFIS1
AT4G25550.1	CFIS2
AT3G04680.2	CLPS3
AT5G39930.1	CLPS5
AT5G23880.1	CPSF100
AT5G51660.1	CPSF160
AT1G30460.1	CPSF30
AT1G61010.3	CPSF73
AT5G60940.1	CSTF50
AT1G71800.1	CSTF64
AT1G17760.1	CSTF77
AT5G01400.1	ESP4
AT3G66652.1	FIPS3
AT5G58040.1	FIPS5
AT5G13480.1	FY
AT5G51120.2	PABN1
AT5G65260.1	PABN2
AT5G10350.1	PABN3
AT1G17980.1	PAPS1
AT2G25850.1	PAPS2
AT3G06560.1	PAPS3
AT4G32850.10	PAPS4
AT1G66500.1	PCFS1
AT2G36485.1 + AT2G36480.3	PCFS2
AT4G04885.1	PCFS4
AT5G43620.1	PCFS5
AT1G27590.1 + AT1G27595.1	SYM
AT4G17410	PQT3
AT5G47430	Mpe1

2.3 Genome re-assemblies

To reassemble and analyze the genomes of selected accessions, the respective short reads were downloaded from SRA (SRP056687). SRA accessions used in this study were SRR1946375 (for *Arabidopsis* accession 9812), SRR1945601 (accession 5984), SRR1946283 (accession 9705), and SRR1946188 (accession 9596). *De novo* assembly for each set of reads was done using the *De Novo* Assembly tool in the CLC Genomics Workbench (versions 20-23 were used in the course of this research), using the default parameters

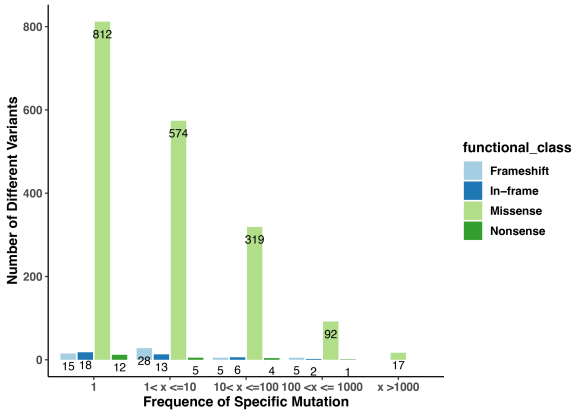


FIGURE 1
The numbers of variants in the collection of *Arabidopsis* ecotypes in the *Arabidopsis* 1001 Genomes database that affect genes encoding PAC subunits. The numbers on the plot indicate the counts of variants with frequencies falling within the specified ranges. The x-axis displays the different ranges, while the y-axis represents the counts of missense, nonsense, frameshift, and in-frame (insertion or deletion) mutations observed in each category. Distinct colors are used to denote four different functional classes.

(Automatic Word Size, Automatic Bubble Size, minimum contig length = 200 bp). The results of each assembly are given in Supplementary Files 4 and 5. The sets of contigs for each sample were used to create blast databases, which were then used in TBLASTN searches. TBLASTN searches were conducted using the respective Col-0 amino acid sequences.

2.4 PCFS1, PCFS5, and CLPS5 poly(A) site profiles

To confirm APA profiles for the PCFS1, PCFS5, and CLPS5 genes, 3' end profiling (Poly(A) Tag Sequences, or PATSeq) datasets from four published studies were downloaded and analyzed; for this, only Col-0 control data were used. These datasets were from studies of *Arabidopsis* mutants affected in various PAC subunits (CstF77 and CstF64 (Zeng et al., 2019), FY (Yu et al., 2019) and CPSF30 (Hou et al., 2021) and from a characterization of poly(A) site choice in plants subjected to hypoxic conditions (de Lorenzo et al., 2017). SRA accessions are given in Supplementary File 8. PATSeq reads were mapped to the *Arabidopsis* genome (TAIR10 reference) using the read mapping tool in the CLC Genomics Workbench package. For this, genomic regions adjacent to tracts of 8 or more A's were masked. The mapping parameters used were: Match score = 1, Mismatch cost = 2, Cost of insertions and deletions = Linear gap cost, Insertion cost = 3, Deletion cost = 3, Length fraction = 0.9, Similarity fraction = 0.9, Global alignment = No, Non-specific match handling = Map randomly, Execution mode = Standard, Minimum seed length = 15.

An additional Col-0 PATSeq dataset was generated for this study. Col-0 seeds were surface-sterilized by incubating in 70% ethanol for 1 min, followed by treatment with 10% bleach for 10 min, and then rinsed with distilled water five times. After the sterilization, seeds were suspended in a 0.1% agar solution and sowed onto ½ MS medium with 1% (w/v) sucrose, 0.8% (w/v) agar. Following stratification for 2 d in the dark at 4°C, plates were transferred to a growth chamber and incubated under long-day conditions (16 h light/8 h dark) at 22°C for 3 weeks. At this time, whole seedlings were removed, frozen in liquid nitrogen, the frozen tissue ground in a mortar and pestle, and RNA extracted using TRIzol RNA Isolation Reagents (Life Technologies) as recommended by the manufacturer. Short read sequencing libraries that query the mRNA-poly(A) junction (PAT-Seq libraries) were prepared as described previously (Ma et al., 2014; Pati et al., 2015). 1-3 µg of total RNA brought to 50 µL in 1X NEB RNA Fragmentation Buffer and incubated at 95°C for 2 min. Fragmented RNAs bearing poly(A) tracts were purified using the NEB Poly(A) RNA Isolation Kit and eluted in a final volume of 14 µL. The entire sample was then used as a template for SMART cDNA synthesis using Smartscribe (Takara); for this, the poly(A)-enriched

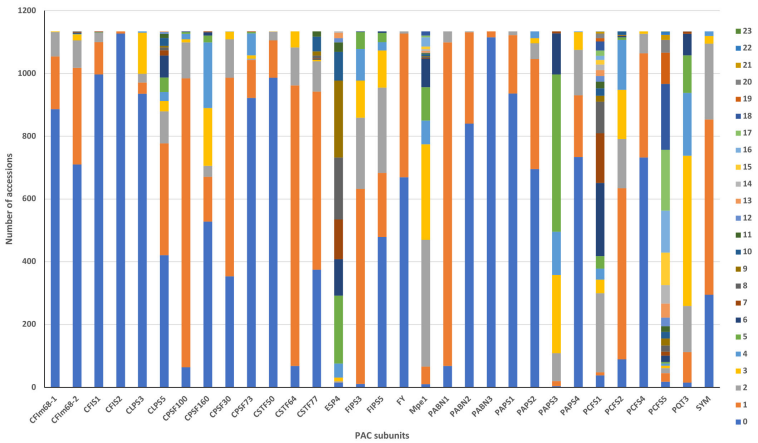


FIGURE 2
Summary of the numbers of ecotypes that possess missense changes in genes encoding PAC subunits. Colors denote the numbers of variants found in each accession, with the code indicated in the legend. The x axis displays the name of the PAC subunits, and the y axis represents the number of accessions bearing the indicated numbers of changes.

RNA was incubated in 25 µL of 1X RT buffer (prepared from the 5X stock provided by the manufacturer) containing 1 mM dNTPs, 1 mM DTT, 4 µM RT primer (see Supplementary File 7), and 1 µL of enzyme as supplied by the manufacturer. After 30 min at 42°C, 100 pmol of the strand-switching primer (SMART7.5; see Supplementary File 7) and an additional 1 µL of enzyme were added and reactions incubated for an additional 30 min at 42°C. After a subsequent incubation at 70°C for 5 min, 16.25 µL of SPRI beads (HighPrep PCR, Magbio Genomics, Inc.) was added, the solution completely mixed, and incubated for 8 min at room temperature. Beads were collected on a magnetic stand, washed twice with 100 µL 80% ethanol, air-dried for 5 min, and bound cDNA eluted with 25 µL water. 1 µL of the eluted cDNA was used for a limited PCR amplification using Phire Hot Start II DNA Polymerase (Thermo Fisher) and PE-PCR1 and PE-PCR2 primers. The cycle temperatures and durations were 95°C for 15 seconds, 60 °C for 15 seconds, and 72°C for 60 seconds. Reactions were run for 15 cycles. PCR products were separated on 1.5% agarose gels and products ranging between 300 and 500 bp excised and purified using a Qiagen gel purification kit. The gel-purified fragments were re-amplified using the same PCR conditions; PCR products at this point were purified using SPRI beads as described above. This final library was quantified using a Qubit and submitted for sequencing on an Illumina HiSeq2500 instrument at the University of Kentucky HealthCare Genomics Core Laboratory. PATSeq reads were analyzed using the pipeline described in the preceding paragraph and elsewhere (Thomas et al., 2012; Thomas, 2015; de Lorenzo et al., 2017). These sequencing data are available under Bioproject PRJNA1023006.

3 Results

3.1 Naturally-occurring sequence variation affecting the *Arabidopsis* polyadenylation complex – an overview

To study possible variability in the *Arabidopsis* PAC, genetic variants in a large collection of *Arabidopsis* accessions (Alonso-Blanco et al., 2016) that affect different subunits of the PAC were compiled and tabulated. The PAC subunits, notations used in this report, and corresponding gene identifiers are listed in Table 1. Earlier compilations of plant PAC subunits (Hunt et al., 2008; Hunt et al., 2012) lacked mention of possible orthologs of RBBP6/Mpe1, a scaffold protein that coordinates processing and polyadenylation activities of the mammalian and yeast complexes (Di Giammartino et al., 2014; Lee and Moore, 2014; Hill et al., 2019; Lee et al., 2020; Boreikaite et al., 2022; Rodríguez-Molina et al., 2022; Schmidt et al., 2022). For the sake of completeness, *Arabidopsis* RBBP6/Mpe1 orthologs were identified with BLASTP; this analysis yielded two possible counterparts, encoded by AT4G17410 and AT5G47430 (Supplementary Figure 1). One of these proteins (AT5G47430) is present in nuclear complexes containing CstF77 (Antosz et al., 2017); for this reason, these two proteins are included in this compilation and analysis. To facilitate subsequent analyses, the gene designations for these subunits that are in the ViVa (Hamm et al., 2019) database were retained.

1814 non-redundant missense SNPs were identified in genes encoding PAC subunits. Of these, 55% (1002/1814) were observed in at least two accessions, with 17 variants found in more than 1000 accessions (Figure 1). These 17 variants affected 10 of the 31 genes of interest (Table 2). Additionally, 22 nonsense mutations (affecting 9 of the 31 genes) and 53 frameshift variants (affecting 13 of the 31 genes) were found (Figure 1; Tables 3, 4; Supplementary File 1). Notably, one of the nonsense mutations, affecting the CLPS5 gene, occurred in almost half of the accessions (Table 3; Supplementary File 1). Specific frameshift mutations in three genes (PCFS5, PQT3, FIPS3) were observed in more than 100 accessions (Table 4). Seven genes with frameshift variants were also among those with nonsense variants (Tables 3, 4). These findings indicate that the 31 genes of interest exhibit distinct amino acid sequences in *Arabidopsis* strains, and some of them may lose function in specific strains.

All 1134 of the strains in the 1001 Genomes collection possessed variations (compared to the Col-0 reference) that affect the amino acid sequences of PAC subunits. The numbers of such variants in specific strains ranged from 9 (in Lan-0) to 106 (in IP-Vis-0) (Supplementary File 2). Many accessions had multiple missense variants in different PAC subunits; the range of variants in particular subunits ranged from 1 to 23 (Figure 2, Supplementary File 2). For 9 genes, the Col-0 reference sequence was the one seen in >70% of accessions (Figure 2; Supplementary File 2). For another 10 genes, either the Col-0 reference or a single amino acid substitution was seen in >70% of accessions (Figure 2; Supplementary File 2). For the remaining genes, the range and frequency of substitutions was broad.

A subset of genes showed a striking extent of variation, indicated by numerous accessions with multiple substitutions in each gene (Figure 2; Supplementary File 2); this subset consisted of the CSTF77, MPE1, CLPS5, ESP4, PAPS3, PCFS1, and PCFS5 genes. The scope of variation in PCFS5 was especially striking, with 86% of the accessions having more than 10 missense substitutions in this gene (Figure 2; Supplementary File 2).

3.2 Purifying and diversifying selection in *Arabidopsis* genes encoding PAC subunits

To further assess the variation affecting PAC subunits, the ratio of collective non-synonymous to synonymous diversity (π_N/π_S) for each gene was determined using the tool provided in the ViVa package (Hamm et al., 2019). Analogous to determinations of the rates of non-synonymous and synonymous substitutions, the π_N/π_S ratio derived from ViVa provides information about the overall conservation of amino acid sequence and consequently of functional diversity in the collection (Hughes, 1999; Hughes et al., 2000). Among the information is that concerning the tendencies towards purifying or diversifying evolution for specific genes. This tool has been shown useful in lending new and interesting insights into the nuclear auxin signaling pathway, identifying ARF members subjected to differing extents of purifying or diversifying evolution (Hamm et al., 2019). Demarcation of PAC subunits along these lines could be informative. Accordingly, the PAC-associated genes listed in Table 1 were analyzed using this tool.

10.3389/fpls.2023.1303398

TABLE 2 Missense mutations seen in more than 1000 accessions.

Gene_Name	Wild_Type	Position	SNP	Frequency
PCFS1	Asp	364	Ala	1077
PCFS1	Leu	217	Ser	1042
CSTF64	Phe	363	Tyr	1066
PAPS3	Ile	180	Ser	1100
PAPS3	Thr	297	Asn	1013
PAPS3	Leu	312	Arg	1098
FIPS3	Tyr	273	Asp	1111
ESP4	Asp	1398	Glu	1096
ESP4	Met	791	Lys	1096
ESP4	Gly	585	Val	1062
CPSF100	Ile	441	Val	1066
PCFS5	Val	115	Gly	1009
PCFS5	Asn	215	Ser	1013
PABN1	Lys	89	Glu	1065
PQT3	Ser	715	Pro	1050
Mpe1	Pro	809	Ser	1048
Mpe1	Thr	693	Pro	1044

As shown in Figure 3, the range of π_N/π_S ratios in PAC-associated genes ranged from 0.064 to 6.4. PAC-associated genes could be loosely divided into three groups (Figure 3; Supplementary File 3) – those with π_N/π_Sratios less than 0.8, those with ratios between 0.8 and 1.5, and with ratios greater than 1.5 (Figure 3; Supplementary File 3). The various PAC subcomplexes (CPSF, CstF, etc.) and other functional groups (scaffold proteins, poly(A) polymerases, PABNs) have members with low and high π_N/π_S ratios (Figure 3). All but one of the known essential PAC subunits have ratios less than 0.8. The exception (FY) has a ratio greater than 1.5 (Figure 3; Supplementary File 3). For this protein, the majority of missense mutations affect the C-terminus of the protein (Figure 4A; Supplementary Figure 2A).

Many PAC subunits are encoded by more than one gene. For several of these – PABN, CFIS, CFIm-68, FIPS, PAPS, and symplekin (SYM/ESP4) – one or more genes had ratios greater than 1.5 and others had π_N/π_S ratios less than 0.8 (Figure 3). For some pairs, the contrast between genes was striking. Specifically, for CFIS, CFIm68, FIPS, and SYM, one of the respective duplicate genes (CFIS2, CFIm68-1, FIPS5, and SYM) had low π_N/π_S ratios, while the other member of each duplicate set had ratios greater than 2 (Figure 3).

Low π_N/π_S may be reflective of purifying evolution and conservation of sequence and function, while high π_N/π_Sratios perhaps suggestive of a trend towards diversification. π_N/π_Sratios nearer 1 might reflect a more neutral mode of evolution, and thus of a protein not subject to strong selective pressures. Five of the set of genes associated with the PAC has this feature – PCFS1, PABN3, CstF77, CPSF30, and PAPS2. One of these, PCFS1, is a probable pseudogene (see the following). Two, CstF77 and CPSF30, are single-copy genes whose proteins have core functions in the PAC. However, these two genes are also non-essential (Zeng et al., 2019), a feature that may be related to their possible neutral evolution. PAPS2 is one of three nuclear PAPS isoforms in *Arabidopsis* and other plants. The *Arabidopsis* isoforms show a degree of functional specialization that may be attributed to the divergent C-termini of the proteins. The missense variants in these genes are largely clustered near the 3' ends of the respective coding regions (Figure 4B; Supplementary Figure 2B).

3.3 The distributions of nonsense and frameshift variants provide novel insights into the functions of several PAC-associated genes

Included in the variability that affects genes encoding PAC subunits are 22 nonsense mutations and 53 frameshift variants (insertions or deletions). These variants affect 13 genes (Tables 3, 4). In several of these genes, the changes fall near the C-termini of the corresponding coding regions, and likely do not affect the functionality of the respective gene (Supplementary Figure 3). Others, however, are predicted to have a large impact on gene functionality, due to severe truncations of the respective protein-coding regions (Supplementary Figure 3). Several of these affect members of small gene families; included in this set are genes encoding PABN1, PABN3, PQT3, CFIS1, ESP4, FIPS3, and PCFS2. Still others affect genes that are not members of families, or are

10.3389/fpls.2023.1303398

TABLE 3 Nonsense Mutations in genes encoding PAC subunits.

Gene_Name	Wild_Type	Position	Frequency
PCFS1	Gly	5	1
PAPS2	Arg	307	2
PAPS3	Trp	54	2
PAPS3	Glu	198	2
PAPS3	Leu	209	1
PAPS3	Tyr	492	13
PABN3	Glu	33	1
CLPS5	Trp	421	1
CLPS5	Gln	259	1
CLPS5	Arg	209	1
CLPS5	Arg	98	21
CLPS5	Gln	97	547
CLPS5	Gln	95	1
CLPS5	Arg	18	18
PCFS5	Tyr	83	9
PCFS5	Ser	95	1
PCFS5	Leu	99	1
PCFS5	Arg	281	1
PCFS5	Gln	284	3
FIPS5	Arg	1186	28
CSTF50	Arg	212	1
PQT3	Gln	453	1

unique to the *Arabidopsis* lineage. These latter genes – encoding CstF50, PAPS3, PCFS1, PCFS5, and CLPS5 – are interesting and provocative and are discussed in the following subsections.

3.3.1 The *Arabidopsis* CstF50 gene is not required for growth and development

One *Arabidopsis* accession (CS77397) had a premature stop codon within the CstF50 gene, and three others (CS78771, CS78772 and CS76987) had frameshift variants (Tables 3, 4; Supplementary File 1). The locations of these changes (Supplementary Figure 3) imply an inactivation of this gene in the respective accessions. This was unexpected, as CstF50 is essential in mammals and yeast and the *Arabidopsis* CstF50 gene (At5g60940) is a single copy gene. To confirm these suggestions, the CS77397 line was further characterized. Soil-grown plants had typical appearances, flowering behaviors, and fertility (Figure 5A; Supplementary Figure 4). The DNA sequence of the affected site was determined after PCR amplification and cloning. The results confirmed the presence of the mutation in a homozygous state (Figure 5B), with no suggestion of an additional copy of the gene that might encode a wild-type copy of the gene. To test the possibility that the CstF50

TABLE 4 Frameshift Variants in genes encoding PAC subunits.

gene_name	sequence_change	Frequency
CFIS1	p.Ser14_Asp15fs/c.42_43insC	6
CFIS1	p.Ser14_Asp15fs/c.40_41insAT	4
CLPS5	p.Tyr401fs/c.1202_1203delAT	5
CLPS5	p.Ile399fs/c.1195delA	12
CLPS5	p.Arg398_Ile399fs/c.1194_1195ins	1
CLPS5	p.Val322_Lys323fs/c.966_967insT	1
CLPS5	p.Gly238fs/c.713delG	12
CLPS5	p.Val235fs/c.705_708delTCGT	1
CLPS5	p.Lys182_Ala183fs/c.545_546insA	14
CLPS5	p.Phe147_Val148fs/c.441_442insAA	4
CLPS5	p.Lys140fs/c.420_432delAGATGGTTG	1
CLPS5	p.Ser79fs/c.235delT	1
CLPS5	p.Ala36fs/c.106delG	1
CLPS5	p.Glu16_Leu17fs/c.46_47insCG	10
CLPS5	p.Gly3fs/c.9delT	3
CSTF50	p.Ser120fs/c.358delT	2
CSTF50	p.Val249fs/c.746_753delTAAACACA	1
ESP4	p.Gly585fs/c.1754_1755delGG	2
ESP4	p.Asp584fs/c.1751delA	3
FIPS3	c.2993_2994insA	807
PABN1	p.Glu124fs/c.371delA	1
PAPS2	p.Asn697_Glu698fs/c.2090_2091ins	6
PAPS3	p.Asp36fs/c.108delT	2
PAPS3	p.Ile76_Leu77fs/c.226_227insA	3
PAPS3	p.Asp108_Phe109fs/c.324_325insT	2
PAPS3	p.Asn221_Gly222fs/c.661_662insA	3
PAPS3	p.Phe430fs/c.1290delC	1
PAPS3	p.Leu441fs/c.1321_1325delCTTGT	3
PAPS3	p.Lys460fs/c.1378delA	20
PAPS4	p.His739fs/c.2217_2218delTG	2
PCFS1	p.Thr168fs/c.504_505delTC	1
PCFS1	p.Ser142fs/c.424delT	9
PCFS1	p.Gly118_Asn119fs/c.352_353insA	2
PCFS1	p.Ser80fs/c.240_241delTC	1
PCFS2	p.Thr876fs/c.2626delA	1
PCFS2	p.His866fs/c.2597delA	9
PCFS2	p.Ser141_Cys142fs/c.421_422insT	1
PCFS5	p.Asp89fs/c.267_268delTG	15
PCFS5	p.Ala98_Leu99fs/c.294_295insT	3

(Continued)

TABLE 4 Continued

gene_name	sequence_change	Frequency
PCFS5	p.Asn118fs/c.354delC	5
PCFS5	p.Asn175fs/c.525_526delCA	129
PCFS5	p.Met176fs/c.528_534delGGTTTCA	246
PCFS5	p.Asn226fs/c.678delT	4
PCFS5	p.Ile238fs/c.713delT	7
PCFS5	p.Gln331fs/c.993_994delAC	3
PCFS5	p.Val334_Pro335fs/c.1000_1001ins	8
PCFS5	p.Ala344fs/c.1031delC	6
PCFS5	p.Leu345fs/c.1033delT	5
PQT3	p.Trp418_Ala419fs/c.1252_1253ins	430
PQT3	p.Trp418_Ala419fs/c.1253_1254ins	311
PQT3	p.Glu664fs/c.1990_1991delGA	1
PQT3	p.Arg665fs/c.1995delT	1
PAPS3	p.Cys507fs/c.1519_1528delTGTTAGG	10

gene in this line has been duplicated, the raw re-sequencing data for this accession were re-assembled and the assembly searched to identify all contigs that may possess CstF50-related sequences. This exercise yielded a single contig that could encode a polypeptide with substantial identity to CstF50 (Supplementary File 4). While this experiment does not rule out large-scale (chromosome-sized) structural variants, it does indicate that there are no additional CstF50 genes that lack a stop codon in this accession (CS77397). These results indicate that CstF50 is not essential for *Arabidopsis* growth and development.

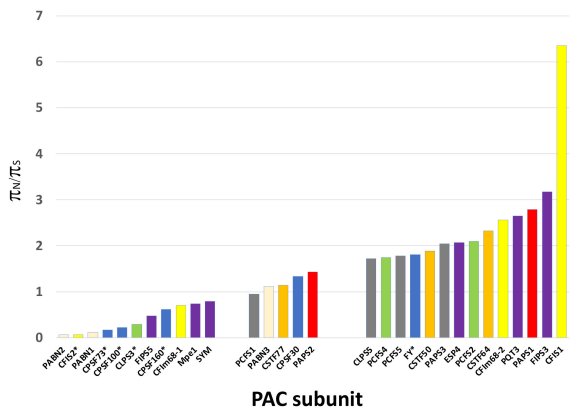


FIGURE 3
Results of the Visualizing Variation (ViVa) analysis. The π_N/π_S ratio for each gene is plotted as shown. Genes are grouped according to their π_N/π_S values - π_N/π_S ratios less than 0.8, ratios between 0.8 and 1.5, and with ratios greater than 1.5. The y axis represents the π_N/π_S value. Stars indicate genes that are essential. Data for this plot is provided in Supplementary File 3.

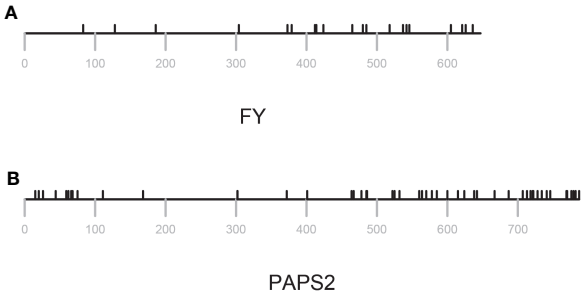


FIGURE 4
Distribution of missense changes in FY and PAPS2. Locations of variants are indicated by black tics above the lines that represent the amino acid sequence. Position numberings are shown with gray tics beneath the sequence representation. **(A)** FY **(B)** PAPS2.

3.3.2 Three genes that encode putative CFIIm subunits are pseudogenes

The canonical mammalian factor CFIIm consists of two subunits, Pcf11 and Clp1. *Arabidopsis* possesses four possible Pcf11-encoding genes and two Clp1 genes [termed as PCFS and CLPS in this report, as suggested by others (Hunt et al., 2008; Hunt et al., 2012)]. Six nonsense and 18 frameshift mutations affecting three of the PCFS genes were found in the collection of *Arabidopsis* accessions (Tables 3, 4; Supplementary Figure 3). These mutations occur in a large number of accessions. Specifically, fifteen accessions contain premature termination codons in PCFS5 and one accession has a premature termination codon in PCFS1 (Table 3). However, no accessions possess premature termination codons in both PCFS1 and PCFS5. Numerous other accessions possess frameshifts in either PCFS1 or PCFS5 (but not both; Table 4; Supplementary Figure 3).

These two genes are distinctive in other ways. As noted above (Figure 3; Supplementary File 1), there is extensive missense variation in these two genes (109 missense found in PCSF1, and 123 missense mutations found in PCFS5). The predicted polypeptides lack important functional domains that are seen in the other PCFS orthologs (PCFS2 and PCFS4; Figure 6A). These observations raise the possibility that these two genes may not be functional, even in accessions with no clear debilitating changes. Other reports and data support this conclusion. The PCFS1 gene was among those noted in an earlier study as being affected by APA, with a majority of mRNAs encoded by this gene ending well within the protein coding region of the gene (Parker et al., 2021). Such APA products would lack translation termination codons and thus would be substrates for non-stop RNA decay. To confirm that this is the case, different poly(A) site-profiling datasets were analyzed. These data sets include four published ones as well as one independently-generated, hitherto unpublished set of data (see Methods). The results showed that, in every dataset analyzed, a large majority of PCFS1-encoded RNA isoforms end within the protein-coding region of the gene (Figure 6B). Similar results were seen in mappings of reads to the PCFS5 gene (Figure 6C). These results indicate that most PCFS1- and PCFS5- encoding transcripts are non-stop RNAs. These collective features – the large numbers of missense variants, the occurrence of premature termination codons

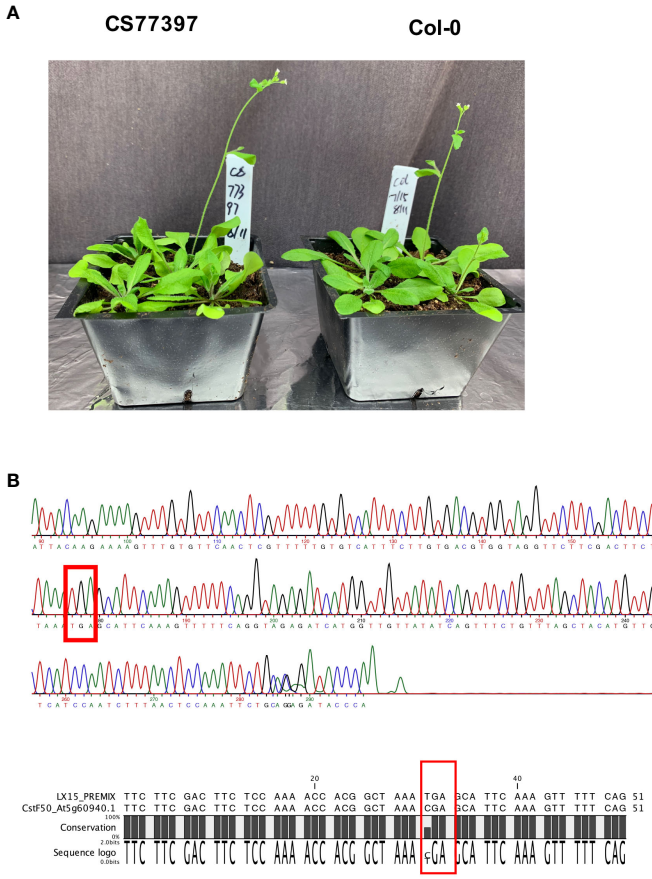


FIGURE 5
Characterization of *Arabidopsis* line CS77397. **(A)** Photograph of soil-grown CS77397 and Col-0 plants. **(B)** Sanger sequencing trace of the region encompassing the nonsense variant seen in CS77397. The location of the change is highlighted with a red box. The alignment beneath the trace shows a comparison of the Col-0 and CS77397 sequences, confirming the alteration that is noted in the *Arabidopsis* 1001 Genomes database.

and frameshift variants, and the prominence of nonstop RNAs encoded by these two genes – strongly suggest that these two *Arabidopsis* genes are not functional, and probable pseudogenes.

Most plants have single genes that encode the Clp1 ortholog, but *Arabidopsis* possesses two such genes, termed CLPS3 and CLPS5 (Hunt et al., 2012). The CLPS3 gene is orthologous to ones that are ubiquitous in plants. This gene is essential (Xing et al., 2008a) and exhibits a very small π_N/π_S ratio (Figure 3). In contrast, the CLPS5 gene seen only in the *Arabidopsis* lineage (Hunt et al., 2012). In the collection of *Arabidopsis* accessions, seven nonsense mutations were found in CLPS5 genes. One nonsense variant (Q97*) was seen in 547 lines (Table 3). The expression level of the CLPS5 gene in *Arabidopsis* is very low (Figure 6D). Moreover, the *Arabidopsis* CLPS5 is not essential (Xing et al., 2008a). Together, these results suggest that, as with the PCFS1 and PCFS5 genes, CLPS5 is a pseudogene.

3.3.3 PAPS3 – a novel plant poly(A) polymerase borne of paradoxes

Plants possess a conserved set of poly(A) polymerase isoforms, typified by the *Arabidopsis* PAPS1, PAPS2, PAPS3, and PAPS4 proteins (Hunt et al., 2008; Meeks et al., 2009; Hunt et al., 2012; Trost et al., 2014; Kappel et al., 2015; Czesnick and Lenhard, 2016; Zhang et al., 2019). PAPS1, PAPS2, and PAPS4 are all nucleus-localized proteins that play roles in poly(A) tail length control as related to aspects of plant growth and development (Vi et al., 2013; Trost et al., 2014; Kappel et al., 2015; Czesnick and Lenhard, 2016; Zhang et al., 2019; Ramming et al., 2023). These proteins, while related, are functionally-specialized, with specific roles attributed to novel C-terminal domains (Czesnick and Lenhard, 2016). Consistent with this, most of the missense changes in these proteins lie within the respective C-termini (Figure 4B; Supplementary Figures 2B, 5); this distribution helps to explain the elevated π_N/π_S ratios seen with PAPS1 and PAPS2 (Figure 3; note that PAPS4 could not be analyzed using the ViVa tool).

In the AtPAPS3 gene, premature stop codons can be found at four different locations in the collection of *Arabidopsis* accessions (Figure 7A); these variations are seen (collectively) in 18 different ecotypes (Table 3). These stop codons are predicted to severely truncate the encoded proteins and would be null mutations. This was unexpected, as it had been reported that other *Arabidopsis* mutants with PAPS3 null mutations were not viable (Meeks et al., 2009). To explore this, three of these accessions (CS76822, CS76769, CS78841) were grown and characterized. All three accessions had normal growth habits, flowering behaviors, and fertility (Figure 7B; Supplementary Figure 4). The DNA sequences of the affected sites

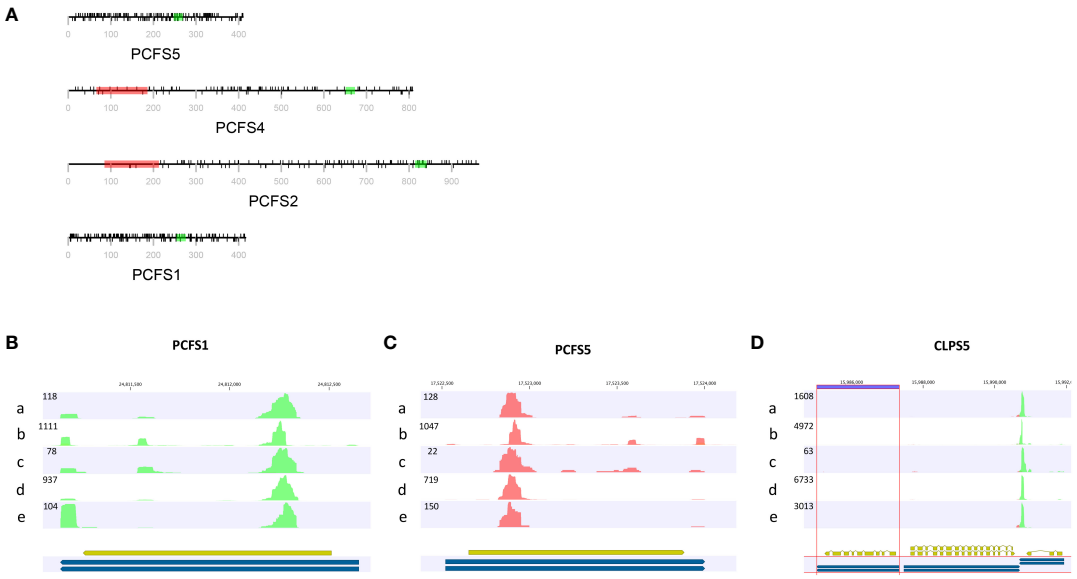


FIGURE 6
Feature of genes encoding CFIIm subunits. **(A)** The domain structures of *Arabidopsis* Pcf11 orthologs (PCFS). Red rectangles in represent the CID domain, green rectangles the zinc finger motif. The black lines above the protein representation are missense mutations, and those below are silent mutations. The grey line and text represent the amino acid positions and overall polypeptide length. **(B–D)** PATSeq mapping tracks showing the 3′ ends (as well as overall expression levels) of the PCFS1, PCSF5, and CLPS5 genes. Tracks a-d were derived from a re-analysis of four different PATSeq experiments as described in Methods and in Supplementary File 8. Track e was generated in this study as described in Methods. Blue and gold bars beneath each track show the respective gene annotations. Numbers above each track show the respective chromosome positions. Numbers at the upper left of each track denote the sizes (mapped reads) of the largest peaks. Green reads indicate those that read right-to-left (5′-3′) and red reads those that read left-to-right (5′-3′). Poly(A) sites are the left-most (for green reads) or right-most (for red reads) parts of each peak.

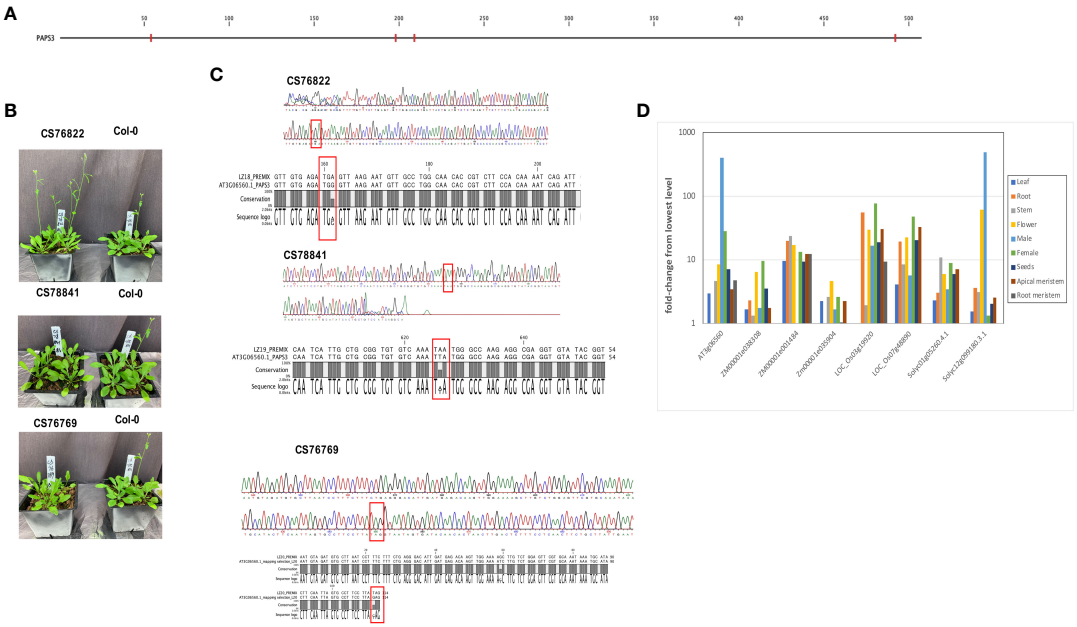


FIGURE 7
Characterization of ecotypes with nonsense mutations in the PAPS3 gene. **(A)** Locations of pre-mature stop codons seen the *Arabidopsis* 1001 Genomes collection of *Arabidopsis* ecotypes. Nonsense mutations are denoted with red tics. **(B)** Appearances of three ecotypes that harbor nonsense mutations in their respective PAPS3 genes. **(C)** Sanger sequencing traces obtained from DNA isolated from the CS76822, CS78841 and CS76769 ecotypes. Regions from the respective PAPS3 genes that include the stop codons are shown. Pre-mature stop codons are highlighted with red rectangles. The alignments beneath the traces show comparisons of the Col-0 and different ecotype sequences, confirming the alterations that are noted in the *Arabidopsis* 1001 Genomes database. **(D)** Expression characteristics of PAPS3-like genes in different plant species. Normalized expression of the indicated genes in different tissues were determined using data downloaded from the ENDOREPRO database as described in Methods. Normalized values were plotted as shown. Descriptions of the gene designations and the raw data for the plot is provided in Supplementary File 6.

were subsequently determined after PCR amplification and cloning. The results confirmed the presence of the mutation in a homozygous state (Figure 7C). To test the possibility that the PAPS3 gene in these lines had been duplicated, the raw re-sequencing data for these accessions were re-assembled and the assembly searched to identify all contigs that may possess PAPS3-related sequences. This exercise yielded a single contig for each accession (Supplementary File 5). These results indicate that there are no additional PAPS3 genes that lack stop codons in these accessions. Several accessions also possess frameshift mutations in the PAPS3 gene - see Table 4; Supplementary File 1, and Supplementary Figure 3. The locations of many of these would dramatically truncate the protein. Given the results seen with the CS76822, CS76769, CS78841 accessions, none of these additional frameshift lines further analyzed.

An earlier bioinformatics analysis reported that the expression of the *Arabidopsis* PAPS3 gene was pollen-specific (Hunt et al., 2008), suggestive of a novel role for this protein in male gametophyte development. Given the features of the PAPS3 gene seen in the collection of *Arabidopsis* accessions, this issue was revisited. For this, PAPS-like genes in a select set of plant species were identified by BLASTP and their expression characteristics compared. The results corroborated the earlier report (Hunt et al., 2008) showing a strong pollen-specific expression of the *Arabidopsis* PAPS3 in pollen (Figure 7D; Supplementary File 6). One of the two *S. lycopersicum* PAPS3 isoforms (*Solyc12g099180.3.1*) also showed a strong preferential expression in male reproductive tissues as well as in flowers (Figure 7D; Supplementary File 6). However, the other *S. lycopersicum* PAPS3 isoform, as well as five PAPS3 isoforms present in the maize and rice genomes, did not exhibit strong tissue-specificity in their expression characteristics (Figure 7D; Supplementary File 6). Therefore, the novel tissue specific expression of the *Arabidopsis* PAPS3 gene is not a universal feature of PAPS3 genes in plants; rather, different genes exhibit different expression properties.

4 Discussion

4.1 Conservation and variation in subunits of the plant polyadenylation complex – general considerations

For the most part, the patterns of sequence variation that are seen in the *Arabidopsis* genes that encode PAC subunits are consistent with the functioning of these subunits in a fundamental step in gene expression. With the interesting exceptions discussed in the following subsections, the paucity of nonsense and frameshift mutations that would eliminate gene function also supports important roles for these proteins. In some cases, residues seen in most *Arabidopsis* accessions but absent in Col-0 are identical to residues seen in most other plants. This reinforces the point that the Col-0 reference sequence is not necessarily the universal one for *Arabidopsis* or plants in general.

The ratios of non-synonymous to synonymous substitutions (π_N/π_S) in PAC-associated genes are interesting. As explained in Hamm et al. (Hamm et al., 2019), the π_N/π_S ratio is a useful proxy for the K_a/K_s metric; thus, low π_N/π_S ratios may be taken as evidence for purifying selection, and high π_N/π_S ratios for positive (diversifying) selection. Many PAC-associated genes exhibit high π_N/π_S ratios. This suggests that many PAC subunits may be under diversifying selection, perhaps evolving in ways to add or alter protein-protein interactions. This is a feature often seen in genes encoding regulatory proteins. Alternative polyadenylation is an important determinant of gene regulation, and the distinctive distribution of PAC-associated genes suggests that a surprising number of these subunits may play roles in alternative polyadenylation.

4.2 Multi-purposing is reflected in patterns of variation and sequence diversity

Genes that encode ten PAC subunits possess π_N/π_S ratios of less than 0.8. These genes encode CPSF160, CPSF100, CPSF73, CLPS3, CFIS2, FIPS5, CFIm68-1, symplekin, PABN1, and PABN2. Interestingly, all but one of the genes that have been reported as essential for plant growth are members of this set. For the one exception, FY, the bulk of divergence resides in the plant-specific C-terminal domain, with the evolutionarily-conserved core of the protein exhibiting a similar paucity of divergence (Figure 4A).

A recurring theme in the functioning of many of these relatively invariant PAC subunits is their demonstrated or hypothetical involvement in different RNA processing or metabolic activities apart from their roles in mRNA polyadenylation. For example, CLPS3 is the *Arabidopsis* ortholog of Clp1, a subunit of CFIIm (Xing et al., 2008a; Hunt et al., 2012). Clp1 also plays vital roles in tRNA maturation (Weitzer and Martinez, 2007; Ramirez et al., 2008). Clp1 may act as a more general RNA kinase, as it has been reported to be the kinase responsible for 5'-phosphorylation of siRNAs in mice (Fujinami et al., 2020). CFIS2 is one of two *Arabidopsis* orthologs of CFIm25. CFIS2 (but not CFIS1) also has roles in ribosomal RNA processing (Palm et al., 2019).

CPSF73 is among the most widely-conserved of all subunits of the PAC, being readily identifiable in virtually all eukaryotic genomes. CPSF73 is the enzyme that processes the pre-mRNA prior to polyadenylation. Its activity is tightly regulated, and access to the RNA substrate is controlled through a network of interactions with other PAC subunits. Chief among these subunits is CPSF100, also a protein that shows limited variability in *Arabidopsis* accessions. These two subunits form the core of the endonuclease module of the yeast and mammalian PACs. This module has other functions. For example, in yeast, it also mediates 3'-end processing of snoRNAs (Larochelle et al., 2018). In mammals, it mediates 3'-end formation of histone mRNAs, in conjunction with a dedicated complex that includes the snRNP U7 (Dominski et al., 2005; Kolev and Steitz, 2005). In plants, CPSF73 also plays roles in 3' end formation of snRNAs (Liu et al., 2016), perhaps analogous to the functioning of the endonuclease module in snoRNA processing in yeast.

Two of the three genes encoding PABN (PABN1 and PABN2) exhibit among the lowest ratios of all PAC subunit-encoding genes. In mammals, PABN helps to control the lengths of poly(A) tracts added to the newly-processed mRNA in the nucleus. PABN also

functions in alternative polyadenylation. In *Arabidopsis*, PABN2 has been reported to bind to the C-terminal extension telomerase-reverse transcriptase (TERT), a metallothionein (MTA2), Modifier Of Snc1 (MOS1), a nuclear DNA-binding protein (GP2), Oxidation Related Zinc Finger 2 (OZF2), and a Heat Shock 70 Cognate protein (HSP70-1) (Lee et al., 2012; Dokladal et al., 2015). All three *Arabidopsis* PABN isoforms interact with the *Arabidopsis* Cold Shock Domain 3 (AtCSD3) protein (Kim et al., 2013). The significance of these interactions is not known, but the different interacting partners are not 3' end processing factors (as far as has been reported). Thus, as is the case with CLPS3, CFIS2, CPSF100, and CPSF73, PABN1 and PABN2 may well have roles apart from those in mRNA polyadenylation.

Based on these considerations, it is tempting to speculate multifunctionality may impose stringent constraints on the abilities of proteins to explore sequence space, such that even modest missense changes may be selected against sufficiently to preclude fixation of variants in populations. This could suggest similar multifunctionality for the other two proteins whose diversity metrics are low. These two proteins, CPSF160 and FIPS5, are scaffolds of sorts. In mammals, CPSF160 coordinates the binding of two other CPSF subunits – CPSF30 and FY – to the polyadenylation signal and serves as a bridge between the PAS-binding module and the so-called cleavage module that consists of CPSF100 and CPSF73. FIP1 (the mammalian and yeast counterpart of FIPS5) recruits poly(A) polymerase to the PAC through interactions with both PAP and CPSF30 (Kumar et al., 2021; Muckenfuss et al., 2022). Analogous interactions have been reported for FIPS5 (Forbes et al., 2006; Hunt et al., 2008), as has a FIPS5-RNA interaction similar to that seen with the mammalian FIP1 ortholog (Forbes et al., 2006). In all three model organisms, the FIP1-CPSF30 interaction involves a conserved zinc finger motif (the C terminal most of the three such motifs in the *Arabidopsis* protein). The *Arabidopsis* FIPS5 protein has a stimulatory effect on the non-specific activity of recombinant PAPS2 (Forbes et al., 2006), and also inhibits a novel endonuclease activity associated with the third zinc finger motif of CPSF30 (Addepalli and Hunt, 2007). Different domains of FIPS5 are associated with interactions with PAP and CPSF30 and with RNA (Forbes et al., 2006). This multiplicity of interactions and activities may impose constraints that limit the sequence diversity seen in the FIPS5 gene in *Arabidopsis* accessions.

4.3 Proteins exhibiting high sequence diversity – suggestions of functional specialization

At the other end of the spectrum of sequence diversity are 13 PAC subunits whose π_N/π_S ratios are greater than 1.5 (Figure 3). Three of these (CLPS5, PCFS5, and PAPS3) are also affected by frameshifts and nonsense mutations (Tables 3, 4) and are discussed in following subsections. For the other 10, the patterns of diversity raise intriguing possibilities. These arise because, as indicated by Hughes (Hughes, 1999; Hughes et al., 2000), π_N/π_S ratios substantially greater than 1 are indicators of positive evolution. Positive evolution is often associated with diversification of protein function, as might be expected for protein isoforms derived from duplicated genes.

Four subunits with π_N/π_S ratios greater than 1.5 – CFIm68-2, PAPS1, FIPS3, and CFIS1 – are encoded by members of small gene families. For CFIm68-2, FIPS3, and CFIS1, the other members of the gene families (CFIm68-1, FIPS5, and CFIS2) show very low diversity (Figure 3). These observations are consistent with the hypothesis that the three genes with high π_N/π_S ratios encode proteins that possess functions apart, or differently, from their invariant counterparts. PAPS1 is one of three nuclear PAP isoforms; one of these (PAPS4) could not be assessed using the ViVa tool, but the other (PAPS2) showed a moderate degree of diversity, with a π_N/π_S ratio close to 1. The various PAPS isoforms have been shown to be functionally specialized, with these specialized roles being attributable to the C-termini of the respective proteins. The patterns of diversity seen in the three nuclear PAP isoforms are consistent with this, in that most of the variation seen in *Arabidopsis* accessions is localized to the respective C-terminal domains (Figure 4B; Supplementary Figures 2B, 5). These prior demonstrations of specialization amongst *Arabidopsis* nuclear PAP isoforms are consistent with the possibility raised by the high π_N/π_S ratios seen in the PAPS1 and PAPS2 genes. This in turn lends credence to the proposal that CFIm68-2, FIPS3, and CFIS1 also have distinct (if as yet unknown) roles.

Two of the proteins whose genes exhibit high π_N/π_S ratios (PCFS2 and PCFS4) encode isoforms of Pcf11. In contrast to the gene pairs represented by CFIS1/CFIS2, FIPS3/FIPS5, and CFIm68-1/CFIm68-2, both *Arabidopsis* Pcf11 isoforms are encoded by genes that exhibit high sequence diversity. This diversity falls outside of the parts of the proteins that are conserved and comprise functional domains (the polII CTD-interacting domain, or CID, and a zinc finger domain; Figure 6A). In mammals and yeast, Pcf11 functions in 3' end formation, transcription termination, and mRNA export (Birse et al., 1998; Grzechnik et al., 2015; Kamieniarz-Gdula et al., 2019). In *Arabidopsis*, PCSF4 has been implicated in transcription termination (de Felippes et al., 2020), and both PCFS2 and PCFS4 are found in nuclear complexes that include the bulk of the polyadenylation complex (Parker et al., 2021). Beyond these reports, little is known about the full scope of functioning of either Pcf11 isoform in plants. Since the CID and zinc finger domains mediate interactions between Pcf11 and the transcription/polyadenylation machineries in mammals and yeast, the paucity of diversity in these domains in PCFS2 and PCFS4 suggest that these two isoforms perform similar, overlapping functions in concert with the plant transcription/polyadenylation machineries. The patterns of diversity in PCFS2 and PCFS4 suggest that these two proteins engage in additional interactions that are more specific for the two isoforms; these specialized interactions might be attributed to the large swaths of each protein that are unique to the respective isoform. Moreover, given the association of high π_N/π_S ratios with positive selection during evolution, these two sets of specialized functions may be rapidly evolving. Of course, this is at the moment highly speculative. However, it is of interest to note that, in *Populus euphratica*, QTLs that encompass PCFS4 are associated with variation in shoot length (Zhang et al., 2017). Thus,

variation in PCFS4 may be causal for an important crop phenotype. This would lend credence to the proposition that PCFS function may be rapidly evolving in plants.

The *Arabidopsis* genome possesses two genes that encode orthologs of symplekin, a scaffold upon which other subcomplexes assemble. One of these genes – ESP4 – has a relatively high π_N/π_S ratio, while the other – SYM (At1g27595) – has a π_N/π_S ratio less than 0.8 (Figure 3). Neither of these two are, individually, essential for plant growth and development. ESP4 was first identified as a gene mutant of which exhibit increased transcriptional read-through and altered posttranscriptional gene silencing (Herr et al., 2006); these properties are consistent with functioning in mRNA 3' end formation and transcription termination. SYM mutants have altered responses to sugars (Zheng et al., 2015). The connection between a presumptive role for SYM in mRNA polyadenylation and sugar responses is not clear, and this protein has not been studied in the context of polyadenylation. Interestingly, ESP4 is present in complexes isolated by affinity purification of CPSF100 (Herr et al., 2006), FPA (Parker et al., 2021), CstF77 (Antosz et al., 2017), TFIIS (Antosz et al., 2017), and SPT4 (Antosz et al., 2017); in contrast, SYM is only seen in complexes containing SPT4 (Antosz et al., 2017). While the absence of a protein in a copurification analysis may be due many factors, this difference raises the possibility that the two symplekin orthologs may have somewhat different associations or roles.

Two of the high-diversity PAC subunits are CstF64 and CstF50 (Figure 3). In contrast to the subunits discussed in the preceding paragraphs, these proteins are encoded by single genes in *Arabidopsis*. As noted in this report (Figure 5) and elsewhere (Hunt, 2020), the plant CstF complex is curiously different from its mammalian counterpart; specifically, whereas CstF is essential in mammals, it is dispensable for viability in *Arabidopsis*. CstF64 and CstF77 play general roles in poly(A) site choice in *Arabidopsis*, but are dispensable for large numbers of poly(A) sites (Zeng et al., 2019). For example, these two proteins promote usage of a proximal poly(A) site associated with the COOLAIR antisense RNA but do not seem to have roles in usage of the poly(A) sites that define the 3' ends of the "sense" FLC transcripts (Liu et al., 2010). The possibility that CstF64 and CstF50 may be subjected to positive (diversifying) selection raises the possibility natural variation in these proteins may be a source for new or altered regulatory behavior.

4.4 PAC-encoding genes that possess premature translation termination codons

Several *Arabidopsis* PAC genes are affected by the occurrence in one or more accessions of premature stop codons and/or frame-shift mutations that severely truncate predicted protein products. This result was unexpected; by way of comparison, none of the 1815 human genes found to be tolerant of biallelic variation impact any of the known subunits of the human PAC (Karczewski et al., 2020). These instances pose questions regarding the structure and functioning of the plant polyadenylation complex. The implications of these results are discussed in the following.

4.4.1 CstF50 is not essential in *Arabidopsis*

The apparent absence of a functional CstF50 gene in one *Arabidopsis* accession (CS77397) that otherwise has a normal growth habit is interesting, since CstF50 is a subunit of a heteromeric complex (CstF) that in mammals is required for mRNA polyadenylation. The mammalian CstF recognizes functional RNA sequences (the DownStream Element, or DSE) 3' of the cleavage/polyadenylation site. As part of the complex, CstF50 serves to fine-tune the association of the complex with G/U-containing RNAs (that comprise the downstream element) (Yang et al., 2018) and links 3' end processing with DNA repair (Kleiman and Manley, 1999). In plants, CstF50 is present in nuclear complexes affinity-purified using tagged CstF77 or CstF64 (Antosz et al., 2017), suggestive of a presence in an analogous heteromeric complex. However, it does not seem to interact with the other two CstF subunits in pairwise interaction assays (Yao et al., 2002; Hunt et al., 2008). Beyond these reports, little is known about possible functions of CstF50 in polyadenylation in plants, or the architecture that links CstF50 with CstF77/CstF64-containing nuclear complexes in plants.

These considerations aside, the seeming dispensability of CstF50 in the CS77397 accession aligns with reports indicating that CstF77 and CstF64 are not required for *Arabidopsis* growth and development. Specifically, it has been shown that *Arabidopsis* (Col-0) mutants with null mutations in genes that encode these CstF subunits are viable, if diminished in stature and general growth habit (Zeng et al., 2019). These mutants exhibit a range of phenotypes that may be linked with altered responses to auxin. They also possess a molecular phenotype in which mRNA poly(A) site choice is altered on a genome-wide scale; this phenotype is consistent with the presumed functions of the proteins in polyadenylation. However, the dispensability of these two proteins suggests that CstF may not be needed for the basic functionality of the PAC, namely recognition of the pre-mRNA, endonucleolytic cleavage, and addition of the poly(A) tract.

These considerations notwithstanding, there are some distinctions that may be made. The CstF77 and CstF64 null mutants in the Col-0 background have profound growth phenotypes. CS77397, in contrast, has a growth habit that is as unremarkable as most other *Arabidopsis* accessions, and shows no hints of having strongly-altered auxin responses. Moreover, if one grants a cause-and-effect relationship between the growth phenotypes, altered auxin responses, and global changes in poly (A) site choice in the CstF77 and CstF64 mutants, it stands to reason that global poly(A) site choice is probably not affected by the absence of CstF50 in the CS77397 accession. This in turn suggests a modest role for CstF50 in the functioning of the PAC. Other eukaryotic lineages lack CstF50 orthologs; these lineages include yeast, in which two other orthologs of CstF subunits (Rna14 and Rna15) function as part of a complex (CF1A) that lacks a CstF50 ortholog. It may be that the plant PAC may be more akin to the yeast than the mammalian complex, and the plant CstF50 may be an accessory rather than a core PAC subunit.

Clearly, the functioning of CstF50 in polyadenylation is largely undefined, with much remaining to be learned. Whatever its role(s), the variability beyond the singular premature termination codon in

CS77397 raises some interesting possibilities. In particular, the high π_N/π_S ratio seen in the CstF50 gene suggests that this protein may be subject to diversifying evolutionary change. Such a possibility is consistent with a role as an accessory protein in the complex, one whose activity (or even presence) may vary in the plant and over evolutionary time.

The existence of an *Arabidopsis* accession that has a nonfunctional CstF50 gene raises questions as to how CstF50 might persist over evolutionary time in the plant lineage. The durability of the plant CstF50, even after many millions of years of evolution, strongly suggests that the protein has important functions that are targets of natural selection. This possibility is not consistent with the dispensability of the protein in the CS77397 accession. It is difficult to resolve this paradox at the moment. However, this curious result magnifies the possibility that the plant CstF50 has unexpected roles, either in mRNA polyadenylation or perhaps other aspects of plant growth and development.

4.4.2 PCFS1/PCFS5 (Pcf11) and CLPS5 (Clp1) are pseudogenes

Arabidopsis possesses three genes encoding PAC subunits that are not seen in other plants; these are genes that encode novel orthologs of Pcf11 and Clp1, subunits of CFIIm. Two of these genes, termed PCFS1 and PCFS5, encode novel Pcf11-related proteins. In a survey of 11 well-characterized plant genomes, these two genes were only seen *Arabidopsis thaliana*, *Arabidopsis lyrata*, and perhaps *Populus trichocarpa* (Hunt et al., 2012). The genes encoding PCFS1 and PCSF5 are similar in gene structure. They lack introns and a majority of RNAs specified by the *Arabidopsis* genes terminate at distinct poly(A) sites situated well within the respective protein-coding regions (Figure 6). As such, these RNAs would likely be substrates for nonstop RNA degradation. The PCFS1 and PCFS5 genes are also impacted by frameshift and nonsense mutations in the collection of *Arabidopsis* accessions. The polypeptides encoded by the full-length PCFS1 and PCFS5 mRNAs are truncated when compared with other *Arabidopsis* Pcf11 orthologs (PCFS2 and PCFS4 mentioned in the preceding) and lack the CID domains seen in the other Pcf11 orthologs. Collectively, these data raise the possibility that the genes that encode PCFS1 and PCFS5 are likely to be pseudogenes.

Like PCFS1 and PCFS5, the novel Clp1-related isoform CLPS5 seems to be specific for the *Arabidopsis* lineage, and is not found in other plant genomes (Hunt et al., 2012). In contrast to CLPS3, *Arabidopsis* mutants with T-DNA insertions that would disrupt the CLPS5 gene are viable, indicating that this protein is dispensable for growth and development (Xing et al., 2008a). 590 *Arabidopsis* accessions were found to possess CLPS5 genes with premature termination codons, most (or all) of which would dramatically truncate translated polypeptides (Table 3; Figure 6). In addition, surveys of gene expression indicate that the CLPS5 gene is expressed at very low levels, if at all (Hunt et al., 2008). Taken together, these observations suggest that the *Arabidopsis* CLPS5 gene is a lineage-specific duplicate that lacks function and is likely a pseudogene.

4.4.3 PAPS3 - an enigma

Of the polyadenylation-associated genes in the *Arabidopsis* genome, PAPS3 is perhaps the most perplexing. PAPS3-like proteins, enzymes that lack C-terminal domains that are associated with nuclear polyadenylation and specialized functions of nuclear PAPs in *Arabidopsis*, are widespread in the plant lineage (Hunt et al., 2012). However, the variation seen the *Arabidopsis* PAPS3 gene is extensive, with a high π_N/π_S ratio and premature termination codons in many accessions (Figure 3; Table 3). These features are similar to those seen in the PCFS1, PCFS5, and CLPS5 genes, and thus raise the possibility that PAPS3 genes may be non-functional, and perhaps pseudogenes. These observations are cause to re-visit other aspects of PAPS3 genes in plants. For example, in contrast to PCFS1, PCFS5, and CLPS5, all of which are seen only in *Arabidopsis*, PAPS3-like genes are found widely in angiosperms (Hunt et al., 2012). However, their occurrence is not universal, as some species (for example, *Glycine max*) lack identifiable PAPS3-like genes (Hunt et al., 2012). Therefore, PAPS3-like proteins are likely not an essential part of the plant proteomic toolkit. The observation that the expression of the *Arabidopsis* PAPS3 gene is strongly pollen-specific suggested a role for the protein in some aspect of male gametophyte development (Hunt et al., 2012). However, male-specific expression is not a general feature of plant PAPS3-like genes (Figure 7D).

As is the case with CstF50, it is challenging to reconcile the evolutionary conservation of PAPS3 genes in plants with the dispensability of the protein in *Arabidopsis* (documented in this report) and its seeming absence in other plants. If one assumes that PAPS3 was present in the common ancestor of higher plants, its absence in species such as *Glycine max* supports the contention that, absent selectable roles, this gene is subject to evolutionary forces (random mutant, chiefly) that over the course of time would eliminate the gene. Given that the *Arabidopsis* PAPS3 gene is not essential (this study) and can be inactivated without obvious phenotypic impacts, it is reasonable to expect that this gene should not persist, but rather should be lost in higher plants. However, this is not the case. Along with the distinct and different PAPS3 expression patterns noted here (Figure 7D), these results raise the possibility that PAPS3 orthologs may have evolved lineage-specific functions that are both dispensable (at least in some cases, as is seen in *Arabidopsis*) and subject to natural selection (so as to preserve the genes over evolutionary time scales).

5 Summary

We have compiled and studied the range of variation in *Arabidopsis thaliana* that affects the different subunits of the polyadenylation complex. The results suggest that a sizable number of PAC subunits exhibit variation that is suggestive of a degree of diversifying selection, and may indicate expanded roles for different subunits in the regulation of alternative polyadenylation. At least three genes, all *Arabidopsis*-specific, are likely to non-functional, based on both the widespread occurrence of disruptive (e.g., nonsense) mutations and gene expression patterns that are consistent with a lack of function. Most interestingly, two genes (CstF50 and PAPS3) that are widely-conserved in plants are affected in some accessions by disruptive mutations. The seeming dispensability of these genes is difficult to reconcile by their broad evolutionary conservation, and

poses new questions regarding the composition and functioning of the plant polyadenylation complex.

Data availability statement

The high throughput sequencing data generated in the course of this research may be found under Bioproject PRJNA1023006. Accessions for the publicly available datasets analyzed in this work are provided in Supplementary File 8. The original contributions presented in the study are included in the article and Supplementary Material; further inquiries can be directed to the corresponding authors.

Author contributions

LZ: Data curation, Formal analysis, Investigation, Methodology, Validation, Writing – original draft, Writing – review & editing. KL: Data curation, Investigation, Methodology, Writing – review & editing. AH: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was supported by Hatch Project KY006118. Lichun Zhou was supported by the Department of Plant and Soil Sciences at the University of Kentucky.

Acknowledgments

The authors thank Carol Von Lanken for support and suggestions. They are grateful to Dr. Clay Wright for guidance regarding the use of the Visualizing Variation (ViVa) analysis package.

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.2023.1303398/full#supplementary-material

References

Addepalli, B., and Hunt, A. G. (2007). A novel endonuclease activity associated with the Arabidopsis ortholog of the 30-kDa subunit of cleavage and polyadenylation specificity factor. *Nucleic Acids Res.* 35 (13), 4453–4463. doi: 10.1093/nar/gkm457

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 (2), 481–491. doi: 10.1016/j.cell.2016.05.063

Antosz, W., Pfab, A., Ehrnsberger, H. F., Holzinger, P., Kollen, K., Mortensen, S. A., et al. (2017). The composition of the arabidopsis RNA polymerase II transcript elongation complex reveals the interplay between elongation and mRNA processing factors. *Plant Cell* 29 (4), 854–870. doi: 10.1105/tpc.16.00735

Bard, J., Zhelkovsky, A. M., Helmling, S., Earnest, T. N., Moore, C. L., and Bohm, A. (2000). Structure of yeast poly(A) polymerase alone and in complex with 3 '-dATP. *Science* 289 (5483), 1346–1349. doi: 10.1126/science.289.5483.1346

Birse, C. E., Minvielle-Sebastia, L., Lee, B. A., Keller, W., and Proudfoot, N. J. (1998). Coupling termination of transcription to messenger RNA maturation in yeast. *Science* 280 (5361), 298–301. doi: 10.1126/science.280.5361.298

Boreikaitė, V., and Passmore, L. A. (2023). 3'-end processing of eukaryotic mRNA: machinery, regulation, and impact on gene expression. *Annu. Rev. Biochem.* 92, 199–225. doi: 10.1146/annurev-biochem-052521-012445

Boreikaite, V., Elliott, T. S., Chin, J. W., and Passmore, L. A. (2022). RBBP6 activates the pre-mRNA 3' end processing machinery in humans. *Genes Dev.* 36 (3-4), 210–224. doi: 10.1101/gad.349223.121

Bruggeman, Q., Garmier, M., de Bont, L., Soubigou-Taconnat, L., Mazubert, C., Benhamed, M., et al. (2014). The polyadenylation factor subunit CLEAVAGE AND POLYADENYLATION SPECIFICITY FACTOR30: A key factor of programmed cell death and a regulator of immunity in arabidopsis. *Plant Physiol.* 165 (2), 732–746. doi: 10.1104/pp.114.236083

Chakrabarti, M., de Lorenzo, L., Abdel-Ghany, S. E., Reddy, A. S., and Hunt, A. G. (2020). Wide-ranging transcriptome remodelling mediated by alternative polyadenylation in response to abiotic stresses in Sorghum. *Plant J.* 102 (5), 916–930. doi: 10.1111/tpj.14671

Chakrabarti, M., and Hunt, A. G. (2015). CPSF30 at the interface of alternative polyadenylation and cellular signaling in plants. *Biomolecules* 5 (2), 1151–1168. doi: 10.3390/biom5021151

Chan, S. L., Huppertz, I., Yao, C. G., Weng, L. J., Moresco, J. J., Yates, J. R., et al. (2014). CPSF30 and Wdr33 directly bind to AAUAAA in mammalian mRNA 3 ' processing. *Genes Dev.* 28 (21), 2370–2380. doi: 10.1101/gad.250993.114

Clerici, M., Faini, M., Muckenfuss, L. M., Aebersold, R., and Jinek, M. (2018). Structural basis of AAUAAA polyadenylation signal recognition by the human CPSF complex. *Nat. Struct. Mol. Biol.* 25 (2), 135–13+. doi: 10.1038/s41594-017-0020-6

Czesnick, H., and Lenhard, M. (2016). Antagonistic control of flowering time by functionally specialized poly(A) polymerases in Arabidopsis thaliana. *Plant J.* 88 (4), 570–583. doi: 10.1111/tpj.13280

de Felippes, F., McHale, M., Doran, R. L., Roden, S., Eamens, A. L., Finnegan, E. J., et al. (2020). The key role of terminators on the expression and post-transcriptional gene silencing of transgenes. *Plant J.* 104 (1), 96–112. doi: 10.1111/tpj.14907

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

Di Giammartino, D. C., Li, W., Ogami, K., Yashinskie, J. J., Hoque, M., Tian, B., et al. (2014). RBBP6 isoforms regulate the human polyadenylation machinery and modulate expression of mRNAs with AU-rich 3' UTRs. *Genes Dev.* 28 (20), 2248–2260. doi: 10.1101/gad.245787.114

Dokladal, L., Honys, D., Rana, R., Lee, L. Y., Gelvin, S. B., and Sykorova, E. (2015). cDNA library screening identifies protein interactors potentially involved in non-telomeric roles of arabidopsis telomerase. *Front. Plant Sci.* 6. doi: 10.3389/fpls.2015.00985

Dominski, Z., Yang, X. C., and Marzluff, W. F. (2005). The polyadenylation factor CPSF-73 is involved in histone-pre-mRNA processing. *Cell* 123 (1), 37–48. doi: 10.1016/j.cell.2005.08.002

Edmonds, M. (2002). A history of poly A sequences: from formation to factors to function. *Prog. Nucleic Acid Res. Mol. Biol.* 71, 285–389. doi: 10.1016/S0079-6603(02)71046-5

Forbes, K. P., Addepalli, B., and Hunt, A. G. (2006). An Arabidopsis Fip1 homolog interacts with RNA and provides conceptual links with a number of other polyadenylation factor subunits. *J. Biol. Chem.* 281 (1), 176–186. doi: 10.1074/jbc.M510964200

Fu, H., Yang, D., Su, W., Ma, L., Shen, Y., Ji, G., et al. (2016). Genome-wide dynamics of alternative polyadenylation in rice. *Genome Res.* 26 (12), 1753–1760. doi: 10.1101/gr.210757.116

Fujinami, H., Shiraishi, H., Hada, K., Inoue, M., Morisaki, I., Higa, R., et al. (2020). CLP1 acts as the main RNA kinase in mice. *Biochem. Biophys. Res. Commun.* 525 (1), 129–134. doi: 10.1016/j.bbrc.2020.02.066

Ghazy, M. A., He, X., Singh, B. N., Hampsey, M., and Moore, C. (2009). The essential N terminus of the Pta1 scaffold protein is required for snoRNA transcription termination and Ssu72 function but is dispensable for pre-mRNA 3'-end processing. *Mol. Cell Biol.* 29 (8), 2296–2307. doi: 10.1128/mcb.01514-08

Ghosh, S., Ataman, M., Bak, M., Börsch, A., Schmidt, A., Buczak, K., et al. (2022). CFIm-mediated alternative polyadenylation remodels cellular signaling and miRNA biogenesis. *Nucleic Acids Res.* 50 (6), 3096–3114. doi: 10.1093/nar/gkac114

Grozdanov, P. N., Masoumzadeh, E., Latham, M. P., and MacDonald, C. C. (2018). The structural basis of CstF-77 modulation of cleavage and polyadenylation through stimulation of CstF-64 activity. *Nucleic Acids Res.* 46 (22), 12022–12039. doi: 10.1093/nar/gky862

Grzechnik, P., Gdula, M. R., and Proudfoot, N. J. (2015). Pcf11 orchestrates transcription termination pathways in yeast. *Genes Dev.* 29 (8), 849–861. doi: 10.1101/gad.251470.114

Hamm, M. O., Moss, B. L., Leydon, A. R., Gala, H. P., Lanctot, A., Ramos, R., et al. (2019). Accelerating structure-function mapping using the ViVa webtool to mine natural variation. *Plant Direct* 3 (7), e00147. doi: 10.1002/pld3.147

Helmling, S., Zhelkovsky, A., and Moore, C. L. (2001). Fip1 regulates the activity of Poly(A) polymerase through multiple interactions. *Mol. Cell Biol.* 21 (6), 2026–2037. doi: 10.1128/mcb.21.6.2026-2037.2001

Henderson, I. R., Liu, F. Q., Drea, S., Simpson, G. G., and Dean, C. (2005). An allelic series reveals essential roles for FY in plant development in addition to flowering-time control. *Development* 132 (16), 3597–3607. doi: 10.1242/dev.01924

Herr, A. J., Molnar, A., Jones, A., and Baulcombe, D. C. (2006). Defective RNA processing enhances RNA silencing and influences flowering of Arabidopsis. *Proc. Natl. Acad. Sci. U.S.A.* 103 (41), 14994–15001. doi: 10.1073/pnas.0606536103

Hill, C. H., Boreikaite, V., Kumar, A., Casanal, A., Kubik, P., Degliesposti, G., et al. (2019). Activation of the Endonuclease that Defines mRNA 3' Ends Requires Incorporation into an 8-Subunit Core Cleavage and Polyadenylation Factor Complex. *Mol. Cell* 73 (6), 1217–1231 e1211. doi: 10.1016/j.molcel.2018.12.023

Hou, Y. F., Sun, J., Wu, B. X., Gao, Y. Y., Nie, H. B., Nie, Z. T., et al. (2021). CPSF30-L-mediated recognition of mRNA m(6)A modification controls alternative polyadenylation of nitrate signaling-related gene transcripts in Arabidopsis. *Mol. Plant* 14 (4), 688–699. doi: 10.1016/j.molp.2021.01.013

Hughes, A. L. (1999). *Adaptive evolution of genes and genomes* (New York: Oxford University Press).

Hughes, A. L., Green, J. A., Garbayo, J. M., and Roberts, R. M. (2000). Adaptive diversification within a large family of recently duplicated, placentally expressed genes. *Proc. Natl. Acad. Sci. U.S.A.* 97 (7), 3319–3323. doi: 10.1073/pnas.97.7.3319

Hunt, A. G. (2014). The Arabidopsis polyadenylation factor subunit CPSF30 as conceptual link between mRNA polyadenylation and cellular signaling. *Curr. Opin. Plant Biol.* 21, 128–132. doi: 10.1016/j.pbi.2014.07.002

Hunt, A. G. (2020). mRNA 3' end formation in plants: Novel connections to growth, development and environmental responses. *Wiley Interdiscip Rev. RNA* 11 (3), e1575. doi: 10.1002/wrna.1575

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

Hunt, A. G., Xu, R., Addepalli, B., Rao, S., Forbes, K. P., Meeks, L. R., et al. (2008). Arabidopsis mRNA polyadenylation machinery: comprehensive analysis of protein-protein interactions and gene expression profiling. *BMC Genomics* 9, 220. doi: 10.1186/1471-2164-9-220

Kadirjan-Kalbach, D. K., Turmo, A., Wang, J., Smith, B. C., Chen, C., Porter, K. J., et al. (2019). Allelic variation in the chloroplast division gene ftsZ2-2 leads to natural variation in chloroplast size. *Plant Physiol.* 181 (3), 1059–1074. doi: 10.1104/pp.19.00841

Kamieniarz-Gdula, K., Gdula, M. R., Panser, K., Nojima, T., Monks, J., Wisniewski, J. R., et al. (2019). Selective roles of vertebrate PCF11 in premature and full-length transcript termination. *Mol. Cell* 74 (1), 158–172 e159. doi: 10.1016/j.molcel.2019.01.027

Kappel, C., Trost, G., Czesnick, H., Ramming, A., Kolbe, B., Vi, S. L., et al. (2015). Genome-wide analysis of PAPS1-dependent polyadenylation identifies novel roles for functionally specialized poly(A) polymerases in arabidopsis thaliana. *PloS Genet.* 11 (8), e1005474. doi: 10.1371/journal.pgen.1005474

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alfoldi, J., Wang, Q., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581 (7809), 434–443. doi: 10.1038/s41586-020-2308-7

Kaufmann, I., Martin, G., Friedlein, A., Langen, H., and Keller, W. (2004). Human Fip1 is a subunit of CPSF that binds to U-rich RNA elements and stimulates poly(A) polymerase. *EMBO J.* 23 (3), 616–626. doi: 10.1038/sj.emboj.7600070

Kennedy, S. A., Frazier, M. L., Steiniger, M., Mast, A. M., Marzluff, W. F., and Redinbo, M. R. (2009). Crystal structure of the HEAT domain from the Pre-mRNA processing factor Symplekin. *J. Mol. Biol.* 392 (1), 115–128. doi: 10.1016/j.jmb.2009.06.062

Kerwitz, Y., Kühn, U., Lilie, H., Knoth, A., Scheuermann, T., Friedrich, H., et al. (2003). Stimulation of poly(A) polymerase through a direct interaction with the nuclear poly(A) binding protein allosterically regulated by RNA. *EMBO J.* 22 (14), 3705–3714. doi: 10.1093/emboj/cdg347

Kim, M. H., Sonoda, Y., Sasaki, K., Kaminaka, H., and Imai, R. (2013). Interactome analysis reveals versatile functions of Arabidopsis COLD SHOCK DOMAIN PROTEIN 3 in RNA processing within the nucleus and cytoplasm. *Cell Stress Chaperones* 18 (4), 517–525. doi: 10.1007/s12192-012-0398-3

Kim, M., Swenson, J., McLoughlin, F., and Vierling, E. (2023). Mutation of the polyadenylation complex subunit CstF77 reveals that mRNA 3' end formation and HSP101 levels are critical for a robust heat stress response. *Plant Cell* 35 (2), 924–941. doi: 10.1093/plcell/koac351

Kleiman, F. E., and Manley, J. L. (1999). Functional interaction of BRCA1-associated BARD1 with polyadenylation factor CstF-50. *Science* 285 (5433), 1576–1579. doi: 10.1126/science.285.5433.1576

Kolev, N. G., and Steitz, J. A. (2005). Symplekin and multiple other polyadenylation factors participate in 3'-end maturation of histone mRNAs. *Genes Dev.* 19 (21), 2583–2592. doi: 10.1101/gad.1371105

Kühn, U., Gündel, M., Knoth, A., Kerwitz, Y., Rüdel, S., and Wahle, E. (2009). Poly(A) tail length is controlled by the nuclear poly(A)-binding protein regulating the interaction between poly(A) polymerase and the cleavage and polyadenylation specificity factor. *J. Biol. Chem.* 284 (34), 22803–22814. doi: 10.1074/jbc.M109.018226

Kuhn, U., and Wahle, E. (2004). Structure and function of poly(A) binding proteins. *Biochim. Et Biophys. Acta-Gene Structure Expression* 1678 (2-3), 67–84. doi: 10.1016/j.bbaexp.2004.03.008

Kumar, A., Yu, C. W. H., Rodriguez-Molina, J. B., Li, X. H., Freund, S. M. V., and Passmore, L. A. (2021). Dynamics in Fip1 regulate eukaryotic mRNA 3' end processing. *Genes Dev.* 35 (21-22), 1510–1526. doi: 10.1101/gad.348671.121

Larochelle, M., Robert, M. A., Hebert, J. N., Liu, X., Matteau, D., Rodrigue, S., et al. (2018). Common mechanism of transcription termination at coding and noncoding RNA genes in fission yeast. *Nat. Commun.* 9 (1), 4364. doi: 10.1038/s41467-018-06546-x

Lee, S. D., Liu, H. Y., Graber, J. H., Heller-Trulli, D., Kaczmarek Michaels, K., Cerezo, J. F., et al. (2020). Regulation of the Ysh1 endonuclease of the mRNA cleavage/polyadenylation complex by ubiquitin-mediated degradation. *RNA Biol.* 17 (5), 689–702. doi: 10.1080/15476286.2020.1724717

Lee, S. D., and Moore, C. L. (2014). Efficient mRNA polyadenylation requires a ubiquitin-like domain, a zinc knuckle, and a RING finger domain, all contained in the Mpe1 protein. *Mol. Cell Biol.* 34 (21), 3955–3967. doi: 10.1128/MCB.00077-14

Lee, L. Y., Wu, F. H., Hsu, C. T., Shen, S. C., Yeh, H. Y., Liao, D. C., et al. (2012). Screening a cDNA library for protein-protein interactions directly in planta. *Plant Cell* 24 (5), 1746–1759. doi: 10.1105/tpc.112.097998

Li, Y., Chen, F., Yang, Y., Han, Y., Ren, Z., Li, X., et al. (2023). The Arabidopsis pre-mRNA 3' end processing related protein FIP1 promotes seed dormancy via the DOG1 and ABA pathways. *Plant J.* 115 (2), 494–509. doi: 10.1111/tpj.16239

Li, Z. H., Wang, R. C., Gao, Y. Y., Wang, C., Zhao, L. F., Xu, N., et al. (2017). The Arabidopsis CPSF30-L gene plays an essential role in nitrate signaling and regulates the nitrate transceptor gene NRT1.1. *New Phytol.* 216 (4), 1205–1222. doi: 10.1111/nph.14743

Lin, J. C., Xu, R. W., Wu, X. H., Shen, Y. J., and Li, Q. S. Q. (2017). Role of cleavage and polyadenylation specificity factor 100: anchoring poly(A) sites and modulating transcription termination. *Plant J.* 91 (5), 829–839. doi: 10.1111/tpj.13611

Liu, Y., Li, S., Chen, Y., Kimberlin, A. N., Cahoon, E. B., and Yu, B. (2016). snRNA 3' End processing by a CPSF73-containing complex essential for development in arabidopsis. *PloS Biol.* 14 (10), e1002571. doi: 10.1371/journal.pbio.1002571

Liu, F., Marquardt, S., Lister, C., Swiezewski, S., and Dean, C. (2010). Targeted 3 ' Processing of antisense transcripts triggers arabidopsis FLC chromatin silencing. *Science* 327 (5961), 94–97. doi: 10.1126/science.1180278

Liu, M., Xu, R., Merrill, C., Hong, L., Von Lanken, C., Hunt, A. G., and Li, Q. (2014). Q Integration of developmental and environmental signals via a polyadenylation factor in Arabidopsis. *PloS One* 9 (12), e115779. doi: 10.1371/journal.pone.0115779

Ma, H., Cai, L., Lin, J., Zhou, K., and Li, Q. Q. (2022). Divergence in the Regulation of the Salt Tolerant Response Between Arabidopsis thaliana and Its Halophytic Relative Eutrema salsugineum by mRNA Alternative Polyadenylation. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.866054

Ma, L., Pati, P. K., Liu, M., Li, Q. Q., and Hunt, A. G. (2014). High throughput characterizations of poly(A) site choice in plants. *Methods* 67 (1), 74–83. doi: 10.1016/j.ymeth.2013.06.037

Martin, G., Gruber, A. R., Keller, W., and Zavolan, M. (2012). Genome-wide analysis of pre-mRNA 3' end processing reveals a decisive role of human cleavage factor I in the regulation of 3' UTR length. *Cell Rep.* 1 (6), 753–763. doi: 10.1016/j.celrep.2012.05.003

Meeks, L. R., Addepalli, B., and Hunt, A. G. (2009). Characterization of genes encoding poly(A) polymerases in plants: evidence for duplication and functional specialization. *PloS One* 4 (11), e8082. doi: 10.1371/journal.pone.0008082

Mirkin, N., Fonseca, D., Mohammed, S., Cevher, M. A., Manley, J. L., and Kleiman, F. E. (2008). The 3' processing factor CstF functions in the DNA repair response. *Nucleic Acids Res.* 36 (6), 1792–1804. doi: 10.1093/nar/gkn005

Muckenfuss, L. M., Migenda Herranz, A. C., Boneberg, F. M., Clerici, M., and Jinek, M. (2022). Fip1 is a multivalent interaction scaffold for processing factors in human mRNA 3' end biogenesis. *Elife* 11, e80332. doi: 10.7554/eLife.80332

Palm, D., Streit, D., Shanmugam, T., Weis, B. L., Ruprecht, M., Simm, S., et al. (2019). Plant-specific ribosome biogenesis factors in Arabidopsis thaliana with essential function in rRNA processing. *Nucleic Acids Res.* 47 (4), 1880–1895. doi: 10.1093/nar/gky1261

Parker, M. T., Knop, K., Zacharaki, V., Sherwood, A. V., Tome, D., Yu, X., et al. (2021). Widespread premature transcription termination of Arabidopsis thaliana NLR genes by the spen protein FPA. *Elife* 10, e65537. doi: 10.7554/eLife.65537

Pati, P. K., Ma, L., and Hunt, A. G. (2015). Genome-wide determination of poly(A) site choice in plants. *Methods Mol. Biol.* 1255, 159–174. doi: 10.1007/978-1-4939-2175-1_14

Pérez Cañadillas, J. M., and Varani, G. (2003). Recognition of GU-rich polyadenylation regulatory elements by human CstF-64 protein. *EMBO J.* 22 (11), 2821–2830. doi: 10.1093/emboj/cdg259

Ramirez, A., Shuman, S., and Schwer, B. (2008). Human RNA 5'-kinase (hClp1) can function as a tRNA splicing enzyme *in vivo*. *RNA* 14 (9), 1737–1745. doi: 10.1261/rna.1142908

Ramming, A., Kappel, C., Kanaoka, M. M., Higashiyama, T., and Lenhard, M. (2023). Poly(A) polymerase 1 contributes to competence acquisition of pollen tubes growing through the style in Arabidopsis thaliana. *Plant J.* 114 (3), 651–667. doi: 10.1111/tpj.16162

Rodríguez-Molina, J. B., O'Reilly, F. J., Fagarasan, H., Sheekey, E., Maslen, S., Skehel, J. M., et al. (2022). Mpe1 senses the binding of pre-mRNA and controls 3' end processing by CPF. *Mol. Cell* 82 (13), 2490–2504.e2412. doi: 10.1016/j.molcel.2022.04.021

Ruepp, M. D., Schweingruber, C., Kleinschmidt, N., and Schümperli, D. (2011). Interactions of CstF-64, CstF-77, and symplekin: implications on localisation and function. *Mol. Biol. Cell* 22 (1), 91–104. doi: 10.1091/mbc.E10-06-0543

Sadowski, M., Dichtl, B., Hubner, W., and Keller, W. (2003). Independent functions of yeast Pcf11p in pre-mRNA 3 ' end processing and in transcription termination. *EMBO J.* 22 (9), 2167–2177. doi: 10.1093/emboj/cdg200

Schäfer, P., Tüting, C., Schönemann, L., Kühn, U., Treiber, T., Treiber, N., et al. (2018). Reconstitution of mammalian cleavage factor II involved in 3' processing of mRNA precursors. *Rna* 24 (12), 1721–1737. doi: 10.1261/rna.068056.118

Schmidt, M., Kluge, F., Sandmeir, F., Kuhn, U., Schafer, P., Tuting, C., et al. (2022). Reconstitution of 3' end processing of mammalian pre-mRNA reveals a central role of RBBP6. *Genes Dev.* 36 (3-4), 195–209. doi: 10.1101/gad.349217.121

Schonemann, L., Kuhn, U., Martin, G., Schafer, P., Gruber, A. R., Keller, W., et al. (2014). Reconstitution of CPSF active in polyadenylation: recognition of the polyadenylation signal by WDR33. *Genes Dev.* 28 (21), 2381–2393. doi: 10.1101/gad.250985.114

Shen, Y., Venu, R. C., Nobuta, K., Wu, X., Notibala, V., Demirci, C., et al. (2011). Transcriptome dynamics through alternative polyadenylation in developmental and environmental responses in plants revealed by deep sequencing. *Genome Res.* 21 (9), 1478–1486. doi: 10.1101/gr.114744.110

Simpson, G. G., Dijkwel, P. P., Quesada, V., Henderson, I., and Dean, C. (2003). FY is an RNA 3' end-processing factor that interacts with FCA to control the Arabidopsis floral transition. *Cell* 113 (6), 777–787. doi: 10.1016/s0092-8674(03)00425-2

Tellez-Robledo, B., Manzano, C., Saez, A., Navarro-Neila, S., Silva-Navas, J., de Lorenzo, L., et al. (2019). The polyadenylation factor FIP1 is important for plant development and root responses to abiotic stresses. *Plant J.* 99 (6), 1203–1219. doi: 10.1111/tpj.14416

Thomas, P. E. (2015). Analysis of poly(A) site choice using a Java-based clustering algorithm. *Methods Mol. Biol.* 1255, 49–56. doi: 10.1007/978-1-4939-2175-1_5

Thomas, P. E., Wu, X., Liu, M., Gaffney, B., Ji, G., Li, Q. Q., et al. (2012). Genome-wide control of polyadenylation site choice by CPSF30 in arabidopsis. *Plant Cell* 24 (11), 4376–4388. doi: 10.1105/tpc.112.096107

Trost, G., Vi, S. L., Czesnick, H., Lange, P., Holton, N., Giavalisco, P., et al. (2014). Arabidopsis poly(A) polymerase PAPS1 limits founder-cell recruitment to organ primordia and suppresses the salicylic acid-independent immune response downstream of EDS1/PAD4. *Plant J.* 77 (5), 688–699. doi: 10.1111/tpj.12421

Vi, S. L., Trost, G., Lange, P., Czesnick, H., Rao, N., Lieber, D., et al. (2013). Target specificity among canonical nuclear poly(A) polymerases in plants modulates organ growth and pathogen response. *Proc. Natl. Acad. Sci. U.S.A.* 110 (34), 13994–13999. doi: 10.1073/pnas.1303967110

Weitzer, S., and Martinez, J. (2007). The human RNA kinase hClp1 is active on 3' transfer RNA exons and short interfering RNAs. *Nature* 447 (7141), 222–226. doi: 10.1038/nature05777

West, S., and Proudfoot, N. J. (2008). Human Pcf11 enhances degradation of RNA polymerase II-associated nascent RNA and transcriptional termination. *Nucleic Acids Res.* 36 (3), 905–914. doi: 10.1093/nar/gkm1112

Whittaker, C., and Dean, C. (2017). The FLC locus: A platform for discoveries in epigenetics and adaptation. *Annu. Rev. Cell Dev. Biol.* 33, 555–575. doi: 10.1146/annurev-cellbio-100616-060546

Xiang, K., Nagaike, T., Xiang, S., Kilic, T., Beh, M. M., Manley, J. L., et al. (2010). Crystal structure of the human symplekin-Ssu72-CTD phosphopeptide complex. *Nature* 467 (7316), 729–733. doi: 10.1038/nature09391

Xing, D., Zhao, H., and Li, Q. Q. (2008a). Arabidopsis CLP1-SIMILAR PROTEIN3, an ortholog of human polyadenylation factor CLP1, functions in gametophyte, embryo, and postembryonic development. *Plant Physiol.* 148 (4), 2059–2069. doi: 10.1104/pp.108.129817

Xing, D., Zhao, H., Xu, R., and Li, Q. Q. (2008b). Arabidopsis PCFS4, a homologue of yeast polyadenylation factor Pcf11p, regulates FCA alternative processing and promotes flowering time. *Plant J.* 54 (5), 899–910. doi: 10.1111/j.1365-313X.2008.03455.x

Xu, R. Q., Zhao, H. W., Dinkins, R. D., Cheng, X. W., Carberry, G., and Li, Q. Q. (2006). The 73 kD Subunit of the cleavage and polyadenylation specificity factor (CPSF) complex affects reproductive development in Arabidopsis. *Plant Mol. Biol.* 61 (4-5), 799–815. doi: 10.1007/s11103-006-0051-6

Yan, C., Wang, Y., Lyu, T., Hu, Z., Ye, N., Liu, W., et al. (2021). Alternative Polyadenylation in response to temperature stress contributes to gene regulation in Populus trichocarpa. *BMC Genomics* 22 (1), 53. doi: 10.1186/s12864-020-07353-9

Yang, W., Hsu, P. L., Yang, F., Song, J.-E., and Varani, G. (2018). Reconstitution of the CstF complex unveils a regulatory role for CstF-50 in recognition of 3 '-end processing signals. *Nucleic Acids Res.* 46 (2), 493–503. doi: 10.1093/nar/gkx1177

Yao, Y., Song, L., Katz, Y., and Galili, G. (2002). Cloning and characterization of Arabidopsis homologues of the animal CstF complex that regulates 3' mRNA cleavage and polyadenylation. *J. Exp. Bot.* 53 (378), 2277–2278. doi: 10.1093/jxb/erf073

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

Yu, Z. B., Lin, J. C., and Li, Q. S. Q. (2019). Transcriptome analyses of FY mutants reveal its role in mRNA alternative polyadenylation. *Plant Cell* 31 (10), 2332–2352. doi: 10.1105/tpc.18.00545

Zan, Y., and Carlborg, Ö. (2019). A polygenic genetic architecture of flowering time in the worldwide arabidopsis thaliana population. *Mol. Biol. Evol.* 36 (1), 141–154. doi: 10.1093/molbev/msy203

Zeng, W., Dai, X., Sun, J., Hou, Y., Ma, X., Cao, X., et al. (2019). Modulation of auxin signaling and development by polyadenylation machinery. *Plant Physiol.* 179 (2), 686–699. doi: 10.1104/pp.18.00782

Zhang, M., Bo, W., Xu, F., Li, H., Ye, M., Jiang, L., et al. (2017). The genetic architecture of shoot-root covariation during seedling emergence of a desert tree, Populus euphratica. *Plant J.* 90 (5), 918–928. doi: 10.1111/tpj.13518

Zhang, Z., Fu, J., and Gilmour, D. S. (2005). CTD-dependent dismantling of the RNA polymerase II elongation complex by the pre-mRNA 3'-end processing factor, Pcf11. *Genes Dev.* 19 (13), 1572–1580. doi: 10.1101/gad.1296305

Zhang, X., Nomoto, M., Garcia-Leon, M., Takahashi, N., Kato, M., Yura, K., et al. (2022). CFI 25 subunit of cleavage factor I is important for maintaining the diversity of 3' UTR lengths in arabidopsis thaliana (L.) heynh. *Plant Cell Physiol.* 63 (3), 369–383. doi: 10.1093/pcp/pcac002

Zhang, Y., Ramming, A., Heinke, L., Altschmied, L., Slotkin, R. K., Becker, J. D., et al. (2019). The poly(A) polymerase PAPS1 interacts with the RNA-directed DNA-methylation pathway in sporophyte and pollen development. *Plant J.* 99 (4), 655–672. doi: 10.1111/tpj.14348

Zhang, Y. X., Sun, Y. D., Shi, Y. S., Walz, T., and Tong, L. (2020). Structural insights into the human pre-mRNA 3 '-end processing machinery. *Mol. Cell* 77 (4), 800–80+. doi: 10.1016/j.molcel.2019.11.005

Zhao, J., Hyman, L., and Moore, C. (1999). Formation of mRNA 3 ' ends in eukaryotes: Mechanism, regulation, and interrelationships with other steps in mRNA synthesis. *Microbiol. Mol. Biol. Rev.* 63 (2), 405–40+. doi: 10.1128/mmbr.63.2.405-445.1999

Zheng, L., Shang, L., Chen, X., Zhang, L., Xia, Y., Smith, C., et al. (2015). TANG1, encoding a symplekin_C domain-contained protein, influences sugar responses in arabidopsis. *Plant Physiol.* 168 (3), 1000–1012. doi: 10.1104/pp.15.00288

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

Zhu, Y., Wang, X., Forouzmand, E., Jeong, J., Qiao, F., Sowd, G. A., et al. (2018). Molecular mechanisms for CFIm-mediated regulation of mRNA alternative polyadenylation. *Mol. Cell* 69 (1), 62–74.e64. doi: 10.1016/j.molcel.2017.11.031

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 24 January 2024
DOI 10.3389/fpls.2023.1258042

OPEN ACCESS

EDITED BY
Manohar Chakrabarti,
The University of Texas Rio Grande Valley, United States

REVIEWED BY
Aamir W. Khan,
University of Missouri, United States
Dinakaran Elango,
Iowa State University, United States

*CORRESPONDENCE
Tusar Kanti Behera
tusar@rediffmail.com

RECEIVED 13 July 2023
ACCEPTED 29 December 2023
PUBLISHED 24 January 2024

CITATION
Vinay ND, Singh K, Ellur RK, Chinnusamy V, Jaiswal S, Iquebal MA, Munshi AD, Matsumura H, Boopalakrishnan G, Jat GS, Kole C, Gaikwad AB, Kumar D, Dey SS and Behera TK (2024) High-quality *Momordica balsamina* genome elucidates its potential use in improving stress resilience and therapeutic properties of bitter gourd.
Front. Plant Sci. 14:1258042.
doi: 10.3389/fpls.2023.1258042

High-quality *Momordica balsamina* genome elucidates its potential use in improving stress resilience and therapeutic properties of bitter gourd

N. D. Vinay[1], Kalpana Singh[2], Ranjith Kumar Ellur[3], Viswanathan Chinnusamy[4], Sarika Jaiswal[2], Mir Asif Iquebal[2], Anilabha Das Munshi[1], Hideo Matsumura[5], G. Boopalakrishnan[1], Gograj Singh Jat[1], Chittaranjan Kole[6], Ambika Baladev Gaikwad[7], Dinesh Kumar[2], Shyam Sundar Dey[1] and Tusar Kanti Behera[1,8]*

[1]Division of Vegetable Science, Indian Council of Agricultural Research (ICAR)-Indian Agricultural Research Institute, New Delhi, India, [2]Division of Agricultural Bioinformatics, Indian Council of Agricultural Research (ICAR)-Indian Agricultural Statistics Research Institute, New Delhi, India, [3]Division of Genetics, Indian Council of Agricultural Research (ICAR)-Indian Agricultural Research Institute, New Delhi, India, [4]Division of Plant Physiology, Indian Council of Agricultural Research (ICAR)-Indian Agricultural Research Institute, New Delhi, India, [5]Gene Research Centre, Shinshu University, Ueda, Japan, [6]Foundation for Science and Society, Kolkata, India, [7]Division of Genomic Resources, Indian Council of Agricultural Research (ICAR)-National Bureau of Plant Genetic Resources, New Delhi, India, [8]Indian Council of Agricultural Research (ICAR)-Indian Institute of Vegetable Research, Varanasi, Uttar Pradesh, India

Introduction: *Momordica balsamina* is the closest wild species that can be crossed with an important fruit vegetable crop, *Momordica charantia*, has immense medicinal value, and placed under II subclass of primary gene pool of bitter gourd. *M. balsamina* is tolerant to major biotic and abiotic stresses. Genome characterization of *Momordica balsamina* as a wild relative of bitter gourd will contribute to the knowledge of the gene pool available for improvement in bitter gourd. There is potential to transfer gene/s related to biotic resistance and medicinal importance from *M. balsamina* to *M. charantia* to produce high-quality, better yielding and stress tolerant bitter gourd genotypes.

Methods: The present study provides the first and high-quality chromosome-level genome assembly of *M. balsamina* with size 384.90 Mb and N50 30.96 Mb using sequence data from 10x Genomics, Nanopore, and Hi-C platforms.

Results: A total of 6,32,098 transposons elements; 2,15,379 simple sequence repeats; 5,67,483 transcription factor binding sites; 3,376 noncoding RNA genes; and 41,652 protein-coding genes were identified, and 4,347 disease resistance, 67 heat stress–related, 05 carotenoid-related, 15 salt stress–related, 229 cucurbitacin-related, 19 terpenes-related, 37 antioxidant activity, and 06 sex determination–related genes were characterized.

Conclusion: Genome sequencing of *M. balsamina* will facilitate interspecific introgression of desirable traits. This information is cataloged in the form of

webgenomic resource available at http://webtom.cabgrid.res.in/mbger/. Our finding of comparative genome analysis will be useful to get insights into the patterns and processes associated with genome evolution and to uncover functional regions of cucurbit genomes.

KEYWORDS

cucurbits, genome assembly, genomic resources, medicinal value, *Momordica balsamina*, stress tolerance

1 Introduction

Momordica balsamina (2n = 2x = 22), commonly referred to as Balsam apple, Southern Balsam pear, or African pumpkin, is a wild plant belonging to the *Momordica* genus within the *Cucurbitaceae* family (Bharathi and John, 2013). It is native to tropical regions of Africa, Asia, and Australia (Jeffrey, 1967; Mishra et al., 1986). *M. balsamina* has an annual to perennial life cycle and grows as a trailing herb (John, 2005; Behera et al., 2010). It grows better in hot, humid climates and prefers acidic soil (pH 5.0–6.5) (Mishra et al., 1986). Ellipsoid-shaped immature fruits of *M. balsamina* are rich in essential vitamins (A and C) and vital minerals (iron and calcium) (Wehner et al., 2020). Additionally, its leaves are abundant in carotenoids (Mashiane et al., 2022). These nutritionally and medicinally enriched fruits and leaves are consumed in rural areas of Africa and Asia (Flyman and Afolayan, 2007; Bharathi and John, 2013). It is one of the four *Momordica* species cultivated in India, primarily in the dry regions of the Northwest plains, Eastern Ghats, and Western Ghats (Peter and Abraham, 2007).

Balsam apple has the reputation of a "gifted plant" due to its richness in bioactive compounds, which offer diverse therapeutic benefits. These compounds exhibit wide spectrum of medicinal values, including anti-septic, anti-microbial, anti-bacterial, anti-viral (including anti-HIV), anti-inflammatory, anti-plasmodial, antioxidant, and analgesic properties (Hassan and Umar, 2006; Thakur et al., 2009). The extensive range of medicinal properties exhibited by *M. balsamina* can be attributed to its diverse array of terpenoid compounds, such as balsaminol, balsaminoside, balsaminagenins, karavilagenin, cucurbalsaminol, and balsaminapentaol (Ramalhete et al., 2009; Ramalhete et al., 2010; Ramalhete et al., 2011a; Ramalhete et al., 2011b). Numerous researches have been conducted on these compounds in order to highlight their potential medical uses. Additionally, "cucurbitacins" derived from *M. balsamina* were found to have selective antiproliferative activity against multidrug resistant cancer cells (Ramalhete et al., 2022). Furthermore, Balsam apple contains ribosomal-inactivating proteins (RIPs) such as Momordin II and Balsamin, which possess remarkable antiviral, anticancer, and antibacterial properties. These RIPs have found practical applications in the development of commercial drug preparations (Khare, 2007; Kaur et al., 2012; Ajji, 2016; Ajji et al., 2017). The findings from these aforementioned studies justify the immense potential of *M. balsamina* within the pharmaceutical industry, thus making it a subject of intense scientific research in the field of cucurbitaceous vegetable crops.

Momordica charantia, commonly known as Bitter gourd, is the most widely cultivated vegetable within the *Momordica* genus, renowned for its distinctive bitter taste, attributed by cucurbitane-type tri-terpenoids (Chen et al., 2005). The fruits of Bitter gourd are abundant in vitamin C and iron and exhibit high antioxidant activity (Behera et al., 2010). Beyond its culinary use, it finds extensive application in traditional medicine, alleviating stomach pain, anemia, malaria, coughs, and fever, and it is a renowned source of anti-diabetic drug in pharmaceutical industry (Tan et al., 2008; Krawinkel et al., 2018). Despite its biological and economic significance, the crop improvement and varietal development program in Bitter gourd have been hindered by the limited genetic diversity found in natural populations (Dhillon et al., 2016). Furthermore, bitter gourd, being a crop of tropics and subtropics, is affected by various biotic and abiotic stresses. To overcome these obstacles, there is a critical need for diverse and valuable genetic resources to facilitate the development of elite high-yielding and resilient bitter gourd varieties (Cui et al., 2020).

Among the seven *Momordica* species found in India, *M. charantia* and *M. balsamina* are the only two species with monoecious sex expression. These two species share same basic chromosome number of $x = 11$ and exhibit similar frequencies and ranges of bivalent and chiasmata formation. This high karyo-morphological similarity indicates a close ancestral relationship between these two species (Trivedi and Roy, 1972; Singh, 1990; Bharathi et al., 2011). *M. balsamina*, in particular, is considered the closest wild relative that can be crossed with Bitter gourd, falling under the II subclass of the primary gene pool of Bitter gourd (Bharathi et al., 2012). *M. balsamina* also possesses a high level of tolerance to like pests such as ladybird beetle (*Epilacna septima*), pumpkin caterpillar (*Margaronia indica*), red pumpkin beetle (*Aulocophora fevicoli*), gall fly (*Lasioptera falcata*), root-knot nematode (*Meladogyne incognita*), and diseases such as yellow mosaic and little leaf disease, making it an invaluable genetic resource for the improvement of *M. charantia* (Rathod et al., 2021). Hence, in addition to medicinal attributes, *M. balsamina* can serve as a potent genetic source of biotic stress resistance.

Interspecific hybridization has proven to be a successful method for harnessing natural genetic variation and transferring desirable

genes from wild relatives to cultivated crops (Bowley and Taylor, 1987; Dempewolf et al., 2017). In the Cucurbitaceae family, successful inter-specific hybrids have been developed within and between wild and cultivated taxa (Weeden and Robinson, 1986; Singh, 1991; Robinson and Decker-Walters, 1997). Likewise, there is great potential for the transfer of beneficial genes from *M. balsamina* to *M. charantia* for the genetic improvement of Bitter gourd. Previous studies have reported partial cross-compatibility between *M. charantia* and *M. balsamina*, resulting in progenies exhibiting normal meiosis (Singh, 1990; Bharathi et al., 2012). Recently, a detailed study on crossability involving 116 diverse Bitter gourd genotypes demonstrated success in six cross-combinations (Rathod et al., 2021). The study also confirmed the partial introgression of chromosome segments from *M. balsamina* into the Bitter gourd genome through morpho-cytological and molecular analysis of interspecific hybrids between *M. charantia* cv. Pusa Aushadhi × *M. balsamina* and their advanced generations (F_2 and backcross generations). These findings suggest the possibility of transferring genes or traits related to biotic resistance and medicinal properties from *M. balsamina* to *M. charantia*, producing high-quality and resistant Bitter gourd varieties.

The era of genomics-assisted vegetable breeding commenced with the completion of the cucumber whole genome assembly in 2009 (Huang et al., 2009). In 2016, the first draft genome of Bitter gourd was published (Urasaki et al., 2017), followed by subsequent high-quality, chromosome-level assemblies (Cui et al., 2020; Matsumura and Urasaki, 2020). With advancements in sequencing technologies and bioinformatics tools, genomic data for flowering plants has been expanding rapidly (Chen et al., 2018), and genome assemblies for most cultivated cucurbits are now available in the public domain. Presently, there is a focus on genome characterization of closely related cross-compatible crop wild relatives (CWRs).

CWRs serve as a dynamic gene pool to access vital genetic diversity needed for crop improvement. Earlier, molecular techniques were used to characterize CWR (Dillon et al., 2007a; Sotowa et al., 2013). Now, advanced next-generation sequencing (NGS) platforms can be utilized for genome characterization of CWR to study phylogeny and discover useful genes in order to support agriculture and food security (Brozynska et al., 2016). Several wild relatives of tomato (Sato et al., 2012), brinjal (Gramazio et al., 2019), potato (Aversano et al., 2015), and sweet potato (Wu et al., 2018) have already been sequenced. In the current study, we present first high-quality genome assembly of *M. balsamina a*, close relative of bitter gourd that can be a vital genetic resource to improve medicinal value and stress resistance in bitter gourd.

2 Material and method

2.1 Sample collection and DNA extraction

Young leaf samples of *M. balsamina* (IC-467683) weighing around 10 g were collected for DNA isolation from 30-day-old seedlings at the active vegetative stage during the early morning hours. The collected leaf samples were packed immediately in aluminium foil, frozen into liquid nitrogen and stored at −80°C. Total DNA was isolated using the modified cetyl trimethyl ammonium bromide (CTAB) method (Saghai-Maroof et al., 1984). The genomic DNA samples were adjusted to 50 ng DNA/µL and stored at 4°C until used for sequencing. The quality and quantity of the extracted DNA were estimated with an Eppendorf Biospectrometer confirmed by running on 0.8% w/v agarose gel.

2.2 10x genomics sequencing and library preparation

High-molecular weight DNA (1.25 ng) was loaded onto a Chromium Controller chip, along with 10x Chromium reagents and gel beads following manufacturers recommended protocols. Initial library construction occurred within droplets containing Gel Beads-in-Emulsion (GEMs) beads with unique barcodes. The library construction incorporated a unique barcode adjacent to read one. All molecules within a GEM got tagged with the same barcode. However, because of the limiting dilution of the genome (roughly 300 haploid genome equivalents), the probability that two molecules from the same region of the genome were partitioned in the same GEM was minimal. Thus, the barcodes were used to associate short reads with their source long molecule statistically. The resulting library was sequenced on Illumina HiSeq X Ten sequencer (San Diego, CA, USA) as per the manufacturer's protocol to produce 2 × 150 paired-end sequences. The entire process was performed on four replicates; thus, four pair-end libraries were prepared.

2.3 NanoPore sequencing and library preparation

First, 05-µg genomic DNA was sheared to approximately 15,000 bp by centrifugation at 5,200 rpm in a gTUBE. DNA was repaired with damage repair reagent and end-repaired using end-repair mix before ligation to nanopore blunt end adapter. Unligated material was digested with Exo III and Exo VII. Then, 12–25 Kb library fragments were purified via two consecutive Ampure cleanups, and size selection was done on Blue Pippin (SageScience, Beverley, MA, USA) with a 0.75% agarose cassette. An aliquot of 20 picomol of the final library was loaded onto the flow cell and sequenced on machine MinION (Oxford Nanopore Technologies, Oxford Science Park, United Kingdom) using Oxford Nanopore sequencing kit 2.0 and improved instrument workflow (Instrument Control Software 4.0).

2.4 Hi-C sequencing and library preparation

Fresh and young leaf samples were collected and cross-linked for 10 min with a 1% final concentration of fresh formaldehyde and quenched with a 0.2 M final concentration of glycine for 5 min. The cross-linked cells were subsequently lysed in lysis buffer. The extracted

nuclei were re-suspended with a 150-µL 0.1% Sodium dodecyl sulfate (SDS) and incubated at 65°C for 10 min. Furthermore, they were quenched by adding 120 µL of water and 30 µL of 10% Triton X-100 and incubated at 37°C for 15 min. The DNA in the nuclei was digested by adding 30 µL of 10x NEB buffer 2.1 and 150 U of MboI and incubated at 37°C for 12h. This was followed by inactivation of MboI enzyme at 65°C for 20 min and filling of cohesive ends by adding 1 µL of each 10 mM deoxythymidine triphosphate (dTTP), deoxyadenosine triphosphate (dATP), and deoxyguanosine triphosphate (dGTP), 2 µL of 5 mM biotin-14-deoxycytidine triphosphate (dCTP), and 4 µL (40 U) Klenow and after that incubated at 37°C for 2h. To start proximity ligation, 120 pL 10x blunt-end ligation buffer, 100 pL 10% Triton X-100, and 20U T4 DNA ligase were added and held at 16°C for 4h. This was followed by reversing of the cross-linking with 200 ug/mL proteinase K (Thermo Fisher Scientific) at 65°C for 12h. Furthermore, chromatin DNA manipulations were performed using a method described by Belaghzal et al. (2017), followed by DNA purification using QIAamp DNA Mini Kits (Qiagen) and shearing of purified DNA in length of 400 bp. Dynabeads MyOne Streptavidin C1 (Thermo Fisher Scientific) was used to pull down point ligation junctions. NEB Next Ultra II DNA library Prep Kit for Illumina (NEB) was used to prepare Hi-C library for Illumina sequencing. The final library was sequenced on the Illumina HiSeq X Ten platform (San Diego, CA, USA) as per the manufacturer's protocol with 2 × 150 paired-end mode.

2.5 Data pre-processing and genome assembly

All the raw reads of 10x Genomics, Nanopore and HiC libraries used in the present study have been submitted in National Center for Biotechnology Information (NCBI) with SRA IDs SRR21495983, SRR21495982, and SRR21495981, respectively. Figure 1 shows the outline followed during the present study. Prior to assembly, reads of these libraries were cleaned using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc: Andrews, 2010) by removing low quality reads at < 20 phred score, followed by adapter cleaning using TrimGalore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). *De-novo* genome assembly was performed using all the 10x Genomics libraries of four replicates using Supernova v2.1.1 (Weisenfeld et al., 2017). After this, Nanopore libraries were mapped on *de-novo* genome assembly for further scaffolding using npScarf (Cao et al., 2017). Finally, HiC libraries were mapped on improved genome assembly using Juicerv1.5 (Durand et al., 2016) to obtain the de-duplicated alignment file. Furthermore, scaffolding, editing, and polishing of assembly was performed using 3dDNA v180419 (Dudchenko et al., 2017). Finally, identification of chromosomes and editing of miss-assembly was performed using JuiceBox v1.11.08 (Robinson et al., 2018) to construct contact maps for chromosomes. Genome polishing was performed on final assembly using Pilon (Walker et al., 2014).

2.6 Validation of chromosome level assembly

To assess the quality of the assembled genome, assembly statistics were calculated using QUAST (Gurevich et al., 2013). Furthermore, validation of assembly was performed using BUSCO (Simao et al., 2015) to find the completeness and contamination within genome assembly. A comparative study of *M. balsamina* genome assembly with other related species, such as *Momordica charantia*, *Citrullus lanatus*, *Cucumis sativus*, and *Cucumis melo* was also performed.

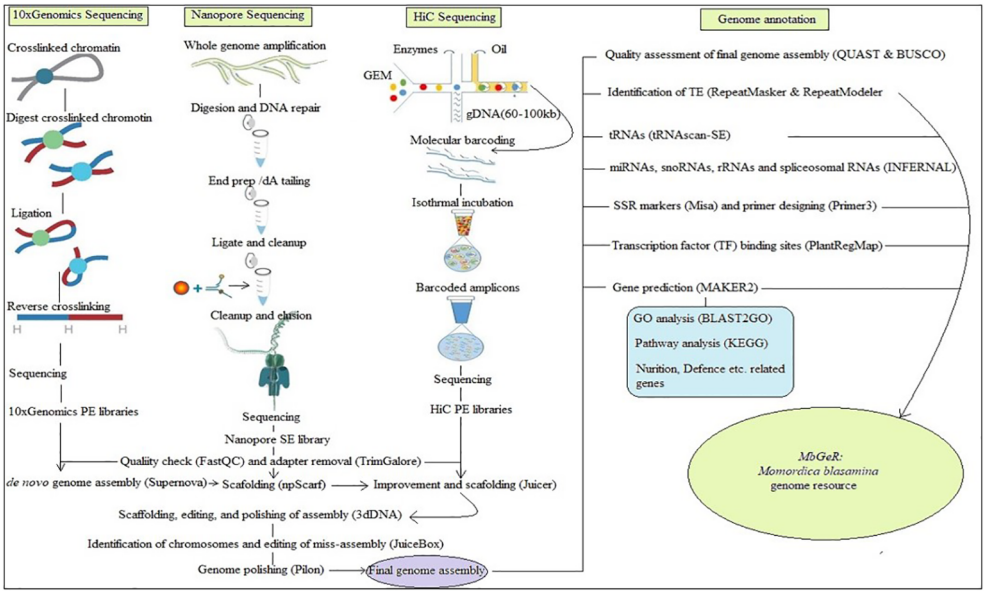


FIGURE 1
Over-view of pipeline adopted in the study.

2.7 Genome annotation

For genome annotation, a series of bioinformatics tools were employed. First, repeat regions of the assembled genome were masked using RepeatMasker v4.1.0 (http://www.repeatmasker.org/RMDownload.html). This was followed by the identification of transposable elements (TEs) using RepeatModeler (http://www.repeatmasker.org/RepeatModeler/) to find LINE, SINE, Simple Repeats, LTR elements, DNA elements, and so forth. The ncRNA-encoding genes were also identified from the assembled genome. Furthermore, tRNAs were identified using tRNA scan-SE v.1.3.1 (Chan and Lowe, 2019) with < 1 false positive per 15 gigabases. Other ncRNAs, such as microRNAs, snRNAs, rRNAs, and spliceosomal RNAs, were also identified using INFERNAL v1.1.4 (Nawrocki and Eddy, 2013) at default parameters., Protein-encoding genes were predicted using SEQing v0.1.45 (Lewinski et al., 2020), which is an automated pipeline of self-trained hidden Markov models (HMM) models and transcriptomic data for gene prediction by Glimmer HMM, SNAP, and AUGUSTUS and combining their results by MAKER2 in association with transcriptomic evidence of *Momordica charantia*. Finally, the predicted genes passed through Cluster Database at High Identity with Tolerance (CD-HIT) (Limin et al., 2012), clustering at 90% sequence similarity to extract non-redundant genes. Extraction of Single Sequence Repeat (SSR) markers was performed using MIcroSAtellites (MISA) (Beier et al., 2017), considering mononucleotide repeats motif with at least 10 repeats, dinucleotide with six, tri-, tetra-, penta-, and hexa-nucleotide with five repeats (Thiel et al., 2003). Compound microsatellites were defined as those with the interval between two repeats motifs ≤100 nucleotides in the previous reports (Zhao et al., 2017). Furthermore, primers were also designed for each of the SSR makers using Primer3 (Untergasser et al., 2012) with parameters 18–27 bp primer length, 57°C–63°C melting temperature, 30%–70% GC content, and 100–300 bp product size. Transcription factor (TF) binding sites were extracted using PlantRegMap (Jin et al., 2017).

2.8 Functional annotation of protein-coding genes

The predicted protein-coding genes were mapped against the NR database (updated May 2020) and the plant TF database (version 5.0) using NCBI blast (version 2.2.29+) (Lipman and Pearson, 1985) for functional annotation. Furthermore, gene ontology (GO) analysis was performed on predicted genes using Blast2GO (Conesa et al., 2005). Pathway analysis was performed using Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (Erxleben and Grüning, 2020).

2.9 Disease resistant, defence, stress, and sex expression–related genes

Disease resistant genes were identified by mapping proteins against the PRGDB database v.4.0 (García et al., 2021) with e-value cutoff of 1e-10 using BLAST (NCBI 2.2.29+) (Lipman and Pearson, 1985). Along with resistance genes, genes related to salt stress, heat stress and sex expression were also extracted.

2.10 Orthologous genes, phylogenetic, and synteny analysis

M. balsamina genes, orthologous with *M. charantia*, *Citrullus lanatus*, *Cucumis sativus*, and *Cucumis melo*, were predicted using OrthoMCL (Chen et al., 2006) based on a Markov Cluster algorithm to group (putative) orthologs utilizing all-against-all BLAST (Lipman and Pearson,1985) comparisons among protein sequences of considered species. For the detection of synteny between *M. balsamina* genome assembly and four other genome assemblies of abovementioned species was performed by SyMAP v4.2 (Soderlund et al., 2011). Synteny blocks shown as colored ribbons between the chromosomes arranged in circle were visualized in the circular plots using Circos (Krzywinski et al., 2009). Furthermore, micro-synteny, a synteny in small regions, identified between two or more genomic regions was performed between of *M. balsamina* and *M. charantia* genomes using McScan python version (Tang and Krishnakuar, 2015). Furthermore, a phylogenetic study was also performed among genome assemblies of *M. balsamina*, *M. charantia*, *Cucumis melo*, *Citrullus lanatus*, and *Cucumis sativus*. First, a multi-sequence alignment (MSA) was performed among genome assemblies using Multiple Alignment using Fast Fourier Transform (MAFFT) (Katoh et al., 2002). Later, a distance matrix was calculated among assemblies using MSA with Poisson correction method, >70% site coverage, and <30% alignment gaps, missing data, and ambiguous bases by ClustalW2 (Thompson et al., 1994). Finally, a phylogenetic tree was constructed using Neighbor-Joining method by ClustalW2.

2.11 Development of *M. balsamina* genomic resource

A web-genomic resource for *M. balsamina*, named MbGeR, was developed using all the results obtained from the genomic data analyses performed in the present study. MbGeR catalogs the information related to molecular markers such as SSRs, transposons elements (TEs), TF sites, ncRNAs and genes. It is based on a three-tier architecture, namely, client tier, middle tier, and database tier, developed using PHP, MySQL, HTML, and Apache. Web pages are developed using PHP and HTML in order to browse MbGeR and put up queries by users in client tier. All the information regarding transcripts, Differentially Expressed Genes (DEGs), markers, and so forth. are placed in different tables in MySQL database in the database tier. The scripting of client query page was done in PHP and HTML for execution and fetching in the middle tier. The web hosting was performed using Apache server. The bitter melon web resources are available at http://webtom.cabgrid.res.in/mbger/.

3 Result

3.1 Data pre-processing, genome assembly, and comparative analysis

In the present study, the whole genome of *M. balsamina* was assembled using reads obtained from three different platforms: Oxford Nanopore, 10 X and Hi-C. A combination of multiple technologies is reported to improve the quality and completeness of genome assembly (Wang et al., 2023). An average of 27,767,526; 2,331,456; and 168,098,715 reads were accessed in 10x Genomics, Nanopore, and Hi-C libraries, respectively after pre-processing and quality check. Supplementary Table S1 shows the detailed read statistics in different replicates and their average length in all three libraries. GC% was 39 for 10x Genomics and Hi-C read libraries, while Nanopore reads had 35% GC content.

De-novo genome assembly was generated using 10x Genomics libraries followed by mapping of Nanopore libraries onto *de-novo* genome assembly for further scaffolding. The nanopore raw read size ranged from 1000 bp to 222917 bp, with N50 (minimum length representing half of the total length of the assembly) as 26.08 Kb and 15.29 Mb for raw reads and scaffolds, respectively. Then, reads from HiC libraries were used for chromosome-level scaffolding, which is considered as the best choice for capturing the longest range DNA connectedness (Wang et al., 2023).

The genome assembly of *M. balsamina* and its assessment was found to have 3,710 scaffolds of 384,902,967 bp length and N50 of 30,984,295 bp (Table 1). BUSCO analysis, which uses universal single-copy orthologs, is considered as high-resolution quantifications of genomes, which facilitate informative comparisons and provides suggestions for improvements to assemblies or annotations (Simao et al., 2015). Assessment of this generated assembly shows 2,266 (97.4%) of 2,326 BUSCO to be complete and single copy (Table 1). The comparative statistics of *M. balsamina* assembly with other assemblies of related species showed the assembly size to be comparable with others while the N50 value (30.96 mb) was much improved than other assemblies (Table 2).

TABLE 1 *M. balsamina* assembly statistics.

Assembly parameters	Statistics
# contigs	3,710
# contigs (≥ 1000 bp)	3,702
# contigs (≥ 10000 bp)	569
# contigs (≥ 100000 bp)	67
# contigs (≥ 1000000 bp)	11
Largest contig (bp)	40,892,414
Average length (bp)	103,747
Smallest contig (bp)	957
Total length (bp)	384,902,967
Total length (≥ 1000 bp)	384,902,967
Total length (≥ 10000 bp)	384,902,967
Total length (≥ 100000 bp)	353,358,081
Total length (≥ 1000000 bp)	353,358,081
N50 (bp)	30,984,295
N75 (bp)	27,371,744
L50	6
L75	9
Total GC Content	35.43%
BUSCO	
Complete BUSCOs (C)	97.4%
Complete and single-copy BUSCOs (S)	95.3%
Complete and duplicated BUSCOs (D)	2.1%
Fragmented BUSCOs (F)	0.7%
Missing BUSCOs (M)	1.9%
Total BUSCO groups searched (*n*)	2326
Noncoding RNA	
tRNA	1,823
rRNA (large + small subunit)	270
sRNA	1
miRNA	150
spliceosomal RNA	129
snoRNA	961
Antisense RNA	15
SRP RNA	27

3.2 Annotation of genome assembly

Genome annotation is crucial to facilitate the utilization of assembled genomes in genetic studies. In the current study, homology-based inference, *in-silico* prediction techniques and merged transcriptomics data (of *Momordica charantia*) are merged into a single concordant annotation (Yandell and Ence, 2012). Genome annotation was done to identify TEs, ncRNA encoding genes, tRNAs, ncRNAs, SSR makers, TF binding sites and protein-encoding genes in the assembled genome.

Out of the total 384,902,967 bp length of 3,710 scaffolds of the assembled genome, 218,862,155 (56.73%) bases were masked. Frequencies of various classes of predicted TEs in genome assembly are delineated in Table 3. A significant proportion of TE class belonged to LTR elements, while 22.29% were found to be unclassified. The frequency of SINEs was the least (0.05%), while it was 3.02% for SINEs. A sum of 567,483 TF binding sites were predicted in *M. balsamina* genome and Figure 2A is showing chromosome wide distribution of TF binding sites. Maximum number of TF binding sites were observed in chromosome

TABLE 2 Comparative statistics of *M.balsamina* genome assembly with genome assemblies of related species.

Assembly statistics	*Momordica. balsamina* (Current study)	Bitter gourd (OHB3-1)	Bitter gourd (Dali-1)	Bitter gourd (long read assembly)	Cucumber	Musk melon	Water melon	Bottle gourd
Genome size (Mb)	384.9	285.5	293.6	302.9	243.5	375	353.5	313.4
Chromosomes (Mb)	349.27	172.0	251.3	291	177.3	316.3	330.0	308.1
Unknown scaffolds (Mb)	35.63	60.2	85.5	96.27	72.8	87.5	93.5	98.3
N50 (Mb)	30.96	1.1	3.3	25	1.1	4.7	2.4	8.7
GC content (%)	35.4	36.4	35.4	–	32.2	33.2	32.8	–
Predicted genes	41652	45859	26427	–	26682	27427	23440	22472
Masked (%)	56.73	34.7	41.5	52.52	20.8	35.4	39.8	46.9
LTR content (%)	26.82	27.4	31.8	23.97	11.5	25.0	30.5	39.8

number 2 (~12%), followed by chromosome number 1 (~9%) and chromosome number 11 (~9%). Almost ~12% of TF binding sites were associated with the remaining unknown scaffolds (Figure 2A).

A total of 2,15,379 SSR markers were mined from the assembled genome. The highest number of SSR belonged to motif type mono-nucleotide (~69%), followed by di (~13%) and tri (~6%). A total of 29,618 (~9%) SSRs were compound type (Figure 2B). A total 3,376 different non-coding RNA genes were predicted in *M. balsamina* assembly, out of which 1,823 tRNA, 270 rRNA, 150 microRNA, 961 snoRNA, 27 SRP RNA, and 129 spliceosomal RNA genes were predicted (Table 1). Out of the total 1,823 predicted tRNA genes in *M. balsamina* assembly, their frequency distribution over chromosome 1 was highest, followed by chromosomes 3 and 2. A minimum number of tRNA genes were observed in chromosome 4 (Figure 2C). Apart from the chromosomes, higher number of tRNA genes were found localized on unknown scaffolds. Figure 2D shows the frequencies of protein-coding genes distributed over various chromosomes along with 74 pseudogenes predicted in *M. balsamina* assembly. It was observed that a higher number of protein-coding genes were found on chromosomes 1 (4,592), followed by chromosome 2 (4,410) and 3 (3,909).

TABLE 3 Frequencies and proportion of various classes of TEs predicted in *M. balsamina* assembly.

Classes of TEs	Frequency	Length (bp)	Proportion
SINEs	746	181264	0.05%
LINEs	29,398	11,643,197	3.02%
LTR elements	101,274	103,473,791	26.82%
DNA transposons	18,021	9,880,729	2.56%
Simple repeat	125,651	4,585,002	1.19%
Low complexity	26,644	1,304,816	0.34%
Unclassified	322,593	85,978,857	22.29%

3.3 Functional annotation of protein-coding genes

Functional annotation of protein-coding genes yielded a total of 33,450 genes that were annotated with NR database. GO analysis of these annotated genes showed 52 GO terms to be associated with 20,525 genes, of which 16, 12, and 25 were from cellular component, molecular function, and biological process classes, respectively. The GO terms were categorized into three classes, namely, molecular function, biological functions and cellular components. Figure 3A shows the GO terms associated with more than five protein-coding genes predicted in *M. balsamina* assembly. It was found that the GO terms named binding activities (11,892) followed by the catalytic activities (9,604) and transporter activities (889) were associated with most genes in molecular function class. In biological processes, cellular processes GO term (8,650) was the most frequent in genes, followed by metabolic processes (8,458) and biological regulations (1,252). Cell (5,026), cell part (5,026), and membrane (4,850) GO terms were the most frequent in cellular component class (Figure 3A). Figure 3B shows the top 10 KEGG pathways associated with 3,414 annotated genes in *M. balsamina* assembly. It was found that metabolic pathways (>1,500 genes involved) were the most abundant pathway, followed by biosynthesis of secondary metabolites (~700 genes involved) and microbial metabolism (~250 genes involved) in diverse environments.

3.4 Genes related to plant defence, medicinal properties, and sex expression

M. balsamina is well-known for its biotic and abiotic stress tolerance and medicinal properties. In the *M. balsamina* assembly, a total of 4,347 important disease resistance genes (R genes) were identified, out of which 1,174 genes encoded for nucleotide-binding site–leucine-rich repeat (NBS-LRR) domains along with 858 *RLP* and 273 *RLK* encoding genes, which are well known in resistance

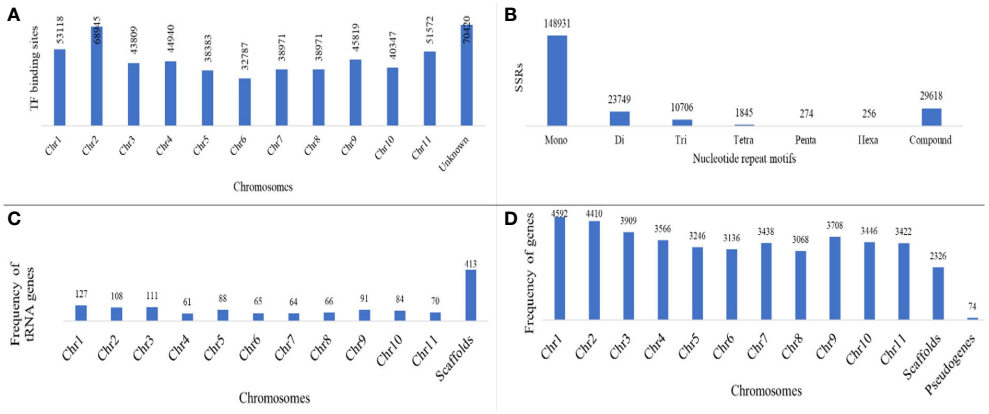


FIGURE 2
(A) Distribution of TF binding sites on different chromosomes; **(B)** frequency of SSRs of different nucleotide repeat motifs; **(C)** frequency of predicted tRNA genes and **(D)** protein coding genes distributed over chromosomes along with pseudogenes predicted in *M. balsamina* assembly.

response in plants. In our study, we identified 67 heat stress-related genes, including a total of 34 heat stress factor genes (HSFs), contribute to thermo-tolerance through the regulation of heat shock proteins (HSPs). In addition, 29 *HSP* genes, predominantly encoding HSP70, and 17 small heat proteins (HSP20) were identified in the *M. balsamina* assembly. Similarly, 15 genes encoding proteins related to salinity tolerance in the *M. balsamina* assembly, including alkaline ceramidase (ACER), S-acyltransferase, salt stress root protein RS1-like, and protein RICE SALT SENSITIVE 3 isoforms were identified. Cucurbit crops are considered as models for deciphering the mechanism of sex determination in monoecious plant species and ethylene is considered to be the core regulator. To shed more light on this, in the current study, 06 genes related to ethylene biosynthesis were extracted. *M. balsamina* contains a diverse array of Cucurbitacin terpenoid compounds exhibiting anti-septic, anti-microbial, anti-bacterial, anti-viral (including anti-HIV), anti-inflammatory, anti-plasmodial, antioxidant, and analgesic properties (Thakur et al., 2009; Ramalhete et al., 2022). The genes related to terpenoid biosynthesis were searched in the genome to elucidate the mechanism behind the medicinal property exhibited by this species. Thirty-seven antioxidant activity related and 229 genes related to the biosynthesis of cucurbitacin, the key factors behind medicinal attributes of the *M. balsamina*, were detected. Table 4 shows the frequencies of genes extracted with provided functions. GO terms of pathogenesis-related genes, heat tolerant genes, salt tolerance-related genes, sex determination-related genes, triterpenoid-related genes, cucurbitin-related genes, nutrition-related genes, and phloem-related genes are graphically represented in Supplementary Figure S1.

3.5 Orthologous genes, phylogenetic, and synteny analysis

Comparative genetic parameters such as orthology, synteny, and phylogeny were utilized in the study to understand the genome composition, evolution and relatedness among the members of a

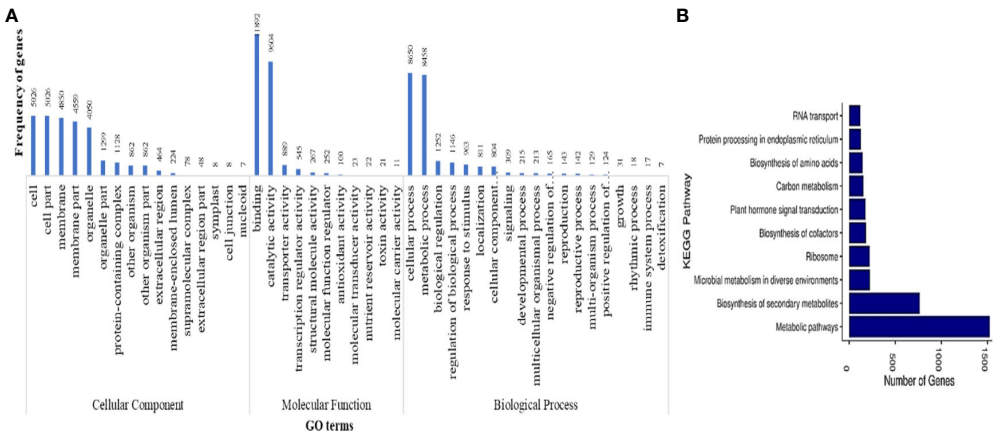


FIGURE 3
(A) GO terms associated with predicted protein coding genes and **(B)** top 10 KEGG pathways associated with annotated protein coding genes in *M. balsamina* assembly.

TABLE 4 Frequencies of genes associated with disease resistance, defence, salt stress, heat stress, sex determination, and secondary metabolite synthesis identified in *M. balsamina* assembly.

Function	Class	Description	Genes
Disease resistance	C	Coiled-coil domain	01
	CK	Coiled-coil and Kinase domains	379
	CL	Coiled-coil and LRR domains	59
	CLK	Coiled-coil, LRR and Kinase domains	23
	CN	Coiled-coil and NBS domains	197
	CNK	Coiled-coil, NBS and Kinase domains	129
	CNL	Coiled-coil domain, a nucleotide binding site and a leucine-rich repeat (CC-NB-LRR)	107
	CNT	Coiled-coil, NBS and TIR domains	02
	CT	Coiled-coil and TIR domains	24
	CTL	Coiled-coil, TIR, LRR domains	06
	CTNL	Coiled-coil, TIR, NBS and LRR domains	16
	KIN	Kinase domain	1376
	L	LRR domain	32
	N	NBS domain only, lack of LRR	388
	NL	NBS domain at N-terminal and LRR st the C-terminal, and lack of the CC domain	408
	RLK	Kinase domain, and an extracellular leucine-rich repeat (Kin-LRR)	273
	RLP	Receptor-like protein, groups those with a receptor serine-threonine kinase-like domain, and an extracellular leucine-rich repeat (ser/thr-LRR)	858
	T	TIR domain only, lack of LRR or NBS	46
	TN	TIR and NBS domains	12
	TNL	Toll-interleukin receptor-like domain, a nucleotide binding site and a leucine-rich repeat (TIR-NB-LRR)	62
	TRAN	Transmembrane helix domain	17
	Others	–	06
Heat stress tolerance	–	Hsp20, hsp70, hsfb1, hsfa2, hsfa4, hsfb4, hsf-a6	67
Salt stress tolerance	–	Alkaline ceramidase (ACER), S-acyltransferase, Salt stress root protein RS1-like, Protein RICE SALT SENSITIVE 3 isoforms	15
Sex determination	–	ACS (1-aminocyclopropane-1-carboxylate synthase)-1, ACS-7, ACS-CMA101, ACS-CMW-33	06
Secondary metabolite sysnthesis	Carotenoids (Nutrition)	Chloroplast specific lycopene beta cyclase, Phytoene destaurase/phytoene dehydrogenase, Prolycopene isomerase, Zeta-carotene desaturase, Lycopene epsilon cyclase.	05
	Cucurbitacins (Defence)	Oxidosqualene cyclase (OSC), Cytochrome P450 (CYP), Acetyltransferase (ACT), UDP-glucosyltransferase (UGT)	229
	Trierpenoids (Medicinal use)	Balsaminol, Balsaminoside, Balsaminagenin, Karavilagenin, Cucurbalsaminol, Balsaminapentaol, Megastigmane-type nor-isoprenoid, Pimarane-type diterpenes	19
Antioxidant activity	Abiotic stress tolerance	Glutatione S transferase (GST)	37

family or clade at the nucleotide/molecular level. A total of 1,542 genes of *M. balsamina* were found orthologous with other related species considered in the present study. Frequencies of these genes are provided in Table 5 along with the species with which these are found orthologous. The unique and overlapping *M. balsamina* genes found orthologous in other related species are delineated in Figure 4A. It is observed that 165, 159, 953, and136 *M. balsamina* genes were orthologous in *Cucumis melo*, *Citrullus lanatus*, *M. charantia*, and *Cucumis sativus*, respectively, only and the rest of the genes were orthologous in more than two species.

The syntenic relationship analyses of *M. balsamina* with other species were performed. In the synteny analysis, the sequences of related species were aligned, and conserved genes between the two genomes were identified as anchors, and then regions with more than seven anchors connecting two species were considered as synteny blocks. Frequencies of orthologous genes and syntenic blocks of *M. balsamina* with related species, *M. charantia*, *Citrullus lanatus*, *Cucumis sativus*, and *Cucumis melo* were found to be (8845, 306), (8308, 264), (8265, 245), and (8092, 282), respectively (Table 5). Also, the diagrammatic representation of syntenic blocks in the form of Circos figures is provided for synteny between *M. balsamina* and *Cucumis sativus*, *M. balsamina* and *Cucumis melo*, *M. balsamina* and *Citrullus lanatus*, *M. balsamina* and *M. charantia* (all scaffolds), and *M. balsamina* and *M. charantia* (scaffolds >100Mb), respectively (Supplementary Figures S2A–E). A general absence of a one-to-one relationship in the chromosomes between the *Momordica balsamina* and other cucurbit genomes was observed. However, syntenic loci of one chromosome of *Momordica balsamina* chromosome exhibited a syntenic relationship between one or two chromosomes of studied

TABLE 5 Frequencies of *M. balsamina* orthologous genes and syntenic blocks found in other related species.

Species	*M. balsamina* orthologous genes	*M. balsamina* syntenic blocks
Momordica charantia	8,845	306
Citrullus lanatus	8,308	264
Cucumis sativus	8,265	245
Cucumis melo	8,092	282

cucurbits. *Momoridica balsamina* Chr11 was syntenic to Chr6 and Chr7 of *Cucumis sativus* and Chr5 was syntenic to Chr3 and Chr4 of *C. sativus*. Similarly, Chr8 of *Momordica balsamina* was syntenic to Chr 11 of *C. melo*. Furthermore, Chr7 was colinear to Chr 2 and 12 of Melon. Synteny between *M. balsamina* Chr 7 and Chr2 of watermelon was observed. Furthermore, Chr 5 was syntenic to Chr5 and Chr7 of watermelon.

Maximum number of genes on each chromosome of *M. Balsamina* found homologous with genes on corresponding scaffolds of *M. charantia* are shown in Supplementary Table 2. In addition, the Supplementary Figures S3A–K show homologous genes on chromosomes 1–11 of *M. balsamina* with syntenic relationship with corresponding scaffolds *M. charantia*. The rooted phylogenetic tree was constructed to represent the phylogenetic relationship of *M. balsamina* with other related species, namely, *M. charantia*, *Cucumis melo*, *Cucumis sativus*, and *Citrullus lanatus* (Figure 4B). *M. balsamina* was observed to be more closely related to *M. charantia*.

3.6 Development of *M. balsamina* web-genomic resource

A web genomic resource for *M. balsamina*, named MbGeR, was developed from the output obtained after genomic data analyses of *M. balsamina* genome in the present study. Its web interface includes a home page with an introduction to MbGeR with horizontal and vertical tabs including statistics, SSRs, TEs, TF sites, ncRNAs, genes and team, each of which is linked to their respective pages (Figure 5). The statistics page provides summary statistics of data provided in genome resources in the form of histograms. Users are provided with flexible options to select SSR data on the desired 11 chromosomes of *M. balsamina* along with desired motifs on SSRs page. Users can choose TEs from the TEs page according to their desired types and chromosome numbers. TF sites provide options to choose TF binding sites on the desired chromosome. On the ncRNAs page, users can select non-coding RNAs among the various types. Gene's page has two options: (i) selection of chromosomes for all genes extracted from the genome and (ii) choice of extracted genes associated with a certain function. Once the desired options are submitted on each of the mentioned the page, the output is displayed in tabular form in desired combinations of options. The Team page provides information and hyperlinked profiles of the team members involved in the study. The bitter gourd web resources, MbGeR is available for non-commercial use for research community at http://webtom.cabgrid.res.in/mbger/.

4 Discussion

CWRs are the primary source of diversity for utilization in crop improvement. Specifically, in crops with narrow genetic bases, the lack of diversity becomes the major bottlenecks in breeding program. To address the issue, close wild relatives inter-fertile with the cultivated crop species can be used as extended gene pool in crop improvement (Brozynska et al., 2016). CWRs evolve continuously in the natural environment and, hence, serve as a dynamic resource to access desirable genes to overcome several challenges in agriculture posed by increasing human population and climate change. Several workers have documented the wide-scale use of CWR to enhance agriculture production

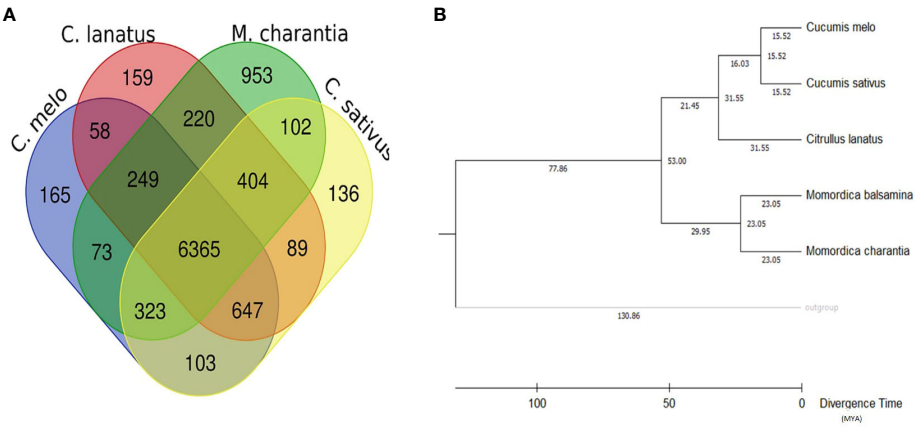


FIGURE 4
(A) Unique and overlapping *M. balsamina* genes found orthologous in other related species (*Cucumis sativus*, *Cucumis melo*, *Citrullus lanatus*, and *M. charantia*); **(B)** rooted phylogenetic tree represented in terms of divergence time (MYA: million years ago) based on whole genome assemblies of *M. balsamina* and other related species (*Cucumis sativus*, *Cucumis melo*, *Citrullus lanatus*, and *M. charantia*).

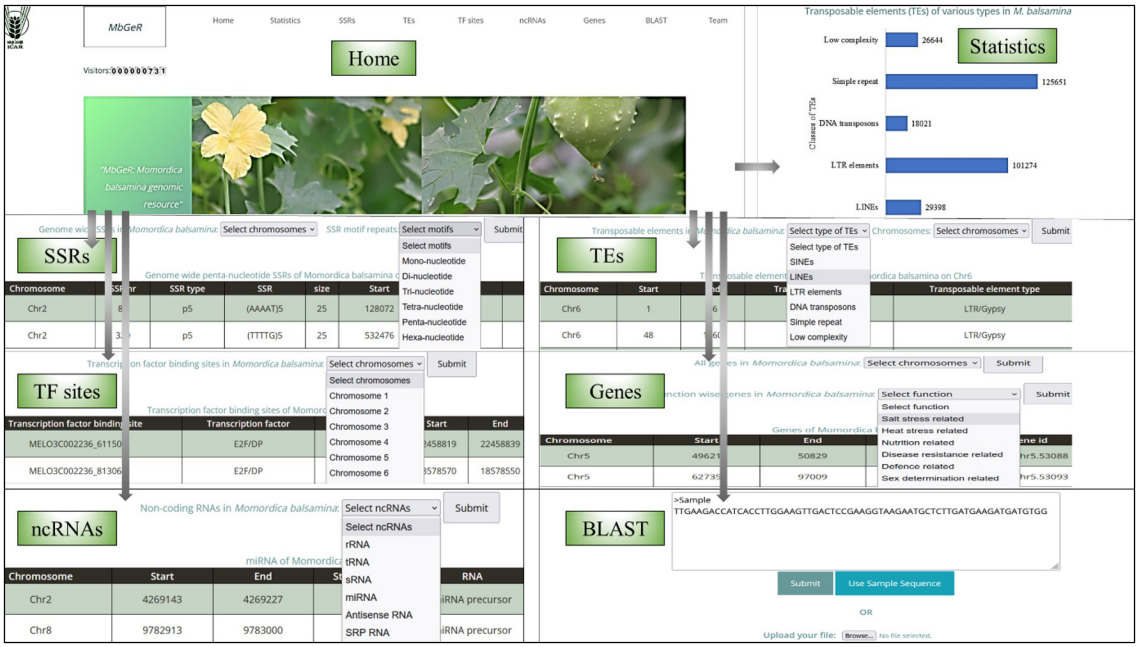


FIGURE 5
Various interfaces of *Momordica balsamina* web-genomic resource (*MbGeR*).

(Maxted et al., 2012; Fitzgerald, 2013; Dempewolf et al., 2014; Kell et al., 2015). It was estimated that about 30% of increased crop yields in the late 20th century can be attributed to the use of CWR in plant breeding programs (Pimentel et al., 1997). Hence, there is an increased need for the conservation and characterization of wild germplasm to utilize in crop improvement programmes. Molecular tools [e.g., simple sequence repeat (SSR) markers or microsatellites] were used in the past to characterize the CWR and to establish the relationship between wild and domesticated species (Dillon et al., 2007a; Sotowa et al., 2013). Recent DNA sequencing technology advancements increase opportunities to understand species at the whole-genome level (Edwards and Henry, 2011). Hence, genomic tools serve as the best strategy to characterize CWR and elucidate phylogenetic relationships between species, so that wild genetic diversity can be used in crop improvement (Kasem et al., 2010).

M. balsamina, Balsam apple is the closest wild species with cross-compatibility with *M. charantia*, exhibits greater tolerance to biotic stress, and possesses medicinal qualities (Rathode et al., 2021). Therefore, it is an ideal candidate species for harnessing natural variation within the primary gene pool and transferring desirable genes to cultivated *M. charantia*. Hence, genome characterization of this species proved vital for its usage in future breeding programs. In this study, we present the world's first high-quality chromosome-level genome assembly of *M. balsamina*, with a genome size estimate of 384.90 Mb and an N50 of 30.96 Mb. This study used reads from multiple platforms (Oxford Nanopore, 10 X and Hi-C), which facilitates chromosome-level scaffolding with improved base accuracy. This assembly will facilitate targeted gene introgression between *M. balsamina* and *M. charantia*, enhancing tolerance and medicinal properties. Furthermore, this assembly, a combination of multiple technologies, can be used to improve further the quality and completeness of genome assembly of related species (Wang et al., 2023).

Approximately 89.44% (345 Mb) of the assembly was anchored on 11 chromosomes, while the remaining scaffolds remained un-localized. The quality of this assembly, based on the N50 value, surpasses that of previously published assemblies for other members of the *Cucurbitaceae* family, such as cucumber (Huang et al., 2009), melon (Garcia-Mas et al., 2012), watermelon (Guo et al., 2013), and bitter gourd (Cui et al., 2020 and Matsumura and Urasaki, 2020). Additionally, the BUSCO analysis revealed that the *M. balsamina* assembly contains 97.4% conserved core genes, a higher percentage compared to other *M. charantia* assemblies [M. cDali-11 (95.9%), TR (95.5%), and OHB3-1 (82.20%)] and related species: *C. lanatus* (86.50%), *C. melo* (86.9%), *Cucurbita pepo* (92.8%), *C. sativus* (94.8%), and *Lagenaria siceraria* (88.2%) assemblies (Cui et al., 2020). Gene space completeness as measured by single-copy standards, including universal single-copy orthologs (BUSCOs) and core gene families (CoreGFs) are widely used for evaluating genome assembly and annotation for its completeness and quality (Vaattovaara et al., 2019). Using estimates of gene content from hundreds of species and guided by evolution, BUSCO assessments provide comprehensible metrics to assess the completeness of genome and. hence it is considered high-resolution quantifications of the genomes (Simao et al., 2015). Therefore, with a high BUSCO score (97.4%), this assembly provides a comprehensive representation of the *M. balsamina* genome and serves as a valuable reference for studying the genome architecture and evolution of related cucurbits, including its closest cultivated species, *M. charantia*. The assembled genome of *M. balsamina* will aid in the identification of a greater number of genome-wide markers, allowing for the specific and accurate tracing of

introgressed segments, which is crucial in interspecific introgression breeding, as reported by Qin et al. (2021). The assembly also revealed the presence of 632,098 TEs; 215,379 SSRs; 3,376 noncoding RNAs (ncRNAs); 567,483 TF binding sites; and 41,652 protein-coding genes. Many of these genes are associated with disease resistance (4421), heat stress tolerance (67), salt stress tolerance (15), carotenoid biosynthesis (05), cucurbitacin biosynthesis (229), terpenes related (19). antioxidant activity (37), and sex determination (06). Identifying these genes provides insights into the defence mechanisms, nutritional properties, and stress responses of *M. balsamina*.

TEs are well recognized for their role in genome evolution and regulation, providing alternative promoters, novel exons, neo-functionalization, and extensive rearrangements (Hoen & Bureau, 2015). A Comparison of our study's assembly with recent studies on *M. charantia* assemblies by Cui et al., 2020, and Matsumura and Urasaki, 2020, revealed an improvement in the genome size of approximately 95 Mb and 84 Mb, respectively. This enhancement could be attributed to a higher repeat content in the *M. balsamina* genome than *M. charantia*. Our findings supported this hypothesis, as we observed that 56.73% (218.86 Mb) of the *M. balsamina* assembly was masked as TEs, which was higher than the percentages reported for *M. charantia* (52.52%), cucumber (20.8%), watermelon (39.8%), and muskmelon (35.4%) assemblies. LTR repeat content (26.82%) was the most abundant in *M. balsamina* genome. Higher LTR repeats are a common feature of cucurbit genomes evident from genomes of cucumber, melon, and watermelon (Huang et al., 2009; Garcia-Mas et al., 2012; Guo et al., 2013). In addition to this, the findings of the current experiment support the results of the previous studies on genome characterization of bitter gourd done by Urasaki et al., 2017 and Cui et al., 2020, which reported a higher accumulation of repeat content in the *Momordica* genus compared to *Cucumis* and *Citrullus*, particularly LTR repeats. However, LTR repeat content in the *M. balsamina* genome was less than in the Watermelon (30.5) and Bottle gourd (39.8). Earlier studies also speculated a differential rate of LTR retro transposon accumulation in the cucurbits as the reason behind the difference in the genome size among cucurbits. For instance, a higher accumulation of LTR retrotransposons is found in sponge gourd (Wu et al., 2020) and watermelon genome (Guo et al., 2013) than in cucumber (Huang et al., 2009). Hence, with absence of WGD (whole genome duplication) in cucurbits, TE might be playing vital role in genome expansion (Wu et al., 2020).

In our study, 3,376 noncoding RNA genes were annotated in the *M. balsamina* assembly. Extracted miRNAs, tRNAs, rRNAs, and other noncoding genes can be important resources for further studies. Additionally, we predicted 41,652 protein-coding genes in the *M. balsamina* assembly, a number comparable to the *M. charantia* OHB3-1 assembly (45859) by Urasaki et al. (2017), and significantly higher than the assemblies of *M. charantia* Dali-1 (26,427) by Cui et al., 2020, cucumber (26,682) by Huang et al. (2009), melon (27,427) by Garcia-Mas et al. (2012), and watermelon (23,440) by Guo et al. (2013). The variation in gene numbers could be attributed to the utilization of different transcript information during the annotation of genome assemblies or the loss of genetic diversity due to the domestication of cucurbits. Functional annotation of the protein-coding genes in our study revealed the presence of essential genes associated with detoxification, antioxidant activity, toxin activity, response to stimuli, immune system processes, defence, nutrient reservoir activity, and nutritional properties. These genes were also associated with pathways such as biosynthesis of secondary metabolites, plant hormone signal transduction, and protein processing in the endoplasmic reticulum.

M. balsamina is resistant to significant pest and diseases affecting cucurbits (Rathod et al., 2021). To understand the molecular basis for pest and pathogen resistance three major classes of *R*/resistance genes were searched in the genome. In the *M. balsamina* assembly, we identified 4,347 disease resistance genes (R genes), out of which 1,174 genes encoded NBS–LRR domains. These genes were grouped into two subfamilies based on the presence of either the toll/interleukin-1 receptor (TIR) domain or the coiled-coil (CC) domain at the N-terminal region, as described by Tameling et al. (2002). Additionally, we identified 858 *RLP* and 273 *RLK* encoding genes involved in conferring resistance response. These genes, such as *Cf* family proteins in tomatoes conferring resistance against *Cladosporium fulvum* fungus (Jones et al., 1994; Thomas et al., 1997) and *HcrVf2* in apples conferring apple scab resistance (Belfanti et al., 2004), were found in lower numbers compared to melon and cucumber. The number of R genes identified in *M. balsamina* was much higher than reported in bottle gourd, watermelon, cucumber, and melon. However, cucurbits generally have fewer NBS-LRR encoding genes than *Arabidopsis* (Baumgarten et al., 2003) and rice (Goff et al., 2002). Only 61 NBS containing resistance were found in the cucumber genome (Huang et al., 2009). Likewise, out of 411 genes associated with disease resistance in melon only 81 disease resistance genes encoded *NBS*, the *LRR* and the *TIR* domains (Garcia-Mas et al., 2012). Similarly, only 44 *NBS-LRR* genes were found in watermelon genome (Guo et al., 2013). So, in general, C*ucurrbitaceae* genomes possess comparatively a smaller number of R genes encoding *NBS-LRR* proteins (Lin et al., 2013). Hence, other mechanisms might be involved in stress response. For instance, in cucumber and *LOX* gene family expansion is speculated as the possible complementary mechanism to cope with pathogen invasion (Huang et al., 2009). However, in *M. balsamina*, it seems the defence mechanisms works through the involvement of "R" genes like the majority of crop plants. The variation in the number of *R* genes in cucurbits suggests that they are not conserved, and the differential expansion of NBS-encoding families could be attributed to segmental and whole-genome duplications during the evolution of plant species, as suggested by Wang et al. (2009). The higher number of *R* genes in the *M. balsamina* assembly suggests their potential use in improving resistance to a wide variety of prevalent biotic stresses in its closest relative, *M. charantia*.

In our study, we identified a total of 34 HSFs in the *M. balsamina* assembly, which was higher than the numbers reported for rice (25) by Chauhan et al. (2011), Arabidopsis (21) by Nover et al. (2001), and cucumber (23) by Chen et al. (2021). Among these genes, the primary heat stress factors identified were *HSFB1* (01), *HSFA2* (03), *HSFA4* (04), *HSFB4* (04), and *HSF-A6* (04), which contribute to thermo-tolerance through regulating HSPs as

described by Ohama et al. (2017). Additionally, we identified 29 *HSP* genes, predominantly encoding HSP70, and 17 small heat proteins (HSP20) in the *M. balsamina* assembly. *HSP*s play an essential role in the regulation of HSFs and, subsequently, the expression of heat-responsive genes associated with heat tolerance. *HSP20* has been reported to contribute to heat stress tolerance in melon (Zheng et al., 2021), watermelon (He et al., 2019), cucumber (Chen et al., 2021), and pumpkin (Hu et al., 2021). Over-expression of *HSP70* has also been reported to significantly increase heat tolerance in watermelon, cabbage, and chilli (Park et al., 2013; Guo et al., 2015; Usman et al., 2015; Zhao et al., 2018; He et al., 2019). Therefore, the thermo-tolerance capacity of *M. balsamina* can be attributed to the identified important HSPs, which can be further functionally validated for future use. Similarly, we identified 15 genes encoding proteins related to salinity tolerance in the *M. balsamina* assembly, including ACER, S-acyltransferase, salt stress root protein RS1-like, and protein RICE SALT SENSITIVE 3 isoforms. These proteins have previously been reported to play a role in salinity tolerance in Arabidopsis by Wu et al. (2015) and in wheat by Kang et al. (2012). However, their role in salt tolerance in cucurbits has yet to be well documented. These identified genes with a possible role in salt tolerance can be further studied to understand the detailed physiological and molecular network associated with salt tolerance and improve the salt tolerance of related species through inter-specific introgression. Additionally, we found 37 glutathione S-transferase (GST) family genes in *M. balsamina*, which are vital antioxidant enzymes involved in reducing the damage caused by reactive oxygen species during abiotic stress (salt, drought, and cold) tolerance mechanisms (Venkateswarlu et al., 2012; Chan and Lam, 2014; Islam et al., 2019; and Song et al., 2021). GSTs are also involved in detoxification processes and protection against damage from various environmental factors (Dixon et al., 1998; (Esmaeili et al., 2009). The large number of identified GST family genes in *M. balsamina* suggests its high tolerance to abiotic stress, which can be harnessed to improve abiotic stress tolerance in *M. charantia*.

In the *M. balsamina* assembly, we identified five genes related to carotenoid biosynthesis, including chloroplast-specific lycopene beta-cyclase, phytoene desaturase/phytoene dehydrogenase, pro-lycopene isomerase, zeta-carotene desaturase, and lycopene epsilon cyclase. The overexpression of one or more carotenoid biosynthesis genes to produce carotene-rich varieties has been successfully employed in advanced vegetable improvement programs for crops such as tomatoes (Fraser et al., 2001), carrot (Fraser and Bramley, 2004), and potatoes (Diretto et al., 2007). Carotenoids contribute to color, serve as precursors of vitamin A, and have various health benefits, including reducing the risk of cancers and cardiovascular diseases (Paine et al., 2005; Aluru et al., 2008). Therefore, the transfer of these carotenoid biosynthesis genes from *M. balsamina* to *M. charantia* could be utilized to improve its nutritional value. Furthermore, we identified 229 genes related to cucurbitacin biosynthesis in the *M. balsamina* assembly. Cucurbitacins are signature bioactive compounds of the *Cucurbitaceae* family and confer a bitter taste to cucurbits (Chen et al., 2005). The identified genes encoding enzymes such as oxidosqualene cyclase, cytochromes P450, and acyltransferases are essential for cucurbitacin biosynthesis. Similar pathways and mechanisms are involved in the production of terpenoids across the genera of the *Cucurbitaceae* family (Huang et al., 2009; Shang et al., 2014). Moreover, we identified 19 genes related to the biosynthesis of other triterpenoids in the *M. balsamina* assembly. These triterpenoids have diverse medicinal properties, namely, anticancer, antidiabetic, anti-HIV, antimalarial, anti-inflammatory, and antimicrobial activities (Ramalhete et al., 2022). Many of these triterpenoids such as balsaminol, balsaminoside, balsaminagenin, karavilagenin, cucurbalsaminol, and balsaminapentaol (Ramalhete et al., 2009a; Ramalhete et al., 2009; Ramalhete et al., 2010; Ramalhete et al., 2011a; and Ramalhete et al., 2011b) have been previously isolated from *M. balsamina*, highlighting its potential as a source of bioactive compounds. These results confirm the value of *M. baslamina* in terms of its nutritional and therapeutic properties.

M. balsamina is a monoecious plant with separate male and female flowers on the same plant. Sex determination and expression in cucurbits have been extensively studied, and various phytohormones and their cross talk have been identified as key regulators (Chen et al., 2016; Wang et al., 2019). Ethylene, in particular, is considered a core regulator of sex expression in cucurbits (Yin and Quinn, 1995; Boualem et al., 2015; Chen et al., 2016). In the *M. balsamina* assembly, we identified six genes related to ethylene biosynthesis, including *ACS* (1-aminocyclopropane-1-carboxylate synthase) *ACS-7*, *ACS-CMA101*, and *ACS-CMW-33* genes. These genes are involved in the production of ethylene, which regulates sex expression in cucurbits. In *Cucumis sativus, ACS-1* is encoded by the *F* locus and is known to promote female sex expression by suppressing stamen development in bisexual flower primordial (Trebitsh et al., 1997; Mibus and Tatlioglu, 2004). Likewise, *ACS-7* is encoded by *A* locus (orthologue of the cucumber *M* gene) and is known to promote femaleness in monoecious melon lines, and a miss-sense mutation in *CmACS-7* led to andromonoecy, the predominant sex type of commercial melon (Boualem et al., 2008; Boualem et al., 2009). Similarly, two genes (*MOMC46*_189, *MOMC518*_1) encoding CmAcs-7 like protein and a gene (*MOMC3*_649) encoding CmACS 11 like protein were identified in *M. charantia* (Urasaki et al., 2017). *ACS* encoding genes for sex determination in *M. balsamina* and *M. charantia* were found orthologous by synteny analysis as well. This suggests the possible involvement ethylene regulated sex expression like all other cucurbits in *Momoridica* genus. The orthologous relationship of these *ACS* genes with those identified in *M. charantia* and other cucurbits suggests a highly conserved nature of sex-regulating genes across the *Cucurbitaceae* family. Additionally, our study revealed a high number of conserved genes (approximately 8,500) between *M. balsamina* and *M. charantia*, *Cucumis sativus*, *Cucumis melo*, and *Citrullus lanatus*, indicating a substantial level of genetic similarity and potential for comparative genomics studies among cucurbits.

Comparative plant genomics investigates the distinctiveness and differences among plant genomes. By comparing the genomes of closely and distantly related species, researchers can gain insights into the patterns and processes associated with plant genome evolution and identify functional regions within genomes

(Caicedo and Purugganan, 2005). In this particular study, we conducted a genome comparison of *Momordica balsamina* with other related cucurbit species, namely, *Momordica charantia* (Bitter gourd), *Cucumis sativus* (Cucumber), *Cucumis melo* (Musk melon), and *Citrullus lanatus* (Watermelon), in order to identify syntenic and phylogenetic relationships. Our analysis revealed that *Momordica balsamina* shared the highest number of orthologous pairs (8,845) with *Momordica charantia*, followed by 8,265 orthologous pairs between *Momordica balsamina* and *Cucumis sativus*. Previous research by Garcia-Mas et al. (2012) identified 19,377 one-to-one ortholog pairs between *Cucumis melo* and *Cucumis sativus*.

Furthermore, we detected paralogous and orthologous relationships between the five studied *Cucurbitaceae* genomes, which can serve as a guide for translational research and facilitate the study of conserved economic traits. By utilizing conserved BUSCO genes (orthologous genes), we identified the evolutionary relationship between *Momordica balsamina*, *Momordica charantia*, *Citrullus lanatus*, *Cucumis sativus*, and *Cucumis melo*. Phylogenetic analysis done using *Vitis vinifera* as an outgroup classified *Momordica balsamina* and *Momordica charantia* to the same clade, indicating a close genetic relationship between these two species with a speciation/separation event estimated to have occurred 23 million years ago. Additionally, *Momordica* was found to be closer to *Citrullus* (Watermelon) than to *Cucumis*, suggesting a divergence around 53 million years ago. Previous studies by Urasaki et al. (2017); Jobst et al. (1998), and Schaefer et al. (2009) also reported a closer genetic association between bitter gourd and watermelon compared to cucumber or melon in phylogenetic and genetic analyses.

We performed synteny analysis to elucidate variations at the nucleotide level arising from mutations, duplications, chromosomal rearrangements, and gene family expansion or loss (Alkan et al., 2011). Synteny blocks, which identify regions of chromosomes shared between genomes that have a common order of homologous genes from a common ancestor, were identified to shed light on evolutionary relationships between species (Vergara and Chen, 2010). Previous synteny analysis in members of the *Cucurbitaceae* family helped to clarify the reason behind differences in basic chromosome number between *Cucumis sativus* and *C. melo* (Huang et al., 2009; Li et al., 2011). In our current study, we found the highest number of syntenic blocks between *Momordica balsamina* and *Momordica charantia* (306), followed by 282 syntenic blocks between *Momordica balsamina* and *Citrullus lanatus*, indicating a high level of synteny between *M. balsamina* and *M. charantia*, followed by watermelon (*Citrullus lanatus*). Previous synteny analyses also reported a high level of co-linearity between *Momordica* and *Citrullus* genomes (Urasaki et al., 2017; Cui et al., 2020). Our findings revealed a general absence of one-to-one relationships in the chromosomes between *Momordica balsamina* and other cucurbit genomes. This observation aligns with most of the synteny analyses conducted in cucurbits (Matsusmura and Urasaki, 2020); Wu et al., 2020, except for the study by Wu et al. (2017), which identified chromosome-level synteny between bottle gourd and melon (*C. melo*) and watermelon (*Citrullus lanatus*) genomes. The findings of our study, along with the synteny analysis by Matsusmura and Urasaki (2020), support the fact that most *Cucurbitaceae* genomes belong to a different clade than the genus *Momordica* (Renner and Schaefer, 2016). Therefore, the absence of one-to-one chromosome synteny between *Momordica* (*balsamina* and *charantia*) and other cucurbits may be attributed to higher structural re-arrangement in chromosomes after speciation.

In addition to the genome comparison and synteny analysis, we identified 215,379 SSRs and 567,483 TF binding sites (TFBSs). These data were incorporated into a genomic web resource called MbGeR, developed to provide access to the data extracted during this study. Characterizing the *M. balsamina* genome contributes to our understanding of the available gene pool that can be utilized to improve *M. charantia* through advanced plant breeding techniques. Due to the significant therapeutic values, resilience to biotic and abiotic stress and nutritional value of *M. balsamina*, this study offers valuable insights and a high-quality assembly and annotation of its genome, thereby assisting in the development of high-yielding and resistant varieties of this promising vegetable crop.

5 Conclusion

M. balsamina is the closest wild species of *M. charantia*, with higher resilience to biotic and abiotic stresses and greater medicinal and nutritional qualities. The present study provides the first high-quality chromosome-level genome assembly of *M. balsamina* with size 384.90 Mb and N5030.96 Mb using sequence data from 10x Genomics, Nanopore, and Hi-C platforms. Annotation of the provided assembly identified 215,379 SSRs; 632,098 TEs; 567,483 TF binding sites; 3,376 noncoding RNAs (tRNA, miRNA, snoRNA, and so forth) genes, and 41,652 protein coding genes. A sum of 4,347 disease resistance, 67 heat stress–related, 15 salt stress related, 229 cucurbitacin related, 19 terpenes related, 37 antioxidant activity, 05 carotenoid related, and 06 sex determination related genes were identified in *M. balsamina* assembly. Because of stress tolerance and better therapeutic values, *M. balsamina* will serve as a potential genomic resource, and provided assembly will help to boost the targeted gene introgression between *M. balsamina* and *M. charantia* species in developing high-yielding climate-smart and stress-resilient crop varieties. In addition, this high-quality genome assembly done using reads from multiple sequencing platforms can be used to improve further the quality and completeness of genome assembly of related species. The SSR markers obtained in this study would assist in linkage mapping, QTL and gene discovery, population genetics, evolutionary studies and gene regulation. The provided assembly will also help in identifying a higher number of genome-wide markers with greater specificity and accuracy to trace the introgressed segments during advanced breeding programs to improve resistance and medicinal values to high-yielding *M. charantia* varieties, which is significantly lost due to domestication of bitter gourd. Furthermore, the finding of comparative genome analysis (phylogeny and synteny) will be helpful to get insights into the patterns and processes associated with genome evolution and to uncover functional regions of cucurbit genomes.

Data availability statement

Whole genome sequencing of Momordica balsamina: BioProject: PRJNA877043: First and high-quality assembly of Momordica balsamina, a potential genetic resource to improve tolerance and medicinal properties in bitter gourd, BioSample: SAMN30678163: Momordica balsamina genome; SRA: SRR21495983, SRR21495982 and SRR21495981. The assembly is submitted in NCBI with ID: SUB13995037.

Author contributions

VN: Writing – original draft, Investigation. KS: Data curation, Writing – original draft. RE: Supervision, Writing – review & editing. VC: Writing – review & editing, Resources. SJ: Writing – review & editing, Data curation, Writing – original draft. MI: Data curation, Writing – review & editing. AM: Writing – review & editing, Resources, Supervision. HM: Supervision, Writing – review & editing. BG: Writing – review & editing, Investigation. GJ: Writing – review & editing. CK: Writing – review & editing, Supervision. AG: Writing – review & editing, Supervision. DK: Supervision, Writing – review & editing, Data curation, Investigation, Writing – original draft. SD: Supervision, Writing – review & editing, Data curation, Investigation, Writing – original draft. TB: Resources, Supervision, Writing – original draft, Writing – review & editing, Conceptualization.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. The research work was funded by the NAHEP-CAAST programme of Indian Council of Agricultural Research (ICAR). This work was also supported by the Indian Council of Agricultural Research, Ministry of Agriculture and Farmers' Welfare, Govt. of India, who provided financial assistance in the form of a CABin grant (F. no. Agril. Edn.4-1/2013-A&P), as well as Advanced Super Computing Hub for Omics Knowledge in Agriculture (ASHOKA) facility at ICAR-IASRI, New Delhi, India.

Acknowledgments

Authors are thankful to the ICAR-Indian Agricultural Research Institute, New Delhi for providing financial support and conduct of the research program of the PhD student, Mr. VN. We are thankful to the Indian Council of Agricultural Research, Ministry of Agriculture and Farmers' Welfare, Govt. of India for Advanced Super Computing Hub for Omics Knowledge in Agriculture (ASHOKA) facility at ICAR-IASRI, New Delhi, India created under National Agricultural Innovation Project, funded by World Bank at ICAR-IASRI, New Delhi.

Conflict of interest

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

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1258042/full#supplementary-material

References

Ajji, P. K. (2016). *Functional characterization of a novel ribosome inactivating protein from Momordica balsamina* (Deakin University, Australia: Doctoral dissertation, Deak in University). Available at: https://dro.deakin.edu.au/eserv/DU:30103049/ajji-functionalcharacterization-2017.pdf.

Ajji, P. K., Binder, M. J., Walder, K., and Puri, M. (2017). Balsamin induces apoptosis in breast cancer cells via DNA fragmentation and cell cycle arrest. *Mol. Cell. Biochem.* 432 (1), 189–198. doi: 10.1007/s11010-017-3009-x

Alkan, C., Coe, B. P., and Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* 12 (5), 363–376. doi: 10.1038/nrg2958

Aluru, M., Xu, Y., Guo, R., Wang, Z., Li, S., White, W., et al. (2008). Generation of transgenic maize with enhanced provitamin A content. *J. Exp. Botany.* 59 (13), 3551–3562. doi: 10.1093/jxb/ern212

Andrews, S. (2010) *FASTQC. A quality control tool for high throughput sequence data*. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Aversano, R., Contaldi, F., Ercolano, M. R., Grosso, V., Iorizzo, M., Tatino, F., et al. (2015). The *Solanum commersonii* genome sequence provides insights into adaptation to stress conditions and genome evolution of wild potato relatives. *Plant Cell.* 27 (4), 954–968. doi: 10.1105/tpc.114.135954

Baumgarten, A., Cannon, S., Spangler, R., and May, G. (2003). Genome-level evolution of resistance genes in *Arabidopsis thaliana*. *Genetics* 165 (1), 309–319. doi: 10.1093/genetics/165.1.309

Behera, T. K., Behera, S., Bharathi, L. K., John, K. J., Simon, P. W., and Staub, J. E. (2010). Bitter gourd: botany, horticulture, breeding. *Hortic. Rev.* 37, 101–141. doi: 10.1002/9780470543672.ch2

Beier, S., Thiel, T., Münch, T., Scholz, U., and Mascher, M. (2017). MISA-web: a web server for microsatellite prediction. *Bioinformatics* 33 (16), 2583–2585. doi: 10.1093/bioinformatics/btx198

Belaghzal, H., Dekker, J., and Gibcus, J. H. (2017). Hi-C 2.0: An optimized Hi-C procedure for high-resolution genome-wide mapping of chromosome conformation. *Methods* 123, 56–65. doi: 10.1016/j.ymeth.2017.04.004

Belfanti, E., Silfverberg-Dilworth, E., Tartarini, S., Patocchi, A., Barbieri, M., Zhu, J., et al. (2004). The *HcrVf2* gene from a wild apple confers scab resistance to a

transgenic cultivated variety. *Proc. Natl. Acad. Sci.* 101 (3), 886–890. doi: 10.1073/pnas.0304808101

Bharathi, L. K., and John, K. J. (2013). *Momordica Genus in Asia-An Overview* Vol. p (New Delhi: Springer), 147). doi: 10.1007/978-81-322-1032-0

Bharathi, L. K., Munshi, A. D., Behera, T. K., Vinod, J. K.J., Das, A. B., Bhat, K. V., et al. (2012). Production and preliminary characterization of inter-specific hybrids derived from *Momordica* species. *Curr. Sci.* 103 (2), 178–186.

Bharathi, L. K., Munshi, A. D., Chandrashekaran, S., Behera, T. K., Das, A. B., and John, K. J. (2011). Cytotaxonomical analysis of *Momordica* L. (Cucurbitaceae) species of Indian occurrence. *J. Genet.* 90 (1), 21–30.

Boualem, A., Fergany, M., Fernandez, R., Troadec, C., Martin, A., Morin, H., et al. (2008). A conserved mutation in an ethylene biosynthesis enzyme leads to andromonoecy in melons. *Science* 321 (5890), 836–838. doi: 10.1126/science.1159023

Boualem, A., Troadec, C., Camps, C., Lemhemdi, A., Morin, H., Sari, A., et al. (2015). A cucurbit androecy gene reveals how unisexual flowers develop and dioecy emerges. *Science* 350 (6261), 688–691. doi: 10.1126/science.aac8370

Boualem, A., Troadec, C., Kovalski, I., Sari, M. A., Perl-Treves, R., and Bendahmane, A. (2009). A conserved ethylene biosynthesis enzyme leads to andromonoecy in two *Cucumis* species. *PloS One* 4 (7), e6144. doi: 10.1371/journal.pone.0006144

Bowley, S. R., and Taylor, N. L. (1987). "Introgressive hybridization," in *CRC handbook of plant science in agriculture*, vol. 1 . Ed. B. R. Christie (Boca Raton: CRC Press), 23–59.

Brozynska, M., Furtado, A., and Henry, R. J. (2016). Genomics of crop wild relatives: expanding the gene pool for crop improvement. *Plant Biotechnol. J.* 14 (4), 1070–1085. doi: 10.1111/pbi.12454

Caicedo, A. L., and Purugganan, M. D. (2005). Comparative plant genomics. Frontiers and prospects. *Plant Physiol.* 138 (2), 545–547. doi: 10.1104/pp.104.900148

Cao, M. D., Nguyen, S. H., Ganesamoorthy, D., Elliott, A. G., Cooper, M. A., and Coin, L. J. (2017). Scaffolding and completing genome assemblies in real-time with nanopore sequencing. *Nat. Commun.* 8 (1), 1–10. doi: 10.1038/ncomms14515

Chan, C., and Lam, H. M. (2014). A putative lambda class glutathione S-transferase enhances plant survival under salinity stress. *Plant Cell Physiol.* 55 (3), 570–579. doi: 10.1093/pcp/pct201

Chan, P. P., and Lowe, T. M. (2019). tRNAscan-SE: searching for tRNA genes in genomic sequences. *Methods Mol. Biol. (Clifton NJ)* 1962, 1–14. doi: 10.1007/978-1-4939-9173-0_1

Chauhan, H., Khurana, N., Agarwal, P., and Khurana, P. (2011). Heat shock factors in rice (*Oryza sativa* L.): genome-wide expression analysis during reproductive development and abiotic stress. *Mol. Genet. Genomics* 286 (2), 171–187. doi: 10.1007/s00438-011-0638-8

Chen, J. C., Chiu, M. H., Nie, R. L., Cordell, G. A., and Qiu, S. X. (2005). Cucurbitacins and cucurbitane glycosides: structures and biological activities. *Natural product Rep.* 22 (3), 386–399. doi: 10.1039/B418841C

Chen, F., Dong, W., Zhang, J., Guo, X., Chen, J., Wang, Z., et al. (2018). The sequenced angiosperm genomes and genome databases. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.00418

Chen, F., Mackey, A. J., Stoeckert, C. J. Jr., and Roos, D. S. (2006). OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.* 34 (suppl_1), D363–D368. doi: 10.1093/nar/gkj123

Chen, H., Sun, J., Li, S., Cui, Q., Zhang, H., Xin, F., et al. (2016). An ACC oxidase gene essential for cucumber carpel development. *Mol. Plant* 9 (9), 1315–1327. doi: 10.1016/j.molp.2016.06.018

Chen, X., Wang, Z., Tang, R., Wang, L., Chen, C., and Ren, Z. (2021). Genome-wide identification and expression analysis of *Hsf* and *Hsp* gene families in cucumber (*Cucumis sativus* L.). *Plant Growth Regul.* 95 (2), 223–239. doi: 10.1007/s10725-021-00739-z

Conesa, A., Gotz, S., García-Gomez, J. M., Terol, J., Talón, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21 (18), 3674–3676. doi: 10.1093/bioinformatics/bti610

Cui, J., Yang, Y., Luo, S., Wang, L., Huang, R., Wen, Q., et al. (2020). Whole-genome sequencing provides insights into the genetic diversity and domestication of bitter gourd (*Momordica* spp.). *Horticulture Res.* 7 (1), 85. doi: 10.1038/s41438-020-0305-5

Dempewolf, H., Baute, G., Anderson, J., Kilian, B., Smith, C., and Guarino, L. (2017). Past and future use of wild relatives in crop breeding. *Crop Sci.* 57 (3), 1070–1082. doi: 10.2135/cropsci2016.10.0885

Dempewolf, H., Eastwood, R. J., Guarino, L., Khoury, C. K., Müller, J. V., and Toll, J. (2014). Adapting agriculture to climate change: a global initiative to collect, conserve, and use crop wild relatives. *Agroecology Sustain. Food Syst.* 38 (4), 369–377. doi: 10.1080/21683565.2013.870629

Dhillon, N. P., Sanguansil, S., Schafleitner, R., Wang, Y. W., and McCreight, J. D. (2016). Diversity among a wide Asian collection of bitter gourd landraces and their genetic relationships with commercial hybrid cultivars. *J. Am. Soc. Hortic. Sci.* 141 (5), 475–484. doi: 10.21273/JASHS03748-16

Dillon, S. L., Lawrence, P. K., Henry, R. J., and Price, H. J. (2007). Sorghum resolved as a distinct genus based on combined ITS1, ndh F and Adh 1 analyses. *Plant Systematics Evol.* 268, 29–43. doi: 10.1007/s00606-007-0571-9

Diretto, G., Al-Babili, S., Tavazza, R., Papacchioli, V., Beyer, P., and Giuliano, G. (2007). Metabolic engineering of potato carotenoid content through tuber-specific overexpression of a bacterial mini-pathway. *PloS One* 2 (4), e350. doi: 10.1371/journal.pone.0000350

Dixon, D. P., Cummins, I., Cole, D. J., and Edwards, R. (1998). Glutathione-mediated detoxification systems in plants. *Curr. Opin. Plant Biol.* 1 (3), 258–266. doi: 10.1016/S1369-5266(98)80114-3

Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., et al. (2017). *De novo* assembly of the *Aedes aEgypti* genome using Hi-C yields chromosome-length scaffolds. *Science* 356 (6333), 92–95. doi: 10.1126/science.aal3327

Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S., Huntley, M. H., Lander, E. S., et al. (2016). Juicer provides a one-click system for analysing loop-resolution Hi-C experiments. *Cell Syst.* 3 (1), 95–98. doi: 10.1016/j.cels.2016.07.002

Edwards, M. A., and Henry, R. J. (2011). DNA sequencing methods contributing to new directions in cereal research. *J. Cereal Sci.* 54 (3), 395–400. doi: 10.1016/j.jcs.2011.07.006

Erxleben, A., and Grüning, B. (2020) *Genome Annotation (Galaxy Training Materials)*. Available at: https://training.galaxyproject.org/training-material/topics/genome-annotation/tutorials/genome-annotation/tutorial (Accessed May 24 2022).

Esmaeili, M., Shahrtash, M., Moosavi, F., Mohsenzadeh, S., and Mohabatkar, H. (2009). Plant glutathione S-transferase function. *Paper Presentation Proc. 6th Natl. Biotechnol. Congress Iran Tehran Iran*.

Fitzgerald, H. (2013).

Fraser, P. D., and Bramley, P. M. (2004). The biosynthesis and nutritional uses of carotenoids. *Prog. Lipid Res.* 43 (3), 228–265. doi: 10.1016/j.plipres.2003.10.002

Flyman, M. V., and Afolayan, A. J. (2007). Proximate and mineral composition of the leaves of Momordica balsamina L.: an under-utilized wild vegetable in Botswana. *Int. J. Food Sci. Nutr.* 58 (6), 419–423. doi: 10.1080/09637480701253417

Fraser, P. D., Romer, S., Kiano, J. W., Shipton, C. A., Mills, P. B., Drake, R., et al. (2001). Elevation of carotenoids in tomato by genetic manipulation. *J. Sci. Food Agric.* 81 (9), 822–827. doi: 10.1002/JSFA.908

García, J. C., Guadagno, A., Paytuvi-Gallart, A., Saera-Vila, A., Amoroso, C. G., D'Esposito, D., et al. (2021). PRGdb 4.0: an *updated database dedicated to genes involved in plant disease resistance process*. *Nucleic Acids Res.* 50 (D1), D1483–D1490. doi: 10.1093/nar/gkab1087

Garcia-Mas, J., Benjak, A., Sanseverino, W., Bourgeois, M., Mir, G., González, V. M., et al. (2012). The genome of melon (*Cucumis melo* L.). *Proc. Natl. Acad. Sci.* 109 (29), 11872–11877. doi: 10.1073/pnas.1205415109

Goff, S. A., Ricke, D., Lan, T. H., Presting, G., Wang, R., Dunn, M., et al. (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* 296 (5565), 92–100. doi: 10.1126/science.1068275

Gramazio, P., Yan, H., Hasing, T., Vilanova, S., Prohens, J., and Bombarely, A. (2019). Whole-genome resequencing of seven eggplant (*Solanum melongena*) and one wild relative (*S. incanum*) accessions provides new insights and breeding tools for eggplant enhancement. *Front. Plant science.* 1220. doi: 10.3389/fpls.2019.01220

Guo, M., Lu, J. P., Zhai, Y. F., Chai, W. G., Gong, Z. H., and Lu, M. H. (2015). Genome-wide analysis, expression profile of heat shock factor gene family (*CaHsfs*) and characterisation of *CaHsfA2* in pepper (*Capsicum annuum* L.). *BMC Plant Biol.* 15 (1), 151. doi: 10.1186/s12870-015-0512-7

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 (1), 51–58. doi: 10.1038/ng.2470

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 29 (8), 1072–1075. doi: 10.1093/bioinformatics/btt086

Hassan, L. G., and Umar, K. J. (2006). Nutritional value of Balsam Apple (*Momordica balsamina* L.) leaves. *Pakistan J. Nutr.* 5 (6), 522–529. doi: 10.3923/pjn.2006.522.529

He, Y., Fan, M., Sun, Y., and Li, L. (2019). Genome-wide analysis of watermelon *HSP20s* and their expression profiles and subcellular locations under stresses. *Int. J. Mol. Sci.* 20 (1), 12. doi: 10.3390/ijms20010012

Hoen, D. R., and Bureau, T. E. (2015). Discovery of novel genes derived from transposable elements using integrative genomic analysis. *Mol. Biol. Evol.* 32 (6), 1487–1506. doi: 10.1093/molbev/msv042

Hu, Y., Zhang, T., Liu, Y., Li, Y., Wang, M., Zhu, B., et al. (2021). Pumpkin (*Cucurbita moschata*) *HSP20* Gene Family Identification and Expression under Heat Stress. *Front. Genet.* 2062. doi: 10.3389/fgene.2021.753953

Huang, S., Li, R., Zhang, Z., Li, L. I., Gu, X., Fan, W., et al. (2009). The genome of the cucumber, *Cucumis sativus* L. *Nat. Genet.* 41 (12), 1275–1281. doi: 10.1038/ng.475

Islam, S., Sajib, S. D., Jui, Z. S., Arabia, S., Islam, T., and Ghosh, A. (2019). Genome-wide identification of glutathione S-transferase gene family in pepper, its classification, and expression profiling under different anatomical and environmental conditions. *Sci. Rep.* 9 (1), 1–15. doi: 10.1038/s41598-019-45320-x

Jeffrey, C. (1967). "Cucurbitaceae," in *Flora of tropical East Africa*. Eds. C. E. Milne-Redhead and R. M. Polhill (London, UK: Crown Agents for Overseas Governments and Administrations), 1–156.

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

Jobst, J., King, K., and Hemleben, V. (1998). Molecular evolution of the internal transcribed spacers (ITS1 and ITS2) and phylogenetic relationships among the species of the family Cucurbitaceae. *Mol. Phylo. Evol.* 9, 204–219. doi: 10.1006/mpev.1997.0465

John, K. J. (2005). *Studies on ecogeography and genetic diversity of the genus Momordica L. @ in India* (Kottayam, Kerala: Dissertation, Mahatma Gandhi University).

Jones, D. A., Thomas, C. M., Hammond-Kosack, K. E., Balint-Kurti, P. J., and Jones, J. D. (1994). Isolation of the tomato *Cf-9* gene for resistance to *Cladosporium fulvum* by transposon tagging. *Science* 266 (5186), 789–793. doi: 10.1126/science.7973631

Kang, G., Li, G., Zheng, B., Han, Q., Wang, C., Zhu, Y., et al. (2012). Proteomic analysis on salicylic acid-induced salt tolerance in common wheat seedlings (*Triticum aestivum* L.). *Biochim. Biophys. Acta (BBA)-Proteins Proteomics.* 1824 (12), 1324–1333. doi: 10.1016/j.bbapap.2012.07.012

Kasem, S., Waters, D. L., Rice, N., Shapter, F. M., and Henry, R. J. (2010). Whole grain morphology of Australian rice species. *Plant Genet. Resour.* 8 (1), 74–81. doi: 10.1017/S1479262109990189

Katoh, K., Misawa, K., Kuma, K. I., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30 (14), 3059–3066. doi: 10.1093/nar/gkf436

Kaur, I., Yadav, S. K., Hariprasad, G., Gupta, R. C., Srinivasan, A., Batra, J. K., et al. (2012). Balsamin, a novel ribosome-inactivating protein from the seeds of Balsam apple *Momordica balsamina*. *Amino Acids* 43 (2), 973–981. doi: 10.1007/s00726-011-1162-1

Kell, S., Qin, H., Chen, B., Ford-Lloyd, B., Wei, W., Kang, D., et al. (2015). China's crop wild relatives: diversity for agriculture and food security. *Agriculture Ecosyst. Environ.* 209, 138–154. doi: 10.1016/j.agee.2015.02.012

Khare, C. (2007). "*Momordica balsamina* Linn," in *Indian Medicinal Plants*. Ed. C. Khare (New York, NY: Springer). doi: 10.1007/978-0-387-70638-2_1027

Krawinkel, M. B., Ludwig, C., Swai, M. E., Yang, R. Y., Chun, K. P., and Habicht, S. D. (2018). Bitter gourd reduces elevated fasting plasma glucose levels in an intervention study among prediabetics in Tanzania. *J. ethnopharmacology* 216, 1–7. doi: 10.1016/j.jep.2018.01.016

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., et al. (2009). Circos: an information aesthetic for comparative genomics. *Genome Res.* 19 (9), 1639–1645. doi: 10.1101/gr.092759.109

Lewinski, M., Bramkamp, Y., Köster, T., and Staiger, D. (2020). SEQing: web-based visualization of iCLIP and RNA-seq data in an interactive python framework. *BMC Bioinf.* 21 (1), 113. doi: 10.1186/s12859-020-3434-9

Li, D., Cuevas, H. E., Yang, L., Li, Y., Garcia-Mas, J., Zalapa, J., et al. (2011). Syntenic relationships between cucumber (*Cucumis sativus* L.) and melon (C. melo L.) chromosomes as revealed by comparative genetic mapping. *BMC Genomics* 12 (1), 1–14. doi: 10.1186/1471-2164-12-396

Limin, F., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next generation sequencing data. *Bioinformatics* 28 (23), 3150–3152. doi: 10.1093/bioinformatics/bts565

Lin, X., Zhang, Y., Kuang, H., and Chen, J. (2013). Frequent loss of lineages and deficient duplications accounted for low copy number of disease resistance genes in cucurbitaceae. *BMC Genomics* 14, 1–13.

Lipman, D. J., and Pearson, W. R. (1985). Rapid and sensitive protein similarity searches. *Science* 227 (4693), 1435–1441. doi: 10.1126/science.2983426

Mashiane, P., Shoko, T., Manhivi, V., Slabbert, R., Sultanbawa, Y., and Sivakumar, D. (2022). A Comparison of bioactive metabolites, antinutrients, and bioactivities of african pumpkin leaves (*Momordica balsamina* L.) cooked by different culinary techniques. *Molecules* 27 (6), 1901. doi: 10.3390/molecules27061901

Matsumura, H., and Urasaki, N. (2020). "Genome sequence of bitter Gourd and Its Comparative Study with Other Cucurbitaceae Genomes," in *The Bitter Gourd Genome. Compendium of plant Genomes*. Eds. C. Kole, H. Matsumura and T. Behera (Cham: Springer), 113–123). doi: 10.1007/978-3-030-15062-4_10

Maxted, N., Kell, S., Ford-Lloyd, B., Dulloo, E., and Toledo, Á. (2012). Toward the systematic conservation of global crop wild relative diversity. *Crop Sci.* 52 (2), 774–785. doi: 10.2135/cropsci2011.08.0415

Mibus, H., and Tatlioglu, T. (2004). Molecular characterization and isolation of the *F/f* gene for femaleness in cucumber (*Cucumis sativus* L.). *Theor. Appl. Genet.* 109 (8), 1669–1676. doi: 10.1007/s00122-004-1793-7

Mishra, K. C., Sahu, P. R., and Jha, U. C. (1986). Balsam apple for your vegetable garden. *Indian Horticulture J.* 13.

Nawrocki, E. P., and Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* 29 (22), 2933–2935. doi: 10.1093/bioinformatics/btt509

Nover, L., Bharti, K., Döring, P., Mishra, S. K., Ganguli, A., and Scharf, K. D. (2001). Arabidopsis and the heat stress transcription factor world: how many heat stress transcription factors do we need? *Cell Stress chaperones* 6 (3), 177. doi: 10.1379/1466-1268(2001)006<0177:aathst>2.0.co;2

Ohama, N., Sato, H., Shinozaki, K., and Yamaguchi-Shinozaki, K. (2017). Transcriptional regulatory network of plant heat stress response. *Trends Plant Sci.* 22 (1), 53–65. doi: 10.1016/j.tplants.2016.08.015

Paine, J. A., Shipton, C. A., Chaggar, S., Howells, R. M., Kennedy, M. J., Vernon, G., et al. (2005). Improving the nutritional value of Golden Rice through increased pro-vitamin A content. *Nat. Biotechnol.* 23 (4), 482–487. doi: 10.1038/nbt1082

Park, H. J., Jung, W. Y., Lee, S. S., Song, J. H., Kwon, S. Y., Kim, H., et al. (2013). Use of heat stress responsive gene expression levels for early selection of heat tolerant cabbage (*Brassica oleracea* L.). *Int. J. Mol. Sci.* 14 (6), 11871–11894. doi: 10.3390/ijms140611871

Peter, K. V., and Abraham, Z. (2007). *Biodiversity in horticultural crops* Vol. 1 (New Delhi, India: Daya Publisher).

Pimentel, D., Wilson, C., McCullum, C., Huang, R., Dwen, P., Flack, J., et al. (1997). Economic and environmental benefits of biodiversity. *BioScience* 47 (11), 747–757. doi: 10.2307/1313097

Qin, X., Zhang, Z., Lou, Q., Xia, L., Li, J., Li, M., et al. (2021). Chromosome-scale genome assembly of *Cucumis hystrix*—a wild species interspecifically cross-compatible with cultivated cucumber. *Horticulture Res.* 8 (1), 40. doi: 10.1038/s41438-021-00475-5

Ramalhete, C., da Cruz, F. P., Lopes, D., Mulhovo, S., Rosario, V. E., Prudêncio, M., et al. (2011a). Triterpenoids as inhibitors of erythrocytic and liver stages of Plasmodium infections. *Bioorganic medicinal Chem.* 19 (24), 7474–7481. doi: 10.1016/j.bmc.2011.10.044

Ramalhete, C., Gonçalves, B. M., Barbosa, F., Duarte, N., and Ferreira, M. J. U. (2022). *Momordica balsamina*: phytochemistry and pharmacological potential of a gifted species. *Phytochem. Rev.* 21 (2), 617–646. doi: 10.1007/s11101-022-09802-7

Ramalhete, C., Lopes, D., Molnár, J., Mulhovo, S., Rosário, V. E., and Ferreira, M. J. U. (2011b). Karavilagenin C derivatives as antimalarial. *Bioorganic medicinal Chem.* 19 (1), 330–338. doi: 10.1016/j.bmc.2010.11.015

Ramalhete, C., Lopes, D., Mulhovo, S., Molnar, J., Rosário, V. E., and Ferreira, M. J. U. (2010). New antimalarial with a triterpenic scaffold from *Momordica balsamina*. *Bioorganic medicinal Chem.* 18 (14), 5254–5260. doi: 10.1016/j.bmc.2010.05.054

Ramalhete, C., Mansoor, T. A., Mulhovo, S., Molnár, J., and Ferreira, M. J. U. (2009). Cucurbitane-type triterpenoids from the African plant *Momordica balsamina*. *J. Natural products* 72 (11), 2009–2013. doi: 10.1021/np900457u

Rathod, V., Behera, T. K., Munshi, A. D., Gaikwad, A. B., Singh, S., Vinay, N. D., et al. (2021). Developing partial interspecific hybrids of *Momordica charantia*× *Momordica balsamina* and their advance generations. *Scientia Hortic.* 281, 109985. doi: 10.1016/j.scienta.2021.109985

Renner, S. S., and Schaefer, H. (2016). "Phylogeny and Evolution of the Cucurbitaceae," in *Genetics and Genomics of Cucurbitaceae. Plant Genetics and Genomics: Crops and Models*, vol. 20 . Eds. R. Grumet, N. Katzir and J. Garcia-Mas (Cham: Springer). doi: 10.1007/7397_2016_14

Robinson, R. W., and Decker-Walters, D. S. (1997). "Interspecific hybridization," in *Cucurbits*. Eds. R. Robinson and D. S. Decker-Walters, (Oxon, U.K: CAB Intl.) 51–55. doi: 10.1073/pnas.81.24.8014

Robinson, J. T., Turner, D., Durand, N. C., Thorvaldsdottir, H., Mesirov, J. P., and Aiden, E. L. (2018). Juicebox. js provides a cloud-based visualization system for Hi-C data. *Cell Syst.* 6 (2), 256–258. doi: 10.1016/j.cels.2018.01.001

Saghai-Maroof, M. A., Jorgensen, R. A., and Allard, R. W. (1984). Ribosomal DNA spacer-length polymorphisms in barley: Mendelian inheritance, chromosomal location and population dynamics. *Proc. Natl. Acad. Sci. U.S.A.* 81, 8014–8018.

Sato, S., Tabata, S., Hirakawa, H., Asamizu, E., Shirasawa, K., Isobe, S., et al. (2012). The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* 485 (7400), 635–641. doi: 10.1038/nature11119

Schaefer, H., Heibl, C., and Renner, S. S. (2009). Gourds afloat: a dated phylogeny reveals an Asian origin of the gourd family (Cucurbitaceae) and numerous oversea dispersal events. *Proc. R. Soc. B: Biol. Sci.* 276 (1658), 843–851. doi: 10.1098/rspb.2008.1447

Shang, Y., Ma, Y., Zhou, Y., Zhang, H., Duan, L., Chen, H., et al. (2014). Biosynthesis, regulation, and domestication of bitterness in cucumber. *Science* 346 (6213), 1084–1088. doi: 10.1126/science.1259215

Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. 31 (19), 3210–3212. doi: 10.1093/bioinformatics/btv351

Singh, A. K. (1990). "Cytogenetics and evolution in the cucurbitaceae," in *Biology and Utilization of Cucurbitaceae*. Eds. D. M. Bates, R. W. Robinson and C. Jeffrey (Ithaca, New York, London: Comstock Publishing Associates, Cornell University Press), 10–28.

Singh, B. P. (1991). Interspecific hybridization in between new and old-world species of *Luffa* and its phylogenetic implication. *Cytologia* 56 (3), 359–365. doi: 10.1508/cytologia.56.359

Soderlund, C., Bomhoff, M., and Nelson, W. M. (2011). SyMAP v3. 4: a turnkey synteny system with application to plant genomes. *Nucleic Acids Res.* 39 (10), e68–e68. doi: 10.1093/nar/gkr123

Song, W., Zhou, F., Shan, C., Zhang, Q., Ning, M., Liu, X., et al. (2021). Identification of Glutathione S-Transferase Genes in Hami Melon (*Cucumis melo* var. *saccharinus*) and Their Expression Analysis under Cold Stress. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.672017

Sotowa, M., Ootsuka, K., Kobayashi, Y., Hao, Y., Tanaka, K., Ichitani, K., et al. (2013). Molecular relationships between Australian annual wild rice, Oryza meridionalis, and two related perennial forms. *Rice* 6, 1–19. doi: 10.1186/1939-8433-6-26

Tameling, W. I., Elzinga, S. D., Darmin, P. S., Vossen, J. H., Takken, F. L., Haring, M. A., et al. (2002). The tomato R gene products I-2 and MI-1 are functional ATP binding proteins with ATPase activity. *Plant Cell* 14 (11), 2929–2939. doi: 10.1105/tpc.005793

Tan, M., Ye, J., Turner, N., Hohnen-Behrens, C., Ke, C., Tang, C., et al. (2008). Antidiabetic activities of triterpenoids isolated from bitter melon associated with activation of the AMPK pathway. *Chem. Biol.* 15 (3), 263–273. doi: 10.1016/j.chembiol.2008.01.013

Tang, H., Krishnakuar, V., and Li, J. (2015). jcvi: JCVI utility libraries. *Zenodo*. doi: 10.105281/zenodo31631

Thakur, G. S., Bag, M., Sanodiya, B. S., Bhadauriya, P., Debnath, M., Prasad, G. B. K. S., et al. (2009). *Momordica balsamina*: a medicinal and neutraceutical plant for health care management. *Curr. Pharm. Biotechnol.* 10 (7), 667–682. doi: 10.2174/138920109789542066

Thiel, T., Michalek, W., Varshney, R., and Graner, A. (2003). Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theor. Appl. Genet.* 106, 411–422. doi: 10.1007/s00122-002-1031-0

Thomas, C. M., Jones, D. A., Parniske, M., Harrison, K., Balint-Kurti, P. J., Hatzixanthis, K., et al. (1997). Characterization of the tomato *Cf-4* gene for resistance to *Cladosporium fulvum* identifies sequences that determine recognition specificity in *Cf-4* and *Cf-9*. *Plant Cell* 9 (12), 2209–2224. doi: 10.1105/tpc.9.12.2209

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22 (22), 4673–4680. doi: 10.1093/nar/22.22.4673

Trebitsh, T., Staub, J. E., and O'Neill, S. D. (1997). Identification of a 1-aminocyclopropane-1-carboxylic acid synthase gene linked to the female (*F*) locus that enhances female sex expression in cucumber. *Plant Physiol.* 113 (3), 987–995. doi: 10.1104/pp.113.3.987

TrimGalore (The Babraham Institute by @ FelixKrueger). Available at: https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.

Trivedi, R. N., and Roy, R. P. (1972). Cytological studies in some species of Momordica. *Genetica* 43 (2), 282–291. doi: 10.1007/BF00123635

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., et al. (2012). Primer3—new capabilities and interfaces. *Nucleic Acids Res.* 40 (15), e115–e115. doi: 10.1093/nar/gks596

Urasaki, N., Takagi, H., Natsume, S., Uemura, A., Taniai, N., Miyagi, N., et al. (2017). Draft genome sequence of bitter gourd (*Momordica charantia*), a vegetable and medicinal plant in tropical and subtropical regions. *DNA Res.* 24 (1), 51–58. doi: 10.1093/dnares/dsw047

Usman, M. G., Rafii, M. Y., Ismail, M. R., Malek, M. A., and Latif, M. A. (2015). Expression of target gene *Hsp70* and membrane stability determine heat tolerance in chili pepper. *J. Am. Soc Hortic. Sci.* 140 (2), 144–150. doi: 10.21273/JASHS.140.2.144

Vaattovaara, A., Leppälä, J., Salojärvi, J., and Wrzaczek, M. (2019). High-throughput sequencing data and the impact of plant gene annotation quality. *J. Exp. Bot.* 70 (4), 1069–1076.

Venkateswarlu, B., Shanker, A. K., Shanker, C., and Maheswari, M. (2012). *Crop stress and its management: perspectives and strategies* (DORDRECHT, Netherlands: Springer Science & Business Media). doi: 10.1007/978-94-007-2220-0

Vergara, I. A., and Chen, N. (2010). Large synteny blocks revealed between *Caenorhabditis elegans* and *Caenorhabditis briggsae* genomes using OrthoCluster. *BMC Genomics* 11 (1), 1–13. doi: 10.1186/1471-2164-11-516

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., et al. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS One* 9 (11), e112963. doi: 10.1371/journal.pone.0112963

Wang, R., Jin, Q., Yao, C., Zhong, Y., and Wu, T. (2019). RNA-Seq analysis of gynoecious and weak female cucumber revealing the cell cycle pathway may regulate sex determination in cucumber. *Gene* 687, 289–297. doi: 10.1016/j.gene.2018.11.071

Wang, X., Tang, H., Bowers, J. E., and Paterson, A. H. (2009). Comparative inference of illegitimate recombination between rice and sorghum duplicated genes produced by polyploidization. *Genome Res.* 19 (6), 1026–1032. doi: 10.1101/gr.087288.108

Wang, J., Veldsman, W. P., Fang, X., Huang, Y., Xie, X., Lyu, A., et al. (2023). Benchmarking multi-platform sequencing technologies for human genome assembly. *Briefings Bioinf.* 24 (5), bbad300. doi: 10.1093/bib/bbad300

Weeden, N. F., and Robinson, R. W. (1986). Allozyme segregation ratios in the interspecific cross *Cucurbita maxima* x *C. Ecuadorensis* suggest that hybrid breakdown is not caused by minor alterations in chromosome structure. *Genetics* 114 (2), 593–609. doi: 10.1093/genetics/114.2.593

Wehner, T. C., Naegele, R. P., Myers, J. R., Narinder, P. S., and Crosby, K. (2020). *Cucurbits. 2nd ed* (Parlier, CA, USA: CABI). Available at: https://www.ars.usda.gov/research/publications/publication/?seqNo115=360003.

Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M., and Jaffe, D. B. (2017). Direct determination of diploid genome sequences. *Genome Res.* 27 (5), 757–767.

Wu, S., Lau, K. H., Cao, Q., Hamilton, J. P., Sun, H., Zhou, C., et al. (2018). Genome sequences of two diploid wild relatives of cultivated sweet potato reveal targets for genetic improvement. *Nat. Commun.* 9 (1), 1–12. doi: 10.1038/s41467-018-06983-8

Wu, X., Li, J., Liu, Z., Yin, J., Chang, Y., Rong, C., et al. (2015). The Arabidopsis ceramidase *AtACER* functions in disease resistance and salt tolerance. *Plant J.* 81 (5), 767–780. doi: 10.1111/tpj.12769

Wu, S., Shamimuzzaman, M. D., Sun, H., Salse, J., Sui, X., Wilder, A., et al. (2017). The bottle gourd genome provides insights into Cucurbitaceae evolution and facilitates mapping of a Papaya ring-spot virus resistance locus. *Plant J.* 92 (5), 963–975. doi: 10.1111/tpj.13722

Wu, H., Zhao, G., Gong, H., Li, J., Luo, C., He, X., et al. (2020). A high-quality sponge gourd (*Luffa cylindrica*) genome. *Horticulture Res.* 7 (1), 128. doi: 10.1038/s41438-020-00350-9

Yandell, M., and Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. *Nat. Rev. Genet.* 13 (5), 329–342. doi: 10.1038/nrg3174

Yin, T., and Quinn, J. A. (1995). Tests of a mechanistic model of one hormone regulating both sexes in *Cucumis sativus* (Cucurbitaceae). *Am. J. Bot.* 82 (12), 1537–1546. doi: 10.1002/j.1537-2197.1995.tb13856.x

Zhao, Q., Chen, W., Bian, J., Xie, H., Li, Y., Xu, C., et al. (2018). Proteomics and phosphoproteomics of heat stress-responsive mechanisms in spinach. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.00800

Zhao, C., Qiu, J., Agarwal, G., Wang, J., Ren, X., Xia, H., et al. (2017). Genome-Wide Discovery of Microsatellite Markers from Diploid Progenitor Species, *Arachis duranensis* and A. ipaensis, and Their Application in Cultivated Peanut (*A. hypogaea*). *Front. Plant Science.* 8. doi: 10.3389/fpls.2017.01209

Zheng, Y., Chen, B., Zhi, C., Qiao, L., Liu, C., Pan, Y., et al. (2021). Genome-wide identification of small heat shock protein (*HSP20*) homologs in three cucurbit species and the expression profiles of *CsHSP20s* under several abiotic stresses. *Int. J. Biol. Macromolecules* 190, 827–836. doi: 10.1016/j.ijbiomac.2021.08.222

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 19 March 2024
DOI 10.3389/fpls.2024.1342359

OPEN ACCESS

EDITED BY
Manohar Chakrabarti,
The University of Texas Rio Grande Valley, United States

REVIEWED BY
Sujan Mamidi,
HudsonAlpha Institute for Biotechnology, United States
Mehboob-ur- Rahman,
National Institute for Biotechnology and Genetic Engineering, Pakistan

*CORRESPONDENCE
Bahram Heidari
bheidari@shirazu.ac.ir
Maryam Salami
marysalami666@gmail.com

RECEIVED 21 November 2023
ACCEPTED 26 February 2024
PUBLISHED 19 March 2024

CITATION
Salami M, Heidari B, Alizadeh B, Batley J, Wang J, Tan X-L, Dadkhodaie A and Richards C (2024) Dissection of quantitative trait nucleotides and candidate genes associated with agronomic and yield-related traits under drought stress in rapeseed varieties: integration of genome-wide association study and transcriptomic analysis. *Front. Plant Sci.* 15:1342359. doi: 10.3389/fpls.2024.1342359

Dissection of quantitative trait nucleotides and candidate genes associated with agronomic and yield-related traits under drought stress in rapeseed varieties: integration of genome-wide association study and transcriptomic analysis

Maryam Salami[1*], Bahram Heidari[1*], Bahram Alizadeh[2], Jacqueline Batley[3], Jin Wang[4], Xiao-Li Tan[4], Ali Dadkhodaie[1] and Christopher Richards[5]

[1]Department of Plant Production and Genetics, School of Agriculture, Shiraz University, Shiraz, Iran, [2]Oil Crops Research Department, Seed and Plant Improvement Institute, Agricultural Research Education and Extension, Organization, (AREEO), Karaj, Iran, [3]School of Biological Sciences, University of Western Australia, Perth, WA, Australia, [4]School of Life Sciences, Jiangsu University, Zhenjiang, China, [5]United States Department of Agriculture (USDA), Agricultural Research Service (ARS), National Laboratory for Genetic Resources Preservation, Fort Collins, CO, United States

Introduction: An important strategy to combat yield loss challenge is the development of varieties with increased tolerance to drought to maintain production. Improvement of crop yield under drought stress is critical to global food security.

Methods: In this study, we performed multiomics analysis in a collection of 119 diverse rapeseed (*Brassica napus* L.) varieties to dissect the genetic control of agronomic traits in two watering regimes [well-watered (WW) and drought stress (DS)] for 3 years. In the DS treatment, irrigation continued till the 50% pod development stage, whereas in the WW condition, it was performed throughout the whole growing season.

Results: The results of the genome-wide association study (GWAS) using 52,157 single-nucleotide polymorphisms (SNPs) revealed 1,281 SNPs associated with traits. Six stable SNPs showed sequence variation for flowering time between the two irrigation conditions across years. Three novel SNPs on chromosome C04 for plant weight were located within drought tolerance-related gene *ABCG16*, and their pleiotropically effects on seed weight per plant and seed yield were characterized. We identified the C02 peak as a novel signal for flowering time, harboring 52.77% of the associated SNPs. The 288-kbps LD decay distance analysis revealed 2,232 candidate genes (CGs) associated with traits. The CGs *BIG1-D*, *CAND1*, *DRG3*, *PUP10*, and *PUP21* were involved in phytohormone signaling and pollen development with significant effects on seed number, seed weight, and grain yield in drought conditions. By integrating GWAS and RNA-seq, 215 promising CGs were

associated with developmental process, reproductive processes, cell wall organization, and response to stress. GWAS and differentially expressed genes (DEGs) of leaf and seed in the yield contrasting accessions identified *BIG1-D*, *CAND1*, and *DRG3* genes for yield variation.

Discussion: The results of our study provide insights into the genetic control of drought tolerance and the improvement of marker-assisted selection (MAS) for breeding high-yield and drought-tolerant varieties.

KEYWORDS

drought, gene ontology, linkage disequilibrium, QTN, RNA-sequencing, single-nucleotide polymorphism, yield

1 Introduction

Rapeseed (*Brassica napus* L.) is a member of the *Brassicaceae* family and is ranked second in global oilseed production (Meyer and Purugganan, 2013; FAO, 2018; Raboanatahiry et al., 2018). It is utilized almost entirely for oil production, food, feedstock, and biodiesel production. Rapeseed, originated from interspecific hybridization between turnip rape (*B. rapa*, AA, 2n = 20) and cabbage (*B. oleracea*, CC, 2n = 18), is the most widespread oilseed crop in various climates due to the ability to germinate and grow at low temperatures (Ren et al., 2000; Chalhoub et al., 2014; Koh et al., 2017; Wozniak et al., 2019).

Understanding the genetic bases of yield-related trait is of great significance for breeding high-yield rapeseed (Shi et al., 2011; Khan et al., 2018; Raboanatahiry et al., 2018; Zhu et al., 2021). Although both spring and winter growth habit forms have been identified in rapeseed, the winter form has shown a higher grain yield (Fordonski et al., 2016). The grain yield of rapeseed can be directly increased through selection for fecundity and by indirect selection for phenological traits which show complicated genetic control in plants (Nowosad et al., 2016; Marjanovic-Jeromela et al., 2019). Nevertheless, the polygenic nature of the genetic control of yield and its components and the influence of environmental variables complicates mining genetic loci contributing to stress tolerance in plants including rapeseed. Water-deficit stress in the period between the flowering stage and pod formation stages causes up to around 30% loss in grain yield (Elferjani and Soolanayakanahally, 2018).

Drought stress in arid and semiarid areas restricts plant growth and production in agriculture (Haq et al., 2014; Seleiman et al., 2021). At the early vegetative growth stage, drought slows down the rapeseed growth by inhibiting cell expansion and division (Yosefi and Heidari, 2022). However, drought at the reproductive development stage could drastically reduce rapeseed yield by producing small and medium-sized grains (Hatzig et al., 2018). Several studies have been conducted to evaluate the drought tolerance of rapeseed (Yarnia et al., 2011; Liu et al., 2015; Zhou et al., 2021). However, progress in improvement for drought tolerance is slow because of the complex genetic architecture of drought stress tolerance controlled by several minor and major genes (Bernardo, 2008). Understanding the genetic control of drought tolerance mechanisms can significantly accelerate the development of drought-tolerant varieties through marker-assisted selection (MAS) and genomic selection (GS).

Genome-wide association studies (GWASs) are currently the powerful tool to detect marker–trait associations (MTAs) and can be applied to mapping and identifying linked markers and candidate genes contributing to drought tolerance (Xiao et al., 2017). In rapeseed, GWAS has been used to identify loci and candidate genes for drought stress tolerance (Zhang et al., 2015; Khanzada et al., 2020; Shahzad et al., 2021; Salami et al., 2024). In a study consisting of 157 genotyped inbred rapeseed cultivars, GWAS was used to identify 320 SNPs linked with both seed weight (SW) and silique length (SL) traits and mapped to the gene *BnaA.ARF18* (Dong et al., 2018). In another study, 197 candidate genes were detected for budding, bolting, day to flowering (DTF), and the interval between DTF and bolting in rapeseed of which *FRIGIDIA* (*FRI*), *FLOWERING LOCUS C* (*FLC*), and *AGAMOUS-like 16* (*AGL16*) showed significant contribution to flowering time (Helal et al., 2021). Raman et al. (2019) conducted an association analysis in canola accessions using 11,804 SNPs under normal irrigation and water-stressed conditions and identified 47 SNPs on chromosome A02, and an additional 13 SNPs on chromosome C03 were associated with flowering close to *FLOWERING LOCUS T* (*FT)* and *FLOWERING LOCUS C* (*FLC*) genes for the drought avoidance mechanism (Raman et al., 2020). In the Hu et al. (2022) study, 628 SNPs were identified for 56 agronomically important traits through GWAS in a panel of diverse rapeseed accessions. A whole-genome resequencing and multilocus genome-wide association study (ML-GWAS) in rapeseed accessions revealed that 908 SNPs for agronomic and phenological traits of which 79 candidate genes were associated with *BnaA09g39790D* (*RNA helicase*), *BnaA09g39950D* (*Lipase*), and *BnaC09g25980D* (*SWEET7*) genes (Zhang et al., 2023a).

Transcriptomics and RNA-seq technology exploits transcript sequences to estimate patterns of gene expression, alternative

splicing, and allele-specific expression (Marguerat and Bahler, 2010; Zhang et al., 2023b). Transcript analysis can complement QTLs identified. In one study, RNA-seq was performed on eight tissues of extremely high- and low-harvest index (HI) rapeseed accessions and demonstrated that 33 functional candidate genes were located within the confidence intervals of significant SNPs associated with HI-related traits (Lu et al., 2016). In another study, Zhang et al. (2023a) performed GWAS and transcriptome analysis for seed yield and yield-related traits in *Brassica napus* for identification of differentially expressed genes (DEGs) in the seed of contrasting seed size/seed weight accessions.

Expanding rapeseed production areas through cultivation of drought-tolerant varieties could be an efficient strategy to alleviate the adverse effects of drought. Genomic studies focused on drought tolerance can be translated into breeding objectives for varietal development. In the present study, we aimed to identify novel SNPs and key genes for drought tolerance related traits in rapeseed. First, we investigated the effects of drought stress on agronomic, grain yield-related, and yield-related traits. Then, a GWAS approach was used to determine novel SNPs/genes for traits in the well-watered and drought stress conditions in 3 years. The RNA-Sequencing (RNA-Seq) experiment was performed in the leaf and seed of the contrasting high- and low-yielding varieties to validate the genes associated with the position of the identified linked SNPs. Differentially expressed genes (DEGs) in the leaf and seed of the accessions in the yield contrasting varieties to explore the genetic basis of drought tolerance can be facilitated by integrated functional genomic approaches.

2 Materials and methods

2.1 Plant materials, field experiment, and drought treatments

The plant materials used in this study consisted of 119 rapeseed varieties including breeding lines, hybrids, and commercial cultivars provided by the Institute of Seed and Plant Improvement (SPII), Iran (Additional File 1). The field experiment was performed at the Research Farm of Plant Production and Genetics, Shiraz, Iran in a 3-year trial in 2017, 2018, and 2020 growing seasons. Plants growth was incomplete in a 2019 trial which was due to spring frost damage. The texture of the soil was silty loam, and concentrations of micro- and macronutrients are shown in Additional File 2. Seasonal temperature, relative humidity, and mean precipitation are shown in Supplementary Figure S1. The experimental design was a lattice by patterning 11 × 11-unit cells (11 varieties and 11 units per block) with three replicates per watering regime [well-watered (WW) and drought stress (DS)]. The plot size was 1 m^2 with the between-plot distances of 0.5 m. Each plot was composed of four 1-m-long rows, each with 25 plants per 4-cm spaces. The seeds were sown on four rows 1 m in length on 17 September for the three seasons.

The fertilizers were applied at the rates of 250 kg N ha^{-1}, 100 kg P ha^{-1}, and 100 kg K ha^{-1}. Phosphorus and potassium fertilizers were incorporated to the soil prior to sowing, and nitrogen was used as top dressed in different growth stages of rapeseed. In the well-watered condition, irrigation was carried out from planting till seed physiological maturity (maximum seed dry weight) as previously described by Ozer (2003). In the drought stress treatment, irrigation continued till the 50% pod development stage when irrigation then stopped till end of the growing season, which made three irrigation practices less in the drought-stressed plants than in the well-watered plants throughout the growing season. The numbers of irrigations in WW and DS treatments were 9 and 6, respectively.

For weed control, 3 L ha^{-1} Treflan® HFP herbicide was sprayed at sowing and hand weeding was also followed during the growing season. The Pirimor 50 pesticide at a rate of 2 L ha^{-1} was used for the aphids on rapeseed at the flowering and early podding periods. Harvesting time was the first week of July when the siliquae in terminal raceme turned creamy white in color.

2.2 Measurement of phenotypic characteristics

Days to flowering (DTF) was measured as the interval between the time of sowing and the time when the first flowers opened on 50% of the plants followed by Matar et al. (2021) description. Days to silique development (DTSD) was recorded as a number of days from the sowing to the time that first pods appeared on 50% of the plants. Days to ripening (DTR) was measured as the interval between the dates of sowing and the time when pods were dried. Plant height (PH) (cm) was recorded from the ground to the tip of the main pod at the ripening stage. For branch number/plant (BNPP) and yield components including silique length (SL; cm), seed number/silique (SNPS), seed weight/plant (SWPP; g), and thousand seed weight (THSW; g), 10 randomly plants were harvested from the middle rows to avoid border effects in each plot. At the harvesting time, plants in two middle rows were cut for plant weight (PW; g), seed yield (SY; kg ha^{-1}), and harvest index (HI; %) measurements. The grain weight of 10 spikes was used as grain weight per spike. The HI was calculated by dividing the grain yield by the biological yield.

2.3 Analysis of variance and estimation of genetic variation and genetic gain

Descriptive parameters including mean and standard deviation (SD) were calculated for each treatment in SAS software (version 9.4). Box and whisker plots were used for the graphical presentation of the descriptive statistics. The packages ggplot2 in R (version 4.3.2) for win (http://CRAN.R-project.org/, accessed on 23 February 2021) and RStudio (version 1.3.1093) (https://rstudio.com/, accessed on 23 February 2021) were used for analysis of boxplots (McGill et al., 1978; Wickham, 2016). The correlation matrix between variables and a constructing heat map of correlation coefficients were computed using packages plotly, heatmaply, and ggcorrplot (https://cran.rproject.org/web/packages/ggcorrplot/index.html) in the R software.

The PROC GLM procedure was used for combined analysis of variances (ANOVA) in the Statistical Analysis System (SAS)

software (SAS Institute Inc., Cary, NC, USA, version 9.3) (Dodig et al., 2008). The RANDOM statement with the TEST as option procedure was used to define the year as a random effect and water treatment and genotype as fix effects in ANOVA.

Phenotypic and genotypic variances were calculated as shown in Equation 1 using the expected mean squares (EMS) of sources of variations in ANOVA as follows (Lush, 1949),

$$\sigma_g^2 = \left(\frac{(\mathrm{MSG} - \mathrm{MSE})}{r}\right) \times (100) \quad (1)$$

where σg^2 is the genotypic variance, MSG is the mean square for genotype, MSE is the error mean square in ANOVA, and r is the number of replications.

$$\sigma_p^2 = \sigma_g^2 + \sigma_e^2 \quad (2)$$

where σ_p^2 and σ_e^2 are the phenotypic and environmental variances, respectively (Equation 2).

The environmental, genotypic, and phenotypic coefficients of variation were calculated as shown in Equations 3, 4, and 5 using as follows (Burton, 1952):

$$\mathrm{ECV} = \frac{MSE}{\bar{x}} \times (100) \quad (3)$$

$$\mathrm{GCV} = \frac{(\sigma_g)}{\bar{x}} \times (100) \quad (4)$$

$$\mathrm{PCV} = \frac{(\sigma_p)}{\bar{x}} \times (100) \quad (5)$$

where ECV is the environmental coefficient of variation, GCV is the genotypic coefficient of variation, PCV is the phenotypic coefficient of variation, σ_g is the root of genotypic variance, σ_p is the root of phenotypic variance, and $\bar{x}$ is the trait mean.

Variance components were used to calculate the broad-sense heritability (h^2) in Equation 6 as follows (Marwede et al., 2004):

$$h^2 = \frac{\sigma_g^2}{\sigma_p^2} \times (100) \quad (6)$$

Simple genetic advance (GA) and GA over means (GAM) were calculated as shown in Equations 7 and 8 as follows::

$$\mathrm{GA} = (k \times h^2 \times \sigma_p / \bar{x}) \quad (7)$$

$$\mathrm{GAM} = (GA/\bar{x}) \times (100) \quad (8)$$

where k is the selection intensity which was 1.76 denoting selection of 10% of top-ranked varieties, h^2 is heritability in a broad sense, and $\bar{x}$ is the trait mean.

2.4 Reference mapping and variant calling

Pseudo-genome sequences of the diploid A (283.8 Mb) and C genomes (488.6 Mb) were combined and used as the reference sequences for mapping analyses. Reads for each genotype were aligned independently to the reference genome using CLC Genomics Workbench (version 7.0.4). The mapped reads were interrogated for sequence variation using the CLC Bio probabilistic variant calling tool. A minimum depth of coverage of 3× for 454 and 8× for Illumina data was required for SNP calling. Mapping data and variant calls were exported from CLC and combined using a custom Perl script to determine reference, or variant call for every genotype at all variant positions.

2.5 DNA extraction and single SNP-based association mapping

Genomic DNA was isolated from the fresh leaves collected from a bulk of five randomly chosen plants per variety in a greenhouse (Murray and Thompson, 1980). The genomic DNA was quantified using a NanoDrop ND-1000 Spectrophotometer (NanoDrop Technologies, Inc., Wilmington, DE, United States). The DNA samples were used for genotyping *Brassica* 60 K Infinium array as described in the manufacturer's protocol (Illumina Inc., San Diego, CA). Quality preprocessing of 52,157 SNPs obtained from 60K chips was done by using TASSEL software v5 (Bradbury et al., 2007). The SNPs were filtered for site coverage (90%), minimum minor allele frequency (MAF) of 0.05 with only biallelic markers, and low rates of missing data (≤10%) using the TASSEL (version 5).

2.6 Population genetics analysis, linkage disequilibrium, and LD decay

The polymorphism information content (PIC) value of each SNP locus in all varieties and the PIC values on each chromosome were calculated by PowerMarker (version 3.2.5) (Liu and Muse, 2005). We generated 29,310 SNPs involving 119 varieties derived from Iran, Germany, France, America, Australia, Hungary, Serbia, and Russia (Additional File 1). For population structure analysis, the natural logarithms of probability data (LnP(K)) and the *ad hoc* statistic ΔK were calculated (Su et al., 2017). The population structure underlying the collection of rapeseed varieties was analyzed using STRUCTURE (version 2.3.4) (Hubisz et al., 2009). The model-based Bayesian cluster analysis program was used to identify subpopulations. A total 10,000 burn-in periods followed by 100,000 Markov Chain Monte Carlo (MCMC) iterations from K = 3 to K = 10 were used to identify the optimal number of clusters (K). Three independent runs were generated for each K. The results were collated by the Structure Harvester tool (Earl and VonHoldt, 2011), and the best K-value was identified based on the delta K method (Evanno et al., 2005). A neighbor-joining (NJ) tree was created to validate population stratification with the software TASSEL (version 5). A PCA was done on the significant SNPs data using R software (version 4.3.2) (R Core Team, 2018) with the ggplot2 (Wickham, 2016) and ape (Paradis and Schliep, 2018) packages, respectively. To investigate chromosome-wide and the genome-specific patterns of linkage disequilibrium (LD) (r^2), the software TASSEL (version 5) (www.maizegenetics.net/) with 1000 permutations was used. After quality control processing, a total of 29,310 high-quality SNPs with MAF ≥0.05 and pairwise r^2 values

were used to determine the extent of LD decay across genome and among chromosomes.

2.7 Genome-wide association analysis

The marker–trait associations (MTAs) were analyzed using the program TASSEL (version 5). Four models, namely, general linear model (GLM) with the Q matrix of population structure (GLM + Q), mixed linear model (MLM) with both the kinship (K) as a random effect and Q matrices (MLM + Q), GLM model with the major principal components (PC) matrix (GLM + PC) and MLM with both the PC and K matrices (MLM + K + PC), were used to identify the MTAs. In GWAS, five PCs based on their cumulative eigenvalue contribution were used in population structure analysis. The Q matrix obtained from structure analysis and the relative kinship and PC matrices were calculated by TASSEL software. The phenotypic variation explained by significant SNP marker (R^2) was calculated in TASSEL (version 5) (Bradbury et al., 2007). Quantile–quantile (QQ) plots were shown with −log10(P) of each SNP and expected P-value using the R package qqman (https://cran.rproject.org/web/packages/qqman/index.html). Manhattan plots were drawn in TASSEL software.

2.8 Screening candidate genes overlapped with the SNP position

To identify candidate genes (CGs) related to the SNPs of traits under the well-watered and drought-stressed conditions, the flanking sequences of the linked SNPs obtained from the "Darmor-bzh" reference genome (http://www.genoscope.cns.fr/brassicanapus/data/) was used to search in the rapeseed genome by Ensembl Plants (https://plants.ensembl.org/). Consequently, all the genes underlying the genomic region of each SNPs were functionally annotated by Ensembl Plants (https://plants.ensembl.org/) and online resources (https://genome.ucsc.edu/ and https://www.ncbi.nlm.nih.gov/). The CGs were identified based on their putative function in rapeseed or closely related species.

2.9 Allele effect and haplotype analyses

The allele effects for the linked significant SNPs were analyzed as previously described by Alemu et al. (2021). Varieties were divided into two different groups according to their specific SNP alleles, and the means were compared using Turkey's honest significant difference (HSD) test. Exploring and harnessing haplotype diversity helps in the detection of CGs for improvement of target traits in crops (Qian et al., 2017). A haplotype association test was performed to investigate the combined effect of the linked significant SNPs. The SNPs in the same haploblock and the LD of significant SNPs were determined using Haploview (version 4.2) (Barrett et al., 2005). A standardized disequilibrium coefficient (D′) was used to evaluate the LD between markers and generate the LD heatmap. Haploid blocks were detected based on LD using the confidence intervals (CI) method in Haploview (version 4.2) (Gabriel et al., 2002).

2.10 RNA extraction, library construction, and RNA-sequencing

To validate the GWAS-identified SNPs contributed to drought tolerance, the expression of candidate genes associated with the position of the related SNPs was analyzed in two drought-tolerant and two drought-sensitive varieties. The experiment consisted of the RNA-Seq analysis in two top high- (G19 and G41) and two top low- (G111 and G114) yielding varieties showing contrasting yield under the WW and DS conditions. The plants were grown in 20-cm-diameter pots in the greenhouse under 12-h light/12-h dark conditions with normal experimental management. There were six plants in each pot. When the flower buds became visible, plants were randomly divided into two groups, each with three plants: the control group and the drought stress treatment group and each experiment underwent three biological replicates. During irrigation, drought-treated flower pots maintain a soil moisture content of 10% (irrigated with PEG6000 at a concentration of 20%), whereas well-watered flower pots maintain a soil moisture content of 30% (irrigated with sterile water of equal volume). Thirty days after flowering, the leaf and mature seeds were harvested from plants in each replicate group and each condition (the well-watered and drought stress treatments). All harvested seeds and leaves were immediately frozen using liquid nitrogen and transferred to a deep freezer (−80°C) for storage.

A total of 24 samples (control and treatments with three biological replicates, respectively) were prepared for RNA-Seq. Total RNA was extracted from the seeds and leaves using a Plant RNA Mini Kit (Tiangen, Inc., China) according to the manufacturer's instruction. Four cDNA libraries were constructed, and RNA-Seq was performed on a DNBSEQ-G400 platform.

Low-quality reads were filtered out using the NGS QC toolkit (version 2.2.3) (https://omictools.com/ngs-qc-toolkit-tool) (Patel and Jain, 2012). High-quality reads from the raw sequencing reads were matched to the *B. napus* reference genome of "Darmor-bzh" (http://www.genoscope.cns.fr/brassicanapus/). The identified genes in the previous step were quantitatively analyzed using Cluffquant and Cluffnorm of Cufflinks 2.0.0 (http://cole-trapnell-lab.github.io/cufflinks/releases/v2.0.0/).

2.11 Identification of differentially expressed genes

The DEGs were identified based on FPKM (fragments per kilo base of transcript per million mapped fragment) and Q value (<0.05) (Q value: error-corrected value after multiple testing), and a $\log_2$ (fold change) ≥1 was set as the threshold to identify the significance of gene expression differences. Furthermore, to verify the statistical significance and hierarchical clustering of DEGs, a heat map was generated using R software (version 4.3.2).

2.12 Enrichment analyses of Gene Ontology and Kyoto Encyclopedia of Genes and Genomes pathways

To further understand the function of DEGs, we performed Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene Ontology (GO) analyses on the identified DEGs. The sequence file of each gene was used as input into Eggnog software (version 2.0.1) to identify annotation of genes (Huerta-Cepas et al., 2017). GO and KEGG analyses were conducted using the ClusterProfiler (version 4.0.0) R package. Only GO terms or KEGG pathways with *P*-value <0.05 verified for subsequent analyses. The REVIGO program (http://revigo.irb.hr/) was used to remove redundant GO terms (Supek et al., 2011).

2.13 Integration of genome-wide association study and transcriptome data

The RNA-seq data were used for the ratios of genome-wide up- and downregulated DEGs. In addition, we calculated the ratios of up- and downregulated DEGs within the 288-kbp intervals corresponding to the significant SNPs of the GWAS analysis. Then, we compared the DEG ratios for all genome-wide genes and for potential drought tolerance-related genes detected in the GWAS of the high- and low-yield contrasting varieties under drought stress.

3 Results

3.1 Phenotypic variation and heritability of traits

The results of ANOVA for the main effects of treatments and the interactions are shown in Additional File 3. The effects of year (Y), environment (E), and genotype (G) were statistically significant (*P*-value < 0.01), which shows variation among varieties for traits over years and irrigation treatments. The G × Y, G × E, Y × E, and Y × G × E interactions were significant for all traits. Phenotypic variation for traits under normal irrigation and drought stress conditions is shown in Table 1; Figures 1A–M. The results showed that PW, SWPP, HI, and SY had high variation among the 119 rapeseed varieties in two irrigation regimes.

The SWPP, SY, HI, and PW traits showed significant differences between normal irrigation and drought stress conditions. Compared with normal irrigation, drought stress significantly reduced SWPP, SY, HI, PW, BNPP, THSW, PH, SL, SNPS, and DTR by 67.08%, 53.62%, 44.09%, 44.09%, 31.92%, 24.58%, 22.85%, 11.51%, 4.03%, 1.73%, and 1.03%, respectively (Table 1; Supplementary Figure S2).

Analysis of genetic variation showed that the traits were divided into three groups with PCV and GCV above 20% as high, 10%–20% as moderate, and below 10% as low (Table 1). In normal irrigation conditions, SWPP, SY, and PW had high PCV and GCV in 3 years (Table 1). Moderate PCV and GCV were recorded for SL and SNPS in 3 years. The DTF, DTSD, and DTR traits in 3 years and PH and BNPP in two years showed low PCV and GCVs (Table 1). Under drought conditions, the PCV estimates of various characters varied from 0.48% for DTR to 51.08% for SY and SWPP (Table 1). The GCV estimates varied from 0.61% for DTR to 80.77% for SWPP. Both GCV and PCV were high (>20%) for PW, SY, SWPP, and HI in 3 years and for SNPS in 2018 and 2020, THSW in 2017 and 2018. DTF, DTSD, and DTR had lower variation (<6%) in drought stress condition in 3 years.

Heritability estimates for most of the traits were higher under drought compared with well-watered conditions. The heritability estimates for traits in the WW condition ranged from 3.24% for HI to 99.18% for SNPS and from 35.28% for SWPP to 99.18% for SNPS in drought stress. The estimates of heritability were moderate for PW and SY in 3 years (Table 1). Genetic advances (GA) ranged from 0.33 for HI to 2106.05 for SY in normal irrigation conditions, and it ranged from 0.77 for THSW to 1503.2 for SY in drought stress. High heritability values coupled with high GA were recorded for PW, SWPP, and SY under well-watered conditions in three growing seasons. Genetic advance normalized based on the trait mean (GAM) under well-watered conditions ranged from 63.19% for SY to 0.75% for DTR and from 0.65% for DTR to 85.34% for SY in drought stress. Among the tested traits, SNPS showed high heritability (>80%) coupled with the high GAM (>20%) under both irrigation conditions across 3 years.

3.2 Interrelationship of agronomic and yield-related traits

Analysis of correlation of traits helps in indirect selection for yield improvement. Under WW conditions, significant and relatively high correlations were identified for DTF with DTSD and DTR (0.85**, 0.60**) and DTR with DTSD (0.61**) (Figure 1N). Correlations of SY with PW (0.62**) and SWPP (0.83**) were significant. Under drought conditions, a positive and significant correlation coefficient was found for SWPP with each SY, HI, and PW traits (0.65**, 0.60**, and 0.54**) whereas DTSD had a strong positive correlation (0.95**) with DTF (Figure 1O).

3.3 Morphological variations between high- and low-yield varieties

Owing to the lower complexity of yield components and lower influence of environmental effects compared with yield, use of yield-related traits with high heritability as indirect selection for improvement of grain yield is preferred. We assessed the difference of agronomic traits and yield components in the high- and low-grain yield varieties. The results indicated that BNPP, SWPP, and PW were significantly larger in the high-yield varieties than in the low-yield varieties (*P*-value < 0.01**). However, PH, THSW, and HI were relatively similar between the two contrasting groups (Figures 2A–L).

10.3389/fpls.2024.1342359

TABLE 1 Phenotypic variation of agronomic and yield related traits in 119 rapeseed (*Brassica napus* L.) varieties.

Trait	Watering regime	Mean	SD	Min	Max	ECV (%)	PCV (%)	GCV (%)	h^2 (%)	GA	GAM (%)
DTF	WW17	183.01	5.82	168	189	2.66	3.18	4.15	58.73	7.84	4.29
	DS17	186.00	6.34	171	192	1.20	3.39	3.59	88.91	10.45	5.62
	WW18	180.64	3.57	171	188	0.89	1.98	2.17	83.26	5.74	3.18
	DS18	180.62	4.13	166	187	0.48	2.23	2.28	95.65	6.93	3.83
	WW20	183.54	9.03	171	201	0.54	4.95	4.98	98.82	15.91	8.67
	DS20	184.89	9.14	172	202	0.83	4.98	5.05	97.31	16.00	8.65
DTSD	WW17	190.73	4.97	179	197	2.02	2.56	3.26	61.70	6.75	3.54
	DS17	193.04	5.81	182	190	1.15	3.01	3.22	87.19	9.54	4.94
	WW18	187.73	3.18	179	198	0.56	1.67	1.76	89.81	5.22	2.78
	DS18	186.88	3.45	173	195	0.55	1.79	1.87	91.39	5.62	3.01
	WW20	192.47	10.02	179	213	0.92	5.22	5.30	96.97	17.40	9.04
	DS20	193.72	9.71	181	212	0.95	5.01	5.10	96.52	16.77	8.66
DTR	WW17	256.84	2.27	253	273	0.79	0.87	1.17	54.99	2.91	1.13
	DS17	255.31	2.15	252	262	0.56	0.84	1.01	69.02	3.12	1.22
	WW18	268.80	1.46	266	274	0.42	0.54	0.68	62.46	2.01	0.75
	DS18	267.80	1.28	265	271	0.39	0.48	0.61	60.14	1.74	0.65
	WW20	278.58	4.84	265	289	0.33	1.75	1.78	96.64	8.41	3.02
	DS20	272.82	5.60	254	285	0.28	2.03	2.05	98.14	9.64	3.53
PH	WW17	108.20	19.35	41.33	157.55	6.11	18.04	19.05	89.73	32.55	30.08
	DS17	86.58	14.25	61.14	141.33	12.16	16.56	20.54	64.97	20.33	23.49
	WW18	178.24	2.37	123.45	141.33	0.84	0.78	1.15	46.38	2.80	0.94
	DS18	170.60	5.09	111.11	130.45	1.06	1.77	2.06	73.47	7.74	2.67
	WW20	142.31	12.50	113.34	173.12	2.09	8.81	9.05	94.69	21.47	15.09
	DS20	122.21	8.80	100.43	154.32	2.15	7.18	7.50	91.80	14.81	12.11
BNPP	WW17	84.21	9.92	29	92	7.09	11.76	13.73	73.31	14.92	17.72
	DS17	28.10	4.16	23	47	8.19	14.69	16.82	76.28	6.35	22.58
	WW18	168.53	12.20	125	193	4.70	7.37	8.75	71.06	18.43	10.94
	DS18	143.06	12.13	115	172	4.75	8.55	9.78	76.43	18.83	13.16
	WW20	181.23	12.35	149	218	1.85	6.94	7.18	93.34	21.39	11.80
	DS20	156.10	10.48	129	186	2.58	6.66	7.15	86.96	17.07	10.94
PW	WW17	1,599.65	543.50	575.76	3,575.42	27.57	33.64	43.50	59.82	732.57	45.80
	DS17	472.43	256.80	120.45	1,535.34	41.01	53.89	67.72	63.33	356.62	75.49
	WW18	2,421.35	530.42	665.45	3,900.32	13.29	21.93	25.64	73.14	799.14	33.00
	DS18	2,302.02	480.84	345.42	3,395.33	13	20.73	24.47	71.77	711.41	30.90
	WW20	2,726.34	589.07	835.24	4,130.35	11.95	21.65	24.73	76.66	909.55	33.36
	DS20	1,818.87	456.85	865.32	3,195.46	19.35	25.12	31.71	62.75	636.94	35.02
SL	WW17	7.87	0.85	5.32	10.33	7.23	10.68	12.89	68.55	1.22	15.56
	DS17	7.81	0.89	6.14	10.33	4.18	17.45	17.94	94.56	2.33	29.86
	WW18	6.85	0.83	5.21	9.53	2.98	11.05	11.45	93.23	1.29	18.78

(Continued)

10.3389/fpls.2024.1342359

TABLE 1 Continued

Trait	Watering regime	Mean	SD	Min	Max	ECV (%)	PCV (%)	GCV (%)	h^2 (%)	GA	GAM (%)
	DS18	6.69	0.64	5.32	9.50	3.75	8.37	9.17	83.26	0.90	13.44
	WW20	7.10	0.99	5.12	9.32	1.99	12.76	12.91	97.63	1.58	22.18
	DS20	6.44	0.90	4.52	9.39	2.71	12.96	13.24	95.81	1.44	22.32
SNPS	WW17	23.53	3.82	19	32	4.55	16.19	16.82	92.68	6.46	27.44
	DS17	24.50	4.08	17	39	2.19	16.65	16.79	98.30	7.12	29.06
	WW18	20.52	4.49	11	31	1.97	21.68	21.77	99.18	7.80	38.00
	DS18	19.73	4.47	11	31	11.15	22.56	25.16	80.38	7.02	35.60
	WW20	23.40	4.93	11	36	2.63	20.82	20.98	98.43	8.51	36.35
	DS20	22.05	4.56	11	39	2.47	20.59	20.74	98.58	7.93	35.97
SWPP	WW17	155.71	79.35	20.33	398.45	72.79	51.25	89.02	33.15	80.87	51.93
	DS17	35.24	16.84	12.45	122.38	64.98	47.97	80.77	35.28	17.67	50.15
	WW18	135.47	38.37	50.22	235.25	29.71	28.62	41.25	48.14	47.35	34.95
	DS18	109.01	39.86	33.11	219.42	34.88	36.60	50.56	52.41	50.84	46.63
	WW20	511.78	143.88	125.43	845.30	23.05	27.98	36.25	59.58	194.53	38.01
	DS20	120.03	71.82	30.54	375.64	54.77	59.44	80.83	54.08	92.34	76.93
THSW	WW17	2.93	1.07	1.32	6.11	31.29	35.36	47.22	56.09	1.37	46.62
	DS17	2.10	0.85	1.24	4.39	20.19	40.78	45.50	80.32	1.35	64.32
	WW18	3.48	0.68	1.09	4.89	7.79	18.10	19.70	84.38	1.02	29.26
	DS18	2.05	0.65	1.05	3.97	20.30	26.86	33.67	63.64	0.77	37.71
	WW20	4.57	0.57	2.36	6.43	2.60	10.20	10.53	93.89	0.79	17.40
	DS20	4.32	0.59	2.15	7.42	2.81	10.80	11.16	93.67	0.79	18.40
HI	WW17	10.91	6.55	1.32	32.34	51.86	9.5	52.73	3.24	0.33	3.01
	DS17	8.53	3.43	3.76	21.46	50.98	39.84	64.70	37.91	3.68	43.17
	WW18	5.78	1.75	3.31	13.17	34.31	30.09	45.63	43.48	2.02	34.92
	DS18	4.88	1.87	2.70	15.75	40.95	37.74	55.69	45.92	2.20	45.01
	WW20	18.73	3.33	7.45	27.33	21.39	17.64	27.72	40.48	3.70	19.75
	DS20	6.39	2.70	2.45	16.35	46.58	41.3	62.26	44.01	3.08	48.22
SY	WW17	2,594.18	1,096.08	396.34	7,666.34	25.77	42.10	49.36	72.74	1,639.3	63.19
	DS17	1,918.77	986.37	630.71	7,378.25	45.85	57.09	73.22	60.79	1503.2	78.34
	WW18	1,352.23	383.93	500.22	2,350.21	24.32	28.67	37.60	58.16	520.46	38.49
	DS18	1,088.11	398.32	330.32	2,190.25	30.76	36.72	47.90	58.77	539.12	49.55
	WW20	5,125.42	1,443.27	1,250.37	8,450.23	18.61	28.02	33.64	69.41	2,106.05	41.09
	DS20	1,200.34	718.25	300.34	3,750.25	42.13	59.44	72.85	66.55	1,024.36	85.34

DTF, DTSD, DTR, PH, BNPP, PW, SL, SNPS, SWPP, THSW, HI, and SY are the abbreviations of days to flowering, days to silique development, days to ripening, plant height, branch number/plant, plant weight, silique length, seed number/silique, seed weight/plant, thousand seed weight, harvest index, and seed yield, respectively. WW17, DS17, WW18, DS18, WW20, and DS20 are the codes of the two watering regimes during 3 years: well-watered in 2017, drought stress in 2017, well-watered in 2018, drought stress in 2018, well-watered in 2020, and drought stress in 2020. SD, ECV, PCV, GCV, h^2, GA, and GAM are the abbreviations of standard deviation, environmental coefficient of variation, phenotypic coefficient of variation, genotypic coefficient of variation, heritability in the broad sense, genetic advance, and genetic advance as the percentage of the mean of the studied traits at two watering regimes under 3 years.

10.3389/fpls.2024.1342359

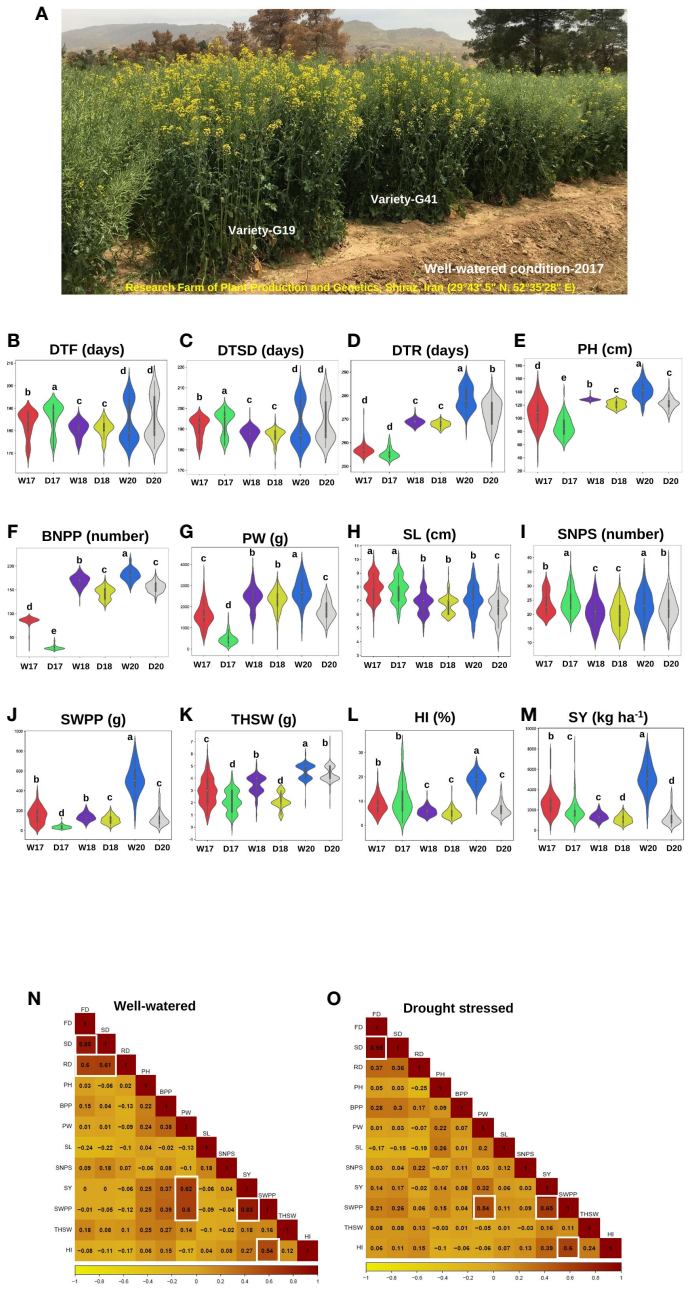


FIGURE 1
Agronomic and yield-related traits phenotyping in rapeseed (*Brassica napus* L.) varieties under two watering regimes (well-watered and drought stress conditions) across three years, 2017, 2018, and 2020. **(A)** A diverse collection of rapeseed varieties was assessed for agronomic and yield-related traits under well-watered conditions in 2017 in Research Farm of Plant Production and Genetics, Shiraz, Iran (29°43′ 5″ N, 52° 35′28″ E), highlighting the range of phenotypic diversity within the panel. G19 and G41 were high-yield varieties. **(B-M)** The violin plots illustrating the changes in agronomic and yield-related traits under well-watered and drought stress conditions from the individual data sets of 3 years (2017, 2018, and 2020) of 119 rapeseed varieties. The width of the violin plot represents the density of the distribution. The white dot in the box plot shows the median value, and the upper and lower boxes in the box represent the upper and lower quartiles of the data set. Data are means ± SD, *P*-value < 0.05, as determined by multiple comparison testing by one-way ANOVA. Traits represent as **(B)** days to flowering (DTF), **(C)** days to silique development (DTSD), **(D)** days to ripening (DTR), **(E)** plant height (PH), and **(F)** branch number/plant (BNPP), **(G)** plant weight (PW), **(H)** silique length (SL), **(I)** seed number/silique (SNPS), **(J)** seed weight/plant (SWPP), **(K)** thousand seed weight (THSW), **(L)** harvest index (HI), and **(M)** seed yield (SY). WW17, DS17, WW18, DS18, WW20, and DS20 were the codes of the two watering regimes during 3 years: well-watered in 2017, drought stress in 2017, well-watered in 2018, drought stress in 2018, well-watered in 2020, and drought stress in 2020, respectively. **(N, O)** Heat map showing the correlation between the agronomic and yield-related traits under two watering regimes in three growing seasons (2017, 2018, and 2020). **(N)** Well-watered condition, **(O)** drought stress condition. Traits represent as flowering date (FD), silique date (SD), ripening date (RD), plant height (PH), branches per plant (BPP), plant weight (PW), silique length (SL), seed number/silique (SNPS), seed yield (SY), seed weight/plant (SWPP), thousand seed weight (THSW), and harvest index (HI). A color scale showing the correlation values ranging from dark yellow, –1, to orange, 0, to 1, dark red is shown below the heat map.

10.3389/fpls.2024.1342359

3.4 Distribution of SNPs, LD, LD decay, and population structure

After filtering low-quality SNPs (call rate <90% and minor allele frequency <0.05) in TASSEL software, a set of 29,310 high-quality SNPs was used for genetic variation and GWAS analyses. The SNP markers were not evenly distributed across the whole genome with the A subgenome having a higher number of SNP markers (14,925; 50.92%) compared to the C subgenomes (Supplementary Figures S3A, D). However, the density of SNPs in the C subgenome (42.96 SNPs/kbps) was higher than that in the A subgenome (19.76 SNPs/kbps) (Additional File 4). Among all chromosomes, C04 (2,624 SNPs) and C05 (723 SNPs) had the highest and lowest numbers of markers (Supplementary Figure S3A). The PIC values for chromosome ranged from 0.24 to 0.38 (Additional File 4). The mean PIC values of the A and C subgenomes were 0.32 and 0.32, respectively.

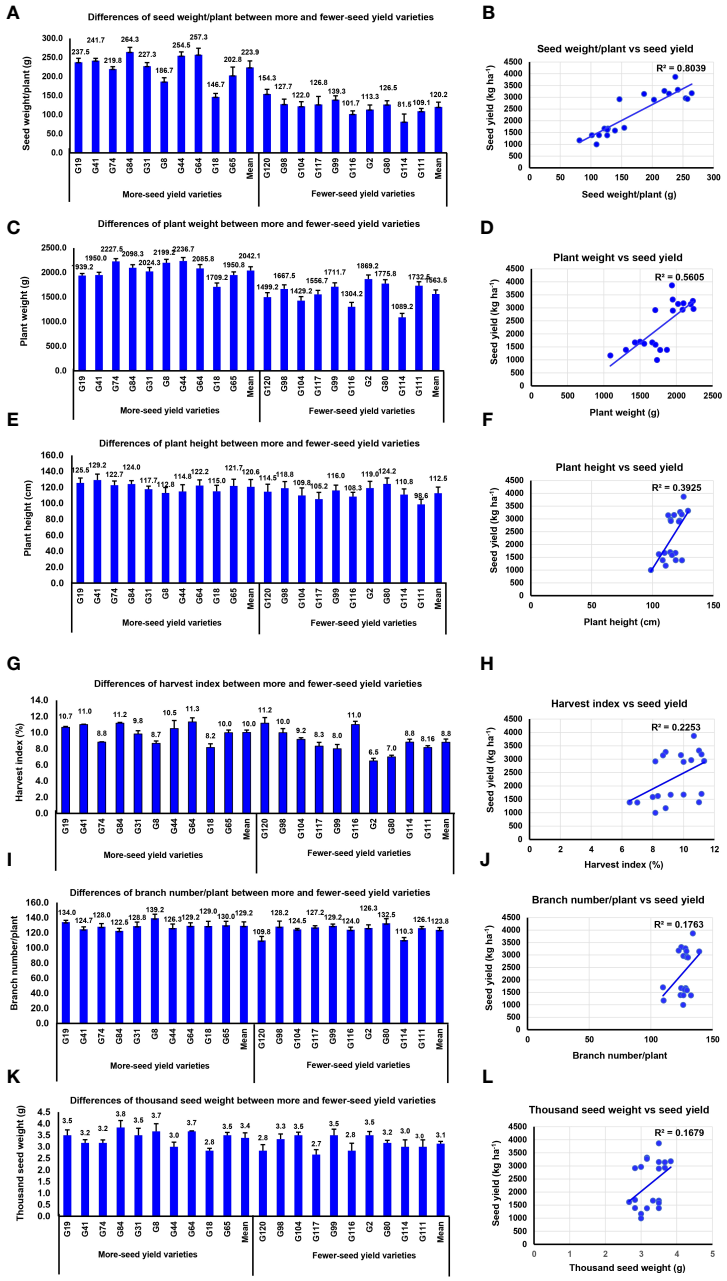


FIGURE 2
The seed yield variation was caused by agronomic and yield components. **(A)** Differences of seed weight/plant between high- and low-yield varieties. **(B)** Correlation analysis between seed yield and seed weight/plant. **(C)** Differences of plant weight between high- and low-yield varieties. **(D)** Correlation analysis between seed yield and plant weight. **(E)** Differences of plant height between high- and low-yield varieties. **(F)** Correlation analysis between seed yield and plant height. **(G)** Differences of harvest index between high- and low-yield varieties. **(H)** Correlation analysis between seed yield and harvest index. **(I)** Differences of branch number/plant between high- and low-yield varieties. **(J)** Correlation analysis between seed yield and branch number/plant. **(K)** Differences of thousand seed weight between high- and low-yield varieties. **(L)** Correlation analysis between seed yield and thousand seed weight.

Linkage disequilibrium was examined as the squared Pearson correlation coefficient (r^2) between all pairs of SNP markers. The LD in the C subgenome was significantly higher LD (0.031) than in the A subgenome (0.025). The C05 chromosome showed the highest LD (0.088) among chromosomes (Additional File 4). The LD decay with an average of 288 kbps in the whole genome ranged from 101.61 in A10 to 953.78 kbps in A08. However, the LD decay distance for C03 was 942.6 kbps, which was large compared with those for other chromosomes in the C subgenome (Additional File 4).

In analysis of population structure, the peak of the broken line was observed at k = 7 suggesting that the tested population can be divided into seven distinct groups and one mixed group partly correlated with their origins (Supplementary Figure S3B). Neighbor-joining (NJ) cluster analysis was performed to explore the relatedness among the rapeseed varieties. The NJ tree showed that our varieties could be divided into four groups (Supplementary Figure S3C). The first group (group A) was composed of 30 breeding lines originated from Iran and Australia, group B was composed of 10 varieties from Iran and Germany, 55 breeding lines in group C originated from Iran, and Group D had 24 varieties composed of hybrids and cultivars originated in France and America. The PCA based on the genome-wide SNPs supported the results of population structure and phylogenetic tree (Supplementary Figure S3E).

3.5 Significant SNPs and candidate genes associated with traits

We measured 12 traits including plant architecture, phenological and agronomic traits, and yield components in 119 rapeseed varieties grown under two irrigation conditions over 3 years. Using 29,310 high-quality SNPs, our GWAS for these 12 traits revealed 1,281 unique SNPs linked with the traits at the *P*-value < 10^{-4} threshold (Supplementary Figures S3F, S4-S6; Additional Files 5-8). Higher numbers of significant SNPs associated with traits were identified in the GLM + Q and MLM + K + PC models. However, the result of the MLM model was preferred, which showed fewer false positives than the GLM model. False positives are often controlled by incorporating covariates for the kinship and PCs matrices in the MLM model (Liu et al., 2016). Accordingly, the results of the MLM + K + PC model were used for further analysis. The number of significant SNPs linked with traits was variable. Higher SNPs were found for the DTF, BNPP, and SY traits than others. Based on statistical significance and the repeatability of the linked SNPs in two irrigation treatments across years, several important SNPs are shown in Tables 2, 3. Generally, the effective candidate regions with significant GWAS signals were defined as the LD blocks surrounding the signal peak (Yano et al., 2016). Based on the 288-kbps LD decay distance and CG analysis, we identified 2,232 genes as GWAS-associated CGs (Additional Files 9, 10). The SNPS (557) trait followed by DTF (524) showed the highest number of CGs. The key genes related to four agronomically important traits were selected for further functional verifications.

3.5.1 SNPs and candidate genes linked with phenological and agronomic traits under drought conditions

3.5.1.1 Days to flowering

A total of 324 significant SNPs were significantly associated with the DTF (Additional Files 5-8). The SNP Chr10:222927 linked with flowering time in both WW and DS conditions across 2 years (Table 2). Six linked SNPs, (Chr10:222927, Chr10:8893827, Chr10:8998128, Chr10:16091976, Chr10:16091976, Chr10:16139771) were significant across 2 years under drought conditions (Table 2). We found that the C02 peak was a major associated signal, harboring 52.77% of the associated SNPs (171) (Additional File 5; Figures 3A–D), which had not been reported previously. Accordingly, we focused on the 159 SNPs positioned in the Chr12:23345227–32168120 genomic region and extracted all CGs within 200 kbps of the most significant SNPs. A number of 146 genes, including transcription factors, enzymes, and transporters that represent plausible candidates for the causal gene of the flowering time, were identified (Additional File 10). Information of seven enriched gene ontology groups for DTF candidate genes is shown in Figure 3E. The SNP Bn-scaff_18507_1-p354053 (A/G) located 12.7 kbps downstream of *LIPOXYGENASE 4* (*LOX4*) plays important roles in flower development and male fertility regulation (Klepikova et al., 2016). The Chr12:24697490 (A/G) and Chr12:24697925 (A/G) SNPs in a LD block were located within the position of candidate flowering gene *GDSL ESTERASE/LIPASE* (Figures 3F, G). Our results showed that the varieties with the alleles GG in SNP Chr12:24697490 showed significantly late flowering compared with those with the alleles AG (Figure 3H). The 1-kbps LD block surrounding Chr12:24986073 (A/C) contains the candidate flowering gene *FAR1-RELATED SEQUENCE (FRS)*. The 3.2-kbps LD block surrounding Chr12:28382413 (A/G) contains the candidate flowering gene *9-CIS-EPOXTCAROTENOID DIOXYGENASE* (*NCED9*), which is a critical gene in the regulation of abscisic acid (ABA) synthesis. The 60.8-kbps LD block surrounding the Chr12:29519291 (A/G) SNP contains *CINNAMOYL-COA REDUCTASE 2* (*CCR2*), which plays an important role in pollen development by regulating the programmed cell death (PCD) of tapetum cells (Zhang et al., 2023c). Chr12:30219143 (A/G), which explained 11.62% of the phenotypic variance located 3.6 kbps downstream of pollen-specific gene *RALFL14*. Another pollen-specific gene, *DEFENSIN-LIKE 7* (*DEFL7*), is located 36.71 kbps downstream of Chr12:30219143 ($P = 6.58 \times 10^{-4}$). *DEFL* genes are involved in pollen tube guidance and pollen tube reception and are responsible for the failure of double fertilization events (Takeuchi and Higashiyama, 2012).

3.5.1.2 Days to silique development

GWAS identified 30 TASs on chromosomes A01, A02, A03, and A10 for DTSD (Additional Files 5-8). Chr1:18351464 and Chr1:5217408 had the strongest signals for DTSD in the WW condition in 2017, which explained 43.50% and 37.75% of the phenotypic variance, respectively (Additional File 5). Three SNPs under drought conditions, Chr3:25485861 (A/G), Chr3:25524140 (A/G), and Chr3:25525060 (A/C), identified on chromosome A03 were located ~20 kbps downstream and upstream of the *AGL19*

10.3389/fpls.2024.1342359

TABLE 2 Repetitive significant SNP in the current study.

				2017		2018		2020	
Traits	SNP name	Chr	Position	WW	DS	WW	DS	WW	DS
DTF	Bn-A04-p1865434	A04	1588506	ns	4.67E-04	ns	9.61E-04	ns	ns
	Bn-A04-p2568394	A04	2279208	ns	3.62E-04	ns	8.59E-05	ns	ns
	Bn-A10-p15330596	A10	16164284	ns	4.46E-04	ns	9.33E-04	ns	ns
	Bn-A10-p15361519	A10	16139771	ns	2.15E-05	ns	8.88E-04	ns	ns
	Bn-A10-p15405149	A10	16091976	ns	1.37E-04	ns	4.87E-04	ns	ns
	Bn-A10-p15542820	A10	222927	1.77E-04	2.80E-04	ns	9.52E-04	ns	ns
	Bn-A10-p7252424	A10	8893827	ns	4.81E-04	ns	6.67E-05	ns	ns
	Bn-A10-p7347530	A10	8998128	ns	5.02E-04	ns	1.94E-04	ns	ns
	Bn-scaff_22728_1-p744551	C03	5904544	ns	9.02E-04	ns	2.97E-04	ns	ns
THSW	Bn-scaff_15818_1-p427676	C06	15724425	ns	ns	ns	4.40E-04	1.60E-04	2.40E-04
	Bn-scaff_15818_1-p453625	C06	15751833	ns	ns	ns	4.40E-04	1.60E-04	2.40E-04
	Bn-scaff_15818_1-p469375	C06	15766868	ns	ns	ns	4.40E-04	1.60E-04	2.40E-04
	Bn-scaff_15818_1-p471106	C06	15768599	ns	ns	ns	4.40E-04	1.60E-04	2.40E-04
	Bn-scaff_18702_1-p263991	C02	16251517	ns	ns	ns	5.53E-04	7.18E-04	9.53E-04
	Bn-scaff_18702_1-p270197	C02	16260356	ns	ns	ns	5.53E-04	7.18E-04	9.53E-04
	Bn-scaff_18702_1-p288731	C02	16279001	ns	ns	ns	5.53E-04	7.18E-04	9.53E-04
	Bn-scaff_18702_1-p323368	C02	16314729	ns	ns	ns	5.53E-04	7.18E-04	9.53E-04
	Bn-scaff_18702_1-p361232	C02	16355013	ns	ns	ns	5.53E-04	7.18E-04	9.53E-04
	Bn-scaff_18702_1-p365839	C02	16359619	ns	ns	ns	5.53E-04	7.18E-04	9.53E-04

DTF and THSW are the abbreviations of days to flowering and thousand seed weight. WW and DS are the codes of the two watering regimes; well-watered and drought stress during 3 years: 2017, 2017, and 2020.

TABLE 3 Details of 49 pleiotropic SNPs of agronomic and yield-related traits detected from genome-wide association study (GWAS).

SNP name	Chr	Allele			Number			Phenotype			Traits	Near locus previously reported in the same chromosome
Bn-A02-p8191099	A02	AA	AC	CC	15	20	83	186.0	182.3	183.0	DTF	New
								193.8	189.4	190.7	DTSD	
Bn-A02-p8284992	A02	AA	AG	GG	15	18	84	186.0	181.5	183.2	DTF	New
								193.8	189.1	190.8	DTSD	
Bn-A02-p8323616	A02	AA	AC	CC	16	19	84	185.8	181.8	183.1	DTF	New
								193.5	189.0	190.8	DTSD	
Bn-A02-p8440451	A02	AA	AG	GG	24	31	54	185.1	181.3	183.5	DTF	New
								192.7	188.9	191.1	DTSD	
Bn-A02-p8660632	A02	AA	AG	GG	18	25	76	185.4	182.1	184.4	DTF	*BnaA02g12130D*, *BnaA02g12260D* (Zheng et al., 2017)
								186.8	189.5	191.6	DTSD	
Bn-A02-p8934537	A02	AA	AG	GG	18	20	81	185.2	182.5	183.0	DTF	Bn-A02-p3539297 (Xu et al., 2015)
								193.1	190.0	190.6	DTSD	

(Continued)

TABLE 3 Continued

SNP name	Chr	Allele			Number			Phenotype			Traits	Near locus previously reported in the same chromosome
Bn-A02-p8999771	A02	AA	AC	CC	19	21	79	185.1	182.4	183.0	DTF	Bn-A02-p3539297 (Xu et al., 2015)
								192.9	189.8	190.6	DTSD	
Bn-A02-p9000921	A02	AA	AG	GG	79	21	19	183.0	182.4	185.1	DTF	Bn-A02-p3539297 (Xu et al., 2015)
								190.6	189.8	192.9	DTSD	
Bn-A02-p8190375	A02	AA	AG	GG	15	20	83	168.0	182.3	183.0	DTF	New
								193.8	189.4	190.7	DTSD	
Bn-A10-p13390065	A10	AA	AG	GG	6	14	98	179.5	178.96	184.0	DTF	*BnaA10g18420D*, *BnaA10g18480D*, *BnaA10g22080D*, *BnaA10g24300D* (Helal et al., 2021)
								187.3	187.26	191.5	DTSD	
Bn-A10-p15668415	A10	AA	AG	GG	12	17	89	179.8	180.5	184.1	DTF	New
								187.6	188.7	191.7	DTSD	
Bn-A01-p17377721	A01	AA	AC	CC	4	25	88	187.0	190.7	191.2	DTSD	New
								121.0	116.2	118.2	PH	
Bn-A01-p27079797	A01	AA	AG	GG	33	83	2	192.0	190.4	191.9	DTSD	New
								117.8	118.1	121.3	PH	
Bn-A01-p21758046	A01	AA	AG	GG	51	42	25	191.2	190.4	191.0	DTSD	*BnaA01g26410D-BnaA01g26530D* (Lu et al., 2017)
								118.3	117.6	117.7	PH	
								9.5	8.7	9.1	HI	
Bn-A01-p5715141	A01	AA	AG	GG	27	28	62	191.9	190.7	190.5	DTSD	*Bna.QRT3* (*BnaA01g10390D*) (Lu et al., 2017)
								118.4	117.5	117.7	PH	
								9.3	8.8	9.1	HI	
Bn-A01-p6678914	A01	AA	AG	GG	93	16	6	266.8	266.5	266.3	DTR	Bn-A01-p7430311 (Sun et al., 2016b)
								117.8	116.8	118.2	PH	
Bn-A01-p9203096	A01	AA	AG	GG	8	19	90	265.6	266.1	266.8	DTR	New
								118.1	117.8	117.9	PH	
Bn-scaff_18936_1-p102755	C03	AA	AG	GG	75	27	13	266.8	266.5	266.4	DTR	New
								118.2	116.6	117.1	PH	
Bn-scaff_18936_1-p240670	C03	AA	AG	GG	16	29	72	266.4	266.4	266.8	DTR	New
								116.9	117.4	118.3	PH	
Bn-scaff_18936_1-p269153	C03	AA	AC	CC	95	19	4	266.7	266.4	267.7	DTR	New
								118.3	116.6	113.7	PH	
Bn-scaff_18936_1-p472353	C03	AA	AG	GG	17	36	65	266.9	266.4	266.7	DTR	New
								116.8	118.2	118.0	PH	
Bn-scaff_18936_1-p93643	C03	AA	AG	GG	13	27	76	266.4	266.5	266.8	DTR	New
								117.1	117.0	118.2	PH	
Bn-scaff_18936_1-p97644	C03	AA	AC	CC	76	28	13	266.8	266.5	266.4	DTR	New

(Continued)

TABLE 3 Continued

SNP name	Chr	Allele			Number			Phenotype			Traits	Near locus previously reported in the same chromosome
								118.2	117.0	117.1	PH	
Bn-scaff_18936_1-p274133	C03	AA	AG	GG	5	17	95	267.3	266.4	266.7	DTR	New
								114.4	116.5	118.3	PH	
Bn-A01-p7619726	A01	AA	AG	GG	102	12	5	265.6	265.7	266.8	DTR	New
								117.8	117.0	122.0	PH	
								9.1	9.0	9.8	HI	
Bn-A01-p8014995	A01	AA	AG	GG	4	9	104	266.9	265.6	266.7	DTR	BnvaA0107152286 (Han et al., 2022)
								118.7	115.6	118.0	PH	
								9.1	8.7	9.2	HI	
Bn-scaff_18936_1-p439378	C03	AA	AG	GG	5	17	96	267.3	265.8	266.8	DTR	New
								114.4	117.1	118.2	PH	
								1,762.8	1,875.1	1,905.8	PW	
Bn-scaff_18936_1-p440619	C03	AA	AG	GG	5	17	96	266.8	265.8	267.3	DTR	New
								114.4	117.1	118.2	PH	
								1,762.8	1,875.1	1,905.8	PW	
Bn-scaff_18936_1-p559490	C03	AA	AC	CC	3	20	95	267.4	265.9	266.8	DTR	New
								112.8	116.3	118.4	PH	
								1,539.0	1,878.7	1908.4	PW	
Bn-scaff_18936_1-p610540	C03	AA	AG	GG	94	20	4	267.8	266.1	266.7	DTR	New
								118.4	116.2	114.4	PH	
								1,911.3	1,867.0	1662.0	PW	
Bn-scaff_18936_1-p611810	C03	AA	AC	CC	94	20	4	267.8	266.1	266.7	DTR	New
								118.4	116.2	114.4	PH	
								1,911.3	1,867.0	1662.0	PW	
Bn-scaff_18936_1-p618378	C03	AA	AG	GG	4	20	94	267.8	266.1	266.7	DTR	*BnaC03g45540D* (Zhang et al., 2023a)
								114.4	116.2	118.4	PH	
								1,662.0	1,867.0	1911.3	PW	
Bn-scaff_18936_1-p622547	C03	AA	AC	CC	94	20	4	266.7	266.1	267.8	DTR	New
								118.4	116.2	114.4	PH	
								1,911.3	1,867.0	1662.0	PW	
Bn-scaff_18936_1-p643990	C03	AA	AG	GG	94	20	4	266.7	266.1	267.8	DTR	New
								118.4	116.2	114.4	PH	
								1,911.3	1,867.0	1662.0	PW	
Bn-A08-p12555227	A08	AA	AG	GG	11	20	86	1,936.4	1,903.4	1882.5	PW	New
								197.6	177.1	174.9	SWPP	
								9.9	9.0	9.0	HI	
								2,507.1	2,172.5	2196.8	SY	

(Continued)

TABLE 3 Continued

SNP name	Chr	Allele			Number			Phenotype			Traits	Near locus previously reported in the same chromosome
Bn-A08-p15782077	A08	AA	AG	GG	70	29	20	1,822.2	1954.1	2041.1	PW	New
								168.1	182.7	204.7	SWPP	
								9.1	9.0	9.5	HI	
								2,127.5	2,255.8	2520.0	SY	
Bn-A08-p15782229	A08	AA	AG	GG	70	28	20	1,822.2	1,947.9	2041.1	PW	New
								168.1	182.1	204.7	SWPP	
								9.1	9.0	9.5	HI	
								2,127.5	2,234.1	2520.0	SY	
Bn-A08-p12556455	A08	AA	AG	GG	88	10	11	1,887.5	1,832.9	2098.3	PW	New
								175.0	179.7	264.3	SWPP	
								2,201.0	2,151.8	3177.8	SY	
Bn-A08-p13626189	A08	AA	AG	GG	72	31	15	1,890.8	1,870.0	1932.4	PW	New
								175.3	175.7	189.7	SWPP	
								2,191.0	2,198.2	2368.3	SY	
Bn-A08-p13626982	A08	AA	AG	GG	16	32	70	1,906.0	1,895.2	1885.0	PW	New
								188.7	175.6	175.4	SWPP	
								2,339.1	2,190.0	2198.8	SY	
Bn-A08-p13638847	A08	AA	AC	CC	69	33	16	1,889.1	1,886.3	1906.0	PW	New
								175.8	174.8	188.7	SWPP	
								2,201.5	2,184.7	2339.1	SY	
Bn-A08-p13670107	A08	AA	AG	GG	15	28	72	1,932.4	1,880.5	1890.8	PW	New
								189.7	176.0	175.3	SWPP	
								2,368.3	2,217.4	2191.0	SY	
Bn-A08-p14538807	A08	AA	AG	GG	87	19	8	1,873.6	1,955.0	1930.9	PW	New
								171.6	193.4	185.4	SWPP	
								2,163.2	2,330.7	2371.0	SY	
Bn-A08-p15994149	A08	AA	AC	CC	29	25	63	1,988.2	1,901.6	1839.1	PW	*Bna.BBX20* (*BnaA08g16780D*, *AT4G39070*) (Lu et al., 2017) *Bna.BBX15* (*BnaA08g19420D*) (Lu et al., 2017)
								198.4	178.7	167.3	SWPP	
								2,432.6	2,196.9	2132.8	SY	
Bn-scaff_19208_1-p78898	C04	AA	AG	GG	66	32	20	1,883.9	1,828.8	2002.8	PW	New
								178.4	161.4	200.9	SWPP	
								2,196.2	2,098.2	2489.5	SY	
Bn-scaff_19208_1-p82535	C04	AA	AG	GG	66	31	20	1,883.9	1,827.1	2002.8	PW	New
								178.4	161.5	200.9	SWPP	
								2,196.2	2,103.9	2489.5	SY	
Bn-scaff_19208_1-p93814	C04	AA	AC	CC	66	31	21	1,883.9	1,827.1	2002.8	PW	New
								178.4	161.5	200.9	SWPP	
								2,196.2	2,103.9	2489.5	SY	

(Continued)

TABLE 3 Continued

SNP name	Chr	Allele			Number			Phenotype			Traits	Near locus previously reported in the same chromosome
Bn-scaff_19208_1-p94498	C04	AA	AC	CC	66	32	21	1,883.9	1,828.8	2002.8	PW	New
								178.4	161.4	200.9	SWPP	
								2,196.2	2,098.2	2489.5	SY	
Bn-scaff_19208_1-p94501	C04	AA	AG	GG	20	31	66	2,002.8	1,827.1	1883.9	PW	New
								200.9	161.5	178.4	SWPP	
								2,489.5	2,103.9	2196.2	SY	

DTF, DTSD, DTR, PH, PW, SWPP, HI, and SY are the abbreviations of days to flowering, days to silique development, days to ripening, plant height, plant weight, seed weight/plant, harvest index, and seed yield, respectively.

gene, respectively. This gene is involved in seed formation, silique maturity, and seed desiccation (Shah et al., 2022).

3.5.1.3 Days to ripening

A total of 97 significant associations were identified for DTR (Additional Files 5-8). The 55 and 42 TASs identified in WW and DS conditions explained 63.20% and 36.79% of the phenotypic variance. There were 77 CGs associated with 21 significant SNPs in the C03:2498421–3217123 intervals (Additional Files 5, 10; Figures 3I–L). Chr13:2852351 ($P = 2.11 \times 10^{-5}$) and Chr13:2878346 ($P = 1.36 \times 10^{-8}$) were located ~10 kbps downstream and upstream of aspartic proteinase oryzasin-1, respectively. In *B. napus*, *BnaAP36s* and *BnaAP39s* genes play a critical role in pollen tube growth (Wang et al., 2023). Chr13:3058428 ($P = 1.36 \times 10^{-8}$) explained 42.99% of the phenotypic variance located within the position of the panicle architecture-related gene *LAX PANICLE 2* (*LAX2*) (Figures 3M, N). The varieties with the AA and GG alleles in this SNP showed significantly late maturity compared with those with the AG alleles (Figure 3O). Another specific gene, *FY*, located 19.96 kbps upstream of the Chr13:3140112 ($P = 1.07 \times 10^{-8}$) SNP, plays a role in the regulation of flowering time in the autonomous flowering pathway through repression of the FLOWERING LOCUS C (FLC) (Kyung et al., 2022). Chr13:3182947 which explained 43.65% of the phenotypic variance was located <1 kbps upstream of two genes from MADS-box *AGAMOUS* (*AG*) genes: *AGL15* and *AGL16*. MADS-box genes play an important role in regulating floral carpel and ovule development (Sheng et al., 2019; Shah et al., 2022).

3.5.1.4 Plant height

A total of 70 significantly associated SNPs were identified for PH, of which 59 and 11 were identified in the WW and DS conditions, respectively (Additional Files 5-8). Among the linked SNPs, 16 SNPs identified in the WW conditions were located on chromosome C03 (2736068–3217123 bp), which explained 31.69% of the phenotypic variance (Additional File 6; Figures 3P–S). The 14.28-kbps LD block surrounding Chr13:3023049 ($P = 8.96 \times 10^{-6}$) contains the *TIFY9* gene (Figure 3R). The *TIFY* family is a plant-specific gene involved in accelerated cell division, leaf flatness, and lateral organ development (Zhang et al., 2020). Chr13:3058428 (A/G) which explained 27.34% of the phenotypic variance of PH was located 12.78 kbps downstream of gene *GH3.12* responsible for plant stem growth. Three SNPs, Chr13:3140112 (A/G), and Chr13:3182947 (A/C), and Chr13:3184218 (A/G), were located ~20 kbps downstream and upstream of *PAO1*.

3.5.2 SNPs and candidate genes linked with yield and yield-related traits under drought conditions

3.5.2.1 Branch number/plant

Of 250 SNPs detected for BNPP, 28 and 222 SNPs explained 10.58% and 89.14% of the BNPP variance in the WW and DS conditions, respectively (Additional Files 5-8). There were 22 of the 222 SNPs identified under DS condition located in the C05:11200372–1417644-bp region, which had not been reported previously (Additional Files 5, 10; Figures 3T–W). Five SNPs, namely, Chr15:125848, Chr15:135624, Chr15:135780, Chr15:136834, and Chr15:136975, within an LD block were located <5 kbps downstream and upstream of the *SBT1.1* gene (Figure 3V), a gene which contributed to elongation of the main shoot, increasing inflorescence branching and biomass (Martinez et al., 2015). Two SNPs, Chr15: 136834 ($P = 1.31 \times 10^{-4}$) and Chr15: 136975 ($P = 3.63 \times 10^{-5}$), in an LD block were mapped to 11 kbps upstream of a member of *SKP1-Like* gene family, *ASK3*.

3.5.2.2 Plant weight

Of the 56 significant SNPs for PW, 34 and 22 explained 58.10% and 41.89% of PW variance in the WW and DS conditions, respectively (Additional Files 5-8). There were 12 TASs of the WW condition located on the 35,474,299–35,698,370 bp interval in chromosome C04 (Additional File 7; Figures 4A–D). This genomic region contained four genes of ABC transporter G superfamily (*ABCG16*, *ABCG17*, *ABCG18*, and *ABCG19*) which contribute to cytokinin transport in the shoot and enhance the tiller number, grain number per panicle, and grain yield (Wu et al., 2022a). Four SNPs, Chr14:35642269 (A/G), Chr14:35642321 (A/C), Chr14:35642324 (A/C), and Chr14:35643023 (A/G), in an LD block, were located within the *ABCG16* sequence (Figures 4C, E, F). The varieties with the alleles AA and CC in these SNPs showed significantly higher plant weight compared with varieties with other alleles (Figure 4G). We identified four haplotypes/markers

10.3389/fpls.2024.1342359

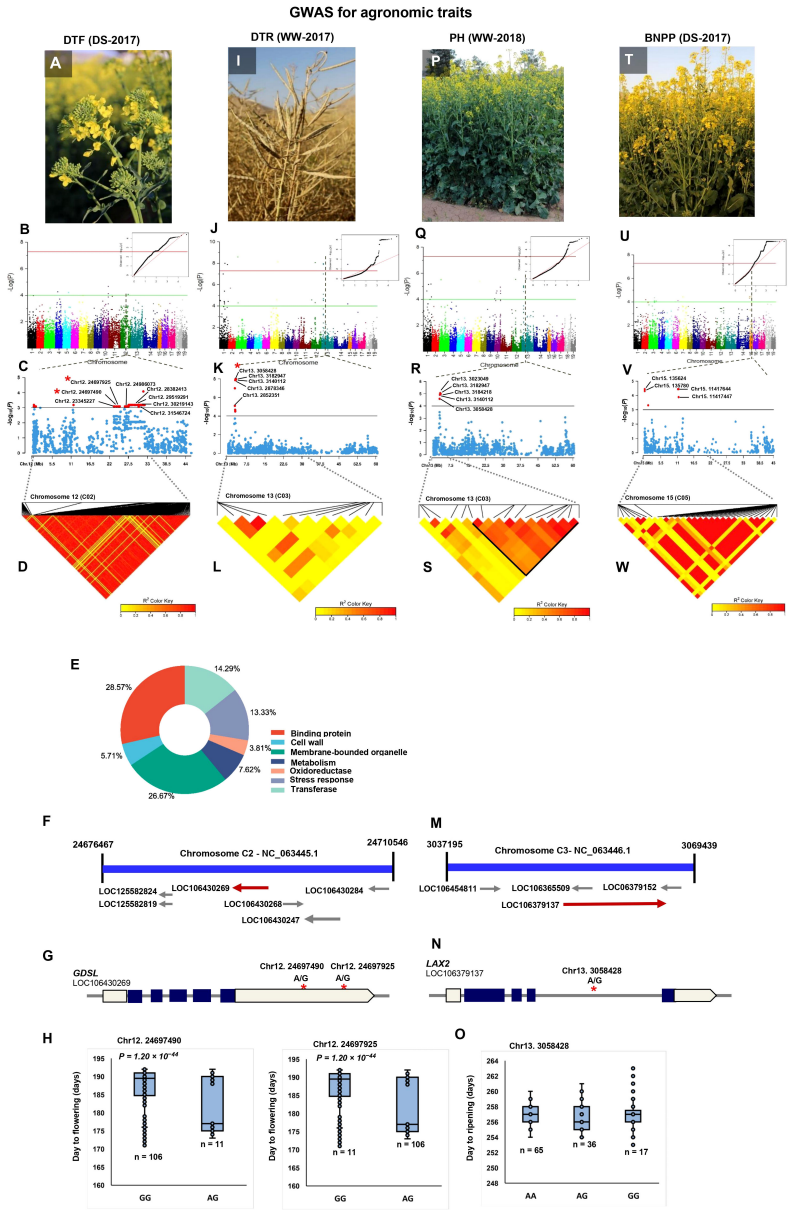


FIGURE 3
Genome-wide association study (GWAS) reveals the genetic basis of agronomic traits. **(A)** GWAS of day to flowering (DTF) using the 119 rapeseed (*Brassica napus*) varieties under drought stress conditions in 2017. **(B)** Manhattan plot and corresponding quantile–quantile (QQ) plot displaying the GWAS result of DTF in 19 chromosomes (1–10 stand for rapeseed chromosome of A01–A10, and 11–19 stand for rapeseed chromosome of C01–C09 at the horizontal axis). SNPs on different chromosomes are denoted by different colors. **(C)** Locus zoom plot for DTF associations in chromosome 12 (C02). **(D)** A representation of pairwise r^2 value (displayed as percentages) among polymorphic sites of chromosome 12 (C02) for DTF. **(E)** CirGO visualization of GO enrichment analysis of significant genes identified by GWAS for DTF. **(F)** Schematic diagram of the region of chromosome 12 (C02) genotyped in this study, showing the associated gene *GDSL* (LOC106430269). Chromosomal position based on the National Center for Biotechnology Information (NCBI). **(G)** Gene model of *GDSL*. Solid boxes indicate exons, open boxes indicate untranslated regions (UTRs), and lines connecting the exons indicate introns. The red stars mark the position of Chr12:24697490 and Chr12:24697925. **(H)** The influence of Chr12:24697490 and Chr12:24697925 on day to flowering. The significance of difference between two varieties was evaluated using Student's t-test. **(I)** GWAS of day to ripening (DTR) under well-watered conditions in 2017. **(J)** Manhattan plot and corresponding quantile–quantile (QQ) plot displaying the GWAS result of DTR in 19 chromosomes. **(K)** Locus zoom plot for DTR associations in the chromosome 13 (C03). **(L)** A representation of pairwise r^2 value (displayed as percentages) among polymorphic sites of chromosome 13 (C03) for DTR. **(M)** Schematic diagram of the region of chromosome 13 (C03) genotyped in this study, showing the associated gene *LAX2* (LOC106379137). Chromosomal position based on NCBI. **(N)** Gene model of *LAX2*. Solid boxes indicate exons, open boxes indicate untranslated regions (UTRs), and lines connecting the exons indicate introns. The red star marks the position of Chr13:3058428. **(O)** The influence of Chr13:3058428 on day to ripening. **(P)** GWAS of plant height (PH) under well-watered conditions in 2018. **(Q)** Manhattan plot and corresponding quantile–quantile (QQ) plot displaying the GWAS result of PH in 19 chromosomes. **(R)** Locus zoom plot for PH associations in the chromosome 13 (C03). **(S)** A representation of pairwise r^2 value (displayed as percentages) among polymorphic sites of chromosome 13 (C03) for PH. **(T)** GWAS of branch number/plant (BNPP) under drought stress conditions in 2017. **(U)** Manhattan plot and corresponding quantile–quantile (QQ) plot displaying the GWAS result of BNPP in 19 chromosomes. **(V)** Locus zoom plot for BNPP associations in the chromosome 15 (C05). **(W)** A representation of pairwise r^2 value (displayed as percentages) among polymorphic sites of chromosome 15 (C05) for BNPP.

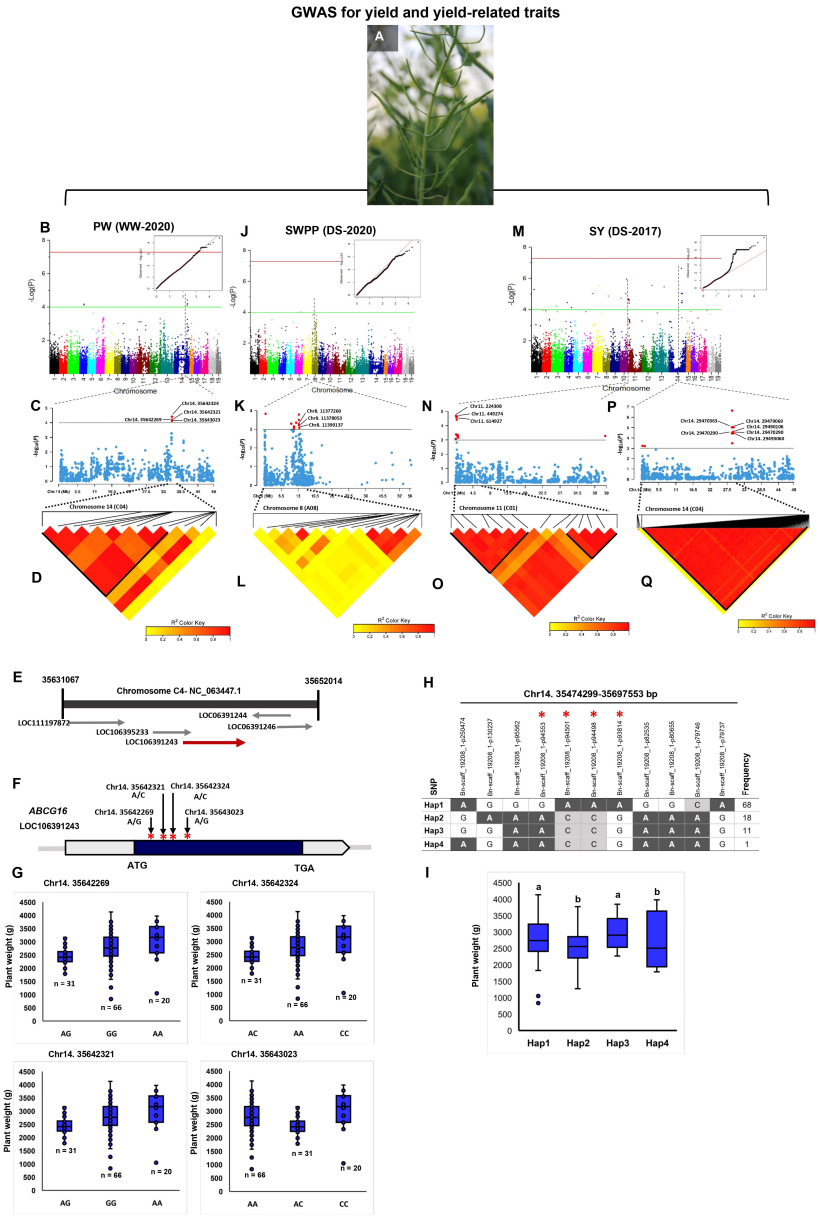


FIGURE 4
Genome-wide association study (GWAS) reveals the genetic basis of seed yield and yield-related traits in rapeseed (*Brassica napus*). **(A)** Phenotypes of siliques in rapeseed which are considered the major contributing factors for increasing rapeseed yield. **(B–D)** GWAS of plant weight (PW) using the 119 rapeseed varieties under well-watered conditions in 2020. **(B)** Manhattan plot and corresponding quantile–quantile (QQ) plot displaying the GWAS result of PW in 19 chromosomes (1–10 stand for rapeseed chromosomes of A01–A10, and 11–19 stand for rapeseed chromosomes of C01–C09 at the horizontal axis). SNPs on different chromosomes are denoted by different colors. **(C)** Locus zoom plot for PW associations in the chromosome 14 (C04). **(D)** A representation of pairwise r^2 value (displayed as percentages) among polymorphic sites of chromosome 14 (C04) for PW. **(E)** Schematic diagram of the region of chromosome 14 (C04) genotyped in this study, showing the associated gene *ABCG16* (LOC106391243). Chromosomal position based on the National Center for Biotechnology Information (NCBI). **(F)** Gene model of *ABCG16*. Solid boxes indicate exons, and open boxes indicate untranslated regions (UTRs). The red stars mark the positions of Chr14:35642269, Chr14:35642321, Chr14:35642324, and Chr14:35643023. **(G)** The influence of Chr14:35642269, Chr14:35642321, Chr14:35642324, and Chr14:35643023 on plant weight. **(H)** Haplotype block based on 11 significant SNPs on chromosome 14 (C04). **(I)** Four different haplotype variants (Hap1–Hap4) are present at different frequencies in the analyzed population. Boxplots for plant weight indicate the phenotype values corresponding to the four different haplotype groups. Significant differences among haplotypes were identified using one-way ANOVA. Different letters indicate distinct groups. **(J–L)** GWAS of seed weight/plant (SWPP) under drought stress conditions in 2020. **(J)** Manhattan plot and corresponding quantile–quantile (QQ) plot displaying the GWAS result of SWPP in 19 chromosomes. **(K)** Locus zoom plot for SWPP associations in chromosome 8 (A08). **(L)** A representation of the pairwise r^2 value (displayed as percentages) among polymorphic sites of chromosome 8 (A08) for SWPP. **(M–Q)** GWAS of seed yield (SY) under drought stress conditions in 2017. (**M**) Manhattan plot and corresponding quantile–quantile (QQ) plot displaying the GWAS result of SY in 19 chromosomes. **(N)** Locus zoom plot for SY associations in chromosome 11 (C01). **(O)** A representation of pairwise r^2 value (displayed as percentages) among polymorphic sites of chromosome 11 (C01) for SY. **(P)** Locus zoom plot for SY associations in the chromosome 14 (C04). **(Q)** A representation of pairwise r^2 value (displayed as percentages) among polymorphic sites of chromosome 14 (C04) for SY.

associated with PW located on chromosome C04 (35,474,299–35,697,553 bp) (Figure 4H) with an average plant weight of 2984.37 g in Hap3 significantly greater than in other three Haps (Figure 4I).

3.5.2.3 Seed number/silique

There were 10 drought-related TASs identified on chromosome A02 (750,194–791,056 bp) that had not been reported in drought stress in rapeseed, previously (Additional File 7). The Chr2:751843 (A/G), Chr2:751923 (A/G), and Chr2:752016 (A/G) SNPs located within a MADS-box gene, *FLOWERING LOCUS C* (*FLC*), involved in flowering time control, inflorescence architecture, floral organ identity determination, and seed development (Soppe et al., 2021). Chr2:989812 (A/C) located 9.10 kbps downstream of seed plant-specific *BIG GRAIN LIKE* gene family (*BG1-D*) regulates grain number per plant, grain size with both bigger length and width, and finally grain yield (Liu et al., 2015; Lo et al., 2020; Gao et al., 2022). Chr7:36944365 (A/C) located within *CAND1* encodes cullin-associated Nedd8-dissociated protein 1 (Additional File 7).

3.5.2.4 Seed weight/plant

Of the significant TASs identified in drought stress, 10 were located on 10,335,313 bp–13,452,346 bp of chromosome A08 (Additional Files 7, 8). Chr8:11377260 and Chr8:11378053 ($P = 7 \times 10^{-4}$) were located within the position of a member of the purine permease (PUP)-type transporter gene family, *PUP21*. In addition, Chr8:11390137 (A/G) was located within other members of the PUP-like family gene, *PUP10* (Figures 4J–L).

3.5.2.5 Thousand seed weight

Among the THSW-associated SNPs, 10 were stable overs 2 years under drought stress conditions (Table 2). There were 13 TASs on chromosome C02 (16,251,517 bp–16,902,605 bp) and eight on chromosome C06 (15,724,425 bp–15,768,599 bp) that explained 26.08% and 19.7% of the THSW variance in the WW and DS treatments, respectively (Additional File 7).

3.5.2.6 Seed yield

Of the SY SNPs, 34 were positioned in the 130,644-bp–676,816-bp interval on chromosome C01 (Figures 4M–O). Chr11:224308 ($P = 2.58 \times 10^{-5}$; Figure 4N), which explained 24.34% of the phenotypic variance located within the position of the *DRG3* gene belonging to the G-protein family. Chr11:449274 ($P = 4.56 \times 10^{-4}$; Figure 4N), which explained 17.69% of the SY variance located 7.78 kbps downstream of *CYP79B1*. Chr11:614927 ($P = 4.68 \times 10^{-4}$) was located within the sequence of the *AGAMOUS LIKE21* (*AGL21*) gene, which has been shown to upregulate in siliques and dry seeds in rapeseed (Yu et al., 2017). There were 112 drought-related SY–SNPs located in the 29,030,148-bp–29,607,326-bp interval on chromosome C04 and explained 67.56% of the phenotypic variance of SY (Additional File 5; Figures 4P, Q). Six SNPs on chromosome C04 with complete LD ($r^2 = 1$, Figure 4Q), namely, Chr14:29470290, Chr14:29470363, Chr14:29479069, Chr14:29490106, Chr14:29492969, and Chr14:29493060, located <13 kbps downstream and upstream of *CYP78A9*, which is another member of the cytochrome P450 superfamily.

3.5.2.7 Harvest index

The 38 SNPs significantly associated with HI contributed to 38.34% and 61.65% of HI variance in the WW and DS conditions, respectively (Additional Files 5-8). Among the HI SNPs, seven SNPs were located on chromosome C06 (24,703,167 bp–34,271,977 bp) and five on chromosome C08 (28,759,659 bp–30,043,480 bp) in the WW condition in 2017 (Additional File 5).

3.6 Identification of novel pleiotropic SNPs associated with more than one trait

In the current study, 19 novel repetitive SNPs linked to both DTF and THSW traits were identified on chromosomes A04, A10, C02, C03, and C06. Analysis of SNPs for pairwise traits revealed that 49 SNPs associated with two or more than two traits and the differences in phenotypic values between varieties with two alleles at each of these SNPs were significant. Of the 49 pleiotropic SNPs, 11 SNPs were overlapped with those reported in previous studies only and the rest were unique in this study (Table 3). There were 11 SNPs on chromosome A02 and A10 that were associated with both DTF and DTSD simultaneously. A pleiotropic SNP Bn-A02-p8660632 for the phenotype of both DTF and DTSD mapped on chromosome A02 and was 2,186 kbps downstream of the position of the *BnaA02g12130D* and *BnaA02g12260D* genes. These genes affected DTF and PH in *B. napus* in the Zheng et al. (2017) study. Pleiotropic SNPs were identified for DTSD and PH; DTSD, PH, and HI; DTR and PH; DTR, PH, and HI; DTR, PH, and PW; PW, SWPP, HI, and SY; and PW, SWPP, and SY (Table 3). The position of the pleiotropic SNP Bn-A08-p15994149 for PW, SWPP, and SY on chromosome A08 was 3,734 kbps downstream of the candidate gene *BnaA08g16780D* and was 69 kbps downstream of the region (13,520,923 bp–13,598,303 bp) that affected branch pod number and pod number per plant in the Lu et al. (2016) study.

3.7 Estimating the effect of major pleiotropic SNPs on traits

As shown in Table 3, varieties with allele AA in the Bn-A01-p21758046 and Bn-A01-p5715141 SNPs showed higher DTSD, PH, and HI. Varieties with the allele GG in Bn-A01-p7619726 had higher DTR, PH, and HI. The allele GG in Bn-scaff_18936_1-p439378 and Bn-scaff_18936_1-p440619 presented higher DTR, PH, and PW. The allele AA in Bn-scaff_18936_1-p610540 and Bn-scaff_18936_1-p611810 increased DTR, PH, and PW. Varieties with the allele AA in Bn-A08-p12555227 showed higher PW, SWPP, HI, and SY, whereas varieties with the allele GG in Bn-A08-p15782077 and Bn-A08-p15782229 had higher PW, SWPP, HI, and SY.

3.8 Comparative transcriptome analysis between seed and leaf of low and high grain yield varieties

The transcriptome analysis of mature seeds in the rapeseed varieties differing in their seed yield can provide crucial systems-level insights into molecular mechanisms underlying seed development and seed yield. We selected four rapeseed varieties, namely, G111 and G114 as low-seed yield and G19 and G41 as high-seed yield varieties, to investigate the transcriptional differences in the two contrasting groups. In total, 2,906 DEGs (1,441 up- and 1,465 downregulated) of both tissues (seed and leaf) in low-seed yield and 7,243 (3,519 up- and 3,724 downregulated) of both tissues in high-seed yield varieties with $|\log2FC| \geq 1$ and padj < 0.05 were identified (Figures 5A, B). Of the DEGs, 994 upregulated and 1,008 downregulated genes shared between the two contrasting groups. The Gene Ontology (GO) enrichment analyses of all the genes showed that most of the genes were related to various developmental process, reproductive processes, cell wall organization, cell cycle and cell division, metabolic processes, response to stress/hormone, and regulation of transcription (Figure 5C). These processes are well known to be involved during various aspects of seed development. At least, 321 transcription factor (TF)-encoding genes belonging to 66 families exhibited stage-specific expression in one or more than one cultivar. The members of MYB, bHLH, ERF, WRKY, bZIP, and ARF families were highly represented in these varieties (Figure 5D). The expression profiles of key gene families and individual genes involved in cell division, cell size determination, cell wall modification, carbohydrate metabolism, and grain filling were analyzed. We observed a higher expression of several members of these gene families in high-yield varieties (Figure 5E). A higher transcriptional activity of cyclin-encoding genes was identified in high-yield varieties, which is almost related to higher mitotic activity and an extended period of cell division. The genes encoding glucan synthases and xyloglucan endotransglucosylases/hydrolases exhibited higher transcriptional activity in high-yield varieties. These enzymes are involved in the synthesis and remodeling of cell wall components and production of energy (Miedes et al., 2013; Perrot et al., 2022; Zhang et al., 2022). Furthermore, the transcript abundance of genes involved in cell expansion (expansins), seed storage proteins (e.g., vicilin-like storage protein), and lipid transfer proteins was also significantly higher in the high-seed yield rapeseed varieties (Figure 5E). It has been shown that these proteins contribute to various aspects of seed development and seed maturation (Pagnussat et al., 2012; Wang et al., 2015a; Yaqoob et al., 2020; Rahman et al., 2021; Fang et al., 2023).

3.9 Validation of candidate genes associated with SNPs by transcriptome analysis

Transcriptome sequencing was performed for further analysis of the identified CGs associated with the linked SNPs. We found that 215 CGs were significantly expressed under drought stress (Additional File 11; Figure 6A). The identified GCs were mainly associated with DTF (47.90% of the CGs), DTR (41.86%), PH (4.18%), and yield components (4.65%). The results of the KEGG pathway analysis are shown in Figure 6B. For the upregulated DEGs, 16 KEGG pathways were enriched according to *P*-value < 0.01 and FDR < 0.01. Our gene expression analysis identified 23 DTF-associated genes surrounding the SNP peaks in chromosome C02 of which 19 genes were upregulated under the DS condition (Figure 6B). One of these genes with 14.32-fold change between the two contrasting varieties, *GDSL ESTERASE/LIPASE* (LOC106430269) in C02, contained two significant DTF SNPs (Figure 6C). Six genes, *LOX4*, *FAR1*, *NCED9*, *CCR2*, *RALFL14*, and *DEFL7*, were located near our SNPs; Bn-scaff_18507_1-p354053, Bn-scaff_22749_1-p574003, Bn-scaff_16328_1-p636786, Bn-scaff_16485_1-p1575966, Bn-scaff_18245_1-p84866, and Bn-scaff_18406_1-p183669 were, respectively, upregulated in the DS condition and in the high-seed yield varieties (Table 4; Figure 6C). Among these genes, the *LOX4* gene showed 11.56-fold change between the two contrasting groups. In the KEGG analysis, we found that *GDSL ESTERASE/LIPASE* and *LOX4* are involved in the lipid metabolism pathways (Additional File 12; Figure 6B). Transcriptome analysis in both leaf and seed samples showed variable responses to drought stress. Higher DEGs and larger absolute changes were observed in expression of the genes in seed extracts than in leaf (Figures 6D–J).

We found that 20 genes surrounding the peak SNPs for DTR were significantly upregulated in the DS condition (Figure 7B). Expression of two genes, *ORYZASIN1* and *FY*, near the SNPs Bn-scaff_18936_1-p269153 and Bn-scaff_18936_1-p559490 for DTR, respectively, which was significantly higher in seed and drought than in leaf and WW conditions was significant between the two contrasting low- and high-yield varieties (Figures 7B, C, E). In the KEGG analysis, we found that *ORYZASIN1* and *FY* were from the lipid metabolism and biosynthesis of other secondary metabolites (Figure 7A; Additional File 13). The *LAX2* gene which belonged to transport and catabolism pathways was upregulated in drought treatment and contained a significant SNP in C03 for DTR (Figures 7B, D; Table 4; Additional File 13). Similar results were observed for gene expression level under drought stress for two members of the MADS domain family, *AGL15* and *AGL16*, that were near the significant SNP loci on C03 (Table 4). We observed that the expression of the *LAX2*, *AGL15*, and *AGL16* genes was significantly higher in seed than in leaf samples under drought (Figure 7F).

The *TIFY9*, *GH3.12*, and *PAO1* genes associated with PH SNPs on C03 showed significant differential expression between high- and low-yield varieties in the DS condition (Figures 8A, B, E–H). *GH3.12* had a higher expression in leaf than in seed (Figure 8D), whereas the expression of *TIFY9* was higher in seed (Figure 8C).

There were 10 DEGs (*BG1*, *FLC*, *DRG3*, *CAND1*, *PUP10*, *PUP21*, *ABCG16*, *AGL21*, *CYP79B1*, and *CYP78A9*) associated with the SY SNPs which showed significant differential expression between high- and low-yield varieties (Table 4; Figures 8I–K). These DEGs are known to regulate seed yield by affecting anther and pollen development, seed ripening, seed size, and seed weight regulation (Ma et al., 2015; Castelan-Munoz et al., 2019; Xu et al., 2019; Li et al., 2022). The *ABCG16* gene with 33.03-fold change in the seed sample showed a higher expression in the high-yield variety G19 under

TABLE 4 SNPs and candidate genes significantly associated with agronomic and yield-related traits integrating genome-wide association and transcriptome studies.

Traits	Lead SNP	Chr[a]	Position (bp)[b]	Allele	*P* value[c]	R^2 (%)[d]	Watering regime	Candidate Gene[e]	Annotation
DTF	Bn-scaff_18507_1-p354053	C02	23345227	A/G	8.15E-04	11.28	D17	LOC106381296	Lipoxygenase 4, chloroplastic (LOX4)
	Bn-scaff_18199_1-p285821	C02	24697490	A/G	8.15E-04	11.28	D17	LOC106430269	GDSL esterase/lipase At1g74460-like (GDSL)
	Bn-scaff_18199_1-p286255	C02	24697925	A/G	8.15E-04	11.28	D17		
	Bn-scaff_22749_1-p574003	C02	24986073	A/C	8.15E-04	11.28	D17	LOC125582273	Protein FAR1-RELATED SEQUENCE 3-like (FAR1)
	Bn-scaff_16328_1-p636786	C02	28382413	A/G	6.58E-04	11.62	D17	LOC106441091	9-cis-epoxycarotenoid dioxygenase NCED9, chloroplastic-like (NCED9)
	Bn-scaff_16485_1-p1575966	C02	29519291	A/G	6.58E-04	11.62	D17	LOC106381558	Cinnamoyl-CoA reductase 2-like (CCR2)
	Bn-scaff_18245_1-p84866	C02	30219143	A/G	6.58E-04	11.62	D17	LOC106387383	Protein RALF-like 14 (RALFL14)
	Bn-scaff_18406_1-p183669	C02	31546724	A/C	6.58E-04	11.62	D17	LOC125581930	Defensin-like protein 7 (DEFL7)
DTSD	Bn-A03-p27256355	A03	25485861	A/G	5.76E-04	16.57	D18	LOC111214287	Agamous-like MADS-box protein AGL19 (AGL19)
	Bn-A03-p27288521	A03	25524140	A/G	5.50E-04	16.67	D18		
	Bn-A03-p27289437	A03	25525060	A/C	7.95E-04	15.87	D18		
DTR	Bn-scaff_18936_1-p240670	C03	2852351	A/G	2.11E-05	24.55	W17	LOC125584437	Aspartic proteinase oryzasin-1-like (ORYZASIN1)
	Bn-scaff_18936_1-p269153	C03	2878346	A/G	1.36E-08	23.68	W17	LOC125584078	Aspartic proteinase oryzasin-1-like (ORYZASIN1)
	Bn-scaff_18936_1-p472353	C03	3058428	A/G	1.36E-08	42.99	W17	LOC106379137	Protein LAX PANICLE 2 (LAX2)
	Bn-scaff_18936_1-p559490	C03	3140112	A/G	1.07E-08	43.65	W17	LOC111204276	Flowering time control protein FY (FY)
	Bn-scaff_18936_1-p610540	C03	3182947	A/C	1.07E-08	43.65	W17	LOC106431084	Agamous-like MADS-box protein AGL15 (AGL15)
	Bn-scaff_18936_1-p610540	C03	3182947	A/C	1.07E-08	43.65	W17	LOC106431084	Agamous-like MADS-box protein AGL16 (AGL16)
PH	Bn-scaff_18936_1-p440619	C03	3023049	A/G	8.96E-06	27.75	W18	LOC106454811	Protein TIFY 9 (TIFY9)
	Bn-scaff_18936_1-p472353	C03	3058428	A/G	1.06E-05	27.34	W18	LOC106454850	4-substituted benzoates-glutamate ligase GH3.12 (GH3.12)
	Bn-scaff_18936_1-p559490	C03	3140112	A/G	1.06E-05	27.32	W18	LOC106431049	Polyamine oxidase 1 (PAO1)
	Bn-scaff_18936_1-p610540	C03	3182947	A/C	1.02E-05	27.42	W18		
	Bn-scaff_18936_1-p611810	C03	3184218	A/G	1.02E-05	27.42	W18		
BNPP	Bn-scaff_20901_1-p276384	C05	125848	A/G	4.90E-05	23.31	D17	LOC106358392	Subtilisin-like protease SBT1.1 (SBT1.1)

(Continued)

10.3389/fpls.2024.1342359

TABLE 4 Continued

Traits	Lead SNP	Chr[a]	Position (bp)[b]	Allele	*P* value[c]	R^2 (%)[d]	Watering regime	Candidate Gene[e]	Annotation
	Bn-scaff_21821_1-p122253	C05	135624	A/G	3.63E-05	20.96	D17		
	Bn-scaff_21821_1-p128045	C05	135780	A/G	3.63E-05	20.96	D17		
	Bn-scaff_20125_1-p116436	C05	136834	A/G	3.63E-05	20.96	D17		
	Bn-scaff_20125_1-p114833	C05	136975	A/G	3.63E-05	20.96	D17		
	Bn-scaff_20125_1-p110307	C05	11417447	A/G	1.31E-04	20.99	D17	LOC106347076	SKP1-like protein 3 (ASK3)
	Bn-scaff_20125_1-p110480	C05	11417644	A/G	3.63E-05	20.96	D17		
PW	Bn-scaff_19208_1-p94553	C04	35642269	A/G	6.23E-05	22.40	W20	LOC106391243	ABC transporter G family member 16-like (ABCG16)
	Bn-scaff_19208_1-p94501	C04	35642321	A/C	6.23E-05	22.40	W20		
	Bn-scaff_19208_1-p94498	C04	35642324	A/C	3.83E-05	23.55	W20		
	Bn-scaff_19208_1-p93814	C04	35643023	A/G	6.23E-05	22.40	W20		
SNPS	Bn-A02-p2124245	A02	751843	A/G	3.71E-04	18.61	D20	LOC106383096	MADS-box protein FLOWERING LOCUS C (FLC)
	Bn-A02-p2124328	A02	751923	A/G	5.18E-04	17.83	D20		
	Bn-A02-p2124421	A02	752016	A/G	2.88E-04	19.20	D20		
	Bn-A02-p2364479	A02	989812	A/C	2.66E-04	19.38	D20	LOC106383766	Protein BIG GRAIN 1-like D (BIG1-D)
	Bn-scaff_16069_1-p607537	C07	36944365	A/C	4.07E-04	15.68	D20	LOC106410807	Cullin-associated NEDD8-dissociated protein 1 (CAND1)
SWPP	Bn-A08-p13626189	A08	11377260	A/G	7.02E-04	17.23	D20	LOC125575103	Purine permease 21-like (PUP21)
	Bn-A08-p13626982	A08	11378053	A/G	7.76E-04	17.00	D20		
	Bn-A08-p13638847	A08	11390137	A/G	9.51E-04	16.53	D20	LOC106360860	Probable purine permease 10 (PUP10)
SY	Bn-scaff_19244_1-p283272	C01	224308	A/G	2.58E-05	24.345	D17	LOC106439041	Developmentally-regulated G-protein 3 (DRG3)
	Bn-scaff_19244_1-p517124	C01	449274	A/G	4.56E-04	17.693	D17	LOC125580497	Cytochrome P450 79B1 (CYP79B1)
	Bn-scaff_19244_1-p683537	C01	614927	A/G	4.68E-04	17.63	D17	LOC106426830	Agamous-like MADS-box protein AGL21 (AGL21)
	Bn-scaff_22148_1-p400066	C04	29470290	A/C	3.55E-05	23.587	D17	LOC106390665	Cytochrome P450 78A9 (CYP78A9)
	Bn-scaff_22148_1-p400140	C04	29470363	A/G	9.25E-06	23.587	D17		
	Bn-scaff_18776_1-p18005	C04	29479069	A/G	9.25E-06	23.587	D17		
	Bn-scaff_18776_1-p33555	C04	29490106	A/G	9.25E-06	23.587	D17		
	Bn-scaff_18776_1-p36364	C04	29492969	A/C	9.25E-06	23.587	D17		
	Bn-scaff_18776_1-p36455	C04	29493060	A/G	3.55E-05	23.587	D17		

DTF, DTR, PH, BNPP, PW, SWPP, and SY are the abbreviations of days to flowering, days to ripening, plant height, branch number/plant, plant weight, seed weight/plant, and seed yield, respectively. WW17, DS17, WW18, DS18, WW20, and DS20 are the codes of two watering regimes in 3 years: well-watered in 2017, drought stress in 2017, well-watered in 2018, drought stress in 2018, well-watered in 2020, and drought stress in 2020. [a]Chromosome. [b]Position in base pairs for the lead SNP according to version 4 of the rapeseed reference sequence. [c]P-value of the corresponding agronomic and yield-related traits calculated by MLM (Mixed linear model). [d]The phenotypic variance explained by the corresponding locus. [e]A plausible candidate gene in the locus.

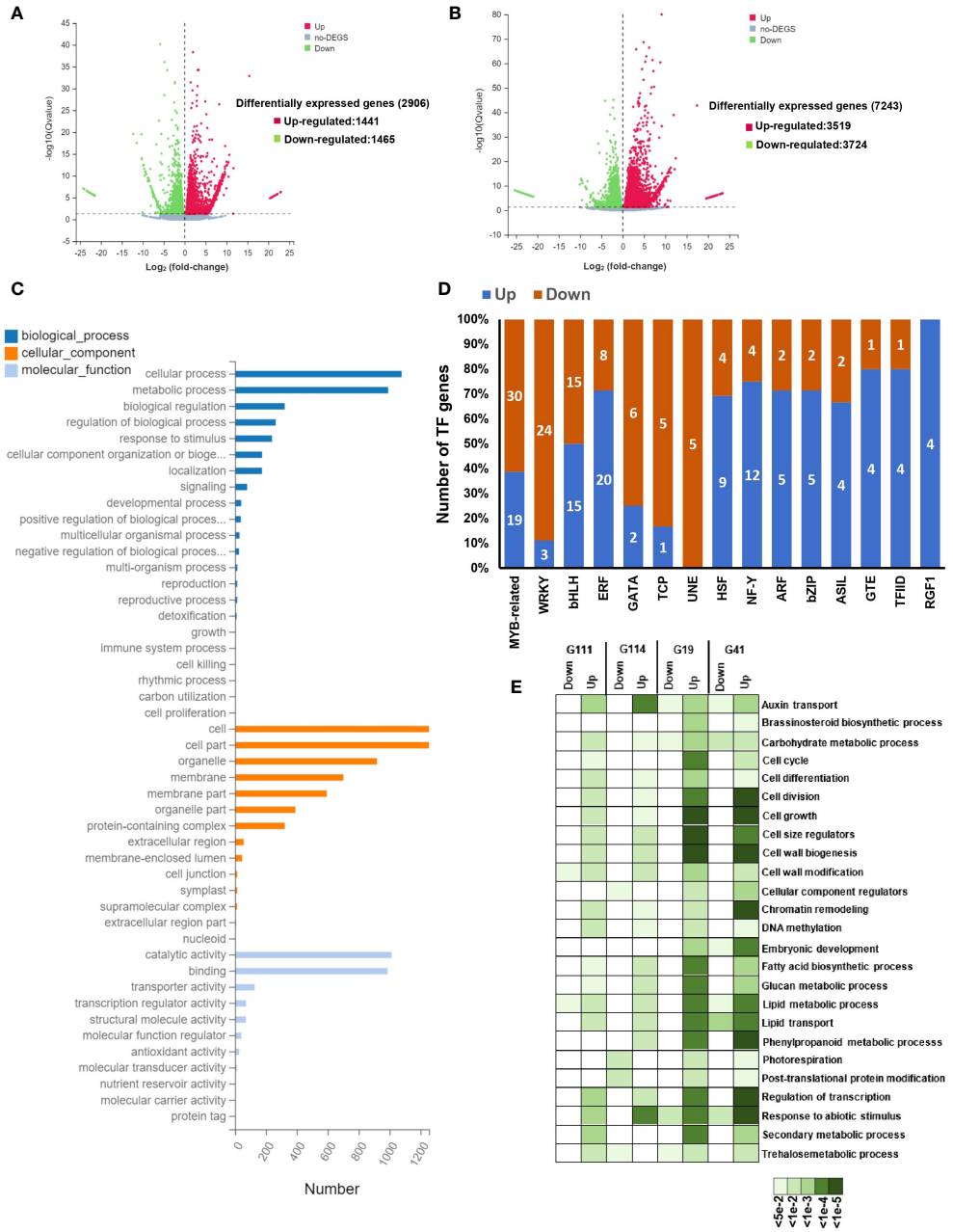


FIGURE 5
Drought-responsive transcriptional changes in rapeseed (*Brassica napus*). **(A, B)** Volcano plots showing the comparison of differentially expressed genes (DEGs) between the treatment and control groups in low-yield varieties **(A)** and high-yield varieties **(B)**. The red scatters indicate upregulated DEGs, green scatters indicate downregulated DEGs, and black scatters indicate no DEGs between the uniconazole-treated and untreated samples. Datasets were filtered to remove genes with low expression levels (dotted line from −1 to 1 on the x-axis), and a significance cutoff (*P*-value < 0.01) was applied (dotted line on the y-axis). The x-axis shows log2 fold changes in expression and the y-axis −log10 of *q*-values of a transcript being differentially expressed. **(C)** The Gene Ontology (GO) enrichment analysis of differentially expression genes. The top 23 enriched GO terms (*P*-value < 1.0E-5) were used for diagram. Gene ontology categories included biological process, cellular component, and molecular function. **(D)** The number of genes from different transcription factor (TF) families showing up- or downregulation in high-yield varieties of mature seed. **(E)** Enriched gene ontology (GO) terms of mature seed in down- and upregulated genes in high-yield varieties and low-yield varieties. The color scale at the bottom represents significance (corrected *P*-value).

drought than in the WW condition (Figures 8I–K). The *AGL21* gene with 33.42-fold change showed higher expression in leaf under drought than in the WW condition in G19 (Figure 8K). The *CAND1* and *DRG3* genes had higher expression in seed than in leaf samples in both low- and high-yield varieties under drought conditions (Figures 8I, K–N).

3.10 Validation of major drought-related genes by combining GWAS and RNA-seq results

We identified the DEGs in the selected yield contrasting varieties in the drought compared with the well-watered treatment. A gene

10.3389/fpls.2024.1342359

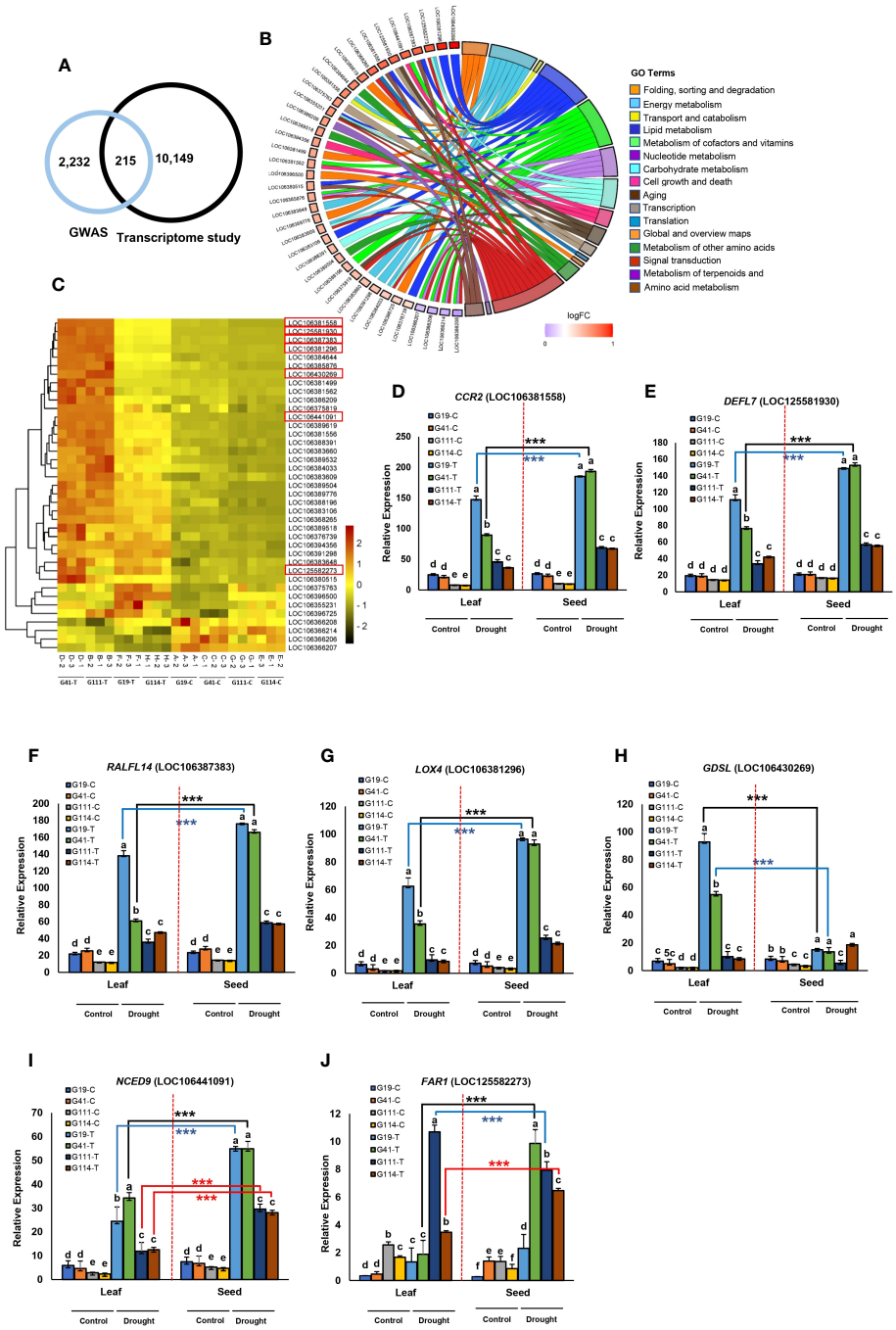


FIGURE 6

Analysis of transcriptomes between high- and low-yielding varieties selected in drought stress treatment in the field study. Transcriptome study for days to flowering (DTF). **(A)** Venn diagram of overlapped genes significant in GWAS and transcriptome analysis. **(B)** Detailed analysis of 16 enriched gene ontology groups selected using Circos plots for candidate genes in chromosomes A02, C01, C02, and C04 related to DTF. Symbols of DEG from each of the analyzed comparisons are displayed on the left side of the graph with their logFC values, mapped by color scale (red = higher expression; violet = lower expression). The white color corresponds to expression levels below the cutoff value for the given comparisons. Colored connecting lines determine gene involvement in the GO terms. **(C)** Heat map of the expression of 40 genes associated with DTF in chromosomes A02, C01, C02, and C04 among low-yield varieties (G111 and G114) and high-yield varieties (G19 and G41) under drought stress. The red box indicates the key genes in the associated region of C02 related to DTF. Heatmap color represents the expression level of each gene (rows) under drought treatments (fold change > 1, *P*-value < 0.05). Red bars: upregulation; green bars: downregulation; (C): control and (T): drought treatment. **(D-J)** Tissue-specific expression of seven candidate genes in the peak regions of C02 from four varieties of rapeseed (low-yield varieties: G111 and G114; high-yield varieties: G19 and G41) under drought stress; (C): control and (T): drought treatment. Tissue-specific expression of the gene *CCR2* **(D)**, *DEFL7* **(E)**, *RALFL14* **(F)**, *LOX4* **(G)**, *GDSL* **(H)**, *NCED9* **(I)**, and *FAR1* **(J)** between low-yield varieties and high-yield varieties under drought stress. Data are means ± SD, *P*-value < 0.05, as determined by multiple comparison testing by one-way ANOVA. Different letters indicate distinct groups. Red dotted line indicates division expression of genes between seed and leaf. Asterisks indicate significant difference between seed and leaf (Student's *t*-test, ***P*-value < 0.001).

10.3389/fpls.2024.1342359

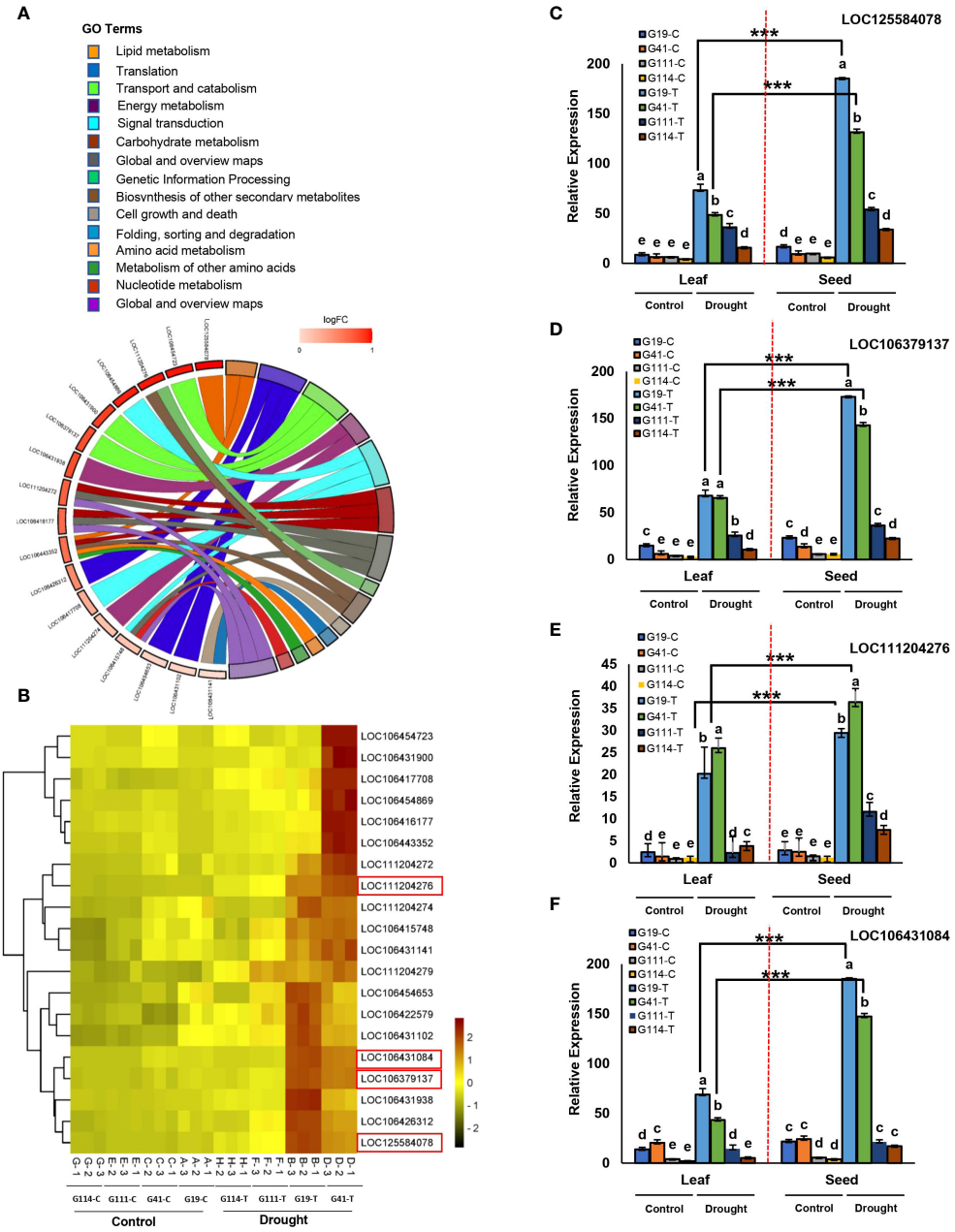


FIGURE 7
Analysis of transcriptomes between high- and low-yielding varieties selected in drought stress treatment in the field study. Transcriptome study for days to ripening (DTR). **(A)** Detailed analysis of 15 enriched gene ontology groups selected using Circos plots for candidate genes in chromosomes A01, C03, and C06 related to DTR. Symbols of DEG from each of the analyzed comparisons are displayed on the left side of the graph with their logFC values, mapped by color scale (dark red = higher expression; light red = lower expression). Colored connecting lines determine gene involvement in the GO terms. **(B)** Heat map of the expression of 20 genes associated with DTR in chromosomes A01, C03, and C06 among low-yield varieties (G111 and G114) and high-yield varieties (G19 and G41) under drought stress. The red box indicates the key genes in the associated region of C03 related to DTR. Heatmap color represents the expression level of each gene (rows) under drought treatments (fold change >1, *P*-value <0.05). Red bars: upregulation; green bars: downregulation; (C): control and (T): drought treatment. **(C-F)** Tissue-specific expression of seven candidate genes in the peak regions of C03 from four varieties of rapeseed (low-yield varieties: G111 and G114; high-yield varieties: G19 and G41) under drought stress; (C): control and (T): drought treatment. Tissue-specific expression of the gene LOC125584078 **(C)**, LOC106379137 **(D)**, LOC111204276 **(E)**, and LOC106431084 **(F)** between low-yield varieties and high-yield varieties under drought stress. Data are means ± SD, *P*-value < 0.05, as determined by multiple comparison testing by one-way ANOVA. Different letters indicate distinct groups. Red dotted line indicates division expression of genes between seed and leaf. Asterisks indicate significant difference between seed and leaf (Student's *t*-test, *** *P*-value < 0.001).

ontology enrichment analysis was also performed to identify the functional roles of DEGs and variety-specific responses under drought stress conditions. Transcriptomic analysis showed that 10 DEGs (*BG1*, *FLC*, *DRG3*, *CAND1*, *PUP10*, *PUP21*, *ABCG16*, *AGL21*, *CYP79B1*, and *CYP78A9*) had significantly higher expression in the high- than in low-yield varieties. These CGs were upregulated in the low-yield varieties under the drought stress compared with the well-watered conditions. Consistent with the higher number of DEGs

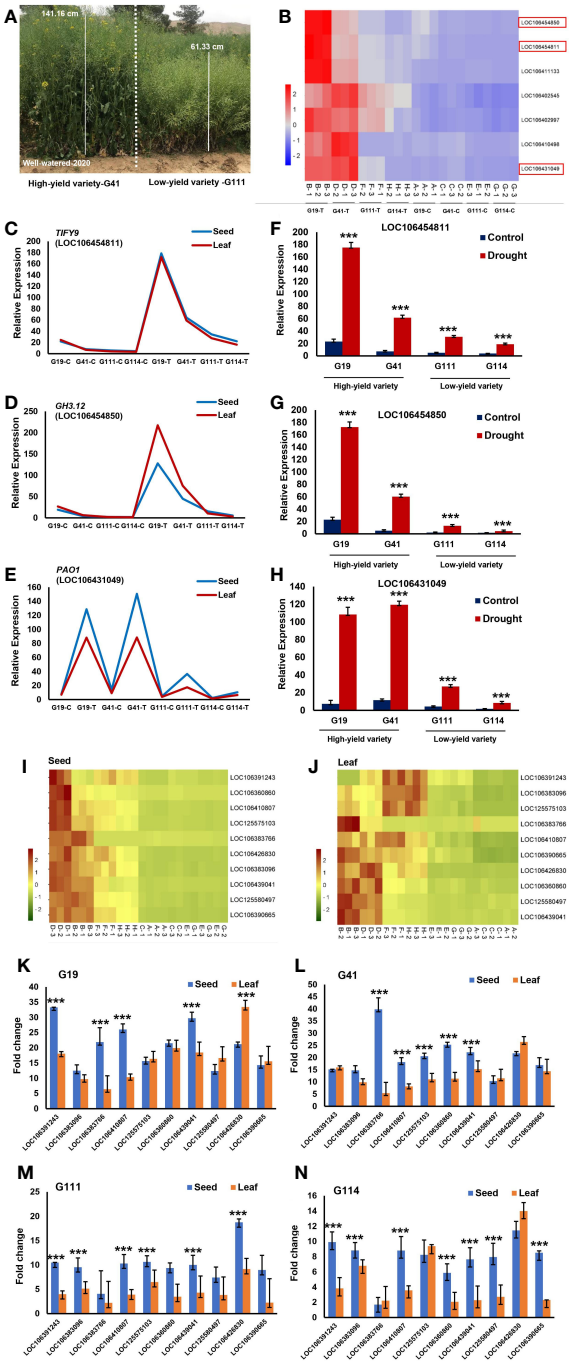


FIGURE 8
Transcriptome study for plant height (PH) and seed yield (SY). **(A)** Plant height phenotypes of two rapeseed varieties; G41 as a high-yield varieties and G111 as a low-yield varieties were grown in field under well-watered conditions in 2020. **(B)** Heat map of the expression of seven genes associated with PH in chromosomes A07, C03, C06, and C07 among low-yield varieties (G111 and G114) and high-yield varieties (G19 and G41) under drought stress. The red box indicates the key genes in the associated region of C03 related to PH. Heatmap color represents the expression level of each gene (rows) under drought treatments (fold change > 1, *P*-value < 0.05). Red bars: upregulation; blue bars: downregulation; (C): control and (T): drought treatment. **(C-E)** Tissue-specific expression of three candidate genes in the peak regions of C03 from four varieties of rapeseed (low-yield varieties: G111 and G114; high-yield varieties: G19 and G41) between seed and leaf under drought stress; (C): control and (T): drought treatment. Tissue-specific expression of the gene *TIFY9* **(C)**, *GH3.12* **(D)**, and *PAO1* **(E)** between seed and leaf from four varieties. **(F)** Expression pattern of the gene *TIFY9* between low-yield varieties and high-yield varieties under drought stress. **(G)** Expression pattern of the gene *GH3.12* between low-yield varieties and high-yield varieties under drought stress. **(H)** Expression pattern of the gene *PAO1* between low-yield varieties and high-yield varieties under drought stress. **(I, J)** Heat map of the expression of 10 key genes associated with SY in chromosomes A02, A08, C01, C04, and C07 among low-yield varieties (G111 and G114) and high-yield varieties (G19 and G41) from seed **(I)** and leaf **(J)** under drought stress. Heatmap color represents the expression level of each gene (rows) under drought treatments (fold change >1, *P*-value < 0.05). Red bars: upregulation; green bars: downregulation; (C): control and (T): drought treatment. **(K-N)** Tissue-specific expression of 10 genes associated with SY in four varieties: G19 **(K)**, G41 **(L)**, G111 **(M)**, and G114 **(N)** under drought stress. Asterisks indicate significant difference between seed and leaf (Student's *t*-test, ***P*-value < 0.001).

observed in the leaf or seed of the high-yield varieties in response to drought, higher enriched GO terms unique to the high-yield varieties, particularly in seed, were detected. Most of the enriched GO terms exclusive to the high-yield varieties in seeds were those processes that were associated with abiotic stress, including the responses to ABA, osmotic adjustment, and regulation of stomatal movements. A number of 25 upregulated DEGs were identified in the seeds of the high-yield compared with 20 DEGs in the low-yield varieties (Supplementary Figure S7). In the high-yield varieties, xyloglucan endotransglucosylase/hydrolases (*XTH13*), triacylglycerol lipase (*SDP1*), serine/threonine-protein kinase (*SRK2F*), ABC transporter F family (*ABCF1*), ethylene-responsive transcription factor (*ERF113*), zinc-finger proteins (*ZC3H49*), F-box protein (*FBD*), NAC domain-containing proteins (*NAC072*), and acetyl-CoA acetyltransferase (*FadA*) showed above fivefold change (Supplementary Figures S7C, E). Upregulation of the *XTH13, NAC072* and *FadA* genes was observed in the two contrasting varieties (Supplementary Figures S7C–F). The *XTH* gene family is involved in various physiological processes in plants, especially in abiotic stress responses and cell elongation (Ma et al., 2022). The *XTH* genes have been regulated by TF NAC by directly binding to the promoter region of the *XTH* genes improving stress tolerance (Tao et al., 2022).

4 Discussion

4.1 Traits variations in rapeseed under drought stress

Breeding rapeseed which is an important source of oil and protein for food and industrial applications is challenging due to environmental stresses worldwide (Elferjani and Soolanayakanahally, 2018; Secchi et al., 2023). Results of evaluation of a rapeseed population over three growing seasons showed large genetic diversity for morphological, phenological, and yield traits in both drought and well-watered treatments. Our data showed a significant treatment by trait value interaction. Phenotyping under two irrigation regimes revealed differences for the traits among varieties over treatments and years that were higher than those in earlier studies (Korber et al., 2016; Luo et al., 2017; Raman et al., 2020; Menendez et al., 2021; Qin et al., 2022; Yang et al., 2023). Higher heritability and genetic gain from selection for several traits in this study under two contrasting irrigation regimes indicated selection potential for the improvement of drought tolerance in the tested rapeseed population. Our study showed that the 119 varieties could be clustered into seven supported genetic lineages that roughly reflected their geographical origin consistent with previous studies (Gazave et al., 2016; Bird et al., 2017).

4.2 Novel and stable loci identified for agronomic and yield-related traits under drought conditions

Due to the sensitivity of complex traits to the environmental effects, the integration of GWAS and RNA-seq across different tissues might assist identification of candidate genes responsible for stress tolerance. Our study uncovered numerous loci correlated with variation in agronomic and yield-related traits. This study identified a set of candidate genes that could be exploited to alter agronomic traits and yield components to improve grain yield in rapeseed varieties. Life cycle timing is critical for yield and productivity of *Brassica napus* cultivars grown in different environments. Timing of flowering is crucial for optimal pollination, survival in specific environments, high seed quality, and seed yield and maintaining seed propagation in crop rotation systems (Kirkegaard et al., 2018). In the present study, we identified six stable SNPs for flowering time that were common between the two irrigation treatments across years. In addition, 171 novel signals significantly associated with flowering time were identified under drought conditions. Numerous loci of flowering time in *B. napus* have been identified on chromosomes A02, C02, and C03 by QTL mapping and GWAS in other studies (Akhatar et al., 2021; Vollrath et al., 2021; Han et al., 2022). We identified 21 novel MTAs on chromosome C03 for ripening time in rapeseed that were not previously reported. In temperate regions such as Iran, early flowering and maturity are important breeding targets in *B. napus* because drought restricts the growth season (Helal et al., 2021). Identification of candidate loci involved in drought tolerance can be used for marker-assisted selection and helps develop drought-tolerant rapeseed for dry regions. In a meta-QTL analysis, co-localized QTLs for flowering and maturity times were identified on chromosomes A01, A02, A03, A05, and C09 (Zhou et al., 2014).

Plant architecture (PA), which refers to the spatial distribution pattern of aboveground parts including plant height (PH) and number of aerial branches, is influenced by both genetic and environmental factors (Cai et al., 2016; Wang et al., 2019; Dong et al., 2022). In the present study, we identified 70 unique SNPs on chromosomes A01, A02, A04, A07, A09, A10, C03, and C05 associated with PH under two irrigation regimes of which several SNPs were close to those identified in previous studies (Sun et al., 2016a; Zhang et al., 2023a). In line with results of the Liu et al. (2021) and Zheng et al. (2017) studies, the highest number of significant SNPs for PH was identified on chromosome C03. In the current study, we identified 22 unique SNPs associated with branch number under drought stress condition that were not discovered in previous studies in rapeseed. These SNPs provide new information for understanding the establishment of ideal PA and developing breeding strategies for yield improvement in rapeseed. The molecular genetic mechanisms underlying branch number have been analyzed using GWAS in other rapeseed studies (Sun et al., 2016b; Dong et al., 2022; Hu et al., 2022). In the Sun et al. (2016b) study, 56 unique loci significantly associated with branch number explained up to 51.1% of the phenotypic variation. Hu et al. (2022) identified two significant SNP signals on chromosome C07 associated with branch trait in three environments.

Increasing the yield potential is a major goal for rapeseed breeding. We identified SNPs linked with yield and yield components of which several were stable over years and unique to this study. We identified 10 significant association signals for each THSW and SNPS traits that were stable across 2 of 3 years under drought stress conditions. Among the linked SNPs, 37

associated with two or more traits. Our pleiotropic SNP Bn-A02-p3539297 affecting both DTF and DTSD traits in our study has also been detected in a GWAS in the Xu et al. (2016) study. The position of the pleiotropic SNP Bn-A10-p13390065 affected DTF and DTSD in our study was close to the position of four flowering time-related genes (*BnaA10g18420D*, *BnaA10g18480D*, *BnaA10g22080D*, and *BnaA10g24300D*) identified in the Helal et al. (2021) study. Markers stable over environments and those with pleiotropic effects are preferred for use in marker-assisted selection (MAS) programs. According to the SNP-trait associations, we identified 188 SNPs for SY of which the SNPs in the 34,329-kbps–34,381-kbps interval on chromosome C06 were adjacent to the position of previously reported SNPs associated with yield components (Pal et al., 2021). In addition, the SNP Bn-A08-p14538807 on chromosome A08 was close to the SNP Chr18:12100271 detected for siliqua length in Pal et al. (2021). The SNPs in the 131-kbps–677-kbps interval on chromosome C01 overlapped with the position of the linked SNPs detected by Zhang et al. (2011) and Dong et al. (2018). In this study, 38 SNPs for HI were identified under two irrigation regimes across 3 years of which the SNPs Bn-scaff_16361_1-p2541607 and Bn-scaff_16361_1-p2545776 on chromosome C08 were close to the SNP Chr18:33618188 detected for HI in Qin et al. (2022).

4.3 Candidate genes involved in rapeseed growth under drought stress

Information about genetic control of architecture- and phenology-related characters helps in breeding for drought tolerance and avoidance. Integration of GWAS and RNA-seq revealed 59 DEGs associated with flowering and maturity times of which the *LOX4, GDSL*, *FAR1, NCED9, CCR2, RALFL14, DEFL7*, *ORYZASIN1*, *LAX2*, *FY*, *AGL15*, *AGL16*, and *AGL19* genes were key flowering and maturity genes in our study. In the hormone pathway, the *LOX4*, *NCED9*, *ORYZASIN1*, and *LAX2* genes play important roles in flower development and pollen tube growth (Caldelari et al., 2011; Chauvin et al., 2013; Huang and Han, 2014; Gomez et al., 2015; Li et al., 2018; Zuniga-Mayo et al., 2018; Park et al., 2019; Tran et al., 2023; Zhang et al., 2023d). From the auxin carrier AUX/LAX family, *LAX2* regulates the floral organ development by modulating auxin polar transport (Cardarelli and Cecchetti, 2014). Auxin, which is required for floral meristem initiation, regulates floral organ initiation, growth, and patterning that ensure reproductive success of the mature flower (Iqbal et al., 2017). In the present study, we also identified three members of MADS-box genes, *AGL15*, *AGL16*, and *AGL19*, that were involved in flowering and maturity in rapeseed. *FAR1, CCR2*, *AGL15*, *AGL16*, and *AGL19* regulate several events during pollen development such as tapetal degradation, the formation of anther cuticle and pollen exine, central vacuole development, and flowering transition (Fang and Fernandez, 2002; Lee and Lee, 2010; Whittaker and Dean, 2017; Ma et al., 2020; Zhao et al., 2020; Li et al., 2021; Chen et al., 2022). It has been shown that *FAR1* involved in floral bud differentiation interacts with proteins of flowering promoting SQUAMOSA-PROMOTER BINDING PROTEIN-LIKE (SPL) transcription factors (Xie et al., 2020). *FAR1* is even involved in branching regulation and floral bud sex differentiation (Hiraoka et al., 2013). Plant GDSL lipases form a large gene family, and members have been identified in *Arabidopsis* (105), *Oryza sativa* (114), and *Brassica rapa* L. (121) (Lai et al., 2017; Shen et al., 2022), which regulate seed development (Ma et al., 2018; Ding et al., 2019; Zhang et al., 2020; Liu et al., 2023).

Seven DEGs including *TIFY9*, *GH3.12*, and *PAO1* were identified as the candidate genes of plant height in the present study. *TIFY* genes play an important role in leaf and stem growth and responses to environmental stresses (Vanholme et al., 2007; Cai et al., 2014; Baekelandt et al., 2018; Liu et al., 2020a; Singh and Mukhopadhyay, 2021). The overexpression of *TIFY1* in *Arabidopsis* resulted in elongated petioles and hypocotyls due to increased cell elongation (Shikata et al., 2004). Previous studies have shown that the *BrGH3.12*, one of the CGs identified in our study, was highly expressed in the leaf apical region, which regulate flowering in *B. rapa* (Gu et al., 2017). It has been shown that upregulation of *BnaC03.GH3-12* may improve stress tolerance ability in *B. napus* (Wang et al., 2019). Polyamine oxidase (PAO) has been detected in many actively growing tissues (roots, stems, leaves, and floral organs) and plays a key role in maintaining normal growth and resisting the adverse environmental stresses in plants (Murray Stewart et al., 2018; Yu et al., 2019; Wu et al., 2022b; Samanta et al., 2023).

4.4 Genes responsible for yield variation between the high- and low-yield varieties under drought conditions

The number of seeds per silique and seed size/seed weight traits are the important goals of *B. napus* breeding and development of new high-density seeding varieties (Zhu et al., 2023). In the present study, 10 DEGs were identified as the yield component-related CGs of which *BIG GRAIN 1* (*BG1*) governs seed size in legume species (Ge et al., 2016) and rice (Lo et al., 2020). The expression level of *BG1* and *PUP10* in the high-yield varieties was higher than in low-yield varieties, which suggests that *BG1* and *PUP10* might positively regulate the grain yield. The *BG3* gene encoding a purine permease regulates grain size *via* modulating cytokinin (CK) transport in rice (Xiao et al., 2019). Among the CGs detected in our study, long- and short-distance transporters including *ABCG16*, *PUP10*, and *PUP21* are from the ABC transporter and PUP families that are involved in CK traffic (Gillissen et al., 2000; Tsai et al., 2012; Xiao et al., 2019; Zhao et al., 2019; Liu et al., 2020b).

We identified *CYP78A9* and *CYP79B1* as two members of the CYP family genes that play highly conserved roles in facilitating organ growth including floral organ, seeds, embryos, and endosperm in *Arabidopsis*, maize, soybean, and rice (Sotelo-Silveira et al., 2013; Wang et al., 2015b; Xu et al., 2015; Sun et al., 2017; Yeon et al., 2021). Shi et al. (2019) reported that the

expression level of *BnaA9.CYP78A9* in silique valves of the long-silique variety was much higher than that in the regular-silique variety and that the long-silique plants showed higher concentrations of auxin in the developing silique, suggesting that *BnaA9.CYP78A9* contributes to the silique elongation phase in rapeseed.

The expression level of *cullin-associated NEDD8-dissociated protein 1* (*CAND1*) and *developmentally regulated G-protein 3* (*DRG3*) genes associated with yield components in our study was much higher in seed than in leaf of high-yield varieties. One of the critical elements of plant reproduction is the production of functional pollen grains. The expression of *CAND1* which is one of the key regulators of the SKP1-CUL1/RBX1-F-BOX (SCF) complex is imperative for fertility (Feng et al., 2004; Li et al., 2022; Rani et al., 2023). *CAND1* regulates the dynamic and functionality of the SCF complex, which is required for pollen and grain development) Hong et al., 2014(. Analysis of the *cand1-3* mutants in the Li et al. (2022) study has shown that *CUL1* from the SCF complex was expressed in pollen at a level significantly higher than any other tissue in *Arabidopsis* that could be due to the high concentration of the SCF substrates in pollen. *DRG3*, one of the CGs in our study, is a member of the family of GTP binding protein (G-proteins). G-proteins regulate plant growth and development pathways especially phytohormone signaling and cross-talk and defense responses (Kumar et al., 2014; Pandey, 2020). G-proteins are also known to regulate key agronomical traits such as seed size and yield (Cui et al., 2020). Roy Choudhury et al. (2014) identified the correlation between plant-specific G protein expression in seed tissue with higher seed size, seed mass, and seed number per plant, effectively resulting in significantly higher seed yield in *Camelina sativa*.

4.5 Candidate genes associated with enhanced drought tolerance in rapeseed

Breeding drought-tolerant rapeseed has always been a tough challenge for plant breeders. The use of grain yield as a selection criterion has proved to be a boon in this area of research. In addition, understanding response to drought stress in high-yielding crops at the molecular level is useful for developing drought tolerance. In this study, we found the *XTH13*, *SDP1*, *SRK2F*, *ABCF1*, *ERF113* genes that were upregulated in the high-yield varieties under drought conditions. Xyloglucan endotransglucosylase/hydrolase (XTH) is one of the critical enzymes which contribute to the development and strengthening of cell walls (Shi et al., 2015). The constitutive expression of *XTH* that increased stomatal closures was conferred by the increased cell-wall remodeling activity of *XTH* in guard cells, which may reduce transpirational water loss in response to dehydration stress (Choi et al., 2011; Han et al., 2017). In plants, the SnRK family comprises 38 members, which can be subdivided into three subfamilies: SnRK1, SnRK2, and SnRK3. Compelling evidence suggests that SnRK2s are involved in ABA and/or stress signaling pathways. The SnRK2 protein kinase is a central regulator of abscisic acid (ABA)-dependent stomatal closure (Kamiyama et al., 2021; Hasan et al., 2022). Some SnRK2 are also activated by hyperosmotic stress (Julkowska et al., 2015; Hasan et al., 2022). *ERF113* is a member of the ethylene response factor (ERF) family that is induced by salt stress and drought stresses (Debbarma et al., 2019). Additionally, *ERF113* transcription is responsive to jasmonic acid, salicylic acid, ABA, and ethylene hormones for stress tolerance (Krishnaswamy et al., 2011). In the Fang et al. (2022) study, transgenic soybean plants overexpressing *ERF113* showed significantly slower water loss in the leaves than wild type and plants with RNAi silencing of the gene under drought stress. These results reveal that the *XTH13*, *SDP1*, *SRK2F*, *ABCF1*, and *ERF113* genes identified in our study might improve drought tolerance in rapeseed, providing a theoretical basis for the molecular breeding of drought-tolerant varieties.

5 Conclusion

In the present study, multiomics analysis with the use of phenotypic, genomic, and transcriptomic data was performed to identify the major QTNs/genes associated with the plant responses to the effects of drought stress in rapeseed. We found high variation and linked SNPs for agronomic and seed yield-related traits that can be used in marker-assisted breeding (MAB) and development of improved rapeseed varieties with high yield under drought stress conditions. Two SNPs (Bn-scaff_18936_1-p610540 and Bn-scaff_18936_1-p472353) on chromosome C03 explained 43.65% and 42.99% of the phenotypic variance of ripening time, which is an important drought-adaptive trait. There were 10 drought tolerance-related genomic regions located on chromosome A02 (750,194 bp–791,056 bp) that had not been reported in previous studies in rapeseed. Four novel SNPs on chromosome C04 for plant weight were located within the sequence of the drought tolerance-related gene *ABCG16*, and their pleiotropic effects on seed weight per plant and seed yield were characterized. The GWAS analysis showed that 49 pleiotropic SNPs on chromosomes A01, A02, A08, A10, C03, and C04 were associated with at least two traits and 10 SNPs had a stable association with thousand seed weight overs 2 years under drought stress conditions. Overall, our study provides supporting information for three research areas. One is an irrigation treatment by trait value interaction and the interrelationship of novel candidate genes and drought-adaptive traits that help to unravel the drought tolerance mechanisms in rapeseed. The other is comparative transcriptomics that proved that seed weight/plant and plant weight at the phenotypic level and stress signaling pathway genes at the molecular level had higher contributions in response to the high-yield varieties to drought stress. The third is a large SNP data set correlated with drought tolerance, which will accelerate future efforts aiming at the development of drought-tolerant varieties through a MAS program.

Data availability statement

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

Author contributions

MS: Investigation, Writing – original draft, Data curation, Formal analysis, Methodology, Software. BH: Investigation, Writing – original draft, Conceptualization, Project administration, Supervision, Visualization, Writing – review & editing. BA: Investigation, Methodology, Writing – review & editing. JB: Data curation, Investigation, Methodology, Writing – review & editing. JW: Data curation, Investigation, Visualization, Writing – review & editing. XT: Data curation, Investigation, Visualization, Writing – review & editing. AD: Writing – review & editing. CR: Visualization, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

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

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

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2024.1342359/full#supplementary-material

References

Akhatar, J., Goyal, A., Kaur, N., Atri, C., Mittal, M., Singh, M. P., et al. (2021). Genome wide association analyses to understand genetic basis of flowering and plant height under three levels of nitrogen application in *Brassica juncea* (L.) Czern and Coss. *Sci. Rep.* 11, 4278. doi: 10.1038/s41598-021-83689-w

Alemu, A., Brazauskas, G., Gaikpa, D. S., Henriksson, T., Islamov, B., Jorgensen, L. N., et al. (2021). Genome-wide association analysis and genomic prediction for adult-plant resistance to Septoria tritici blotch and powdery mildew in winter wheat. *Front. Genet.* 12. doi: 10.3389/fgene.2021.661742

Baekelandt, A., Pauwels, L., Wang, Z., Li, N., De Milde, L., Natran, A., et al. (2018). Arabidopsis leaf flatness is regulated by PPD2 and NINJA through repression of *CYCLIN D3* genes. *Plant Physiol.* 178, 217–232. doi: 10.1104/pp.18.00327

Barrett, J. C., Fry, B., Maller, J. D. M. J., and Daly, M. J. (2005). Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21, 263–265. doi: 10.1093/bioinformatics/bth457

Bernardo, R. (2008). Molecular markers and selection for complex traits in plants: Learning from the last 20 years. *Crop Sci.* 48, 164916–164964. doi: 10.2135/cropsci2008.03.0131

Bird, K. A., An, H., Gazave, E., Gore, M. A., Pires, J. C., Robertson, L. D., et al. (2017). Population structure and phylogenetic relationships in a diverse panel of *Brassica rapa* L. *Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.00321

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

Burton, G. W. (1952). "Quantitative inheritance in grasses," in *Proceedings of the 454 International Grassland Congress*, Vol. 1 (State College, PA: Pennsylvania State College), 277–283.

Cai, G., Yang, Q., Chen, H., Yang, Q., Zhang, C., Fan, C., et al. (2016). Genetic dissection of plant architecture and yield-related traits in *Brassica napus*. *Sci. Rep.* 6, 21625. doi: 10.1038/srep21625

Cai, Q., Yuan, Z., Chen, M., Yin, C., Luo, Z., Zhao, X., et al. (2014). Jasmonic acid regulates spikelet development in rice. *Nat. Commun.* 5, 3476. doi: 10.1038/ncomms4476

Caldelari, D., Wang, G., Farmer, E. E., and Dong, X. (2011). Arabidopsis lox3 lox4 double mutants are male sterile and defective in global proliferative arrest. *Plant Mol. Biol.* 75, 25–33. doi: 10.1007/s11103-010-9701-9

Cardarelli, M., and Cecchetti, V. (2014). Auxin polar transport in stamen formation and development: how many actors? *Front. Plant Sci.* 5. doi: 10.3389/fpls.2014.00333

Castelan-Munoz, N., Herrera, J., Cajero-Sanchez, W., Arrizubieta, M., Trejo, C., Garcia-Ponce, B., et al. (2019). MADS-box genes are key components of genetic regulatory networks involved in abiotic stress and plastic developmental responses in plants. *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.00853

Chalhoub, B., Denoeud, F., Liu, S., Parkin, I. A., Tang, H., Wang, X., et al. (2014). Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. *Science* 345, 950–953. doi: 10.1126/science.1253435

Chauvin, A., Caldelari, D., Wolfender, J. L., and Farmer, E. E. (2013). Four 13-lipoxygenases contribute to rapid jasmonate synthesis in wounded *Arabidopsis thaliana* leaves: a role for lipoxygenase 6 in responses to long-distance wound signals. *New Phytol.* 197, 566–575. doi: 10.1111/nph.12029

Chen, W. H., Lin, P. T., Hsu, W. H., Hsu, H. F., Li, Y. C., Tsao, C. W., et al. (2022). Regulatory network for *FOREVER YOUNG FLOWER*-like genes in regulating *Arabidopsis* flower senescence and abscission. *Commun. Biol.* 5, 662. doi: 10.1038/s42003-022-03629-w

Choi, J. Y., Seo, Y. S., Kim, S. J., Kim, W. T., and Shin, J. S. (2011). Constitutive expression of *CaXTH3*, a hot pepper xyloglucan endotransglucosylase/hydrolase, enhanced tolerance to salt and drought stresses without phenotypic defects in tomato plants (*Solanum lycopersicum* cv. Dotaerang). *Plant Cell Rep.* 30, 867–877. doi: 10.1007/s00299-010-0989-3

Cui, Y., Jiang, N., Xu, Z., and Xu, Q. (2020). Heterotrimeric G protein are involved in the regulation of multiple agronomic traits and stress tolerance in rice. *BMC Plant Biol.* 20, 1–13. doi: 10.1186/s12870-020-2289-6

Debbarma, J., Sarki, Y. N., Saikia, B., Boruah, H. P. D., Singha, D. L., and Chikkaputtaiah, C. (2019). Ethylene response factor (ERF) family proteins in abiotic stresses and CRISPR–Cas9 genome editing of ERFs for multiple abiotic stress tolerance in crop plants: a review. *Mol. Biotechnol.* 61, 153–172. doi: 10.1007/s12033-018-0144-x

Ding, L. N., Guo, X. J., Li, M., Fu, Z. L., Yan, S. Z., Zhu, K. M., et al. (2019). Improving seed germination and oil contents by regulating the GDSL transcriptional level in *Brassica napus*. *Plant Cell Rep.* 38, 243–253. doi: 10.1007/s00299-018-2365-7

Dodig, D., Zoric, M., Knezevic, D., King, S. R., and Surlan-Momirovic, G. (2008). Genotype × environment interaction for wheat yield in different drought stress conditions and agronomic traits suitable for selection. *Aust. J. Agric. Res.* 59, 536–545. doi: 10.1071/AR07281

Dong, H., Tan, C., Li, Y., He, Y., Wei, S., Cui, Y., et al. (2018). Genome-wide association study reveals both overlapping and independent genetic loci to control seed weight and silique length in *Brassica napus*. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.00921

Dong, Z., Tang, M., Cui, X., Zhao, C., Tong, C., Liu, Y., et al. (2022). Integrating GWAS, linkage mapping and gene expression analyses reveal the genetic control of first branch height in *Brassica napus* L. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.1080999

Earl, D. A., and VonHoldt, B. M. (2011). STRUCTURE HARVESTER: a website and a program for vizualizing STRUCTURE output and implementing the Evanno method. *Conserv. Genet. Resour.* 4, 359–361. doi: 10.1007/s12686-011-9548-7

Elferjani, R., and Soolanayakanahally, R. (2018). Canola responses to drought, heat, and combined stress: shared and specific effects on carbon assimilation, seed yield, and oil composition. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.01224

Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol.* 14, 2611–2620. doi: 10.1111/j.1365-294X.2005.02553.x

Fang, S. C., and Fernandez, D. E. (2002). Effect of regulated overexpression of the MADS domain factor *AGL15* on flower senescence and fruit maturation. *Plant Physiol.* 130, 78–89. doi: 10.1104/pp.004721

Fang, X., Ma, J., Guo, F., Qi, D., Zhao, M., Zhang, C., et al. (2022). The AP2/ERF GmERF113 positively regulates the drought response by activating *GmPR10-1* in soybean. *Int. J. Mol. Sci.* 23, 8159. doi: 10.3390/ijms23158159

Fang, C., Wu, S., Li, Z., Pan, S., Wu, Y., An, X., et al. (2023). A systematic investigation of lipid transfer proteins involved in male fertility and other biological processes in maize. *Int. J. Mol. Sci.* 24, 1660. doi: 10.3390/ijms24021660

Food and Agriculture Organization of the United Nations (2018). *Oilcrops, oils and meals market assessment*. Available at: http://faostat.fao.org.

Feng, S., Shen, Y., Sullivan, J. A., Rubio, V., Xiong, Y., Sun, T. P., et al. (2004). Arabidopsis CAND1, an unmodified CUL1-interacting protein, is involved in multiple developmental pathways controlled by ubiquitin/proteasome-mediated protein degradation. *Plant Cell* 16, 1870–1882. doi: 10.1105/tpc.021949

Fordonski, G., Pszczokowska, A., Okorski, A., Olszewski, J., Zaluski, D., and Gorzkowska, A. (2016). The yield and chemical composition of winter oilseed rape seeds depending on different nitrogen fertilization doses and the preceding crop. *J. Elem.* 21, 1225–1234. doi: 10.5601/jelem.2016.21.2.1122

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., et al. (2002). The structure of haplotype blocks in the human genome. *Science* 296, 2225–2229. doi: 10.1126/science.1069424

Gao, X., Zhang, J., Li, J., Wang, Y., Zhang, R., Du, H., et al. (2022). The phosphoproteomic and interactomic landscape of qGL3/OsPPKL1-mediated brassinosteroid signaling in rice. *Plant J.* 109, 1048–1063. doi: 10.1111/tpj.15613

Gazave, E., Tassone, E. E., Ilut, D. C., Wingerson, M., Datema, E., Witsenboer, H. M., et al. (2016). Population genomic analysis reveals differential evolutionary histories and patterns of diversity across subgenomes and subpopulations of *Brassica napus* L. *Front. Plant Sci.* 7. doi: 10.3389/fpls.2016.00525

Ge, L., Yu, J., Wang, H., Luth, D., Bai, G., Wang, K., et al. (2016). Increasing seed size and quality by manipulating *BIG SEEDS1* in legume species. *Proc. Natl. Acad. Sci.* 113, 12414–12419. doi: 10.1073/pnas.1611763113

Gillissen, B., Burkle, L., Andre, B., Kuhn, C., Rentsch, D., Brandl, B., et al. (2000). A new family of high-affinity transporters for adenine, cytosine, and purine derivatives in Arabidopsis. *Plant Cell* 12, 291–300. doi: 10.1105/tpc.12.2.291

Gomez, J. F., Talle, B., and Wilson, Z. A. (2015). Anther and pollen development: a conserved developmental pathway. *J. Integr. Plant Biol.* 57, 876–891. doi: 10.1111/jipb.12425

Gu, A., Meng, C., Chen, Y., Wei, L., Dong, H., Lu, Y., et al. (2017). Coupling Seq-BSA and RNA-Seq analyses reveal the molecular pathway and genes associated with heading type in Chinese cabbage. *Front. Genet.* 8. doi: 10.3389/fgene.2017.00176

Han, Y., Han, S., Ban, Q., He, Y., Jin, M., and Rao, J. (2017). Overexpression of persimmon *DkXTH1* enhanced tolerance to abiotic stress and delayed fruit softening in transgenic plants. *Plant Cell Rep.* 36, 583–596. doi: 10.1007/s00299-017-2105-4

Han, X., Tang, Q., Xu, L., Guan, Z., Tu, J., Yi, B., et al. (2022). Genome-wide detection of genotype environment interactions for flowering time in *Brassica napus*. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.1065766

Haq, T., Ali, A., Nadeem, S. M., Maqbool, M. M., and Ibrahim, M. (2014). Performance of canola cultivars under drought stress induced by withholding irrigation at different growth stages. *Soil Environ.* 33, 43–50.

Hasan, M. M., Liu, X. D., Waseem, M., Guang-Qian, Y., Alabdallah, N. M., Jahan, M. S., et al. (2022). ABA activated SnRK2 kinases: An emerging role in plant growth and physiology. *Plant Signal. Behav.* 17, 2071024. doi: 10.1080/15592324.2022.2071024

Hatzig, S. V., Nuppenau, J. N., Snowdon, R. J., and Schießl, S. V. (2018). Drought stress has transgenerational effects on seeds and seedlings in winter oilseed rape (*Brassica napus* L.). *BMC Plant Biol.* 18, 1–13. doi: 10.1186/s12870-018-1531-y

Helal, M. M. U., Gill, R. A., Tang, M., Yang, L., Hu, M., Yang, L., et al. (2021). SNP- and haplotype-based GWAS of flowering-related traits in *Brassica napus*. *Plants* 10, 2475. doi: 10.3390/plants10112475

Hiraoka, K., Yamaguchi, A., Abe, M., and Araki, T. (2013). The florigen genes FT and TSF modulate lateral shoot outgrowth in *Arabidopsis thaliana*. *Plant Cell Physiol.* 54, 352–368. doi: 10.1093/pcp/pcs168

Hong, M. J., Kim, D. Y., and Seo, Y. W. (2014). Cullin, a component of the SCF complex, interacts with TaRMD5 during wheat spike development. *Biol. Plant* 58, 218–230. doi: 10.1007/s10535-013-0383-4

Hu, J., Chen, B., Zhao, J., Zhang, F., Xie, T., Xu, K., et al. (2022). Genomic selection and genetic architecture of agronomic traits during modern rapeseed breeding. *Nat. Genet.* 54, 694–704. doi: 10.1038/s41588-022-01055-6

Huang, X., and Han, B. (2014). Natural variations and genome-wide association studies in crop plants. *Annu. Rev. Plant Boil.* 65, 531–551. doi: 10.1146/annurev-arplant-050213-035715

Hubisz, M. J., Falush, D., Stephens, M., and Pritchard, J. K. (2009). Inferring weak population structure with the assistance of sample group information. *Mol. Ecol. Resour.* 9, 1322–1332. doi: 10.1111/j.1755-0998.2009.02591.x

Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., Von Mering, C., et al. (2017). Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol. Biol. Evol.* 34, 2115–2122. doi: 10.1093/molbev/msx148

Iqbal, N., Khan, N. A., Ferrante, A., Trivellini, A., Francini, A., and Khan, M. I. R. (2017). Ethylene role in plant growth, development and senescence: interaction with other phytohormones. *Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.00475

Julkowska, M. M., McLoughlin, F., Galvan-Ampudia, C. S., Rankenberg, J. M., Kawa, D., Klimecka, M., et al. (2015). Identification and functional characterization of the *Arabidopsis* Snf 1-related protein kinase SnRK 2.4 phosphatidic acid-binding domain. *Plant Cell Environ.* 38, 614–624. doi: 10.1111/pce.12421

Kamiyama, Y., Hirotani, M., Ishikawa, S., Minegishi, F., Katagiri, S., Rogan, C. J., et al. (2021). *Arabidopsis* group C Raf-like protein kinases negatively regulate abscisic acid signaling and are direct substrates of SnRK2. *Proc. Natl. Acad. Sci.* 118, e2100073118. doi: 10.1073/pnas.2100073118

Khan, S., Anwar, S., Kuai, J., Noman, A., Shahid, M., Din, M., et al. (2018). Alteration in yield and oil quality traits of winter rapeseed by lodging at different planting density and nitrogen rates. *Sci. Rep.* 8, 1–12. doi: 10.1038/s41598-017-18734-8

Khanzada, H., Wassan, G. M., He, H., Mason, A. S., Keerio, A. A., Khanzada, S., et al. (2020). Differentially evolved drought stress indices determine the genetic variation of *Brassica napus* at seedling traits by genome-wide association mapping. *J. Adv. Res.* 24, 447–461. doi: 10.1016/j.jare.2020.05.019

Kirkegaard, J. A., Lilley, J. M., Brill, R. D., Ware, A. H., and Walela, C. K. (2018). The critical period for yield and quality determination in canola (*Brassica napus* L.). *Field Crops Res.* 222, 180–188. doi: 10.1016/j.fcr.2018.03.018

Klepikova, A. V., Kasianov, A. S., Gerasimov, E. S., Logacheva, M. D., and Penin, A. A. (2016). A high resolution map of the *Arabidopsis thaliana* developmental transcriptome based on RNA-seq profiling. *Plant J.* 88, 1058–1070. doi: 10.1111/tpj.13312

Koh, J. C., Barbulescu, D. M., Norton, S., Redden, B., Salisbury, P. A., Kaur, S., et al. (2017). A multiplex PCR for rapid identification of *Brassica* species in the triangle of U. *Plant Methods* 13, 1–8. doi: 10.1186/s13007-017-0200-8

Korber, N., Bus, A., Li, J., Parkin, I. A., Wittkop, B., Snowdon, R. J., et al. (2016). Agronomic and seed quality traits dissected by genome-wide association mapping in *Brassica napus*. *Front. Plant Sci.* 7. doi: 10.3389/fpls.2016.00386

Krishnaswamy, S., Verma, S., Rahman, M. H., and Kav, N. N. (2011). Functional characterization of four APETALA2-family genes (*RAP2.6*, *RAP2.6L*, *DREB19* and *DREB26*) in *Arabidopsis*. *Plant Mol. Biol.* 75, 107–127. doi: 10.1007/s11103-010-9711-7

Kumar, R., Arya, G. C., and Bisht, N. C. (2014). Differential expression and interaction specificity of the heterotrimeric G-protein family in *Brassica nigra* reveal their developmental-and condition-specific roles. *Plant Cell Physiol.* 55, 1954–1968. doi: 10.1093/pcp/pcu126

Kyung, J., Jeon, M., and Lee, I. (2022). Recent advances in the chromatin-based mechanism of *FLOWERING LOCUS C* repression through autonomous pathway genes. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.964931

Lai, C. P., Huang, L. M., Chen, L. F. O., Chan, M. T., and Shaw, J. F. (2017). Genome-wide analysis of GDSL-type esterases/lipases in *Arabidopsis*. *Plant Mol. Biol.* 95, 181–197. doi: 10.1007/s11103-017-0648-y

Lee, J., and Lee, I. (2010). Regulation and function of *SOC1*, a flowering pathway integrator. *J. Exp. Bot.* 61, 2247–2254. doi: 10.1093/jxb/erq098

Li, L., Garsamo, M., Yuan, J., Wang, X., Lam, S. H., Varala, K., et al. (2022). CAND1 is required for pollen viability in *Arabidopsis thaliana*—a test of the adaptive exchange hypothesis. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.866086

Li, Y., Tao, H., Zhang, B., Huang, S., and Wang, P. (2018). Timing of water deficit limits maize kernel setting in association with changes in the source-flow-sink relationship. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.01326

Li, Y., Zhang, L., Hu, S., Zhang, J., Wang, L., Ping, X., et al. (2021). Transcriptome and proteome analyses of the molecular mechanisms underlying changes in oil storage under drought stress in. *Brassica napus L. Gcb Bioenergy* 13, 1071–1086. doi: 10.1111/gcbb.12833

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

Liu, J., Liu, J., Wang, H., Khan, A., Xu, Y., Hou, Y., et al. (2023). Genome wide identification of GDSL gene family explores a novel *GhirGDSL26* gene enhancing

drought stress tolerance in cotton. *BMC Plant Biol.* 23, 1–18. doi: 10.1186/s12870-022-04001-0

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

Liu, L., Tong, H., Xiao, Y., Che, R., Xu, F., Hu, B., et al. (2015). Activation of *Big Grain1* significantly improves grain size by regulating auxin transport in rice. *Proc. Natl. Acad. Sci.* 112, 11102–11107. doi: 10.1073/pnas.1512748112

Liu, H., Wang, J., Zhang, B., Yang, X., Hammond, J. P., Ding, G., et al. (2021). Genome-wide association study dissects the genetic control of plant height and branch number in response to low-phosphorus stress in *Brassica napus*. *Ann. Bot.* 128, 919–930. doi: 10.1093/aob/mcab115

Liu, L., Zhao, L., Chen, P., Cai, H., Hou, Z., Jin, X., et al. (2020b). ATP binding cassette transporters *ABCG1* and *ABCG16* affect reproductive development via auxin signaling in *Arabidopsis*. *Plant J.* 102, 1172–1186. doi: 10.1111/tpj.14690

Liu, X., Zhao, C., Yang, L., Zhang, Y., Wang, Y., Fang, Z., et al. (2020a). Genome-wide identification, expression profile of the TIFY gene family in *Brassica oleracea* var. *capitata*, and their divergent response to various pathogen infections and phytohormone treatments. *Genes* 11, 127. doi: 10.3390/genes11020127

Lo, S. F., Cheng, M. L., Hsing, Y. I. C., Chen, Y. S., Lee, K. W., Hong, Y. F., et al. (2020). *Rice Big Grain 1* promotes cell division to enhance organ development, stress tolerance and grain yield. *J. Plant Biotechnol.* 18, 1969–1983. doi: 10.1111/pbi.13357

Lu, K., Peng, L., Zhang, C., Lu, J., Yang, B., Xiao, Z., et al. (2017). Genome-wide association and transcriptome analyses reveal candidate genes underlying yield-determining traits in *Brassica napus*. *Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.00206

Lu, K., Xiao, Z., Jian, H., Peng, L., Qu, C., Fu, M., et al. (2016). A combination of genome-wide association and transcriptome analysis reveals candidate genes controlling harvest index-related traits in *Brassica napus*. *Sci. Rep.* 6, 36452. doi: 10.1038/srep36452

Luo, X., Ding, Y., Zhang, L., Yue, Y., Snyder, J. H., Ma, C., et al. (2017). Genomic prediction of genotypic effects with epistasis and environment interactions for yield-related traits of rapeseed (*Brassica napus* L.). *Front. Genet.* 8. doi: 10.3389/fgene.2017.00015

Lush, J. L. (1949). "Heritability of quantitative characters in farm animals," in *Heritability of Quantitative Characters in Farm Animals* (CABI, Wallingford, UK), 356–375. doi: 10.1111/j.1601-5223.1949.tb03347.x

Ma, X., Chen, Y., Liu, M., Xue, X., Zhang, X., Xu, L., et al. (2022). Genome-wide analysis of the XTH gene family and functional analysis of *DlXTH23.5/25* during early longan somatic embryogenesis. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.1043464

Ma, M., Wang, Q., Li, Z., Cheng, H., Li, Z., Liu, X., et al. (2015). Expression of *TaCYP78A3*, a gene encoding cytochrome P450 CYP78A3 protein in wheat (*Triticum aestivum* L.), affects seed size. *Plant J.* 83, 312–325. doi: 10.1111/tpj.12896

Ma, J. Q., Xu, W., Xu, F., Lin, A., Sun, W., Jiang, H. H., et al. (2020). Differential alternative splicing genes and isoform regulation networks of rapeseed (*Brassica napus* L.) infected with *Sclerotinia sclerotiorum*. *Genes* 11, 784. doi: 10.3390/genes11070784

Ma, R., Yuan, H., An, J., Hao, X., and Li, H. (2018). A *Gossypium hirsutum* GDSL lipase/hydrolase gene (*GhGLIP*) appears to be involved in promoting seed growth in *Arabidopsis*. *PloS One* 13, e0195556. doi: 10.1371/journal.pone.0195556

Marguerat, S., and Bahler, J. (2010). RNA-seq: from technology to biology. *Cell. Mol. Life Sci.* 67, 569–579. doi: 10.1007/s00018-009-0180-6

Marjanovic-Jeromela, A., Terzic, S., Jankulovska, M., Zoric, M., Kondic-Spika, A., Jockovic, M., et al. (2019). Dissection of year related climatic variables and their effect on winter rapeseed (*Brassica napus* L.) development and yield. *Agronomy* 9, 517. doi: 10.3390/agronomy9090517

Martinez, D. E., Borniego, M. L., Battchikova, N., Aro, E. M., Tyystjärvi, E., and Guiamet, J. J. (2015). SASP, a Senescence-Associated Subtilisin Protease, is involved in reproductive development and determination of silique number in *Arabidopsis*. *J. Exp. Bot.* 66, 161–174. doi: 10.1093/jxb/eru409

Marwede, V., Schierholt, A., Mollers, C., and Becker, H. C. (2004). Genotype × environment interactions and heritability of tocopherol contents in canola. *Crop Sci.* 44, 728–731. doi: 10.2135/cropsci2004.7280

Matar, S., Kumar, A., Holtgrawe, D., Weisshaar, B., and Melzer, S. (2021). The transition to flowering in winter rapeseed during vernalization. *Plant Cell Environ.* 44, 506–518. doi: 10.1111/pce.13946

McGill, R., Tukey, J. W., and Larsen, W. A. (1978). Variations of box plots. *Am. Stat.* 32, 12–16. doi: 10.1080/00031305.1978.10479236

Menendez, Y. C., Sanchez, D. H., Snowdon, R. J., Rondanini, D. P., and Botto, J. F. (2021). Unraveling the impact on agronomic traits of the genetic architecture underlying plant-density responses in canola. *J. Exp. Bot.* 72, 5426–5441. doi: 10.1093/jxb/erab191

Meyer, R. S., and Purugganan, M. D. (2013). Evolution of crop species: genetics of domestication and diversification. *Nat. Rev. Genet.* 14, 840–852. doi: 10.1038/nrg3605

Miedes, E., Suslov, D., Vandenbussche, F., Kenobi, K., Ivakov, A., Straeten., V. D. D., et al. (2013). Xyloglucan endotransglucosylase/hydrolase (XTH) overexpression affects growth and cell wall mechanics in etiolated *Arabidopsis* hypocotyls. *J. Exp. Bot.* 64, 2481–2497. doi: 10.1093/jxb/ert107

Murray, M. G., and Thompson, W. (1980). Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Res.* 8, 4321–4326. doi: 10.1093/nar/8.19.4321

Nowosad, K., Liersch, A., Popławska, W., and Bocianowski, J. (2016). Genotype by environment interaction for seed yield in rapeseed (*Brassica napus* L.) using additive main effects and multiplicative interaction model. *Euphytica* 208, 187–194. doi: 10.1007/s10681-015-1620-z

Ozer, H. (2003). Sowing date and nitrogen rate effects on growth, yield and yield components of two summer rapeseed cultivars. *Eur. J. Agron.* 19, 453–463. doi: 10.1016/S1161-0301(02)00136-3

Pagnussat, L., Burbach, C., Baluska, F., and Canal, L. D. L. (2012). An extracellular lipid transfer protein is relocalized intracellularly during seed germination. *J. Exp. Bot.* 63, 6555–6563. doi: 10.1093/jxb/ers311

Pal, L., Sandhu, S. K., Bhatia, D., and Sethi, S. (2021). Genome-wide association study for candidate genes controlling seed yield and its components in rapeseed (*Brassica napus* subsp. *napus*). *Physiol. Mol. Biol.* 27, 1933–1951. doi: 10.1007/s12298-021-01060-9

Pandey, S. (2020). Plant receptor-like kinase signaling through heterotrimeric G-proteins. *J. Exp. Bot.* 71, 1742–1751. doi: 10.1093/jxb/eraa016

Paradis, E., and Schliep, K. (2018). Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35, 526–528. doi: 10.1093/bioinformatics/bty633

Park, H. Y., Lee, H. T., Lee, J. H., and Kim, J. K. (2019). Arabidopsis *U2AF65* regulates flowering time and the growth of pollen tubes. *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.00569

Patel, R. K., and Jain, M. (2012). NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PloS One* 7, e30619. doi: 10.1371/journal.pone.0030619

Perrot, T., Pauly, M., and Ramirez, V. (2022). Emerging roles of β-glucanases in plant development and adaptative responses. *Plants* 11, 1119. doi: 10.3390/plants11091119

Qian, L., Hickey, L. T., Stahl, A., Werner, C. R., Hayes, B., Snowdon, R. J., et al. (2017). Exploring and harnessing haplotype diversity to improve yield stability in crops. *Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.01534

Qin, M., Song, J., Guo, N., Zhang, M., Zhu, Y., Huang, Z., et al. (2022). Genome-wide association analyses reveal candidate genes controlling harvest index and related agronomic traits in *Brassica napus* L. *Agronomy* 12, 814. doi: 10.3390/agronomy12040814

Raboanatahiry, N., Chao, H., Dalin, H., Pu, S., Yan, W., Yu, L., et al. (2018). QTL alignment for seed yield and yield related traits in *Brassica napus*. *Front. Plant Sci.* 9, 1127. doi: 10.3389/fpls.2018.01127

Rahman, M., Guo, Q., Baten, A., Mauleon, R., Khatun, A., Liu, L., et al. (2021). Shotgun proteomics of *Brassica rapa* seed proteins identifies vicilin as a major seed storage protein in the mature seed. *PloS One* 16, e0253384. doi: 10.1371/journal.pone.0253384

Raman, H., Raman, R., Mathews, K., Diffey, S., and Salisbury, P. (2020). QTL mapping reveals genomic regions for yield based on an incremental tolerance index to drought stress and related agronomic traits in canola. *Crop Pasture Sci.* 71, 562–577. doi: 10.1071/CP20046

Raman, H., Raman, R., Qiu, Y., Yadav, A. S., Sureshkumar, S., Borg, L., et al. (2019). GWAS hints at pleiotropic roles for *FLOWERING LOCUS T* in flowering time and yield-related traits in canola. *BMC Genom.* 20, 1–18. doi: 10.1186/s12864-019-5964-y

Rani, V., Sengar, R. S., Garg, S. K., Mishra, P., and Shukla, P. K. (2023). Physiological and molecular role of strigolactones as plant growth regulators: A review. *Mol. Biotechnol.* 1–27. doi: 10.1007/s12033-023-00694-2

R Core Team (2018). *R: A language and environment for statistical computing* (Vienna, Austria: R Foundation for Statistical Computing).

Ren, J. P., Dickson, M. H., and Earle, E. D. (2000). Improved resistance to bacterial soft rot by protoplast fusion between *Brassica rapa* and *B. oleracea*. *Theor. Appl. Genet.* 100, 810–819. doi: 10.1007/s001220051356

Roy Choudhury, S., Riesselman, A. J., and Pandey, S. (2014). Constitutive or seed-specific overexpression of Arabidopsis *G-protein γ subunit 3* (*AGG3*) results in increased seed and oil production and improved stress tolerance in *Camelina sativa*. *Plant Biotechnol. J.* 12, 49–59. doi: 10.1111/pbi.12115

Salami, M., Heidari, B., Batley, J., Wang, J., Tan, X. L., Richards, C., et al. (2024). Integration of genome-wide association studies, metabolomics, and transcriptomics reveals phenolic acid-and flavonoid-associated genes and their regulatory elements under drought stress in rapeseed flowers. *Front. Plant Sci.* 14. doi: 10.3389/fpls.2023.1249142

Samanta, I., Roy, P. C., Das, E., Mishra, S., and Chowdhary, G. (2023). Plant peroxisomal polyamine oxidase: A ubiquitous enzyme involved in abiotic stress tolerance. *Plants* 12, 652. doi: 10.3390/plants12030652

Secchi, M. A., Fernandez, J. A., Stamm, M. J., Durrett, T., Prasad, P. V., Messina, C. D., et al. (2023). Effects of heat and drought on canola (*Brassica napus* L.) yield, oil, and protein: A meta-analysis. *Field Crops Res.* 293, 108848. doi: 10.1016/j.fcr.2023.108848

Seleiman, M. F., Al-Suhaibani, N., Ali, N., Akmal, M., Alotaibi, M., Refay, Y., et al. (2021). Drought stress impacts on plants and different approaches to alleviate its adverse effects. *Plants* 10, 259. doi: 10.3390/plants10020259

Shah, L., Sohail, A., Ahmad, R., Cheng, S., Cao, L., and Wu, W. (2022). The roles of MADS-Box genes from root growth to maturity in *Arabidopsis* and rice. *Agronomy* 12, 582. doi: 10.3390/agronomy12030582

Shahzad, A., Qian, M., Sun, B., Mahmood, U., Li, S., Fan, Y., et al. (2021). Genome-wide association study identifies novel loci and candidate genes for drought stress tolerance in rapeseed. *Oil Crop Sci.* 6, 12–22. doi: 10.1016/j.ocsci.2021.01.002

Shen, G., Sun, W., Chen, Z., Shi, L., Hong, J., and Shi, J. (2022). Plant GDSL esterases/lipases: Evolutionary, physiological and molecular functions in plant development. *Plants* 11, 468. doi: 10.3390/plants11040468

Sheng, X. G., Zhao, Z. Q., Wang, J. S., Yu, H. F., Shen, Y. S., Zeng, X. Y., et al. (2019). Genome wide analysis of MADS-box gene family in *Brassica oleracea* reveals conservation and variation in flower development. *BMC Plant Biol.* 19, 1–15. doi: 10.1186/s12870-019-1717-y

Shi, J., Li, R., Zou, J., Long, Y., and Meng, J. (2011). A dynamic and complex network regulates the heterosis of yield-correlated traits in rapeseed (*Brassica napus* L.). *PloS One* 6, e21645. doi: 10.1371/journal.pone.0021645

Shi, L., Song, J., Guo, C., Wang, B., Guan, Z., Yang, P., et al. (2019). A CACTA-like transposable element in the upstream region of *BnaA9.CYP78A9* acts as an enhancer to increase silique length and seed weight in rapeseed. *Plant J.* 98, 524–539. doi: 10.1111/tpj.14236

Shi, Y. Z., Zhu, X. F., Miller, J. G., Gregson, T., Zheng, S. J., and Fry, S. C. (2015). Distinct catalytic capacities of two aluminium-repressed *Arabidopsis thaliana* xyloglucan endotransglucosylase/hydrolases, XTH15 and XTH31, heterologously produced in *Pichia*. *Phytochem.* 112, 160–169. doi: 10.1016/j.phytochem.2014.09.020

Shikata, M., Matsuda, Y., Ando, K., Nishii, A., Takemura, M., Yokota, A., et al. (2004). Characterization of *Arabidopsis ZIM*, a member of a novel plant-specific GATA factor gene family. *J. Exp. Bot.* 55, 631–639. doi: 10.1093/jxb/erh078

Singh, P., and Mukhopadhyay, K. (2021). Comprehensive molecular dissection of TIFY Transcription factors reveal their dynamic responses to biotic and abiotic stress in wheat (*Triticum aestivum* L.). *Sci. Rep.* 11, 9739. doi: 10.1038/s41598-021-87722-w

Soppe, W. J., Vinegra, D. L. T. N., and Albani, M. C. (2021). The diverse roles of FLOWERING LOCUS C in annual and perennial Brassicaceae species. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.627258

Sotelo-Silveira, M., Cucinotta, M., Chauvin, A. L., Chavez Montes, R. A., Colombo, L., Marsch-Martinez, N., et al. (2013). Cytochrome P450 *CYP78A9* is involved in Arabidopsis reproductive development. *Plant Physiol.* 162, 779–799. doi: 10.1104/pp.113.218214

Stewart, T. M., Dunston, T. T., Woster, P. M., and Casero, R. A. (2018). Polyamine catabolism and oxidative damage. *J. Biol. Chem.* 293, 18736–18745. doi: 10.1074/jbc.TM118.003337

Su, W., Wang, L., Lei, J., Chai, S., Liu, Y., Yang, Y., et al. (2017). Genome-wide assessment of population structure and genetic diversity and development of a core germplasm set for sweet potato based on specific length amplified fragment (SLAF) sequencing. *PloS One* 12, e0172066. doi: 10.1371/journal.pone.0172066

Sun, X., Cahill, J., Van Hautegem, T., Feys, K., Whipple, C., Novak, O., et al. (2017). Altered expression of maize *PLASTOCHRON1* enhances biomass and seed yield by extending cell division duration. *Nat. Commun.* 8, 14752. doi: 10.1038/ncomms14752

Sun, C., Wang, B., Wang, X., Hu, K., Li, K., Li, Z., et al. (2016a). Genome-wide association study dissecting the genetic architecture underlying the branch angle trait in rapeseed *(Brassica napus* L.). *Sci. Rep.* 6, 33673. doi: 10.1038/srep33673

Sun, C., Wang, B., Yan, L., Hu, K., Liu, S., Zhou, Y., et al. (2016b). Genome-wide association study provides insight into the genetic control of plant height in rapeseed (*Brassica napus* L.). *Front. Plant Sci.* 7. doi: 10.3389/fpls.2016.01102

Supek, F., Bosnjak, M., Skunca, N., and Smuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. *PloS One* 6, e21800. doi: 10.1371/journal.pone.0021800

Takeuchi, H., and Higashiyama, T. (2012). A species-specific cluster of defensin-like genes encodes diffusible pollen tube attractants in *Arabidopsis*. *PloS Biol.* 10, e1001449. doi: 10.1371/journal.pbio.1001449

Tao, Y., Wan, J. X., Liu, Y. S., Yang, X. Z., Shen, R. F., and Zhu, X. F. (2022). The NAC transcription factor ANAC017 regulates aluminum tolerance by regulating the cell wall-modifying genes. *Plant Physiol.* 189, 2517–2534. doi: 10.1093/plphys/kiac197

Tran, L. T., Sugimoto, K., Kasozi, M., Mitalo, O. W., and Ezura, H. (2023). Pollination, pollen tube growth, and fertilization independently contribute to fruit set and development in tomato. *Front. Plant Sci.* 14. doi: 10.3389/fpls.2023.1205816

Tsai, Y. C., Weir, N. R., Hill, K., Zhang, W. J., Kim, H. J., Shiu, S. H., et al. (2012). Characterization of genes involved in cytokinin signaling and metabolism from rice. *Plant Physiol.* 158, 1666–1684. doi: 10.1104/pp.111.192765

Vanholme, B., Grunewald, W., Bateman, A., Kohchi, T., and Gheysen, G. (2007). The tify family previously known as ZIM. *Trends Plant Sci.* 12, 239–244. doi: 10.1016/j.tplants.2007.04.004

Vollrath, P., Chawla, H. S., Schiessl, S. V., Gabur, I., Lee, H., Snowdon, R. J., et al. (2021). A novel deletion in *FLOWERING LOCUS T* modulates flowering time in winter oilseed rape. *Theor. Appl. Genet.* 134, 1217–1231. doi: 10.1007/s00122-021-03768-4

Wang, R., Li, M., Wu, X., and Wang, J. (2019). The gene structure and expression level changes of the *GH3* gene family in *Brassica napus* relative to its diploid ancestors. *Genes* 10, 58. doi: 10.3390/genes10010058

Wang, X. B., Li, Y. H., Zhang, H. W., Sun, G. H., Zhang, W. M., and Qiu, L. J. (2015b). Evolution and association analysis of *GmCYP78A10* gene with seed size/weight and pod number in soybean. *Mol. Biol. Rep.* 42, 489–496. doi: 10.1007/s11033-014-3792-3

Wang, L., Liang, X., Dou, S., Yi, B., Fu, T., Ma, C., et al. (2023). Two aspartic proteases, BnaAP36s and BnaAP39s, regulate pollen tube guidance in *Brassica napus*. *Mol. Breed.* 43, 27. doi: 10.1007/s11032-023-01377-1

Wang, X., Zhou, W., Lu, Z., Ouyang, Y., and Yao, J. (2015a). A lipid transfer protein, OsLTPL36, is essential for seed development and seed quality in rice. *Plant Sci.* 239, 200–208. doi: 10.1016/j.plantsci.2015.07.016

Whittaker, C., and Dean, C. (2017). The *FLC* locus: a platform for discoveries in epigenetics and adaptation. *Annu. Rev. Cell Dev. Biol.* 33, 555–575. doi: 10.1146/annurev-cellbio-100616-060546

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis* (New York, NY, USA: Springer). doi: 10.1007/978-3-319-24277-4

Wozniak, E., Waszkowska, E., Zimny, T., Sowa, S., and Twardowski, T. (2019). The rapeseed potential in Poland and Germany in the context of production, legislation, and intellectual property rights. *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.01423

Wu, S., Fang, C., Li, Z., Wang, Y., Pan, S., Wu, Y., et al. (2022a). ATP-Binding Cassette G Transporters and their multiple roles especially for male fertility in *Arabidopsis*, Rice and Maize. *Int. J. Mol. Sci.* 23, 9304. doi: 10.3390/ijms23169304

Wu, J., Liu, W., Jahan, M. S., Shu, S., Sun, J., and Guo, S. (2022b). Characterization of polyamine oxidase genes in cucumber and roles of *CsPAO3* in response to salt stress. *Environ. Exp. Bot.* 194, 104696. doi: 10.1016/j.envexpbot.2021.104696

Xiao, Y., Liu, H., Wu, L., Warburton, M., and Yan, J. (2017). Genome-wide association studies in maize: Praise and stargaze. *Mol. Plant* 10, 359–374. doi: 10.1016/j.molp.2016.12.008

Xiao, Y., Liu, D., Zhang, G., Gao, S., Liu, L., Xu, F., et al. (2019). *Big Grain3*, encoding a purine permease, regulates grain size via modulating cytokinin transport in rice. *J. Integr. Plant Biol.* 61, 581–597. doi: 10.1111/jipb.12727

Xie, Y., Liu, Y., Ma, M., Zhou, Q., Zhao, Y., Zhao, B., et al. (2020). *Arabidopsis* FHY3 and FAR1 integrate light and strigolactone signaling to regulate branching. *Nat. Commun.* 11, 1955. doi: 10.1038/s41467-020-15893-7

Xu, F., Fang, J., Ou, S., Gao, S., Zhang, F., Du, L., et al. (2015). Variations in *CYP78A13* coding region influence grain size and yield in rice. *Plant. Cell Environ.* 38, 800–811. doi: 10.1111/pce.12452

Xu, L., Hu, K., Zhang, Z., Guan, C., Chen, S., Hua, W., et al. (2016). Genome-wide association study reveals the genetic architecture of flowering time in rapeseed (*Brassica napus* L.). *DNA Res.* 23, 43–52. doi: 10.1093/dnares/dsv035

Xu, R., Li, N., and Li, Y. (2019). Control of grain size by G protein signaling in rice. *J. Integr. Plant Biol.* 61, 533–540. doi: 10.1111/jipb.12769

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

Yaqoob, A., Shahid, A. A., Imran, A., Sadaqat, S., Liaqat, A., and Rao, A. Q. (2020). Dual functions of Expansin in cell wall extension and compression during cotton fiber development. *Biol.* 75, 2093–2101. doi: 10.2478/s11756-020-00514-x

Yarnia, M., Arabifard, N., Khoei, F. R., and Zandi, P. (2011). Evaluation of drought tolerance indices among some winter rapeseed cultivars. *Afr. J. Biotechnol.* 10, 10914–10922. doi: 10.5897/AJB

Yeon, M. H., Park, C. H., Lee, Y. E., Roh, J., and Kim, S. K. (2021). Seed-specifically overexpressed *Arabidopsis cytochrome P450 85A2* promotes vegetative and reproductive growth and development of *Arabidopsis thaliana*. *J. Plant Biol.* 65. doi: 10.3389/fpls.2021.639508

Yosefi, M., and Heidari, H. (2022). Evaluation of wheat tolerance during germination and early growth stages to detergent-contaminated water. *Tenside Surfactants Deterg* 1), 95–103. doi: 10.1515/tsd-2021-2380

Yu, Z., Jia, D., and Liu, T. (2019). Polyamine oxidases play various roles in plant development and abiotic stress tolerance. *Plants* 8, 184. doi: 10.3390/plants8060184

Yu, L. H., Wu, J., Zhang, Z. S., Miao, Z. Q., Zhao, P. X., Wang, Z., et al. (2017). *Arabidopsis* MADS-box transcription factor AGL21 acts as environmental surveillance of seed germination by regulating *ABI5* expression. *Mol. Plant* 10, 834–845. doi: 10.1016/j.molp.2017.04.004

Zhang, C., Gong, R., Zhong, H., Dai, C., Zhang, R., Dong, J., et al. (2023a). Integrated multi-locus genome-wide association studies and transcriptome analysis for seed yield and yield-related traits in *Brassica napus*. *Front. Plant Sci.* 14. doi: 10.3389/fpls.2023.1153000

Zhang, J., Mason, A. S., Wu, J., Liu, S., Zhang, X., Luo, T., et al. (2015). Identification of putative candidate genes for water stress tolerance in canola (*Brassica napus*). *Front. Plant Sci.* 6. doi: 10.3389/fpls.2015.01058

Zhang, X., Ran, W., Zhang, J., Ye, M., Lin, S., Li, X., et al. (2020). Genome-wide identification of the Tify gene family and their expression profiles in response to biotic and abiotic stresses in tea plants (*Camellia sinensis*). *Int. J. Mol. Sci.* 21, 8316. doi: 10.3390/ijms21218316

Zhang, X., Song, J., Wang, L., Yang, Z. M., and Sun, D. (2022). Identification of a DEAD-box RNA helicase BnRH6 reveals its involvement in salt stress response in rapeseed (*Brassica napus*). *Int. J. Mol. Sci.* 24, 2. doi: 10.3390/ijms24010002

Zhang, Y., Wang, K., Wang, Z., Li, X., Li, M., Zhu, F., et al. (2023c). The lipoxygenase gene *AfLOX4* of *Amorpha fruticosa* L. is a potential regulator of drought stress tolerance pathways under saline and alkaline conditions. *Acta Physiol. Plant* 45, 72. doi: 10.1007/s11738-023-03542-7

Zhang, Y., Xu, J., Li, R., Ge, Y., Li, Y., and Li, R. (2023d). Plants' Response to abiotic stress: mechanisms and strategies. *Int. J. Mol. Sci.* 24, 10915. doi: 10.3390/ijms241310915

Zhang, L., Yang, G., Liu, P., Hong, D., Li, S., and He, Q. (2011). Genetic and correlation analysis of silique-traits in *Brassica napus* L. by quantitative trait locus mapping. *Theor. Appl. Genet.* 122, 21–31. doi: 10.1007/s00122-010-1419-1

Zhang, L., Zheng, L., Wu, J., Liu, Y., Liu, W., He, G., et al. (2023b). *OsCCRL1* is essential for phenylpropanoid metabolism in rice anthers. *Rice* 16, 1–20. doi: 10.1186/s12284-023-00628-1

Zhao, J., Long, T., Wang, Y., Tong, X., Tang, J., Li, J., et al. (2020). *RMS2* encoding a GDSL lipase mediates lipid homeostasis in anthers to determine rice male fertility. *Plant Physiol.* 182, 2047–2064. doi: 10.1104/pp.19.01487

Zhao, J., Yu, N., Ju, M., Fan, B., Zhang, Y., Zhu, E., et al. (2019). ABC transporter OsABCG18 controls the shootward transport of cytokinins and grain yield in rice. *J. Exp. Bot.* 70, 6277–6291. doi: 10.1093/jxb/erz382

Zheng, M., Peng, C., Liu, H., Tang, M., Yang, H., Li, X., et al. (2017). Genome-wide association study reveals candidate genes for control of plant height, branch initiation height and branch number in rapeseed (*Brassica napus* L.). *Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.01246

Zhou, X., Dai, L., Wang, P., Liu, Y., Xie, Z., Zhang, H., et al. (2021). Mining favorable alleles for five agronomic traits from the elite rapeseed cultivar Zhongshuang 11 by QTL mapping and integration. *Crop J.* 9, 1449–1459. doi: 10.1016/j.cj.2020.12.008

Zhou, Q., Fu, D., Mason, A. S., Zeng, Y., Zhao, C., and Huang, Y. (2014). *In silico* integration of quantitative trait loci for seed yield and yield-related traits in *Brassica napus*. *Mol. Breed.* 33, 881–894. doi: 10.1007/s11032-013-0002-2

Zhu, J., Cai, D., Wang, J., Cao, J., Wen, Y., He, J., et al. (2021). Physiological and anatomical changes in two rapeseed (*Brassica napus* L.) genotypes under drought stress conditions. *Oil Crop Sci.* 6, 97–104. doi: 10.1016/j.ocsci.2021.04.003

Zhu, J., Lei, L., Wang, W., Jiang, J., and Zhou, X. (2023). QTL mapping for seed density per silique in *Brassica napus*. *Sci. Rep.* 13, 772. doi: 10.1038/s41598-023-28066-5

Zuniga-Mayo, V. M., Banos-Bayardo, C. R., Diaz-Ramirez, D., Marsch-Martinez, N., and Folter, D. S. (2018). Conserved and novel responses to cytokinin treatments during flower and fruit development in *Brassica napus* and *Arabidopsis thaliana*. *Sci. Rep.* 8, 6836. doi: 10.1038/s41598-018-25017-3

frontiers | Frontiers in Plant Science

TYPE Methods
PUBLISHED 28 May 2024
DOI 10.3389/fpls.2024.1370618

OPEN ACCESS

EDITED BY
Manohar Chakrabarti,
The University of Texas Rio Grande Valley, United States

REVIEWED BY
Haibo Liu,
University of Massachusetts Medical School, United States
Eduardo D. Munaiz,
UniLaSalle, France

*CORRESPONDENCE
Pao-Yang Chen
paoyang@gate.sinica.edu.tw
Chung-Ju Rachel Wang
rwang@gate.sinica.edu.tw

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

RECEIVED 15 January 2024
ACCEPTED 07 May 2024
PUBLISHED 28 May 2024

CITATION
Hsieh J-WA, Lin P-Y, Wang C-T, Lee Y-J, Chang P, Lu RJ-H, Chen P-Y and Wang C-JR (2024) Establishing an optimized ATAC-seq protocol for the maize. *Front. Plant Sci.* 15:1370618. doi: 10.3389/fpls.2024.1370618

Establishing an optimized ATAC-seq protocol for the maize

Jo-Wei Allison Hsieh[1,2†], Pei-Yu Lin[1†], Chi-Ting Wang[1], Yi-Jing Lee[1], Pearl Chang[1,3], Rita Jui-Hsien Lu[1], Pao-Yang Chen[1,2*] and Chung-Ju Rachel Wang[1*]

[1]Institute of Plant and Microbial Biology, Academia Sinica, Taipei, Taiwan, [2]Genome and Systems Biology Degree Program, Academia Sinica and National Taiwan University, Taipei, Taiwan, [3]Department of Tropical Agriculture and International Cooperation/Department of Biological Science and Technology, National Pingtung University of Science and Technology, Pingtung, Taiwan

The advent of next-generation sequencing in crop improvement offers unprecedented insights into the chromatin landscape closely linked to gene activity governing key traits in plant development and adaptation. Particularly in maize, its dynamic chromatin structure is found to collaborate with massive transcriptional variations across tissues and developmental stages, implying intricate regulatory mechanisms, which highlights the importance of integrating chromatin information into breeding strategies for precise gene controls. The depiction of maize chromatin architecture using Assay for Transposase Accessible Chromatin with high-throughput sequencing (ATAC-seq) provides great opportunities to investigate cis-regulatory elements, which is crucial for crop improvement. In this context, we developed an easy-to-implement ATAC-seq protocol for maize with fewer nuclei and simple equipment. We demonstrate a streamlined ATAC-seq protocol with four key steps for maize in which nuclei purification can be achieved without cell sorting and using only a standard bench-top centrifuge. Our protocol, coupled with the bioinformatic analysis, including validation by read length periodicity, key metrics, and correlation with transcript abundance, provides a precise and efficient assessment of the maize chromatin landscape. Beyond its application to maize, our testing design holds the potential to be applied to other crops or other tissues, especially for those with limited size and amount, establishing a robust foundation for chromatin structure studies in diverse crop species.

KEYWORDS

maize ATAC-seq, chromatin accessibility, chromatin structure, ATAC-seq protocol, next-generation sequencing

1 Introduction

In eukaryotes, genomic DNA is packaged with histone proteins, forming nucleosomes that constitute the structural basis of chromatin (Kornberg and Lorch, 1999). The density of nucleosomes determines chromatin compactness and DNA accessibility, which regulate various cellular and chromosomal functions (Hsieh and Fischer, 2005). Genomic regions

with dense nucleosomes are tightly packed (i.e., "closed"), whereas nucleosome-depleted regions with exposed DNA are more accessible (i.e., "open"), so the dynamic structures of chromatin provide different levels of availability of DNA binding sites in regulatory regions to transcription factors (Wolffe, 1997). It is now evident that chromatin structure is intimately linked to the activity of underlying genes, thus influencing proper development and the ability to adapt to an ever-changing environment.

Over the last decade, the development of a wide range of methods that utilize nuclease enzymes such MNase and DNase I to target open DNA regions, combined with next-generation sequencing (NGS), has enabled genome-wide investigations of chromatin accessibility (Zhang and Pugh, 2011; Meyer and Liu, 2014). For example, MNase digests the linker DNA between nucleosomes, so subsequent NGS reads largely represent the footprints of nucleosomes, which can be used to assess nucleosome occupancy (Chodavarapu et al., 2010). Another method utilizes an optimized concentration of non-specific endonuclease DNase I that generates DNA fragments by liberating open chromatin regions. The resulting NGS reads are characterized as DNase I hypersensitive sites (DHSs), which represent accessible regions of chromatin (Valouev et al., 2011). These methods have been implemented in a number of organisms, ranging from yeast to plants and human. However, they usually require millions of cells, empirical enzymatic titrations, as well as several purification steps, rendering them challenging for reproducible evaluation of chromatin status and consequently, they are infeasible for some cell types (Song and Crawford, 2010).

An improved method for identifying accessible chromatin is the Assay for Transposase-Accessible Chromatin with high-throughput sequencing (ATAC-seq) (Buenrostro et al., 2013, 2015). This method takes advantage of the engineered prokaryotic Tn5 transposase (Tn5p) that cleaves DNA in open chromatin regions and simultaneously integrates adapters into cleavage sites *in vivo*. Due to the integrated adapters loaded by Tn5p, the ligation and additional purification steps prior to sequencing necessary in other methodologies are eliminated. Using simple library amplification steps, a very small amount of starting material has been successfully used to profile genome-wide chromatin accessibility (Buenrostro et al., 2013). Several studies have shown that 500–50,000 nuclei are sufficient for NGS analyses, considerably less than the nuclei required for other methods (Buenrostro et al., 2013; Gray et al., 2017). Thus, the small amount of cell numbers needed and its high resolution for mapping open chromatin regions make ATAC-seq an excellent tool for genome accessibility profiling.

ATAC-seq has been used to profile the chromatin structure of various animal samples (Corces et al., 2017; Gray et al., 2017; Shashikant and Ettensohn, 2019), but such analyses on plants remain more challenging (Wilkins et al., 2016; Lu et al., 2017; Maher et al., 2018). Unlike animal samples, where cells can be directly used as input for the transposition reaction immediately after lysis with a gentle detergent, plant tissues require physical disruption to release cell contents from the rigid cell walls. Thus, various methods were attempted to break down plant cell walls for plant ATAC-seq, such as grinding samples in liquid nitrogen (Galli et al., 2018; Deschamps et al., 2021; Wu et al., 2023), digesting cell walls for protoplast preparation (Dong et al., 2017; Dai et al., 2022), or disrupting plant tissues using homogenizer or razor blade (Lu et al., 2017; Ricci et al., 2019). In addition to cell walls, another obstacle in plants is the higher quantity of organelles (i.e. mitochondrial and chloroplast) compared to animal cells. As Tn5p targets not only nuclear DNA but also organellar DNA, the presence of a significant amount of organellar DNA can lead to up to 50% of sequencing reads being unusable (Montefiori et al., 2017). To purify plant nuclei from organelles, flow cytometry is commonly used, where nuclei stained with DAPI or expressing fluorescent proteins are sorted (Lu et al., 2017; Galli et al., 2018; Noshay et al., 2019; Ricci et al., 2019; Wu et al., 2023). Alternatively, Deal and Henikoff developed a method, named the isolation of nuclei tagged in specific cell types (INTACT), which involves applying streptavidin-coated magnetic beads to isolate biotin-labeled nuclei extracted from plant samples (Deal and Henikoff, 2011). Although this method allows for the purification of large numbers of nuclei, it requires generating transgenic plants, which could be time-consuming and not practical for many species, especially crops with low transformation efficiency.

Maize (*Zea mays*) is one of the most important crops globally, being widely consumed as staple food and livestock feed, as well as for its industrial uses. The maize genome is relatively large (~2.3 Gbp) and has a complex organization of interspersed genes that are separated by transposable elements (Schnable et al., 2009). In fact, 85% of genes are positioned within 1 kb of transposons nearby (West et al., 2014). How maize cells coordinate this complicated chromatin structure with the transcription program in different tissues and developmental stages is a fascinating biological question. Many studies have explored massive transcriptional variation in different maize tissues and among different maize inbred lines (Dong et al., 2012; Chen et al., 2014; Walley et al., 2016). Interestingly, epigenetic analyses—such as DNA methylation, siRNA profiling, nucleosome occupancy and histone modification profiles—have suggested that epigenetics plays regulatory roles in gene expression and various aspects of maize development (Wang et al., 2009; Gent et al., 2013; Makarevitch et al., 2013; Rodgers-Melnick et al., 2016; Hsu et al., 2017; Oka et al., 2017; Zhao et al., 2018). Its genome complexity and intricate interplay between different levels of epigenetic control make maize an interesting model. Although the maize genome has been sequenced and related epigenetic features are being actively studied, its chromatin accessibility profile remains less investigated.

ATAC-seq has been employed in maize genome research, where most ATAC-seq results derive from tens of thousands of nuclei isolated by flow cytometry or prepared from mesophyll protoplasts (Dong et al., 2017; Galli et al., 2018; Noshay et al., 2019; Ricci et al., 2019; Crisp et al., 2020; Wu et al., 2023). To make the ATAC-seq method more applicable easily, we optimize critical steps in this study and develop a robust and efficient methodology for ATAC-seq analysis in maize. We meticulously evaluated four key components (as illustrated in Figure 1), including (1) the isolation of maize nuclei; (2) the optimization of nuclei number and Tn5p transposition efficiency; (3) determination of library amplification conditions; and (4) the assessment of library quality. All tested parameters and their corresponding results are presented, and the

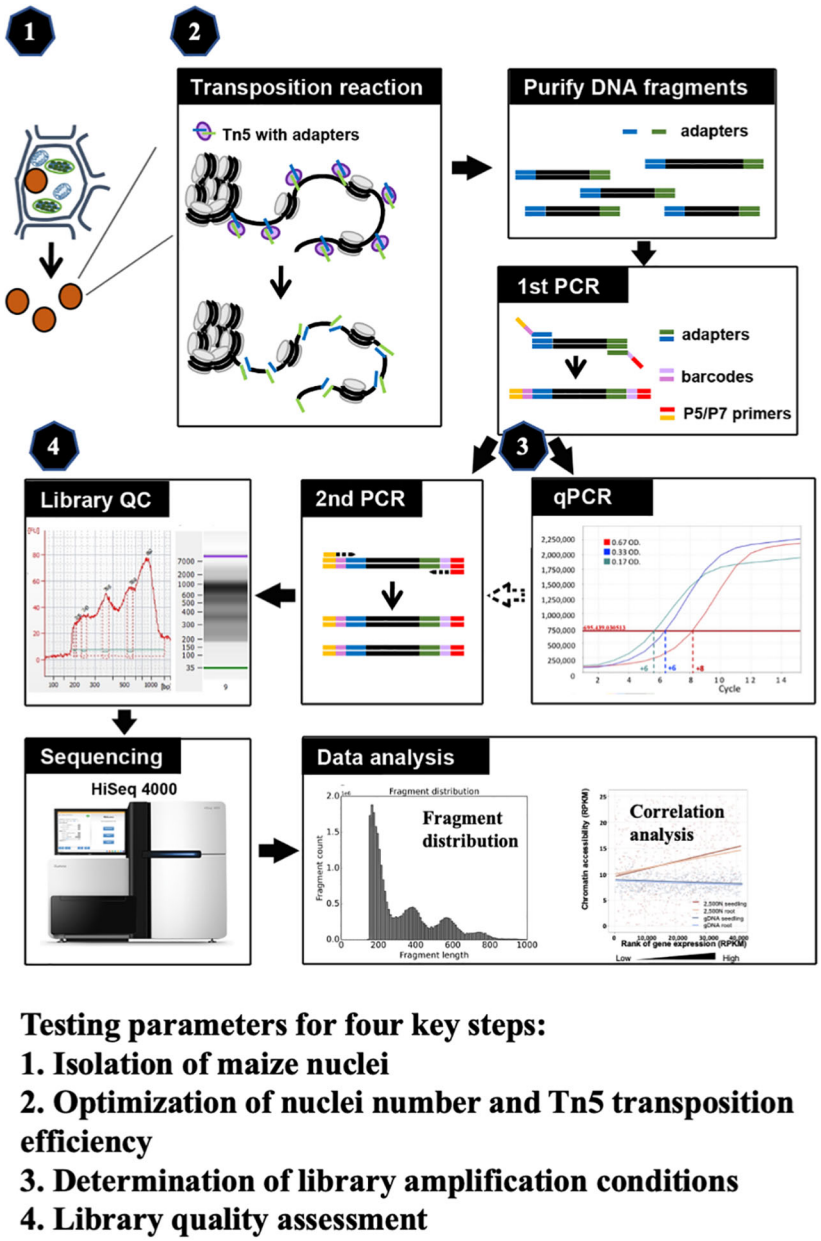


FIGURE 1
The schematic of maize ATAC-seq protocol, highlighting four key steps that were optimized: (1) isolation of maize nuclei with reduced organelle contamination; (2) optimization of nuclei number and Tn5p transposition efficiency to mitigate over-reaction and increase effectiveness; (3) determination of library amplification conditions to minimize excessive PCR duplications; and (4) library quality assessment prior to the costly and time-consuming sequencing process.

optimized protocol is described in the step-by-step method details. Finally, we validated our ATAC-seq results by examining fragment size distribution, key metrics, and the correlation with RNA-seq data. Through optimization of critical steps and quality assessments, this refined protocol provides an efficient assessment of the maize chromatin landscape and can be adapted for other plant species.

2 Materials and equipment

2.1 Biological materials

- 10-day-old maize B73 seedlings

2.2 Reagents

2.2.1 Chemicals and enzymes

- Percoll (Sigma-Aldrich, Cat#P7828)
- 3-(N-morpholino) propanesulfonic acid, 4-morpholinepropanesulfonic acid (MOPS) (Merck, Cas#1132612)
- 10N NaOH (Merck, Cat# 1310732)
- 5M NaCl (Merck, Cat#7647145)
- 3M KCl (Merck, Cat# 7447407)
- 500 mM EDTA (Invitrogen, Cat#15575020)
- 100 mM EGTA (Sigma-Aldrich, Cas#13368133)
- 400 mM spermine (Sigma-Aldrich, Cat#85590)
- 400 mM spermidine (Sigma-Aldrich, Cas#124209)
- cOmplete™, EDTA-free Protease Inhibitor Cocktail (Roche, Cat#04693159001)
- Sucrose (Sigma-Aldrich, Cas#57501)
- Tris base (Sigma-Aldrich, Cat#77861)
- 100 mM MgCl2 (Invitrogen, Cat#AM9530G)
- Triton X-100 (Sigma-Aldrich, Cat#9036195)
- 4',6-diamidino-2-phenylindole (DAPI)
- Ethanol (Merck, Cat#1.00983.2500)
- UltraPure DNase/RNase-Free Distilled Water (Thermo Fisher Scientific, Cat#10977015)
- Buffer EB (Qiagen, Part#19086)

2.2.2 Critical commercial assays

- Illumina Tagment DNA TDE1 Enzyme and Buffer Kits (Illumina, Cat#20034197)
- Qubit dsDNA BR Assay Kit (Thermo Fisher Scientific, Cat#Q32853)
- NEBNext High-Fidelity 2x PCR Master Mix (New England Labs, Cat#M0541S)
- Nextera Index Kit (Illumina, Cat#FC-121–1011)
- MinElute PCR Purification Kit (Qiagen, Cat#28006)
- AMPure beads (Beckman, Cat#A63880)
- KAPA SYBR® FAST qPCR Master Mix (Roche, Cat#07960204001)
- Qubit dsDNA High Sensitivity Assay Kit (Thermo Fisher Scientific, Cat#Q32851)
- High Sensitivity DNA Kit (Agilent Technologies, Part#5067–4626)

2.3 Solutions

1 ml of Protease inhibitor solution (50x)

Mix by vertexing and store at -20°C for up to 12 weeks.

- complete™, EDTA-free protease inhibitor cocktail tablet - 1 tablet
- Milli-Q water - 1 ml

100 ml of 20% (v/v) Triton X-100

Stir for 30 minutes until fully mixed. Store in the dark at room temperature for a maximum of 2 months.

- Triton X-100 - 20 ml
- Milli-Q water - 80 ml

10 ml of Organelle Removal Buffer (ORB)

Store at 4°C for 1 week. Add the protease inhibitor stock solution right before use.

- 2.5 M Sucrose stock solution - 1 ml (250 mM)
- 1M, PH8.0 Tris-HCl - 100 µl (10 mM)
- 100 mM $MgCl_2$ - 1 ml (10 mM)
- 20% Triton X-100 - 500 µl (1%)
- 50x Protease inhibitor stock solution - 200 µl
- Milli-Q water - 7.2 ml

10 ml of Sucrose Cushion Buffer (SCB)

Store at 4°C for 1 week. Add the protease inhibitor stock solution right before use.

- 2.5 M Sucrose stock solution - 6.8 ml (1.7 M)
- 1M, PH8.0 Tris-HCl - 100 µl (10 mM)
- 100 mM $MgCl_2$ - 200 µl (2 mM)
- 20% Triton X-100 - 50 µl (0.1%)
- 50x Protease inhibitor stock solution - 200 µl
- Milli-Q water - 2.65 ml

100 ml of MOPS (1M)

Autoclave and store at 4°C for a maximum of 3 months.

- 3-(N-morpholino) propanesulfonic acid, 4-morpholinepropanesulfonic acid (MOPS) - 20.93 g (1 M)
- Milli-Q water - 80 ml
- NaOH (10N) - Adjust to pH7.0

10 ml of Nucleus Extraction Buffer (NEB)

Store at 4°C for 1 week. Add protease inhibitor, spermine, and spermidine stock solutions right before use.

- 1 M, pH 7.0 MOPS - 200 µl (20 mM)
- 5 M NaCl -80 µl (40 mM)
- 3 M KCl - 300 µl (90 mM)
- 500 mM EDTA - 50 µl (2.5 mM)
- 100 mM EGTA - 50 µl (0.5 mM)
- 400 mM Spermine - 5 µl (0.2 mM)
- 400 mM Spermidine - 13 µl (0.52 mM)
- 50x Protease inhibitor solution - 200 µl

- Milli-Q water - 9.1 ml

DAPI stock solution

Dissolve DAPI with vigorous shaking. To prepare the working DAPI solution, dilute the stock solution 1000 x to 1 µg/ml. For long-term storage, aliquot the stock solution and store at -20°C, where it will remain stable for at least six months.

- DAPI dilactate - 10 mg
- Milli-Q water - 1 ml

2.4 Equipment and materials

- Long gel-loading pipette tips (Labcon, Part#1034–960-008)
- Double-edged stainless steel razor blade (Electron Microscopy Sciences)
- Glass Petri dish
- Parafilm
- MiraCloth
- CellTrics cell strainers (10 and 20 µm mesh) (Sysmex Partec)
- Hemocytometer
- 1-ml and 2-ml Eppendorf low-binding tubes
- 1.7ml Microtube, Clear, Maxymum Recovery (Axygen, Part#MCT-175-L-C)
- Fine nylon paintbrush (size 0)
- Thermomixer Comfort (Eppendorf, Cat#5355)
- T100™ Thermal Cycler (BIO-RAD, Cat#186–1096)
- Centrifuge: MiniSpin Plus (Eppendorf, Cat#5453)
- Applied Biosystems QuantStudio™ 12K Flex Real-Time PCR System (Thermo Fisher Scientific, Cat#4470050)
- DYNAL MPC-S magnetic stand (Applied Biosystems)
- Qubit™ 2.0 (Thermo Fisher Scientific, Cat#Q32866)
- Agilent 2100 Bioanalyzer Instrument (Agilent Technologies, Cat#G2939BA)

2.5 Software and algorithms

- R version 3.4.4 (https://www.r-project.org)
- FastQC v0.11.8 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/)
- TrimGalore v0.4.1 (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/)
- HISAT2 v2.1.0 (http://daehwankimlab.github.io/hisat2/)
- edgeR v3.20.9 (https://bioconductor.org/packages/release/bioc/html/edgeR.html)
- Bowtie2 v2.3.5.1 (http://bowtie-bio.sourceforge.net/bowtie2/)
- ATACgraph (https://github.com/RitataLU/ATACgraph)

3 Methods

3.1 Step-by-step method details

3.1.1 Nuclei isolation

A crucial requirement for generating a good ATAC-seq library is to use pure and intact nuclei (Lu et al., 2017; Marand et al., 2021). Notably, since Tn5p targets not only nuclear DNA but also organellar DNA, the sequencing reads from organellar DNA fragments can often account for up to 50% of the total sequencing reads (Montefiori et al., 2017). Therefore, the removal of organelles prior to the transposition reaction can thus greatly improve ATAC-seq efficiency. In the process of isolating plant nuclei, two steps are typically involved: breaking plant cell walls to release nuclei, and removing the organelles.

Timing 1–2 h

1. Cool the centrifuge to 4°C. Keep reagents and solutions on ice throughout the process.

2. Prepare the Percoll-Sucrose gradient buffer. Mix 400 µl Organelle Removal Buffer (ORB) with 600 µl Percoll to make 60% Percoll solution. Add 200 µl Sucrose Cushion Buffer (SCB) to a new 1.5-ml microcentrifuge tube. Carefully overlay with 400 µl 60% Percoll solution. Be very careful not to mix the sucrose and Percoll layers. Keep the gradient on ice. Perform this step at least 1 hour before using the Percoll-Sucrose gradient.

Note: We found that one-hour equilibration of the Percoll-Sucrose gradient buffer resulted in better nuclei recovery than immediate use, as incorrect Percoll concentration may affect gradient formation and stability, leading to irregular sample sedimentation.

3. Place a piece of 5x5 cm^2 parafilm on a glass Petri dish on ice. Pipette 500 µl NEB onto the parafilm and finely slice the above-ground parts of 10-day-old maize seedlings (approximately 5 seedlings) in the NEB buffer with a double-edged razor blade. After slicing, proceed to chop further in 2–3 separate batches until the tissue appears like a coarse mixture (Figure 2). Each batch of chopping takes about 2–5 minutes to complete. Add more NEB buffer if needed. Approximately 5 ml NEB is required for 5 seedlings.

Note: Chopping with a sharp razor blade is a gentler method to release nuclei, minimizing potential damage compared to grinding tissues directly in liquid nitrogen. To assist free-hand chopping, leaves were first rolled up and sliced finely across the veins into thin strips. Subsequently, these strips are then chopped into a coarse mixture.

4. Transfer the mixture to a small petri dish on ice. Incubate for 5 minutes with gentle agitation. Place 4 layers of Miracloth on a pre-chilled glass funnel. Filter the mixture through the funnel into a 15-ml Falcon tube. Gently squeeze the remaining tissue against the funnel to extract more nuclei.

5. Filter the crude nuclei suspension through a 20-µm CellTrics twice to remove large debris. Divide the filtrate into 1.5-ml microcentrifuge tubes and centrifuge at 800 rcf for 10 minutes at 4°C.

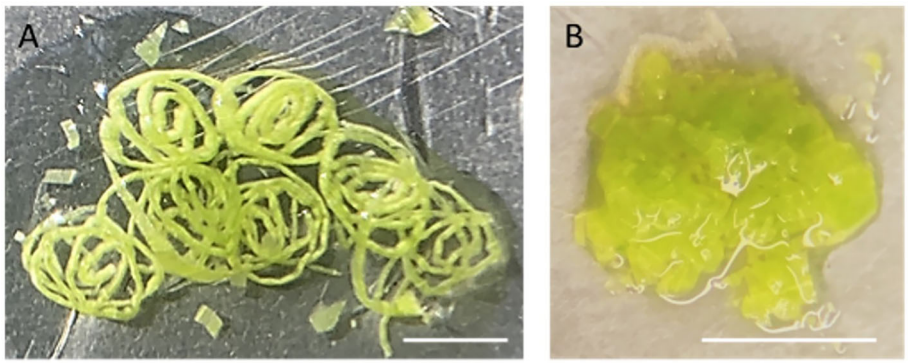


FIGURE 2
Chopped leaf tissue on a piece of parafilm. **(A)** The rolled leaves were first sliced into thin strips. **(B)** After fine chopping, the leaf tissue has the appearance of a coarse mixture. Scaled bar represents 5mm.

6. Gently remove and discard the supernatant without disturbing the pellet. The pellet should comprise a pale-white layer at the bottom and a green layer on top.

Note: The pale-white layer predominantly consists of starch grains, and the top layer contains nuclei, chloroplasts and other organelles.

7. Slowly pipette 200 µl ORB (with 1% Triton X-100) into each tube from step 6 and incubate for 2 minutes so the pellet loosens. Use a fine nylon paintbrush to gently remove the green layer from the pellet, leaving the pale-white bottom layer undisturbed. Gently transfer the green pellets with ORB to a new tube by pipetting. Resuspend the green pellet material in the ORB using the paintbrush and by gentle pipetting.

Note: Chloroplasts are lysed in 1% Triton X-100 in the ORB buffer. If this leftover solution is centrifuged, the pellet will be white and lack the uppermost green layer as chlorophyll is released in the supernatant from broken chloroplasts.

8. Filter the nuclei suspension from step 7 through a 10-µm CellTrics twice to remove aggregate nuclei and debris.

9. Carefully load approximately 400 µl of the filtered green-colored nuclei suspension on top of the previously prepared Percoll-Sucrose gradient buffer. Centrifuge at 1,000 rcf for 15 minutes at 4°C. Prepare an additional Percoll-Sucrose gradient, if more than 400 µl nuclei suspension solution is being used.

10. Remove the green-colored supernatant on the top. Collect the second layer of 60% Percoll buffer that contains intact nuclei and transfer the nuclear fraction into a new 1.5-ml microcentrifuge tube. Be careful not to disturb the interface between Percoll and sucrose layers.

Note: As we have not tested this protocol in other organisms, users may need to verify the presence of intact nuclei under a microscope when applying it to different species. Although untested, the flowchart and concept of our protocol could serve as a reference to help scientists conduct ATAC-seq in other organisms.

11. Add 1X volume of NEB to dilute the nuclear fraction suspension. Centrifuge at 500 rcf for approximately 10 minutes to pellet the nuclei.

Note: Check the pellet condition every 2 minutes and stop centrifugation when a small translucent pellet is first visible. We found that excessive centrifugation can result in nuclear damage. The nuclear pellet is translucent, different from the pale white color of the starch grain pellet from step 6.

12. Dissolve the pellet in 20 µl NEB to obtain the final isolated nuclei solution.

13. Take 2 µl of the isolated nuclei solution from step 12 and dilute in NEB. Stain with DAPI (final concentration: 0.3 µg/ml) (Figure 3). Examine the nuclei integrity under a microscope and calculate nuclei density of the isolated nuclei suspension using a hemocytometer.

Note: The expected concentration of nuclei in the isolated nuclei solution from step 12 is approximately 5–20 million/ml.

14. Transfer a volume equivalent to 2,500 or 5,000 nuclei to a new 1.5-ml low-binding tube.

Note: According to the manufacturer's manual of Nextera DNA Library Preparation Kit, the input volume of nuclei for the transposition reaction should be < 5 µl.

3.1.2 Tn5p Transposition

The Nextera DNA Library Preparation Kit (FC-121–1031, Illumina) was used to perform the transposition reaction according to the manufacturer's manual. In addition to nuclear samples, 50 ng genomic DNA (gDNA) was used as a negative control input. A brief description of the procedure is as follows.

Timing 40 minutes

15. Prepare the transposition reaction mix and gently resuspend the nucleus pellet in the transposition reaction mix as follows.

Note: For maize, we found that only about 2,500 and 5,000 nuclei are good starting materials are required for a single reaction. As we have not tested this protocol in other organisms, when adapting this protocol for different species, it is recommended to optimize the number of nuclei used in one reaction.

- Purified intact nuclei (2,500 or 5000 nuclei) or 50 ng gDNA - 2 µl

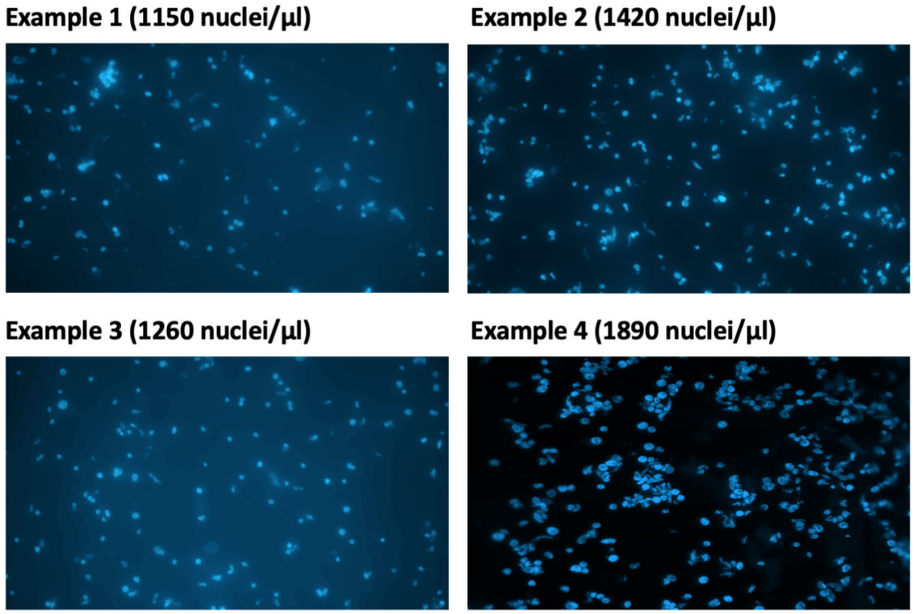


FIGURE 3
Estimation of DAPI-stained nuclei under microscope.

- 2 x TD reaction buffer - 25 µl
- TDE1 (Nextera Tn5 Transposase) - 2.5 µl
- Nuclease free water - 20.5 µl

16. Incubate the transposition reaction at 37°C for 30 minutes in an Eppendorf ThermoMixer Comfort with agitation at 2 rcf for 2 minutes, followed by a rest period of 2 minutes (i.e., occasionally mixing).

Note: Occasional gentle mixing may facilitate uniform reaction and increase fragment yield.

17. Immediately following transposition, purify the Tn5p transposed DNA fragments with a Qiagen MinElute PCR Purification Kit.

3.1.3 Purification of DNA fragments

Use a Qiagen MinElute PCR Purification kit to purify the transposed DNA fragments according to the manufacturer's manual. A brief description of the procedure is as follows.

Timing 15 minutes

18. Add 5X volume of PB Buffer into the transposition reaction and vortex to mix.

Note: Check for the yellow color of PB Buffer. If the buffer turns orange or violet, the pH is too high. A small volume of 3M sodium acetate can be added to adjust the pH before proceeding.

19. Transfer the mix into a MinElute spin column provided by the Qiagen MinElute PCR Purification kit.

20. Centrifuge for 1 minute at 16,000 rcf. Discard all the flow-through.

21. Add 750 µl of PE Buffer, and centrifuge for 1 minute at 16,000 rcf. Discard the flow-through.

22. Centrifuge for 1 minute at 16,000 rcf. Rotate the column by 180°C and centrifuge for 1 minute.

Note: This step removes any remaining ethanol in the column.

23. Place the MinElute spin column in a new low-binding tube.

24. Elute in a new tube with 12 µl of EB buffer (10 mM Tris, pH 8.5) and stand at room temperature for 1 minute.

25. Centrifuging for 1 minute at 16,000 rcf. Rotate the column 180°C and centrifuge for 1 minute.

26. Discard the column and store the purified DNA at -20°C if necessary.

Note: This is a convenient stopping point for the protocol.

3.1.4 Preparation of sequencing libraries

This phase consists of four major steps. First, indexing barcodes are added onto each transposed DNA fragment by 5 cycles of PCR amplification ('the first PCR'). Second, a small aliquot from the product of the first PCR is subjected to quantitative PCR (qPCR) to assess the amount of DNA template, and the result will be used to determine the number of cycles used in the secondary PCR. Third, these indexed fragments from the first PCR are further amplified with sequencing primers (i.e., 'the secondary PCR') to generate a sufficient amount of DNA for sequencing. Fourth, the resulting library from the secondary PCR is purified.

Timing 3–4 hours

27. First PCR amplification with indexing barcode primers.

Use the Nextera Index Kit and NEBNext High-Fidelity 2x PCR Master Mix to amplify transposed DNA fragments. A brief description of the procedure is as follows.

a. Prepare a PCR mixture by combining the First PCR Amplification Reagent with 10 µl purified transposition product in a 0.2 ml PCR tube as follows.

Note: Different index primers (barcodes) should be selected for distinct samples, so that these samples can be pooled together for sequencing. After sequencing, reads from these samples can be identified and separated according to their barcodes.

- Transposed DNA - 10 µl
- NEBNext High-Fidelity 2x PCR Master Mix - 25 µl
- 25 µM Index 1 (i7) - 2.5 µl
- 25 µM Index 2 (i5) - 2.5 µl
- Nuclease free water - 10 µl

b. The thermal cycle for PCR amplification is as follows: 72°C for 5 minutes, 98°C for 30 seconds, 5 cycles (98°C for 10 seconds, 63°C for 30 seconds, and 72°C for 1 minute).

c. Transfer the PCR product to a new microcentrifuge tube.

28. Determination of cycle number for the secondary PCR by qPCR.

The KAPA Library Quantification Kit is used for the quantitative PCR. A brief description of the procedure is as follows.

a. Take 1 µl of the amplification product from the first PCR reaction and mix with qPCR reagents as follows.

- first PCR product amplification product - 1 µl
- 2X KAPA SYBR® FAST qPCR Master Mix with 10X
- Primer Premix -12 µl
- Nuclease free water - 7 µl

b. Perform the qPCR amplification on an Applied Biosystems QuantStudioTM 12K Flex Real-Time PCR System with the following conditions: 98°C for 5 minutes, 20 cycles (98°C for 10 seconds, 63°C for 30 seconds, and 72°C for 1 minute).

c. Calculate the required number of cycles (i.e., 'N' cycles) for the secondary PCR step by plotting linear Rn versus the cycle number. The N cycle number corresponds to the cycle number on the qPCR plot, where the fluorescent intensity reaches one-third of the maximum value.

29. Secondary PCR amplification.

Conduct the secondary PCR using the remaining product from the first PCR as the template, mixed with the NEBNext High-Fidelity 2x PCR Master.

a. Prepare the Secondary PCR Amplification reagents and add 40 µl of the first PCR product from step 27 in a 0.2 ml PCR tube as follows.

- first PCR product amplification product - 40 µl
- NEBNext High-Fidelity 2x PCR Master Mix - 50 µl
- PCR Primer Cocktail - 5 µl
- Nuclease free water - 5 µl

b. Perform the secondary PCR amplification as follows: 98°C for 30 seconds, and N cycles of (98°C for 10 seconds, 63°C for 30 seconds, and 72°C for 1 minute). The N cycle number is determined based on the qPCR analysis in Step 28.

Note: While the original protocol suggests using 5 µl of the first PCR product as the template for qPCR, we have found that 1 µl is sufficient (step 28). Consequently, approximately one-fifth of the DNA template is used to estimate the "N" cycle number in our assay, which equates roughly to two cycles of PCR amplification. Therefore, in our experience, successful libraries can be generated using a cycle number lower than what is determined by qPCR (step 28). This adjustment helps in avoiding over-amplification.

30. Purification of the amplified library.

Use the AMPure XP kit to purify the amplified DNA from step 29. A brief description of the procedure is as follows.

a. Allow the AMPure XP beads to reach room temperature for at least 30 minutes. Vortex the AMPure XP Beads until they are well dispersed. Add 100 µl (1X volume) of well-mixed AMPure XP Beads into a new 1.7 ml tube.

b. Transfer each DNA library from step 29 to the 1.7 ml tube containing the AMPure XP Beads. Gently pipette the entire volume up and down 10 times or vortex gently to mix thoroughly.

c. Incubate the tubes at room temperature for 5 minutes. Place the tubes on a magnetic stand at room temperature for 3 minutes or until the liquid appears clear. Carefully remove the supernatant from each tube.

Note: Some liquid may remain in each tube. Do not disturb the beads.

d. Wash the bead-bound DNA pellet twice with 200 µl of 80% ethanol. Let the tubes open at room temperature for 10 minutes, enabling the pellet to dry.

e. Resuspend the dried pellet in each tube with 22.5 µl EB Buffer. Vortex to mix thoroughly and incubate the tube at room temperature for 2 minutes.

f. Place the tube on a magnetic stand at room temperature for 3 minutes or until the liquid appears clear. Transfer all of the clear supernatant, which contains the purified DNA, to a new 1.7 ml tube. The samples can be stored at -20°C.

3.1.5 Checking library quality and sequencing

31. Use the Qubit High Sensitivity Assay Kit to determine the concentration of each ATAC-seq library.

32. Assess the fragment size distribution and peak pattern of each library using either the Agilent Bioanalyzer 2100 system or BiOptic Qsep400.

Note: A successful ATAC-seq library should exhibit a pattern of periodicity in fragment sizes with an interval of around 200 bp.

33. We used an Illumina platform HiSeq 4000 for 150-bp paired-end sequencing to generate 30 million raw reads for each maize library.

Note: Other Illumina sequencing platforms (e.g., Hiseq X Ten, and NovaSeq 6000) can be used. Choose ones capable of paired-end sequencing with at least read length of 50 bp.

3.2 Quantification and statistical analysis pipeline

3.2.1 RNA-seq analysis pipeline

The RNA-seq reads derived from maize seedling and root samples (accession numbers SRR7548392 and SRR2043190) are used for analysis. The bioinformatics pipeline for the basic analysis of the RNA-seq dataset is outlined as follows, including quality control to the calculation of Fragments Per Kilobase Million (FPKM).

1. Each sample is quality checked by FASTQC v0.11.8.
2. The adapter and quality were trimmed using TrimGalore v0.4.1 (Krueger et al., 2016), and at least 35 bp long reads were retained, obtaining clean FASTQ files.
3. Reads were aligned to the Maize B73 reference genome (AGPv4) using HISAT2 v2.1.0 (Kim et al., 2019) (Supplementary Table 1).
4. Read counts were normalized using the TMM function in edgeR package v3.20.9 (Chen et al., 2016) under R v3.4.4, and then FPKM values were obtained by calling function of rpkm. The ranks of transcript abundance are listed by its $\log_2$ FPKM from the lowest to the highest for its correlation analysis with chromatin accessibility.

3.2.2 ATAC-seq analysis pipeline

The pipeline for the ATAC-seq data analysis is adopted from ATACgraph (Lu et al., 2020). The code used here is available at GitHub: https://github.com/beritlin/ATACgraph2 (Lin, 2023). It is specifically designed for ATAC-seq data analysis and is able to remove mitochondria and plastid DNA, identify open regions by peak calling, plot the fragment length distribution, compute the periodicity of fragment length distribution using FFT algorithms as well as show heatmaps showing accessibility around all genes.

1. The raw ATAC-seq reads were underwent trimming, duplication removal, and aligned to the maize reference genome (AGPv4) using Bowtie2 v2.3.5.1 (Langmead and Salzberg, 2012) (Supplementary Table 2).
2. Following the steps in ATACgraph, mitochondria and plastid DNA were removed by command line 00_rmChr with Mt and Pt as the option, an aligned bam file as input and obtained clean bam files as output.

```
> ATACgraph 00_rmChr atac_sample.bam
atac_sample_rmM.bam Mt, Pt
```

3. The clean bam file and its matched gDNA bam file were then used for peak calling. Running the function 03_callPeak with the provided gene body bed file, will generate the peak location narrowPeak files, intensity bigwig (bw) files, as well as a genes list of overlapping with peak locations txt file as the selected output name in the command. The script is updated to MACS3 and the parameter here is using p-value < 0.05 as a cutoff for peaks calling due to the small amount of nuclei used in our protocol.

```
> ATACgraph 03_callPeak atac_sample_rmM.bam
atac_sample_rmM_peakcall Maize_gene_body_
bed6.bed -c gDNA.bam
```

4. Figures were also plotted by ATACgraph, including the period of fragment length distribution and FFT by implementing 01_calFragDist by using clean bam file as input. Heatmaps around genes was regenerated by command line 03_genePlot with narrowPeak file, bigwig (bw) and the bed files.

```
> ATACgraph 01_calFragDist atac_sample_
rmM.bamatac_sample_fragment atac_sample_FFT
> ATACgraph 03_genePlot atac_sample_rmM_
peakcall.narrowpeak atac_sample_rmM_peakcall_
coverage.bw Maize
```

5. The read counts at the peak regions were generated by computeMatrix v3.3.2 from the bigwig (bw) file and normalized using the TMM function in edgeR package v3.20.9 under R v3.4.4.
6. The evaluation of libraries was also performed by running the following commands, providing the peaks files, bam files, and gene annotation which returns the log file of each score.

```
> ATACgraph 04_IDR.py atac_sample_rmM_R1_
peakcall.narrowpeak
atac_sample_rmM_R2_peakcall.narrowpeak
atac_sample_idr
> ATACgraph 04_frip.py atac_sample_rmM_
peakcall.narrowpeak
atac_sample_frip atac_sample_rmM.bam
> ATACgraph 04_tssEnrich.py atac_sample_rmM_
peakcall.narrowpeak Maize atac_sample_
tss atac_sample_rmM.bam
```

4 Results

Expected outcomes at each critical step of the protocol are described. Additionally, comparisons of various parameters were carried out to ensure optimized parameters and the effectiveness of ATAC-seq.

4.1 Pure and intact nuclei isolation

The isolation of plant nuclei typically involves the mechanical disruption of cell walls by grinding in liquid nitrogen or breaking tissues using polytrons or razor blades. The objective of this step is to break down the cell wall and release cellular contents, including nuclei. Since the integrity of nuclei is essential for obtaining qualified data of ATAC-seq, the method ought to be gentle enough to preserve nuclei and chromatin conformation. We found that chopping with a sharp razor blade is a gentler method to release nuclei, minimizing potential damage. To compare nucleus morphology released by different methods, we stained nuclei with DAPI and observed them under a microscope. As shown in Figures 4A and B, the nuclei extracted by grinding in liquid nitrogen or disrupting by a polytron homogenizer were of poor quality. The stringy mess or misconfigured nuclei suggested that cellular structures are destroyed, and nuclei are damaged. Chromatin released from damaged nuclei exhibited irregular DAPI staining, often seen as scrambled mess with cell debris (Figure 4E), so the native chromatin structure likely has been altered. We found that it was difficult to conduct either of these two methods without damaging maize nuclei. In our adapted protocol, where maize fresh tissues are sliced and chopped with a sharp stainless steel razor blade, we successfully obtained high yields of round-shaped nuclei with distinct chromatin morphology (Figures 4C, D), in contrast to the damaged nuclei isolated with two other methods (Figures 4A, B, E). Interestingly, after collecting nuclei by centrifugation, we found that a large amount of starch grains appeared in a pale white layer underneath a green layer. Starch grains do not stain with DAPI nor show autofluorescence (Figure 4F). To minimize starch grains in our samples, we found that a fine nylon paintbrush is very useful to separate the green layer hosting nuclei and most organelles from the underlying starch grain layer (step 7 in the 3.1 section).

To purify nuclei effectively, it is crucial to consider factors such as purity, the required duration, efficiency, and ease of access to equipment. Flow cytometry offers an effective method for isolating intact nuclei from organelles and broken nuclei as it utilizes fluorescent signals (i.e. DAPI-stained nuclei or nucleus-expressed fluorescence) to distinguish nuclei based on their DNA content and size. It is most suitable for collecting hundreds of thousands of nuclei from various debris and organelles. As we wanted to develop a reliable and easy-to-implement ATAC-seq protocol, we optimized the use of the sucrose-Percoll centrifugation method in combination with Triton X-100 treatment, which requires only a bench-top centrifuge.

Prior to the sucrose-Percoll centrifugation (step 9 in the 3.1 section), the non-ionic detergent, Triton X-100, is used to eliminate organelles. We tested different concentrations (0.5–2%) of Triton X-100 to lyze chloroplasts while nuclei remained intact. Figure 5 shows that both chloroplasts and nuclei remained intact after 0.5% Triton X-100 treatment, but most chloroplasts were lysed in 1% Triton X-100 buffer. However, at higher concentrations of Triton X-100, there was a noticeable decrease of intact nuclei, suggesting that an optimal concentration for lysing chloroplasts, the most abundant organelles, without compromising nuclear integrity, is 1% Triton X-100.

The sucrose-Percoll centrifugation method takes advantage of differential buoyant density, size, and shape of subjects, which allows separation of subcellular compartments during centrifugation in high-viscosity media such as sucrose and/or Percoll (a colloidal silica). We tested 60% Percoll with 2.5 M sucrose as a cushion in a bench-top centrifuge (Sikorskaite et al., 2013) (Figure 6). After centrifuging at 1,000 rcf for 15 minutes, we

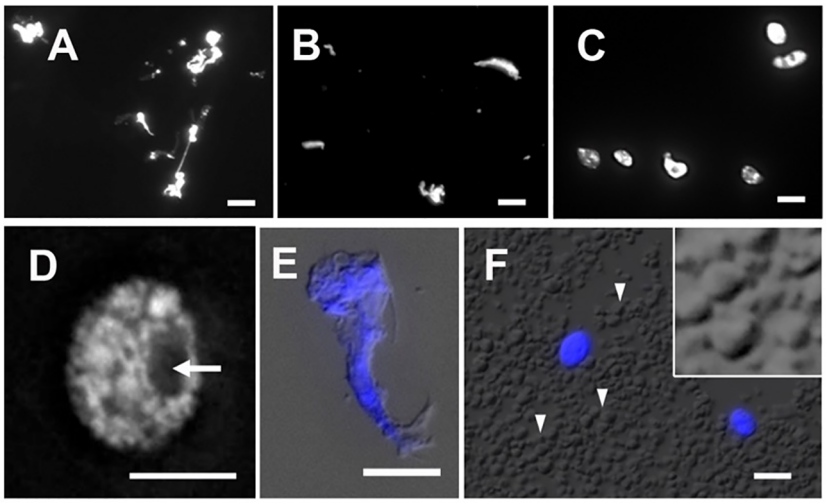


FIGURE 4
Nuclei stained with DAPI after extraction by **(A)** grinding in liquid nitrogen, **(B)** polytron homogenizer and, **(C)** manual chopping. **(D)** A magnified image of a single nucleus isolated by chopping showing a round-shaped nucleus with distinct chromatin morphology and a visible nucleolus (arrow). **(E)** An example of cell debris demonstrates distorted cell structures and scrambled chromatin with DAPI stain (blue), shown with the merged DIC image. **(F)** In the crude homogenate, a large amount of starch grains (arrowheads) can be observed by the DIC microscope. Note that DAPI-stained round-shape nuclei (blue) are scattered. The magnified inset shows polyhedral starch grains. Scale bars represent 5 µm.

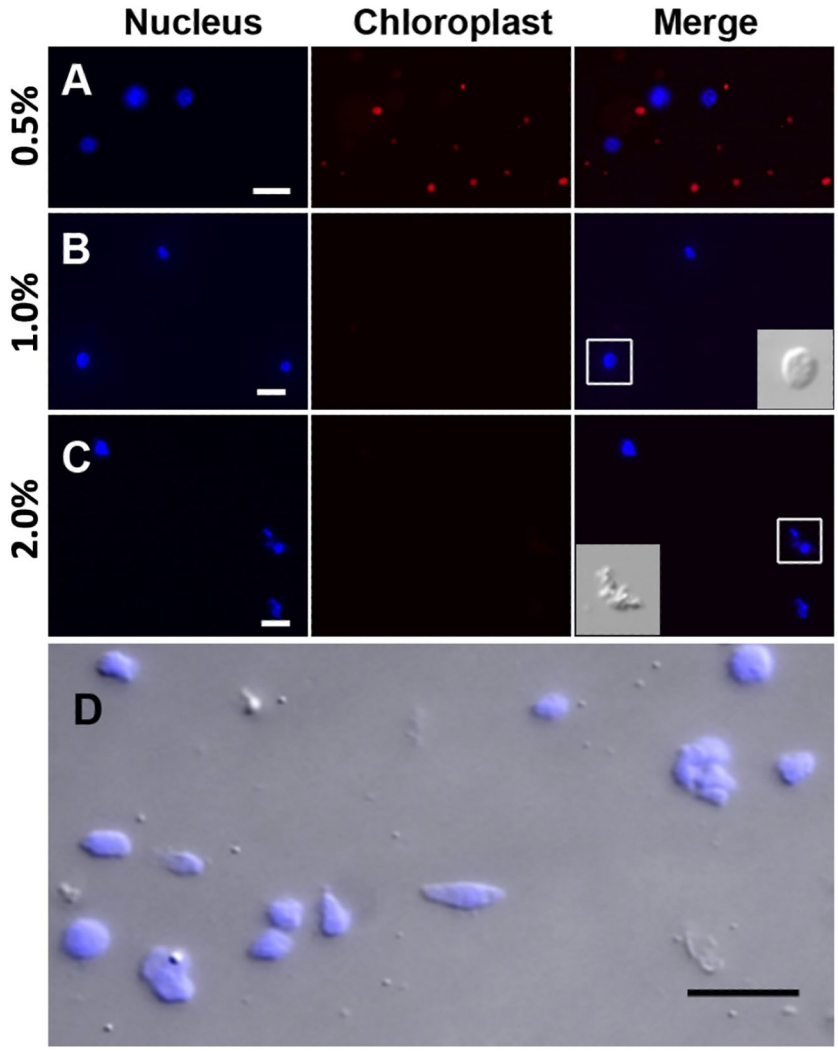


FIGURE 5
Nuclei treated with different Triton X-100 concentrations: **(A)** 0.5%, **(B)** 1.0%, **(C)** 2.0% Triton X-100 and stained with DAPI (blue). Chloroplasts exhibit strong red autofluorescence. Note that a majority of chloroplasts are lysed in Triton X-100 above 1.0%. When treated with 2.0% Triton X-100, nuclei integrity is severely compromised, as shown in the inset in C by DIC. **(D)** After the sucrose-Percoll gradient centrifugation, most nuclei remain intact. Scale bars represent 10 µm. DAPI-stained nuclei were detected using excitation wavelength 353 nm and emission wavelength 465 nm. Chloroplast autofluorescence was captured using 638 nm/646 nm (excitation/emission).

observed that the upper layer of supernatant is colored green due to chlorophyll content, and nuclei are suspended in the 60% Percoll layer. A brownish-white layer is sometimes deposited at the sucrose-Percoll interface, which contains mostly contaminating material and intact cells. Nuclei isolated from the Percoll layer were intact with clear nucleoli in microscopic examination. Although some nuclei appeared elongated after centrifugation, their chromatin was well-contained within the nuclear boundary (Figure 5D). Taken together, we were able to collect approximately one hundred thousand pure nuclei from five seedlings using this method, which includes manual chopping, 1% Triton X-100 treatment, and the 60% Percoll:2.5 M sucrose gradient separation. Before Tn5p transposition, the isolated nuclei were examined under a microscope to check their morphology (Figure 7). From four independent analyses, the average percentage of intact nuclei was 89.08% with an S.D. of 1.85.

4.2 Optimization of nuclei number and Tn5p transposition efficiency

In previous studies of maize ATAC-seq, 50,000 nuclei are commonly used (Lu et al., 2019; Ricci et al., 2019; Crisp et al., 2020). Several Arabidopsis ATAC-seq studies have shown that a variable number of nuclei, ranging from 500–50,000, were used in their experiments (Buenrostro et al., 2013; Gray et al., 2017). To evaluate the appropriate number of maize nuclei for one reaction in our protocol, we tested 500, 2,500, 5,000, and 50,000 nuclei and used genomic DNA as a control. After the transposition reaction, DNA fragments were purified using PCR purification kit for the next amplification step. During the first round of PCR amplification, indexing barcodes were added onto each transposed DNA fragment. After five cycles of PCR, we paused the reaction and took a small aliquot (1 µl) of the PCR product for a qPCR assay to

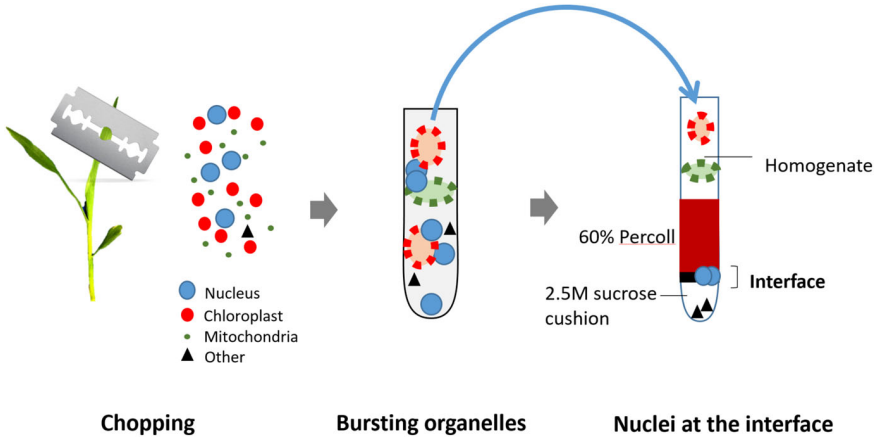


FIGURE 6
Illustration of the nuclei isolation method from fresh seedlings. The seedlings are first manually chopped in a buffer using a sharp blade to release nuclei and organelles. Next, organelles are lysed using 1% Triton X-100, but nuclei remain intact. The crude exacts are then loaded on the top of Percoll-Sucrose gradient. After centrifugation, maize nuclei are enriched in the 60% Percoll layer. Isolated nuclei are transferred to a new tube and collected by centrifugation for the next transposition reaction. are transferred to a new tube and collected by centrifugation for the next transposition reaction.

estimate the relative amount of effective DNA template. Higher amounts of DNA template detected by qPCR (i.e. a lower N number) indicate more successful transposition events. This should correlate with the initial number of nuclei used in the reaction.

As shown in Figure 8A, qPCR amplification plots from transposition reactions with 500 and 5,000 nuclei showed positive correlations with effective DNA template. However, reactions with 50,000 nuclei did not further increase effective transposition events. In contrast, the naked genomic DNA exhibited a steep amplification curve, suggesting that free DNA is much more easily targeted by the Tn5p. Moreover, we tested whether using frozen nuclei could enhance transposition events (Corces et al., 2017). We froze approximately 2,500 purified nuclei at -20°C for 30 minutes before proceeding with the transposition reaction. In comparison with 2,500 fresh (non-frozen) nuclei, qPCR analysis revealed a notable increase in transposed DNA fragments within the frozen samples, even exceeding the numbers generated in the reaction using 50,000 nuclei (Figure 8A). We reasoned that the freezing process may damage nuclei and chromatin structures, thus more susceptible to Tn5p transposition. In addition, we tested whether a mild detergent in the transposition reaction could increase nuclear permeability for Tn5p transposition. Our results showed that SDS, CHAPS and NP40 did not result in significantly improved outcomes (Figure 8B).

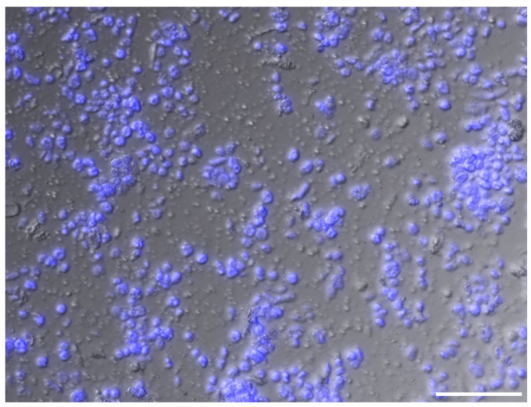


FIGURE 7
An image of isolated nuclei in high density. Intact nuclei account for approximately 90% of the population. Nuclei are visualized with DAPI stain (blue). Unstained matters include cell debris, a small number of chloroplast, and starch grains. Scale bar represents 50 µm.

4.3 Determination of PCR cycle number for amplifying the NGS library

In our libraries of 2,500 and 5,000 fresh maize nuclei, we estimated N cycles of 14 and 13, respectively, based on the cycle number reaching one-third of the maximum value in our qPCR plots (Table 1; Figure 8). The fact that these N cycle numbers are higher than we expected may suggest that the number of tagged fragments in maize is lower than achieved for other organisms, which may reflect that the maize genome contains a smaller proportion of open chromatin (Lu et al., 2017; Bajic et al., 2018; Maher et al., 2018) or alternatively reflect the less potential complexity due to fewer nuclei. Indeed, a previous study using MNase-seq showed that only a small portion (<1%) of the maize genome resides in open chromatin (Rodgers-Melnick et al., 2016) though MNase-seq is not fully comparable with ATAC-seq. To avoid excessive duplication and PCR bias, we tested the secondary PCR using a reduced number of additional cycles (four fewer than initially estimated). Accordingly, we generated sequencing libraries using 10 and 9 cycles for the 2,500 and 5,000 nuclei samples, respectively, so that the DNA concentrations of the libraries were sufficient for sequencing with limited duplicates (Table 1).

10.3389/fpls.2024.1370618

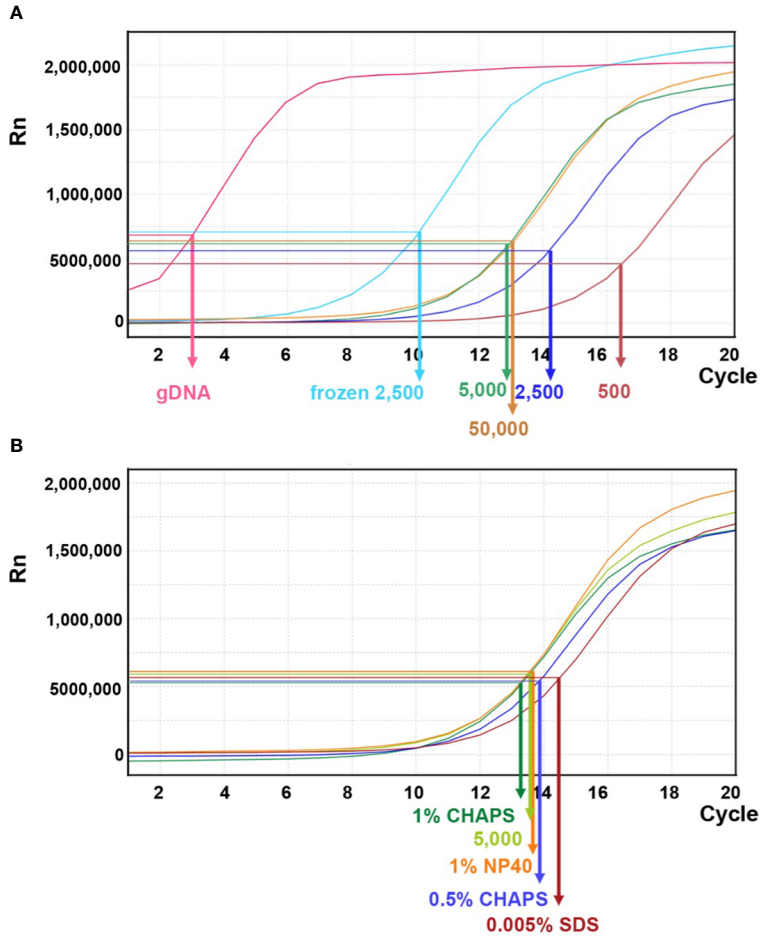


FIGURE 8
Linear amplification plot of quantitative PCR analysis for different transposition reactions. **(A)** The linear amplification plot versus PCR cycles shows that reactions with 500, 2,500, or 5,000 fresh nuclei (dark brown, dark blue, and green lines, respectively) positively correlate with nuclei numbers. However, the reaction with 50,000 fresh nuclei did not result in an improved outcome. When 2,500 frozen nuclei (light blue line) were used as input, the amplification plot showed a significant increase of DNA template. Genomic DNA (gDNA) serves as a control for the transposition reaction. The arrows indicate the N cycle number of each sample, corresponding to one-third of the maximum signal intensity. **(B)** Quantitative PCR analysis for transposition reactions with different mild detergents using 5,000 fresh nuclei.

TABLE 1 PCR cycles and DNA concentrations of ATAC-seq libraries.

Sample	Treatment	Predicted N cycle	Additional cycles of 2nd PCR	Library concentration (ng/ul)	Total DNA (ng)
50 ng gDNA	NA	3	3	19.50	390
500 nuclei	fresh	16	13	1.97	39
5,000 nuclei	fresh	13	9	1.96	39
50,000 nuclei	fresh	13	9	3.52	70
5,000 nuclei	0.5% CHAPS	14	10	2.02	40
5,000 nuclei	1% CHAPS	14	10	2.44	49
5,000 nuclei	1% NP40	14	10	2.24	45
5,000 nuclei	0.005% SDS	14	10	1.38	28
2,500 nuclei	fresh	14	10	1.51	30
2,500 nuclei	frozen	10	8	8.68	174

10.3389/fpls.2024.1370618

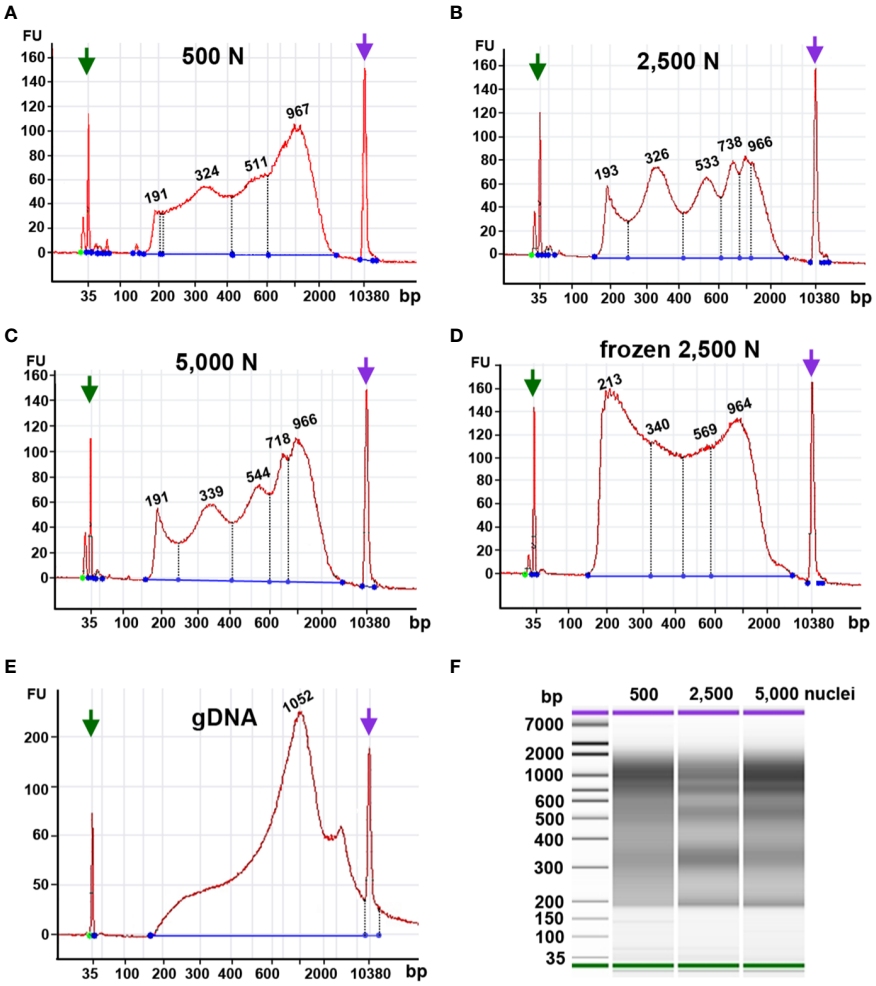


FIGURE 9
Fragment size distribution of ATAC-seq libraries determined by an Agilent 2100 Bioanalyzer. The X-axis represents the product size in base pairs and the Y-axis is the arbitrary fluorescence intensity (FU). **(A–E)** Fragment size distribution of samples with **(A)** 500 fresh nuclei (500 N), **(B)** 2,500 fresh nuclei (2,500 N), **(C)** 5,000 fresh nuclei (5,000 N), **(D)** 2,500 frozen nuclei, or **(E)** 50 ng genomic DNA (gDNA). Numbers indicate fragment sizes of peaks, which may represent mono- or multiple nucleosomes in the 2,500 nuclei and 5,000 nuclei samples. Purple and green arrows represent upper and lower size markers, respectively. **(F)** A gel-like picture of the fragment size distributions of ATAC-seq libraries generated from 500, 2,500, or 5,000 fresh nuclei.

4.4 Library quality assessment

Tn5p may insert open chromatin regions loosely packed chromatin, which gives rise to transposed DNA fragments with lengths corresponding to mono- or multiple nucleosomes. Thus, a successful ATAC-seq library should exhibit a pattern of fragment size periodicity with intervals of around 200 bp. Additionally, transposed DNA fragments can result from two insertions within nucleosome-free DNA regions. In contrast, if chromatin structures are damaged or perturbed, random transposition can result in a library with various DNA fragment lengths lacking a consistent periodicity. Our Bioanalyzer analysis showed that using 2,500 or 5,000 fresh maize nuclei as input yielded fragment size distributions with the expected ATAC-seq characteristics, i.e., size intervals of ~200 bp (Figures 9B, C). The characteristic periodicity is a measurement to evaluate the Tn5p insertion. Although the fragment periodicity of the library may not guarantee the quality of the sequencing data, this serves as a preliminary indicator before sequencing. The reaction using 500 nuclei exhibited a less pronounced periodicity of peak, implying that Tn5p may become oversaturated in reactions with only 500 nuclei (Figure 9A).

Although our qPCR results suggested that frozen nuclei resulted in more efficient Tn5p transposition (Figure 9A), our Bioanalyzer data revealed that the frozen nuclei sample lacked a clear pattern of periodicity in fragment sizes (Figure 9D), suggesting that the chromatin structure may be disrupted by freezing or thawing. While in Arabidopsis ATAC-seq has been successfully performed using frozen samples, our result (Figure 9) supports that fresh maize samples showed a better periodic pattern than the maize nuclei frozen in liquid nitrogen, suggesting that frozen samples may not be suitable for maize ATAC-seq experiments. Similarly, Bioanalyzer data indicated that transposed fragments from the naked genomic DNA completely lack size periodicity (Figure 9E).

Taken together, these results suggest that 2,500 and 5,000 fresh maize nuclei represent a good starting material for ATAC-seq. In addition, maize ATAC-seq libraries can be generated successfully

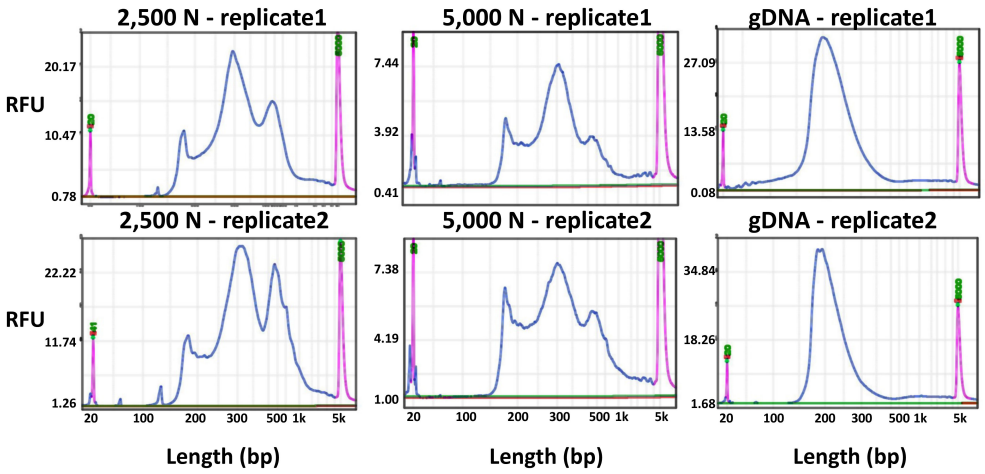


FIGURE 10
Fragment size distribution of ATAC-seq libraries determined by a BiOptic Qsep400 before NGS sequencing. The X-axis represents the product size in base pairs and the Y-axis is the relative fluorescence unit (RFU). Purple peaks indicate size markers.

with an N cycle number lower than typically recommended, as long as the fragment length distribution indicated by Bioanalyzer analysis exhibits the desired characteristic periodicity.

4.5 Validation of ATAC-seq

To evaluate the results of our ATAC-seq protocols, we sequenced ATAC-seq libraries constructed from fresh 2,500 nuclei, 5,000 maize nuclei (two optimized nuclei conditions), and genomic DNA isolated from maize seedlings as a control, each with two biological replicates (Figure 10). We assessed our libraries using three methods: (1) analysis of fragment size distribution by plotting histograms to confirm the hallmark of the characteristic size periodicity of ATAC-seq libraries (Buenrostro et al., 2013) (Figures 10, 11); (2) calculation of three standard metrics of ATAC-seq to assess our libraries' quality (Figures 12, 13), and (3) correlation analysis with maize RNA-seq data from seedling and root tissues (Figure 14).

High quality ATAC-seq data is expected to display a clear fragment size periodicity of approximately 200 bp (Buenrostro et al., 2013), a pattern presumably resulting from two Tn5 transposase insertions flanking single or multiple nucleosomes. To compare fragment size distributions, we aligned the paired-end reads to the maize reference genome (AGPv4) using Bowtie2 as described in step 3.2.2 (Langmead and Salzberg, 2012), and then plotted read frequencies against fragment size. As shown in the top and middle graphs of Figure 11A, both libraries generated from 2,500 and 5,000 nuclei exhibited clear size periodicity, indicating successful ATAC-seq libraries. The fast Fourier transform (FFT) analysis (Figure 11B) confirmed that the periodicity of the fragment length distribution was approximately 192 bp, similar to the expected length of two Tn5p insertions spanning mono-nucleosome protected sequences. In contrast, the library generated from naked genomic DNA lacked the distinct size periodicity (bottom of Figures 11A, B), indicating various fragment lengths were derived from random Tn5p insertions into genomic DNA.

To assess the quality of our ATAC-seq libraries, we examined three established metrics commonly applied in ATAC-seq library evaluation (Landt et al., 2012; ENCODE, 2017, 2020; Schmitz et al., 2022). First, the enrichment analysis at Transcription Start Sites (TSS) +/- 1kb is a useful method to validate the efficacy of ATAC-seq in identifying open chromatin in regulatory regions. The fraction of reads in peaks (FRiP) metric assesses the proportion of all mapped reads located in peak regions, indicating the signal-to-noise ratio. Additionally, the irreproducible discovery rate (IDR) is used to calculate the number of irreproducible peaks between replicates (Li et al., 2011) (Figure 12). For these metrics, we referenced the ENCODE ATAC-seq standards (ENCODE, 2017, 2020) and guidelines from Schmitz (Schmitz et al., 2022), which suggest that for maize, the FRiP score should exceed 45%, and the TSS enrichment value should be no less than 5. In our analysis, we included 16 published maize ATAC-seq libraries for comparison. Our libraries of 2,500 and 5,000 nuclei maintain above-standard scores and either meet or exceed the averages of published maize ATAC-seq datasets (Figure 12). Moreover, we profiled the abundance of ATAC-seq reads around genes. Notably, in the libraries derived from 2,500 and 5,000 nuclei, there is a peak in read abundance at TSS sites (Figures 13A, B), which is apparently different from the library generated from naked gDNA control (Figure 13C).

Since gene expression is partially affected by chromatin accessibility (Hsu et al., 2017; Chang et al., 2018), examining the association between these two factors serves as a valuable assessment for the ATAC-seq data generated using our protocol. As shown in Figure 14A, chromatin accessibility analyzed from our 2,500-nuclei library is positively correlated to RNA transcript abundance obtained from maize seeding and root samples. Interestingly, we observed a slightly stronger correlation between

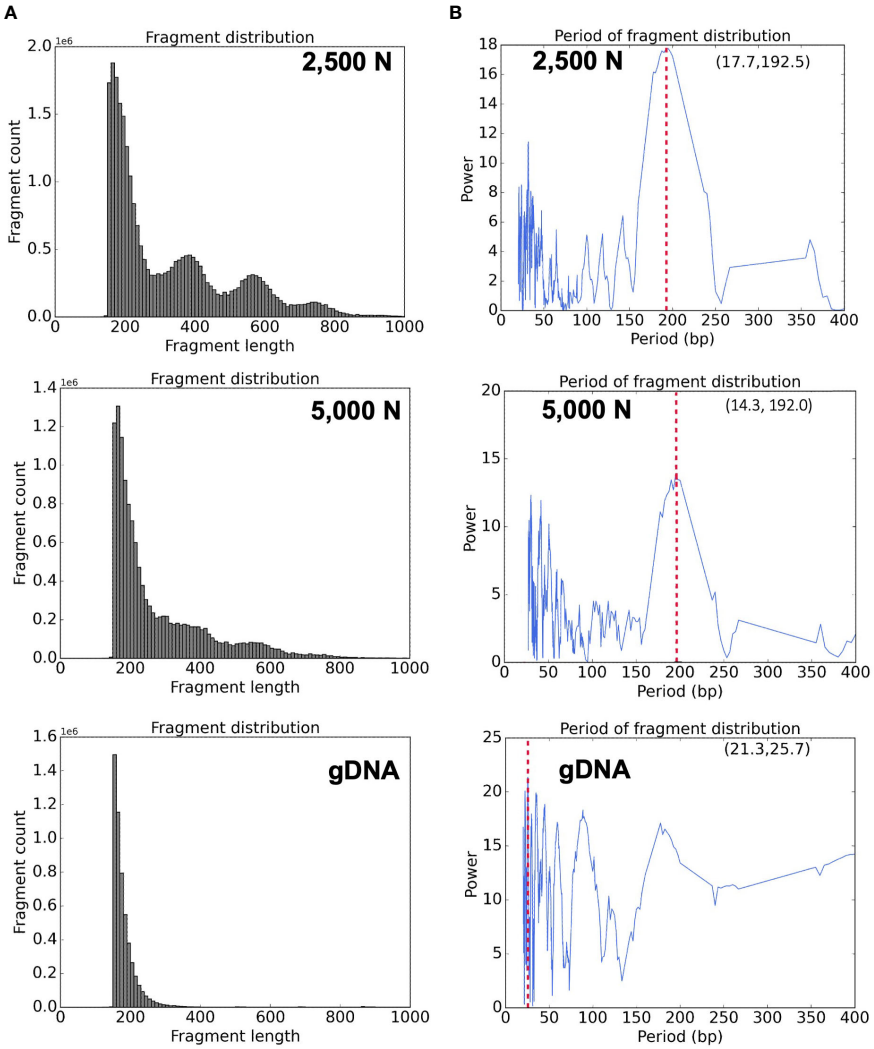


FIGURE 11
Bioinformatic analyses of ATAC-seq library reads. Fragment size distribution of the **(A)** top: 2,500-nuclei library, middle: 5,000-nuclei library, and bottom: gDNA control. The period of fragment length distribution using a fast Fourier transform (FFT) algorithm for the **(B)** top: 2,500-nuclei library, middle: 5,000-nuclei library, and bottom: gDNA control. The red dashed lines in the graph represent the peaks in the period of fragment length distribution.

chromatin accessibility in promoter regions with the transcriptome of seedling sample compared to roots, likely reflecting that our ATAC-seq libraries were generated from seedlings (Figure 14A).

The correlation analysis in promoter and gene body regions (Figures 14A, B) exhibit a similar pattern, suggesting that the open chromatin in promoter regions and gene body are both associated with higher transcript abundance. In contrast, the gDNA ATAC-seq control showed that chromatin accessibility at the promoter and gene body regions exhibited no correlation to gene expression levels.

To have a better sense of the ATAC-seq data from 2500 or 5000 maize nuclei, we showed four genes (lipoxygenases 10, S-adenosyl methionine decarboxylase 2, plasma-membrane H+ATPase2, and nudix hydroxylase 4), including their expression levels and ATAC-seq abundance in our 2500 N and 5000 N libraries using a genome browser (Supplementary Figure 1). These genes, or their family members, have been linked to plant defense and early development (Suzuki and Hirasawa, 1980; Christensen et al., 2013; Liu et al., 2022), suggesting that their open state may enhance gene transcription, leading to improved development and adaptation. We generated peaks distribution across genomic features using ChIPSeeker and ATACgraph. Both tools consistently indicated that peaks in our maize ATAC-seqs occur in promoters, gene body, and intergenic regions (Supplementary Figure 2).

5 Discussion

The most critical step in ATAC-seq protocols is the Tn5p transposition reaction, during which the Tn5p fragments the DNA *in vivo* and tags it with unique Illumina library adaptors. Open chromatin regions have a higher probability of being targeted by Tn5p during the reaction, which is influenced by several critical factors.

First, high purity of nuclei is important, as the Tn5p targets not only genomic DNA, but also mitochondrial and chloroplast DNA.

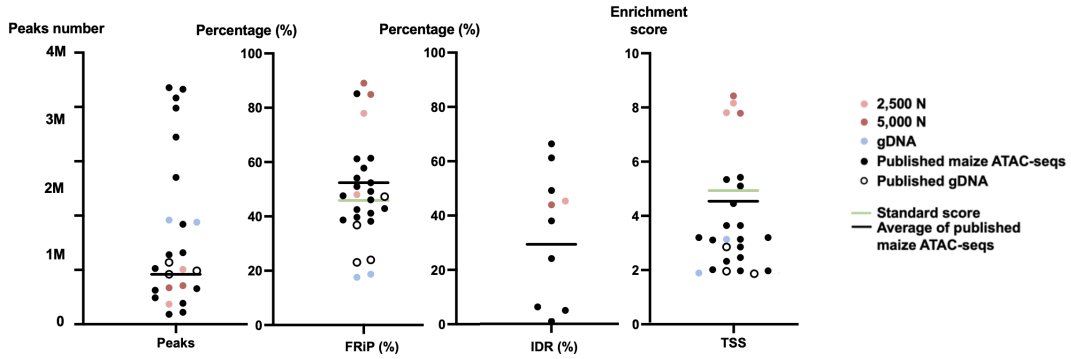


FIGURE 12
The quality metrics of 2,500-nuclei library, 5,000-nuclei library, published maize ATAC-seq libraries and their matched gDNA control. Four new generated ATAC-seq libraries and 16 published ATAC-seq libraries from six seedlings, six leaves, and two ears, six roots, as well as six matched gDNA are included in the comparison. The accession numbers of the included published ATAC-seq libraries are SRR12321693, SRR12321694, SRR12321695, SRR7889829, SRR7889830, SRR7889827, SRR7889828, SRR7889831, SRR7889832, SRR7904001, SRR7904002, SRR7904003, SRR13920264, SRR13920265, SRR6761057, SRR6761058, SRR6761060, SRR6761061, and SRR11955290. The green lines indicate the ENCODE standards for the metrics and the black lines are the average values for the published maize ATAC-seq libraries (excluding gDNA) of each metric. IDR, irreproducible discovery rate; FRiP, fraction of reads in peaks; TSS, transcription start site; 2,500 N, 2,500-nuclei library; 5,000 N, 5,000-nuclei library; gDNA, genomic DNA.

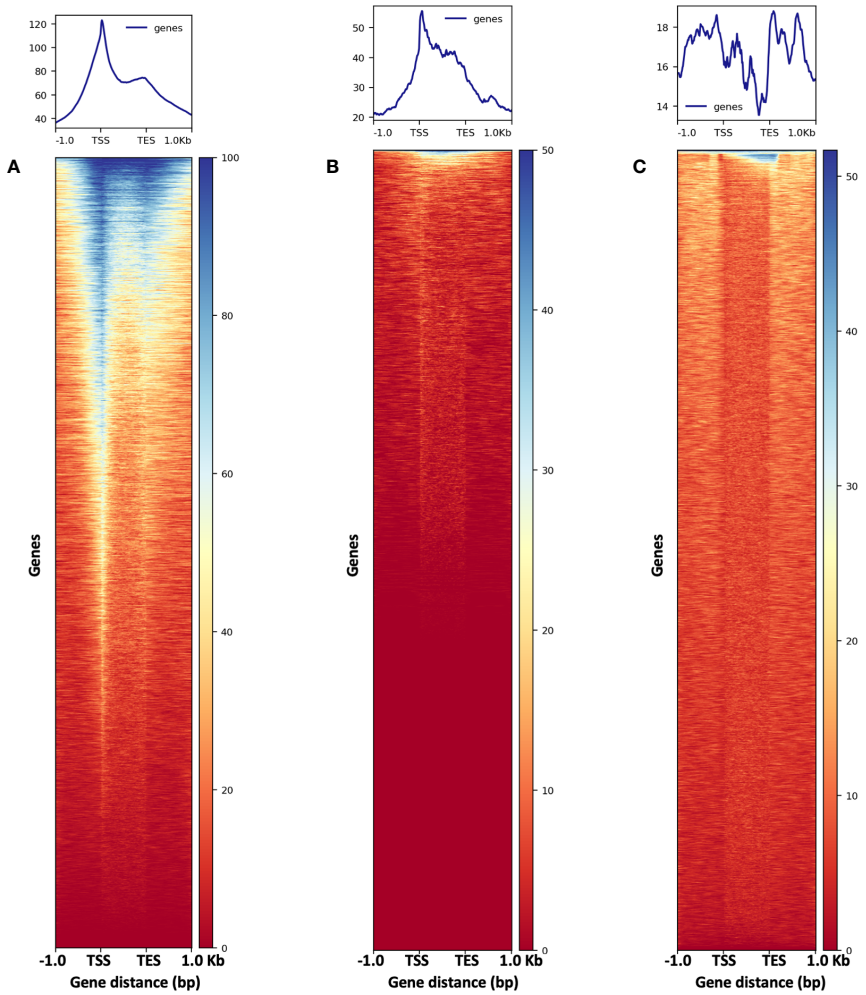


FIGURE 13
The metaplots (top) and heatmaps (bottom) illustrate read abundance around maize genes. ATAC-seq libraries from **(A)** 2,500-nuclei and **(B)** 5,000-nuclei showing peaks at TSS, in contrast to results from gDNA control **(C)**. TSS, transcription start site; TES, transcription end sites.

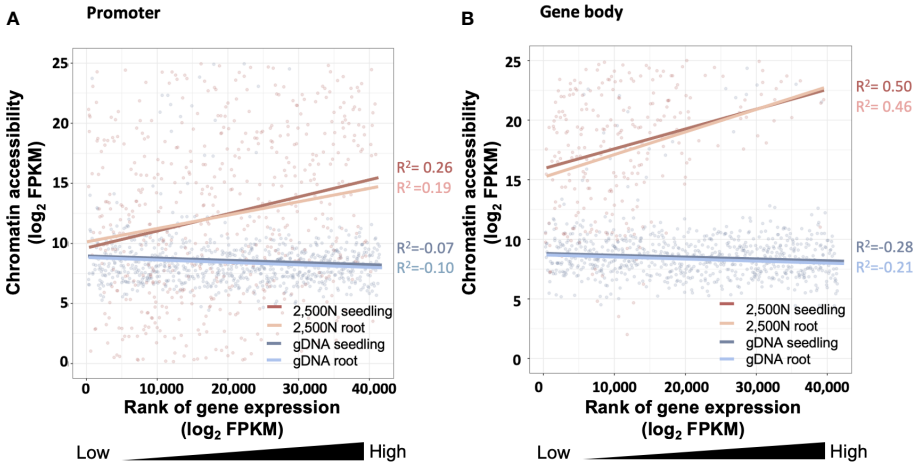


FIGURE 14
The correlation between chromatin accessibility and transcript abundance for the 2,500-nuclei library and gDNA control at the region of **(A)** promoter and **(B)** gene body. The x-axis is the rank of transcript abundance (log2FPKM) and the y-axis represents the chromatin accessibility analyzed from our ATAC-seq data. The lines with different colors are linear regressions of correlations for ATAC-seq libraries with seedling or root (accession number SRR7548392 and SRR2043190) transcriptome data. The coefficient of correlation is tested by Spearman correlation. FPKM, Fragments Per Kilobase Million.

Therefore, it is better to minimize the presence of organelles. In addition, an effective reaction requires Tn5p to be able to attack chromatin efficiently, so the elimination of cell walls and cell debris is necessary to facilitate the accessibility. Second, the native chromatin conformation must be preserved during isolation, as damaged nuclei and disordered chromatin structure could result in distortive results. Last, the ratio of transposase to the number of nuclei should be optimized based on genome size and percentage of open chromatin (i.e., frequency of accessible regions). After an effective and efficient transposition reaction, the resulting fragments tagged with adapters are subjected to two rounds of PCR to generate an ATAC-seq library. In the first round of PCR, these fragments are amplified only by five PCR cycles, during which distinct sequencing barcodes are added. A fraction of this first PCR product is then subjected to quantitative PCR (qPCR) to estimate the relative amount of successfully tagged DNA fragments and then to determine the optimal amplification cycle number for the second round of PCR. Under favorable conditions, an ATAC-seq library is ready for sequencing, when it contains sufficient DNA with appropriate sequence complexity after the second PCR. Since NGS is relatively expensive, applying a proper method for assessing library quality before sequencing can potentially reduce the sequencing expense.

In our testing design, we found that liquid nitrogen-based homogenization is not suitable for maize ATAC-seq analysis, so we instead implement manual chopping for nucleus extraction. This protocol is designed for fresh plant materials using manual chopping, followed by 1% Triton X-100 treatment and 60% Percoll:2.5 M sucrose gradient separation for nucleus extraction. If the materials are stored at −80°C (or in liquid nitrogen) or incorrect concentration of extraction buffer are used, it will lead to a high amount of tissue debris, which can interfere with isolation of intact nuclei.

To conclude, we demonstrated here a protocol that only required a small amount of nuclei without any cell sorting process to identify the open chromatin regions in the plant genome. This strategy shows reliable results even compared to other published maize ATAC-seq libraries. Moreover, this method may be adaptable to other plant species, including crops, thereby contributing to research in plant epigenomics and agriculture.

Data availability statement

The data presented in the study are deposited in the NCBI repository, accession number GSE252638 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE252638).

Author contributions

J-WAH: Formal Analysis, Investigation, Validation, Writing – original draft, Writing – review & editing. P-YL: Methodology, Writing – review & editing, Investigation, Software, Validation, Visualization, Writing – original draft. C-TW: Investigation, Methodology, Writing – original draft. Y-JL: Investigation, Methodology, Writing – original draft. PC: Investigation, Methodology, Writing – original draft. RJ-HL: Investigation, Methodology, Writing – original draft. P-YC: Methodology, Conceptualization, Funding acquisition, Resources, Supervision, Writing – review & editing. C-JRW: Conceptualization, Data curation, Methodology, Supervision, Validation, 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 National Taiwan University and Academia Sinica (NTU-AS Innovative Joint Program: AS-NTU-112-12) to P-YC, and the Ministry of Science and Technology, Taiwan, grant no. 103-2311-B-001-014 and 107-2923-B-002-001-MY4 to C-JRW and grant no. 104-2923-B-001-003-MY2, 106-2311-B-001-035-MY3 and 111-2311-B-001-030 to P-YC.

Conflict of interest

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

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

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2024.1370618/full#supplementary-material

References

Bajic, M., Maher, K. A., and Deal, R. B. (2018). Identification of open chromatin regions in plant genomes using atac-seq. *Methods Mol. Biol.* 1675, 183–201. doi: 10.1007/978-1-4939-7318-7_12

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., and Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, dna-binding proteins and nucleosome position. *Nat. Methods* 10, 1213–1218. doi: 10.1038/nmeth.2688

Buenrostro, J. D., Wu, B., Chang, H. Y., and Greenleaf, W. J. (2015). Atac-seq: A method for assaying chromatin accessibility genome-wide. *Curr. Protoc. Mol. Biol.* 109, 21 29 1–9. doi: 10.1002/0471142727.mb2129s109

Chang, P., Gohain, M., Yen, M. R., and Chen, P. Y. (2018). Computational methods for assessing chromatin hierarchy. *Comput. Struct. Biotechnol. J.* 16, 43–53. doi: 10.1016/j.csbj.2018.02.003

Chen, J., Zeng, B., Zhang, M., Xie, S., Wang, G., Hauck, A., et al. (2014). Dynamic transcriptome landscape of maize embryo and endosperm development. *Plant Physiol.* 166, 252–264. doi: 10.1104/pp.114.240689

Chen, Y., Lun, A. T., and Smyth, G. K. (2016). From reads to genes to pathways: differential expression analysis of rna-seq experiments using rsubread and the edger quasi-likelihood pipeline. *F1000res* 5, 1438. doi: 10.12688/f1000research.8987.2

Chodavarapu, R. K., Feng, S., Bernatavichute, Y. V., Chen, P. Y., Stroud, H., Yu, Y., et al. (2010). Relationship between nucleosome positioning and dna methylation. *Nature* 466, 388–392. doi: 10.1038/nature09147

Christensen, S. A., Nemchenko, A., Borrego, E., Murray, I., Sobhy, I. S., Bosak, L., et al. (2013). The maize lipoxygenase, zmlox10, mediates green leaf volatile, jasmonate and herbivore-induced plant volatile production for defense against insect attack. *Plant J.* 74, 59–73. doi: 10.1111/tpj.12101

Corces, M. R., Trevino, A. E., Hamilton, E. G., Greenside, P. G., Sinnott-Armstrong, N. A., Vesuna, S., et al. (2017). An improved atac-seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods* 14, 959–962. doi: 10.1038/nmeth.4396

Crisp, P. A., Marand, A. P., Noshay, J. M., Zhou, P., Lu, Z., Schmitz, R. J., et al. (2020). Stable unmethylated dna demarcates expressed genes and their cis-regulatory space in plant genomes. *Proc. Natl. Acad. Sci. U.S.A.* 117, 23991–24000. doi: 10.1073/pnas.2010250117

Dai, X., Tu, X., Du, B., Dong, P., Sun, S., Wang, X., et al. (2022). Chromatin and regulatory differentiation between bundle sheath and mesophyll cells in maize. *Plant J.* 109, 675–692. doi: 10.1111/tpj.15586

Deal, R. B., and Henikoff, S. (2011). The intact method for cell type-specific gene expression and chromatin profiling in arabidopsis thaliana. *Nat. Protoc.* 6, 56–68. doi: 10.1038/nprot.2010.175

Deschamps, S., Crow, J. A., Chaidir, N., Peterson-Burch, B., Kumar, S., Lin, H., et al. (2021). Chromatin loop anchors contain core structural components of the gene expression machinery in maize. *BMC Genomics* 22, 23. doi: 10.1186/s12864-020-07324-0

Dong, P., Tu, X., Chu, P. Y., Lu, P., Zhu, N., Grierson, D., et al. (2017). 3d chromatin architecture of large plant genomes determined by local A/B compartments. *Mol. Plant* 10, 1497–1509. doi: 10.1016/j.molp.2017.11.005

Dong, Z., Danilevskaya, O., Abadie, T., Messina, C., Coles, N., and Cooper, M. (2012). A gene regulatory network model for floral transition of the shoot apex in maize and its dynamic modeling. *PloS One* 7, E43450. doi: 10.1371/journal.pone.0043450

ENCODE (2017) Terms And Definitions. Available at: Https://Www.Encodeproject.Org/Data-Standards/Terms/ (Accessed 26 2022).

ENCODE (2020) Atac-Seq Data Standards And Processing Pipeline. Available at: Https://Www.Encodeproject.Org/Atac-Seq/ (Accessed 26 2023).

Galli, M., Khakhar, A., Lu, Z., Chen, Z., Sen, S., Joshi, T., et al. (2018). The dna binding landscape of the maize auxin response factor family. *Nat. Commun.* 9, 4526. doi: 10.1038/s41467-018-06977-6

Gent, J. I., Ellis, N. A., Guo, L., Harkess, A. E., Yao, Y., Zhang, X., et al. (2013). Chh islands: *de novo* dna methylation in near-gene chromatin regulation in maize. *Genome Res.* 23, 628–637. doi: 10.1101/gr.146985.112

Gray, L. T., Yao, Z., Nguyen, T. N., Kim, T. K., Zeng, H., and Tasic, B. (2017). Layer-specific chromatin accessibility landscapes reveal regulatory networks in adult mouse visual cortex. *Elife* 6, e21883. doi: 10.7554/eLife.21883.061

Hsieh, T. F., and Fischer, R. L. (2005). Biology of chromatin dynamics. *Annu. Rev. Plant Biol.* 56, 327–351. doi: 10.1146/annurev.arplant.56.032604.144118

Hsu, F. M., Yen, M. R., Wang, C. T., Lin, C. Y., Wang, C. R., and Chen, P. Y. (2017). Optimized reduced representation bisulfite sequencing reveals tissue-specific mchh islands in maize. *Epigenet. Chromatin* 10, 42. doi: 10.1186/s13072-017-0148-y

Kim, D., Paggi, J. M., Park, C., Bennett, C., and Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with hisat2 and hisat-genotype. *Nat. Biotechnol.* 37, 907–915. doi: 10.1038/s41587-019-0201-4

Kornberg, R. D., and Lorch, Y. (1999). Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell* 98, 285–294. doi: 10.1016/S0092-8674(00)81958-3

Krueger, F., James, F., Ewels, P. E., E., A., Weinstein, M., Schuster-Boeckler, B., et al. (2016) *Felixkrueger/Trimgalore: V0.6.10 - Add Default Decompression Path*. Available at: Https://Github.Com/Felixkrueger/Trimgalore.

Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., et al. (2012). Chip-seq guidelines and practices of the encode and modencode consortia. *Genome Res.* 22, 1813–1831. doi: 10.1101/gr.136184.111

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

Li, Q. H., Brown, J. B., Huang, H. Y., and Bickel, P. J. (2011). Measuring reproducibility of high-throughput experiments. *Ann. Of Appl. Stat* 5, 1752–1779. doi: 10.1214/11-AOAS466

Lin, P.-Y. (2023) Atacgraph2. Available at: Https://Github.Com/Beritlin/Atacgraph2 (Accessed 26 2023).

Liu, Y., Zhang, W., Wang, Y., Xie, L., Zhang, Q., Zhang, J., et al. (2022). Nudix hydrolase 14 influences plant development and grain chalkiness in rice. *Front. Plant Sci.* 13, 1054917. doi: 10.3389/fpls.2022.1054917

Lu, R. J., Liu, Y. T., Huang, C. W., Yen, M. R., Lin, C. Y., and Chen, P. Y. (2020). Atacgraph: profiling genome-wide chromatin accessibility from atac-seq. *Front. Genet.* 11, 618478. doi: 10.3389/fgene.2020.618478

Lu, Z., Hofmeister, B. T., Vollmers, C., Dubois, R. M., and Schmitz, R. J. (2017). Combining atac-seq with nuclei sorting for discovery of cis-regulatory regions in plant genomes. *Nucleic Acids Res.* 45, E41. doi: 10.1093/nar/gkw1179

Lu, Z., Marand, A. P., Ricci, W. A., Ethridge, C. L., Zhang, X., and Schmitz, R. J. (2019). The prevalence, evolution and chromatin signatures of plant regulatory elements. *Nat. Plants* 5, 1250–1259. doi: 10.1038/s41477-019-0548-z

Maher, K. A., Bajic, M., Kajala, K., Reynoso, M., Pauluzzi, G., West, D. A., et al. (2018). Profiling of accessible chromatin regions across multiple plant species and cell types reveals common gene regulatory principles and new control modules. *Plant Cell* 30, 15–36. doi: 10.1105/tpc.17.00581

Makarevitch, I., Eichten, S. R., Briskine, R., Waters, A. J., Danilevskaya, O. N., Meeley, R. B., et al. (2013). Genomic distribution of maize facultative heterochromatin marked by trimethylation of H3k27. *Plant Cell* 25, 780–793. doi: 10.1105/tpc.112.106427

Marand, A. P., Chen, Z., Gallavotti, A., and Schmitz, R. J. (2021). A cis-regulatory atlas in maize at single-cell resolution. *Cell* 184, 3041–3055 E21. doi: 10.1016/j.cell.2021.04.014

Meyer, C. A., and Liu, X. S. (2014). Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nat. Rev. Genet.* 15, 709–721. doi: 10.1038/nrg3788

Montefiori, L., Hernandez, L., Zhang, Z., Gilad, Y., Ober, C., Crawford, G., et al. (2017). Reducing mitochondrial reads in atac-seq using crispr/cas9. *Sci. Rep.* 7, 2451. doi: 10.1038/s41598-017-02547-w

Noshay, J. M., Anderson, S. N., Zhou, P., Ji, L., Ricci, W., Lu, Z., et al. (2019). Monitoring the interplay between transposable element families and dna methylation in maize. *PloS Genet.* 15, E1008291. doi: 10.1371/journal.pgen.1008291

Oka, R., Zicola, J., Weber, B., Anderson, S. N., Hodgman, C., Gent, J. I., et al. (2017). Genome-wide mapping of transcriptional enhancer candidates using dna and chromatin features in maize. *Genome Biol.* 18, 137. doi: 10.1186/s13059-017-1273-4

Ricci, W. A., Lu, Z., Ji, L., Marand, A. P., Ethridge, C. L., Murphy, N. G., et al. (2019). Widespread long-range cis-regulatory elements in the maize genome. *Nat. Plants* 5, 1237–1249. doi: 10.1038/s41477-019-0547-0

Rodgers-Melnick, E., Vera, D. L., Bass, H. W., and Buckler, E. S. (2016). Open chromatin reveals the functional maize genome. *Proc. Natl. Acad. Sci. U.S.A.* 113, E3177–E3184. doi: 10.1073/pnas.1525244113

Schmitz, R. J., Marand, A. P., Zhang, X., Mosher, R. A., Turck, F., Chen, X., et al. (2022). Quality control and evaluation of plant epigenomics data. *Plant Cell* 34, 503–513. doi: 10.1093/plcell/koab255

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

Shashikant, T., and Ettensohn, C. A. (2019). Genome-wide analysis of chromatin accessibility using atac-seq. *Methods Cell Biol.* 151, 219–235. doi: 10.1016/bs.mcb.2018.11.002

Sikorskaite, S., Rajamaki, M. L., Baniulis, D., Stanys, V., and Valkonen, J. P. (2013). Protocol: optimised methodology for isolation of nuclei from leaves of species in the solanaceae and rosaceae families. *Plant Methods* 9, 31. doi: 10.1186/1746-4811-9-31

Song, L., and Crawford, G. E. (2010). Dnase-seq: A high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb. Protoc.* 2010, Pdb Prot5384. doi: 10.1101/pdb.prot5384

Suzuki, Y., and Hirasawa, E. (1980). S-adenosylmethionine decarboxylase of corn seedlings. *Plant Physiol.* 66, 1091–1094. doi: 10.1104/pp.66.6.1091

Valouev, A., Johnson, S. M., Boyd, S. D., Smith, C. L., Fire, A. Z., and Sidow, A. (2011). Determinants of nucleosome organization in primary human cells. *Nature* 474, 516–520. doi: 10.1038/nature10002

Walley, J. W., Sartor, R. C., Shen, Z., Schmitz, R. J., Wu, K. J., Urich, M. A., et al. (2016). Integration of omic networks in A developmental atlas of maize. *Science* 353, 814–818. doi: 10.1126/science.aag1125

Wang, X. F., Elling, A. A., Li, X. Y., Li, N., Peng, Z. Y., He, G. M., et al. (2009). Genome-wide and organ-specific landscapes of epigenetic modifications and their relationships to mrna and small rna transcriptomes in maize. *Plant Cell* 21, 1053–1069. doi: 10.1105/tpc.109.065714

West, P. T., Li, Q., Ji, L., Eichten, S. R., Song, J., Vaughn, M. W., et al. (2014). Genomic distribution of H3k9me2 and dna methylation in A maize genome. *PloS One* 9, E105267. doi: 10.1371/journal.pone.0105267

Wilkins, O., Hafemeister, C., Plessis, A., Holloway-Phillips, M. M., Pham, G. M., Nicotra, A. B., et al. (2016). Egrins (Environmental gene regulatory influence networks) in rice that function in the response to water deficit, high temperature, and agricultural environments. *Plant Cell* 28, 2365–2384. doi: 10.1105/tpc.16.00158

Wolffe, A. P. (1997). Histones, nucleosomes and the roles of chromatin structure in transcriptional control. *Biochem. Soc. Trans.* 25, 354–358. doi: 10.1042/bst0250354

Wu, H., Galli, M., Spears, C. J., Zhan, J., Liu, P., Yadegari, R., et al. (2023). Naked endosperm1, naked endosperm2, and opaque2 interact to regulate gene networks in maize endosperm development. *Plant Cell* 36, 19–39. doi: 10.1093/plcell/koad247

Zhang, Z., and Pugh, B. F. (2011). High-resolution genome-wide mapping of the primary structure of chromatin. *Cell* 144, 175–186. doi: 10.1016/j.cell.2011.01.003

Zhao, H., Zhang, W., Chen, L., Wang, L., Marand, A. P., Wu, Y., et al. (2018). Proliferation of regulatory dna elements derived from transposable elements in the maize genome. *Plant Physiol.* 176, 2789–2803. doi: 10.1104/pp.17.01467

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 01 August 2024
DOI 10.3389/fpls.2024.1391248

OPEN ACCESS

EDITED BY
Umesh K. Reddy,
West Virginia State University, United States

REVIEWED BY
Wang Huasen,
Qingdao Agricultural University, China
Zilhas Ahmed Jewel,
Sejong University, Republic of Korea

*CORRESPONDENCE
Xiaomeng Hao
haoxiaomeng@mail.jnmc.edu.cn

RECEIVED 25 February 2024
ACCEPTED 15 July 2024
PUBLISHED 01 August 2024

CITATION
He S, Xu S, He Z and Hao X (2024) Genome-wide identification, characterization and expression analysis of the *bZIP* transcription factors in garlic (*Allium sativum* L.). *Front. Plant Sci.* 15:1391248. doi: 10.3389/fpls.2024.1391248

Genome-wide identification, characterization and expression analysis of the *bZIP* transcription factors in garlic (*Allium sativum* L.)

Shutao He[1,2], Sen Xu[1], Zhengjie He[3] and Xiaomeng Hao[1]*

[1]Institute of Neurobiology, Jining Medical University, Jining, China, [2]Institute of Biotechnology and Health, Beijing Academy of Science and Technology, Beijing, China, [3]Rehabilitation Department, Traditional Chinese Medicine Hospital of Yanzhou District of Jining City, Jining, China

Introduction: The *bZIP* genes (*bZIPs*) are essential in numerous biological processes, including development and stress responses. Despite extensive research on *bZIPs* in many plants, a comprehensive genome-wide analysis of *bZIPs* in garlic has yet to be undertaken.

Methods: In this study, we identified and classified 64 *AsbZIP* genes (*AsbZIPs*) into 10 subfamilies. A systematic analysis of the evolutionary characteristics of these *AsbZIPs*, including chromosome location, gene structure, conserved motifs, and gene duplication, was conducted. Furthermore, we also examined the nucleotide diversity, cis-acting elements, and expression profiles of *AsbZIPs* in various tissues and under different abiotic stresses and hormone treatments.

Results and Discussion: Our findings revealed that gene replication plays a crucial role in the expansion of AsbZIPs, with a minor genetic bottleneck observed during domestication. Moreover, the identification of cis-acting elements suggested potential associations of *AsbZIPs* with garlic development, hormone, and stress responses. Several *AsbZIPs* exhibited tissue-preferential and stress/hormone-responsive expression patterns. Additionally, *Asa7G01972* and *Asa7G01379* were notably differentially expressed under various stresses and hormone treatments. Subsequent yeast two-hybridization and yeast induction experiments validated their interactions with *Asa1G01577*, a homologue of ABI5, reinforcing their importance in hormone and abiotic stress responses. This study unveiled the characteristics of the *AsbZIP* superfamily and lays a solid foundation for further functional analysis of *AsbZIP* in garlic.

KEYWORDS

genome-wide, garlic, bZIP, abiotic stress, expression profiles

Introduction

Transcription factors (TFs) play pivotal roles in regulating gene expression by recognizing and binding to promoters, which is essential for plant development and stress responses (Singh et al., 2002; Li et al., 2013). The basic leucine zipper (bZIP) family is one of the TF families and is widely present in eukaryotes (Rodriguez-Uribe and O'Connell, 2006; Nijhawan et al., 2008). The bZIP family is characterized by a conserved bZIP domain, usually comprising 60 to 80 amino acids, which contains a basic region and a leucine zipper region (Nijhawan et al., 2008). The basic region, positioned at the N-terminal end of the region, contains a conserved N-x7-R/K motif that is related to nuclear localization and binds to target DNA (Suckow et al., 1994; Correa et al., 2008). The leucine zipper region exhibits a relatively lower level of conservation, consisting of a repetitive sequence comprising leucine or other hydrophobic amino acids (Ile, Val, Phe, or Met). Specifically, 9 of these amino acids are positioned at the C-terminus, forming an amphipathic helix (Jakoby et al., 2002; Wei et al., 2012; Liu and Chu, 2015; Hu et al., 2016b).

The *bZIP* gene family has been comprehensively characterized and investigated at the genomic level across various plant species, including *Arabidopsis thaliana* (Jakoby et al., 2002), *Oryza sativa* (Nijhawan et al., 2008), *Glycine max* (Liao et al., 2008b), *Sorghum bicolor* (Wang et al., 2011), *Zea mays* (Wei et al., 2012), *Vitis vinifera* (Liu et al., 2014b), *Cucumis sativus* (Baloglu et al., 2014), *Maninot esculenta* (Hu et al., 2016b), *Malus domestica* (Zhao et al., 2016), *Brassica napus* (Zhou et al., 2017), *Fragaria ananassa* (Wang et al., 2017), *Daucus carota* (Que et al., 2015) and *Hordeum vulgare* (Pourabed et al., 2015). Previous studies have demonstrated the important role of *bZIPs* in diverse crucial biological processes, such as organ and tissue differentiation (Gangappa et al., 2013; Lozano-Sotomayor et al., 2016; Zhang et al., 2016; Tan et al., 2020), seed storage (Bombarely et al., 2012; Edwards et al., 2017), metabolic activity (Baena-Gonzalez et al., 2007), photomorphogenesis and light signal regulation (Joo et al., 2014; Babitha et al., 2015), salt and drought tolerance (Ying et al., 2012; Liu et al., 2014a), and hormone and sugar signaling processes (Fan et al., 2019; Liu et al., 2020). Specifically, *HY5*, a key transcription factor in light signal transduction, encodes a bZIP protein that regulates root and hypocotyl development in *Arabidopsis thaliana* (Oyama et al., 1997). *AtZIP1* is involved in sugar signaling and influences seed growth and development (Wigge et al., 2005). Additionally, AtbZIP53 can form heterodimers with AtbZIP1, 10 or 25 to promote the transcriptional activation of seed maturation genes (Alonso et al., 2009). *AtZIP17* and *AtZIP24* are crucial in the salt stress response (Liu et al., 2008; Yang et al., 2009). In rice, *OsbZIP46* is strongly upregulated under drought, heat, and abscisic acid (ABA) stresses (Tang et al., 2012). *OsbZIP72* is induced by drought and ABA treatments (Lu et al., 2009). In addition, *SlbZIP33* is involved in stress-induced response and plays a vital role in fruit ripening (Orellana et al., 2010; Bastias et al., 2014). *GmbZIP15* increases the sensitivity of soybean to salt and drought stresses by negatively regulating the gene expression levels of *GmWRKY12* and *GmABF1* (Zhang et al., 2020a). *IbbZIP1* strongly responds to ABA and is related to salt and drought tolerances in sweet potato (Kang et al., 2019). *TabZIP6* is involved in cold tolerance by forming a dimer with two other bZIP proteins belonging to the S subfamily (Cai et al., 2018).

Garlic, originating from Central Asia, the Mediterranean and the Caucasus, has been cultivated for over 5,000 years. Like onion, which is the largest crop in the *Allium* genus, garlic is not only an economically important vegetable and spice, but is also widely applied in the pharmaceutical and nutritional industries (ETOH et al., 2001; Martin and Ernst, 2003; Kamenetsky et al., 2015). At present, due to global warming and the shrinking availability of cultivated land, enhancing the quality of key traits is particularly imperative for garlic breeding. *bZIP* genes play vital roles in numerous physiological processes; thus, comprehensive identification and analysis of *bZIP* gene family members are essential. However, a notable research gap exists in the investigation of *bZIP* genes in garlic. The successful assembly of the garlic chromosome genome provides a solid foundation for these endeavors (Sun et al., 2020).

In this work, 64 *AsbZIP* genes (*AsbZIPs*) were identified and separated into 10 groups on the basis of phylogenetic relationships; these groups were compared with those of *Arabidopsis thaliana* and *Oryza sativa*, which are representative species of dicotyledonous and monocotyledonous plants, respectively. Further comprehensive analyses of *AsbZIPs*, including gene structure, motif analysis, chromosome distribution, evolutionary characteristics and *cis* −acting elements, were conducted. Additionally, we investigated the variation atlas of single-nucleotide polymorphisms (SNPs) in *AsbZIPs*. Finally, we explored the expression profiles of *AsbZIPs* in different garlic tissues and under multiple stresses via quantitative real-time polymerase chain reaction (qRT-PCR). Our study provides a solid foundation for further functional investigations of *AsbZIPs*.

Materials and methods

Identification of *bZIP* genes in the garlic genome

Whole garlic genome data were downloaded from https://doi.org/10.6084/m9.Figshare.12570947.v1, and those of *Arabidopsis thaliana* and *Oryza sativa* were retrieved from the TAIR database (https://www.Arabidopsis.org/) and the RGAP database (https://rice.plantbiology.msu.edu/), respectively. The bZIP domains (PF00170 and PF07716) were retrieved from the PFAM database (http://pfam.xfam.org) (El-Gebali et al., 2019) and used to perform an HMM (hidden Markov model) search via the HMMER 3.0 program. NCBI-CDD (https://www.ncbi.nlm.nih.gov/cdd/) and SMART (http://smart.embl.de/) were used to further confirm the bZIP domain of potential *bZIP* gene family members. The ExPASy proteomics server (https://web.expasy.org/computepi/) was used to calculate the molecular weight (MW), isoelectric point (pI), instability index (I.I.), aliphatic index (A.I.), total number of negatively charged residues (Asp + Glu, n.c.r, %), total number of positively charged residues (Arg + Lys, p.c.r, %), and grand average hydropathicity (GRAVY) of the bZIP proteins in garlic. The subcellular localizations of the bZIP proteins were assessed via the Cell-PLoc 2.0 web server (http://www.csbio.sjtu.edu.cn/bioinf/plant-multi/).

Phylogenetic analysis

The bZIP protein sequences from *Allium sativum*, *Arabidopsis thaliana*, and *Oryza sativa* were aligned via MUSCLE with default parameters. A phylogenetic tree for these proteins was established via the neighbor-joining (NJ) method in MEGA 7.0 software (Kumar et al., 2018), with 1000 iterations. The iTOL online software tool (https://itol.embl.de/) was used to output visual images.

Gene structure and conserved motif analysis

The Gene Structure Display Server (GSDS) (http://gsds.gao-lab.org/) (Hu et al., 2015) was employed to identify the exon and intron structures of all *AsbZIPs*. The MEME program (https://meme-suite.org/meme/) (Bailey et al., 2009) was used to investigate conserved motifs. The minimum and maximum lengths of the conserved motifs were set to 6 and 50, respectively, with a maximum of 10 conserved motifs.

Chromosomal localization and gene duplication analysis of *AsbZIPs*

The chromosomal positions of the *AsbZIP* genes were visualized via MapChart software (Stajich et al., 2002) according to the garlic genome annotation (https://doi.org/10.6084/m9.Figshare.12570947.v1). Using the Multiple Collinear Scan Toolkit (MCScanX), we investigated the gene duplication events of the *AsbZIP* genes (Wang et al., 2012). A tandem duplication event is a chromosome region with two or more adjacent genes within a 200 kb range, and duplicated pairs positioned on different chromosomes are defined as segmental duplication events (Cannon et al., 2004). Syntenic analysis of *bZIP* genes between garlic and four plant species (*Arabidopsis thaliana*, *Oryza sativa*, *Zea mays*, and *Theobroma cacao*) was conducted via Dual Synteny Plotter software (https://github.com/CJChen/TBtools) (Chen et al., 2020). The calculation of synonymous (Ks) and nonsynonymous (Ka) substitution rates and Ka/Ks ratios of each duplicated gene pair were carried out based on the coding sequences (CDS) alignments of *bZIP* genes via Ka/Ks calculator 2.0 software, and a Ka/Ks ratio > 1 was interpreted as positive selection, < 1 as purified selection, and = 1 as neutral evolution (Nei and Gojobori, 1986; Wang et al., 2010).

Analysis of *cis*-acting elements

Cis-acting elements within the 2000 bp region upstream of the transcriptional start site of each *AsbZIP* gene were identified via the PlantCARE database (http://bioinformatics.psb.ugent.be/webtools/plantcare/html) (Chow et al., 2019).

Nucleotide variation and population structure

The resequencing data of 233 garlic samples were retrieved from the Genome Variation Map (project accession: PRJCA006629). Supplementary Table 1 presents the geographic distribution and detailed material information. SnpEff v4.3 was utilized for the annotation of SNPs (Cingolani et al., 2012). Additionally, the population structure was analyzed via ADMIXTURE v1.3.0, with K values ranging from 2 to 5. The phylogenetic tree was established by Treebest v1.9.2, and the Smartpca within EIGENSOFT v4.2 was used for the analysis of principal component analysis (PCA). Nucleotide diversity (π) and Wright's F statistic (Fst) were calculated using VCFtools v0.1.16.

Plant materials and various stresses and hormone treatments

Garlic cloves (cv. Ershuizao) were planted in pots and cultivated in a chamber (16 h/8 h of light/dark, 30°/22° day/night). The treatment experiments were performed as described previously (Zhang et al., 2022b; Munim Twaij et al., 2023; Yao et al., 2023; Shi et al., 2024) with some modifications. The salt, cold and heat stress experiments were simulated with a 200 mM NaCl solution, at 4° and 30°, respectively. Five hormone treatments, including ABA (1 mg/L), GA3 (gibberellic acid, 200mg/L), MeJA (methyl jasmonic acid, 100 μM), IAA (indoleacetic acid, 100 μM), and SA (salicylic acid, 100 μM)), were conducted. Root and leaf samples were collected at 0, 6, 12, 24 and 48 hours after each treatment and promptly frozen in liquid nitrogen. For the drought treatment, irrigation was ceased for the seedlings in the treatment group, and those in the control group were irrigated normally. Roots and leaves were harvested at 0, 7 and 14 days after treatment. Additionally, freshly harvested garlic cloves were stored at 4°C and collected at 0, 10, 15, and 40 days after treatment. To investigate the tissue-specific expression profiles of the *AsbZIPs*, leaves, stems, pseudostems, roots, buds, bulbs and flowers were harvested at 192 days after planting. All samples were promptly frozen in liquid nitrogen and then preserved at -80°C for subsequent mRNA extraction.

RNA isolation and qRT-PCR analysis

RNA extraction kits (Vazyme, Nanjing, China) were used to extract total RNA according to the manufacturer's instructions. Subsequently, reverse transcription of two micrograms of RNA was conducted via HiScript III RT SuperMix for qPCR (Vazyme, Nanjing, China). The qRT-PCR assay was conducted as previously described (Xu et al., 2013). Supplementary Table 2 shows the primers designed with Primer Express software (v3.0). The *tubulin* gene was used as the internal control gene, and the expression level of each gene was determined via the $2^{-\Delta\Delta Ct}$ method (Livak and Schmittgen, 2001).

Gene regulatory network analysis

The co-expression network was constructed via 185 RNA-Seq datasets retrieved from the Gene Expression Omnibus (GEO) database (accession codes GSE211495, GSE186042, and GSE145455) and the Sequence Read Archive (SRA) database (accession codes PRJNA682570, PRJNA472416, and PRJNA683607). The high-quality reads were filtered via the NGS QC Toolkit (v2.3). TopHat (v2.0.0) was used to align these filtered reads to the garlic genome with default parameters. Then, the FPKM values and read counts of each garlic gene were calculated via Cufflinks (v2.0.2). To identify genes co-expressed with *AsbZIP* genes, weighted gene co-expression network analysis (WGCNA) was used to construct a co-expression network. For further investigation, genes within the top 5% highest weighted values associated with *AsbZIP* genes were selected.

Furthermore, to determine directional interactions in the transcriptional regulatory network related to *AsbZIP* genes, 2-kb upstream sequences of co-expressed genes of each *AsbZIP* gene were analyzed, and FIMO (Grant et al., 2011) was utilized to screen genes whose promoters included a significantly enriched motif of corresponding *AsbZIP* gene according to the high-quality TF binding motifs retrieved from the PlantTFDB database. A motif was regarded as present in a promoter if it had at least one match at a P value ≤ 10^{-4}. The clusterProfiler package in R was used to perform GO enrichment analysis of the filtered co-expressed genes of each *AsbZIP* gene.

Prediction of the AsbZIP protein interaction network

STRING (https://cn.string-db.org/) was used to construct an interaction network for AsbZIP proteins according to their orthologous proteins in *Arabidopsis thaliana*.

Heterologous expression in yeast and yeast two-hybrid

To investigate the functions of *Asa7G01972* and *Asa7G01379* in salt stress, a specific primer pair for each gene (Supplementary Table 2) was used to amplify the DNA fragment, which was cloned in frame into the pYES2 recombinant vector. The recombinant and empty vector plasmids were separately transformed into yeast cells (BY4741) following the manufacturer's protocol for the Yeastmaker™ Yeast Transformation System 2 (Clontech Laboratories, Inc., Palo Alto, CA, USA). Yeast transformants were cultivated in SD/-Ura liquid medium at 30° until the OD_{600} reached 0.5. The preculture was transferred to SG (-Ura, 2% galactose) and diluted to an OD_{600} of 0.4, followed by incubation with shaking for an additional 24 hours at 30°C to induce gene expression. Subsequently, the yeast cells were harvested for subsequent stress treatments (Ibrahim et al., 2001).

The Y2H experiment was conducted according to the manufacturer's instructions (Clontech Laboratories, Inc., Palo Alto, CA, USA). The CDS of *Asa7G01972* or *Asa7G01379* was inserted into the pGBKT7 vector to form the bait construct, whereas that of *AsaABI5* (geneID: *Asa1G01577*) was cloned and inserted into the pGADT7 vector to generate the prey construct. The primers used for the Y2H assay are presented in Supplementary Table 2. These plasmids were then transformed into yeast cells (strain AH109), and the resulting yeast transformants were cultured on SD medium supplemented with 3-aminotriazole and X-α-gal but lacking tryptophan, leucine, histidine, and adenine at 30°C for 2–3 days.

Statistical analysis

The samples were harvested from three independent plants. Data from at least three replicates are shown as the means ± SDs. The statistical analysis, including Student's t test, was conducted via SPSS software (version 17, SPSS Inc., Chicago, IL, USA). A significance criterion of $P < 0.05$ indicated statistical significance.

Results

Identification and characterization of *bZIP* genes in garlic

A total of 64 *AsbZIP* genes were identified via hmm search and confirmed via the NCBI-CDD and SMART databases (Supplementary Table 3). Apart from 3 genes (*Asa0G04894*, *Asa0G01277*, and *Asa0G02642*) located on unassigned scaffoldings, the remaining 61 genes were randomly distributed across 8 chromosomes. Specifically, chromosome 4 had the greatest number of genes, with 12, whereas chromosome 3 had the lowest number of genes, with only 4.

Gene characteristics were further analyzed. The CDS lengths of the *AsbZIP* genes ranged from 240 bp (*Asa2G04074*) to 1758 bp (*Asa8G00330*), with predicted molecular weights (MWs) ranging from 9.24256 to 64.00282 kDa. Asa7G01972 and Asa7G01871 presented the extremes in terms of isoelectric point (pI), with values of 4.89 and 11.24, respectively, indicating the lowest and highest pIs within the set. All the *AsbZIP* proteins had consistently negative GRAVY values, indicating their hydrophilic nature, which was coincident with their presumed roles as transcription factors. Apart from 8 genes (*Asa3G03771*, *Asa4G00594*, *Asa4G00875*, *Asa4G04645*, *Asa5G05838*, *Asa7G02467*, *Asa7G05774*, and *Asa8G00330*), the majority of proteins exhibited instability. Furthermore, all the AsbZIP proteins were localized in the nucleus.

Phylogenetic analysis and classification of *AsbZIPs*

bZIP protein sequences from garlic (64), *Arabidopsis thaliana* (78), and *Oryza sativa* (88) were used to construct a phylogenetic tree

10.3389/fpls.2024.1391248

via the NJ method, delineating their evolutionary relationships (Figure 1). A total of 230 bZIPs from these species were divided into 10 subfamilies, labelled as Groups A, B, C, D, E, F, G, H, I, and S, on the basis of their classification in *Arabidopsis thaliana* (Jakoby et al., 2002). The subfamilies varied in size, with the largest encompassing 47 members (S group) and the smallest containing 9 members (B group). Notably, all the species contributed members to each identified subfamily.

Gene structure and conserved motif analysis

The structure of exons and introns provides pivotal evidence for discerning phylogenetic relationships within gene families (Li et al., 2006). The number and distribution of exons and introns of *AsbZIPs* were investigated (Figures 2A, B). The results revealed that 17 *AsbZIPs*, constituting 27% of the total, lacked introns, and Groups D and I had the most abundant intron-lacking *AsbZIP* genes, with four. Among *AsbZIPs* with introns, the number of introns ranged from 1 to 11. The greatest number of introns among the *AsbZIP* genes was 11, and these genes belonged to Groups A, D, H, S and I. The *AsbZIPs* in Groups C and G contained either 0 or 3 introns, whereas those in Group E had either 0 or 1 introns. Generally, *AsbZIP* genes within the same subfamily presented similar gene structures.

Additionally, to gain insight into the divergence and characterization of AsbZIP proteins, 10 conserved motifs were identified (Figures 2A, C). These motifs varied in length, ranging from 21 (motif 6) to 50 (motifs 2, 3, 4, 5, 8 and 9) (Supplementary Table 4). The majority (89.06%) of the AsbZIP proteins presented the prevalent presence of motifs 1 and 6. Motif 8 was present in Groups A, B, D and S; motif 9 was found in Groups H, I and S; and motif 10 occurred in Groups A, B and S. Numerous motifs were found in particular groups, suggesting potential associations with distinct biological functions.

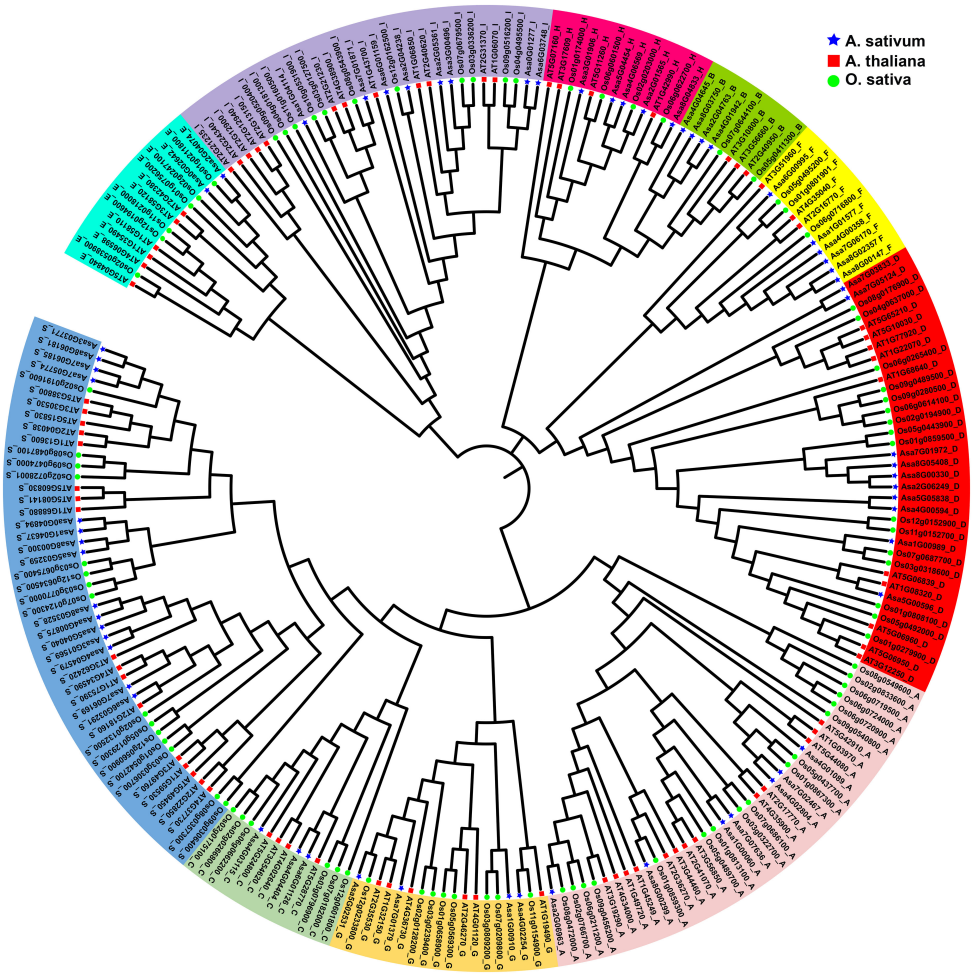


FIGURE 1
Phylogenetic relationships of the identified bZIP proteins in *Allium sativum*, *Arabidopsis thaliana*, and *Oryza sativa*. The colored regions represent *bZIP* genes of different subfamilies A to I, S. The blue stars indicate the AsbZIP proteins, the red squares represent the AtbZIP proteins, the green circles represent the OsbZIP proteins.

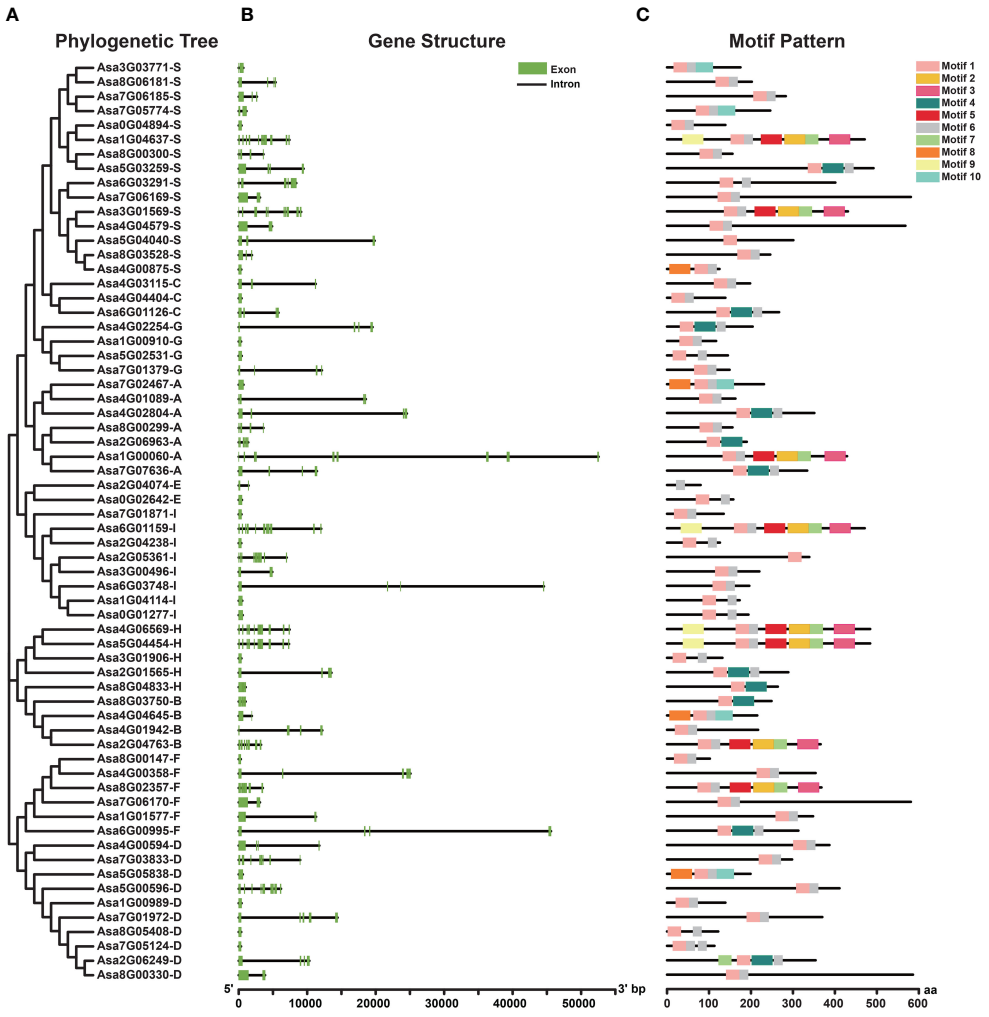


FIGURE 2
Phylogenetic relationships, gene structures and conserved protein motifs of the *AsbZIP* superfamily in garlic. **(A)** Phylogenetic relationships of 64 AsbZIP proteins. **(B)** Exon-intron structure of *AsbZIP* genes. Green boxes indicate exons, and black lines indicate introns. **(C)** The motif profile of AsbZIP proteins. These motifs are presented in differently colored boxes numbered 1-10. The sequence information of each motif is provided in Supplementary Table 4. The protein length can be estimated using the scale at the bottom.

Chromosomal distribution, gene duplication events and synteny analysis of *bZIPs* in garlic

A total of 61 *AsbZIP* genes were unevenly distributed on 8 chromosomes, and three *AsbZIP* genes (*Asa0G01277*, *Asa0G02642*, and *Asa0G04894*) were present on the scaffolds (Figure 3). Chromosome 4 had the most genes, with 12 genes, followed by chromosome 7 (11 genes), chromosome 8 (10 genes) and chromosome 2 (7 genes). Chromosome 1 and chromosome 5 each featured an identical number of 6 *AsbZIP* genes. Chromosome 3 had the lowest number of *AsbZIP* TFs (4 genes).

To explore the evolutionary patterns of the *AsbZIP* genes, tandem and segmental duplication events were analyzed. Two tandem duplication regions, *Asa7G06170* and *Asa7G06169*, *Asa8G00299* and *Asa8G00300*, are located on chromosomes 7 and 8, respectively (Figure 3; Supplementary Table 5). Furthermore, seven pairs of segmental duplicated genes (*Asa4G00594* and *Asa7G01379*, *Asa5G00596* and *Asa7G01871*, *Asa6G03748* and *Asa7G01871*, *Asa4G03115* and *Asa8G04833*, *Asa5G00596* and *Asa8G04833*, *Asa5G04040* and *Asa8G00330*, *Asa8G03528* and *Asa8G00330*) were identified; these pairs were associated with chromosomes 4 and 7, chromosomes 5 and 7, chromosomes 6 and 7, chromosomes 4 and 8, chromosomes 5 and 8, chromosomes 5 and 8, and chromosome 8, respectively (Figure 4; Supplementary Table 5). These observations strongly suggest that tandem and segmental duplication played a significant role in the expansion of *AsbZIP* genes.

To delve into the evolutionary restrictions of *AsbZIP* genes, the Ka *vs*. Ks substitution ratios were analyzed (Supplementary Table 5). The Ka/Ks ratios between *Asa7G06170* and *Asa7G06169*, *Asa8G00299* and *Asa8G00300*, *Asa8G00330* and *Asa8G03528*, *Asa8G00330* and *Asa5G04040*, *Asa7G01871* and *Asa6G03748*, and *Asa7G01871* and *Asa5G00596* were less than 1, implying purifying selection, whereas those between *Asa8G04833* and *Asa5G00596*, *Asa8G04833* and *Asa4G03115*, and *Asa7G01379* and *Asa4G00594* were greater than 1, suggesting positive selection.

10.3389/fpls.2024.1391248

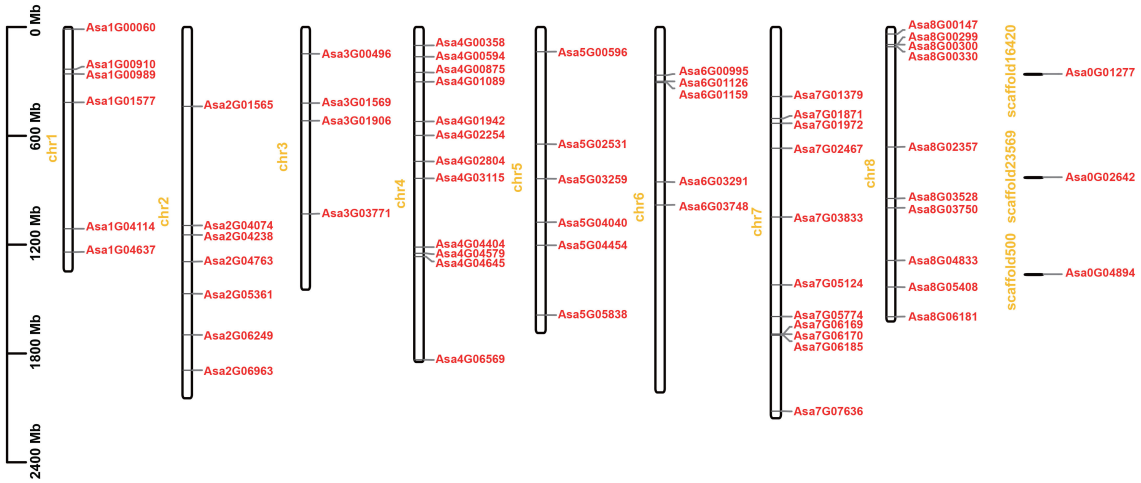


FIGURE 3
Chromosomal distribution of *bZIP* genes in garlic. The chromosome number is marked on the left of each chromosome in yellow. Chromosome lengths and gene positions can be inferred from the scale on the left.

Evolutionary analysis of *AsbZIP* genes and several other species

To further elucidate the phylogenetic mechanisms of the *AsbZIP* family, we established syntenic maps of garlic compared with those of four other representative species: two monocotyledonous plants (*Oryza sativa* and *Zea mays*) and two dicotyledonous plants (*Arabidopsis thaliana* and *Theobroma cacao*) (Figure 5). Twenty-two *bZIP* genes in garlic exhibited collinearity associations with those in *Oryza sativa* (19 pairs), *Zea mays* (12 pairs), *Arabidopsis thaliana* (15 pairs), and *Theobroma cacao* (2 pairs). Additionally, 19 gene pairs consisting of only 6 *AsbZIPs* were detected between garlic and *Oryza sativa*, and *Zea mays*, indicating that these orthologous pairs appeared after the divergence between monocotyledonous and dicotyledonous plants.

Nucleotide variation and population structure of *AsbZIP* genes

Publicly available garlic resequencing data were used to explore *AsbZIP*-related SNPs, revealing the intricate sequence diversity of *AsbZIP* genes. The dedicated SNP calling pipeline successfully generated 2941 SNPs, each with high confidence (Supplementary Table 6). A predominant portion of *AsbZIP*-associated SNPs were enriched in intergenic regions, whereas others occurred within genic regions, including 4 missense, 40 intron, 1 synonymous, 43 downstream, and 70 upstream variants. The overall transition/transversion (Ts/Tv) ratio was 1.938, with G/A (21.73%) and C/T (19.25%) as the predominant allelic substitution patterns, indicating a higher frequency of mutations involving purine-to-purine or pyrimidine-to-pyrimidine transitions than mutations that switch pyrimidines to purines or purines to pyrimidines.

PCA was performed using *AsbZIP*-related SNPs to explore the interrelationship among the origin group and three distinct garlic cultivars (Figure 6A). The first eigenvector, which explained 54.12% of the total genetic variance, indicated divergence within these populations. The second and third eigenvectors, contributing 17.32% and 9.06% to genetic variation, respectively, facilitated the differentiation among these groups. The phylogenetic tree displayed consistent population affinities (Figure 6B). In accordance with the phylogenetic tree, ADMIXTURE analysis also validated the consistent group relatedness. At K=4, a clear demarcation according to geographical origin was observed. The existence of a genetic mixture between wild and landrace garlic indicated a possible domestication of cultivated garlic and the ongoing gene flow between wild and landrace garlic.

To assess the occurrence of genetic bottlenecks in *AsbZIPs* in garlic during acclimation, the population-based nucleotide diversity of *AsbZIP* genes was calculated. *AsbZIP* genes presented a slight decrease of only 0.00003% in nucleotide diversity from wild garlic to local garlic (Supplementary Figure 1), suggesting a weak genetic bottleneck during domestication. Additionally, we measured the degree of population differentiation via Wright's F statistic. The calculated Fst index of 0.2235 between wild and landrace garlic within the *AsbZIPs* indicated that this gene family experienced relatively mild selective pressures during the domestication of garlic.

Analysis of *Cis*–acting elements in *AsbZIP* promoters

Regulatory mechanisms potentially regulating *AsbZIP* genes were investigated by analyzing the *cis*-elements present in their promoters. The results revealed that these *cis*-elements could be classified into three groups: development, hormone and stress-associated elements (Figure 7; Supplementary Figure 2). Several development-related *cis*-elements, such as the CAT-box (associated with meristem development), O2-site (associated with zein metabolism regulation), circadian (associated with circadian control), GCN4_motif (associated with endosperm development), and MSA-like (associated with cell cycle regulation), were located on the promoters of *AsbZIPs*. Ten hormone-responsive regulatory

10.3389/fpls.2024.1391248

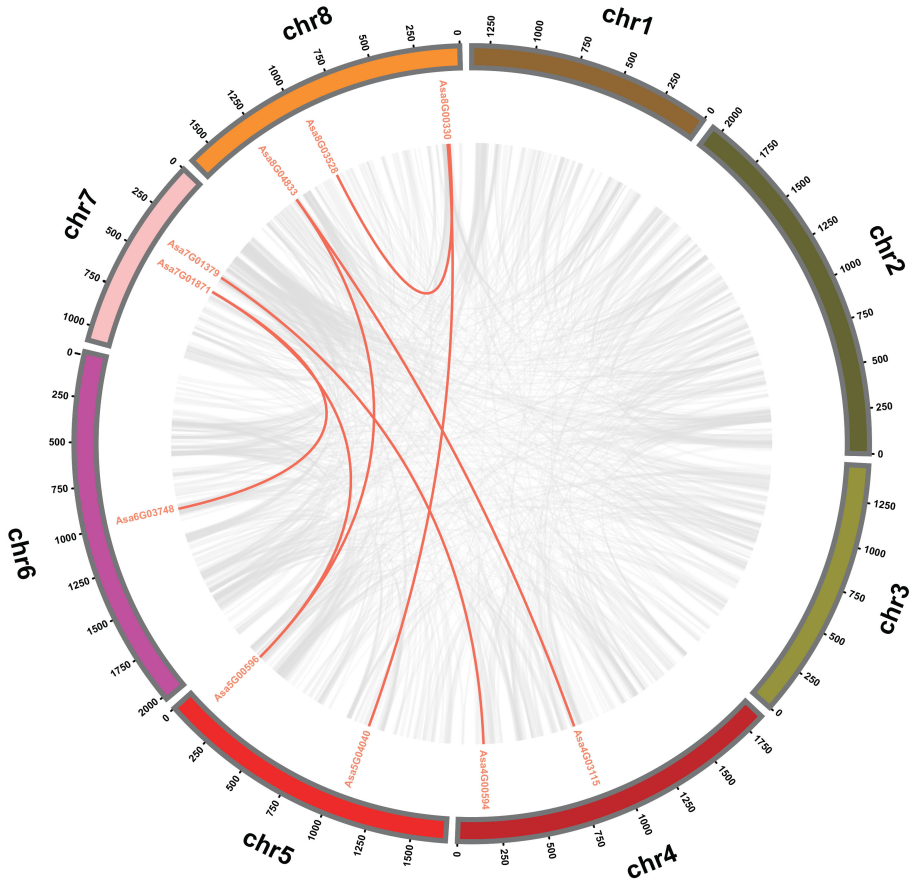


FIGURE 4
Interchromosomal relationships of the *bZIP* genes in garlic. Different-colored circles represent distinct chromosomes, with grey lines within the circles representing all syntenic blocks in the garlic genome and red lines within the circles indicating collinear blocks among *AsbZIP* genes.

elements associated with ABA, IAA, GA, MeJA, and SA responses, including ABRE (29.48%), TGA-element (8.25%), AuxRR-core (2.59%), TGA-box (0.71%), P-box (5.19%), GARE-motif (3.54%), TATC-box (2.83%), TGACG-motif (19.81%), CGTCA-motif (19.81%) and TCA-element (7.78%), were discovered in the promoters of *AsbZIPs*. Additionally, seven stress-responsive regulatory elements that respond to various stress conditions, such as anaerobic induction, anoxic-specific inducibility, defense and stress responsiveness, drought inducibility, low-temperature responsiveness, salt responsiveness and wound-responsiveness, were identified, including ARE (47.74%), GC-motif (3.01%), TC-rich repeats (8.65%), MBS (20.3%), LTR (19.17%), DRE (0.38%) and WUN-motif (0.75%). These findings indicate that the transcriptional regulation of *AsbZIPs* is associated with development, hormone, and stress.

Expression patterns of *AsbZIPs* in multiple tissues of garlic

To investigate the function of the *AsbZIP* genes in growth and development, 14 *AsbZIP* genes were randomly selected, and their expression levels in various tissues (bulb, bud, leaf, pseudostem, sprout, root and flower) were determined. As shown in the Supplementary Figures 3A, 4 (*Asa6G00995*, *Asa8G02357*, *Asa2G06963* and *Asa4G03115*), and 2 (*Asa8G06181* and *Asa7G01379*) *AsbZIP* genes were relatively highly expressed in roots and leaves, respectively. *Asa6G00995* presented high expression levels in roots, whereas its expression was relatively low in buds. The expression trend of *Asa3G01906* was opposite to that of *Asa6G00995*; *Asa3G01906* presented elevated expression levels in buds, with lower expression levels observed in roots. Additionally, *Asa8G04833* exhibited high expression in flowers, whereas the expression level of *Asa4G00594* was relatively high in bulbs. These distinct expression patterns suggest various roles of these genes in garlic development.

Expression patterns of *AsbZIPs* under abiotic stress

Given the crucial regulatory roles of bZIP transcription factors in response to adverse conditions, it is imperative to study the expression of *AsbZIPs* under various abiotic stress conditions. Figures 8A–D illustrate the expression profiles of *AsbZIP* genes under salt, drought, heat and cold stresses, respectively. Under salt

10.3389/fpls.2024.1391248

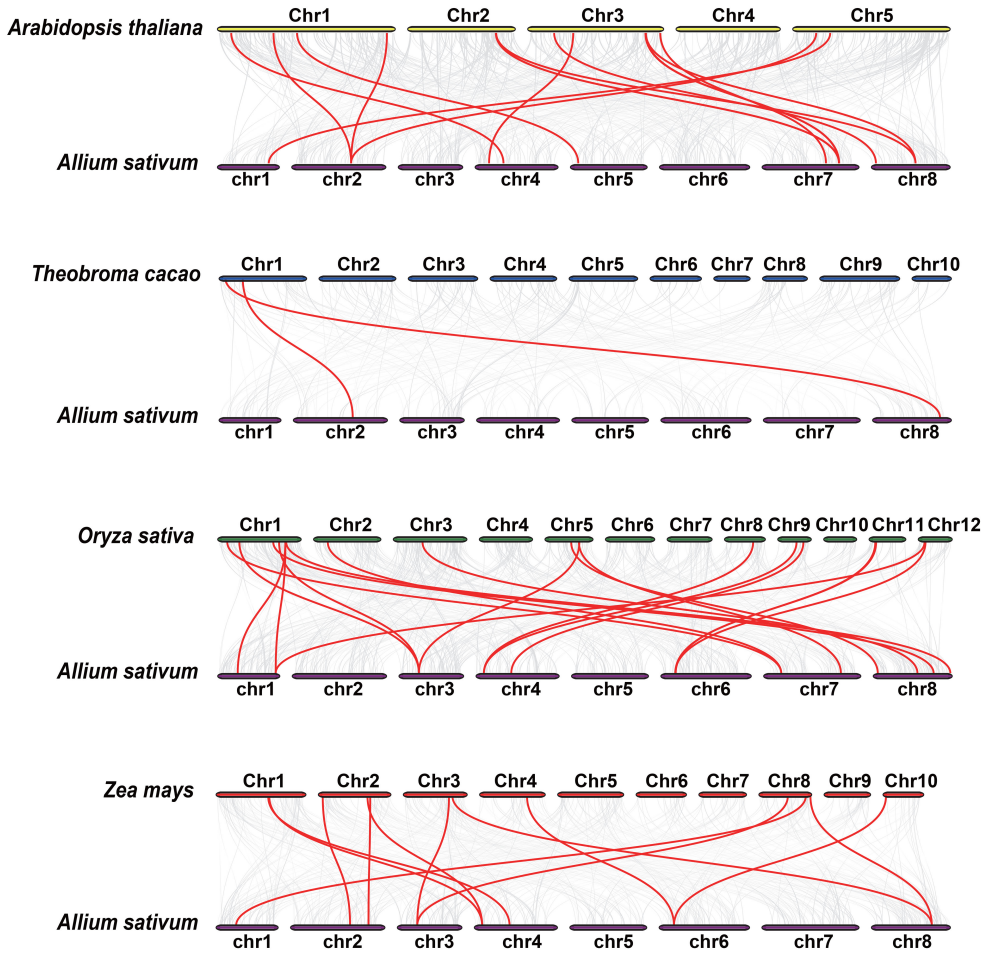


FIGURE 5
Syntenic analysis of *bZIP* genes between garlic and four representative plant species. Gray lines in the background represent collinear blocks between garlic and other plant genomes, whereas red lines highlight syntenic *bZIP* gene pairs.

stress, *Asa4G00594*, *Asa7G01379*, *Asa2G06963* and *Asa7G01972* were significantly upregulated in both leaves and roots at different time points. *Asa6G00995*, *Asa8G02357*, *Asa2G04763*, *Asa7G02467* and *Asa5G05838* presented induced expression and peaked at 12 h in leaves but presented suppressed expression in roots under salt stress. In contrast, *Asa3G01906*, *Asa8G04833* and *Asa4G03115* were downregulated in leaves and upregulated in roots at different time points. Additionally, *Asa7G01972* presented the greatest change in gene expression level in both leaves (peaked at 12 h) and roots (peaked at 6 h) under salt stress, followed by *Asa7G01379*, which peaked at 12 h in both leaves and roots. Under drought stress, *Asa4G00594*, *Asa3G01906*, *Asa7G01379*, *Asa7G01972* and *Asa5G05838* were upregulated and peaked at 14 d, whereas *Asa6G00995*, *Asa8G06181*, *Asa2G06963*, *Asa7G02467*, *Asa4G03115* and *Asa8G05408* were downregulated at distinct time points in both roots and leaves, implying their pivotal roles in response to drought stress in distinct tissues. Furthermore, *Asa7G01972* and *Asa7G01379* were the genes whose expression was most predominantly induced in leaves and roots, respectively. Under heat stress, with the exception of *Asa8G04833* and *Asa2G06963*, whose expression decreased, the other genes were induced in the leaves. Moreover, *Asa8G06181* and *Asa7G01379* presented the greatest induction at 24 h in the leaves and roots, respectively. Under cold treatment, *Asa4G00594*, *Asa6G00995*, *Asa3G01906*, *Asa7G01379*, *Asa2G04763*, *Asa8G04833*, *Asa7G02467*, *Asa7G01972* and *Asa8G05408* presented increased expression in the leaves. Among these genes, the expression of *Asa4G00594*, *Asa6G00995*, *Asa7G01379*, *Asa8G04833*, *Asa7G02467* and *Asa7G01972* peaked at 6 h. Seven genes (*Asa4G00594*, *Asa6G00995*, *Asa3G01906*, *Asa7G01379*, *Asa8G04833*, *Asa7G02467* and *Asa8G05408*) were upregulated in the roots at all time points after cold treatment, with *Asa3G01906* exhibiting the highest expression levels. In addition, under cold stress, the majority of *AsbZIPs* were induced in garlic cloves, except for *Asa7G01379*, *Asa8G04833*, and *Asa5G05838* (Supplementary Figure 3B). Among these genes, the highest induction level was found in *Asa2G06963* at 40 d, while *Asa5G05838* was the most significantly inhibited gene. Overall, the predominant and distinct expression profiles of these genes suggest their potential roles in the response to abiotic stress.

10.3389/fpls.2024.1391248

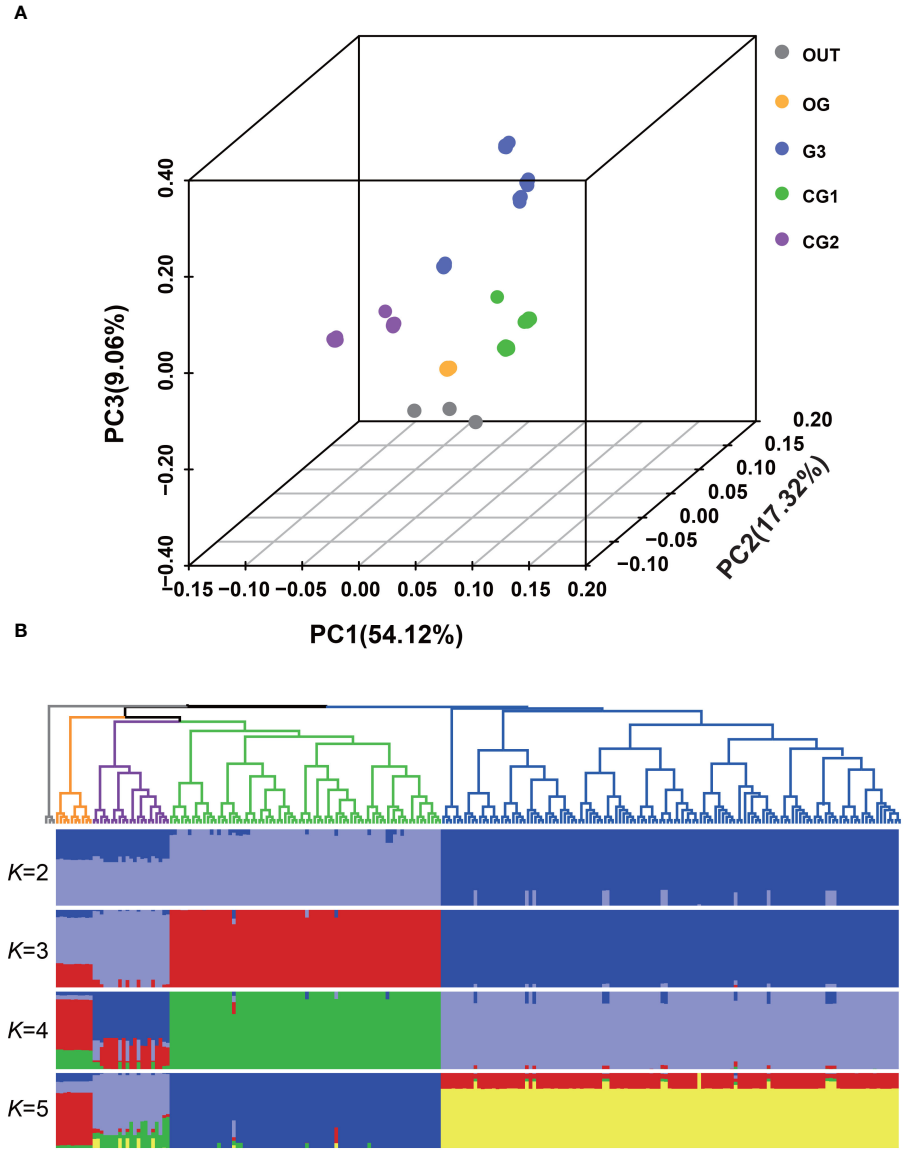


FIGURE 6
Population structure of wild and local garlic based on *AsbZIP*-related SNPs. **(A)** Plots of principal component analysis for the first (PC1), second (PC2), and third (PC3) components. Dot color indicates population and location. **(B)** Phylogenetic tree and population structure with *K* ranging from 2 to 5. The rooted tree was constructed using neighbor-joining method. The orders and positions of sample accessions on the x-axis are consistent with those in the neighbor-joining tree.

Expression analysis of *AsbZIP* genes under hormonal treatment

Numerous *cis*-elements associated with hormone response in the promoters of *AsbZIP* genes indicate pivotal roles of these genes in the plant hormone response. Therefore, the expression profiles of *AsbZIP* genes under hormonal stress conditions (ABA, GA3, IAA, MeJA, and SA) were investigated. The expression levels of 14 *AsbZIPs* in garlic leaves and roots at 6 h, 12 h, 24 h and 48 h after treatment were determined via qRT-PCR (Figures 9A–E). Under ABA stress, the increased expression patterns of *Asa7G01379*, *Asa8G04833*, *Asa7G01972*, and *Asa5G05838* were detected at all time points, whereas other genes presented reduced expression in leaves. Among these genes, *Asa7G01379* had the highest induction level and peaked at 24 h, followed by *Asa7G01972*, which peaked at 12 h; the most significantly inhibited level was obtained in *Asa8G02357* at 12 h. In roots, *Asa4G00594*, *Asa6G00995*, *Asa8G06181*, *Asa7G01379*, *Asa8G04833*, *Asa7G01972* and *Asa4G03115* were predominantly induced by ABA at all time points, whereas other genes were inhibited. Furthermore, *Asa7G01972* was most significantly induced and peaked at 12 h. Under GA3 treatment, most *AsbZIP* genes presented increased expression profiles, except for *Asa8G02357*, whose expression was most significantly induced in *Asa7G01379* at 6 h in leaves. In roots, the expression levels of 9 *AsbZIP* genes, namely, *Asa4G00594*, *Asa6G00995*, *Asa8G06181*, *Asa7G01379*, *Asa2G04763*, *Asa8G04833*, *Asa7G02467*, *Asa7G01972*, and *Asa4G03115*, were induced by GA3 at different

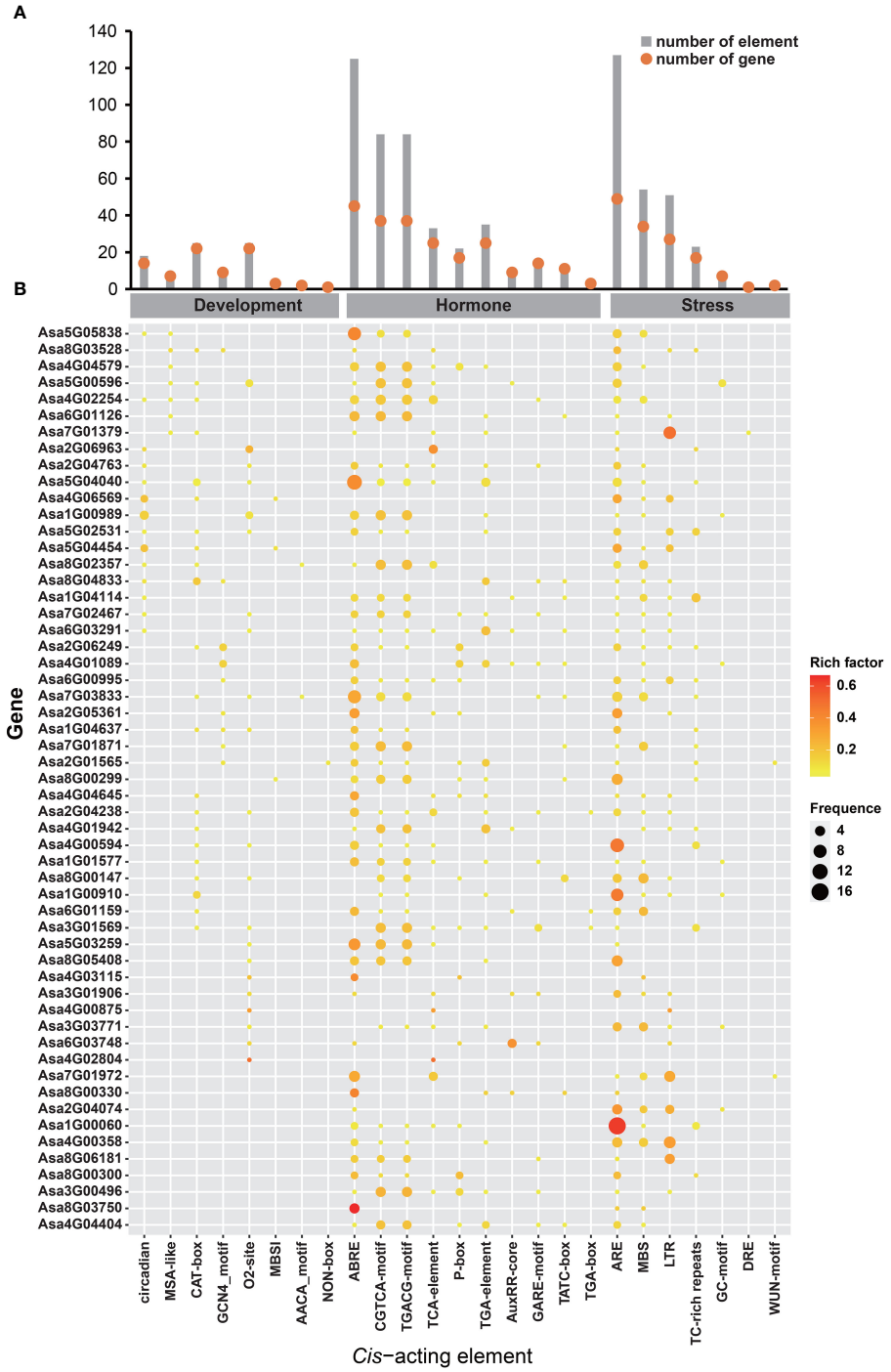


FIGURE 7
Cis-acting elements in the promoter of *AsbZIP* genes. **(A)** Number of each *cis*-acting element in the promoter regions of the *AsbZIP* genes. Red dots indicate the number of genes containing the corresponding *cis*-acting element, while gray boxes represent the total number of *cis*-acting elements. **(B)** Frequency of each *cis*-acting element in the promoters of *AsbZIP* genes. *Cis*-acting elements were categorized into development-related, hormone-related, and stress-related classes.

time points. Furthermore, 9 genes (*Asa8G02357*, *Asa8G06181*, *Asa2G04763*, *Asa2G06963*, *Asa7G02467*, *Asa7G01972*, *Asa4G03115*, *Asa8G05408*, and *Asa5G05838*) were upregulated in leaves after IAA treatment. Notably, *Asa7G01972* was strongly induced by IAA treatment, with its expression peaking at 24 h. In roots treated with IAA, only *Asa2G04763* was upregulated, with its expression peaking at 12 h. MeJA stress led to the upregulation of *Asa8G02357*, *Asa7G01379*, *Asa2G06963*, and *Asa5G05838* in leaves, with *Asa8G02357* peaking at 48 h and the other genes peaking at 24 h. After MeJA treatment, the expression of six genes (*Asa8G02357*, *Asa8G06181*, *Asa2G04763*, *Asa8G04833*, *Asa7G01972*, and *Asa5G05838*) in roots increased, and the

10.3389/fpls.2024.1391248

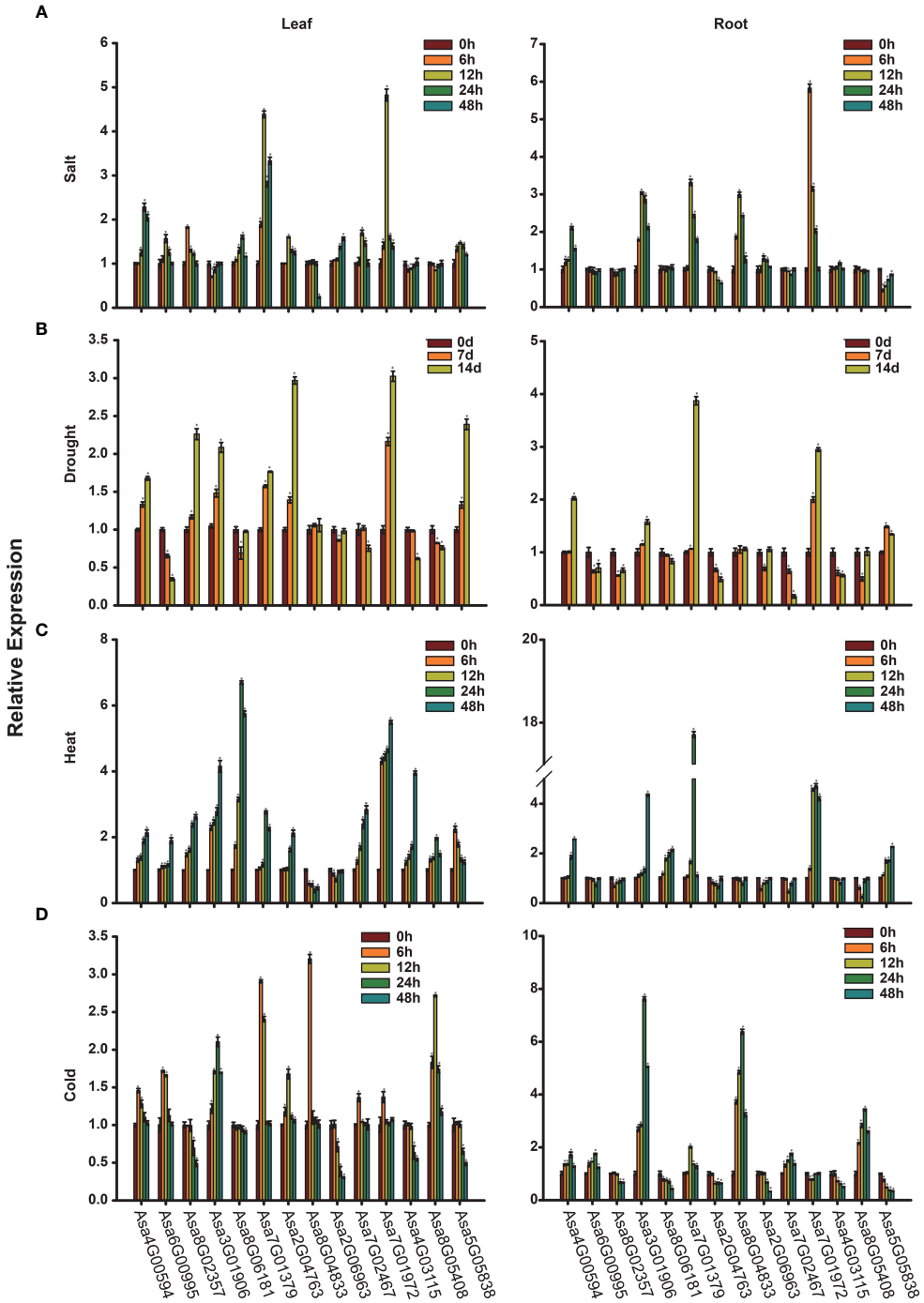


FIGURE 8
Expression analysis of *AsbZIP* genes. **(A-D)** The expression levels of *AsbZIP* genes under salt **(A)**, drought **(B)**, heat **(C)** and cold **(D)** detected by qRT-PCR. Error bars represent standard deviations from biological replicates. One asterisk (*) denotes a significant difference at $P < 0.05$, determined using Student's *t*-test.

expression of *Asa2G04763* was notably increased at 6 h. After SA treatment, the expression of *Asa8G02357*, *Asa7G01379*, *Asa2G04763*, and *Asa5G05838* was upregulated in leaves, and only *Asa2G04763* presented increased expression and peaked at 6 h in roots. In summary, these results provide useful information for further functional investigations of *AsbZIPs*.

Gene regulatory network analysis

Gene regulatory network analysis serves as an effective method for annotating gene function. The regulatory network of *AsbZIP* genes was constructed via WGCNA on the basis of 185 RNA-Seq samples (Supplementary Table 7) and the binding motifs of *bZIP* genes in

10.3389/fpls.2024.1391248

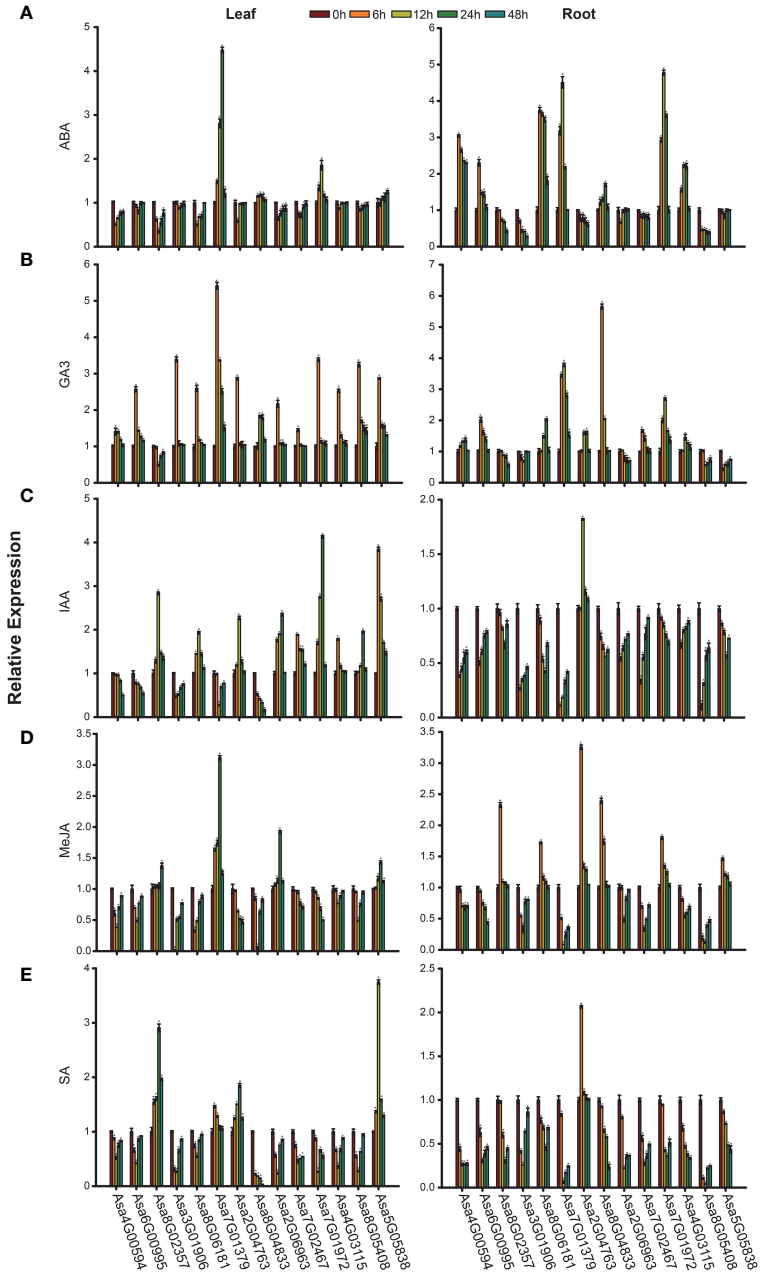


FIGURE 9
Expression patterns of *AsbZIP* genes under hormone treatment were examined by qRT-PCR. **(A)** Expression profiles of *AsbZIP* genes after exogenous ABA treatment. **(B)** Expression profiles of *AsbZIP* genes after exogenous GA3 treatment. **(C)** Expression profiles of *AsbZIP* genes after exogenous IAA treatment. **(D)** Expression profiles of *AsbZIP* genes after exogenous MeJA treatment. **(E)** Expression profiles of *AsbZIP* genes after exogenous SA treatment. The error bars represent standard deviations from biological replicates. A significance level of $P < 0.05$, as determined by Student's *t*-test, is indicated by an asterisk (*).

Arabidopsis thaliana, resulting in 31 gene regulatory networks (Supplementary Table 8). The largest network was centered on *Asa7G01972*, which included 201 genes, whereas the smallest network, was centered on *Asa0G04894*, which included 8 genes.

To delve into the potential biological processes in which these genes might be involved, we conducted GO enrichment analysis. GO terms related to stress responses, such as the cellular response to abiotic stimuli, the cellular response to environmental stimuli and heat acclimation, were present in the majority of the *AsbZIP* genes, indicating their vital roles in the stress response (Figure 10). In addition to *Asa3G01906*, the gene regulatory networks of *AsbZIP* genes were enriched in the regulation of development and growth, indicating crucial roles of these genes in plant growth and development. Additionally, some GO terms associated with the hormone response, including abscisic acid-activated signaling pathway, gibberellin mediated signaling pathway and response to jasmonic acid, were enriched in the gene regulatory network of several *AsbZIP* genes (*Asa0G02642*, *Asa1G00989*, *Asa3G00496*, *Asa3G01569*, *Asa4G1089*, *Asa4G02254*, *Asa4G02804*, *Asa4G04579*, *Asa5G03259*, *Asa6G00995*, *Asa6G01126*, *Asa6G01159*, *Asa6G03748*,

Asa7G01972, *Asa7G02467*, *Asa7G05774*, *Asa8G00299* and *Asa3G01906*), indicating that these genes might play critical roles in hormone response regulation.

AsbZIP protein-protein interaction network prediction

Analyzing the functional associations among AsbZIP proteins is crucial for understanding the regulatory pathways within this protein family. To further elucidate the functions of the AsbZIP proteins, an interaction network of the AsbZIP proteins was constructed (Supplementary Figure 4). Asa1G01577 (homologue of ABI5), Asa4G04645 (homologue of ABF3), Asa5G05838 (homologue of ABF4), Asa4G00594 (homologue of ABF2), Asa7G06169 (homologue of ABF1), Asa8G03528 (homologue of DPBF4), Asa7G06185 (homologue of DPBF3) and Asa4G03115 (homologue of DPBF2) are associated with an abscisic acid-activated signaling pathway. Asa6G03748 (GBF4), Asa5G00596 (BZIP68), Asa7G03833 (GBF1), Asa2G05361 (BZIP16) and AT2G46270 (GBF3) are transcription factors regulated by diverse stimuli, such as light-induction or hormone control. Furthermore, Asa7G01379 (HY5) and Asa4G01089 (HYH) are transcription factors that promote photomorphogenesis in light.

Functional analysis of *Asa7G01379* and *Asa7G01972*

Owing to their highly significant alterations in expression under salt stress and ABA treatment, *Asa7G01379* and *Asa7G01972* were selected for investigation of their biological functions via heterogeneous expression in yeast (Figure 11A). Compared with the yeast carrying the empty pYES2 vector, the growth of the yeast (BY4741) strains containing pYES2-Asa7G01379 and pYES2-Asa7G01972 was not significantly different under the control conditions but was greater under the salt treatment. These results suggest the important roles of *Asa7G01379* and *Asa7G01972* in salt stress.

To confirm the potential protein-protein interaction profiles predicted by STRING, we performed Y2H assays. The results validated the protein interactions of Asa7G01379-Asa1G01577 and Asa7G01972-Asa1G01577 (Figure 11B), demonstrating the reliability of the predicted protein interactions. Owing to the vital role of Asa1G01577 (a homologue of ABI5) in the ABA signal

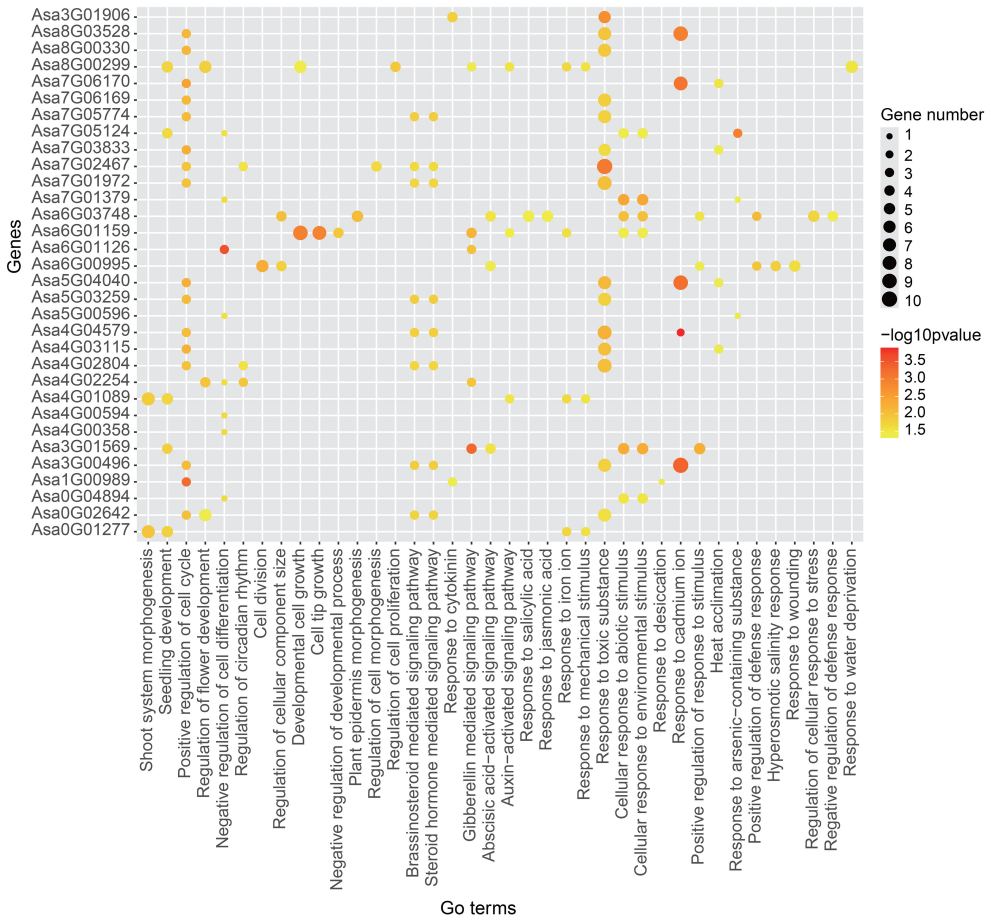


FIGURE 10
GO enrichment analysis of 31 gene sets of gene regulatory network. The size of each spot represents the number of genes enriched in specific term. The color of each spot indicates the enrichment significance level.

transduction pathway, we presume that Asa7G01379 and Asa7G01972 might regulate the response to salt stress through the ABA pathway.

Discussion

Garlic, with its plump bulbs and high allicin content, is not only used as a vegetable and flavoring ingredient, but is also widely used in the pharmaceutical industry. The *bZIP* gene family is one of the most critical families in plants and plays pivotal roles in regulating plant growth, development, and response to both biotic and abiotic stresses (Rodriguez-Uribe and O'Connell, 2006; Prasad et al., 2012; Gangappa et al., 2013). Previous studies on *bZIP* transcription factors have focused mainly on *Arabidopsis thaliana*, crops and fruits (Jakoby et al., 2002; Nijhawan et al., 2008; Liao et al., 2008a; Liu et al., 2020; Kumar et al., 2021). However, a comprehensive study of *AsbZIP* genes at the genome-wide level in garlic has not been conducted thus far. Here, a comprehensive study of the *bZIP* gene family in garlic was performed, and 64 *bZIP* genes were identified in garlic (genome size: 16 GB). The number of *bZIP* genes in garlic was less than that in *Arabidopsis thaliana* (78, genome size: 117 Mb) (Jakoby et al., 2002), rice (89, genome size: 466 Mb) (Nijhawan et al., 2008), maize (125, genome size: 2182 Mb) (Bolle, 2004), soybean (160, genome size: 915 Mb) (Zhang et al., 2018), and Tartary buckwheat (96, genome size: 489 Mb) (Liu et al., 2019) but greater than that in grape (55, genome size: 490 Mb) (Liu et al., 2014b), indicating that there is no positive correlation between the number of *bZIP* genes and genome size. Phylogenetic analysis revealed that *AsbZIPs* can be divided into 10 subfamilies (Figure 1), which was consistent with the findings in *Arabidopsis thaliana* (Jakoby et al., 2002) and grape (Liu et al., 2014b), but lower than those in tartary buckwheat (11) (Liu et al., 2019), Chinese jujube (14) (Zhang et al., 2020b), poplar (12) (Zhao et al., 2021) and potato (11) (Kumar et al., 2021).

Furthermore, gene intron/exon structural analysis revealed variation in the number of introns within *AsbZIP* genes, ranging from 0 to 11 (Figure 2), which is similar to that of *bZIP* genes in other plants, such as rice (0-12) (Nijhawan et al., 2008), wheat (0-14) (Liang et al., 2022) and moso bamboo (0-17) (Pan et al., 2019). However, members within the same subfamily shared similar gene structures, supporting the results of the phylogenetic tree. The *bZIP* genes in subfamilies C/E/G have no more than three introns, supporting the hypothesis that genes with fewer introns may facilitate rapid environmental responsiveness (Shabalina et al., 2010). In addition, although there are variations in the motif components of distinct subfamilies, the motifs encoding the bZIP domain are highly conserved, and the majority of members of the same subfamily present similar motif components, which is coincident with the findings of previous studies (Zhou et al.,

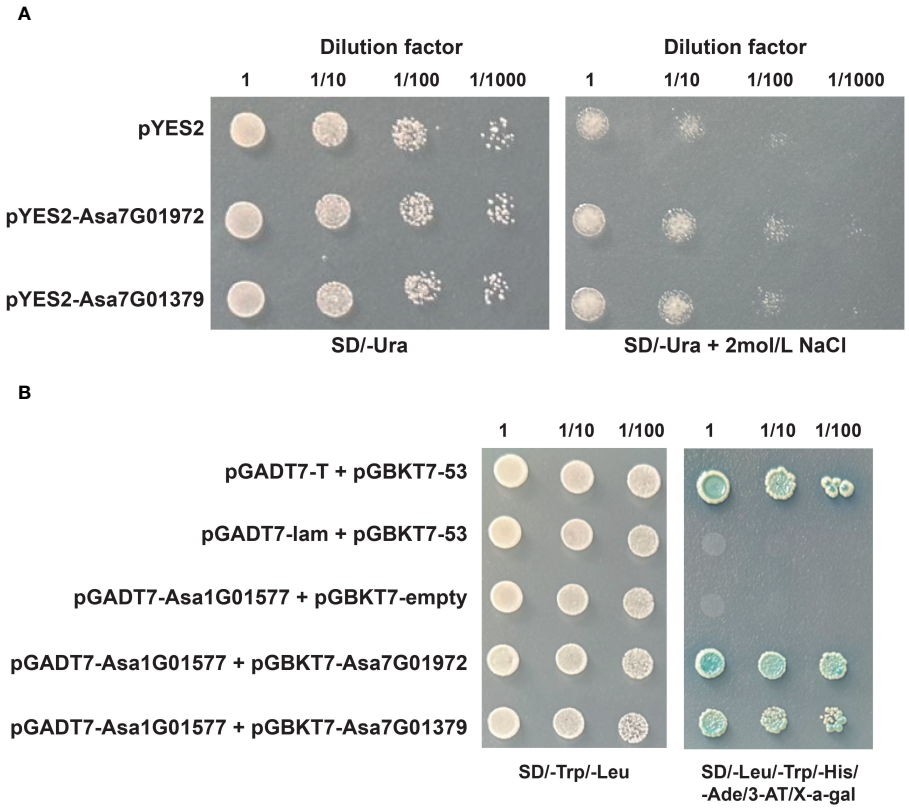


FIGURE 11
The functional analysis of Asa7G01972 and Asa7G01379. **(A)** The growth conditions of recombinant yeast BY4741 (pYES2-Asa7G01972/pYES2-Asa7G01979) and control yeast BY4741 (pYES2) were assessed under 2 mol/L NaCl treatment for 24 h Photographs of the plates were taken after 72 h of incubation at 30°. **(B)** Yeast two-hybrid analysis. 53-pGBKT7/T-pGADT7 was used as the positive control, and 53-pGBKT7/lam-pGADT7 was used as the negative control.

2017). For example, most AsbZIP proteins in subfamily S share motifs 1 and 6, with the exception of Asa5G04040, which only possesses motif 1. Several motifs (such as motifs 8, 9, and 10) found exclusively in specific subfamilies might contribute to the diverse biological functions of *AsbZIP* genes, which is consistent with previous reports (Wang et al., 2018; Liu et al., 2019).

The expansion of gene families through tandem or segmental duplication is crucial for functional diversity in evolution (Cannon et al., 2004). Furthermore, another study indicated that garlic has experienced three whole genome duplication events (Sun et al., 2020). Here, two tandem duplications and seven segmental duplications of *AsbZIPs* were identified in the garlic genome (Figures 3, 4), suggesting the vital role of gene replication in *AsbZIP* gene family expansion, which is consistent with previous reports (Nijhawan et al., 2008; Liu et al., 2014b, 2019). The number of paralogous pairs in garlic (9) was less than that in poplar (31) (Zhao et al., 2021), rice (34) (Nijhawan et al., 2008) and pear (68) (Ma et al., 2021). Additionally, the results of evolutionary analysis between garlic *bZIP* genes and their counterparts in two eudicots and two monocots revealed that there were more collinear gene pairs between garlic and *Oryza sativa* (Figure 5), indicating relatively close evolutionary relationships between these species.

The exposure of wild species to novel selection environments driven by human demands gradually induces different morphological and physiological alterations, eventually leading to their divergence from wild ancestors. This process represents domestication, a form of coevolution between plants and animals (Purugganan and Fuller, 2009). Cultivated garlic, which originates from the wild *Allium longicuspis* and *Allium tuncelianum*, has undergone genetic alterations leading to divergence in plant architecture and growth habits, referred to as the domestication syndrome (Li et al., 2022). Nevertheless, the alterations in *AsbZIPs* resulting from garlic domestication are poorly understood. Here, we identified 2941 *AsbZIP*-associated SNPs and observed their uneven distribution across the genome sequence (Supplementary Table 6), including a total of 2783 intergenic and 158 intragenic variations, which was consistent with the report presenting lower polymorphism of SNPs in intragenic regions than in intergenic regions (Hao and He, 2024). PCA, admixture and phylogenetic analysis successfully segregated all the accessions into two groups: wild garlic and local garlic (Figures 6A, B), which was consistent with the phylogenetic classification between wild and domesticated garlic of *AsHSFs* (Hao and He, 2024). In addition, domestication of garlic led to a bottleneck that decreased nucleotide diversity in alleles. The nucleotide diversity of *AsbZIPs* in wild garlic was slightly higher than that of domesticated garlic, with a decrease of only 0.00003%, which is lower than the average reduction of nucleotide diversity from wild garlic to landraces (Li et al., 2022), indicating a minor genetic bottleneck in *AsbZIP* genes during domestication. This result was confirmed by the Fst result, implying that *AsbZIP* genes did not experience strong selection pressure during domestication.

bZIP genes are related to plant development, and analyzing tissue-specific gene expression profiles facilitates to further comprehension of the biological functions of these genes. In *Arabidopsis thaliana*, HY5 (AtbZIP56) interacts with BBX25 to inhibit seedling photomorphogenesis through regulating the gene expression of *BBX22* (Gangappa et al., 2013). In *Oryza sativa*, overexpression of *OsbZIP49* resulted in the reduction of internode length and plant height, showcasing a tiller-spreading phenotype (Gangappa and Botto, 2016; Ma et al., 2018; Lorenzo, 2019; Ding et al., 2021; Han et al., 2023; Liu et al., 2023). Moreover, suppressed plant growth and decreased number of petals were observed in the *Arabidopsis thaliana* overexpressing the *Capsicum annuum CabZIP1* gene (Lee et al., 2006). Here, we systematically investigated the expression profiles of *AsbZIPs* in several tissues of garlic. As shown in Supplementary Figure 3A, *Asa2G04763*, *Asa7G02467*, *Asa8G05408* and *Asa3G01906* were highly expressed in floral buds, and *Asa6G00995*, *Asa8G02357*, *Asa4G03115* and *Asa2G06963* presented relatively high expression levels in roots, whereas relatively high expression levels in leaves were observed for *Asa8G06181* and *Asa7G01379*, suggesting potential specific functions of these genes in particular organs during the growth and development of garlic, which was comparable with tissue-specific expression profiles of *bZIP* genes in other species including *Malus halliana* (Wang et al., 2021b), *Musa nana* (Hu et al., 2016a) and *Citrullus lanatus* (Yang et al., 2019).

In addition, numerous studies have revealed that *bZIP* TFs are involved in responding to stresses. In wheat, *TabZIP96* played a crucial role in the regulation of cold stress (Liang et al., 2022). In potato, the gene expression of *StbZIP25* was induced under salt stress and could improve salt tolerance (Wang et al., 2021a). In pepper, *CabZIP25*-overexpression resulted in increased fresh weight and root length under salt stress (Gai et al., 2020). In *Oryza sativa*, *OsbZIP71* could regulate the expression of *OsNHX1* and *COR413-TM1* to improve the drought and salinity tolerance (Liu et al., 2014a). In *Vitis vinifera*, the expression of *VvbZIP23* was upregulated by several abiotic stresses, such as cold and drought stresses (Tak and Mhatre, 2013). Our results revealed that most *AsbZIP* genes were specifically induced or repressed under multiple types of stress (Figures 8A, D). In particular, *Asa4G00594* is upregulated under salt and drought stresses, and its homologous gene *AtABF2* is known to contribute to the tolerance to salinity and drought stresses (Fujita et al., 2005; Du et al., 2023). Moreover, *Asa7G01379*, encoding the AsHY5 protein, was upregulated significantly under salt stress, which is coincident with the findings of a previous report (Yang et al., 2023). The most of *AsbZIP* genes were induced under heat stress, except for *Asa8G04833* and *Asa2G06963*. Under cold stress, the expression levels of *Asa7G01379* and *Asa2G06963* increased and decreased, respectively, which is consistent with the key roles of their homologues in cold acclimation in *Arabidopsis thaliana* (Oono et al., 2006; Perea-Resa et al., 2017). The expression profiles of *AsbZIP* genes under abiotic stresses provide potential avenues for garlic breeding to increase their ability to tolerance stresses.

An increasing number of studies have revealed that *bZIP* genes are involved in phytohormone signaling. In rice, *OsbZIP72* and *OsbZIP23* were identified as regulators of ABA response to affect

drought tolerance (Xiang et al., 2008; Lu et al., 2009). In *Arabidopsis thaliana*, *AtbZIP39*, *AtbZIP36*, *AtbZIP38*, *AtbZIP35* and *AtbZIP37* have been found to play vital roles in the ABA signaling pathway (Choi et al., 2000; Uno et al., 2000; Lopez-Molina et al., 2001; Hossain et al., 2010; Liang et al., 2022). In *Artemisia annua*, the *bZIP* gene *AaTGA6* was demonstrated as a key regulator in the SA signaling pathway (Lv et al., 2019), and *AabZIP1* played a vital role in the regulation of ABA signaling (Zhang et al., 2022a). Here, the identification of *cis*-acting elements revealed the potential roles of *AsbZIPs* in phytohormone signal transduction (Figures 7A, B; Supplementary Figure 2). As shown in Figures 9A-E, the majority of *AsbZIP* genes showed significant alteration under several hormone stresses. For example, after ABA treatment, *Asa4G00594* and *Asa5G05838* were induced in roots and leaves, respectively, which is consistent with previous investigations (Choi et al., 2000; Kim et al., 2004). Particularly, *Asa7G01379* and *Asa7G01972* were the most significantly induced *AsbZIP* genes under salt stress, implying crucial functions of these genes for salt stress response. Furthermore, GO enrichment terms of gene regulatory network analysis showed that *Asa7G01379* and *Asa7G01972* might be involved in the stress response (Figure 10), which was confirmed by yeast-induced expression assay (Figure 11A). Besides, the Y2H experiment demonstrated the interactions of Asa7G01379-Asa1G01577 and Asa7G01972-Asa1G01577, as predicted by STRING (Figures 11B; Supplementary Figure 4), which is useful for revealing the biological functions of these AsbZIP proteins involved in the stress response through ABA signaling. According to the above results, *Asa7G01379* and *Asa7G01972* might be potential target locus for breeding stress-resistant garlic.

Conclusions

Here, we systematically investigated the genome-wide *AsbZIP* TFs in garlic. We identified 64 *AsbZIP* genes and conducted a comprehensive analysis of their physical characteristics, evolutionary associations, gene structures, conserved motifs, gene duplication events, nucleotide variation, population structure, expression patterns, and responses to abiotic stress and hormone treatment. On the basis of the above analysis, we speculate that *AsbZIP* genes play vital roles in the development and response of garlic to stress. Furthermore, our study revealed that 2 ABA-responsive genes, *Asa7G01972* and *Asa7G01379*, are closely related to the response to salt stress in garlic. This study could inform subsequent functional analysis of *AsbZIP* genes, contributing to a more in-depth understanding of the molecular mechanisms underlying developmental processes and stress responses in garlic.

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

SH: Conceptualization, Data curation, Funding acquisition, Investigation, Writing – original draft, Writing – review & editing. SX: Data curation, Investigation, Writing – review & editing. ZH: Data curation, Investigation, Writing – review & editing. XH: Conceptualization, Data curation, Funding acquisition, Investigation, 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 Natural Science Foundation of Shandong Province (Grant No. ZR2023QC312), research initiation project of Jining Medical University (Grant No. 600993001) and research initiation project of Beijing Academy of Science and Technology (Grant No. 0420239352KF001-05).

Conflict of interest

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

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2024.1391248/full#supplementary-material

References

Alonso, R., Onate-Sanchez, L., Weltmeier, F., Ehlert, A., Diaz, I., Dietrich, K., et al. (2009). A pivotal role of the basic leucine zipper transcription factor bZIP53 in the regulation of Arabidopsis seed maturation gene expression based on heterodimerization and protein complex formation. *Plant Cell* 21, 1747–1761. doi: 10.1105/tpc.108.062968

Babitha, K. C., Ramu, S. V., Nataraja, K. N., Sheshshayee, M. S., and Udayakumar, M. (2015). EcbZIP60, a basic leucine zipper transcription factor from Eleusine coracana L. improves abiotic stress tolerance in tobacco by activating unfolded protein response pathway. *Mol. Breed.* 35, 1–17. doi: 10.1007/s11032-015-0374-6

Baena-Gonzalez, E., Rolland, F., Thevelein, J. M., and Sheen, J. (2007). A central integrator of transcription networks in plant stress and energy signalling. *Nature* 448, 938–942. doi: 10.1038/nature06069

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

Baloglu, M. C., Eldem, V., Hajyzadeh, M., and Unver, T. (2014). Genome-wide analysis of the bZIP transcription factors in cucumber. *PloS One* 9, e96014. doi: 10.1371/journal.pone.0096014

Bastias, A., Yanez, M., Osorio, S., Arbona, V., Gomez-Cadenas, A., Fernie, A. R., et al. (2014). The transcription factor AREB1 regulates primary metabolic pathways in tomato fruits. *J. Exp. Bot.* 65, 2351–2363. doi: 10.1093/jxb/eru114

Bolle, C. (2004). The role of GRAS proteins in plant signal transduction and development. *Planta* 218, 683–692. doi: 10.1007/s00425-004-1203-z

Bombarely, A., Rosli, H. G., Vrebalov, J., Moffett, P., Mueller, L. A., and Martin, G. B. (2012). A draft genome sequence of Nicotiana benthamiana to enhance molecular plant-microbe biology research*Mol. Plant Microbe Interact* 25, 1523–1530. doi: 10.1094/MPMI-06-12-0148-TA

Cai, W., Yang, Y., Wang, W., Guo, G., Liu, W., and Bi, C. (2018). Overexpression of a wheat (Triticum aestivum L.) bZIP transcription factor gene, TabZIP6, decreased the freezing tolerance of transgenic Arabidopsis seedlings by down-regulating the expression of CBFs. *Plant Physiol. Biochem.* 124, 100–111. doi: 10.1016/j.plaphy.2018.01.008

Cannon, S. B., Mitra, A., Baumgarten, A., Young, N. D., and May, G. (2004). The roles of segmental and tandem gene duplication in the evolution of large gene families in Arabidopsis thaliana. *BMC Plant Biol.* 4, 10. doi: 10.1186/1471-2229-4-10

Chen, C., Chen, H., Zhang, Y., Thomas, H. R., Frank, M. H., He, Y., et al. (2020). TBtools: an integrative toolkit developed for interactive analyses of big biological data. *Mol. Plant* 13, 1194–1202. doi: 10.1016/j.molp.2020.06.009

Choi, H., Hong, J., Ha, J., Kang, J., and Kim, S. Y. (2000). ABFs, a family of ABA-responsive element binding factors. *J. Biol. Chem.* 275, 1723–1730. doi: 10.1074/jbc.275.3.1723

Chow, C. N., Lee, T. Y., Hung, Y. C., Li, G. Z., Tseng, K. C., Liu, Y. H., et al. (2019). PlantPAN3.0: a new and updated resource for reconstructing transcriptional regulatory networks from ChIP-seq experiments in plants. *Nucleic Acids Res.* 47, D1155–D1163. doi: 10.1093/nar/gky1081

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

Correa, L. G., Riano-Pachon, D. M., Schrago, C. G., dos Santos, R. V., Mueller-Roeber, B., and Vincentz, M. (2008). The role of bZIP transcription factors in green plant evolution: adaptive features emerging from four founder genes. *PloS One* 3, e2944. doi: 10.1371/journal.pone.0002944

Ding, C., Lin, X., Zuo, Y., Yu, Z., Baerson, S. R., Pan, Z., et al. (2021). Transcription factor OsbZIP49 controls tiller angle and plant architecture through the induction of indole-3-acetic acid-amido synthetases in rice. *Plant J.* 108, 1346–1364. doi: 10.1111/tpj.15515

Du, J., Zhu, X., He, K., Kui, M., Zhang, J., Han, X., et al. (2023). CONSTANS interacts with and antagonizes ABF transcription factors during salt stress under long-day conditions. *Plant Physiol.* 193, 1675–1694. doi: 10.1093/plphys/kiad370

Edwards, K. D., Fernandez-Pozo, N., Drake-Stowe, K., Humphry, M., Evans, A. D., Bombarely, A., et al. (2017). A reference genome for Nicotiana tabacum enables map-based cloning of homeologous loci implicated in nitrogen utilization efficiency. *BMC Genomics* 18, 448. doi: 10.1186/s12864-017-3791-6

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, D427–D432. doi: 10.1093/nar/gky995

ETOH, T., WATANABE, H., and IWAI, S. (2001). RAPD variation of garlic clones in the center of origin and the westernmost area of distribution. Memoirs of the Faculty of Agriculture Kagoshima University. 37, 21–27. Available at: http://hdl.handle.net/10232/2819.

Fan, L. X., Xu, L., Wang, Y., Tang, M. J., and Liu, L. W. (2019). Genome- and Transcriptome-Wide Characterization of bZIP Gene Family Identifies Potential Members Involved in Abiotic Stress Response and Anthocyanin Biosynthesis in Radish (Raphanus sativus L.). *Int. J. Mol. Sci.* 20, 6334. doi: 10.3390/ijms20246334

Fujita, Y., Fujita, M., Satoh, R., Maruyama, K., Parvez, M. M., Seki, M., et al. (2005). AREB1 is a transcription activator of novel ABRE-dependent ABA signaling that enhances drought stress tolerance in Arabidopsis. *Plant Cell* 17, 3470–3488. doi: 10.1105/tpc.105.035659

Gai, W., Ma, X., Qiao, Y., Shi, B., Ul Haq, S., Li, Q., et al. (2020). Characterization of the bZIP transcription factor family in pepper (L.): positively modulates the salt tolerance. *Front. Plant Sci.* 11, 139. doi: 10.3389/fpls.2020.00139

Gangappa, S. N., and Botto, J. F. (2016). The multifaceted roles of HY5 in plant growth and development. *Mol. Plant* 9, 1353–1365. doi: 10.1016/j.molp.2016.07.002

Gangappa, S. N., Crocco, C. D., Johansson, H., Datta, S., Hettiarachchi, C., Holm, M., et al. (2013). The Arabidopsis B-BOX protein BBX25 interacts with HY5, negatively regulating BBX22 expression to suppress seedling photomorphogenesis. *Plant Cell* 25, 1243–1257. doi: 10.1105/tpc.113.109751

Grant, C. E., Bailey, T. L., and Noble, W. S. (2011). FIMO: scanning for occurrences of a given motif. *Bioinformatics* 27, 1017–1018. doi: 10.1093/bioinformatics/btr064

Han, H., Wang, C., Yang, X., Wang, L., Ye, J., Xu, F., et al. (2023). Role of bZIP transcription factors in the regulation of plant secondary metabolism. *Planta* 258, 13. doi: 10.1007/s00425-023-04174-4

Hao, X., and He, S. (2024). Genome-wide identification, classification and expression analysis of the heat shock transcription factor family in Garlic (Allium sativum L.). *BMC Plant Biol.* 24, 421. doi: 10.1186/s12870-024-05018-3

Hossain, M. A., Cho, J. I., Han, M., Ahn, C. H., Jeon, J. S., An, G., et al. (2010). The ABRE-binding bZIP transcription factor OsABF2 is a positive regulator of abiotic stress and ABA signaling in rice. *J. Plant Physiol.* 167, 1512–1520. doi: 10.1016/j.jplph.2010.05.008

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

Hu, W., Wang, L., Tie, W., Yan, Y., Ding, Z., Liu, J., et al. (2016a). Genome-wide analyses of the bZIP family reveal their involvement in the development, ripening and abiotic stress response in banana. *Sci. Rep.* 6, 30203. doi: 10.1038/srep30203

Hu, W., Yang, H., Yan, Y., Wei, Y., Tie, W., Ding, Z., et al. (2016b). Genome-wide characterization and analysis of bZIP transcription factor gene family related to abiotic stress in cassava. *Sci. Rep.* 6, 22783. doi: 10.1038/srep22783

Ibrahim, H. R., Matsuzaki, T., and Aoki, T. (2001). Genetic evidence that antibacterial activity of lysozyme is independent of its catalytic function. *FEBS Lett.* 506, 27–32. doi: 10.1016/S0014-5793(01)02872-1

Jakoby, M., Weisshaar, B., Droge-Laser, W., Vicente-Carbajosa, J., Tiedemann, J., Kroj, T., et al. (2002). bZIP transcription factors in Arabidopsis. *Trends Plant Sci.* 7, 106–111. doi: 10.1016/S1360-1385(01)02223-3

Joo, J., Lee, Y. H., and Song, S. I. (2014). Overexpression of the rice basic leucine zipper transcription factor OsbZIP12 confers drought tolerance to rice and makes seedlings hypersensitive to ABA. *Plant Biotechnol. Rep.* 8, 431–441. doi: 10.1007/s11816-014-0335-2

Kamenetsky, R., Faigenboim, A., Shemesh Mayer, E., Ben Michael, T., Gershberg, C., Kimhi, S., et al. (2015). Integrated transcriptome catalogue and organ-specific profiling of gene expression in fertile garlic (*Allium sativum* L.). *BMC Genomics* 16, 12. doi: 10.1186/s12864-015-1212-2

Kang, C., Zhai, H., He, S. Z., Zhao, N., and Liu, Q. C. (2019). A novel sweetpotato bZIP transcription factor gene, IbbZIP1, is involved in salt and drought tolerance in transgenic Arabidopsis. *Plant Cell Rep.* 38, 1373–1382. doi: 10.1007/s00299-019-02441-x

Kim, S., Kang, J. Y., Cho, D. I., Park, J. H., and Kim, S. Y. (2004). ABF2, an ABRE-binding bZIP factor, is an essential component of glucose signaling and its overexpression affects multiple stress tolerance. *Plant J.* 40, 75–87. doi: 10.1111/j.1365-313X.2004.02192.x

Kumar, P., Kumar, P., Sharma, D., Verma, S. K., Halterman, D., and Kumar, A. (2021). Genome-wide identification and expression profiling of basic leucine zipper transcription factors following abiotic stresses in potato (Solanum tuberosum L.). *PloS One* 16, e0247864. doi: 10.1371/journal.pone.0247864

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

Lee, S. C., Choi, H. W., Hwang, I. S., Choi, D. S., and Hwang, B. K. (2006). Functional roles of the pepper pathogen-induced bZIP transcription factor, CAbZIP1, in enhanced resistance to pathogen infection and environmental stresses. *Planta* 224, 1209–1225. doi: 10.1007/s00425-006-0302-4

Li, J., Besseau, S., Toronen, P., Sipari, N., Kollist, H., Holm, L., et al. (2013). Defense-related transcription factors WRKY70 and WRKY54 modulate osmotic stress tolerance by regulating stomatal aperture in Arabidopsis. *New Phytol.* 200, 457–472. doi: 10.1111/nph.12378

Li, X., Duan, X., Jiang, H., Sun, Y., Tang, Y., Yuan, Z., et al. (2006). Genome-wide analysis of basic/helix-loop-helix transcription factor family in rice and Arabidopsis. *Plant Physiol.* 141, 1167–1184. doi: 10.1104/pp.106.080580

Li, N., Zhang, X., Sun, X., Zhu, S., Cheng, Y., Liu, M., et al. (2022). Genomic insights into the evolutionary history and diversification of bulb traits in garlic. *Genome Biol.* 23, 188. doi: 10.1186/s13059-022-02756-1

Liang, Y., Xia, J., Jiang, Y., Bao, Y., Chen, H., Wang, D., et al. (2022). Genome-Wide Identification and Analysis of bZIP Gene Family and Resistance of TaABI5 (TabZIP96) under Freezing Stress in Wheat (Triticum aestivum). *Int. J. Mol. Sci.* 23, 2351. doi: 10.3390/ijms23042351

Liao, Y., Zou, H. F., Wang, H. W., Zhang, W. K., Ma, B., Zhang, J. S., et al. (2008a). Soybean GmMYB76, GmMYB92, and GmMYB177 genes confer stress tolerance in transgenic Arabidopsis plants. *Cell Res.* 18, 1047–1060. doi: 10.1038/cr.2008.280

Liao, Y., Zou, H. F., Wei, W., Hao, Y. J., Tian, A. G., Huang, J., et al. (2008b). Soybean GmbZIP44, GmbZIP62 and GmbZIP78 genes function as negative regulator of ABA signaling and confer salt and freezing tolerance in transgenic Arabidopsis. *Planta* 228, 225–240. doi: 10.1007/s00425-008-0731-3

Liu, Y., Chai, M., Zhang, M., He, Q., Su, Z., Priyadarshani, S., et al. (2020). Genome-wide analysis, characterization, and expression profile of the basic leucine zipper transcription factor family in pineapple. *Int. J. Genomics* 2020, 3165958. doi: 10.1155/2020/3165958

Liu, J., Chen, N., Chen, F., Cai, B., Dal Santo, S., Tornielli, G. B., et al. (2014b). Genome-wide analysis and expression profile of the bZIP transcription factor gene family in grapevine (Vitis vinifera). *BMC Genomics* 15, 281. doi: 10.1186/1471-2164-15-281

Liu, X., and Chu, Z. Q. (2015). Genome-wide evolutionary characterization and analysis of bZIP transcription factors and their expression profiles in response to multiple abiotic stresses in Brachypodium distachyon. *BMC Genomics* 16, 227. doi: 10.1186/s12864-015-1457-9

Liu, C., Mao, B., Ou, S., Wang, W., Liu, L., Wu, Y., et al. (2014a). OsbZIP71, a bZIP transcription factor, confers salinity and drought tolerance in rice. *Plant Mol. Biol.* 84, 19–36. doi: 10.1007/s11103-013-0115-3

Liu, J. X., Srivastava, R., and Howell, S. H. (2008). Stress-induced expression of an activated form of AtbZIP17 provides protection from salt stress in Arabidopsis. *Plant Cell Environ.* 31, 1735–1743. doi: 10.1111/j.1365-3040.2008.01873.x

Liu, H., Tang, X., Zhang, N., Li, S., and Si, H. (2023). Role of bZIP transcription factors in plant salt stress. *Int. J. Mol. Sci.* 24, 7893. doi: 10.3390/ijms24097893

Liu, M., Wen, Y., Sun, W., Ma, Z., Huang, L., Wu, Q., et al. (2019). Genome-wide identification, phylogeny, evolutionary expansion and expression analyses of bZIP transcription factor family in tartaty buckwheat. *BMC Genomics* 20, 483. doi: 10.1186/s12864-019-5882-z

Livak, K. J., and Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* 25, 402–408. doi: 10.1006/meth.2001.1262

Lopez-Molina, L., Mongrand, S., and Chua, N. H. (2001). A postgermination developmental arrest checkpoint is mediated by abscisic acid and requires the ABI5 transcription factor in Arabidopsis. *Proc. Natl. Acad. Sci. U.S.A.* 98, 4782–4787. doi: 10.1073/pnas.081594298

Lorenzo, O. (2019). bZIP edgetic mutations: at the frontier of plant metabolism, development and stress trade-off. *J. Exp. Bot.* 70, 5517–5520. doi: 10.1093/jxb/erz298

Lozano-Sotomayor, P., Chavez Montes, R. A., Silvestre-Vano, M., Herrera-Ubaldo, H., Greco, R., Pablo-Villa, J., et al. (2016). Altered expression of the bZIP transcription factor DRINK ME affects growth and reproductive development in Arabidopsis thaliana. *Plant J.* 88, 437–451. doi: 10.1111/tpj.13264

Lu, G., Gao, C. X., Zheng, X. N., and Han, B. (2009). Identification of OsbZIP72 as a positive regulator of ABA response and drought tolerance in rice. *Planta* 229, 605–615. doi: 10.1007/s00425-008-0857-3

Lv, Z., Guo, Z., Zhang, L., Zhang, F., Jiang, W., Shen, Q., et al. (2019). Interaction of bZIP transcription factor TGA6 with salicylic acid signaling modulates artemisinin biosynthesis in Artemisia annua. *J. Exp. Bot.* 70, 3969–3979. doi: 10.1093/jxb/erz166

Ma, M., Chen, Q., Dong, H., Zhang, S., and Huang, X. (2021). Genome-wide identification and expression analysis of the bZIP transcription factors, and functional analysis in response to drought and cold stresses in pear (Pyrus breschneideri). *BMC Plant Biol.* 21, 583. doi: 10.1186/s12870-021-03356-0

Ma, H., Liu, C., Li, Z., Ran, Q., Xie, G., Wang, B., et al. (2018). ZmbZIP4 contributes to stress resistance in maize by regulating ABA synthesis and root development. *Plant Physiol.* 178, 753–770. doi: 10.1104/pp.18.00436

Martin, K. W., and Ernst, E. (2003). Herbal medicines for treatment of bacterial infections: a review of controlled clinical trials. *J. Antimicrob. Chemother.* 51, 241–246. doi: 10.1093/jac/dkg087

Munim Twaij, B., Jameel Ibraheem, L., Al-Shammari, R. H. H., Hasan, M., Akter Khoko, R., Sunzid Ahomed, M., et al. (2023). Identification and characterization of aldehyde dehydrogenase (ALDH) gene superfamily in garlic and expression profiling in response to drought, salinity, and ABA. *Gene* 860, 147215. doi: 10.1016/j.gene.2023.147215

Nei, M., and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* 3, 418–426. doi: 10.1093/oxfordjournals.molbev.a040410

Nijhawan, A., Jain, M., Tyagi, A. K., and Khurana, J. P. (2008). Genomic survey and gene expression analysis of the basic leucine zipper transcription factor family in rice. *Plant Physiol.* 146, 333–350. doi: 10.1104/pp.107.112821

Oono, Y., Seki, M., Satou, M., Iida, K., Akiyama, K., Sakurai, T., et al. (2006). Monitoring expression profiles of Arabidopsis genes during cold acclimation and deacclimation using DNA microarrays. *Funct. Integr. Genomics* 6, 212–234. doi: 10.1007/s10142-005-0014-z

Orellana, S., Yanez, M., Espinoza, A., Verdugo, I., Gonzalez, E., Ruiz-Lara, S., et al. (2010). The transcription factor SlAREB1 confers drought, salt stress tolerance and regulates biotic and abiotic stress-related genes in tomato. *Plant Cell Environ.* 33, 2191–2208. doi: 10.1111/j.1365-3040.2010.02220.x

Oyama, T., Shimura, Y., and Okada, K. (1997). The Arabidopsis HY5 gene encodes a bZIP protein that regulates stimulus-induced development of root and hypocotyl. *Genes Dev.* 11, 2983–2995. doi: 10.1101/gad.11.22.2983

Pan, F., Wu, M., Hu, W., Liu, R., Yan, H., and Xiang, Y. (2019). Genome-Wide Identification and Expression Analyses of the bZIP Transcription Factor Genes in moso bamboo (Phyllostachys edulis). *Int. J. Mol. Sci.* 20, 2203. doi: 10.3390/ijms20092203

Perea-Resa, C., Rodriguez-Milla, M. A., Iniesto, E., Rubio, V., and Salinas, J. (2017). Prefoldins negatively regulate cold acclimation in arabidopsis thaliana by promoting nuclear proteasome-mediated HY5 degradation. *Mol. Plant* 10, 791–804. doi: 10.1016/j.molp.2017.03.012

Pourabed, E., Ghane Golmohamadi, F., Soleymani Monfared, P., Razavi, S. M., and Shobbar, Z. S. (2015). Basic leucine zipper family in barley: genome-wide characterization of members and expression analysis. *Mol. Biotechnol.* 57, 12–26. doi: 10.1007/s12033-014-9797-2

Prasad, V. B., Gupta, N., Nandi, A., and Chattopadhyay, S. (2012). HY1 genetically interacts with GBF1 and regulates the activity of the Z-box containing promoters in light signaling pathways in Arabidopsis thaliana. *Mech. Dev.* 129, 298–307. doi: 10.1016/j.mod.2012.06.004

Purugganan, M. D., and Fuller, D. Q. (2009). The nature of selection during plant domestication. *Nature* 457, 843–848. doi: 10.1038/nature07895

Que, F., Wang, G. L., Huang, Y., Xu, Z. S., Wang, F., and Xiong, A. S. (2015). Genomic identification of group A bZIP transcription factors and their responses to abiotic stress in carrot. *Genet. Mol. Res.* 14, 13274–13288. doi: 10.4238/2015.October.26.24

Rodriguez-Uribe, L., and O'Connell, M. A. (2006). A root-specific bZIP transcription factor is responsive to water deficit stress in tepary bean (Phaseolus acutifolius) and common bean (P-vulgaris). *J. Exp. Bot.* 57, 1391–1398. doi: 10.1093/jxb/erj118

Shabalina, S. A., Ogurtsov, A. Y., Spiridonov, A. N., Novichkov, P. S., Spiridonov, N. A., and Koonin, E. V. (2010). Distinct patterns of expression and evolution of intronless and intron-containing mammalian genes. *Mol. Biol. Evol.* 27, 1745–1749. doi: 10.1093/molbev/msq086

Shi, Y. P., Ding, G. H., Shen, H. T., Li, Z. H., Li, H. B., and Xiao, G. H. (2024). Genome-wide identification and expression profiles analysis of the authentic response regulator gene family in licorice. *Front. Plant Sci.* 14. doi: 10.3389/fpls.2023.1309802

Singh, K., Foley, R. C., and Onate-Sanchez, L. (2002). Transcription factors in plant defense and stress responses. *Curr. Opin. Plant Biol.* 5, 430–436. doi: 10.1016/S1369-5266(02)00289-3

Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., et al. (2002). The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.* 12, 1611–1618. doi: 10.1101/gr.361602

Suckow, M., Schwamborn, K., Kisters-Woike, B., von Wilcken-Bergmann, B., and Muller-Hill, B. (1994). Replacement of invariant bZip residues within the basic region of the yeast transcriptional activator GCN4 can change its DNA binding specificity. *Nucleic Acids Res.* 22, 4395–4404. doi: 10.1093/nar/22.21.4395

Sun, X. D., Zhu, S. Y., Li, N. Y., Cheng, Y., Zhao, J., Qiao, X. G., et al. (2020). A Chromosome-Level Genome Assembly of Garlic (*Allium sativum*) Provides Insights into Genome Evolution and Allicin Biosynthesis. *Mol. Plant* 13, 1328–1339. doi: 10.1016/j.molp.2020.07.019

Tak, H., and Mhatre, M. (2013). Cloning and molecular characterization of a putative bZIP transcription factor VvbZIP23 from Vitis vinifera. *Protoplasma* 250, 333–345. doi: 10.1007/s00709-012-0417-3

Tan, S. L., Yang, Y. J., Liu, T., Zhang, S. B., and Huang, W. (2020). Responses of photosystem I compared with photosystem II to combination of heat stress and fluctuating light in tobacco leaves. *Plant Sci.* 292, 110371. doi: 10.1016/j.plantsci.2019.110371

Tang, N., Zhang, H., Li, X., Xiao, J., and Xiong, L. (2012). Constitutive activation of transcription factor OsbZIP46 improves drought tolerance in rice. *Plant Physiol.* 158, 1755–1768. doi: 10.1104/pp.111.190389

Uno, Y., Furihata, T., Abe, H., Yoshida, R., Shinozaki, K., and Yamaguchi-Shinozaki, K. (2000). Arabidopsis basic leucine zipper transcription factors involved in an abscisic acid-dependent signal transduction pathway under drought and high-salinity conditions. *Proc. Natl. Acad. Sci. U.S.A.* 97, 11632–11637. doi: 10.1073/pnas.190309197

Wang, X. L., Chen, X., Yang, T. B., Cheng, Q., and Cheng, Z. M. (2017). Genome-Wide Identification of bZIP Family Genes Involved in Drought and Heat Stresses in Strawberry (Fragaria vesca). *Int. J. Genomics* 2017, 3981031. doi: 10.1155/2017/3981031

Wang, Q., Guo, C., Li, Z. Y., Sun, J. H., Wang, D., Xu, L. T., et al. (2021a). Identification and analysis of bZIP family genes in potato and their potential roles in stress responses. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.637343

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

Wang, S., Zhang, R., Zhang, Z., Zhao, T., Zhang, D., Sofkova, S., et al. (2021b). Genome-wide analysis of the bZIP gene lineage in apple and functional analysis of MhABF in Malus halliana. *Planta* 254, 78. doi: 10.1007/s00425-021-03724-y

Wang, D., Zhang, Y., Zhang, Z., Zhu, J., and Yu, J. (2010). KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics Proteomics Bioinf.* 8, 77–80. doi: 10.1016/S1672-0229(10)60008-3

Wang, Y., Zhang, Y., Zhou, R., Dossa, K., Yu, J., Li, D., et al. (2018). Identification and characterization of the bZIP transcription factor family and its expression in response to abiotic stresses in sesame. *PloS One* 13, e0200850. doi: 10.1371/journal.pone.0200850

Wang, J., Zhou, J., Zhang, B., Vanitha, J., Ramachandran, S., and Jiang, S. Y. (2011). Genome-wide expansion and expression divergence of the basic leucine zipper transcription factors in higher plants with an emphasis on sorghum. *J. Integr. Plant Biol.* 53, 212–231. doi: 10.1111/j.1744-7909.2010.01017.x

Wei, K., Chen, J., Wang, Y., Chen, Y., Chen, S., Lin, Y., et al. (2012). Genome-wide analysis of bZIP-encoding genes in maize. *DNA Res.* 19, 463–476. doi: 10.1093/dnares/dss026

Wigge, P. A., Kim, M. C., Jaeger, K. E., Busch, W., Schmid, M., Lohmann, J. U., et al. (2005). Integration of spatial and temporal information during floral induction in Arabidopsis. *Science* 309, 1056–1059. doi: 10.1126/science.1114358

Xiang, Y., Tang, N., Du, H., Ye, H., and Xiong, L. (2008). Characterization of OsbZIP23 as a key player of the basic leucine zipper transcription factor family for conferring abscisic acid sensitivity and salinity and drought tolerance in rice. *Plant Physiol.* 148, 1938–1952. doi: 10.1104/pp.108.128199

Xu, J., Duan, X., Yang, J., Beeching, J. R., and Zhang, P. (2013). Enhanced reactive oxygen species scavenging by overproduction of superoxide dismutase and catalase delays postharvest physiological deterioration of cassava storage roots. *Plant Physiol.* 161, 1517–1528. doi: 10.1104/pp.112.212803

Yang, Y., Li, J., Li, H., Yang, Y., Guang, Y., and Zhou, Y. (2019). The bZIP gene family in watermelon: genome-wide identification and expression analysis under cold stress and root-knot nematode infection. *PeerJ* 7, e7878. doi: 10.7717/peerj.7878

Yang, O., Popova, O. V., Suthoff, U., Luking, I., Dietz, K. J., and Golldack, D. (2009). The Arabidopsis basic leucine zipper transcription factor AtbZIP24 regulates complex transcriptional networks involved in abiotic stress resistance. *Gene* 436, 45–55. doi: 10.1016/j.gene.2009.02.010

Yang, J., Qu, X., Li, T., Gao, Y., Du, H., Zheng, L., et al. (2023). HY5-HDA9 orchestrates the transcription of HsfA2 to modulate salt stress response in Arabidopsis. *J. Integr. Plant Biol.* 65, 45–63. doi: 10.1111/jipb.13372

Yao, X., Lai, D., Zhou, M., Ruan, J., Ma, C., Wu, W., et al. (2023). Genome-wide identification, evolution and expression pattern analysis of the GATA gene family in Sorghum bicolor. *Front. Plant Sci.* 14, 1163357. doi: 10.3389/fpls.2023.1163357

Ying, S., Zhang, D. F., Fu, J., Shi, Y. S., Song, Y. C., Wang, T. Y., et al. (2012). Cloning and characterization of a maize bZIP transcription factor, ZmbZIP72, confers drought and salt tolerance in transgenic Arabidopsis. *Planta* 235, 253–266. doi: 10.1007/s00425-011-1496-7

Zhang, Y., Gao, W., Li, H., Wang, Y., Li, D., Xue, C., et al. (2020b). Genome-wide analysis of the bZIP gene family in Chinese jujube (Ziziphus jujuba Mill.). *BMC Genomics* 21, 483. doi: 10.1186/s12864-020-06890-7

Zhang, X., Lin, H. M., Hu, H., Hu, X., and Hu, L. (2016). Gamma-aminobutyric acid mediates nicotine biosynthesis in tobacco under flooding stress. *Plant Divers.* 38, 53–58. doi: 10.1016/j.pld.2016.05.004

Zhang, M., Liu, Y., Cai, H., Guo, M., Chai, M., She, Z., et al. (2020a). The bZIP transcription factor gmbZIP15 negatively regulates salt- and drought-stress responses in soybean. *Int. J. Mol. Sci.* 21, 7778. doi: 10.3390/ijms21207778

Zhang, M., Liu, Y., Shi, H., Guo, M., Chai, M., He, Q., et al. (2018). Evolutionary and expression analyses of soybean basic Leucine zipper transcription factor family. *BMC Genomics* 19, 159. doi: 10.1186/s12864-018-4511-6

Zhang, X., Yang, X., He, Q., Wang, Y., Liang, G., and Liu, T. (2022b). Genome-wide identification and characterization of the GRAS transcription factors in garlic (*Allium sativum* L.). *Front. Plant Sci.* 13, 890052. doi: 10.3389/fpls.2022.890052

Zhang, H., Zhu, J., Gong, Z., and Zhu, J. K. (2022a). Abiotic stress responses in plants. *Nat. Rev. Genet.* 23, 104–119. doi: 10.1038/s41576-021-00413-0

Zhao, K., Chen, S., Yao, W. J., Cheng, Z. H., Zhou, B. R., and Jiang, T. B. (2021). Genome-wide analysis and expression profile of the bZIP gene family in poplar. *BMC Plant Biol.* 21, 122. doi: 10.1186/s12870-021-02879-w

Zhao, J., Guo, R., Guo, C., Hou, H., Wang, X., and Gao, H. (2016). Evolutionary and expression analyses of the apple basic leucine zipper transcription factor family. *Front. Plant Sci.* 7, 376. doi: 10.3389/fpls.2016.00376

Zhou, Y., Xu, D., Jia, L., Huang, X., Ma, G., Wang, S., et al. (2017). Genome-Wide Identification and Structural Analysis of bZIP Transcription Factor Genes in Brassica napus. *Genes (Basel)* 8, 288. doi: 10.3390/genes8100288

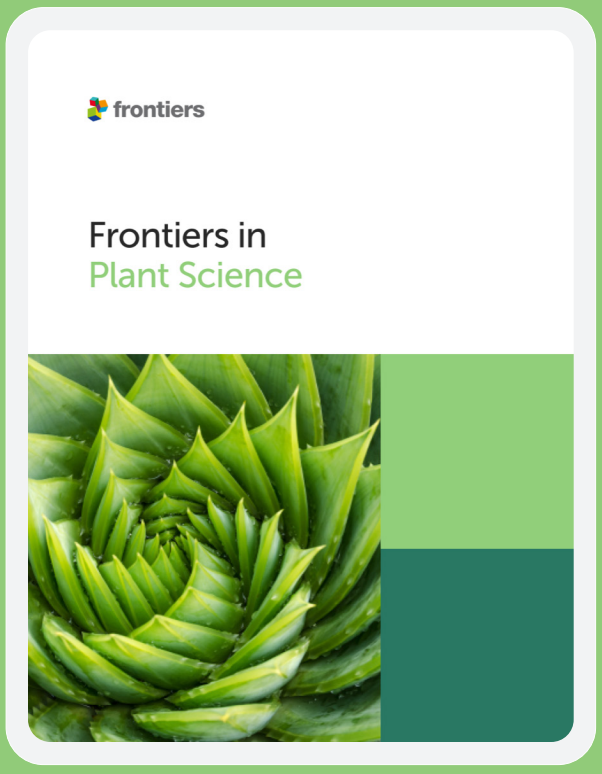
Frontiers in
Plant Science

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