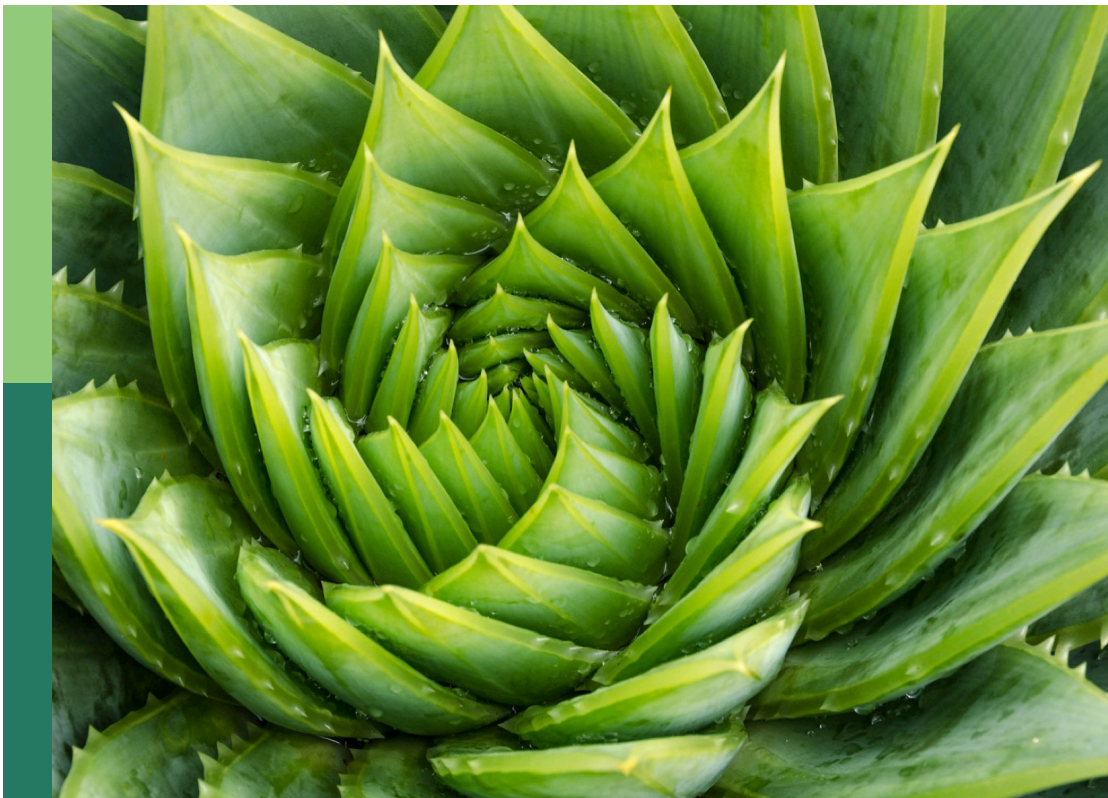
frontiers | Research Topics

Advances and applications of cost-effective, high-throughput genotyping technologies for sustainable agriculture

Edited by
Nisha Singh, Sapna Langyan and Vandna Rai

Published in
Frontiers in Plant Science



ISSN 1664-8714
ISBN 978-2-8325-4186-9
DOI 10.3389/978-2-8325-4186-9

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

Advances and applications of cost-effective, high-throughput genotyping technologies for sustainable agriculture

Topic editors

Nisha Singh — Gujarat Biotechnology University, India

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

Vandna Rai — National Institute for Plant Biotechnology, Indian Council of Agricultural Research, India

Citation

Singh, N., Langyan, S., Rai, V., eds. (2024). *Advances and applications of cost-effective, high-throughput genotyping technologies for sustainable agriculture*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-4186-9

Table of contents

frontiers | Frontiers in Plant Science

TYPE Editorial
PUBLISHED 08 December 2023
DOI 10.3389/fpls.2023.1335417

OPEN ACCESS

EDITED AND REVIEWED BY
Nunzio D'Agostino,
University of Naples Federico II, Italy

*CORRESPONDENCE
Nisha Singh
singh.nisha88@gmail.com

RECEIVED 08 November 2023
ACCEPTED 27 November 2023
PUBLISHED 08 December 2023

CITATION
Singh N, Langyan S and Rai V (2023) Editorial: Advances and applications of cost-effective, high-throughput genotyping technologies for sustainable agriculture. *Front. Plant Sci.* 14:1335417. doi: 10.3389/fpls.2023.1335417

Editorial: Advances and applications of cost-effective, high-throughput genotyping technologies for sustainable agriculture

Nisha Singh[1]*, Sapna Langyan[2] and Vandna Rai[3]

[1]Department of Bioinformatics, Gujarat Biotechnology University, Gandhinagar, India, [2]ICAR-National Bureau of Plant Genetic Resources, New Delhi, India, [3]ICAR-National Institute for Plant Biotechnology, New Delhi, India

KEYWORDS

crop improvement, NGS, genomics, GWAS, high-throughput genotyping technologies, sustainable agriculture

Editorial on the Research Topic
Advances and applications of cost-effective, high-throughput genotyping technologies for sustainable agriculture

Rapid breakthroughs in next-generation sequencing (NGS) technology have greatly aided crop improvement. It became a powerful tool for the production of genomic data, and various other integrated techniques have considerably expanded and deepened our understanding of living organisms' molecular systems. NGS technologies bring novel tools and concepts that can enhance the precision and efficiency of plant breeding, such as the development of cost-effective, high throughput genotyping technologies and its various applications in sustainable agriculture. These technologies enable plant breeders to genotype a large number of samples in a short time span. It is used to implement genomic selection (GS) as a routine practice in breeding programs for fast-track development of superior crop breeds that are stress-resistant while still having a high nutritional value. In conventional plant breeding approaches, it takes a much longer time to create new, improved varieties because this relies on phenotypic selection and breeder experience. Moreover, complex quantitative traits have low heritability and are therefore difficult to select. Modern breeding procedures are advantageous because they are more reliable and efficient. Anand et al. explain that it is more feasible and sustainable to use cutting-edge technology to power agronomic development.

The application of genomic technologies such as genome-wide association studies (GWAS) and genomic prediction analysis (GPA) to durum wheat landrace resources paves the way for next-generation breeding programs to overcome the knowledge gap of these underexplored resources and to identify advantageous alleles lost in modern varieties. Therefore, to address the challenges of climate change, food, and nutritional security, Broccanello et al. emphasized the significance of durum wheat landraces as a valuable

genetic resource for improving the sustainability of Mediterranean agroecosystems, with a focus on resilience to environmental stresses.

GS approaches have proven beneficial for assessing breeding values and phenotypic prediction using data from genome-wide molecular markers. To evaluate individual breeding value, the link between an individual genotype and a phenotype has been simulated using a variety of parametric approaches. To overcome the constraints of distinct parametric and nonparametric models, an integrated model for genomic prediction under additive and non-additive genetic architectures was developed by Budhlakoti et al. It can simultaneously handle both additive and epistatic effects by minimizing their error variance. Standard evaluation criteria such as prediction ability and error variance are used to evaluate the proposed integrated model. In the past few years, there has been an immense advancement in high-throughput phenotyping, phenomics, and computational biology, which has made it possible to explore such enormous datasets.

The significance of high-throughput sequencing and gene editing technologies for molecular breeding applications and trait-related genes in orchids is comprehensively explored by Song et al. This information will facilitate a scientific reference and theoretical basis for orchid genome studies. Yasir et al. used high-density single nucleotide polymorphism (SNP) arrays and DNA sequencing to disclose the bulk of the genotypic space for several crops, including cotton. The importance of GWAS and its various applications is emphasized. It was developed and first used in the field of human disease genetics. It connects the dots between a phenotype and its underlying genetics across population genomes and provides thoughtful insight into GWAS studies in cotton crop. Fiber yield and quality traits, GWAS status, prospects, and bottlenecks are discussed through case studies on both biotic and abiotic tolerance, a thought-provoking discussion. Exploring GWAS for dissecting competitive traits in major legume crops is well described by Susmitha et al. The study shed further light on advancements in biotechnology, sequencing, and several bioinformatics tools to estimate linkage disequilibrium (LD)–based associations. By computing genomic estimated breeding values (GEBVs), GWAS markers could be used for GS to forecast superior lines. However, it is yet to be employed in minor legumes where germplasm/population is available. Upadhyay et al. used a GWAS approach to evaluate the association for the acid phosphatase activity at various phosphorus levels in 280 mustard genotypes in two environmental conditions. A total set of 44 SNPs was identified that were significantly associated with two traits at three Pi (inorganic phospate) levels, for acid phosphatase (Apase) activity in the root (RApase) and leaf (LApase).These findings laid a solid foundation to improve the phosphorus use efficiency (PUE) of Indian mustard using marker-assisted selection in the future. This will redeem the crop by increasing the yield in the face of limited Pi reserves and deteriorating agro-environments.

Breeding parthenocarpic lines based on molecular markers is a faster and more efficient method. In cucumber, since selection may be based on genotypes rather than phenotypes, the identification of gynoecious traits and sex expression in cucurbitaceous vegetable crops has greatly aided the hybrid breeding program and has improved production, as reported by D. et al. The study of F2 progenies of PVGy-201 Pusa Do Mousami indicated that the genotype's gynoecious sex expression is governed by a single recessive gene. The gynoecious sex expression is shown to be related to 1.31 Mb areas located on chromosome 1. A large number of variants were discovered in the QTL-region, which will aid in the precise mapping of the gynoecious characteristic.

Moreover, the parthenocarpic fruit set is a crucial characteristic in cucumbers, enabling large-scale, protected farming on a global scale. Devi et al. provide insight into genomic regions, closely associated markers, and possible candidate genes associated with parthenocarpy in Pusa Parthenocarpic Cucumber-6 (PPC-6), which will be instrumental for functional genomics study and better understanding of parthenocarpy in cucumber. The authors identified a major QTL, Parth6.1, associated with parthenocarpic fruit development in the slicing cucumber genotype PPC-6 using QTL-seq analysis in combination with conventional mapping using the F2:3 population.

The information gathered about genetic links and QTLs will be extremely beneficial for mango breeding and for increasing our understanding of the genetics of these traits. This insight was investigated by Srivastav et al. to enable mango breeders to develop trait-specific breeding methods. An 80K high density genic SNP chip array was used for genotyping to construct high resolution mapping of QTLs for fruit color and firmness in Amrapali/Sensation mango hybrids. This cross-produced 92 bi-parental offspring was utilized to create a high-density linkage map and to identify QTLs.

Systematic evaluation and cataloging of genetic variation at the morphological and phytochemical levels is particularly beneficial for effective conservation, utilization, and optimal genetic improvement of allelic and genotypic variability. There is a dearth of detailed information on genetic diversity in Kalmegh. Chaturvedi et al. identified 91 metabolic pathway-specific EST-SSR markers, of which, 32 EST-SSR primer pairs were chosen randomly for genetic diversity analysis within the six populations (ecotypes) of 24 kalmegh accessions (*Andrographis paniculata* (Burm. f.) Nee) for genetic improvement, germplasm conservation, and maximizing genetic gain.

In conclusion, this Research Topic focuses on recent advancements in NGS-related technologies, mainly the development of cost-effective, high-throughput genotyping platforms with a wide range of bioinformatics tools, and possible translational multi-omics applications in crop breeding programs for sustainable agriculture. We included a total of 13 publications in this Research Topic, which comprised of both original research and review papers. The Research Topic provides insights into cutting-edge research on various facets of emerging NGS, high-throughput genotyping technologies, GWAS, GS, and QTL mapping for identification and molecular breeding applications in a diverse array of crops such as cereals, legumes, and horticultural and

medicinal plants. It is well proven that, over the last two decades, scientists have been able to uncover most of the genotypic space for diverse crops by using high-density SNP arrays and DNA sequencing. In this context, GWAS is considered a very powerful tool to identify key genes to unravel the mechanisms that will help devise efficient strategies for crop improvement and breeding programs. Furthermore, this will facilitate the investigation of the relationship between natural genetic variations and trait mapping in large populations. This Research Topic highlights the most recent genetic tools and statistical approaches ideal for the discovery of beneficial genes/alleles and associated with the most important traits in diverse crops for marker-assisted selection. The information gathered about genetic links and QTLs will be extremely beneficial for crop breeding and for increasing our understanding of the genetics of key agronomically important traits.

Author contributions

NS: Conceptualization, Data curation, Formal Analysis, Investigation, Project administration, Resources, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. SL: Formal Analysis, Visualization, Writing – review & editing. VR: Formal Analysis, Investigation, Validation, Visualization, Writing – review & editing.

Acknowledgments

We thank all the authors and reviewers who contributed to this Research Topic entitled “Advances and Applications of Cost-Effective, High-Throughput Genotyping Technologies for Sustainable 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.

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

Publisher’s note

frontiers | Frontiers in Plant Science

TYPE Mini Review
PUBLISHED 23 September 2022
DOI 10.3389/fpls.2022.1018029

OPEN ACCESS

EDITED BY
Nisha Singh,
Gujarat Biotechnology University, India

REVIEWED BY
Junpeng Shi,
Sun Yat-sen University, China
Ping Li,
Agricultural University of Hebei, China

*CORRESPONDENCE
Fucheng Zhu
fucheng323@163.com
Yunpeng Cao
xfcypeng@126.com

†These authors have contributed equally to this work

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

RECEIVED 12 August 2022
ACCEPTED 05 September 2022
PUBLISHED 23 September 2022

CITATION
Song C, Wang Y, Manzoor MA, Mao D, Wei P, Cao Y and Zhu F (2022) In-depth analysis of genomes and functional genomics of orchid using cutting-edge high-throughput sequencing. *Front. Plant Sci.* 13:1018029. doi: 10.3389/fpls.2022.1018029

In-depth analysis of genomes and functional genomics of orchid using cutting-edge high-throughput sequencing

Cheng Song[1†], Yan Wang[1†], Muhammad Aamir Manzoor[2], Di Mao[3], Peipei Wei[1], Yunpeng Cao[4*] and Fucheng Zhu[1*]

[1]College of Biological and Pharmaceutical Engineering, West Anhui University, Lu'an, China, [2]School of Life Science, Anhui Agricultural University, Hefei, China, [3]Albrecht Daniel Thaer Institute for Agricultural and Horticultural Sciences, Humboldt University of Berlin, Berlin, Germany, [4]Chinese Academy of Sciences (CAS) Key Laboratory of Plant Germplasm Enhancement and Specialty Agriculture, Wuhan Botanical Garden, Chinese Academy of Sciences, Wuhan, China

High-throughput sequencing technology has been facilitated the development of new methodologies and approaches for studying the origin and evolution of plant genomes and subgenomes, population domestication, and functional genomics. Orchids have tens of thousands of members in nature. Many of them have promising application potential in the extension and conservation of the ecological chain, the horticultural use of ornamental blossoms, and the utilization of botanical medicines. However, a large-scale gene knockout mutant library and a sophisticated genetic transformation system are still lacking in the improvement of orchid germplasm resources. New gene editing tools, such as the favored CRISPR-Cas9 or some base editors, have not yet been widely applied in orchids. In addition to a large variety of orchid cultivars, the high-precision, high-throughput genome sequencing technology is also required for the mining of trait-related functional genes. Nowadays, the focus of orchid genomics research has been directed to the origin and classification of species, genome evolution and deletion, gene duplication and chromosomal polyploidy, and flower morphogenesis-related regulation. Here, the progressing achieved in orchid molecular biology and genomics over the past few decades have been discussed, including the evolution of genome size and polyploidization. The frequent incorporation of LTR retrotransposons play important role in the expansion and structural variation of the orchid genome. The large-scale gene duplication event of the nuclear genome generated plenty of recently tandem duplicated genes, which drove the evolution and functional divergency of new genes. The evolution and loss of the plastid genome, which mostly affected genes related to photosynthesis and autotrophy, demonstrated that orchids have experienced more separate transitions to heterotrophy than any other terrestrial plant. Moreover, large-scale resequencing provide useful SNP markers for constructing genetic maps, which will facilitate the breeding of novel orchid varieties. The significance of high-throughput sequencing and gene editing technologies in the identification and molecular breeding of the trait-related genes in orchids provides us with a representative trait-improving gene as well as some

mechanisms worthy of further investigation. In addition, gene editing has promise for the improvement of orchid genetic transformation and the investigation of gene function. This knowledge may provide a scientific reference and theoretical basis for orchid genome studies.

KEYWORDS

third-generation sequencing, orchid, genome assembly, polyploidy, functional genomics, molecular breeding

Introduction

The Orchidaceae family of monocotyledonous plants have the second-largest members after Compositae. This family contains over 750 genera and nearly 28,000 species (Zhang et al., 2017). Conventional orchids could be classified into five subfamilies (*Apostasioideae*, *Vanilloideae*, *Cypripedioideae*, *Epidendroideae*, and *Orchidoideae*) by their morphology and anatomy (Lu et al., 2019). The habitat of wild orchids has been gravely affected by natural and manmade factors. Many endangered species are on the edge of extinction due to indiscriminate gathering. The current protection efforts for orchids include the construction of nature reserves and genetic resource nurseries, as well as seed-preservation and *in vitro* tissue culture (Williams et al., 2018). Although this act ensures a huge number of original germs, the seedlings degenerate and eventually lose their ability to differentiate during the subculture processes, which makes it difficult to maintain the original genetic background. Besides, most orchids are cross-pollinated, and artificial pollination is considered essential in most cases (Suetsugu, 2015). Because of their huge species diversity and significant economic value, orchids have been the focus of study in botany and ecology for many years. China has a long history of cultivating orchids and has bred numerous varieties. So far, 187 genera and 1500 species of wild orchids have been recorded, including some subspecies and varieties (Chase et al., 2015). There are still several ornamental wild orchids to be created, preserved, and exploited in nature. In addition to its high economic and ornamental value, the orchid also has a profound historic origin. In Chinese traditional culture, the orchid referred to be one of the "four gentlemen among the flowers," the others being the *Prunus mume*, *Chrysanthemum morifolium*, and *Sasa pygmaea* (Li et al., 2021).

Before the emergence of molecular-assisted breeding, distant hybridization was one of the most commonly used methodology for fertilizing orchids. In recent years, high-throughput sequencing technology and gene editing have been widely applied in the molecular biology, genomics, and discovery of trait-related genes in orchids, as well as modern genetic engineering breeding (Paun et al., 2010; Hsiao et al., 2021; Hsu et al., 2022; Li et al., 2022a). Whole genome sequencing of non-model organisms is now common due to the rapid advancement and lower cost of next-generation sequencing. The draft genome of *Phalaenopsis equestris*, a tropical epiphytic orchid that is normally utilized as a parent species in orchid breeding, was the first real achievement (Cai et al., 2015). Due to the fast development of ultralong sequencing and new assembly algorithms, whole-genome shotgun sequencing and single molecule sequencing have been done on even more orchid species, such as *Dendrobium officinale*, *Dendrobium catenatum*, *Dendrobium huoshanense*, *Phalaenopsis* 'KHM190, *Phalaenopsis aphrodite*, *Gastrodia elata*, *Vanilla planifolia*, *Apostasia shenzhenica*, *etc.* (Yan et al., 2015; Huang et al., 2016; Zhang et al., 2016; Zhang et al., 2016; Zhang et al., 2017; Chao et al., 2018; Yuan et al., 2018; Hu et al., 2019; Han et al., 2020; Niu et al., 2021). The growing number of orchid species with high-quality genomes and the use of advanced genetic analysis tools make it much easier to study the functional genes, especially those that are of interest for molecular breeding. The new advancement of genome editing technologies, such as the CRISPR/Cas9 system, is beneficial to this continuing endeavor (Wang et al., 2021). Depending on many defined gene transformation systems in orchids, the CRISPR/Cas9 tool has been effectively implemented in *P. equestris* by having to take tiny insertion/deletion or reversal mutations into target genes or perhaps the establishing kilobase-scale deletions of genes of interest. (Kui et al., 2017; Tong et al., 2020; Li et al., 2022b).

The market for orchids has expanded in size and diversity as a result of economic globalization, driving scientists and biologists to develop new varieties with distinctive looks, improved adaptability, and premium features (Li et al., 2021). Traditional breeding, despite being time- consuming, is always the predominant means of orchid cultivation. Because of the limitations and inefficiencies of the traditional approaches, hybridization and mutagenesis can not be used to get some desirable traits, like the spotted blooms and foliage of a single plant. Agrobacterium-mediated transformation and particle bombardment methods have been routinely used in transgenic

molecular breeding, leading to significant progress in horticultural development (Liau et al., 2003; Men et al., 2003; Hsing et al., 2016; Khumkarjorn et al., 2017; Lin et al., 2018; Kayika Febryanti et al., 2020; Setiawati et al., 2020). Our understanding of orchid reproductive biology will undoubtedly change as a result of these efforts to enhance orchid genome-editing tools and the power of large-scale genome sequencing, which will enable us to better understand the inherent roles of orchid genes and changes to genes of interest for desired blooming and floral features (Molla and Yang 2019; Nopitasari et al., 2020; Tong et al., 2020; Guo et al., 2022). Here, we systematically summarized the studies on orchid genomes, including plastid genomes, especially the molecular evolution of orchids based on high-throughput sequencing technology and the identification and functional studies of trait-related genes. In addition, the application of gene editing and genetic transformation technologies in orchids was also discussed in detail.

Genome size and ploidy analysis of the orchid

Ten years ago, only bacterial artificial chromosome (BAC) end sequences were used in genetic investigations of *Phalaenopsis* orchids. Short sequences can be used as molecular markers to assist in gene mapping and the construction of genetic maps. These sequences contained several repetitive DNA and SSR markers (Hsu et al., 2011). Cytogenetic evidence is only available for few orchid species (Felix and Guerra, 2010). *Cattleya*, *Cymbidium*, *Dendrobium*, *Oncidium*, *Phalaenopsis*, *Paphiopedilum*, *Vanilla*, and *Vanda* are examples of commercially significant genera that are valuable in floriculture, medicinal, and food condiments (da Rocha Perini et al., 2016; Vilcherrez-Atoche et al., 2022). Chromosomal counting and nuclear DNA content estimation with flow cytometry (FCM) are the most popular techniques employed for polyploid identification in these orchids (Younis et al., 2013; Mohammadi et al., 2021). Using flow cytometry, the genetic traits and types of endoreplication of 149 orchid species were compared. The variations in genome size and particularly in GC contents were inextricably bound with evolutionary transitions from the conventional mode of endoreplication to partial endoreplication (Trávníček et al., 2019). In eukaryotic species, nuclear genome size is an inherited quantitative feature with both biological and practical relevance. Genome size, karyotype, and nucleobase composition vary significantly across angiosperms, with potential adaptive consequences. A systematic analysis of the major plant families could help us understand the biological significance of the huge differences in genome size within plants. Several studies have assessed C-values in 48 orchid species in order to analyze the distributions of nuclear DNA quantities and identify tissues suited for accurate estimations of nuclear DNA content (Trávníček et al., 2015; Rewers et al., 2021). Additional analysis on the size of the genomes of *Pleurothallidinae* species showed that those with partial endoreplication (PE) had much bigger genomes and that the number of genomic repeats was closely linked to the size of the non-endoreplicated part of the genome (Chumová et al., 2021). According to previous investigations on the variation of Apostasioideae genome size, the predicted 1C-values vary from 0.38 pg in *Apostasia nuda* to 5.96 pg in *Neuwiedia zollingeri* var. javanica, a roughly 16-fold difference. The genome sizes of the two genera did not overlap. *Apostasia* had much smaller genomes than *Neuwiedia*, which suggested that smaller genomes were common in the Apostasioideae subfamily (Jersáková et al., 2013). The genome of *Apostasia ramifera* showed the population size histories of many orchid species, as well as a continual fall in population size in seven orchid genomes (Zhang W. et al., 2021). Some research had shown that the incorporation of LTR retrotransposons Orchid-rt1 and Gypsy1 into *Phalaenopsis* genomes might be linked to genome size growth (Hsu et al., 2020). Genome size is also linked to cellular and developmental characteristics. The evolutionary connection between genome size, floral lifespan, and labellum epidermal cell size in *Paphiopedilum* revealed that genome size was connected to floral duration but negatively relevant to labellum epidermal cell size (Zhang and Zhang, 2021).

In addition to flow cytometry, k-mer analysis-based genome survey sequencing is also a common method for estimating genome size. It has the advantages of high-throughput sequencing, high speed, and large amounts of data, which can quickly determine the size and heterozygosity of the genome (Lee et al., 2017). The k-mer depth values are often derived from the curves used to estimate genome size. Through the distribution of the k-mer curve, the genomic characteristics are estimated, and the ratio of the heterozygous peak to the homozygous peak is calculated to obtain the heterozygous rate (Jersáková et al., 2013). For determining the size of orchid genomes, k-mer analysis based on the Illumina Hiseq sequencing platform has been widely applied. The genome of *C. ensifolium* was evaluated using 17-mer analysis, which indicated the genome size and heterozygozity to be 3.56 Gb and 1.40%, respectively (Ai et al., 2021). The estimated genome size of *G. menghaiensis* based on k-mers is 0.98 Gb, with 0.1% heterozygosity and high repeats. The 17-mer distribution is Poisson-distributed and is dependent on the properties of the genome (Jiang Y. et al., 2022). Using k-mer distribution analysis, the genome size, heterozygozity, and repetitive ratio of *D. officinale* were determined. The largest peak of 17 k-mer frequency was seen at a depth of 90, allowing the determination of the genome size, heterozygosity, and repetitive ratio (Niu et al., 2021).

The development of the orchid industry benefits greatly from the ploidy identification of orchid germplasm resources. Chromosomal and cytological investigations revealed that *Cymbidium* species contained a prevalence of 40 chromosomes

along with variations found in *C. serratum* (41, 43, 60, and 80). From the earliest polyploids recorded at the beginning of the 20th century, it has been feasible to create a number of *Cymbidium* polyploid cultivars through biological and artificial approaches (Xie et al., 2017). Since then, *Cymbidium* cultivars have been known to be diploids, triploids, and tetraploids with distinct chromosomal morphology (Younis et al., 2013). About 75.8% of *C. hybridum* cultivars harbor polyploids, indicating a link between the intentional or unintentional selection of polyploids instead of diploids for superior features (Vilcherrez-Atoche et al., 2022). The majority of *Dendrobium* species contained 38 chromosomes, with the exception of *D. leonis* and *D. dixanthum*, which both have 40 chromosomes (Zheng et al., 2018). The majority of *Phalaenopsis* species have 38 chromosomes, with the exception of the Aphyllae, which has only 34 or 36 chromosomes (Lee et al., 2017). However, a significant heterogeneity of genome size was detected among species and hybrids within this genus (Chang et al., 2006; Lee et al., 2017). *Phalaenopsis* cultivars have a wide range of chromosomal numbers (38, 57, and 76 more), indicating polyploidy. Flower gardening traditionally employs *Phalaenopsis* hybrid cultivars. Only one diploid cultivar has been documented, whereas over 80% of tetraploid cultivars have 76 chromosomes (Lee S. Y. et al., 2020). The domination of commercial tetraploid cultivars demonstrates the relevance of polyploidy in the development of better *Phalaenopsis* cultivars. These tetraploid species are implemented as parentals to create subgroups of *Phalaenopsis* cultivars with the goal of achieving desirable colors for commercial purposes (Bolaños-Villegas and Chen, 2007; Li et al., 2012). *Vanda*, like *Dendrobium* and *Phalaenopsis*, has 38 chromosomes and naturally occurs in tetraploid and hexaploid species (Khan et al., 2019; Liu et al., 2020). In *Oncidium*, it is assumed that x = 7 is the basic number of chromosomes, but unlike other genera, there is a huge chromosomal variation across species, with the majority exhibiting polyploidy (Su et al., 2013).

Evaluation of gene duplication events under high-quality genome sequencing in orchid

The continuity and integrity of model plant genomes have also been greatly improved due to the continuous development of genome research and the improvement of sequencing technology. The orchid genome has gone through the draft genome obtained by ordinary next-generation sequencing, to the chromosome-level genome assembled by PacBio or ONT sequencing technology combined with Hi-C, and then to the near complete genome obtained by ONT (N50>50Kb) assembly (Figure 1). By combining ONT ultra-long and PacBio HiFi techniques, those gap-free genomes assembled at telomere to telomere (T2T) level will be a new direction in the future. The whole genome sequencing of the *A. shenzhenica* helps us better understand the origins and evolution within subfamilies (Zhang et al., 2017). The whole genome duplication (WGD) that has occurred more than once in plant genomes is a noteworthy feature (Clark and Donoghue, 2018). Angiosperm genome sequences provide information regarding polyploidy and genome evolution. By evaluating the prevalence of synonymous substitutions per synonymous site (Ks) throughout all paralogous genes and duplicated genes situated in synteny blocks based on the *Phalaenopsis* and *Dendrobium* genome sequences, two WGDs were projected to have evolved in the *D. catenatum* lineage. (Zhang et al., 2016). The nearest WGD event is shared by *Dendrobium*, *Phalaenopsis*, and *Apostasia*, and it could have occurred near the Cretaceous-Paleogene

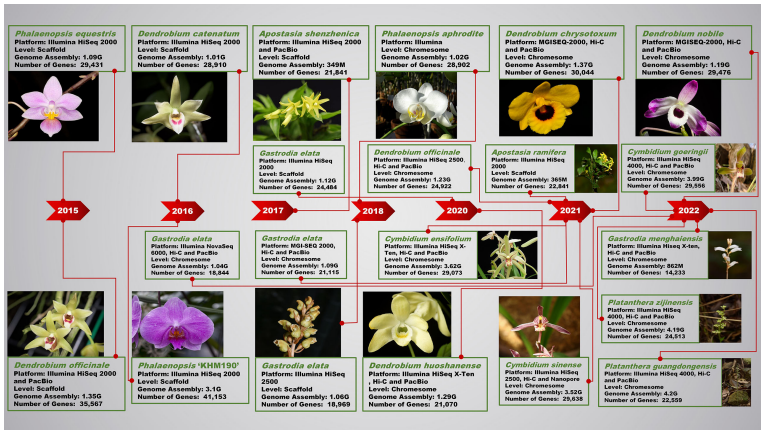


FIGURE 1
Research progress of next-generation sequencing and third-generation sequencing technology in orchid genomes.

(K/Pg) boundary. Peaks in older Ks distributions are thought to be an additional ancient WGD event shared by monocot ancestors (Cai et al., 2015; Zhang et al., 2016; Zhang et al., 2017). The draft genome sequencing revealed compelling evidence of a whole-genome duplication that all orchids share and that came right before their divergence (Zhang et al., 2017). The MADS-box family members may govern a wide spectrum of developmental events during orchid evolution. A chromosomal-scale genome and chromosome linkage groups of *P. aphrodite* were first created, which contributed to the variation in labellum and pollinium morphology and structures (Chao et al., 2018). A chromosome-scale genome assembly of *C. goeringii* suggested several new gene families, resistance-related homologs and variations within the *MADS-box* genes may regulate a wide set of developmental processes during adaptive evolution (Chung et al., 2022). A haplotype-resolved genome of *Bletilla striata* reveals its evolutionary relationship with other orchids, which have experienced an ancient WGD event shared with monocots and a recent WGD event within all orchids. The biochemical machinery of *B. striata* polysaccharide (BSP) biosynthesis indicated that MYB2 interacted physically with some BSP-regulated genes (Jiang L. et al., 2022). Partial endoreplication has been discovered across all *Vanilla* species. A chromosome-scaled genome of *Vanilla planifolia* showed that the genome size discrepancy was driven by the presence of PE (Piet et al., 2022).

Mycoheterotrophic and parasitic plants get some or all of the nutrients they need from other organisms. *Gastrodia* fungi are typically perennial, achlorophyllous orchids with a unique evolutionary mechanism for adaptability to a non-photosynthetic lifestyle. The genome of *G. elata* reveals the genetic basics of most adaptive changes in photosynthesis, leaf development, and plastid division (Chen S. et al., 2020). Comparative genomics studies revealed that *G. elata* and other completely heterotrophic species dropped nearly 10% of the conserved orthogroups, including those important for autotrophs (Xu et al., 2021). Photosynthesis, circadian clock, flowering control, immunity, food intake, and root and leaf growth are all governed by these orthogroups. Recent assembly of the *G. elata* genome also showed a strong contraction of genes which involved in multiple biosynthetic processes and cellular components but also an expansion of genes for some metabolic processes and mycorrhizal interactions (Bae et al., 2022). Many genes involved in arbuscular mycorrhizae colonization and biological interaction between *Gatrodia* and symbiotic microbes were identified in the genome of *G. menghaiensis* (Jiang Y. et al., 2022). The loss and conservation of symbiotic genes associated with the evolution of unique symbionts in plants were determined by analyzing a broad array of plant genome and transcriptomics data. A shared symbiosis network progressed at the same time as intracellular endosymbioses, from the primitive arbuscular mycorrhiza to the more recent ericoid and orchid mycorrhizae in angiosperms and ericoid-like connections in bryophytes (Radhakrishnan et al., 2020). The comparison of *Platanthera zijinensis* and *Platanthera guangdongensis* genomes indicated that mycoheterotrophy is linked to higher rates of gene loss and alternation, and that the deletion of most photoreceptor and auxin transporter genes might explain how fully mycoheterotrophic orchids look so different from other orchids. Some trehalase genes have grown, which makes sense since orchid non-endosperm seeds need carbohydrates from fungi to sprout when they are in the protocorm stage (Li M-H. et al., 2022; Minasiewicz et al., 2022).

Dendrobium is the second biggest genus in Orchidaceae. The first genome of a lithophytic orchid, *D. catenatum* (now recognized as *D. officinale*), showed wide duplication of genes associated with glucomannan synthase (Yan et al., 2015; Zhang et al., 2016). Recent assembly of the *D. officinale* genome has brought new insights into the evolution of this *Dendrobium* spp. (Niu et al., 2021). Our previous study released a chromosome-level assembly of the *D. huoshanense* genome with PacBio sequencing and Hi-C method (Han et al., 2020). A chromosome-scale reference genome of *D. chrysotoxum* was also obtained based on PacBio sequencing and Hi-C methods. The phylogeny of the *SWEET* gene family implied that gene expansion occurred in clade II may associated with fleshy stems rich in polysaccharides (Zhang Y. et al., 2021). *Cymbidium* is famous for its distinctive leaves, flower morphology, and pleasant aroma (Yang L. et al., 2021). The genome of *C. ensifolium* has undergone two WGD events, and the abnormal expression of *MADS-box* genes might be related to flower development and shape mutations (Ai et al., 2021). A chromosome- scale genome of *D. nobile* showed two polyploidization events occurred. The expression profile of *TPS* and *CYP450* genes suggested that the distinct distribution of *TPS-b* subclade may contribute to the species-specific alkaloid biosynthesis pathways (Xu et al., 2022). Finally, a phylogenetic tree was constructed based on single-copy genes to better demonstrate the evolutionary relationship between orchid species (Figure S1).

The associated mapping method performed statistical analyses to discover the importance of the relationship between genetic variants and polymorphism in a group of individuals with genetic variations (Ogura and Busch, 2015). Large-scale resequencing has been broadly used for gene mapping of crop quality traits and differential analysis of SNP loci within genes. However, investigations for genome-wide association studies (GWAS) based on genotyping-by-sequencing (GBS) have received less attention in orchids. Through NGS technology, a large number of SNP markers have been found through sequencing to create a high-density genetic map. A total of 691,532 SNP sites were identified to generate a genetic linkage map for marker-assisted selection breeding by resequencing *Phalaenopsis pulcherrima* and denovo sequencing of *Phalaenopsis* 'KHM190' (Huang et al., 2016). Species-specific markers could help to identify unknown intraspecies and validate the parentage of interspecifc hybrid

offspring. Genomics-based diversity analysis of *Vanilla* species indicated that the value of the GBS approach to interpret diversity in *Vanilla* collections has been demonstrated to be the paternal parent of hybrids more efficiently than other methods (Hu et al., 2019). The interspecific hybridization of *D. nobile* and *Dendrobium wardianum* was used to construct a population with 100 F1 individuals (Li J. et al., 2019). A total of 331,642 SNP markers were obtained, 9645 of which were used to build a high-density genetic map with 19 linkage groups, and three QTLs identified may be associated with stem length and diameter. The genetic diversity and variations among *Dendrobium* mutants and common *Dendrobium* cultivars were compared based on SNPs by GBS (Ryu et al., 2019). A total of 517,660 SNPs were identified, 37,721 of which were used to discriminate the differences across *Dendrobium* genotypes. 129 accessions were collected from 10 wild cultivated populations to explore the genetic diversity and population structure of *D. nobile* in China (He et al., 2022). Approximately 830,000 SNPs were obtained and used for genetic variation analysis. The recent completion of the chromosome-level assembly of the *D. officinale* genome provides a reliable data basis for its genetic background and breeding improvement. Niu and his colleagues performed *D. officinale* resequencing to conduct a GWAS investigation on 38 cultivars and five related species (Niu et al., 2021). A total of 13 GWAS loci were identified to associate with some morphologic traits.

Sequencing and evolution of the chloroplast genome in orchid

The chloroplast genome (cp) contains more conserved structures than the nuclear and mitochondrial genomes, which is beneficial for systematics and species identification. Studies on the chloroplast genomes of Orchidaceae have remained prominent in recent years (Table 1). The chloroplast genomes of *D. officinale* and *Cypripedium macranthos* were compared, and there were parallels in structure as well as gene order and content, but there were differences in the organization of the inverted repeat/small single-copy junction and *ndh* genes (Luo et al., 2014). Since *ndh* genes are truncated or excluded in the cp genomes of some autotrophic Epidendroideae orchids, some studies had mentioned that these gene deletion events are independent (Lin et al., 2015). By comparing 53 cp genomes, it was indicated that the expansion of inverted repeats in *Paphiopedilum* and *Vanilla* is also associated with a loss of *ndh* genes (Niu et al., 2017b). *Bulbophyllum Thou.* is one of the biggest genera with over 2,000 species, found in rainforest regions (Gamisch and Comes, 2019). Long-term geographic isolation exposed Asian and South American *Bulbophyllum* cp genomes to varying selective pressures (Yang et al., 2022). Besides the *Bulbophyllum* orchids, plastid genome sequencing has been reported for a large number of *Dendrobium* species, which are commonly used for phylogenetic studies and variety authentication (Zhang et al., 2018; Wu X.-Y. et al., 2019; Liu et al., 2021). *Phalaenopsis* orchids are another orchid species that has received significant interest (Chang et al., 2006; Kim et al., 2016; Wang et al., 2019; Xia et al., 2021). *Paphiopedilum*, also known as slipper orchid, is well-known for its large, specialized lip, as well as its lovely flowers and colors. The cp genome of many *Paphiopedilum* orchids was investigated to provide the phylogenomic analysis of this species and its relatives (Zhao et al., 2019; Tang F. L. et al., 2020; Hu et al., 2022). Furthermore, the cp genomes of some other orchid genera or subtribes have been published, including *Pelatantheria scolopendrifolia*, *Cymbidium ensifolium*, *Eulophia flava*, *Calanthe arcuata*, and *Coelogyne fimbr* (Yun et al., 2018; Bertrand et al., 2019; Jiang et al., 2019; Li H. et al., 2019; Zhong et al., 2019). These results are important for figuring out how chloroplasts have changed over time and how gene structures vary in orchids (Zeng et al., 2007). A phylogenetic tree of 58 representative orchid species was constructed to investigate the relationship of cp genomes within subfamilies or subtribes (Figure S2). The results also revealed that these varieties could be classified into five subfamilies, with the majority of individuals belonging to the Epidendroideae and Orchidoideae.

Orchids have undergone more independent transitions to heterotrophy than any other land plants. Another interesting fact is that some heterotrophic orchids lose photosynthesis and autotrophy-related genes on chloroplasts throughout evolution, which provides an excellent opportunity to explore the effects of shifting selective regimes on genome evolution (Li M.-H. et al., 2022). As a consequence of the relaxation of functional restrictions on photosynthesis, certain heterotrophic plants, such as mycoheterotrophs and parasites, exhibit enormous gene losses. The comparative genomics of 12 tribe *Neottieae* orchids indicated that genes related to the NAD(P)H dehydrogenase complex, photosystems, and RNA polymerase were functionally lost many times (Feng et al., 2016). A phylogenetic analysis of 26 full plastome sequences from *Epidendreae* suggested that photosynthesis-related genes such as the atp complex had undergone severe gene loss (Lee S. Y. et al., 2020). Numerous investigation have identified evidence of fast plastome degradation in heterotrophic orchids based on the accumulation of pseudogenes and substantial deletions (Barrett and Kennedy, 2018; Barrett et al., 2019; Kim et al., 2019). Infraspecific analysis of the plastome evolution of leafless *Corallorhiza* revealed that considerable changes in plastome size and functional gene composition occurred in just a few million years as a consequence of decreasing selection constraints on photosynthesis (Barrett et al., 2018).

TABLE 1 Features of representative plastid genomes in orchidaceae.

Subfamily	Taxon	Total length (bp)	Large single copy (LSC)	Inverted repeat (IR)	Small single copy (SSC)	Protein-coding genes	Accession	Reference
Epidendroideae	*Dendrobium officinale*	152,221	85,109	26,298	14,516	76	KC771275	Luo et al., 2014
	Pelatantheria scolopendrifolia	146,971	86,096	24,570	11,735	72	MG752972	Yun et al., 2018
	Dendrobium bellatulum	152,107	85,061	26,297	14,503	83	MG595965	Zhang et al., 2018
	Dendrobium comatum	158,008	85,592	27,032	18,352	87	MZ666386	Liu et al., 2021
	Dendrobium nobile	152,018	84,944	26,285	14,504	79	KX377961	Konhar et al., 2019
	Cymbidium ensifolium	150,257	85,110	25,692	13,761	78	MK841484	Jiang et al., 2019
	Cymbidium mastersii	155,362	84,465	25,125	20,647	80	MK848042	Zheng et al., 2019
	Cymbidium floribundum	153,998	84,725	25,132	19,009	80	MK848043	Zhang G. Q. et al., 2019
	Cymbidium hookerianum	155,447	84,186	26,711	17,839	78	MT800927	Wei et al., 2021
	Cymbidium aloifolium	157,328	85,793	26,829	17,877	78	MN641752	Chen J. et al., 2020
	Cymbidium floribundum var. *pumilum*	155,291	84,415	26,696	17,484	80	MN173778	Ai et al., 2019a
	Cymbidium sinense x *C. goeringii*	150,149	84,987	25,691	13,780	75	MN532117	Choi et al., 2020
	Cymbidium dayanum	155,408	84,189	26,614	17,991	76	MW160431	Du et al., 2021
	Cymbidium bicolor	156,528	85,907	26,703	17,215	78	MN654912	Hu et al., 2020
	Dendrobium longicornu	160,024	88,075	25,403	21,143	80	MN227146	Wu X.-Y. et al., 2019
	Calanthe arcuata	158,735	87,348	26,449	18,489	88	MK934523	Zhong et al., 2019
	Danxiaorchis singchiana	87,931	42,575	13,762	17,831	36	MN584923	Lee S. Y. et al., 2020
	Coelogyne fimbriata	158,935	87,444	26,374	18,743	91	MT548043	Yue et al., 2020
	Pleione maculata	158,394	86,603	26,646	18,499	89	MW699846	He et al., 2021
	Pleione bulbocodioides	159,269	87,125	26,716	18,712	81	KY849819	Shi et al., 2018
	Pleione chunii	158,880	87,259	26,465	18,691	87	MK792342	Wu S. et al., 2019
	Hexalectris warnockii	119,057	66,903	17,332	17,490	38	MH444822	Barrett and Kennedy, 2018
	Arundina graminifolia	159,482	87,285	26,813	18,581	88	MN171408	Ai et al., 2019b
	Eulophia zollingeri	145,201	81,566	25,272	13,091	86	MG181954	Huo et al., 2018
	Dendrobium thyrsiflorum	160,123	88,001	25,490	21,142	80	MN306203	Pan et al., 2019
	Liparis vivipara	158,329	85,950	27,043	18,293	77	MK862100	Zhang D. et al., 2019
	Liparis bootanensis	158,325	86,584	26,700	18,341	83	MN627759	Liu, 2020
	Tainia dunnii	158,305	86,819	25,244	20,998	88	MN641754	Xie et al., 2020
	Gomesa flexuosa	147,764	83,579	25,757	12,671	73	OL692830	Mo et al., 2022
	Geodorum densiflorum	149,468	85,070	25,554	13,290	76	MT153204	Tang J. M. et al., 2020
Orchidoideae	*Phalaenopsis aphrodite*	148,964	85,957	25,732	11,543	65	AY916449	Chang et al., 2006

(Continued)

TABLE 1 Continued

Subfamily	Taxon	Total length (bp)	Large single copy (LSC)	Inverted repeat (IR)	Small single copy (SSC)	Protein-coding genes	Accession	Reference
	Phalaenopsis 'Tiny Star'	148,918	85,885	25,755	11,523	70	KJ944326	Kim et al., 2016
	Phalaenopsis equestris	148,959	85,967	25,846	11,300	75	JF719062	Jheng et al., 2012
	Phalaenopsis wilsonii	145,096	84,688	24,787	10,834	73	MW194929	Fan et al., 2021
	Ophrys aveyronensis	146,816	80,495	16,309	16,309	79	MN120441	Bertrand et al., 2019
	Phalaenopsis lowii	146,834	84,469	25,944	10,477	76	MN385684	Wang J. Y. et al., 2019
	Vanda subconcolor	149,490	85,691	25,912	11,975	74	MT180955	Liu et al., 2020
	Phalaenopsis wilsoniii	145,373	84,996	24,855	10,668	76	MW218959	Xia et al., 2021
	Habenaria ciliolaris	154,544	84,032	25,455	19,602	133	MN495954	Chen et al., 2019
	Satyrium nepalense var. *ciliatum*	154,418	83,475	26,715	17,513	79	MN497244	Ma et al., 2019
	Spiranthes sinensis	152,786	83,446	25,701	17,938	78	MK936427	Fan and Huang, 2019
	Anoectochilus roxburghii	152,802	82,641	26,364	17,433	81	KP776980	Yu et al., 2016
	Nothodoritis zhejiangensis	143,522	83,830	24,464	10,764	74	MW646088	Yang L. et al., 2021
	Goodyera foliosa	154,008	83,248	25,045	20,670	80	MN443774	Zhou et al., 2019
Cypripedioideae	*Cypripedium macranthos*	157,050	85,292	26,777	18,285	79	KF925434	Luo et al., 2014
	Paphiopedilum hirsutissimum	154,569	85,198	34,344	683	79	MN153815	Zhao et al., 2019
	Paphiopedilum emersonii	162,590	87,852	36,934	870	81	MT648789	Tang F. L. et al., 2020
	Paphiopedilum gratrixianum	157,292	87,252	34,106	1,828	68	MW284890	Hu et al., 2022
	Paphiopedilum barbigerum	156,329	86,056	34,214	1,845	80	MN153814	Li M. et al., 2019
	Paphiopedilum parishii	154,689	86,863	32,690	2,446	82	MW528213	Kao et al., 2021
	Paphiopedilum bellatulum	156,567	88,243	32,336	3,652	76	MN315107	Peng et al., 2020
	Paphiopedilum spicerianum	157,292	87,252	34,106	1,828	71	MT683624	Ge et al., 2020
Apostasioideae	*Apostasia wallichii*	156,126	83,035	26,452	20,187	79	LC199394	Niu et al., 2017a
	Apostasia ramifera	157,518	86,353	27,360	16,445	87	MT864006	Zheng et al., 2021
	Apostasia shenzhenica	153,164	86,167	27,510	11,977	75	MK370661	Li Y. et al., 2019
	Neuwiedia singapureana	161,068	89,031	26,991	18,058	79	LC199503	Niu et al., 2017a
Vanilloideae	*Cyrtosia septentrionalis*	96,859	58,085	10,414	17,946	38	MH615835	Kim et al., 2019
	Vanilla shenzhenica	151,537	87,487	22,439	19,172	69	MK962478	Li T. Z. et al., 2019
	Vanilla pompona	148,009	86,358	29,807	2,037	75	MF197310	Amiryousefi et al., 2017

Functional genomics study of orchid development and breeding

Orchid genome sequencing initiatives and other cuttingedge technologies, such as genome editing tools are undoubtedly facilitating molecular genetic studies on orchid reproductive development. The genome sequencing of the tropical epiphytic orchid *P. equestris*, which provide an important resource for beginning to explore orchid diversity and evolution at the genome level, was a significant step forward in orchid genome study (Cai et al., 2015). It is now possible to identify and compare gene families that might have new functions across the whole genome with the availability of whole genome sequences (Lin et al., 2016; Cao et al., 2019; Chen T. C. et al., 2020; Song et al., 2021). As most orchid plants contain both C4 metabolism and CAM, phosphoenolpyruvate carboxylase (PEPC) plays an important role in photosynthetic performance and CO_2 efficiency. For green plants, especially CAM plants, little is known about the evolutionary history of the *PEPC* gene family. Using high-throughput sequencing and comprehensive phylogenetic analysis, the results indicated that CAM or C4-related PEPC may originate from the PPC-1M1 clade. The WGD event was responsible for the increment of *PEPC* gene copies (Deng et al., 2016). The plant-specific YABBY TFs regulate leaf polarity. Two *DROOPING LEAF/CRABS CLAW* (*DL/CRC*) genes were discovered in *P. equestris*, where *PeDLs* have demonstrated conserved function in floral meristem and carpel development (Chen et al., 2021). Protocorm-like bodies (PLBs) are commonly utilized in orchid micropropagation (Ren et al., 2020). According to certain research, SHOOT MERISTEMLESS (STM)-dependent organogenesis is required for PLB formation (Fang et al., 2022). Overexpression of *PaSTM* improved the regeneration from vegetative tissue-based explants of *Phalaenopsis*.

Moreover, many studies have demonstrated that *MADS-box* family genes control flower formation and morphogenesis (Teo et al., 2019). So far, a total of 51, 56, and 63 putative ones have been noticed in *P. equestris*, *P. aphrodite* and *D. catenatum*, respectively (Cai et al., 2015; Zhang et al., 2016; Chao et al., 2018). Despite having fewer *MADS-box* genes than *Arabidopsis* (107 genes) and rice (80 genes), orchids have more *MADS-box* genes involved in floral organ production (Leseberg et al., 2006). This distinction suggests that higher *MADS-box* gene diversity might be connected with highly specific floral morphological traits in orchids (Cai et al., 2015; Chao et al., 2018). This hypothesis is backed further by the fact that the number of *MADS-box* genes differs across Apostasioideae and the other orchid subfamilies. *A. shenzhenica*, a member of the Apostasioideae subfamily, yields solanum-type flowers with undifferentiated lips and somewhat simple gynostemia (Chen et al., 2012). *A. shenzhenica* contains fewer B-class AP3-clade and E-class *MADS-box* genes than *Dendrobium* and *Phalaenopsis* (Cai et al., 2015; Zhang et al., 2016). Notably, all modern orchids have shared a WGD event, which may be related to their diversification (Zhang et al., 2017; Yuan et al., 2018). The B-class AP3-clade and E-class genes may have increased just after WGD in the common ancestor of all orchids. Nevertheless, their paralogous genes may have been eliminated in *Apostasia*, culminating in a reversion to an earlier form with the plesiomorphic bloom (Zhang et al., 2017).

In the long term, the orchid breeding paradigm has seen the transition from conventional selection to cross-breeding, from molecular-assisted breeding to gene editing breeding (Li et al., 2021). Except for some self-incompatible species, the hybrid progeny preserve the parents' superior genetic features (Niu et al., 2018). However, the fertility of the hybrid combination and the genetic instability of the embryo after fertilization, the mapping of important agronomic traits and the selection of homozygotes are challenges (Su et al., 2019). Among them, seed germination is closely related to hybridization efficiency. When hybrid seeds are obtained, a proper cultivation technique is required to maintain the population. *In vitro* cultivation is a common method of seed propagation that has been used in the cultivation of numerous orchid species (Gao et al., 2020). The major goals of *in vitro* propagation are hybrid gex and a reduced breeding cycle. Mutagenesis breeding is also broadly applied for selecting elite crop and horticultural plant varieties. Many orchid varieties, including *Dendrobium*, *Phalaenopsis*, *Cymbidium*, *Oncidium*, *etc.*, have successfully undergone polyploid breeding by colchicine induction (Vilcherrez-Atoche et al., 2022). The high heterozygosity of orchids can lead to an increase in the perceived mutation rate and result in a flurry of good mutation types. However, unpredictable mutations can occur throughout the genome, and those negative mutations may occur, with only minor changes frequently achieved (Su et al., 2019). Molecular marker-assisted breeding is fast, efficient, and independent of environmental factors. Techniques such as AFLP, RFLP, SSR, RAPD, *etc.* are regularly employed to identify trait-related differential sequences (Poczai et al., 2013). These markers, when combined with function annotation given by unigenes, enable the identification of candidates with specific roles. Moreover, the completion of large-scale chromosome-level genomes lays the foundation for gene editing breeding and precise breeding based on features.

Discussion

Polyploidy is the driving force behind species adaptation, diversity, and genome evolution. Some superior orchid cultivars are produced through chromosomal polyploidy in the domain of horticulture (Vilcherrez-Atoche et al., 2022). Domestication and polyploidy have a close link since polyploid plants are randomly selected for their greater vigor, and consequently, polyploid species are more profitable and attractive for domestication than wild ones. The size of a genome is mostly determined by endoreplication and

LTR retrotransposon insertion during expansion (Chumová et al., 2021). Initially, FCM and k-mer analysis was used to calculate the size of these genomes. Large-scale tandem duplication and segmental duplication within the chromosome drive the generation of new genes and species evolution (Clark and Donoghue, 2018). In most cases, orchids underwent WGD more than once, including a historical WGD event and a recent WGD event shared by all orchids. There are both mycoheterotrophic and parasitic orchids, in addition to the vast majority of ornamental orchids. The loss and survival of symbiotic genes related to the evolution of specific symbionts span from the ancestral arbuscular mycorrhiza to the recent ericoid and orchid mycorrhizae (Barrett et al., 2019; Gao et al., 2020). Fully mycoheterotrophic orchids look very different from other orchids. This might be due to the loss of most of their photoreceptor and auxin transporter genes. Large-scale resequencing has been utilized to pinpoint key genes or chromosomal regions linked with some trait characteristics. GWAS based on GBS has sparked a lot of interest in several orchids. Some valuable SNP markers are widely applied to discriminate against orchid varieties (Kumagai et al., 2019). Furthermore, a small single-copy region in the cp genome of *Paphiopedilum* lost a large number of sequences, implying its significance in adaptive evolution (Trávníček et al., 2015). In this study, a phylogenetic tree of 58 orchid species was constructed to investigate the relationship of cp genomes within five subfamilies. The major sequenced species are those designated as Epidendroideae and Orchidoideae. MADS-box transcriptional factors are one of the most studied gene families in orchids, with evidence that they are involved in the regulation of various developmental processes as well as responses to environmental stimuli (Teo et al., 2019). The biological functions of these MADS-box proteins and the mechanisms by which they contribute to flowering or floral organ development are detailed. The molecular mechanisms underpinning flowering and floral development can be exploited for both traditional orchid breeding and targeted manipulation for desired blooming features.

Despite recent advancements in the field of orchid reproductive development, molecular genetic studies of flowering initiation and development continue to lag behind those in other model plants due to a number of bottlenecks. These included the prolonged vegetative stage, the inefficiency of established genetic transformation systems, and available data on genome sequences (Wang et al., 2017). Consequently, the majority of studies on orchid reproductive development have concentrated on genes that are homologs of other well-known genes in model plants.The duplication of genes in the genomes of some orchids may be beneficial for the inheritance of specific characteristics that contribute to the adaption to various environments. Furthermore, clarifying the inherent roles of the key genes in homologous orchid transgenic systems is critical (Hsing et al., 2016; Zhang et al., 2017). This technique involves the ongoing development of a few orchid-specific technical platforms, such as *in vitro* tissue culture, gene transformation, and genome editing tools (Hsiao et al., 2011; Li et al., 2022b). Many recent studies on the crop pan-genome have successfully identified core genes, individual-specific genes, and structural variation between many subspecies, providing new insights into the genetic underpinning of intricate biological characteristics (Liao et al., 2004; Li et al., 2021). A pan-genome encompasses more genetic variation within plants than a single reference genome. Therefore, another research hotspot of orchids may be concentrated on pan-genome and next-generation breeding technologies under the genetic background of different species (Tsai et al., 2017). Together, these efforts and the ever-improving use of multi-omics techniques to find specific molecular markers linked with morphological changes in orchid reproductive development will pave the way to figure out the molecular basis of specialized orchid reproductive processes.

Author contributions

CS, FZ, and YC discussed the writing plan. CS, YW and DM wrote the draft manuscript. CS, MM, and PPW edited the manuscript. FZ and CS acquired the funding. All the authors have read and approved the submitted version.

Funding

This work was supported by Anhui Province Postdoctoral Fund (2020B454), High-level Talents Research Initiation Fund of West Anhui University (WGKQ2022025), Postdoctoral Fund of West Anhui University (WXBSH2019001), and Anhui Provincial Administration of Traditional Chinese Medicine Project (2020zcyb09).

Acknowledgments

We apologize to those authors whose excellent work could not be cited because of space restrictions.

Conflict of interest

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

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.1018029/full#supplementary-material

SUPPLEMENTARY FIGURE 1
The phylogenetic tree of 11 orchid species with publicly available protein sequences based on the identified single-copy genes. *A. thaliana* was regarded as an outgroup. The tree was visulized by the iTOL online service (https://itol.embl.de/).

SUPPLEMENTARY FIGURE 2
The maximum-likelihood (ML) tree of 58 Orchidaceae species based on the chloroplast genomes. Alignments of the cp genomes were performed using MAFFT (*v*7.505) based on the FFT-NS-2 method (https://mafft.cbrc.jp/alignment/software/). The Archaeopteryx.js tool was used to display the ML tree (https://sites.google.com/site/cmzmasek/home/software/archaeopteryx).

References

Ai, Y., Li, Z., Sun, W. H., Chen, J., Zhang, D., Ma, L., et al. (2021). The cymbidium genome reveals the evolution of unique morphological traits. *Hortic. Res.* 8, 1–15. doi: 10.1038/s41438-021-00683-z

Ai, Y., Xie, T. X., Chen, J., Zhou, J., Chen, M. K., and Liu, Z. J. (2019a). The complete chloroplast genome of cymbidium floribundum var. pumilum (Orchidaceae). *Mitochondrial DNA Part B. Resour.* 4, 3648–3649. doi: 10.1080/23802359.2019.1678419

Ai, Y., Xie, T. X., Liu, D. K., Tu, X., Zhou, J., and Liu, Z. J. (2019b). Complete chloroplast genome of arundina graminifolia (Orchidaceae). *Mitochondrial DNA Part B. Resour.* 4, 2898–2899. doi: 10.1080/23802359.2019.1660281

Amiryousefi, A., Hyvonen, J., and Poczai, P. (2017). The plastid genome of vanillon (Vanilla pompona, orchidaceae). *Mitochondrial DNA Part B. Resour.* 2, 689–691. doi: 10.1080/23802359.2017.1383201

Bae, E. K., An, C., Kang, M. J., Lee, S. A., Lee, S. J., Kim, K. T., et al. (2022). Chromosome-level genome assembly of the fully mycoheterotrophic orchid gastrodia elata. *G3 Genes. Genomes. Genet.* 12, 1–11. doi: 10.1093/g3journal/jkab433

Barrett, C. F., and Kennedy, A. H. (2018). Plastid genome degradation in the endangered, mycoheterotrophic, north American orchid hexalectris warnockii. *Genome Biol. Evol.* 10, 1657–1662. doi: 10.1093/gbe/evy107

Barrett, C. F., Sinn, B. T., Kennedy, A. H., and Pupko, T. (2019). Unprecedented parallel photosynthetic losses in a heterotrophic orchid genus. *Mol. Biol. Evol.* 36, 1884–1901. doi: 10.1093/molbev/msz111

Barrett, C. F., Wicke, S., and Sass, C. (2018). Dense infraspecific sampling reveals rapid and independent trajectories of plastome degradation in a heterotrophic orchid complex. *New Phytol.* 218, 1192–1204. doi: 10.1111/nph.15072

Bertrand, J. A. M., Gibert, A., Llauro, C., and Panaud, O. (2019). Characterization of the complete plastome of ophrys aveyronensis, a Euro-Mediterranean orchid with an intriguing disjunct geographic distribution. *Mitochondrial DNA Part B. Resour.* 4, 3256–3257. doi: 10.1080/23802359.2019.1670748

Bolaños-Villegas, P., and Chen, F. C. (2007). Cytological identification of chromosomal rearrangements in doritaenopsis and phalaenopsis. *J. Int. Coop.* 2, 1–11. doi: 10.13140/RG.2.1.1641.8007

Cai, J., Liu, X., Vanneste, K., Proost, S., Tsai, W.-C., Liu, K.-W., et al. (2015). The genome sequence of the orchid phalaenopsis equestris. *Nat. Genet.* 47, 65–72. doi: 10.1038/ng.3149

Cao, Y., Meng, D., Han, Y., Chen, T., Jiao, C., Chen, Y., et al. (2019). Comparative analysis of b-BOX genes and their expression pattern analysis under various treatments in *Dendrobium officinale*. *BMC Plant Biol.* 19, 1–16. doi: 10.1186/s12870-019-1851-6

Chang, C. C., Lin, H. C., Lin, I. P., Chow, T. Y., Chen, H. H., Chen, W. H., et al. (2006). The chloroplast genome of phalaenopsis aphrodite (Orchidaceae): Comparative analysis of evolutionary rate with that of grasses and its phylogenetic implications. *Mol. Biol. Evol.* 23, 279–291. doi: 10.1093/molbev/msj029

Chao, Y. T., Chen, W. C., Chen, C. Y., Ho, H. Y., Yeh, C. H., Kuo, Y. T., et al. (2018). Chromosome-level assembly, genetic and physical mapping of phalaenopsis aphrodite genome provides new insights into species adaptation and resources for orchid breeding. *Plant Biotechnol. J.* 16, 2027–2041. doi: 10.1111/pbi.12936

Chase, M. W., Cameron, K. M., Freudenstein, J. V., Pridgeon, A. M., Salazar, G., van den Berg, C., et al. (2015). An updated classification of orchidaceae. *Bot. J. Linn. Soc* 177, 151–174. doi: 10.1111/boj.12234

Chen, J., Chen, M.-K., Zheng, Q.-D., Ma, S.-H., Liu, Z.-J., and Ai, Y. (2020). Chloroplast characterization and phylogenetic relationship of cymbidium aloifolium (Orchidaceae). *Mitochondrial DNA Part B.* 5, 478–479. doi: 10.1080/23802359.2019.1704656

Chen, M. K., Chen, J., Zhou, J., Ma, S. H., Zheng, Q. D., Xie, T. X., et al. (2019). The complete chloroplast genome sequence of habenaria ciliolaris (Orchidaceae). *Mitochondrial DNA Part B. Resour.* 4, 4132–4133. doi: 10.1080/23802359.2019.1692727

Chen, Y. Y., Hsiao, Y. Y., Li, C. I., Yeh, C. M., Mitsuda, N., Yang, H. X., et al. (2021). The ancestral duplicated DL/CRC orthologs, PeDL1 and PeDL2, function in orchid reproductive organ innovation. *J. Exp. Bot.* 72, 5442–5461. doi: 10.1093/jxb/erab195

Chen, Y. Y., Lee, P. F., Hsiao, Y. Y., Wu, W. L., Pan, Z. J., Lee, Y. I., et al. (2012). C-and d-class MADS-box genes from phalaenopsis equestris (orchidaceae) display functions in gynostemium and ovule development. *Plant Cell Physiol.* 53, 1053–1067. doi: 10.1093/pcp/pcs048

Chen, T. C., Liu, Y. C., Wang, X., Wu, C. H., Huang, C. H., and Chang, C. C. (2017). Whole plastid transcriptomes reveal abundant RNA editing sites and differential editing status in phalaenopsis aphrodite subsp. formosana. *Bot. Stud.* 58, 1–14. doi: 10.1186/s40529-017-0193-7

Chen, T. C., Su, Y. Y., Wu, C. H., Liu, Y. C., Huang, C. H., and Chang, C. C. (2020). Analysis of mitochondrial genomics and transcriptomics reveal abundant RNA edits and differential editing status in moth orchid, phalaenopsis aphrodite subsp. formosana. *Sci. Hortic. (Amsterdam).* 267, 1–13. doi: 10.1016/j.scienta.2020.109304

Chen, S., Wang, X., Wang, Y., Zhang, G., Song, W., Dong, X., et al. (2020). Improved *de novo* assembly of the achlorophyllous orchid gastrodia elata. *Front. Genet.* 11. doi: 10.3389/fgene.2020.580568

Choi, H., Lyu, J., Lee, H. O., Kim, J. B., and Kim, S. H. (2020). Complete chloroplast genome sequence of an orchid hybrid cymbidium sinense (♀) × c. goeringii (♂). *Mitochondrial DNA Part B. Resour.* 5, 3802–3803. doi: 10.1080/23802359.2020.1839367

Chumová, Z., Záveská, E., Hloušková, P., Ponert, J., Schmidt, P. A., Čertner, M., et al. (2021). Repeat proliferation and partial endoreplication jointly shape the patterns of genome size evolution in orchids. *Plant J.* 107, 511–524. doi: 10.1111/tpj.15306

Chung, O., Kim, J., Bolser, D., Kim, H. M., Jun, J. H., Choi, J. P., et al. (2022). A chromosome-scale genome assembly and annotation of the spring orchid (Cymbidium goeringii). *Mol. Ecol. Resour.* 22, 1168–1177. doi: 10.1111/1755-0998.13537

Clark, J. W., and Donoghue, P. C. J. (2018). Whole-genome duplication and plant macroevolution. *Trends Plant Sci.* 23, 933–945. doi: 10.1016/j.tplants.2018.07.006

da Rocha Perini, V., Leles, B., Furtado, C., and Prosdocimi, F. (2016). Complete chloroplast genome of the orchid cattleya crispata (Orchidaceae:Laeliinae), a Neotropical rupiculous species. *Mitochondrial DNA Part A. DNA Mapping. Seq. Anal.* 27, 4075–4077. doi: 10.3109/19401736.2014.1003850

Deng, H., Zhang, L. S., Zhang, G. Q., Zheng, B. Q., Liu, Z. J., and Wang, Y. (2016). Evolutionary history of PEPC genes in green plants: Implications for the evolution of CAM in orchids. *Mol. Phylogenet. Evol.* 94, 559–564. doi: 10.1016/j.ympev.2015.10.007

Du, Z., Yang, X., Tan, G., and Chen, Z. (2021). The complete chloroplast genome of cymbidium dayanum (Orchidaceae). *Mitochondrial DNA Part B. Resour.* 6, 1897–1898. doi: 10.1080/23802359.2021.1934173

Fang, S.-C., Chen, J.-C., Chang, P.-Y., and Lin, H.-Y. (2022). Co-Option of the SHOOT MERISTEMLESS network regulates protocorm-like body development in phalaenopsis aphrodite. *Plant Physiol.* 1–19, 1–19. doi: 10.1093/plphys/kiac100

Fan, J., and Huang, M. Y. (2019). Chloroplast genome structure and phylogeny of spiranthes sinensis, an endangered medicinal orchid plant. *Mitochondrial DNA Part B. Resour.* 4, 2994–2996. doi: 10.1080/23802359.2019.1664345

Fan, Z. F., Yu, D. Y., and Ma, C. (2021). The complete chloroplast genome sequence of phalaenopsis wilsonii Rolfe, a vulnerable wild moth orchid species (Orchidaceae). *Mitochondrial DNA Part B. Resour.* 6, 2903–2905. doi: 10.1080/23802359.2021.1923420

Felix, L. P., and Guerra, M. (2010). Variation in chromosome number and the basic number of subfamily epidendroideae (Orchidaceae). *Bot. J. Linn. Soc* 163, 234–278. doi: 10.1111/j.1095-8339.2010.01059.x

Feng, Y. L., Wicke, S., Li, J. W., Han, Y., Lin, C. S., Li, D. Z., et al. (2016). Lineage-specific reductions of plastid genomes in an orchid tribe with partially and fully mycoheterotrophic species. *Genome Biol. Evol.* 8, 2164–2175. doi: 10.1093/gbe/evw144

Gamisch, A., and Comes, H. P. (2019). Clade-age-dependent diversification under high species turnover shapes species richness disparities among tropical rainforest lineages of bulbophyllum (Orchidaceae). *BMC Evol. Biol.* 19, 1–16. doi: 10.1186/s12862-019-1416-1

Gao, Y., Zhao, Z., Li, J., Liu, N., Jacquemyn, H., Guo, S., et al. (2020). Do fungal associates of co-occurring orchids promote seed germination of the widespread orchid species gymnadenia conopsea? *Mycorrhiza* 30, 221–228. doi: 10.1007/s00572-020-00943-1

Ge, L. P., Tang, L., Li, L., and Luo, Y. (2020). The complete chloroplast genome of an endangered orchid paphiopedilum spicerianum. *Mitochondrial DNA Part B. Resour.* 5, 3594–3595. doi: 10.1080/23802359.2020.1830727

Guo, M., Chen, H., Dong, S., Zhang, Z., and Luo, H. (2022). CRISPR-cas gene editing technology and its application prospect in medicinal plants. *Chin. Med. (United. Kingdom).* 17, 1–19. doi: 10.1186/s13020-022-00584-w

Han, B., Jing, Y., Dai, J., Zheng, T., Gu, F., Zhao, Q., et al. (2020). A chromosome-level genome assembly of dendrobium huoshanense using long reads and hi-c data. *Genome Biol. Evol.* 12, 2486–2490. doi: 10.1093/GBE/EVAA215

He, L. F., Qiang, S. J., and Zhang, Y. H. (2021). The complete chloroplast genome of pleione maculata, an orchid with important ornamental value and medicinal value. *Mitochondrial DNA Part B. Resour.* 6, 2263–2265. 69, 1–16. doi: 10.1080/23802359.2021.1948366

He, T., Ye, C., Zeng, Q., Fan, X., and Huang, T. (2022). Genetic diversity and population structure of cultivated dendrobium nobile lindl. in southwest of China based on genotyping-by-sequencing. *Genet. Resour. Crop Evol.* doi: 10.1007/s10722-022-01401-x

Hsiao, Y. Y., Fu, C. H., Ho, S. Y., Li, C. I., Chen, Y. Y., Wu, W. L., et al. (2021). OrchidBase 4.0: a database for orchid genomics and molecular biology. *BMC Plant Biol.* 21, 1–11. doi: 10.1186/s12870-021-03140-0

Hsiao, Y. Y., Pan, Z. J., Hsu, C. C., Yang, Y. P., Hsu, Y. C., Chuang, Y. C., et al. (2011). Research on orchid biology and biotechnology. *Plant Cell Physiol.* 52, 1467–1486. doi: 10.1093/pcp/pcr100

Hsing, H. X., Lin, Y. J., Tong, C. G., Li, M. J., Chen, Y. J., and Ko, S. S. (2016). Efficient and heritable transformation of phalaenopsis orchids. *Bot. Stud.* 57. doi: 10.1186/s40529-016-0146-6

Hsu, C. C., Chen, S. Y., Chiu, S. Y., Lai, C. Y., Lai, P. H., Shehzad, T., et al. (2022). High-density genetic map and genome-wide association studies of aesthetic traits in phalaenopsis orchids. *Sci. Rep.* 12, 1–15. doi: 10.1038/s41598-022-07318-w

Hsu, C. C., Chen, S. Y., Lai, P. H., Hsiao, Y. Y., Tsai, W. C., Liu, Z. J., et al. (2020). Identification of high-copy number long terminal repeat retrotransposons and their expansion in phalaenopsis orchids. *BMC Genomics* 21, 1–13. doi: 10.1186/s12864-020-07221-6

Hsu, C. C., Chung, Y. L., Chen, T. C., Lee, Y. L., Kuo, Y. T., Tsai, W. C., et al. (2011). An overview of the phalaenopsis orchid genome through BAC end sequence analysis. *BMC Plant Biol.* 11, 1–13. doi: 10.1186/1471-2229-11-3

Huang, J. Z., Lin, C. P., Cheng, T. C., Huang, Y. W., Tsai, Y. J., Cheng, S. Y., et al. (2016). The genome and transcriptome of phalaenopsis yield insights into floral organ development and flowering regulation. *PeerJ* 2016, 1–17. doi: 10.7717/peerj.2017

Hu, C., Jiang, K., Zeng, X., and Huang, W. (2022). The complete chloroplast genome sequence of a critically endangered orchid paphiopedilum gratrixianum (Orchidaceae). *Mitochondrial DNA Part B.* 7, 609–610. doi: 10.1080/23802359.2021.1891005

Huo, X., Zhao, Y., Qian, Z., and Liu, M. (2018). Characterization of the complete chloroplast genome of eulophia zollingeri, an endangered orchid in China. *Conserv. Genet. Resour.* 10, 817–819. doi: 10.1007/s12686-017-0938-3

Hu, Y., Resende, M. F. R., Bombarely, A., Brym, M., Bassil, E., and Chambers, A. H. (2019). Genomics-based diversity analysis of vanilla species using a vanilla planifolia draft genome and genotyping-By-Sequencing. *Sci. Rep.* 9, 1–16. doi: 10.1038/s41598-019-40144-1

Hu, G., Zhou, H., Zhang, S., and Zhao, P. (2020). Characterization of the complete chloroplast genome of orchid family species cymbidium bicolor (Orchidaceae). *Mitochondrial DNA Part B. Resour.* 5, 323–324. doi: 10.1080/23802359.2019.1703591

Jersáková, J., Trávníček, P., Kubátová, B., Krejčíková, J., Urfus, T., Liu, Z. J., et al. (2013). Genome size variation in orchidaceae subfamily apostasioideae: Filling the phylogenetic gap. *Bot. J. Linn. Soc* 172, 95–105. doi: 10.1111/boj.12027

Jheng, C. F., Chen, T. C., Lin, J. Y., Chen, T. C., Wu, W. L., and Chang, C. C. (2012). The comparative chloroplast genomic analysis of photosynthetic orchids and developing DNA markers to distinguish phalaenopsis orchids. *Plant Sci.* 190, 62–73. doi: 10.1016/j.plantsci.2012.04.001

Jiang, Y., Hu, X., Yuan, Y., Guo, X., Chase, M. W., Ge, S., et al. (2022). The gastrodia menghaiensis (Orchidaceae) genome provides new insights of orchid mycorrhizal interactions. *BMC Plant Biol.* 22, 1–14. doi: 10.1186/s12870-022-03573-1

Jiang, Y. T., Lin, R. Q., Liu, B., Zeng, Q. M., Liu, Z. J., and Chen, S. P. (2019). Complete chloroplast genome of cymbidium ensifolium(Orchidaceae). *Mitochondrial DNA Part B. Resour.* 4, 2236–2237. doi: 10.1080/23802359.2019.1624637

Jiang, L., Lin, M., Wang, H., Song, H., Zhang, L., Huang, Q., et al. (2022). Haplotype-resolved genome assembly of bletilla striata (thunb .) reichb . f . to elucidate medicinal value. *Plant J.* 111, 1–14. doi: 10.1111/tpj.15892

Kao, H., Zhao, Y., Yang, M., Sun, Y., and Cheng, J. (2021). The complete chloroplast genome sequences of an endangered orchid species paphiopedilum parishii (Orchidaceae). *Mitochondrial DNA Part B. Resour.* 6, 2521–2522. doi: 10.1080/23802359.2021.1959437

Kayika Febryanti, N. L. P., Nurliana, S., Gutierrez-Marcos, J., and Semiarti, E. (2020). The expression analysis of AtRKD4 transgene in dendrobium lineale rolfe transgenic orchid carrying 35S::GR::AtRKD4 for micropropagation. *AIP. Conf. Proc.* 2260, 1–6. doi: 10.1063/5.0015876

Khan, H., Marya,, Belwal, T., Mohd Tariq,, Atanasov, A. G., and Devkota, H. P. (2019). Genus vanda: A review on traditional uses, bioactive chemical constituents and pharmacological activities. *J. Ethnopharmacol.* 229, 46–53. doi: 10.1016/j.jep.2018.09.031

Khumkarjorn, N., Thanonkeo, S., Yamada, M., Klanrit, P., and Thanonkeo, P. (2017). Agrobacterium-mediated transformation of dendrobium orchid with the flavanone 3-hydroxylase gene. *Turk. J. Bot.* 41, 442–454. doi: 10.3906/bot-1701-13

Kim, Y. K., Jo, S., Cheon, S. H., Joo, M. J., Hong, J. R., Kwak, M. H., et al. (2019). Extensive losses of photosynthesis genes in the plastome of a mycoheterotrophic orchid, cyrtosia septentrionalis (Vanilloideae: Orchidaceae). *Genome Biol. Evol.* 11, 565–571. doi: 10.1093/gbe/evz024

Kim, G. B., Kwon, Y., Yu, H. J., Lim, K. B., Seo, J. H., and Mun, J. H. (2016). The complete chloroplast genome of phalaenopsis "Tiny star". *Mitochondrial DNA* 27, 1300–1302. doi: 10.3109/19401736.2014.945566

Konhar, R., Debnath, M., Vishwakarma, S., Bhattacharjee, A., Sundar, D., Tandon, P., et al. (2019). The complete chloroplast genome of dendrobium nobile, an endangered medicinal orchid from north-east India and its comparison with related dendrobium species. *PeerJ* 2019, 1–28. doi: 10.7717/peerj.7756

Kui, L., Chen, H., Zhang, W., He, S., Xiong, Z., Zhang, Y., et al. (2017). Building a genetic manipulation tool box for orchid biology: Identification of constitutive promoters and application of CRISPR/Cas9 in the orchid, dendrobium officinale. *Front. Plant Sci.* 7. doi: 10.3389/fpls.2016.02036

Kumagai, M., Nishikawa, D., Kawahara, Y., Wakimoto, H., Itoh, R., Tabei, N., et al. (2019). Tasuke1: A web-based platform for exploring GWAS results and large-scale resequencing data. *DNA Res.* 26, 445–452. doi: 10.1093/dnares/dsz022

Lee, Y. I., Chung, M. C., Kuo, H. C., Wang, C. N., Lee, Y. C., Lin, C. Y., et al. (2017). The evolution of genome size and distinct distribution patterns of rDNA in phalaenopsis (Orchidaceae). *Bot. J. Linn. Soc* 185, 65–80. doi: 10.1093/BOTLINNEAN/BOX049

Lee, S. Y., Kaikai, M., Zhou, R., and Liao, W. (2020). Severe plastid genome size reduction in a mycoheterotrophic orchid, danxiaorchis singchiana, reveals heavy gene loss and gene relocations. *Plants* 9, 521. doi: 10.3390/plants9040521

Lee, Y. I., Tseng, Y. F., Lee, Y. C., and Chung, M. C. (2020). Chromosome constitution and nuclear DNA content of phalaenopsis hybrids. *Sci. Hortic. (Amsterdam).* 262, 109089. doi: 10.1016/j.scienta.2019.109089

Leseberg, C. H., Li, A., Kang, H., Duvall, M., and Mao, L. (2006). Genome-wide analysis of the MADS-box gene family in populus trichocarpa. *Gene* 378, 84–94. doi: 10.1016/j.gene.2006.05.022

Liao, L. J., Pan, I. C., Chan, Y. L., Hsu, Y. H., Chen, W. H., and Chan, M. T. (2004). Transgene silencing in phalaenopsis expressing the coat protein of cymbidium mosaic virus is a manifestation of RNA-mediated resistance. *Mol. Breed.* 13, 229–242. doi: 10.1023/B:MOLB.0000022527.68551.30

Liau, C. H., You, S. J., Prasad, V., Hsiao, H. H., Lu, J. C., Yang, N. S., et al. (2003). Agrobacterium tumefaciens-mediated transformation of an oncidium orchid. *Plant Cell Rep.* 21, 993–998. doi: 10.1007/s00299-003-0614-9

Li, T. Z., Chen, L. J., Wang, M., Chen, J. B., and Huang, J. (2019). The complete chloroplast genome of vanilla shenzhenica (Orchidaceae). *Mitochondrial DNA Part B. Resour.* 4, 2610–2611. doi: 10.1080/23802359.2019.1642165

Li, C., Dong, N., Zhao, Y., Wu, S., Liu, Z., and Zhai, J. (2021). A review for the breeding of orchids: Current achievements and prospects. *Hortic. Plant J.* 7, 380–392. doi: 10.1016/j.hpj.2021.02.006

Li, Y., Li, Z., Hu, Q., Zhai, J., Liu, Z., and Wu, S. (2019). Complete plastid genome of apostasia shenzhenica (Orchidaceae). *Mitochondrial DNA Part B. Resour.* 4, 1388–1389. doi: 10.1080/23802359.2019.1591192

Li, M.-H., Liu, K.-W., Li, Z., Lu, H.-C., Ye, Q.-L., Zhang, D., et al. (2022). Genomes of leafy and leafless platanthera orchids illuminate the evolution of mycoheterotrophy. *Nat. Plants* 8, 373–388. doi: 10.1038/s41477-022-01127-9

Li, H., Li, R., Zhang, J., Yin, Q., and Zhang, Y. (2019). Complete chloroplast genome of ornamental orchids eulophia flava. *Mitochondrial DNA Part B. Resour.* 4, 2997–2998. doi: 10.1080/23802359.2019.1664946

Li, R., Wang, A., Sun, S., Liang, S., Wang, X., Ye, Q., et al. (2012). Functional characterization of FT and MFT ortholog genes in orchid (Dendrobium nobile Lindl) that regulate the vegetative to reproductive transition in Arabidopsis. *Plant Cell. Tissue Organ Cult.* 111, 143–151. doi: 10.1007/s11240-012-0178-x

Lin, B. Y., Chang, C. D., Huang, L. L. H., Liu, Y. C., Su, Y. Y., Chen, T. C., et al. (2016). The mitochondrial DNA markers for distinguishing phalaenopsis species and revealing maternal phylogeny. *Biol. Plant* 60, 68–78. doi: 10.1007/s10535-015-0566-2

Lin, H. Y., Chen, J. C., and Fang, S. C. (2018). A protoplast transient expression system to enable molecular, cellular, and functional studies in phalaenopsis orchids. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.00843

Lin, C. S., Chen, J. J. W., Huang, Y. T., Chan, M. T., Daniell, H., Chang, W. J., et al. (2015). The location and translocation of ndh genes of chloroplast origin in the orchidaceae family. *Sci. Rep.* 5, 1–10. doi: 10.1038/srep09040

Liu, J. F. (2020). The complete chloroplast genome sequence of liparis bootanensis (Orchidaceae). *Mitochondrial DNA Part B. Resour.* 5, 2058–2059. doi: 10.1080/23802359.2020.1763866

Liu, D. K., Tu, X., Zhang, S., and Li, M. H. (2020). The complete plastid genome of vanda subconcolor (Orchidaceae, aeridinae). *Mitochondrial DNA Part B. Resour.* 5, 1712–1713. doi: 10.1080/23802359.2020.1749169

Liu, F., Xiao, X., An, M., and Li, Z. (2021). The complete chloroplast genome of dendrobium comatum (Orchidaceae). *Mitochondrial DNA Part B. Resour.* 6, 3229–3230. doi: 10.1080/23802359.2021.1990152

Li, J., Xu, Y.c., and Wang, Z.h. (2019). Construction of a high-density genetic map by RNA sequencing and eQTL analysis for stem length and diameter in dendrobium (Dendrobium nobile × dendrobium wardianum). *Ind. Crops Prod.* 128, 48–54. doi: 10.1016/j.indcrop.2018.10.073

Li, Y., Zhang, B., and Yu, H. (2022a). Kilobase-scale genomic deletion of DOTFL1 in dendrobium orchids. *J. Genet. Genomics* 49, 81–84. doi: 10.1016/j.jgg.2021.07.008

Li, Y., Zhang, B., and Yu, H. (2022b). Molecular genetic insights into orchid reproductive development. *J. Exp. Bot.* 73, 1841–1852. doi: 10.1093/jxb/erac016

Li, M., Zhao, Z., He, J., Cheng, J., and Xie, L. (2019). The complete chloroplast genome sequences of a highly endangered orchid species paphiopedilum barbigerum (Orchidaceae). *Mitochondrial DNA Part B. Resour.* 4, 2928–2929. doi: 10.1080/23802359.2019.1660269

Lu, H., Liu, Z., and Lan, S. (2019). Genome sequencing reveals the role of MADS-box gene families in the floral morphology evolution of orchids. *Hortic. Plant J.* 5, 247–254. doi: 10.1016/j.hpj.2019.11.005

Luo, J., Hou, B. W., Niu, Z. T., Liu, W., Xue, Q. Y., and Ding, X. Y. (2014). Comparative chloroplast genomes of photosynthetic orchids: Insights into evolution of the orchidaceae and development of molecular markers for phylogenetic applications. *PloS One* 9, 1–15. doi: 10.1371/journal.pone.0099016

Ma, X., Lin, H., Chen, Y., Lan, S., and Ming, R. (2019). The complete chloroplast genome of a gynodioecious deciduous orchid satyrium ciliatum (Orchidaceae) female. *Mitochondrial DNA Part B. Resour.* 4, 3876–3877. doi: 10.1080/23802359.2019.1687359

Men, S., Ming, X., Wang, Y., Liu, R., Wei, C., and Li, Y. (2003). Genetic transformation of two species of orchid by biolistic bombardment. *Plant Cell Rep.* 21, 592–598. doi: 10.1007/s00299-002-0559-4

Minasiewicz, J., Krawczyk, E., Znaniecka, J., Vincenot, L., Zheleznaya, E., Korybut-Orlowska, J., et al. (2022). Weak population spatial genetic structure and low infraspecific specificity for fungal partners in the rare mycoheterotrophic orchid epipogium aphyllum. *J. Plant Res.* 135, 275–293. doi: 10.1007/s10265-021-01364-7

Mohammadi, M., Kaviani, B., and Sedaghathoor, S. (2021). *In vivo* polyploidy induction of phalaenopsis amabilis in a bubble bioreactor system using colchicine. *Ornam. Hortic.* 27, 204–212. doi: 10.1590/2447-536X.V27I2.2275

Molla, K. A., Sretenovic, S., Bansal, K. C., and Qi, Y. (2021). Precise plant genome editing using base editors and prime editors. *Nat. Plants* 7, 1166–1187. doi: 10.1038/s41477-021-00991-1

Molla, K. A., and Yang, Y. (2019). CRISPR/Cas-mediated base editing: Technical considerations and practical applications. *Trends Biotechnol.* 37, 1121–1142. doi: 10.1016/j.tibtech.2019.03.008

Mo, P., Zhou, J., Zhou, F., Chen, Y., Huang, K., Mo, P., et al. (2022). The complete chloroplast genome sequence of gomesa flexuosa. *Mitochondrial DNA Part B.* 7, 1237–1239. doi: 10.1080/23802359.2022.2093670

Niu, S. C., Huang, J., Xu, Q., Li, P. X., Yang, H. J., Zhang, Y. Q., et al. (2018). Morphological type identification of self-incompatibility in dendrobium and its phylogenetic evolution pattern. *Int. J. Mol. Sci.* 19, 1–18. doi: 10.3390/ijms19092595

Niu, Z., Pan, J., Zhu, S., Li, L., Xue, Q., Liu, W., et al. (2017a). Comparative analysis of the complete plastomes of apostasia wallichii and neuwiedia singapureana (Apostasioideae) reveals different evolutionary dynamics of IR/SSC boundary among photosynthetic orchi. *Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.01713

Niu, Z., Xue, Q., Zhu, S., Sun, J., Liu, W., and Ding, X. (2017b). The complete plastome sequences of four orchid species: Insights into the evolution of the orchidaceae and the utility of plastomic mutational hotspots. *Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.00715

Niu, Z., Zhu, F., Fan, Y., Li, C., Zhang, B., Zhu, S., et al. (2021). The chromosome-level reference genome assembly for dendrobium officinale and its utility of functional genomics research and molecular breeding study. *Acta Pharm. Sin. B.* 11, 2080–2092. doi: 10.1016/j.apsb.2021.01.019

Nopitasari, S., Setiawati, Y., Lawrie, M. D., Purwantoro, A., Widada, J., Sasongko, A. B., et al. (2020). Development of an agrobacterium-delivered crispr/cas9 for phalaenopsis amabilis (L.) blume genome editing system. *AIP. Conf. Proc.* 2260, 1–10. doi: 10.1063/5.0015868

Ogura, T., and Busch, W. (2015). From phenotypes to causal sequences: Using genome wide association studies to dissect the sequence basis for variation of plant development. *Curr. Opin. Plant Biol.* 23, 98–108. doi: 10.1016/j.pbi.2014.11.008

Pan, Y. Y., Li, T. Z., Chen, J. B., Huang, J., and Rao, W. H. (2019). Complete chloroplast genome of dendrobium thyrsiflorum (Orchidaceae). *Mitochondrial DNA Part B. Resour.* 4, 3192–3193. doi: 10.1080/23802359.2019.1667918

Paun, O., Bateman, R. M., Fay, M. F., Hedrén, M., Civeyrel, L., and Chase, M. W. (2010). Stable epigenetic effects impact adaptation in allopolyploid orchids (Dactylorhiza: Orchidaceae). *Mol. Biol. Evol.* 27, 2465–2473. doi: 10.1093/molbev/msq150

Peng, X., Ye, H., Liu, H., Zhao, Z., Hu, G., and Zhao, P. (2020). Characterization of the complete chloroplast genome of orchid family species cymbidium bicolor (Orchidaceae). *Mitochondrial DNA Part B. Resour.* 5, 323–324. doi: 10.1080/23802359.2019.1703591

Piet, Q., Droc, G., Marande, W., Sarah, G., Bocs, S., Klopp, C., et al. (2022). A chromosome-level, haplotype-phased vanilla planifolia genome highlights the challenge of partial endoreplication for accurate whole-genome assembly. *Plant Commun.* 3, 100330. doi: 10.1016/j.xplc.2022.100330

Poczai, P., Varga, I., Laos, M., Cseh, A., Bell, N., Valkonen, J. P. T., et al. (2013). Advances in plant gene-targeted and functional markers: A review. *Plant Methods* 9, 1–31. doi: 10.1186/1746-4811-9-6

Radhakrishnan, G. V., Keller, J., Rich, M. K., Vernié, T., Mbadinga Mbadinga, D. L., Vigneron, N., et al. (2020). An ancestral signalling pathway is conserved in intracellular symbioses-forming plant lineages. *Nat. Plants* 6, 280–289. doi: 10.1038/s41477-020-0613-7

Ren, R., Gao, J., Lu, C., Wei, Y., Jin, J., Wong, S. M., et al. (2020). Highly efficient protoplast isolation and transient expression system for functional characterization of flowering related genes in cymbidium orchids. *Int. J. Mol. Sci.* 21, 1–18. doi: 10.3390/ijms21072264

Rewers, M., Jedrzejczyk, I., Rewicz, A., and Jakubska-Busse, A. (2021). Genome size diversity in rare, endangered, and protected orchids in Poland. *Genes (Basel).* 12, 1–19. doi: 10.3390/genes12040563

Ryu, J., Kim, W. J., Im, J., Kang, K. W., Kim, S. H., Jo, Y. D., et al. (2019). Single nucleotide polymorphism (SNP) discovery through genotyping-by-sequencing (GBS) and genetic characterization of dendrobium mutants and cultivars. *Sci. Hortic. (Amsterdam).* 244, 225–233. doi: 10.1016/j.scienta.2018.09.053

Setiawati, Y., Nopitasari, S., Lawrie, M. D., Purwantoro, A., Widada, J., Sasongko, A. B., et al. (2020). Agrobacterium-mediated transformation facillitates the CRISPR/Cas9 genome editing system in dendrobium macrophyllum a. rich orchid. *AIP. Conf. Proc.* 2260, 1Њ9. doi: 10.1063/5.0016200

Shi, Y., Yang, L., Yang, Z., and Ji, Y. (2018). The complete chloroplast genome of pleione bulbocodioides (Orchidaceae). *Conserv. Genet. Resour.* 10, 21–25. doi: 10.1007/s12686-017-0753-x

Song, C., Li, G., Dai, J., and Deng, H. (2021). Genome-wide analysis of PEBP genes in dendrobium huoshanense: Unveiling the antagonistic functions of FT/TFL1 in flowering time. *Front. Genet.* 12. doi: 10.3389/fgene.2021.687689

Su, C. L., Chao, Y. T., Yen, S. H., Chen, C. Y., Chen, W. C., Chang, Y. C. A., et al. (2013). Orchidstra: An integrated orchid functional genomics database. *Plant Cell Physiol.* 54, 1–19. doi: 10.1093/pcp/pct004

Suetsugu, K. (2015). Autonomous self-pollination in the nectarless orchid pogonia minor. *Plant Species Biol.* 30, 37–41. doi: 10.1111/1442-1984.12037

Su, J., Jiang, J., Zhang, F., Liu, Y., Ding, L., Chen, S., et al. (2019). Current achievements and future prospects in the genetic breeding of chrysanthemum: a review. *Hortic. Res.* 6. doi: 10.1038/s41438-019-0193-8

Tang, F. L., Deng, L. L., Qin, H. Z., and Shi, Y. C. (2020). Complete chloroplast genome of paphiopedilum emersonii (Orchidaceae). *Mitochondrial DNA Part B. Resour.* 5, 3518–3519. doi: 10.1080/23802359.2020.1827069

Tang, J. M., Tang, F. L., and Shi, Y. C. (2020). The complete chloroplast genome of geodorum densiflorum (Orchidaceae). *Mitochondrial DNA Part B. Resour.* 5, 2056–2057. doi: 10.1080/23802359.2020.1763865

Teo, Z. W. N., Zhou, W., and Shen, L. (2019). Dissecting the function of MADS-box transcription factors in orchid reproductive development. *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.01474

Tong, C. G., Wu, F. H., Yuan, Y. H., Chen, Y. R., and Lin, C. S. (2020). High-efficiency CRISPR/Cas-based editing of phalaenopsis orchid MADS genes. *Plant Biotechnol. J.* 18, 889–891. doi: 10.1111/pbi.13264

Trávníček, P., Čertner, M., Ponert, J., Chumová, Z., Jersáková, J., and Suda, J. (2019). Diversity in genome size and GC content shows adaptive potential in orchids and is closely linked to partial endoreplication, plant life-history traits and climatic conditions. *New Phytol.* 224, 1642–1656. doi: 10.1111/nph.15996

Trávníček, P., Ponert, J., Urfus, T., Jersáková, J., Vrána, J., Hřibová, E., et al. (2015). Challenges of flow-cytometric estimation of nuclear genome size in orchids, a plant group with both whole-genome and progressively partial endoreplication. *Cytom. Part A.* 87, 958–966. doi: 10.1002/cyto.a.22681

Tsai, W. C., Dievart, A., Hsu, C. C., Hsiao, Y. Y., Chiou, S. Y., Huang, H., et al. (2017). Post genomics era for orchid research. *Bot. Stud.* 58, 1–22. doi: 10.1186/s40529-017-0213-7

Vilcherrez-Atoche, J. A., Iiyama, C. M., and Cardoso, J. C. (2022). Polyploidization in orchids: From cellular changes to breeding applications. *Plants* 11, 1–21. doi: 10.3390/plants11040469

Wang, J. Y., Liu, Z. J., Zhang, G. Q., and Peng, C. C. (2019). The complete chloroplast genome sequence of phalaenopsis lowii (Orchidaceae). *Mitochondrial DNA Part B. Resour.* 4, 3569–3570. doi: 10.1080/23802359.2019.1674715

Wang, H. M., Tong, C. G., and Jang, S. (2017). Current progress in orchid flowering/flower development research. *Plant Signal. Behav.* 12, 1–6. doi: 10.1080/15592324.2017.1322245

Wang, M., Wang, Z., Mao, Y., Lu, Y., Yang, R., Tao, X., et al. (2019). Optimizing base editors for improved efficiency and expanded editing scope in rice. *Plant Biotechnol. J.* 17, 1697–1699. doi: 10.1111/pbi.13124

Wang, T., Zhang, C., Zhang, H., and Zhu, H. (2021). CRISPR/Cas9-mediated gene editing revolutionizes the improvement of horticulture food crops. *J. Agric. Food Chem.* 69, 13260–13269. doi: 10.1021/acs.jafc.1c00104

Wei, Z., Chen, B., Cao, Y., Zheng, Y., Zhang, Y., Zhao, K., et al. (2021). The complete chloroplast genome of cymbidium hookerianum (Orchidaceae): genome structure and basic analysis. *Mitochondrial DNA Part B. Resour.* 6, 36–37. doi: 10.1080/23802359.2020.1845996

Williams, S. J., Gale, S. W., Hinsley, A., Gao, J., and St. John, F. A. V. (2018). Using consumer preferences to characterize the trade of wild-collected ornamental orchids in China. *Conserv. Lett.* 11, 1–8. doi: 10.1111/conl.12569

Wu, X.-Y., Li, T.-Z., Chen, G.-Z., Xu, Q., Pan, Y.-Y., and Chen, L. J. (2019). The complete chloroplast genome of dendrobium longicornu (Orchidaceae). *Mitochondrial DNA Part B.* 4, 3776–3777. doi: 10.1080/23802359.2019.1666049

Wu, S. S., Shen, L. M., Ling, R., Dai, Z. W., Liu, Z. J., and Lan, S. R. (2019). Next-generation sequencing yields the complete chloroplast genome of pleione chunii, a vulnerable orchid in China. *Mitochondrial DNA Part B. Resour.* 4, 2576–2578. doi: 10.1080/23802359.2019.1640642

Xia, K., Liu, D. K., and Wang, J. Y. (2021). The complete chloroplast genome sequence of phalaenopsis wilsoniii (Orchidaceae). *Mitochondrial DNA Part B. Resour.* 6, 3303–3305. doi: 10.1080/23802359.2021.1994889

Xie, T.-X., Yu, X., Zheng, Q.-D., Ma, S.-H., Liu, Z.-J., and Ai, Y. (2020). The complete chloroplast genome of tainia dunnii (Orchidaceae): genome structure and evolution. *Mitochondrial DNA Part B.* 5, 3–4. doi: 10.1080/23802359.2019.1693935

Xie, L., Zhou, S. S., Wang, M. G., Zeng, R. Z., Guo, H. R., and Zhang, Z. S. (2017). Creation and micropropagation of polyploids in cymbidium hybridum. *Acta Hortic.* 1167, 107–114. doi: 10.17660/ActaHortic.2017.1167.16

Xu, Y., Lei, Y., Su, Z., Zhao, M., Zhang, J., Shen, G., et al. (2021). A chromosome-scale gastrodia elata genome and large-scale comparative genomic analysis indicate convergent evolution by gene loss in mycoheterotrophic and parasitic plants. *Plant J.* 108, 1609–1623. doi: 10.1111/tpj.15528

Xu, Q., Niu, S.-C., Li, K.-L., Zheng, P.-J., Zhang, X.-J., Jia, Y., et al. (2022). Chromosome-scale assembly of the dendrobium nobile genome provides insights into the molecular mechanism of the biosynthesis of the medicinal active ingredient of dendrobium. *Front. Genet.* 13. doi: 10.3389/fgene.2022.844622

Yang, F. X., Gao, J., Wei, Y. L., Ren, R., Zhang, G. Q., Lu, C. Q., et al. (2021). The genome of cymbidium sinense revealed the evolution of orchid traits. *Plant Biotechnol. J.* 19, 2501–2516. doi: 10.1111/pbi.13676

Yang, L., Wu, Q., Yang, M., Zhang, D., Dong, S., and Cheng, J. (2021). The complete chloroplast genome sequence of the endemic and rare orchid nothodoritis zhejiangensis (Orchidaceae) in China. *Mitochondrial DNA Part B. Resour.* 6, 2931–2932. doi: 10.1080/23802359.2021.1972867

Yang, J., Zhang, F., Ge, Y., Yu, W., Xue, Q., Wang, M., et al. (2022). Effects of geographic isolation on the bulbophyllum chloroplast genomes. *BMC Plant Biol.* 22, 1–14. doi: 10.1186/s12870-022-03592-y

Yan, L., Wang, X., Liu, H., Tian, Y., Lian, J., Yang, R., et al. (2015). The genome of dendrobium officinale illuminates the biology of the important traditional Chinese orchid herb. *Mol. Plant* 8, 922–934. doi: 10.1016/j.molp.2014.12.011

Younis, A., Ryu, K. B., Co, V. T., Hwang, Y.-J., Jee, S. O., Kim, M.-S., et al. (2013). Analysis of chromosomes and nuclear DNA content in nine genotypes of cymbidium. *Korean. Soc Floric. Sci.* 21, 158–161. doi: 10.11623/frj.2013.21.4.31

Yuan, Y., Jin, X., Liu, J., Zhao, X., Zhou, J., Wang, X., et al. (2018). The gastrodia elata genome provides insights into plant adaptation to heterotrophy. *Nat. Commun.* 9, 1–11. doi: 10.1038/s41467-018-03423-5

Yue, Z., Kao, H., Zhang, Y., Wang, T., Dong, S., and Cheng, J. (2020). The complete chloroplast genome sequence of a medicinal orchid species coelogyne fimbriata (Orchidaceae). *Mitochondrial DNA Part B. Resour.* 5, 3458–3460. doi: 10.1080/23802359.2020.1823264

Yu, C. W., Lian, Q., Wu, K. C., Yu, S. H., Xie, L. Y., and Wu, Z. J. (2016). The complete chloroplast genome sequence of anoectochilus roxburghii. *Mitochondrial DNA* 27, 2477–2478. doi: 10.3109/19401736.2015.1033706

Yun, S. A., Son, H. D., Im, H. T., and Kim, S. C. (2018). Two complete chloroplast genomes of an endangered orchid species, pelatantheria scolopendrifolia (Orchidaceae), in Korea. *Mitochondrial DNA Part B. Resour.* 3, 225–226. doi: 10.1080/23802359.2018.1437815

Zeng, W. H., Liao, S. C., and Chang, C. C. (2007). Identification of RNA editing sites in chloroplast transcripts of phalaenopsis aphrodite and comparative analysis with those of other seed plants. *Plant Cell Physiol.* 48, 362–368. doi: 10.1093/pcp/pcl058

Zhang, D., Liu, D. K., Hao, Y., Lan, S. R., and Liu, Z. J. (2019). The complete chloroplast genome sequence of liparis vivipara (Orchidaceae). *Mitochondrial DNA Part B. Resour.* 4, 2223–2224. doi: 10.1080/23802359.2019.1624638

Zhang, G. Q., Liu, Z. J., and Lan, S. R. (2019). Complete chloroplast genome of an orchid species cymbidium floribundum lindl. *Mitochondrial DNA Part B. Resour.* 4, 2940–2941. doi: 10.1080/23802359.2019.1662743

Zhang, G. Q., Liu, K. W., Li, Z., Lohaus, R., Hsiao, Y. Y., Niu, S. C., et al. (2017). The apostasia genome and the evolution of orchids. *Nature* 549, 379–383. doi: 10.1038/nature23897

Zhang, Y. J., Ma, C., Feng, Y., Cheng, X., and Song, J. (2018). The complete chloroplast genome sequence of an endangered orchid species dendrobium bellatulum (Orchidaceae). *Mitochondrial DNA Part B. Resour.* 3, 233–234. doi: 10.1080/23802359.2018.1437811

Zhang, G. Q., Xu, Q., Bian, C., Tsai, W. C., Yeh, C. M., Liu, K. W., et al. (2016). The dendrobium catenatum lindl. genome sequence provides insights into polysaccharide synthase, floral development and adaptive evolution. *Sci. Rep.* 6, 1–10. doi: 10.1038/srep19029

Zhang, F. P., and Zhang, S. B. (2021). Genome size and labellum epidermal cell size are evolutionarily correlated with floral longevity in paphiopedilum species. *Front. Plant Sci.* 12 1–14. doi: 10.3389/fpls.2021.793516

Zhang, W., Zhang, G., Zeng, P., Zhang, Y., Hu, H., Liu, Z., et al. (2021). Genome sequence of apostasia ramifera provides insights into the adaptive evolution in orchids. *BMC Genomics* 22, 1–12. doi: 10.1186/s12864-021-07852-3

Zhang, Y., Zhang, G. Q., Zhang, D., Liu, X. D., Xu, X. Y., Sun, W. H., et al. (2021). Chromosome-scale assembly of the dendrobium chrysotoxum genome enhances the understanding of orchid evolution. *Hortic. Res.* 8. doi: 10.1038/s41438-021-00621-z

Zhao, Z., Li, M., He, J., Cheng, J., and Xie, L. (2019). Complete chloroplast genome sequences of an important horticultural orchid: Paphiopedilum hirsutissimum (Orchidaceae). *Mitochondrial DNA Part B. Resour.* 4, 2950–2951. doi: 10.1080/23802359.2019.1662752

Zheng, F., Chen, J. B., Liu, W. R., and Wang, M. (2021). The complete chloroplast genome of an endangered species apostasia ramifera (Orchidaceae). *Mitochondrial DNA Part B. Resour.* 6, 470–471. doi: 10.1080/23802359.2021.1872429

Zheng, S.g., Hu, Y.d., Zhao, R., Yan, S., Zhang, X.q., Zhao, T.m., et al. (2018). Genome-wide researches and applications on dendrobium. *Planta* 248, 769–784. doi: 10.1007/s00425-018-2960-4

Zheng, F., Pan, Y. Y., Li, T. Z., Chen, G. Z., Wu, X. Y., and Li, L. Q. (2019). The complete chloroplast genome of an endangered species cymbidium mastersii (Orchidaceae). *Mitochondrial DNA Part B. Resour.* 4, 3068–3069. doi: 10.1080/23802359.2019.1666047

Zhong, H., Shen, L. M., Liu, H. P., Liu, Z. J., Wu, S. S., and Zhai, J. W. (2019). The complete chloroplast genome of calanthe arcuata, an endemic terrestrial orchid in China. *Mitochondrial DNA Part B. Resour.* 4, 2629–2630. doi: 10.1080/23802359.2019.1639561

Zhou, J., Xie, T.-X., Ma, S.-H., Chen, M.-K., Zheng, Q.-D., and Ai, Y. (2019). The complete chloroplast genome sequence of goodyera foliosa (Orchidaceae. *Mitochondrial DNA Part B.* 4, 3477–3478. doi: 10.1080/23802359.2019.1674728

frontiers | Frontiers in Plant Science

TYPE Review
PUBLISHED 18 October 2022
DOI 10.3389/fpls.2022.1019347

OPEN ACCESS

EDITED BY
Nisha Singh,
Gujarat Biotechnology University, India

REVIEWED BY
Hafiz Ghulam Muhu-Din Ahmed,
Islamia University of Bahawalpur, Pakistan
GaoFei Sun,
Anyang Institute of Technology, China

*CORRESPONDENCE
Yurong Jiang
yurongjiang746@126.com
Junkang Rong
junkangrong@126.com

†These authors share first authorship

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

RECEIVED 15 August 2022
ACCEPTED 08 September 2022
PUBLISHED 18 October 2022

CITATION
Yasir M, Kanwal HH, Hussain Q, Riaz MW, Sajjad M, Rong J and Jiang Y (2022) Status and prospects of genome-wide association studies in cotton.
Front. Plant Sci. 13:1019347.
doi: 10.3389/fpls.2022.1019347

Status and prospects of genome-wide association studies in cotton

Muhammad Yasir[1†], Hafiza Hamrah Kanwal[2†], Quaid Hussain[3], Muhammad Waheed Riaz[3], Muhammad Sajjad[3], Junkang Rong[1*] and Yurong Jiang[1*]

[1]The Key Laboratory for Quality Improvement of Agricultural Products of Zhejiang Province, College of Advanced Agricultural Sciences, Zhejiang A&F University, Hangzhou, China, [2]School of Computer Science, Chongqing University of Posts and Telecommunications, Chongqing, China, [3]State Key Laboratory of Subtropical Silviculture, Zhejiang A&F University, Hangzhou, China

Over the last two decades, the use of high-density SNP arrays and DNA sequencing have allowed scientists to uncover the majority of the genotypic space for various crops, including cotton. Genome-wide association study (GWAS) links the dots between a phenotype and its underlying genetics across the genomes of populations. It was first developed and applied in the field of human disease genetics. Many areas of crop research have incorporated GWAS in plants and considerable literature has been published in the recent decade. Here we will provide a comprehensive review of GWAS studies in cotton crop, which includes case studies on biotic resistance, abiotic tolerance, fiber yield and quality traits, current status, prospects, bottlenecks of GWAS and finally, thought-provoking question. This review will serve as a catalog of GWAS in cotton and suggest new frontiers of the cotton crop to be studied with this important tool.

KEYWORDS

GWAS, cotton, SNP, linkage disequilibrium, fiber

Introduction

Cotton is one of the most important cash crops, accounting for approximately 35% of total fiber consumption worldwide. Most of the world's cotton production comes from upland cotton (*Gossypium hirsutum*), which has a wide range of adaptability and high yield. Owing to a key component of the global textile industry, cotton fiber traits have gained more attention of researchers as compared to other traits. Cotton is grown in more than 30 countries as a major commercial commodity for food, feed, and renewable fiber (Ullah et al., 2017). It has evolved through a long process of polyploidization followed by diploidization. It is an ideal crop for studying a wide range of biological

phenomena such as genome evolution, diversification, single-celled biological processes, biotic and abiotic stress tolerance (Qin and Zhu, 2011; Shan et al., 2014; Du et al., 2018).

Natural and human selection pressure generates natural variants of crops from their wild progenitors. The decoding of the cotton genome has provided valuable insight into the role of natural variations and polyploidy in the improvement of agronomic traits in the genus *Gossypium* (Zhang et al., 2008). About 5–10 million years ago, the 'A' diploid genome diverged from the eudicot progenitor along with the 'D' diploid genome (Wendel, 1989; Wendel and Albert, 1992). A transoceanic dispersal of an A-genome ancestor (*Gossypium arboretum* L.) to the New World crossed with a D-genome ancestor (*Gossypium raimondii* L.), resulting in allotetraploid cotton around 1–2 million years ago (Wendel, 1989). This strong evolutionary nexus between allotetraploid and diploid cotton genomes will assist us in better understanding the role of natural variations in association with yield, resilience to climate change, better adaptation, neofunctionalization and subfunctionalization of genes through the period of evolution of gene expression, as most of the wild, diploid and tetraploid cotton species have shared genetic functions (Wendel, 1989). Modern crop breeding necessitates a thorough understanding of the genetic basis of the origin, adaptation, and occurrence of natural genotypic and phenotypic variations to develop successful target-oriented breeding programs. The GWAS has enabled us to comprehend all of these multiple genetic patterns mediating complex traits.

Genome-wide association study (GWAS) is an experimental design to dissect the association of natural genetic variants and traits in samples representing a big population (Visscher et al., 2017). GWAS traces ancient genetic crossovers, allowing researchers to identify genetic loci underlying traits at a much higher resolution than previously possible. This technique got much attention and success in human genetics, coupled with advancements in sequencing technologies. It has become a powerful tool for identifying natural variations underlying the complex attributes of crop plants (Gupta et al., 2014). Compared to GWAS in humans, GWAS in crops have an advantage of a permanent population (natural population of diverse background or homozygous varieties) resource to be genotyped for once, that can be phenotyped for different traits (Atwell et al., 2010). A schematic flow of GWAS has been illustrated in Figure 1. Figure 1A, yellow shaded area of the map, shows major cotton-growing countries in the world; Figure 1B shows genotyping/sequencing platform, phenotypic diversity in cotton crop is sown in Figure 1C. GWAS, combined with advanced phenotyping systems has made it possible to investigate genetic diversity at the nucleotide scale precision.

Association mapping usually based on single nucleotide polymorphisms (SNPs) markers and phenotype of interest is used to find causative locus as shown in Figure 1C; DNA sequencing detects SNPs in different individuals/plants, comparing DNA sequences reveals common genome variations. Some DNA variations (haplotypes) are more common in certain traits. Haplotypes associated with a trait can pinpoint its gene(s). DNA sequences closer together on a chromosome tend to get inherited together and will often stay together over time. A haplotype is a set of close-together DNA

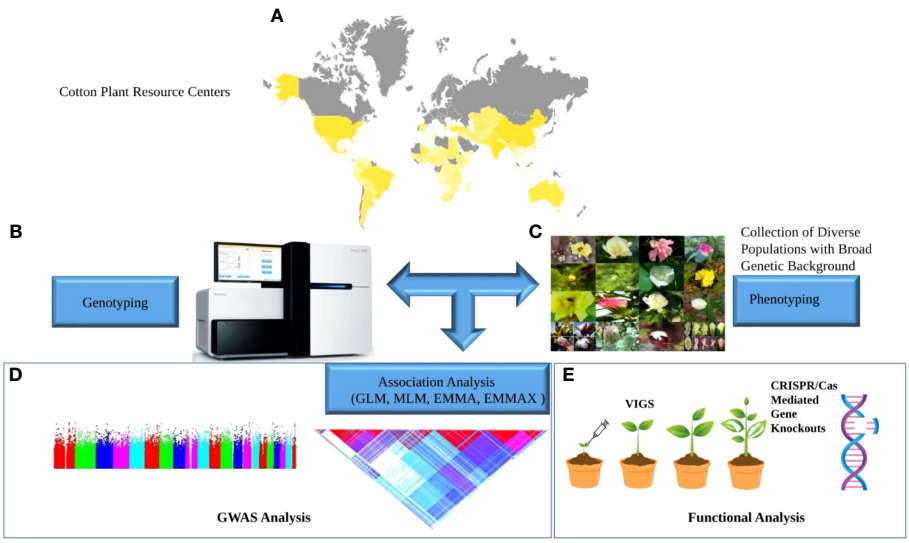


FIGURE 1
Schematic flow of GWAS. Collection of diverse germplasm; **(A)** Yellow shaded area on the world map depicts key cotton-growing countries; **(B)** Genotyping platform **(C)** Phenotypic diversity in cotton crop **(D)** GWAS analysis to find associated SNPs. **(E)** Functional validation of genes associated with traits under study.

variations (SNPs) on a chromosome that are inherited together. Because they're so close, there aren't many crossovers or recombination between these SNP groups. A haplotype can be a single gene's alleles or multiple genes' alleles. In-between-gene SNPs are possible too. Basically, it just means that these are variations in the DNA that are so close together that they tend not to recombine, and therefore tend to be passed down through the generations together. the colored triangle shows linkage disequilibrium LD block Figure 1D, small red triangles shows the presence of SNPs or group of SNPs having a strong association with other SNPs. Association analysis of the phenotype of interest and SNPs are shown in the Manhattan plot (to show association statistical significance –log10 *P*-value) Figure 1D. GWAS relies on linkage disequilibrium (LD) between markers and functional variations of causative genes. The probability of gametic-phase separation of two closely located loci by recombination is relatively less often than the loci located apart from each other. This nonrandom association or linkage of two loci is known as LD. SNPs near the causative loci can be in high association with the functional variation and thus associated with the phenotype under study. GWAS identify these associations and genomic regions containing these significant SNPs and the implicated genes. One of the interesting aspects of GWAS is the identification of pleiotropic genes (the genes controlling more than one trait at a time). There are several types of pleiotropy, however GWAS studies report pleiotropy regardless of the specific type. Identification of large effect pleiotropic regions can assist in modification the modification of multiple traits with less effort and resources. However, functional validation of associated genomic regions SNPs/genes is necessary for the fruitful utilization of GWAS results; for that purpose, virus-induced gene silencing VIGS or the most advanced and precise genome editing CRISPR/Cas technique can be used, Figure 1E.

A genome-wide association study's success and robustness lies in four sound bases: Genetic diversity, trait acquisition veracity, marker density, and statistical methodologies. GWAS has now been successfully applied in several crops, and a considerable number of studies have also been conducted on the cotton crop (Rice et al., 2020; Li et al., 2021b; Li et al., 2021c; Somegowda et al., 2021; Zhang et al., 2021; Zhong et al., 2021). In this review, we aim to provide an overview of GWAS in cotton, the success of GWAS, shortcomings, future perspectives, and finally, thought-provoking questions.

GWAS in cotton

GWAS, also known as linkage disequilibrium (LD) mapping, uses the phenotypic and genotypic variations within a species to identify the genetic underpinnings of the traits of interest. Figure 2 provides an overview of the year-wise number of GWAS publications on the cotton crop.

GWAS comes up with an opportunity to investigate the relationship between natural variations and major agronomic attributes of the cotton crop. Researchers have focused mainly on fiber, yield and abiotic stress in cotton crop, as shown in Figure 3A, however many other important aspects such as plant architecture, oil content, colored cotton, diversification, adaptation, insect pest resistance, viral disease tolerance and weed stress tolerance still need considerable attention, as shown in the lower portion of Figure 3A. Till now, 82 GWAS studies have been reported in cotton crop comprising 72 studies in *hirsutum*, 8 in *arboreum* and 2 in *barbadense*. GWAS studies for different traits in different species have been shown in Figure 3B. In the next section, we will discuss important GWAS studies conducted on different aspects of the cotton crop.

Fiber quality traits (FQTs)

Cotton is an important industrial crop because of its natural spinnable fiber used in the textile industry. Diverse interspecific variations of fiber quality traits (FQTs) particularly fiber length trait (fuzzless to extra-long fiber), depicts the genetic complexity of this trait. Low FQTs lowers the market value of cotton lint. Considerable research and breeding efforts have been expended on identifying and utilizing allelic variance that contribute to FQTs improvement. The development of cotton varieties with improved FQTs is highly desirable. More than 30 GWAS studies have been conducted on FQTs till now. Ma et al. (2018) resequenced a core collection of 419 genotypes of a natural population for 13 fiber traits in 12 diverse environments and evaluated for genomic variations with 3.96 million SNPs. 7383 SNPs depicted association with fiber traits, elucidating sufficient genomic variation present in the population under study. Single nucleotide polymorphisms (SNPs) associated with fiber traits were more prevalent than any other trait. Association of more number of SNPs with fiber traits than any other traits shows that genes involved in fiber traits improvement were under positive selection pressure. Consequently, positive selection pressure increased the proportion of fiber quality attributes. The identification of association of several important genes like *Gh_D03G0728* coding for COP1 interactive protein *CIP1* regulating the flowering time, *GhUCE* involved in fiber initiation, *GhFL2* a pleiotropic gene-regulating both fiber length and strength and some previously unreported genes associated with fiber length and other FQTs presents an excellent example of the power of GWAS. Most SNPs were found on the *D* subgenome. *COP1* in Arabidopsis is involved in mediating light-dependent growth and development of floral organs. The function of COP*1* (designated as *GhCIP1)* in *G.hirutum* was unknown, so GWAS was found useful tool to identify genes with previously unknown function. The results of Virus-induced gene silencing (VIGS) for functional validation of *Gh_D03G0728* containing different haplotype non-synonymous

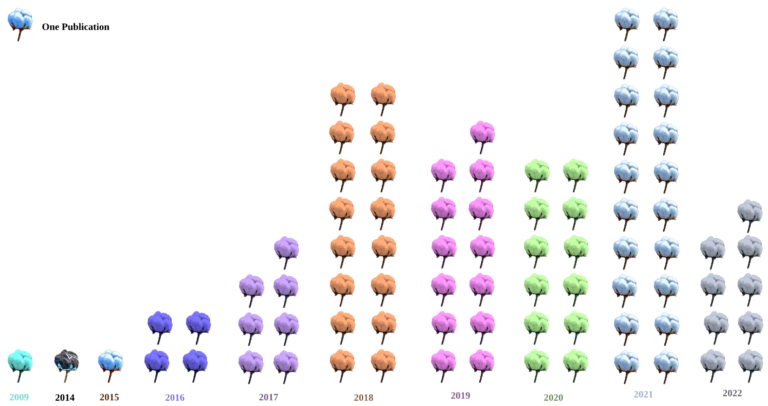


FIGURE 2
An increasing number of GWAS publications from 2009-22. Each opened boll represents one publication, specific color depicts publications of a specific year.

SNPs further corroborated the GWAS results. Silenced plants (plants with silenced gene under study) had not produced any flowering or squares; however, non-silenced (non-VIGS or CK) plants produced three fruiting branches having squares at the same growth stage, this functional analysis validated the exact effect of gene associated with flowering, suggesting it to be a key lead gene regulating flower development. A team of researchers conducted GWAS study and examined 196 cotton accessions with 41,815 SNPs. Two different GWAS models, a single-locus model [MLM (Q+K)] and five multi-locus models, were employed in six different environments for five fiber-related traits. Both the models commonly found 40 SNPs, 38 quantitative trait loci QTLs and 89 candidate genes for fiber traits. Moreover, 13 QTLs and 5 putative genes were reported with probable pleiotropic effects (Yuan et al., 2019). Single locus GWAS model and multi locus GWAS models have their own uses and limitations, in single locus model a key concern is the high false positive rate especially in large field experiments. To reduce false positive rate, Bonferroni correction is frequently applied in the single-locus methods which results in many important loci associated with the target traits being eliminated because they do not satisfy the stringent criterion of the significance test. However multi-locus models do not require Boneferroni correction thus more marker-trait associations are identified (Li et al., 2018a). Concurrent application of different GWAS models in cotton crop reveal the suitability of GWAS in cotton crop to decipher natural variants associated with complex traits.

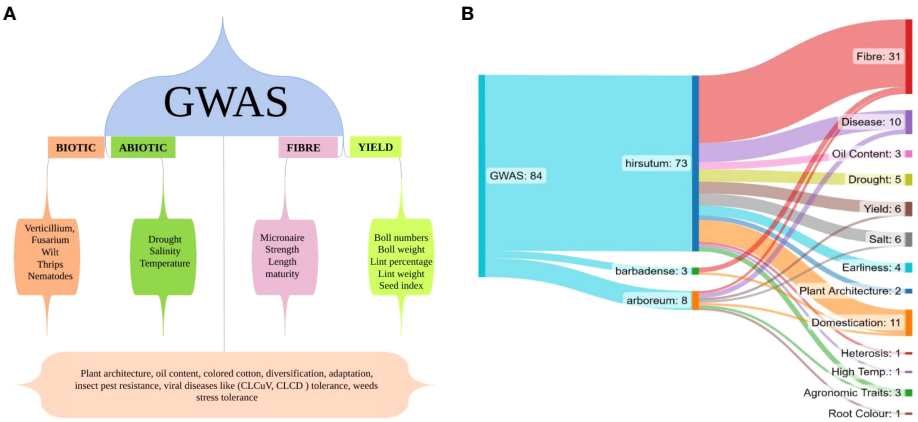


FIGURE 3
(A) This figure illustrates the GWAS conducted for different aspects of cotton, the bottom part shows less studied traits of cotton crop with GWAS tool; **(B)** This illustration shows total number of GWAS studies in all Gossypium species till now and node ends show total number of studies on a particular trait in different Gossypium species.

Most of the studies have reported an inverse relation between FQTs and the final yield of cotton. However, an increase in lint percentage is directly proportional to cotton yield. *Gh_D05G0313* lycopene cyclase family gene, located on chromosome D05 coding for phytoene synthase (a key regulator enzyme in the biosynthesis of carotenoids), was found to be associated with lint percentage (Song et al., 2019). The same gene has been well characterized and reported to be involved in enhanced provitamin A production in cottonseed oil (Yao et al., 2018). Fine mapping of genomic region containing SNP associated with lint percentage and oil content can assist in breeding cotton varieties with enhanced oil content and lint percentage. Utilization of genomic regions containing pleiotropic genes can be a cost effective method to foster breeding programs. In addition, another gene, *Gh_D05G1124*, Probable protein phosphatase 2C 21, has been found to be associated with lint percentage. Moreover, GWAS deciphered 23 SNPs, and 15 QTLs with a strong lint percentage association in a panel of 276 cotton accessions genotyped with Cotton60k SNP array. Most of the loci were located on the *Dt* subgenome than *At* subgenome (Song et al., 2019). It can be inferred that GWAS can provide complementary evidence about the pleiotropic effect of genes interaction.

Two different studies conducted by Wang et al. (2021) and Su et al. (2019) scoured FQTs and lint percentage using restriction site-associated DNA sequencing (Rad-seq) and specific locus amplified fragment sequencing (SLAF-seq) in 316 and 160 accessions of cotton, respectively. As a result of GWAS findings, 231 SNPs were associated with Fiber and yield-related traits in 27 genomic regions. Interestingly four genomic regions held favorable pleiotropic loci and 6 genes, D11 chromosome was found to be having most of the loci related to FQTs. Su et al. (2019) applied one single locus and six multi-locus GWAS models in four different environments. It was found that 4 and 45 quantitative traits nucleotides (QTNs) were associated with lint percentage in single-locus model SLM-GWAS and multi-locus model MLM-GWAS, respectively. Half of the identified QTNs were located on the *D* genome. Although several SNPs have been identified using different GWAS models, however the dire need is to validate genomic regions/genes containing highly associated SNPs as we know that GWAS only reports associated SNPs not causal SNPs, so we further need to confirm associated SNPs to declare them as the causal SNPs.

Cotton fibers are single celled highly thickened epidermal cell extensions of cotton seeds that make it a suitable model to understand single cell processes. Hence, FQTs are the most important and interesting area of research in the cotton crop, as fiber traits depend on the primary cell wall differentiation followed by pure cellulose synthesis during secondary cell wall thickening/maturation. In this section we have seen that most of the significant SNPs and candidate QTLs related to FQTs were located on the *D* subgenome. Later we will discuss the possible reasons for this feature. In this section, it has also been observed that *D* subgenome have more SNPs associated with FQTs than *A* subgenome. This finding suggest that *D* subgenome holds the promise of providing more useful knowledge to understand genetic patterns mediating FQTs.

None of the studies described above have employed large scale germplasm > 1000 accessions for the exploration of genetic diversity to connect with breeding programs. It has been proposed that GenBank genomics can serve as a link between genetic diversity and breeding in agricultural crops. Recent studies in different crops such as rice, wheat and watermelon have utilized large scale germplasm >1000 accessions to find complete set of novel variation, role of introgression in shaping adaptability and candidate genes for important traits respectively (Wang et al., 2018; He et al., 2019; Zhao et al., 2019). To bridge the gap of Genbank genomics study in cotton He et al. (2021) explored 3248 tetraploid cotton accessions genomes to decipher the genomic basis of geographic differentiation and fiber improvement. The genomes of 2500 cultivars were compared with their most probable exotic donors to find introgression events. The most introgressed fragments were found in *G.barbadense* widely distributed on all the chromosomes. Chromosome A09 was enriched with introgressed fragments in case of *G.arboreum*. Diploid specie *G. thurberi* contained highest introgressed fragments on D08 chromosome. Two large effect pleiotropic alleles *FL3* and *FS2* associated with fiber length and fiber strength overlapped with 61.8-62.2 Mb introgressed region of A09 chromosome named GaIR_A09 in *G. arboreum.* The accessions carrying GaIR_A09 showed significantly improved fiber length and strength than those lacking this fragment. GaIR A09 may be unique candidate locus introgressed from *G. arboreum* responsible for fiber quality in contemporary cultivars.

As described above, GWAS conducted by Ma et al. (2018) has provided reliable information about genes mediating flowering and fruiting branches. Gene expression was not only confirmed with RNA expression data but also functionally validated with virus induced gene silencing method. The association information in this study provides an opportunity to select potential genotypes/cultivars with exact haplotype that can be inducted in breeding program for improved fiber quality cultivars. Moreover, these findings suggest hotspots for molecular selection and genetic manipulation in cotton fiber quality improvement breeding programs. The identification of pleiotropic genes with GWAS further diversifies its application in a broader spectrum.

Cotton breeding programs depend on predefined breeding objectives, and the textile industry needs fine fiber, whereas farmers' desire is to get high-yielding varieties. However, most of the FQTs are negatively associated with fiber yield traits. Here we will see the genetic variations underpinning the yield attributes that will help design cotton breeding programs better.

Fiber yield traits

Due to its superior yield and adaptability, *Gossypium hirsutum* L. outperforms other cultivated species of cotton in lint production and meets more than 95% of total fiber demand (Chen et al., 2007). Cotton breeders have long sought to contrive high-yielding cultivars. Cotton yield improvement is hampered by a narrow genetic background and conventional breeding practices (Zhang et al., 2008). Investigation and pyramiding the elite quantitative trait loci (QTLs)/genes related to yield components for molecular breeding to boost cotton output is crucial. Molecular markers and biparental linkage mapping analysis have found many QTLs for cotton yield-related traits (Shen et al., 2007; Liu et al., 2012). However, due to the restricted number of markers and the huge QTL regions, it isn't easy to utilize QTLs with marker-assisted breeding. An unprecedented advancement in sequencing technologies and statistical techniques have made it possible to exploit genomic variations at single nucleotide level. GWAS have emerged as an important tool to utilize the genomic variations to explore the genetic underpinnings of cotton yield traits employing SNPs. In the previous decade, assembly, sequencing and fine mapping of the cotton genome have accelerated the pace of gene mapping for key traits of cotton improvement (Li et al., 2015; Liu et al., 2015; Zhang et al., 2015; Hu et al., 2019; Wang et al., 2019). GWAS has utilized reference genome to identify several SNPs and candidate genes related to cotton yield (Fang et al., 2017; Wang et al., 2017; Ma et al., 2018; Song et al., 2019).

Genotypic data comprising more than 56 thousand SNPs explored the genetic factors associated with yield- traits in a population of 242 cotton cultivars (Zhu et al., 2021). GWAS identified 560 QTNs associated with yield traits in multiple locations. In total, 95 stable QTLs (sQTLs) (spanning two or more environments) with 12,23,45 and 33 QTLs associated with boll numbers (BN), boll weight (BW), lint percentage (LP) and seed index (SI) were found respectively. Identification of several sQTLs across broad range of agrometeorological environments and multiple years are considered important for breeders to utilize in marker assisted selection. One of the proposed mechanisms of *TPR* in cotton fiber development is that it forms a complex with actins to control fiber growth. Zhu et al. (2021) studied 242 cotton accessions in 13 different locations and reported 92 high-quality QTLs associated with four fiber yield-related traits, including 12 sQTLs and an important gene, *Gh_A07G1389*, controlling short fiber development in the stable QTL19 region encoding a tetratricopeptide repeat (TPR)-like superfamily protein (Zhu et al., 2021). Understanding how fiber cell elongates using a fiber mutant has been a fascinating subject of investigation (Hinchliffe et al., 2011; Ding et al., 2014; Gilbert et al., 2014). A separate study conducted by Fang et al. (2020) found a mutation induced in tetratricopeptide repeat (TPR)-like protein gene causing the short fiber phenotype in cotton. So we can infer that the results of GWAS study were in compliance with the study based on functional validation of gene coding same protein as found in GWAS. Pentatricopeptide repeat gene *PPR* similar to *TPR* gene has been well characterized in cotton fiber development (Thyssen et al., 2016). The amino acid composition and structure of Pentatricopeptide proteins are similar to those of tetratricopeptide repeat proteins. Few studies have been reported on the role of *TPR* gene in fiber development, however GWAS have reported association of TPR gene with fiber development.

A separate study explored 719 accessions of cotton genotyped with an SNP63K array with 63,058 SNPs. 62 SNP loci located mostly on *D* subgenome were associated with different fiber traits. Furthermore, two pleiotropic genes *(Gh_D03G1064* and *Gh_D12G2354*) that increased lint yield were also identified (Sun et al., 2018b). The protein coding gene *Gh D03G1064* encodes the FRIGIDA protein; a developmental protein in terms of molecular function, whereas differentiation and flowering protein in terms of biological process, involved in the regulation of flowering time in the late-flowering phenotype. It contributes to the enrichment of a WDR5A-containing COMPASS-like complex at the 'FLOWERING LOCUS C' that trimethylates histone H3 'Lys-4,' resulting in FLC up-regulation and increased RNA levels (Jiang et al., 2009). A GWAS has successfully elucidated the role of two ethylene-pathway-related genes linked to higher lint yield and have a pleiotropic relationship with two fiber characteristics, LP and NB at A02:79153947 and D08:3040023 association signals, respectively (*GhLYI-A02* and *GhLYI-D08*). One non-synonymous SNP significantly associated with a trait of interest located in the gene *Gh_A02G1392*, a homolog of the *AP2/ETHYLENE RESPONSE FACTOR (ERF)*-type transcription-factor-encoding gene *AINTEGUMENTA-like 6 (AIL6)* in Arabidopsis. Comparative RNA-seq and (qRT-PCR) data showed high expression of this gene during seed development from 20 to 25 days post-anthesis, more than tenfold higher in TM-1 (lower lint percentage and number of bolls per plant) than in the corresponding Chinese improved cultivar ZMS12, which has a higher lint percentage and increased number of bolls per plant, in developing ovules at 10 and 20 DPA. The orthologous gene *AIL6* has a major role in shoot and flower meristem maintenance, flower initiation, organ size and floral organ identity, and cell proliferation in Arabidopsis. It has been reported that *ethylene insensitive 3 like protein* (*EIL*) serves as a key downstream regulator, *EIN3*, in the ethylene-initiated signaling network and binds to the promoters of *ERF* genes such as *AIL6*, stimulating their transcription in an ethylene-dependent manner. When a plant responds to both biotic and abiotic challenges, ethylene is a key phytohormone for orchestrating the integration of environmental signals into a specific phenotype (Dubois et al., 2018). However, ethylene also participates in cotton fiber development (Shi et al., 2006; Li et al., 2007; Qin et al., 2007). Higher accumulation of 3 transcripts of

ethylene biosynthesis gene at fiber elongation stage, enhanced fiber elongation by exogenous application of ethylene, elimination of fiber elongation inhibition caused by 2-chloro-N-[ethoxymethyl]-N-[2-ethyl-6-methyl-phenyl]-acetamide (Qin et al., 2007), with ethylene elaborates the key role of ethylene in cotton fiber development. These findings imply that the equivalent multiple-functionality alleles linked to the ethylene pathway were rigorously selected during breeding of higher lint percentage (Fang et al., 2017).

A recent GWAS of 103 cotton accessions (breeding lines, released and obsolete cultivars) across three seasons identified the associations of SNPs to a wide range of cotton yield components, including fiber quality, plant architecture, and stomatal conductance. Out of 63,058 markers, only 28,480 SNP markers > 5% MAF were used for GWAS. For fiber length and micronaire, MLM revealed 17 and 50 significant SNP associations, respectively (Gapare et al., 2017). The possible reasons for less number of associated SNPs suitable for GWAS analysis in this study could be less DNA polymorphism or low resolution provided by Chip-seq in cotton population under study, because the chip seq platform is limited to detect only those SNPs which are exclusively present on the chip. Few SNPs associated with multiple traits revealed that GWAS was actually unable to find rare variants, genetic interactions and genetic variance heterogeneity. GWAS detects common variants in a germplasm collection but has limited power to detect rare variants. Plant breeders are sometimes interested in finding rare variants. A possible solution for such germplasm populations with less genetic diversity is to use whole-genome selection (WGS) prediction instead of GWAS. WGS forecasts the performance of new quantitative trait candidates by combining hundreds or thousands of SNP markers with prior phenotypic data.

Another GWAS of 231 recombinant inbred lines RILs genotyped with 122 SSR and 4,729 SNP markers in nine locations was conducted for fiber and yield quality characteristics. The cottonSNP80K array platform, having 77,774 SNPs, was used for genotyping. There were 134 QTLs for fiber traits and 122 QTLs for yield parameters; 57 of these QTLs were sQTLs. Four multi-locus GWAS models identified a total of 209 and 139 (QTNs) linked with fiber yield components, respectively. A total of 51, 50, and 38 QTNs were found to be linked with BW, LP, and SI, respectively, with 82, 83, and 65 candidate genes. KEGG analysis revealed two yield-related candidate genes involved in six pathways. The putative genes found in 57 sQTLs were matched to those found in QTN, 35 common candidate genes were reported with four possibly pleiotropic genes (Liu et al., 2018). Yield can either be increased by improving complementary factors involved in lint production or by decreasing the effect of limiting factors like biotic and abiotic stresses. GWAS application in multiple seasons and locations have provided several clues about candidate SNPs and genes involved in yield improvement.

Identification of *TPR* gene associated with fiber yield traits using GWAS and functional validation of *TPR* transgene of Arabidopsis in *G. hirsutum* in a different study confirms the results of GWAS. So, this study supported by previous studies of molecular characterization of *TPR* gene, suggests the application of GWAS for the prioritization of candidate genes to develop transgenic plants with increased fiber yield. Previous studies have generally described the role of ethylene in abiotic stress tolerance; however Fang et al. (2017) deciphered the role of ethylene phytohormone in fiber elongation using the GWAS tool. This study promises to utilize a single gene involved in a cascade of biological pathways to develop abiotic stress tolerant genotypes with long fiber characteristics. The need of the hour is to do functional validation of the potential candidate genes identified with GWAS and field evaluation of the germplasm containing causal/associated SNPs/genes.

Biotic and abiotic stresses

Plants have evolved intricate molecular pathways to cope with biotic and abiotic stresses. The dynamic climatic conditions have led to exceptional weather patterns resulting in augmented crop losses. Similarly, disease dissemination of pathogens in an era of increased international commerce has resulted in pathogen introduction and adaption to new places, resulting in recurrent outbreaks. A deep apprehension of the molecular processes underpinning plant stress responses to generate climate-smart crop varieties is becoming increasingly important (Prasad et al., 2021). Environmental stresses can be categorized into biotic and abiotic stresses. GWAS has provided a tremendous opportunity to explore the complex genetic patterns and genes involved in stress tolerance biological pathways. Here we will elucidate some studies that describe the application of GWAS to decipher the molecular mechanism of biotic and abiotic stress resilience in cotton crops.

Verticillium wilt is a plant disease caused by a soil-borne fungus, *Verticillium dahliae*. It is a severe disease in cotton (Sink and Grey, 1999; Klosterman et al., 2009). *V. dahliae* outbreaks in cotton have become more common in China, the disease is prevalent on nearly all cotton acreage, resulting in massive economic losses every year (Mo et al., 2016). To identify potential loci implicated in wilt resistance, GWAS in 215 Chinese *Gossypium arboreum* accessions inoculated at seedling stage with *V. dahliae* have traced 309 loci significantly associated with *Verticillium wilt* resistance with the strongest signal Ca3 in a 74 kb haplotype block. *CG05*, designated as *GaGSTF9*, was located close to the most significant SNP Ca3 23037225 (14 kb); this gene has been more responsive when treated with *verticillium dahilae* and salicylic acid (SA). Therefore, it was inferred that *CG05* might respond to invasion by *V. dahliae via* an SA-related signaling pathway. *GaGSTF9* was found to be a

positive regulator of *Verticillium wilt* through the use of (VIGS) and overexpression in Arabidopsis. Genes identified in GWAS studies followed by functional validation provides robust and valid results in terms of molecular markers that can be utilized in the development of genetically modified plants with improved characteristics.

Moreover, the Arabidopsis mutant *gstf9* of (GST) was more susceptible to *Verticillium wilt* than wild-type plants. The endogenous SA and H_2O_2 expressed a significant effect on Arabidopsis that overexpressed *GaGSTF9*, depicting that GST may regulate reactive oxygen species (ROS) content *via* catalytic reduction of the (GSH), subsequently affecting SA content. Plant *glutathione S-transferase* (*GSTs*) genes have been classified into several classes, including dehydroascorbate reductase (DHAR), metaxin, phi, tau, zeta, theta, iota, lambda, and others. In cotton, few types of glutathione S-transferase GSTs have been identified. These findings elucidate the success of GWAS in identifying *GaGSTF9* gene, a key regulator mediating cotton responses to *V. dahliae* and a potential candidate gene for cotton genetic improvement (Gong et al., 2018).

Verticillium wilt foliar disease severity ratings (DSR) indexed, multi-parent advanced generation inter-cross (MAGIC) population of 550 recombinant inbred lines (RILs) coupled with 11 upland cotton parents with a total of 473,516 polymorphic SNP markers used to identify chromosomal regions for VW resistance, GWAS identified three QTLs on chromosome A01, D02 and D08 in common and three QTL clusters detected on chromosome A01, A13, and D01. Potential candidate genes for VW resistances were found in a narrow region of three common resistance QTLs (Zhang et al., 2020). Identification of common QTLs on different chromosomes link certain complex phenotypes to specific regions of chromosomes. Common QTL regions associated with particular trait of interest can be skewed towards fewer QTLs with major effects (Hayes and Goddard, 2001).

A combined study of GWAS, QTL-seq and transcriptomic analysis showed the role of the flavonoid biosynthesis pathway in VW resistance in cotton crop. Gene *Ghir_A01G022140* encodes a pyridoxal 5′-phosphate synthase subunit PDX1 involved in vitamin B6 biosynthesis, which is important for disease resistance in tomato and *Arabidopsis thaliana* (Zhao et al., 2021b). A comparative study of Combined results of GWAS, QTL-seq and transcriptome sequencing detected basal defense-related genes showing gDNA sequence and expression variation in VW tolerant and sensitive cotton lines, which might be the molecular mechanisms of VW resistance in *G. hirsutum*. Vitamin B6 is an assembly of six vitamers: pyridoxine (PN), pyridoxal (PL), pyridoxamine (PM), and their phosphorylated derivatives. The *de novo* biosynthesis of VB6 vitamers requires two highly conserved pyridoxal protein families (PDX1 and PDX2) in plants. Vitamin B6 has a crucial role in stress responses as it is involved in pathways of starch metabolism, glucosinolate biosynthesis, ethylene and auxin biosynthesis (Mooney and Hellmann, 2010). In previous studies, plants with complete PDX knockouts have depicted increased sensitivity with B6 reduction to high levels of salt, sucrose, mannose, polyethylene glycol (PEG), UV and different diseases (Titiz et al., 2006). In contrast, transgenic plants with overexpression of PDX showed higher tolerance to different biotic and abiotic stresses, which corroborates the role of B6 in plant's defense systems (Raschke et al., 2011). These findings depict that multilayered (GWAS, QTL, transcriptomic) identification of genes for traits under study provides more robust results.

G.hir_A01G022110, a negative-regulated gene coding for 2-oxoglutarate-dependent dioxygenase involved in the flavonoid biosynthesis pathway, has been linked to VW resistance by the enrichment of flavonoids in a spontaneous cotton mutant with red coloration, which results in a significant increase in resistance to *Verticillium dahliae*, these findings, supported by metabolomics studies, revealed the success of GWAS in figuring out the multifaceted genetic pattern of disease resistance mechanism in cotton crop.

Cotton is an evergreen perennial crop attracting numerous insect pests. Insect pests cause great damage by attacking leaves, bolls and consequently causing reduction in yield. Only a few GWAS studies have been done to find genetic association for insect resistance mechanisms in cotton crop. Evaluation of an association mapping panel of 376 Upland cotton accessions identified quantitative trait loci (QTL) for Thrips resistance in two replicated tests. Based on a GWAS using 26,301 polymorphic SNPs Eight QTL were identified for Thrips resistance on five chromosomes, with most SNPs on the *D* subgenome (A09, D01, D02, D03 and D11) (Abdelraheem et al., 2021a). Exclusive presence of several QTLs on D genome is an indication of persistence of genomic regions on D genome over the course of selection and evolution of cotton crop. These genomic regions can help us to trace out signatures of selection for Thrips resistance in cotton crop.

Bacterial blight (BB), caused by *Xanthomonas citri* pv. *malvacearum* (Xcm) is a destructive disease of cotton crop in many countries. Breeding BB-resistant cotton cultivars is the most effective strategy for controlling this disease. A current study of GWAS in 335 cotton accessions has been reported. GWAS, based on 26,301 polymorphic (SNP) markers, detected 11 QTLs associated with 79 SNPs, including three QTLs located on each of the three chromosomes, A01, A05 and D02, and one QTL on each of D08 and D10. Once again, these results found more associated SNPs on the *D* subgenome. Several studies concurrent findings of genomic regions associated with multiple traits on *D* genome offer some clues to focus on genomic hotspots associated with several traits. These results will assist in breeding cotton for BB resistance and facilitate further genomic studies in fine-mapping resistance genes to enhance the understanding of the genetic basis of BB resistance in cotton (Elassbli et al., 2021).

Previous studies of abiotic stress have not dealt simultaneously with the genetics and genomics of drought and salt tolerance (DT, ST). A MAGIC population of 550 RILs and their 11 Upland cotton parents used 473,516 SNPs to pinpoint QTLs for (DT) and (ST) at the seedling stage. Transgressive segregation in the MAGIC-RILs indicated tolerant and sensitive alleles recombination during the crossing-over process for the population development. A total of 20 QTL were detected for DT, including 13 and 7 QTL related to plant height (PH) and dry shoot weight (DSW), respectively, whereas 23 QTL were detected for ST, including 12 and 11 QTL for PH and DSW, respectively. Four QTLs were reported on chromosome A13, three QTLs on A01 for DT, four QTLs on D08 and three QTLs on A11 for ST. Nine QTL were shared by DT and ST, showing possible implication that the two stresses have a common genetic base. Salt and drought stress-responsive 53 candidate genes were identified. However, these findings need further validation with functional analysis of putative candidate genes. The QTL discovered for both DT and ST have significantly increased the number of QTL for abiotic stress tolerance that can be used for marker-assisted selection (MAS) to develop DT, ST cultivars and further genomic studies to identify drought and salt responsive genes in cotton (Abdelraheem et al., 2021b). Common SNPs identified with GWAS for drought and salt stress tolerance provide time and cost effective opportunity to focus on common QTLs for breeding drought and salt-tolerant cultivars simultaneously.

319 upland cotton accessions genotyped by more than 55 thousand (SNPs) for nine traits related to drought tolerance, using two datasets to identify QTNs with multi-locus random-SNP-effect MLM (mrMLM) reported a total of 20 QTNs distributed on 16 chromosomes. A compendium of 205 genes was induced after drought stress, combined results of (GWAS), RNA-seq and qRT-PCR verification, proposed four genes to be potential candidate genes for drought tolerance, *RD2* encoding response to desiccation 2 protein, *HAT22* encoding a homeobox-leucine zipper protein, *PIP2* encoding a plasma membrane intrinsic protein 2, and *PP2C* encoding a protein phosphatase 2C. *PP2C* is an important regulator of stress signal transduction pathway and is an excellent candidate gene to decipher complex stress tolerance traits (Schweighofer et al., 2004; Hussain et al., 2021). Silencing of *GhPIP2;7* gene in *Gossypium hirsutum* has shown decreased chlorophyll content, superoxide dismutase (SOD) and peroxidase activity, it depicts positive regulatory role of *PIP2;7* in cotton under salt stress (Guo et al., 2022). Zheng et al. (2021a) reported cloning of *PIP2* gene from *Canavalia rosea* and overexpression in *A.thaliana* increased the recovery and survival rate under elevated drought and salt stresses. The most surprising aspect of the study was *PIP2* mode of action in drought and salt stress tolerance by mediating water homeostasis instead of reactive oxygen species (ROS) transporter. This shows that GWAS findings are in compliance with functional studies of the same genes in model plants and cotton plants. These results will contribute to a better knowledge of the genetic basis of drought stress tolerance in cotton and prospective markers for breeding drought-tolerant cotton cultivars (Hou et al., 2018).

Previous studies have not utilized advanced image-based automatic phenotyping platforms to bridge the gap between phenotypes and genotypes with the GWAS tool. A team of researchers has recently utilized an auto-phenotyping platform to examine 119 digital traits to decipher genes regulating drought stress tolerance in a population of 200 upland cotton accessions at the seedling stage. The phenomics data identified 390 genetic loci by GWAS using 56 morphological and 63 texture i-traits. Some digital traits identified drought-responsive genes, including *GhRD2, GhNAC4, GhHAT22* and *GhDREB2*. Moreover, potential candidate genes *Gh_A04G0377* and *Gh_ A04G0378* were negative regulators of cotton drought response. Comparative analysis of phenomics, GWAS and transcriptomic data provides an exceptional resource to characterize key genetic loci with an unprecedented resolution that can predict future genome-based breeding for improved drought tolerance in cotton (Li et al., 2020). It is of great interest that two different GWAS studies on same traits found *RD2* and *HAT22* as common genes involved in drought stress tolerance as mentioned above. Together these results provide important insights into drought stress tolerance mechanism. Advanced phenotyping platforms utilizing artificial intelligence-based image analysis are potential future resources for precise and accurate phenotyping. Although phenotyping is as important as genotyping yet, most of the time, researchers give more importance to genotyping accuracy. This study is a classic example of utilizing cutting-edge technologies for both genotyping and phenotyping coupled with GWAS.

A population of 316 upland cotton accessions studied with GWAS using GLM and a factored spectrally transformed LMM. A total of eight, three, and six SNPs were linked to the euphylla withering score (EWS), cotyledon wilting score (CWS), and leaf temperature (LT), respectively. Interestingly, combined results of GWAS and RNA seq depicted that DEGs *WRKY70, GhCIPK6, SnRK2.6*, and *NET1A* induced by drought stress were found in the candidate region, regulating ABA, the mitogen-activated protein kinase (MAPK) signaling and the calcium transduction pathways (Li et al., 2019). RNA seq analysis further corroborated these findings by elucidating differential expression of these genes under normal and drought conditions; moreover, the expression of these genes was found to be induced by the drought stress.

Soil salinity is a serious threat affecting more than 800 million hectares of arable land worldwide. Plant growth and development is significantly hindered under salt stress. GWAS has been successfully utilized to identify SNPs/genes associated with salt stress tolerance mechanisms. A GWAS in 217 cotton accessions for salt tolerance related (ST) traits at the seedling

stage performed on 2 years of phenotypic data and more than 50 thousand SNPs identified 27 SNPs scattered over 12 chromosomes associated with three ST traits. Among these, the associations on chromosomes A13 and D08 for relative PH, A07 for relative shoot fresh matter weight (RSFW), A08 and A13 for relative shoot dry matter weight (RSDW) were stable SNPs, indicating that they were likely to be sQTLs. A total of 12 salt-induced candidate genes were identified by GWAS and transcriptome analysis (Xu et al., 2021).

A genome-wide association conducted in a core collection of 419 accessions of cotton representing a vast genetic background, subjected to GWAS for various agronomic traits. Phenotypic variations depicted significant variation between salt-sensitive and tolerant accessions. 17264 significantly associated SNPs were found distributed on multiple chromosomes. Twenty potential candidate genes discovered around SNPs A10_95330133 and D10_61258588, linked to relative water content under 150mM NaCl salt stress (RWC_150) and leaf fresh weight (FW_150). Fine mapping of these important genomic region can unveil candidate genes for functional validation. Differential expression of candidate genes under normal and drought stress conditions revealed the involvement of the genes in the salinity tolerance mechanism. Further study on the functional validation of candidate genes development of transgenic lines can provide useful knowledge about the genetic control of salt tolerance at the seedling stage (Yasir et al., 2019).

Zhu et al. (2020) examined three major components of lint yield across 316 *G. hirsutum* accessions over two years under four salt conditions. GWAS analysis reported 57,413 SNPs above *P*-value threshold. A total of 42, 91 and 25 sQTLs were associated with single boll weight SBW, lint percentage LP, and number of bolls per plant NBPP, respectively. Eight sQTLs discovered concurrently in four different salt environments in the case of LP, whereas SBW and BNPP had fewer sQTLs. According to gene ontology (GO) analysis, their regulatory mechanisms were also quite different. The transcriptomic analysis defined 8 genes associated with LP under salt stress; Haplotype analysis showed an MYB gene *GhMYB103* with two SNP variants in cis-regulatory and coding areas substantially linked with LP. Moreover, 40 candidate genes from NBPP QTLs were salt stress-responsive (Zhu et al., 2020).

Here we describe another GWAS study in which a high-throughput CottonSNP80K array was used to identify genotyping in various cotton accessions. 77,774 SNP loci were synthesized on the array. In 288 *G. hirsutum* accessions for GWAS, more than 54 thousand SNPs (MAFs >0.05) related to 10 salt stress attributes were found, with eight significant SNPs connected with three salt stress variables. Two loci on D5 were significantly associated with chlorophyll content, one on A2 and four on D9 were significantly associated with melondialdehyde content, and one on A12 significantly associated with germination rate (Cai et al., 2017). Similarly, in another study, two ST characteristics were assessed in a population of 713 upland cotton seedlings. Infinium CottonSNP63K array discovered marker-trait relationships under salt stress. Across seven chromosomes, A01, A10, D02, D08, D09, D10, and D11, 23 SNPs were linked with a relative survival rate of seedlings and salt tolerance level. The *D* subgenome had most of the candidate genes. The D09 SNPs i46598Gh and i47388Gh were linked to both traits. 280 potential genes showed differential expression under salt stress. Many of these genes have been implicated in salt tolerance in plants *via* transcription factors, transporters, and enzymes. The differential expression of six candidate genes in salt-tolerant and sensitive cotton cultivars confirmed their role in salt tolerance (Sun et al., 2018a). This study elucidated that most of the SNPs were skewed on the *D* subgenome. These findings are important for improving our understanding of the complex salt tolerance mechanisms in *G. hirsutum*.

Many GWAS studies have been reported for abiotic stress tolerance in cotton crop at seedling stage, however number of studies under field conditions are limited. Dynamic gene expression is highly influenced by environmental variables. It is imperative to study GWAS results at field level to believe on the stable expression of genes under changing environmental conditions. However, phenotyping of crops under certain stress is challenging under field conditions because of inhomogeneity of certain stress, difficult differentiation of plant response against concurrent effect of different stresses. We have learned from the above descriptions that GWAS has been successfully conducted on different biotic and abiotic stresses. These findings provide valuable information about the genetic underpinnings of the stress tolerance mechanism and prove the robustness of the GWAS approach. However, in some studies, we have seen that GWAS could not find the unprecedented number of associated SNPs with traits under study. The possible reasons for this inability could be the less natural genetic variation enrichment of the population or the low resolution of the sequencing platform, e.g. Chip_seq. If the population under study doesn't have enough natural variation common in different lines/landraces/RILs or varieties but has rare causal genetic variants, then GWAS shouldn't be preferred over WGS as WGS have more power to find out natural genetic variants than GWAS. This knowledge suggest that we should carefully select populations to perform GWAS analysis. One of the suggestions to overcome this shortcoming is to use core collections of populations to make a representative population from a big germplasm. The genes identified in GWAS studies are very much important in terms of their annotation and pathways involved in stress tolerance mechanism. Genetic characterization, transformation and functional validation of these genes in model plants and confirmation in cotton plants can provide promising genes that can be utilized for the breeding programs of stress-tolerant cultivars.

Domestication and traits improvement

Gossypium hirsutum L. adapted during polyploidization to generate a larger fiber production and endure harsh environments better than *Gossypium barbadense* L., which yields superior-quality fibers. The genetic and molecular underpinnings of these interspecies divergences were unknown globally. However, GWAS has opened new avenues to trace genetic footprints that lead to speciation, interspecific divergence, spatio temporal adaptation and the development of modern cotton cultivars.

Here we discuss the findings of Hu et al. (2019) about the diversification of *G. hirsutum* and *G. barbadense* species. Whole-genome comparative analyses revealed that driving forces of speciation and the evolutionary history of these species were species-specific alterations in structural variations, gene expression, and gene family expansion. These findings aid in understanding cotton genome evolution and domestication history. The genetics behind local adaptation and domestication can be found in interspecific genomic diversity. A recent study addressed the role of interspecific haplotypes and introgression in improving the agronomic traits of the cotton crop. Two allotetraploid Gossypium species (*Gb*) and (*Gh*) were cultivated independently. A combined result of three GWAS panels (one panel of 229 *Gb* accessions and two panels of 491 *Gh* panels) revealed that most functional haplotypes related to agronomic traits were highly divergent, depicting strong divergent improvement between *Gb* and *Gh* accessions. According to the interspecific haplotype map, six interspecific introgressions from *Gh* to *Gb* were strongly related to *Gb* phenotypic performance, accounting for 5%–40% of phenotypic diversity in yield and fiber quality. In addition, three introgressions in *Gb* overlapped with six linked loci, indicating that these introgression sites were under selection and stabilization during the course of improvement. A single interspecific introgression might increase production while lowering fiber quality, or vice versa, making it challenging to raise yield and fiber quality simultaneously (Fang et al., 2021). A similar study found 315 introgression events from *G.hirsutum* to *G.barbadense* causing population divergence and agronomic trait variations. Moreover pleiotropic gene controlling traits was found (Wang et al., 2022).

Cotton's development and domestication are fascinating from economic and evolutionary perspectives. An intraspecific QTL mapping population of 466 F_2 individuals from an intraspecific cross between the wild *Gossypium hirsutum* var. *yucatanense* (TX2094) and the elite cultivar *G. hirsutum* cv. Acala Maxxa, in two environments targeting domesticated cotton phenotypes, found only 22 stable QTLs (sQTLs) associated with phenotypic changes during domestication. Even though around half of the QTL were found in the *A*-subgenome, numerous critical fiber QTL were found in the *D*-subgenome, inherited from a species with unspinnable fiber. Many QTLs were environment-specific, with few shared across the two environments, demonstrating that QTLs related to *G. hirsutum* domestication were genomically clustered yet environmentally mutable. It was concluded that the evolutionary dynamics shaping sympatric speciation divergence and domestication in cotton are complicated and that phenotypic alteration was likely influenced by several interacting and environmentally sensitive variables (Grover et al., 2020; Li et al., 2021a).

Natural and artificial selection pressure in crop plants leads to modifications in genotypes and phenotypes for better adaptation to changing conditions. A chromosomal inversion phenomenon is supposed to be an efficient source of accumulation of favorable alleles to fine-tune population genetic architecture. Many studies have reported this phenomenon in population adaptation in dynamic environmental conditions (Jones et al., 2012; Lamichhaney et al., 2016 ;Sinclair-Waters et al., 2018). Recently a GWAS study to identify adaptive loci involved in cotton crop subgroup differentiation concluded similar results. Loci associated with environment adaptation had locus on divergent chromosomal regions of A06 and A08. Collinearity analysis of several assembled genomes of cotton proved the evidence of chromosome inversion. Inverted sequences on chromosomes suppressed homologous recombination, allowing desirable alleles to persist in the subsequent populations. This study revealed the cause of population divergence and the consequences of variation in its environmental adaptability. These findings shed light on the genetic basis of environmental adaptability in Upland cotton, potentially speeding up the designing of molecular markers for climate change adaptation in future cotton breeding programs (Dai et al., 2020).

A variation map for 352 wild and domesticated cotton accessions scanned 93 domestication sweeps encompassing 74 Mb and 104 Mb of the *A D* subgenomes, respectively; moreover, GWAS found 19 potential loci for fiber-quality characteristics. Asymmetric subgenome domestication was reported as the responsible agent for long fiber directional selection. Global investigations of DNase I–hypersensitive sites and 3D genome architecture, which relate functional variations to gene transcription, showed the consequences of domestication on cis-regulatory divergence. This study provides new insights into the evolution of gene organization, regulation and adaptation in cotton crop and should serve as a rich resource for genome-based cotton improvement (Wang et al., 2017). However, directional selection results in the loss of desirable variations for important traits of interest.

An important study conducted by Nazir et al. (2021) re-evaluated a landrace of *Gossypium hirsutum*, formerly known as

Gossypium purpurascens to understand the genomic structure, variation, and breeding potential, providing potential insights into the biogeographic history and genomic variations likely associated with domestication. Paucity of large number of cultivars/accessions used in all the GWAS studies on domestication and adaptability makes the results less representatives of the entire story. However, He et al. (2021) resequenced and mapped the data of 3278 cotton accessions to the reference genome. Based on phylogenetic and principal component analysis classification of more than 3K accessions into eight subgroups shows the enrichment of variation lead by natural and artificial selection. Fixation statistics (F_{st}) is the proportion of the total genetic variance contained in a subpopulation (the S subscript) relative to the total genetic variance (the T subscript). Low F_{ST} values within improved subgroups (G3–G6) demonstrated that the genetic divergence was low (0.019–0.067) within improved cultivars, However, the average pairwise F_{ST} values were much higher (0.425– 0.552) between improved cultivars and landraces when using G1 (landraces group) as comparison pair than other. Higher *Fst* values between improved cultivars and landraces depicts noticeable genetic differentiation. Maximum divergence and major haplotypes were traced on chromosome A06 and A08 of *G.hirsutum*. *De novo* assembled genome carrying haplotypes of interests showed large scale chromosomal inversions causing haplotype polymorphism. Divergence on chromosomes A06 and A08 is the most notable genomic signature in cultivated G. *hirsutum*. Population differentiation in animals and plants has been linked to large chromosomal inversions. This evolutionary mechanism allows species to adapt to new environments by repressing recombination to sustain favorable genotypes. These findings support the theory of chromosomal inversion-population differentiation in crops and define the haplotypes associated with geographic differentiation in cotton cultivars. Here we enlist GWAS studies conducted on cotton crop (Table 1) for different traits, identified SNPs and QTLs.

Pervasive pleiotropy

Several GWAS studies described above have found some common SNPs/genes associated with more than one trait called pleiotropic SNPs/genes. A single gene controlling more than one apparently unrelated distinct traits/phenotypes is called pleiotropic gene. Several studies have revealed pleiotropic genes regulating multiple variables in cotton crop, so we will outline the likely causes and advantages of pleiotropic SNPs/genes. One possible explanation for pleiotropy is that the product of a single gene can be utilized in different cell types or can participate in cascade-like signaling to different targets. It's difficult to discern actual biological pleiotropy from mediated and pseudo pleiotropy because genes normally work in intricate pathways/networks causing associated phenotypes. Mutual pathway sharing of two associated phenotypes leads to mediated pleiotropy, while spurious pleiotropy occurs when identified SNPs lies in a small region of high LD, where two tightly packed distinct genes lie, that regulate two different variables/phenotypes. However, no GWAS studies have reported such a deep level of understanding of pleiotropic genes. Pleiotropic genes can be considered hot spots for precise genome editing with the most advanced techniques like CRISPR/Cas system to explore the potential effects on associated phenotypes. Cotton genome editing at loci having pleiotropic genes can provide an opportunity to fine-tune genomic regions controlling more than one trait with little effort. Collective regulation of multiple metabolic processes can be modified in single gene editing event instead of editing several genomic regions. Moreover, editing of pleiotropic genes can produce large effect novel phenotypes. We have discussed several GWAS studies reporting pleiotropic genes based on their action on seemingly unrelated phenotypes. Still, there is not a single in-depth study stating the validation of pleiotropic genes identified with the GWAS tool. As the number of GWAS reports of pleiotropic genes increases over time, cotton breeders will pay greater attention to it. There are no reports on the functional characterization of pleiotropic genes or SNPs found with the help of GWAS. It's possible that with the development of new statistical models and genomic techniques for GWAS investigations, we will be able to learn more about pleiotropic loci.

Genomic basis of fiber development and occurrence of more SNPs on *D* subgenome

It has been anticipated that the spinnable fibers were formed only once in the progenitor of the *A* genome. Although this long morphology evolved in the *A*-genome progenitor, it has been found that the *Dt* genome was also important in tetraploid cotton fiber formation. Even though the *D* genome ancestor does not produce spinnable fiber, nonreciprocal DNA exchanges from *At* to *Dt* have resulted in the recruitment of *Dt* subgenome genes into the regulatory processes of tetraploid cotton fiber development. *Dt* has 1.4 times (104 Mb) as many sequences containing domestication signals as *At*, demonstrating a case of asymmetric domestication for the two coexisting subgenomes. The *Dt* subgenome also has 2.2 times more regions with selection sweep signatures than the *At* subgenome, indicating that it has been subjected to more selection pressure than the *At*

TABLE 1 A summary of GWAS aimed at identifying SNPs/QTLs/Genes or quantitative trait nucleotides QTN, that contribute to cotton improvement.

Cotton species	Population Size (accessions)	Years/location/ environment	Traits	No. of unique SNPS QTLs/sQTLs	Chromosomal location	Reference
G.hirsutum L	419	6 locations 2 years	13 fiber traits	7,383	A10, A07, A08, and D11	(Ma et al., 2018)
G.hirsutum L	196	6 environments	5 fiber traits	23 QTLs	A2, A6, A7, A9, A10, A13, D1, D5, D6, D7, D8, D10, D11, and D12	(Yuan et al., 2019)
G.hirsutum L	276	6 locations 2 years	1 fiber trait	23 SNPs and 15 QTLs	D05	(Song et al., 2019)
G.hirsutum L	316	9 environments	8 fiber traits	231 loci	A06, A07 and D11	(Wang et al., 2021)
G.hirsutum L	242	13 environments	4 yield traits	95 QTLs	A08	(Zhu et al., 2021)
G.hirsutum L	719	8 environments	5 fiber quality traits	62 QTLs	Dt11 and At07	(Sun et al., 2017)
G.hirsutum L	231 RILs	9 environments	Fiber and yield traits	32 and 25 sQTLS		(Liu et al., 2018)
G.arboreum L	215	Green house	*Verticilium dahiliae*	309 SNPs	A03	(Gong et al., 2018)
G.hirsutum L	550	Green house	*Verticillium dahliae*	3 sQTLs, 3 QTLs	A01, D02, D08, A13, D01	(Zhang et al., 2020)
G.hirsutum L	376	Green house	Thrips	8 QTLs	A09, D01, D02, D03 and D11	(Abdelraheem et al., 2021a)
G.hirsutum L	335	Green house	Bacterial blight	7 QTLs and 4 sQTLs	A01, A05, D02, D08 and D10	(Elassbli et al., 2021)
G.hirsutum L	550RILs	Green house	Drought, salt	20, 23 QTLs	A13, A01, D08	(Abdelraheem et al., 2021b)
G.hirsutum L	319	Green house	Drought	20 QTNs		(Hou et al., 2018)
G.hirsutum L	200	Green house	Drought	390 QTLs, 71 sQTLs	A04, D11, D13	(Li et al., 2020)
G.hirsutum L	316	Green house	Drought	7 QTLs	A01, A05, A11, D03	(Li et al., 2019)
G.arboreum L.	215	Green house	Salt	143 SNPs	A02, A07,A09, A11	(Dilnur et al., 2019)
G.hirsutum L	217	Green house	SALT	27 SNPs, 2 sQTLs	A07,A08,A13, D08	(Xu et al., 2021)
G.hirsutum L	419	Green house	SALT		A10, D10	(Yasir et al., 2019)
G.hirsutum L	316	4 environment 2 years	Salt	158 sQTLs	A05, A08, A11, A12, D06, D07	(Zhu et al., 2020)
G.hirsutum L	713	Greenhouse	Salt	8 SNPs	A01, A10, D02, D08, D09, D10, and D11,	(Sun et al., 2018a)
G.hirsutum L	149	Greenhouse	Salt	27 SNPs	A01, D01, D08	(Zheng et al., 2021b)
G.hirsutum L, *G.barbadense*	229 and 491	3 environment 3 years	Domestication	111 and 119 Loci	A03, A05,A08, D05, D12	(Fang et al., 2021)
Yucatenese x Acala Maxa	466	2 environments	Domestication	120 QTLs	Multiple chromosomes	(Grover et al., 2020)
G.hirsutum L	419	3 environments	Domestication	45 SNPs	A06, A08	(Dai et al., 2020)
G.hirsutum L	352	3 environments	Domestication	19 loci	A12, D04	(Wang et al., 2017)
G.hirsutum L	288	12 environments 2 years	Heterosis	271 QTNs	D09	(Sarfraz et al., 2021)
G.hirsutum L	550	12 environments 4 location 4 years	Fibre traits	25 sQTLs, 14 hQTLs	A07, D11	(Wang et al., 2022)
G.arboreum L.	215	3 environment	Fiber traits	177 SNPs	A05, A14	(Iqbal et al., 2021)
G.hirsutum	185	2 environment 2 years	Senescence	63 QTNs	A02, D03, D13	(Liu et al., 2022)
G.barbadense L.	336	6 years 4 locations	Fiber and fusarium wilt	6241 SNPs	D03, A05, D11	(Zhao et al., 2021a)
G.arboreum L.	215	Greenhouse	Biomass	83 SNPs	A7, A11	(Hu et al., 2022)
G.hirsutum L.	196	7 environments	Fiber	84 SNPs	A10	(Xing et al., 2019)
G.arboreum L.	246	Greenhouse	Nematode	15 SNPs	A01, A02, A03, A05, A06, A07, A09, A12	(Li et al., 2018b)

(Continued)

TABLE 1 Continued

Cotton species	Population Size (accessions)	Years/location/ environment	Traits	No. of unique SNPS QTLs/sQTLs	Chromosomal location	Reference
G.hirsutum L.	169	2 locations 2 years	Early maturation	29 SNPs	A6, A7, A8, D01, D02, D09	(Li et al., 2018a)
G.arboreum L.	215	Greenhouse	Root color	225 SNPs	A02, A04, A08, A09, A13	(Zhao et al., 2021b)
G.hirsutum L.	196	2 years field location	Oil content	47 SNPs, 28 QTLs	D12	(Yuan et al., 2018)

subgenome. As demonstrated by the large number of fiber quality-related QTL hotspots discovered in the *Dt* subgenome. Similarly, findings have reported that there were more fiber-related genes in the *Dt* subgenome than in the *At* subgenome (Xu et al., 2015). It stands to reason that a collaboration between the two coexisting subgenomes has improved the fiber quality and yield of current tetraploid cotton cultivars.

As it is evident from our previous description of SNPs that, most of the candidate SNPs or loci associated with different yield and FQTs have been found on the *D* subgenome. The probable reason for these findings may be that the D genome provides many fiber genes after merging with another parental diploid cotton (*Gossypium arboreum*) *A* genome during evolution and domestication, even though the *D* genome does not produce any spinnable fiber.

Benefits and limitations of GWAS

GWAS have revolutionized the field of intricate genetic architectures of important agronomic traits over the past decade, providing numerous compelling associations for complex quantitative traits. Despite numerous successes in identifying novel genes and biological pathways and in translating these findings into breeding programs in different crops, GWAS has not been without controversy. Here we provide an overview of the benefits and limitations of GWAS (Table 2).

Meta-GWAS a way forward in genome-wide association studies

Genome-wide association study estimates the statistical association of single nucleotide variants with variables of interest. Although the number of GWASs has exponentially increased in the previous decade, the results have low reproducibility. Different GWASs on the same trait are not in compliance with each other, which means different GWAS results report different genomic regions for the same trait. That might be because of insufficient genetic variation in mapping populations, and different algorithmic and statistical approaches. The potential reason behind this conundrum could be the following factors.

- Low accuracy of phenotyping
- Phenotyping of non/low heritable traits
- Inaccurate sample size in terms of variation
- Wrong statistical analysis
- Population stratification
- Technical biasedness

Meta-GWAS is a statistical technique to combine the results across the GWA studies. It has garnered considerable attention from academics to resolve disparities in genome-wide associations. Meta-GWAS has been widely employed in human studies with a large amount of available data. Such work is becoming more readily available for commonly used

TABLE 2 A comparison of GWAS benefits and limitations.

BENEFITS	LIMITATIONS
No need to have prior knowledge of biological pathways of the traits under study.	The findings must be replicated in independent samples from diverse populations.
Possibility of discovering novel candidate genes that have not been identified using previous methods.	It is necessary to have a large study population.
Encourages the formation of collaborative consortia to recruit sufficient numbers of participants for analysis, which tend to continue their collaboration for subsequent analyses.	GWAS studies look for correlation rather than causation.
Rules out specific genetic associations.	GWAS pinpoints a specific site rather than entire genes.
Provides data on the ancestry of each subject, which assists in matching case subjects with control subjects.	Many of the variations (SNPs) found in GWAS aren't near a protein-coding gene
Data on two types of structural variants, sequence and copy number variations, is provided, resulting in more robust data.	The combined impact of many SNPs typically only explain a modest fraction for any given characteristic.
Identify genetic contributors to common, complex traits for which each gene may only have a small effect.	Finding related variations doesn't always reveal the trait's underlying biology.

populations, such as haplotype and regional mapping populations. Online tools such as MetaGenyo (http://bioinfo.genyo.es/metagenyo/), AraGWAS (https://aragwas.1001genomes.org/) and GWApp (http://gwas.gmi.oeaw.ac.at) are available for Meta-GWAS analysis. Meta-GWAS have been reported in *A. thaliana* and soybean crop, but as per our updated knowledge, no Meta-GWAS has been reported in cotton crop. With the rapidly expanding volume of plant-GWAS data (even that of plant salt tolerance), meta-GWAS will be more popular for studying plants.

Open questions

How feasible is mapping the causative genes/SNPs at exceptionally high resolution in order to grasp the utility of pleiotropy and how it can be applied in crop genetic improvement in the foreseeable future?

Can we predict crop responses to continuously changing ambient conditions by studying the genes and processes that contribute to phenotypic/physiological alterations?

How far are we from designing new climate-smart cultivars precisely by incorporating natural variants revealed by GWAS?

If a trait of interest is controlled by a rare variant, then why is that trait common in a large population but GWAS cannot find that associated rare variant?

Concluding remarks and future perspectives

Cotton crop has evolved and adapted over millions of years, resulting in the accumulation of genetic based natural variations SNPs from various environments over time, resulting in many phenotypic functional variations. Recent advances in DNA sequencing have allowed for in-depth characterization of plant natural variations. These resources can be exploited with GWAS to link phenotypic variations to relevant genes or functional polymorphisms, providing insight into complex traits. Recent developments in the quantitative omics phenotypes (transcriptomics, metabolomics, and epigenomics) give rise to veritable alphabets of association studies TWAS, MWAS and EWAS collectively known as OWAS (omics wide association studies) coupled with highthroughput phenotyping platforms are beginning to yield unprecedented insights for the associated genes underlying agronomically important traits, expediting genomics assisted breeding. Consequently, we believe that developing more efficient GWAS computational algorithms would be highly desirable in this context. A better understanding of genetic variability at the SNPs level will help in *in-situ* conservation, characterization and utilization of diverse germplasm. GWAS will foster the breeding programs by expanding the accessibility to desired germplasm collections. Cotton GWAS's ability to investigate the genetic architecture of complex traits has been established in several studies, and this number is projected to expand rapidly. However, most studies are of limited scope to genetic architecture's main (additive) effect. This complexity is not only the result of differences in gene action, but also determined by ontogenic gene networks or even epigenetic effects, and the interaction with surroundings, and greatly changing environmental conditions. It will soon be possible to identify all underlying genes and their activities for major cotton traits, that will significantly speed up molecular breeding strategies. Although GWAS remains a valuable technique for fully using NGS technology breakthroughs, advances in statistical methods and genetic design can help to accomplish this goal. Diverse statistical methods can help us understand the genetic basis of complex features, but their differing assumptions remind us to careful selection of analysis methodology for each study or to combine allied methods. Improved population designs/core collections and new sequencing and statistical approaches may help identify and manipulate genetic elements generating quantitative variation. These novel designs will improve detection precision and accuracy by restructuring allelic spectra and diminishing confounding effects, which will help overcome intrinsic statistical hurdles.

Author contributions

MY: Writing and idea development; HK: Proof reading and figures; QH: Language and proof reading; MR: Data compilation; MS: Data curation; YJ: Supervision; JR: Supervision. All authors contributed to the article and approved the submitted version.

Funding

Zhejiang Provincial Natural Science Foundation of China (Grant No. Y21C130006) to YJ.

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

Abdelraheem, A., Kuraparthy, V., Hinze, L., Stelly, D., Wedegaertner, T., and Zhang, J. (2021a). Genome-wide association study for tolerance to drought and salt tolerance and resistance to thrips at the seedling growth stage in US upland cotton. *Ind. Crops Prod.* 169. doi: 10.1016/j.indcrop.2021.113645

Abdelraheem, A., Thyssen, G. N., Fang, D. D., Jenkins, J. N., Mccarty, J. C., Wedegaertner, T., et al. (2021b). GWAS reveals consistent QTL for drought and salt tolerance in a MAGIC population of 550 lines derived from intermating of 11 upland cotton (*Gossypium hirsutum*) parents. *Mol. Genet. Genomics* 296, 119–129. doi: 10.1007/s00438-020-01733-2

Atwell, S., Huang, Y. S., Vilhjalmsson, B. J., Willems, G., Horton, M., Li, Y., et al. (2010). Genome-wide association study of 107 phenotypes in arabidopsis thaliana inbred lines. *Nature* 465, 627–631. doi: 10.1038/nature08800

Cai, C., Zhu, G., Zhang, T., and Guo, W. (2017). High-density 80 K SNP array is a powerful tool for genotyping g. hirsutum accessions and genome analysis. *BMC Genomics* 18, 654. doi: 10.1186/s12864-017-4062-2

Chen, Z. J., Scheffler, B. E., Dennis, E., Triplett, B. A., Zhang, T., Guo, W., et al. (2007). Toward sequencing cotton (Gossypium) genomes. *Plant Physiol.* 145, 1303–1310. doi: 10.1104/pp.107.107672

Dai, P., Sun, G., Jia, Y., Pan, Z., Tian, Y., Peng, Z., et al. (2020). Extensive haplotypes are associated with population differentiation and environmental adaptability in upland cotton (*Gossypium hirsutum*). *Theor. Appl. Genet.* 133, 3273–3285. doi: 10.1007/s00122-020-03668-z

Dilnur, T., Peng, Z., Pan, Z., Palanga, K. K., Jia, Y., Gong, W., et al. (2019). Association analysis of salt tolerance in Asiatic cotton (Gossypium arboretum) with SNP markers. *Int. J. Mol. Sci.* 20, 2168. doi: 10.3390/ijms20092168

Ding, M., Jiang, Y., Cao, Y., Lin, L., He, S., Zhou, W., et al. (2014). Gene expression profile analysis of ligon lintless-1 (Li1) mutant reveals important genes and pathways in cotton leaf and fiber development. *Gene* 535, 273–285. doi: 10.1016/j.gene.2013.11.017

Dubois, M., Van Den Broeck, L., and Inze, D. (2018). The pivotal role of ethylene in plant growth. *Trends Plant Sci.* 23, 311–323. doi: 10.1016/j.tplants.2018.01.003

Du, X., Huang, G., He, S., Yang, Z., Sun, G., Ma, X., et al. (2018). Resequencing of 243 diploid cotton accessions based on an updated a genome identifies the genetic basis of key agronomic traits. *Nat. Genet.* 50, 796–802. doi: 10.1038/s41588-018-0116-x

Elassbli, H., Abdelraheem, A., Zhu, Y., Teng, Z., Wheeler, T. A., Kuraparthy, V., et al. (2021). Evaluation and genome-wide association study of resistance to bacterial blight race 18 in U.S. upland cotton germplasm. *Mol. Genet. Genomics* 296, 719–729. doi: 10.1007/s00438-021-01779-w

Fang, D. D., Naoumkina, M., Thyssen, G. N., Bechere, E., Li, P., and Florane, C. B. (2020). An EMS-induced mutation in a tetratricopeptide repeat-like superfamily protein gene (Ghir_A12G008870) on chromosome A12 is responsible for the liy short fiber phenotype in cotton. *Theor. Appl. Genet.* 133, 271–282. doi: 10.1007/s00122-019-03456-4

Fang, L., Wang, Q., Hu, Y., Jia, Y., Chen, J., Liu, B., et al. (2017). Genomic analyses in cotton identify signatures of selection and loci associated with fiber quality and yield traits. *Nat. Genet.* 49, 1089–1098. doi: 10.1038/ng.3887

Fang, L., Zhao, T., Hu, Y., Si, Z., Zhu, X., Han, Z., et al. (2021). Divergent improvement of two cultivated allotetraploid cotton species. *Plant Biotechnol. J.* 19, 1325–1336. doi: 10.1111/pbi.13547

Gapare, W., Conaty, W., Zhu, Q.-H., Liu, S., Stiller, W., Llewellyn, D., et al. (2017). Genome-wide association study of yield components and fibre quality traits in a cotton germplasm diversity panel. *Euphytica* 213, 66. doi: 10.1007/s10681-017-1855-y

Gilbert, M. K., Kim, H. J., Tang, Y., Naoumkina, M., and Fang, D. D. (2014). Comparative transcriptome analysis of short fiber mutants ligon-lintless 1 and 2 reveals common mechanisms pertinent to fiber elongation in cotton (*Gossypium hirsutum* l.). *PloS One* 9, e95554. doi: 10.1371/journal.pone.0095554

Gong, Q., Yang, Z., Chen, E., Sun, G., He, S., Butt, H. I., et al. (2018). A phi-class glutathione s-transferase gene for verticillium wilt resistance in gossypium arboreum identified in a genome-wide association study. *Plant Cell Physiol.* 59, 275–289. doi: 10.1093/pcp/pcx180

Grover, C. E., Yoo, M. J., Lin, M., Murphy, M. D., Harker, D. B., Byers, R. L., et al. (2020). Genetic analysis of the transition from wild to domesticated cotton (*Gossypium hirsutum* l.). *G3 (Bethesda Md.)* 10, 731–754. doi: 10.1534/g3.119.400909

Guo, A., Hao, J., Su, Y., Li, B., Zhao, N., Zhu, M., et al. (2022). Two aquaporin genes, GhPIP2;7 and GhTIP2;1, positively regulate the tolerance of upland cotton to salt and osmotic stresses. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.780486

Gupta, P. K., Kulwal, P. L., and Jaiswal, V. (2014). "Association mapping in crop plants: Opportunities and challenges," in *Advances in genetics*, vol. 85 . Eds. T. Friedmann, J. C. Dunlap and S. F. Goodwin. (Cambridge: Elsevier), 109–147.

Hayes, B., and Goddard, M. E. (2001). The distribution of the effects of genes affecting quantitative traits in livestock. *Genet. Sel. Evol.* 33, 209–229. doi: 10.1186/1297-9686-33-3-209

He, F., Pasam, R., Shi, F., Kant, S., Keeble-Gagnere, G., Kay, P., et al. (2019). Exome sequencing highlights the role of wild-relative introgression in shaping the adaptive landscape of the wheat genome. *Nat. Genet.* 51, 896–904. doi: 10.1038/s41588-019-0382-2

He, S., Sun, G., Geng, X., Gong, W., Dai, P., Jia, Y, et al. (2021). The genomic basis of geographic differentiation and fiber improvement in cultivated cotton. *Nat. Genet.* 53, 916–924. doi: 10.1038/s41588-021-00844-9

Hinchliffe, D. J., Turley, R. B., Naoumkina, M., Kim, H. J., Tang, Y., Yeater, K. M., et al. (2011). A combined functional and structural genomics approach identified an EST-SSR marker with complete linkage to the ligon lintless-2 genetic locus in cotton (*Gossypium hirsutum* l.). *BMC Genom.* 12, 445. doi: 10.1186/1471-2164-12-445

Hou, S., Zhu, G., Li, Y., Li, W., Fu, J., Niu, E., et al. (2018). Genome-wide association studies reveal genetic variation and candidate genes of drought stress related traits in cotton (*Gossypium hirsutum* l.). *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.01276

Hu, Y., Chen, J., Fang, L., Zhang, Z., Ma, W., Niu, Y., et al. (2019). Gossypium barbadense and *Gossypium hirsutum* genomes provide insights into the origin and evolution of allotetraploid cotton. *Nat. Genet.* 51, 739–748. doi: 10.1038/s41588-019-0371-5

Hu, D., He, S., Jia, Y., Nazir, M. F., Sun, G., Geng, X., et al. (2022). Genome-wide association study for seedling biomass-related traits in. *BMC Plant Biol.* 22, 54. doi: 10.1186/s12870-022-03443-w

Hussain, Q., Asim, M., Zhang, R., Khan, R., Farooq, S., and Wu, J. (2021). Transcription factors interact with ABA through gene expression and signaling pathways to mitigate drought and salinity stress. *Biomolecules* 11. doi: 10.3390/biom11081159

Iqbal, M. S., Tang, S., Sarfraz, Z., Iqbal, M. S., Li, H., He, S., et al. (2021). Genetic factors underlying single fiber quality in a-genome donor Asian cotton (Gossypium arboreum). *Front. Genet.* 12. doi: 10.3389/fgene.2021.758665

Jiang, D., Gu, X., and He, Y. (2009). Establishment of the winter-annual growth habit *via* FRIGIDA-mediated histone methylation at FLOWERING LOCUS c in arabidopsis. *Plant Cell* 21, 1733–1746. doi: 10.1105/tpc.109.067967

Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J., et al. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* 484, 55–61. doi: 10.1038/nature10944

Klosterman, S. J., Atallah, Z. K., Vallad, G. E., and Subbarao, K. V. (2009). Diversity, pathogenicity, and management of verticillium species. *Annu. Rev. Phytopathol.* 47, 39–62. doi: 10.1146/annurev-phyto-080508-081748

Lamichhaney, S., Fan, G., Widemo, F., Gunnarsson, U., Thalmann, D. S., Hoeppner, M. P., et al. (2016). Structural genomic changes underlie alternative reproductive strategies in the ruff (Philomachus pugnax). *Nat. Genet.* 48, 84–88. doi: 10.1038/ng.3430

Li, B., Chen, L., Sun, W., Wu, D., Wang, M., Yu, Y., et al. (2020). Phenomics-based GWAS analysis reveals the genetic architecture for drought resistance in cotton. *Plant Biotechnol. J.* 18, 2533–2544. doi: 10.1111/pbi.13431

Li, R., Erpelding, J. E., and Stetina, S. R. (2018b). Genome-wide association study of gossypium arboreum resistance to reniform nematode. *BMC Genet.* 19, 52. doi: 10.1186/s12863-018-0662-3

Li, F., Fan, G., Lu, C., Xiao, G., Zou, C., Kohel, R. J., et al. (2015). Genome sequence of cultivated upland cotton (*Gossypium hirsutum* TM-1) provides insights into genome evolution. *Nat. Biotechnol.* 33, 524–530. doi: 10.1038/nbt.3208

Li, Z.-J., Jia, G.-Q., Li, X.-Y., Li, Y.-C., Zhi, H., Tang, S., et al. (2021c). Identification of blast-resistance loci through genome-wide association analysis in foxtail millet (Setaria italica (L.) beauv.). *J. Integr. Agric.* 20, 2056–2064. doi: 10.1016/s2095-3119(20)63196-3

Li, H. M., Liu, S. D., Ge, C. W., Zhang, X. M., Zhang, S. P., Chen, J., et al. (2019). Analysis of drought tolerance and associated traits in upland cotton at the seedling stage. *Int. J. Mol. Sci.* 20(16), 3888. doi: 10.3390/ijms20163888

Li, H. B., Qin, Y. M., Pang, Y., Song, W. Q., Mei, W. Q., and Zhu, Y. X. (2007). A cotton ascorbate peroxidase is involved in hydrogen peroxide homeostasis during fibre cell development. *New Phytol.* 175, 462–471. doi: 10.1111/j.1469-8137.2007.02120.x

Liu, R., Gong, J., Xiao, X., Zhang, Z., Li, J., Liu, A., et al. (2018). GWAS analysis and QTL identification of fiber quality traits and yield components in upland cotton using enriched high-density SNP markers. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.01067

Liu, Q., Li, L., Feng, Z., and Yu, S. (2022). Uncovering novel genomic regions and candidate genes for senescence-related traits by genome-wide association studies in upland cotton (*Gossypium hirsutum* l.). *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.809522

Liu, R., Wang, B., Guo, W., Qin, Y., Wang, L., Zhang, Y., et al. (2012). Quantitative trait loci mapping for yield and its components by using two immortalized populations of a heterotic hybrid in *Gossypium hirsutum* l. *Mol. Breed.* 29, 297–311. doi: 10.1007/s11032-011-9547-0

Liu, X., Zhao, B., Zheng, H. J., Hu, Y., Lu, G., Yang, C. Q., et al. (2015). Gossypium barbadense genome sequence provides insight into the evolution of extra-long staple fiber and specialized metabolites. *Sci. Rep.* 5, 14139. doi: 10.1038/srep14139

Li, C., Wang, Y., Ai, N., Li, Y., and Song, J. (2018a). A genome-wide association study of early-maturation traits in upland cotton based on the CottonSNP80K array. *J. Integr. Plant Biol.* 60, 970–985. doi: 10.1111/jipb.12673

Li, J., Yuan, D., Wang, P., Wang, Q., Sun, M., Liu, Z., et al. (2021a). Cotton pan-genome retrieves the lost sequences and genes during domestication and selection. *Genome Biol.* 22, 119. doi: 10.1186/s13059-021-02351-w

Li, L., Zhang, C., Huang, J., Liu, Q., Wei, H., Wang, H., et al. (2021b). Genomic analyses reveal the genetic basis of early maturity and identification of loci and candidate genes in upland cotton (*Gossypium hirsutum* l.). *Plant Biotechnol. J.* 19, 109–123. doi: 10.1111/pbi.13446

Ma, Z., He, S., Wang, X., Sun, J., Zhang, Y., Zhang, G., et al. (2018). Resequencing a core collection of upland cotton identifies genomic variation and loci influencing fiber quality and yield. *Nat. Genet.* 50, 803–813. doi: 10.1038/s41588-018-0119-7

Mooney, S., and Hellmann, H. (2010). Vitamin B6: Killing two birds with one stone? *Phytochemistry* 71, 495–501. doi: 10.1016/j.phytochem.2009.12.015

Mo, H. J., Sun, Y. X., Zhu, X. L., Wang, X. F., Zhang, Y., Yang, J., et al. (2016). Cotton s-adenosylmethionine decarboxylase-mediated spermine biosynthesis is required for salicylic acid- and leucine-correlated signaling in the defense response to verticillium dahliae. *Planta* 243, 1023–1039. doi: 10.1007/s00425-015-2463-5

Nazir, M. F., He, S., Ahmed, H., Sarfraz, Z., Jia, Y., Li, H., et al. (2021). Genomic insight into the divergence and adaptive potential of a forgotten landrace g. hirsutum l. purpurascens. *J. Genet. Genomics* 48, 473–484. doi: 10.1016/j.jgg.2021.04.009

Prasad, A., Senthil-Kumar, M., and Prasad, M. (2021). Complex molecular mechanisms determine fitness of plants to biotic and abiotic stresses. *J. Plant Biochem. Biotechnol.* 30, 633–635. doi: 10.1007/s13562-021-00751-4

Qin, Y. M., Hu, C. Y., Pang, Y., Kastaniotis, A. J., Hiltunen, J. K., and Zhu, Y. X. (2007). Saturated very-long-chain fatty acids promote cotton fiber and arabidopsis cell elongation by activating ethylene biosynthesis. *Plant Cell* 19, 3692–3704. doi: 10.1105/tpc.107.054437

Qin, Y. M., and Zhu, Y. X. (2011). How cotton fibers elongate: a tale of linear cell-growth mode. *Curr. Opin. Plant Biol.* 14, 106–111. doi: 10.1016/j.pbi.2010.09.010

Raschke, M., Boycheva, S., Crevecoeur, M., Nunes-Nesi, A., Witt, S., Fernie, A. R., et al. (2011). Enhanced levels of vitamin B(6) increase aerial organ size and positively affect stress tolerance in arabidopsis. *Plant J.* 66, 414–432. doi: 10.1111/j.1365-313X.2011.04499.x

Rice, B. R., Fernandes, S. B., and Lipka, A. E. (2020). Multi-trait genome-wide association studies reveal loci associated with maize inflorescence and leaf architecture. *Plant Cell Physiol.* 61, 1427–1437. doi: 10.1093/pcp/pcaa039

Sarfraz, Z., Iqbal, M. S., Geng, X., Iqbal, M. S., Nazir, M. F., Ahmed, H., et al. (2021). GWAS mediated elucidation of heterosis for metric traits in cotton (*Gossypium hirsutum* l.) across multiple environments. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.565552

Schweighofer, A., Hirt, H., and Meskiene, I. (2004). Plant PP2C phosphatases: emerging functions in stress signaling. *Trends Plant Sci.* 9, 236–243. doi: 10.1016/j.tplants.2004.03.007

Shan, C. M., Shangguan, X. X., Zhao, B., Zhang, X. F., Chao, L. M., Yang, C. Q., et al. (2014). Control of cotton fibre elongation by a homeodomain transcription factor GhHOX3. *Nat. Commun.* 5, 5519. doi: 10.1038/ncomms6519

Shen, X., Guo, W., Lu, Q., Zhu, X., Yuan, Y., and Zhang, T. (2007). Genetic mapping of quantitative trait loci for fiber quality and yield trait by RIL approach in upland cotton. *Euphytica* 155, 371–380. doi: 10.1007/s10681-006-9338-6

Shi, Y. H., Zhu, S. W., Mao, X. Z., Feng, J. X., Qin, Y. M., Zhang, L., et al. (2006). Transcriptome profiling, molecular biological, and physiological studies reveal a major role for ethylene in cotton fiber cell elongation. *Plant Cell* 18, 651–664. doi: 10.1105/tpc.105.040303

Sinclair-Waters, M., Bradbury, I. R., Morris, C. J., Lien, S., Kent, M. P., and Bentzen, P. (2018). Ancient chromosomal rearrangement associated with local adaptation of a postglacially colonized population of Atlantic cod in the northwest Atlantic. *Mol. Ecol.* 27, 339–351. doi: 10.1111/mec.14442

Sink, K. C., and Grey, W. E. (1999). A root-injection method to assess verticillium wilt resistance of peppermint (mentha × piperita l.) and its use in identifying resistant somaclones of cv. black mitcham. *Euphytica* 106, 223–230. doi: 10.1023/a:1003591908308

Somegowda, V. K., Rayaprolu, L., Rathore, A., Deshpande, S. P., and Gupta, R. (2021). Genome-wide association studies (GWAS) for traits related to fodder quality and biofuel in sorghum: Progress and prospects. *Protein Pept. Lett.* 28, 843–854. doi: 10.2174/0929866528666210127153103

Song, C., Li, W., Pei, X., Liu, Y., Ren, Z., He, K., et al. (2019). Dissection of the genetic variation and candidate genes of lint percentage by a genome-wide association study in upland cotton. *Theor. Appl. Genet.* 132, 1991–2002. doi: 10.1007/s00122-019-03333-0

Sun, Z., Li, H., Zhang, Y., Li, Z., Ke, H., Wu, L., et al. (2018a). Identification of SNPs and candidate genes associated with salt tolerance at the seedling stage in cotton (*Gossypium hirsutum* l.). *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.01011

Sun, Z., Wang, X., Liu, Z., Gu, Q., Zhang, Y., Li, Z., et al. (2017). Genome-wide association study discovered genetic variation and candidate genes of fibre quality traits in *Gossypium hirsutum* l. *Plant Biotechnol. J.* 15, 982–996. doi: 10.1111/pbi.12693

Sun, Z., Wang, X., Liu, Z., Gu, Q., Zhang, Y., Li, Z., et al. (2018b). A genome-wide association study uncovers novel genomic regions and candidate genes of yield-related traits in upland cotton. *Theor. Appl. Genet.* 131, 2413–2425. doi: 10.1007/s00122-018-3162-y

Su, J., Wang, C., Hao, F., Ma, Q., Wang, J., Li, J., et al. (2019). Genetic detection of lint percentage applying single-locus and multi-locus genome-wide association studies in Chinese early-maturity upland cotton. *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.00964

Thyssen, G. N., Fang, D. D., Zeng, L., Song, X., Delhom, C. D., Condon, T. L., et al. (2016). The immature fiber mutant phenotype of cotton (*Gossypium hirsutum*) is linked to a 22-bp frame-shift deletion in a mitochondria targeted pentatricopeptide repeat gene. *G3 (Bethesda Md.)* 6, 1627–1633. doi: 10.1534/g3.116.027649

Titiz, O., Tambasco-Studart, M., Warzych, E., Apel, K., Amrhein, N., Laloi, C., et al. (2006). PDX1 is essential for vitamin B6 biosynthesis, development and stress tolerance in arabidopsis. *Plant J.* 48, 933–946. doi: 10.1111/j.1365-313X.2006.02928.x

Ullah, A., Sun, H., Yang, X., and Zhang, X. (2017). Drought coping strategies in cotton: increased crop per drop. *Plant Biotechnol. J.* 15, 271–284. doi: 10.1111/pbi.12688

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., Mccarthy, M. I., Brown, M. A., et al. (2017). 10 years of GWAS discovery: Biology, function, and translation. *Am. J. Hum. Genet.* 101, 5–22. doi: 10.1016/j.ajhg.2017.06.005

Wang, P., Dong, N., Wang, M., Sun, G., Jia, Y., Geng, X., et al. (2022). Introgression from gossypium hirsutum is a driver for population divergence and genetic diversity in gossypium barbadense. *Plant J.* 110, 764–780. doi: 10.1111/tpj.15702

Wang, P., He, S., Sun, G., Pan, Z., Sun, J., Geng, X., et al. (2021). Favorable pleiotropic loci for fiber yield and quality in upland cotton (*Gossypium hirsutum*). *Sci. Rep.* 11, 15935. doi: 10.1038/s41598-021-95629-9

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

Wang, M., Qi, Z., Thyssen, G. N., Naoumkina, M., Jenkins, J. N., Mccarty, J. C., et al. (2022). Genomic interrogation of a MAGIC population highlights genetic

factors controlling fiber quality traits in cotton. *Commun. Biol.* 5, 60. doi: 10.1038/s42003-022-03022-7

Wang, M., Tu, L., Lin, M., Lin, Z., Wang, P., Yang, Q., et al. (2017). Asymmetric subgenome selection and cis-regulatory divergence during cotton domestication. *Nat. Genet.* 49, 579–587. doi: 10.1038/ng.3807

Wang, M., Tu, L., Yuan, D., Zhu, D., Shen, C., Li, J., et al. (2019). Reference genome sequences of two cultivated allotetraploid cottons, *Gossypium hirsutum* and gossypium barbadense. *Nat. Genet.* 51, 224–229. doi: 10.1038/s41588-018-0282-x

Wendel, J. F. (1989). New world tetraploid cottons contain old world cytoplasm. *Proc. Natl. Acad. Sci. U.S.A.* 86, 4132–4136. doi: 10.1073/pnas.86.11.4132

Wendel, J. F., and Albert, V. A. (1992). Phylogenetics of the cotton genus (Gossypium): Character-state weighted parsimony analysis of chloroplast-DNA restriction site data and its systematic and biogeographic implications. *Syst. Bot.* 17, 115–143. doi: 10.2307/2419069

Xing, H., Yuan, Y., Zhang, H., Wang, L., Mao, L., Tao, J., et al. (2019). Multi-environments and multi-models association mapping identified candidate genes of lint percentage and seed index in *Gossypium hirsutum* l. *Mol. Breed.* 39. doi: 10.1007/s11032-019-1063-7

Xu, P., Guo, Q., Meng, S., Zhang, X., Xu, Z., Guo, W., et al. (2021). Genome-wide association analysis reveals genetic variations and candidate genes associated with salt tolerance related traits in *Gossypium hirsutum*. *BMC Genomics* 22, 26. doi: 10.1186/s12864-020-07321-3

Xu, Z., Yu, J., Kohel, R. J., Percy, R. G., Beavis, W. D., Main, D., et al. (2015). Distribution and evolution of cotton fiber development genes in the fibreless gossypium raimondii genome. *Genomics* 106, 61–69. doi: 10.1016/j.ygeno.2015.03.002

Yao, D., Wang, Y., Li, Q., Ouyang, X., Li, Y., Wang, C., et al. (2018). Specific upregulation of a cotton phytoene synthase gene produces golden cottonseeds with enhanced provitamin a. *Sci. Rep.* 8, 1348. doi: 10.1038/s41598-018-19866-1

Yasir, M., He, S., Sun, G., Geng, X., Pan, Z., Gong, W., et al. (2019). A genome-wide association study revealed key SNPs/Genes associated with salinity stress tolerance in upland cotton. *Genes (Basel)* 10(10), 829. doi: 10.3390/genes10100829

Yuan, Y., Wang, X., Wang, L., Xing, H., Wang, Q., Saeed, M., et al. (2018). Genome-wide association study identifies candidate genes related to seed oil composition and protein content in *Gossypium hirsutum* l. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.01359

Yuan, Y., Zhang, H., Wang, L., Xing, H., Mao, L., Tao, J., et al. (2019). Candidate quantitative trait loci and genes for fiber quality in *Gossypium hirsutum* l. detected using single- and multi-locus association mapping. *Indust. Crops Prod.* 134, 356–369. doi: 10.1016/j.indcrop.2019.04.010

Zhang, J., Abdelraheem, A., Thyssen, G. N., Fang, D. D., Jenkins, J. N., Mccarty, J. C., et al. (2020). Evaluation and genome-wide association study of verticillium wilt resistance in a MAGIC population derived from intermating of eleven upland cotton (*Gossypium hirsutum*) parents. *Euphytica* 216, (9). doi: 10.1007/s10681-019-2547-6

Zhang, T., Hu, Y., Jiang, W., Fang, L., Guan, X., Chen, J., et al. (2015). Sequencing of allotetraploid cotton (*Gossypium hirsutum* l. acc. TM-1) provides a resource for fiber improvement. *Nat. Biotechnol.* 33, 531–537. doi: 10.1038/nbt.3207

Zhang, H. B., Li, Y., Wang, B., and Chee, P. W. (2008). Recent advances in cotton genomics. *Int. J. Plant Genomics* 2008, 742304. doi: 10.1155/2008/742304

Zhang, Y., Chen, B., Sun, Z., Liu, Z., Cui, Y., Ke, H, et al (2021). A large-scale genomic association analysis identifies a fragment in Dt11 chromosome conferring cotton Verticillium wilt resistance. *Plant Biotechnol. J.* 19, 2126–2138. doi 10.1111/pbi.13650

Zhao, Y., Chen, W., Cui, Y., Sang, X., Lu, J., Jing, H., et al. (2021a). Detection of candidate genes and development of KASP markers for verticillium wilt resistance by combining genome-wide association study, QTL-seq and transcriptome sequencing in cotton. *Theor. Appl. Genet.* 134, 1063–1081. doi: 10.1007/s00122-020-03752-4

Zhao, Z., Hu, D., Azhar, M. T., Li, H., Ma, C., He, S., et al. (2021b). Genome-wide association and transcriptome analysis of root color-related genes in. *Planta* 253, 95. doi: 10.1007/s00425-021-03622-3

Zhao, G., Lian, Q., Zhang, Z., Fu, Q., He, Y., Ma, S., et al. (2019). A comprehensive genome variation map of melon identifies multiple domestication events and loci influencing agronomic traits. *Nat. Genet.* 51, 1607–1615. doi: 10.1038/s41588-019-0522-8

Zhao, N., Wang, W., Grover, C. E., Jiang, K., Pan, Z., Guo, B., et al. (2022). Genomic and GWAS analyses demonstrate phylogenomic relationships of gossypium barbadense in China and selection for fibre length, lint percentage and fusarium wilt resistance. *Plant Biotechnol. J.* 20, 691–710. doi: 10.1111/pbi.13747

Zheng, J., Lin, R., Pu, L., Wang, Z., Mei, Q., Zhang, M., et al. (2021a). Ectopic expression of CrPIP2;3, a plasma membrane intrinsic protein gene from the halophyte canavalia rosea, enhances drought and salt-alkali stress tolerance in arabidopsis. *Int. J. Mol. Sci.* 22(2), 565. doi: 10.3390/ijms22020565

Zheng, J., Zhang, Z., Gong, Z., Liang, Y., Sang, Z., Xu, Y., et al. (2021b). Genome-wide association analysis of salt-tolerant traits in terrestrial cotton at seedling stage. *Plants (Basel)* 11, 97. doi: 10.3390/plants11010097

Zhong, H., Liu, S., Sun, T., Kong, W., Deng, X., Peng, Z., et al. (2021). Multi-locus genome-wide association studies for five yield-related traits in rice. *BMC Plant Biol.* 21, 364. doi: 10.1186/s12870-021-03146-8

Zhu, G., Gao, W., Song, X., Sun, F., Hou, S., Liu, N., et al. (2020). Genome-wide association reveals genetic variation of lint yield components under salty field conditions in cotton (*Gossypium hirsutum* l.). *BMC Plant Biol.* 20, 23. doi: 10.1186/s12870-019-2187-y

Zhu, G., Hou, S., Song, X., Wang, X., Wang, W., Chen, Q., et al. (2021). Genome-wide association analysis reveals quantitative trait loci and candidate genes involved in yield components under multiple field environments in cotton (*Gossypium hirsutum*). *BMC Plant Biol.* 21, 250. doi: 10.1186/s12870-021-03009-2

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 07 November 2022
DOI 10.3389/fpls.2022.1042222

OPEN ACCESS

EDITED BY
Nisha Singh,
Gujarat Biotechnology University, India

REVIEWED BY
Mahesh Rao,
National Institute for Plant Biotechnology (ICAR), India
Alkesh Hada,
Agricultural Research Organization (ARO), Israel

*CORRESPONDENCE
Gunjan Tiwari
gunjantiwari@cimap.res.in

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

RECEIVED 12 September 2022
ACCEPTED 05 October 2022
PUBLISHED 07 November 2022

CITATION
Chaturvedi T, Gupta AK, Shanker K, Dubey BK and Tiwari G (2022) Maximizing genetic gain through unlocking genetic variation in different ecotypes of kalmegh (*Andrographis paniculata* (Burm. f.) Nee). *Front. Plant Sci.* 13:1042222. doi: 10.3389/fpls.2022.1042222

Maximizing genetic gain through unlocking genetic variation in different ecotypes of kalmegh (*Andrographis paniculata* (Burm. f.) Nee)

Trishna Chaturvedi[1], Anil Kumar Gupta[1], Karuna Shanker[2], Basant Kumar Dubey[3] and Gunjan Tiwari[1]*

[1]Division of Plant Breeding and Genetic Resource Conservation, Central Institute of Medicinal and Aromatic Plants, Council of Scientific and Industrial Research, Lucknow, Uttar Pradesh, India, [2]Phytochemistry Division, Central Institute of Medicinal and Aromatic Plants, Council of Scientific and Industrial Research, Lucknow, Uttar Pradesh, India, [3]Biotechnology Division, Central Institute of Medicinal and Aromatic Plants, Council of Scientific and Industrial Research, Lucknow, Uttar Pradesh, India

Andrographis paniculata, commonly known as kalmegh is among the most popular medicinal herbs in Southeast Asia. It is widely cultivated for medicinal purposes. The bioactive molecule, Andrographolide accumulated in herb leaves has immense therapeutic and economic potential. However, comprehensive information regarding genetic diversity is very limited in this species. The present study assessed genetic diversity between and within the six populations (ecotypes) of twenty-four kalmegh accessions using multiple datasets (agro-morphological traits, phytochemical traits, and genic markers). This is the established report where EST-SSR (Expressed sequence tags-Simple Sequence Repeat) markers have been used to unlock genetic variation in kalmegh. Here, we identified and developed ninety-one metabolic pathway-specific EST-SSR markers. Finally, 32 random EST-SSR primer pairs were selected for genetic diversity assessment. Multivariate analysis to unveil the agro-morphological, phytochemical and genotypic variability was helpful in discriminating various germplasms studied in the present study. Among all the morphological discriptors used in present study, days to fifty percent flowering and dry herb yield were found as potential selection index for AP genetic improvement. Hierarchical cluster analysis built with agro-morphological data identified three major groups. However, corresponding analysis with phytochemical and molecular data generated two clear-cut groups among the studied individuals. Moreover, the grouping of individuals into different clusters using multiple datasets was geographically independent, and also showed inconsistency in grouping among agromorphological, phytochemical and molecular dataset based clusters. However, joint analysis using agro-morphological, phytochemical and genotypic information generated two genetic groups, which could be a valuable resource for identifying complementary crossing panels in the kalmegh breeding

program. The accessions AP7, AP13, AP5, AP3 belong to cluster I and accessions AP17, AP18 belong to cluster II could be utilized as potential donors for high dry herb yield and andrographolide content, respectively in different selective breeding programs of AP. Thus, our results provided useful information about the overall genetic diversity and variation in economic traits useful for initiating selective breeding programs for contrasting traits of interest and maximizing genetic gain in kalmegh.

KEYWORDS

genetic diversity, agro-morphological, phytochemical, EST-SSR markers, *Andrographis paniculata*

1 Introduction

Integrated approaches are required to accelerate the genetic gain over time in crop improvement programs concerning various economic traits and to realize the improved gain in farmers' fields. One of the critical components required to maximize genetic gain is the enhanced genetic variance (σ^2g) which can be achieved by unlocking or creating desirable alleles/genotypes and deploying them for improving key traits. It is desirable to assess new alleles for target traits and introgress them in breeding populations while maintaining genetic diversity. Comprehensive approaches involving phenotyping, chemotyping, genotyping, and bioinformatics tools present an enormous opportunity to measure novel alleles precisely and review the breeder's equations for genetic improvement over time.

Kalmegh (*Andrographis paniculata* (Burm. f.) Wall. ex Nees.) is a self-pollinated, annual, diploid (2n=50), and highly traded medicinal herb of the family Acanthaceae. It is well known with other vernacular names like Green Chirata, bhui-neem, king of bitter, etc. There are roughly 40 species in the genus *Andrographis*, with kalmegh (*Andrographis paniculata*) being the most popular medicinal plant species (Boopathi, 2000) of the genus. It is said to have originated in the southern parts of India and Sri Lanka and has a broad geographical distribution in tropical and subtropical regions of the country. Several bioactive specialized metabolites have been identified from *Andrographis paniculata* (AP), including ent-labdane-related diterpenes (ent-LRDs), phenylpropanoids, xanthones, and flavonoids (Koteswara et al., 2004; Li et al., 2007). Ent-LRDs that accumulated in the leaves and thought to be the major bioactive ingredients of kalmegh are Andrographolide (AG), Neoandrographolide (NAG), 14-deoxy-11,12-didehydroandrographolide (DDAG) and Andographanin (AN) (Ooi et al., 2011; Valdiani et al., 2012; Mondal et al., 2013; Lin et al., 2014; Raghavan et al., 2014; Wang et al., 2014) among which, Andrographolide is the most prevalent ones and has been widely researched for various pharmacological activities (Chopra, 1956; Akbar, 2011; Garg et al., 2015; Chauhan et al., 2019). The andrographolide based drugs are reported to have numerous biological activities such as hepatoprotective, anti-diabetic, anti-oxidant, immune-modulatory, anti-allergic, anti-pyretic, antidiarrhoeal, and anti-HIV activity (Garg et al., 2015). Recent experiments have also shown its anti-cancerous activity in human cancer cells (Luo et al., 2014). Andrographis paniculata has blood purifying action and is suggested for curing gonorrhea, leprosy, and several skin disorders (Pandey and Rao, 2018). The herb derived from leaves or aerial parts of the Kalmegh is known as Chuaxinlian, Lanhelian, or Yijianxi in the Chinese system of medicine. It possesses similar properties as described in the traditional system of medicine in India. Various preparations and formulations of the Kalmegh have been used to treat infectious and non-infectious diseases with significant efficacy reported in case of epidemic encephalitis B, vaginitis, pelvic inflammation, herpes zoster, neonatal subcutaneous annual ulcer, chickenpox, mumps, neurodermatitis, eczema, and burns (Lim et al., 2012). Besides andrographolides, flavonoids, caffeic acid, and chlorogenic acid are also produced in this plant (Rao et al., 2004). Kalmegh is reported to show high efficacy against chronic malaria and is often used as an alternative to *Swertia chirata* (Valdiani et al., 2012). Recent *in-silico* analysis suggested the potential role of Andrographolide against SARS-CoV-2 main protease (Mpro) (Enmozhi et al., 2020). Thus, *Andrograhis paniculata* (AP) has immense therapeutic and economic potential. Quality dry herbs of the plant are sold for as much as Rs. 17-30/kg (source: e-Charak, 2008). The costs of Andrographis powder, with varied diterpenoid content, ranged from US$0.12 per gram to US$ 0.70 per gram in July 2016. Also, the price offered by Sigma-Aldrich for 100 and 500gm packages of pure Andrographolide (98%) was 44.2USD and 162.50 USD, respectively, in the same year (Pandey and Rao, 2018). However, there is a considerable gap between the demand and supply of quality raw herbs on national and international platforms. The heavy demand for diterpene andrographolide has motivated Indian farmers to commercialize kalmegh cultivation. However, to meet global needs, most raw herbs are

rigorously collected from wild habitats causing massive mutilation of genetic diversity and shifting this species on the verge of extinction. Growing kalmegh under captive cultivation is the only way to prevent the loss of natural diversity from the wild source and meet global demand. Plant breeding and biotechnology are potential tools to bring kalmegh into captive cultivation, which entails great genetic variations and maximizes genetic gain by utilizing more selection programs effectively. Although few commercial cultivars are available, they are ecotype specific, minimal, and insufficient to meet national and international demands. The major bottleneck for the resulting yield gap is the narrow genetic base of the existing cultivars. Sustainable yield increase can only be achieved by introducing new sources of favorable alleles from different ecotypes into the rapid breeding cycle and attaining kalmegh production challenges.

Systematic evaluation and cataloging of genetic diversity at morphological and phytochemical levels are extremely useful for effective conservation and optimum genetic amelioration of allelic and genotypic variability. In the recent past, the introgression of molecular markers has augmented the accumulation of genetic gains achieved by morpho-chemical descriptors based characterization. In kalmegh, an array of genetic diversity studies has been done using agro-morphological traits (Nagvanshi and Tirkey, 2016), phytochemical traits (Sabu et al., 2001; Archana et al., 2016), and molecular markers, including RAPD, ISSR, SCoT, CBDP and Genomic SSRs (Padmesh et al., 1999; Maison et al., 2005; Kumar and Shekhawat, 2009; Wijarat et al., 2011; Ghosh et al., 2014; Tiwari et al., 2016; Kumar et al., 2020). Compared with dominant markers, the reproducibility and reliability of SSR (simple sequence repeat) markers are high due to co-dominance, high polymorphism, uniform distribution throughout the genome, and multi-allelic nature (Varshney et al., 2005). However, the number of SSR markers reported in kalmegh is limited, and most of them are genomic SSRs, the construction of which is a tedious and costly affair.

In contrast, the advent of modern genomics-based *denovo* transcriptome assembly in kalmegh has rapidly paved the way to mine and develop a large number of high throughput unigene-based SSRs (EST-SSRs) at low cost. These SSRs are possibly linked with particular transcriptional regions that contribute to agronomic traits and are well suited for marker-assisted breeding in *Andrographis paniculata*. To date, EST-SSR based genetic characterization has not been done in AP. Moreover, comprehensive data regarding genetic diversity studies are scanty and provide the rationale for this study (Lattoo et al., 2008; Sharma et al., 2009; Valdiani et al., 2014; Hiremath et al., 2020). This information would be essential to exploit beneficial genes present in indigenous genetic resources of different ecotypes to increase the selection efficiency in kalmegh breeding and for adequate biodiversity protection and management. Moreover, it would also be intriguing to see if there is any conceptual or empirical agreement between agro-phytochemical features and molecular markers to accelerate kalmegh breeding and maximize genetic gain. Thus, with this backup, the present study was designed to (i) evaluate the genetic diversity of various germplasm of kalmegh at agro-morphological and phytochemical levels to identify superior individuals/genotypes for the breeding purpose (ii) assess molecular diversity and population structure by employing metabolic pathways specific EST-SSRs and (iii) finally comparison and joint analysis of agro-phytochemical traits and molecular information to provide in-depth insight into the genetic variability present in studied germplasm.

2 Materials and methods

2.1 Plant materials

The experimental material covered twenty-four diverse accessions of Kalmegh (*A. paniculata)*, including one released and cultivated variety as a local check (Supplementary Table S1). All the accessions were procured from different states of India, covering six agroecological regions of the country (Figure 1), and conserved and maintained over the years at the National Gene bank of CSIR-Central Institute of Medicinal and Aromatic Plants (CIMAP), Lucknow (India).

2.2 Experimental designs

The present study was initiated at the Field Research Centre of CSIR- CIMAP, Lucknow, India (80.50°C E longitude and 26.5°C N latitude), where the annual temperature varies between 5°C to 45 °C. For the genetic diversity study, nursery planting was done under outdoor conditions using individual open-pollinated seed lots of twenty-four accessions in earthen pots having a combination of sandy loam soil and vermicompost in the ratio of 4:1 in the month of June 15th, 2020, and June 15th, 2021. After 45 days of nursery germination, transplanting was done into 4.5 x 3.5m plots in a randomized complete block design with 3 replications at 35 x 30 cm spacing in both years. Standard cultivation practices were followed to raise healthy populations in both years, i.e., 2020 and 2021.

2.3 Evaluation of agro-morphological data

Nine agro-morphological descriptors were measured at the reproductive stage [120-130 days after transplanting (DAT)]. Data on days to 50% flowering (DFF) and days to maturity (DM) was scored on a plot basis. However, in each replication, ten competitive plants of each accession were selected to measure other traits, i.e., plant height (PH) (cm), Number of nodes per plant (NNP), Number of secondary branches per plant (NSBP),

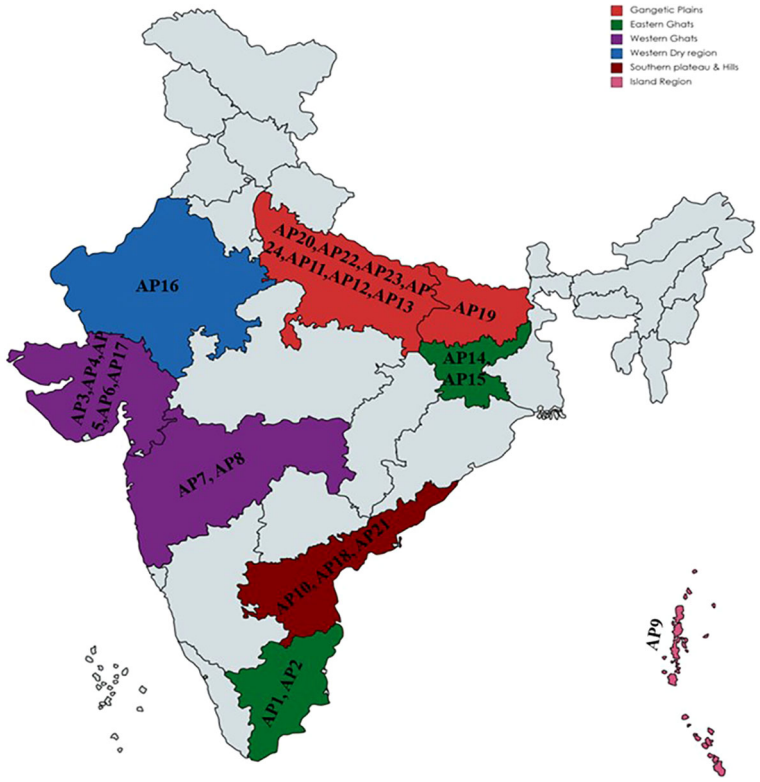


FIGURE 1
Geographical distribution of twenty-four kalmegh accessions used in the present study.

leaf length (LL) (cm), leaf width (LW) (cm), Inflorescence length (IL) (cm), and dry herbage yield per plant (DHY) (g)

2.4 Extraction of Ent-LRDs and high-performance liquid chromatography (HPLC) analysis

Aerial parts of the plant of each accession were picked at the maturity stage and were shade dried at room temperature for seven days. Dried samples were ground into a fine powder and stored under -20°C for a short period. About 100mg of powdered samples were extracted three times in 10ml of analytical grade methanol on a sonicator at 30mins intervals. All the filtrates were combined to obtain a total volume of 30ml and evaporated to dryness under a water bath. For HPLC analysis, the dried filtrate of each sample was dissolved in 1ml HPLC grade methanol (Merck, Germany). The extraction process for each accession was done using five biological replicates.

A 20ml solvent extract of individual accession was filtered into HPLC vials using disposable polypropylene syringe filters and injected into the HPLC-UV (Shimadzu LC-10A, Tokyo, Japan) system. The analysis was carried out as defined earlier by Tewari et al., 2010. The stock solutions of the standard compounds of AG, NAG, DDAG, and AN were also prepared in HPLC grade methanol at a concentration of 1mg/ml for standard curve preparation and quantifying of the experimental samples.

2.5 DNA extraction and genotyping with EST-SSR markers

The genomic DNA of each accession was isolated from 0.2g frozen leaves with a modified CTAB (Cetyl triethyl ammonium bromide) extraction method (Khanuja et al., 1998) (Supplementary Table S2). DNA quality was checked, and quantity was estimated using 0.8% agarose gel, and nanodrop (Thermo Fisher, USA) at 260/280 and 260/230 OD ratios, respectively. After that, a working concentration of 10ng/µL was made for each accession and kept at 4°C for further use.

The 53 non-redundant combined master control transcripts annotated and reported to be involved in specialized metabolic pathways in kalmegh variety "CIM-Megha" by Garg et al. (2015)

were rescanned for the presence of SSR motifs using the Batch Primer3 v1.0 tool (http://probes.pw.usda.gov/cgibin/batchprimer3/batchprimer3.cgi) developed by You et al. (2008). Similar repeat motifs were identified using this tool, as reported earlier by Garg et al. (2015). The in-built Primer3 core program of Batch Primer3 v1.0 was used to design primer pairs using flanking sequences. Out of 53, a total of 91 SSR primers were identified from 50 master control transcripts. To test the polymorphism of EST-SSR primers in kalmegh,32 primer pairs were randomly selected. The target amplicon size was set as 100-300bp, melting temperature (T^M) as 50-60°C, GC percentage as 40-60%, and optimal primer length between 18-25bp.

PCR reactions were carried out in a10μL volumes containing 5μL of dye mixed with One PCR^{TM} supermix (GeneDirex, Taiwan), 0.5μL of forward primer (5pmol), 0.5μL of reverse primer (5pmol), and 40ng of template DNA. PCR reactions were performed in a Thermal Cycler (Bio-Rad, California, USA) using the following cycling conditions: initial denaturation for 5 min at 94°C followed by 35cycles of denaturation for 1min at 94 °C, annealing for 1min at temperatures ranging between 50.5 °C to 57.8 °C depending upon the primer T^M and extension for 2mins at 72 °C followed by final extension for 7mins at 72 °C. PCR products were stored at 4 °C for gel electrophoresis. Metaphor agarose gel (2%) was used to separate amplified products using a power supply of 80V for 2.5-3h and visualized under Gel documentation System (UVP Bioimaging system, Analytik-Jena, Germany). DNA ladder of 100bp (Gene Direx, Taiwan) was used as standard.

2.6 Statistical analysis of agro-morphological and phytochemical data

All the quantitative data were statistically analyzed in software R-studio v4.0.3. The 'variability' package was used to perform an analysis of variance (ANOVA) on mean values of agro-morphological traits across replications (Popat et al., 2020). Pooled data of both years were used to analyze various genetic parameters such as Genotypic and Phenotypic coefficient of variation (Singh & Chaudhary, 1985), heritability, and genetic advance as percent of the mean (Johnson et al., 1955) using the same package in R-studio. The principal component analysis (PCA) for both types of traits was done in the 'pca3d' package of R software (Weiner and Weiner, 2016). To identify the divergence among different ecotypes, agglomerative hierarchical clustering methods were employed to create a tree diagram using 'ggplot2' package (Wickham et al., 2016). The K-mean methods of function 'factoextra' were applied to determine the number of k-groups to explain the agronomic and chemical variation among tested populations. Calculating pair-wise Euclidean distance and cluster analysis was performed using function 'NbClust' of package 'ggplot2'. Further, a heatmap was also prepared to visualize data more clearly using the function 'heatmap.2' in package 'gplots.'

2.7 Analysis of genotyping data

A binary data matrix was prepared to perform molecular data analysis by scoring the presence (1) or absence (0) of amplified bands produced after gel electrophoresis. The discriminating power of primers was assessed by measuring four parameters; PIC (Polymorphic Information Content), Rp (Resolving Power), MI (Marker Index), and EMI (Effective Marker Index). The PIC was calculated following Botstein et al. (1980) as: PIC=1-Σfi^2, where, 'fi' = frequency of the 'i^{th}' allele (band present). Similarly, the resolving power of each primer was measured following Prevost and Wilkinson (1999) method as: Rp= Σ Ib=1 − [2 × |0.5 − p|], where, Ib= band informativeness and p= number of individuals containing band. Further, MI provides a convenient estimate of marker utility and is calculated as: MI=EMR x PIC, where EMR (Effective multiplex ratio)= number of polymorphic band × fraction of polymorphic band (Milbourne et al., 1997; Prevost and Wilkinson, 1999). EMR determines the number of polymorphic loci analyzed per experiment in the germplasm set of interest. Varshney et al. (2007) provided an index called the Effective marker Index (EMI) to accelerate the practicability of the marker system to plant breeders. EMR is calculated as: EMR=MI x QND or, EMR= MI x DC x QM x PR, where QND= Qualitative nature of data, DC=Documentation capability, QM= Quality of Marker, and PR= Percent Reproducibility of the band/fragment of the marker. DC and PR represent the constant value and are set as 0.75 and 1.0 for SSR markers, respectively. However, the QM value varies with different primer pairs and is defined as per the scale (0.25 to 1.0) given by Varshney et al. (2007). Here, we also measured the EMI of all the polymorphic markers and took a scale of 1.0 as a QM value to estimate QND since all the amplified bands were single and clear.

The genetic similarity matrix and phenetic analysis of binary datasets were performed by software NTSYS v2.02e (Rohlf, 2000). Genetic relatedness among the twenty-four kalmegh accessions was estimated using the SIMQUAL module of Jaccard's similarity coefficient (Jaccard, 1908). The UPGMA (Un-weighted Pair Group Method with Arithmetic Mean) algorithm along with the SAHN (Sequential agglomerative hierarchical and nested clustering method) module of the same software was also used to compute a dendrogram demonstrating genetic association among all the accessions. Moreover, a Model-based population structure study was carried out to study the genetic association in the twenty-four accessions of kalmegh using polymorphic EST-SSR primers. STRUCTURE software version 2.3.4 (Pritchard et al., 2000) was used to perform this analysis. The analysis was performed

without incorporating the population information and considering both the admixture model and correlated allele frequencies between the populations. Here, accessions from same ecotypes were considered as single populations, thereby forming six populations. The K values were set to 1-10, and the software was run three times (r=3) for each K (number of populations). The number of Markov Chain Monte Carlo (MCMC) replications and burn-in-period was set to 100,000 for each run for all the twenty-four accessions to evaluate the number of populations. The plateau of the ΔK values was plotted using Ln(PD), which was derived for each K (Evanno et al., 2005). The online program "structure harvester" was used (http://taylor0.biology.ucla.edu) to compute the final number of K in population structure. The analysis of molecular variance (AMOVA) and Principal Coordinate Analysis (PCoA) was performed using GenAlEx6.501 software (Peakall and Smouse, 2012) to partition the genetic variation in studied kalmegh accessions. A Mantel test was also done to reveal the correlation between phytochemical and genotypic distance matrices using the same software. Finally, the dendextend R package was used to assess the correlation between two dendrograms generated for agro-phytochemical and molecular datasets (Tal, 2015). A joint cluster analysis was also executed by combining the distance matrices of all datasets generated in the present study using the R package (Garnier et al., 2018).

3 Results

The present study conducted with twenty-four accessions of *A. paniculata* were collected from thirteen states and belong to India's six ecotypes (agro ecological regions) (Figure 1). Out of twenty-four, 23 (AP1 to AP23) were germplasm accessions, and one was a cultivated variety, CIM-Megha (AP24) (Supplementary Table S1). To unlock the genetic variation present in *A.paniculata* accessions, three different tools, including phenotypic, phytochemical, and molecular markers (EST-SSRs) were used.

3.1 Agro-morphological diversity

The phenotypic diversity was assessed among experimental sets using nine agro-morphological (quantitative) traits. The analysis of variance (ANOVA) results are mentioned in Supplementary Table S3(A). The ANOVA results showed substantial variation among all the accessions for the maximum number of traits studied except leaf length. The results of the mean comparison and genetic variability parameters for all studied traits are mentioned in Supplementary Tables S3(B) and S3(C). For all the metric traits studied (Figure 2), the PCV (phenotypic coefficient of variation) was consistently greater than the GCV (genotypic coefficient of variation), and ranged from 2.03% (DM) to 31.92% (LL). However, the estimates of GCV varied from 1.74%(DM) to 15.33% (LW). The selection efficiency parameters, heritability (h^2_{bs}) and genetic advance (GA) ranged from 8.86% (LL) to 87.59% (DFF) and 0.37% (LW) to 12.05% (DFF), respectively. The highest heritability with moderate genetic advance was estimated for days to fifty percent flowering (DFF) trait. However, moderate heritability with moderate genetic advance was observed for dry herb yield (DHY).

Different clustering methods were used to execute phenetic analysis considering agro-morphological traits to gain reliable and precise estimates of genetic diversity present in the experimental sets. K-mean clustering, Euclidean distance-based

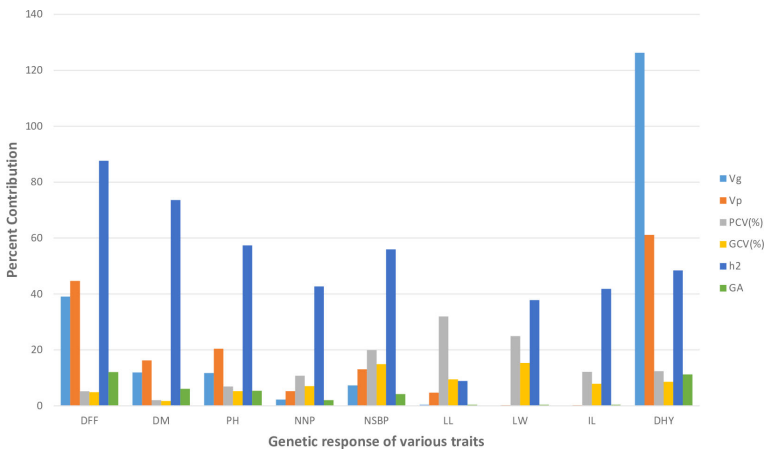


FIGURE 2
Graphical representation of genetic variability parameters estimated for nine quantitative traits in the present study.

agglomerative hierarchically clustered heatmap, and Eigen value-based Principal Component Analysis (PCA) divided all the studied germplasm into three clusters (Figures 3A–C). However, the grouping of individuals into different clusters was entirely different. The results of Euclidean distance-based clustering are shown in Figure 3B. Three accessions (AP6, AP2, and AP17) were included in Cluster I. These accessions were grouped since they had a similar range of dry herb yield and leaf length. Out of the three accessions, two were from ecotype Western Ghats (WG), and one was from ecotype Eastern Ghats (EG). In Cluster II, accession AP7 was grouped alone. This accession was superior to other accessions concerning dry herb yield and days to fifty percent flowering and belonged to the WG ecotype. Cluster III was observed as the largest group on the heat map and divided into 3 sub-groups III(a), III(b), and III(c). Sub-group III(a) covered eight accessions and was observed to be superior for plant height but not for dry herb yield. Among eight accessions, three (AP22, AP24, and AP 19) were associated with

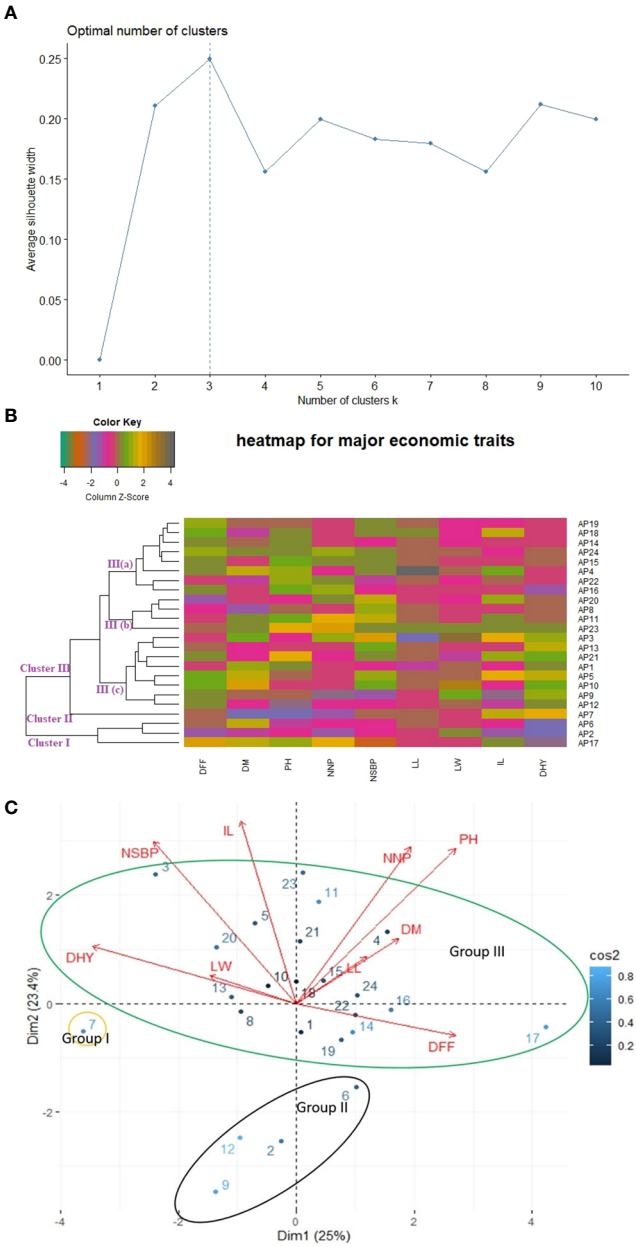


FIGURE 3
Relationship among twenty-four accessions of kalmegh based on agronomic traits using **(A)** k-mean **(B)** Euclidean distance **(C)** PCA biplot.

Gangetic Plains (GP), two (AP15 and AP14) to EG, one individually to WG (AP4), Western dry regions (WDR) (AP16) and Southern Plains and Hills (SP&H) (AP18). Likewise, Cluster III (b) grouped four accessions in which three (AP23, AP11, and AP20) were related to ecotype GP and one to ecotype WG (AP8). All the accessions of Cluster III(b) were superior to other accessions for the number of secondary branches per plant and found in a favorable position for dry herb yield. Cluster III(c) also grouped eight accessions, of which two belong to ecotype GP (AP12 & AP13), two separately to WG (AP3 and AP5), and SP&H ecotypes (AP10 and AP21), one individually to EG(AP1) and Island Region (IR) ecotype (AP9). All the accessions of Cluster III (c) were found superior to other accessions for dry herb yield except accession AP7, which showed the highest dry herb yield compared to all other accessions.

PCA clustered accessions into different groups based on their eigenvalues. The PCA results showed that the first two principal components (PC1 and PC2) collectively elucidated 48.40% of the total variation, as mentioned in Supplementary Table S4(A). Based on results shown in Figure 3C, the first area of the biplot, which included positive values of both components, covered accessions AP23, AP11, AP21, AP18, AP15, AP4, and AP24 and associated with traits like the number of nodes per plant, days to maturity, plant height, and leaf length. Likewise, the second area of the biplot, which comprised positive values of the second component and negative values of the first component, positioned accessions AP1, AP6, AP22, AP19, AP14, AP16, and AP17 on the biplot and associated with days to fifty percent flowering trait. The third area of the biplot, which has negative values for both components, covered five accessions (AP2, AP7, AP8, AP9, and AP12), and no significant association with studied traits was observed. Lastly, in the fourth area of the biplot, which contained a positive value of the PC1 and a negative value of the PC2, accessions AP3, AP5, AP10, AP13, and AP20 were located and linked with dry herb yield, number of secondary branches/plant, leaf width, and inflorescence length. Overall, all accessions were classified into three clear-cut groups on the biplot display and clustered with four accessions in group I, one accession in group II and nineteen accessions in group III, somewhat similar to cluster analysis.

3.2 Phytochemical diversity

A phytochemical dataset of twenty-four accessions was also used to estimate the level of variability present among them (Supplementary Table S4B). K-mean clustering divided the whole accessions into two clusters, confirmed further by agglomerative hierarchical clustering based heatmap and PCA (Figures 4A–C). The one-way heatmap clustered the whole germplasm into two main groups (Figure 4B): Cluster I covered nine accessions, and Cluster II comprised fifteen accessions. Cluster I was divided further into two sub-clusters-Sub-cluster I(a) and Sub-cluster I(b). The accession AP23 was grouped alone in Sub-cluster I(a) and found in a favorable position for 14-deoxy-11,12-didehydro-andrographolide (DDAG) content. Sub-cluster I(b) consisted of eight accessions (AP14, AP8, AP9, AP15, AP9, AP22, AP20, and AP13) and showed no significant association with any phytochemical trait. Likewise, Cluster II was further divided into two sub-clusters-Sub-cluster II(a) and Sub-cluster II(b) and showed superior association with andrographolide content (AG) compared to other studied accessions. Sub-cluster II (a) grouped six accessions (AP2, AP4, AP5, AP6, AP10, and AP24) and Sub-cluster II (b) covered nine accessions (AP1, AP3, AP7, AP11, AP12, AP16, AP17, AP18, and AP21).

In PCA, the first two principal components (PC1 and PC2) collectively described 73.28% of the total variability, as shown in Supplementary Table S4(C). As per the PCA biplot (Figure 4C), the first area of the biplot covered two accessions (AP2 and AP12) and showed an association with DDAG and neoandrographolide (NAG) content. However, eight accessions (AP1, AP3, AP7, AP10, AP11, AP16, AP18, and AP21) were grouped in the second area of the biplot and encompassed Andrographolide (AG) and Andrographanin (AN) content. The third and fourth areas of the biplot included ten (AP4, AP6, AP13, AP14, AP15, AP17, AP19, AP20, AP22, and AP24), and three (AP5, AP8, AP9) accessions, respectively, and showed no significant association with the studied phytochemical trait. Overall, on the PCA biplot, two groups were visible, encompassing maximum accessions in group I and single accession (AP2) in group II.

3.3 Molecular diversity

3.3.1 SSR primer designing

In total, 91 primer pairs were designed from 50 unique transcripts (ESTs), 23 (46%) of which comprised more than one SSR loci, as shown in Supplementary Table S5(A). Of these 91 EST-SSR primers, 32 random primer pairs were finally selected for validation and genetic variation study in *A. paniculata*. Of these 32 primer pairs, 23 could amplify unambiguous bands, and thirteen showed polymorphic and reproducible bands. Supplementary Table S5(B) shows detailed information on 23 primer pairs and their probable gene functions.

3.3.2 SSR analysis

Finally, 13 polymorphic EST-SSR primer pairs were utilized to fingerprint twenty-four accessions of *A.paniculata*. An average of two alleles/primer pairs were spotted, with a total of 26 alleles at 13 marker loci. The percentage polymorphism across all the primer pairs was 100%. Three representing

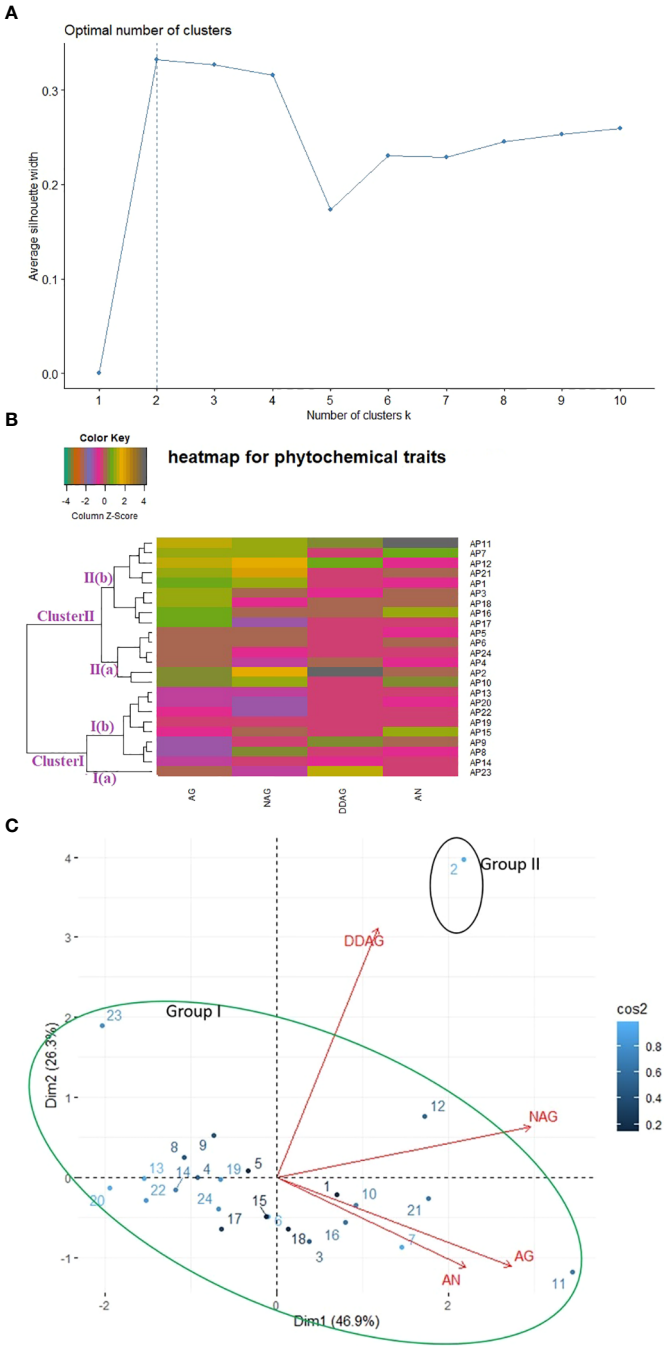


FIGURE 4
Relationship among twenty-four accessions of kalmegh based on phytochemical traits using **(A)** k-mean **(B)** Euclidean distance **(C)** PCA biplot.

profiles [primer pair ID APSSR 6, APSSR 16 & APSSR 2] are displayed in Supplementary Image 1(A–C). The estimates of various genetic parameters representing the discriminating power of polymorphic SSR primers are shown in Table 1. The PIC value varied from 0.149 to 0.465, averaging 0.322 per primer. The resolving power (Rp) ranged from 0.667 to 2.917, with an average value of 1.802/primer. The marker index (MI) and effective marker index (EMI) values also varied from 0.299 to 0.899 and 0.224 to 0.674 per primer, averaging 0.644 and 0.483/primer, respectively.

TABLE 1 Genetic parameters of thirteen polymorphic EST-SSRs used in the study.

S.No.	Primer name	TNB	NPB	PP (%)	PIC	RP	EMR	MI	QM	QND	EMI
1.	APSSR 1	2	2	100	0.326	1.833	2	0.653	1	0.75	0.490
2.	APSSR3	2	2	100	0.345	1.750	2	0.691	1	0.75	0.518
3.	APSSR5	2	2	100	0.283	1.583	2	0.566	1	0.75	0.424
4.	APSSR6	2	2	100	0.450	1.750	2	0.899	1	0.75	0.674
5.	APSSR7	2	2	100	0.149	1.833	2	0.299	1	0.75	0.224
6.	APSSR12	2	2	100	0.465	1.500	2	0.931	1	0.75	0.698
7.	APSSR14	2	2	100	0.299	1.500	2	0.597	1	0.75	0.448
8.	APSSR16	2	2	100	0.352	2.917	2	0.705	1	0.75	0.529
9.	APSSR17	2	2	100	0.387	1.417	2	0.774	1	0.75	0.581
10.	APSSR18	2	2	100	0.247	1.500	2	0.493	1	0.75	0.370
11.	APSSR19	2	2	100	0.247	0.667	2	0.493	1	0.75	0.370
12.	APSSR29	2	2	100	0.274	1.833	2	0.549	1	0.75	0.411
13.	APSSR33	2	2	100	0.361	1.667	2	0.722	1	0.75	0.542
	Average	2	2	100	0.322	1.673	2	0.644	1	0.75	0.483

TNB, NPB, PP, PIC, RP, EMR, MI, QM, QND and EMI refer total number of bands, number of polymorphic bands, percent polymorphism, polymorphic information content, resolving power, effective multiplex ratio, marker index, quality of marker, qualitative nature of data and effective marker index, respectively.

3.3.3. Genetic diversity and relationships study

Genetic diversity was studied using binary data matrices produced by thirteen polymorphic marker loci. The pair-wise genetic similarity coefficient showed 100% genetic similarity among accessions AP4, AP17, and AP18; between AP5 & AP7; among accessions AP8, AP9, AP10 & AP12; between AP6 & AP11, and between AP13 & AP14. However, the minimum genetic similarity (40%) was observed between accession AP1 & AP10 (Supplementary Table S6).

We also constructed a UPGMA tree using the corresponding genetic similarity coefficient among the studied accessions (Supplementary Image 2). The UPGMA-based dendrogram grouped twenty-four accessions into two distinct clusters, wherein three accessions (AP1, AP2, and AP3) were grouped in Cluster I and twenty-one in Cluster II. Further grouping was observed in cluster II with two sub-clusters- Sub-cluster II (a) and Sub-cluster II (b). Sub-cluster II (a) encompassed three accessions (AP13, AP14, AP22), and cluster II (b) included eighteen accessions (AP4, AP17, AP18, AP23, AP5, AP7, AP8, AP9, AP10, AP12, AP6, AP11, AP15, AP24, AP16, AP19, AP20, and AP21) of different ecotypes.

3.3.4 Analysis of molecular variance (AMOVA)

The AMOVA was used to calculate the variability across and within the *A. paniculata* accessions procured from various agro ecological regions of India. Six populations were considered in the current study depending on agro ecological zones (Figure 1). The SSR data showed 7% variation among six populations and 93% variation within a population, as shown in Figure 5A. Significant genetic variation was found among and within the accessions of kalmegh (Supplementary Table S7A).

3.3.5 Principal coordinate analysis (PCoA)

The PCoA was also computed to get an alternative view of phylogenetic relationships among the twenty-four accessions of *A. paniculata*. The cumulative percentage of variation elucidated by the first three coordinates was 56.05%, with PCo1 contributing 28.66%, PCo2 contributing 16.18%, and PCo3 contributing 11.21%, respectively (Supplementary Table S7B). The grouping of accessions on the PCoA biplot was not in accordance with cluster analysis. However, the accession AP18 from Southern Plateau and Hills was distinct and grouped alone on the biplot display (Figure 5B).

3.3.6 Population structure-based study

Genetic structure of the studied germplasms was also evaluated using Evanno's method based STRUCTURE software. The total number of genetic populations (k) indicated a clear peak at two with an optimum delta k value, indicating the distribution of two populations across all the studied accessions (Figure 6A). As shown in Figure 6B, the population I comprised three pure accessions and one admixed accession. However, population II displayed sixteen pure accessions and four admixed accessions. Grouping of studied accessions into different populations was ecotype independent.

3.4 Joint analysis of agro-phytochemical and genotypic data

The phenotypic and genotypic information-based distance matrices were used to generate separate hierarchical clusters that

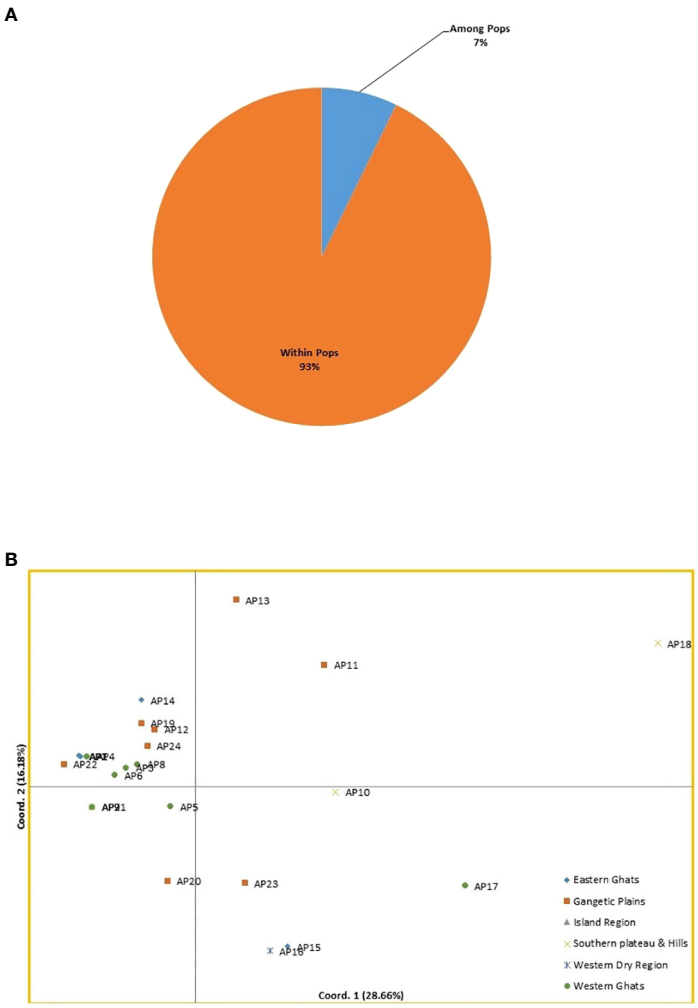


FIGURE 5
(A) Analysis of molecular variance (AMOVA) of twenty-four kalmegh accessions based on EST-SSR primers. **(B)** Two-dimensional distribution of twenty-four kalmegh accessions on PCA biplot.

were finally compared with each other. We did not find any individual to be clustered in the same position across the two phylogenetic trees (Figure 7A). Also, the Mantel test performed between phytochemical and genotypic datasets showed no significant correlation (R^2 = 0.0071, P>0.05) besides forming two clear-cut clusters with each dataset (Figure 7B). The combined clustering analysis of agro-phytochemical and genotyping data revealed two well-defined clusters (Cluster I and Cluster II) in the current set of materials (Figure 7C). A total of twenty-two and two accessions were grouped in Cluster I and ClusterII, respectively. The Cluster I was further subgrouped with five accessions in sub-cluster I(a), sixteen accessions in sub-cluster I(b), and one accession in sub-cluster I(c). No geographical regions or trait-specific grouping of individuals was observed in different clusters and sub-clusters.

4 Discussion

A. paniculata has gained much attention as the sole plant producing Andrographolide in the last several decades. Commercial cultivars available in kalmegh are ecotype specific and minimal in number to meet global demand. The increased demand for high-quality raw materials warrants the exploration of its genetic potential from wild sources to develop an operative breeding program in kalmegh, following suitable selection techniques. Also, assessing genetic variation present in the indigenous genetic resources may provide the foundation for effective selection response and genetic gain in kalmegh breeding. Moreover, evaluating the genetic mechanism of different indigenous ecotypes of *A. paniculata* are central to its sustainable cultivation. Considering the lack of enough

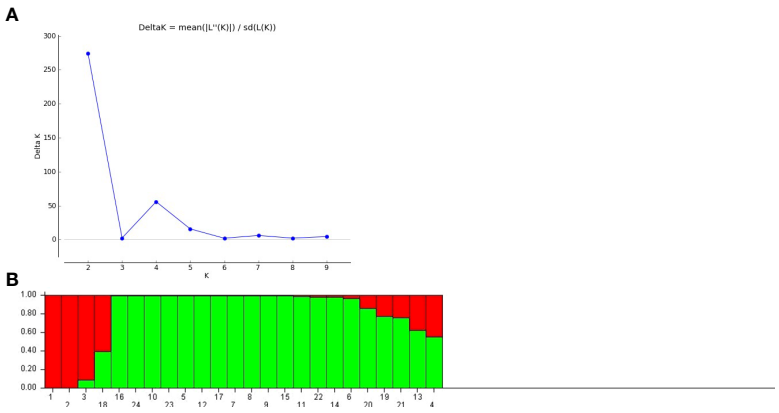


FIGURE 6
(A) Estimation of population using LnP(D) derived delta k value **(B)** Population structure at K= 2 by Evanno table in *A.paniculata* accessions.

comprehensive information on the genetic diversity in Indian kalmegh ecotypes, the present study exploited both agro-morphological and phytochemical systems to evaluate the genetic diversity in *A. paniculata* ecotypes since the variation in the agro-phytochemical traits in different plant populations may be primarily caused by genetic variation and interaction of environmental conditions. In the present study, specialized metabolic pathway-specific genic SSRs (EST-SSRs) were introduced for the first time to determine the genetic variability among kalmegh accessions and to analyze the inter-relationship between agro-phytochemical traits and EST-SSR markers to speed-up trait-specific breeding in kalmegh.

By analyzing ANOVA for nine agro-morphological traits of studied accessions, we demonstrated that all the accessions have sufficient phenotypic diversity for all traits except leaf length (LL), which could provide the basis of selection in the kalmegh breeding program. A broad range of variation was seen in kalmegh accessions against dry herb yield, plant height, number of secondary branches per plant, days to fifty percent flowering, and andrographolide content, suggesting the diverse genetic makeup of these accessions. These promising accessions from different ecological regions may be valuable resources for augmenting the genetic gains of cultivated varieties (Supplementary Table S3B). Similar results were shown by Latto et al. (2008) in *A. paniculata* for morphometric traits.

Various genetic variability parameters were analyzed to determine the value of genetic diversity among experimental sets. The PCV was consistently higher than the GCV, showing the role of environment in the trait expression. However, the difference between PCV and GCV values was estimated to be narrow (<10%) for all traits studied except leaf length (LL), showing significant genetic control and little environmental influence on trait expression (Supplementary Table S3C, Figure 2). The moderate GCV was observed for leaf width (LW), and the number of secondary branches/plant (NSBP), enhancing the scope of selection due to strict genetic control on the expression of these traits. Heritability estimates predict the breeding value and strengthen the reliability of the phenotypic value of traits in any crop improvement program (Kumar et al., 2014). Simple selection can quickly improve traits with high heritability. Heritability, on the other hand, has proven to be useless without the association of genetic advancement. Heritability estimations and genetic advance (GA) together predict how well selection will work by choosing better genotypes. High heritability with high GA describes the involvement of additive gene action in quantitative traits expression, whereas low GA with high heritability defines the influence of non-additive gene action in the manifestation of quantitative traits. As revealed in Supplementary Table S3(C) & Figure 2, none of the traits showed high estimates of GA and heritability. However, moderate estimates of GA and high to moderate estimates of heritability were calculated for traits, namely days to fifty percent flowering and dry herb yield, anticipating the efficacy of direct selection in the studied accessions for these traits. Little efforts have been undertaken to understand the role of genetic factors in manifesting quantitative traits in kalmegh. In 2000, Mishra et al. (2000) observed high GCV, heritability, and GA for dry herb yield and plant height in 22 morphologically diverse accessions of kalmegh. Thus, in the present study, dry herb yield and the number of secondary branches per plant could be appropriate selection indexes for parent selection in the hybridization program of kalmegh accessions.

Different multivariate clustering methods were used to classify twenty-four kalmegh accessions based on agro-morphological and phytochemical traits. The silhouette algorithm-based k-mean clustering, Euclidean distance-based clustering, and PCA divided the experimental set into three groups based on agro-morphological attributes (Figures 3A–C). Among the nine quantitative traits, the main distinguishing trait

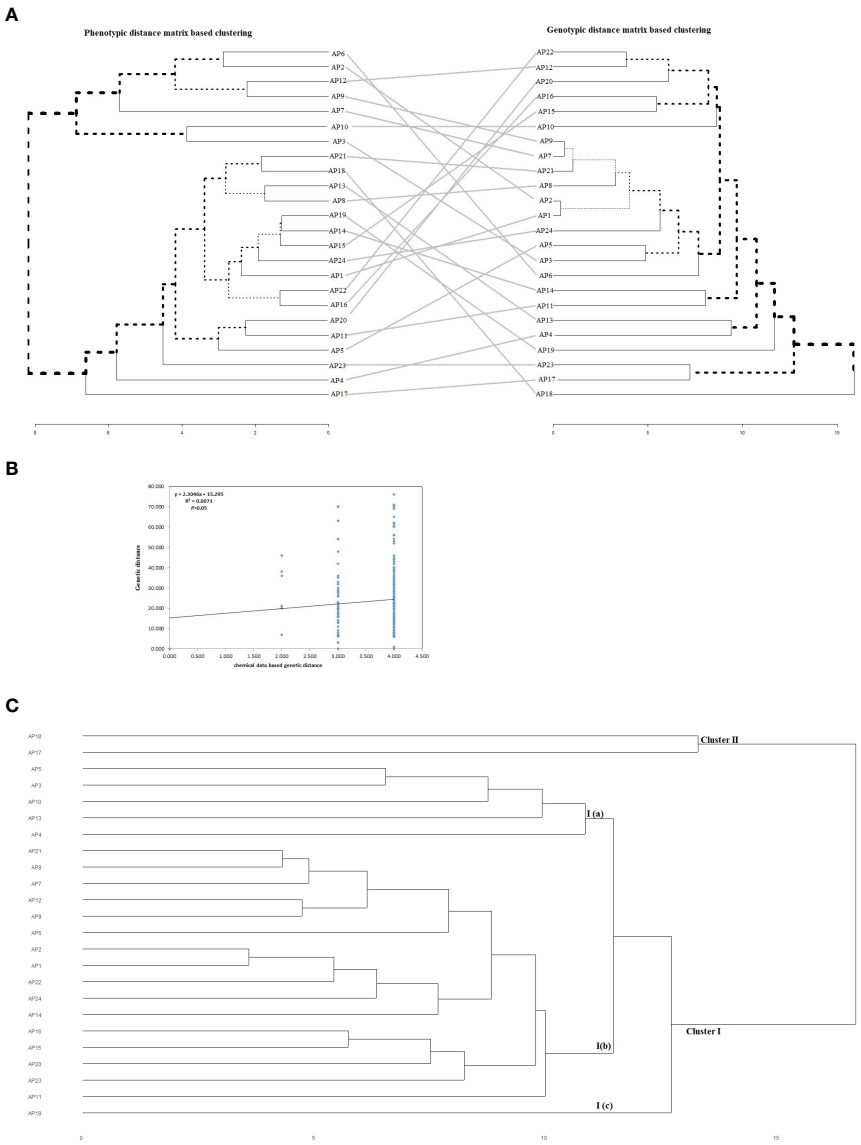


FIGURE 7
(A) Comparative assessment of phenotypic and genotypic data based hierarchical cluster dendrograms. The black lines represent mismatched accessions in-between the two dendrograms from the phenotypic to the genotypic cluster. **(B)** Mantel test representing the relationship between phytochemical and molecular data sets. **(C)** Hierarchical clustering of twenty-four accession of kalmegh using combined agro-phytochemical and molecular datasets.

that contributed to diversity among different groups in cluster analysis was the proportion of dry herb yield. Overall, the grouping pattern was slightly similar in both heatmap clustering and PCA biplot showing considerable diversity among experimental sets, which might be attributed to gene ontology, soil, and environmental conditions. The hybridization between the accessions with the least genetic similarities could be effective for the production of superior genotypes containing desirable traits. Likewise, all the three clustering methods (k-mean clustering, Euclidean distance-based clustering, and PCA) divided twenty-four accessions into two clear-cut groups based on phytochemical traits (Figures 4A–C), demonstrating considerable variance among experimental sets based on chemical concentrations. The distinctive characteristic contributing to diversity among different groups was the share of andrographolide content. Thus, despite growing in the same soil and temperature regime, principal component analysis and cluster analysis allowed the classification of *A.paniculata* accessions into different groups and revealed significant variability in chemical concentrations among experimental

sets, which appears to be associated with genetic factors accumulated during selection.

Moreover, our clustering results based on agro-morphological traits and phytochemical traits showed no specific grouping pattern of geographically closer accessions of *A. paniculata*. The propensity to produce such a clustering pattern suggests that geographic segregation might not always dilute the genetic makeup of introductions which leads to diversity in natural populations. In light of this, it seems that accessions' genetic makeup, as opposed to their eco-geographic origin, significantly influences clustering. This may result from the unrestricted flow of seed from the native place to the area of their domestication. Our results align with the reports of Lattoo et al. (2008) for morphometric traits and Hiremath et al. (2020) for morphometric and chemotypic traits in *A.paniculata.*

Characterization of diversity based on morpho-metric traits may not be reliable as these are vulnerable to ontology and environmental factors. Therefore, it is desirable to assess diversity based on molecular markers and compare them with morphometric traits to attain more realistic results. According to Varshney et al. (2005), molecular markers have become a frequent and crucial tool for assessing genetic relationships and diversity, cultivar identification and development, efficient gene mapping, tagging, and early generation detection of superior genotypes in many crops. In recent years, microsatellites (SSRs) have become a marker of choice due to their high abundance, hyper-variable, reproducible, co-dominance, and discriminatory nature. Simple Sequence Repeats (SSRs) in the transcribed regions are perceived to be more conserved, significant, and transferable across taxonomic borders than the anonymous SSRs (Pashley et al., 2006; Ellis and Burke, 2007). Transcriptome sequencing is nowadays a quick and affordable way to obtain the EST (Expressed sequence tags) sequences required to isolate a massive set of functional SSRs associated with novel genes. Many EST-SSRs have been developed recently in diverse plant species using transcriptome sequences (Wu et al., 2014). In *A. paniculata*, this is the first established report where transcriptome-based EST-SSR markers were developed and validated. Thus, in the present study, we aimed to generate new EST-SSR markers from the already available transcriptome data of kalmegh in our lab through Illumina paired-end RNA-seq technology (published by Garg et al., 2015). The non-redundant combined transcript extracted from the leaf and root transcriptome and annotated to be involved in different metabolic pathways were used for marker development. The markers obtained from these sequences would be more beneficial than genomic sequence data for trait-specific breeding and detection in *A. paniculata.*

In the present study, thirty-two SSRs primer pairs discovered *via* 50 non-redundant transcripts were selected randomly for experimental validation to build a working maker set for kalmegh genetic improvement. Of these 32 primer pairs, over 71.87% (23primer pairs) successfully amplified genomic DNA and over 40% (13primer pairs) produced polymorphic and reproducible bands across twenty-four accessions (Supplementary Table S5A &B). Our success rate is comparable to other efforts done in medicinal herbs, where EST-SSRs amplification rate of 60-80% was reported (Gong & Deng, 2010; Sathyanarayana et al., 2017; Vidya et al., 2021). Primer development across intron/exon splice sites, alternate splice sites, or chimeric transcripts could cause marker dropout in genic-SSRs. In the current study, 13 polymorphic EST-SSRs were finally selected to evaluate genetic relationships and diversity among studied accessions. The discriminatory power of these primer pairs was analyzed using different parameters (Table 1). The PIC value is assessed by considering both the allelic numbers as well as their frequency distribution across the experimental set. It quantifies the polymorphism for a marker locus (Guo and Elston, 1999). All the kalmegh SSRs validated in the current study showed a moderate PIC value (<0.5), with a mean of 0.322. Reports around the globe suggested that the EST-SSR primers show less polymorphism than genomic SSRs in crop plants due to larger sequence conservation in transcribed regions (Varshney et al., 2005). Our result of PIC was lower than *Coriandrum sativum* (0.38) (Tulsani et al., 2019) *Docynia delavayi* (0.587) (Peng et al., 2021), *Ginkgo biloba* (0.781) (Zhou et al., 2019), and *Quercus petraea* (0.787) (Lupini et al., 2019) identified with EST-SSR markers. The moderate estimates of PIC might be due to the development of SSRs from the metabolic pathway-specific, highly conserved transcribed region of kalmegh. The PIC estimated in our study was comparatively greater than that of the plant species *Rhododendron arboretum* where PIC was reported to be low (0.195) (Sharma et al., 2020). This indicates that the *A. paniculata* loci examined here had a significant level of discernment, reflecting the complexity of genetic diversity and structure (Peng et al., 2021). The Rp, MI and EMI of 13 EST-SSRs were also estimated, indicating the high efficiency of these primers in kalmegh genetic diversity assessment. Varshney et al. (2007) also analyzed the discriminating nature of EST-SSR primers during the evaluation of different species and cultivars of barley (*Hordeum vulgare*). Sahoo et al. (2021) also reported the average MI of 4.84 in Indian Curcuma species using EST-SSR markers which is significantly higher than our study.

The cluster analysis and population structure analysis placed all the twenty-four accession of *A.paniculata* into two groups showing the presence of reasonable variability among them that could be potential sources for selecting parents for breeding purposes. In UPGMA-based clustering (Supplementary Image 2), thirteen accessions of five different sub-groups were found to be genetically identical, showing the inability of EST-SSR primers to differentiate them at the genetic level. This might be due to limited sampling and similar topography of the regions represented by these accessions, as seven of these thirteen accessions belong to the western regions of the country. However, this could also be due to the low genetic variability present among accessions of represented regions (Shiferaw et al., 2012). Bayesian model-based population structuring considered individuals with a probability score of >0.80 as genetically pure

and a score of<0.80 as admixed type (Figure 6B). Mixing of pure individuals with few admixed accessions was observed in both the populations derived from model-based study, which could be due to the breeding behavior of the studied plant (Kumar et al., 2020). AMOVA and PCoA also explained substantial genetic diversity among the studied accessions of six agroecologically grouped populations. However, ANOVA based genetic differentiation showed maximum variation within agroecological regions rather than between agroecological regions indicating frequent gene flow through seed or out-crossing across different agroecological populations (Tiwari et al., 2016) (Figure 5A). Positioning of studied accessions on PCoA biplot was not in congruence with Cluster and STRUCTURE analysis (Figure 5B). The accession AP18 from Southern Plateau and Hills was very distinct on the biplot otherwise no specific grouping pattern was observed in PCoA analysis. Overall, different clustering methods used in the present study could not able to classify studied accession with their geographical distribution. The low genetic differentiation among different (six) agroecological populations could be interpreted as genetic drift due to seed dispersal, human intervention, or cross-pollination. Seed dispersal or seed exchange may result in an increase in allelic diversity among diverse populations regardless of their geographical isolation causing enhanced genetic diversity in local germplasm (Louette et al., 1997). Also, the reproductive biology of the plant might have contributed to the distribution of alleles across the regionally isolated population. Although the anthecology of *A. paniculata* favors self-pollination, there are records of substantial outcrossing (around 4%) through insect-pollination. (Shiferaw et al., 2012; Tiwari et al., 2016). Poor sampling size of different agroecological populations could be another reason for the low genetic diversity among different populations. Thus, more detailed studies with larger sampling sizes from extended geographical regions could draw a concrete inference (Sahoo et al., 2021). Previously, evaluation of genetic diversity among the geographically isolated germplasm of *A.paniculata* was carried out using various dominant molecular markers (Lattoo et al., 2008; Minz et al., 2013; Tiwari et al., 2016; Kumar et al., 2020; Hiremath et al., 2020) the outcomes of which are in line with our results. However, in the present study, metabolic pathway specific EST-SSR markers were designed and used to determine the genetic diversity among twenty-four ecotypes of *A. paniculata*. These noval SSR loci displayed relatively high polymorphism levels and could be an important tool for investigating genetic diversity and assessing effective strategies for selective breeding and conservation in *A. paniculata*.

Further, the inconsistency observed between hierarchical clusters identified by phenotypic and genotypic distance matrices could be due to the negligible correlation observed between them and enormous genotype x environment (GxE) interaction effects observed for quantitatively inherited agro-phytochemical traits (Figure 7A). This observation was also supported by the Mantel test drawn between the phytochemical and genotypic distance matrix (Figure 7B). The negligible correlation could also be due to the non-adaptive nature of variation created by EST-SSR markers, unlike quantitative agronomic or phytochemical traits (Singh et al., 1991). Similar results showing the discrepancy between phenotypic and genotypic datasets were reported by several workers in different crops (Hartings et al., 2008; Soriano et al., 2016; Agre et al., 2019; Darkwa et al., 2020). Therefore, an approach using combined datasets of genotypic and phenotypic information to capture entire genetic variability present in the plant populations and assess genetic diversity was suggested by Alves et al. (2013) and da Silva et al. (2017). The joint cluster analysis performed with agro-phytochemical and genotypic datasets generated two genetic groups in the experimental sets of kalmegh with regrouping of individuals in different clusters, unlike separate hierarchical clustering computed by different datasets in the present study (Figure 7C). The genetic diversity assessed by joint cluster analysis could have significant implications for *A. paniculata* genetic improvement. The genetic groups identified in different clusters could be better utilized as trait progenitors in the different selection and hybridization programs, thereby enlarging the genetic base of the kalmegh breeding program and maximizing genetic gain. Our results are in line with the finding of Agre et al. (2019); Darkwa et al. (2020), and Alves et al. (2013), who unlocked genetic diversity in their studies using phenotypic, molecular, and combined datasets.

5 Conclusion

Adequate genetic diversity is crucial to project appropriate breeding programs and develop improved varieties in kalmegh genetic improvement. In the present study, genic EST-SSR markers along with agronomic and phytochemical traits, were used to assess genetic diversity among twenty-four kalmegh accessions collected from diverse agroecological zones of India. Our results on genetic diversity revealed sufficient genetic variation in the studied population, which can be exploited for kalmegh genetic improvement and germplasm conservation. The agro-morphological descriptors, days to fifty percent flowering and dry herb yield identified as potential selection index in the present study could be utilized further in kalmegh genetic improvement program. The low genetic differentiation observed among the different agroecological populations could be improved by increasing the sample size from extended geographical regions. In our study, the inconsistency observed between genotypic and phenotypic information could be resolved by enhancing genome-wide information with more number of functional EST-SSR markers to obtain concrete outcomes from them. Our results on combined datasets

expanded the scope of selective breeding in kalmegh by utilizing different trait-specific parental lines grouped in different genetic clusters generated by phenotypic and genotypic information. However, further research on economic traits using more genetic and genomic resources can complement the current study and generate more reliable information on Indian kalmegh ecotypes.

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

GT: Conceived and designed the whole research. Designed EST-SSR markers and performed all statistical analyses. Wrote the manuscript; TC: Collected the phenotype data and performed molecular genotyping; AG: Provided experimental materials from the gene bank; KS: Performed chemical analysis; BK: Supported in lab activity required to perform whole research. All authors contributed to the article and approved the submitted version.

Acknowledgments

We are grateful to the Late Dr. Hari Om Misra, Ex-Senior Principal Scientist, Division of Plant Breeding and Genetic Resource Conservation, Council of Scientific and Industrial Research-Central Institute of Medicinal and Aromatic Plants (CIMAP), Lucknow who collected the kalmegh germplasm from different parts of India. The authors are also thankful to Dr. Sumit Ghosh, Senior Principal Scientist, Division of Plant Biotechnology, CSIR-CIMAP for providing combined transcript sequences of AP variety 'CIM-Megha' and giving guidance for SSR marker development. The institutional communication number of the publication is CIMAP/PUB/2022/99.

Conflict of interest

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

Publisher's note

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.1042222/full#supplementary-material

References

Agre, P., Asibe, F., Darkwa, K., Edemodu, A., Bauchet, G., Asiedu, R., et al. (2019). Phenotypic and molecular assessment of genetic structure and diversity in a panel of winged yam *(Dioscorea alata)* clones and cultivars. *Sci. Rep.* 9, 1–11. doi: 10.1038/s41598-019-54761-3

Akbar, S. (2011). *Andrographis paniculata*: a review of pharmacological activities and clinical effects. *Altern. Med. Rev.* 16, 66–77.

Alves, A. A., Bhering, L. L., Rosado, T. B., Laviola, B. G., Formighieri, E. F., and Cruz, C. D. (2013). Joint analysis of phenotypic and molecular diversity provides new insights on the genetic variability of the Brazilian physic nut germplasm bank. *Genet. Mol. Biol.* 36, 371–381. doi: 10.1590/S1415-47572013005000033

Archana, P. R., Sivaraj, N., and Kumar, A. (2016). Chemical diversity among *Andrographis paniculata* nees (Kalmegh) and assessing climate suitable regions for elite germplasm distribution in India. *Medicinal Plants.* 8, 294–302. doi: 10.5958/0975-6892.2016.00036.8

Boopathi, C. A. (2000). *Andrographis spp,*: a source of bitter compounds for medicinal use. *Anc Sci. Life.* 19, 164–168.

Botstein, D., White, R. L., Skolnick, M., and Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* 32, 314–331.

Chauhan, E. S., Sharma, K., and Bist, R. (2019). *Andrographis paniculata*: A review of its phytochemistry and pharmacological activities. *Res. J. Pharm. Tech.* 12, 891–900. doi: 10.5958/0974-360X.2019.00153.7

Chopra, R. N. (1956). *Glossary of Indian medicinal plants* (agris.fao.org).

Darkwa, K., Agre, P., Olasanmi, B., Iseki, K., Matsumoto, R., Powell, A., et al. (2020). Comparative assessment of genetic diversity matrices and clustering methods in white Guinea yam (*Dioscorea rotundata*) based on morphological and molecular markers. *Sci. Rep.* 10, 1–14. doi: 10.1038/s41598-020-69925-9

da Silva, M. J., Pastina, M. M., de Souza, V. F., Schaffert, R. E., Carneiro, P. C. S., Noda, R. W., et al. (2017). Phenotypic and molecular characterization of sweet sorghum accessions for bioenergyproduction. *PloSone* 12, e0183504. doi: 10.1371/journal.pone.0183504

Ellis, J. R., and Burke, J. M. (2007). EST-SSRs as a resource for population genetic analyses. *Heredity* 99, 125–132. doi: 10.1038/sj.hdy.6801001

Enmozhi, S. K., Raja, K., Sebastine, I., and Joseph, J. (2020). Andrographolide as a potential inhibitor of SARS-CoV-2 main protease: an *in silico* approach. *J. Biomol. Struct. Dyn.* 39, 3092–3098. doi: 10.1080/07391102.2020.1760136

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

e-Charak (2008) *National Medicinal Plants Board, Ministry of AYUSH, Government of India*. Available at: https://echarak.in/echarak/marketprice.do

Garg, A., Agrawal, L., Misra, R. C., Sharma, S., and Ghosh, S. (2015). *Andrographis paniculata* transcriptome provides molecular insights into tissue-

specific accumulation of medicinal diterpenes. *BMC Genomics* 16, 1–16. doi: 10.1186/s12864-015-1864-y

Garnier, S., Ross, N., Rudis, B., Sciaini, M., and Scherer, C. (2018). *Viridis: default color maps from'matplotlib'. R package version 051. CRAN: the comprehensive R archive network*.

Ghosh, B. K., Mandal, A., Datta, A. K., and Das, D. (2014). RAPD analysis in *Andrographis paniculata* (Burm. f.) ness plant types. *Int. J. Res. Ayurveda Pharm.* 5, 84–88. doi: 10.7897/2277-4343.05117

Gong, L., and Deng, Z. (2010). EST-SSR markers for gerbera (*Gerbera hybrida*). *Mol. Breed.* 26, 125–132. doi: 10.1007/s11032-009-9380-x

Guo, X., and Elston, R. (1999). Linkage information content of polymorphic genetic markers. *Hum.Hered* 49, 112–118. doi: 10.1159/000022855

Hartings, H., Berardo, N., Mazzinelli, G. F., Valoti, P., Verderio, A., and Motto, M. (2008). Assessment of genetic diversity and relationships among maize (*Zea mays l.*) Italian landraces by morphological traits and AFLP profiling. *Theor. Appl. Genet.* 117, 831–842. doi: 10.1007/s00122-008-0823-2

Hiremath, C., Greeshma, M., Gupta, N., Kuppusamy, B., Shanker, K., and Sundaresan, V. (2020). Morphometric, chemotypic, and molecular diversity studies in *Andrographis paniculata*. *J. Herbs Spices Med. Plants*. 27, 109–122. doi: 10.1080/10496475.2020.1787290

Jaccard, P. (1908). Nouvelles researches sur la distribution florale. *Bull. Soc Vaud. Sci. Nat.* 44, 223–270. doi: 10.1270/jsbbs.56.107

Johnson, H. W., Robinson, H. F., and Comstock, R. E. (1955). Estimates of genetic and environmental variability in soybeans. *Agron. J.* 47, 314–318. doi: 10.2134/agronj1955.00021962004700070009x

Khanuja, S. P. S., Shasany, A. K., Dhawan, S., and Kumar, S. (1998). Rapid procedure for isolating somaclones of altered genotypes in *Mentha arvensis*. *J. Med. Aroma Plant Sci.* 20, 359–361.

Koteswara, R.Y., Vimalamma, G., Rao, C.V., and Tzeng, Y (2004). Flavonoids and andrographolides from Andrographis paniculata. *Phytochemistry* 65, 2317–2321. doi: 10.1016/j.phytochem.2004.05.008

Kumar, R., Kumar, C., Paliwal, R., Choudhury, D. R., Singh, I., Kumar, A., et al. (2020). Development of novel genomic simple sequence repeat (g-SSR) markers and their vvlidation for genetic diversity analyses in kalmegh [*Andrographis paniculata* (Burm. f.) nees]. *Plants* 9, 17–34. doi: 10.3390/plants9121734

Kumar, N., Markar, S., and Kumar, V. (2014). Studies on heritability and genetic advance estimates in timely sown bread wheat (*Triticum aestivum l.*). *Biosci. Discv.* 5, 64–69.

Kumar, A., and Shekhawat, N. S. (2009). *Plant tissue culture and molecular markers: Their role in improving crop productivity* (India: I. K. International Publishing House).

Lattoo, S. K., Dhar, R. S., Khan, S., Bamotra, S., Bhan, M. K., Dhar, A. K., et al. (2008). Comparative analysis of genetic diversity using molecular and morphometric markers in *Andrographis paniculata* (Burm. f.) nees. *Genet. Resour. Crop Evol.* 55, 33–43. doi: 10.1007/s10722-007-9212-y

Lim, J. C. W., Chan, T. K., Ng, D. S., Sagineedu, S. R., Stanslas, J., and Wong, W. F. (2012). Andrographolide and its analogues: versatile bioactive molecules for combating inflammation and cancer. *Clin. Exp. Pharmacol. Physiol.* 39, 300–310. doi: 10.1111/j.1440-1681.2011.05633.x

Lin, H. H., Shi, M. D., Tseng, H. C., and Chen, J. H. (2014). Andrographolide sensitizes the cytotoxicity of human colorectal carcinoma cells toward cisplatin *via* enhancing apoptosis pathways *in vitro* and *in vivo*. *Tox. Sci.* 139, 108–120. doi: 10.1093/toxsci/kfu032

Li, W., Xu, X., Zhang, H., Ma, C., Fong, H., van Breemen, R., et al. (2007). Secondary metabolites from *Andrographis paniculata*. chem. *Pharm. Bull.* 55, 455–458. doi: 10.1248/cpb.55.455

Louette, D., Charrier, A., and Berthaud, J. (1997). *In-situ* conservation of maize in Mexico: Genetic diversity and maize seed management in a traditional community. *Econ Bot.* 51, 20–38. doi: 10.1007/BF02910401

Luo, X., Luo, W., Lin, C., Zhang, L., and Li, Y. (2014). Andrographolide inhibits proliferation of human lung cancer cells and the related mechanisms. *Int. J. Clin. Exp. Med.* 7, 4220–4225.

Lupini, A., Aci, M. M., Mauceri, A., Luzzi, G., Bagnato, S., Menguzzato, G., et al. (2019). Genetic diversity in old populations of sessile oak from calabria assessed by nuclear and chloroplast SSR. *J. Mt. Sci.* 16, 1111–1120. doi: 10.1007/s11629-018-5335-1

Maison, T., Volkaert, H., Boonprakob, U., and Paisooksantivatana, Y. (2005). Genetic diversity of *Andrographis paniculata* wall. ex nees as revealed by morphological characters and molecular markers. kasetsart J. *Nat. Sci.* 39, 388–399.

Milbourne, D., Meyer, R., Bradshaw, J. E., Baird, E., Bonar, N., Provan, J., et al. (1997). Comparison of PCR-based marker systems for the analysis of genetic relationships in cultivated potato. *Mol. Breed.* 3, 127–136. doi: 10.1023/A:1009633005390

Minz, P. L., Singh, N., Mishra, S. K., and Koche, V. (2013). Genetic variability among *Andrographis paniculata* in chhattisgarh region assessed by RAPD markers. *Afr. J. Biotechnol.* 12, 5174–5722. doi: 10.5897/AJB2012.2970

Mishra, H. O., Sharma, J. R., Lal, R. K., and Shukla, N. (2000). Pattern of genetic variability for different traits in a collection of kalmegh (*Andrographis paniculata*) genotypes. conference title: Proceedings of the national seminar on the frontiers of research and development in medicinal plants. *J. Med. Aromat. Plants.* 22, 348–351.

Mondal, S., Roy, P., Das, S., Halder, A., Mukherjee, A., and Bera, T. (2013). *In vitro* susceptibilities of wild and drug resistant *Leishmania donovani* amastigote stages to andrographolide nanoparticle: role of vitamin e derivative TPGS for nanoparticle efficacy. *PloS One* 8, e81492. doi: 10.1371/journal.pone.0081492

Nagvanshi, D., and Tirkey, A. (2016). Studies on genetic diversity in various quantitative characters in kalmegh (*Andrographis paniculata*) germplasm. *Adv. Res. J. Crop Improv.* 7, 60–64.

Ooi, J. P., Kuroyanagi, M., Sulaiman, S. F., Muhammad, T. S. T., and Tan, M. L. (2011). Andrographolide and 14-deoxy-11, 12-didehydroandrographolide inhibit cytochrome P450s in HepG2 hepatoma cells. *Life Sci.* 88, 447–454. doi: 10.1016/j.lfs.2010.12.019

Padmesh, P., Sabu, K. K., Seeni, S., and Pushpangadan, P. (1999). The use of RAPD in assessing genetic variability in *Andrographis paniculata nees*, a hepatoprotective drug. *Curr. Sci.* 76, 833–835.

Pandey, G., and Rao, C. H. (2018). Andrographolide: its pharmacology, natural bioavailability and current approaches to increase its content in *Andrographis paniculata*. *Int. J. Complement Alt Med.* 11, 355–360. doi: 10.15406/ijcam.2018.11.00425

Pashley, C. H., Ellis, J. R., McCauley, D. E., and Burke, J. M. (2006). EST databases as a source for molecular markers: lessons from helianthus. *J. Hered.* 9, 381–388. doi: 10.1093/jhered/esl013

Peakall, R., and Smouse, P. E. (2012). GenAlEx V6.5: genetic analysis in excel. population genetic software for teaching and research–an update. *Bioinformatics* 28, 2537–2539. doi: 10.1111/j.1471-8286.2005.01155.x

Peng, J., Shi, C., Wang, D., Li, S., Zhao, X., Duan, A., et al. (2021). Genetic diversity and population structure of the medicinal plant *Docynia delavayi* (Franch.) schneid revealed by transcriptome-based SSR markers. *J. Appl. Res. Med. Aromat. Plants*. 21, 100294. doi: 10.1016/j.jarmap.2021.100294

Popat, R., Patel, R., and Parmar, D. (2020). *Variability: Genetic variability analysis for plant breeding research*. Available at: https://cran.rproject.org/web/packages/variability/variability.pdf.

Prevost, A., and Wilkinson, M. J. (1999). A new system of comparing PCR primers applied to ISSR fingerprinting of potato cultivars. *Theor. Appl. Genet.* 98, 107–112. doi: 10.1007/s001220051046

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959. doi: 10.1534/genetics.116.195164

Raghavan, R., Cheriyamundath, S., and Madassery, J. (2014). 14-Deoxy-11, 12-didehydroandrographolide inhibits proliferation and induces GSH-dependent cell death of human promonocytic leukemic cells. *J. Nat. Med.* 68, 387–394. doi: 10.1007/s11418-014-0815-2

Rao, Y. K., Vimalamma, G., Rao, C. V., and Tzeng, Y. M. (2004). Flavonoids and andrographolides from *Andrographis paniculata*. *Phytochemistry* 65, 2317–2321. doi: 10.1016/j.phytochem.2004.05.008

Rohlf, F. J. (2000). *NTSYSpc: Numerical taxonomy and multivariate analysis system version 2.11x* (Setauket, New York: Exeter Software).

Sabu, K. K., Padmesh, P., and Seeni, S. (2001). Intraspecific variation in active principle content and isozymes of *Andrographis paniculata* nees (Kalmegh): a traditional hepatoprotective medicinal herb of India. *J. Med. Aromat Plant Sci.* 23, 637–647.

Sahoo, A., Behura, S., Singh, S., Jena, S., Ray, A., Dash, B., et al. (2021). EST-SSR marker-based genetic diversity and population structure analysis of Indian *Curcuma species*: significance for conservation. *Rev. Bras. Bot.* 44, 411–428. doi: 10.1007/s40415-021-00711-1

Sathyanarayana, N., Pittala, R. K., Tripathi, P. K., Chopra, R., Singh, H. R., Belamkar, V., et al. (2017). Transcriptomic resources for the medicinal legume *Mucuna pruriens: de novo* transcriptome assembly, annotation, identification and validation of EST-SSR markers. *BMC Genom.* 18, 409. doi: 10.1186/s12864-017-3780-9

Sharma, H., Kumar, P., Singh, A., Aggarwal, K., Roy, J., Sharma, V., et al. (2020). Development of polymorphic EST-SSR markers and their applicability in genetic diversity evaluation in *Rhododendron arboretum*. *Mol. Biol. Rep.* 47, 2447–2457. doi: 10.1007/s11033-020-05300-1

Sharma, S. N., Sinha, R. K., Sharma, D. K., and Jha, Z. (2009). Assessment of intra-specific variability at morphological, molecular and biochemical level of *Andrographis paniculata* (Kalmegh). *Curr. Sci.* 96, 402–408.

Shiferaw, E., Pe, M. E., Porceddu, E., and Ponnaiah, M. (2012). Exploring the genetic diversity of Ethiopian grass pea (*Lathyrus sativus l.)* using EST-SSR markers. *Mol. Breed.* 30, 789–797. doi: 10.1007/s11032-011-9662-y

Singh, R. K., and Chaudhary, B. D. (1985). *Biometrical method in quantitative genetics analysis* (New Delhi: Kalyani Publishers), 225–252.

Singh, S. P., Nodari, R., Gepts, P., and Singh, S. P. (1991).). genetic diversity in cultivated common bean: I. allozymes. *Crop Sci.* 31, 19–23. doi: 10.2135/cropsci1991.0011183X003100010004x

Soriano, J. M., Villegas, D., Aranzana, M. J., García del Moral, L. F., and Royo, C. (2016). Genetic structure of modern durum wheat cultivars and Mediterranean landraces matches with their agronomic performance. *PloS One* 11, e0160983. doi: 10.1371/journal.pone.0160983

Tal, G. (2015). Dendextend: an r package for visualizing, adjusting, and comparing trees of hierarchical clustering. *Bioinformatics* 31, 3718–3720. doi: 10.1093/bioinformatics/btv428

Tewari, S. K., Niranjan, A., and Lehri, A. (2010). Variations in yield, quality, and antioxidant potential of kalmegh (*Andrographis paniculata nees*) with soil alkalinity and season. *J. Herbs Spices Med. Plants.* 16, 41–50. doi: 10.1080/10496475.2010.481926

Tiwari, G., Singh, R., Singh, N., Chaudhary, D. R., and Paliwal, R. (2016). Study of arbitrarily amplified (RAPD and ISSR) and gene targeted (Scot and CBDP) markers for genetic diversity and population structure in kalmegh *[Andrographis paniculata* (Burm. f.) nees]. *Ind. Crops Prods.* 86, 1–11. doi: 10.1016/j.indcrop.2016.03.031

Tulsani, N. J., Hamid, R., Jacob, F., Umretiya, N. G., Nandhae, A. K., Tomarf, R. S., et al. (2019). Transcriptome landscaping for gene mining and SSR marker development in coriander (*Coriandrum sativum* l.). *Genomics* 112, 1545–1553. doi: 10.1016/j.ygeno.2019.09.004

Valdiani, A., Kadir, M. A., Tan, S. G., Talei, D., Abdullah, M. P., and Nikzad, S. (2012). Nain-e havandi *Andrographis paniculata* present yesterday, absent today: a plenary review on underutilized herb of iran's pharmaceutical plants. *Mol. Biol. Rep.* 39, 5409–5424. doi: 10.1007/s11033-011-1341-x

Valdiani, A., Talei, D., Javanmard, A., Tan, S. G., Kadir, M. A., and Maziah, M. (2014). Morpho-molecular analysis as a prognostic model for repulsive feedback of the medicinal plant "*Andrographis paniculata*" to allogamy. *Gene* 542, 156–167. doi: 10.1016/j.gene.2014.03.039

Varshney, R. K., Chabane, K., Hendre, P. S., Aggarwal, R. K., and Graner, A. (2007). Comparative assessment of EST-SSR, EST-SNP and AFLP markers for evaluation of genetic diversity and conservation of genetic resources using wild, cultivated and elite barleys. *Plant Sci.* 173, 638–649. doi: 10.1016/j.plantsci.2007.08.010

Varshney, R. K., Graner, A., and Sorrells, M. E. (2005). Genic microsatellite markers in plants: features and applications. *Trends Biotechnol.* 23, 1237–1248. doi: 10.1016/j.tibtech.2004.11.005

Vidya, V., Prasath, D., Snigdha, M., Gobu, R., Sona, C., and Maiti, C. S. (2021). Development of EST-SSR markers based on transcriptome and its validation in ginger (*Zingiber officinale rosc.*). *PloS One* 16, e0259146. doi: 10.1371/journal.pone.0259146

Wang, J., Tan, X. F., and Nguyen, V. S. (2014). A quantitative chemical proteomics approach to profile the specific cellular targets of andrographolide, a promisinganticancer agent that suppresses tumor metastasis. *Mol Cell Proteomics* 13, 876–886.

Weiner, J., and Weiner, M. J (2020). *Package 'pca3d'. ver.0.10.2 .Three dimensional PCA plots.*

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., et al. (2016). *ggplot2: Create elegant data visualisations using the grammar of graphics. R package version. 2.2.1.*

Wijarat, P., Keeratinijakal, V., Toojinda, T., Vanavichit, A., and Tragoonrung, S. (2011). Genetic diversity and inbreeder species of andrographis paniculata (Burm. f.) nees by randomly amplified polymorphic deoxyribonucleic acid (RAPD) and floral architecture analysis. *J. Plant Breed Crop Sci.* 3, 327–334. doi: 10.5897/JPBCS11.066

Wu, J., Cai, C., Cheng, F., Cui, H., and Zhou, H. (2014). Characterization and development of EST-SSR markers in tree peony using transcriptome sequences. *Mol. Breed.* 34, 1853–1866. doi: 10.1007/s11032-014-0144-x

You, F.M., Huo, N., Gu, Y.Q., Luo, M.C., and Ma, Y. (2008). BatchPrimer3: a high throughput web application for PCR and sequencing primer design. *BMC bioinformatics* 9 (1), 1–13. doi: 10.1186/1471-2105-9-253

Zhou, Q., Mua, K., Nia, Z., Liub, X., Lib, Y., and Xua, L. (2019). Analysis of genetic diversity of ancient ginkgo populations using SSR markers. *Ind. Crops Prod.* 154, 112687. doi: 10.1016/j.indcrop.2019.111942

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 30 November 2022
DOI 10.3389/fpls.2022.1027558

OPEN ACCESS

EDITED BY
Nisha Singh,
Gujarat Biotechnology University, India

REVIEWED BY
Zitong Li,
Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australia
Abhinandan Surgonda Patil,
Agharkar Research Institute, India

*CORRESPONDENCE
Dwijesh Chandra Mishra
Dwijesh.Mishra@icar.gov.in
Anil Rai
anil.rai@icar.gov.in

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

RECEIVED 25 August 2022
ACCEPTED 11 October 2022
PUBLISHED 30 November 2022

CITATION
Budhlakoti N, Mishra DC, Majumdar SG, Kumar A, Srivastava S, Rai SN and Rai A (2022) Integrated model for genomic prediction under additive and non-additive genetic architecture.
Front. Plant Sci. 13:1027558.
doi: 10.3389/fpls.2022.1027558

Integrated model for genomic prediction under additive and non-additive genetic architecture

Neeraj Budhlakoti[1], Dwijesh Chandra Mishra[1]*, Sayanti Guha Majumdar[1], Anuj Kumar[2], Sudhir Srivastava[1], S. N. Rai[3] and Anil Rai[1]*

[1]Division of Agricultural Bioinformatics, ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India, [2]Department of Microbiology and Immunology, Dalhousie University, Halifax, NS, Canada, [3]Bioinformatics and Biostatistics Department, University of Louisville, Louisville, KY, United States

Using data from genome-wide molecular markers, genomic selection procedures have proved useful for estimating breeding values and phenotypic prediction. The link between an individual genotype and phenotype has been modelled using a number of parametric methods to estimate individual breeding value. It has been observed that parametric methods perform satisfactorily only when the system under study has additive genetic architecture. To capture non-additive (dominance and epistasis) effects, nonparametric approaches have also been developed; however, they typically fall short of capturing additive effects. The idea behind this study is to select the most appropriate model from each parametric and nonparametric category and build an integrated model that can incorporate the best features of both models. It was observed from the results of the current study that GBLUP performed admirably under additive architecture, while SVM's performance in non-additive architecture was found to be encouraging. A robust model for genomic prediction has been developed in light of these findings, which can handle both additive and epistatic effects simultaneously by minimizing their error variance. The developed integrated model has been assessed using standard evaluation measures like predictive ability and error variance.

KEYWORDS

GBLUP, GEBVs, k-RCV, nonparametric, parametric, SVM, RCV

1 Introduction

Genomic selection is a form of marker-assisted selection (MAS) in which genomic markers covering the whole genome are used to identify quantitative trait loci (QTL) which are in linkage disequilibrium (LD) with at least one marker (Meuwissen et al., 2001). Genomic selection predicts the breeding values of individuals or lines in a population by analyzing their phenotypes and high-density marker scores. The genomic selection process starts with building a statistical model from individuals having both genotypic and phenotypic information (i.e., training set); this model is further used for estimation of breeding value of the individuals in the breeding population/validation set (i.e., Genomic Estimated Breeding Value (GEBVs) for individuals having only genotypic information). Individuals are then ranked on the basis of GEBVs and subsequently superior individuals are selected. Genomic selection methods have been successfully applied for various plants (Jannink et al., 2010; Spindel et al., 2015; Zhao et al., 2015; Crossa et al., 2016; Liu et al., 2019) and animals (Hayes et al., 2009; Daetwyler et al., 2010; Daetwyler et al., 2012; Wang et al., 2013; Wolc et al., 2015; Lu et al., 2016; Wiggans et al., 2017; Liu et al., 2019), and reason behind this success is that it incorporates all information on genome wide markers into the prediction model.

As a choice of model, different methods that may be parametric, nonparametric, and semiparametric can be used for genomic selection. But, in general, it was observed that performance of parametric methods were considerably better than nonparametric methods in case of additive genetic architectures (Gianola et al., 2006; Crossa et al., 2010; Daetwyler et al., 2010; Heslot et al., 2012; Howard et al., 2014; Sahebalam et al., 2019). The practical use of genomic selection includes efforts such as appropriate statistical model selection, training and testing data proportions, marker density, etc., which requires resource-based decision-making. Prediction accuracy of a model can also be affected by factors like span of LD, heritability of trait under observation, and genetic architecture of individual under study. Due to the complexity of plant genetics, some genomic selection techniques perform very poorly as they are unable to model marker variance. Further, due to the huge number of epistatic interactions, it becomes challenging to practice parametric methods (Moore and Williams, 2009). In epistatic interactions, a number of loci are involved and also the possibility of interaction cannot be ignored. Epistatic interaction may play a crucial role for explaining genetic variation for quantitative traits, as ignoring these kinds of interaction in the model may result in lower genomic prediction accuracy (Gianola et al., 2006; Cooper et al., 2009). In such cases, performance of model free i.e. nonparametric methods were found to be more impressive (Gianola et al., 2006).

Although some semiparametric (Gianola et al., 2006; Campos et al., 2010; Legarra and Reverter, 2018) and other robust approaches (Tanaka, 2018; Budhlakoti et al., 2020a; Majumdar et al., 2020b; Sehgal et al., 2020; Mishra et al., 2021) have also been proposed and implemented for this purpose, there is still room for improvement. To overcome the limitation of individual parametric and nonparametric models, the current study has been designed to develop a robust model by integrating the best model from each category that can handle diverse genetic architecture.

2 Materials and method

In GS, our main objective is to select superior individuals by modelling the relationship between individual genotypic and phenotypic information. One of the simplest models for modeling this relationship is simple linear regression model. One problem with linear regression is that, generally, the number of markers (genotype) is greater than the number of individual (phenotype), that is, there exists a problem of large p and small n i.e., $p > n$. In such a case, it may not be possible to estimate parameters of regression model. Therefore, variable selection approach i.e., Ridge Regression (RR) and Least absolute Shrinkage and Selection Operator (LASSO), are alternatives to this situation. Some other improved methods include Best Linear Unbiased Prediction (BLUP) (Henderson, 1949), Genomic BLUP (GBLUP) (Endelman and Jannink, 2012), Bayesian methods, and their derivatives i.e. Bayes A, Bayes B, Bayes C π and D π (Meuwissen et al., 2001; Gianola et al., 2009; Habier et al., 2009; Habier et al., 2010). However, assumptions of parametric models do not always hold (e.g., normality, linearity, independent explanatory variables), which further suggests the use of nonparametric methods. Various nonparametric based methods, i.e. Reproducing Kernel Hilbert Space (RKHS), Support Vector Machine (SVM), Artificial Neural Network (ANN), and Random Forest (RF), have been proposed and successfully used for genomic prediction in plants and animals. A detailed comparison of various parametric and nonparametric methods has been provided by Howard et al., 2014; Budhlakoti et al., 2020b, in context to genomic selection.

2.1 Integrated estimation of GEBVs

The best model from each parametric and nonparametric methods was identified. Under parametric methods performance of GBLUP was found to be the best, whereas for nonparametric method, SVM was found to be best using appropriate evaluation measures. An integrated estimator for GEBVs (more formally GEBVs from parametric methods and EGV i.e. estimated genomic values from nonparametric methods) has been developed for genomic selection by combining estimates from the best parametric and nonparametric methods (Majumdar et al., 2020a and

Majumdar et al., 2020b). For better understanding, details of both the methods have been given below.

2.2 Best linear unbiased prediction

BLUP is based on the theory of mixed random effect model. Statistical formulation of the BLUP model can be written as follows:

$$Y = X\beta + Zm + e$$

where, β is a p × 1 vector of fixed effects, m is q × 1 vector of random effects, m~N (0,G) and e is n × 1vector of residuals, e ~N (0, R). The estimator of fixed effect β is called Best Linear Unbiased Estimator (BLUE) and random effects m is known as BLUP. Estimation of BLUE and BLUP (β, m) by maximizing the joint likelihood function is given below (Henderson, 1949):

$$f(Y, m) = f(Y|m)f(m)$$

$$= \frac{1}{2\pi^{n/2}|R|^{1/2}}\left[-\frac{1}{2}(Y - X\beta - Zm)' R^{-1}(Y - X\beta - Zm)\right]$$

$$\times \frac{1}{2\pi^{p/2}|G|^{1/2}}\left[-\frac{1}{2}m' G^{-1}m\right]$$

The estimate of (β, m) could be obtained by maximizing the log of the above likelihood function and equating it to zero, which could be written as the famous Henderson mixed model equation:

$$\begin{pmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} \end{pmatrix}\begin{pmatrix} \hat{\beta} \\ \hat{m} \end{pmatrix} = \begin{pmatrix} X'R^{-1}Y \\ Z'R^{-1}Y \end{pmatrix}$$

where G = var (m) and R = var (e). The solution to the Henderson equation is BLUE of β, BLUP of m, where m and e are normally distributed and maximizes f (Y, m) over unknown parameters β and m.

GBLUP is an improved version of BLUP where additive genomic relationship matrix (G) is used as a variance-covariance matrix of random effect in the model.

2.3 Support vector machine

SVM is based on the principle of maximum separating hyperplane. It constructs a hyperplane with the objective of separating data into different classes. In case our problem is based on regression instead of classification, i.e., when output data is continuous in nature, then the Support Vector Regression can be used. Support Vector Regression (SVR) is an important application of SVM technique and has been used interchangeably in the literature. In order to understand this, consider a mapping function $f(X)$: $R^p \rightarrow R$, given the set of training data

$$(X_1, Y_1), (X_2, Y_2) \ , \ . \ . \ , (X_n, Y_n), \ X_i \in R^p, \ Y_i \in R$$

Let us assume a simple linear function of the following form: f (X) = w'X + b, where, w is vector of weight to be estimated (i.e. regression coefficients) and b denotes bias. f (X) is minimized by the following problem formulation:

$$\min_{w,b} \varnothing(w, b) = \frac{1}{2}||w||^2 + c\sum_{i=1}^{n} e_i^k$$

where $e_i = Y_i - f(X_i)$, is error of i^{th} data point from training set, also known as loss function L(.) which measures quality of estimation, and c represents regularization parameter which handles trade-off between margin and error.

2.4 Proposed estimator

The integrated estimator for estimated breeding or genomic value can be expressed as

$$Y_{Est} = wY_{GBLUP} + (1 - w)Y_{SVR} \quad (1)$$

where, Y_{Est} is new predicted phenotype from integrated model, w is $\frac{\sigma^2_{SVR}}{\sigma^2_{GBLUP}+\sigma^2_{SVR}}$, where σ^2_{SVR} and σ^2_{GBLUP} are the error variance of models SVR and GBLUP respectively, Y_{GBLUP} is the predicted GEBV from GBLUP, whereas Y_{SVR} is the predicted EGV from SVR model. Let us assume that error variance of Y_{Est} is represented by σ^2_{EST}, then by optimizing w, σ^2_{EST} can be obtained as:

$$\sigma^2_{Est} = \left(\frac{\sigma^2_{SVR}}{\sigma^2_{GBLUP}+\sigma^2_{SVR}}\right)^2 \sigma^2_{GBLUP} + \left(\frac{\sigma^2_{GBLUP}}{\sigma^2_{GBLUP}+\sigma^2_{SVR}}\right)^2 \sigma^2_{SVR}$$
$$\sigma^2_{Est} = \frac{\sigma^2_{GBLUP}\sigma^2_{SVR}}{\sigma^2_{GBLUP}+\sigma^2_{SVR}} \quad (2)$$

2.5 Estimation of error variance for proposed estimator

In order to develop the integrated genomic selection model, estimate of error variances for GBLUP (σ^2_{GBLUP}) and SVR (σ^2_{SVR}) models have been obtained using two different methods i.e. Refitted Cross Validation (RCV) and k fold Refitted Cross Validation (k-RCV). RCV method was originally given by Fan et al., 2012, for the estimation of error variance in ultrahigh dimensional regression procedure. The basic procedure behind RCV and k-RCV is the same except that data is split into two equal halves for RCV and k equal sizes for k-RCV respectively. Algorithm of both RCV and k-RCV methods are depicted through the flow diagrams in Figure 1.

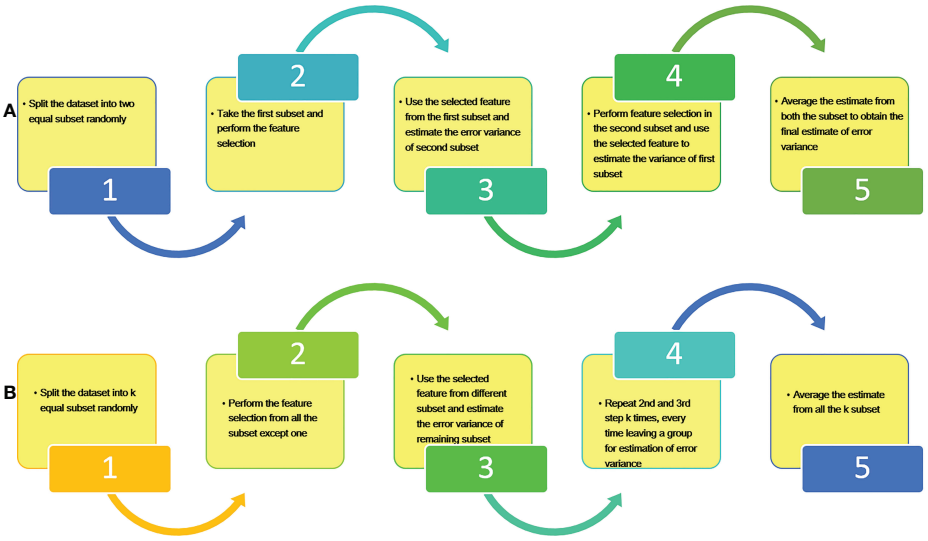


FIGURE 1
Basic steps for estimation of error variance using **(A)** RCV and **(B)** k-RCV.

2.6 Data simulation

In order to check the performance of the model, data was simulated using QTL Bayesian interval mapping method implemented in R based package "*qtlbim*" (Yandell et al., 2007). R is open source and freely available at http://www.r-project.org (R Core Team, 2019). Package "*qtlbim*" is based on Cockerham's model which is a standard model for simulation of marker data and has been followed in many studies (Bedo et al., 2008; Piao et al., 2011; Howard et al., 2014; Budhlakoti et al., 2020a; Budhlakoti et al., 2020b; Li et al., 2020).

Statistical formulation of Cockerham's model is given as follows:

$$Y_{ijk} = G_{ij} + e_{ijk}$$

$$= \mu + a_1 x_1 + d_1 z_1 + a_2 x_2 + d_2 z_2 + i_{aa} w_{aa} + i_{ad} w_{ad} + i_{da} w_{da} + i_{dd} w_{dd} + e_{ijk} \quad (3)$$

where μ is the mean, a_1 and a_2 are additive genetic effects at locus A & B, d_1 and d_2 are dominance effects at locus A & B, e_{ijk} is a residual. i_{aa} is additive × additive effect of loci A and B, i_{ad} is additive × dominance of loci A and B, i_{da} is dominance × additive of loci A and B, and i_{dd} is dominance × dominance of loci A and B.

We have simulated a total of five data sets for genotypic and phenotypic information using the Cockerham's model described above (Eq. 3) with diversified genetic architecture (additive and epistasis) at various levels of heritability (ranges from low heritability 0.3 to medium 0.5 and high heritability 0.7 for F2 population). For the additive data, there is one QTL in each chromosome with either a positive or negative additive effect and no epistatic interaction say it as (a, e_0). For non-additive/epistatic data, we assumed two QTLs on each of the five, seven, and ten chromosomes respectively; remaining chromosomes have no QTL. So, a total 5, 7, and 10 two-way epistatic interactions are considered for the non-additive datasets. So in each dataset, there is a combination of one of the five different levels of heritability (viz. 0.3, 0.5, 0.7) and four levels of epistatic effects (viz. 0, 5, 7, 10) denoted as e_0, e_1, e_2, e_3. So, finally, we have four different combinations of datasets with additive and epistatic effects i.e. (a, e_0), (a, e_1), (a, e_2) and (a, e_3). For each genetic architecture we have simulated the data for 200 individuals with 2000 SNPs each. Simulated data have 10 chromosomes with 200 SNPs in each with specified length. A total of 2000 markers are distributed over all 10 chromosomes in such a way that each marker is equi-spaced over the chromosome. No missing genotypic values and no missing phenotypic values are considered in the datasets.

2.7 Real data set

In order to check the robustness of our approach the same has been validated using real data. We have used a total of six datasets in the current study. A detailed discussion regarding each of the dataset is given below.

2.7.1 Dataset 1: Wheat

Wheat lines were genotyped using 1447 Diversity Array Technology markers generated by Triticarte Pty. Ltd. (Canberra, Australia; http://www.triticarte.com.au). Markers are coded for two different values i.e. their presence (1) or absence (0). This data set includes 599 lines phenotyped for trait grain yield (GY) for four mega environments. However, for matter of

convenience we have just considered GY for the first mega environment. The final number of DArT markers after quality control and final editing was 1279 and the same was used in the current study (Crossa et al., 2010; Cuevas et al., 2016).

2.7.2 Dataset 2: Maize

The maize dataset is generated by CIMMYT's Global Maize Program (Crossa et al., 2010). It originally included 300 maize line with 1148 SNP markers. Markers with the highest frequency are coded as 0 and lowest frequency as 1. Here also the trait under study is GY, evaluated under drought and watered conditions. After final editing, 264 maize lines with 1135 SNPs markers were available for final study (Crossa et al., 2010).

2.7.3 Dataset 3-6: Wheat

This wheat dataset is generated from CIMMYT semiarid wheat breeding program, which is comprised of 254 advanced wheat breeding lines genotyped for 1726 DArt markers (Poland et al., 2012). Dataset is recorded for four different phenotypic traits: Days to Heading (DTH), Thousand Kernel Weight (TKW), Yield (under irrigated condition hence denoted as Y_I), and Yield (under draught condition i.e. Y_D). For convenience, here trait DTH is considered as Dataset-3, trait TKW as Dataset-4, trait Y_I as Dataset-5, and trait Y_D as Dataset-6.

2.8 Evaluation measure

Predictive Ability and Prediction Error were used for evaluation of the different models. Predictive ability can be defined as Pearson correlation coefficient (r) between observed phenotypic value and predicted phenotypic value. The same can be expressed as (Eq. 4)

$$r = \frac{S_{Y,\hat{Y}}}{S_Y S_{\hat{Y}}} \tag{4}$$

where $S_{Y,\hat{Y}}$ denotes the covariance between observed and predicted phenotypic value, S_Y is standard deviation of observed phenotype, and $S_{\hat{Y}}$ denotes standard deviation of predicted phenotype. Prediction error can be simply defined as mean sum of square error (MSE) between observed phenotypic value and predicted phenotypic value. The same can be expressed using the following formula (Eq. 5)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{5}$$

where Y_i is observed response, $\hat{Y}_i$ is predicted phenotype value of i^{th} individual, and n denotes total number of individuals in the training set.

To compare the performance of methods under study, a cross-validation technique is used. Data is divided into two parts, i.e., training and validation sets, in such a way that the training set comprises 70% of data and the rest of the data is in the validation set. The former is used for model building and the latter for model evaluation. The whole procedure is repeated 100 times and predictive ability and prediction error were calculated. For better understanding, a brief flowchart of the whole procedure followed in the current study is provided in Figure 2.

In order to implement all the methods under study, R programming platform (R Core Team (2019). R: A language and environment for statistical computing, R foundation for statistical computing Vienna - Google Search) was used; to fit different models under study, R package STGS was used (Budhlakoti et al., 2019).

3 Results and discussion

3.1 Comparative study of existing parametric methods

Here, using a simulation analysis, the most popular methods (i.e., Stepwise Regression, BLUP, LASSO, Bayesian LASSO, and GBLUP) for genomic selection under diverse genetic architectures were examined. Each method was evaluated at different heritability levels (i.e. 0.3, 0.5, and 0.7). Cross-validation technique was used to assess the performance of various models, and results of the same are presented in Table 1.

The following critical observations can be made from the results (Table 1).

i. At low heritability (0.3), the performance of GBLUP was found to be the highest and reasonable. However, performance of BLUP and Bayesian LASSO were also quite impressive. It can also be observed that as heritability increases, the performance of LASSO in comparison to other methods quickly improves.
ii. At moderate heritability (0.5), performance of GBLUP is highest in comparison to the others. However, an important thing to note is that there is not much difference in the performance of all the methods except stepwise regression.
iii. At high heritability (0.7), consistency in the performance of GBLUP is still maintained, with the performance of other methods (BLUP, LASSO, and Bayesian LASSO) also at par with GBLUP.
iv. Performance of stepwise regression is very low throughout at all levels of heritability. This makes this method unsuitable for genomic selection studies.
v. LASSO can also be used as one of the preferable statistical models for genomic selection studies, especially when additive effects are present, but only for high heritable traits.

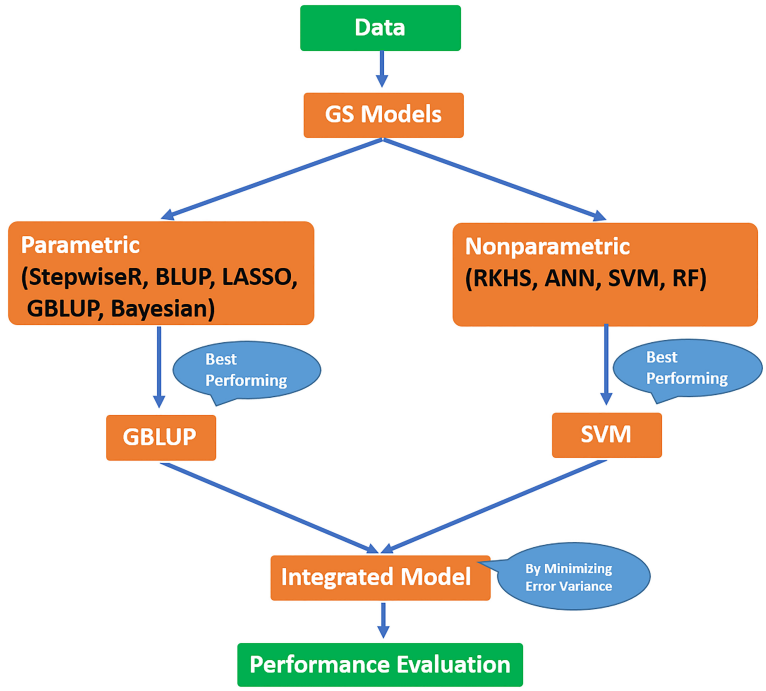


FIGURE 2
Flow diagram for the procedure followed to develop the integrated model in the current study.

vi. For real-time scenarios (e.g., agriculture field data) where trait heritability is generally low (for most commonly studied yield related traits), GBLUP can be quite good for genomic selection studies. Results indicates that GBLUP has better predictive ability of estimating GEBVs of individuals over their counterparts.

3.2 Comparative study of existing nonparametric methods

This section summarizes the performance of different nonparametric methods under study, i.e., RKHS, SVR, ANN, and RF, at diverse levels of heritability. Predictive ability and prediction error were used as evaluation measures for different models. Results of the same are presented in Table 2.

On the basis of results obtained (Table 2), the following inferences can be drawn:

i. Performance of SVR was consistent throughout different levels of heritability with respect to its predictive ability and MSE.
ii. However, ANN also performed quite well, almost at par with SVR. Performance of random forest was poor at low heritability, however it improved gradually with high heritability.
iii. Performance of RKHS and RF were not found to be encouraging in comparison to their counterparts throughout the study.

TABLE 1 Predictive ability and MSE of GEBVs for different parametric methods using simulated dataset at different levels of heritability (h^2).

h^2/Parameters		GBLUP	BayseianLasso	StepwiseR	BLUP	LASSO
0.3	**PA**	0.74	0.72	0.52	0.70	0.72
	MSE	0.26	0.18	0.92	0.26	0.21
0.5	**PA**	0.86	0.83	0.42	0.84	0.83
	MSE	0.23	0.25	0.90	0.24	0.23
0.7	**PA**	0.89	0.86	0.48	0.87	0.86
	MSE	0.32	0.32	0.88	0.32	0.24

TABLE 2 Predictive ability and MSE of EGVs for different nonparametric methods under study using simulated dataset for various levels of heritability (h^2).

h^2/Parameters		RKHS	ANN	SVR	RF
0.3	**PA**	0.53	0.72	0.75	0.63
	MSE	0.67	0.60	0.47	0.78
0.5	**PA**	0.55	0.82	0.85	0.70
	MSE	0.86	0.55	0.55	0.60
0.7	**PA**	0.62	0.84	0.88	0.72
	MSE	0.93	0.58	0.54	0.71

From the above discussion, two models, GBLUP and SVR, each from parametric and nonparametric respectively, can been considered as the best model based on their performances in terms of estimating GEBVs and EGVs respectively for selection of individuals. Using these results, a robust model has been developed by integrating GBLUP and SVR by minimizing their error variance. Detailed results regarding error variance estimated using different methods is given below.

3.3 Comparison of error variances for GBLUP, SVM and integrated model

Here, two different methods for estimation of error variances, i.e., RCV and k-RCV, have been used for GBLUP, SVR, and Integrated model. Results of the same have been presented one by one in the tables given below.

3.3.1 Refitted cross validation

Error variance estimated using Refitted Cross Validation (RCV) for GBLUP, SVR, and Integrated model is presented in Table 3.

From Table 3, it has been observed that error variance of the integrated model is less than the error variance of GBLUP and SVR at diverse genetic architectures i.e., irrespective of levels of heritability and genetic effects.

3.3.2 k-fold refitted cross validation

Error variances estimated using k-fold refitted cross validation (i.e., k-RCV) for GBLUP, SVR, and Integrated model were given in Table 4.

From Table 4, it has also been observed that the error variance of the integrated model was found to be less than GBLUP and SVR across all levels of heritability using k-RCV approach.

In order to compare and better understand the results obtained through different methods of estimations for error variance (i.e., RCV, k-RCV), the same has been presented graphically in Figure 3.

The following important findings can be drawn from the results (Figure 3).

i. The error variance estimated through RCV and k-RCV is almost similar. However, variance estimated through RCV is slightly lower than k-RCV; this difference may be caused by the reduced sample size in case of k-RCV.
ii. Our proposed method is robust to both architecture (i.e., additive and epistatic) as evidenced from error variance obtained through RCV and k-RCV.
iii. Error variance obtained through RCV and k-RCV is highest for SVR in comparison to BLUP and the integrated model.
iv. In general, error variance increases with increase in heritability level across the various methods.

3.4 Performance of error variance estimation methods for integrated model

Here we have presented the results of different error variance estimation methods (RCV and k-RCV) in terms of their capability and how accurately it gives GEBVs or EGVs. The same has been calculated using each approach, i.e., GBLUP,

TABLE 3 Error variance for different GS models at different heritability using RCV.

h^2	GBLUP	SVR	Integrated Model
0.3	1.12	4.57	0.90
0.5	0.94	9.39	0.85
0.7	1.10	22.84	1.05

TABLE 4 Error variance for different GS models at different heritability using k-RCV.

h^2	GBLUP	SVR	Integrated Model
0.3	1.04	4.57	0.85
0.5	1.05	9.91	0.95
0.7	1.28	26.72	1.22

SVR, and integrated model, and the predictive ability of each of them was observed.

3.4.1 Refitted cross validation

Predictive ability for GBLUP, SVR, and the integrated model using RCV variance is given below in the table at different levels of heritability and genetic effect.

3.4.2 k-fold refitted cross validation

Predictive ability for GBLUP, SVR, and the integrated model using k-RCV variance is given below in the table at different levels of heritability and genetic effect.

The following important findings can be drawn from the results obtained in Tables 5, 6:

i. Performance of GBLUP is good when data have only additive architecture, while SVR performs equally well with diverse genetic architecture (with and without epistasis), especially at low heritability.
ii. At low heritability, the performance of the integrated model is consistent and robust.
iii. However, at high heritability (i.e. h^2 = 0.5 & 0.7), the performance of all the models in terms of prediction accuracy are at par.
iv. With increasing levels of epistasis and heritability, the predictive ability of the integrated model is still maintained

In order to support the facts obtained from the results of the simulation study, the same has also been tested on real datasets. Results obtained from the real dataset also tells the same story; here prediction accuracy for the integrated model is either at par or better than GBLUP and SVR model. However, here also the performance of k-RCV is slightly better than RCV. Graphical representation of the same is given below (Figure 4).

From the above discussion, two models, GBLUP and SVR, each of parametric and nonparametric respectively, can be considered the best models based on their performance in terms of reduced error variance and improved estimation of GEBVs and EGVs, respectively, for the selection of individuals. On the basis of this result, a robust model has been developed in this study by integrating GBLUP and SVR based on suitable weightage according to their error variance.

3.5 Practical deployment to the breeding programs

Here we present the R script as supplementary information for estimating the GEBVs of an individual using the integrated

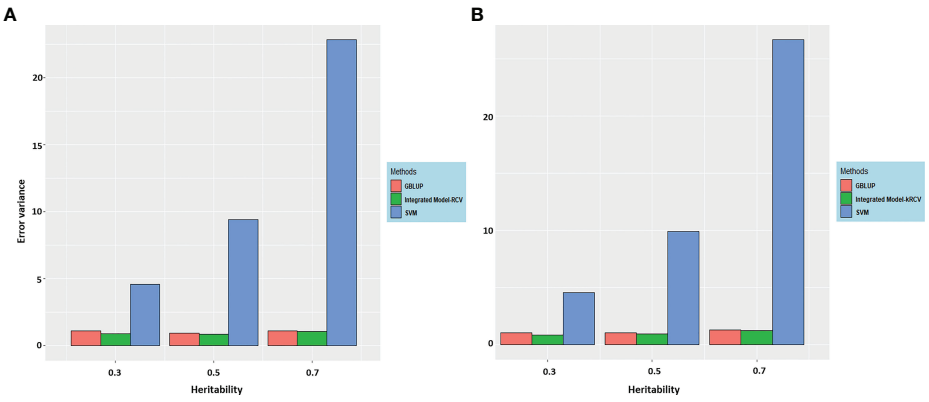


FIGURE 3
Error variance for integrated GS model at different heritability using various methods **(A)** RCV and **(B)** k-RCV in comparison to the error variance of best methods from both parametric and nonparametric i.e. GBLUP and SVR (Results from Tables 3, 4).

10.3389/fpls.2022.1027558

TABLE 5 Predictive ability (PA) with its standard error (SE) for different genomic selection models on mixed architecture (additive and epistatic effects) using RCV variance.

h^2		GBLUP (a,e_0)	GBLUP (a,e_1)	SVR(a, e_1)	SVR(a, e_0)	Integrated Model (a,e_0)	Integrated Model (a,e_1)	Integrated Model (a,e_2)	Integrated Model (a,e_3)
0.3	**PA**	0.74	0.68	0.75	0.74	0.75	0.74	0.70	0.64
	SE (PA)	0.045	0.072	0.062	0.063	0.059	0.056	0.055	0.06
0.5	**PA**	0.86	0.84	0.85	0.81	0.88	0.85	0.82	0.79
	SE (PA)	0.032	0.045	0.041	0.048	0.035	0.027	0.03	0.03
0.7	**PA**	0.89	0.87	0.88	0.85	0.90	0.89	0.84	0.81
	SE (PA)	0.030	0.039	0.032	0.041	0.024	0.020	0.02	0.02

TABLE 6 Predictive ability (PA) with its standard error (SE) for different GS models at different heritability using k-RCV variance.

h^2		GBLUP (a,e_0)	GBLUP (a,e_1)	SVR(a, e_1)	SVR(a, e_0)	Integrated Model (a,e_0)	Integrated Model (a,e_1)	Integrated Model (a,e_2)	Integrated Model (a,e_3)
0.3	**PA**	0.74	0.68	0.75	0.74	0.76	0.74	0.70	0.66
	SE (PA)	0.045	0.072	0.062	0.063	0.048	0.045	0.04	0.042
0.5	**PA**	0.86	0.84	0.85	0.81	0.88	0.87	0.82	0.80
	SE (PA)	0.032	0.045	0.041	0.048	0.029	0.026	0.032	0.041
0.7	**PA**	0.89	0.87	0.88	0.85	0.91	0.89	0.85	0.82
	SE (PA)	0.030	0.039	0.032	0.041	0.027	0.024	0.032	0.03

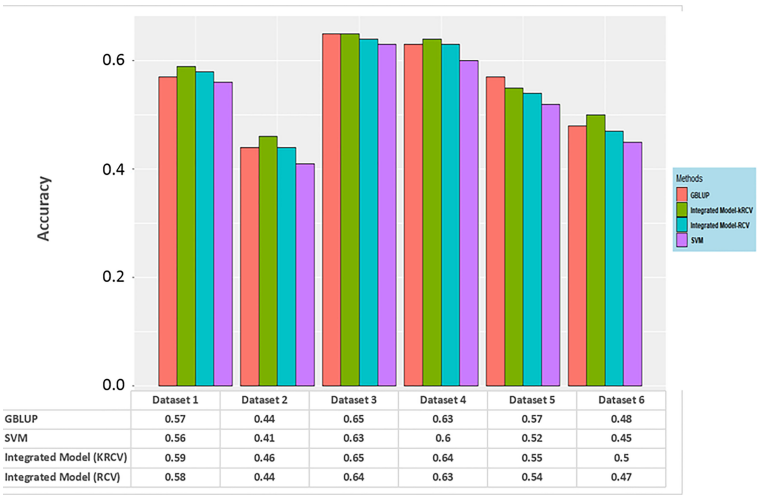


FIGURE 4
Prediction accuracy for different genomic selection models on a real dataset.

model (Supplementary File S1). The user may also run different GS-based models using a variety of other publicly accessible R tools & packages. In the future, GS-based tools or R packages may be developed that incorporate advanced and other GS-based models for hassle-free implementation.

4 Conclusion

In the current study, an effort has been made to develop a comprehensive methodology that addresses both the advantages and disadvantages of each parametric and nonparametric model. The performance of the GBLUP and SVR models was determined to be the best among its counterparts for both the parametric and nonparametric frameworks, respectively. The predictive ability and error variance of the developed integrated model were assessed, and it was found that our proposed approach performs either better or at par with existing models. It has also been observed that our proposed model is good at handling the diverse genetic architecture, i.e., additive and epistatic, in terms of reducing the error variance and enhancing the predictive ability. As a future directive, developed methodology could be evaluated by measuring the impact of within and across family predictive ability and other cross validation schemes.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: Crossa et al., 2010; Cuevas et al., 2016.

Author contributions

Conceived and designed the study: AR, NB, and DM. Developed the methodology: DM and NB. Performed the experiments: NB. Analyzed the data: NB and SM. Contributed materials: NB, DM, SS, SM, and AK. Drafted the manuscript: NB. Corrected the manuscript: AR, DM, NB, SS, and SR. All authors contributed to the article and approved the submitted version.

Funding

The funding support received from ICAR sponsored CABin scheme network project entitled "Agricultural Bioinformatics and Computational Biology".

Acknowledgments

The authors sincerely acknowledge the fellowship support received from PG School ICAR-IARI and ICAR-IASRI to conduct this research and analysis. Authors duly acknowledge the computational support received from Advanced Supercomputing Hub for OMICS Knowledge in Agriculture (ASHOKA).

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

Bedo, J., Wenzl, P., Kowalczyk, A., and Kilian, A. (2008). Precision-mapping and statistical validation of quantitative trait loci by machine learning. *BMC Genet.* 9, 1–18. doi: 10.1186/1471-2156-9-35

Budhlakoti, N., Mishra, D. C., Rai, A., Chaturvedi, K. K., and Budhlakoti, N. (2019) *Title genomic selection using single trait version 0.1.0*. Available at: https://cran.r-project.org/package=brnn.

Budhlakoti, N., Rai, A., and Mishra, D. C. (2020a). Statistical approach for improving genomic prediction accuracy through efficient diagnostic measure of influential observation. *Sci. Rep.* 10, 1–11. doi: 10.1038/s41598-020-65323-3

Budhlakoti, N., Rai, A., Mishra, D. C., Jaggi, S., Kumar, M., and Rao, A. R. (2020b). Comparative study of different non-parametric genomic selection methods under diverse genetic architecture. *Indian J. Genet. Plant Breed* 80, 395–401. doi: 10.31742/IJGPB.80.4.4

Campos, G. D. L., Gianola, D., Rosa, G. J. M., Weigel, K. A., and Crossa, J. (2010). Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. *Genet. Res. (Camb).* 92, 295–308. doi: 10.1017/S0016672310000285

Cooper, M., Podlich, D. W., Micallef, K. P., Smith, O. S., Jensen, N. M., Chapman, S. C., et al. (2009). Complexity, quantitative traits and plant breeding: a role for simulation modelling in the genetic improvement of crops. *Quant. Genet. Genomics Plant Breed.*, 143–166. doi: 10.1079/9780851996011.0143

Crossa, J., De Los Campos, G., Pérez, P., Gianola, D., Burgueño, J., Araus, J. L., et al. (2010). Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics* 186, 713–724. doi: 10.1534/genetics.110.118521

Crossa, J., Jarquín, D., Franco, J., Pérez-Rodríguez, P., Burgueño, J., Saint-Pierre, C., et al. (2016). Genomic prediction of gene bank wheat landraces. *G3 Genes Genomes Genet.* 6, 1819–1834. doi: 10.1534/G3.116.029637

Cuevas, J., Crossa, J., Soberanis, V., Pérez-Elizalde, S., Pérez-Rodríguez, P., de los Campos, G., et al. (2016). Genomic prediction of genotype × environment interaction kernel regression models. *Plant Genome* 9, 1–20. doi: 10.3835/plantgenome2016.03.0024

Daetwyler, H. D., Hickey, J. M., Henshall, J. M., Dominik, S., Gredler, B., van der Werf, J. H. J., et al. (2010). Accuracy of estimated genomic breeding values for wool and meat traits in a multi-breed sheep population. *Anim. Prod. Sci.* 50, 1004–1010. doi: 10.1071/AN10096

Daetwyler, H. D., Kemper, K. E., van der Werf, J. H. J., and Hayes, B. J. (2012). Components of the accuracy of genomic prediction in a multi-breed sheep population. *J. Anim. Sci.* 90, 3375–3384. doi: 10.2527/jas.2011-4557

Endelman, J. B., and Jannink, J. L. (2012). Shrinkage estimation of the realized relationship matrix. *G3 Genes Genomes Genet.* 2, 1405–1413. doi: 10.1534/g3.112.004259

Fan, J., Guo, S., and Hao, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *J. R. Stat. Soc B* 74, 37–65. doi: 10.1111/j.1467-9868.2011.01005.x

Gianola, D., de los Campos, G., Hill, W. G., Manfredi, E., and Fernando, R. (2009). Additive genetic variability and the Bayesian alphabet. *Genetics* 183, 347–363. doi: 10.1534/GENETICS.109.103952

Gianola, D., Fernando, R. L., and Stella, A. (2006). Genomic-assisted prediction of genetic value with semiparametric procedures. *Genetics* 173, 1761–1776. doi: 10.1534/genetics.105.049510

Habier, D., Fernando, R. L., and Dekkers, J. C. M. (2009). Genomic selection using low-density marker panels. *Genetics* 182, 343–353. doi: 10.1534/GENETICS.108.100289

Habier, D., Tetens, J., Seefried, F.-R., Lichtner, P., and Thaller, G. (2010). The impact of genetic relationship information on genomic breeding values in German Holstein cattle. *Genet. Sel. Evol.* 42, 1–12. doi: 10.1186/1297-9686-42-5

Hayes, B. J., Bowman, P. J., Chamberlain, A. J., and Goddard, M. E. (2009). Invited review: Genomic selection in dairy cattle: Progress and challenges. *J. Dairy Sci.* 92, 433–443. doi: 10.3168/JDS.2008-1646

Henderson, C. R. (1949). Estimation of changes in herd environment. *J. Dairy Sci.* 32 (8), 706–706.

Heslot, N., Yang, H.-P., Sorrells, M. E., and Jannink, J.-L. (2012). Genomic selection in plant breeding: A comparison of models. *Crop Sci.* 52, 146–160. doi: 10.2135/CROPSCI2011.06.0297

Howard, R., Carriquiry, A. L., and Beavis, W. D. (2014). Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. *G3 Genes Genomes Genet.* 4, 1027–1046. doi: 10.1534/g3.114.010298

Jannink, J.-L., Lorenz, A. J., and Iwata, H. (2010). Genomic selection in plant breeding: from theory to practice. *Brief. Funct. Genomics* 9, 166–177. doi: 10.1093/BFGP/ELQ001

Legarra, A., and Reverter, A. (2018). Semi-parametric estimates of population accuracy and bias of predictions of breeding values and future phenotypes using the LR method. *Genet. Sel. Evol.* 50, 1–18. doi: 10.1186/S12711-018-0426-6

Li, N., Shang, J., Wang, J., Zhou, D., Li, N., and Ma, S. (2020). Discovery of the genomic region and candidate genes of the scarlet red flesh color (Yscr) locus in watermelon (Citrullus lanatus l.). *Front. Plant Sci.* 0. doi: 10.3389/FPLS.2020.00116

Liu, X., Wang, H., Hu, X., Li, K., Liu, Z., Wu, Y., et al. (2019). Improving genomic selection with quantitative trait loci and nonadditive effects revealed by empirical evidence in maize. *Front. Plant Sci.* 0. doi: 10.3389/FPLS.2019.01129

Lu, D., Akanno, E. C., Crowley, J. J., Schenkel, F., Li, H., De Pauw, M., et al. (2016). Accuracy of genomic predictions for feed efficiency traits of beef cattle using 50K and imputed HD genotypes. *J. Anim. Sci.* 94, 1342–1353. doi: 10.2527/jas.2015-0126

Majumdar, S. G., Mishra, D. C., and Rai, A. (2020a). Effect of genotype imputation on integrated model for genomic selection. *J. Crop Weed* 16 (1), 133–137. doi: 10.22271/09746315.2020.v16.i1.1283

Majumdar, S. G., Rai, A., and Mishra, D. C. (2020b). Integrated framework for selection of additive and nonadditive genetic markers for genomic selection. *J. Comput. Biol.* 27, 845–855. doi: 10.1089/CMB.2019.0223

Meuwissen, T. H., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157 (4), 1819–1829. doi: 10.1093/genetics/157.4.1819

Mishra, D. C., Budhlakoti, N., Majumdar, S. G., and Rai, A. (2021). *Innovations in genomic Selection: Statistical perspective*. Special Proceedings: ISBN #: 978-81-950383-0-5 101–111.

Moore, J. H., and Williams, S. M. (2009). Epistasis and its implications for personal genetics. *Am. J. Hum. Genet.* 85, 309. doi: 10.1016/J.AJHG.2009.08.006

Piao, Z., Li, M., Li, P., Zhang, J., Zhu, C., Wang, H., et al. (2011). Bayesian Dissection for genetic architecture of traits associated with nitrogen utilization efficiency in rice. *Afr. J. Biotechnol.* 8, 6834–6839. doi: 10.4314/ajb.v8i24.68760

Poland, J., Endelman, J., Dawson, J., Rutkoski, J., Wu, S., Manes, Y., et al. (2012). Genomic selection in wheat breeding using genotyping-by-Sequencing. *Plant Genome* 5, 1–11. doi: 10.3835/PLANTGENOME2012.06.0006

R Core Team (2019). *R: A language and environment for statistical computing, r foundation for statistical computing*. Vienna, Austria. Available at: https://www.R-project.org/.

Sahebalam, H., Gholizadeh, M., Hafezian, H., and Farhadi, A. (2019). Comparison of parametric, semiparametric and nonparametric methods in genomic evaluation. *J. Genet.* 98, 1–8. doi: 10.1007/S12041-019-1149-3

Sehgal, D., Rosyara, U., Mondal, S., Singh, R., Poland, J., and Dreisigacker, S. (2020). Incorporating genome-wide association mapping results into genomic prediction models for grain yield and yield stability in CIMMYT spring bread wheat. *Front. Plant Sci.* 0. doi: 10.3389/FPLS.2020.00197

Spindel, J., Begum, H., Akdemir, D., Virk, P., Collard, B., Redoña, E., et al. (2015). Genomic selection and association mapping in rice (Oryza sativa): Effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. *PloS Genet.* 11, e1004982. doi: 10.1371/JOURNAL.PGEN.1004982

Tanaka, E. (2018) *Simple robust genomic prediction and outlier detection for a multi-environmental field trial*. Available at: https://arxiv.org/abs/1807.07268v1.

Wang, L., Sørensen, P., Janss, L., Ostersen, T., and Edwards, D. (2013). Genome-wide and local pattern of linkage disequilibrium and persistence of phase for 3 Danish pig breeds. *BMC Genet.* 14, 1–11. doi: 10.1186/1471-2156-14-115

Wiggans, G. R., Cole, J. B., Hubbard, S. M., and Sonstegard, T. S. (2017). Genomic selection in dairy cattle: The USDA experience∗. *Annu. Rev. Anim. Biosci.* 5, 309–327. doi: 10.1146/annurev-animal-021815-111422

Wolc, A., Zhao, H. H., Arango, J., Settar, P., Fulton, J. E., O'Sullivan, N. P., et al. (2015). Response and inbreeding from a genomic selection experiment in layer chickens. *Genet. Sel. Evol.* 47, 1–12. doi: 10.1186/S12711-015-0133-5

Yandell, B. S., Mehta, T., Banerjee, S., Shriner, D., Venkataraman, R., Moon, J. Y., et al. (2007). R/qtlbim: QTL with Bayesian interval mapping in experimental crosses. *Bioinformatics* 23, 641–643. doi: 10.1093/bioinformatics/btm011

Zhao, Y., Mette, M. F., and Reif, J. C. (2015). Genomic selection in hybrid breeding. *Plant Breed* 134, 1–10. doi: 10.1111/PBR.12231

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 14 December 2022
DOI 10.3389/fpls.2022.1064556

OPEN ACCESS

EDITED BY
Nisha Singh,
Gujarat Biotechnology University, India

REVIEWED BY
Abhinandan Surgonda Patil,
Agharkar Research Institute, India
Umesh K. Reddy,
West Virginia State University,
United States

*CORRESPONDENCE
Shyam Sundar Dey
shyam.iari@gmail.com;
shyam.dey@icar.gov.in
Anilabha Das Munshi
anilabhm@yahoo.co.in
Tusar Kanti Behera
tusar@rediffmail.com

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

RECEIVED 08 October 2022
ACCEPTED 17 November 2022
PUBLISHED 14 December 2022

CITATION
Devi S, Sharma PK, Behera TK, Jaiswal S, Boopalakrishnan G, Kumari K, Mandal NK, Iquebal MA, Gopala Krishnan S, Bharti, Ghosal C, Munshi AD and Dey SS (2022) Identification of a major QTL, *Parth6.1* associated with parthenocarpic fruit development in slicing cucumber genotype, Pusa Parthenocarpic Cucumber-6. *Front. Plant Sci.* 13:1064556. doi: 10.3389/fpls.2022.1064556

Identification of a major QTL, *Parth6.1* associated with parthenocarpic fruit development in slicing cucumber genotype, Pusa Parthenocarpic Cucumber-6

Shilpa Devi[1], Parva Kumar Sharma[2], Tusar Kanti Behera[1,3]*, Sarika Jaiswal[2], G. Boopalakrishnan[1], Khushboo Kumari[1], Neha Kumari Mandal[1], Mir Asif Iquebal[2], S. Gopala Krishnan[4], Bharti[5], Chandrika Ghosal[1], Anilabha Das Munshi[1]* and Shyam Sundar Dey[1]*

[1]Division of Vegetable Science, ICAR-Indian Agricultural Research Institute, New Delhi, India, [2]Centre for Agricultural Bioinformatics, ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India, [3]ICAR-Indian Institute of Vegetable Research, Varanasi, India, [4]Division of Genetics, ICAR-Indian Agricultural Research Institute, New Delhi, India, [5]Division of Sample Survey, ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India

Parthenocarpy is an extremely important trait that revolutionized the worldwide cultivation of cucumber under protected conditions. Pusa Parthenocarpic Cucumber-6 (PPC-6) is one of the important commercially cultivated varieties under protected conditions in India. Understanding the genetics of parthenocarpy, molecular mapping and the development of molecular markers closely associated with the trait will facilitate the introgression of parthenocarpic traits into non-conventional germplasm and elite varieties. The F_1, F_2 and back-crosses progenies with a non-parthenocarpic genotype, Pusa Uday indicated a single incomplete dominant gene controlling parthenocarpy in PPC-6. QTL-seq comprising of the early parthenocarpy and non-parthenocarpic bulks along with the parental lines identified two major genomic regions, one each in chromosome 3 and chromosome 6 spanning over a region of 2.7 Mb and 7.8 Mb, respectively. Conventional mapping using $F_{2:3}$ population also identified two QTLs, *Parth6.1* and *Parth6.2* in chromosome 6 which indicated the presence of a major effect QTL in chromosome 6 determining parthenocarpy in PPC-6. The flanking markers, SSR01148 and SSR 01012 for *Parth6.1* locus and SSR10476 and SSR 19174 for *Parth6.2* locus were identified and can be used for introgression of parthenocarpy through the marker-assisted back-crossing programme. Functional annotation of the QTL-region identified two major genes, *Csa_6G396640* and *Csa_6G405890* designated as probable indole-3-pyruvate monooxygenase YUCCA11 and Auxin response factor 16, respectively associated with auxin biosynthesis as potential candidate genes.

Csa_6G396640 showed only one insertion at position 2179 in the non-parthenocarpic parent. In the case of *Csa_6G405890*, more variations were observed between the two parents in the form of SNPs and InDels. The study provides insight about genomic regions, closely associated markers and possible candidate genes associated with parthenocarpy in PPC-6 which will be instrumental for functional genomics study and better understanding of parthenocarpy in cucumber.

KEYWORDS

cucumber, parthenocarpy, inheritance, QTL-seq, molecular mapping, candidate genes

Introduction

Cucumber (*Cucumis sativus* L.) is grown commercially in tropical and subtropical climates around the world (Pradeepkumara et al., 2022). In the Indian sub-continent, cucumber is grown from the highlands to the plains under open fields and protected conditions, including riverbeds. India is considered as the home of cucumber and has a wide range of genetic diversity and variation depending on growth habits, fruit size, fruit composition, and skin color besides several other agronomically important traits (Staub et al., 1997), but this variation has never been fully utilized in the crop improvement. The cultivated cucumber has a narrow genetic base with only 3-8% polymorphism within the cultivated genotypes, and 10-25% between plant species (Behera et al., 2011). The small, diploid genome (367 Mb), annual growth pattern, autogamous mating system, and relatively short life cycle (~ 3 months from generation to offspring) provide important genetic benefits (Wang et al., 2020) and detailed genomics-based studies in cucumber.

In most of the Angiosperms, fruit formation usually occurs after successful pollination followed by fertilization of eggs, which results in ovary growth, however, fruit development without pollination and fertilization is referred to as parthenocarpy (Li et al., 2014). Parthenocarpic fruits are seedless as ovules fertilization is disrupted due to changes in the basic genetic makeup involved in fertilization processes. Sources of genetic parthenocarpy are either obligate or facultative by nature. In sexually transmitted species, parthenocarpic genotypes need to be facultative in nature for successful development of fruits when pollinated. Alternatively, the obligate parthenocarpy can be found in asexually propagated plants (Gorguet et al., 2005). From a consumer perspective, parthenocarpy is a possible way to improve fruit quality and total productivity in several fruit crops. If seed setting fails, flower mortality is a common way to avoid wasting of resources. However, parthenocarpic genotypes are also found in wild or non-fruit species, indicating that there may be a variety of factors that cause the formation of seedless fruit in higher plants. The viability and permanence of parthenocarpy in a variety of plants is mainly the result of human selection (Varoquaux et al., 2000). Parthenocarpy in cucumber is determined by a complex interaction between several genetic factors and phytohormones (Sharif et al., 2022; Gou et al., 2022). Various phytohormones, especially gibberellins, cytokinins and auxins are involved in the processes that follow pollination and fertilization and these are essential factors for fruit and seed development (Fos et al., 2001). Growing seeds are major source of phytohormones that stimulate fruit growth and development (Ozga et al., 2002). The use of gynoecy in combination with parthenocarpy is necessary as cucumber exhibits facultative parthenocarpy as seeded fruit set can occur in parthenocarpic varieties when fertilised with viable pollen source. Gynoecious varieties are advantageous because of increased numbers of pistillate flowers, and thus greater opportunities for higher fruit set and per unit production. Parthenocarpic cucumber varieties offer several advantages over conventional seeded varieties. Parthenocarpic varieties are able to set fruits sequentially without suffering from first-fruit inhibition (Denna, 1973; Sun et al., 2006a). Parthenocarpy should be combined with stable gynoecious habit, because the fruits formed after fertilization of parthenocarpic plants become misshapen, have no economic value and lead to loss of productivity in case the female flowers received viable pollen. Selection of diverse genotypes to be used as a parent in the development of the F_1 hybrid to achieve higher yield, uniformity and suitability for protected cultivation (2016; 2017; Jat et al., 2015) is necessary in crop improvement programme. Therefore, the development of molecular markers closely associated with parthenocarpy and its marker-assisted introgression into diverse back-grounds is necessary to facilitate hybrid breeding programme.

Marker assisted selection (MAS) can enhance the efficiency of traditional breeding. In cucumber, breeding of parthenocarpic lines based on molecular markers provides a faster and more efficient way as selection can be based on genotypes itself rather than the phenotypes. There are two key requirements in successful of MAS, *i.e.* markers should be closely linked to target genes and a moderately saturated or high density genetic linkage map (Miao et al., 2011). The development of genetic linkage maps in cucumber have made possible for molecular characterization of important economic traits which includes fruit quality (Wenzel et al., 1995), resistance to diseases (Park et al., 2000; Zhang et al., 2010), yield (Serquen et al., 1997b; Fazio et al., 2003a), gynoecious sex and fruit colour (Miao et al., 2011) and yellow fruit flesh (Lu et al., 2015). Genetic studies have been largely inconsistent on the mode of inheritance for parthenocarpy in cucumber and have ranged from proposals of a single gene to complex multigenic inheritance (Pike and Peterson, 1969; De Ponti and Garretsen, 1976; El-Shawaf and Baker, 1981; Kim et al., 1992b). In the past, parthenocarpy has been studied by several workers to unravel the genetic and physiological basis of this extremely important trait (Fu et al., 2008; Li et al., 2014; Su et al., 2021; Gou et al., 2022; Mandal et al., 2022). The parthenocarpic genotypes of cucumber can set fruit without pollination however normal seed formation happens with successful pollination with viable pollen grains. This typical phenomenon in cucumber is attributed to the facultative parthenocarpic nature. The majority of the studies in the last two decades suggested that multiple QTLs across the genome are responsible for parthenocarpic fruit development in cucumbers. In a European greenhouse-slicing cucumber genotype, EC-1 parthenocarpy was found to be determined by one major and stable QTL in chromosome 2 (*Parth 2.1*) revealed through a $F_{2:3}$ population (Wu et al., 2015). In north American pickling type cucumber 2A, seven QTL were detected for parthenocarpy and one QTL each on chromosomes 5 and 7 (*parth5.1* and *parth7.1*) and two on chromosome 6 (*parth6.1* and *parth6.2*) were found govern parthenocarpy (Lietzow et al., 2016). Besides, in a south China ecotype cucumber, 4 novel QTLs associated with parthenocarpy were detected (Niu et al., 2020).

There is a broad consensus based on the available reports that parthenocarpic fruit set is complex in nature and genomic regions in different chromosomes are responsible for induction of parthenocarpic fruit development in cucumbers. The present study was conducted to identify and map the genomic regions associated with parthenocarpy in one of the commercially cultivated gynoecious parthenocarpic genotype, Pusa Parthenocarpic Cucumber-6 (PPC-6) through QTL-seq approach. Identification of closely linked PCR-based markers and possible identification of candidate genes associated with parthenocarpy in PPC-6 would facilitate the marker-assisted back-cross breeding and characterization of parthenocarpic trait in cucumber.

Materials and methods

Plant materials

The commercially cultivated parthenocarpic genotype, PPC-6 is cultivated widely in India under protected condition and was used as one of the parents for studying the inheritance and development of mapping population for parthenocarpic trait. In contrast, the non-parthenocarpic parent, Pusa Uday (PU), an Indian type cultivar suitable for cultivation under open field conditions was taken for the study. The F_1 progeny was developed by crossing the PU with PPC-6 under protected conditions. Development of F_1, F_2 and back-cross progenies were undertaken under protected conditions. The plants of inbreds and developed progenies were grown under protected conditions using the standard agronomic practices developed by the Division of Vegetable Science, ICAR-Indian Agricultural Research Institute, New Delhi.

Inheritance of parthenocarpy

The parthenocarpic line, PPC-6 was crossed with the non-parthenocarpic cultivar, Pusa Uday (PU). The resulting F_1 generation was selfed to obtain a sufficient number of F_2 population. The female flowers were covered with a butter paper bag, one day prior to anthesis to avoid cross-pollination, and pollens collected from the freshly opened male flowers were used for pollination. The observations were recorded for development of parthenocarpic fruits up to the 20^{th} node. If a plant produced parthenocarpic fruits up to the 1^{st} to 5^{th} nodes, it was considered an early parthenocarpic plant while if the fruit set occurred beyond the 10^{th} node, plants were categorized as late parthenocarpic. A total of 498 F_2 progenies were grown for recording observation on parthenocarpy and observation was recorded from 400 plants. The observations were recorded separately for early parthenocarpy, late parthenocarpy, and non-parthenocarpy. The goodness of fit of the observed values to the expected segregation ratio for parthenocarpic and non-parthenocarpic plants was tested using the classical Chi-square (χ^2) test as expressed below (Panse and Sukhatme, 1985):

$$\chi^2 = \frac{(Observed - Expected)^2}{Expected}$$

DNA extraction and whole genome resequencing

Approximately 15g of leaf samples were collected for DNA isolation from 30-35 day-old seedlings at the active vegetative stage during early morning hours. The collected leaf samples were packed in aluminum foil and labeled properly, then frozen

into liquid nitrogen and stored at -80°C for further use. Total DNA was isolated from the individual parental lines, F_1 hybrids, and mapping populations using the modified cetyl trimethyl ammonium bromide (CTAB) method (Saghai-Maroof et al., 1984). The genomic DNA samples were adjusted to 50ng DNA/µl and stored at 4°C until used as the templates for PCR amplification and sampling for sequencing. The quality and quantity of the extracted DNA were estimated with an Eppendrof Biospectrometer confirmed by running on 0.8% w/v agarose gel.

QTL-seq for identification of genomic regions associated with parthenocarpy

For QTL-seq analysis, 498 F_2 progenies derived from the crossing of the PU × PPC-6, were grown under polyhouse with a partially controlled environment along with their parents. Observation on parthenocarpy was recorded 45 days after sowing when the plants were in the full reproductive stage. One plant from each of the parents, PU and PPC-6 along with two extreme bulks constituting twenty plants each from the parthenocarpic and non-parthenocarpic types were used for sampling (Figure 1). Young leaves from each selected plant were used for the isolation of genomic DNA. After DNA isolation and purification, quantification was done using a Qubit (Thermofisher Scientific, USA). An equal quantity of DNA from each plant taken for bulking to constitute the final bulks.

Pre-processing of reads

Paired-end Illumina reads were obtained for both the parents and the bulks in duplicates. All the reads were 2*150 in length. FastQC version 0.11.8 (Andrews, 2010) was used for visualization of various read parameters and the presence of low-quality bases and of Illumina adapters. Based on the FastQC report the reads were cleaned and trimmed using Trimmomatic v0.39 (Bolger et al., 2014) with command parameters 'ILLUMINACLIP:TruSeq3-PE.fa:2:30:10:2:keepBothReads SLIDINGWINDOW:4:15 MINLEN:50'. Here, TruSeq3-PE.fa is a fasta file containing Illumina adapter sequences. The obtained high-quality reads were used further for the identification of QTL regions.

QTL-Seq analysis

QTL-Seq is a fast and efficient method to identify loci related to agronomically important traits (Takagi et al., 2013) in plants. Bulked Segregation Analysis (BSA) is the root of QTL-Seq that detects genomic location(s) showing significant variations in contrasting parents and progenies, produced by the contrasting parents. BSA relies on two main parameters, *namely*, SNP-index and Δ (SNP-index) (Abe et al., 2012; Takagi et al., 2013). SNP-index is the ratio of a number of reads having a variation, to the total number of reads at a particular position. While Δ (SNP-index) is the difference in the SNP indices of contrasting bulks. The range of SNP index varies from 0 to 1 depending on the parent chosen as a reference. For instance, if parent 1 is used as a reference, and the reads aligned at a particular locus do not have any variation this means that all the loci have been contributed from parent 1 and hence SNP-index = 0, but if all the reads aligned at a particular locus show variation, then it means that the loci have been contributed by another parent and therefore SNP-index = 1. For calculating the SNP-index, a sliding window approach was used with window size of 2000 kb and 100 kb increment followed by their averaging. Graphs were plotted for the Δ (SNP-index) against the chromosomal location. A Δ (SNP-

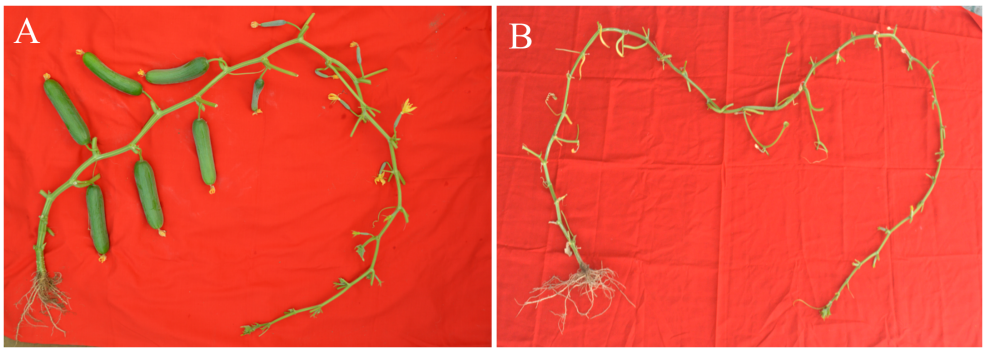


FIGURE 1
Fruit set in the contrasting parental lines **(A)** Pusa Parthenocarpic Cucumber-6 with parthenocarpic fruit development **(B)** Pusa Uday with no parthenocarpic fruit set under protected condition.

index) value close to zero, indicates that no significant QTL is present at that locus for the studied trait. To obtain significant results, statistical confidence intervals of Δ (SNP-index) were also plotted for all SNP positions with read depth assuming that there are no QTLs as null hypothesis at 95% level of significance (Takagi et al., 2013). We used QTL-seq version 2.2.2 for the detection of significant QTL regions for parthenocarpy in cucumber (Sugihara et al., 2020). The parameters used were [-n1 20 -n2 20 -o qtlseq_results -F 2 -e *Cucumis sativus*] where n1 and n2 are the numbers of individuals in each bulk, -o is the output directory, F is the filial population (here we had F_2) and -e is the dataset of reference genome for identification of the effects due to SNPs. All the other parameters were kept as default. This software uses BWA for the read alignment, SAMtools for filtering and BCF tools for variant calling (Figure 2).

Conventional mapping using $F_{2:3}$ progenies

Phenotyping was conducted in $F_{2:3}$ population developed through selfing of the individual F_2 progenies for conventional molecular mapping of the parthenocarpy. An single F_1 plant was selfed to obtain the F_2 population. An $F_{2:3}$ population comprising of 94 progenies derived from the cross between parthenocarpic and non-parthenocarpic parents was used to construct linkage map of cucumber. The $F_{2:3}$ progeny rows along with parental lines were raised under an insect-proof net house during Kharif, July-October, 2021. The $F_{2:3}$ progenies were grown in two replications with 10 plants in each replication for recording the observation on parthenocarpy. Eight female flowers were bagged one day prior to anthesis from the fifth node onwards on the main stem and eight more from the laterals. At 10 days after anthesis, well-developed and malformed fruits were counted as parthenocarpic fruit, whereas aborted ones were recorded as non-parthenocarpic (Figure 3) as suggested by Wu et al. (2016).

Genotyping was done using a large set of PCR-based markers uniformly distributed all throughout the cucumber genome. For the parental polymorphic survey, 1285 SSRs, Indels, CAPS markers were selected from the *Cucumis sativus* genome representing 7 linkage groups. In the present experiment, previously reported markers were used for the polymorphic study between the two genotypes, PPC-6 and PU (Miao et al., 2011; Zhu et al., 2016). Linkage analysis was performed using identified 123 polymorphic SSRs, Indels, CAPS markers for the construction of linkage map by IciMapping 4.1.0.0 at LOD threshold of 3.0 (Lander et al., 1987). Segregation of 123 SSRs, Indels, CAPS markers, and parthenocarpy was analyzed and genetic distance between markers was calculated using the Haldane and Kosambi map function (Kosambi, 1944). Parthenocarpy locus was mapped using Inclusive Composite Interval Mapping (ICIM) in ICIMapping 4.1.0.0 software (Wang et al., 2016).

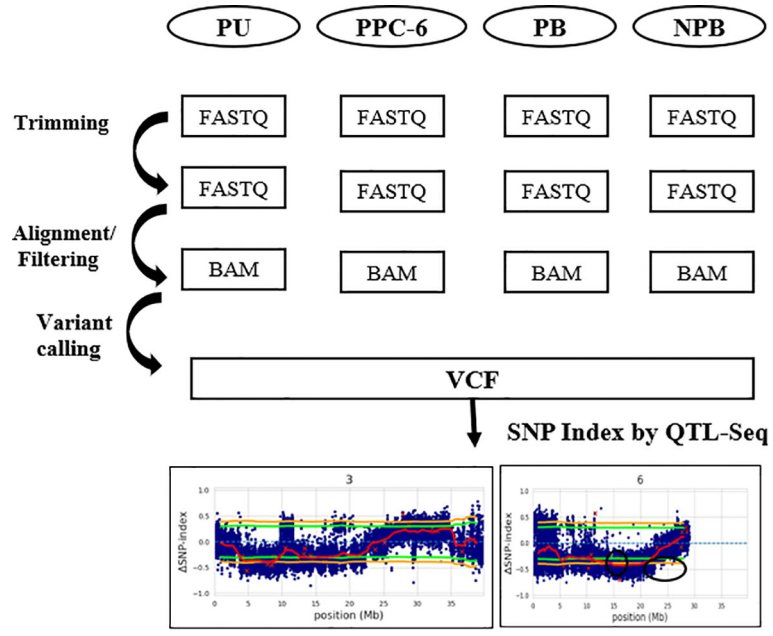


FIGURE 2
Schematic representation of the pipeline used for the QTL- Seq analysis.

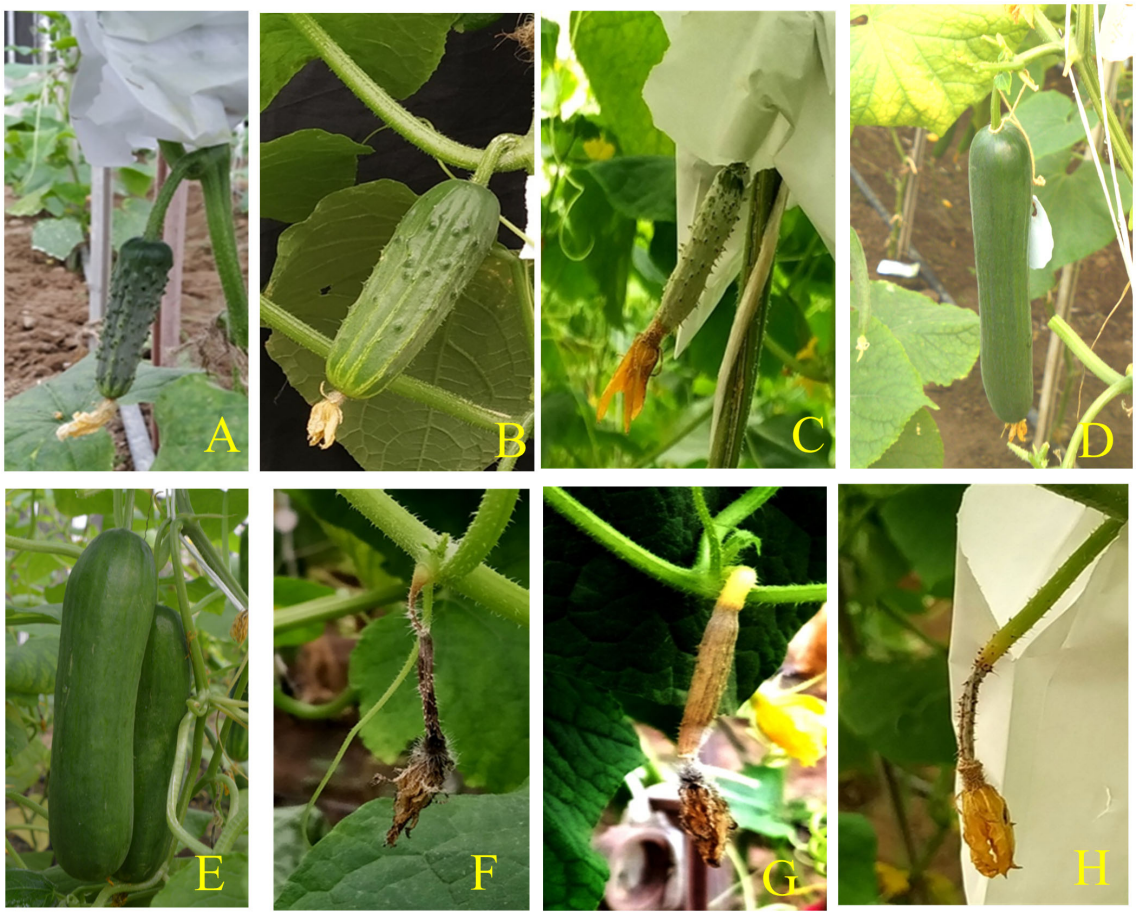


FIGURE 3
The pattern of fruit set in the segregating population at 10 days after anthesis **(A–H)**.

Results

Inheritance of parthenocarpy

In the present study, the inheritance pattern of parthenocarpy was studied based on the classical dominant-recessive Mendelian model by grouping the cucumber plants into three categories of their fruit development i.e. early parthenocarpic, late parthenocarpic, and non-parthenocarpic fruit development. This information would facilitate the adoption of appropriate breeding strategies for the development of stable parthenocarpic cucumber lines and will improve the efficiency of selection procedures. The genotype, PU produced non-parthenocarpic fruits and it was considered to be homozygous for non-parthenocarpic fruit development. The development of parthenocarpic fruits in PPC-6 is characterised by early parthenocarpy with fruit setting from the beginning or from the base of the plant. Therefore, PPC-6 was used as a homozygous genotype for parthenocarpic fruit development. The F_1 hybrid derived from the cross of PU × PPC-6 with heterozygous conditions produced some parthenocarpic fruits on the lower nodes, *i.e.* 10^{th} node and above (Supplementary Figure 1). In segregating F_2 individuals, early, late and non-parthenocarpic fruits were recorded. Out of 400 plants, 307 produced either early parthenocarpic or late parthenocarpic fruits and 93 plants were recorded as non-parthenocarpic plants. The $\chi 2$ value indicated a good fit for segregation of parthenocarpy (early, late and non-parthenocarpy) in the F_2 population populations confirmed with the expected ratio of 1:2:1 for early parthenocarpy, later parthenocarpy and non-parthenocarpy, respectively (Table 1). In the back cross progeny with the non-parthenocarpic genotype, PU the segregation for late parthenocarpy and early parthenocarpy were in the ratio of 1:1. Similarly, segregation of the plants for early parthenocarpy and late parthenocarpy was in the ratio of 1:1 for early and late parthenocarpy in the back-cross progenies with the parthenocarpic parent (Table 2).

TABLE 1 Evaluation of the parents along with F_1 and F_2 and back-cross progenies for studying the inheritance of parthenocarpy.

Crosses / Parents	Total number of plants	EP	LP	NP	Expected Ratio	χ^2 -value	P-Value
PU	10	0	0	10	–	–	–
PPC-6	10	10	0	0	–	–	–
PU × PPC-6 (F_1)	10	0	10	0	–	–	–
PU × PPC-6 (F_2)	400	87	220	93	1:2:1	4.18	0.12
(PU × PPC-6) × PU	60	0	26	34	1:1	1.06	0.31
(PU × PPC-6) × PPC-6	60	27	33	0	1:1	0.60	0.43

EP, Early parthenocarpy; LP, late parthenocarpy; NP, Non-parthenocarpy.

Pre-processing of reads

The raw reads for both the parents and bulks were subjected to quality check and removal of adapter sequences. After pre-processing, the non-parthenocarpic parent (NPP), PU retained 41383222 clean reads while parthenocarpic parent (PP), PCC-6 retained 41255039 clean reads. In case of both the bulks, *i.e.*, Parthenocarpic bulk (PB) and Non-Parthenocarpic bulk (NPB), 99904977 and 103354754 cleans reads were obtained, respectively (Table 1).

Identification of candidate genes in the QTL regions

After alignment and filtering of clean reads followed by variant calling using BWA, SAM tools and BCF tools, two QTL regions related to parthenocarpy were detected. The major QTL was detected on chromosome 6 while a minor QTL region was detected on chromosome 3 (Figure 4; Supplementary Table 1). SNP index of the parthenocarpic and non-parthenocarpic bulks is presented in Supplementary Figure 2. For both the regions, 99% confidence interval was considered. The region covered under the QTL region of chromosome 6 expanded from 13,500,000 till 21,300,000 and 7,000,000 till 9,700,000 for chromosome 3 (Figure 5; Table 3). A total of 5714 variants (SNP and Indels) were identified in chromosome 6 while we found 1129 variants on chromosome 3 (Supplementary Table 2). To view the effect of these variations on the protein sequence, we used the SNPeff software using *Cucumis sativus* database. All the identified variations were divided into several categories based on the impact on the protein. These categories are High, Moderate, Low and Modifier. Chromosome 3 has 4, 67, 110 and 948 SNPs in each category, respectively while chromosome 6 has 12, 183, 348 and 5171 SNPs, respectively (Figure 6; Supplementary Table 2).

Molecular mapping of parthenocarpy through conventional approaches

A total of 1285 SSR, Indel and CAPS markers were screened for parthenocarpy in polymorphic survey between PU and PPC-6 parental lines to identify polymorphic markers with ability to distinguish the parental lines. The markers were selected for all linkage groups of cucumber. A total of 1285 markers were used for parental polymorphic survey and among them 123 (11.28%) were polymorphic among the parental lines and produced clear and easily identifiable amplicons (Supplementary Table 3). The F_2 mapping population comprising of 94 individuals, were genotyped using selected polymorphic markers and respective $F_{2:3}$ progenies were evaluated for parthenocarpy. A total of 123 polymorphic markers were used for linkage analysis and others were rejected, due to non-amplification, missing data and difficulty in scoring. To confirm the parthenocarpic locus, specificity of the markers genotyping data of these 123 polymorphic markers were used for the construction of linkage map of parthenocarpic locus using Inclusive Composite Interval Mapping (ICIM) method of Ici Mapping (4.1.0.0) software at LOD threshold of 3.0 (Lander et al., 1987) (Figure 7). The linkage map of these polymorphic 123 SSRs,

TABLE 2 Summary of reads' statistics of the re-sequenced samples of the parents along with the contrasting bulks.

Sample	Input Read Pairs	Clean reads	% Clean reads
PU (NPP)	43632883	41383222	94.84412
PPC-6 (PP)	43799473	41255039	94.19072
Parthenocarpic bulk (Bulk 1)	105030659	99904977	95.11982
Non-parthenocarpic bulk (Bulk 2)	107216823	103354754	96.39789

NPP, non-parthenocarpic parent; PP, parthenocarpic parent.

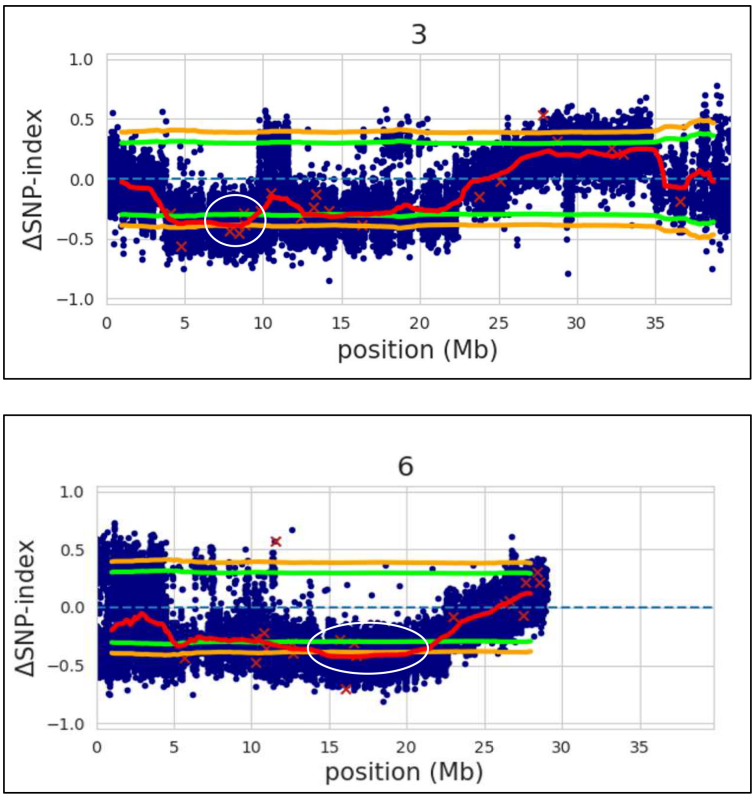


FIGURE 4
Δ(SNP-index) with statistical confidence intervals (orange, 99%; green, 95%). The major QTL identified on chromosome 6 and a minor QTL region on chromosome 3 of cucumber by QTL-seq (shown in white outline).

Indels, CAPS markers is presented in Supplementary Figure 3. Out of total 123 markers,10 markers were mapped on 1st linkage group, 20 on 2nd linkage group, 15 on 3rd linkage group, 12 on 4th linkage group, 17 on 5th linkage group, 30 on 6th linkage group and 19 on 7th linkage group (Supplementary Figure 4). Two major effect QTLs associated with parthenocarpy (*Parth6.1* and *Parth6.2*) were mapped to chromosome 6 (Figure 7). These two QTLs had LOD scores of 5.06, 4.59 and phenotypic variance of 16.69% and 12.93%, respectively. The additive effects of *Parth6.1* and *Parth6.2* were -13.71 and -12.49, respectively indicating the contribution of the parthenocarpy trait from male parent PPC-6. The markers flanking *Parth6.1* locus were, SSR 01148 and SSR 01012, spanning a distance of 5.0 cM. The markers flanking *Parth6.2* locus were SSR10476 and SSR 19174, spanning a distance of 5.0 cM. The results from this experiment depicted that markers SSR 01148, SSR 01012, SSR10476, and

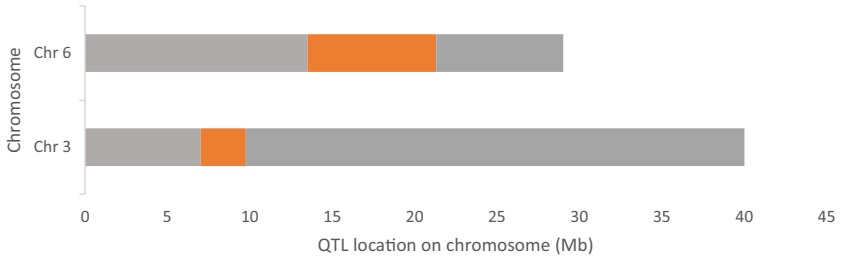


FIGURE 5
Chromosome 3 and 6 of cucumber showing the identified QTL region for parthenocarpy trait (in orange-coloured regions).

TABLE 3 Identified QTL regions in cucumber for parthenocarpy and their distribution in chromosome 3 and 6.

Chromosome	QTL	Start	End	Length (bp)	nSNPs
Parth3.1	1	7,000,000	9,700,000	2,700,000	1129
Parth6.1	2	13,500,000	21,300,000	7,800,000	5714

SSR 19174 on chromosome 6 are closely associated with parthenocarpic traits in cucumber.

Functional annotation of the identified QTL regions

To identify the major genes, present in these regions, a reference genome annotation file of cucumber (http://ftp.ebi.ac.uk/ensemblgenomes/pub/release-53/plants/gff3/cucumis_sativus) was used. A total of 998 genes were present in chromosome 6, while 485 genes were present in chromosome 3 of the QTL region (Supplementary Sheet 4). Among the identified genes, majority of them were under the category of hypothetical protein. Two genes, *Csa_6G396640* and *Csa_6G405890* designated as probable indole-3-pyruvate monooxygenase YUCCA11 and Auxin response factor 16, respectively were the potential candidate genes associated with auxin biosynthesis in plants which is crucial in parthenocarpic fruit development in cucumber. *Csa_6G396640* gene showed only one variation (insertion) at position 2179 in PU while in case of *Csa_6G405890*, more variations were observed between the two parents which includes both SNPs and few INDELs (Figure 8).

Discussion

Parthenocarpic fruit development in cucumber is extremely important for its cultivation under protected condition. Studies on inheritance of parthenocarpy by different workers depicted its complex genetics and number of genes/QTLs associated with resistance to parthenocarpy. In cucumber, parthenocarpy is facultative in nature and extent of parthenocarpy varies across different developmental stages of the plants (Joldersma and Liu, 2018). Pike and Peterson (1969) have reported that an incomplete dominant gene, *P* determines parthenocarpy in cucumber. They have postulated that development of early parthenocarpic fruits in the lower nodes is controlled by the dominant homozygous state, *PP* while late parthenocarpy and lower extent of parthenocarpy are represented by the heterozygous state, *Pp*. Whereas, homozygous recessive state, *pp* is responsible for non-parthenocarpic fruit development. In the homozygous condition, *PP* produces parthenocarpic fruits early, with the first developing generally by the fifth node. Heterozygous *Pp* plants produce parthenocarpic fruits later than homozygous plants and are fewer in number. The homozygous recessive *pp* produces no parthenocarpic fruits. Besides, several other studies have also reported monogenic control of parthenocarpy in cucumber (Hawthorn and Wellington, 1930; Kvasnikov et al., 1970; Juldasheva, 1973;

FIGURE 6
Distribution of SNP according to the effect in chromosome 3 and 6 of cucumber.

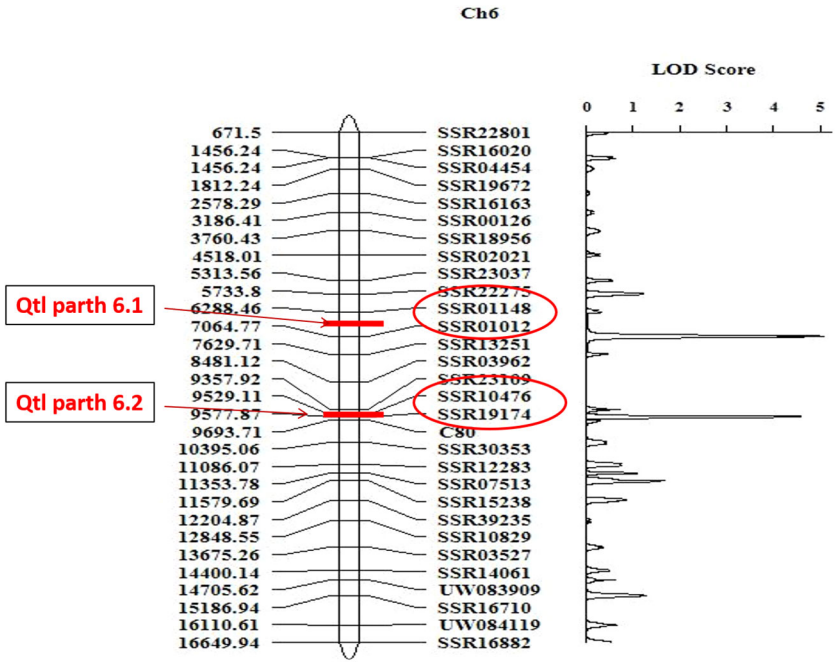


FIGURE 7
Linkage map of parthenocarpic locus on chromosome 6, constructed using SSR markers. Marker names, LOD score are depicted on the right side of the estimated map and the genetic distances shown in cM.

Meshcherov and Juldasheva, 1974). Later on, most of the studies reported polygenic control of parthenocarpy (El-Shawaf and Baker, 1981; Sun et al., 2006; Yan et al., 2009; Lietzow et al., 2016; Wu et al., 2016). In the present study, we have recorded that there was varied types of parthenocarpy in the F_2 progenies although the parthenocarpic parent, PPC-6 was early parthenocarpic type and started fruiting from 3-5 nodes onwards. We have recorded the plants as early parthenocarpic when fruiting started from 5th node onwards and late parthenocarpy when parthenocarpic fruit set was recorded after 10th node and considered all the early and late-type plants as parthenocarpic. Based on these observations, the parthenocarpy was found to be controlled by single recessive gene. The present information supported the earlier observation by Pike and Peterson (1969) which postulated that homozygous dominant, heterozygous and homozygous recessive forms are responsible early parthenocarpy, late parthenocarpy and non-parthenocarpy, respectively. However, it was evident there was a significant contribution of the back-ground evidenced from the

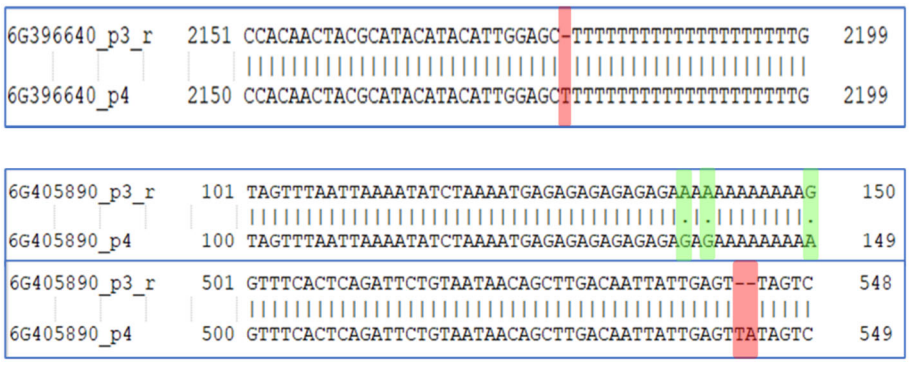


FIGURE 8
Variation in the *Csa_6G396640* and *Csa_6G405890* genes associated with indole-3-pyruvate monooxygenase YUCCA11 and Auxin response factor 16, respectively in the parents PU (p3_r) and PPC-6 (p4).

extent of parthenocarpy and therefore, it is possible to introgress parthenocarpy traits in different elite and non-conventional genotypes of cucumber through marker-assisted back-cross breeding with identification and development of molecular markers closely associated with parthenocarpic trait.

In the recent times, discovery of molecular markers and physical map construction is greatly facilitated by advancement in next-generation sequencing technology (Cao et al., 2021). QTL-seq combines the next-generation sequencing technology with BSA for rapid detection of QTLs for any particular trait and facilitate development of closely associated molecular markers and identification of candidate genes. Thereafter, QTL-seq has been widely used for the detection of QTLs, identification of closely linked molecular markers and identification of candidate genes for number of traits in different crops (Singh et al., 2016; Wang et al., 2016; Wei et al., 2016; Chen et al., 2017; Wen et al., 2019; Arikit et al., 2019; Li et al., 2020). In cucumber, QTL-seq has been used successfully for the identification of QTL for early flowering traits (Lu et al., 2014), flesh thickness (Xu et al., 2015), sub-gynoecy sex expression (Win et al., 2019), pre-harvest sprouting of the seeds (Cao et al., 2021) and resistance to powdery mildew (Zhang et al., 2021). Based on the QTL-seq results, two major QTLs, one each in chromosomes 3 and 6 were identified based on the Δ(SNP-index). The QTL, *Parth3.1* was *Parth6.1* were spanned 2.7 Mb on chromosome 3 and 7.8 Mb on chromosome 6, respectively. Besides, a large number of variants were detected in both the genomic regions in the form of SNPs and InDels. The identified SNPs and InDels in the genomic regions detected through QTL-seq will be extremely useful in the development of molecular markers and fine mapping of the genomic region associated with parthenocarpy in cucumbers. However, the QTL region identified through QTL-Seq are often not very precise and needs further validation and authentication through additional method of molecular mapping (Xu et al., 2017). Therefore, we have employed mapping of parthenocarpy through conventional $F_{2:3}$ population. Based on the QTL-seq results, two major QTLs, one each in chromosomes 3 and 6 were identified.

In cucumber, systematic efforts have been made for molecular mapping of number of qualitative traits but for quantitative traits like parthenocarpy progress is slow and hence very few public sector parthenocarpic varieties/hybrids are available in market. Now-a-days, role of marker-assisted selection (MAS) is increasing in conventional plant breeding (Miao et al., 2011). Due to narrow genetic base and low polymorphism, around 30 linkage maps have been constructed (Kennard et al., 1994). These linkage maps involved use of RAPDs (Randomly Amplified Polymorphic DNA) or AFLP (Amplified Fragment Length Polymorphism) (Fukino et al., 2008 and Yuan et al., 2008) that are not breeder friendly. Hence, co-dominant markers like SSR/InDel are best suited for marker-assisted breeding and are breeder friendly.

Two major QTLs for parthenocarpy at chromosome 6 (*Parth6.1* & *Parth 6.2*) were identified using the $F_{2:3}$ mapping population. Locus *Parth6.1* was flanked by SSR01148 and SSR01012 with a LOD score of 5.06 and 16.69% of PVE (Phenotypic Variance Explained) reflecting that this locus is a major effect QTL. Second QTL *Parth 6.2* was flanked by SSR10476 and SSR19174 primers and LOD value was 4.59 with 12.93% of PVE explaining another major effect QTL. Previously, Lietzow et al. (2016) identified seven QTLs associated with parthenocarpic fruit set, one on each chromosomes 5 and 7 (*parth5.1* and *parth7.1*) and two on chromosome 6 (*parth 6.1* and *parth 6.2*) were consistently identified in all experiments. Wu et al. (2016) identified seven novel QTLs on chromosomes 1, 2, 3, 5 and 7. The identification of QTLs is a valuable resource for cucumber breeders for the development of parthenocarpic cultivars (Dey et al., 2022). Molecular markers flanking major effect parthenocarpy QTLs can prove useful in the Marker Assisted Breeding (MAB) programme. However, there is need to further saturate linkage map to narrow-down genetic distance between flanking molecular markers to get markers better suited for foreground selection in endeavour of higher/quality production of cucumber.

Parthenocarpy is a complex trait and determined by interaction of large number of metabolic pathways interlinked with each other. Among the different metabolism, auxin, gibberellins and cytokinins are reported to play key role determining parthenocarpic fruit set in cucumber (Li et al., 2014; Su et al., 2021; Sharif et al., 2022; Gou et al., 2022). Cross-talk between the important phytohomones in determining parthenocarpy in PPC-6 was recently reported by Mandal et al. (2022). The QTL region identified through QTL-seq had two important genes with a possible association with parthenocarpic phenomenon. Indole-3-pyruvate monooxygenase YUCCA11 (*Csa_6G396640*) was found to be have one SNP in the non-parthenocarpic parent, PU when compared with the parthenocarpic reference genotype. Besides, the auxin response factor 16 (*Csa_6G405890*) present in the QTL region also showed variation in terms of Indels and SNPs in the parental lines. These two identified genes with key role in auxin biosynthesis could be possible candidate genes for induction of parthenocarpy in cucumber. Auxin, through its influence in cell division and expansion is key determinant in development of fleshy fruits and reported to be integral part in the initial signal for fertilisation and increased fruit (Godoy et al., 2021). After parthenocarpic fruit set their further development is influenced by auxins which was evidenced by the upregulation of the auxin biosynthesis-related genes in the later stages of fruit development in parthenocarpic genotypes in our earlier study (Mandal et al., 2022). Indole-3-pyruvate is one of the important routes for tryptophan-dependent auxin biosynthesis which is believed to be common in all plants (De Smet et al., 2011). Auxin is the key phytohormone besides gibberellins and cytokinin reported to play important role in the induction of parthenocarpy (Sharif et al., 2022). Among the different auxin biosynthesis pathways, the role of *Trp-IPyA* (tryptophan-indole-3-pyruvic acid) in

parthenocarpic fruit development has been reported by several workers. The role of the *YUCCA10*, *PavYUCCA10*, *SlTAR1*, *ToFZY2*, *ToFZY3* and *PARENTAL ADVICE-1 (PAD-1)* genes in parthenocarpic fruit development of loquat, tomato and eggplants have been reviewed in details by Sharif et al. (2022). However, narrow down of the QTL region through fine mapping is required for the precise identification of candidate genes associated with parthenocarpy in cucumber.

Conclusion

In cucumber parthenocarpic fruit set is extremely important trait facilitated large scale protected cultivation worldwide. In one commercially cultivated parthenocarpic genotype, PPC-6, it was found that, single incomplete dominant gene control this trait in spite of significant effect of genetic back-ground in expression of parthenocarpy. QTL-seq analysis in combination with conventional mapping using $F_{2:3}$ population identified one major effect QTLs, *Parth6.1*. The flanking markers, SSR01148 and SSR 01012 for Parth6.1 locus were identified for their use in marker-assisted back-crossing programme. Two major genes, *Csa_6G396640* and *Csa_6G405890* designated as probable indole-3-pyruvate monooxygenase YUCCA11 and Auxin response factor 16, respectively associated with auxin biosynthesis as potential candidate genes. The study provides insight about the genetics and genomic regions, closely associated markers and possible candidate genes associated with parthenocarpy in PPC-6 for functional genomics studies and future fine mapping.

Data availability statement

The data presented in the study are deposited in the NCBI repository, accession number PRJNA885599, https://www.ncbi.nlm.nih.gov/bioproject/PRJNA885599.

Author contributions

Conceived theme of the study and designed experiment: ShyD. Data curation: PS, ShyD, MI, SJ, GS. Investigation: ShiD, ShyD, KK, BG. Resources: ShyD, TB, AM. Supervision: ShyD, AM, TB, GS. Visualization: ShyD, TB, AM. Writing original draft: ShiD, ShyD, SJ. Review and editing: ShyD, TB, GS, MI. All authors contributed to the article and approved the submitted version.

Funding

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, Ms. Shilpa Devi. Authors are thankful to the CRP Molecular Breeding (Cucumber; 12-143 H) programme for funding the research. The research work was partially funded by the NAHEP-CAAST programme of Indian Council of Agricultural Research (ICAR).

Conflict of interest

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

Publisher's note

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.1064556/full#supplementary-material

SUPPLEMENTARY FIGURE 1
Fruit setting pattern in the F_1 hybrid involving Pusa Uday × Pusa Parthenocarpic Cucumber-6.

SUPPLEMENTARY FIGURE 2
SNP indices of **(A)** Parthenocarpic bulk (PB) and **(B)** Non-partyhenocarpic bulk (NPB) for each chromosomes.

SUPPLEMENTARY FIGURE 3
Linkage map of parth locus on chromosome 6, constructed using SSR markers. Marker names, LOD score are depicted on the right side of the estimated map and the genetic distances shown in cM on the left are calculated using software IciMapping ver. 4.1.0.

SUPPLEMENTARY FIGURE 4
Distribution of primers across the seven linkage groups and frequency of the amplified and polymorphic markers used for the study.

SUPPLEMENTARY FIGURE 5
Amplification pattern of the selected markers closely associated with parthenocarpy in cucumber genotype, PPC-6.

SUPPLEMENTARY FILE 1
Identified QTL regions in chromosome 3 and 6 through QTL-seq.

SUPPLEMENTARY FILE 2
Identified SNPs and InDels in the QTL regions of chromosome 3 and 6, their impact and associated genes.

SUPPLEMENTARY TABLE 3
List of the polymorphic markers used for construction of linkage map and molecular mapping of parthenocarpy using F2:3 population.

SUPPLEMENTARY FILE 4
Functional annotation of the genes present in the QTL regions in chromosome 3 and 6.

References

Abe, A., Kosugi, S., Yoshida, K., Natsume, S., Takagi, H., Kanzaki, H., et al. (2012). Genome sequencing reveals agronomically important loci in rice using MutMap. *Nat. Biotechnol.* 30 (2), 174–178. doi: 10.1038/nbt.2095

Andrews, S. (2010). *FastQC: a quality control tool for high throughput sequence data*. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Arikit, S., Wanchana, S., Khanthong, S., Saensuk, C., Thianthavon, T., Vanavichit, A., et al. (2019). QTL-seq identifies cooked grain elongation QTLs near soluble starch synthase and starch branching enzymes in rice (*Oryza sativa* l.). *Sci. Rep.* 9 (1), 1–10. doi: 10.1038/s41598-019-44856-2

Behera, T. K., Staub, J. E., Behera, S., Delannay, I. Y., and Chen, J. F. (2011). Marker-assisted backcross selection in an interspecific cucumis population broadens the genetic base of cucumber (*Cucumis sativus l.*). *Euphytica* 178, 261–272. doi: 10.1007/s10681-010-0315-8

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A flexible trimmer for illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Cao, M., Li, S., Deng, Q., Wang, H., and Yang, R. (2021). Identification of a major-effect QTL associated with pre-harvest sprouting in cucumber (*Cucumis sativus* l.) using the QTL-seq method. *BMC Genomics* 22, 249. doi: 10.1186/s12864-021-07548-8

Chen, F., Fu, B., Pan, Y., Zhang, C., Wen, H., Weng, Y., et al. (2017). Fine mapping identifies CsGCN5 encoding a histone acetyltransferase as putative candidate gene for tendril-less1 mutation (td-1) in cucumber. *Theor. Appl. Genet.* 130 (8), 1549–1558. doi: 10.1007/s00122-017-2909-1

Denna, D. W. (1973). Effect of genetic parthenocarpy and gynoecious flowering habit on fruit production and growth of cucumber (*Cucumis sativus* l.). *J. Am. Soc. Hortic. Sci.* 98, 602–604. doi: 10.21273/JASHS.98.6.602

De Ponti, O. M. B., and Garretsen, F. (1976). Inheritance of parthenocarpy in pickling cucumbers (*Cucumis sativus* l.) and linkage with other characters. *Euphytica* 25 (1), 633–642. doi: 10.1007/BF00041600

De Smet, I., Voß, U., Lau, S., Wilson, M., Shao, N., Timme, R. E., et al. (2011). Unraveling the evolution of auxin signaling. *Plant Physiol.* 155 (1), 209–221. doi: 10.1104/pp.110.168161

Dey, S. S., Singh, S., Munshi, A. D., and Behera, T. K. (2022). "Classical Genetics and Traditional Breeding. In: *The Cucumber Genome. Compendium of Plant Genomes*. Eds. Pandey, S., Weng, Y., Behera, T. K., and Bo, K. Cham: Springer. doi: 10.1007/978-3-030-88647-9_12

El-Shawaf, I. I. S., and Baker, L. R. (1981). Inheritance of parthenocarpic yield in gynoecious pickling cucumber for once-over mechanical harvest by diallel analysis of six gynoecious lines. *J. Am. Soc Hortic. Sci.* 106, 359–364. doi: 10.21273/JASHS.106.3.359

Fazio, G., Staub, J. E., and Stevens, M. R. (2003a). Genetic mapping and QTL analysis of horticultural traits in cucumber (*Cucumis sativus* l.) using recombinant inbred lines. *Theor. Appl. Genet.* 107, 864–874. doi: 10.1007/s00122-003-1277-1

Fos, M., Proaño, K., Nuez, F., and García-Martínez, J. L. (2001). Role of gibberellins in parthenocarpic fruit development induced by the genetic system pat-3/pat-4 in tomato. *Physiol. Plant.* 111 (4), 545–550. doi: 10.1034/j.1399-3054.2001.1110416.x

Fukino, N., Ohara, T., Monforte, A. J., Sugiyama, M., Sakata, Y., Kunihisa, M., et al. (2008). Identification of QTLs for resistance to powdery mildew and SSR markers diagnostic for powdery mildew resistance genes in melon (*Cucumis melo* l.). *Theor. Appl. Genet.* 118 (1), 165–175. doi: 10.1007/s00122-008-0885-1

Fu, F. Q., Mao, W. H., Shi, K., Zhou, Y. H., Asami, T., and Yu, J. Q. (2008). A role of brassinosteroids in early fruit development in cucumber. *J. Exp. Bot.* 59, 2299–2308. doi: 10.1093/jxb/ern093

Godoy, F., Kühn, N., Muñoz, M., Marchandon, G., Gouthu, S., Deluc, L., et al. (2021). The role of auxin during early berry development in grapevine as revealed by transcript profiling from pollination to fruit set. *Hortic. Res.* 8, 140. doi: 10.1038/s41438-021-00568-1

Gorguet, B., Van Heusden, A. W., and Lindhout, P. (2005). Parthenocarpic fruit development in tomato. *Plant Biol.* 7 (2), 131–139. doi: 10.1055/s-2005-837494

Gou, C., Zhu, P., Meng, Y., Yang, F., Xu, Y., Xia, P., et al. (2022). Evaluation and genetic analysis of parthenocarpic germplasms in cucumber. *Genes (Basel)* 13 (2), 225. doi: 10.3390/genes13020225

Hawthorn, L. R., and Wellington, R. (1930). Geneva, A greenhouse cucumber that develops fruit without pollination. *Agr. Exp. Stat. Bull.* 580, 1–11. Available at: https://hdl.handle.net/1813/4556.

Jat, G. S., Munshi, A. D., Behera, T. K., and Bharadwaj, C. (2017). Inheritance of parthenocarpy in gynoecious cucumber (*Cucumis sativus* l.) cultivar PPC-2. *J. Hortic. Sci.* 12 (2), 193–197. Available at: https://jhs.iihr.res.in/index.php/jhs/article/view/23.

Jat, G. S., Munshi, A. D., Behera, T. K., Choudhary, H., and Dev, B. (2015). Exploitation of heterosis in cucumber for earliness, yield and yield components utilizing gynoecious lines. *Indian J. Hortic.* 72 (4), 494–499. doi: 10.5958/0974-0112.2015.00112.7

Jat, G. S., Munshi, A. D., Behera, T. K., and Tomar, B. S. (2016). Combining ability estimation of gynoecious and monoecious hybrids for yield and earliness in cucumber (*Cucumis sativus*). *Indian J. Agric. Sci.* 86 (3), 399–403. Available at: https://epubs.icar.org.in/index.php/IJAgS/article/view/57033.

Joldersma, D., and Liu, Z. (2018). The making of virgin fruit: the molecular and genetic basis of parthenocarpy. *J. Exp. Bot.* 69 (5), 955–962. doi: 10.1093/jxb/erx446

Juldasheva, L. M. (1973). Inheritance of the tendency towards parthenocarpy in cucumbers. *Byull. Vseso yuznogo ordena Lenina Inst. Rastwni evodstva Imeni N.I. Vavilova* 32, 58–59. doi: 10.1007/BF00041600

Kennard, W. C., Slocum, M. K., Figdore, S. S., and Osborn, T. C. (1994). Genetic analysis of morphological variation in brassica oleracea using molecular markers. *Theor. Appl. Genet.* 87, 721–732. doi: 10.1007/BF00222898

Kim, S., Okubo, H., and Fujieda, K. (1992b). Endogenous levels of IAA in relation to parthenocarpy in cucumber (*Cucumis sativus* l.). *Scientia Hortic.* 52, 1–8. doi: 10.1016/0304-4238(92)90002-T

Kosambi, D. D. (1944). The estimation of map distance from recombination values. *Ann. Eugen* 12, 172–175. doi: 10.1111/j.1469-1809.1943.tb02321.x

Kvasnikov, B. V., Rogova, N. T., Tarakanova, S. I., and Ignatova, S. I. (1970). Methods of breeding vegetable crops under the covered ground. *Tmtdyprikl. Bot. Genet. Selek.* 42, 45–57. Available at: https://agris.fao.org/agris-search/search.do?recordID=US201302372125.

Lander, E. S., Green, P., Abrahamson, J., Barlow, A., Daly, M. J., and Lincoln, S. E. (1987). MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics* 1, 174–181. doi: 10.1016/0888-7543(87)90010-3

Lietzow, C. D., Zhu, H., Pandey, S., Havey, M. J., and Weng, Y. (2016). QTL mapping of parthenocarpic fruit set in north American processing cucumber. *Theor. Appl. Genet.* 129 (12), 2387–2401. doi: 10.1007/s00122-016-2778-z

Li, Z., Han, Y., Niu, H., Wang, Y., Jiang, B., and Weng, Y. (2020). Gynoecy instability in cucumber (*Cucumis sativus* l.) is due to unequal crossover at the copy number variation-dependent femaleness (F) locus. *Hortic. Res.* 7, 32. doi: 10.1038/s41438-020-0251-2

Li, J., Wu, Z., Cui, L., Zhang, T., Guo, Q., Xu, J., et al. (2014). Transcriptome comparison of global distinctive features between pollination and parthenocarpic fruit set reveals transcriptional phytohormone crosstalk in cucumber (*Cucumis sativus* l.). *Plant Cell Physiol.* 55, 1325–1342. doi: 10.1093/pcp/pcu051

Lu, H., Lin, T., Klein, J., Wang, S., Qi, J., Zhou, Q., et al (2014). QTL-seq identifies an early flowering QTL located near Flowering Locus T in cucumber. *Theor. Appl. Genet.* 127 (7), 1491–1499. doi: 10.1007/s00122-014-2313-z

Lu, H., Miao, H., Tian, G., Wehner, T., Gu, X., and Zhang, S. (2015). Molecular mapping and candidate gene analysis for yellow fruit flesh in cucumber. *Molecular Breeding* 35, 1–8. doi: 10.1007/s11032-015-0263-z

Mandal, N. K., Kumari, K., Kundu, A., Arora, A., Bhowmick, P. K., Iquebal, M. A., et al. (2022). Cross-talk between the cytokinin, auxin, and gibberellin regulatory networks in determining parthenocarpy in cucumber. *Front. Genet* 13, 957360. doi: 10.3389/fgene.2022.957360

Meshcherov, E. T., and Juldasheva, L. W. (1974). Parthenocarpy in cucumber. *Proc. Appl. Bot. Plant Breed.* 51, 204–213.

Miao, H., Zhang, S., Wang, X., Zhang, Z., Li, M., Mu, S., et al. (2011). A linkage map of cultivated cucumber (*Cucumis sativus* l.) with 248 microsatellite marker loci and seven genes for horticulturally important traits. *Euphytica* 182, 167–176. doi: 10.1007/s10681-011-0410-5

Niu, Z. H., Song, X. F., Li, X. L., Guo, X. Y., He, S. Q., He, L. J. Z., et al. (2020). Inheritance and QTL mapping for parthenocarpy in cucumber. *Sci. Agric. Sin.* 53, 160–171. doi: 10.3864/j.issn.0578-1752.2020.01.015

Ozga, J. A., Van, H. R., and Reinecke, D. M. (2002). Hormone and seed-specific regulation of pea fruit growth. *Plant Physiol.* 128 (4), 1379–1389. doi: 10.1104/pp.010800

Panse, V. G., and Sukhatme, P. V. (1985). *Statistical methods for agricultural workers* (New Delhi: ICAR).

Park, Y., Sensoy, S., Wye, C., Antonise, R., Peleman, J., and Havey, M. J. (2000). A genetic map of cucumber composed of RAPDs, RFLPs, AFLPs, and loci conditioning resistance to papaya ringspot and zucchini yellow mosaic viruses. *Genome* 43, 1003–1010. doi: 10.1139/g00-075

Pike, L. M., and Peterson, C. E. (1969). Inheritance of parthenocarpy in the cucumber (*Cucumis sativus* l.). *Euphytica* 18, 101–105. doi: 10.1007/BF00021987

Pradeepkumara, N., Sharma, P. K., Munshi, A. D., Behera, T. K., Bhatia, R., Kumari, K., et al. (2022). Fruit transcriptional profiling of the contrasting genotypes for shelf life reveals the key candidate genes and molecular pathways regulating post-harvest biology in cucumber. *Genomics* 114 (2), 110273. doi: 10.1016/j.ygeno.2022.110273

Saghai-Maroof, M. A., Solimann, K. M., 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.* 81, 8014–8018. doi: 10.1073/pnas.81.24.8014

Serquen, F. C., Bacher, J., and Staub, J. E. (1997b). Mapping and QTL analysis of a narrow cross in cucumber (*Cucumis sativus* l.) using random amplified polymorphic DNA markers. *Mol. Breed.* 3, 257–268. doi: 10.1023/A:1009689002015

Sharif, R., Su, L., Chen, X., and Qi, X. (2022). Hormonal interactions underlying parthenocarpic fruit formation in horticultural crops. *Hortic. Res.* 9, uhab024. doi: 10.1093/hr/uhab024

Singh, V. K., Khan, A. W., Jaganathan, D., Thudi, M., Roorkiwal, M., Takagi, H., et al. (2016). QTL-seq for rapid identification of candidate genes for 100-seed weight and root/total plant dry weight ratio under rainfed conditions in chickpea. *Plant Biotechnol. J.* 14 (11), 2110–2119. doi: 10.1111/pbi.12567

Staub, J. E., Serquen, F. C., and Mccreight, J. D. (1997). Genetic diversity in cucumber (Cucumis sativus l.): III. an evaluation of Indian germplasm. *Genet. Resour. Crop Evol.* 44 (4), 315–326. doi: 10.1023/A:1008639103328

Sugihara, Y., Young, L., Yaegashi, H., Natsume, S., Shea, D. J., Takagi, H., et al. (2020). High performance pipeline for MutMap and QTL seq. *PeerJ* 10, e13170. doi: 10.7717/peerj.13170

Sun, Z., Lower, R. L., and Staub, J. E. (2006a). Analysis of generation means and components of variance for parthenocarpy in cucumber (*Cucumis sativus* l.). *Plant Breed.* 123 (3), 277–280. doi: 10.1111/j.1439-0523.2006.01224.x

Sun, Z., Staub, J. E., Chung, S. M., and Lower, R. L. (2006). Identification and comparative analysis of quantitative trait loci associated with parthenocarpy in processing cucumber. *Plant Breed.* 125, 281–287. doi: 10.1111/j.1439-0523.2006.01225

Su, L., Rahat, S., Ren, N., Kojima, M., Takebayashi, Y., Sakakibara, H., et al. (2021). Cytokinin and auxin modulate cucumber parthenocarpy fruit development. *Sci. Hortic.* 282, 110026. doi: 10.1016/j.scienta.2021.110026

Takagi, H., Abe, A., Yoshida, K., Kosugi, S., Natsume, S., Mitsuoka, C., et al. (2013). QTL-seq: rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations. *Plant J.* 74 (1), 174–183. doi: 10.1111/tpj.12105

Varoquaux, F., Blanvillain, R., Delseny, M., and Gallois, P. (2000). Less is better: new approaches for seedless fruit production. *Trends Biotechnol.* 18 (6), 233–242. doi: 10.1016/S0167-7799(00)01448-7

Wang, J., Li, H., Zhang, L., and Meng, L. (2016). Users' Manual of QTL IciMapping. The Quantitative Genetics Group, Institute of Crop Science, Chinese Academy of Agricultural Sciences (CAAS), Beijing 100081, China and Genetic Resources Program, International Maize and Wheat Improvement Center (CIMMYT). (Mexico, D.F., Mexico) *Apdo. Postal* 6-641, 06600.

Wang, Y., Bo, K., Gu, X., Pan, J., Li, Y., Chen, J., et al. (2020). Molecularly tagged genes and quantitative trait loci in cucumber with recommendations for QTL nomenclature. *Hortic. Res.* 7, 3. doi: 10.1038/s41438-019-0226-3

Wei, Q. Z., Fu, W. Y., Wang, Y. Z., Qin, X. D., Wang, J., Li, J., et al. (2016). Rapid identification of fruit length loci in cucumber (*Cucumis sativus* l.) using next-generation sequencing (NGS)-based QTL analysis. *Sci. Rep.* 6 (1), 1–11. doi: 10.1038/srep27496

Wenzel, G., Kennard, W. C., and Havey, M. J. (1995). Quantitative trait analysis of fruit quality in cucumber: QTL detection, confirmation, and comparison with mating-design variation. *Theor. Appl. Genet.* 91, 53–61. doi: 10.1007/BF00220858

Wen, C., Zhao, W., Liu, W., Yang, L., Wang, Y., Liu, X., et al. (2019). CsTFL1 inhibits determinate growth and terminal flower formation through interaction with CsNOT2a in cucumber. *Development* 146 (14), 180166. doi: 10.1242/dev.180166

Win, K. T., Zhang, C., Silva, R. R., Lee, J. H., Kim, Y. C., and Lee, S. (2019). Identification of quantitative trait loci governing subgynoecy in cucumber. *Theor. Appl. Genet.* 132 (5), 1505–1521. doi: 10.1007/s00122-019-03295-3

Wu, Z., Li, L., Zhang, T., Zhang, T. L., Li, J., Lou, Q. F., et al. (2015). QTL mapping for parthenocarpy in cucumber. *Sci. Agri. Sinica* 48, 112–119.

Wu, Z., Zhang, T., Li, L., Xu, J., Qin, X., Zhang, T., et al. (2016). Identification of a stable major-effect QTL (Parth 2.1) controlling parthenocarpy in cucumber and associated candidate gene analysis via whole genome re-sequencing. *BMC Plant Biol.*, 16–182. doi: 10.1186/s12870-016-0873-6

Xu, X., Lu, L., Zhu, B., Xu, Q., Qi, X., and Chen, X. (2015). QTL mapping of cucumber fruit flesh thickness by SLAF-seq. *Sci. Rep.* 5 (1), 1–9. doi: 10.1038/srep15829

Xu, Q., Xu, X., Shi, Y., Qi, X., and Chen, X. (2017). Elucidation of the molecular responses of a cucumber segment substitution line carrying *Pm5.1* and its recurrent parent triggered by powdery mildew by comparative transcriptome profiling. *BMC Genomics* 18, 21. doi: 10.1186/s12864-016-3438-z

Yan, L., Luo, L., Feng, Z., Li, X., Lou, Q., and Chen, J. (2009). Analysis on mixed major gene and polygene inheritance of parthenocarpy in monoecious cucumber (*Cucumis sativus* l). *Acta Botanica Boreali Occidentalia Sin.* 29 (6), 1122–1126.

Yuan, X. J., Li, X. Z., Pan, J. S., Wang, G., Jiang, S., Li, X. H., et al. (2008). Genetic linkage map construction and location of QTLs for fruit-related traits in cucumber. *Plant Breed.* 127 (2), 180–188. doi: 10.1111/j.1439-0523.2007.01426.x

Zhang, S. P., Miao, H., Gu, X. F., Yang, Y. H., Xie, B. Y., Wang, X. W., et al. (2010). Genetic mapping of the scab resistance gene *Ccu* in cucumber. *J. Am. Soc Hortic. Sci.* 135, 53–58. doi: 10.21273/JASHS.135.1.53

Zhang, C., Badri Anarjan, M., Win, K. T., Begum, S., and Lee, S. (2021). QTL-seq analysis of powdery mildew resistance in a Korean cucumber inbred line. *Theor. Appl. Genet.* 134 (2), 435–451. doi: 10.1007/s00122-020-03705-x

Zhu, W. Y., Huang, L., Chen, L., Yang, J. T., Wu, J. N., Qu, M. L., et al. (2016). A high-density genetic linkage map for cucumber (Cucumis sativus l.): based on specific length amplified fragment (SLAF) sequencing and QTL analysis of fruit traits in cucumber. *Front. Plant Sci.* 7, 437. doi: 10.3389/fpls.2016.00437

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 20 December 2022
DOI 10.3389/fpls.2022.1056028

OPEN ACCESS

EDITED BY
Nisha Singh,
Gujarat Biotechnology University, India

REVIEWED BY
Nazia Rehman,
National Agricultural Research Centre, Pakistan
Maharajan Theivanayagam,
Rajagiri College of Social Sciences, India
Reetika Mahajan,
Sher-e-Kashmir University of Agricultural Sciences and Technology, India

*CORRESPONDENCE
Gurpreet Kaur
gurpreetkaur@pau.edu

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

RECEIVED 28 September 2022
ACCEPTED 29 November 2022
PUBLISHED 20 December 2022

CITATION
Upadhyay P, Gupta M, Sra SK, Sharda R, Sharma S, Sardana VK, Akhatar J and Kaur G (2022) Genome wide association studies for acid phosphatase activity at varying phosphorous levels in *Brassica juncea* L.
Front. Plant Sci. 13:1056028.
doi: 10.3389/fpls.2022.1056028

Genome wide association studies for acid phosphatase activity at varying phosphorous levels in *Brassica juncea* L

Priyanka Upadhyay[1], Mehak Gupta[1], Simarjeet Kaur Sra[1], Rakesh Sharda[2], Sanjula Sharma[1], Virender K. Sardana[1], Javed Akhatar[1] and Gurpreet Kaur[1]*

[1]Department of Plant Breeding and Genetics, Punjab Agricultural University, Ludhiana, India, [2]Department of Soil & Water Engineering, Punjab Agricultural University, Ludhiana, India

Acid phosphatases (Apases) are an important group of enzymes that hydrolyze soil and plant phosphoesters and anhydrides to release Pi (inorganic phosphate) for plant acquisition. Their activity is strongly correlated to the phosphorus use efficiency (PUE) of plants. Indian mustard (*Brassica juncea* L. Czern & Coss) is a major oilseed crop that also provides protein for the animal feed industry. It exhibits low PUE. Understanding the genetics of PUE and its component traits, especially Apase activity, will help to reduce Pi fertilizer application in the crop. In the present study, we evaluated 280 genotypes of the diversity fixed foundation set of Indian mustard for Apase activity in the root (RApase) and leaf (LApase) tissues at three- low (5μM), normal (250μM) and high (1mM) Pi levels in a hydroponic system. Substantial effects of genotype and Pi level were observed for Apase activity in both tissues of the evaluated lines. Low Pi stress induced higher mean RApase and LApase activities. However, mean LApase activity was relatively more than mean RApase at all three Pi levels. JM06016, IM70 and Kranti were identified as promising genotypes with higher LApase activity and increased R/S at low Pi. Genome-wide association study revealed 10 and 4 genomic regions associated with RApase and LApase, respectively. Annotation of genomic regions in the vicinity of peak associated SNPs allowed prediction of 15 candidates, including genes encoding different family members of the acid phosphatase such as *PAP10* (purple acid phosphatase 10), *PAP16*, *PNP* (polynucleotide phosphorylase) and *AT5G51260* (HAD superfamily gene, subfamily IIIB acid phosphatase) genes. Our studies provide an understanding of molecular mechanism of the Apase response of *B. juncea* at varying Pi levels. The identified SNPs and candidate genes will support marker-assisted breeding program for improving PUE in Indian mustard. This will redeem the crop with enhanced productivity under restricted Pi reserves and degrading agro-environments.

KEYWORDS

Indian mustard, acid phosphatase, phosphorus use efficiency, SNP genotyping, marker trait associations

1 Introduction

Phosphorus (P) is one of the most crucial macronutrients required for the optimal growth and development of plants (Han et al., 2022; Bhadouria and Giri, 2022). This element is a prime structural constituent of cell biomolecules, including ATP, NADPH, phospholipids, and nucleic acids (Kohli et al., 2020). Significant yield losses in cereal (Plenet et al., 2000; Wissuwa and Ae, 2001; Lazaro et al., 2010), pulse (Bonser et al., 1996; Mahamood et al., 2009), fodder (Camacho et al., 2002; Ceasar et al., 2014 and oilseed crops (Bastani and Hajiboland, 2017; Grzebisz et al., 2018) have been reported under P deficient conditions (Maharajan et al., 2018). Surprisingly, majority (71%) of the global cropland area has a surplus of P, whereas 29% is in a state of P deficiency (Thudi et al., 2021). However, soils that carry ample P to support plant growth also tend to show P deficiency (Raghothama and Karthikeyan, 2005). It is due to the low (1-10 μM) availability of soluble inorganic phosphate (Pi; $H_2PO_4^-$ or HPO_4^{2-}), the only forms that plants can absorb and assimilate (Shen et al., 2011). Pi has high propensity to form insoluble complexes with metal cations such as aluminum and iron in acidic soils and calcium and magnesium in alkaline soils, which render Pi unavailable to plants (Bhadouria and Giri, 2022). Also, an abundant amount of P (30–80% of total P) remains fixed in soil as Po (organophosphate) and becomes unavailable for root acquisition unless hydrolyzed by enzymes to liberate absorbable Pi (Richardson and Simpson, 2011; Dissanayaka et al., 2018). Phosphate fertilizers are routinely applied to alleviate Pi deficiency and maintain crop yields and quality. It is estimated that there is a requirement for 51–86% more Pi inputs by 2050 to sustain global food production (Mogollon et al., 2018). At the same time, phosphate rock, a non-renewable Pi resource, is diminishing at a faster rate and may be completely depleted in the next 100–400 years (Gilbert et al., 2009; Desmidt et al., 2015). Furthermore, excessive use of inorganic fertilizers has led to potential environmental problems as most of the applied Pi is not recovered by crops (Syers et al., 2008). Most crop plants absorb less than 20% of applied Pi (Plaxton and Tran, 2011). A significant amount of Pi coprecipitates and may run off from the soil to surface waters, resulting in aquatic eutrophication (Zak et al., 2018). In view of these concerns, there is a need to divert efforts towards engineering crop cultivars that acquire and utilize Pi more efficiently to produce higher yields under Pi limited conditions (Cong et al., 2020).

Abbreviations: Apase, acid phosphatase; LApase, leaf acid phosphatase; RApase, root acid phosphatase; MTAs, marker trait associations; P, phosphorus; Pi, inorganic phosphate; PUE, phosphorus use efficiency; LP, low Pi level; NP, normal Pi level; HP, high Pi level; PCA, principal component analysis; PCs, principal components.

Plants elicit a series of alterations at morphological, physio-chemical and molecular levels such as changes in root architecture, root to shoot ratio, membrane structure, anthocyanin accumulation, secretion of organic acids (malate, citrate, oxalate etc.) and hydrolases (phospholipases, ribonucleases, acid phosphatases etc.) and activation of various Pi stress response genes to increase Pi acquisition and utilization to sustain plant growth under Pi limited conditions (Plaxton and Tran, 2011; Ryan et al., 2014; Li et al., 2016). Collectively, these response mechanisms induced to maintain plant Pi homeostasis are known as Pi starvation response (PSR). Amongst these adaptive responses, induction and secretion of acid phosphatase (Apase) enzymes is a universal response. They catalyze hydrolysis of organic P complexes (phosphoesters and anhydrides) in acidic soils and plant tissues to release soluble Pi for plant utilization (Raghothama, 1999; Nannipieri et al., 2011; Gu et al., 2016). Significant positive correlations between Apase activity and PUE (phosphorus use efficiency) have been recorded (Chen et al., 2003; Radersma and Grierson, 2004; Zhang et al., 2010). In rice, overexpression of Apase genes (*OsPAP10a, OsPAP10c, and OsPAP21b*) significantly increased the hydrolysis and utilization of externally supplied Po and ATP under low Pi conditions (Tian et al., 2012; Lu et al., 2016; Mehra et al., 2017; Deng et al., 2020). Thus, genetic manipulation of Apase activity is of high interest for improving the PUE of plants (Han et al., 2022). Purple acid phosphatases (PAPs) are the most studied class of Apases in relation to P homeostasis. To date, the identification of PAPs has been completed for several plant species (Bhadouria and Giri, 2022). They are also known to be present in bacteria and animals, executing a similar function (Wang et al., 2021; Bhadouria and Giri, 2022). In plants, PAPs occur as a multigene family. There are 29 members of PAPs in *Arabidopsis thaliana*, 26 in *Oryza sativa*, 33 in *Zea mays ssp. mays* var. B73, 38 in *Glycine max*, 25 in *Cicer arietinum*, 19 in *Camellia sinensis* and 25 in *Jatropha curcas* (Li et al., 2002; Zhang et al., 2011; Li et al., 2012; Gonzalez-Munoz et al., 2015; Bhadouria et al., 2017; Venkidasamy et al., 2019; Yin et al., 2019; Srivastava et al., 2020). In contrast, other Apases such as halogenated acid dehalogenase (HAD), polynucleotide phosphorylase (PNPase) and protein phosphatases (PP2C) are less investigated in plants in response to Pi variations (Marchive et al., 2010; Khan et al., 2018; Su et al., 2021; Bhadouria and Giri, 2022). The molecular regulation of Apase expression is found to be complex in nature. A number of transcription factors like *PHR1 (PHOSPHATE STARVATION RESPONSE 1)* and its homologues (*PHR-like- PHL1, PHL2*, and *PHL3*), *WRKY75, ZAT6 (ZINC FINGER OF ARABIDOPSIS THALIANA 6), OsMYB2P-1 (Oryza sativa MYB2 phosphate-responsive gene 1), StMYB44 (Solanum tuberosum MYB transcription factor)*, and *AP2/ERF (APETALA 2/ethylene-responsive element binding factor)* regulate Apase activity (Dai et al., 2012;

Wang et al., 2013; Zhou et al., 2017). SPX proteins (*SYG1, PHO81* and *Xpr1*) indirectly control PAP activity by binding to PHR/PHL transcription factors. In monocots, SPX proteins show low affinity to PHR/PHL transcription factors, thus inducing PAPs under Pi deficient conditions (Wang et al., 2014; Puga et al., 2014). The trend is opposite in dicots, where PAPs are positively regulated by SPX proteins. In addition, phytohormones (auxins, cytokinins and ethylene) and sugar signalling, miRNA399 expression, and some post-translation modifications such as glycosylation strongly influence Apase activity (Puga et al., 2017). GWAS (genome wide association study) is a classical method for studying the genetic basis of complex quantitative traits (Pal et al., 2021). It takes full advantage of historical recombination events coupled with high allelic diversity of the association panels for fine mapping of genetic loci (Rafalski, 2010; Huang and Han, 2014). Indian mustard (*Brassica juncea* L. Czern & Coss, 2n = 4x = 36, genome AABB) is an important crop species that provides oil for human consumption and protein rich extraction meal for the animal industry (Goel et al., 2018). Its yield and quality are severely affected by low Pi availability in the soil (Zhang et al., 2009; Yao et al., 2011). So, breeding mustard varieties with enhanced PUE is imperative for sustainable agriculture. Unveiling the molecular mechanisms of different players of plant adaptation to low Pi will help to design Pi efficient cultivars. In the present study, we analyzed a wide germplasm set (280 genotypes) of Indian mustard for Apase activity in root and leaf tissues at three Pi levels in a hydroponic system. GWAS enabled us to study the association of SNP markers with genetic variation for Apase activity. The identified marker-trait associations (MTAs) and candidate genes in the present investigation will support the development of P efficient cultivars *via* marker assisted breeding.

2 Materials and methods

2.1 Plant materials

A diversity fixed foundation set of 280 genotypes of *B. juncea*, including landraces, historical varieties, cultivars, resynthesized and determinate *B. juncea*, alloplasmic lines and introgression lines was evaluated for Apase activity in the root (RApase) and leaf (LApase) tissues at three Pi levels: low (LP; 5μM), normal (NP; 250μM) and high (HP; 1mM) in a hydroponic system (Supplementary Table 1). The diversity fixed foundation set collection was established at Punjab Agricultural University, Ludhiana, under the ICAR (Indian Council of Agricultural Research) funded NASF (National Agricultural Science Fund) project: "Creating a fully characterized genetic resources pipeline for mustard improvement".

2.2 Plant growth conditions and Apase activity measurement

The hydroponic experiment was conducted twice at an experimental farm of the Department of Soil and Water Engineering, Punjab Agricultural University, Ludhiana, from July 2020 to September 2020. An in-house developed hydroponic system was deployed for the current study. For this, seven PVC pipes (length: 609.6 cm; diameter: 10 cm) were installed on an angle iron frame in a pyramidal arrangement. The bottom pipes were maintained at the height of 75 cm above the ground. 20 holes/pipe were drilled at a spacing of 30 cm to retain pots of diameter- 7.5 cm for plant growth. The growth conditions of the polyhouse were maintained at 25°/18°C day/night temperature with relative humidity of 70 ± 2%. Two seeds of uniform size from each of 280 genotypes were sown in portray, after surface sterilization in a 0.5% (w/v) sodium hypochlorite solution for 15 minutes, followed by three washings with deionized water. Virgin plasticware and glasswares were used in the whole experiment to avoid Pi contamination, if any. After seed germination, seedlings were watered with a quarter strength of modified Hoagland's solution twice a day. The modified full-strength Hoagland's solution was comprised of 4.5 mM $Ca(NO_3)_2 \cdot 4H_2O$, 1 mM KH_2PO_4, 4 mM KNO_3, and 2 mM $MgSO_4 \cdot 7H_2O$ as macronutrients, and 0.32 μM $CuSO_4 \cdot 5H_2O$, 46 μM H_3BO_3, 50 μM EDTA-Fe, 0.37 μM $NaMoO_4 \cdot 2H_2O$, 9.14 μM $MnCl_2 \cdot 4H_2O$ and 0.77 μM $ZnSO_4 \cdot 7H_2O$ as micronutrients (Hoagland and Arnon, 1950). Three levels of Pi were adjusted using KH_2PO_4 as low (5μM), normal (250μM) and high (1mM). Under LP and NP conditions, 5 μM and 250 μM KH_2PO_4 were supplied along with 0.50 mM and 0.26 mM KCl respectively. Upon the emergence of true leaves, seedlings were shifted to the hydroponic system with one seedling/pot. After five days of shifting, the quarter-strength nutrient solution was progressively increased to a full-strength solution once a week until tissue sampling. The pH and EC of the nutrient solution were adjusted to 5.7 ± 0.2 and 1.5–2.5 ds/m, respectively (Greenway and Munns, 1980). Twenty-eight days old seedlings were taken for root and leaf sampling. Fresh weights of root and leaf tissues were estimated before measuring Apase activity. The standard protocol was followed to measure Apase activity in root and leaf tissue (Tabatabai and Bremner, 1969). 100 mg of root or leaf tissue was ground into a fine powder in liquid nitrogen and then homogenized in 3 ml of 0.1 M sodium acetate buffer, pH 5.0. The homogenate was centrifuged at 10,000 rpm for 15 min at 4°C. Further, 0.1 mL of the supernatant containing enzyme extract/intracellular proteins was mixed with 1.9 mL of sodium acetate buffer and 1 mL of p-nitrophenyl phosphate (p-NPP) as the substrate. Reactions proceeded for 15 min at 37°C and was terminated using one ml of 2N NaOH. p-nitrophenol (pNP) accumulation was read at 410 nm wavelength in the spectrophotometer

(Techcomp UV 2600) (McLachlan et al., 1987). Apase activity was recorded as µmol of p-NP liberated min^{-1} g^{-1} fresh tissue weight (FW) from p-NPP.

2.3 Statistical analysis

Pooled analysis of variance was performed using Minitab v 19.0 software to assess the significance of variance due to genotype, Pi-levels, and environment, and all possible interactions between these parameters (genotype x environment, genotype x Pi-level, and Pi-level x environment) for estimated traits. Descriptive statistics and Pearson's correlation coefficients (r) between the estimated traits were calculated using Minitab v 19.0 software.

2.4 Genotyping by sequencing and genome wide association study

Genotyping by sequencing data of the NASF diversity set was generated under National Agricultural Science Fund aided project "Creating a fully characterized genetic resources pipeline for mustard improvement" (Bio-Project INRP000037). The clean reads of genotypes were aligned to the reference genome of an oleiferous type of *B. juncea* variety Varuna (NCBI bioproject: PRJNA550308) using BWA software (Li and Durbin, 2009). SNP calling was then performed by employing the NGSEP-GBS (Next Generation Sequencing Experience Platform-GBS) pipeline (Duitama et al., 2014). From this marker dataset, SNPs showing minor allele frequency < 0.05, missing data >30% and heterozygosity >10% were removed. The resultant 3, 72,285 SNPs were used for GWAS analysis. BLUPs (Best linear unbiased predictions) datasets across two environments for the estimated traits (RApase and LApase) along with 3, 72,285 filtered SNPs were used as input data for GWAS analysis. BLUPs were estimated using META-R (Multi Environment Trial Analysis with Version 6.0) (https://data.cimmyt.org/dataset.xhtml?persistentId=hdl: 11529/10201) (Alvarado et al., 2020). Principal component analysis (PCA) of genotypic data was performed in R. We used different algorithms GLM (general linear model), MLM (mixed linear model), MLMM (multiple loci mixed linear model), Farm CPU (Fixed and random model Circulating Probability Unifcation) and BLINK (Bayesian-information and Linkage disequilibrium Iteratively Nested Keyway) installed in the GAPIT3 (Genome Association Predict Integrate Tools v3.0) package of R software, incorporating principal components (PCs) and kinship matrices as covariates, to execute association analysis (Wang and Zhang, 2021). Best fit algorithm was predicted using multiple Quantile-quantile (Q-Q) plots for the estimated traits. An arbitrary threshold of -log10 (P) = 3.00 was used as the suggestive threshold to term an association between SNP and trait as significant (To et al., 2019; Mai et al., 2021). The GAPIT3 package was also used to construct Manhattan plots. The genomic regions around the identified peak SNPs (50-kb upstream and 50-kb downstream of the peak SNP) were annotated to scrutinize potential candidate genes pertinent to Apase activity using Blast2GO v5.2.5 tool (Gotz et al., 2008).

3 Results

3.1 Phenotypic variation for Apase activity

Significant effects of genotype and Pi level were observed for the estimated traits (RApase, LApase and R/S) among the tested genotypes. Genotype × Pi level interactions were also significant, whereas genotype × environment was found to be non-significant (Table 1). Traits showed near normal distribution at all three Pi levels, with variations across levels (Figure 1). Descriptive statistics of the examined traits under three Pi levels are given in Table 2 and Figure 2. Low Pi induced higher Apase activity by 67% and 29% in root and 11% and 58% in leaf tissues as compared to NP and HP, respectively. However, LApase exhibited a comparatively higher mean value than RApase by 51%, 82% and 18% at LP, NP and HP levels respectively. The

TABLE 1 Analysis of variance for RApase, LApase and R/S at three Pi levels.

Source of variation	DF	Adjusted mean square		
		RApase	LApase	R/S ratio
Genotype	279	0.255***	1.526***	0.009***
Pi level	2	71.779***	252.245***	3.708***
Environment/experiment	1	0.029	0.004	0.0024
Genotype×Pi level	558	0.232***	1.056***	0.007***
Genotype×Environment/experiment	279	0.012	0.017	0.00228
Pi level×Environment/experiment	2	0.025	0.004	0.00136
Error	558	0.012	0.018	0.00277

*Significance at p< 0.05, **Significance at p<0.01 and ***Significance at p<0.001

TABLE 2 Descriptive statistics of RApase, LApase and R/S ratio at three Pi levels.

Pi level	Variables	Mean ± SE	Range	StDev	CV (%)
	RApase	1.06 ± 0.03	0.15-2.2	0.5	47.11
LP	LApase	2.18 ± 0.05	0.05-4.31	0.88	40.10
	R/S	0.25 ± 0.004	0.08-0.55	0.08	29.43
NP	RApase	0.35 ± 0.012	0.05-0.83	0.19	55.34
	LApse	1.94 ± 0.06	0.03-4.24	0.93	47.78
	R/S	0.12 ± 0.002	0.04-0.22	0.04	32.14
	RApase	0.75 ± 0.02	0.05-1.14	0.27	35.65
HP	LApase	0.92 ± 0.03	0.02-1.93	0.44	48.23
	R/S	0.11 ± 0.002	0.05-0.23	0.04	30.83

mean LApase ranged from 0.05 to 4.31, 0.03 to 4.24 and 0.02 to 1.93 µmole p-NP $min^{-1}g^{-1}FW$ in LP, NP and HP applications, respectively. Genotypes Pusa-Mahak, JM-06016, IM-170 and Kranti were found with high LApase activity (>3.6 µmole p-NP $min^{-1}g^{-1}FW$) at low Pi dose. Mean RApase ranged from 0.15 to 2.19, 0.05 to 0.83 and 0.05 to 1.14 µmole p-NP $min^{-1}g^{-1}FW$ in LP, NP and HP doses, respectively. Genotypes RB-50, RH-9308-1, IM-127 and JT-1 possessed high RApase values (> 2.1 µmole p-NP $min^{-1}g^{-1}FW$) at low Pi supply. The R/S increased significantly by 52% and 56% at LP in comparison to NP and HP, respectively. IM-170, Kranti, JM-06016 and JT-1 were the promising genotypes, depicting higher Apase activity and increased R/S ratio under LP condition. IM-170, Kranti and JM-06016 showed greater LApase, while JT-1 was identified with higher RApase activity. The coefficient of variation (CV) was highest for RApase at NP (55.34) followed by LApase at HP (48.23%) and NP (47.78%). Pairwise Pearson's correlation coefficients revealed a negative correlation between LApase and RApase at HP ($P<0.05$), while at LP and NP level they exhibited no correlation to each other (Table 3). The R/S exhibited a positive correlation with LApase under LP condition.

3.2 Principal component analysis and GWAS for Apase activity

To reduce false positives due to population structure, we used principal components (2 PCs) and kinship matrix as covariates in GWAS analysis. Two PCs depicted clear separation among populations (Figure 3). In total, four groups appeared. Groups I and II were less diverse than groups III and IV. Group I (shown in turquoise blue color) comprised only 5 genotypes. Groups II (magenta) and III (royal blue) included seven and twenty-five genotypes, respectively. Group IV (red color) was the largest group, with 243 genotypes. Groups I and II primarily comprised of exotic mustard genotypes. Most of the introgression and advanced breeding lines were in group III. Group IV was the mix of varieties, resynthesized genotypes and exotic *B. juncea*. GWAS was conducted to identify MTAs for RApase and LApase under LP, NP and HP levels (Table 4). A total of 3, 72,285 quality SNPs were used as marker dataset. The GAPIT3 package (Wang and Zhang, 2021) implemented with five algorithms (GLM, MLM, Farm CPU, MLMM and Blink) in R software was run for trait-SNP association analysis. An ideal model is supposed to show a fair degree of correspondence between the observed and expected p-values in the quantile–quantile (QQ) plots. We compared p values [observed – log10

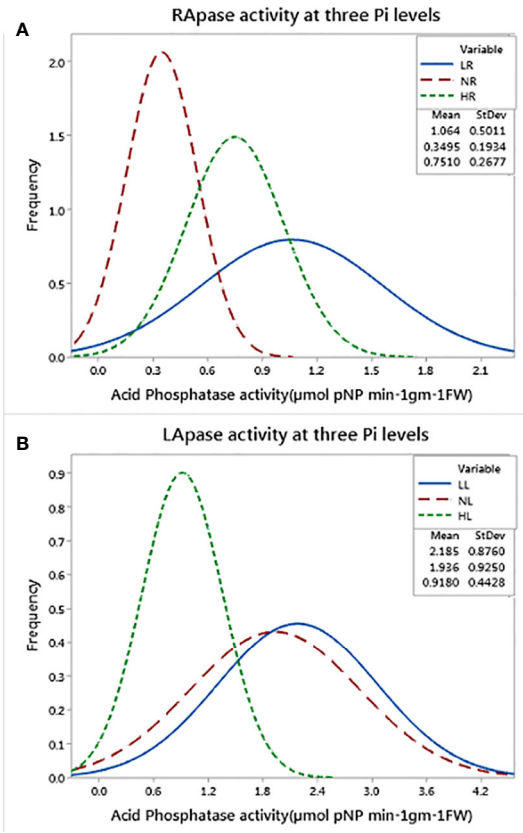


FIGURE 1
Frequency distribution of **(A)** RApase and **(B)** LApase at three Pi levels.

10.3389/fpls.2022.1056028

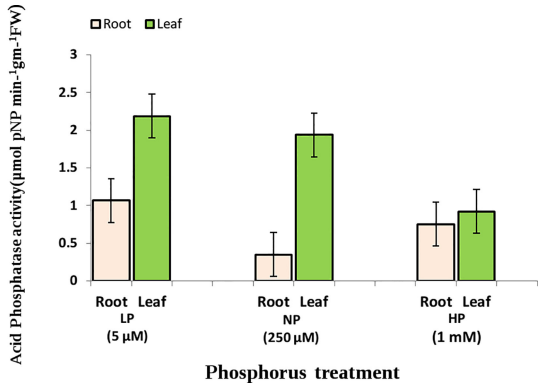


FIGURE 2
Acid phosphatase activity (in root and leaf) at three Pi levels.

(p-value)] and their expected ranked values [expected – log10 (p-value)] through Q-Q plots to test the predictability of applied GWAS models, over all environments. MLMM and BLINK were identified as the best fit models over all environments (Supplementary Figure 1). They showed minimum deviations from uniform distribution in multiple Q-Q plots. The estimated Bonferroni threshold value was 6.87. However, the value was found to be highly stringent to detect MTAs for a complex trait like Apase activity which might be controlled by several genes of minor effects. So, we used an arbitrary threshold value of -log 10 (p) ≥ 3.0 to identify MTAs for Apase activity. Manhattan plots depicting the associated SNPs for Apase activity in root and leaf tissues under three environments (LP, NP and HP) are presented in Figures 4 and 5. Associated SNPs and surrounding genomic regions were further annotated to decipher trait related genes. A total of 14 genomic regions involving 44 unique MTAs were envisioned for Apase activity (including both RApase and LApase) under LP, NP and HP levels on chromosomes A01, A03, A04, A05, A08, A09, B05, B06, B07 and B08 of *B. juncea* (Table 4). Ten associated regions were predicted for A genome chromosomes, while four for B genome chromosomes. Chromosome A01 revealed the maximum number of MTAs (21 SNPs). The identified QTLs accounted for 4.15 to 7.21% of phenotypic variation. Among 14, 10 regions were recorded for RApase and 4 for LApase. The number of QTLs detected explicitly under LP, NP and HP were 6, 3 and 6, respectively. We could not find any common MTA for Apase activity across three Pi doses. Functional annotation of 100 kbp region (50 kbp on both sides) of peak-associated SNPs, facilitated the identification of 15 candidate genes with diverse roles in the molecular regulation of Apase activity. Four genes encode diverse family members of Apases. Other genes have roles in phytohormone or sugar signalling pathways and root modulation. Genes *PAP10* and *PAP16*, which belong to PAP family (central family of Apase) of Apase, were predicted in association with RApase and LApase activities, respectively, under the NP environment. Also, genes, *AT5G51260* (HAD superfamily gene) and *PNP* (*polyribonucleotide nucleotidyltransferase* family gene), encoding Apases, were envisaged 21.48 and 27.84 kbp away from the peak SNPs B06_33902675 and A05_33349252, respectively, influencing RApase under LP condition. In the present study, gene *ARF5* (*AUXIN RESPONSE FACTOR5)* was visualized 4.83 kbp away from the SNP B07_55242497 governing LApase activity at the HP level. Gene *ABR1 (APETALA2 like ABA repressor 1*), associated with RApase activity at LP, was envisaged 36.31kbp away from the peak SNP A09_5080851. We annotated genes *PLC2, PLAIVA* and P*LAIVC*, encoding phospholipases that hydrolyze phospholipids and release secondary messengers for phytohormone signalling. At LP dose, the gene *PLC2* was predicted for the genomic region B05_52982160-52982181 associated with RApase. Seven SNPs (A01_874995-876161) were present near *PLAIVA* and P*LAIVC* depicting 7% variation for RApase under HP condition. Another important gene *HXK1* (*HEXOKINASE1*) involved in the sugar signalling pathway, was located 15.48 Kbp away from peak SNP A09_6643946. This gene explained 7% of the phenotypic variation for LApase activity under Pi deficit condition. Four genes (*RGF1, BRXL1, SHR*, and *SAUR 41*) controlling root architecture were also recorded in the close surroundings of associated regions. Under Pi sufficient conditions, a SNP (A03_20686760) on A03 was linked with a signalling peptide *RGF1* (*ROOT GROWTH FACTOR1*) controlling variation for RApase. Gene *BRXL1* (*BREVIS RADIX LIKE 1*) was predicted near the cluster of six SNPs (A04_18578091-18578642), influencing LApase activity in Pi deficit condition. Gene *CIPK6* (*CBL-interacting protein kinase 6*) was present in the vicinity of two SNPs (A08_19805050, A08_19805060) located on the A08 chromosome. *CIPK6* codes for a CBL (Calcineurin B-like proteins)-interacting protein kinase with role in Pi deficiency and ABA signalling pathways.

TABLE 3 Correlation analysis of BLUP estimated traits at three Pi levels.

Trait 1	Trait 2	LP (5 µM)	NP (250 µM)	HP (1 mM)
LApase	RApase	-0.057	0.05	-0.14*
R/S	RApase	-0.016	-0.074	-0.014
R/S	LApase	0.14*	-0.026	0.074

* Significance at p< 0.05.

4 Discussion

Breeding for enhanced PUE in Indian mustard is imperative for yield increment at low production costs. The Apase class of enzymes is crucial to improve PUE, as their activity is largely associated to P remobilization within the plant and acquisition from the soil by hydrolyzing P rich organic compounds (Duff et al., 1994; Bhadouria and Giri, 2022). There are two groups of APases depending upon their site of action-extracellular (secreted; SAPs) and intracellular (IAPs) Apases (Tian et al., 2012; Tian and Liao, 2015). SAPs act upon external organic P complexes in rhizosphere to liberate Pi, whereas IAPs are involved in Pi recycling from internal P reservoirs of plant cells (Sanchez-Calderon et al., 2010; Chiou and Lin, 2011; Swetha and Padmavathi, 2016). Some Apases possess both SAP and IAP properties (Robinson et al., 2012; Deng et al., 2020). Additionally, Apase enzymes are revealed to regulate diverse plant processes like seed development, flowering, senescence, carbon metabolism, response to biotic and abiotic stresses, cellular signalling pathways, symbiotic association, and root development. In the present study, we studied the genetics of Apase activity using GWAS methodology on a diversity set in *B.juncea*. Enzyme activity was estimated in two plant tissues (leaf and root) at three doses of Pi application in a hydroponic system. Significant differences were observed for both LApase (Apase activity in leaf) and RApase (Apase activity in root) over three Pi levels. The mean Apase activity increased in both tissues with a decrease in Pi input that emphasized the enzyme involvement under Pi deprived condition. This was in correspondence with previous reports in Indian mustard, common bean crops and rapeseed (Haran et al., 2000; Yan et al., 2001; Zhang et al., 2010). However, at all Pi levels, LApase activity was higher than RApase. This indicted the greater or earlier response of LApase than RApase to Pi supply. Zhang et al. (2010) has studied the contributions of root secreted Apase and leaf intracellular APase to PUE in *B. napus* and reported a significant contribution of leaf Apase activity towards PUE whereas root secreted Apase has revealed no direct correlation with PUE. Intracellular Apases are believed to be synthesized prior to the secreted Apases under Pi deprived condition (Bozzo et al., 2004; Bozzo et al., 2006). Apase activity in leaf enables the plant to remobilize Pi from P rich biomolecules present in older tissues (Duff et al., 1994; Garcia et al., 2004; Zhang et al., 2010). In our study, LApase was observed with a significant positive correlation with R/S at LP level. This may be due to the additional role of Apase in modulating root architecture by induction of Pi signalling pathways (Wang et al., 2018; Cai et al., 2021). At only Pi sufficiency, RApase was negatively correlated to LApase. It indicated the differential response of Apase activity in root and leaf to Pi status.

We identified several candidate genes (15) on chromosomes (A01, A03, A04, A05, A08, A09, B05, B06, B07 and B08) that might affect Apase activity in the evaluated genotypes of *B. juncea*. Important among them were: *PAP10, PAP16, PNP* and *AT5G51260*, which encode different family members of Apases. In our study, *PNP* and *AT5G51260* were observed for RApase activity under LP environment, whereas, *PAP10* and *PAP16* governed the variation at NP dose for RApase and LApase, respectively. Overexpression of PAP10 (that shows both SAP and IAP properties) in Arabidopsis and rice have been found to significantly increase the plant's ability to degrade Po and tolerance to low Pi stress (Wang et al., 2011; Deng et al., 2020). *PNP* and *AT5G51260* regulate Pi tolerance by releasing Pi during polynucleotide synthesis from nucleotide diphosphates or triphosphates (Marchive et al., 2010; Deng et al., 2021). We also predicted several genes with roles in phytohormone and sugar signalling pathways and root modulation. *ARF5*, a gene located 4.8 kbps away from the SNP

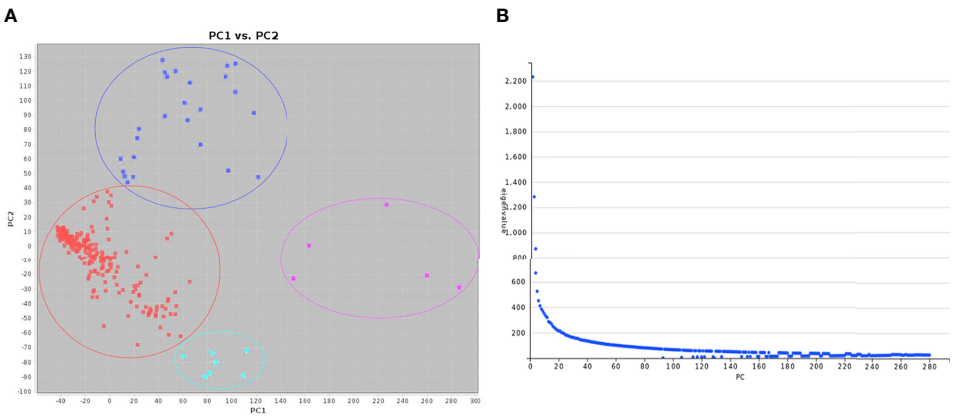


FIGURE 3
Principal component analysis of the diversity fixed foundation set of *Brassica juncea*: **(A)** Principal components and **(B)** Scree plot.

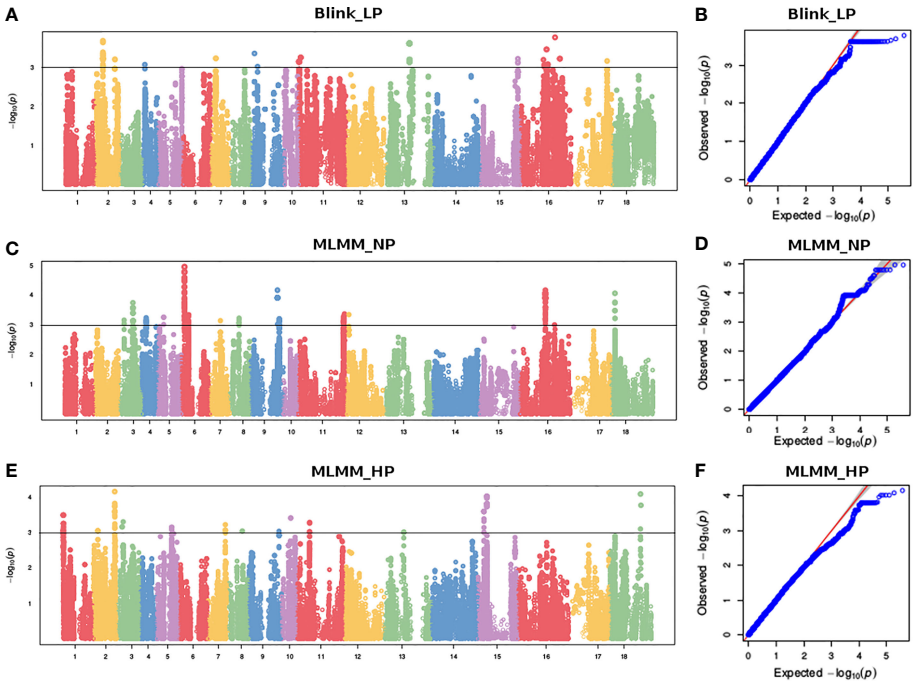


FIGURE 4
Manhatton and Q-Q plots showing marker trait associations for RApase at **(A, B)** LP, **(C, D)** NP and **(E, F)** HP levels.

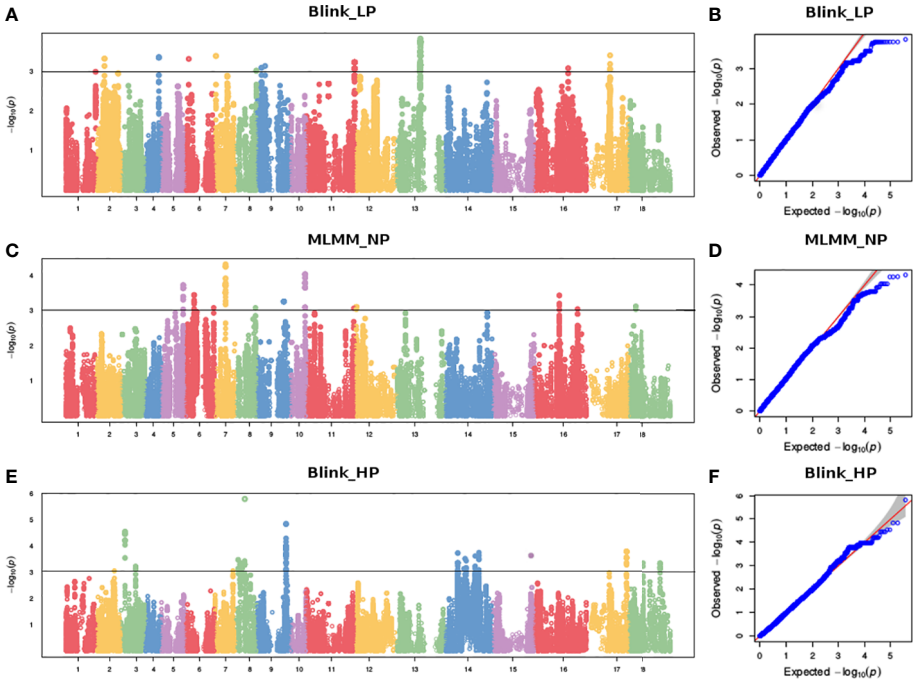


FIGURE 5
Manhatton and Q-Q plots showing marker trait associations for LApase at **(A, B)** LP, **(C, D)** NP and **(E, F)** HP levels.

TABLE 4 Summary of marker trait associations identified for RApase and LApase at three Pi levels.

Trait	Pi level	Chr	Position and number of SNPs	Candidate Gene	Distance from peak SNP(Kb)	-log10(p)	R^2
RApase	LP	A05	33349252	*PNP(AT3G03710)*	27.84	3.27	4.47
		A09	5080851	*ABR1(AT5G64750)*	36.31	3.36	4.61
		B05	52982160-52982181(3)	*PLC2(AT3G08510)*	47.15	3.16	4.31
		B06	33902675	*AT5G51260* (*HAD* superfamily)	21.48	3.05	4.15
	NP	A03	20686760	*RGF1(AT5G60810)*	33.93	3.46	6.35
		B08	4817280-4817319(3)	*PAP10(AT2G16430)*	38.60	3.48	6.37
	HP	A01	874995-876161(7)	*PLAIVA(AT4G37070)*	12.25	3.23	7.00
				PLAIVC(AT4G37050)	9.00	3.23	7.00
		A01	651481-655735(14)	*SHR(AT4G37650)*	26.91	3.21	7.00
		A08	19805050, 19805060(2)	*CIPK6(AT4G30960)*	31.63	3.04	7.00
		A09	43265797	*SAUR41(AT1G16510)*	35.82	3.04	7.00
LApase	LP	A04	18578091-18578642(6)	*BRXL1(AT2G35600)*	49.26	3.35	7.21
		A09	6643946-6643948(2)	*HXK1(AT4G29130)*	15.48	3.09	7.00
	NP	A05	30105013	*PAP16(AT3G10150)*	21.21	3.15	5.00
	HP	B07	55242497	*ARF5(AT1G19850)*	4.83	3.18	6.13

B07_55242497, explained variation in Apase activity in leaf under Pi sufficient conditions. It codes for an auxin response factor 5 which transcriptionally regulates almost one-half of Aux/IAA genes (Krogan et al., 2014). Another auxin induced gene *SAUR 41*(*SMALL AUXIN UP RNA41*), appeared close to SNP A09_43265797. It is responsible for developing auxin-related phenotypes including root meristem repatterning in Arabidopsis (Kong et al., 2013). Under low Pi conditions, the gene *ABR1* (*APETALA2 like ABA repressor 1*) encodes a negative regulator of ABA-regulated gene expression (Pandey et al., 2005). A SNP B05_52982181 was linked to a phosphatidylinositol-specific phospholipase C2 encoded by gene *PLC2* for RApase activity at low Pi. This gene shows hydrolytic activity against phosphoinositides to release secondary messengers, myo-inositol-1,4,5-trisphosphate (InsP3) and diacylglycerol (DAG) that are known to improve plant tolerance against various types of biotic and abiotic stresses (Nokhrina et al., 2014). Another two patatin-related phospholipase A genes (*PLAIVA and PLAIVC*) were also predicted for RApase activity but at the high Pi level. They hydrolyze phospholipids and galactolipids and generate free fatty acids and lysolipids as secondary messengers that participate in phytohormone signalling for root development under normal and Pi stress conditions (Rietz et al., 2010). *HXK1* gene encoding *HEXOKINASE 1*, the first enzyme of glycolysis, which converts glucose into glucose 6-phosphate. Besides this phosphorylation activity, it mediates sugar signalling pathway to influence pant architecture (Barbier et al., 2021). Here, this gene explained 7% variation for LApase activity at LP level. For RApase activity at high Pi level, SNP A01_55242497 corresponded to gene *CIPK6*. This gene (*CIPK6*) codes for a CBL (Calcineurin B-like proteins) interacting protein kinase that functions in Pi starvation signalling pathway to enhance the plant tolerance to low Pi stress (Chen et al., 2012). Gene *BRXL1* associated with LApase activity under Pi deficit condition in the present study, determines the extent of cell proliferation and elongation in the plant roots (Mouchel et al., 2006; Beuchat et al., 2010). The SHR gene, which controls 7% of the variation in RApase at HP, plays an important role in the specification and maintenance of the root stem-cell niche (Niu et al., 2015). Gene *RGF1* encoding a signalling peptide is known to influence the circumferential cell number in the root meristem in response to low Pi environment (Cederholm and Benfey, 2015). We could associate this gene with RApase activity at NP dose. The present investigation is the first attempt at genome wide association mapping for Apase enzyme activity in root and leaf tissues of Indian mustard. Predicted genes could serve as potential targets for improving PUE in *Brassica juncea*, after due validation.

5 Conclusion

In this study, 280 mustard genotypes were examined for root and leaf Apase activities in two environments under three doses of Pi. GWAS analysis revealed a total of 44 SNPs significantly associated with two traits at three Pi levels. Functional annotation of genomic regions in or around SNPs facilitated the prediction of genes encoding diverse Apase family members, root modulators, signalling peptides, phytohormone-induced factors, and secondary messenger releasing enzymes. These findings provided useful information to improve the PUE of Indian mustard by marker-assisted selection in the future.

Data availability statement

The data presented in the study are deposited in the Indian Biological Data Center (http://ibdc.rcb.res.in/) repository, accession number INRP000037

Author contributions

GK designed and supervised the whole experiment. PU performed phenotypic evaluations. RS and VS helped in conducting the experiment under hydroponic condition. SS assisted in the biochemical analysis. PU compiled the results and performed the statistical analysis. JA and SKS performed bioinformatics. PU, SKS and MG carried out annotation and wrote the manuscript. All authors have read and approved the published version of the manuscript.

Funding

Genotyping by sequencing of the NASF diversity set was supported by grants received under National Agricultural Science Fund aided project "Creating a fully characterized genetic resources pipeline for mustard improvement." Phenotyping of the diversity set was conducted with financial assistance from the "All India Coordinated Research Project on Oilseeds-rapeseed mustard" funded by the Indian Council of Agricultural Research (ICAR). PU is also thankful to the ICAR for financial assistance in the form of an ICAR Senior Research Fellowship (ICAR JRF/SRF-PGS).

Acknowledgments

Thanks are expressed to Prof. S.S. Banga, then ICAR National Professor, for providing access to the diversity set used in the study along with the SNP genotyping data.

Conflict of interest

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

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.1056028/full#supplementary-material

References

Alvarado, G., Lopez, M., Vargas, M., Pacheco, A., Rodriguez, F., Burgueno, J., et al. (2020). META-R: a software to analyze data from multi-environment plant breeding trials. *Crop J* 8 (5), 745–756.

Barbier, F. F., Cao, D., Fichtner, F., Weiste, C., Perez-Garcia, M. D., Caradeuc, M., et al. (2021). HEXOKINASE1 signalling promotes shoot branching and interacts with cytokinin and strigolactone pathways. *New Phytol.* 231 (3), 1088–1104. doi: 10.1111/nph.17427

Bastani, S., and Hajiboland, R. (2017). Uptake and utilization of applied phosphorus in oilseed rape (Brassica napus l. cv. hayola) plants at vegetative and reproductive stages: Comparison of root with foliar phosphorus application. *Soil Sci. Plant Nutr.* 63 (3), 254–263. doi: 10.1080/00380768.2017.1321471

Beuchat, J., Li, S., Ragni, L., Shindo, C., Kohn, M. H., and Hardtke, C. S. (2010). A hyperactive quantitative trait locus allele of arabidopsis BRX contributes to natural variation in root growth vigor. *Proc. Natl. Acad. Sci. U.S.A.* 107 (18), 8475–8480. doi: 10.1073/pnas.0913207107

Bhadouria, J., and Giri, J. (2022). Purple acid phosphatases: roles in phosphate utilization and new emerging functions. *Plant Cell Rep.* 41, 33–51. doi: 10.1007/s00299-021-02773-7

Bhadouria, J., Singh, A. P., Mehra, P., Verma, L., Srivastawa, R., Parida, S. K., et al. (2017). Identification of purple acid phosphatases in chickpea and potential roles of CaPAP7 in seed phytate accumulation. *Sci. Rep.* 7, 1–12. doi: 10.1038/s41598-017-11490-9

Bonser, A. M., Lynch, J., and Snapp, S. (1996). Effect of phosphorus deficiency on growth angle of basal roots in phaseolus vulgaris. *New Phytol.* 132, 281–288. doi: 10.1111/j.1469-8137.1996.tb01847.x

Bozzo, G. G., Dunn, E. L., and Plaxton, W. C. (2006). Differential synthesis of phosphate-starvation inducible purple acid phosphatase isozymes in tomato (Lycopersicon esculentum) suspension cells and seedlings. *Plant Cell Environ.* 29 (2), 303–313. doi: 10.1111/j.1365-3040.2005.01422.x

Bozzo, G. G., Raghothama, K. G., and Plaxton, W. C. (2004). Structural and kinetic properties of a novel purple acid phosphatase from phosphate-starved tomato (Lycopersicon esculentum) cell cultures. *Biochem. J.* 377 (2), 419–428. doi: 10.1042/bj20030947

Cai, Y., Qi, J., Li, C., Miao, K., Jiang, B., Yang, X., et al. (2021). Genome-wide analysis of purple acid phosphatase genes in brassica rapa and their association with pollen development and phosphorus deprivation stress. *Horticulturae* 7 (10), 363. doi: 10.3390/horticulturae7100363

Camacho, R., Malavolta, E., Guerrero Alves, J., and Camacho, T. (2002). Vegetative growth of grain sorghum in response to phosphorus nutrition. *J. Sci. Food Agric.* 59, 771–776. doi: 10.1590/S0103-90162002000400022

Ceasar, S. A., Hodge, A., Baker, A., and Baldwin, S. A. (2014). Phosphate concentration and arbuscular mycorrhizal colonisation influence the growth, yield and expression of twelve PHT1 family phosphate transporters in foxtail millet (Setaria italica). *PloS One* 9, 1–12. doi: 10.1371/journal.pone.0108459

Cederholm, H. M., and Benfey, P. N. (2015). Distinct sensitivities to phosphate deprivation suggest that RGF peptides play disparate roles in arabidopsis thaliana root development. *New Phytol.* 207 (3), 683–691. doi: 10.1111/nph.13405

Chen, C. R., Condron, L. M., Davis, M. R., and Sherlock, R. R. (2003). Seasonal changes in soil phosphorus and associated microbial properties under adjacent grassland and forest in new Zealand. *For. Ecol. Manage.* 177 (1-3), 539–557. doi: 10.1016/S0378-1127(02)00450-4

Chen, L., Ren, F., Zhou, L., Wang, Q. Q., Zhong, H., and Li, X. B. (2012). The brassica napus calcineurin b-like 1/CBL-interacting protein kinase 6 (CBL1/CIPK6) component is involved in the plant response to abiotic stress and ABA signalling. *J. Exp. Bot.* 63 (17), 6211–6222. doi: 10.1093/jxb/ers273

Chiou, T. J., and Lin, S. I. (2011). Signaling network in sensing phosphate availability in plants. *Annu. Rev. Plant Biol.* 62, 185–206. doi: 10.1146/annurev-arplant-042110-103849

Cong, W. F., Suriyagoda, L. D., and Lambers, H. (2020). Tightening the phosphorus cycle through phosphorus-efficient crop genotypes. *Trends Plant Sci.* 25 (10), 967–975. doi: 10.1016/j.tplants.2020.04.013

Dai, X., Wang, Y., Yang, A., and Zhang, W. H. (2012). OsMYB2P-1, an R2R3 MYB transcription factor, is involved in the regulation of phosphate-starvation responses and root architecture in rice. *Plant Physiol.* 159 (1), 169–183. doi: 10.1104/pp.112.194217

Deng, S., Li, J., Du, Z., Wu, Z., Yang, J., Cai, H., et al. (2021). Rice ACID PHOSPHATASE 1 regulates Pi stress adaptation by maintaining intracellular Pi homeostasis. *Plant Cell Environ* 45(1), 191–205. doi: 10.1111/pce.1419

Deng, S., Lu, L., Li, J., Du, Z., Liu, T., Li, W., et al. (2020). Purple acid phosphatase 10c encodes a major acid phosphatase that regulates plant growth under phosphate-deficient conditions in rice. *J. Exp. Bot.* 71 (14), 4321–4332. doi: 10.1093/jxb/eraa179

Desmidt, E., Ghyselbrecht, K., Zhang, Y., Pinoy, L., van der Bruggen, B., Verstraete, W., et al. (2015). Global phosphorus scarcity and full-scale p-recovery techniques: a review. *Crit. Rev. Environ. Sci. Technol.* 45 (4), 336–384. doi: 10.1080/10643389.2013.866531

Dissanayaka, D. M. S. B., Plaxton, W. C., Lambers, H., Siebers, M., Marambe, B., and Wasaki, J. (2018). Molecular mechanisms underpinning phosphorus-use efficiency in rice. *Plant Cell Environ.* 41 (7), 1483–1496. doi: 10.1111/pce.13191

Duff, S. M., Sarath, G., and Plaxton, W. C. (1994). The role of acid phosphatases in plant phosphorus metabolism. *Physiol. Plant* 90 (4), 791–800. doi: 10.1111/j.1399-3054.1994.tb02539.x

Duitama, J., Quintero, J. C., Cruz, D. F., Quintero, C., Hubmann, G., Foulquie-Moreno, M. R., et al. (2014). An integrated framework for discovery and genotyping of genomic variants from high-throughput sequencing experiments. *Nucleic Acids Res.* 42 (6), e44–e44. doi: 10.1093/nar/gkt1381

Garcia, N. A. T., Olivera, M., Iribarne, C., and Lluch, C. (2004). Partial purification and characterization of a non-specific acid phosphatase in leaves and root nodules of phaseolus vulgaris. *Plant Physiol. Biochem.* 42 (7-8), 585–591. doi: 10.1016/j.plaphy.2004.04.004

Gilbert, J., Gowing, D., and Wallace, H. (2009). Available soil phosphorus in semi-natural grasslands: assessment methods and community tolerances. *Biol. Conserv.* 142 (5), 1074–1083. doi: 10.1016/j.biocon.2009.01.018

Goel, P., Sharma, N. K., Bhuria, M., Sharma, V., Chauhan, R., Pathania, S., et al. (2018). Transcriptome and co-expression network analyses identify key genes regulating nitrogen use efficiency in brassica juncea l. *Sci. Rep.* 8 (1), 1–18. doi: 10.1038/s41598-018-25826-6

Gonzalez-Munoz, E., Avendano-Vazquez, A. O., Montes, R. A. C., de Folter, S., Andres-Hernandez, L., Abreu-Goodger, C., et al. (2015). The maize (Zea mays ssp. mays var. B73) genome encodes 33 members of the purple acid phosphatase family. *Front. Plant Sci.* 6. doi: 10.3389/fpls.2015.00341

Gotz, S., García-Gómez, J. M., Terol, J., Williams, T. D., Nagaraj, S. H., Nueda, M. J., et al. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.* 36 (10), 3420–3435. doi: 10.1093/nar/gkn176

Greenway, H., and Munns, R. (1980). Mechanisms of salt tolerance in nonhalophytes. *Annu. Rev. Plant Physiol.* 31 (1), 149–190. doi: 10.1146/annurev.pp.31.060180.001053

Grzebisz, W., Szczepaniak, W., Barlog, P., Przygocka-Cyna, K., and Potarzycki, J. (2018). Phosphorus sources for winter oilseed rape (Brassica napus l.) during reproductive growth–magnesium sulfate management impact on p use efficiency. *Arch. Agron. Soil Sci.* 64 (12), 1646–1662. doi: 10.1080/03650340.2018.1448389

Gu, R., Chen, F., Long, L., Cai, H., Liu, Z., Yang, J., et al. (2016). Enhancing phosphorus uptake efficiency through QTL-based selection for root system architecture in maize. *J. Genet. Genomics* 43 (11), 663–672. doi: 10.1016/j.jgg.2016.11.002

Han, Y., White, P. J., and Cheng, L. (2022). Mechanisms for improving phosphorus utilization efficiency in plants. *Ann. Bot.* 129 (3), 247–258. doi: 10.1093/aob/mcab145

Haran, S., Logendra, S., Seskar, M., Bratanova, M., and Raskin, I. (2000). Characterization of arabidopsis acid phosphatase promoter and regulation of acid phosphatase expression. *Plant Physiol.* 124 (2), 615–626. doi: 10.1104/pp.124.2.615

Hoagland, D. R., and Arnon, D. I. (1950). *The water-culture method for growing plants without soil.* (California, USA: Circular, California Agricultural Experiment Station), 347(2nd edit))

Huang, X., and Han, B. (2014). Natural variations and genome-wide association studies in crop plants. *Annu. Rev. Plant Biol.* 65, 531–551. doi: 10.1146/annurev-arplant-050213-035715

Khan, N., Bano, A., and Zandi, P. (2018). Effects of exogenously applied plant growth regulators in combination with PGPR on the physiology and root growth of chickpea (Cicer arietinum) and their role in drought tolerance. *J. Plant Interact.* 13 (1), 239–247.924. doi: 10.1080/17429145.2018.1471527

Kohli, P. S., Kumar Verma, P., Verma, R., Parida, S. K., Thakur, J. K., and Giri, J. (2020). Genome-wide association study for phosphate deficiency responsive root hair elongation in chickpea. *Funct. Integr. Genomics* 20 (6), 775–786. doi: 10.1007/s10142-020-00749-6

Kong, Y., Zhu, Y., Gao, C., She, W., Lin, W., Chen, Y., et al. (2013). Tissue-specific expression of SMALL AUXIN UP RNA41 differentially regulates cell expansion and root meristem patterning in arabidopsis. *Plant Cell Physiol.* 54 (4), 609–621. doi: 10.1093/pcp/pct028

Krogan, N. T., Yin, X., Ckurshumova, W., and Berleth, T. (2014). Distinct subclades of Aux/IAA genes are direct targets of ARF 5/MP transcriptional regulation. *New Phytol.* 204 (3), 474–483. doi: 10.1111/nph.12994

Lazaro, L., Abbate, P., Cogliatti, D., and Andrade, F. (2010). Relationship between yield, growth and spike weight in wheat under phosphorus deficiency and shading. *J. Agric. Sci.* 148, 83–93. doi: 10.1017/S0021859609990402

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with burrows–wheeler transform. *Bioinform* 25 (14), 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, C., Gui, S., Yang, T., Walk, T., Wang, X., and Liao, H. (2012). Identification of soybean purple acid phosphatase genes and their expression responses to phosphorus availability and symbiosis. *Ann. Bot.* 109 (1), 275–285. doi: 10.1093/aob/mcr246

Li, Y., Niu, S., and Yu, G. (2016). Aggravated phosphorus limitation on biomass production under increasing nitrogen loading: a meta-analysis. *Glob. Change Biol.* 22 (2), 934–943. doi: 10.1111/gcb.13125

Li, D., Zhu, H., Liu, K., Liu, X., Leggewie, G., Udvardi, M., et al. (2002). Purple acid phosphatases of arabidopsis thaliana: comparative analysis and differential regulation by phosphate deprivation. *J. Biol. Chem.* 277 (31), 27772–27781. doi: 10.1074/jbc.M204183200

Lu, L., Qiu, W., Gao, W., Tyerman, S. D., Shou, H., and Wang, C. (2016). OsPAP10c, a novel secreted acid phosphatase in rice, plays an important role in the utilization of external organic phosphorus. *Plant Cell Environ.* 39 (10), 2247–2259. doi: 10.1111/pce.12794

Mahamood, J., Abayomi, Y., and Aduloju, M. (2009). Comparative growth and grain yield responses of soybean genotypes to phosphorous fertilizer application. *Afr. J. Biotechnol.* 8, 1030–1036.

Maharajan, T., Ceasar, S. A., Ajeesh krishna, T. P., Ramakrishnan, M., Duraipandiyan, V., Naif Abdulla, A. D., et al. (2018). Utilization of molecular markers for improving the phosphorus efficiency in crop plants. *Plant Breed.* 137 (1), 10–26. doi: 10.1111/pbr.12537

Mai, N. T., Mai, C. D., Van Nguyen, H., Le, K. Q., Duong, L. V., Tran, T. A., et al. (2021). Discovery of new genetic determinants of morphological plasticity in rice roots and shoots under phosphate starvation using GWAS. *J. Plant Physiol.* 257, 153340. doi: 10.1016/j.jplph.2020.153340

Marchive, C., Yehudai-Resheff, S., Germain, A., Fei, Z., Jiang, X., Judkins, J., et al. (2010). Abnormal physiological and molecular mutant phenotypes link chloroplast polynucleotide phosphorylase to the phosphorus deprivation response in arabidopsis. *Plant Physiol.* 151 (2), 905–924. doi: 10.1104/pp.109.145144

McLachlan, K. D., Elliot, D. E., Marco, D., Garran, J. H., and De Marco, D. G. (1987). Leaf acid phosphatase isozymes in the diagnosis of phosphorus status in field-grown wheat. *Aust. J. Agric. Res.* 38 (1), 1–13. doi: 10.1071/AR9870001

Mehra, P., Pandey, B. K., and Giri, J. (2017). Improvement in phosphate acquisition and utilization by a secretory purple acid phosphatase (OsPAP21b) in rice. *Plant Biotechnol. J.* 15 (8), 1054–1067. doi: 10.1111/pbi.12699

Mogollon, J. M., Beusen, A. H. W., Van Grinsven, H. J. M., Westhoek, H., and Bouwman, A. F. (2018). Future agricultural phosphorus demand according to the shared socioeconomic pathways. *Glob. Environ. Change* 50, 149–163. doi: 10.1016/j.gloenvcha.2018.03.007

Mouchel, C. F., Osmont, K. S., and Hardtke, C. S. (2006). BRX mediates feedback between brassinosteroid levels and auxin signalling in root growth. *Nature* 443 (7110), 458–461. doi: 10.1038/nature05130

Nannipieri, P., Giagnoni, L., Landi, L., and Renella, G. (2011). "Role of phosphatase enzymes in soil," in *Phosphorus in action* (Berlin, Heidelberg: Springer), 215–243. doi: 10.1007/978-3-642-15271-9_9

Niu, Y., Jin, G., Li, X., Tang, C., Zhang, Y., Liang, Y., et al. (2015). Phosphorus and magnesium interactively modulate the elongation and directional growth of primary roots in arabidopsis thaliana (L.) heynh. *J. Exp. Bot.* 66 (13), 3841–3854. doi: 10.1093/jxb/erv181

Nokhrina, K., Ray, H., Bock, C., and Georges, F. (2014). Metabolomic shifts in brassica napus lines with enhanced BnPLC2 expression impact their response to low temperature stress and plant pathogens. *GM Crops Food.* 5 (2), 120–131. doi: 10.4161/gmcr.28942

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. Plants.* 27 (9), 1933–1951. doi: 10.1007/s12298-021-01060-9

Pandey, G. K., Grant, J. J., Cheong, Y. H., Kim, B. G., Li, L., and Luan, S. (2005). ABR1, an APETALA2-domain transcription factor that functions as a repressor of ABA response in arabidopsis. *Plant Physiol.* 139 (3), 1185–1193. doi: 10.1104/pp.105.066324

Plaxton, W. C., and Tran, H. T. (2011). Metabolic adaptations of phosphate-starved plants. *Plant Physiol.* 156 (3), 1006–1015. doi: 10.1104/pp.111.175281

Plenet, D., Etchebest, S., Mollier, A., and Pellerin, S. (2000). Growth analysis of maize field crops under phosphorus deficiency. *Plant Soil.* 223, 119–132. doi: 10.1023/A:1004877111238

Puga, M. I., Mateos, I., Charukesi, R., Wang, Z., Franco-Zorrilla, J. M., de Lorenzo, L., et al. (2014). SPX1 is a phosphate-dependent inhibitor of phosphate starvation response 1 in arabidopsis. *Proc. Natl. Acad. Sci. U.S.A.* 111 (41), 14947–14952. doi: 10.1073/pnas.1404654111

Puga, M. I., Rojas-Triana, M., de Lorenzo, L., Leyva, A., Rubio, V., and Paz-Ares, J. (2017). Novel signals in the regulation of pi starvation responses in plants: facts and promises. *Curr. Opin. Plant Biol.* 39, 40–49. doi: 10.1016/j.pbi.2017.05.007

Radersma, S., and Grierson, P. F. (2004). Phosphorus mobilization in agroforestry: organic anions, phosphatase activity and phosphorus fractions in the rhizosphere. *Plant Soil.* 259 (1), 209–219. doi: 10.1023/B:PLSO.0000020970.40167.40

Rafalski, J. A. (2010). Association genetics in crop improvement. *Curr. Opin. Plant Biol.* 13 (2), 174–180. doi: 10.1016/j.pbi.2009.12.004

Raghothama, K. G. (1999). Phosphate acquisition. *Annu. Rev. Plant Biol.* 50, 665. doi: 10.1146/annurev.arplant.50.1.665

Raghothama, K. G., and Karthikeyan, A. S. (2005). Phosphate acquisition. *Plant Soil.* 274 (1), 37–49. doi: 10.1007/s11104-004-2005-6

Richardson, A. E., and Simpson, R. J. (2011). Soil microorganisms mediating phosphorus availability update on microbial phosphorus. *Plant Physiol.* 156 (3), 989–996. doi: 10.1104/pp.111.175448

Rietz, S., Dermendjiev, G., Oppermann, E., Tafesse, F. G., Effendi, Y., Holk, A., et al. (2010). Roles of arabidopsis patatin-related phospholipases a in root development are related to auxin responses and phosphate deficiency. *Mol. Plant* 3 (3), 524–538. doi: 10.1093/mp/ssp109

Robinson, W. D., Park, J., Tran, H. T., Del Vecchio, H. A., Ying, S., Zins, J. L., et al. (2012). The secreted purple acid phosphatase isozymes AtPAP12 and AtPAP26 play a pivotal role in extracellular phosphate-scavenging by arabidopsis thaliana. *J. Exp. Bot.* 63 (18), 6531–6542. doi: 10.1093/jxb/ers309

Ryan, P. R., James, R. A., Weligama, C., Delhaize, E., Rattey, A., Lewis, D. C., et al. (2014). Can citrate efflux from roots improve phosphorus uptake by plants? testing the hypothesis with near-isogenic lines of wheat. *Physiol. Plant* 151 (3), 230–242. doi: 10.1111/ppl.12150

Sanchez-Calderon, L., Chacon-Lopez, A., Perez-Torres, C. A., and Herrera-Estrella, L. (2010). "Phosphorus: plant strategies to cope with its scarcity," in *Cell biology of metals and nutrients* (Berlin, Heidelberg: Springer), 173–198. doi: 10.1007/978-3-642-10613-2_8

Shen, J., Yuan, L., Zhang, J., Li, H., Bai, Z., Chen, X., et al. (2011). Phosphorus dynamics: from soil to plant. *Plant Physiol.* 156 (3), 997–1005. doi: 10.1104/pp.111.175232

Srivastava, R., Parida, A. P., Chauhan, P. K., and Kumar, R. (2020). Identification, structure analysis, and transcript profiling of purple acid phosphatases under pi deficiency in tomato (Solanum lycopersicum l.) and its wild relatives. *Int. J. Biol. Macromol.* 165, 2253–2266. doi: 10.1016/j.ijbiomac.2020.10.080

Su, M., Meng, L., Zhao, L., Tang, Y., Qiu, J., Tian, D., et al. (2021). Phosphorus deficiency in soils with red color: Insights from the interactions between minerals and microorganisms. *Geoderma* 404, 115311. doi: 10.1016/j.geoderma.2021.115311

Swetha, S., and Padmavathi, T. (2016). Study of acid phosphatase in solubilization of inorganic phosphates by piriformospora indica. *Pol. J. Microbiol.* 65 (4), 7. doi: 10.5604/17331331.1227666

Syers, J. K., Johnston, A. E., and Curtin, D. (2008). Efficiency of soil and fertilizer phosphorus use. *FAO Fertilizer Plant Nutr. Bull.* 18 (108), 5–14.

Tabatabai, M. A., and Bremner, J. M. (1969). Use of p-nitrophenyl phosphate for assay of soil phosphatase activity. *Soil Biol. Biochem.* 1 (4), 301–307. doi: 10.1016/0038-0717(69)90012-1

Thudi, M., Chen, Y., Pang, J., Kalavikatte, D., Bajaj, P., Roorkiwal, M., et al. (2021). Novel genes and genetic loci associated with root morphological traits, phosphorus-acquisition efficiency and phosphorus-use efficiency in chickpea. *Front. Plant Sci.* 12, 636973. doi: 10.3389/fpls.2021.636973

Tian, J., and Liao, H. (2015). The role of intracellular and secreted purple acid phosphatases in plant phosphorus scavenging and recycling. *Annu. Rev. Plant Biol.* 48, 265–287. doi: 10.1002/9781118958841.ch10

Tian, J., Wang, C., Zhang, Q., He, X., Whelan, J., and Shou, H. (2012). Overexpression of OsPAP10a, a root-associated acid phosphatase, increased extracellular organic phosphorus utilization in rice. *J. Integr. Plant Biol.* 54, 631–639. doi: 10.1111/j.1744-7909.2012.01143.x

To, H. T. M., Nguyen, H. T., Dang, N. T. M., Nguyen, N. H., Bui, T. X., Lavarenne, J., et al. (2019). Unraveling the genetic elements involved in shoot and root growth regulation by jasmonate in rice using a genome-wide association study. *Rice* 12 (1), 1–18. doi: 10.1186/s12284-019-0327-5

Venkidasamy, B., Selvaraj, D., and Ramalingam, S. (2019). Genome-wide analysis of purple acid phosphatase (PAP) family proteins in jatropha curcas l. *Int. J. Biol. Macromol.* 123, 648–656. doi: 10.1016/j.ijbiomac.2018.11.027

Wang, X., Bai, J., Liu, H., Sun, Y., Shi, X., and Ren, Z. (2013). Overexpression of a maize transcription factor ZmPHR1 improves shoot inorganic phosphate content and growth of arabidopsis under low-phosphate conditions. *Plant Mol. Biol.* 31 (3), 665–677. doi: 10.1007/s11105-012-0534-3

Wang, X., Balamurugan, S., Liu, S. F., Ji, C. Y., Liu, Y. H., Yang, W. D., et al. (2021). Hydrolysis of organophosphorus by diatom purple acid phosphatase and sequential regulation of cell metabolism. *J. Exp. Bot.* 72 (8), 2918–2932. doi: 10.1093/jxb/erab026

Wang, X., Chen, Y., Thomas, C. L., Ding, G., Xu, P., Shi, D., et al. (2018). Genetic variants associated with the root system architecture of oilseed rape (Brassica napus l.) under contrasting phosphate supply. *DNA Res.* 24 (4), 407–417. doi: 10.1093/dnares/dsx013

Wang, L., Li, Z., Qian, W., Guo, W., Gao, X., Huang, L., et al. (2011). The arabidopsis purple acid phosphatase AtPAP10 is predominantly associated with the root surface and plays an important role in plant tolerance to phosphate limitation. *Plant Physiol.* 157 (3), 1283–1299. doi: 10.1104/pp.111.183723

Wang, L., Lu, S., Zhang, Y., Li, Z., Du, X., and Liu, D. (2014). Comparative genetic analysis of arabidopsis purple acid phosphatases AtPAP10, AtPAP12, and AtPAP26 provides new insights into their roles in plant adaptation to phosphate deprivation. *J. Integr. Plant Biol.* 56 (3), 299–314. doi: 10.1111/jipb.12184

Wang, J., and Zhang, Z. (2021). GAPIT version 3: boosting power and accuracy for genomic association and prediction. *Genom. Proteom. Bioinf.* 19 (4), 629–640. doi: 10.1016/j.gpb.2021.08.005

Wissuwa, M., and Ae, N. (2001). Further characterization of two QTLs that increase phosphorus uptake of rice (Oryza sativa l.) under phosphorus deficiency. *Plant Soil.* 237, 275–286. doi: 10.1023/A:1013385620875

Yan, X., Liao, H., Trull, M. C., Beebe, S. E., and Lynch, J. P. (2001). Induction of a major leaf acid phosphatase does not confer adaptation to low phosphorus availability in common bean. *Plant Physiol.* 125 (4), 1901–1911. doi: 10.1104/pp.125.4.1901

Yao, Y., Sun, H., Xu, F., Zhang, X., and Liu, S. (2011). Comparative proteome analysis of metabolic changes by low phosphorus stress in two brassica napus genotypes. *Planta* 233 (3), 523–537. doi: 10.1007/s00425-010-1311-x

Yin, C., Wang, F., Fan, H., Fang, Y., and Li, W. (2019). Identification of tea plant purple acid phosphatase genes and their expression responses to excess iron. *Int. J. Mol. Sci.* 20 (8), 1954. doi: 10.3390/ijms20081954

Zak, D., Kronvang, B., Carstensen, M. V., Hoffmann, C. C., Kjeldgaard, A., Larsen, S. E., et al. (2018). Nitrogen and phosphorus removal from agricultural runoff in integrated buffer zones. *Environ. Sci. Technol.* 2 (11), 6508–6517. doi: 10.1021/acs.est.8b01036

Zhang, D., Cheng, H., Geng, L., Kan, G., Cui, S., Meng, Q., et al. (2009). Detection of quantitative trait loci for phosphorus deficiency tolerance at soybean seedling stage. *Euphytica* 167 (3), 313–322. doi: 10.1007/s10681-009-9880-0

Zhang, H., Huang, Y., Ye, X., and Xu, F. (2010). Analysis of the contribution of acid phosphatase to p efficiency in brassica napus under low phosphorus conditions. *Sci. China Life Sci.* 53 (6), 709–717. doi: 10.1007/s11427-010-4008-2

Zhang, Q., Wang, C., Tian, J., Li, K., and Shou, H. (2011). Identification of rice purple acid phosphatases related to posphate starvation signalling. *Plant Biol.* 13 (1), 7–15. doi: 10.1111/j.1438-8677.2010.00346.x

Zhou, X., Zha, M., Huang, J., Li, L., Imran, M., and Zhang, C. (2017). StMYB44 negatively regulates phosphate transport by suppressing expression of PHOSPHATE1 in potato. *J. Exp. Bot.* 68 (5), 1265–1281. doi: 10.1093/jxb/erx026

frontiers | Frontiers in Plant Science

TYPE Mini Review
PUBLISHED 19 January 2023
DOI 10.3389/fpls.2022.1076094

OPEN ACCESS

EDITED BY
Vandna Rai,
National Institute for Plant Biotechnology (ICAR), India

REVIEWED BY
Shyam Sundar Dey,
Indian Agricultural Research Institute (ICAR), India

*CORRESPONDENCE
Debasish Kar
debasish.bios@gmail.com

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

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

RECEIVED 21 October 2022
ACCEPTED 28 December 2022
PUBLISHED 19 January 2023

CITATION
Anand A, Subramanian M and Kar D (2023) Breeding techniques to dispense higher genetic gains.
Front. Plant Sci. 13:1076094.
doi: 10.3389/fpls.2022.1076094

Breeding techniques to dispense higher genetic gains

Achala Anand†, Madhumitha Subramanian† and Debasish Kar*

Department of Biotechnology, Ramaiah University of Applied Sciences, Bangalore, India

Plant breeding techniques encompass all the processes aimed at improving the genetic characteristics of a crop. It helps in achieving desirable characteristics like resistance to diseases and pests, tolerance to environmental stresses, higher yield and improved quality of the crop. This review article aims to describe and evaluate the current plant breeding techniques and novel methods. This qualitative review employs a comparative approach in exploring the different plant breeding techniques. Conventional plant breeding techniques were compared with modern ones to understand the advancements in plant biotechnology. Backcross breeding, mass selection, and pure-line selection were all discussed in conventional plant breeding for self-pollination and recurrent selection and hybridisation were employed for cross-pollinated crops. Modern techniques comprise of CRISPR Cas-9, high-throughput phenotyping, marker-assisted selection and genomic selection. Further, novel techniques were reviewed to gain more insight. An in-depth analysis of conventional and modern plant breeding has helped gain insight on the advantages and disadvantages of the two. Modern breeding techniques have an upper hand as they are more reliable and less time consuming. It is also more accurate as it is a genotype-based method. However, conventional breeding techniques are cost effective and require less expertise. Modern plant breeding has an upper hand as it uses genomics techniques. Techniques like QTL mapping, marker assisted breeding aid in selection of superior plants right at the seedling stage, which is impossible with conventional breeding. Unlike the conventional method, modern methods are capable of selecting recessive alleles by using different markers. Modern plant breeding is a science and therefore more reliable and accurate.

KEYWORDS

plant breeding, CRISPR Cas-9, marker-assisted selection, genomic selection, genotypic technology

Introduction

Plant breeding can be defined as a process wherein, specific heritable changes are induced in plants through human efforts (Orton, 2020). It is an ongoing attempt in developing plants of superior phenotypes which produce more yield, resistance to diseases and abiotic stressors, synchronous maturity etc. (Bhargava and Srivastava, 2019). The scientific and technological advancements during the 20th century fastened the process of plant breeding and it was no longer just a skill, people took it up as a profession. Mendel's laws were of utmost value during

the 20th century and provided a sturdy framework within which breeding was extensively practiced. Watson and Crick's discovery of DNA as genetic material in the 1940s paved way for a novel era of biological discoveries which were appropriately incorporated by several plant breeders. During the late 20th century, the Mendelian laws along with the upcoming cell and molecular biology approaches, hastened the plant breeding industry. Currently, with techniques like genome sequencing and editing, the plant breeding industry has further been improved and has seen huge profits (Orton, 2020).

Higher yield, pest resistance in crops, etc. is required to keep up with the rapid growth and demand globally for food crops. In recent times, this has been achieved using advanced plant breeding techniques. Although they are extremely reliable and consistent with their results, they are a method of artificial growth. The potential risks, drawbacks and challenges of modern breeding methods are yet to be explored.

Conventional plant breeding

Plant breeding can be classified into two main types based on the methods and available tools, conventional and unconventional plant breeding. Conventional breeding can be defined as a process which involves the development of new varieties by using natural methods. In conventional breeding, desirable traits are put together from different gene pools but closely related by a process of cross hybridisation and therefore the product of conventional breeding is one in which pre-existing traits are mixed and matched to give rise to desirable crops (Acquaah, 2015).

Conventional plant breeding follows a particular sequence of steps. It begins with setting a clear objective, creating and assembling variation, selection, evaluation, release, multiplication and finally distribution of the new plant.

The species and intended application of the cultivar being created will determine the breeding goals. Prior to starting the breeding programme, the breeder must clearly identify its goals while taking the demands of end users into consideration. Thinking about it from the standpoint of cultivating the plant profitably (e.g. need for high yield, disease resistance, early maturity, lodging resistance, etc.). The processes should be taken into account when considering how effectively and affordably to use the cultivar as a source of raw materials for creating new goods (e.g. canning qualities, fibre strength, wood quality, mechanised production) (Figure 1).

The next stage is to gather the necessary germplasm for starting the breeding programme after deciding on the breeding objective(s). If, for instance, the goal is to breed for disease resistance, the gene causing the disease must be present when the base population is created. By artificially crossing suitable parents, the targeted gene is most frequently introduced into the base population. Breeders may try to incorporate the gene if it isn't already present. Mutagenesis is a widely used traditional technique for generating a gene that does not exist (Acquaah, 2015).

Breeders have created standardised breeding techniques for different species based on geneti characteristics. Species can be selected or bred using techniques based on their reproductive strategies, genetics, or whether the final result should be uniform or varied. A small number of genotypes emerge from the breeding process' final selection cycle as possible candidates for development into cultivars and release to farmers. These genotypes are put through a rigorous assessment process, which must take place in environments similar to those in which the cultivars will be produced for sale (Acquaah, 2015).

Advances in conventional breeding technologies and techniques

Breeders engage in two fundamental tasks when developing elite cultivars: they build or assemble germplasm and then they distinguish amongst (select) variability to find and promote suitable individuals who match the breeding objectives. These two actions account for a sizable portion of a breeding program's efficacy and efficiency. Breeders thus look for new or improved technology and procedures that support these efforts. Here are a few of the most important, each of which may also have a related approach.

Selection

The most basic method for crop enhancement is this, and it is employed by both skilled scientists and inexperienced farmers. In essence, it is the process of selecting and advancing suitable plants by differentiating among variety.

Artificial pollination

This kind of controlled pollination is used for a variety of purposes, including genetic research, breeding stock development, enhancing fruit set, and seed production.

Hybridization

It entails the use of managed pollination, which may be accomplished through artificial techniques. Breeders may create hybridization blocks where controlled pollination takes place, depending on the programme.

Wide crosses

Wide crosses for cultivated species are those that use components from outside the primary gene pool. They may entail a cross between two species or even two genera (intergeneric cross). The likelihood of genetic difficulties leading to infertility and poor success increases with the genetic distance between the parents.

Embryo culture

Because of infertility issues, the embryos resulting from especially wide crosses do not develop normally and need to be extracted prematurely.

Chromosome doubling

Wide interspecific crosses include parents with various chromosomal counts. Due to meiotic incompatibility, the hybrid arising from such crossings is reproductively sterile.

Doubled haploids

When a haploid cell experiences chromosomal doubling, which can occur *in vivo* or *in vitro* and be naturally occurring or purposefully generated in plant breeding, the outcome is a doubled haploid genotype. Doubled haploid methods are currently applicable to many hundreds of plants, however the technique has benefits and drawbacks. The main drawback is that the populace cannot be forced to participate in selection. Hybrid maize breeding has successfully used doubled haploid technology.

Bridge crossing

Bridge crossing is a method for crossing two parents with varying ploidy that is employed in broad crossings or to make a transitional or intermediate cross. The intermediate hybrid is given chromosomal doubling to make it fertile because it is reproductively sterile.

Protoplast fusion

This method is very beneficial for wide crossing. In situations where pollination and regular fertilisation are difficult or impossible, protoplasts may be fused in a laboratory setting to produce

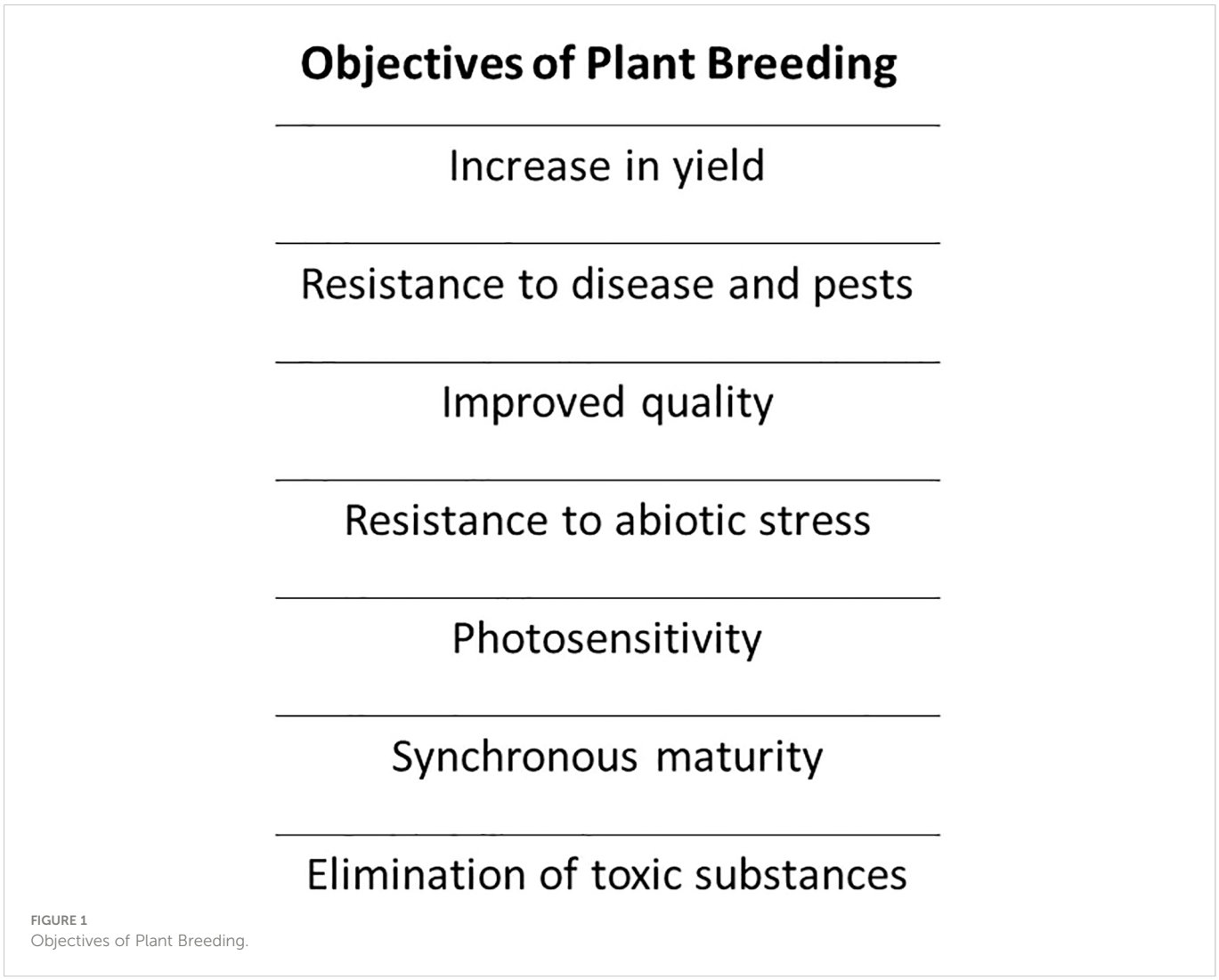


FIGURE 1
Objectives of Plant Breeding.

hybridization. In order to produce potato plants that were resistant to the potato leaf roll disease, somatic fusion was employed.

Modern plant breeding techniques

Modern plant breeding techniques came into being when molecular techniques was integrated along with conventional breeding techniques in order to achieve higher genetic gains. This was done by identifying the desired traits of the crop and their respective phenotypes and genotypes. On application of molecular techniques and genomics, the crop is enhanced.

There was a need for Modern breeding techniques due to the extensive time taken for Conventional breeding techniques. Sexual incompatibility and the sexual barriers in the form of pre and post fertilization were also major concerns (Bhargava and Srivastava, 2019). To achieve unique and more specific traits like higher absorption of nutrients, resistance to weeds, prevention of pests, decreased time of harvest, etc., conventional breeding techniques were combined with other branches of science (Lamichhane and Thapa, 2022).

In order to achieve higher genetic gains, genomic selection, enviromics and High Throughput Phenotyping (HTP) were utilised. It was also said that "Modern plant- breeding Triangle" consists of genomics, phenomics and enviromics (Crossa et al., 2017). Several staple crops like rice, wheat, sorghum and maize have successfully been bred using the above techniques (Table 1).

Marker assisted selection

In Marker Assisted Selection, genotypic markers are utilised in determining the phenotypic markers responsible for the desired trait with the help of bioinformatics tools. This method has proven to be much more efficient and accurate in comparison to the conventional method of direct phenotypic selection, which can be highly time consuming, more labour intensive and also less accurate. (Figure 2)

Markers are broadly categorised into 4 types- Morphological, Biochemical, Cytological and Molecular (DNA) based markers (Kumawat et al., 2020).

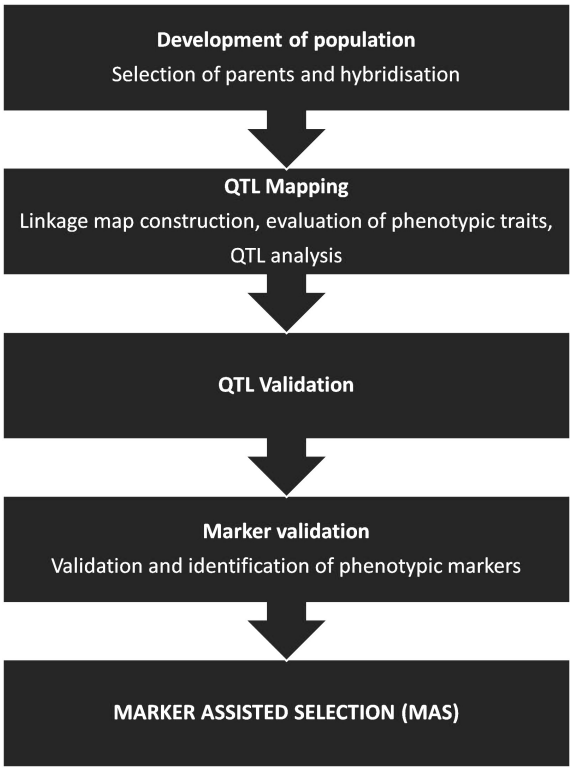


FIGURE 2
Marker development pipeline for Marker Assisted Selection (MAS).

Morphological markers

Morphological markers are phenotype- based markers. They are meant for traits that reflect the qualitative characteristics of a plant like colour of flower, shape of seed, height of plant, etc.

Biochemical markers

Biochemical markers are commonly known as isozymes. They are multi- molecular forms of enzymes that perform the same function but are coded by non- identical genes. These markers have a property of co-dominance. They have been proved successful in several

TABLE 1 Successful examples of modern plant breeding techniques.

Name of Technique	Mechanism of Technique	Example
Marker assisted selection Marker Assisted Selection (MAS)	GS works on employing DNA markers that are responsible for the expression of desired characteristics in crops	The DNA marker (9871.T7E2b) linked to the blast resistance phenotype in the presence of the Pi40 gene in a 70-kb chromosomal region was obtained from NBS-LRR disease resistance motif sequences. (Jeung et al., 2007)
Crispr Cas-9	CRISPR/Cas9 edits genes by accurately slicing DNA, which is then repaired by the body's own mechanisms. The Cas9 enzyme and a guide RNA make up the system's two components	CsLOB1 gene of Citrus was edited using Crispr Cas-9 to provide resistance to Citrus canker disease. (Nerkar et al., 2022)
High Throughput Phenotyping (HTP)	It is an advanced technology which generates phenotypic data of desired plant traits on utilising automated trait analysis and also involves automated sensing, data acquisition and data analysis.	SD1, Hd1, and OsGH3-2 genes of Rice (Oryza sativa L.) were modified to improve the crop yield and quality. (Xiao et al., 2022)

applications like detection of population structure, population subdivision, genetic diversity and gene flow, although, they don't have a large variety and don't identify polymorphism to a great extent.

Cytological markers

Cytological markers are chromosome- based markers that show variations in distribution of euchromatin and heterochromatin along with position, shape, size, number and order of chromosomes. They are widely used in physical mapping.

Molecular (DNA- based markers)

Molecular markers are most commonly used in plant breeding. These markers, not only identify the gene responsible for the desired trait, but also flag the gene such that it can be identified even in future generations. There are mainly 3 types of Molecular markers:

a. Hybridization-based markers: Examples: Restriction Fragment Length Polymorphism (RFLP)
b. Polymerase chain reaction (PCR)-based markers: Examples: RAPD, AFLP, SSR, chloroplast microsatellites (cpSSRs), randomly amplified microsatellite polymorphisms (RAMP), and intersimple sequence repeat (ISSR)
c. Sequence-based markers: Examples: Single nucleotide polymorphism (SNP) that were developed by the introduction of the DNA sequencing technologies like next-generation sequencing (NGS) and genotyping by sequencing (GBS) resulting in high polymorphism (Bhargava and Srivastava, 2019).

Genomic selection

In recent times, climate change has been quite drastic. Constant change in rainfall, soil acidity, increased temperature, etc. can significantly affect the growth and yield of crops.

Selection of crop variants through the method of phenotypic selection is laborious, time- consuming and does not always promise accurate results. Some crops even take a few years before they can express themselves. This method also provides room for considerable experimental error (Crossa et al., 2017).

Genomic selection uses DNA- based markers on a training population that express their genotypes as well as their phenotypes in order to predict the superior genes with desired traits.

The first a set of genotypes to be phenotyped is called as a training population, which are identified and phenotyped. A regression model is then trained to predict GEBVs for individuals which were not phenotyped. The GEBV (Genomic Estimated Breeding Values) are determined by adding the impacts of the genetic markers, or haplotypes of these markers, throughout the whole genome, that captures all quantitative trait loci (QTL) that affect any variation in trait (Ibeagha-Awemu and Khatib, 2017).

High throughput phenotyping

The rapid expansion of the food industry has in turn brought the need for the agronomic industry to produce more crops. HTP works on the basis of phenotyping (morphological, physiological, biochemical and molecular factors) as it is the initial and more important step of plant breeding. It is an advanced technology which generates phenotypic data of desired plant traits on utilising automated trait analysis and also involves automated sensing, data acquisition and data analysis. On using this technology, there is no need to wait for a plant to mature and express itself later at its life cycle, as it can be analysed and phenotyped at its initial stage of growth (Jangra et al., 2021).

Imaging technologies are used for plant phenotyping, which is mainly carried out by electromagnetic radiation. The properties of this radiation (absorption, emission, transmission, reflection and fluorescence) reflect the health of the plant, which cannot be detected by the naked eye. The spectral information is obtained through this advanced process of imaging. The ratio of intensity of reflected light to intensity of illuminated light is measured in different wavelengths. The reflectance is high on interaction with photoactive compounds like anthocyanins and chlorophyll when compared to protein ad water- rich regions which have low reflectance. There are different types of Imaging which are widely used- Visible Light Imaging, Fluorescence Imaging, Thermal Imaging and Tomographic Imaging (Jangra et al., 2021).

TABLE 2 Comparison between conventional and modern breeding techniques.

Conventional plant breeding	modern plant breeding
This breeding process places a greater emphasis on phenotype	This breeding process places a greater emphasis on genotype
It is labor- intensive and highly time consuming to develop a new hybrid	It requires lesser time, but needs more expertise to develop a new hybrid
Dominant genes are typically the only ones chosen. Recessive allele selection is a more process	Newly used technologies like Genomic selection, enviromics and High Throughput Phenotyping (HTP) Through the use of markers and the identification of particular gene locations, recessive alleles may be chosen in a very quick process
It is carried out physically by local breeders using basic tools, making it less accurate	Advanced technologies like Marker Assisted Selection (MAS), Genomic selection and High Throughput Phenotyping (HTP) are used

Novel plant breeding techniques

Genome editing is a technology used to manipulate genes precisely as Clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR associated protein (Cas) system. There have been some milestones in plant breeding using CRISPR Cas9 technology- i) *Solanum pimpinellifolium*, a progenitor of the tomato, may be rapidly domesticated by using CRISPR/Cas-mediated multiplex editing of "domestication genes" (e.g. loci linked to desired features). ii) Somatic homologous recombination (HR) utilising a homologous chromosome as a template can be used to repair targeted double-strand breaks (DSBs) caused by CRISPR/Cas (Rönspies et al., 2021).

Discussions and conclusion

The benefit of traditional plant breeding is that it increases the genetic resources that may be used to enhance crops by introducing the desired features. However, certain plants run the danger of losing genetic variety and becoming vulnerable to environmental stress. Therefore, the problems with global food security cannot be solved by using traditional agricultural techniques. Several conventional and molecular methods, such as genetic selection, mutagenic breeding, somaclonal variations, whole-genome sequence-based methods, physical maps, and functional genomic tools, have been used to improve agronomic traits related to yield, quality, and resistance to biotic and abiotic stresses in crop plants. Modern plant breeding techniques has, however, been brought about by recent developments in genome editing technologies utilising programmable nucleases, clustered regularly interspaced short palindromic repeats (CRISPR), and CRISPR-associated (Cas) proteins (Table 2).

In the current scenario of constant development, rapid climate change, unpredictable rainfall, etc., modern breeding techniques is required to produce the crops with consistent produce. There is also a need for increased yield with the expanding food industry. Therefore, using latest and advanced technology to the power of agronomic development is more practical and sustainable.

Author contributions

AA and MS have contributed equally to the development of this manuscript with guidance from DK. All authors contributed to the article and approved the submitted version.

Conflict of interest

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

Publisher's note

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

References

Acquaah, G. (2015). "Conventional plant breeding principles and techniques," in *Advances in plant breeding strategies: Breeding, biotechnology and molecular tools*. Eds. J. M. Al-Khayri, S. M. Jain and D. V. Johnson (Cham: Springer International Publishing), 115–158. doi: 10.1007/978-3-319-22521-0_5

Bhargava, A., and Srivastava, S. (2019). "Plant breeding," in *Participatory plant breeding: Concept and applications*. Eds. A. Bhargava and S. Srivastava (Singapore: Springer Singapore), 29–68. doi: 10.1007/978-981-13-7119-6_2

Collard, B., and Mackill, D. (2007). Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Philos. Trans. R. Soc. B.: Biol. Sci.* 363 (1491), 557–572. doi: 10.1098/rstb.2007.2170

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., Campos Los, G., et al. (2017). Genomic selection in plant breeding: Methods, models, and perspectives. *Trends Plant Sci.* 22 (11), 961–975. doi: 10.1016/j.tplants.2017.08.011

Ibeagha-Awemu, E. M. (2017). "Chapter 29 - epigenetics of livestock breeding," in *Epigenetic epidemiology* (Canada: Livestock Genomics), 23.

Jangra, S., Chaudhary, V., Yadav, R.C., and Yadav, N.R.. (2021). High-throughput phenotyping: A platform to accelerate crop improvement. *Phenomics* 1 (2), 31–53. doi: 10.1007/s43657-020-00007-6

Jeung, J. U., Kim, B. R., Cho, Y. C., Han, S. S., Moon, H. P., Lee, Y. T., et al. (2007). A novel gene, Pi40(t), linked to the DNA markers derived from NBS-LRR motifs confers broad spectrum of blast resistance in rice. *Theor. Appl. Genet.* 115 (8), 1163–1177. doi: 10.1007/s00122-007-0642-x

Kumawat, G., et al. (2020). Plant breeding - current and future views 175–195. doi: 10.5772

Lamichhane, S., and Thapa, S.. Advances from conventional to modern plant breeding methodologies. plant breed. biotech. 10(1):1~000. doi: 10.9787/PBB.2022.10.1.1

Nadeem, M. A., Nawaz, M. A., Shahid, M. A., Doğan, Y., Comertpay, G., Yıldız, M., et al. (2018). DNA Molecular markers in plant breeding: current status and recent advancements in genomic selection and genome editing. *Biotechnol. Biotechnol. Equip.* 32 (2), 261–285. doi: 10.1080/13102818.2017.1400401

Newell, M. A., and Jannink, J.-L. (2014). "Genomic selection in plant breeding," in *Crop breeding*. Eds. D. Fleury and R. Whitford (New York, NY: Springer New York (Methods in Molecular Biology), 117–130. doi: 10.1007/978-1-4939-0446-4_10

Nerkar, G., Devarumath, S., Purankar, M., Kumar, A., Valarmathi, R., Devarumath, R., et al. (2022). Advances in crop breeding through precision genome editing. *Front. Genet.* 13. doi: 10.3389/fgene.2022.880195

Orton, T. J. (2020). "Introduction," in *Horticultural plant breeding* (United States: Elsevier), 3–7. doi: 10.1016/B978-0-12-815396-3.09986-8

Rönspies, M., Schindele, P., and Puchta, H. (2021). *J. Exp. Bot.* 722, 177–183. doi: 10.1093/jxb/eraa463

Xiao, Q., et al. (2022). Advanced high-throughput plant phenotyping techniques for genome-wide association studies: A review. *J. Adv. Res.* 35, 215–230. doi: 10.1016/j.jare.2021.05.002

frontiers | Frontiers in Plant Science

TYPE Review
PUBLISHED 27 January 2023
DOI 10.3389/fpls.2023.1101271

OPEN ACCESS

EDITED BY
Nisha Singh,
Gujarat Biotechnology University, India

REVIEWED BY
Sivakumar Sukumaran,
The University of Queensland, Australia
Abhinandan Surgonda Patil,
Agharkar Research Institute, India
Puja Srivastava,
Punjab Agricultural University, India

*CORRESPONDENCE
Chiara Broccanello
chiara.broccanello@univr.it
Diana Bellin
diana.bellin@univr.it

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

RECEIVED 17 November 2022
ACCEPTED 09 January 2023
PUBLISHED 27 January 2023

CITATION
Broccanello C, Bellin D, DalCorso G, Furini A and Taranto F (2023) Genetic approaches to exploit landraces for improvement of *Triticum turgidum* ssp. *durum* in the age of climate change. *Front. Plant Sci.* 14:1101271. doi: 10.3389/fpls.2023.1101271

Genetic approaches to exploit landraces for improvement of *Triticum turgidum* ssp. *durum* in the age of climate change

Chiara Broccanello[1]*, Diana Bellin[1]*, Giovanni DalCorso[1], Antonella Furini[1] and Francesca Taranto[2]

[1]Department of Biotechnology, University of Verona, Verona, Italy, [2]Institute of Biosciences and Bioresources, (CNR-IBBR), Bari, Italy

Addressing the challenges of climate change and durum wheat production is becoming an important driver for food and nutrition security in the Mediterranean area, where are located the major producing countries (Italy, Spain, France, Greece, Morocco, Algeria, Tunisia, Turkey, and Syria). One of the emergent strategies, to cope with durum wheat adaptation, is the exploration and exploitation of the existing genetic variability in landrace populations. In this context, this review aims to highlight the important role of durum wheat landraces as a useful genetic resource to improve the sustainability of Mediterranean agroecosystems, with a focus on adaptation to environmental stresses. We described the most recent molecular techniques and statistical approaches suitable for the identification of beneficial genes/alleles related to the most important traits in landraces and the development of molecular markers for marker-assisted selection. Finally, we outline the state of the art about landraces genetic diversity and signature of selection, already identified from these accessions, for adaptability to the environment.

KEYWORDS

***triticum turgidum* ssp. durum, landraces, genotyping, breeding, molecular markers, climate change, abiotic stress**

Introduction

Durum wheat (DW) (*Triticum turgidum* L. ssp. *durum*) (Desf.) is the 10th most cultivated cereal worldwide, with a total production of about 38 million tons (Xynias et al., 2020). DW is grown primarily in the Mediterranean basin, accounting for 75% of global production, supported mainly by Italy, Spain, France, Greece, Morocco, Algeria, Tunisia, Turkey and Syria (Table 1). In addition, outside the Mediterranean basin, the major producers are Canada, Mexico, the USA, Russia, Kazakhstan, Azerbaijan, and India (De Vita and Taranto, 2019; Martínez-Moreno et al., 2022). Although DW production constitutes only 7% of total wheat production, the rest is produced from bread wheat (*Triticum aestivum*), its importance for the countries of the Mediterranean basin is pivotal. DW is considered a staple food as it

TABLE 1 Durum wheat of the major producer countries in European Union: area, production, yield and growing seasons (source: DG Agriculture and Rural Development based on Eurostat crop production annual data).

Country	Durum wheat area (thousand hectares)			Durum wheat production (thousand tonnes)			Durum wheat yield (tonnes/hectare)			Growing seasons
	2019	2020	2021	2019	2020	2021	2019	2020	2021	
EU	**2,145**	**2,112**	**2,206**	**7,476**	**7,420**	**7,822**	**3**	**4**	**4**	
Italy	1,224	1,210	1,229	3,849	3,885	4,065	3	3	3	July and August
Spain	267	251	298	704	787	744	3	3	2	June and July
Greece	254	263	220	684	794	573	3	3	3	From June to August
France	246	252	294	1,566	1,321	1,581	6	5	5	From June to August
Slovakia	44	34	49	188	174	292	4	5	6	From June to August
Hungary	37	27	30	162	121	160	4	4	5	From June to August
Germany	32	34	38	155	183	207	5	5	6	From June to August
Austria	17	17	19	81	79	88	5	5	5	From June to August
Others EU	26	25	29	86	75	113	n.a.	n.a.	n.a.	

constitutes the dominant part of the diet for many populations in this area. The main products derived from DW include pasta, cous cous, burghul, and bakery products. Durum wheat-based products have a low glycemic index which makes them healthy products that can be recommended for low-carb diets (Di Pede et al., 2021). Furthermore, DW constitutes the main food source for 1.2 billion poor people, providing 20/50% of daily calories, 20% of protein, and is considered a strategic crop for food security. Regarding the economic importance of DW, Italy is the world's largest producer of pasta with over 3.36 million tons/year of pasta produced, and the leading country for exports with 1.9 million tons/year (Altamore et al., 2020). On a cultural level, DW and its ancestor wild emmer (*Triticum turgidum* ssp. *dicoccoides*) have been at the foundations of food, from the Neolithic period to the Greeks and the Roman Empire, up to the present day (Martínez-Moreno et al., 2020). Its cultivation and processing constitutes a cultural heritage.

However, the on-going climate change threatens DW production and is subjecting this crop to stresses rarely experienced. In the Mediterranean area and western Europe, climate projections for the 2040-2070 interval warn that extreme drought events will become more frequent and severe due to decreased winter precipitation and increasingly dry springs (Spinoni et al., 2018). In a recent study, it was estimated that global warming may reduce the world's suitable areas for DW cultivation by 19% in 2050 and by 48% in 2100, with the greatest losses occurring in the Mediterranean basin which is recognized as a climate change hotspot (Shayanmehr et al., 2020; Martínez-Moreno et al., 2022). The main environmental constraints influencing the yield of DW in this area are drought, high temperatures, and salinity (De Vita and Taranto, 2019). These stresses, if occurring in growth stages such as flowering, pollination, and grain-filling, can strongly affect crop productivity.

This review aims to highlight the important role of durum wheat landraces as a useful genetic resource to improve the sustainability of Mediterranean agroecosystems, with a focus on adaptation to environmental stresses (Figure 1). We described the most recent molecular techniques and statistical approaches suitable for the identification of beneficial genes/alleles related to the most important traits in landraces and the development of molecular markers for marker-assisted selection. Finally, we outline the state of the art about landraces genetic diversity and signature of selection, already identified from these accessions, for adaptability to the environment.

Durum wheat cultivation: An historical overview

Durum wheat (tetraploid, genome AABB, 2n=4x=48) is a cereal grain evolved from the tetraploid domesticated emmer wheat *Triticum turgidum* ssp. *dicoccum* (Schrank ex Schübl.) Thell (Özkan et al., 2002). Domestication of wild emmer (*Triticum turgidum* L. ssp. *dicoccoides*) occurred in the Fertile Crescent (Israel, Jordan, Lebanon, Syria, south-eastern Turkey, northern Iraq, and western Iran) about 8000 years BCE (Before Common Era) from limited founder lineages (Özkan et al., 2002; Wang et al., 2022). Emmer wheat has been the main cereal together with einkorn and barley during the Neolithic period and the Bronze age. Then, it was progressively replaced by the tetraploid naked DW, spread by land through the Balkans and the maritime routes to the Mediterranean regions of Southern Italy, France, Spain, and Greece (Martínez-Moreno et al., 2020). Finally, DW became a prominent crop at about 300 BCE during the Hellenistic period. DW was commonly cultivated in the Roman Republic where the writers started to call it *Triticum* in Latin. Later, the early Islamic world and then the Arab empire further promoted the spread of the cultivation of DW through the Mediterranean areas by introducing several food types based on semolina (dry pasta and couscous). Until 1950-1955 most of the DW Mediterranean areas were cultivated with DW landraces, local accessions adapted to their place of origin (Martínez-Moreno et al., 2020). However, the first breeding activities were started in Italy in the early 1900s, resulting in the release of the most renowned cultivar Senatore Cappelli in 1915, by the breeder Nazareno Strampelli (Scarascia Mugnozza, 2005).

10.3389/fpls.2023.1101271

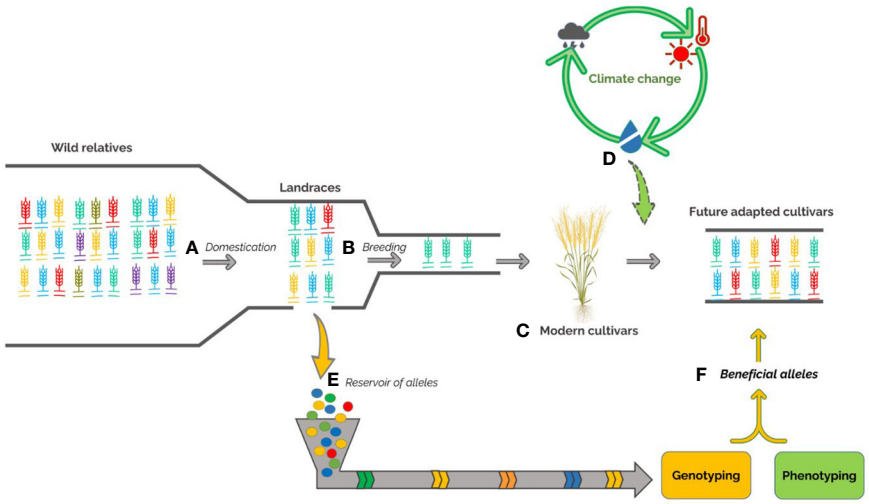


FIGURE 1
Development of durum wheat over time, including the loss of the diversity through the genetic bottlenecks of domestication **(A)** from wild relatives to the selection of landraces and plant breeding activities **(B)** moving from landraces to modern cultivars **(C)**. The emergence of climate change **(D)** requires to broad the genetic basis of modern cultivars. The exploration and exploitation of genetic variability within landrace germplasm **(E)** coupled to -omics approaches will be useful to discover beneficial alleles **(F)** and develop new modern cultivars best adapted to environmental changes.

Since then, the cv. Senatore Cappelli appeared in the pedigree of almost all new DW varieties due to its repeated use in DW programs until the end of the 1960s (Laidò et al., 2013). Afterward, the introduction of the law n. 580/67 for Pasta Pureness gave the impulse to start the development of private seed companies and international research centers, such as CIMMYT and ICARDA. In particular, the development of the CIMMYT-derived semi-dwarf wheat varieties led to the Green Revolution in several countries. During the period between 1960'-90', breeding activities coupled with mutagenesis generated new genetic variability (Xynias et al., 2020) and efforts were made to release every year improved varieties with high yield potential and other interesting traits. Such more productive cultivars were preferred for cultivation over large areas. Therefore, the multitude of DW landraces planted for centuries was progressively replaced. This replacement has led to an important erosion of the environmental adaptation traits evolved during the years by the landraces.

Environmental challenges for durum wheat cultivation

The climate characteristics of the Mediterranean region have played a significant role in shaping the phenotypic (and the genotypic) configuration of both DW wild relatives and cultivated varieties. This basin is characterized by hot and dry summers, followed by cold and wet winters. Climate change, particularly important in the last decades, points to an increased variability, in which drought events, often coupled with heat waves, can hamper growth and development, eventually affecting crop yield. For example, yield is reduced of about 5% per Celsius degree of increase in temperature, with a gross loss reaching 24% under a growth temperature of 31°C during flowering (Liu et al., 2016). Clearly, the negative effect of the abiotic stress depends on its duration and the phenological phase of the plant. For instance, sudden and extremely high temperature (T > 32°C) for a short duration (3 to 5 days) is referred to as a *heat shock*, while moderately high maximum temperature (20 to 30°C) for a longer duration is known as *chronic heat stress* (Li et al., 2013). In DW, the most sensitive stages to heat stress are anthesis and grain filling (Chaparro-Encinas et al., 2021). Heat stress alters membrane fluidity and enzyme activity which in turn hamper respiration and photosynthesis, and related processes (e.g. electron flow and carbon fixation metabolism, starch accumulation and stability, architecture and functioning of thylakoids), as well as water assimilation and nutrient absorption and allocation in the plant body. After phase transition, this results in compromised pollen viability, aberrant macrosporogenesis, starch synthesis, and grain filling, responsible for the reduction in yield. Reduced water availability, due to both erratic or deficient rainfall and limited irrigation, worsens the negative effect of heat stress, hindering grain yield (in terms of seed number and weight) and technological quality and protein composition (Flagella et al., 2010). Drought stress is induced also by soil physical characteristics, which significantly affect water holding capacity and supply, influencing water and nutrient absorption by roots. Plants respond to drought and heat stress by enacting similar physiological mechanisms. Transcriptome analysis of heat susceptible and tolerant wheat revealed the involvement of multiple processes associated with tolerance to heat shock and drought stress, including the formation of deeper roots, synthesis of heat shock proteins, stomatal control, coordination of transpiration rate, and enhancement of osmoprotective response (Kulkarni et al., 2017). Also, the use of genome wide mapping approaches is providing abundant information about genomic regions associated to heat tolerance (Sukumaran et al., 2018).

Soil geo-biochemistry, geographical localization (sea proximity, with seawater intrusion into freshwater aquifers), and events of rising groundwater table can increase the amount of salts in soils. Moreover,

anthropogenic activities, such as inappropriate irrigation and drainage practices or irrigation with brackish water, determine salt accumulation in the soil surface or within the solum, causing salinity stress in plants (Annunziata et al., 2017). Soil salinity is usually referred to the increased amount of Na^+/Cl^- in the soil upper layer, but a variety of ions, mainly K^+, Ca^{2+}, Mg^{2+}, and SO_4^{2-}, HCO_3^-, NO_3^-, can also accumulate. On one hand, salinity leads to permanent modifications of the soil structure by decreasing soil aeration, leaching, and infiltration rate, and increasing runoff and soil erosion (Edelstein et al., 2010). On the other hand, salinity affects plant physiology and growth. Stress effects harbored by salinity are usually due to both (*i*) cell toxicity of the accumulated ions, which often results in nutrient imbalance and enhanced oxidative stress, and (*ii*) osmotic stress, due to the extremely low water potential of soil along with a reduction in water assimilation. Therefore, cellular, and metabolic processes involved to counteract salt stress are comparable to those induced by drought (Munns et al., 2006). Interestingly, DW is conventionally considered a tolerant crop enduring up to 5.9 dS/m (De Santis et al., 2021). In field conditions with 10 dS/m NaCl, DW produced a reduced yield, compared with rice that died before maturity (Munns et al., 2006). Lower level of salinity reduces the leaf area and shoot biomass, while grain yield is not affected. Tolerance to salinity is associated with low rate of Na^+ root-to-shoot transport and higher selectivity for K^+ than for Na^+. Indeed, a correlation between grain yield and Na^+ exclusion from the vegetative organs, together with the enhanced K^+/Na^+ discrimination in root absorption and xylem loading has been identified in tolerant genotypes (Munns et al., 2000), which confirmed xylem transport, as one of the main discriminants between sensitive and tolerant species (Davenport et al., 2005). Tolerance to salinity is a quantitative trait controlled by many genes. Moreover, it appears that the expression of genes responsible for salt tolerance depends on plant age and ontogeny. Also, environmental factors contribute to the large phenotypic variation reported, enhancing the difficulty of breeding programs aimed to improve salt tolerance (De Santis et al., 2021). In wheat, a QTL mapping approach has identified the locus Nax1 (involved in limiting Na^+ translocation to the above-ground tissues and inducing salt tolerance), mapped to the long arm of chromosome 2A, responsible for almost 38% of phenotypic variation in low Na accumulation in the mapping population, and this locus, together with closely linked markers, are commonly adopted to select salt tolerant durum genotypes (Lindsay et al., 2004). Other characteristics of salt-tolerant genotypes include differential ion partitioning between aged and young leaves, cell osmotic adjustment contrasting osmotic stress, and early phase-transition, leading to a shorter growing season (Colmer et al., 2005).

Drought, heat, and salt stress, being linked to each other, induce the generation of reactive oxygen species (ROS), including hydrogen peroxide, superoxide, or hydroxyl radicals, which are continuously formed mainly in the cytosol, chloroplasts, and mitochondria (Laus et al., 2022). ROS have a significant role in signaling but, under stress conditions, their over-accumulation may be responsible for the oxidative stress characterized by membrane peroxidation, protein degradation, and DNA mutation, eventually leading to the death of the plant cell. Plant cells are usually equipped with a great variety of ROS scavenging enzymes including superoxide dismutase, catalase, and glutathione peroxidase, and antioxidants, such as ascorbic acid, tocopherol or glutathione, which also contribute to ROS detoxification (Dvorak et al., 2021). Interestingly, the tolerance of DW genotypes to environmental stresses leading to ROS production has been widely associated with higher activities of scavenging enzymes, pointing to a role of these mechanisms in the drought and salt tolerance in particular genotypes (Laus et al., 2022). Therefore, they are a good candidate to be considered in DW breeding programs. Anyway, it should be stressed that as the DW sensitivity to stress is influenced by the phenology, also the antioxidant performance depends on the stress characteristics (severity and duration), on the stage of development at which the stress acts, and on the plant organs targeted. Finally, (as shown by the increasing literature on the topic, Li et al., 2013; De Santis et al., 2021), breeding programs should also keep the attention on the stress effect on quality traits of DW grains. Indeed, changes in grain content and composition, affecting technological and health quality (e.g. protein, starch and dietary fiber accumulation and composition, phytochemical, and health-related micronutrient accumulation), are incredibly susceptible to environmental clues and stress and must be taken into account when implementing the breeding programs.

Durum wheat landraces: An endless treasure

Many efforts are made by researchers and breeders to constantly look for new sources of genetic variability to improve the elite varieties for adaptation traits, to face climate change. The exploitation of the existing genetic variability, still available in landraces, is one of the best approaches to adopt. According to Camacho Villa et al. (2005), a landrace is "a dynamic population of a cultivated plant that has a historical origin, distinct identity, and lacks formal crop improvement, as well as often being genetically diverse, locally adapted, and associated with traditional farming systems". It is the result of natural and/or farmer-mediated evolutionary forces that generated plants better adapted to the local climate/environmental conditions (Zeven, 1999).

They are considered a reservoir of useful alleles that can be exploited to broaden the genetic basis of important adaptation traits. Landraces are rich in micronutrients and have high concentrations in total phenol and antioxidant content, as well as in tocols, carotenoids, and lutein (Azeez et al., 2018). Since the landraces can tolerate abiotic and biotic stresses, their yield is lower than modern varieties (Tan, 2002). For this reason, landraces are no longer cultivated over large areas where the more productive cultivars are preferred. Anyhow, several landraces have been rediscovered and reused thanks to their adaptation to sustainable and low-input cropping systems. They produce a great amount of straw, which, when used for animals, can make them economically more convenient than modern varieties, or preferable for organic farming because of their greater competitive ability against weeds (Lemerle et al., 2001; Annicchiarico et al., 2005).

Indeed, thanks to the efforts of farmers and scientists, wheat landraces and old cultivars have been collected and conserved by *in-situ* or ex-situ strategies. The *in-situ* strategy relies on individual farmers who traditionally cultivate landraces for their production or are sponsored by the government or private companies. The ex-situ

conservation is managed by international institutions such as CIMMYT, ICARDA, and USDA or by national projects led by local universities (Adhikari et al., 2022). With the advance of modern technologies, phenotyping and genotyping are extremely affordable, and the landraces can be studied both for their conservation and for molecular markers development (Nazco et al., 2012; Marone et al., 2021). The exploration of genetic variability in landrace germplasm has become an issue of great global interest, mainly during the last two decades. Figure 2 shows how the number of publications related both to "plant breeding" and "landraces", and "plant breeding" and "climate change", have had a strong increase since 2005, when "The 2005 United Nations Climate Change Conference" took place and marked the entry into force of the Kyoto Protocol. From 2005 to 2021 the number of publications related to the plant breeding strategies, to cope with climate change, has grown exponentially. Also for "durum wheat" and "landraces", it can be noted a similar trend, although the number of publications is scarce. Some works aimed to characterize specific DW landraces are reported in Table 2, which includes a list of the most important European landraces specially studied both for stresses and quality related traits.

Because of the genetically heterogeneous nature of landraces, which are in a constant state of evolution due to different factors such as ecogeographical area and conventional or modern breeding techniques (Casañas et al., 2017), the establishment of core collections represents an affordable cost approach to reduce the degree of co-ancestry redundancy and the genetic stratification in the landraces whole collections. The goal of creating core collections is to maximize the allelic richness at molecular markers and best represent variation at phenotypic traits, in order to define the smallest possible number of individuals that represents a more compact and manageable population. Core collections were made for Spanish, Indian, Iranian, and Israeli-Palestinian landraces (Etminan et al., 2017; Ruiz et al., 2012 and 2013; Frankin et al., 2021; Phogat et al., 2019; Chacón et al., 2020). In addition, a core set of landraces was developed starting from the Global Panel of Durum Wheat (GPD), reducing their number from 416 to 192 (Mazzucotelli et al., 2020 and Kabbaj et al., 2017). This approach is useful not only to represent whole genetic diversity but also to enquire and identify new sources for interesting traits. For example, SNP markers associated to resistance to leaf rust, tan spot and *Stagonospora nodorum* blotch were identified using the core collection created from the Watkins collection (Martínez-Moreno et al., 2021, Halder et al., 2019). Moreover, the core collections have also been used for unlocking the genetic and morpho-physiological adaptation traits to semi-arid environments (Abu-Zaitoun et al., 2018) and to study agronomic and quality traits, such as root architecture, stem cross section, height, heading date and carotenoid content (Ruiz et al., 2018; Ávila et al., 2021; Requena-Ramírez et al., 2021).

Durum wheat genome and pangenome assemblies

The DW sequencing project has been carried out by Maccaferri et al. (2019) using the modern cultivar Svevo. The annotation led to the identification of 66,559 genes, where the gene density distribution reflects the QTL density distribution. The comparison between wild emmer Zavitan and Svevo genomes identified putative loci under domestication and selection, and localized the reduction in diversity mainly in the pericentromeric regions of the chromosomes (Maccaferri et al., 2019). However, in the evolution of DW landraces, the reduction of diversity was more spread along the genome as a consequence of breeding programs.

Resequencing techniques, such as whole genome resequencing, are not suitable for species with complex genomes, for which a reduction of genomic complexity prior to NGS-based SNP discovery is preferred (Borrill et al., 2019). In polyploid species such as DW, with a large genome size (about 10.45Gb) and a large proportion of repetitive sequences (> 85%), the presence of paralogous and multi-copy sequences adds complexity in classifying the correct scoring of SNPs at a single locus for SNP discovery (Sandve et al., 2010).

In the last decades, the number of sequenced genomes in crop species has continued to increase exponentially, highlighting the

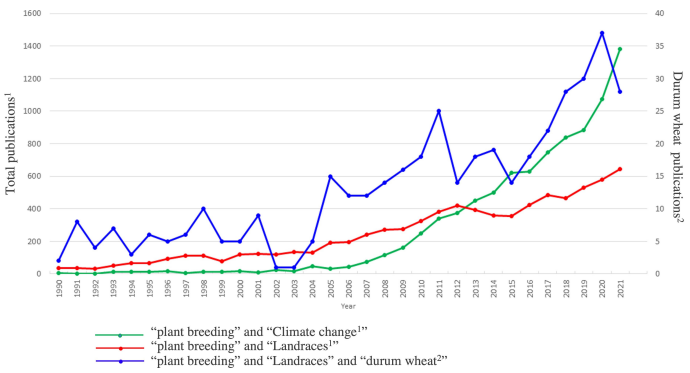


FIGURE 2
Number of publications in which climate change, landraces, and durum wheat are associated to plant breeding. The scale of the primary vertical axis shows the values for the associated data series in green: complete query = ("plant breeding" AND "climate change") and red: complete query = ("plant breeding" AND "landraces"); the scale of the secondary vertical axis shows the values for the associated data series in blue: ("plant breeding" AND "landraces" AND "durum wheat"). The analysis is based on the information available in the Web of Science database (www.webofknowledge.com), category "Plant science", considering the time interval of 1990–2021. Different keywords (i.e., "plant breeding", "landraces", "climate change" and "durum wheat") and Boolean operators were used to query the database.

10.3389/fpls.2023.1101271

TABLE 2 List of the most important durum wheat landraces.

Country	Common name/Accession	Trait	References
Italy	Tumminia SG3	polyphenols profile and content	Lo Bianco et al., 2017
Italy	Scavuzza	polyphenols profile and content	Lo Bianco et al., 2017
Italy	Russello SG8	polyphenols profile and content	Lo Bianco et al., 2017
Italy	Ruscia	polyphenols profile and content	Lo Bianco et al., 2017
Italy	Manto di Maria	polyphenols profile and content	Lo Bianco et al., 2017
Italy	Margherito	polyphenols profile and content	Lo Bianco et al., 2017
Italy	Biancuccia	polyphenols profile and content	Lo Bianco et al., 2017
Italy	Bidì	suitable characteristics for malting and brewing	Alfeo et al., 2018
Italy	Francesa	suitable characteristics for malting and brewing	Alfeo et al., 2018
Italy	Gioia	suitable characteristics for malting and brewing	Alfeo et al., 2018
Italy	Giustalisa	suitable characteristics for malting and brewing	Alfeo et al., 2018
Italy	Inglesa	suitable characteristics for malting and brewing	Alfeo et al., 2018
Italy	Martinella	suitable characteristics for malting and brewing	Alfeo et al., 2018
Italy	Bufala Bianca	malt charaterisrics suitable for brewing	Alfeo et al., 2021
Italy	Bufala nera corta	malt charaterisrics suitable for brewing	Alfeo et al., 2021
Italy	Bufala rossa lunga	malt charaterisrics suitable for brewing	Alfeo et al., 2021
Italy	Russello	high content of antioxidant phenolic compounds	Visioli et al., 2021
Italy	Trentino	suitable characteristics for malting and brewing; polyphenols profile and content	Alfeo et al., 2018; Lo Bianco et al., 2017
Italy	Tripolino	polyphenols profile and content	Alfeo et al., 2018; Lo Bianco et al., 2017
Italy	Urria	polyphenols profile and content	Alfeo et al., 2018; Lo Bianco et al., 2017
Italy	Timilia	high content of antioxidant phenolic compounds	Taranto et al., 2022
Portugal	PI 192051	stem rust resistance sources	Aoun et al., 2019
Portugal	Aus26582	leaf rust resistance	Qureshi et al., 2018
Portugal	Aus26579	leaf rust resistance	Qureshi et al., 2018
Cyprus	IG-82549	glutenin protein composition	Moragues et al., 2006
Portugal	Lobeiro de Grao Escuro	having high EU quality index and high protein quality	Nazco et al., 2014a
France	Trigo Glutinoso	having high EU quality index and a high sedimentation index	Nazco et al., 2014b
Spain	BGE-013614	glutenin protein composition	Moragues et al., 2006
Spain	BGE018675	higher zeaxanthin relative content	Requena-Ramírez et al., 2021
Spain	BGE045643	higher zeaxanthin relative content	Requena-Ramírez et al., 2021
Spain	BGE045657	higher zeaxanthin relative content	Requena-Ramírez et al., 2021
Spain	BGE018321	higher relative β-carotene content	Requena-Ramírez et al., 2021
Spain	BGE045628	higher relative β-carotene content	Requena-Ramírez et al., 2021
Spain	BGE045633	higher relative β-carotene content	Requena-Ramírez et al., 2021
Spain	BGE030921	highest α-carotene content	Requena-Ramírez et al., 2021

presence of large structural variations between individuals of the same species. Therefore, relying on the single reference genome cannot represent the entire sequence diversity present in a population (Golicz et al., 2016). This observation led to the concept of "pangenome", that describes the landscape of genetic variation within a species, in order to capture a comprehensive view of genetic diversity that include the entire crop gene pool (Bayer et al., 2020; Khan et al., 2020). In the pangenome development, it is pivotal to consider the genetic variation

brought by the crop wild relatives, in order to include as much genetic variability as possible (Khan et al., 2020). In fact, wild species preserve important genes related to tolerance to various types of stress that were lost during the domestication process. The increasingly less expensive sequencing approaches have allowed to deepen the genetic architecture of the crop wild relatives leading, in the past decade, to several *de novo* sequencing projects developed also in crop wild relatives. In soybean, 14 cultivated and 17 wild accessions were resequenced, confirming the great allele diversity present in wild accessions (Zhou et al., 2015). In maize, 75 lines including cultivated, wild and landrace accessions were resequenced, highlighting the genes linked to selection and providing evidence for introgression from wild relatives (Hufford et al., 2012).

In wheat, the first pangenome has been constructed using the bread cv. Chinese Spring as suitable reference assembly, followed by the expansion of this reference with 16 additional sequences from other bread wheat varieties (Montenegro et al., 2017). Graph pangenomes based on 16 public assemblies (Wheat Panache) was developed with the aim to discover genome variation between cultivars and to mine the diversity present in the large and complex wheat genome (Bayer et al., 2022). However, the mathematical modeling of pangenome expansion revealed that all these wheat varieties have a closed pangenome; therefore, the inclusion of more distant accessions such as wild relatives and landraces could harbor yet unexplored sequence variants that may further increase the gene content of the pangenome. The use of divergent individuals may increase the number of novel gene variants as well as improve the accuracy of the read mapping for SNP discovery. The use of landraces can support the breeding of cultivars better adapted to diverse environments and more resilient to climate change; indeed, plant pangenome assemblies have shown that genetic variations are often associated with biotic or abiotic stress.

No pangenome has yet been assembled for durum wheat, although several projects are underway. Indeed, the use of Svevo genome as suitable reference may accelerate the sequencing of new durum cultivars enabling the pangenome construction. As far as is known, at the moment the only reference genome for durum wheat remains Svevo. Given the growing interest in some European landraces (Table 2), the release of new genome assemblies from landraces is expected in the next few years.

High-throughput genotyping techniques

Exome and RNA sequencing

A widely employed method for *de novo* SNP discovery and genotyping in large genome-size species is the exome sequencing. The workflow involves the fragmentation of high-quality genomic DNA, repair ends, adenylation, adapter ligation, and the selective hybridization of probes for target enrichment. Then two consecutive captures of the hybridization probes ensure high specificity of target region before the sequencing step (Kaur and Gaikwad, 2017). Ready to use exome kits and their customization are available for many crop species such as wheat (Harrington et al., 2019). Exome sequencing techniques have been used to identify polymorphisms and genes in the tetraploid wheat genome, also in combination with bulk segregant analysis (Saintenac et al., 2011; Mo et al., 2018). In the last few years, thanks to the release of wheat reference genomes and annotations, this approach is increasingly used.

An alternative to exome capture is the high-throughput RNA sequencing which analyzes sequence variations in the transcribed portion of the genome. RNA-seq is the technique of choice for the identification of new genes and isoforms, and for the detection of variants including expressed SNPs and INDELs. Furthermore, this approach allows the identification of differentially expressed genes in plants under stress conditions. However, this technique is expensive since the reagents for the sequencing of the entire transcriptome are required. To overcome this problem, a new technique that is emerging also in plants (it is already widely used in human genetics) is the target RNA sequencing. This technique allows very high precision in the discovery and quantification of genes because it sequences only those of interest. The most important step of target RNA-seq is the design of the specific probes that can be customized to meet the specific needs of each experiment. Since only the genes of interest are sequenced, the coverage can also be very high, allowing to increase the power to assemble low expression transcripts (Ostezan et al., 2021).

Reduced representation sequencing (RRS)

With the decrease in the NGS cost, sequencing techniques are increasingly used as genotyping tools. However, to afford sequencing in large genome size species, reduced representation sequencing (RRS) approaches can be considered (Davey et al., 2011). Genotyping-by Sequencing (GBS) is the most used technique to greatly reduce genome complexity using restriction enzyme(s) digestion (Elshire et al., 2011; Rasheed and Xia, 2019). The digestion occurs in the presence of a specific combination of enzymes recognizing rare-rare, frequent-rare, or frequent-frequent restriction sites. In wheat, since the complexity of the genome is very high, it is normally used a double enzyme digestion (Poland et al., 2012). The digested DNA is then ligated with adapters, amplified through PCR, and sequenced. The generated data are directly used for genotyping (Deschamps et al., 2012). Typically, the sequencing involves 100-150 bp. This technique has a simple protocol, specific and reproducible, with a reduced sample handling, without the need of a reference genome (Davey et al., 2011). These properties make GBS a genotyping technique suitable for a number of species and genetic studies (Chung et al., 2017), such as genomic selection (Poland et al., 2012), SNP marker development (Forrest et al., 2014), and genetic diversity (Alipour et al., 2017). Diversity Array Technology (DArT), developed by Diversity Arrays Technology Pty Ltd. (Canberra, Australia) originally with the microarray technology platform, is one of the GBS-based techniques widely adopted in wheat, thanks to its versatility, accuracy, and low cost (Colasuonno et al., 2021).

SNP Array

NGS technology has also created the basis for the establishment of high-density SNP arrays and the related high-throughput platforms capable of genotyping large numbers of samples (Ganal et al., 2012).

Currently, the most widely used genotyping platforms for large scale genotyping are the Infinium platform from Illumina (San Diego,

CA, USA) and the Axiom technology from Thermo Fisher Scientific (Waltham, MA, USA) (Scheben et al., 2018). Technically, the Illumina technology is based on spheres covered with specific oligos adapted to the microwells and the amplification takes place on a single-base two-color extension with a single probe SNP marker (Steemers and Gunderson 2007). On the other side, the GeneChip® array of Affymetrix uses photolithographic oligos on an array and the target SNP amplification involves assays with 30-mer probes (Thomson, 2014).

In wheat, the first SNP array developed was the 9K Infinium SNP Array by Cavanagh et al. (2013) used for genotyping 2,994 lines of bread wheat. Later, the array with 90K SNPs was fine-tuned (Wang et al., 2014). However, both of these chip had a greater representation in cultivated varieties, thus their use was very limited in the study of landraces (Rasheed et al., 2018). This problem was overcome by the development of the 820K Affymetrix Axiom SNP array which relied on exomes sequencing of 43 bread wheat and wild species accessions (Winfield et al., 2016). Axiom 35K SNP was then developed from this array, capable of analyzing even wild accessions at a more limited cost (Allen et al., 2017). In parallel, the Chinese Academy of Agricultural Sciences (CAAS) also developed an array containing 660K SNPs (Jin et al., 2016). More recently, in 2019, Wheat 50K (Triticum TraitBreed array, Rasheed and Xia, 2019) and 15K SNP arrays were developed (Muqaddasi, 2017) containing a selection of SNPs from Wheat 35K, 90K, and 660K SNP arrays.

SNP arrays have the advantage of facilitating high-density SNP scanning, have a high call rate, and are also cost effective when there is a need to genotype a high number of markers on many samples. However, a disadvantage, is that the set of SNPs is fixed and cannot be changed; they are also developed in hexaploid wheat, thus the SNPs present in the D genome will be unknown if it is applied in DW.

Genotyping for marker assisted breeding

Among the most competitive technologies used for marker assisted breeding, with a medium/low throughput, there are TaqMan (Applied Biosystems, Foster City, CA), KASP (Kompetitive allele specific PCR, Hoddesdon, UK), and rhAmp (Integrated DNA Technology technologies, Redwood City, CA), widely used in many plant species such as wheat and sugar beet (Broccanello et al., 2018; Ayalew et al., 2019). TaqMan chemistry is based on fluorescently-tagged, allele-specific probes detected using real-time PCR-based assays, while KASP technology adopts an endpoint fluorescence detection to discriminate tagged SNP alleles. The most recent method is rhAMP, that uses RNase H2 to activate primers after successful binding to their target site. All these chemistries are suitable for use on a variety of real-time PCR instruments with different throughputs. For example, the TaqMan assays can be applied in Real-Time PCR, but also can be used with the Open Array technology (Thermo Fisher Scientific, Carlsbad, CA) which allows the simultaneous analysis of 4 OpenArray plates, each composed of 3,072 through-holes allowing the genotyping analysis of 16, 32, 64, 128, 192, 256 SNPs at a time (Broccanello et al., 2020). Perry and Lee, (2017) developed an OpenArray plate composed of 16 SNP markers able to discriminate 47 DW varieties registered for production in Canada.

These chemistries have the advantage of being highly reproducible, sensitive, and cost effective; moreover, they can be freely customized both for the number of samples and SNPs that can be analyzed, adapting perfectly to marker assisted selection for crop improvement (Broccanello et al., 2018).

Advancements in trait genetic dissection and breeding

Genome wide association study (GWAS)

Genome wide association study is a powerful tool to study the genetic base of complex traits and detect relationships between phenotypic variations and the associated genetic polymorphisms (Taranto et al., 2018). The statistical methods for the analysis of associations have improved over the years, going from a classic ANOVA, that generates many false positives to the development of the mixed model framework, which increases computational speed and improves statistical power (Pavan et al., 2020; Tibbs Cortes et al., 2021). Subsequent advancement in statistical analysis methods led the association analysis of all markers simultaneously. This approach is based on Bayesian methods which are normally used in genomic prediction (Fernando and Garrick, 2013). However, the most common actual methods include TASSEL (Bradbury et al., 2007), GAPIT (Lipka et al., 2012), and GEMMA (Zhou and Stephens, 2012). 58 candidate genes associated with salt tolerance have been found, in bread wheat, performing 5 multi locus GWAS models that include mrMLM, FASTmrMLM, FASTmrEMMA, pLARmEB, and ISIS EMBLASSO (Chaurasia et al., 2020; Esposito et al., 2022). In DW genome wide association studies often involve the use of landraces to identify new causative SNPs. Moreover, the genomic regions linked to wheat blast resistance were identified in Indian genotypes with a MLM (mixed linear mode) in TASSEL. A novel GWAS approach is the environmental GWAS (envGWAS) that associates the single nucleotide polymorphisms with the geographic information system (GIS) of the original samples collection sites. In this context, the genome wide association was performed to study the local adaptation of Iranian and Pakistani bread wheat landraces using an EigenGWAS approach and a fixed and random model circulating probability unification (FarmCPU) (Hanif et al., 2021).

Genomic selection

In a context of climate change, a technique developed to accelerate breeding procedures and speed up the selection of superior genotypes is genomic selection (GS) (Crossa et al., 2017). This statistical model uses SNP molecular markers for a genomic prediction of genotype performance. The aim of GS is to predict breeding and/or genetic values. GS uses genotypic and phenotypic data for the constitution of a training population and then the predictive equation is used to select candidates that have been genotyped but not phenotyped. GS has the advantage of being able to rapidly improve complex and low heritability traits and reduce the cost of hybrid development. This technique can also be used for less complex traits with high inheritance and, for this scenario, high

genomic prediction (GP) accuracy is expected. However, when a trait is controlled by a high number of loci there are several factors influencing the prediction accuracy such as the size and genetic diversity and how distant is the training from the testing population. Moreover, for complex traits with large numbers of markers that are not in linkage disequilibrium (LD) with the QTL, GP accuracy is lower (Daetwyler et al., 2010).

In general, the statistical models developed for GP are based on single-environment assessments. However, in plant breeding the presence of a Genotype x Environment (G x E) interaction may complicate the selection of stable lines. Hence, some genomic prediction models, considering the G x E interaction, help breeder select lines with optimal overall performance across different environments and in a specific target environment. Specifically, a reaction norm model, which is an extension of the random effect Genomic Best Linear Unbiased Predictor (GBLUP) model, was developed by Jarquín et al., 2014. In this model, the main effect of lines, the main effect of environments, the main effect of markers, the main effect of pedigree, and their interactions with environments, are modeled using random covariance structures that are functions of marker or pedigree genotypes and environmental covariates (Jarquín et al., 2014). Appropriate cross-validation schemes are designed to obtain valid and unbiased estimates of the predictive ability obtainable from the developed genomic prediction models (Roberts et al., 2017). The reaction norm model has already been applied for the genomic prediction of 8,416 Mexican wheat landrace accessions and 2,403 Iranian wheat landrace accessions from the CIMMYT by Crossa et al. (2016). In this work, the authors evaluated two traits in two different environments and some heritable traits in a single optimal environment. The accuracy of the prediction for some traits such as maturity, quality traits, and grain yield and yield components was around 50-70%. The most used traits of study in genome selection experiments are related to quality improvement involving the use of different prediction models divided between parametric and non-parametric. Michel et al. (2018) proved the benefit of GS over marker assisted selection investigating the prediction of dough rheological traits in early generations and adopting the parametric RRBLUP, W-BLUP function. Genomic prediction models are routinely used in the CIMMYT spring bread wheat program since 2013 (Guzman et al., 2016). These models have also been successfully applied in genomic predictions for Fusarium head blight resistance in a DW panel (Moreno-Amores et al., 2020).

Landscape genomics

The selective pressure of abiotic stresses often varies in space, causing the evolution of advantageous mutations under their local environment, leading that genotype to have better fitness than the same genotype originating elsewhere. This differentiation process is called local adaptation and is driven by spatially divergent natural selection (Kawecki and Ebert, 2004). To trigger local adaptation, spatially divergent selection needs to overwhelm the homogenizing effect of gene flow. The study of the genetic bases of plant adaptation is crucial for the conservation and management of wild and cultivated species. Climatic stresses require a rapid evolution of populations to quickly adapt to new conditions and avoid extinction. Furthermore, to identify the genetic basis of adaptation, it is necessary to distinguish the under-selection (adaptive) genes from the pool of neutral genes. One possible approach is landscape genomics which aims to identify the gene-environment association, in particular loci associated with certain environmental variables. The landscape genomics analyses are providing unprecedented insight into the evolutionary processes and molecular basis that govern environmental adaptation. In the context of climate change, this type of analysis investigates how the species are adapting to the various types of stress they are subjected to but also could help to identify the wild relative introgression and its contribution to local adaptation (He et al., 2019). Landscape genomics integrates molecular analyses with climatic and geographic data in which samples have been collected to identify adaptive genes (Sork et al., 2013). This association analysis can be considered a valid alternative to GWAS when working with wild accessions or landraces, as they are naturally adapted to the place of origin. Moreover, the relationship between phenotypic variation and climatic factors, in DW, has been widely studied and confirmed (Annicchiarico et al., 1995; Royo et al., 2014). Recently, landscape genomics has been applied in many species such as *Populus tricocarpa*, *Beta vulgaris* spp. *maritima*, and *Arabidopsis tahaliana*. He et al. (2019) used the landscape genomics approach to find genomic windows associated with environmental adaptation in hexaploid wheat underlining the contribution to local adaptation given by wild emmer. In addition, 93 rice landraces from sub-Saharan regions were used to study adaptation to the local environment (Meyer et al., 2016). In landscape genomics, environmental information is screened for association with genetic variations through univariate or multivariate gene-environment association (GEA) analysis (Rellstab et al., 2015; Forester et al., 2018). Many statistical models have been developed for association analysis. For example, some methods involve a logistic association model such as the Spatial Analysis Method (SAM or SAMβADA), multiple logistic regression, and Generalized Estimating Equations (GEEs). Other methods involve a linear association model such as General linear models, redundancy analysis (RDA), bayenv, Spatial Generalized Linear Mixed Model (SGLMM), Latent Factor Mixed Models (LFMMs), and GWAS mixed models (Rellstab et al., 2015). However, to make the results more reliable, it would be a good practice to compare results coming from different association models.

There are 19 bioclimatic variables that can be screened, which can be downloaded from the WorldClim database, concerning the period 1970-2000; moreover, data reporting global soil salinity layers for the years 1986, 1992, 2000, 2002, 2005, 2009, and 2016 are also available. Using this association model, it is possible to detect candidate genes associated with salinity, thanks to the historical data available on the Global Salinity Soil Map website, as it was done in *Medicago truncatula* (Guerrero et al., 2018).

The relationship between genotype and environment could be also used to predict the spatial distribution of adaptive genetic variants in future climates and the future maladaptation or genomic offset that provide a direct estimate of the expected genomic vulnerability of the species toward ongoing climate change (Capblancq et al., 2020; Cavanagh et al., 2020; Capblancq and Forester, 2021).

Speed breeding

Recent advances in high-throughput phenotyping techniques have greatly increased the accuracy of breeding programs, having the advantage of being non-destructive and large-scale methods. However, the classic breeding programs, that have allowed the improvement of varieties, have the disadvantage of being extremely long and articulated and they take 10-15 years to release a variety. A technique that allows a rapid advancement of breeding generations is known as 'speed breeding'. This technique allows up to 6 generations of wheat per year and involves the use of fully enclosed, controlled-environment growth chambers with the addition of supplementary lighting (Watson et al., 2018). Several protocols for rapid and high-throughput phenotyping have been developed for the characterization of several important traits related to biotic and abiotic stresses in bread wheat. In durum wheat, a protocol has been developed providing multi-trait phenotyping and trying to accelerate even more the breeding cycles by using early filial generations (Alahmad et al., 2018). These 'speed breeding' techniques integrate perfectly with the new technologies of high-throughput genotyping and genomic selection.

Genetic diversity and signature of divergence in landrace germplasm

Until a few years ago, DW was well adapted to the Mediterranean environment. More recently, due to the climate crisis, drought, salinity, and low nutrient inputs occurring during flowering, pollination, and grain-filling represent the major stresses which adversely affect crop yield and quality, thus hampering agricultural productivity. Landraces coming from the Mediterranean basin are considered a particularly important group of genetic resources thanks to their high variability and tolerance to drought, pests, and adaptability to low farming systems (Lopes et al., 2015). Nowadays, the recovery, conservation, and enhancement of landraces are becoming central to increase the resilience of agricultural systems. However, how to exploit the genetic diversity of landraces to deal with environmental stress resilience is unclear and scattered (Lopes et al., 2015).

With the advance in genomic sequencing technologies and the release of the DW genome (Maccaferri et al., 2019), there has been a growing interest in comparing the patterns of genetic variation observed in landraces and modern varieties. These analyses were often focused on panels of landraces with a specific geographical origin. Population structure analysis was conducted for example in Iranian, Ethiopian, Tunisian, Turkish, and Italian germplasm revealing that, in most cases, landraces clustered separately from modern cultivars (Fayaz et al., 2019; Alemu et al., 2020; Alsaleh et al., 2022; Miazzi et al., 2022). Interestingly, a high level of genetic variation within landrace populations was detected, according to their geographical and climate of origin, revealing the importance of these factors in shaping wheat genome (Alipour et al., 2017; Taranto et al., 2022). On the other hand, cases of synonyms or homonyms as well as the presence of higher admixture of accessions between different populations of landraces were discovered, probably due to the exchange of seeds associated with human migration over time.

Examining wider collections, including landraces from different geographic origins, opens the possibility of investigating relationships on a wider global level and provides also a more precise estimation of the genetic diversity within each group. Genotyping the already mentioned core GPD by means of the iSelect 90K SNPChip followed by structure analysis showed comparatively limited genetic diversity in modern cultivar and a closer relationship to specific landrace population (North Africa and Transcaucasia). Landraces from Ethiopia appeared instead as the more isolated and distant to modern cultivars, while a high admixture level within landrace populations was confirmed (Maccaferri et al., 2019).

In addition, genome-wide population structure uncovers divergent selection during modern wheat breeding, suggesting the existence of untapped gene pools which will provide a basis for DW improvement in the next future. Many hotspots of selection were detected in the genomic regions where there are located the genes for adaptation, quality, grain yield, and stress response (Taranto et al., 2020; Soriano et al., 2021). These hotspots included important loci such as the photoperiod (*Ppd*), the vernalization (*Vrn*), and the dwarfing (*Rht*) genes, as well as loci associated with nitrogen use efficiency, plant architecture and grain yield (*TaASN3, asparagine synthetase 3*; *NR1, nitrate reductase 1*; *Fd-GOGAT, ferredoxin-dependent glutamate synthase*; *GS, glutamine synthetase*; *Sus2, sucrose synthase 2* and *TEF*, *transcript elongation factor*). In addition, genes related to quality such as pasta-making quality and color of semolina and other durum wheat-end products were also divergent between landraces and modern cultivars. In detail, loci for gluten composition (HMW/LMW, high/low molecular weight, and α, β, γ, and ώ gliadins), as well as loci involved in the carotenoid pathway (*Psy*) and polyphenol oxidase reaction (*Ppo*) were identified in hotspot regions (Requena-Ramírez et al., 2022; Taranto et al., 2022). Other divergent loci with implications in disease resistance, plant-microbe interactions, abiotic stresses, and plant development corresponded to gene models involved in important biological functions (Soriano et al., 2021). However, the identification of these genes and their allelic variants in the germplasm of indigenous DW varieties has been mainly carried out by *in-silico* analysis, at least for now. Future studies will be needed to validate the potential of the new allelic variants discovered in landraces.

Environmental adaptation traits from durum wheat landraces

In DW, domestication and, more lately, selection and fixation of favorable alleles had led to genetic erosion, lowering the buffering capacity of modern elite cultivars towards varied climatic conditions and strongly reducing potential for improvement (Tanksley and McCouch, 1997; Lopes et al., 2015). The systematic search and discovery of genetic resources from landraces by means of genotypic and innovative phenotypic profiling of genetic resource collections and the introduction in elite crops through pre-breeding efforts are being currently implemented in bread wheat and could be a promising strategy also for DW (Reynolds et al., 2021; Sharma et al.,

2021; Schulthess et al., 2022). Improvement of yield potential of landraces, by considering them as recipients, can be undertaken as an alternative strategy for developing better cultivars adapted to climate change. Inherent agronomic inferiority and disease susceptibility would hinder the direct utilization of landraces for breeding programs. However, the genomic tools and approaches we have described (i.e. genomic prediction and selection) could also strengthen pre-breeding efforts aiming at the improvement of genetic backgrounds of landraces, thereby attempting to achieve agronomic superiority starting directly from landraces as recipients and making this second alternative approach a possible and viable option (Adhikari et al., 2022). Finally, more complex breeding approaches, to develop new "synthetic" wheat crops exploiting genetic resources from wild species instead of landraces, also exist and have been already undertaken in the past (Reynolds et al., 2007; Balla et al., 2022).

Desirable genes were already identified in landraces and exploited for DW cultivars improvement by classical breeding. The most renowned DW cultivar is Cappelli, which is assumed to have been selected from the North-African landraces (De Cillis, 1942). However, the cv. Cappelli has a height of about 1.80 meters; therefore, several breeding activities were focused to create new variability by crossing the cv. Cappelli with Syriacum landraces (Aziziah, Eiti, Sinai, Tripolino). The result was the introduction of cultivars such as Capeiti 8 and Patrizio 6 which had slightly lower height, higher yield, earliness, and lodging resistance compared to Capelli, while preserving grain quality. Other similar examples of superior DW varieties, obtained by introgressing traits from landraces, were released in the frame of other breeding programs conducted in different countries (Kabbaj et al., 2017; Martínez-Moreno et al., 2020).

Molecular mapping technologies such as bulk segregant analysis (BSA), gene/QTL mapping, and genome-wide association studies (GWAS), supported by the high-throughput genotyping tools and strategies previously described, importantly increased the rate discovery of genes/QTLs regulating biotic, abiotic stress resistance, agronomic and quality traits from landraces. Several studies have been already undertaken so far to identify new genes/traits in landraces useful for breeding (see Table 3 for a comprehensive but not exhaustive list). As can be clearly appreciated, in most recent studies traits and genes have been often also mapped by taking advantage of the newly released genomic tools for DW here described, to further support their prompt exploitation in breeding.

Besides introducing new abiotic stresses, climate changes are shaping the dynamics of plants and pathogens resulting in more complex biological interactions difficult to predict and characterized by new outbreaks. Therefore, landraces are being widely explored as a potential source of new resistance traits, in regions where plant and pathogens co-evolved. Resistance sources to *Fusarium* head blight, rust, common bunt, stem sawfly, tan spot, and *Septoria tritici* blotch disease have been discovered within different DW landrace collections and majority of these have also been successfully mapped in the latest years. As pivotal examples leaf and stem rust resistance sources were mapped in the Portuguese DW landraces PI 192051 and Aus26582, by developing RIL mapping populations, beside alternative contribution from other sources (Qureshi et al., 2017; Qureshi et al., 2018; Aoun et al., 2019). Similarly, a resistance gene to *Zymoseptoria tritici* was mapped in the Tunisian DW landrace 'Agili 39' (Ferjaoui et al., 2022).

Concerning abiotic stress resistance, it is well recognized that Mediterranean DW landraces represent a particularly important group because of their documented better adaptation to drought (Moragues et al., 2006; Royo et al., 2014). Therefore, landraces from this region potentially include adaptive traits that could be exploited to boost the breeding for heat/drought tolerance and promote cultivars adaptation to stress-prone environments. Up to now, drought, heat, and salinity resistance traits have been studied in landraces coming from the Mediterranean basin, such as in Jordanian, Israeli-Palestinian, Tunisian, Italian, and Spanish, but also Afghan landraces (Al Khateeb et al., 2017; Shamaya et al., 2017; Hamdi et al., 2020; Frankin et al., 2021; Naranjo et al., 2022; Taranto et al., 2022). However, works concerning the mapping analysis of resistance traits to abiotic stresses are less in number than the biotic stresses ones. Just in two cases, abiotic stress resistance sources have been genetically mapped. A salinity resistance trait has been identified in an Afghan DW landrace and mapped. Moreover, using a GWAS approach on a worldwide collection of DW landraces, drought stress tolerance was associated to a locus of DW genome found to be collinear with a previously identified QTL in bread wheat (Shamaya et al., 2017; Wang et al., 2019). In order to improve the genetic characterization of abiotic resistance traits from landraces, an important contribution is expected from the development of adequate protocols for the abiotic stress evaluation, following similar strategies to those applied in bread wheat for high-throughput and accurate stress response phenotyping in large collections (Rasheed and Xia, 2019; Langridge et al., 2021; Shan et al., 2022).

Landraces are typically low yielding and can show lower agronomic attributes. Therefore, several studies enquired agronomic traits variability in landraces, focusing mainly on yield, phenology and morphological traits (Table 3). GWAS studies, based on high-throughput genotyping tools, helped in defining genomic regions affecting such agronomic traits in landraces highlighting available superior alleles. Among others, the contribution to the yield of root system architecture traits and phenology were highlighted (Mengistu et al., 2016; Kidane et al., 2017a; Desiderio et al., 2019; Gupta et al., 2020; Alemu et al., 2021; Ávila et al., 2021; Royo et al., 2021).

Quality traits of durum wheat landraces

DW semolina is considered the ideal raw material for the production of pasta or macaroni products, especially in Italy, which is the first producer and consumer of DW in Europe (http://www.internationalpasta.org, accessed on 12 November 2022). The aptitude of the raw material to be transformed into a high quality end-product mainly depends on grain protein content (GPC) and composition that directly affect wheat's market price and end-use value (Shewry, 2019).

Grain protein content, mainly above 12-13%, is highly related to the amount and composition of glutenins and gliadins proteins, that are the principal components of gluten and are responsible for the viscoelastic properties and extensibility of semolina, respectively. Past breeding activities aimed at improving grain yield resulted in a loss of

10.3389/fpls.2023.1101271

TABLE 3 Publication list regarding resistance/tolerance traits to biotic and abiotic stresses, morpho-agronomic and quality traits identified in durum wheat landraces.

Category	Gene/QTL Trait	Landrace (Origin)	Analysis Type	References
Biotic stress	Fusarium head blight resistance	Tunisia	GWAS	Ghavami et al., 2011
	Fusarium head blight resistance	Syria	Trait phenotyping	Talas et al., 2011
	Leaf and stem rust resistance	diverse	Gene/QTL mapping	Aoun et al., 2017
	Leaf and stem rust resistance	Portugal	Gene/QTL mapping	Aoun et al., 2019
	Leaf and stem rust resistance	Kazakhstan	GWAS	Genievskaya et al., 2022
	Leaf rust resistance	diverse	GWAS	Aoun et al., 2016
	Leaf rust resistance	Portugal	Gene/QTL mapping	Qureshi et al., 2017
	Leaf rust resistance	Portugal	Gene/QTL mapping	Qureshi et al., 2018
	Leaf rust resistance	Middle Est	Gene/QTL mapping	Kthiri et al., 2019
	Resistance to common bunt	Syria	Trait phenotyping	Mamluk and Nachit, 1994
	Septoria tritici blotch disease resistance	Tunisia	Gene/QTL mapping	Medini et al., 2014
	Septoria tritici blotch disease resistance	Tunisia	Trait phenotyping	Ferjaoui et al., 2015
	Septoria tritici blotch disease resistance	Ethiopia	GWAS	Kidane et al., 2017b
	Septoria tritici blotch disease resistance	Tunisia	Gene/QTL mapping	Aouini, 2018
	Septoria tritici blotch disease resistance	Tunisia	Trait phenotyping	Ouaja et al., 2020
	Septoria tritici blotch disease resistance	diverse	Trait phenotyping	Ben M'Barek et al., 2022
	Septoria tritici blotch disease resistance	Tunisia	Gene/QTL mapping	Ferjaoui et al., 2022
	Stem rust resistance	Italy	GWAS	Laidò et al., 2015
	Stem rust resistance	diverse	GWAS	Chao et al., 2017
	Stem rust resistance	Ethiopia	GWAS	Liu et al., 2017
	Stem rust resistance	Italy	GWAS	Saccomanno et al., 2018
	Stem rust resistance	diverse	Trait phenotyping	Olivera et al., 2021
	Stem rust resistance	Ethiopia	Trait phenotyping	Chiko et al., 2022
	Stem rust resistance	Iran	GWAS	Mehrabi et al., 2022
	Stem sawfy resistance	diverse	Gene/QTL mapping	Varella et al., 2019
	Tan spot resistance	diverse	Trait phenotyping	Laribi et al., 2021
	Yellow/stripe leaf and stem rust resistance	diverse	Association analysis/single marker scan	Bansal et al., 2013
	Yellow/stripe rust and common bunt resistance	diverse	Trait phenotyping	Annicchiarico et al., 1995
	Yellow/stripe rust resistance	Ethiopia	GWAS	Alemu et al., 2021
Abiotic stress	Allelophaty	Italy	Trait phenotyping	Scavo et al., 2022
	Cold	Iran	Trait phenotyping	Mohammadi et al., 2014
	Drought	diverse	Trait phenotyping	Pecetti et al., 1994
	Drought	diverse	Trait phenotyping	Annicchiarico et al., 1995
	Drought	Jordania	Trait phenotyping	Al Khateeb et al., 2017

(Continued)

10.3389/fpls.2023.1101271

TABLE 3 Continued

Category	Gene/QTL Trait	Landrace (Origin)	Analysis Type	References
	Drought	Israeli Palestina	Trait phenotyping	Frankin et al., 2021
	Drought	diverse	GWAS	Wang et al., 2019
	Heat	diverse	Trait phenotyping	Sareen et al., 2020
	Heat	Spain	Trait phenotyping	Naranjo et al., 2022
	Heat	Italy	Trait phenotyping	Taranto et al., 2022
	Salinity	Afganistan	Trait phenotyping	Shavrukov et al., 2011
	Salinity	Afganistan	QTL Mapping	Shamaya et al., 2017
	Salinity	Italy	Trait phenotyping	Maucieri et al., 2018
	Salinity	Tunisia	Gene functional characterization	Hamdi et al., 2020
	Salinity	Jordania	Trait phenotyping	Al Khateeb et al., 2020
Agronomic traits	Agromorphological traits	Marocco	Trait phenotyping	Taghouti et al., 2013
	Agromorphological traits, phenology	Italy	Trait phenotyping	Fiore et al., 2019
	Agromorphological traits (phenology, yield and morphology)	Spain	Trait phenotyping	Ruiz et al., 2012
	Agronomic (plant height, yield traits and phenology) and physiology trait	diverse	GWAS	Royo et al., 2021
	Agronomic trait (phenology, biomass and yield plant height)	Ethiopia	GWAS	Mengistu et al., 2016
	Agronomic trait (phenology, biomass and yield plant height)	diverse	Trait phenotyping	Royo et al., 2014
	Flowering time	diverse	GWAS	Gupta et al., 2020
	Flowering time, yield	diverse	GWAS	Royo et al., 2020
	Heading date, seed weight, and morphology	Iran	Gene/QTL mapping	Desiderio et al., 2019
	Morphology, phenology, yield component, GXE interaction	Ethiopia	Trait phenotyping	Mulugeta et al., 2022
	Morphology and yield	diverse	GWAS	Wang et al., 2019
	Morphology and yield, descriptors pigmentation, phenology	Oman	Trait phenotyping	Al Lawati et al., 2021
	Phenology, plant height, yield, and yield components	Ethiopia	GWAS	Kidane et al., 2017a
	Root system architecture traits	Ethiopia	GWAS	Alemu et al., 2021
	Spike height and shape	Marocco	Trait phenotyping	Sahri et al., 2014
	Stem cross section height and heading date	Spain	GWAS	Ávila et al., 2021
	Yield component	India	GWAS	Sukumaran et al., 2018
	Yield component	diverse	Trait phenotyping	Pecetti et al., 1994
	Yield component (kernel and spikes) and heading date	Italy	Trait phenotyping	Marzario et al., 2018
	Yield component, plant height, phenology and biomass	diverse	GWAS	Soriano et al., 2018
	Yield phenology lodging resistance	Israeli Palestina	Trait phenotyping	Frankin et al., 2021
	Yield vigour, plant height, phenology	diverse	Trait phenotyping	Annicchiarico et al., 1995
	Yield vigour, plant height, phenology	Algeria	Trait phenotyping	Annicchiarico et al., 2009
Quality	Arabinoxylan iron zinc phytate and phenolic acids content	Iran Mexico	Trait phenotyping	Hernandez-Espinosa et al., 2020
	Carotenoid content	Spain	Trait phenotyping	Requena-Ramírez et al., 2021

(Continued)

10.3389/fpls.2023.1101271

TABLE 3 Continued

Category	Gene/QTL Trait	Landrace (Origin)	Analysis Type	References
	Carotenoid content	Spain	GWAS	Requena-Ramírez et al., 2022
	Carotenoid content, color characteristics, chemical composition and starch digestibility	Italy	Trait phenotyping	Melini et al., 2021
	Gliadins content	Bulgaria	Trait phenotyping	Melnikova et al., 2010
	Gluten strength	Spain, CIMMYT, Italy, France and US	Trait phenotyping	Nazco et al., 2014b
	Glutenin protein composition	diverse	Trait phenotyping	Moragues et al., 2006
	Low molecular weight glutenin	Spain	Allelic variation	Ruiz et al., 2018
	High molecular weight glutenin	Italy	Mass spectrometry	Visioli et al., 2021
	Grain morphology and color	diverse	GWAS	Chou et al., 2022
	Grain quality	Marocco	Trait phenotyping	Taghouti et al., 2013
	Grain quality	Mexico	Trait phenotyping	Hernández-Espinosa et al., 2018
	Grain quality, yield, protein content, gluten strength and yellow color index	diverse	Trait phenotyping	Nazco et al., 2012
	Malting brewing related traits	Italy	Trait phenotyping	Alfeo et al., 2018
	Morpho-physiological characters	diverse	Trait phenotyping	Annicchiarico et al., 1995
	Phenolic and flavonoid content	Italy	Trait phenotyping	Dinelli et al., 2009
	Phenolic content	Italy	Trait phenotyping	Lo Bianco et al., 2017
	Physico-chemical traits, malt related traits, sugars	Italy	Trait phenotyping	Alfeo et al., 2021
	Phytochemical, antioxidant capacity and phenolic acids	Italy	Trait phenotyping	Di Loreto et al., 2018
	Polyphenolic content and antioxidants	Tunisia	Trait phenotyping	Boukid et al., 2019
	Prolamins	Spain	Trait phenotyping	Chacón et al., 2020
	Protein content, dry gluten gluten index, yellow index, ash P/L W G baking aptitude	Italy	Trait phenotyping	Ruisi et al., 2021
	Proteomic profiling (Metabolic and CM-protein fraction)	Italy	Trait phenotyping	Di Francesco et al., 2020
	Quality and rheological traits	diverse	Trait phenotyping	Ladhari et al., 2022
	Quality traits	Ethiopia	Trait phenotyping	Dagnaw et al., 2022
	Quality traits	Spain	Meta-QTL analysis	Roselló et al., 2018
	Quality traits and nitrogen use efficency	Tunisia	Trait phenotyping	Ayadi et al., 2022
	Quality traits	Italy	Trait phenotyping	Fiore et al., 2019
	Rheological parameters	Italy	Trait phenotyping	Spina et al., 2021
	Volatile organic compounds proteins	Italy	Trait phenotyping	Vita et al., 2016

genetic variability for quality-related traits, probably because of the negative relationship between yield and GPC (Nazco et al., 2014a; Subira et al., 2014). As proof of this, Roncallo et al. (2021) observed a decreasing trend in GPC over the last 85 years using a DW collection including accessions representative of the Argentina, Italy, Chile, France, CIMMYT and other countries breeding programs.

Previous studies suggested the potential quality-enhancing landraces as reservoir of new allelic variants for gluten quality improvement (Table 3) (Moragues et al., 2006; Nazco et al., 2014b; Roselló et al., 2018; Ruiz et al., 2018; Hernández-Espinosa et al., 2018). Roselló et al. (2018) performed a pasta-making quality QTLome using a Mediterranean collection of DW landraces and observed how landraces had higher GPC than modern cultivars but lower gluten strength. This result is due to very few allelic combinations of glutenin subunit loci in modern cultivars (Nazco et al., 2014b), while landraces showed a higher genetic variability useful to recovering and broadening allelic variation of gluten composition.

Other parameters can affect pasta production such as color (Table 3). Semolina and pasta color are constituted by yellow (desirable) and brown (undesirable) pigments (Colasuonno et al., 2019). Usually, DW landraces showed lower total carotenoid contents and higher values of browning compounds compared to commercial cultivars (Digesù et al., 2009; Subira et al., 2014; Taranto et al., 2015). However, the first DW landraces with carotenoid esterification ability were identified by Requena-Ramírez et al. (2021) and could represent donor sources in DW biofortification programs.

Although DW is mostly used for pasta production, it is an ingredient in typical breads in some areas of Southern Italy. It is the case of "Pane nero di Castelvetrano" and "Pane di Monreale" which are two traditional breads constituted by two Sicilian landraces, Timilia and Russello (Palumbo et al., 2008; Melini et al., 2021;Visioli et al., 2021). The most notable characteristic of Timilia is the dark color of semolina, due to the high content of antioxidant phenolic compounds (Giancaspro et al., 2016; Bianco et al., 2017; Taranto et al., 2022). To preserve these landraces and derived-products, a traceability approach was developed using the high molecular weight glutenins, suggesting a method to verify the varietal identity from the seed to the final product (Visioli et al., 2021).

Conclusions

In conclusion, recent findings unveiled the strategic role of landraces in the genetic improvement of durum wheat. Studies on genomic divergence among *T. turgidum* sub-species indicated that the allelic variations of domesticated accessions and their wild relatives, lost during the domestication and breeding processes, were and will be recovered by exploring and exploiting landraces genetic diversity. In particular, in a context of climate changes, understanding the environmental and genetic factors behind the adaptation of landraces can help to introduce beneficial alleles in elite varieties to overcome stress and increase yield. The availability of durum wheat reference genome and the increasingly precise molecular techniques at affordable costs are giving a big boost to accurately identify the genetic determinants underpinning resistance/tolerance against biotic and abiotic stresses.

The recent application of genomic technologies (i.e. genome-wide association and genomic prediction analysis) on durum wheat landrace resources paves the way to accelerate the next-generation breeding programs to overcome the gap of knowledge of these underexplored resources and identify advantageous alleles that have been lost in modern varieties.

Author contributions

CB: writing original draft. DB and FT: conceptualization and writing. FT, GD, AF: writing, review, and editing. All authors contributed to the article and approved the submitted version.

Funding

The authors would like to thank the PON "Ricerca e Innovazione" 2014-2020, Asse IV - Azione IV.6 for its support of this research.

Acknowledgments

This study was carried out within the Agritech National Research Center and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.4 – D.D. 1032 17/06/2022, CN00000022). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

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.

Author disclaimer

This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

References

Abu-Zaitoun, S. Y., Chandrasekhar, K., Assili, S., Shtaya, M. J., Jamous, R. M., Mallah, O. B., et al. (2018). Unlocking the genetic diversity within a middle-East panel of durum wheat landraces for adaptation to semi-arid climate 233. doi: 10.3390/agronomy8100233

Adhikari, S., Kumari, J., Jacob, S. R., Prasad, P., Gangwar, O. P., Lata, C., et al. (2022). Landraces-potential treasure for sustainable wheat improvement. *Genet. Resour. Crop Evol.* 69, 499–523. doi: 10.1007/s10722-021-01310-5

Alahmad, S., Dinglasan, E., Leung, K. M., Riaz, A., Derbal, N., Voss-Fels, K. P., et al. (2018). Speed breeding for multiple quantitative traits in durum wheat. *Plant Methods* 14, 1–15. doi: 10.1186/s13007-018-0302-y

Alemu, Y. A., Anley, A. M., and Abebe, T. D. (2020). Genetic variability and association of traits in Ethiopian durum wheat (*Triticum turgidium* l. var. *durum*) landraces at dabat research station, north gondar. *Cogent. Food Agric.* 6, 1778604. doi: 10.1080/23311932.2020.1778604

Alemu, S. K., Huluka, A. B., Tesfaye, K., Haileselassie, T., and Uauy, C. (2021). Genome-wide association mapping identifies yellow rust resistance loci in Ethiopian durum wheat germplasm. *PloS One* 16 (5), e0243675. doi: 10.1371/journal.pone.0243675

Alfeo, V., De Francesco, G., Sileoni, V., Blangiforti, S., Palmeri, R., Aerts, G., et al. (2021). Physicochemical properties, sugar profile, and non-starch polysaccharides characterization of old wheat malt landraces. *J. Food Compos. Anal.* 102, 103997. doi: 10.1016/j.jfca.2021.103997

Alfeo, V., Jaskula-Goiris, B., Venora, G., Schimmenti, E., Aerts, G., and Todaro, A. (2018). Screening of durum wheat landraces (*Triticum turgidum* subsp. *durum*) for the malting suitability. *J. Cereal Sci.* 83, 10. doi: 10.1016/j.jcs.2018.08.001

Alipour, H., Bihamta, M. R., Mohammadi, V., Peyghambari, S. A., Bai, G., and Zhang, G. (2017). Genotyping-by-sequencing (GBS) revealed molecular genetic diversity of Iranian wheat landraces and cultivars. *Front. Plant Sci.* 8, 1293. doi: 10.3389/fpls.2017.01293

Al Khateeb, W., Al Shalabi, A., Schroeder, D., and Musallam, I. (2017). Phenotypic and molecular variation in drought tolerance of Jordanian durum wheat (*Triticum durum* desf.) landraces. *Physiol. Mol. Biol.* 23, 311–319. doi: 10.1007/s12298-017-0434-y

Al Lawati, A. H., Nadaf, S. K., AlSaady, N. A., Al Hinai, S. A., Almamari, A., Al Adawi, M. H., et al. (2021). Genetic diversity of omani durum wheat (sub sp.) landraces. *Open Agric.*. 15, 21–32. doi: 10.2174/1874331502115010021

Allen, A. M., Winfield, M. O., Burridge, A. J., Downie, R. C., Benbow, H. R., Barker, G. L., et al. (2017). Characterization of a wheat breeders' array suitable for high-throughput SNP genotyping of global accessions of hexaploid bread wheat (*Triticum aestivum*). *Plant Biotechnol. J.* 15 (3), 390–401. doi: 10.1111/pbi.12635

Alsaleh, A., Bektas, H., Baloch, F. S., Nadeem, M. A., and Özkan, H. (2022). Turkish Durum wheat conserved ex-situ and *in situ* unveils a new hotspot of unexplored genetic diversity. *Crop Sci.* 62, 1200–1212. doi: 10.1002/csc2.20723

Altamore, L., Ingrassia, M., Columba, P., Chironi, S., and Bacarella, S. (2020). Italian Consumers' preferences for pasta and consumption trends: Tradition or innovation? *J. Int. Food Agribus.* 32, 337–360. doi: 10.1080/08974438.2019.1650865

Annicchiarico, P., Abdellaoui, Z., Kelkouli, M., and Zerargui, H. (2005). Grain yield, straw yield and economic value of tall and semi-dwarf durum wheat cultivars in Algeria. *J. Agric. Sci.* 143, 57–64. doi: 10.1017/S0021859605004855

Annicchiarico, P., Pecetti, L., and Damania, A. B. (1995). Relationships between phenotypic variation and climatic factors at collecting sites in durum wheat landraces. *Hereditas* 122, 163–167. doi: 10.1111/j.1601-5223.1995.00163.x

Annicchiarico, P., Royo, C., Bellah, F., and Moragues, M. (2009). Relationships among adaptation patterns, morphophysiological traits and molecular markers in durum wheat. *Plant Breed.* 128, 164–171. doi: 10.1111/j.1439-0523.2008.01557.x

Annunziata, M. G., Ciarmiello, L. F., Woodrow, P., Maximova, E., Fuggi, A., and Carillo, P. (2017). Durum wheat roots adapt to salinity remodeling the cellular content of nitrogen metabolites and sucrose. *Front. Plant Sci.* 7, 2035. doi: 10.3389/fpls.2016.02035

Aouini, L. (2018). Durum wheat and septoria tritici blotch: genes and prospects for breeding (Order No. 28232876) Available from ProQuest Dissertations & Theses Global. (2564078719). Retrieved from https://www.proquest.com/dissertations-theses/durum-wheat-septoria-tritici-blotch-genes/docview/2564078719/se-2.

Aoun, M., Breiland, M., Kathryn Turner, M., Loladze, A., Chao, S., Xu, S. S., et al. (2016). Genome-wide association mapping of leaf rust response in a durum wheat worldwide germplasm collection. *Plant Genome* 9 (3). doi: 10.3835/plantgenome2016.01.0008

Aoun, M., Kolmer, J. A., Rouse, M. N., Chao, S., Bulbula, W. D., Elias, E. M., et al. (2017). Inheritance and bulked segregant analysis of leaf rust and stem rust resistance in durum wheat genotypes. *Phytopathology* 107, 1496–1506. doi: 10.1094/PHYTO-12-16-0444-R

Aoun, M., Kolmer, J. A., Rouse, M. N., Elias, E. M., Breiland, M., Bulbula, W. D., et al. (2019). Mapping of novel leaf rust and stem rust resistance genes in the Portuguese durum wheat landrace PI 192051. *G3: Genes Genomes Genet.* 9, 2535–2547. doi: 10.1534/g3.119.400292

Ávila, C. M., Requena-Ramírez, M. D., Rodríguez-Suárez, C., Flores, F., Sillero, J. C., and Atienza, S. G. (2021). Genome-wide association analysis for stem cross section properties, height and heading date in a collection of spanish durum wheat landraces. *Plants* 10, 1123. doi: 10.3390/plants10061123

Ayadi, S., Jallouli, S., Chamekh, Z., Zouari, I., Landi, S., Hammami, Z., et al. (2022). Variation of grain yield, grain protein content and nitrogen use efficiency components under different nitrogen rates in mediterranean durum wheat genotypes. *Agriculture* 12, 916. doi: 10.3390/agriculture12070916

Ayalew, H., Tsang, P. W., Chu, C., Wang, J., Liu, S., Chen, C., et al. (2019). Comparison of TaqMan, KASP and rhAmp SNP genotyping platforms in hexaploid wheat. *PloS One* 14 (5), e0217222. doi: 10.1371/journal.pone.0217222

Azeez, A. M., Adubi, O. A., and Durodola, F. A. (2018). "Landraces and crop genetic improvement," in *Rediscovery of landraces as a resource for the future*, IntechOpen, London, UK, 1–19.

Balla, M. Y., Gorafi, Y. S. A., Kamal, N. M., Abdalla, M. G. A., Tahir, I. S. A., and Tsujimoto, H. (2022). Harnessing the diversity of wild emmer wheat for genetic improvement of durum wheat. *Theor. Appl. Genet.* 135, 1671–1684. doi: 10.1007/s00122-022-04062-7

Bansal, U. K., Arief, V. N., DeLacy, I. H., and Bariana, H. S. (2013). Exploring wheat landraces for rust resistance using a single marker scan. *Euphytica* 194, 219–233. doi: 10.1007/s10681-013-0940-0

Bayer, P. E., Golicz, A. A., Scheben, A., Batley, J., and Edwards, D. (2020). Plant pan-genomes are the new reference. *Nat. Plants* 6, 914–920. doi: 10.1038/s41477-020-0733-0

Bayer, P. E., Petereit, J., Durant, E., Monat, C., Rouard, M., Hu, H., et al. (2022). Bread wheat genomes graph pangenome. *Zenodo*. doi: 10.1101/2022.02.23.481560

Ben M'Barek, S., Laribi, M., Kouki, H., Castillo, D., Araar, C., Nefzaoui, M., et al. (2022). Phenotyping Mediterranean durum wheat landraces for resistance to *Zymoseptoria tritici* in Tunisia. *Genes* 13, 355. doi: 10.3390/genes13020355

Bianco, M. L., Siracusa, L., Dattilo, S., Venora, G., and Ruberto, G. (2017). Phenolic fingerprint of sicilian modern cultivars and durum wheat landraces: a tool to assess biodiversity. *Cereal Chem.* 94, 1045–1051. doi: 10.1094/CCHEM-06-17-0125-R

Borrill, P., Harrington, S. A., and Uauy, C. (2019). Applying the latest advances in genomics and phenomics for trait discovery in polyploid wheat. *Plant J.* 97, 56–72. doi: 10.1111/tpj.14150

Boukid, F., Dall'Asta, M., Bresciani, L., Mena, P., Del Rio, D., Calani, L., et al. (2019). Phenolic profile and antioxidant capacity of landraces, old and modern Tunisian durum wheat. *Eur. Food Res. Technol.* 245, 73–82. doi: 10.1007/s00217-018-3141-1

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

Broccanello, C., Chiodi, C., Funk, A., McGrath, J. M., Panella, L., and Stevanato, P. (2018). Comparison of three PCR-based assays for SNP genotyping in plants. *Plant Methods* 14, 1–8. doi: 10.1186/s13007-018-0295-6

Broccanello, C., Gerace, L., and Stevanato, P. (2020). "QuantStudio™ 12K flex OpenArray® system as a tool for high-throughput genotyping and gene expression analysis," in *Quantitative real-time PCR* (New York, NY: Humana), 199–208.

Camacho Villa, T. C., Maxted, N., Scholten, M., and Ford-Lloyd, B. (2005). Defining and identifying crop landraces. *Plant Genet. Resour. Charact. Util.* 3, 373–384. doi: 10.1079/PGR200591

Capblancq, T., and Forester, B. R. (2021). Redundancy analysis: A Swiss army knife for landscape genomics. *Methods Ecol. Evol.* 12, 2298–2309. doi: 10.1111/2041-210X.13722

Casañas, F., Simó, J., Casals, J., and Prohens, J. (2017). Toward an evolved concept of landrace. *Front. Plant Sci.* 8, 145. doi: 10.3389/fpls.2017.00145

Capblancq, T., Fitzpatrick, M. C., Bay, R. A., Exposito-Alonso, M., and Keller, S. R. (2020). Genomic prediction of (mal) adaptation across current and future climatic landscapes. *Annu. Rev. Ecology Evolution Systematics* 51 (1). doi: 10.1146/annurev-ecolsys-020720-042553

Cavanagh, C. R., Chao, S., Wang, S., Huang, B. E., Stephen, S., Kiani, S., et al. (2013). Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. *Proc. Natl. Acad. Sci. U.S.A.* 110, 8057–8062. doi: 10.1073/pnas.1217133110

Cavanagh, C. R., Chao, S., Wang, S., Huang, B. E., Stephen, S., Kiani, S., et al. (2020). Genomic prediction of (Mal)adaptation across current and future climatic landscapes. *Annu. Rev. Ecol. Evol.* 51, 245–271. doi: 10.1146/annurev-ecolsys-020720-042553

Chacón, E. A., Vázquez, F. J., Giraldo, P., Carrillo, J. M., Benavente, E., and Rodríguez-Quijano, M. (2020). Allelic variation for prolamins in Spanish durum wheat landraces and its relationship with quality traits. *Agronomy* 10, .136. doi: 10.3390/agronomy10010136

Chao, S., Rouse, M. N., Acevedo, M., Szabo-Hever, A., Bockelman, H., Bonman, J. M., et al. (2017). Evaluation of genetic diversity and host resistance to stem rust in USDA NSGC durum wheat accessions. *Plant Genome.* 10 (2). doi: 10.3835/plantgenome2016.07.0071

Chaparro-Encinas, L. A., Santoyo, G., Peña-Cabriales, J. J., Castro-Espinoza, L., Parra-Cota, F. I., and Santos-Villalobos, S. D. L. (2021). Transcriptional regulation of metabolic and cellular processes in durum wheat (*Triticum turgidum* subsp. *durum*) in the face of temperature increasing. *Plants* 10, 2792. doi: 10.3390/plants10122792

Chaurasia, S., Singh, A. K., Songachan, L. S., Sharma, A. D., Bhardwaj, R., and Singh, K. (2020). Multi-locus genome-wide association studies reveal novel genomic regions associated with vegetative stage salt tolerance in bread wheat (*Triticum aestivum* l.). *Genomics* 112, 4608–4621. doi: 10.1016/j.ygeno.2020.08.006

Chiko, S., Kebede Gessese, M., Shimelash, D., Haile, W. T., Melo, B. Y., Wassie, A. S., et al. (2022). Diversity of Ethiopian durum wheat landraces for resistance to stem rust seedling resistance renes. *Adv. Agric.* 2022. doi: 10.1155/2022/3023427

Chou, C. H., Lin, H. S., Wen, C. H., and Tung, C. W. (2022). Patterns of genetic variation and QTLs controlling grain traits in a collection of global wheat germplasm revealed by high-quality SNP markers. *BMC Plant Biol.* 22, 455. doi: 10.1186/s12870-022-03844-x

Chung, Y. S., Choi, S. C., Jun, T. H., and Kim, C. (2017). Genotyping-by-sequencing: a promising tool for plant genetics research and breeding. *Horticult. Environ. Biotechnol.* 58, 425–431. doi: 10.1007/s13580-017-0297-8

Colasuonno, P., Marcotuli, I., Blanco, A., Maccaferri, M., Condorelli, G. E., Tuberosa, R., et al. (2019). Carotenoid pigment content in durum wheat (*Triticum turgidum l.* var *durum*): An overview of quantitative trait loci and candidate genes. *Front. Plant Sci.* 10, 1347. doi: 10.3389/fpls.2019.01347

Colasuonno, P., Marcotuli, I., Gadaleta, A., and Soriano, J. M. (2021). From genetic maps to QTL cloning: an overview for durum wheat. *Plants* 10, 315. doi: 10.3390/plants10020315

Colmer, T. D., Munns, R., and Flowers, T. J. (2005). Improving salt tolerance of wheat and barley: future prospects. *Aust. J. Exp. Agric.* 45, 1425–1443. doi: 10.1071/EA04162

Crossa, J., Jarquín, D., Franco, J., Pérez-Rodríguez, P., Burgueño, J., Saint-Pierre, C., et al. (2016). Genomic prediction of gene bank wheat landraces. *G3: Genes Genomes Genet.* 6, 1819–1834. doi: 10.1534/g3.116.029637

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

Daetwyler, H. D., Pong-Wong, R., Villanueva, B., and Woolliams, J. A. (2010). The impact of genetic architecture on genome-wide evaluation methods. *Genetics* 185, 1021–1031. doi: 10.1534/genetics.110.116855

Dagnaw, T., Mulugeta, B., Haileselassie, T., Geleta, M., and Tesfaye, K. (2022). Phenotypic variability, heritability and associations of agronomic and quality traits in cultivated Ethiopian durum wheat (*Triticum turgidum* l. ssp. *durum*, desf.). *Agronomy* 12, 1714. doi: 10.3390/agronomy12071714

Davenport, R., James, R. A., Zakrisson-Plogander, A., Tester, M., and Munns, R. (2005). Control of sodium transport in durum wheat. *Plant Physiol.* 137, 807–818. doi: 10.1104/pp.104.057307

Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., and Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.* 12, 499–510. doi: 10.1038/nrg3012

De Cillis, U. (1942). I frumenti Siciliani. Stazione Sperimentale di Granicoltura "Benito Mussolini" per la Sicilia - Catania. Pubblicazione n. 9 p. 1–323.

Desiderio, F., Zarei, L., Licciardello, S., Cheghamirza, K., Farshadfar, E., Virzi, N., et al. (2019). Genomic regions from an Iranian landrace increase kernel size in durum wheat. *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.00448

De Santis, M. A., Soccio, M., Laus, M. N., and Flagella, Z. (2021). Influence of drought and salt stress on durum wheat grain quality and composition: A review. *Plants* 10, 2599. doi: 10.3390/plants10122599

Deschamps, S., Llaca, V., and May, G. D. (2012). Genotyping-by-sequencing in plants. *Biology* 1, 460–483. doi: 10.3390/biology1030460

De Vita, P., and Taranto, F. (2019). "Durum wheat (Triticum turgidum ssp. durum) breeding to meet the challenge of climate change," in *Advances in plant breeding strategies: cereals* (Switzerland: Springer, Cham).

Di Francesco, A., Saletti, R., Cunsolo, V., Svensson, B., Muccilli, V., De Vita, P., et al. (2020). Qualitative proteomic comparison of metabolic and CM-like protein fractions in old and modern wheat Italian genotypes by a shotgun approach. *J. Proteomics.* 211, 103530. doi: 10.1016/j.jprot.2019.103530

Digesù, A. M., Platani, C., Cattivelli, L., Mangini, G., and Blanco, A. (2009). Genetic variability in yellow pigment components in cultivated and wild tetraploid wheats. *J. Cereal Sci.* 50 (2), 210–218. doi: 10.1016/j.jcs.2009.05.002

Di Loreto, A., Bosi, S., Montero, L., Bregola, V., Marotti, I., Sferrazza, R. E., et al. (2018). Determination of phenolic compounds in ancient and modern durum wheat genotypes. *Electrophoresis* 39, 2001–2010. doi: 10.1002/elps.201700439

Dinelli, G., Carretero, A. S., Di Silvestro, R., Marotti, I., Fu, S., Benedettelli, S., et al. (2009). Determination of phenolic compounds in modern and old varieties of durum wheat using liquid chromatography coupled with time-of-flight mass spectrometry. *J. Chromatogr. A.* 1216, 7229–7240. doi: 10.1016/j.chroma.2009.08.041

Di Pede, G., Dodi, R., Scarpa, C., Brighenti, F., Dall'Asta, M., and Scazzina, F. (2021). Glycemic index values of pasta products: An overview. *Foods* 10, 2541. doi: 10.3390/foods10112541

Dvorak, P., Krasylenko, Y., Zeiner, A., Šamaj, J., and Takác, T. (2021). Signaling toward reactive oxygen species-scavenging enzymes in plants. *Front. Plant Sci.* 11, 618835. doi: 10.3389/fpls.2020.618835

Edelstein, M., Plaut, Z., and Ben-Hur, M. (2010). Water salinity and sodicity effects on soil structure and hydraulic properties. *Adv. Hortic. Sci.* 24, 154–160.

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., et al. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS One* 6, 19379. doi: 10.1371/journal.pone.0019379

Esposito, S., Taranto, F., Vitale, P., Ficco, D. B. M., Colecchia, S. A., Stevanato, P., et al. (2022). Unlocking the molecular basis of wheat straw composition and morphological traits through multi-locus GWAS. *BMC Plant Biol.* 22, 1–19. doi: 10.1186/s12870-022-03900-6

Etminan, A., Pour-Aboughadareh, A., Mohammadi, R., Ahmadi-Rad, A., Moradi, Z., Mahdavian, Z., et al. (2017). Evaluation of genetic diversity in a mini core collection of Iranian durum wheat germplasms. *J. Anim. Plant Sci.* 27, 1582–1587.

Fayaz, F., Aghaee Sarbarzeh, M., Talebi, R., and Azadi, A. (2019). Genetic diversity and molecular characterization of Iranian durum wheat landraces (*Triticum turgidum durum* (Desf.) husn.) using DArT markers. *Biochem. Genet.* 57, 98–116. doi: 10.1007/s10528-018-9877-2

Ferjaoui, S., Aouini, L., Slimane, R. B., Ammar, K., Dreisigacker, S., Schouten, H. J., et al. (2022). Deciphering resistance to *Zymoseptoria tritici* in the Tunisian durum wheat landrace accession 'Agili39'. *BMC Genom.* 23, 1–20. doi: 10.1186/s12864-022-08560-2

Ferjaoui, S., M'Barek, S. B., Bahri, B., Slimane, R. B., and Hamza, S. (2015). Identification of resistance sources to septoria tritici blotch in old Tunisian durum wheat germplasm applied for the analysis of the *Zymoseptoria tritici*-durum wheat interaction. *J. Plant Pathol.* 97, 1–11. doi: 10.4454/JPP.V97I3.028

Fernando, R. L., and Garrick, D. (2013). "Bayesian Methods applied to GWAS," in *Genome wide association studies and genomic prediction*, vol. 1019 . Eds. C. Gondro, J. van der Werf and B. Hayes (Totowa, NJ: Humana Press), 237–274.

Fiore, M. C., Mercati, F., Spina, A., Blangiforti, S., Venora, G., Dell'Acqua, M., et al. (2019). High-throughput genotype, morphology, and quality traits evaluation for the assessment of genetic diversity of wheat landraces from Sicily. *Plants* 8, 116. doi: 10.3390/plants8050116

Flagella, Z., Giuliani, M. M., Giuzio, L., Volpi, C., and Masci, S. (2010). Influence of water deficit on durum wheat storage protein composition and technological quality. *Eur. J. Agron.* 33, 197–207. doi: 10.1016/j.eja.2010.05.006

Forester, B. R., Lasky, J. R., Wagner, H. H., and Urban, D. L. (2018). Comparing methods for detecting multilocus adaptation with multivariate genotype–environment associations. *Mol. Ecol.* 27, 2215–2233. doi: 10.1111/mec.14584

Forrest, K., Pujol, V., Bulli, P., Pumphrey, M., Wellings, C., Herrera-Foessel, S., et al. (2014). Development of a SNP marker assay for the *Lr67* gene of wheat using a genotyping by sequencing approach. *Mol. Breed.* 34, 2109–2118. doi: 10.1007/s11032-014-0166-4

Frankin, S., Roychowdhury, R., Nashef, K., Abbo, S., Bonfil, D. J., and Ben-David, R. (2021). In-field comparative study of landraces vs. modern wheat genotypes under a mediterranean climate. *Plants* 10, 2612. doi: 10.3390/plants10122612

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

Genievskaya, Y., Pecchioni, N., Laidò, G., Anuarbek, S., Rsaliyev, A., Chudinov, V., et al. (2022). Genome-wide association study of leaf rust and stem rust seedling and adult resistances in tetraploid wheat accessions harvested in kazakhstan. *Plants* 11, 1904. doi: 10.3390/plants11151904

Ghavami, F., Elias, E. M., Mamidi, S., Ansari, O., Sargolzaei, M., Adhikari, T., et al. (2011). Mixed model association mapping for fusarium head blight resistance in Tunisian-derived durum wheat populations. *G3: Genes| Genomes| Genet.* 1, 209–218. doi: 10.1534/g3.111.000489

Giancaspro, A., Colasuonno, P., Zito, D., Blanco, A., Pasqualone, A., and Gadaleta, A. (2016). Varietal traceability of bread 'Pane Nero di castelvetrano' by denaturing high pressure liquid chromatography analysis of single nucleotide polymorphisms. *Food Control* 59, 809–817. doi: 10.1016/j.foodcont.2015.07.006

Golicz, A. A., Bayer, P. E., Barker, G. C., Edger, PP, Kim, H, Martinez, PA, et al. (2016). The pangenome of an agronomically important crop plant brassica oleracea. *Nat. Commun.* 7, 13390. doi: 10.1038/ncomms13390

Guerrero, J., Andrello, M., Burgarella, C., and Manel, S. (2018). Soil environment is a key driver of adaptation in *Medicago truncatula*: new insights from landscape genomics. *New Phytol.* 219, 378–390. doi: 10.1111/nph.15171

Gupta, P. K., Balyan, H. S., Sharma, S., and Kumar, R. (2020). Genetics of yield, abiotic stress tolerance and biofortification in wheat (*Triticum aestivum* l.). *Theor. Appl. Genet.* 133, 1569–1602. doi: 10.1007/s00122-020-03583-3

Guzman, C., Peña, R. J., Singh, R., Autrique, E., Dreisigacker, S., Crossa, J., et al. (2016). Wheat quality improvement at CIMMYT and the use of genomic selection on it. *Appl. Trans. Genomics* 11, 3–8. doi: 10.1016/j.atg.2016.10.004

Halder, J., Zhang, J., Ali, S., Sidhu, J. S., Gill, H. S., Talukder, S. K., et al. (2019). Mining and genomic characterization of resistance to tan spot, stagonospora nodorum blotch (SNB), and fusarium head blight in Watkins core collection of wheat landraces. *BMC Plant Biol.* 19, 480. doi: 10.1186/s12870-019-2093-3

Hamdi, K., Brini, F., Kharrat, N., Masmoudi, K., and Yakoubi, I. (2020). Abscisic acid, stress, and ripening (*Tt*ASR1) gene as a functional marker for salt tolerance in durum wheat. *BioMed. Res. Int.* 31, 7876357. doi: 10.1155/2020/7876357

Hanif, U., Alipour, H., Gul, A., Jing, L., Darvishzadeh, R., Amir, R., et al. (2021). Characterization of the genetic basis of local adaptation of wheat landraces from Iran and Pakistan using genome-wide association study. *TPG* 14, 20096. doi: 10.1002/tpg2.20096

Harrington, S. A., Cobo, N., Karafiatova, M., Dolezel, J., Borrill, P., and Uauy, C. (2019). Identification of a dominant chlorosis phenotype through a forward screen of the *Triticum turgidum* cv. kronos TILLING population. *Front. Plant Sci.* 10, 963. doi: 10.3389/fpls.2019.00963

He, F., Pasam, R., Shi, F., Kant, S., Keeble-Gagnere, G., Kay, P., et al. (2019). Exome sequencing highlights the role of wild-relative introgression in shaping the adaptive landscape of the wheat genome. *Nat. Genet.* 5, 896–904. doi: 10.1038/s41588-019-0382-2

Hernandez-Espinosa, N., Laddomada, B., Payne, T., Huerta-Espino, J., Govindan, V., Ammar, K., et al. (2020). Nutritional quality characterization of a set of durum wheat landraces from Iran and Mexico. *LWT* 124, 109198. doi: 10.1016/j.lwt.2020.109198

Hernández-Espinosa, N., Mondal, S., Autrique, E., Gonzalez-Santoyo, H., Crossa, J., Huerta-Espino, J., et al. (2018). Milling, processing and end-use quality traits of CIMMYT spring bread wheat germplasm under drought and heat stress. *Field Crops Res.* 215, 104–112. doi: 10.1016/j.fcr.2017.10.003

Hufford, M. B., Xu, X., Van Heerwaarden, J., Pyhäjärvi, T., Chia, J. M., Cartwright, R. A., et al. (2012). Comparative population genomics of maize domestication and improvement. *Nat. Genet.* 44 (7), 808–811. doi: 10.1038/ng.2309

Jarquín, D., Crossa, J., Lacaze, X., Du Cheyron, P., Daucourt, J., Lorgeou, J., et al. (2014). A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theor. Appl. Genet.* 127, 595–607. doi: 10.1007/s00122-013-2243-1

Jin, H., Wen, W., Liu, J., Zhai, S., Zhang, Y., Yan, J., et al. (2016). Genome-wide QTL mapping for wheat processing quality parameters in a gaocheng 8901/Zhoumai 16 recombinant inbred line population. *Front. Plant Sci.* 7, 1032. doi: 10.3389/fpls.2016.01032

Kabbaj, H., Sall, A. T., Al-Abdallat, A., Geleta, M., Amri, A., Filali-Maltouf, A., et al. (2017). Genetic diversity within a global panel of durum wheat (*Triticum durum*) landraces and modern germplasm reveals the history of allele exchange. *Front. Plant Sci.* 8, 1277. doi: 10.3389/fpls.2017.01277

Kaur, P., and Gaikwad, K. (2017). From genomes to GENE-omes: exome sequencing concept and applications in crop improvement. *Front. Plant Sci.* 8, 2164. doi: 10.3389/fpls.2017.02164

Kawecki, T. J., and Ebert, D. (2004). Conceptual issues in local adaptation. *Ecol. Lett.* 7, 1225–1241. doi: 10.1111/j.1461-0248.2004.00684.x

Khan, A. W., Garg, V., Roorkiwal, M., Golicz, A. A., Edwards, D., and Varshney, R. K. (2020). Super-pangenome by integrating the wild side of a species for accelerated crop improvement. *Trends Plant Sci.* 25 (2), 148–158. doi: 10.1016/j.tplants.2019.10.012

Kidane, Y. G., Mancini, C., Mengistu, D. K., Frascaroli, E., Fadda, C., Pè, M. E., et al. (2017a). Genome wide association study to identify the genetic base of smallholder farmer preferences of durum wheat traits. *Front. Plant Sci.*, 1230. doi: 10.3389/fpls.2017.01230

Kidane, Y. G., Hailemariam, B. N., Mengistu, D. K., Fadda, C., Pè, M. E., and Dell'Acqua, M. (2017b). Genome-wide association study of *Septoria tritici* blotch resistance in Ethiopian durum wheat landraces. *Front. Plant Sci.* 8, 1586. doi: 10.3389/fpls.2017.01586

Kthiri, D., Loladze, A., N'Diaye, A., Nilsen, K. T., Walkowiak, S., Dreisigacker, S., et al. (2019). Mapping of genetic loci conferring resistance to leaf rust from three globally resistant durum wheat sources. *Front. Plant Sci.* 10, 1247. doi: 10.3389/fpls.2019.01247

Kulkarni, M., Soolanayakanahally, R., Ogawa, S., Uga, Y., Selvaraj, M. G., and Kagale, S. (2017). Drought response in wheat: Key genes and regulatory mechanisms controlling root system architecture and transpiration efficiency. *Front. Chem.* 5, 106. doi: 10.3389/fchem.2017.00106

Ladhari, A., Corrado, G., Rouphael, Y., Carella, F., Nappo, G. R., Di Marino, C., et al. (2022). Chemical, functional, and technological features of grains, brans, and semolina from purple and red durum wheat landraces. *Foods* 11, 1545. doi: 10.3390/foods11111545

Laidò, G., Mangini, G., Taranto, F., Gadaleta, A., Blanco, A., Cattivelli, L., et al. (2013). Genetic diversity and population structure of tetraploid wheats (*Triticum turgidum* l.) estimated by SSR, DArT and pedigree data. *PloS One* 8 (6), e67280. doi: 10.1371/journal.pone.0067280

Laidò, G., Panio, G., Marone, D., Russo, M. A., Ficco, D. B., Giovanniello, V., et al. (2015). Identification of new resistance loci to African stem rust race TTKSK in tetraploid wheats based on linkage and genome-wide association mapping. *Front. Plant Sci.* 6. doi: 10.3389/fpls.2015.01033

Langridge, P., and Reynolds, M. (2021). Breeding for drought and heat tolerance in wheat. *Theor. Appl. Genet.* 134, 1753–1769. doi: 10.1007/s00122-021-03795-1

Laribi, M., Ben M'Barek, S., Fakhfakh, M., Yahyaoui, A. H., and Sassi, K. (2021). Durum wheat mediterranean landraces: a valuable source for resistance to tan spot disease. *Agriculture* 11, 1148. doi: 10.3390/agriculture11111148

Laus, M. N., De Santis, M. A., Flagella, Z., and Soccio, M. (2022). Changes in antioxidant defence system in durum wheat under hyperosmotic stress: A concise overview. *Plants* 11, 98. doi: 10.3390/plants11010098

Lemerle, D., Gill, G. S., Murphy, C. E., Walker, S. R., Cousens, R. D., Mokhtari, S., et al. (2001). Genetic improvement and agronomy for enhanced wheat competitiveness with weeds. *Aust. J. Agric. Res.* 52, 527–548. doi: 10.1071/AR00056

Lindsay, M. P., Lagudah, E. S., Hare, R. A., and Munns, R. (2004). A locus for sodium exclusion (*Nax1*), a trait for salt tolerance, mapped in durum wheat. *Funct. Plant Biol.* 31, 1105–1114. doi: 10.1071/FP04111

Lipka, A. E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P. J., et al. (2012). GAPIT: genome association and prediction integrated tool. *Bioinformatics* 28, 2397–2399. doi: 10.1093/bioinformatics/bts444

Liu, W., Maccaferri, M., Rynearson, S., Letta, T., Zegeye, H., Tuberosa, R., et al. (2017). Novel sources of stripe rust resistance identified by genome-wide association mapping in Ethiopian durum wheat (*Triticum turgidum* ssp. *durum*). *Front. Plant Sci.* 8, 774. doi: 10.3389/fpls.2017.00774

Liu, B., Asseng, S., Müller, C., Ewert, F., Elliott, J., Lobell, D. B., et al. (2016). Similar estimates of temperature impacts on global wheat yield by three independent methods. *Nat. Clim. Change.* 6, 1130–1136. doi: 10.1038/nclimate3115

Li, Y. F., Wu, Y., Hernandez-Espinosa, N., and Peña, R. J. (2013). Heat and drought stress on durum wheat: Responses of genotypes, yield, and quality parameters. *J. Cereal Sci.* 57, 398–404. doi: 10.1016/j.jcs.2013.01.005

Lopes, M. S., El-Basyoni, I., Baenziger, P. S., Singh, S., Royo, C., Ozbek, K., et al. (2015). Exploiting genetic diversity from landraces in wheat breeding for adaptation to climate change. *JXB* 66, 3477–3486.ù. doi: 10.1093/jxb/erv122

Maccaferri, M., Harris, N. S., Twardziok, S. O., Pasam, R. K., Gundlach, H., Spannagl, M., et al. (2019). Durum wheat genome highlights past domestication signatures and future improvement targets. *Nat. Genet.* 51, 885–895. doi: 10.1038/s41588-019-0381-3

Mamluk, O. F., and Nachit, M. M. (1994). Sources of resistance to common bunt (*Tilletia foetida* and t. caries) in durum wheat. *J. Phytopathol.* 142, 122–130. doi: 10.1111/j.1439-0434.1994.tb04522.x

Marone, D., Russo, M. A., Mores, A., Ficco, D. B., Laidò, G., Mastrangelo, A. M., et al. (2021). Importance of landraces in cereal breeding for stress tolerance. *Plants* 10, 1267. doi: 10.3390/plants10071267

Martínez-Moreno, F., Ammar, K., and Solís, I. (2022). Global changes in cultivated area and breeding activities of durum wheat from 1800 to date: a historical review. *Agronomy* 12, 1135. doi: 10.3390/agronomy12051135

Martínez-Moreno, F., Solís, I., Noguero, D., Blanco, A., Özberk, İ., Nsarellah, N., et al. (2020). Durum wheat in the Mediterranean rim: Historical evolution and genetic resources. *Genet. Resour.* 67, 1415–1436. doi: 10.1007/s10722-020-00913-8

Martínez-Moreno, F., Giraldo, P., Cátedra, M. D. M., and Ruiz, M. (2021). Evaluation of leaf rust resistance in the Spanish core collection of tetraploid wheat landraces and association with ecogeographical variables. *Agriculture* 11, 277. doi: 10.3390/agriculture11040277

Marzario, S., Logozzo, G., David, J. L., Zeuli, P. S., and Gioia, T. (2018). Molecular genotyping (SSR) and agronomic phenotyping for utilization of durum wheat (*Triticum durum* desf.) ex situ collection from southern Italy: a combined approach including pedigreed varieties. *Genes* 9, 465. doi: 10.3390/genes9100465

Maucieri, C., Caruso, C., Bona, S., Borin, M., Barbera, A. C., and Cavallaro, V. (2018). Influence of salinity and osmotic stress on germination process in an old sicilian landrace and a modern cultivar of *Triticum durum* desf. *Cereal Res. Commun.* 46, 253–262. doi: 10.1556/0806.46.2018.07

Mazzucotelli, E., Sciara, G., Mastrangelo, A. M., Desiderio, F., Xu, S. S., Faris, J., et al. (2020). The global durum wheat panel (GDP): An international platform to identify and exchange beneficial alleles. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.569905

Medini, M., Ferjaoui, S., Bahri, B., Mhri, W., Hattab, S., and Hamza, S. (2014). Bulk segregant analysis and marker-trait association reveal common AFLP markers for resistance to septoria leaf blotch in Tunisian old durum wheat. *BASE.*

Mehrabi, A. A., Steffenson, B. J., Pour-Aboughadareh, A., Matny, O., and Rahmatov, M. (2022). Genome-wide association study identifies two loci for stripe rust resistance in a durum wheat panel from Iran 4963. doi: 10.3390/app12104963

Melini, V., Melini, F., and Acquistucci, R. (2021). Nutritional characterization of an Italian traditional bread from ancient grains: The case study of the durum wheat bread "Pane di monreale". *Eur. Food Res. Technol.* 247 (1), 193–200. doi: 10.1007/s00217-020-03617-6

Melnikova, N. V., Mitrofanova, O. P., Liapounova, O. A., and Kudryavtsev, A. M. (2010). Global diversity of durum wheat *Triticum durum* desf. for alleles of gliadin-coding loci. *Russ J. Genet.* 46, 43–49. doi: 10.1134/S1022795410010072

Mengistu, D. K., Kidane, Y. G., Catellani, M., Frascaroli, E., Fadda, C., Pè, M. E., et al. (2016). High-density molecular characterization and association mapping in Ethiopian durum wheat landraces reveals high diversity and potential for wheat breeding. *Plant Biotechnol. J.* 14, 1800–1812. doi: 10.1111/pbi.12538

Meyer, R. S., Choi, J. Y., Sanches, M., Plessis, A., Flowers, J. M., Amas, J., et al. (2016). Domestication history and geographical adaptation inferred from a SNP map of African rice. *Nat. Genet.* 48, 1083–1088. doi: 10.1038/ng.3633

Miazzi, M. M., Babay, E., De Vita, P., Montemurro, C., Chaabane, R., Taranto, F., et al. (2022). Comparative genetic analysis of durum wheat landraces and cultivars widespread in Tunisia. *Front. Plant Sci.* 13:939609. doi: 10.3389/fpls.2022.939609

Michel, S., Kummer, C., Gallee, M., Hellinger, J., Ametz, C., Akgöl, B., et al. (2018). Improving the baking quality of bread wheat by genomic selection in early generations. *Theor. Appl. Genet.* 131, 477–493. doi: 10.1007/s00122-017-2998-x

Mohammadi, R., Haghparast, R., Sadeghzadeh, B., Ahmadi, H., Solimani, K., and Amri, A. (2014). Adaptation patterns and yield stability of durum wheat landraces to highland cold rainfed areas of Iran. *Crop Sci.* 54, 944–954. doi: 10.2135/cropsci2013.05.0343

Mo, Y., Howell, T., Vasquez-Gross, H., De Haro, L. A., Dubcovsky, J., and Pearce, S. (2018). Mapping causal mutations by exome sequencing in a wheat TILLING population: a tall mutant case study. *Mol. Genet. Genomics* 293, 463–477. doi: 10.1007/s00438-017-1401-6

Moragues, M., Zarco-Hernandez, J., Moralejo, M. A., and Royo, C. (2006). Genetic diversity of glutenin protein subunits composition in durum wheat landraces [*Triticum turgidum* ssp. *turgidum* convar. *durum* (Desf.) MacKey] from the Mediterranean basin. *Genet. Resour. Crop Evol.* 53 (5), 993–1002. doi: 10.1007/s10722-004-7367-3

Moreno-Amores, J., Michel, S., Miedaner, T., Longin, C. F. H., and Buerstmayr, H. (2020). Genomic predictions for fusarium head blight resistance in a diverse durum wheat panel: An effective incorporation of plant height and heading date as covariates. *Euphytica* 216, 1–19. doi: 10.1007/s10681-019-2551-x

Mulugeta, B., Tesfaye, K., Geleta, M., Johansson, E., Hailesilassie, T., Hammenhag, C., et al. (2022). Multivariate analyses of Ethiopian durum wheat revealed stable and high yielding genotypes. *PloS One* 17, 0273008. doi: 10.1371/journal.pone.0273008

Munns, R., Hare, R. A., James, R. A., and Rebetzke, G. J. (2000). Genetic variation for improving the salt tolerance of durum wheat. *Aust. J. Agric. Res.* 51, 69–74. doi: 10.1071/AR99057

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

Muqaddasi, Q. H. (2017). ""15k SNP chip data for spring and winter wheat [Data set]," in *Plant genomics and phenomics research data repository (PGP)* (Germany: IPK Gatersleben, Seeland OT Gatersleben, Corrensstraße 3).

Naranjo, T., Cuñado, N., and Santos, J. L. (2022). Assessing the heat tolerance of meiosis in spanish landraces of tetraploid wheat *Triticum turgidum*. *Plants* 11, 1661. doi: 10.3390/plants11131661

Nazco, R., Peña, R. J., Ammar, K., Villegas, D., Crossa, J., Moragues, M., et al. (2014a). Variability in glutenin subunit composition of Mediterranean durum wheat germplasm and its relationship with gluten strength. *J. Agric. Sci.* 152 (3), 379–393. doi: 10.1017/S0021859613000117

Nazco, R., Peña, R. J., Ammar, K., Villegas, D., Crossa, J., and Royo, C. (2014b). Durum wheat (*Triticum durum* desf.) Mediterranean landraces as sources of variability for allelic combinations at glu-1/Glu-3 loci affecting gluten strength and pasta cooking quality. *Genet. Resour. Crop Evol.* 61, 1219–1236. doi: 10.1007/s10722-014-0104-7

Nazco, R., Villegas, D., Ammar, K., Pena, R. J., Moragues, M., and Royo, C. (2012). Can Mediterranean durum wheat landraces contribute to improved grain quality attributes in modern cultivars? *Euphytica* 185 (1), 1–7. doi: 10.1007/s10681-011-0588-6

Olivera, P. D., Bulbula, W. D., Badebo, A., Bockelman, H. E., Edae, E. A., and Jin, Y. (2021). Field resistance to wheat stem rust in durum wheat accessions deposited at the USDA national small grains collection. *Crop Sci.* 61, 2565–2578. doi: 10.1002/csc2.20466

Ostezan, A., McDonald, S. C., Tran, D. T., Souza, R. S. E., and Li, Z. (2021). Target region sequencing and applications in plants. *JCSB* 24, 13–26.

Ouaja, M., Aouini, L., Bahri, B., Ferjaoui, S., Medini, M., Marcel, T. C., et al. (2020). Identification of valuable sources of resistance to *Zymoseptoria tritici* in the Tunisian durum wheat landraces. *Eur. J. Plant Pathol.* 156, 647–661. doi: 10.1007/s10658-019-01914-9

Özkan, H., Brandolini, A., Schäfer-Pregl, R., and Salamini, F. (2002). AFLP analysis of a collection of tetraploid wheats indicates the origin of emmer and hard wheat domestication in southeast Turkey. *MBE* 19, 1797–1801. doi: 10.1093/oxfordjournals.molbev.a004002

Palumbo, M., Blangiforti, S., Cambrea, M., Gallo, G., Licciardello, S., and Spina, A. (2008). "Sicilian Durum wheat landraces for production of traditional breads," in *Proceedings of the International Durum Wheat Symposium "From seed to pasta: the durum wheat chain"*, Bologna, Italy, 132.

Pavan, S., Delvento, C., Ricciardi, L., Lotti, C., Ciani, E., and D'Agostino, N. (2020). Recommendations for choosing the genotyping method and best practices for quality control in crop genome-wide association studies. *Front. Genet.* 11, 447. doi: 10.3389/fgene.2020.00447

Pecetti, L., Boggini, G., and Gorham, J. (1994). Performance of durum wheat landraces in a Mediterranean environment (eastern Sicily). *Euphytica* 80, 191–199. doi: 10.1007/BF00039650

Perry, D. J., and Lee, S. J. (2017). Durum wheat variety identification by OpenArray analysis. *Can. J. Plant Sci.* 97, 403–407. doi: 10.1139/cjps-2016-0300

Poland, J., Endelman, J., Dawson, J., Rutkoski, J., Wu, S., Manes, Y., et al. (2012). Genomic selection in wheat breeding using genotyping-by-sequencing. *TPG* 5. doi: 10.3835/plantgenome2012.06.0006

Qureshi, N., Bariana, H., Kolmer, J. A., Miah, H., and Bansal, U. (2017). Genetic and molecular characterization of leaf rust resistance in two durum wheat landraces. *Phytopathol* 107, 1381–1387. doi: 10.1094/PHYTO-01-17-0005-R

Qureshi, N., Bariana, H., Kumran, V. V., Muruga, S., Forrest, K. L., Hayden, M. J., et al. (2018). A new leaf rust resistance gene *Lr79* mapped in chromosome 3BL from the durum wheat landrace Aus26582. *Theor. App. Genet.* 131, 1091–1098. doi: 10.1007/s00122-018-3060-3

Rasheed, A., Mujeeb-Kazi, A., Ogbonnaya, F. C., He, Z. H., and Rajaram, S. (2018). Wheat genetic resources in the post-genomics era: promise and challenges. *Ann. Bot-London* 121, 603–616. doi: 10.1093/aob/mcx148

Rasheed, A., and Xia, X. (2019). From markers to genome-based breeding in wheat. *Theor. App. Genet.* 132, 767–784. doi: 10.1007/s00122-019-03286-4

Rellstab, C., Gugerli, F., Eckert, A. J., Hancock, A. M., and Holderegger, R. (2015). A practical guide to environmental association analysis in landscape genomics. *Mol. Ecol.* 24, 4348–4370. doi: 10.1111/mec.13322

Requena-Ramírez, M. D., Hornero-Méndez, D., Rodríguez-Suárez, C., and Atienza, S. G. (2021). Durum wheat (*Triticum durum* l.) landraces reveal potential for the improvement of grain carotenoid esterification in breeding programs. *Foods* 10, 757. doi: 10.3390/foods10040757

Requena-Ramírez, M. D., Rodríguez-Suárez, C., Flores, F., Hornero-Méndez, D., and Atienza, S. G. (2022). Marker-trait associations for total carotenoid content and individual carotenoids in durum wheat identified by genome-wide association analysis. *Plants* 11, 2065. doi: 10.3390/plants11152065

Reynolds, M. P., Hobbs, P. R., and Braun, H. J. (2007). Challenges to international wheat improvement. *J. Agric. Sci.* 145, 223. doi: 10.1017/S0021859607007034

Reynolds, M. P., Lewis, J. M., Ammar, K., Basnet, B. R., Crespo-Herrera, L., Crossa, J., et al. (2021). Harnessing translational research in wheat for climate resilience. *JXB* 72, 5134–5157. doi: 10.1093/jxb/erab256

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* 40, 913–929. doi: 10.1111/ecog.02881

Roncallo, P. F., Guzmán, C., Larsen, A. O., Achilli, A. L., Dreisigacker, S., Molfese, E., et al. (2021). Allelic variation at glutenin loci (*Glu-1*, glu-2 and *Glu-3*) in a worldwide durum wheat collection and its effect on quality attributes. *Foods* 11, 2845. doi: 10.3390/foods10112845

Roselló, M., Royo, C., Álvaro, F., Villegas, D., Nazco, R., and Soriano, J. M. (2018). Pasta-making quality QTLome from Mediterranean durum wheat landraces. *Front. Plant Sci.* 9, 1512. doi: 10.3389/fpls.2018.01512

Royo, C., Nazco, R., and Villegas, D. (2014). The climate of the zone of origin of Mediterranean durum wheat (*Triticum durum* desf.) landraces affects their agronomic performance. *Genet. Resour.* 61, 1345–1358. doi: 10.1007/s10722-014-0116-3

Royo, C., Ammar, K., Villegas, D., and Soriano, J. M. (2021). Agronomic, physiological and genetic changes associated with evolution, migration and modern breeding in durum wheat. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.674470

Royo, C., Dreisigacker, S., Soriano, J. M., Lopes, M. S., Ammar, K., Villegas, D., et al (2020). Allelic variation at the vernalization response (Vrn-1) and photoperiod sensitivity (Ppd-1) genes and their association with the development of durum wheat landraces and modern cultivars. *Front. Plant Sci.* 11-838.

Ruisi, P., Ingraffia, R., Urso, V., Giambalvo, D., Alfonzo, A., Corona, O., et al. (2021). Influence of grain quality, semolinas and baker's yeast on bread made from old landraces and modern genotypes of Sicilian durum wheat. *Int. Food Res. J.* 140, 110029. doi: 10.1016/j.foodres.2020.110029

Ruiz, M., Bernal, G., and Giraldo, P. (2018). An update of low molecular weight glutenin subunits in durum wheat relevant to breeding for quality. *J. Cereal Sci.* 83, 236–244. doi: 10.1016/j.jcs.2018.09.005

Ruiz, M., Giraldo, P., Royo, C., Villegas, D., Aranzana, M. J., and Carrillo, J. M. (2012). Diversity and genetic structure of a collection of Spanish durum wheat landraces. *Crop Sci.* 52, 2262–2275. doi: 10.2135/cropsci2012.02.0081

Saccomanno, A., Matny, O., Marone, D., Laidò, G., Petruzzino, G., Mazzucotelli, E., et al. (2018). Genetic mapping of loci for resistance to stem rust in a tetraploid wheat collection. *Int. J. Mol. Sci.* 19, 3907. doi: 10.3390/ijms19123907

Sahri, A., Chentoufi, L., Arbaoui, M., Ardisson, M., Belqadi, L., Birouk, A., et al. (2014). Towards a comprehensive characterization of durum wheat landraces in Moroccan traditional agrosystems: analysing genetic diversity in the light of geography, farmers' taxonomy and tetraploid wheat domestication history. *BMC evol. Biol.* 14, 1–18. doi: 10.1186/s12862-014-0264-2

Saintenac, C., Jiang, D., and Akhunov, E. D. (2011). Targeted analysis of nucleotide and copy number variation by exon capture in allotetraploid wheat genome. *Genome Biol.* 12, R88. doi: 10.1186/gb-2011-12-9-r88

Sandve, S. R., Rudi, H., Dørum, G., Vigeland, M. D., Berg, P. R., and Rognli, O. A. (2010). "Genotyping unknown genomic terrain in complex plant genomes," in *Sustainable use of genetic diversity in forage and turf breeding*. Ed. C. Huyghe (New Mexico: Springer), 455–459.

Sareen, S., Bhusal, N., Kumar, M., Bhati, P. K., Munjal, R., Kumari, J., et al. (2020). Molecular genetic diversity analysis for heat tolerance of indigenous and exotic wheat genotypes. *J. Plant Biochem. Biotechnol.* 29, 15–23. doi: 10.1007/s13562-019-00501-7

Scarascia Mugnozza, G. T. (2005). '*The contribution of Italian wheat geneticists: from Nazareno Strampelli to Francesco D'Amato*'. Rome: Accademia Nazionale delle Scienze, 53–75.

Scavo, A., Pandino, G., Restuccia, A., Caruso, P., Lombardo, S., and Mauromicale, G. (2022). Allelopathy in durum wheat landraces as affected by genotype and plant part. *Plants* 11, 1021. doi: 10.3390/plants11081021

Scheben, A., Batley, J., and Edwards, D. (2018). Revolution in genotyping platforms for crop improvement. *Plant Genet. Mol. Biol.* 164, 37–52. doi: 10.1007/10_2017_47

Schulthess, A. W., Kale, S. M., Liu, F., Zhao, Y., Philipp, N., Rembe, M., et al. (2022). Genomics-informed prebreeding unlocks the diversity in genebanks for wheat improvement. *Nat. Genet.* 54, 1544–1552. doi: 10.1038/s41588-022-01189-7

Shamaya, N. J., Shavrukov, Y., Langridge, P., Roy, S. J., and Tester, M. (2017). Genetics of na+ exclusion and salinity tolerance in Afghani durum wheat landraces. *BMC Plant Biol.* 17, 1–8. doi: 10.1186/s12870-017-1164-6

Shan, D., Ali, M., Shahid, M., Arif, A., Waheed, M. Q., Xia, X., et al. (2022). Genetic networks underlying salinity tolerance in wheat uncovered with genome-wide analyses and selective sweeps. *Theor. Appl. Genet.* 135, 2925–2941. doi: 10.1007/s00122-022-04153-5

Sharma, S., Schulthess, A. W., Bassi, F. M., Badaeva, E. D., Neumann, K., Graner, A., et al. (2021). Introducing beneficial alleles from plant genetic resources into the wheat germplasm. *Biology* 10, 982. doi: 10.3390/biology10100982

Shavrukov, Y., Shamaya, N., Baho, M., Edwards, J., Ramsey, C., Nevo, E., et al. (2011). Salinity tolerance and na+ exclusion in wheat: variability, genetics, mapping populations and QTL analysis. *Czech J. Genet. Plant Breed* 47, 85–93. doi: 10.17221/3260-CJGPB

Shayanmehr, S., Henneberry, S. R., Sabouhi, M. S., and Foroushani, N. S. (2020). Drought, climate change, and dryland wheat yield response: An econometric approach. *Int. J. Environ. Res. Public Health* 17, 5264. doi: 10.3390/ijerph17145264

Shewry, P. (2019). What is gluten–why is it special? *Front. Nutr.* 101. doi: 10.3389/fnut.2019.00101

Soriano, J. M., Villegas, D., Sorrells, M. E., and Royo, C. (2018). Durum wheat landraces from east and west regions of the mediterranean basin are genetically distinct for yield components and phenology. *Front. Plant Sci.* 9, 80. doi: 10.3389/fpls.2018.00080

Sork, V. L., Aitken, S. N., Dyer, R. J., Eckert, A. J., Legendre, P., and Neale, D. B. (2013). Putting the landscape into the genomics of trees: Approaches for understanding local adaptation and population responses to changing climate. *Tree Genet. Genomes* 9, 901–911. doi: 10.1007/s11295-013-0596-x

Soriano, J. M., Colasuonno, P., Marcotuli, I., and Gadaleta, A. (2021). Meta-QTL analysis and identification of candidate genes for quality, abiotic and biotic stress in durum wheat. *Sci. Rep.* 11, 11877. doi: 10.1038/s41598-021-91446-2

Spina, A., Dinelli, G., Palumbo, M., Whittaker, A., Cambrea, M., Negri, L., et al. (2021). Evaluation of standard physico-chemical and rheological parameters in predicting bread-making quality of durum wheat (*Triticum turgidum* l. ssp. *durum* [Desf.] husn.). *Int. J. Food Sci.* 56, 3278–3288. doi: 10.1111/ijfs.15018

Spinoni, J., Vogt, J. V., Naumann, G., Barbosa, P., and Dosio, A. (2018). Will drought events become more frequent and severe in Europe? *Int. J. Climatol.* 38, 1718–1736. doi: 10.1002/joc.5291

Steemers, F. J., and Gunderson, K. L. (2007). Whole genome genotyping technologies on the BeadArray™ platform. *Biotechnol. Journal: Healthcare Nutr. Technol.* 2, 41–49. doi: 10.1002/biot.200600213

Subira, J., Peña, R. J., Álvaro, F., Ammar, K., Ramdani, A., and Royo, C. (2014). Breeding progress in the pasta-making quality of durum wheat cultivars released in Italy and Spain during the 20th century. *Crop Pasture Sci.* 65 (1), 16–26. doi: 10.1071/CP13238

Sukumaran, S., Reynolds, M. P., and Sansaloni, C. (2018). Genome-wide association analyses identify QTL hotspots for yield and component traits in durum wheat grown under yield potential, drought, and heat stress environments. *Front. Plant Sci.* 9, 81. doi: 10.3389/fpls.2018.00081

Taghouti, M., Rhrib, K., and Gaboun, F. (2013). "Exploiting landrace genetic diversity for germplasm enhancement in durum wheat breeding in Morocco," in: E. Porceddu, A. B. Damania and C. O. Qualset (ed.). *Proceedings of the International Symposium on Genetics and Breeding of Durum Wheat*. Bari: CIHEAM. p. 109–119 (Options Méditerranéennes : Série A. Séminaires Méditerranéens; n. 110)

Talas, F., Longin, F., and Miedaner, T. (2011). Sources of resistance to fusarium head blight within Syrian durum wheat landraces. *Plant Breed.* 130, 398–400. doi: 10.1111/j.1439-0523.2011.01867.x

Tan, A. (2002). *In situ on-farm conservation of landraces grown in north-Western transitional zone of Turkey (in Turkish), Sonuc Raporu (final report).* Tubitak-Togtag-2347. Tübitak, Ankara.

Tanksley, S. D., and McCouch, S. R. (1997). Seed banks and molecular maps: unlocking genetic potential from the wild. *Science* 277, 1063–1066. doi: 10.1126/science.277.5329.1063

Taranto, F., D'Agostino, N., Rodriguez, M., Pavan, S., Minervini, A. P., Pecchioni, N., et al. (2020). Whole genome scan reveals molecular signatures of divergence and selection related to important traits in durum wheat germplasm. *Front. Genet.* 11, 217. doi: 10.3389/fgene.2020.00217

Taranto, F., Di Serio, E., Miazzi, M. M., Pavan, S., Saia, S., De Vita, P., et al. (2022). Intra-and inter-population genetic diversity of "Russello" and "Timilia" landraces from Sicily: A proxy towards the identification of favorable alleles in durum wheat. *Agronomy* 12, 1326. doi: 10.3390/agronomy12061326

Taranto, F., Mangini, G., Pasqualone, A., Gadaleta, A., and Blanco, A. (2015). Mapping and allelic variations of *Ppo-B1* and *Ppo-B2* gene-related polyphenol oxidase activity in durum wheat. *Mol. Breed.* 35 (2), 1–10. doi: 10.1007/s11032-015-0272-y

Taranto, F., Nicolia, A., Pavan, S., De Vita, P., and D'Agostino, N. (2018). Biotechnological and digital revolution for climate-smart plant breeding. *Agronomy* 8 (12), 277. doi: 10.3390/agronomy8120277

Thomson, M. J. (2014). High-throughput SNP genotyping to accelerate crop improvement. *Plant Breed. Biotech.* 2, 195–212. doi: 10.9787/PBB.2014.2.3.195

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

Trebbi, D., Maccaferri, M., de Heer, P., Sørensen, A., Giuliani, S., Salvi, S., et al. (2011). High-throughput SNP discovery and genotyping in durum wheat (*Triticum durum* desf.). *Theor. Appl. Genet.* 123, 555–569. doi: 10.1007/s00122-011-1607-7

Van Orsouw, N. J., Hogers, R. C., Janssen, A., Yalcin, F., Snoeijers, S., Verstege, E., et al. (2007). Complexity reduction of polymorphic sequences (CRoPS™): a novel approach for large-scale polymorphism discovery in complex genomes. *PloS One* 2 (11), e1172. doi: 10.1371/journal.pone.0001172

Varella, A. C., Zhang, H., Weaver, D. K., Cook, J. P., Hofland, M. L., et al. (2019). A novel QTL in durum wheat for resistance to the wheat stem sawfly associated with early expression of stem solidness. *G3 (Bethesda).* 9, 1999–2006. doi: 10.1534/g3.119.400240

Visioli, G., Giannelli, G., Agrimonti, C., Spina, A., and Pasini, G. (2021). Traceability of Sicilian durum wheat landraces and historical varieties by high molecular weight glutenins footprint. *Agronomy* 11 (1), 143. doi: 10.3390/agronomy11010143

Vita, F., Taiti, C., Pompeiano, A., Gu, Z., Lo Presti, E., Whitney, L., et al. (2016). Aromatic and proteomic analyses corroborate the distinction between Mediterranean landraces and modern varieties of durum wheat. *Sci. Rep.* 6, 1–15. doi: 10.1038/srep34619

Wang, Z., Wang, W., Xie, X., Wang, Y., Yang, Z., Peng, H., et al. (2022). Dispersed emergence and protracted domestication of polyploid wheat uncovered by mosaic ancestral haploblock inference. *Nat. Commun.* 13, 1–14. doi: 10.1038/s41467-022-31581-0

Wang, S., Wong, D., Forrest, K., Allen, A., Chao, S., Huang, B. E., et al. (2014). Characterization of polyploid wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array. *Plant Biotechnol. J.* 12, 787–796. doi: 10.1111/pbi.12183

Wang, S., Xu, S., Chao, S., Sun, Q., Liu, S., and Xia, G. (2019). A genome-wide association study of highly heritable agronomic traits in durum wheat. *Front. Plant Sci.* 10, 919. doi: 10.3389/fpls.2019.00919

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

Winfield, M. O., Allen, A. M., Burridge, A. J., Barker, G. L., Benbow, H. R., Wilkinson, P. A., et al. (2016). High-density SNP genotyping array for hexaploid wheat and its secondary and tertiary gene pool. *Plant Biotechnol. J.* 14, 1195–1206. doi: 10.1111/pbi.12485

Xynias, I. N., Mylonas, I., Korpetis, E. G., Ninou, E., Tsaballa, A., Avdikos, I. D., et al. (2020). Durum wheat breeding in the Mediterranean region: Current status and future prospects. *Agronomy* 10, 432. doi: 10.3390/agronomy10030432

Zeven, A. C. (1999). The traditional inexplicable replacement of seed and seed ware of landraces and cultivars: a review. *Euphytica* 110, 181–191. doi: 10.1023/A:1003701529155

Zhou, Z., Jiang, Y., Wang, Z., Gou, Z., Lyu, J., Li, W., et al. (2015). Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat. Biotechnol.* 33 (4), 408–414. doi: 10.1038/nbt.3096

Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed model analysis for association studies. *Nat. Genet.* 44, 821–824. doi: 10.1038/ng.2310

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 02 March 2023
DOI 10.3389/fpls.2023.1071648

OPEN ACCESS

EDITED BY
Nisha Singh,
Gujarat Biotechnology University, India

REVIEWED BY
Abhinandan Surgonda Patil,
Agharkar Research Institute, India
Changlong Wen,
Beijing Vegetable Research Center, China

*CORRESPONDENCE
Tusar Kanti Behera
tusar@rediffmail.com
Shyam Sundar Dey
shyam.iari@gmail.com;
shyam.dey@icar.gov.in

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

RECEIVED 16 October 2022
ACCEPTED 07 February 2023
PUBLISHED 02 March 2023

CITATION
N. D. V, Matsumura H, Munshi AD, Ellur RK, Chinnusamy V, Singh A, Iquebal MA, Jaiswal S, Jat GS, Panigrahi I, Gaikwad AB, Rao AR, Dey SS and Behera TK (2023) Molecular mapping of genomic regions and identification of possible candidate genes associated with gynoecious sex expression in bitter gourd. *Front. Plant Sci.* 14:1071648. doi: 10.3389/fpls.2023.1071648

Molecular mapping of genomic regions and identification of possible candidate genes associated with gynoecious sex expression in bitter gourd

Vinay N. D.[1], Hideo Matsumura[2], Anilabha Das Munshi[1], Ranjith Kumar Ellur[3], Viswanathan Chinnusamy[4], Ankita Singh[5], Mir Asif Iquebal[5], Sarika Jaiswal[5], Gograj Singh Jat[1], Ipsita Panigrahi[1], Ambika Baladev Gaikwad[6], A. R. Rao[5], Shyam Sundar Dey[1]* and Tusar Kanti Behera[1,7]*

[1]Division of Vegetable Science, ICAR-Indian Agricultural Research Institute, New Delhi, India, [2]Gene Research Centre, Shinshu University, Ueda, Nagano, Japan, [3]Division of Genetics, ICAR-Indian Agricultural Research Institute, New Delhi, India, [4]Division of Plant Physiology, ICAR-Indian Agricultural Research Institute, New Delhi, India, [5]Centre for Agricultural Bioinformatics, ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India, [6]Division of Genomic Resources, ICAR-National Bureau of Plant Genetic Resources, New Delhi, India, [7]ICAR-Indian Institute of Vegetable Research, Varanasi, Uttar Pradesh, India

Bitter gourd is an important vegetable crop grown throughout the tropics mainly because of its high nutritional value. Sex expression and identification of gynoecious trait in cucurbitaceous vegetable crops has facilitated the hybrid breeding programme in a great way to improve productivity. In bitter gourd, gynoecious sex expression is poorly reported and detailed molecular pathways involve yet to be studied. The present experiment was conducted to study the inheritance, identify the genomic regions associated with gynoecious sex expression and to reveal possible candidate genes through QTL-seq. Segregation for the gynoecious and monoecious sex forms in the F_2 progenies indicated single recessive gene controlling gynoecious sex expression in the genotype, PVGy-201. Gynoecious parent, PVGy-201, Monoecious parent, Pusa Do Mausami (PDM), and two contrasting bulks were constituted for deep-sequencing. A total of 10.56, 23.11, 15.07, and 19.38 Gb of clean reads from PVGy-201, PDM, gynoecious bulk and monoecious bulks were generated. Based on the ΔSNP index, 1.31 Mb regions on the chromosome 1 was identified to be associated with gynoecious sex expression in bitter gourd. In the QTL region 293,467 PVGy-201 unique variants, including SNPs and indels, were identified. In the identified QTL region, a total of 1019 homozygous variants were identified between PVGy1 and PDM genomes and 71 among them were non-synonymous variants (SNPS and INDELs), out of which 11 variants (7 INDELs, 4 SNPs) were classified as high impact variants with frame shift/stop gain effect. In total twelve genes associated with male and female gametophyte development were identified in the QTL-region. Ethylene-responsive transcription factor 12, Auxin response factor 6, Copper-transporting ATPase RAN1, CBL-interacting serine/threonine-protein kinase 23, ABC transporter C family member 2, DEAD-box

ATP-dependent RNA helicase 1 isoform X2, Polygalacturonase QRT3-like isoform X2, Protein CHROMATIN REMODELING 4 were identified with possible role in gynoecious sex expression. Promoter region variation in 8 among the 12 genes indicated their role in determining gynoecious sex expression in bitter gourd genotype, DBGy-1. The findings in the study provides insight about sex expression in bitter gourd and will facilitate fine mapping and more precise identification of candidate genes through their functional validation.

KEYWORDS

bitter gourd (*Momordica charantia*), sex expression, gynoecious, inheritance, QTL-seq, candidate genes

Introduction

Momordica charantia L. (2n=22), often known as bitter gourd or bitter melon, is a prominent vegetable cum medicinal plant grown widely in India, China, Malaysia, Africa, and South America (Heiser, 1979). Bitter gourd is rich source of ascorbic acid and iron, and is renowned for its anti-diabetic, anti-carcinogenic, and anti-HIV properties (Behera, 2004; Behera et al., 2010). Indian bitter gourd germplasm exhibits wide phenotypic diversity for growth habit, maturity, fruit shape, size, colour, surface texture and sex expression (Robinson and Decker-Walters, 1999). To exploit genetic variation for crop improvement, it is essential to understand genetic and molecular basis of the traits under consideration.

In flowering plants, sex determination is a key developmental process of great biological significance (Kater et al., 2001). The family *Cucurbitaceae*, is regarded as model to study the physiological and molecular mechanisms of sex determination (Bhowmick and Jha, 2015). Monoecy, bearing separate male and female unisexual flowers on the same plant, is predominant sex form in Cucurbitaceae than the more typical bisexual flowers in higher plants (Behera, 2004). Genetic and molecular basis of sex determination is well documented in two major *Cucumis* species, musk melon *(C. melo)* and cucumber (*C. sativus*). Sex expression in these two species is regulated by highly orthologous and conserved genes related to ethylene biosynthesis and signalling pathways (Boualem et al., 2008). For instance, *CsACS2* and *CmACS7* correspond to the *m* locus governing andromonoecy in cucumber and melon respectively. Mutation in M locus leads to bisexual flowers (Boualem et al., 2008; Boualem et al., 2009; Li et al., 2009). Mutations in *CsACS11* or *CmACS11* gene, which correspond to *a* (*a*ndroecy) loci, produce male flowers and androecious plants (Boualem et al., 2015), while similar androecious plants in cucumber are produced with mutation of *CsACO2* (Chen et al., 2016). So, it is clearly evident that in *Cucumis*, ethylene plays a major role in sex determination. However, mechanisms associated with sex regulation in *Momordica* is still unknown.

Bitter gourd is primarily a monoecious species; however, gynoecious lines with complete femaleness are reported from China, Japan, and India (Ram et al., 2002; Behera et al., 2006; Iwamoto and Ishida, 2006). Gynoecy has potential applications in heterosis breeding in exploiting earliness, high yield, quality and resistance through hybrid development. Use of gynoecious line as female counterpart has substantially decreased the hybrid seed cost and enhanced genetic purity of hybrids (Dey et al., 2010).

Gyneoecious phenotype in cucurbits arises either due to stamen arrest by enhanced ethylene production (Trebitsh et al., 1997; Mibus and Tatlioglu, 2004; Boualem et al., 2009) or anther specific DNA damage (Gu et al., 2011). Both recessive and dominant gene control of gynoecy is reported in cucurbits; in cucumber gynoecy is controlled by dominant gene named *Acr/stF/AcrF/F* (Shifriss, 1961; Galun, 1962; Kubicki, 1969; Robinson et al., 1976). The *F* locus encodes a duplicated copy of *CsACS 1* gene named as *CsACS1G* (Mibus and Tatlioglu, 2004). So, gynoecious line with additional ethylene synthase gene (*CsACS1G*) produces more ethylene leading to complete femaleness (Atsmon and Tabbak, 1979). On the other hand, Gynoecious sex expression is controlled by single recessive gene in muskmelon (*g)* and watermelon *(gy)* (Poole and Grimball, 1939). Gynoecy in water and musk melon is due to mutation in male determinant *CmWIP/CmWIP1 locus*, which encodes a *C2H2* zinc-finger-type transcription factor accountable for carpel abortion in bisexual flowers (Martin et al., 2009; Zhang et al., 2020). The genetic of gynoecy is well documented in bitter gourd and is known to be controlled a single recessive gene (Behera et al., 2006; Ram et al., 2006; Behera et al., 2009; Matsumura et al., 2014) which is similar to other species is controlled by the WIP gene. First genetic mapping of gynoecy locus was done by Matsumura et al. (2014) to a 12.73-cM genetic interval. Later on Cui et al. (2018), also mapped the gy locus to a 3.5-Mb physical interval between 17,619,724 bp and 21,144,642 bp on MC01.

Recently Fine mapping of gynoecy locus and candidate gene detection was done by Cui et al., 2022 by combining BSA seq and traditional molecular marker linkage analysis. Study mapped gynoecy locus i. e Mcgy1 locus into a 292.70-kb physical interval of 20,851,441 - 21,148,382 bp on MC01. Homologous gene of WIP or CsACS1G was not found Mcgy1 locus, signifying that the casual candidate of Gynoecy sex may be due to a novel gene than the previously reported in other cucurbits. Based on evaluation of candidate mutations in the gynoecy locus and gene expression

studies and gene annotation information (Cui et al., 2020), gene MC01g1681 encoding Cytidine triphosphate synthase (CTPS), was considered as the candidate gene of Mcgy1. CTPS plays rate-limiting role in final step of *de novo* synthesis of cytidine triphosphate (CTP), which is essential for DNA, RNA and phospholipid biosynthesis in all organisms (Daumann et al., 2018). In the same study, RNA seq analysis identified several other candidates genes for gynoecy including, MC02g0607 belonging to AGL1 MADS-box family; MC05g0014 is an uclacyanin3-like gene; and MC04g1310. The identified novel candidate genes doesn't seem to have any direct connection with ethylene biosynthesis/signaling. Thus, a different mechanism might be controlling gynoecy sex expression in bitter gourd. However, several studies have indicated the role of phytohormone perticluarly ethylene in sex expression of bitter gourd. Cui et al., 2022 observed differential expression of ethylene signal transduction genes, MC11g0603, encoding a Constitutive Triple Response 1 (CTR1), and MC04g0109, encoding an Ethylene Insensitive 3 (EIN3), between monoecious and gynoecious lines. CTR1 is known to promote biosynthesis of ethylene (Leclercq et al., 2002), whereas EIN has the inhibitory effect on ethylene biosynthesis (Wang et al., 2020). Use of ethylene inhibitory agents such as, Silver nitrate treatment for altering the sex forms in *Momordica* is reported (Hossain et al., 1996). *In-Silico* gene expression analysis by Gunnaiah et al., 2014 also reported ethylene biosynthesis and regulation genes as putative candidates for gynoecy expression in bitter gourd. These results suggests that ethylene play an important role in the formation of gynoecy in bitter gourd. Cui et al., 2022 also identified several other genes involved in the in plant hormone signal transduction, such as that of gibberellin and auxin. Hence, similar to other cucurbits, phytohormone and their cross talk may be associated with formation of gynoecy in bitter gourd. More elaborated studies need to be conducted to narrow down to exact candidate genes and to identify the molecular mechanism regulating gynoecy in bitter gourd.

Cross-talk among the important phytohormone is the well-known pathway associated with sex differentiation in cucurbits (Mandal et al., 2022). In addition to ethylene, other plant hormones like auxin, brassinosteroids (BRs), gibberillic acid (GA) and ABA also contribute to sex determination either by influencing ethylene biosynthesis and signalling or through ethylene independent pathways (Rudich et al., 1972; Trebitsh et al., 1987; Yin and Quinn, 1995; Papadopoulou and Grumet, 2005; Zhang et al., 2014; Tao et al., 2018). Exogenous auxin has feminizing effect in cucumber sex expression (Rudich et al., 1972; Trebitsh et al., 1987) and *C. maxima* (Wang et al., 2019) through up regulation of ethylene biosynthesis and signalling genes. Similarly, Gibberellic acid (GA) also exhibit male promoting effect on monoecious cucumber (Peterson and Anhder, 1960), musk melon and watermelon (Girek et al., 2013; Zhang et al., 2017). The masculinizing effect of GA is either through inhibition of ethylene biosynthesis (Yin and Quinn, 1995) or ethylene-independent manner (Zhang et al., 2014; Zhang et al., 2017). Besides, ABA is known to promote maleness in cucurbits through inhibition of ethylene biosynthesis by down-regulating of ACO genes (Gao et al., 2015; Lee et al., 2017). However, BRs (brassinosteroids) are known to promote ethylene induced femaleness thus, indirectly participate in cucumber sex determination (Papadopoulou and Grumet, 2005; Pawełkowicz et al., 2012). Similar to BA, cytokinin also acts as regulatory switch to increase the level of the ethylene by enhancing the stability of the ACS proteins (Lee et al., 2017). In brief, ethylene is considered as key regulator in sex determination of cucurbits and influence of other plant hormones in sex expression mainly through cross-talk with ethylene.

Marker assisted breeding has significantly accelerated the crop improvement programme. Success of Marker Assisted Selection (MAS) depends on QTL analysis though construction of high-density linkage map and identification of reliable and tightly linked marker to the trait of interest. However, application of molecular breeding in bitter gourd is limited due to the unavailability of decisive linkage map and scarcity of polymorphic markers (Rao et al, 2018). Therefore, the molecular basis of many economic traits is still unknown, and the use of molecular breeding in bitter gourd improvement programmes is still in its infancy. Advancement in sequencing chemistries and availability of next generation sequencing (NGS) techniques provide cheaper, rapid and efficient methods for high-density SNP discovery and genotyping in large populations (Davey et al., 2011). The availability of whole genome sequence of bitter gourd in public domain has served as an ideal resource for genome-wide identification of SSR and SNP markers *in silico* (Urasaki et al., 2017; Cui et al., 2020; Matsumura et al., 2020). This has encouraged researchers to work on genetic map construction, fine mapping and MAS of bitter gourd (Cui et al., 2018; Rao et al., 2018; Rao et al., 2021; Kaur et al., 2022). Recently, QTL-seq has been used as efficient tool for rapid mapping of QTLs and identification of candidate genes in less time and cost (Takagi et al., 2013). It has been successfully employed in mapping of economic traits in variety of vegetable crops including cucumber and tomato (Lu et al., 2014; Illa-Berenguer et al., 2015; Ruangrak et al., 2018). Hence, in the present study QTL-Seq analysis was performed to identify the candidate gene/s associated gynoecious sex expression in PVGy -201. Although sex expression is a vital developmental process in plant sexual reproduction, it is poorly reported in bitter gourd. Elucidating the mechanism underlying flower development and sex expression serves as a valuable resource for sex manipulation for academic and economic benefits in bitter gourd and related crops.

Materials and methods

Plant materials and phenotyping

Two parental lines (PVGy-201 and Pusa Do Mausami) showing contrasting sex expression patterns were chosen as parents. The line PVGy-201 is a gynoeious line developed by transferring gynoecy trait to the Pusa Vishesh background. It exhibit stable-complete gynoecy (100% femaleness) and high female flower production even in temperature at high as 38-40°C. Furthermore, it is easy to maintain through male inducing chemicals such as silver nitrate and silver thiosulphate. The male parent, monoecious line Pusa Do Mausami (PDM), exhibits high male expression (> 95%) and

delayed female appearance, hence serves as ideal monoecious counterpart. Since androecous lines with complete maleness are not yet reported in bitter gourd (Kole et al., 2020), monoecious lines with high male tendency such as PDM is used for genetic studies. (Figure 1; Table 1). The F_1 plants were generated by the cross PVGy-201 × Pusa Do Mausami during Kharif, 2018 and F_2 generation was obtained by self-fertilization (pollination of female flowers with pollen from the same plant) of the F_1 plants during the spring-summer, 2019. The final experiment with 147 F_2 plants along with parents were grown during Kharif season of 2019. The aforementioned population generation and phenotyping work was conducted at the Vegetable Research Farm of Indian Agricultural Research Institute (IARI), New Delhi, India.

Sex expression of each plant was determined by recording the sex form of every flower produced from each plant. Subsequently using this data, male to female sex ratio of each F_2 plant was computed. The plant with 100% female flowers was considered as gynoecious and those with both male and female flowers are considered as monoecious. The gynoecious plants with highest female flowers were used to formulate gynoecious bulk (G-bulk) and those with higher male flower percentage (>95%) were included in monoecious bulk (M-bulk). Each bulk contained 12 F_2 plants exhibiting contrasting sex expression were taken for QTL-seq analysis.

DNA extraction and whole genome re-sequencing of parents and bulked DNA

Total genomic DNA was extracted from young-fresh leaves of the parents (PVGy-201 and PDM), G- bulk and M-bulk (Doyle and Doyle, 1987). DNA concentration was determined using Nano

FIGURE 1
The parents with contrasting behaviour for sex expression **(A)** gynoecious line, PVGy-201 with only female flower **(B)** Monoecious line, Pusa Do Mousami with < 5% female flowers in any plant.

10.3389/fpls.2023.1071648

TABLE 1 Flowering related traits of parents, PVGy-201 and PDM used in the present study.

Traits	PVGy-201	PDM
Node to first male flower	–	7.8
Node to first female flower	5.7	13.7
Days to first male flower	–	34.9
Days to first female flower	37.6	54.5
Percentage maleness	0%	>95%
Percentage femaleness	100%	< 5%

Drop 8000 (Thermo Fisher Scientific, Waltham, MA), and equal amounts (1000 ng) from each of the 12 individuals constituting a bulk were pooled. Library construction and whole genome re-sequencing of the parents and the two bulks was performed as previously described (Abe et al., 2012; Takagi et al., 2013). Sequencing libraries with 250-600 bp insert sizes were prepared. Paired-end sequence reads (2 × 150 bp) of each library were obtained by Illumina HiSeq X. Adapter and low-quality sequences (<Q20) were trimmed by using fastp program (Chen et al., 2018).

QTL-seq analysis

Sequencing data of parental lines and pooled F_2 individuals were applied to QTL-seq pipeline developed by Takagi et al. (2013) for identifying location of gynoecious QTL. Briefly, in the pipeline, clean reads of PDM (monoecious parent) were aligned to reference genome sequence of OHB3-1 (Matsumura et al., 2020), and SNPs were called. By replacing these SNPs, PDM genome sequence file was developed. Thereafter, PVGy-201 (gynoecious parent) reads were mapped against the PDM genome sequence and SNPs between parental lines were defined. Subsequently, short reads of G-bulk and M-bulk were similarly aligned to PDM genome sequence. For each identified SNP locus between parental lines, SNP index was calculated as an allele frequency based on the sequence reads showing maternal or paternal allele. The SNP index is conferred as 0 if the entire short reads contain the PDM allele, while the SNP-index is 1 if all the short reads represent the PVGy-201-type allele. For clarifying SNP loci linked to gynnoecy, Δ (SNP-index) was then calculated in each locus by subtracting the SNP-index values between M-bulk and G-bulk, and sliding-window (10kb window size) of Δ (SNP-index) values were plotted for visualizing QTL region in the genome.

Annotation of variants in the QTL region

For annotation of variants located in the candidate QTL region, PVGy-201-unique variants were extracted by reference mapping of sequence reads of both parental lines and applied to SnpEff software (version 5.0e, (Cingolani et al., 2012) with previously predicted gene models of reference genome (Matsumura et al., 2020).

Promoter sequence variation analysis of putative candidate genes

To analyses the sequence variation, the sequence of the 12 putative candidate genes from the reference genome OHB3-1 (Matsumura et al., 2020) were extracted from NCBI and the promoter regions were identified through extracting flanking sequence regions for all the desired genes (if gene is forward strand, we take –60 for 5'upstream and +10 for 3'downstream and -10 for upstream and +60 3' downstream for reverse strand gene. The corresponding genes were identified from the both parental assemblies, PDM and PVGy-201 through Blast (https://blast.ncbi.nlm.nih.gov) and the promoter regions of all the genes in both parental assemblies were extracted and analysed for any sequence variation in the parental lines.

Results

Inheritance of the gynoecious sex expression

To determine the inheritance pattern of gynoecy, two parental lines PVGy-201 and PDM exhibiting contrasting sex expression, its F_1 and F_2 population were phenotyped for sex expression. PVGy-201 is a gynoecious line produce only female flowers while PDM is monoeious line bearing both male and female flowers separately with predominant maleness (> 95% male flowers). The F_1 plants expressed monoecious phenotype and in F_2, out of 147 plants, 104 were monoecious and 43 plants were gynoecious which was fit to a segregation ratio of 3:1. The phenotypic expression in F_1 indicates the recessive nature of gynoecy and segregation pattern in F_2 suggest monogenic recessive inheritance of gynoecy in the genotype, PVGy-201.

Whole genome re-sequencing and mapping of reads

Genomic DNA of two parental line (PVGY-201 and PDM) and two extreme bulks (G- bulk and M- bulk) were subjected to whole genome re-sequencing. Two extreme bulks were prepared based on the phenotype data of F_2 population. G-bulk and M-bulk each contained 12 plants with high female and male flower production, respectively. Illumina high-throughput sequencing generated 70.4 million and 154.08 million paired- end short reads (150 bp x 2) from PVGy-201 and PDM, respectively (Table 2). For two extreme bulks, 100.48 and 129.20 million short reads for G-bulk and M-bulk, respectively, from F_2 population were obtained. Quality filtering of these reads was carried out and 97-98% of clean reads were employed for further analysis.

QTL-seq analysis

The two parental lines (PVGy-201 and PDM) and two extreme pools, G-bulk and M-bulk from the F_2 population were paired-end

TABLE 2 Summary of whole genome re-sequencing data used for QTL-seq analysis in the present study.

Genotype	Number of plants for pooling	Sequence data obtained (GB)	Number of raw reads	Number of clean reads (%[a])
PVGy-201	-	10.56	70,445,450	68,707,610 (97.5)
PDM	-	23.11	154,084,080	151,581,250 (98.4)
G-bulk	**12**	15.07	100,480,892	98,278,696 (97.8)
M-bulk	**12**	19.38	129,201,904	126,867,286 (98.2)

[a]Percentage represents the ratio of number of filtered reads from number of raw reads by fastp program.

(150 bp) sequenced with an Illumina HiSeq platform. In total 10.56, 23.11, 15.07 and 19.38 Gb of clean reads from PVGy-201 (30× depth coverage), PDM (67.97× depth coverage), G-bulk (44× depth coverage), and M-bulk (57× depth coverage) were generated, respectively. These short reads were aligned to the "OHB3-1" reference genome for SNP calling.

Using clean sequence reads, QTL-seq analysis for gynoecy was carried out. As the reference genome, pseudomolecule of OHB3-1 genome sequence was employed and 66,661 SNP loci were defined as homozygous qualified SNPs as markers. In the QTL-seq, allele frequency in each bulk sample at each SNP locus was calculated as the SNP-index. Also, for avoiding false positive signals, ΔSNP-index was employed as the difference of SNP-index between two bulk samples. In the present analysis, neutral SNPs in gynoecy were expected to show ΔSNP-index between 0.5 and -0.5 (Supplementary Table 1). Its value of the SNPs linked to gynoecious QTL should distribute around 1 or -1. According to sliding window plots of ΔSNP-index in each chromosome, only the region between 23.46 Mb and 24.7Mb on chromosome 1 exhibited significant unequal contributions (Figure 2). In QTL-Seq analysis the DNA samples of progenies of mapping population showing extreme phenotypic values are bulked and subjected to whole genome re-sequencing. We expect the bulked DNA to contain genomes from both parents in a 1:1 ratio for the majority of genomic regions. However, unequal representation of the genomes from the two parents is observed in the genomic regions harboring QTL for the phenotypic difference between "gynoecy" and "monoecy" bulks (Takagi et al., 2013). Thus, the G-bulk mainly had PVGy1-type genomic segments in the 23.46 Mb and 24.7Mb region of chromosome 1, whereas M-bulk had PDM-type genome in the same region, indicating that there is a major QTL differentiating PVGy1 and PDM located at this genomic region. Therefore, 1.31 Mb region on the chromosome 1 is considered as the candidate QTL associated with gynoecy in bitter gourd.

Variants annotation and identification of candidate gene in in gynoecious locus

In the 1.31 Mb genomic region in chromosome 1 identified through the QTL-seq, the possible candidate genes with variant responsible for gynoecious sex expression in bitter gourd was explored. Therefore, PVGy-201-specific sequence variants in this region were selected and their effects to the structure of encoded genes (proteins) were estimated as annotation of variants. According to reference mapping of PDM and PVGy-201 reads and variant calling, 293,467 PVGy-201 unique variants, including SNPs and Indels identified. By using snpEff program and predicted gene model of the reference genome, impact of variants to the encoded proteins was predicted. After excluding variants located in intronic region and causing synonymous change, non-synonymous variants with high effect (frame-shift or stop-gained) and moderate effects (missense) on genes were selected. In the identified QTL region, a total of 1019 homozygous variants were identified between PVGy1 and PDM genomes. A total of 71 non-synonymous variants (SNPs and Indels) were identified, out of which 11 variants (7 Indels, 4 SNPs) were classified as high impact variants with frame shift/stop gain effect. Among the remaining 61 moderate impact variants (3 Indels, 58 SNPs) majority were missense variants. Variant annotation identified that, 71 non-synonymous variants were located in/associated with 41 protein coding genes (Supplementary Tables 1, 2).

Functional annotation of candidate genes

Genes harbouring these non-synonymous variants were annotated by BLASTx (Altschul et al., 1990) against the non-redundant protein database (http://www.uniprot.org/). Gene ontology classification revealed these 41 genes were mainly associated with biological processes (Figure 3), such as regulation of DNA-templated transcription, protein modification process, generation of precursor metabolites and energy, trans membrane transport, reproductive process, anatomical structure development; and molecular function, such as, transcription regulator activity, transporter activity, transferase activity, ATP-dependent activity etc. (Figure 4).

The biological processes of the 41 candidate genes were retrieved from the available bitter gourd genome data and Uniprot database. Among the 41 genes, 12 genes seemed to be related with flower development and sex expression, genes associated with male and female gametophyte development, male fertility restoration and phytohormone (auxin and ethylene) biosynthesis and signalling genes (Supplementary Table 3).

In the QTL region, 6 genes namely, Ethylene-responsive transcription factor 12 (Gene- LOC111025114), Copper-transporting ATPase RAN1(Gene-LOC111015725), Auxin response factor 6 (Gene-LOC111015731), CBL-interacting serine/threonine-protein kinase 23 (Gene- LOC111015768), LOB domain-containing protein 36-like (Gene- LOC111015703 and ABC transporter C family member 2

10.3389/fpls.2023.1071648

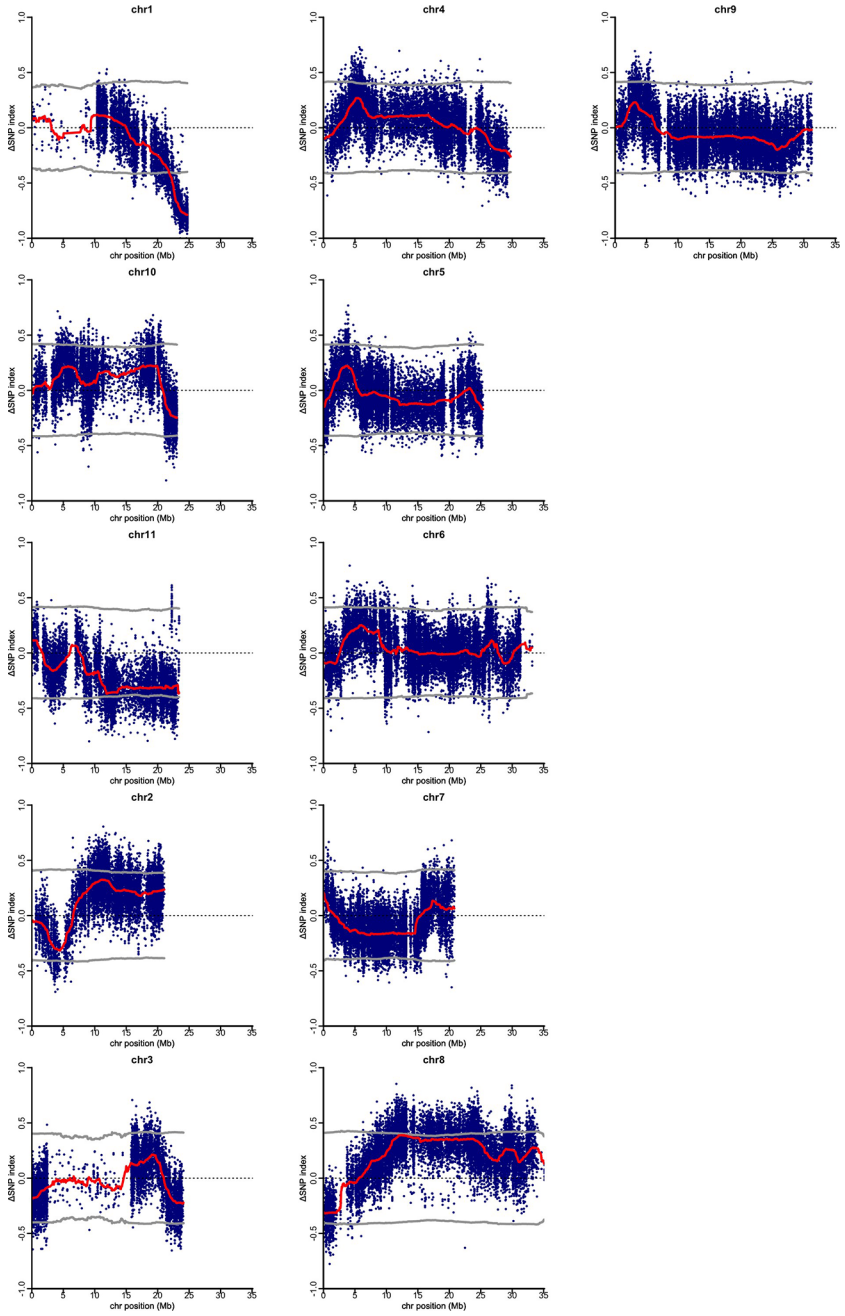


FIGURE 2
Quantitative trait loci (QTL)-seq identifies significant QTL on chromosome 1, for gynious sex expression in bitter gourd. ΔSNP (M-Bulk SNP Index–G-Bulk SNP Index) plotted against the physical position based *Momordica charantia* 11 chromosomes. The dark red line represents a sliding window of 2 Mb moving 500 kb intervals.

(Gene - LOC111015826) associated with development of gametophyte in association with important phytohomenes were identified (Table 3). Besides, another set of genes associated with male female gametophyte development and fertility restoration were identifies with possible role in determination of sex expression in bitter gourd. They were DEAD-box ATP-dependent RNA helicase 1 isoform X2 (Gene - LOC111015817), Polygalacturonase QRT3-like isoform X2 (Gene - LOC111015845), Protein CHROMATIN REMODELING 4 isoform X1 (Gene -LOC111015742), Pentatricopeptide repeat-containing protein (Gene - LOC111015813 and LOC111018538), putative F-box/LRR-repeat protein 23 (Gene -LOC111015857) and DNA replication licensing factor MCM6 isoform (Gene - LOC111015792).

Promoter sequence variation analysis of putative candidate genes

The sequence variation in the promoter region of the 12 putative candidate genes located in the QTL region were analyzed. The four genes LOC111015768, LOC111015725,

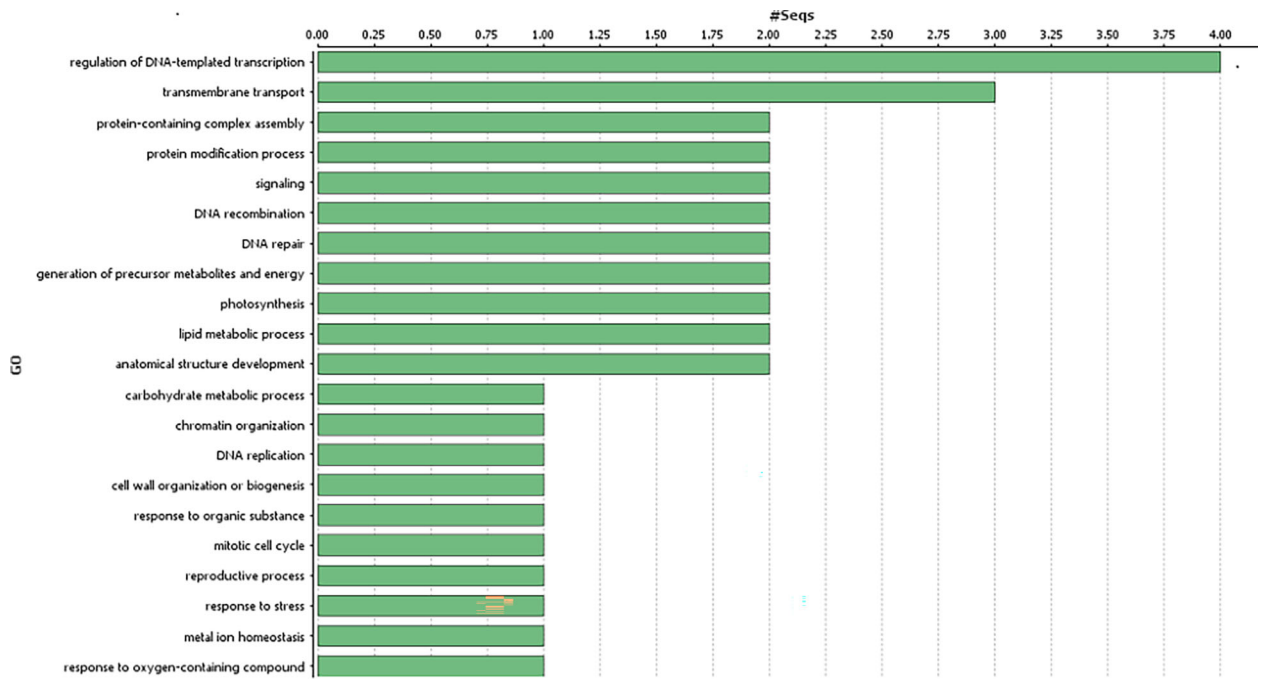


FIGURE 3
Gene ontology analysis of the identified genes in the genomic region associated with gynoecious sex expression in bitter gourd (Biological processes of the candidate genes associated with Non-synonymous variants).

LOC111015857 and LOC111015742 did not exhibit any promoter region variation and rest eight putative candidates namely, ABC transporter C family member 2-like, Polygalacturonase QRT3-like isoform X2, DNA replication licensing factor MCM6 isoform X1, Auxin response factor 6, Pentatricopeptide repeat-containing protein, LOB domain-containing protein 36-like, Ethylene-responsive transcription factor 12-like, DEAD-box ATP-dependent RNA helicase 1 isoform X2 shown sequence variation in the promoter region (Supplementary Table 4).

Discussion

Marker-assisted selection (MAS) is a powerful tool for accelerated breeding program that is quickly replacing tedious, expensive and time-consuming traditional phenotype-based breeding methods (Pandurangan et al., 2022) and applied widely in cucumber (Dey et al., 2020). QTL analysis through construction of high-density linkage map is a fundamental approach for molecular dissection of quantitative traits. Several multi-locus dominant DNA markers including RAPD (Dey et al., 2006; Paul et al., 2010), ISSR (Singh et al., 2007), and AFLP (Gaikwad et al., 2008) have been reported for genetic study of bitter gourd. However, in bitter gourd there is scarcity of genetic markers in public domain for the construction of a genetic map and marker-assisted selection (Tang et al., 2007).

Swift advancement in high-throughput sequencing methods and bioinformatics tools made detection of genome wide genetic polymorphism precise, quick and cheaper (Salvi and Tuberosa, 2005). Therefore, rapid identification of candidate genomic regions associated with a trait of interest through the "QTL-seq" method has been successfully applied in number of crops (Lu et al., 2014; Ruangrak et al., 2018; Shrestha et al., 2021). It combines advantages of bulk segregant analysis (BSA) and whole-genome re-sequencing for efficient genetic mapping (Takagi et al., 2013). Recently, QTL-seq is used widely and preferred over other traditional QTL mapping strategies. It is applied for genetic analysis of various economic traits in cucurbits such as early flowering (Lu et al., 2014), fruit length (Wei et al., 2016), subgynoecy (Win et al., 2019) in cucumber, heat tolerance in bottle gourd (Song et al., 2020) and mosaic resistance in zucchini (Shrestha et al., 2021). However, till now there is no study focusing on the successful application of QTL-seq for trait discovery in bitter gourd is reported.

Genetic analysis of gynoecious trait in bitter gourd genotype, PVGy-201

In bitter gourd, F_1 hybrids are preferred due to earliness, high yield, quality and tolerance to biotic and abiotic stresses (Behera et al., 2009; Dey et al., 2012). However, due to predominance of monoecy sex condition in bitter gourd, manual bagging and hand pollination is practiced for hybrid seed production which is costly and labour-intensive. On the other hand, use of gynoecious lines that produces only female flower as female parent can economise hybrid seed production with increased seed yield and hybrid genetic purity (Dey et al., 2010). Identification of gene(s) controlling gynoecism and (or) tightly linked markers would ease the identification gynoecious lines and hence their utilization in breeding programme. Furthermore, tightly linked marker can

10.3389/fpls.2023.1071648

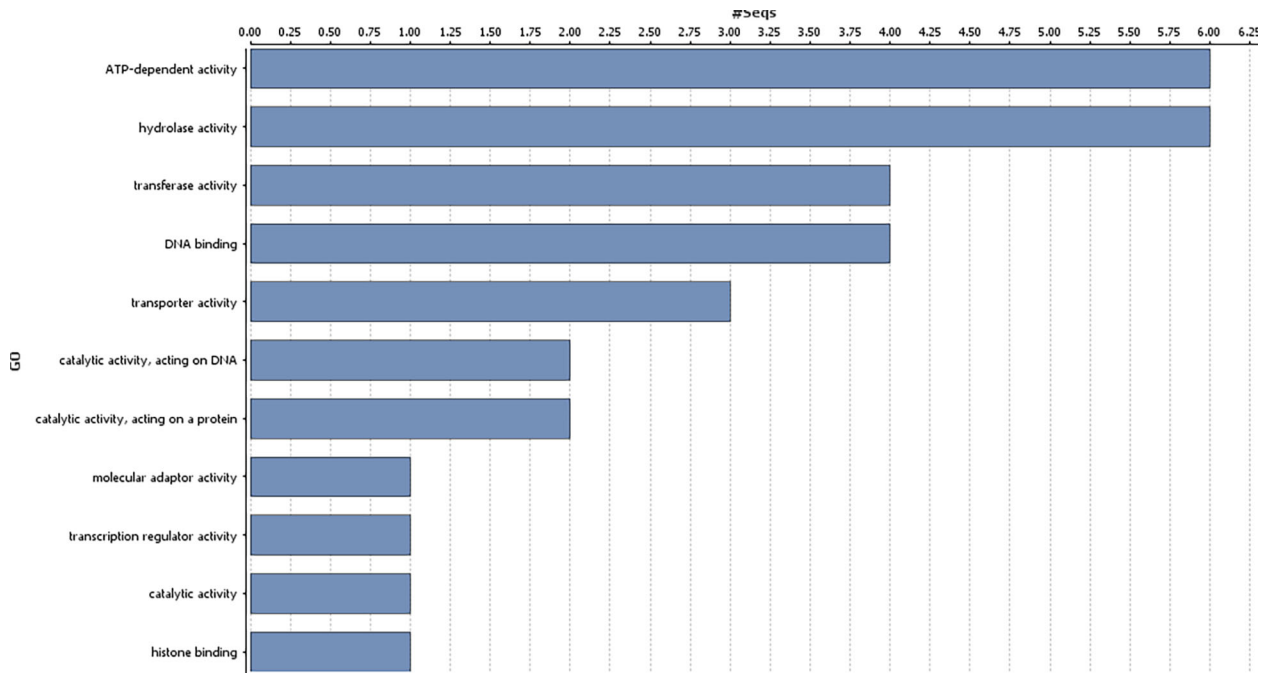


FIGURE 4
Gene ontology analysis of the identified genes in the genomic region associated with gynoecious sex expression in bitter gourd (Molecular functions of the candidate genes associated with Non-synonymous variants).

TABLE 3 List of possible candidate gened associated with gynoecious sex expression in bitter gourd.

Sl no.	Candidate genes	Gene function
1	DEAD-box ATP-dependent RNA helicase 1 isoform X2	RNA metabolism (Li et al., 2011), male gametophyte (Li et al., 2011) and female gametophyte (Li et al., 2011) development
2	CBL-interacting serine/threonine-protein kinase 23	Sex differentiation by altering ethylene biosynthesis and brassinosteroid signaling (Pawełkowicz et al., 2012)
3	Pentatricopeptide repeat-containing protein At1g09900	Carpel development and male fertility restoration in CMS (Igarashi et al., 2016; Liu et al., 2016)
4	Auxin response factor 6	Feminizing effect in sex differentiation by promoting ethylene biosynthesis (Mibus & Tatlioglu, 2004; Liu et al., 2015; Niu et al., 2022)
5	LOB domain-containing protein 36-like	Pollen development (Xu et al., 2016), role in sex differentiation through altering brassinosteroid accumulation (Bell et al., 2012)
6	F-box/LRR-repeat protein 23	Pollen production and male fertility restoration (Gusti et al., 2009; Han et al., 2021)
7	ABC transporter C family member 2-like	Auxin transportation (Geisler et al., 2017; Rea, 2007), pollen development (Yadav et al., 2014; Chen et al., 2017)
8	Polygalacturonase QRT3-like isoform X2	Pollen wall development and pollen maturation (Rhee et al., 2003; Lyu et al., 2015), male fertility restoration (Shi et al., 2021).
9	Copper-transporting ATPase RAN1	Sex differentiation by activation of ethylene receptors in ethylene signal transduction pathway (Binder et al., 2010; Hirayama, N. et al., 1999)
10	Protein CHROMATIN REMODELING 4 isoform X1	Female gametophyte development (Huanca-Mamani et al., 2005) and stamen filament elongation (Zhao et al., 2021).
11	DNA replication licensing factor MCM6 isoform X1	Role in sex differentiation though alternating ethylene biosynthesis (Street et al., 2015; Wang et al., 2019)
12	Ethylene-responsive transcription factor 12-like	Sex differentiation by transcriptional regulation of ethylene biosynthesis genes (Zhang et al., 2009; Li et al., 2016).

help in quick transfer of the gynoecious trait to the desirable parental background. Most genetic studies have reported the single recessive gene (*gy-1*) control of gynoecism (Ram et al., 2006; Behera et al., 2009; Matsumura et al., 2014), whereas involvement of two pairs of genes was reported recently by Cui et al. (2018).

Identification of genomic region associated with gynoecious trait through QTL-seq

In the current study, QTL- seq analysis through comparison of Δ SNP-index graph of G-bulk and M-bulk, 1.31 mb region on chromosome 1 (spanning between 23.46 Mb to 24.7Mb) was identified as the candidate genomic region associated with gynoecious sex expression in bitter gourd. Earlier, using RAD (restriction-associated DNA) based genetic maps in bitter gourd Matsumura et al. (2014) identified a SNP marker, GTFL-1 linked to the gynoecious locus at a distance of 5.46 cM. Later on, Cui et al. (2018) identified QTLs for gynoecy and female flower number using genotyping by sequencing of $F_{2:3}$ population of bitter gourd. Recently, a total of 22 QTLs for four sex expression-related traits namely gynoecy, sex ratio, node and days at first female flower appearance were mapped on 20 Linkage groups (Rao et al., 2018). This study identified a gynoecious (*gy-1*) locus flanked by markers TP_54865 and TP_54890 on LG 12 at a distance of 3.04 cM to TP_54890. In the QTL region identified in present study total of 71 non-synonymous variants (high and moderate impact) associated with 41 protein-coding genes were found. Among the 41 genes, 12 genes seemed to be related to flower development and sex expression; genes associated with male and female gametophyte development, male fertility restoration and phytohormone (Auxin and Ethylene) signalling genes.

Identification candidate genes associated with ethylene biosynthesis

Similar to other cucurbits, in bitter gourd also ethylene is known to play key role in gynoecious sex expression (Matsumura et al., 2014). Analysis of draft genome (monoecious inbred line, OHB3-1) sequence of bitter gourd revealed the presence of orthologous sequences of major ACC synthase genes in the genome. *MOMC3_649* in bitter gourd was presumed to be an ortholog of *CmACS11* (female flower determinant in melon) and two proteins, *MOMC46_189* and *MOMC518_1* were similar to *CmACS-7* that control unisexual flower development in melon (Urasaki et al., 2017). These findings suggest that the sex determination of *M. charantia* is similar to that of *Cucumis melo* and *Cucumis sativus*, which is under the control of ethylene biosynthesis pathways.

Ethylene is considered as a key regulator of sex expression across members of *Cucurbitcea* family (Yin and Quinn, 1995; Boualem et al., 2015). "One hormone hypothesis" explains the dual role of ethylene in deciding sexual morph of individual flower, inhibition of maleness and promotion of femaleness (Yin and Quinn, 1995). Ethylene biosynthesis involves series of enzymatic reactions involving activity of 1-aminocyclopropane-1-carboxylic acid (ACC) synthase (ACS) and 1-aminocyclopropane-1-carboxylic acid oxidase (ACO) (Adams and Yang, 1979; Yang and Hoffman, 1984). Locus *'a'* in melon and *M* locus in cucumber are orthologs, encoding the rate-limiting enzyme in ethylene biosynthesis namely, *CmACS-7* and *CsACS-2*, respectively. In cucumber, locus *A* encodes *CsACS11* a member of ACS gene family and gynoecy governing *F* locus, codes for a duplicated ACS gene *CsACS1G* (Trebitsh et al., 1997; Mibus and Tatlioglu, 2004). Interestingly, exogenous ethylene application is involved in up-regulation of ethylene biosynthesis genes *CsACS7/CmACS7 and CsACS11/CmACS11* in cucumber and melon, respectively (Switzenberg et al., 2014; Tao et al., 2018). To date, except for *WIP1* all the genes reported in cucumber and melon are involved in ethylene biosynthesis, this clearly establishes the significance of ethylene in sex expression. Furthermore, in cucumber, application of exogenous ethylene, ethylene releasing agent (ethephon), or the ethylene precursor ACC promotes the formation of female flowers in monoecious plants (Yamasaki et al., 2003), while interference with ethylene synthesis (with aminoethoxyvinyl-glycine; AVG) or signalling (with $AgNO_3$) induces male flowers in gynoecious plants (Takahashi and Jaffe, 1984).

Ethylene-responsive transcription factor (ERFs) proteins transcriptionally regulate ethylene-responsive genes *via* interaction with *cis*-acting elements or DRE/CRT motif located in the promoter region (Zhang et al., 2009; Li et al., 2016). In cucumber, ethylene response factor, *CsERF110/CmERF110* and *CsERF31* are known to transcriptionally regulate ethylene biosynthesis genes by directly binding to promoters of *CsACS11/CmACS11* and *CsACS2*, respectively (Pan et al., 2018; Tao et al., 2018). These results provide compelling evidence that ERF proteins, which control the transcription of the ACS and ACO genes, play vital role in ethylene biosynthesis in plants. The ethylene receptors *ETR1* and *ETR2* also play an important role in sex determination of cucurbits. The ethylene-insensitive mutant's *etr1a* and *etr2b* of *Cucurbita pepo* both disrupt female flower development (converting monoecy into andromonoecy) and significantly increase the number of male flowers in the plant. This probably indicates that *ETR1* and *ETR2* are able to integrate the two ethylene biosynthesis pathways, perceiving and signalling the ethylene produced by *ACS2/7* as well as that produced by *ACS11* and *ACO2* (García et al., 2020).

Copper-transporting ATPase RAN1 (Gene-LOC111015725) identified in the QTL-region is another critical gene associated with ethylene response. *RAN1* is reported to be essential factor for male flower development in fig and cucumber (Mori et al., 2017; Terefe, 2005). Further, Terefe (2005) indicated that the *CsRAN1* gene is probably linked to the determining A/a gene in cucumber. *RESPONSE TO ANTAGONIST1 (RAN1)*, which encodes copper-transporting *ATPase* enzyme crucial in the first step of ethylene perception (Woeste and Kieber, 2000). RAN1 transports copper ions from the cytoplasm to the Golgi apparatus and plays a vital role in the biogenesis and activation of the ethylene receptors, *ETR1* (ETHYLENE RESISTANCE 1), *ERS1* (ETHYLENE RESPONSE SENSOR 1), *ETR2*, *EIN4* (ETHYLENE INSENSITIVE 4), and *ERS2* in plants (Binder et al., 2010). In an active ethylene signal-

transduction process, the expression of *ACS11* relieves the inhibitory effect of *WIP1* on *ACS2* (Martin et al., 2009; Hu et al., 2017). *ACS2* promotes ethylene synthesis through positive feedback, increases cellular ethylene levels which in turn promotes pistil formation and inhibits stamen development. A mutated *ran1* allows *WIP1* expression through reduced sensitivity of ethylene synthesized by *ACS11* or *ACS7*, leading to stamen formation or male flower induction.

In plants, auxin is known to enhance endogenous ethylene production by promoting the expression of *ACS* genes and *ERF* genes (Trebitsh et al., 1987; El-Sharkawy et al., 2014). For instance, in *Arabidopsis* exogenous auxin treatment can significantly increase the expression of *AtACS4* and induce more ethylene (Stepanova et al., 2007). In cucurbits, auxin along with ethylene plays vital role in flower development and sex determination (Rudich et al., 1972; Friedlander et al. (1977). Auxin exhibits feminizing effect, evident from transformation of male flower buds into female flower buds (Galun, 1962) and increased female flower rate in cucumber (Takahashi and Jaffe, 1984) upon exogenous IAA treatment. Exogenous IAA treatment resulted in up-regulation of ethylene biosynthesis-related genes, ACC synthases (*CsACS1*, *CsACS2* and *CsACS11*) and ACC oxidases (*CsACO1*, *CsACO3*, and *CsACO4*) involved in sex determination in cucumber (Niu et al., 2022). Furthermore, presence of potential auxin-responsive elements (AREs) in the promoter region of *CsACS1* and *CsACS1G* demonstrated the mutual cross-talk in between the auxin and ethylene in the development of female flowers in cucumber (Mibus and Tatlioglu, 2004; Knopf and Trebitsh, 2006). Hence, auxin possibly affects sex determination in cucurbits *via* ethylene promoting ethylene production by enhancing the expression of ethylene biosynthesis and signaling genes (Takahashi and Jaffe, 1984; Trebitsh et al., 1987). Auxin Response Factors (ARFs) are crucial for the growth of pistils in Japanese Apricot (Song et al., 2015) and several auxin response elements (*CpARFs*) are reported to be involved in increased expression of ethylene signalling (*CpETR*) and biosynthesis (*CpACS*, *CpACO*) genes (Liu et al., 2015). In recent study, in cucumber, exogenous IAA treatment increased the transcription of *ERF (CsESR2 and Csa4G630010)* genes and one auxin response factor, *ARF* gene (*Csa2G000030*), suggesting that they may play regulatory roles in this crosstalk (Niu et al., 2022). Two auxin response factors *CsARF13* and *CsARF17* act as upstream regulators of *Cucumber MADS-box 1* (*CUM1)* (Ran et al., 2018), AG homolog in cucumber which expresses specifically in the stamens and carpels (Perl-Treves et al., 1998). *AGAMOUS (AG)* is a MADS-box gene that determines stamen and carpel development in *Arabidopsis*.

Expression of plant MCMs (Mini-chromosome maintenance protein complex) is greatly spread during the entire cycle (Huang et al., 2003). The function of plant MCMs is to prevent extra rounds of DNA replication (Brasil et al., 2017). Gene expression analysis in gynoecious and weak gynoecious cucumber identified correlated expression of many cell cycle pathway genes and ethylene related genes (Wang et al., 2019). An ethylene biosynthesis gene gyneocy determinant gene *CsACS1* (*G*), two ERFs (*CsERF12* and *CsERF118*) and two Ethylene receptors, *CsETR1* and *CsETR2* exhibited consistent expression pattern with cell cycle pathway and the gene (*Cs-MCM6*, *Cs-MCM2*, *Cs-CDC45*, *Cs-CDC20*, and *Cs-Dpri*) involved in the, *CsACS1* (*G*). Interestingly, ethylene also known to regulate cell cycle to inhibit plant growth during environmental stress conditions (Street et al., 2015). Integrating the results of these studies, it seems that the cell cycle genes may be involved in sex differentiation of cucumber initiated by ethylene, so there is a regulation relationship between cell cycle genes (*Cs-MCM6*, *Cs-MCM2*, *Cs-CDC45*, *Cs-CDC20*,and *Cs-Dpri*) and ethylene related genes (*CsERFs*, *CsETRs*, *CsACS1(G)*, *CsACO1* and *CsACO3).*

Identification of other important possible candidate genes

In *Cucumis melo*, dominant allele *Gy/gy* gene can be correlated with the putative serine/threonine kinase gene *CsPSTK1* (Pawełkowicz et al., 2012). Dominant *Gy* allele inhibits *CsPSTK1* gene which in turn negatively affects ethylene biosynthesis. On the other hand, when recessive *gy* is present, the inhibition is removed and the *CsPSTK1* gene has a positive effect on ethylene levels and femaleness is promoted (Pawełkowicz et al., 2012). A correlation exists between *BAK1*, a receptor in the brassinosteroids (BR) signalling pathway, and *CsPSTK1*, which suggests the involvement of *CsPSTK1* in BR signalling (Pawełkowicz et al., 2012). BR phytohormone indirectly take part in cucumber sex determination which increases the number of female flowers through promoting of ethylene production (Papadopoulou and Grumet, 2005; Wu et al., 2010). Auxin response factors (ARFs) and serine/threonine protein kinases were among the 54 genes linked to plant hormone signal transduction which shown differentially expression in male and female floral in the dioecious cucurbit ivy gourd *Coccinia grandis* (Mohanty et al., 2017).

In cucumber LOB protein (encoded by *Cucsa.098680*), contributes to pollen development, as reported by Xu et al. (2016). LOB protein negatively regulates the accumulation of brassinosteroids (BR) (Bell et al., 2012) a phytohormone which indirectly take part in cucumber sex determination, through promotion of ethylene biosynthesis (Papadopoulou and Grumet, 2005; Wu et al., 2010). The ABC transporter family genes are associated with trans-membrane transport of diverse substrates (e.g., lipids, heavy metal ions, sugars, amino acids, peptides, and secondary metabolites) and/or regulating other transporters (Jasinski et al., 2003; Rea, 2007). In *Arabidopsis* ABC transporters genes (*AtPGP1* and *AtPGP19*) are known to regulate auxin transportation (Rea, 2007). ABC transporters genes are also associated with pollen grain development in *Arabidopsis* (*AtABCG1* and *AtABCG16*) (Yadav et al., 2014) and pineapple (*AcABCG38)* (Chen et al., 2017). ABC transporters exhibited differential expression patterns in male, female, and hermaphroditic plants (Pawełkowicz et al., 2019).

RNA helicases are adenosine tri-phosphatases that unwind the secondary structures of RNAs and are required in almost every aspect of RNA metabolism (Liu et al., 2010). Programmed cell death (PCD) in tapetum degeneration is critical for development of male gametophytes in flowering plants. In rice, two putative DEAD-box ATP-dependent RNA helicases (encoded by *AIP1* and *AIP2*) are

involved in tapetum degeneration during pollen development (Li et al., 2011). Furthermore, in *Arabidospis* DEAD/DEAH-box helicases were specifically enriched in the megaspore mother cell and a DEAD-box RNA helicase (encoded by *SWA3*) is known to be involved in female gametogenesis (Liu et al., 2010). These studies imply the role of DEAD/DEAH-box helicases in male and female gametogenesis. Polygalacturonase (PG) is a pectin-digesting enzyme involved in numerous plant developmental processes and is described to be of critical importance in pollen wall development. The *QRT3* gene encodes a divergent class of polygalacturonase (PG) that is transiently expressed in tapetal cells and reported to participate in the pollen maturation through tetrad pectin wall degradation and pollen wall formation (Rhee et al., 2003). *BoMF25* in *Brassica oleracea*, a homologous gene of *At4g35670* is known to encode PG that exhibits pollen specific expression and found to be essential for pollen wall development (Lyu et al., 2015). In *Arabidopsis thaliana* "*res2*" locus which encode QRT3, is associated with restoration of thermo/photoperiod-sensitive genic male sterility (P/TGMS) (Shi et al., 2021). Chromatin remodelling proteins are involved in various biological processes in eukaryotes. In Arabidopsis, chromatin remodelling proteins (CHR11 and CHR17) (Huanca-Mamani et al., 2005) and RINGLET proteins (RLT1 and RLT2) (Li et al., 2012) were identified as the members of the, Imitation of Switch (ISWI) complex. In *Arabidopsis ISWI* complex has role in female gametophyte development (Huanca-Mamani et al., 2005) and in stamen filament elongation by regulating Jasmonic acid (JA) biosynthesis (Zhao et al., 2021).

F-Box LRR is a large subfamily of the plant F-box family known to mediate target protein degradation in response to developmental and hormonal signals (Kuroda et al., 2002). In *Arabidopsis* an F-box/LRR-repeat protein similar to gene04153 is required for pollen mitosis II (Gusti et al., 2009). The homolog of this gene was found in the candidate QTL region associated with male sterility in straw berry *Fregeria vesca*. ssp. *bracteata* (Tennessen et al., 2013). In wheat an F-box/LRR-repeat protein (encoded by TraesCS1B01G085600) was known to be associated with male fertility restoration in TGMS line-YS3038 (Han et al., 2021). Earlier Mutation of gene encoding F-box/LRR (FBL) in *Arabidopsis* affected the fertility of the male gametes due to obstruction in transformation process of microspores from the uninucleate to the binucleate stages.

Promoter sequence variation analysis of putative candidate genes

In cucurbits various sex forms are produced either due mutations in the coding region of the candidate genes and also due to mutation/variation in promoter regions of the candidate gene or even due to the copy number variation of candidate gene (Trebitsh et al., 1997; Mibus and Tatlioglu, 2004; Boualem et al., 2008; Zhang et al., 2015; Tao et al., 2018). In cucumber gynoecy is determined by the copy number variation (CNV)-based, dominant, and dosage-dependent *femaleness* (*F*) locus. Gynoecious plants contained three genes: *CsACS1, CsACS1G*, and *CsMYB*, of which *CsACS1G* is a duplication of *CsACS1* and loss of *CsACS1G* leads monoecy (Trebitsh et al., 1997; Mibus and Tatlioglu, 2004; Zhang et al., 2015). In melon a conserved mutation in the coding region CmACS-7 led to andromonoecious sex form (Boualem et al., 2008). Furthermore, Sex regulation in cucurbits is also due to transcriptional regulation of sex determining genes by various transcription factors that interacts with the with regulatory elements located in the promoters of ethylene biosynthesis and signaling genes (Zhang et al., 2009; Li et al., 2016). For instance in *cucumis* through a conserved mechanism, *CsERF110* and *CmERF110* respond to ethylene signaling, mediating ethylene-regulated transcription of *CsACS11* and *CmACS11* in cucumber and melon, respectively (Tao et al., 2018). These studies strongly indicates that the ethylene biosynthesis gene expressions are modified at transcriptional level by binding of regulatory proteins to promoter region of *ACS* and *ACO* genes. So, in bitter gourd also variation in promoter region might change the sex form. In the current study eight of the twelve putative candidate genes namely, ABC transporter C family member 2-like, Polygalacturonase QRT3-like isoform X2, DNA replication licensing factor MCM6 isoform X1, Auxin response factor 6, Pentatricopeptide repeat-containing protein, LOB domain-containing protein 36-like, Ethylene-responsive transcription factor 12-like, DEAD-box ATP-dependent RNA helicase 1 isoform X2 exhibited sequence variation in the promoter region. Therefore, genes are the most potential candidates determining gynoeciuous sex form in the bitter gourd genotype, DBGy-1. These findings suggests that the transcriptional regulation of sex determination genes play a major role in gynoecy expression in bitter gourd.

Large number of variants identified in the QTL-region will enable to develop molecular markers and fine mapping of the gynoecious sex expression in bitter gourd. The set of the possible candidate genes identified in the study with possible role in sex regulation will be instrumental for future study in bitter gourd and their functional analysis across different plant species.

Conclusion

Gynoecious sex expression is an extremely important trait to facilitate economic hybrid seed production in cucurbits. The present study involving and F_2 progenies of PVGy-201 × Pusa Do Mousami revealed that the gynoecious sex expression in the genotype, PVGy-201 is controlled by a single recessive gene. In the chromosome 1, 1.31 Mb regions was identified to be associated with gynoecious sex expression. A large number of variants were identified in the QTL-region which will be instrumental in fine mapping of gynoecious trait. Among the identified genes in the QTL-region, Ethylene-responsive transcription factor 12, Auxin response factor 6, Copper-transporting ATPase RAN1, CBL-interacting serine/threonine-protein kinase 23, ABC transporter C family member 2, DEAD-box ATP-dependent RNA helicase 1 isoform X2, Polygalacturonase QRT3-like isoform X2, Protein CHROMATIN REMODELING 4 were identified as possible candidate genes associated with gynoecious sex expression in bitter gourd because of their role in development of male and female gametophytes in number of crops. The findings in the study provides insight about sex expression in bitter gourd and will facilitate fine mapping and more precise identification of candidate genes through fine mapping and functional validation

of the identified genes. The present study provides insight into the genetic and molecular basis of gynoecious sex expression in bitter gourd.

Data availability statement

The data presented in the study are deposited in the NCBI repository, accession number PRJNA884851.

Author contributions

Conceived theme of the study and designed experiment: TB. Data curation: HM, AS, SD, MI, SJ, TB. Investigation: VD, SD, KK, Boopalakrishnan G. Resources: SD, TB, VC, AM. Supervision: TB, AR, HM, SD, AM, RE. Visualization: TB, SD, AM, GJ. Writing original draft: VD, IP, TB, SD, SJ. Review and editing: TB, SD, SJ, MI, GJ. A All authors contributed to the article and approved the submitted version.

Funding

The research work was funded by the NAHEP-CAAST programme of Indian Council of Agricultural Research (ICAR).

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, VD.

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.2023.1071648/full#supplementary-material

SUPPLEMENTARY TABLE 1
List of moderate and high effect variants (SNP/INDELs) located in QTL region associated with gynoecious sex expression in bitter gourd.

SUPPLEMENTARY TABLE 2
Functional annotation of candidate genes present in the identified QTL regions associated with gynoecious se expression in bitter gourd.

SUPPLEMENTARY TABLE 3
Effect of non- synonymous variants on candidate gene function/ protein change.

SUPPLEMENTARY TABLE 4
Promoter sequence variation analysis of putative candidate genes.

References

Abe, A., Kosugi, S., Yoshida, K., Natsume, S., Takagi, H., Kanzaki, H., et al. (2012). Genome sequencing reveals agronomically important loci in rice using MutMap. *Nat. Biotechnol.* 30 (2), 174–178. doi: 10.1038/nbt.2095

Adams, D., and Yang, S. (1979). Ethylene biosynthesis. identification of 1-aminocyclopropane-1-carboxylic acid as an intermediate in the conversion of methionine to ethylene. *Proc. Natl. Acad. Sci.* 76 (1), 170–174. doi: 10.1073/pnas.76.1.170

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215 (3), 403–410. doi: 10.1016/S0022-2836(05)80360-2

Atsmon, D., and Tabbak, C. (1979). Comparative effects of gibberellin, silver nitrate and aminoethoxyvinyl glycine on sexual tendency and ethylene evolution in the cucumber plant (Cucumis sativus l.). *Plant Cell Physiol.* 20 (8), 1547–1555. doi: 10.1093/oxfordjournals.pcp.a075957

Behera, T. (2004). Heterosis in bittergourd. *J. New Seeds* 6 (2-3), 217–221. doi: 10.1300/J153v06n02_11

Behera, T. K., Behera, S., Bharathi, L., John, K. J., Simon, P. W., and Staub, J. E. (2010). 2 bitter gourd. botany, horticulture, breeding. *Hortic. Rev.* 37, 101.

Behera, T., Dey, S., Munshi, A., Gaikwad, A. B., Pal, A., and Singh, I. (2009). Sex inheritance and development of gynoecious hybrids in bitter gourd (Momordica charantia l.). *Scientia Hortic.* 120 (1), 130–133. doi: 10.1016/j.scienta.2008.09.006

Behera, T., Dey, S., and Sirohi, P. (2006). DBGy-201 and DBGy-202. two gynoecious lines in bitter gourd (Momordica charantia l.) isolated from indigenous source. *Indian J. Genet. Plant Breed.* 66 (01), 61–62.

Bell, E. M., Husbands, A. Y., Yu, L., Jaganatha, V., Jablonska, B., Mangeon, A., et al. (2012). Arabidopsis lateral organ boundaries negatively regulates brassinosteroid accumulation to limit growth in organ boundaries. *Proc. Natl. Acad. Sci.* 109 (51), 21146–21151. doi: 10.1073/pnas.1210789109

Bhowmick, B. K., and Jha, S. (2015). Dynamics of sex expression and chromosome diversity in cucurbitaceae. a story in the making. *J. Genet.* 94 (4), 793–808. doi: 10.1007/s12041-015-0562-5

Binder, B. M., Rodríguez, F. I., and Bleecker, A. B. (2010). The copper transporter RAN1 is essential for biogenesis of ethylene receptors in arabidopsis. *J. Biol. Chem.* 285 (48), 37263–37270. doi: 10.1074/jbc.M110.170027

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, M.-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

Brasil, J. N., Costa, C. N. M., Cabral, L. M., Ferreira, P. C. G., and Hemerly, A. S. (2017). The plant cell cycle. pre-replication complex formation and controls. *Genet. Mol. Biol.* 40 (1 suppl 1), 276–291. doi: 10.1590/1678-4685-gmb-2016-0118

Chen, P., Li, Y., Zhao, L., Hou, Z., Yan, M., Hu, B., et al (2017). Genome-Wide Identification and Expression Profiling of ATP-Binding Cassette (ABC) Transporter Gene Family in Pineapple (Ananas comosus (L.) Merr.) Reveal the Role of AcABCG38 in Pollen Development. *Front. Plant Sci.* 18, 2150. doi: 10.3389/fpls.2017.02150

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, S., Zhou, Y., Chen, Y., and Gu, J. (2018). Fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34 (17), 884–890. doi: 10.1093/bioinformatics/bty560

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., et al. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. SNPs in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* 6 (2), 80–92. doi: 10.4161/fly.19695

Cui, J., Liu, J., Zhong, C., Hu, F., Dong, J., Cheng, J., et al. (2022). Fine-mapping and candidate gene analysis of the Mcgy1 locus responsible for gynoecy in bitter gourd (Momordica spp.). PREPRINT (Version 1) available at Research Square. doi: 10.21203/rs.3.rs-2103453/v1

Cui, J., Luo, S., Niu, Y., Huang, R., Wen, Q., Su, J., et al. (2018). A RAD-based genetic map for anchoring scaffold sequences and identifying QTLs in bitter gourd (Momordica charantia). *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.00477

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, 85. doi: 10.1038/s41438-020-0305-5

Daumann, M., Hickl, D., Zimmer, D., DeTar, R. A., Kunz, H. H., and Möhlmann, T. (2018). Characterization of filament-forming CTP synthases from arabidopsis thaliana. *Plant J.* 96 (2), 316–328. doi: 10.1111/tpj.14032

Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., and Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.* 12 (7), 499–510. doi: 10.1038/nrg3012

Dey, S. S., Behera, T. K., Bhatia, R., and Munshi, A. D. (2020). "Accelerated breeding in cucumber using genomic approaches," in *Accelerated plant breeding, volume 2, vegetable crops*. (Springer) doi: 10.1007/978-3-030-47298-6_10

Dey, S., Behera, T., Munshi, A., and Pal, A. (2010). Gynoecious inbred with better combining ability improves yield and earliness in bitter gourd (Momordica charantia l.). *Euphytica* 173 (1), 37–47. doi: 10.1007/s10681-009-0097-z

Dey, S. S., Behera, T. K., Munshi, A. D., Rakshit, S., and Bhatia, R. (2012). Utility of gynoecious sex form in heterosis breeding of bitter gourd and genetics of associated vegetative and flowering traits. *Indian J. Horticulture* 69 (4), 523–529.

Dey, S., Singh, A., Chandel, D., and Behera, T. (2006). Genetic diversity of bitter gourd (Momordica charantia l.) genotypes revealed by RAPD markers and agronomic traits. *Scientia Hortic.* 109 (1), 21–28. doi: 10.1016/J.SCIENTA.2006.03.006

Doyle, J. J., and Doyle, J. L. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bull.* 19 (1), 11–15.

El-Sharkawy, I., Sherif, S. M., Jones, B., Mila, I., Kumar, P. P., Bouzayen, M., et al. (2014). TIR1-like auxin-receptors are involved in the regulation of plum fruit development. *J. Exp. Bot.* 65 (18), 5205–5215. doi: 10.1093/jxb/eru279

Friedlander, M., Atsmon, D., and Galun, E. (1977). Sexual differentiation in cucumber. abscisic acid and gibberellic acid contents of various sex genotypes. *Plant Cell Physiol.* 18 (3), 681–691. doi: 10.1093/oxfordjournals.pcp.a075480

Gaikwad, A. B., Behera, T. K., Singh, A. K., Chandel, D., Karihaloo, J. L., and Staub, J. E. (2008). Amplified fragment length polymorphism analysis provides strategies for improvement of bitter gourd (Momordica charantia l.). *HortScience* 43 (1), 127–133. doi: 10.21273/HORTSCI.43.1.127

Galun, E. (1962). Study of the inheritance of sex expression in the cucumber. the interaction of major genes with modifying genetic and non-genetic factors. *Genetica* 32 (1), 134–163. doi: 10.1007/BF01816091

Gao, P., Sheng, Y., Luan, F., Ma, H., and Liu, S. (2015). RNA-Seq transcriptome profiling reveals differentially expressed genes involved in sex expression in melon. *Crop Sci.* 55 (4), 1686–1695. doi: 10.2135/cropsci2014.06.0444

García, A., Aguado, E., Martínez, C., Loska, D., Beltrán, S., Valenzuela, J. L., et al. (2020). The ethylene receptors CpETR1A and CpETR2B cooperate in the control of sex determination in cucurbita pepo. *J. Exp. Bot.* 71 (1), 154–167. doi: 10.1093/jxb/erz417

Geisler, M., Aryal, B., di Donato, M., and Hao, P. (2017). A Critical View on ABC Transporters and Their Interacting Partners in Auxin Transport. *Plant & cell physiology*. 58(10), 1601–1614. doi: 10.1093/pcp/pcx104

Girek, Z., Prodanovic, S., Zdravkovic, J., Zivanovic, T., Ugrinovic, M., and Zdravkovic, M. (2013). The effect of growth regulators on sex expression in melon (Cucumis melo l.). *Crop Breed. Appl. Biotechnol.* 13, 165–171. doi: 10.1590/S1984-70332013000300003

Gu, H. T., Wang, D. H., Li, X., He, C. X., Xu, Z. H., and Bai, S. N. (2011). Characterization of an ethylene-inducible, calcium-dependent nuclease that is differentially expressed in cucumber flower development. *New Phytol.* 192 (3), 590–600. doi: 10.1111/j.1469-8137.2011.03825.x

Gunnaiah, R., Vinod, M. S., Prasad, K., and Elangovan, M. (2014). Identification of candidate genes, governing gynoecy in bitter gourd (Momordica charantia l.) by in-silico gene expression analysis. *Int. J. Comput. Appl.* 2, 5–9.

Gusti, A., Baumberger, N., Nowack, M., Pusch, S., Eisler, H., Potuschak, T., et al. (2009). The arabidopsis thaliana f-box protein FBL17 is essential for progression through the second mitosis during pollen development. *PloS One* 4 (3), e4780. doi: 10.1371/journal.pone.0004780

Han, Y., Zhao, Y., Wang, H., Zhang, Y., Ding, Q., and Ma, L. (2021). Identification of ceRNA and candidate genes related to fertility conversion of TCMS line YS3038 in wheat. *Plant Physiol. Biochem.* 158, 190–207. doi: 10.1016/j.plaphy.2020.10.037

Heiser, C. B. (1979). Origins of some cultivated new world plants. *Annu. Rev. Ecol. Systematics* 10, 309–326. doi: 10.1146/annurev.es.10.110179.001521

Hossain, M. A., Islam, M., and Ali, M. (1996). Sexual crossing between two genetically female plants and sex genetics of kakrol (Momordica dioica roxb.). *Euphytica* 90 (1), 121–125. doi: 10.1007/BF00025168

Hirayama, T., Kieber, J. J., Hirayama, N., Kogan, M., Guzman, P., Nourizadeh, S., et al. (1999). RESPONSIVE-TO-ANTAGONIST1, a Menkes/Wilson disease-related copper transporter, is required for ethylene signaling in Arabidopsis. *Cell.* 97(3), 383–393. doi: 10.1016/s0092-8674 (00)80747-3

Hu, B., Li, D., Liu, X., Qi, J., Gao, D., Zhao, S., et al. (2017). Engineering non-transgenic gynoecious cucumber using an improved transformation protocol and optimized CRISPR/Cas9 system. *Mol. Plant* 10 (12), 1575–1578. doi: 10.1016/j.molp.2017.09.005

Huanca-Mamani, W., Garcia-Aguilar, M., León-Martínez, G., Grossniklaus, U., and Vielle-Calzada, J.-P. (2005). CHR11, a chromatin-remodeling factor essential for nuclear proliferation during female gametogenesis in arabidopsis thaliana. *Proc. Natl. Acad. Sci.* 102 (47), 17231–17236. doi: 10.1073/pnas.0508186102

Huang, X., Springer, P. S., and Kaloshian, I. (2003). Expression of the arabidopsis MCM gene PROLIFERA during root-knot and cyst nematode infection. *Phytopathology* 93 (1), 35–41. doi: 10.1094/phyto.2003.93.1.35

Igarashi, K., Kazama, T., and Toriyama, K. (2016). A Gene Encoding Pentatricopeptide Repeat Protein Partially Restores Fertility in RT98-Type Cytoplasmic Male-Sterile Rice. *Plant & cell physiology*. 57(10), 2187–2193. doi: 10.1093/pcp/pcw135

Illa-Berenguer, E., Van Houten, J., Huang, Z., and van der Knaap, E. (2015). Rapid and reliable identification of tomato fruit weight and locule number loci by QTL-seq. *Theor. Appl. Genet.* 128 (7), 1329–1342. doi: 10.1007/s00122-015-2509-x

Iwamoto, E., and Ishida, T. (2006). Development of gynoecious inbred line in balsam pear (Momordica charantia L.). *Hortic. Res. (Japan)*. doi: 10.2503/hrj.5.101

Jasinski, M., Ducos, E., Martinoia, E., and Boutry, M. (2003). The ATP-binding cassette transporters. structure, function, and gene family comparison between rice and arabidopsis. *Plant Physiol.* 131 (3), 1169–1177. doi: 10.1104/pp.102.014720

Kater, M. M., Franken, J., Carney, K. J., Colombo, L., and Angenent, G. C. (2001). Sex determination in the monoecious species cucumber is confined to specific floral whorls. *Plant Cell* 13 (3), 481–493. doi: 10.1105/tpc.13.3.481

Kaur, G., Pathak, M., Singla, D., Chhabra, G., Chhuneja, P., and Sarao, N. K. (2022). Quantitative trait loci mapping for earliness, fruit, and seed related traits using high density genotyping-by-Sequencing-Based genetic map in bitter gourd (Momordica charantia l.). *Frontiers in Plant Science* 12, 3410. doi: 10.3389/fpls.2021.799932

Knopf, R. R., and Trebitsh, T. (2006). The female-specific Cs-ACS1G gene of cucumber. a case of gene duplication and recombination between the non-sex-specific 1-aminocyclopropane-1-carboxylate synthase gene and a branched-chain amino acid transaminase gene. *Plant Cell Physiol.* 47 (9), 1217–1228. doi: 10.1093/pcp/pcj092

Kole, C., Matsumura, H., and Behera, T. K. (2020). *The bitter gourd genome* (Springer). doi: 10.1007/978-3-030-15062-4

Kubicki, B. (1969). Investigations on sex determination in cucumber (Cucumis sativus l.). IV. multiple alleles of locus acr. *Genetica Polonica* 10 (1-2), 23–68.

Kuroda, H., Takahashi, N., Shimada, H., Seki, M., Shinozaki, K., and Matsui, M. (2002). Classification and expression analysis of arabidopsis f-box-containing protein genes. *Plant Cell Physiol.* 43 (10), 1073–1085. doi: 10.1093/pcp/pcf151

Leclercq, J., Adams-Phillips, L. C., Zegzouti, H., Jones, B., Latché, A., Giovannoni, J. J., et al. (2002). LeCTR1, a tomato CTR1-like gene, demonstrates ethylene signaling ability in arabidopsis and novel expression patterns in tomato. *Plant Physiol.* 130 (3), 1132–1142. doi: 10.1104/pp.009415

Lee, H. Y., Chen, Y. C., Kieber, J. J., and Yoon, G. M. (2017). Regulation of the turnover of ACC synthases by phytohormones and heterodimerization in arabidopsis. *Plant J.* 91 (3), 491–504. doi: 10.1111/tpj.13585

Li, X., Gao, X., Wei, Y., Deng, L., Ouyang, Y., Chen, G., et al. (2011). Rice APOPTOSIS INHIBITOR5 coupled with two DEAD-box adenosine 5′-triphosphate-dependent RNA helicases regulates tapetum degeneration. *Plant Cell.* 23 (4), 1416–1434. doi: 10.1105/tpc.110.082636

Li, Z., Huang, S., Liu, S., Pan, J., Zhang, Z., Tao, Q., et al. (2009). Molecular isolation of the m gene suggests that a conserved-residue conversion induces the formation of bisexual flowers in cucumber plants. *Genetics* 182 (4), 1381–1385. doi: 10.1534/genetics.109.104737

Li, T., Jiang, Z., Zhang, L., Tan, D., Wei, Y., Yuan, H., et al. (2016). Apple (Malus domestica) md ERF 2 negatively affects ethylene biosynthesis during fruit ripening by suppressing md ACS 1 transcription. *Plant J.* 88 (5), 735–748. doi: 10.1111/tpj.13289

Li, G., Zhang, J., Li, J., Yang, Z., Huang, H., and Xu, L. (2012). Imitation switch chromatin remodeling factors and their interacting RINGLET proteins act together in controlling the plant vegetative phase in arabidopsis. *Plant J.* 72 (2), 261–270. doi: 10.1111/j.1365-313X.2012.05074.x

Liu, M., Shi, D. Q., Yuan, L., Liu, J., and Yang, W. C. (2010). SLOW WALKER3, encoding a putative DEAD-box RNA helicase, is essential for female gametogenesis in arabidopsis. *J. Integr. Plant Biol.* 52 (9), 817–828. doi: 10.1111/j.1744-7909.2010.00972.x

Liu, K., Yuan, C., Li, H., Lin, W., Yang, Y., Shen, C., et al. (2015). Genome-wide identification and characterization of auxin response factor (ARF) family genes related to flower and fruit development in papaya (Carica papaya l.). *BMC Genomics* 16 (1), 1–12. doi: 10.1186/s12864-015-2182-0

Liu, Z., Yang, Z., Wang, X., Li, K., An, H., Liu, J., et al. (2016). A Mitochondria-Targeted PPR Protein Restores pol Cytoplasmic Male Sterility by Reducing orf224 Transcript Levels in Oilseed Rape. *Molecular plant* 9(7), 1082–1084. doi: 10.1016/j.molp.2016.04.004

Lu, H., Lin, T., Klein, J., Wang, S., Qi, J., Zhou, Q., et al. (2014). QTL-seq identifies an early flowering QTL located near flowering locus T in cucumber. *Theor. Appl. Genet.* 127 (7), 1491–1499. doi: 10.1007/s00122-014-2313-z

Lyu, M., Liang, Y., Yu, Y., Ma, Z., Song, L., Yue, X., et al. (2015). Identification and expression analysis of BoMF25, a novel polygalacturonase gene involved in pollen development of brassica oleracea. *Plant Reprod.* 28 (2), 121–132. doi: 10.1007/s00497-015-0263-5

Mandal, N., Kumari, K., Kundu, A., Arora, A., Bhowmick, P., Iquebal, M., et al. (2022). Cross-talk between the cytokinin, auxin, and gibberellin regulatory networks in determining parthenocarpy in cucumber. *Front. Genet.* 13. doi: 10.3389/fgene.2022.957360

Martin, A., Troadec, C., Boualem, A., Rajab, M., Fernandez, R., Morin, H., et al. (2009). A transposon-induced epigenetic change leads to sex determination in melon. *Nature* 461 (7267), 1135–1138. doi: 10.1038/nature08498

Matsumura, H., Hsiao, M.-C., Lin, Y.-P., Toyoda, A., Taniai, N., Tarora, K., et al. (2020). Long-read bitter gourd (Momordica charantia) genome and the genomic architecture of nonclassic domestication. *Proc. Natl. Acad. Sci.* 117 (25), 14543–14551. doi: 10.1073/pnas.1921016117

Matsumura, H., Miyagi, N., Taniai, N., Fukushima, M., Tarora, K., Shudo, A., et al. (2014). Mapping of the gynoecy in bitter gourd (Momordica charantia) using RAD-seq analysis. *PloS One* 9 (1), e87138. doi: 10.1371/journal.pone.0087138

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

Mohanty, J. N., Nayak, S., Jha, S., and Joshi, R. K. (2017). Transcriptome profiling of the floral buds and discovery of genes related to sex-differentiation in the dioecious cucurbit coccinia grandis (L.) voigt. *Gene* 626, 395–406. doi: 10.1016/j.gene.2017.05.058

Mori, K., Shirasawa, K., Nogata, H., Hirata, C., Tashiro, K., Habu, T., et al. (2017). Identification of RAN1 orthologue associated with sex determination through whole genome sequencing analysis in fig (Ficus carica L.). *Scientific reports*. 7, 41124. doi: 10.1038/srep41124

Niu, H., Wang, H., Zhao, B., He, J., Yang, L., Ma, X., et al. (2022). Exogenous auxin-induced ENHANCER OF SHOOT REGENERATION 2 (ESR2) enhances femaleness of cucumber by activating the CsACS2 gene. *Horticulture Res.* 9, 85. doi: 10.1093/hr/uhab085

Pan, J., Wang, G., Wen, H., Du, H., Lian, H., He, H., et al. (2018). Differential gene expression caused by the f and m loci provides insight into ethylene-mediated female flower differentiation in cucumber. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.01091

Pandurangan, S., Workman, C., Nilsen, K., and Kumar, S. (2022). "Introduction to marker-assisted selection in wheat breeding," in *Accelerated breeding of cereal crops* (Springer), 77–117.

Papadopoulou, E., and Grumet, R. (2005). Brassinosteriod-induced femaleness in cucumber and relationship to ethylene production. *HortScience* 40 (6), 1763–1767. doi: 10.21273/HORTSCI.40.6.1763

Paul, A., Bandyopadhyay, S., Acharyya, P., and Raychaudhuri, S. (2010). Studies on genetic diversity of twelve accessions of momordica charantia l. using morphological, RAPD and SCAR markers. *Asian J. Plant Sci.* 9 (8), 471–478. doi: 10.3923/ajps.2010.471.478

Pawełkowicz, M., Osipowski, P., Wojcieszek, M., Wóycicki, R., Witkowicz, J., Hincha, D., et al. (2012). Identification and characterization of genes connected with flower morphogenesis in cucumber. *BioTechnologia* 93 (2), 123–134. doi: 10.5114/bta.2012.46588

Pawełkowicz, M., Pryszcz, L., Skarzyńska, A., Wóycicki, R. K., Posyniak, K., Rymuszka, J., et al. (2019). Comparative transcriptome analysis reveals new molecular pathways for cucumber genes related to sex determination. *Plant Reprod.* 32 (2), 193–216. doi: 10.1007/s00497-019-00362-z

Perl-Treves, R., Kahana, A., Rosenman, N., Xiang, Y., and Silberstein, L. (1998). Expression of multiple AGAMOUS-like genes in male and female flowers of cucumber (Cucumis sativus l.). *Plant Cell Physiol.* 39 (7), 701–710. doi: 10.1093/oxfordjournals.pcp.a029424

Peterson, C., and Anhder, L. D. (1960). Induction of staminate flowers on gynoecious cucumbers with gibberellin A3. *Science* 131 (3414), 1673–1674. doi: 10.1126/science.131.3414.1673

Poole, C. F., and Grimball, P. C. (1939). Inheritance of new sex forms in cucumis melo l. *J. Heredity* 30 (1), 21–25. doi: 10.1093/oxfordjournals.jhered.a104626

Ram, D., Kumar, S., Banerjee, M., and Kalloo, G. (2002). Occurrence, identification and preliminary charactrization of gynoecism in bitter gourd (Momordica charantia l.). *Indian Journal of Agricultural Science* 72 (6), 348–349. Available at: http://krishi.icar.gov.in/jspui/handle/123456789/51847.

Ram, D., Kumar, S., Singh, M., Rai, M., and Kalloo, G. (2006). Inheritance of gynoecism in bitter gourd (Momordica charantia l.). *J. Heredity* 97 (3), 294–295. doi: 10.1093/jhered/esj028

Ran, G., Xiaofeng, L., Wensheng, Z., Shuangshuang, Y., Linhan, S., Binning, W., et al. (2018). Functional characterization of the promoter and second intron of CUM1 during flower development in cucumber (Cucumis sativus l.). *Hortic. Plant J.* 4 (3), 103–110. doi: 10.1016/j.hpj.2018.03.004

Rao, P. G., Behera, T. K., Gaikwad, A. B., Munshi, A. D., Srivastava, A., Boopalakrishnan, G., et al (2021). Genetic analysis and QTL mapping of yield and fruit traits in bitter gourd (Momordica charantia L.). *Scientific reports* 11(1), 4109. doi: 10.1038/s41598-021-83548-8

Rao, G. P., Behera, T. K., Gaikwad, A. B., Munshi, A. D., Jat, G. S., and Boopalakrishnan, G. (2018). Mapping and QTL analysis of gynoecy and earliness in bitter gourd (Momordica charantia L.) Using genotyping-by-Sequencing (GBS) technology. *Front. Plant Sci.* 9, 1555. doi: 10.3389/fpls.2018.01555

Rea, P. A. (2007). Plant ATP-binding cassette transporters. *Annu. Rev. Plant Biol.* 58 (1), 347–375. doi: 10.1146/annurev.arplant.57.032905.105406

Rhee, S. Y., Osborne, E., Poindexter, P. D., and Somerville, C. R. (2003). Microspore separation in the quartet 3 mutants of arabidopsis is impaired by a defect in a developmentally regulated polygalacturonase required for pollen mother cell wall degradation. *Plant Physiol.* 133 (3), 1170–1180. doi: 10.1104/pp.103.028266

Robinson, R. W., and Decker-Walters, D. S. (1999). Cucurbits. CAB International Wallingford, Oxon (GB). 226

Robinson, R., and Decker-Walters, D. (1997). Cucurbits crop production science in horticulture series. Wallingford, UK: CAB International, 65–7.

Robinson, R., HM, M., TW, W., and GW, B. (1976). Genes of the cucurbitaceae. *HortScience*. 11 (6), 554–568. doi: 10.21273/HORTSCI.11.6.554

Ruangrak, E., Su, X., Huang, Z., Wang, X., Guo, Y., Du, Y., et al. (2018). Fine mapping of a major QTL controlling early flowering in tomato using QTL-seq. *Can. J. Plant Sci.* 98 (3), 672–682. doi: 10.1139/cjps-2016-0398

Rudich, J., Halevy, A., and Kedar, N. (1972). Ethylene evolution from cucumber plants as related to sex expression. *Plant Physiol.* 49 (6), 998–999. doi: 10.1104/pp.49.6.998

Salvi, S., and Tuberosa, R. (2005). To clone or not to clone plant QTLs. present and future challenges. *Trends Plant Sci.* 10 (6), 297–304. doi: 10.1016/j.tplants.2005.04.008

Shi, Q.-S., Lou, Y., Shen, S.-Y., Wang, S.-H., Zhou, L., Wang, J.-J., et al. (2021). A cellular mechanism underlying the restoration of thermo/photoperiod-sensitive genic male sterility. *Mol. Plant* 14 (12), 2104–2114. doi: 10.1016/j.molp.2021.08.019

Shifriss, O. (1961). Sex control in cucumbers. *J. Heredity* 52 (1), 5–12. doi: 10.1093/oxfordjournals.jhered.a107021

Shrestha, S., Michael, V., Fu, Y., and Meru, G. (2021). Genetic loci associated with resistance to zucchini yellow mosaic virus in squash. *Plants* 10 (9), 1935. doi: 10.3390/plants10091935

Singh, A., Behera, T., Chandel, D., Sharma, P., and Singh, N. (2007). Assessing genetic relationships among bitter gourd (Momordica charantia l.) accessions using inter-simple sequence repeat (ISSR) markers. *J. Hortic. Sci. Biotechnol.* 82 (2), 217–222. doi: 10.1080/14620316.2007.11512222

Song, J., Gao, Z., Huo, X., Sun, H., Xu, Y., Shi, T., et al. (2015). Genome-wide identification of the auxin response factor (ARF) gene family and expression analysis of its role associated with pistil development in Japanese apricot (Prunus mume sieb. et zucc). *Acta Physiologiae Plantarum* 37 (8), 1–13. doi: 10.1007/s11738-015-1882-z

Song, H., Huang, Y., and Gu, B. (2020). QTL-seq identifies quantitative trait loci of relative electrical conductivity associated with heat tolerance in bottle gourd (Lagenaria siceraria). *PloS One* 15 (11), e0227663. doi: 10.1371/journal.pone.0227663

Stepanova, A. N., Yun, J., Likhacheva, A. V., and Alonso, J. M. (2007). Multilevel interactions between ethylene and auxin in arabidopsis roots. *Plant Cell* 19 (7), 2169–2185. doi: 10.1105/tpc.107.052068

Street, I. H., Aman, S., Zubo, Y., Ramzan, A., Wang, X., Shakeel, S. N., et al. (2015). Ethylene inhibits cell proliferation of the arabidopsis root meristem. *Plant Physiol.* 169 (1), 338–350. doi: 10.1104/pp.15.00415

Switzenberg, J. A., Little, H. A., Hammar, S. A., and Grumet, R. (2014). Floral primordia-targeted ACS (1-aminocyclopropane-1-carboxylate synthase) expression in transgenic cucumis melo implicates fine tuning of ethylene production mediating unisexual flower development. *Planta* 240 (4), 797–808. doi: 10.1007/s00425-014-2118-y

Takagi, H., Abe, A., Yoshida, K., Kosugi, S., Natsume, S., Mitsuoka, C., et al. (2013). QTL-seq. rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations. *Plant J.* 74 (1), 174–183. doi: 10.1111/tpj.12105

Takahashi, H., and Jaffe, M. (1984). Further studies of auxin and ACC induced feminization in the cucumber plant using ethylene inhibitors. *Phyton* 44 (1), 81–86.

Tang, R., Gao, G., He, L., Han, Z., Shan, S., Zhong, R., et al. (2007). Genetic diversity in cultivated groundnut based on SSR markers. *J. Genet. Genomics* 34 (5), 449–459. doi: 10.1016/S1673-8527(07)60049-6

Tao, Q., Niu, H., Wang, Z., Zhang, W., Wang, H., Wang, S., et al. (2018). Ethylene responsive factor ERF110 mediates ethylene-regulated transcription of a sex determination-related orthologous gene in two cucumis species. *J. Exp. Bot.* 69 (12), 2953–2965. doi: 10.1093/jxb/ery128

Tennessen, J. A., Govindarajulu, R., Liston, A., and Ashman, T.-L. (2013). Targeted sequence capture provides insight into genome structure and genetics of Male sterility in a gynodioecious diploid strawberry, fragaria vesca ssp. bracteata (Rosaceae). *G3 Genes|Genomes|Genetics* 3 (8), 1341–1351. doi: 10.1534/g3.113.006288

Terefe, D. (2005). "Molecular genetic and physiological studies on the sex-determining m/m and a/a genes in cucumber (Cucumis sativus l.)," in *Doctoral dissertation* (Hannover: Universität).

Trebitsh, T., Rudich, J., and Riov, J. (1987). Auxin, biosynthesis of ethylene and sex expression in cucumber (Cucumis sativus). *Plant Growth Regul.* 5 (2), 105–113. doi: 10.1007/BF00024738

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

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

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, L., Ko, E. E., Tran, J., and Qiao, H. (2020). TREE1-EIN3–mediated transcriptional repression inhibits shoot growth in response to ethylene. *Proc. Natl. Acad. Sci.* 117 (46), 29178–29189. doi: 10.1073/pnas.2018735117

Wei, Q.-z., Fu, W.-y., Wang, Y.-z., Qin, X.-d., Wang, J., Li, J., et al. (2016). Rapid identification of fruit length loci in cucumber (Cucumis sativus l.) using next-generation sequencing (NGS)-based QTL analysis. *Sci. Rep.* 6 (1), 1–11. doi: 10.1038/srep27496

Win, K. T., Zhang, C., Silva, R. R., Lee, J. H., Kim, Y.-C., and Lee, S. (2019). Identification of quantitative trait loci governing subgynoecy in cucumber. *Theor. Appl. Genet.* 132 (5), 1505–1521. doi: 10.1007/s00122-019-03295-3

Woeste, K. E., and Kieber, J. J. (2000). A strong loss-of-function mutation in RAN1 results in constitutive activation of the ethylene response pathway as well as a rosette-lethal phenotype. *Plant Cell* 12 (3), 443–455. doi: 10.1105/tpc.12.3.443

Wu, T., Qin, Z., Zhou, X., Feng, Z., and Du, Y. (2010). Transcriptome profile analysis of floral sex determination in cucumber. *J. Plant Physiol.* 167 (11), 905–913. doi: 10.1016/j.jplph.2010.02.004

Xu, C., Luo, F., and Hochholdinger, F. (2016). LOB domain proteins: Beyond lateral organ boundaries. *Trends Plant Sci.* 21 (2), 159–167. doi: 10.1016/j.tplants.2015.10.010

Yadav, V., Molina, I., Ranathunge, K., Castillo, I. Q., Rothstein, S. J., and Reed, J. W. (2014). ABCG transporters are required for suberin and pollen wall extracellular barriers in arabidopsis. *Plant Cell* 26 (9), 3569–3588. doi: 10.1105/tpc.114.129049

Yamasaki, S., Fujii, N., and Takahashi, H. (2003). Photoperiodic regulation of CS-ACS2, CS-ACS4 and CS-ERS gene expression contributes to the femaleness of cucumber flowers through diurnal ethylene production under short-day conditions. *Plant Cell Environ.* 26 (4), 537–546. doi: 10.1046/j.1365-3040.2003.00984.x

Yang, S. F., and Hoffman, N. E. (1984). Ethylene biosynthesis and its regulation in higher plants. *Annu. Rev. Plant Physiol.* 35 (1), 155–189. doi: 10.1146/annurev.pp.35.060184.001103

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

Zhang, J., Guo, S., Ji, G., Zhao, H., Sun, H., Ren, Y., et al. (2020). A unique chromosome translocation disrupting ClWIP1 leads to gynoecy in watermelon. *Plant J.* 101 (2), 265–277. doi: 10.1111/tpj.14537

Zhang, J., Jianting, S., Gaojie, J., Zhang, H., Guoyi, G., Shaogui, G., et al. (2017). Modulation of sex expression in four forms of watermelon by gibberellin, ethephone and silver nitrate. *Hortic. Plant J.* 3 (3), 91–100. doi: 10.1016/j.hpj.2017.07.010

Zhang, Y., Liu, B., Yang, S., An, J., Chen, C., Zhang, X., et al. (2014). A cucumber DELLA homolog CsGAIP may inhibit staminate development through transcriptional repression of b class floral homeotic genes. *PloS One* 9 (3), e91804. doi: 10.1371/journal.pone.0091804

Zhang, Z., Mao, L., Chen, H., Bu, F., Li, G., Sun, J., et al. (2015). Genome-wide mapping of structural variations reveals a copy number variant that determines reproductive morphology in cucumber. *Plant Cell* 27 (6), 1595–1604. doi: 10.1105/tpc.114.135848

Zhang, Z., Zhang, H., Quan, R., Wang, X.-C., and Huang, R. (2009). Transcriptional regulation of the ethylene response factor LeERF2 in the expression of ethylene biosynthesis genes controls ethylene production in tomato and tobacco. *Plant Physiol.* 150 (1), 365–377. doi: 10.1104/pp.109.135830

Zhao, Y., Jiang, T., Li, L., Zhang, X., Yang, T., Liu, C., et al. (2021). The chromatin remodeling complex imitation of switch controls stamen filament elongation by promoting jasmonic acid biosynthesis in arabidopsis. *J. Genet. Genomics* 48 (2), 123–133. doi: 10.1016/j.jgg.2021.02.003

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 16 March 2023
DOI 10.3389/fpls.2023.1079221

OPEN ACCESS

EDITED BY
Nisha Singh,
Gujarat Biotechnology University, India

REVIEWED BY
Raj Kumar Joshi,
Rama Devi Women's University, India
Meng Wang,
Institute of Genetics and Developmental Biology (CAS), China

*CORRESPONDENCE
Mir Asif Iquebal
ma.iquebal@icar.gov.in

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

RECEIVED 25 October 2022
ACCEPTED 08 February 2023
PUBLISHED 16 March 2023

CITATION
Kumar B, Saha B, Jaiswal S, Angadi UB, Rai A and Iquebal MA (2023) Genome-wide identification and characterization of tissue-specific non-coding RNAs in black pepper (*Piper nigrum* L.). *Front. Plant Sci.* 14:1079221. doi: 10.3389/fpls.2023.1079221

Genome-wide identification and characterization of tissue-specific non-coding RNAs in black pepper (*Piper nigrum* L.)

Baibhav Kumar, Bibek Saha, Sarika Jaiswal, U. B. Angadi, Anil Rai and Mir Asif Iquebal*

Division of Agricultural Bioinformatics, Indian Council of Agricultural Research-Indian Agricultural Statistics Research Institute, New Delhi, India

Long non-coding RNAs (lncRNAs) and circular RNAs (circRNAs) are the two classes of non-coding RNAs (ncRNAs) present predominantly in plant cells and have various gene regulatory functions at pre- and post-transcriptional levels. Previously deemed as "junk", these ncRNAs have now been reported to be an important player in gene expression regulation, especially in stress conditions in many plant species. Black pepper, scientifically known as *Piper nigrum* L., despite being one of the most economically important spice crops, lacks studies related to these ncRNAs. From a panel of 53 RNA-Seq datasets of black pepper from six tissues, *namely*, flower, fruit, leaf, panicle, root, and stem of six black pepper cultivars, covering eight BioProjects across four countries, we identified and characterized a total of 6406 lncRNAs. Further downstream analysis inferred that these lncRNAs regulated 781 black pepper genes/gene products *via* miRNA–lncRNA–mRNA network interactions, thus working as competitive endogenous RNAs (ceRNAs). The interactions may be various mechanisms like miRNA-mediated gene silencing or lncRNAs acting as endogenous target mimics (eTMs) of the miRNAs. A total of 35 lncRNAs were also identified to be potential precursors of 94 miRNAs after being acted upon by endonucleases like Drosha and Dicer. Tissue-wise transcriptome analysis revealed 4621 circRNAs. Further, miRNA–circRNA–mRNA network analysis showed 432 circRNAs combining with 619 miRNAs and competing for the binding sites on 744 mRNAs in different black pepper tissues. These findings can help researchers to get a better insight to the yield regulation and responses to stress in black pepper in endeavor of higher production and improved breeding programs in black pepper varieties.

KEYWORDS

circular RNAs (circRNAs), competitive endogenous RNAs (ceRNAs), gene expression regulation, long non-coding RNAs (lncRNAs), spices

1 Introduction

Spices have been an important part of human food and nutrition for thousands of years. Recent studies confirm their diverse significant potentials against various diseases, in addition to their trivial nutritional benefits as functional foods. Black pepper is one of the most important spices containing *piperine* as the active ingredient. Black pepper also contains various volatile oils, oleoresins, and alkaloids, and has a role in suppressing tumor growth and chemoprevention (Butt et al., 2013). Black pepper consumption also improves nutrients absorption and gastrointestinal function. *Piperine*, the main component of black pepper, inhibits the differentiation of fat cells by lowering PPAR (peroxisome proliferator activated receptor) activity and reducing PPAR expression, which is a potential cure for disorders linked to obesity (Park et al., 2012; Negi et al., 2021).

Studies confirming the existence and important role of short (18-23 nucleotides) and long (>200 nucleotides) RNAs over the past two decades have completely changed the view of the cellular mechanism of gene expression. Non-coding regions of the genome (~98% in mammals) are now no longer considered "junk" and are believed to play important regulatory and tissue-specific expression roles (Azlan et al., 2019). LncRNAs are one such important class of ncRNAs with lengths >200 nucleotides and are predominantly found in the cell's nucleolus, nucleus, or cytoplasm. LncRNAs have little or no coding potential and may contain an ORF only by chance, but are similar to protein-coding mRNAs in aspects like they both are transcribed by RNA polymerase II, spliced and 5′ methyl Guanosine capped and 3′ poly-adenylated. These challenges make the identification and characterization task of lncRNAs a more tedious task (Mattick, 2004; Wang and Chang, 2011).

In a typical mammalian genome, approximately 4%–9% of the genome is transcribed into lncRNAs, which is more than the fraction of the genome that is transcribed into protein-coding mRNAs (~1%) (Amaral et al., 2010). LncRNAs are not actively expressed as in proteins but are involved in a lot of cellular activities. LncRNAs are spread throughout the genome which contains 98%–99% non-coding region and are labeled with respect to the genomic location they are transcribed from, *viz*., those arriving from the intergenic region are called intergenic lncRNAs; intronic lncRNAs are those derived purely from introns, while the exonic lncRNAs are derived from exons of protein-coding genes (Mercer et al., 2009). LncRNAs have been found to be involved in various cellular functions mainly regulating the expression of the genes in *cis*- or *trans*- of their origin. *cis*-Acting lncRNAs may alter the expression of neighboring genes by either blocking the formation of pre-initiation complex (PIC) by attaching to the promoter or interfering the transcription factors or *via* chromatin modifications. The best example of chromatin modification mediated lncRNA function is XIST (X inactive-specific transcript), which is a 19-kb-long human lncRNA which binds to the PRC2 (Polycomb Recessive Complex) to induce H3K27me3 histone modification which leads to transcriptional silencing of genes on the X chromosome. Also, in plants COLDAIR lncRNA (Cold Assisted Intronic Non-coding RNA) works as a necessary repressor of FLC (Flowering Locus C) during vernalization. HOTAIR (Hox Antisense Intergenic RNA) is a trans-acting lncRNA in humans that transports itself to other locations on different chromosomes and regulates the gene expression (Chen and Carmichael, 2010; Song et al., 2016). Apart from the extensive studies done in humans and other mammals, various plant lncRNAs have also been identified including *Arabidopsis thaliana* (Lu et al., 2017), wheat (Lu et al., 2020), rice (Wang et al., 2018; Zhou et al., 2021), maize (Li et al., 2014), tomato (Yang et al., 2019), cucumber (He et al., 2020; Dey et al., 2022), pearl millet (Kumar et al., 2022), and watermelon (Sun et al., 2020). LDMAR (Long Day Specific Male-sterility-associated RNA) which is a 1236 bp long lncRNA can regulate the male sterility sensitive to photoperiod in rice.

Apart from the linear lncRNAs, another important class of non-linear ncRNAs, called circRNAs have emerged, which are formed by the back splicing of 5′ terminus upstream exon of the pre-mRNA with the 3′ downstream exon (Lai et al., 2018). Evidently, circRNAs are more resistant to RNAase degradation due to lack of 5′ cap or 3′ tail-free ends. First reported in 1979 by Hsu, in HeLa cell lines, circRNAs have been studied in many species since then including rice (Lu et al., 2015), wheat (Ren et al., 2018), cucumber (Mu et al., 2016; Zhu et al., 2019), chickpea (Dasmandal et al., 2020) mango (Yang et al., 2022), and watermelon (Sun et al., 2020). Chu et al. (2018) showed that higher expression of the circRNAs has a down regulating effect on its parental genes. Cucumber circRNAs study showed that the circRNAs can also act as miRNA sponges and have a miRNA-circRNA-mRNA network relationship of gene regulation mechanism.

Black pepper (*Piper nigrum* L.) is one of the most widely grown and traded spices in the world and is recognized as the king of spices (Nair et al., 1993). Following the release of the reference genome of black pepper, a lot of studies related to genes of black pepper have been done but no study has been performed till now of the ncRNAs, except for the small miRNAs (Ding et al., 2021), where 128 mature miRNAs and their 1007 target mRNAs were identified. Our work is the first such study to unravel the characteristics and functional roles of larger ncRNAs in black pepper along with circRNAs. For this study, extensive retrieval of black pepper molecular data was made to fetch 53 raw RNA-Seq datasets comprising of >1.2 billion reads covering eight BioProjects and six tissues (flower, fruit, leaf, panicle, root and stem) from 6 black pepper cultivars across four countries, followed by reference transcript assembly for the identification of black pepper lncRNA and circRNAs. This study also aims at having an insight on the tissue-specific nature of lncRNAs and circRNAs and their relationship with miRNAs as competitive endogenous RNAs (ceRNAs), their functional roles in various pathways, development of a freely accessible web resource having the list of lncRNAs and circRNAs, which can be retrieved based on peptide length, sequence length and tissue-wise, and interaction between lncRNA-miRNA and miRNA-circRNA found in this study. This would be helpful to fellow researchers for augmenting related work in the crop.

2 Methods

2.1 Data collection

With the aim to perform a comprehensive study, a total of 53 raw RNA-Seq datasets (>1.2 billion reads) of black pepper were

downloaded from the NCBI database. The dataset covered eight BioProjects, seven institutes across four countries, and six tissues (flower, fruit, leaf, panicle, root, and stem) from six black pepper cultivars *viz.* Reyin-1, Bragantina, Thottumuriyan, IPN No. LK-0-WU-0014181, panniyur-1, and Genotype 4226 (Supplementary Table 1).

2.2 Data pre-processing and transcriptome assembly

Raw reads obtained from NCBI were subjected to quality check using FastQC tool ver. 0.11.8 (Andrews, 2010), which helps us visualize various quality parameters like per base sequence quality, per base sequence content, presence of adaptor sequences, etc. The Trimmomatic ver. 0.39 (Bolger et al., 2014) software was then used to trim out the probable contaminants like adaptor sequences and low-quality reads with a Phred score of less than 30. The reference genome and annotation file of black pepper were downloaded from the GCGI (Group of Cotton Genetic Improvement, http://cotton.hszau.edu.cn/EN/index.php) website (Hu et al., 2019). The HISAT2-build command from HISAT2 ver. 2.2.0 (Kim et al., 2015; Pertea et al., 2015; Pertea et al., 2016) software was used to index the reference genome with splice sites and exons information retrieved from the annotation file. Index files were then used for aligning individual clean reads. Sam files obtained after the reads alignment were converted into binary bam files using Samtools software ver. 1.9 (Li et al., 2009), then the transcriptome assembly of the individual bam files was performed using StringTie ver 2.1.4 (Pertea et al., 2015) software to give out gtf files for each of the reads. Individual files corresponding to each tissue were then merged using StringTie–merge to get a single gtf (Gene Transfer Format) file for each tissue.

2.3 Genome-wide identification of lncRNAs in black pepper

Identification of lncRNA candidate transcripts from the assembled transcripts involved various steps as shown in Figure 1. First the fasta sequences corresponding to each transcript in the merged assembly file were extracted from their respective reference genome fasta files using the gffread program ver 0.12.3. As lncRNAs are RNA transcripts longer than 200 base pairs, using in-house Perl scripts, transcripts shorter than 200 bp were removed. Studies have shown that lncRNAs in general have a lower quality shorter ORFs than the protein-coding mRNAs. ORFPredictor (Min et al., 2005)

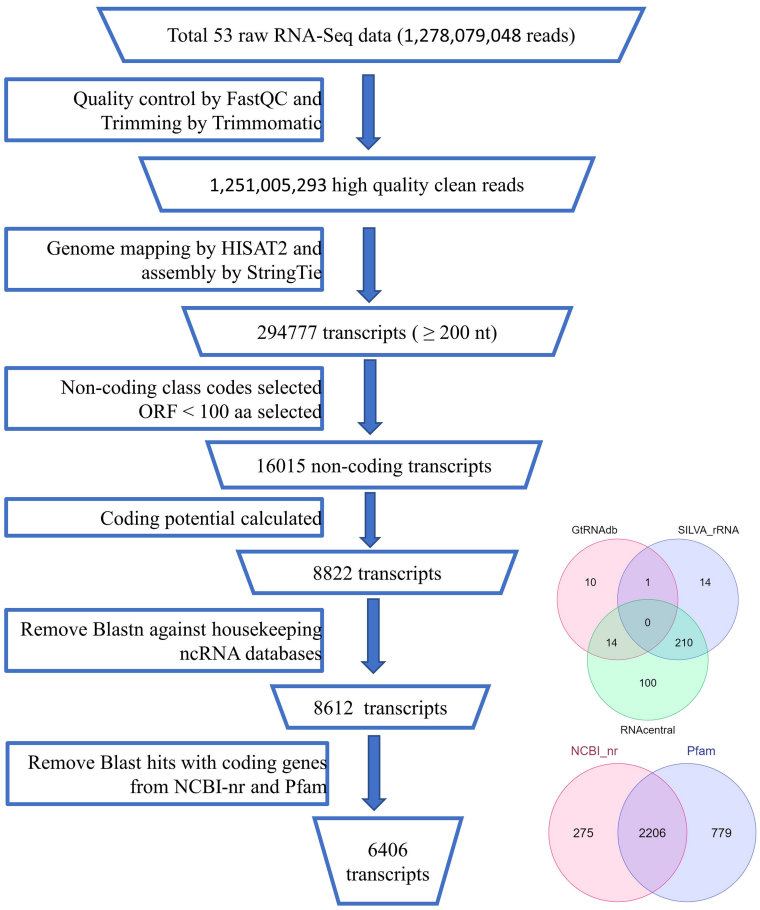


FIGURE 1
Pipeline of lncRNA identification.

was used to predict the ORFs in each transcript and those having ORFs longer than 300 nucleotides were removed. Binary classifiers like CPC2 ver 1.0.1 (Kang et al., 2017) and PLEK (Li et al., 2014) were used to classify the remaining sequences into coding or non-coding. Transcripts showing coding labels in all of them were discarded and kept as coding transcripts. The remaining transcripts may contain some housekeeping RNAs like rRNAs, tRNAs, snoRNAs, and other ncRNAs. So to get novel ncRNAs we did a BlastN search against the databases like SILVA database (https://www.arb-silva.de/download/arb-files/release138.1), GtRNAdb (http://gtrnadb.ucsc.edu/release18.1), and RNACentral (https://rnacentral.org/release16.0) and removed the transcripts showing >=95% identity (Sharma et al., 2017). Transcripts matching any reported proteins or protein families were identified and removed using Blast against NCBI-nr protein and Pfam (http://pfam.xfam.org/release33.1) databases (Altschul et al., 1990).

2.4 Conservation analysis by comparison with known plant lncRNAs in CANTATAdb

LncRNAs are poorly conserved compared to protein-coding mRNAs across species. Yet to check the conservativeness of black pepper's lncRNAs, reported lncRNAs of 38 plant species available at the CANTATAdb ver. 2.0 (http://cantata.amu.edu.pl/) (Szcześniak et al., 2019) database were downloaded and a local BlastN search was performed against our identified black pepper lncRNAs as a query with 10^{-20} e-value as the cutoff.

2.5 LncRNA and circRNA characterization and functional annotation

2.5.1 LncRNAs acting as a precursor of mature miRNAs

Plant cells contain small non-coding RNA transcripts miRNAs endogenously, which can bind to mRNAs and suppress their expression in the cell. These miRNAs are derived from longer primary miRNA transcripts which are converted to comparatively shorter pre-miRNAs (precursor) (Paraskevopoulou and Hatzigeorgiou, 2016). Finally, 18–24 nucleotide long mature miRNAs are produced by endonuclease action upon the pre-miRNAs. These lncRNAs can also act as a source for the biogenesis of mature miRNAs. Identified lncRNAs were matched against precursor miRNAs from the miRBase database using the BlastN software to find such black pepper lncRNAs which can act as a precursor miRNA.

2.5.2 Identification of lncRNAs acting as target mimics of miRNAs

To find out the miRNAs, which may use the identified black pepper lncRNAs as their target mimics, the psRNATarget (Dai et al., 2018) (http://plantgrn.noble.org/psRNATarget/) server was used with black pepper miRNAs (Ding et al., 2021) and lncRNAs as inputs in the options of small RNA sequences and target sequences. Matches with stringent criteria of expectation ≤2 and allowed maximum energy to unpair the target site (UPE) = 25 were considered significant for our study.

2.5.3 Identification of mRNA targets of identified miRNAs and their annotation

For creating the lncRNA-miRNA-mRNA network, mRNA targets of the identified miRNAs are to be found. PsRNATarget was run with identified miRNA sequences and black pepper cDNA sequences as input. Matches with expectation ≤2 and UPE ≥25 were considered significant. Identified mRNAs were annotated using OmicsBox (https://www.biobam.com/omicsbox/) software and GO terms obtained were used to functionally characterize the co-expressed lncRNAs. REVIGO (http://revigo.irb.hr/) server was used to further analyze the GO terms by summarizing the GO terms present and provides a graph based visualization of the GO terms.

2.5.4 Identification of black pepper circRNAs

CircRNA identification pipeline starts from obtaining the clean reads after trimming out the low-quality bases and adaptor sequences. Reads retained from each sample were aligned to the reference genome of black pepper by BWA software ver. 0.7.17 (BWA mem–T 19) after creating an index using BWA–index module. Aligned sam files corresponding to each tissue were merged using Samtools ver. 1.14. The merged alignment files of each tissue were then provided to the CIRI2 ver.2.0 (Gao et al., 2018) tool as input for circRNA prediction. Novel circRNA were identified by comparing the identified circRNA with plant circRNA in PlantcircBase database (http://ibi.zju.edu.cn/plantcircbase/, release 7) (Figure 2).

2.5.5 CircRNAs as miRNA sponges and miRNA-circRNA-mRNA relationship study

For understanding the relationship of the network between the miRNA, circRNA, and mRNA, previously identified and reported black pepper miRNAs by Ding et al. (2021) were collected and used. The circRNA targets of the miRNAs were found using TargetFinder software (Fahlgren and Carrington, 2010). The possible circRNA targets of the miRNAs were found by command line software TargetFinder. After this, the mRNA targets of the identified miRNAs were identified using the webserver psRNAtarget with miRNA sequences as small RNA sequences and CDS (coding sequence) of black pepper as the target sequences.

2.6 Black pepper lncRNA web-resource development

A web-resource, a Black pepper ncRNA database *BPncRDB* was created using the three-tier architecture, *viz*., client, middle, and database tiers that house all the results of this study related to the lncRNAs, circRNAs and their interactions with the miRNAs. The database was developed in MySQL database (https://www.mysql.com/) while the web-interface was prepared in PHP (https://www.php.net/) and HTML while designing was done using CSS and made dynamic using JavaScript. It was hosted on Apache

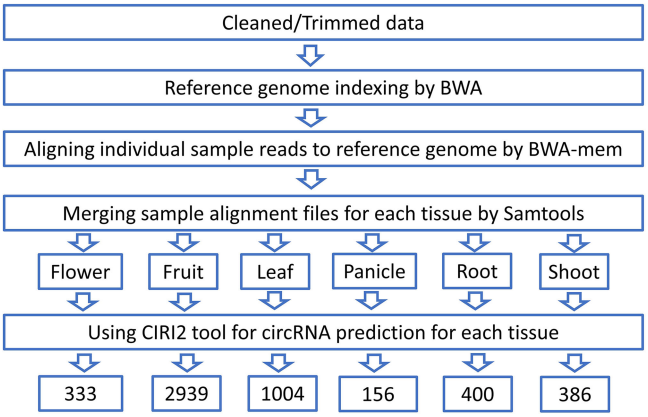


FIGURE 2
Schematic diagram of circRNA identification.

server (https://httpd.apache.org/). XAMPP framework was used to design and deploy the webpage. The user can retrieve data as: (a) request from user to webserver, (b) query sent to MySQL database, (c) response generated by database and sent to web-interface, and (d) response of web-server to user. *BPncRDB* includes information of tissue-specific lncRNAs and circRNAs and their relationship with miRNAs, interaction between lncRNA-miRNA and miRNA-circRNA, etc.

3 Results

3.1 Data pre-processing and transcriptome assembly

The library was sequenced using the Illumina HiSeq X platform and 1,278,079,048 raw reads in 16 samples were collected. After discarding the Illumina platform's adaptor sequences and other low-quality reads using the Trimmomatic software, we obtained 1,251,005,293 (97.88%) clean reads (Supplementary Table 2). The trimmed clean reads were aligned to black pepper's reference genome using HISAT2 and approximately 70%–94% of reads were mapped across the 53 samples (Supplementary Table 2). StringTie software was then used to assemble the mapped reads of individual samples with respect to the reference annotation file of black pepper. Individual assembly files for each tissue were merged using the StringTie-merge module in order to get tissue-wise lncRNAs, and 294,777 transcripts were obtained.

3.2 Identification of long non-coding RNAs in black pepper

A stringent filtering pipeline for the assembled transcripts was developed and used for the identification of those transcripts which fulfill the criteria of lncRNAs (Figure 1). GffCompare program was used to compare the four GTF files with an annotation file of black pepper and annotate the transcripts corresponding to their location on the genome with respect to the known genes (Pertea and Pertea, 2020). Out of the 15 class codes, i (intronic), u (intergenic), and x (natural antisense transcripts) represent the most probable non-coding transcripts and were selected for the downstream analysis, and the rest were discarded. Gffread was used to extract fasta sequences corresponding to the class codes and a total of 41090 sequences were found. In-house developed Perl scripts were used to remove sequences smaller than 200 nucleotides. To remove the potentially coding transcripts, ORFpredictor was used to find out the ORFs in each transcript and those having an ORF length >300 nucleotides were discarded. The coding probability of the remaining transcripts was calculated using CPC2 and PLEK software and taking the intersection of the results. Transcripts with CPC2 score >0.5 and predicted coding by PLEK were considered coding and discarded. Housekeeping RNAs were removed by BlastN against ncRNA databases, *viz*., Silvadb, gtRNAdb, and RNACentral and having percent identity >95%. Transcripts having similarity matching with any of the protein families in Pfam or genes in the NCBI-nr database were removed using BlastX (e-value 10^{-3}). Finally, we identified 6406 novel black pepper long non-coding RNAs in black pepper, out of which 1115, 2621, 2727, 828, 1214, and 1003 were expressed in flower, fruit, leaf, panicle, root, and stem tissues, respectively (Figure 3A; Supplementary Table 3).

3.3 Characterization of identified black pepper lncRNAs

The study presents the first ever identification and characterization of black pepper lncRNAs to understand the functional importance of lncRNAs *via* various mechanisms affecting the crop's yield and stress responses. In consistency with the previous lncRNA studies most (~80%) of the identified black pepper lncRNAs are intergenic in nature (Figure 3B). The length of the identified lncRNAs was distributed in the range of 200–10667 nucleotides which is much shorter than mRNAs which ranges up to 16.6 mega bases in length. More than 80% of the lncRNAs were shorter than 1500 nucleotides while only 10% were longer than 2000 nucleotides (Figure 3C). With

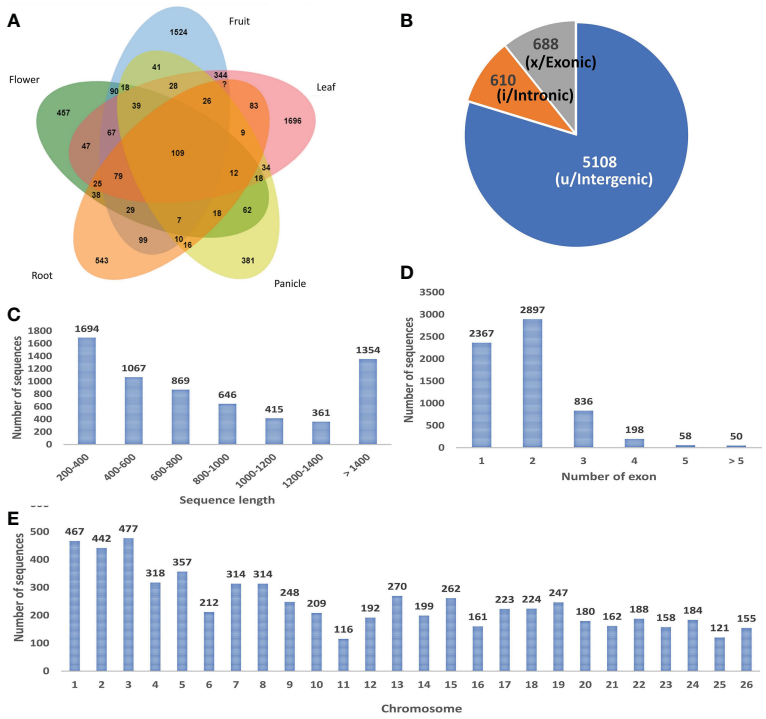


FIGURE 3
(A) Tissue-wise distribution of black pepper lncRNAs. **(B)** Classes of identified lncRNAs. **(C)** Length distribution of lncRNAs. **(D)** Exon distribution of lncRNAs. **(E)** Chromosome-wise distribution of identified lncRNAs.

an average of 1.88 exons per lncRNA, a major proportion (84.85%) of the lncRNAs in all six tissues were derived from 1 or 2 exons while only 1.68% were from ≥5 exons (Figure 3D). Chromosomal distribution of the identified lncRNA was depicted in (Figure 3E) and further visualized using the Circos software (Figure 4).

3.4 Conservation analysis of identified black pepper lncRNAs

To check on the conservation level of identified black pepper lncRNAs, BlastN against the previously known and reported lncRNAs of 38 plant species was performed with 10^{-20} as e-value cutoff. Only 45 high-confidence matches were found where 42 identified black pepper lncRNAs matched with 27 database lncRNAs from 13 plant species with a maximum of five matches to *Oryza barthii* and four matches to *Oryza rufipogon* and *Medicago truncatula* each (Supplementary Table 4). This result agrees with the literature suggesting the very poor conservation levels of lncRNAs compared to protein-coding mRNAs and are species- and tissue-specific.

3.5 Identified black pepper lncRNAs acting as miRNA precursors

lncRNAs are long RNAs present in the nucleus, nucleolus, and or cytoplasm and can function as the precursor for smaller ncRNAs such as snRNAs, snoRNAs, as well as the miRNAs (18–23 nts). To find out the black pepper lncRNAs which can possibly be precursors of known miRNAs, a BlastN search against the mirBase database was done (Supplementary Table 5). A total of 14 pre-miRNAs were matched >=90% with 36 unique lncRNAs, which suggests those lncRNAs can give rise to the mature miRNAs after being acted upon by nuclease enzymes like dicer and/or drosha. Figure 5A shows lncRNA TCONS_00294108 containing the precursor and mature sequences of miRNA vvi-miR156i.

3.6 Identification of lncRNAs and endogenous target mimics of miRNAs and analyzing lncRNA-miRNA-mRNA interaction network

Micro RNAs (miRNAs) are small ncRNAs (18–23 nts) which have the major function of mRNA expression regulation by binding the 3′ UTR of the protein coding mRNAs. The expression is suppressed or silenced depending upon the complementarity of the miRNA-mRNA binding. A full complementarity leads to mRNA degradation thus silencing while a partial binding decreases the mRNA expression level downregulating the genes. LncRNAs sometimes interfere in the process and act as a miRNA sponge and prohibit miRNA-mRNA binding. Identified lncRNAs and known plant miRNAs available at the psRNAtarget server were taken for this analysis. Identified black pepper lncRNAs were submitted to the psRNAtarget server as target sequences against the available plant miRNAs and run with default

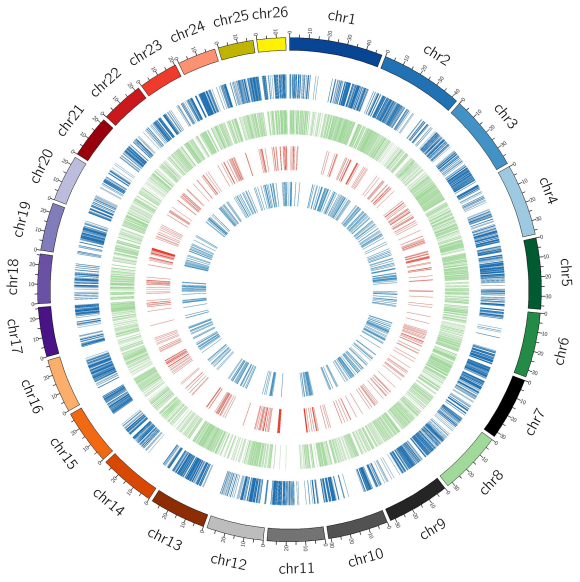


FIGURE 4
Black pepper circRNA/lncRNA shown chromosome wise. Outermost (dark blue) circle represents the black pepper circRNA, green circle represents intergenic lncRNAs, red circle represents intronic lncRNAs, and inner light blue circle represents antisense lncRNAs.

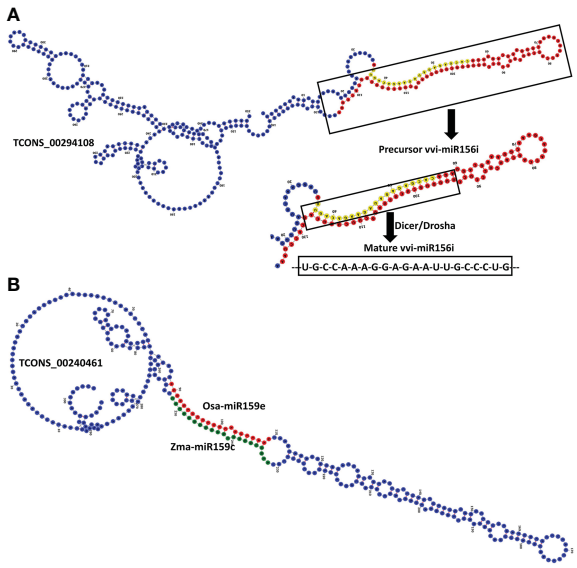


FIGURE 5
(A) RNAFold structure of lncRNA TCONS_00294108 showing the precursor (red) and mature (yellow) sequence of miRNA vvi-miR156i. **(B)** Structure of lncRNA TCONS_00240461 harboring the target sites of two miRNAs zma-miR159c and osa-miR159e.

parameters of max UPE 25 and expectation ≤2 (Supplementary Table 6); 1702 interactions with 1054 unique miRNAs and 396 unique lncRNAs were found (Figure 6). Figure 5B shows the RNAFold structure of lncRNA TCONS00240461 (blue) with the binding sites of miRNAs Osa-miR159e (red) and Zma-miE159c (green). Target mRNAs of the identified miRNAs were also found using psRNAtarget by submitting black pepper CDS (coding sequence) as target and identified 1054 miRNAs as small RNAs (Supplementary Table 7); 3274 total interactions were found between the black pepper genes and miRNAs suggesting the possible gene regulations involved. Individual lncRNA-miRNA and miRNA-mRNA networks were merged and visualized using the Cytoscape software (Figure 7A), where we can see various networks involving miRNA and target mRNAs which can potentially be meddled by lncRNA thus regulating the normal gene regulation mechanism. For instance, miRNA pta-miR156b meant to be targeting mRNA Pn2.1339 is also capable of targeting four lncRNAs *viz.* TCONS_00067586, TCONS_00076072, TCONS_00072122, and TCONS_00076071. (Figure 8A).

3.7 Identification of black pepper circRNAs

After removing the low-quality reads from the raw sequences, the clean reads (>1.2 billion reads) were aligned to the reference

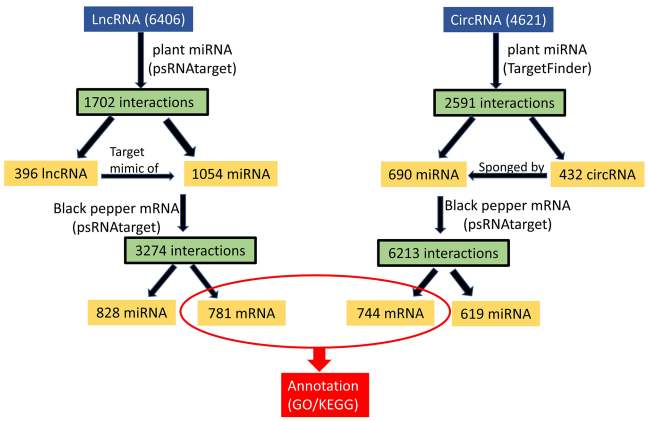


FIGURE 6
ceRNA analysis of lncRNA/circRNA.

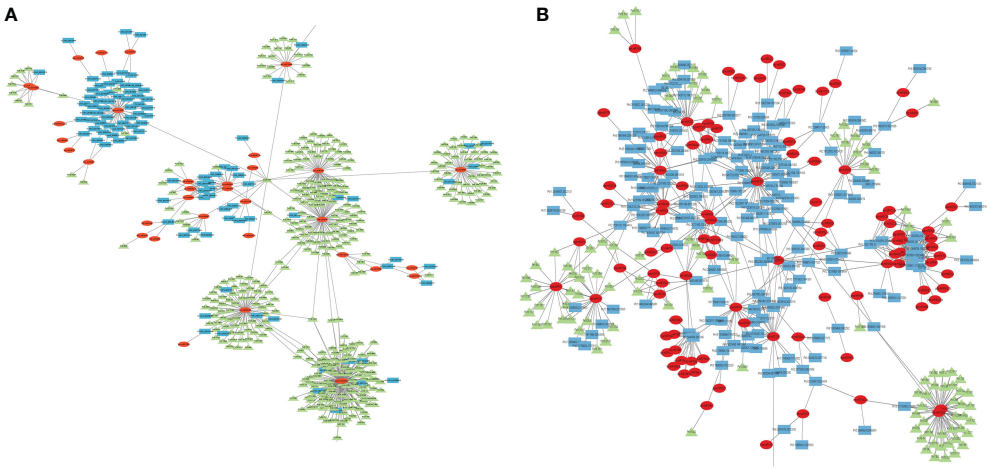


FIGURE 7
Cytoscape diagram showing the ceRNA network relationship between **(A)** lncRNA–miRNA–mRNA and **(B)** circRNA–miRNA–mRNA; green triangles representing mRNA, red ellipses representing miRNA, and blue rectangles representing lncRNA/circRNA.

genome using BWA-mem generating 53 SAM files for each sample which were then merged corresponding to each of the six tissues and submitted to a reliable, sensitive, and widely used command line tool CIRI2. A total of 4621 distinct circRNAs with ≥2 backspliced reads were identified including 3871 novel circRNA. A total of 750 circRNAs were found to be overlapping with known plant circRNAs in PlantcircBase. Tissue-wise analysis revealed 333, 2939, 1004, 156, 400, and 386 circRNAs identified in the flower, fruit, leaf, panicle, root and stem tissues, respectively (Supplementary Table 8). Most the circRNAs were found to be ≤1000 nt (3077 circRNA smaller than 1000 nt) with a median length of approximately 400 nt, which is in consist with previous reports (Figure 9A) (Zheng et al., 2016). As described in previous studies, the abundance of circRNA was found to be lower than that of mRNA and lncRNA. Despite being widely distributed over all the chromosomes, genomic origin analysis showed most of the identified circRNA were coming from intergenic (47.98%) region followed by exonic (43.37%) regions and only few (8.66%) came from the intronic portion of the genome (Figures 9B, C).

3.8 Identification of circRNAs as sponges of miRNAs and analysis of the miRNA–circRNA–mRNA relationship

Various studies suggesting circRNAs can regulate gene expression by acting as ceRNAs for the miRNAs by competing and inhibiting the miRNA binding with the mRNA molecules. The result of TargetFinder revealed 2591 interactions where 690 unique miRNAs were found targeting 432 circRNAs in all tissues combined (Supplementary Table 9). CircRNA Pn4:30202222-30202678 was found to contain the putative miRNA binding site of aly-miR838-3p. mRNA targets of the identified miRNAs were identified using the psRNAtarget web server and a total of 6213 miRNA-mRNA interactions were found in which 619 unique miRNAs were targeting 744 unique black pepper mRNAs. Unique miRNA Pn3.1899 was found to be targeting 15 different mRNAs of black pepper (Supplementary Table 10). The miRNA–circRNA–mRNA network relationship was visualized using Cytoscape software (Figures 7B, 8B).

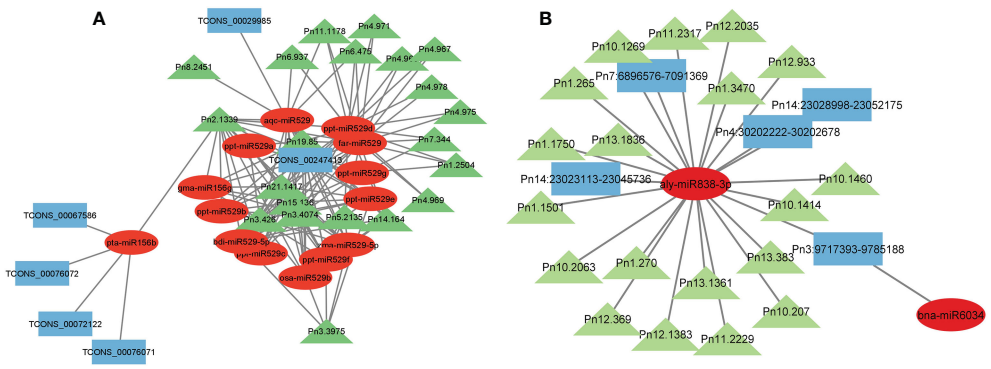


FIGURE 8
Representative figure showing the **(A)** lncRNA–miRNA–mRNA network **(B)** circRNA–miRNA–mRNA network.

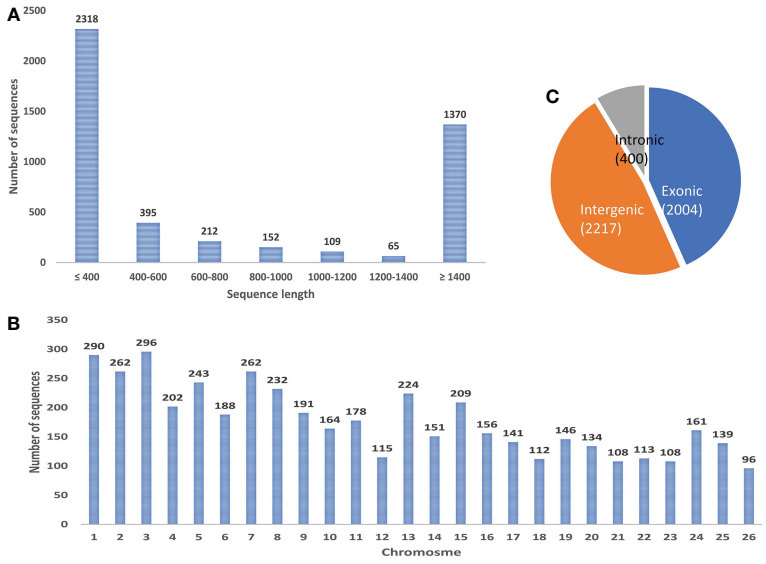


FIGURE 9
(A) Length-wise distribution of circRNA. **(B)** Chromosomal distribution of circRNA. **(C)** Class distribution of circRNA.

3.9 Functional annotation of target genes of competitive endogenous lncRNAs and circRNAs

In this study, we have focused on the trans-regulating action of lncRNAs and circRNAs where they regulate the target gene function by sequestering the miRNAs thus acting as ceRNAs in the cell. A total of 781 and 744 mRNAs were found to have an interaction in 6406 lncRNAs and 4621 circRNAs. To know the functional significance of the identified target mRNAs, GO annotation was performed. Results of the GO analysis revealed important GO subcategories involved, for instance "protein modification process", "transmembrane transport", "anatomical structure development", and "signaling" annotated in the biological process category of GO. Similarly, "transferase activity", "catalytic activity", and "oxidoreductase activity" were annotated in the molecular function category and "nucleus", "membrane", and "plastid" in the cellular component category of GO. KEGG pathway analysis further revealed gene functions. Highly enriched pathways included the ras singaling pathway, diterpenoid biosynthesis, fatty acid biosynthesis, and plant–pathogen interaction. Both the GO and KEGG analyses strongly suggest the potential role of lncRNAs and circRNAs in diverse biological processes in the black pepper plant (Supplementary Tables 11–14).

3.10 Development of web-genomic resource for the black pepper lncRNAs and circRNAs

The black pepper non-coding RNA database, *BPncRDB*, is freely accessible at http://backlin.cabgrid.res.in/bpncrdb/index.php. The website features six tabs including Home, Search lncRNA, Interaction, circRNA, Download and Team (Figure 10). The database catalogues the black pepper's 6406 lncRNAs and 4621 circRNAs along with the miRNAs' interactions involved with them. The database also contains the information regarding the mRNA/genes involved in the ceRNA pathways with the lncRNA/circRNAs. All of the analyzed data that is available in the MySQL database can be downloaded using the links on the "Download" page.

4 Discussion

According to the ENCODE project, only 1%–2% of the human genome codes for proteins, and the vast majority of RNAs are non-coding RNAs such as tRNAs, rRNAs, microRNAs, lncRNAs, circRNAs, and others. LncRNAs and circRNAs are the most abundant non-coding RNAs present in the cell and have been recognized as the key regulators in genetic expression and are actively involved in the plant developmental stages and its response to biotic and abiotic responses. Understanding the functions of non-coding RNAs in biology has received more attention as a result of recent advancements in RNA sequencing technology, epigenomic methodologies, and computational prediction tools. LncRNAs/circRNAs are involved in a wide range of biological activities in all walks of life, which requires additional research (Quinn and Chang, 2016). Low levels of conservation of these ncRNAs between species make their characterization and functional annotation more challenging. Therefore, additional approach is required to create a distinct catalogue of lncRNAs/circRNAs based on numerous datasets for particular species such as black pepper for which there is no such study is available yet. In this study, a stringent methodology was used to identify 6406 lncRNAs and 4621 circRNAs from 53 RNAseq datasets, and they were then classified into three groups based on their position in respect to protein-coding genes. According to the earlier research, the lncRNAs found in this study differ from mRNAs in various unique ways, including having fewer exons, shorter transcript lengths, and

FIGURE 10
Layout of *BPncRDB* web-resource.

lower conversation levels (Wang et al., 2019; Yan et al., 2020). It was found that most of the lncRNAs were between 200-800 base pairs in length and only a few over 2000 base pairs. A similar pattern was found in capsicum (Baruah et al., 2021). Although not evenly distributed, the identified black pepper lncRNAs were found to be distributed across all chromosomes. A similar trend has been reported in crops like rice, wheat and maize (Li et al., 2014; Wang et al., 2018; Lu et al., 2020). Identified lncRNAs belonged to the classes exonic, intergenic and intronic supporting the previous findings suggesting the genome-wide transcription of the lncRNAs (Sun et al., 2020; Baruah et al., 2021).

Studies reporting long non-coding RNAs that act as competitive endogenous RNAs (ceRNAs) have been published, for example, BLIL1 (blue light induced lncRNA), an *Arabidopsis thaliana* lncRNA, was found to be capable of reducing the hypocotyl elongation when subjected to blue light condition by having a competitive relationship with the miRNA miR167 and its mRNA target ARF6/8 (Sun et al., 2020). In rice, an intergenic lncRNA TCONS_00049880 regulates the expression of the SPL gene family by competitively binding to the miRNA osa-miR156. MiRNA-mediated gene regulation by the lncRNAs/circRNAs, also known as target mimicry, miRNA decoy, miRNA sponge, or competitive endogenous RNAs (ceRNAs) is one major interaction between ncRNAs and miRNAs. The first such case of lncRNA-miRNA binding was found in *Arabidopsis thaliana* where miRNA miR399 pairs with lncRNA IPS1 (Induced by Phosphate Starvation 1) leading to inability of miR399 to mediate PHO2 degradation. Similar cases of interactions were also found in our study, for instance, miRNA Osa-miR159e targets lncRNA TCONS_00240461 and sequesters the miRNA, hampering its probable functions in the cell (Meng et al., 2021).

CircRNA studies in plants have also revealed their role as miRNA sponges, where circRNAs with the miRNA binding sites attach and sequester the miRNAs thus regulating the expression of the miRNA targets (Fernandes et al., 2018). In Chinese cabbage circRNA A03:5084249|5089986 has the binding site of miRNA bra-miR5716 known for 15 target mRNAs possibly altering their functions including oxidation-reduction, ATPase activity, electron carrier activity, Myb-type HTH DNA-binding, and transmembrane transport (Wang et al., 2019). These interactions may alter the normal expression of mRNAs thus causing regulatory effect. Similar results were obtained in our study where identified black pepper circRNAs, for instance, miRNA aly-miR838-3p interacts with four circRNAs viz. Pn14:23023113-23045736, Pn7:6896576-7091369, Pn:30202222-30202678, and Pn:23028998-23052175, and is also targeting several mRNAs *viz*. Pn1.1750, Pn10.1460, Pn10.1414, etc., which are according KEGG analysis involved in processes like plant–pathogen interaction (ko04626), ribosome (ko03010), and carotenoid biosynthesis, respectively (ko00906). Plant cells endogenously produce small RNAs called miRNAs from larger precursor transcripts which can interact with the coding mRNAs causing partial transcriptional repression or complete silencing by cleavage, thus acting as a template (Fernandes et al., 2018). For example, lncRNA MSTRG.24217.2 of *Jatropha curcas* act as a precursor of miRNA miR396a which targets growth regulating factors (GRF) which controls the development of plant seeds (Yan et al., 2020). We found four such black pepper lncRNAs that can serve as the precursors for 20 miRNAs which could be involved in multiple cellular processes.

To date, there is no study on lncRNA/circRNAs and ceRNAs network analysis in black pepper. Here we identified genome-wide lncRNA and circRNA as well as explored the lncRNA-miRNA-mRNA and circRNA-miRNA-mRNA networks. BPncRDB, a comprehensive online database of black pepper lncRNAs and circRNAs accessible for academic and research around the world, would provide a platform to

better understand the critical roles that these lncRNAs and circRNAs play in the growth and development as well as the responses towards biotic/abiotic stresses in black pepper.

5 Conclusion

With the goal of creating a comprehensive resource of black pepper lncRNA/circRNA (BPncRDB) and investigating their role as ceRNAs, 53 RNAseq datasets were downloaded by performing a comprehensive search of the publicly available repository NCBI and 6406 and 4621 lncRNAs and circRNAs were each identified using a stringent pipeline. Conservation analysis of the identified lncRNAs revealed 42 identified lncRNAs matching with the previously known plant lncRNAs in the database confirming the weakly conserved nature of lncRNAs. In addition to that 36 lncRNAs were also found to be acting as precursor of 14 miRNAs. A total of 744 and 781 genes were found to be in regulation network *via* lncRNA/circRNA-miRNA-mRNA pathways. Functional enrichment analysis of GO and KEGG revealed pathways such as diterpenoid biosynthesis, RAS signaling pathway, fatty acid biosynthesis, plant-pathogen interactions, etc., suggesting the potential role of lncRNA/circRNA in black pepper growth, development, and resistance against both biotic and abiotic stresses.

Data availability statement

Data presented in the study are included in the article/Supplementary Material. Any further inquiries can be directed to the corresponding author/s.

Author contributions

SJ and MI conceived and designed the study, BK and BS did the data curation and analysis. BK, BS, and UA were involved in database development. BK, BS, SJ and MI wrote the first draft of the manuscript. SJ, MI, UA and AR provided overall guidance and finalized the edited manuscript. All authors contributed to the article and approved the submitted version.

Funding

The authors are thankful to the Indian Council of Agricultural Research, Ministry of Agriculture and Farmers' Welfare, Govt. of India, for providing financial assistance in the form of a CABin grant (F. no. Agril. Edn.4-1/2013-A&P). The grant of the IARI Merit scholarship to BK is duly acknowledged.

Acknowledgments

The financial grants, ICAR- CABin and IARI Merit scholarship to BK are duly acknowledged. The authors further acknowledge the supportive role of the Director, ICAR-IASRI, New Delhi, India.

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.2023.1079221/full#supplementary-material

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215 (3), 403–410. doi: 10.1016/S0022-2836(05)80360-2

Amaral, P. P., Clark, M. B., Gascoigne, D. K., Dinger, M. E., and Mattick, J. S. (2010). lncRNAdb: A reference database for long noncoding RNAs. *Nucleic Acids Res.* 39 (suppl_1), D146–D151. doi: 10.1093/nar/gkq1138

Andrews, S. (2010). FastQC: A quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Azlan, A., Obeidat, S. M., Yunus, M. A., and Azzam, G. (2019). Systematic identification and characterization of aedes aegypti long noncoding RNAs (lncRNAs). *Sci. Rep.* 9 (1), 1–9. doi: 10.1038/s41598-019-47506-9

Baruah, P. M., Krishnatreya, D. B., Bordoloi, K. S., Gill, S. S., and Agarwala, N. (2021). Genome wide identification and characterization of abiotic stress responsive lncRNAs in capsicum annuum. *Plant Physiol. Biochem.* 162, 221–236. doi: 10.1016/j.plaphy.2021.02.031

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A flexible trimmer for illumina sequence data. *Bioinformatics* 30 (15), 2114–2120. doi: 10.1093/bioinformatics/btu170

Butt, M. S., Pasha, I., Sultan, M. T., Randhawa, M. A., Saeed, F., and Ahmed, W. (2013). Black pepper and health claims: A comprehensive treatise. *Crit. Rev. Food Sci. Nutr.* 53 (9), 875–886. doi: 10.1080/10408398.2011.571799

Chen, L. L., and Carmichael, G. G. (2010). Decoding the function of nuclear long non-coding RNAs. *Curr. Opin. Cell Biol.* 22 (3), 357–364. doi: 10.1016/j.ceb.2010.03.003

Chu, Q., Shen, E., Ye, C. Y., Fan, L., and Zhu, Q. H. (2018). Emerging roles of plant circular RNAs. *J. Plant Cell Dev.* 1, 1–14. doi: 10.14302/issn.2832-5311.jpcd-18-1955

Dai, X., Zhuang, Z., and Zhao, P. X. (2018). psRNATarget: A plant small RNA target analysis server, (2017 release). *Nucleic Acids Res.* 46 (W1), W49–W54. doi: 10.1093/nar/gky316

Dasmandal, T., Rao, A. R., and Sahu, S. (2020). Identification and characterization of circular RNAs regulating genes responsible for drought stress tolerance in chickpea and soybean. *Indian J. Of Genet. And Plant Breed.* 80 (01), 1–8. doi: 10.31742/IJGPB.80.1.1

Dey, S. S., Sharma, P. K., Munshi, A. D., Jaiswal, S., Behera, T. K., Kumari, K., et al. (2022). Genome wide identification of lncRNAs and circRNAs having regulatory role in fruit shelf life in health crop cucumber (Cucumis sativus l.). *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.884476/full

Ding, Y., Mao, Y., Cen, Y., Hu, L., Su, Y., Ma, X., et al. (2021). Small RNA sequencing reveals various microRNAs involved in piperine biosynthesis in black pepper (*Piper nigrum* l.). *BMC Genomics* 22 (1), 1–12. doi: 10.1186/s12864-021-08154-4

Fahlgren, N., and Carrington, J. C. (2010). miRNA target prediction in plants. In Plant MicroRNAs, Methods in Molecular Biology, 592, 51–57. B.C. Meyers and P.J. Green (eds.), © Humana Press, a part of Springer Science+Business Media, New Jersey, USA, LLC 2010 doi: 10.1007/978-1-60327-005-2

Fernandes, D. P., Bitar, M., Jacobs, F. M., and Barry, G. (2018). Long non-coding RNAs in neuronal aging. *Non-coding RNA* 4 (2), 12. doi: 10.3390/ncrna4020012

Gao, Y., Zhang, J., and Zhao, F. (2018). Circular RNA identification based on multiple seed matching. *Briefings Bioinf.* 19 (5), 803–810. doi: 10.1093/bib/bbx014

He, X., Guo, S., Wang, Y., Wang, L., Shu, S., and Sun, J. (2020). Systematic identification and analysis of heat stress responsive lncRNAs, circRNAs and miRNAs with associated coexpression and ceRNA networks in cucumber (Cucumis sativus l.). *Physiol. plant.* 168 (3), 736–754. doi: 10.1111/ppl.12997

Hu, L., Xu, Z., Wang, M., Fan, R., Yuan, D., Wu, B., et al. (2019). The chromosome-scale reference genome of black pepper provides insight into piperine biosynthesis. *Nat. Commun.* 10 (1), 1–11. doi: 10.1038/s41467-019-12607-6

Kang, Y. J., Yang, D. C., Kong, L., Hou, M., Meng, Y. Q., Wei, L., et al. (2017). pl2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res.* 45 (W1), W12–W16. doi: 10.1093/nar/gkx428

Kim, D., Langmead, B., and Salzberg, S. L. (2015). HISAT: A fast spliced aligner with low memory requirements. *Nat. Methods* 12 (4), 357. doi: 10.1038/nmeth.3317

Kumar, B., Kumar, A., Jaiswal, S., Iquebal, M. A., Angadi, U. B., Tomar, R. S., et al. (2022). Genome-wide identification of long non-coding RNAs in pearl millet (Pennisetum glaucum (L.)) genotype subjected to drought stress. *Agronomy* 12 (8), 1976. doi: 10.3390/agronomy12081976

Lai, X., Bazin, J., Webb, S., Crespi, M., Zubieta, C., and Conn, S. J. (2018). CircRNAs in plants. *Circular RNAs*, 1087. 329–343. doi: 10.1007/978-981-13-1426-1_26

Li, L., Eichten, S. R., Shimizu, R., Petsch, K., Yeh, C. T., Wu, W., et al. (2014). Genome-wide discovery and characterization of maize long non-coding RNAs. *Genome Biol.* 15 (2), 1–15. doi: 10.1186/gb-2014-15-2-r40

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25 (16), 2078–2079. doi: 10.1093/bioinformatics/btp352

Li, A., Zhang, J., and Zhou, Z. (2014). PLEK: A tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme. *BMC Bioinf.* 15 (1), 1–10. doi: 10.1186/1471-2105-15-311

Lu, T., Cui, L., Zhou, Y., Zhu, C., Fan, D., Gong, H., et al. (2015). Transcriptome-wide investigation of circular RNAs in rice. *Rna* 21 (12), 2076–2087. doi: 10.1261/rna.052282.115

Lu, Q., Guo, F., Xu, Q., and Cang, J. (2020). LncRNA improves the cold resistance of winter wheat by interacting with miR398. *Funct. Plant Biol.* 47 (6), 544–557. doi: 10.1071/FP19267

Lu, Z., Xia, X., Jiang, B., Ma, K., Zhu, L., Wang, L., et al. (2017). Identification and characterization of novel lncRNAs in arabidopsis thaliana. *Biochem. Biophys. Res. Commun.* 488 (2), 348–354. doi: 10.1016/j.bbrc.2017.05.051

Mattick, J. S. (2004). RNA Regulation: A new genetics? *Nat. Rev. Genet.* 5 (4), 316. doi: 10.1038/nrg1321

Meng, X., Li, A., Yu, B., and Li, S. (2021). Interplay between miRNAs and lncRNAs: Mode of action and biological roles in plant development and stress adaptation. *Comput. Struct. Biotechnol. J.* 19, 2567–2574. doi: 10.1016/j.csbj.2021.04.062

Mercer, T. R., Dinger, M. E., and Mattick, J. S. (2009). Long non-coding RNAs: Insights into functions. *Nat. Rev. Genet.* 10 (3), 155. doi: 10.1038/nrg2521

Min, X. J., Butler, G., Storms, R., and Tsang, A. (2005). OrfPredictor: Predicting protein-coding regions in EST-derived sequences. *Nucleic Acids Res.* 33 (suppl_2), W677–W680. doi: 10.1093/nar/gki394

Mu, C., Wang, R., Li, T., Li, Y., Tian, M., Jiao, W., et al. (2016). Long non-coding RNAs (lncRNAs) of sea cucumber: Large-scale prediction, expression profiling, non-coding network construction, and lncRNA-microRNA-gene interaction analysis of lncRNAs in apostichopus japonicus and holothuria glaberrima during LPS challenge and radial organ complex regeneration. *Mar. Biotechnol.* 18 (4), 485–499. doi: 10.1007/s10126-016-9711-y

Negi, A., George Kokkat, J., Jasrotia, R. S., Madhavan, S., Jaiswal, S., Angadi, U. B., et al. (2021). Drought responsiveness in black pepper (*Piper nigrum* l.): Genes associated and development of a web-genomic resource. *Physiol. Plant.* 172 (2), 669–683. doi: 10.1111/ppl.13308

Paraskevopoulou, M. D., and Hatzigeorgiou, A. G. (2016). "Analyzing miRNA–lncRNA interactions," in *Long non-coding RNAs* (New York, NY: Humana Press), 271–286. doi: 10.1007/978-1-4939-3378-5_21

Park, U. H., Jeong, H. S., Jo, E. Y., Park, T., Yoon, S. K., Kim, E. J., et al. (2012). Piperine, a component of black pepper, inhibits adipogenesis by antagonizing PPARγ activity in 3T3-L1 cells. *J. Agric. Food Chem.* 60 (15), 3853–3860. doi: 10.1021/jf204514a

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 (9), 1650. doi: 10.1038/nprot.2016.095

Pertea, G., and Pertea, M. (2020). GFF utilities: GffRead and GffCompare. *F1000Research*, 9:304 doi: 10.12688/f1000research.23297.2

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., and Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* 33 (3), 290. doi: 10.1038/nbt.3122

Quinn, J. J., and Chang, H. Y. (2016). Unique features of long non-coding RNA biogenesis and function. *Nat. Rev. Genet.* 17 (1), 47–62. doi: 10.1038/nrg.2015.10

Ren, Y., Yue, H., Li, L., Xu, Y., Wang, Z., Xin, Z., et al. (2018). Identification and characterization of circRNAs involved in the regulation of low nitrogen-promoted root growth in hexaploid wheat. *Biol. Res.* 51, 43 doi: 10.1186/s40659-018-0194-3

R.R., Nair, B., Sasikumar, and P.N., Ravindran (1993). Polyploidy in a cultivar of black pepper (*Piper nigrum* l.) and its open pollinated progenies. *Cytologia* 58 (1), 27–31. doi: 10.1508/cytologia.58.27

Sharma, S., Taneja, M., Tyagi, S., Singh, K., and Upadhyay, S. K. (2017). Survey of high throughput RNA-seq data reveals potential roles for lncRNAs during development and stress response in bread wheat. *Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.01019/full

Song, X., Sun, L., Luo, H., Ma, Q., Zhao, Y., and Pei, D. (2016). Genome-wide identification and characterization of long non-coding RNAs from mulberry (Morus notabilis) RNA-seq data. *Genes* 7 (3), 11. doi: 10.3390/genes7030011

Sun, Z., Huang, K., Han, Z., Wang, P., and Fang, Y. (2020). Genome-wide identification of arabidopsis long noncoding RNAs in response to the blue light. *Sci. Rep.* 10 (1), 1–10. doi: 10.1038/s41598-020-63187-1

Sun, Y., Zhang, H., Fan, M., He, Y., and Guo, P. (2020). Genome-wide identification of long non-coding RNAs and circular RNAs reveal their ceRNA networks in response to cucumber green mottle mosaic virus infection in watermelon. *Arch. Virol.* 165 (5), 1177–1190. doi: 10.1007/s00705-020-04589-4

Szcześniak, M. W., Bryzghalov, O., Ciomborowska-Basheer, J., and Makałowska, I. (2019). "CANTATAdb 2.0: Expanding the collection of plant long noncoding RNAs," in *Plant long non-coding RNAs* (New York, NY: Humana Press), 415–429. doi: 10.1007/978-1-4939-9045-0_26

Wang, K. C., and Chang, H. Y. (2011). Molecular mechanisms of long noncoding RNAs. *Mol. Cell* 43 (6), 904–914. doi: 10.1016/j.molcel.2011.08.018 doi: 10.1038/s41598-019-51190-0

Wang, Y., Luo, X., Sun, F., Hu, J., Zha, X., Su, W., et al. (2018). Overexpressing lncRNA LAIR increases grain yield and regulates neighbouring gene cluster expression in rice. *Nat. Commun.* 9 (1), 1–9. doi: 10.1038/s41467-018-05829-7

Wang, W., Wang, J., Wei, Q., Li, B., Zhong, X., Hu, T., et al. (2019). Transcriptome-wide identification and characterization of circular RNAs in leaves of Chinese cabbage (Brassica rapa l. ssp. pekinensis) in response to calcium deficiency-induced tip-burn. *Sci. Rep.* 9 (1), 1–9. doi: 10.1038/s41598-019-51190-0

Yan, X., Ma, L., and Yang, M. (2020). Identification and characterization of long non-coding RNA (lncRNA) in the developing seeds of jatropha curcas. *Sci. Rep.* 10 (1), 1–10. doi: 10.1038/s41598-020-67410-x

Yang, Y., Jiang, X., Shi, J., Wang, Y., Huang, H., Yang, Y., et al. (2022). Functional annotation of circRNAs in tea leaves after infection by the tea leaf spot pathogen, lasiodiplodia theobromae. *Phytopathology®* 112 (2), 460–463. doi: 10.1094/PHYTO-05-21-0184-A?rfr_dat=cr_pub++0pubmed&url_ver=Z39.88-2003&rfr_id=ori%3Arid%3Acrossref.org

Yang, Z., Yang, C., Wang, Z., Yang, Z., Chen, D., and Wu, Y. (2019). LncRNA expression profile and ceRNA analysis in tomato during flowering. *PloS One* 14 (1), e0210650. doi: 10.1371/journal.pone.0210650

Zheng, Q., Bao, C., Guo, W., Li, S., Chen, J., Chen, B., et al. (2016). Circular RNA profiling reveals an abundant circHIPK3 that regulates cell growth by sponging multiple miRNAs. *Nat. Commun.* 7, 11215. doi: 10.1038/ncomms11215

Zhou, R., Sanz-Jimenez, P., Zhu, X. T., Feng, J. W., Shao, L., Song, J. M., et al. (2021). Analysis of rice transcriptome reveals the LncRNA/CircRNA regulation in tissue development. *Rice* 14 (1), 1–16. doi: 10.1186/s12284-021-00455-2

Zhu, Y. X., Jia, J. H., Yang, L., Xia, Y. C., Zhang, H. L., Jia, J. B., et al. (2019). Identification of cucumber circular RNAs responsive to salt stress. *BMC Plant Biol.* 19 (1), 1–18. doi: 10.1186/s12870-019-1712-3

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 18 April 2023
DOI 10.3389/fpls.2023.1143111

OPEN ACCESS

EDITED BY
Lida Zhang,
Shanghai Jiao Tong University, China

REVIEWED BY
Shyam Sundar Dey,
Indian Agricultural Research Institute (ICAR), India
Dhananjay K. Pandey,
Amity University, Jharkhand, India

*CORRESPONDENCE
Chakresh Kumar Jain
ckj522@yahoo.com
Nisha Singh
singh.nisha88@gmail.com

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

RECEIVED 12 January 2023
ACCEPTED 22 March 2023
PUBLISHED 18 April 2023

CITATION
Saxena H, Kulshreshtha A, Agarwal A, Kumar A, Singh N and Jain CK (2023) LDRGDb - Legumes disease resistance genes database. *Front. Plant Sci.* 14:1143111. doi: 10.3389/fpls.2023.1143111

LDRGDb - Legumes disease resistance genes database

Harshita Saxena[1], Aishani Kulshreshtha[1], Avinav Agarwal[1], Anuj Kumar[2], Nisha Singh[3]* and Chakresh Kumar Jain[1]*

[1]Department of Biotechnology, Jaypee Institute of Information Technology, Noida, India, [2]Department of Microbiology and Immunology, Dalhousie University, Halifax, NS, Canada, [3]Department of Bioinformatics, Gujarat Biotechnology University, Gandhinagar, India

Legumes comprise one of the world's largest, most diverse, and economically important plant families, known for their nutritional and medicinal benefits. Legumes are susceptible to a wide range of diseases, similar to other agricultural crops. Diseases have a considerable impact on the production of legume crop species, resulting in large yield losses worldwide. Due to continuous interactions between plants and their pathogens in the environment and the evolution of new pathogens under high selection pressure; disease resistant genes emerge in plant cultivars in the field against those pathogens or disease. Thus, disease resistant genes play critical roles in plant resistance responses, and their discovery and subsequent use in breeding programmes aid in reducing yield loss. The genomic era, with its high-throughput and low-cost genomic tools, has revolutionised our understanding of the complex interactions between legumes and pathogens, resulting in the identification of several critical participants in both the resistant and susceptible relationships. However, a substantial amount of existing information about numerous legume species has been disseminated as text or is preserved across fractions in different databases, posing a challenge for researchers. As a result, the range, scope, and complexity of these resources pose challenges to those who manage and use them. Therefore, there is an urgent need to develop tools and a single conjugate database to manage genetic information for the world's plant genetic resources, allowing for the rapid incorporation of essential resistance genes into breeding strategies. Here, developed the first comprehensive database of disease resistance genes named as LDRGDb - LEGUMES DISEASE RESISTANCE GENES DATABASE comprises 10 legumes [Pigeon pea (*Cajanus cajan*), Chickpea (*Cicer arietinum*), Soybean (*Glycine max*), Lentil (*Lens culinaris*), Alfalfa (*Medicago sativa*), Barrelclover (*Medicago truncatula*), Common bean (*Phaseolus vulgaris*), Pea (*Pisum sativum*),Faba bean (*Vicia faba*), and Cowpea (*Vigna unguiculata*)]. The LDRGDb is a user-friendly database developed by integrating a variety of tools and software that combine knowledge about resistant genes, QTLs, and their loci, with proteomics, pathway interactions, and genomics (https://ldrgdb.in/).

KEYWORDS

disease resistance genes, genomics, LDRGDb, legumes, proteomics, QTLs data

Introduction

Legumes are seed plants belonging to the Leguminosae family, spanning more than 13000 species across 600 genera. Legumes are diverse; they are suitable for cultivation in a variety of environments and temperatures (Magrini et al., 2019). They are an excellent source of antioxidants, micronutrients, and proteins, finding use as natural fertilizers, medicines, and animal fodder (Shah et al., 2020; Singh et al., 2022). Apart from their health benefits, legumes naturally have the ability to fix nitrogen from the atmosphere symbiotically, which has positive effects on agriculture and soil enrichment that are both long-lasting and economically viable. Improvement in legume cultivation and their increased usage can help ensure food security and contribute to better soil fertility (Foyer et al., 2016; Singh et al., 2020).

However, one of the main impediments in the way of abundant and quality yield of legume crops is pests and diseases Figure 1. Furthermore, climate changes have made it easier for various species to move freely, which has led to the emergence of new diseases that could grow into uncontrollable epidemics and endanger food security (Piquerez et al., 2014). Thus, crop improvement is a key component of sustainable agriculture and can be accomplished using a variety of techniques, ranging from traditional breeding to genetic engineering. To increase the effectiveness of the breeding process, it is necessary to have a thorough understanding of host-pathogen interactions as well as effective resistance mechanisms at the cellular, genetic, and molecular levels. The most effective, affordable, and environmentally friendly of the various control techniques available is breeding for resistance (Rubiales et al., 2014). These factors drive plant breeders and scientists to focus their efforts on finding mechanisms for plant disease resistance (Osuna-Cruz et al., 2018). Indeed, studies into the genes and genomes of legumes have provided valuable insight into disease resistance components and other agronomic traits (Leal-Bertioli et al., 2009; Khera et al., 2018; Zhuang et al., 2019; Sampaio et al., 2020).

Resistance genes (R genes) are variable plant genes that confer resistance to vast varieties of biotrophic pathogens, including viruses. As they can specifically recognise the matching pathogen effectors or associated protein(s), resistance (R) genes are the most potent defences against pathogen invasion (Liu et al., 2007). This allows plants to mount an effective defence at the site of infection. R genes are capable of expressing PR proteins in response to physical or chemical stimuli (Rodríguez-Sifuentes et al., 2020). As PR proteins have hydrolytic activity, contact toxicity, involvement in defence signalling, and pathogen enzyme inhibition, they constitute a variety of weapons against pathogens. We have seen that plant R genes have been divided into several groups based on their typical domains (Zheng et al., 2016). A class of proteins containing nucleotide binding (NB) and leucine-rich repeat (LRR) domains is encoded by the majority of R genes (Friedman and Baker, 2007).

It is now known that R genes also produce proteins with a series of carboxy-terminal leucine-rich repeats (LRRs), a putative amino-terminal signalling domain, and a nucleotide binding site (NBS). There are two primary classes of "NBS-LRR" proteins: first which encodes an amino-terminal coiled-coiled motif (CC-NBS-LRR or CNL proteins) and second which has an amino-terminal TIR (Toll/interleukin receptor) domain (which are known as TIR-NBS-LRR or TNL proteins) (Meyers et al., 2005). Figure 2 depicts the various classes of R-Gene. Hence, the participation of multiple R genes, post-transcriptional regulators, and biotic and abiotic stressors limits the ability of plants to resist disease.

R gene mechanisms work on the gene-for-gene concept (Kaloshian, 2004). The resistance occurs only when the R gene proteins engage with the pathogen Avr gene in a certain form. They can interact with the pathogen gene in two ways, first by directly interacting with its protein product and second, if it plays a catalytic role by interacting with something created by the Avr gene product (Hulbert et al., 2001). Any attempt of an infection is thus governed by both the genotype of the host and that of the pathogen. Once this recognition has taken place, the defence reactions are triggered. Hypersensitive reaction is frequently characterized in these defence reactions, which results in the death of the initial cell or cells infected and also the local accumulation of antimicrobial compounds (Moffett, 2009). Phytophthora root and stem rot in nonhost common bean has been found to induce a strong hypersensitive response to *Phytophthora sojae* due to upregulation of genes promoting phaseollidin and glyceollin production, which have significant inhibitory effects on oospore production and mycelial formation (Bi et al., 2022). Transcription factors have also been implicated in multilayer defence signalling against *Fusarium* wilt disease in chickpea (Chakraborty et al., 2020).

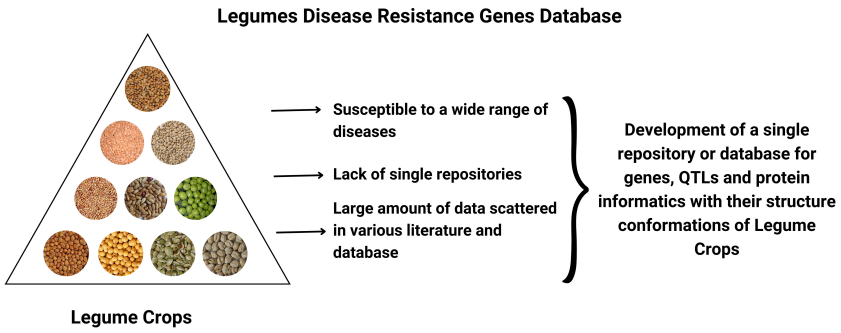


FIGURE 1
Graphical Abstract.

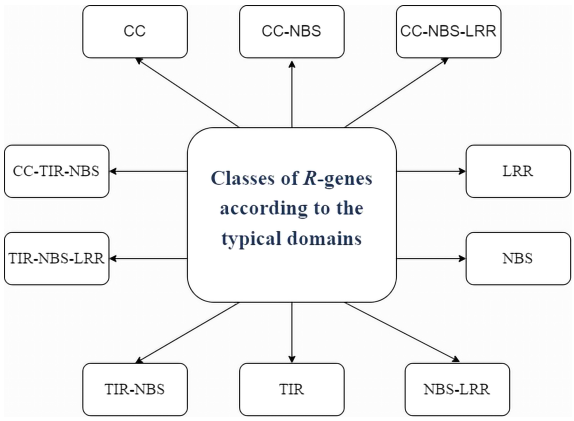


FIGURE 2
Image depicting various classes of R-Gene.

Legumes exhibit resistance through other means as well; a recent study on Faba beans indicated that its resistant genotype withstood the Chocolate spot infection caused by *Botrytis fabae* due to a better Photosystem II Repair mechanism at early stages of the infection (Castillejo et al., 2021). Leaf spot infected alfalfa, when colonized by arbuscular mycorrhizal fungus (AMF), has shown to ameliorate the effects of the infection, thus displaying the potential to serve as a biocontrol strategy for leaf spot infection of alfalfa (Li et al., 2019).

Development of cost-effective next generation sequencing platforms with enhanced performances have given way to copious amounts of data. In order for these data to be effectively analysed and compared, various data management practices are being employed (Bauchet et al., 2019). Many databases exist for comprehensive study and management of the insurmountable data generated by plant genome studies. However, these repositories either only focus individually on model genomes such as soybean (https://soybase.org/) (Grant et al., 2009), *Medicago truncatula* (https://www.legoo.org) (Carrere et al., 2019), (http://www.medicago.org/MtDB) (Lamblin, 2003), or a combination of multiple model species (http://www.plantgdb.org/) (Dong, 2004), (http://plantgrn.noble.org/LegumeIP/) (Li et al., 2011). Currently, there is a lack of repositories that can pool the numerous disease resistance genes with proteomics and facilitate the quick integration of crucial resistance genes into breeding methods. Keeping this in mind, we have established and developed the first comprehensive database of disease resistance genes in legume plants named as LDRGDb, that incorporates knowledge about resistant genes, QTLs, and their loci, with proteomics, pathways interactions as well as structural conformations of proteins. The database serves as a medium for comprehensive search and retrieval of relevant information, and will help researchers understand various biological phenomena with relative ease. Our database has been constructed so as to aid in research of disease resistant common legume cultivars and their underlying mechanisms. LDRGDb spans 10 legume species, with genes, QTL information such as linkage group, neighbouring marker, population, maternal and paternal parent, proteomics, informatics such as molecular weight, theoretical pI, length of amino acids, half-life of protein, and aliphatic index with emphasis on structural conformation, which can aid in a thorough understanding of various pathways and interactions involved in disease resistance, functioning as a single repository for assisting in research.

Methodology

Data collection

The database has major 10 legume crops with their common disease-related genes and QTLs incorporated. QTLs data has been combined from numerous databases for a single model organism, such as soybase (Grant et al., 2009), databases for several species, such as pulsedb (Humann et al., 2019), prgdb (Calle García et al., 2021), and other databases. Further, the QTLs related to diseases were manually curated from the available scientific literature, books and journals. We incorporated gene data from various journals and books using the Pubmed (PubMed Labs, 2021) search box and varied search queries incorporating different diseases and crops from the last 5 years. (Figures 3, 4) show the number of genes and QTLs incorporated into the database.

Annotation of genes and QTLs

We performed functional annotation studies at genes and QTLs levels to develop comprehensive information for the collected data. Uniprot (Bateman et al., 2020) was used to retrieve protein-related data of the genes while protparam (Gasteiger et al., 2005) was used to retrieve individual protein details such as theoretical pI, the total number of negatively and positively charged residues, protein's half-life, aliphatic index, and molecular weight etc. To mine and integrate the biological, molecular, and cellular processes involved in genes Uniport was utilized, (Bateman et al., 2020) whereas Interpro (Paysan-Lafosse et al., 2022) database for protein family and domain retrieval. The Swiss model (Waterhouse et al., 2018)

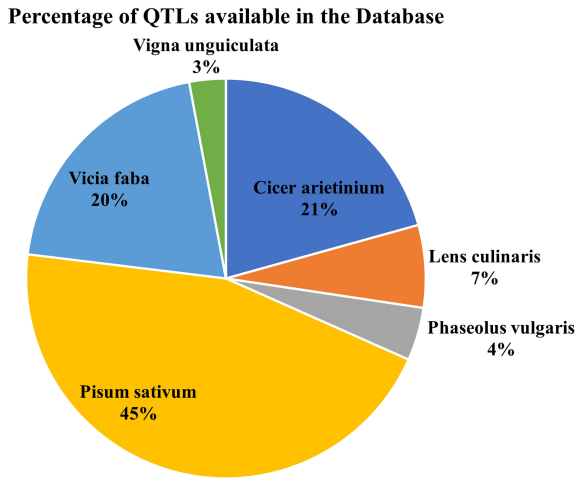


FIGURE 3
Percentage of QTls available in the Database.

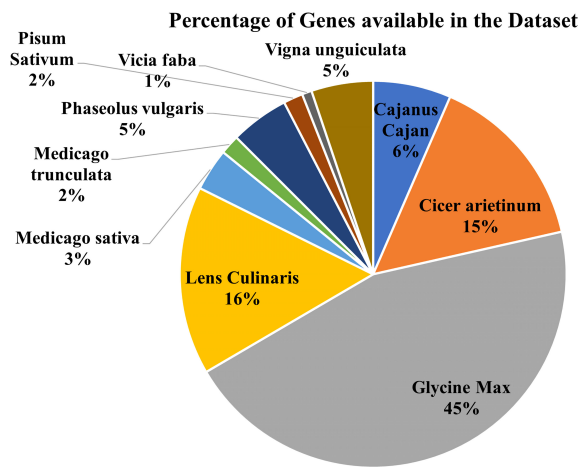


FIGURE 4
Percentage of Genes available in the Database.

was used to incorporate the structure of the protein of a gene with its template included.

Database design

The database is made up of two tables: Genes and QTLs, and one-to-many relationships were utilised to create the complete database structure, allowing for the inclusion of any number of connections. The data's detailed data flow is documented in (Figure 5). For website development, Django v4.1.3 with in-built sqlite3 has been utilized. Django offers speedy development, fast processing, and scalability for website development, while sqlite3 provides a lightweight disk-based database that doesn't require a separate server process and allows accessing the database using a nonstandard variant of the SQL query language. Special care has been taken to protect the database's structure, consistency, and the accuracy of the data stored.

Key features

The LDRGDb is a comprehensive search and retrieval database of significant information that serves as a centralised repository for research purposes. It is the only legume repository that includes not only QTL information such as linkage group, neighbouring marker, population, and maternal and paternal parent, but also genes with proteomics information such as theoretical pI, the total number of negatively and positively charged residues, protein half-life, aliphatic index, and molecular weight, among other things. The inclusion of protein pathway interactions with diverse biological, cellular, and molecular processes aid researchers in understanding a wide range of processes. The incorporation of protein structural conformations with the template annotated through the Swiss model is a major aspect of the LDRGDb that can aid in full knowledge of protein interactions involved in disease resistance. (Figure 6) depicts how the LDRGDb is utilized for the retrieval of significant information.

Conclusion

Disease-resistant mechanisms of plants are constrained by the involvement of various genes, post-transcriptional controls, biotic and abiotic stresses. The understanding of these mechanisms and their impact on plant proteomes and metabolomes is important for improvement of legume production. To obtain knowledge about the mechanisms of resistance, LDRGDb serves as an exemplary one stop database for quick access of disease

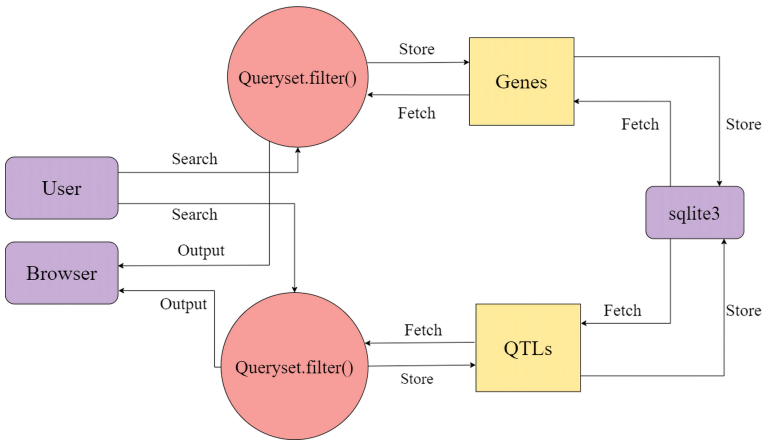


FIGURE 5
Depicting the data flow diagram of the program.

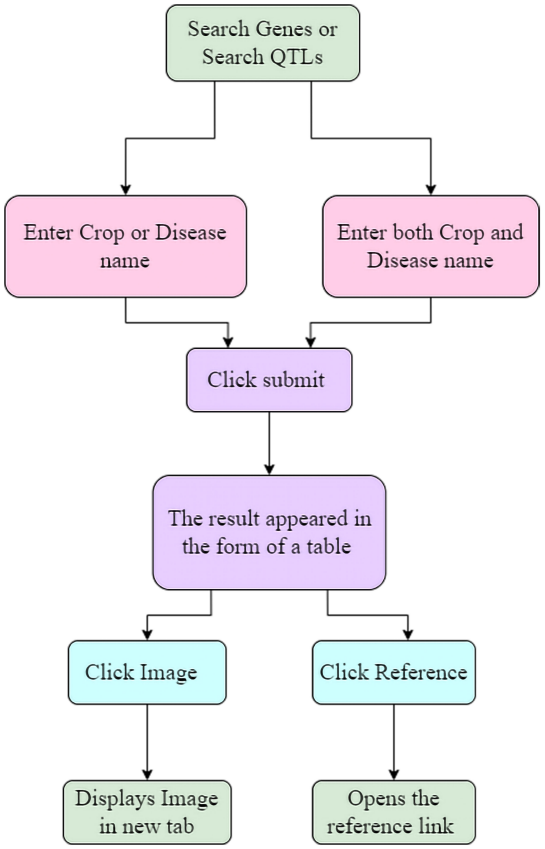


FIGURE 6
Depicting how to utilize the LDRGDb.

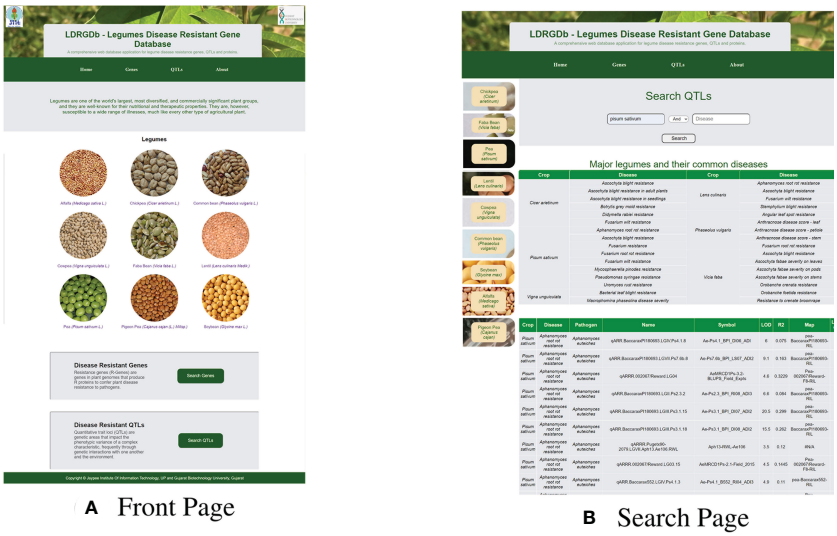


FIGURE 7
Images showing the **(A)** Front Page and **(B)** Search Page of the website.

resistance genes, QTLs, and the proteins and pathways associated with them for various legume species. Users can search using either disease names, crop names, or both, facilitating easy and comprehensive ingress into the data. The data collected from various databases and literature has been meticulously structured to allow for rapid searches for diseases, legumes and the genes associated with them using user-friendly web interfaces. The database has specific sections for particular crop characteristics, prevalent diseases, as well as a FAQ section for frequently asked questions. These are the webpage screenshots (Figure 7).

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

HS contributed to collection of data implementation and initial draft writing. AiK contributed to collection of data, initial draft writing tables, and figures. AA contributed to collection of data and initial draft writing. AK contributed to designing and editing the manuscript. NS conceived, supervised and edited the final manuscript. CJ conceived, supervised and edited the final manuscript. All authors contributed to the article and approved the submitted version.

Acknowledgments

We are thankful for Department of Biotechnology, Jaypee Institute of Information Technology, Noida, India and Department of Bioinformatics, Gujarat Biotechnology University, Gandhinagar, Gujarat, India for providing the necessary support.

Conflict of interest

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

Publisher's note

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

References

Bateman, A., Martin, M.-J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., et al. (2020). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* 49 (D1), D480–D489. doi: 10.1093/nar/gkaa1100

Bauchet, G. J., Bett, K. E., Cameron, C. T., Campbell, J. D., Cannon, E. K. S., Cannon, S. B., et al. (2019). The future of legume genetic data resources: challenges, opportunities, and priorities. *Legume Sci.* 1 (1), e16. doi: 10.1002/leg3.16

Bi, X., Song, G., Yu, H., Zhang, Z., Liu, H., Yang, Z., et al. (2022). Changes in biochemistry and cellular ultrastructure support different resistance mechanisms to phytophthora sojae in nonhost common bean and host soybean. *Plant Pathol.* 71 (4), 917–926. doi: 10.1111/ppa.13527

Calle García, J., Guadagno, A., Paytuvi-Gallart, A., Saera-Vila, A., Amoroso, C., 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

Carrere, S., Verdenaud, M., Gough, C., Gouzy, J., and Gamas, P. (2019). LeGOO: an expertized knowledge database for the model legume medicago truncatula. *Plant Cell Physiol.* 61 (1), 203–211. doi: 10.1093/pcp/pcz177

Castillejo, M.Á., Villegas-Fernández, Á.M., Hernández-Lao, T., and Rubiales, D. (2021). Photosystem II repair cycle in faba bean may play a role in its resistance to botrytis fabae infection. *Agronomy* 11 (11), 2247. doi: 10.3390/agronomy11112247

Chakraborty, J., Sen, S., Ghosh, P., Jain, A., and Das, S. (2020). Inhibition of multiple defense responsive pathways by CaWRKY70 transcription factor promotes susceptibility in chickpea under fusarium oxysporum stress condition. *BMC Plant Biol.* 20 (1), 319. doi: 10.1186/s12870-020-02527-9

Dong, Q. (2004). PlantGDB, plant genome database and analysis tools. *Nucleic Acids Res.* 32 (90001), 354D359. doi: 10.1093/nar/gkh046

Foyer, C. H., Lam, H.-M., Nguyen, H. T., Siddique, K. H. M., Varshney, R. K., Colmer, T. D., et al. (2016). Neglecting legumes has compromised human health and sustainable food production. *Nat. Plants* 2 (8), 16112. doi: 10.1038/nplants.2016.112

Friedman, A. R., and Baker, B. J. (2007). The evolution of resistance genes in multi-protein plant resistance systems. *Curr. Opin. Genet. Dev.* 17 (6), 493–499. doi: 10.1016/j.gde.2007.08.014

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R. D., et al. (2005). Protein identification and analysis tools on the ExPASy server. *Proteomics Protoc. Handb.*, 571–607. doi: 10.1385/1-59259-890-0:571

Grant, D., Nelson, R. T., Cannon, S. B., and Shoemaker, R. C. (2009). SoyBase, the USDA-ARS soybean genetics and genomics database. *Nucleic Acids Res.* 38 (suppl_1), D843–D846. doi: 10.1093/nar/gkp798

Hulbert, S. H., Webb, C. A., Smith, S. M., and Sun, Q. (2001). RESISTANCE GENE COMPLEXES: Evolution and utilization. *Annu. Rev. Phytopathol.* 39 (1), 285–231. doi: 10.1146/annurev.phyto.39.1.285

Humann, J. L., Jung, S., Cheng, C.-H., Lee, T., Zheng, P., Frank, M., et al. (2019). "Cool season food legume genome database: A resource for pea, lentil, faba bean and chickpea genetics, genomics and breeding," in *Proceedings of the international plant and animal genome conference*(San Diego, CA, USA).

Kaloshian, I. (2004). Gene-for-gene disease resistance: Bridging insect pest and pathogen defense. *J. Chem. Ecol.* 30 (12), 2419–2438. doi: 10.1007/s10886-004-7943-1

Khera, P., Pandey, M. K., Mallikarjuna, N., Sriswathi, M., Roorkiwal, M., Janila, P., et al. (2018). Genetic imprints of domestication for disease resistance, oil quality, and yield component traits in groundnut (Arachis hypogaea l.). *Mol. Genet. Genomics* 294 (2), 365–378. doi: 10.1007/s00438-018-1511-9

Lamblin, A.-F. J. (2003). MtDB: A database for personalized data mining of the model legume medicago truncatula transcriptome. *Nucleic Acids Res.* 31 (1), 196–201. doi: 10.1093/nar/gkg119

Leal-Bertioli, S. C., José, A., Alves-Freitas, D. M., Moretzsohn, M. C., Guimarães, P. M., Nielen, S., et al. (2009). Identification of candidate genome regions controlling disease resistance in arachis. *BMC Plant Biol.* 9 (1), 112. doi: 10.1186/1471-2229-9-112

Li, J., Dai, X., Liu, T., and Zhao, P. X. (2011). LegumeIP: an integrative database for comparative genomics and transcriptomics of model legumes. *Nucleic Acids Res.* 40 (D1), D1221–D1229. doi: 10.1093/nar/gkr939

Li, Y., Duan, T., Nan, Z., and Li, Y. (2019). Arbuscular mycorrhizal fungus alleviates alfalfa leaf spots caused by phoma medicaginis revealed by RNA-seq analysis. *J. Appl. Microbiol.* 130 (2), 547–560. doi: 10.1111/jam.14387

Liu, J., Liu, X., Dai, L., and Wang, G. (2007). Recent progress in elucidating the structure, function and evolution of disease resistance genes in plants. *J. Genet. Genomics* 34 (9), 765–776. doi: 10.1016/s1673-8527(07)60087-3

Magrini, M.-B., Cabanac, G., Lascialfari, M., Plumecocq, G., Amiot, M.-J., Anton, M., et al. (2019). Peer-reviewed literature on grain legume species in the WoS, (1980–2018): A comparative analysis of soybean and pulses. *Sustainability* 11 (23), 6833. doi: 10.3390/su11236833

Meyers, B. C., Kaushik, S., and Nandety, R. S. (2005). Evolving disease resistance genes. *Curr. Opin. Plant Biol.* 8 (2), 129–134. doi: 10.1016/j.pbi.2005.01.002

Moffett, P. (2009). Mechanisms of recognition in dominant r gene mediated resistance. *Adv. Virus Res.* 75, 1–33. doi: 10.1016/S0065-3527(09)07501-0

Osuna-Cruz, C. M., Paytuvi-Gallart, A., Di Donato, A., Sundesha, V., Andolfo, G., Aiese Cigliano, R., et al. (2018). PRGdb 3.0: a comprehensive platform for prediction and analysis of plant disease resistance genes. *Nucleic Acids Res.* 46 (D1), D1197–D1201. doi: 10.1093/nar/gkx1119

Paysan-Lafosse, T., Blum, M., Chuguransky, S., Grego, T., Pinto, B. L., Salazar, G., et al. (2022). InterPro in 2022. *Nucleic Acids Res* 51(D1), D418–D427. doi: 10.1093/nar/gkac993

Piquerez, S. J. M., Harvey, S. E., Beynon, J. L., and Ntoukakis, V. (2014). Improving crop disease resistance: lessons from research on arabidopsis and tomato. *Front. Plant Sci.* 5. doi: 10.3389/fpls.2014.00671

PubMed Labs (2021). Available at: https://pubmed.ncbi.nlm.nih.gov/.

Rodríguez-Sifuentes, L., Marszalek, J. E., Chuck-Hernández, C., and Serna-Saldívar, S. O. (2020). Legumes protease inhibitors as biopesticides and their defense mechanisms against biotic factors. *Int. J. Mol. Sci.* 21 (9), 3322. doi: 10.3390/ijms21093322

Rubiales, D., Fondevilla, S., Chen, W., Gentzbittel, L., Higgins, T. J. V., Castillejo, M. A., et al. (2014). Achievements and challenges in legume breeding for pest and disease resistance. *Crit. Rev. Plant Sci.* 34 (1-3), 195–236. doi: 10.1080/07352689.2014.898445

Sampaio, A. M., Araújo, S., de, S., Rubiales, D., and Vaz Patto, M. C. (2020). Fusarium wilt management in legume crops. *Agronomy* 10 (8), 1073. doi: 10.3390/agronomy10081073

Shah, F., Khan, Z., Iqbal, A., Turan, M., and Olgun, M. (2020) *Recent advances in grain crops research*. Available at: https://www.intechopen.com/books/8168.

Singh, N., Jain, P., Ujinwal, M., and Langyan, S. (2022). Escalate protein plates from legumes for sustainable human nutrition. *Front. Nutr.* 9. doi: 10.3389/fnut.2022.977986

Singh, N., Rai, V., and Singh, N. K. (2020). Multi-omics strategies and prospects to enhance seed quality and nutritional traits in pigeonpea. *Nucleus* 63 (3), 249–256. doi: 10.1007/s13237-020-00341-0

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., et al. (2018). SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* 46 (W1), W296–W303. doi: 10.1093/nar/gky427

Zheng, F., Wu, H., Zhang, R., Li, S., He, W., Wong, F. -L., et al. (2016). Molecular phylogeny and dynamic evolution of disease resistance genes in the legume family. *BMC Genomics* 17 (1), 402. doi: 10.1186/s12864-016-2736-9

Zhuang, W., Chen, H., Yang, M., Wang, J., Pandey, M. K., Zhang, C., et al. (2019). The genome of cultivated peanut provides insight into legume karyotypes, polyploid evolution and crop domestication. *Nat. Genet.* 51 (5), 865–876. doi: 10.1038/s41588-019-0402-2

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 07 June 2023
DOI 10.3389/fpls.2023.1135285

OPEN ACCESS

EDITED BY
Nacer Bellaloui,
United States Department of Agriculture, United States

REVIEWED BY
Gehendra Bhattarai,
University of Arkansas, United States
Messias Pereira,
Universidade Estadual do Norte Fluminense Darcy Ribeiro, Brazil

*CORRESPONDENCE
Manish Srivastav
manishfht@gmail.com
Nagendra Kumar Singh
nksingh4@gmail.com

†PRESENT ADDRESS
Ajay Kumar Mahato,
Laboratory of Genome Informatics, Centre for DNA Fingerprinting and Diagnostics, Hyderabad, India

‡These authors have contributed equally to this work

RECEIVED 31 December 2022
ACCEPTED 08 May 2023
PUBLISHED 07 June 2023

CITATION
Srivastav M, Radadiya N, Ramachandra S, Jayaswal PK, Singh N, Singh S, Mahato AK, Tandon G, Gupta A, Devi R, Subrayagowda SH, Kumar G, Prakash P, Singh S, Sharma N, Nagaraja A, Kar A, Rudra SG, Sethi S, Jaiswal S, Iquebal MA, Singh R, Singh SK and Singh NK (2023) High resolution mapping of QTLs for fruit color and firmness in Amrapali/Sensation mango hybrids.
Front. Plant Sci. 14:1135285.
doi: 10.3389/fpls.2023.1135285

High resolution mapping of QTLs for fruit color and firmness in Amrapali/Sensation mango hybrids

Manish Srivastav[1*‡], Nidhi Radadiya[1‡], Sridhar Ramachandra[1], Pawan Kumar Jayaswal[2], Nisha Singh[2], Sangeeta Singh[2], Ajay Kumar Mahato[2†], Gitanjali Tandon[3], Ankit Gupta[1], Rajni Devi[1], Sreekanth Halli Subrayagowda[1], Gulshan Kumar[1], Pragya Prakash[1], Shivani Singh[1], Nimisha Sharma[1], A. Nagaraja[1], Abhijit Kar[4], Shalini Gaur Rudra[4], Shruti Sethi[4], Sarika Jaiswal[3], Mir Asif Iquebal[3], Rakesh Singh[5], Sanjay Kumar Singh[1] and Nagendra Kumar Singh[2*]

[1]Division of Fruits and Horticultural Technology, Indian Council of Agricultural Research (ICAR)- Indian Agricultural Research Institute, New Delhi, India, [2]Genomics Laboratory, Indian Council of Agricultural Research (ICAR)- National Institute for Plant Biotechnology, New Delhi, India, [3]Division of Agricultural Bioinformatics, Indian Council of Agricultural Research (ICAR)- Indian Agricultural Statistics Research Institute, New Delhi, India, [4]Division of Food Science and Postharvest Technology, Indian Council of Agricultural Research (ICAR)- Indian Agricultural Research Institute, New Delhi, India, [5]Division of Genomic Resources, Indian Council of Agricultural Research (ICAR)- National Bureau of Plant Genetic Resources, New Delhi, India

Introduction: Mango (Mangifera indica L.), acclaimed as the 'king of fruits' in the tropical world, has historical, religious, and economic values. It is grown commercially in more than 100 countries, and fresh mango world trade accounts for ~3,200 million US dollars for the year 2020. Mango is widely cultivated in sub-tropical and tropical regions of the world, with India, China, and Thailand being the top three producers. Mango fruit is adored for its taste, color, flavor, and aroma. Fruit color and firmness are important fruit quality traits for consumer acceptance, but their genetics is poorly understood.

Methods: For mapping of fruit color and firmness, mango varieties Amrapali and Sensation, having contrasting fruit quality traits, were crossed for the development of a mapping population. Ninety-two bi-parental progenies obtained from this cross were used for the construction of a high-density linkage map and identification of QTLs. Genotyping was carried out using an 80K SNP chip array.

Results and discussion: Initially, we constructed two high-density linkage maps based on the segregation of female and male parents. A female map with 3,213 SNPs and male map with 1,781 SNPs were distributed on 20 linkages groups covering map lengths of 2,844.39 and 2,684.22cM, respectively. Finally, the integrated map was constructed comprised of 4,361 SNP markers distributed on 20 linkage groups, which consisted of the chromosome haploid number in Mangifera indica (n =20). The integrated genetic map covered the entire genome of Mangifera indica cv. Dashehari, with a total genetic distance of 2,982.75 cM

and an average distance between markers of 0.68 cM. The length of LGs varied from 85.78 to 218.28 cM, with a mean size of 149.14 cM. Phenotyping for fruit color and firmness traits was done for two consecutive seasons. We identified important consistent QTLs for 12 out of 20 traits, with integrated genetic linkages having significant LOD scores in at least one season. Important consistent QTLs for fruit peel color are located at Chr 3 and 18, and firmness on Chr 11 and 20. The QTLs mapped in this study would be useful in the marker-assisted breeding of mango for improved efficiency.

KEYWORDS

Color, firmness, fruit quality QTLs, *Mangifera indica* L., molecular linkage map, SNP markers

Introduction

Mango (*Mangifera indica* L.) belongs to the plant family *Anacardiaceae* and has historical, religious, and economic importance. It is a diploid fruit tree with 20 chromosome pairs and a small haploid genome size of ~439 Mb (Arumuganathan and Earle, 1991). Cytogenetic analysis based on a partial allopolyploid genome for mango has also been suggested (Mukherjee, 1950). In the last two decades, enough evidence has been produced of the inheritance of genetic markers in a disomic fashion, which confirms the diploid nature of mango (Duval et al., 2005; Schnell et al., 2005; Viruel et al., 2005; Schnell et al., 2006; Singh et al., 2016; Kuhn et al., 2017).

World mango production was 57.37 million tons in 2020 from an area of 56.8 million hectares. The majority (76%) of the world production comes from Asia, followed by America (12%) and Africa (11.8%). Mango is commercially cultivated in 102 countries. India's share in the world's mango production is 41.6%, followed by a 10% share in China (Bally et al., 2021). In 2020, the global mango export volume was 2.3 million tons, accounting for a 3.2 billion USD export value (FAOSTAT, 2020). Indian share in the global mango export is very less accounting for <5.0% of the global trade in 2020.

Mango fruit size, firmness, color, and aroma are quality characteristics of this climacteric fruit that need to be investigated at the genomic level to improve mango fruit quality (Bally et al., 2021). These traits are important factors determining the suitability of cultivars for domestic as well as overseas markets. In general, consumers from America and Europe prefer attractive, red-colored fruits having a pleasant aroma, a blend of sweet and sour tastes, and moderate sweetness. However, consumers from Gulf countries and the Indian continent prefer very sweet aromatic mangoes. One of the major objectives in most mango breeding programs globally is the attractive peel color in hybrids, which makes the fruits more attractive and export worthy. In addition, fruit quality parameters like total soluble solids, acidity, sugars, carotenoid contents, flavor compounds, etc. are important traits that decide the superiority of a variety. Mango breeding programs targeting overseas markets consider red peel color and high fruit firmness as important traits to improve. Fruit firmness has great value in the transportation, storage, and processing of mango. However, the genetics of these traits in mango is poorly understood. Despite the recognized high quality of a few well-known mango cultivars, considerable cultivar improvement is needed in most regions of mango culture. Mangoes have a wider adaptability to both tropical and subtropical regions of the world (Singh et al., 2016). Mango originated in the South East Asian or Indo-Myanmar region, having 69 recognized species originating as forest trees with fibrous and resinous fruits (Kostermans and Bompard, 1993). Mango cultivation began at least 4,000 years ago in India (Mukherjee, 1953). The majority of commercial cultivars have originated as chance seedlings and occupied prominent places in domestic as well as overseas markets.

Although domestication and selection of mango varieties have occurred for thousands of years, the systematic breeding of mangoes is relatively recent. Mango is a difficult fruit species to handle in breeding programs due to inherent problems of long juvenility, high heterozygosity, polyembryony, and significant fruit drop, which result in low recovery of hybrid fruits. As a result, the development of meaningful mapping populations and their use in understanding the genetics of horticultural traits has been limited. In India, breeding efforts to develop mango varieties with desirable traits started seven decades ago at the ICAR-Indian Agricultural Research Institute (IARI), New Delhi, and have a notable history of varietal improvement. To date, 10 mango hybrids have been released for commercial cultivation. Mango hybrids, namely Amrapali and Mallika, identified in the 1970s, became the choice of growers at the domestic level and got the attention of mango breeders globally. Amrapali, being a highly regular, dwarf, profuse bearer with excellent fruit quality, is the preferred parent in mango breeding programs. Considering consumer demand and potential overseas markets for mild sweetness, attractive colorful peel, and pleasant aroma, the Floridian cultivar Sensation has been used as a male donor parent to impart the red peel color in Amrapali, which otherwise has light green peel at maturity. These efforts yielded several hundred full-sib bi-parental progeny populations showing phenotypic polymorphism for agronomic traits and served as a core resource for genetic studies in mango (Ramachandra et al., 2021). Full-sib hybrid populations from two known parents chosen for their horticultural traits are more

effective in the construction of genetic linkage maps and are considered a powerful tool to identify linkages between traits and markers for MAS (Ogundiwin et al., 2009; Martínez-García et al., 2013; Harel-Beja et al., 2015; Kuhn et al., 2017).

Despite huge economic significance, genomic resources for mango have been limited. The first genetic map of mango produced by Kashkush et al. (2001) was with a relatively low number of markers, which limited the accuracy and resolution of the resulting linkage map. The first draft of a whole genome sequence of mango was reported in India (Singh et al., 2014; Singh et al., 2016). A high-resolution map reported by Luo et al. (2016) may prove useful in genetic studies. Kuhn et al. (2017) reported a consensus genetic map based on seven populations and significant traits-markers association. However, the marker density across 20 LGs was relatively lower. In the current decade, a wealth of information on the mango genome has been generated (Singh et al., 2014; Luo et al., 2016; Singh et al., 2016; Kuhn et al., 2017; Wang et al., 2020). This advancement in mango genomics has opened a new vista and may contribute to future mango improvement programs. Our group led by Prof. Nagendra Kumar Singh made significant progress in whole genome sequencing of mango using next-generation sequencing technologies and identified millions of SNP markers from the genome sequence data (Mahato et al., 2016; Singh et al., 2016; Iquebal et al., 2017). An increase in the number of unbiased markers and a highly resolved genetic map are essential molecular tools for mango breeders (Kuhn et al., 2017). The demand for new improved cultivars having desirable quality traits is difficult to address by breeders only relying on traditional breeding techniques. The adoption of molecular genomic tools has the potential to identify markers associated with important horticultural traits and, in general, improve the efficiency of mango breeding programs.

In the present study, we generated a high-resolution integrated genetic map based on segregation from female, male, and both parents. This map may serve as a valuable resource that can be used to improve efficiency and overcome the challenges in mango breeding. We also used the genetic linkages to study the association of SNP marker(s) with fruit color and firmness traits. We report here the identification of significant QTLs governing the color and firmness of mango fruit. The findings of this study are novel and have significance in improving the breeding efficiency of mango.

Materials and methods

Mapping population

Ninety-two F_1 bi-parental hybrids obtained from a cross of Amrapali as a female parent and Sensation as a male parent along with the parents were used for the QTL mapping. These parents were selected based on their contrasting features for key fruit quality-related traits including fruit color and firmness. The hybridity of the progenies has been confirmed earlier using SSR markers (Ramachandra et al., 2021). The population was phenotyped for multiple years and shows considerable variation for different fruit quality traits (Figure 1). This mapping population is conserved at the field repository of ICAR-Indian Agricultural Research Institute, New Delhi, India.

Genotyping of the mapping population and parents

Genotyping of the mapping population and parents in duplicate was carried out using an Affymetrix mango genome 80K SNP genotyping chip array (MiSNPnks 96 array, Dr. N. K. Singh, unpublished) having a total of 18,816 genes belonging to five different categories of genes such as single-copy mango genes (SCM), conserved single-copy genes between citrus and mango (CSCCM), cloned horticulturally important mango genes (HIM), disease resistance defense response-like mango genes (DRDRM), and multi-copy genes (MCR). The average SNP density is ~6 per gene, which is highly significant in linkage mapping and QTL identification studies in mango.

FIGURE 1
Variations in mango fruit peel color in mapping population and parents. Fruit No. 1 showing location where peel color and firmness was measured: 1. shoulder, 2. middle, and 3. bottom.

DNA for genotyping was isolated from the young leaves of individual trees of the mapping population and parents following the CTAB method as modified by Ramachandra et al., 2021. Genomic DNA quality was checked by electrophoresis in 1% agarose gel and quantified using a nanodrop spectrophotometer (NanoDrop 2000 Spectrophotometer, Thermo Scientific, USA). For target probe preparation, 50 µl of genomic DNA with a concentration of 10 ng/ul was used according to Affymetrix Axiom® 2.0 Assay Manual. The DNA samples were pre-amplified using Target Prep Protocol QSCB1 (P/N 702990), fragmented and hybridized on the chip, followed by single-base extension through DNA ligation and signal amplification. The Affymetrix GeneTitan® platform was used for staining, washing, and scanning of the chip signals as per manufacturer's protocol. SNP allele calling was done using Axiom™ Analysis Suite version 2.0 using its three workflows, i.e., best practices, sample QC, genotyping, and summary on the Affymetrix Gene Titan. The Axiom Analysis Suite requires stored library files to convert CEL files into genotype calls. SNPs with low call rate across the samples were removed, and only good quality SNPs with a DQC of >0.85 and call rates of >95% were used for further analysis.

SNP data formatting

Genotype calls from all SNP markers generated by the Affymetrix GeneTitan® platform from 92 progenies and parents were appended into a single csv file for export to Excel. A total of 80,816 SNPs were amplified, out of which loci with >5% missing data were filtered out. Markers amplified in only one of the two replications of the parents were also removed. The SNP allelic patterns between parents were taken as a reference for allele assignment in the mapping population. Monomorphic SNP markers were removed because they would not be informative for finding recombination events. Further, homozygous SNP markers for different alleles between the parents (aa, bb, or *vice versa*) were also removed, as there would be no segregation for such markers in the F_1 population. Thus, a total of 32,916 informative polymorphic SNP markers were identified between Amrapali and Sensation (Table 1).

Linkage mapping

Based on analysis of allelic variation between parents for each polymorphic locus, SNP markers were classified as per the expected Mendelian segregation ratio in the mapping population. This set of SNP markers were mapped on the Dashehari physical map (unpublished) to assign their location on the chromosomes, and only those which mapped back on the chromosomes were retained. Polymorphic SNPs identified between parents represented multiple SNPs per gene. Therefore, only single SNPs per gene were selected (Tables 1, 2).

Linkage analysis was performed using homozygous SNP markers in one parent and heterozygous in the other parent (lm × ll or nn × np) and heterozygous in both parents (hk × hk) using JoinMap version 4.1 (Kyazma, Wageningen, The Netherlands). Further, the selection of SNP markers was based on their disomic inheritance and chi-square test of goodness of fit ($p > 0.05$). Initially, for the construction of individual female and male maps, 4,834 markers in lm × ll and hk × hk, and 3,964 markers in nn × np and hk × hk were used, respectively. The initial grouping of markers was based on the independence log of the odds (LOD) tests in step ranging from 1.0 to 7.0. Other parameters were set as default. The regression mapping algorithm and Kosambi's mapping function with minimum LOD of 3.0 were used for calculating the marker's order. Markers showing suspect linkages were excluded in phases. Integration of female and male maps in which genotypes of some or all loci were determined in both populations took place, and the data from the separate populations were combined to calculate the integrated map. The groups that related to the same LG with at least two loci in common were combined by using the combined groups from the map integration function of the Join menu. The recombination frequencies and LOD scores of the selected sets of loci were combined into a combined group node in the navigation tree.

Such a combined group node is identical to a group node of a pairwise data population. The map calculations are based on mean recombination frequencies and combined LOD scores. For each pair of loci, the numbers of recombinant and non-recombinant gametes in the individual populations were calculated from the estimated recombination frequencies and corresponding LOD scores. The total numbers of recombinant and non-recombinant gametes of overall populations were calculated by totaling the numbers of the individual populations. From this, the mean recombination frequency and the combined LOD score were obtained. For map integration, the regression mapping algorithm was used.

Fruit quality measurement

Matured fruits from the F_1 trees of Amrapali/Sensation population and parents were carefully harvested with a 2 cm

TABLE 1 SNP genotyping data formatting.

Type	SNPs
Total SNPs	80,816
SNPs (>95% call rate)	67,188
SNPs homozygous for same allele (aa/aa or bb/bb)	27,226
SNPs homozygous for different alleles (aa/bb or bb/aa)	7,046
SNPs heterozygous in one parent and homozygous in other or heterozygous in both parents (ab/bb or aa/ab or ab/ab)	32,916

TABLE 2 Details of polymorphic SNP markers mapped on Dashehari physical map.

Reference genome	Segregation	Allelic pattern	Expected segregation	SNPs	SNPs mapped	One SNP/ gene	SNPs used for linkage mapping
Dashehari	Amrapali	ab/aa or bb	1:1	14,763	11,631	6,269	3,317
	Sensation	aa or bb/ab	1:1	10,435	8,450	4,886	2,447
	Amrapali/ Sensation	ab x ab	1:2:1	7,718	6,097	3,554	1,517
				32,916	26,178	14,709	7,281

pedicel portion. The fruits were selected from all direction of the tree for evaluating fruit color and firmness traits for two consecutive years, 2019 and 2020. Fruits were washed thoroughly to remove the adhering dirt and dust, and rolled over the blotting paper to remove extra moisture on the surface and air dried. Fruit maturity was determined after harvest, and fruits having a specific gravity of ~1.01 to 1.02 were selected. These fruits were then wrapped in kite paper and placed in wooden boxes to ripen uniformly at room temperature. Initially, 30 fruits per individual genotype were subjected to ripening, of which 12 fruits showing uniform ripening and that were free from damage were selected for further analysis.

Peel and pulp color in terms of L^*, a^* and b^* were measured using a Hunter-Lab Colorimeter (Model No. Miniscan® XE plus 4500 L, Hunter Associates Laboratory, Inc., VA, USA). The instrument (45 / 0 geometry, D 65 optical sensor, 10 observer) was calibrated with black and white reference tiles through the tri-stimulus values X, Y and Z, taking as standard values those of the white background (X = 79.01; Y = 83.96; Z = 86.76) tile. Fruit peel color was measured at three points on the fruit surface, i.e., shoulder, middle, and bottom (Figure 1). Mango pulp collected from ripe fruits was homogenized, and the color of homogenized pulp was measured using a ring and disk attachment.

The firmness of ripe mango fruits was measured with the help of the TA-XT Plus Texture Analyzer (Stable Micro Systems, UK). A 2.0 mm diameter stainless steel cylinder probe was used for the test in compression mode using the load cell of 5 kg capacity. Fruit firmness was expressed in Newton (N). Firmness was measured at three points (shoulder, middle, and bottom) of the fruits of each individual and parents (Figure 1) with pre-test, test, and post-test speeds of 5, 2, and 10 mm/s, respectively (Jha et al., 2006; Jha et al., 2010). During the compression of mango fruit by the cylinder probe, firmness was determined by the highest force recorded in the force-time curve recorded by the Texture Analyzer. The first peak in the texture curve was taken as peel firmness. The average force between the first and second anchors was used to calculate the flesh firmness. Fruit color and firmness concerning individual progenies were observed under three replications having a minimum of three fruits per replication. The data on various parameters were subjected for Qstats analysis to know the basic quantitative statistics, *viz*., mean, variance, standard deviation, skewness, kurtosis, and average deviation. To test the normal distribution of traits in the mapping population, a test of normality was performed, and the critical values for rejection were 5.99 and 9.21 for the tests at the 5% and 9% levels of probability, respectively.

QTL mapping

Phenotypic data of fruit color and firmness for the bi-parental F_1 population was used for QTL mapping with an integrated linkage map using MapQTL®6 (Van Ooijen, 2009; Kyazma B.V.R, Wageningen, The Netherlands) using cross-pollinated (CP) for population type and Multiple QTL model-based MQM mapping for association statistics with mapping step size of 1 cM and regression function. Other calculation parameters were set with MapQTL default. The QTL statistics were reported for those in which the LOD score exceeded the threshold, and LOD peaks were used for determining the position of a significant QTL on chromosomes.

Results

Linkage maps of the 20 mango chromosomes

A total of 3,317 markers heterozygous in female (1:1) and 1,517 heterozygous in both parents (1:2:1) were used for the construction of the female map. Similarly, 2,447 heterozygous in male (1:1) and 1,517 heterozygous in both parents (1:2:1) were used for male linkage mapping.

Finally, two high-density individual linkage maps with 3,213 and 1,781 SNPs distributed on 20 LGs in each segregation category were constructed as female and male maps, respectively (Supplementary Table 1).

The female map had 3,213 markers on 20 chromosomes with an average of 160.65 markers/chromosome. It covered a total map distance of 2,844.39 cM with individual chromosomes ranging from 82.41 cM (Chr 15) to 222.19 cM (Chr 3) with an average length of 142.22 cM. The average interval ranged from 0.43 cM (Chr 15) to 2.50 cM (Chr 3), with an average interval of 1.01 cM. The number of markers on each chromosome ranged from 75 (Chr 17) to 262 (Chr 4). This map is highly dense, and the density of SNPs ranged from 0.40 per cM (Chr 3) to 2.34 per cM (Chr 15) with an average of 1.21 SNPs per cM (Table 3).

A total of 1,781 markers were successfully mapped on the 20 chromosomes in the male map with an average of 89.05 markers/ chromosome. The male map covered a total map distance of 2,684.22 cM. The individual chromosome length ranged from 62.55 cM (Chr 15) to 187.86 cM (Chr 6), with an average length of 134.21 cM. The number of markers ranged from 42 (Chr 7) to 208 (Chr 11). The average interval ranged from 0.64 cM (Chr 15) to

3.68 cM (Chr 6), with an average of 1.74 cM. The number of SNPs per cM ranged from 0.27 (Chr 6) to 1.57 (Chr 15), with an average of 0.69 markers per cM (Table 3).

Map integration was attempted using individual female and male linkage data, and a high-resolution integrated map comprising 4,361 markers mapped on the 20 chromosomes was constructed. Before the construction of integrated map, the LGs of individual maps were matched chromosome-wise, and a minimum of two markers common to individual maps were used for the construction of the integrated map. This map is highly dense, as the number of SNPs ranged from 108 (Chr 17) to 315 (Chr 11) with an average of 218.05 SNPs per chromosome. The integrated map covered a higher total map distance of 2,982.75 cM of mango genome compared to individual female and male maps. The individual chromosomes ranged from 85.78 cM (Chr 15) to 218.29 cM (Chr 4), with an average of 149.14 cM per chromosome. The average map interval on 20 chromosomes ranged from 0.34 cM (Chr 15) to 1.33 cM (Chr 3), with an average interval of 0.73 cM. The number of SNP markers per cM ranged from 0.75 for Chr 3 to 2.96 for Chr 15, with an average of 1.54 (Table 3; Figure 2).

QTLs for the fruit quality traits

Phenotypic data on fruit color and firmness traits were generated for the 92 bi-parental F_1 population for two consecutive years in 2019 and 2020 (Figures 3–5), and their mean value with genetic linkages observed in the integrated map was used for QTL mapping. Color coordinates a^*, b^*, and L^* indicate the red and yellow colors, and brightness on three points on the mango fruit in 2019, 2020, and the mean of two seasons were considered as individual traits. MQM mapping using MapQTL6 resulted in the identification of QTLs for 12 out of 20 phenotypic parameters analyzed for mapping with significant LOD scores in at least one season. Tables 4–6 show the 12 traits with significant LOD scores and their QTL position on the chromosomes.

TABLE 3 Genetic linkage mapping statistics.

Chr	Female map (Amrapali)			Male map (Sensation)			Integrated map		
	SNPs	Map length (cM)	Interval (cM)	SNPs	Map length (cM)	Interval (cM)	SNPs and distribute the width of columns equally	Map length (cM)	Interval (cM)
1	114	212.66	1.87	86	184.84	2.15	175	195.24	1.12
2	246	125.62	0.51	64	132.18	2.07	277	171.58	0.62
3	89	222.19	2.50	82	137.44	1.68	159	211.41	1.33
4	262	216.89	0.83	46	136.39	2.96	297	218.29	0.73
5	143	117.83	0.82	91	121.15	1.33	212	160.07	0.76
6	184	143.22	0.78	51	187.86	3.68	226	144.08	0.64
7	170	172.16	1.01	42	95.19	2.27	184	182.42	0.99
8	136	181.84	1.34	86	149.64	1.74	208	136.36	0.66
9	165	95.29	0.58	82	122.16	1.49	213	164.75	0.77
10	202	145.34	0.72	106	161.44	1.52	243	157.65	0.65
11	229	123.34	0.54	208	137.09	0.66	315	130.09	0.41
12	150	161.34	1.08	162	160.92	0.99	278	115.97	0.42
13	213	146.52	0.69	110	158.14	1.44	275	166.54	0.61
14	144	145.68	1.01	97	165.57	1.71	212	148.16	0.70
15	193	82.41	0.43	98	62.55	0.64	254	85.78	0.34
16	184	102.76	0.56	65	114.52	1.76	235	111.71	0.48
17	75	93.55	1.25	43	88.76	2.06	108	111.97	1.04
18	140	116.05	0.83	92	137.97	1.50	198	134.29	0.68
19	88	128.22	1.46	115	104.89	0.91	182	116.04	0.64
20	86	111.45	1.30	55	125.52	2.28	110	120.33	1.09
Total	3,213	2,844.39		1,781	2,684.22		4,361	2,982.75	
Mean	160.65	142.22	1.01	89.05	134.21	1.74	218.05	149.14	0.73

Peel and pulp color

Chromaticity coordinates observed on the shoulder, middle, and bottom portion of mango fruits revealed that expression of a^* indicative of red color on the shoulder is associated with Chr 3 of integrated genetic linkage map. Four significant QTLs were identified, one at a position of 49.18 cM (6.31 to 6.72 LOD) explains 28.2 to 29.2% of phenotypic variation, the second at a position of 73.79 cM (LOD 7.84 to 8.08) explaining 33.2 to 34.5% of phenotypic variance, the third QTL at a position of 98.75 cM (5.43 to 5.59 LOD) and fourth at 129.83 cM (LOD 5.13 to 5.73) explain around 25.0% of phenotypic variation in the population. These QTLs were observed in both seasons, i.e., 2019 and 2020, as well as with the mean phenotypic values over the years (Table 4; Figure 6A). However, expression of a^* at middle of fruit is associated with Chr 2, 12, and 17. In the present study, peel a^* value observed at fruit bottom did not result in the identification of any significant QTLs. This may be because the observed red blush on the shoulder differed significantly compared to the middle and bottom portions of the fruit. Expression of b^* presenting yellow color is associated with Chr 2, 14, 15, and 18. Results revealed that Chr 2 showed one QTL at a position of 85.75 cM (LOD 4.28; R2 20.1), and Chr 14 showed another QTL at a position of 77.15 cM, explaining 20.3% phenotypic variation in 2019. Similarly, Chr 15 showed one QTL in 2020 with mean value at a position of 28.75 cM and with a LOD score of 4.21 to 4.34, explaining 18.3 to 20.7% phenotypic variations in the population (Table 4). Two consistent QTLs identified on Chr 18 (79.42 and 83.20 cM) explain 19.9 to 25.2% phenotypic variations for b^* of fruit bottom (LOD 4.14-5.56) (Figure 6B). It was also noted that the b^* value observed at the fruit shoulder and middle did not yield any significant QTL. QTLs governing the brightness of fruit were located on Chr 2, 3, 4, 10, 15, and 17 (Table 4; Figure 6C). We did not observe any significant QTL(s) for pulp color in the present study.

Peel and pulp firmness

Peel firmness observed at shoulder and bottom of mango fruit was associated with SNPs located on Chr 6, 11, and 20. Total 12 SNPs on Chr 11 at position ranging from 27.78 to 95.19 cM (R2 16.3-20.4) were identified as having an association with peel firmness at fruit shoulder. Similarly, eight SNPs hosted on Chr 11 at a position 8.69 to 94.82 cM (LOD 3.51 to 4.83; R2 16.6-22.6) consistently appeared with traits observed on fruit bottom (Table 5). One QTL at a position of 18.15 cM (LOD 4.08-5.46; R2 19.6-24.5) on Chr 20 appeared consistently, with peel firmness observed on fruit shoulder and bottom in both seasons (Figure 7A). One minor QTL at Chr 6 at 19.23 cM (LOD 3.59; R2 16.6) was also identified. Fruit firmness measured at the fruit middle did not result in significant QTLs, while firmness observed at the shoulder and bottom of the fruit confirmed significant SNP association. However, average peel firmness of three different positions confirmed association of seven SNPs.

SNPs located on Chr 3, 4, 11, and 19 had an association with pulp firmness (Table 6). SNPs located on Chr 11 at a position of 94.82 to 95.19 cM (LOD 4.54-7.42; R2 20.4-32.4) consistently appeared for pulp firmness observed at three different positions on the fruit and with the mean value (Figure 7B). One QTL on Chr 3 at a position of 99.63 cM (LOD 4.54-4.93), explaining 20.8-23.2% of phenotypic variations, was also identified. Pulp firmness observed at the fruit bottom showed an additional four SNPs on Chr 19 at a position of 54.33-56.47 cM (LOD 4.50-5.33) and explains 20.4-24.8% of phenotypic variations in the population. QTL analysis using a year-wise mean of pulp firmness observed at

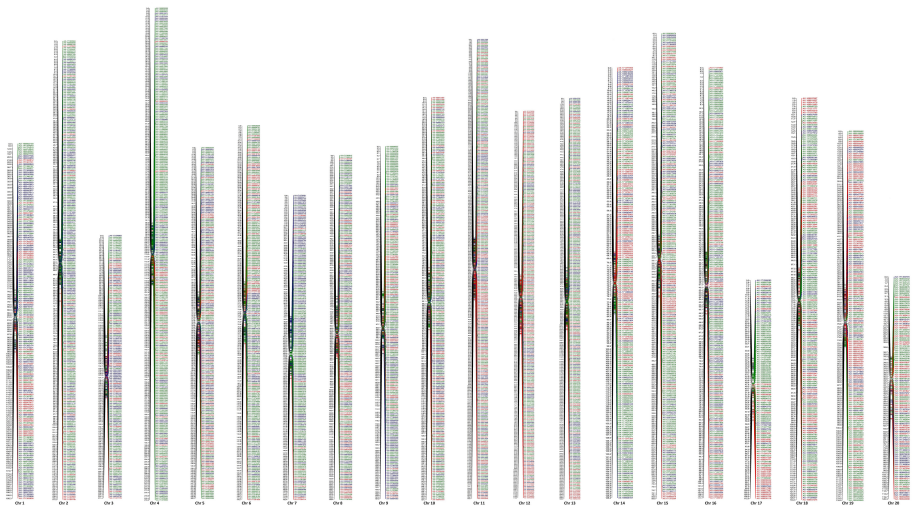


FIGURE 2
Integrated genetic linkage map and distribution of SNP markers on 20 chromosome, green color indicates marker segregating in 1:1 (Amrapali), red in 1:1 (Sensation), and blue in 1:2:1 (Amrapali/ Sensation).

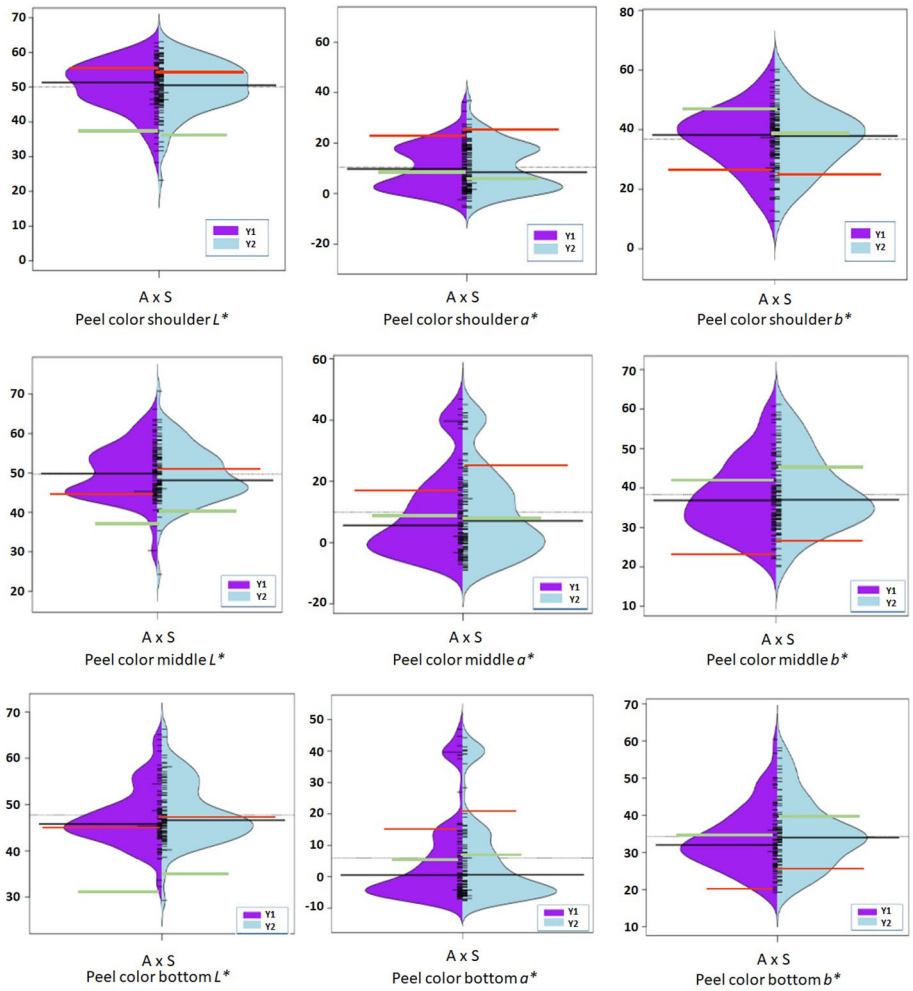


FIGURE 3
Violin-plot distribution and phenotype individual values (bars on Y axis) of peel color observed at shoulder, middle, and bottom portion of fruit in two years. Black bar median values. Green bar indicates 'Amrapali' female, and red bar indicates 'Sensation' male parental values. Symbol * used as standard symbol for chromaticity coordinates *L*, *a*, and *b*.

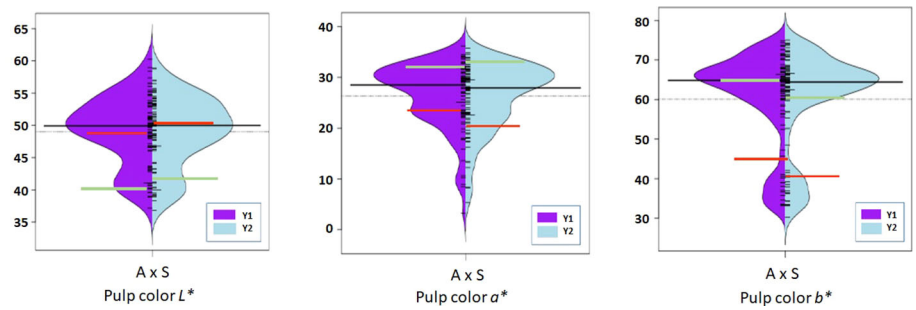


FIGURE 4
Violin-plot distribution and phenotype individual values (bars on Y axis) of fruit pulp color in two years. Black bar median values. Green bar indicates 'Amrapali' female and red bar indicates 'Sensation' male parental values. Symbol * used as standard symbol for chromaticity coordinates *L*, *a*, and *b*.

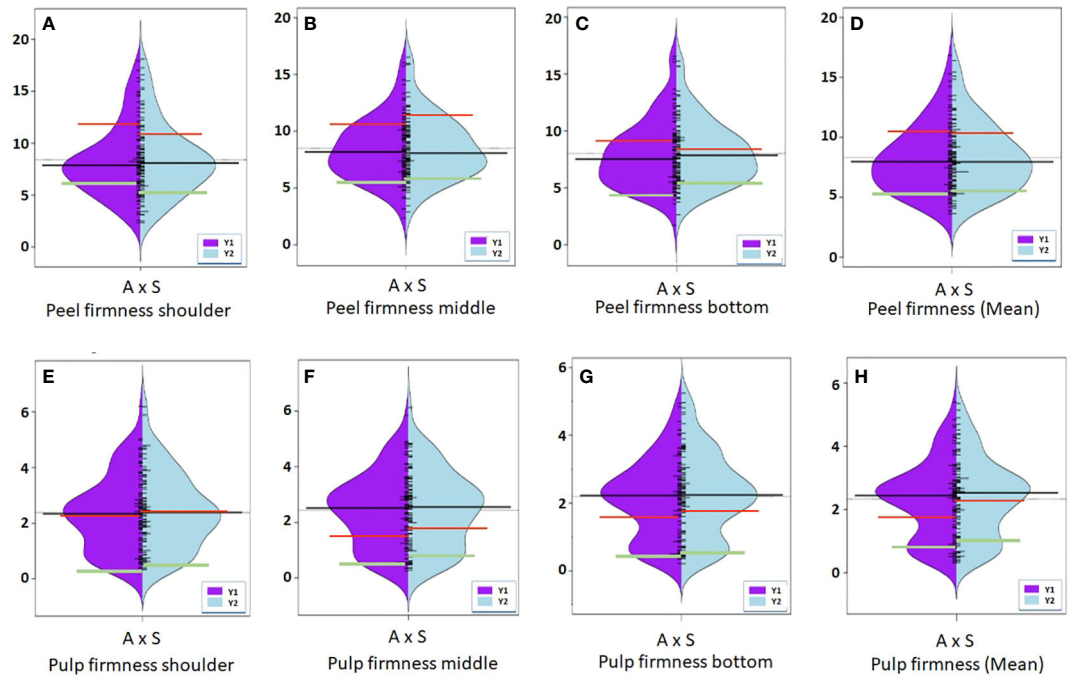


FIGURE 5
Violin-plot distribution and phenotype individual values (bars on Y axis) of peel firmness **(A–D)** and pulp firmness **(E–H)** measured at shoulder, middle, and bottom portion of fruit in two years. Black bar median values. Green bar indicates 'Amrapali' female, and red bar indicates 'Sensation' male parental values.

TABLE 4 Significant QTL(s) identified for mango peel color (*a**, *b**, *L**) using integrated map data.

Season	Chr	Position (cM)	Peaks	LOD	% Expl.
a* fruit shoulder					
2019	3	49.18	AX-171381971	6.61	29.2
2020	3	49.18	AX-171381971	6.31	28.2
Mean	3	49.18	AX-171381971	6.72	28.8
2019	3	73.79	AX-171379053	8.08	34.5
2020	3	73.79	AX-171379053	7.84	34.3
Mean	3	73.79	AX-171379053	7.98	33.2
2019	3	98.75	AX-171375294	5.59	25.4
2020	3	98.75	AX-171375294	5.43	25.2
Mean	3	98.75	AX-171375294	5.58	24.6
2019	3	129.83	AX-169929951	5.73	25.9
2020	3	129.83	AX-169929951	5.13	24.0
Mean	3	129.83	AX-169929951	5.63	24.8
a* fruit middle					
2019	2	8.10	AX-171382092	4.21	19.8
Mean	12	3.40	AX-171383356	4.06	18.6
2020	17	8.64	AX-171386135	4.03	19.4
Mean	17	8.64	AX-171386135	3.98	18.2

(Continued)

10.3389/fpls.2023.1135285

TABLE 4 Continued

Season	Chr	Position (cM)	Peaks	LOD	% Expl.
b* fruit bottom					
2019	2	85.75	AX-169902987	4.28	20.1
2019	14	77.15	AX-169888716	4.33	20.3
2020	15	28.75	AX-169875771	4.34	20.7
Mean	15	28.75	AX-169875771	4.21	18.3
2019	18	79.42	AX-171377288	5.56	25.2
2020	18	79.42	AX-171377288	4.14	19.9
Mean	18	79.42	AX-171377288	4.92	22.0
2019	18	83.20	AX-169935441	4.63	21.5
2020	18	83.20	AX-169935441	4.40	21.0
Mean	18	83.20	AX-169935441	4.89	21.9
L* fruit shoulder					
2019	4	123.72	AX-171379971	4.73	21.9
2020	4	123.72	AX-171379971	4.68	20.2
Mean	4	123.72	AX-171379971	4.33	19.7
2019	10	46.25	AX-169943716	4.84	22.4
2019	10	76.62	AX-171376164	4.62	21.5
2020	10	76.34	AX-171378162	4.56	20.6
Mean	10	76.23	AX-171375190	4.46	20.2
Mean	10	76.62	AX-171376164	4.93	22.1
2020	15	49.25	AX-169946547	4.57	20.6
Mean	15	49.25	AX-169946547	4.57	20.6
L* fruit middle					
2019	2	33.08	AX-171373991	4.44	20.7
2020	2	33.52	AX-169925508	4.47	21.3
Mean	2	33.08	AX-171373991	4.46	20.2
2019	3	75.36	AX-171376560	4.31	20.2
2019	4	123.72	AX-171379971	4.44	20.7
Mean	4	123.72	AX-171379971	4.41	20.0
2020	15	49.25	AX-169946547	4.78	22.6
Mean	15	49.25	AX-169946547	4.39	19.9
2020	17	29.23	AX-169910005	4.77	22.6
2020	17	30.57	AX-171383828	4.34	20.7

10.3389/fpls.2023.1135285

TABLE 5 Significant QTL(s) identified for mango peel firmness using integrated map data.

Season	Chr	Position (cM)	Peaks	LOD	% Expl.
Fruit shoulder					
2020	6	89.28	AX-171383358	3.72	18.1
Mean	11	27.78	AX-169943051	3.6	16.7
2019	11	29.49	AX-169928966	3.84	18.2
2020	11	29.49	AX-169928966	3.55	17.3
Mean	11	29.49	AX-169928966	4.13	18.9
2019	11	29.64	AX-169902238	3.70	17.6
Mean	11	29.64	AX-169902238	3.98	18.2
Mean	11	30.49	AX-171382227	3.51	16.3
2020	11	30.77	AX-171380495	3.56	17.4
Mean	11	30.77	AX-171380495	3.76	17.3
2019	11	30.79	AX-169900539	3.50	16.7
2020	11	30.79	AX-169900539	3.57	17.4
Mean	11	30.79	AX-169900539	3.76	17.3
Mean	11	31.81	AX-169901747	3.75	17.3
2019	11	32.33	AX-169938965	3.76	17.9
2020	11	32.33	AX-169938965	3.70	18.0
Mean	11	32.33	AX-169938965	4.03	18.4
2020	11	32.58	AX-169925476	3.73	18.1
Mean	11	32.58	AX-169925476	3.60	16.6
2020	11	94.82	AX-169878695	4.25	20.4
Mean	11	94.82	AX-169878695	3.75	17.3
2020	11	94.97	AX-169881574	4.03	19.4
2020	11	95.19	AX-171386999	3.69	17.9
2019	20	18.15	AX-171386901	4.17	19.6
2020	20	18.15	AX-171386901	4.08	19.6
Mean	20	18.15	AX-171386901	4.28	20.2
Fruit bottom					
Mean	6	19.23	AX-169909744	3.59	16.6
2019	11	8.69	AX-169936159	3.51	16.8
Mean	11	8.69	AX-169936159	3.69	17.0
2020	11	8.69	AX-169936159	3.89	18.8
2019	11	9.16	AX-169915899	3.85	18.2
Mean	11	9.16	AX-169915899	3.96	18.2
2020	11	9.16	AX-169915899	3.64	17.7
Mean	11	17.91	AX-169900237	3.59	16.6
Mean	11	21.29	AX-169937148	3.58	16.6
Mean	11	27.78	AX-169943051	3.62	16.7

(Continued)

TABLE 5 Continued

Season	Chr	Position (cM)	Peaks	LOD	% Expl.
2019	11	29.68	AX-169924307	4.18	19.7
Mean	11	29.68	AX-169924307	4.83	21.7
2020	11	29.68	AX-169924307	4.79	22.6
2020	11	60.95	AX-169896842	4.08	19.6
2020	11	94.82	AX-169878695	3.73	18.1
2019	20	18.15	AX-171386901	5.06	23.3
2020	20	18.15	AX-171386901	5.25	24.5
Mean	20	18.15	AX-171386901	5.46	24.1
Average of three positions (shoulder, middle, and bottom)					
2020	11	8.69	AX-169936159	3.87	18.7
Mean	11	8.69	AX-169936159	3.51	16.3
2019	11	9.16	AX-169915899	3.58	17.1
2020	11	9.16	AX-169915899	3.85	18.6
Mean	11	9.16	AX-169915899	3.89	17.9
Mean	11	21.29	AX-169937148	3.50	16.2
Mean	11	27.78	AX-169943051	3.43	15.9
2019	11	29.68	AX-169924307	3.85	18.2
2020	11	29.68	AX-169924307	4.17	20.0
Mean	11	29.68	AX-169924307	4.30	19.6
2019	11	94.82	AX-169878695	3.58	17.1
2020	11	94.82	AX-169878695	4.28	20.5
Mean	11	94.82	AX-169878695	3.81	17.5
2020	11	94.97	AX-169881574	4.02	19.4
2019	20	18.15	AX-171386901	4.51	21.0
2020	20	18.15	AX-171386901	4.49	21.4
Mean	20	18.15	AX-171386901	4.71	21.2

different positions of mango fruits confirmed three SNPs on Chr 11 (94.82 to 95.19 cM; LOD 5.47-7.48; R2 24.1-33.0).

Discussion

High density genetic maps from SNP markers

Traditional mango breeding is cumbersome and time-consuming. In the present decade, rapid advancement in DNA sequencing and molecular genetic techniques has generated a wealth of information on mango genomics. Apart from conventional breeding approaches, efforts have been made in the recent past to utilize biotechnological tools for marker-aided breeding. Significant progress has been made in the area of genome sequencing of mango by several researchers (Singh et al., 2014; Luo et al., 2016; Singh et al., 2016; Kuhn et al., 2017; Wang et al., 2020). The wealth of genome resources has also been generated by our group as genome sequences of important mango cultivars (unpublished), and millions of SNPs have been identified, which has immense value in future mango breeding (Singh et al., 2014; Mahato et al., 2016; Singh et al., 2016; Iquebal et al., 2017).

As a key step in genetic linkage mapping, the mapping population is highly important. F_1, F_2, backcross populations, and haploid populations can be used for genetic mapping (Wang, 2001). The F_1 population is an ideal mapping population for highly heterozygous tree species like mango and can be obtained by one-generation hybridization between contrasting parents with high heterozygosity (Lu et al., 2019). There are several examples where genetic maps have been constructed based on the F_1 population such as *Poncirus trifoliata* (Xu et al., 2021), *Ziziphus jujuba* (Wang

10.3389/fpls.2023.1135285

TABLE 6 Significant QTL(s) identified for mango pulp firmness using integrated map data.

Season	Chr	Position (cM)	Peaks	LOD	% Expl.
Fruit shoulder					
2019	11	94.82	AX-169878695	5.08	23.3
2020	11	94.82	AX-169878695	5.80	26.7
Mean	11	94.82	AX-169878695	5.59	24.6
2019	11	94.97	AX-169881574	4.62	21.5
2020	11	94.97	AX-169881574	4.99	23.4
Mean	11	94.97	AX-169881574	4.89	21.9
Fruit middle					
2019	3	99.63	AX-169945503	4.54	21.1
2020	3	99.63	AX-169945503	4.93	23.2
Mean	3	99.63	AX-169945503	4.61	20.8
2019	11	94.82	AX-169878695	7.36	32.0
2020	11	94.82	AX-169878695	7.33	32.4
Mean	11	94.82	AX-169878695	7.42	31.3
2019	11	94.97	AX-169881574	6.78	29.9
2020	11	94.97	AX-169881574	6.53	29.5
Mean	11	94.97	AX-169881574	6.64	28.5
2019	11	95.19	AX-171386999	5.75	26.0
2020	11	95.19	AX-171386999	5.46	25.3
Mean	11	95.19	AX-171386999	5.55	24.5
Fruit bottom					
2020	4	61.59	AX-169919574	4.55	21.6
Mean	4	61.59	AX-169919574	4.79	21.5
Mean	11	21.29	AX-169937148	4.59	20.7
2019	11	94.82	AX-169878695	6.99	30.6
2020	11	94.82	AX-169878695	6.91	30.9
Mean	11	94.82	AX-169878695	7.07	30.1
2019	11	94.97	AX-169881574	6.60	29.2
2020	11	94.97	AX-169881574	6.42	29.1
Mean	11	94.97	AX-169881574	6.50	28.0
2019	11	95.19	AX-171386999	5.61	25.5
2020	11	95.19	AX-171386999	5.53	25.6
Mean	11	95.19	AX-171386999	5.50	24.3
2019	19	54.33	AX-171385518	5.14	23.6
2020	19	54.33	AX-171385518	5.33	24.8
Mean	19	54.33	AX-171385518	5.19	23.1
2020	19	55.04	AX-169902763	4.55	21.6
Mean	19	55.04	AX-169902763	4.50	20.4

(Continued)

TABLE 6 Continued

Season	Chr	Position (cM)	Peaks	LOD	% Expl.
2020	19	56.33	AX-171384203	4.63	22.0
Mean	19	56.33	AX-171384203	4.61	20.8
2019	19	56.47	AX-169937533	4.55	21.2
2020	19	56.47	AX-169937533	4.80	22.6
Mean	19	56.47	AX-169937533	4.72	21.3
Average of three positions (shoulder, middle, and bottom)					
Mean	11	21.29	AX-169937148	4.63	20.9
2019	11	94.82	AX-169878695	6.97	30.6
2020	11	94.82	AX-169878695	7.48	33.0
Mean	11	94.82	AX-169878695	7.28	30.8
2019	11	94.97	AX-169881574	6.42	28.5
2020	11	94.97	AX-169881574	6.65	30.0
Mean	11	94.97	AX-169881574	6.49	28.0
2019	11	95.19	AX-171386999	5.47	24.9
2020	11	95.19	AX-171386999	5.61	26.0
Mean	11	95.19	AX-171386999	5.46	24.1
2020	19	54.33	AX-171385518	4.71	22.3

et al., 2019), *Persea americana* (Rendón-Anaya et al., 2019), *Vitis* (Zhu et al., 2018), *Coryus avellana* (Bhattarai and Mehlenbacher, 2017), and *Theobroma cocoa* (Argout et al., 2011). We chose to generate genetic linkage map using bi-parental F_1 population derived by crossing Amrapali/Sensation contrasting for fruit quality traits. Full-sib populations from two known parents are considered more effective for breeding progress than half-sib populations from open pollinated maternal parents. Genetic maps that are based on segregating full-sib hybrid populations are better for establishing linkages between horticultural traits and molecular markers for MAS (Kuhn et al., 2017).

Moreover, SNP marker-based genetic maps have several advantages, as SNPs are more abundant, easier to identify, easier to score, and unambiguous markers (Kuhn et al., 2017). Studies have shown that SNP loci are ubiquitous in the genome and the most abundant forms of genetic variation between individuals of the same species (Rafalski 2002; Kuhn et al., 2017). In this study, a mango 80K genic-SNP genotyping array was used to genotype the F_1 full-sib (Amrapali/Sensation) population along with parents.

High-density genetic maps are valuable in genetic and genomic studies, illuminating the genetic and molecular mechanisms of plants and providing the necessary framework for QTL analyses, gene cloning, and molecular breeding (Wang et al., 2019; Wu et al., 2019). We used 4,834 and 3,964 SNPs for construction of female and male maps. However, not all the SNPs that expected to segregate in a disomic fashion were able to be assigned to a linkage group. In female, 3,213 SNPs and in male 1,781 SNPs could be successfully mapped on 20 LGs. As the mango has 40 chromosomes with the haploid number of 20, we were successful in identifying 20 LGs for both female and male maps. This suggests the diploid nature of mango, and even though it is a partial allopolyploid, the two ancestral genomes were different enough to be distinguished by SNP markers (Kuhn et al., 2017). Genetic linkage data of female and male maps were used for integration, and a high-density linkage map was constructed, having 4,361 SNPs that covered the entire genome of mango. The map length covered by female, male, and integrated maps slightly differed for number of SNPs. The 3,213 SNPs of the female map covered 2,844.4 cM of the genome while 1,781 SNPs of the male map covered 2,684.2 cM. Integration of female and male maps resulted in 4,361 SNPs distribution across 20 chromosomes covering 2,982.8 cM of map length.

Few genetic maps in *Mangifera indica* have been constructed using different markers (Kashkush et al., 2001; Singh et al., 2014; Luo et al., 2016; Singh et al., 2016; Kuhn et al., 2017; Wang et al., 2020). The integrated genetic linkage map reported here is of high density, as the number of markers is reasonably high. In this study, we used a strategy to make the map that took advantage of the strengths of markers segregating differently. Further, markers corresponding to the same chromosome from all segregation groups were integrated. Our genetic map is significantly better than previous maps, as it is highly dense, based on full-sib (Amrapali/Sensation) population, and covered the entire mango genome.

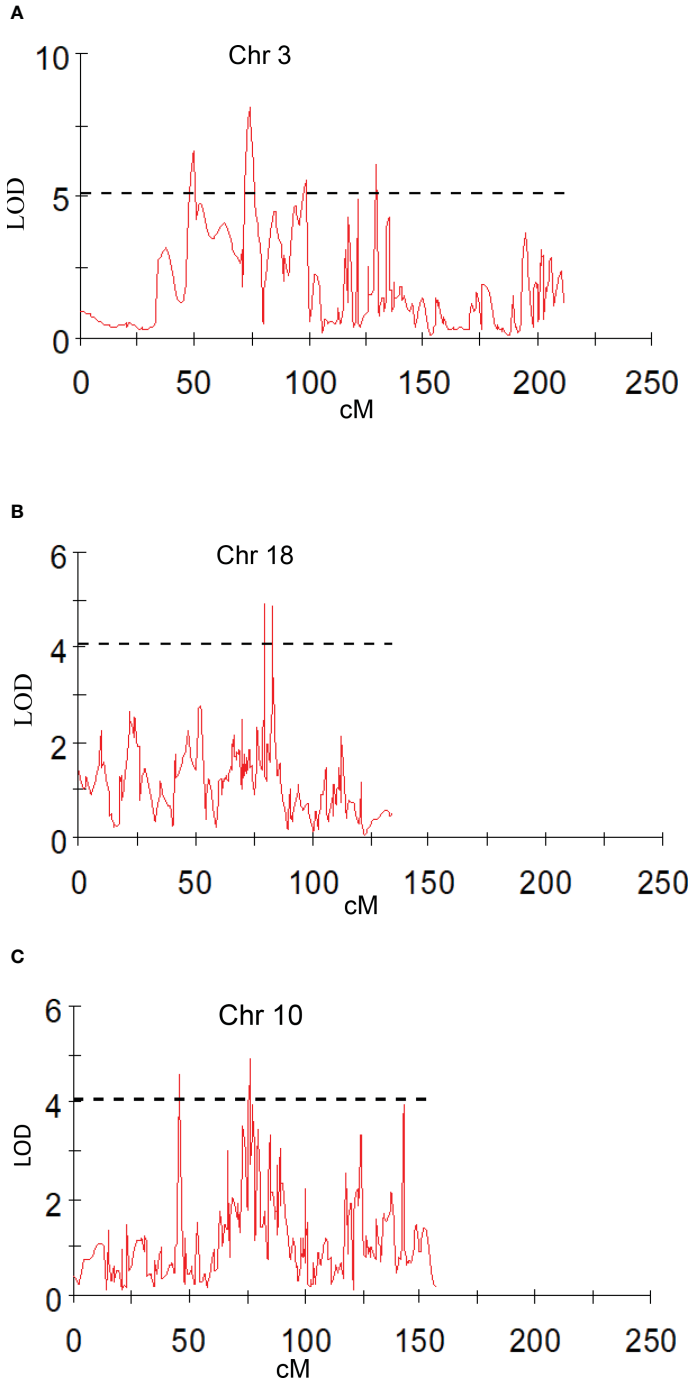


FIGURE 6
Graphs of the plot of the likelihood of the odds that a SNP marker is associated with the trait. **(A)** peel *a** shoulder of fruit; **(B)** peel *b** bottom of fruit; **(C)** peel *L** shoulder of fruit in 2019.

Trait association to the genetic maps

Marker assisted selection provides ample opportunity to reduce the mango breeding cycle and improve efficiency as well. In recent decade, the advancements made in the augmentation of genome resources in mango (Kashkush et al., 2001; Luo et al., 2016; Singh et al., 2016; Kuhn et al., 2017; Wang et al., 2020) provide unprecedented opportunities to breeders for MAS in mango. Mango fruit is adored by people for its taste and nutrition, contributed by color, flavor, and aroma. Among these, peel and pulp color are important traits contributing to fruit quality and market value (Bajpai et al., 2017). Fruit firmness is another important trait for storage, transportation, and disease and insect management. Mapping populations from controlled crosses are not easy to generate in mango due to the high level of technical proficiency required (Bally et al., 2021). The F_1 bi-parental progeny population studied is few in fruit

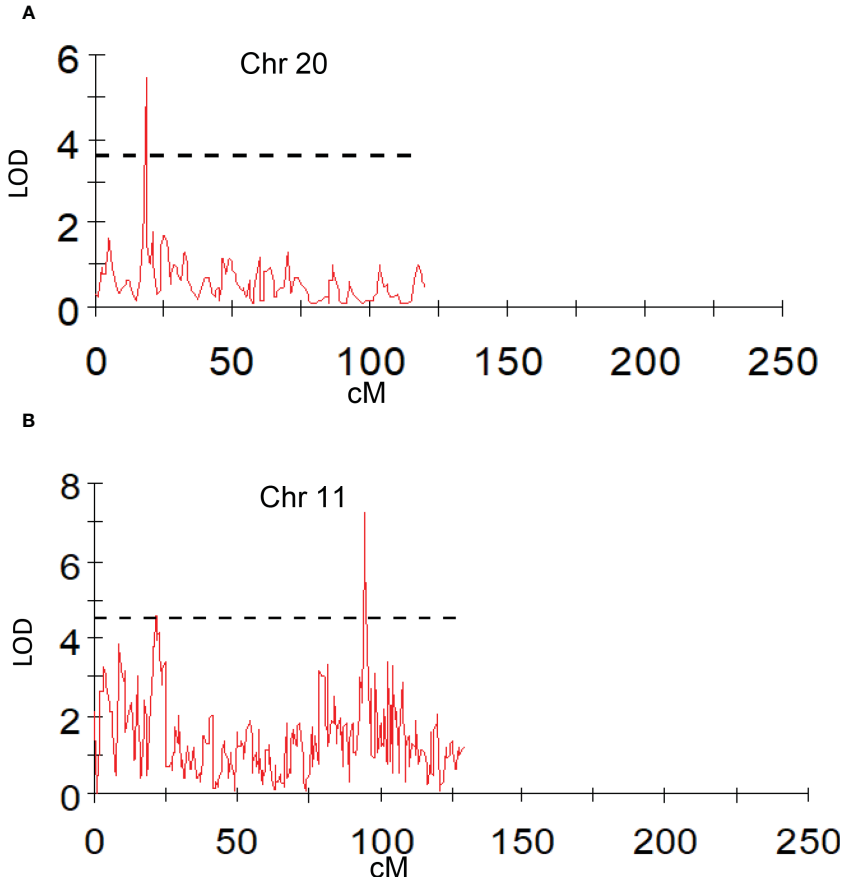


FIGURE 7
Graphs of the plot of the likelihood of the odds that a SNP marker is associated with the trait. **(A)** peel firmness observed at fruit bottom; **(B)** pulp firmness at fruit shoulder (mean of 2019 and 2020).

trees compared to annual crops (Grattapaglia and Sederoff, 1994). In the present study, we attempted to elucidate the regions of mango genome influencing fruit color and firmness using a F_1 bi-parental (Amrapali/Sensation) mapping population. To be useful for marker-aided breeding, it is imperative to have markers showing strong association with the horticultural traits. A map is not necessary to identify markers associated with a trait, but confidence in this association increases as multiple markers near the trait locus on the genetic map also show significant association with the trait (Kuhn et al., 2017). Fruit quality-related traits are quantitative traits that are influenced by multiple genes, and perhaps no single gene shows a significant impact on a trait. In our study, MQM mapping was used. This method has also been used in QTL analyses for quality traits in *Gossypium hirsutum* (Zhang et al., 2019), *Poncirus trifoliata* (Xu et al., 2021), and *Elymus sibiricus* (Zhang et al., 2019). In this study, we could identify SNP association for 12 of 20 traits used for analysis.

Peel and pulp color

Fruit color is a highly variable trait within the fruit and is much influenced by environmental conditions. Precise qualitative evaluation of peel color is influenced by the number of fruits examined and from which part of the tree it is collected. More randomly chosen fruits from all parts of a tree may reduce this variation. In addition, we used another approach for determining the peel color using the Hunter color meter and took multiple observations on different positions of the fruit. This approach provided us with quantifiable data more suitable for QTL mapping compared to data based on scales or grouping. It was evident that red blush was more towards the shoulder compared to the middle and bottom portions of the fruit. An attractive fruit color is one of the most important factors for export markets (Nambi et al., 2016). The accumulation of pigments and their concentration and intensity determine the overall appearance of color in mango fruits (Pervaiz et al., 2017). Red coloration in fruits and other plant tissues has multifarious roles such as conferring plant disease resistance and protection against UV radiation (Bieza and Lois, 2001; Berardini et al., 2005; Sivankalyani et al., 2016). Anthocyanins also provide human health benefits against cancer and cardiovascular and other chronic diseases (Rao and Rao, 2007; Butelli et al., 2008; Singh et al., 2008). We found that fruit peel a^* indicating red blush on the shoulder was associated with Chr 3. Four QTLs (AX-171381971, AX-171379053, AX-171375294, and AX-169929951) with high LOD value of 5.13-8.08 explaining around 35% of phenotypic variation were identified. The a^* value observed on the middle of fruit resulted in associated SNPs located on Chr 2, 12, and 17. However, a^* observed on the bottom portion of the fruit did not result in any QTL(s) in the present study. This may be because the appearance of red blush is more significant on the shoulder

compared to the middle and bottom portion of the fruits. The SNPs associated with peel yellow color (b^*) trait was on Chr 2, 14, 15, and 18. Mango fruits are classified based on peel color into green, yellow, and red types. Mango peel turns from green to yellow or red or retains green colors during ripening. Carotenoids and anthocyanins are the important pigments responsible for the colors of fruits. Pigments in different proportions may have an influence on the expression of different shades of color on mango peel. The genetics of peel color has not been studied in detail, but available reports indicate that it is governed by several genes and regulated in a more complex manner. In agreement with this, we observed several SNP markers housed on different chromosomes that had an influence on peel color development in mango. Consistency of these QTLs over phenotypic values observed in different seasons confirms their involvement in the expression of fruit color.

Peel and pulp firmness

Mango varieties differ considerably for fruit peel and pulp firmness. Fruit firmness is a significant quality aspect of mango for consumers, as it represents ripeness and influences shelf-life, transportation, and processing issues (Jha et al., 2010). Various industries use puncture tests as part of their quality control procedure. Cool store operators monitor firmness throughout the storage period as part of their inventory management. In addition, firmness is sometimes used to predict consumer responses. Stec et al. (1989) found that preferred firmness at eating ripeness varied among assessors (Watkins and Harman, 1981; Harker et al., 1996). Jha et al. (2006) reported that the firmness of the mango fruits remained almost constant over the period of growth, and it decreased after attaining maturity. It was also suggested that the maturity of mango could be predicted by measuring size, color, and firmness. We used the approach of determining the peel and pulp firmness on different positions of fruits. We observed several SNPs associated with the peel and pulp firmness of mango fruits housed on Chr 3, 4, 6, 11, 19, and 20. SNPs on Chr 11 and Chr 20 consistently appeared in all seasons of observation and mean value of peel firmness. SNPs located on Chr 3, 4, 11, and 19 had an association with pulp firmness. SNPs located on Chr 11 at a position of 94.82 to 95.19 cM consistently appeared in different seasons. One QTL on Chr 3 at a position of 99.63 cM was also identified. The work reported here is unique concerning peel and pulp firmness results, where QTLs are reported for the first time in mango.

Conclusions

We demonstrated the usefulness of designing a mapping population from two commercially important mango cultivars, Amrapali and Sensation, that have phenotypic polymorphism for fruit peel color and firmness traits. The integrated genetic recombination map using segregation data of female, male, and both parents reported here is unique and reasonably resolved. Our analysis of the association of SNPs for fruit color and firmness traits is novel and enables us to formulate mango breeding strategies that can improve breeding efficiency in identifying desirable progeny with optimal horticultural traits. The information about genetic linkages and QTLs generated would be highly useful in mango breeding and for broadening the understanding of the genetics of these traits. This knowledge will allow breeders to design trait-specific breeding strategies in mango.

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

MS, NKS, SKS, RS, and NSi conceptualized the project and organized the work plan for the manuscript. MS, SR, PJ, NSi, SaS, and AM contributed to genotyping. MS, SR, AG, RD, NR, SHS, GK, AN, and ShS contributed to phenotyping. MS, NR, GT, PJ, SJ, MI, RS, and PP contributed to data analysis. AK, SRG, and ShSe extended the lab facilities for phenotyping. NSh helped in manuscript writing. All authors contributed to the article and approved the submitted version.

Funding

This work was funded by Department of Biotechnology, Ministry of Science and Technology, Govt. of India under the project 'Identification of QTL(s) for fruit quality trait(s) in mango (*Mangifera indica* L.) Project No. BT/PR27727/AGIII/103/989/2018)' Indian Council of Agricultural Research under the Network Project on Functional Genomics & Genetic Modification (Mango) (NPFGGM- Mango-2001-3064), and SERB-JC Bose National Fellowship to NKS (JCB/2022/000004).

Conflict of interest

The authors declare that the research was conducted in absence of any commercial or financial relationship 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.2023.1135285/full#supplementary-material

References

Argout, X., Salse, J., Aury, J. M., Guiltinan, M. J., Droc, G., Gouzy, J., et al. (2011). The genome of *Theobroma cacao*. *Nat. Genet.* 43, 101–108. doi: 10.1038/ng.736

Arumuganathan, K., and Earle, E. D. (1991). Nuclear DNA content of some important plant species. *Plant Mol. Biol. Rep.* 9, 208–218. doi: 10.1007/BF02672069

Bajpai, A., Khan, K., Muthukumar, M., Rajan, S., and Singh, N. K. (2017). Molecular analysis of anthocyanin biosynthesis pathway genes and their differential expression in mango peel. *Genome* 61, 157–166. doi: 10.1139/gen-2017-0205

Bally, I. S. E., Bombarely, A., Chambers, A. H., Cohen, Y., Dillon, N. L., Innes, D. J., et al. (2021). The 'Tommy atkins' mango genome reveals candidate genes for fruit quality. *BMC Plant Biol.* 21, 108. doi: 10.1186/s12870-021-02858-1

Berardini, N., Knodler, M., Schieber, A., and Carle, R. (2005). Utilization of mango peels as a source of pectin and polyphenolics. *Innovative Food Sci. Emerging Technol.* 6, 442–452. doi: 10.1016/j.ifset.2005.06.004

Bhattarai, G., and Mehlenbacher, S. A. (2017). Development and characterization of tri-nucleotide simple sequence repeat markers in hazelnut (L.). *PloS One* 12 (5), e0178061. doi: 10.1371/journal.pone.0178061

Bieza, K., and Lois, R. (2001). An arabidopsis mutant tolerant to lethal ultraviolet-b levels shows the constitutively elevated accumulation of flavonoids and other phenolics. *Plant Physiol.* 126, 1105–1115. doi: 10.1104/pp.126.3.1105

Butelli, E., Titta, L., Giorgio, M., Mock, H. P., Matros, A., Peterek, S., et al. (2008). Enrichment of tomato fruit with health-promoting anthocyanins by expression of select transcription factors. *Nat. Biotechnol.* 26, 1301–1308. doi: 10.1038/nbt.1506

Duval, M. F., Bunel, J., Sitbon, C., and Risterucci, A. M. (2005). Development of microsatellite markers for mango (*Mangifera indica* l.). *Mol. Ecol. Notes* 5, 824–826. doi: 10.1111/j.1471-8286.2005.01076.x

FATSTAT (2020). https://www.fao. org/ faostat/en/#data/TCL

Grattapaglia, D., and Sederoff, R. (1994). Genetic linkage maps of eucalyptus grandis and *Eucalyptus urophylla* using a pseudo-testcross: mapping strategy and RAPD markers. *Genetics* 137, 1121–1137. doi: 10.1093/genetics/137.4.1121

Harel-Beja, R., Sherman, A., Rubinstein, M., Eshed, R., Bar-Ya'akov, I., Trainin, T., et al. (2015). A novel genetic map of pomegranate based on transcript markers enriched with QTLs for fruit quality traits. *Tree Genet. Genomes* 11, 1–18. doi: 10.1007/s11295-015-0936-0

Harker, F. R., Maindonald, J. H., and Jackson, P. J. (1996). Measurement of apple and kiwifruit texture: operator and instrument differences. *J. Am. Soc. Hortic. Sci.* 121, 927–936. doi: 10.21273/JASHS.121.5.927

Iquebal, M. A., Jaiswal, S., Mahato, A. K., Jayaswal, P. K., Angadi, U. B., Kumar, N., et al. (2017). *MiSNPDb*: a web-based genomic resources of tropical ecology fruit mango (*Mangifera indica* l.) for phylogeography and varietal differentiation. *Sci. Rep.* 7, 14968. doi: 10.1038/s41598-017-14998-2

Jha, S. N., Kingsly, A. R. P., and Chopra, S. (2006). Physical and mechanical properties of mango during growth and storage for determination of maturity. *J. Food Eng.* 72 (1), 73–76.

Jha, S. K., Sethi, S., Srivastav, M., Dubey, A. K., Sharma, R. R., Samuel, D. V. K., et al. (2010). Firmness characteristics of mango hybrids under ambient storage. *J. Food Eng.* 97, 208–212. doi: 10.1016/j.jfoodeng.2009.10.011

Kashkush, K., Jinggui, F., Tomer, E., Hillel, J., and Lavi, U. (2001). Cultivar identification and genetic map of mango (*Mangifera indica*). *Euphytica* 122, 129. doi: 10.1023/A:1012646331258

Kostermans, A. J. G. H., and Bompard, J. M. (1993). *The mangoes: their botany, nomenclature, horticulture, and utilization* (London, UK: Academic Press).

Kuhn, D. N., Bally, I. S. E., Dillon, N. L., Innes, D., Groh, A. M., Rahaman, J., et al. (2017). Genetic map of mango: a tool for mango breeding. *Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.00577

Lu, N., Zhang, M., Xiao, Y., Han, D., Liu, Y., Zhang, Y., et al. (2019). Construction of a high-density genetic map and QTL mapping of leaf traits and plant growth in an interspecific F1 population of *Catalpa bungei* × *Catalpa duclouxii* dode. *BMC Plant Biol.* 19, 596. doi: 10.1186/s12870-019-2207-y

Luo, C., Shu, B., Yao, Q., Wu, H., Xu, W., and Wang, S. (2016). Construction of a high-density genetic map based on large-scale marker development in mango using specific-locus amplified fragment sequencing (SLAF-seq). *Front. Plant Sci.* 7. doi: 10.3389/fpls.2016.01310

Mahato, A. K., Sharma, N., Singh, A., Srivastav, M., Singh, S. K., Singh, A. K., et al. (2016). Leaf transcriptome sequencing for identifying genic-SSR markers and SNP heterozygosity in crossbred mango variety amrapali (*Mangifera indica* l.). *PloS One* 11, e0164325. doi: 10.1371/journal

Martínez-García, P. J., Parfitt, D. E., Ogundiwin, E. A., Fass, J., Chan, H. M., Ahmad, R., et al. (2013). High density SNP mapping and QTL analysis for fruit quality characteristics in peach (*Prunus persica* l.). *Tree Genet. Genomes* 9, 19–36. doi: 10.1007/s11295-012-0522-7

Mukherjee, S. K. (1950). Mango: its allopolyploid nature. *Nature* 166, 196–197. doi: 10.1038/166196b0

Mukherjee, S. K. (1953). The mango-its botany, cultivation, uses, and future improvement, especially as observed in India. *Economic Bot.* 7, 130–162.

Nambi, V. E., Thangavel, K., Shahir, S., and Chandrasekar, V. (2016). Color kinetics during ripening of Indian mangoes. *Int. J. Food Prop.* 19 (10), 2147–2155. doi: 10.1080/10942912.2015.1089281

Ogundiwin, E. A., Peace, C. P., Gradziel, T. M., Parfitt, D. E., Bliss, F. A., and Crisosto, C. H. (2009). A fruit quality gene map of prunus. *BMC Genomics* 10, 587. doi: 10.1186/1471-2164-10-587

Pervaiz, T., Songtao, J., Faghihi, F., Haider, M.S., and Fang, J. (2017). Naturally occurring anthocyanin structure, functions and biosynthetic pathway in fruit plants. *J. Plant Biochem. Physiol.* 5 (2), 187. doi: 10.4172/2329-9029.1000187

Rafalski, J. A. (2002). Novel genetic mapping tools in plants: SNPs and LD-based approaches. *Plant Sci.* 162, 329–333.

Ramachandra, S., Srivastav, M., Singh, S. K., Mahato, A. K., Singh, N., Arumugam, N., et al. (2021). New genomic markers for marker assisted breeding in mango (*Mangifera indica* l.). *J. Horti Sci. Biotech.* 96 (5), 624–633. doi: 10.1080/14620316.2021.1906760

Rao, A. V., and Rao, L. G. (2007). Carotenoids and human health. *Pharmacol. Res.* 55, 207–216. doi: 10.1016/j.phrs.2007.01.012

Rendón-Anaya, M., Ibarra-Laclette, E., and Méndez, A. (2019). The avocado genome informs deep angiosperm phylogeny, highlights introgressive hybridization, and reveals pathogen-influenced gene space adaptation. *PNAS* 116 (34), 17081–17089. doi: 10.1073/pnas.182212911

Schnell, R. J., Brown, J. S., Olano, C. T., Meerow, A. W., Campbell, R. J., and Kuhn, D. N. (2006). Mango genetic diversity analysis and pedigree inferences for Florida cultivars using microsatellite markers. *J. Am. Soc. Hortic. Sci.* 131, 214–224.

Schnell, R. J., Olano, C. T., Quintanilla, W. E., and Meerow, A. W. (2005). Isolation and characterization of 15 microsatellite loci from mango (*Mangifera indica* l.) and cross-species amplification in closely related taxa. *Mol. Ecol. Notes* 5, 625–627. doi: 10.1111/j.1471-8286.2005.01018

Singh, M., Arseneault, M., Sanderson, T., Murthy, V., and Ramassamy, C. (2008). Challenges for research on polyphenols from foods in alzheimer's disease: bioavailability, metabolism, and cellular and molecular mechanisms. *J. Agr. Food Chem.* 56, 4855–4873. doi: 10.1021/jf0735073

Singh, N. K., Mahato, A. K., Sharma, N., Gaikwad, K., Srivastava, M., Tiwari, K., et al. (2014). *"A draft genome of the king of fruit, mango (Mangifera indica L.),"*, , in Plant and animal genome XXII conference (San Diego, CA, USA). (New York: Academic Press)

Singh, N. K., Mahato, A. K., Jayaswal, P. K., Singh, A., Singh, S., Singh, N., et al. (2016). "A draft genome of the king of fruit, mango (*Mangifera indica* l.)" in Plant and animal genome XXII conference (San Diego, CA, USA). (New York: Academic Press). *Indian J. History Sci.* 51, 2. doi: 10.16943/ijhs/2016/v51i2.2/48449

Sivankalyani, V., Feygenberg, O., Diskin, S., Wright, B., and Alkan, N. (2016). Increased anthocyanin and flavonoids in mango fruit peel are associated with cold and pathogen resistance. *Postharvest Biol. Technol.* 111, 132–139. doi: 10.1016/j.postharvbio.2015.08.001

Stec, G. H., Hodgson, J. A., MacRae, E. A., and Triggs, C. M. (1989). Role of fruit firmness in the sensory evaluation of kiwifruit (*Actinidia deliciosa* cv Hayward). *J. Sci. Food Agric.* 47, 417–433. doi: 10.1002/jsfa.2740470404

Van Ooijen, J. W. (2009). MapQTL6, Software for the mapping of quantitative trait loci in experimental population of diploid species. B. V. R Kyazma, (The Netherlands: Wageningen)

Viruel, M. A., Escribano, P., Barbieri, M., Ferri, M., and Hormaza, J. I. (2005). Fingerprinting, embryo type and geographic differentiation in mango (*Mangifera Indica*L.) *Anacardiaceae Breed.*) with microsatellites. *Mol. Breed* 15, 383–393. doi: 10.1007/s11032-004-7982-x

Wang, M. (2001). *Forest genetic and breeding* (Beijing, China: China Forestry Publishing House).

Wang, P., Luo, Y., Huang, J., Gao, S., Zhu, G., Dang, Z., et al. (2020). The genome evolution and domestication of tropical fruit mango. *Genome Bio.* 21, 60. doi: 10.1186/s13059-020-01959-8

Wang, Z., Zhang, Z., Tang, H., Zhang, Q., Zhou, G., and Li, X. (2019). High-density genetic map construction and QTL mapping of leaf and needling traits in *Ziziphus jujuba* mill. *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.01424

Watkins, C., and Harman, J. (1981). Use of penetrometer to measure flesh firmness of the fruit. *Orchardist New Z.* 54, 14–16.

Wu, D., Koch, J., Coggeshall, M., and Carlson, J. (2019). The first genetic linkage map for *Fraxinus pennsylvanica* and syntenic relationships with four related species. *Plant Mol. Biol.* 99, 251–264. doi: 10.1007/s11103-018-0815-9

Xu, Y., Liu, S., Gan, Z., Zeng, R., Zhang, J., and Hu, C. (2021). High-density genetic map construction and identification of QTLs controlling leaf abscission trait in. *Poncirus trifoliata. Int. J. Mol. Sci.* 22, 5723. doi: 10.3390/ijms22115723

Zhang, K., Kuraparthy, V., Fang, H., Zhu, L., Sood, S., and Jones, D. C. (2019). High-density linkage map construction and QTL analyses for fiber quality, yield, and morphological traits using cotton SNP 63K array in upland cotton (*Gossypium hirsutum* l.). *BMC Genomics* 20, 889. doi: 10.1186/s12864-019-6214-z

Zhang, Z., Xie, W., Zhang, J., Wang, N., Zhao, Y., Wang, Y., et al. (2019). Construction of the first high-density genetic linkage map and identification of seed yield-related QTLs and candidate genes in *Elymus sibiricus*, an important forage grass in qinghai-Tibet plateau. *BMC Genomics* 20, 861. doi: 10.1186/s12864-019-6254-4

Zhu, J., Guo, Y., Su, K., Liu, Z., Ren, Z., Li, K., et al. (2018). Construction of a highly saturated genetic map for *Vitis* by next-generation restriction site-associated DNA sequencing. *BMC Plant Biol.* 18, 347. doi: 10.1186/s12870-018-1575-z

frontiers | Frontiers in Plant Science

TYPE Review
PUBLISHED 14 August 2023
DOI 10.3389/fpls.2023.1123631

OPEN ACCESS

EDITED BY
Nisha Singh,
Gujarat Biotechnology University, India

REVIEWED BY
Anuj Kumar,
Dalhousie University, Canada
Priyanka Jain,
Amity University, India

*CORRESPONDENCE
Pusarla Susmitha
pusarlasushmita94@gmail.com
Te Ming Tseng
tt1024@msstate.edu

†These authors have contributed equally to this work

RECEIVED 14 December 2022
ACCEPTED 08 June 2023
PUBLISHED 14 August 2023

CITATION
Susmitha P, Kumar P, Yadav P, Sahoo S, Kaur G, Pandey MK, Singh V, Tseng TM and Gangurde SS (2023) Genome-wide association study as a powerful tool for dissecting competitive traits in legumes. *Front. Plant Sci.* 14:1123631. doi: 10.3389/fpls.2023.1123631

Genome-wide association study as a powerful tool for dissecting competitive traits in legumes

Pusarla Susmitha[1*†], Pawan Kumar[2†], Pankaj Yadav[3], Smrutishree Sahoo[4], Gurleen Kaur[5], Manish K. Pandey[6], Varsha Singh[7], Te Ming Tseng[7*] and Sunil S. Gangurde[8]

[1]Regional Agricultural Research Station, Acharya N.G. Ranga Agricultural University, Andhra Pradesh, India, [2]Department of Genetics and Plant Breeding, College of Agriculture, Chaudhary Charan Singh (CCS) Haryana Agricultural University, Hisar, India, [3]Department of Bioscience and Bioengineering, Indian Institute of Technology, Rajasthan, India, [4]Department of Genetics and Plant Breeding, School of Agriculture, Gandhi Institute of Engineering and Technology (GIET) University, Odisha, India, [5]Horticultural Sciences Department, University of Florida, Gainesville, FL, United States, [6]Department of Genomics, Prebreeding and Bioinformatics, International Crops Research Institute for the Semi-Arid Tropics, Hyderabad, India, [7]Department of Plant and Soil Sciences, Mississippi State University, Starkville, MS, United States, [8]Department of Plant Pathology, University of Georgia, Tifton, GA, United States

Legumes are extremely valuable because of their high protein content and several other nutritional components. The major challenge lies in maintaining the quantity and quality of protein and other nutritional compounds in view of climate change conditions. The global need for plant-based proteins has increased the demand for seeds with a high protein content that includes essential amino acids. Genome-wide association studies (GWAS) have evolved as a standard approach in agricultural genetics for examining such intricate characters. Recent development in machine learning methods shows promising applications for dimensionality reduction, which is a major challenge in GWAS. With the advancement in biotechnology, sequencing, and bioinformatics tools, estimation of linkage disequilibrium (LD) based associations between a genome-wide collection of single-nucleotide polymorphisms (SNPs) and desired phenotypic traits has become accessible. The markers from GWAS could be utilized for genomic selection (GS) to predict superior lines by calculating genomic estimated breeding values (GEBVs). For prediction accuracy, an assortment of statistical models could be utilized, such as ridge regression best linear unbiased prediction (rrBLUP), genomic best linear unbiased predictor (gBLUP), Bayesian, and random forest (RF). Both naturally diverse germplasm panels and family-based breeding populations can be used for association mapping based on the nature of the breeding system (inbred or outbred) in the plant species. MAGIC, MCILs, RIAILs, NAM, and ROAM are being used for association mapping in several crops. Several modifications of NAM, such as doubled haploid NAM (DH-NAM), backcross NAM (BC-NAM), and advanced backcross NAM (AB-NAM), have also been used in crops like rice, wheat, maize, barley mustard, etc. for reliable marker-trait associations (MTAs), phenotyping accuracy is equally important as genotyping. Highthroughput

genotyping, phenomics, and computational techniques have advanced during the past few years, making it possible to explore such enormous datasets. Each population has unique virtues and flaws at the genomics and phenomics levels, which will be covered in more detail in this review study. The current investigation includes utilizing elite breeding lines as association mapping population, optimizing the choice of GWAS selection, population size, and hurdles in phenotyping, and statistical methods which will analyze competitive traits in legume breeding.

KEYWORDS

breeding, genomic selection, linkage, mapping, phenotyping, protein

Introduction

The term legume originated from the Latin word "legumen", which denotes "seeds harvested in pods". During the Neolithic Revolution, which marked the beginning of human farming methods, farmers were accompanied by legumes that belong to the family Fabaceae. It is acknowledged that inadequate protein-energy intake and micronutrient deficits are two of the primary causes of undernutrition. Legumes play a minor but significant role in our food system. They are the superior economical dietary solutions due to their rich protein content (17%–30%) and relevant nutritional richness compared to expensive food sources containing animal-based protein and dairy products that may be difficult to obtain in situations where there is food insecurity (Marinangeli et al., 2017).

Compared with cereals, legumes provide a substantial quantity of protein throughout the complete plant, notably in grains. The incorporation of leguminous crops in cropping systems enabled an enhancement in soil quality (Hasanuzzaman et al., 2020). Legumes' ability to fix atmospheric nitrogen in symbiotic relationships with soil bacteria such as Rhizobium and Brady rhizobium minimizes the requirement for chemical fertilizers during crop growth and contributes to a reduction in greenhouse gas emissions like nitrous oxide (N_2O) and carbon dioxide (CO_2). In addition, they can help to reduce the utilization of fossil-based energy inputs in the chain of agriculture and food production by infusing high-quality organic matter, facilitating nutrient circulation, and promoting water retention in the soil (Stagnari et al., 2017). Legumes are rich in nutraceuticals, such as vitamin B6, calcium, magnesium, sodium, zinc, copper, and manganese. Thus, it is crucial to expand the genetic background and foster the breeding of legume crops, which will serve the needs of the growing human population under changing climatic conditions. Therefore, it is essential to come up with high-yielding cultivars that have enhanced resistance to diseases, higher nitrogen fixation ability, and tolerance to abiotic and biotic stresses, which can be achieved using biotechnological and genomics-assisted breeding approaches.

Genome-wide association study (GWAS) is an effective technique for determining the genes underlying a particular trait. To accomplish this, it is ideal to assess the genomic regions where genotypic and phenotypic variations are correlated with each other. In comparison to standard biparental populations, GWAS offers greater mapping precision for detecting interactions among molecular markers and desirable characteristics in a variety of crops (Liu et al., 2016; Cui et al., 2017). It has become a vital tool in agricultural genetics due to its techniques that build upon the mixed linear model (MLM) framework and deliver radically improved computational speed and statistical power.

Furthermore, improvements can be applied in fields like omic-wide association studies, which utilize GWAS techniques to analyze relationships among desirable morphological traits and other kinds of omics data that include transcriptomic or metabolomic. GWAS requires structuring the population of diverse panels to estimate genetic distinction and minimize the detection of spurious connections (Sul et al., 2016). Breeders can develop new varieties owing to recent innovations in NGS applications and technologies that enable advanced tools to characterize genetic variation at a high resolution (Gali et al., 2019). The ultimate objective of this review is to quantify the genetic diversity, GWASs, and other related aspects or techniques that could be used to break the plateau of yield in legume crop production and can be utilized for further crop improvement.

Mapping population in association studies

Association mapping (AM), an alternative to QTL mapping, is dependent on linkage disequilibrium (LD) and uses collections of genotypes with known or unknown ancestry that have a significant degree of genetic variation due to hundreds of recombination cycles. The ultimate goal of association studies is to find a strong correlation between a genome-wide DNA marker and an interesting attribute that can be highly useful in marker-assisted selection for crop development. GWAS and candidate gene (CG)–based analysis are two important approaches to AM.

The creation of a mapping population that will be tested for the marker–trait relationship is a prerequisite for the GWAS. Both broad-based natural populations and narrow-based breeding populations can be utilized as the mapping population for GWAS

(Figure 1). The sort of mapping population needed for the success of GWAS hangs significantly on the mode of pollination (inbreeding or outbreeding) of the plant species. Both natural diverse germplasm panels and family-based breeding populations can be used for this. Among the breeding population, both biparental and multiparental mapping populations such as Multiparent Advanced Generation Inter-Cross ([MAGIC), Multiline Cross Inbred Lines (MCILs), Recombinant Inbred Advanced Intercross Lines (RIAILs), Nested Association Mapping (NAM), and Random Open- parents Association Mapping (ROAM) are being used for AM in several crop plants. Populations such as doubled haploid NAM (DH-NAM), backcross NAM (BC-NAM), and advanced backcross NAM (AB-NAM) that are modifications of NAM have also been used in recent times. The selection of the mapping population should be taken care of enough to avoid the false-positive marker–trait association. Because of the problematic inconsistent phenotyping scores of segregating lines over the years and location, heterozygote segregating individuals should not be included with the inbred lines as one population when creating the AM panel. When significant features like days to bloom and maturity are influencing the target trait, extreme genotypes should be eliminated from the AM panel for proper scoring of trait data (Kulwal and Singh, 2021). Each population has unique virtues and flaws, which will also be discussed further in the review study.

Natural population and elite breeding lines as association mapping population

Any naturally occurring panmictic population with a significant history of recombination events can undergo AM. Utilizing hundreds of recombination events makes it simple to do an LD analysis of the target characteristic. These populations, however, are not appropriate for QTL mapping. When a germplasm accession collection represents the natural population, it may be a core collection or a sample that is more resilient to environmental changes. The population is excellent for assessing the QTLs for rare alleles that can help develop elite breeding lines or highly heritable domestic features. QTLs for some agronomically key characteristics have been uncovered in germplasms of several crops using GWAS, such as in 135 pea accessions (Gali et al., 2019), 366 sesame accessions (Cui et al., 2017), and 119 accessions in rice (Pawar et al., 2021).

The cultivars and lines created by a deliberate breeding program are known as elite inbred lines. These lines are unbreakable and can be maintained by numerous researchers in various places to identify QTLs using an AM panel. For instance, two AM panels of maize having 306 dent corn and 292 European flint corn inbred lines were individually assessed using single-nucleotide polymorphism (SNP) markers in the cold and control growth chamber conditions to identify genes related to cold tolerance (Revilla et al., 2016). For GWAS research in sorghum, AM panels of 377 tropical accessions from various geographic and climatic zones, significant U.S. breeding lines, and the wild species have been brought together to be used as AM panels (Casa et al., 2008). GWASs in legumes mostly include the natural populations and elite advanced breeding lines (Table 1), whereas GWAS using artificial mapping populations is more or less a recent phenomenon, and they are still underway.

Biparental mapping population and association mapping

Recombinant inbred lines (RILs) and Near Isogenic Lines (NILs) are the most used biparental population, usually used for linkage mapping or Quantitative Trait Locus (QTL) or QTL

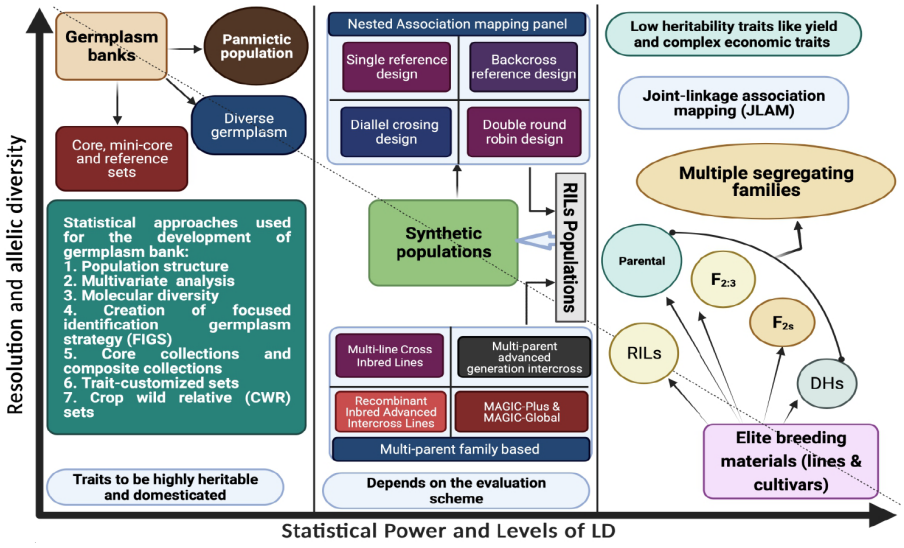


FIGURE 1
Types of mapping population used in GWAS studies along with their brief properties.

TABLE 1 GWAS studies for various traits in different leguminous crops.

Crop	Mapping population	Traits	QTLs/Marker trait associations	References
Gram	132 varieties and Advanced Breeding Lines (ABLs)	Yield traits	38 MTAs (marker trait association)	Li et al. 2018
	192 desi & kabuli accessions	Seed weight	8 MTAs	Bajaj et al., 2016
	182 diverse genotypes	Phenological, physiological and yield traits	14-34 MTAs in different environment condition	Jha et al. 2021
	75 ABLs	Fusarium wilt	3 MTAs	Jha et al. 2021
	165 chickpea genotypes	resistance to *Ascochyta rabiei*	30 MTAs	Farahani et al. 2022
	280 accessions	Grain Nutrient and Agronomic Traits	20 and 46 MTAs for grain nutrient and agronomic traits, respectively	Srungarapu et al., 2022
Arhar	Diverse collection of 142 pigeonpea lines	Flowering related traits	22MTAs	Kumar et al. 2022
	Pangenome based on 89 accessions	9 agronomic traits	229 MTAs	Zhao et al. 2020
Faba beans	481 elite breeding lines	Agronomic Traits	30 MTAs	Keller et al., 2020
Lentil	188 lines of the USDA Lentil Core Collection	Pea aphid	15 candidate genes	Das et al. 2022
Pea	135 pea accessions	Agronomic and Seed Quality Traits	251 MTAs	Gali et al., 2019
	135 pea accessions	Heat and Drought Adaptive Traits	15 MTAs	Tafesse et al. 2021
Mungbean	127 test genotypes	Mungbean yellow mosaic India virus resistance	15 MTAs	Singh et al. 2020
	95 cultivated mung bean genotypes	Seed Mineral content	43 MTAs	Wu et al. 2020
Blackgram	100 diverse genotypes	Agronomic traits	42 QTLs	Singh et al. 2022
	99 diverse genotypes	Agronomic traits	83 MTAs	Nkhata et al. 2021
Soybean		Protein, Oil, unsaturated fatty acid, oleic acid	SNP	(Hwang et al., 2014; Zhang et al., 2019; Zhao et al., 2019; Liu et al., 2020)
		Nematode resistance, Iron deficiency and Canopy wilt, brown stem rot, Diseases resistance	SNP	(Butenhoff, 2015; Vuong et al., 2015; Chang et al., 2016; Rincker et al., 2016; Zhang et al., 2017a; Zhang et al., 2017b; Zatybekov et al., 2018; Do et al., 2019; Ravelombola et al., 2019; Tran et al., 2019; Che et al., 2020; Lin et al., 2020)

(Continued)

TABLE 1 Continued

Crop	Mapping population	Traits	QTLs/Marker trait associations	References
		Salt tolerance, Flood tolerance, Drought tolerance, Water Use Efficiency	SNP	(Dhanapal et al., 2015; Zeng et al., 2017; Chen et al., 2018; Khan et al., 2018; Yu et al., 2019; Assefa et al., 2020)
		Agronomic trait	SNP	(Wen et al., 2015; Zhang et al., 2015; Contreras-Soto et al., 2017; Yan et al., 2017; Zhang et al., 2021; Zatybekov et al., 2017; Pan et al., 2018; Hu et al., 2019a; Li et al., 2019; Kim et al., 2020)
		Physiological traits	SNP	(Sui et al., 2020; Wang et al., 2020; Yang et al., 2020)
Groundnut	170 genotypes	Quality traits	SNP	(Shaibu et al., 2019a)
	125 ICRISAT groundnut mini core collection	Physiological traits	SNP	(Shaibu et al., 2019b; Shaibu et al., 2020)
	158 peanut accessions; 195 peanut accessions	Agronomic traits	SNP	(Zhang et al., 2017c; Wang et al., 2019)
	120 genotypes	Disease resistance	SNP	(Zhang et al., 2019; Zhang et al., 2020)
	249 peanut accessions	Abiotic stress tolerance	SNP	(Zou et al., 2020)
Chickpea		Agronomic traits	SNP	(Bajaj et al., 2015a; Bajaj et al., 2016; Kujur et al., 2015a; Upadhyaya et al., 2015; Upadhyaya et al., 2017; Basu et al., 2018; Orsak et al., 2019; Fayaz et al., 2022; Srungarapu et al., 2022)
		Abiotic stress tolerance	SNP	(Thudi et al., 2014; Li et al., 2018; Kohli et al., 2020; Ahmed et al., 2021; Kalve et al., 2022)
		Physiological traits	SNP	(Basu et al., 2019)
		Quality traits		(Upadhyaya et al., 2016a; Upadhyaya et al., 2016b; Parida et al., 2017; Samineni et al., 2022)
		Biotic stress resistance		(Li et al., 2017; Agarwal et al., 2019; Agarwal et al., 2022; Farahani et al., 2022; Raman et al., 2022)
Beans		Agronomic traits/ Quality	SNP	(Warsame et al., 2019; Rasool et al., 2022)
		Abiotic stress	SNP	(Ali et al., 2016; Li et al., 2017; Hu et al., 2019b; Breria et al., 2020; Abou-Khater et al., 2022; Maalouf et al., 2022; Sallam et al., 2022)
		Biotic stress	SNP	(Faridi et al., 2021)
Lentils		Agronomic traits	SNP, SSR	(Kumar et al., 2018a; Kumar et al., 2018b; Singh et al., 2019; Karthika et al., 2021; Neupane et al., 2022)
		Quality traits, Seed quality	SNP	(Khazaei et al., 2017; Khazaei et al., 2018; Johnson et al., 2021; Hang, 2022; Puspitasari et al., 2022)
		Biotic stress	SNP	(Banoo et al., 2020; Gela et al., 2021)
		Abiotic stress	SSR	(Singh et al., 2017; Kumar et al., 2019; Ma et al., 2020)

mapping. Whereas the power of QTL identification is higher in linkage mapping as compared to AM, the resolution has a reverse relationship with both mapping schemes. The concept of joint linkage AM (JLAM) was introduced to fully exploit the capabilities of both mapping methods. JLAM uses either a biparental population set or one or more multiparental AM panels, or two sets of genotypes consisting of germplasm and biparental mapping populations, which are genotyped utilizing the same set of markers (Myles et al., 2009; Lu et al., 2010; Reif et al., 2010 and Wurschum et al., 2012). Hence, JLAM is also recognized as integrated mapping that identifies more significant marker–trait associations and increases the power of AM. Using JLAM (by combining germplasm accessions and full-sib F2 population of a bioenergy crop Shrub willow (*Salix* sp.) identified several major QTLs along with QTL hotspots (Carlson et al., 2019). Several studies using JLAM include QTL identification and CG identification for drought tolerance in maize (Lu et al., 2010), pleiotropic QTLs for silique length and seed weight in rapeseed (Li et al., 2014), and the epistatic QTLs for agronomically important characters in sugarbeet (Reif et al., 2010). Recent studies claim that

regulating population structure and addressing rare alleles can be accomplished through cofactors and a demographic effect accounting for JLAM, which enhances the predictive power of the methods (Wurschum et al., 2012).

Multiparent mapping population for GWAS

The multiparent populations include several founder parents, which reflect wider genetic diversity. Hence, in AM studies, the use of multiparent mapping populations helps limit the demerit of recombination frequency in biparental populations. Multiparent mapping populations provides tools to control population structure and balance allele frequencies. The historical and artificial recombinational events of the multiparent mapping populations such as NAM and MAGIC populations and their derivatives increase the efficiency of QTL identification in AM. Because of the controlled crosses, NAM population has higher power because of maximized population structure and minimal familial relatedness and accumulated frequency of rare alleles. The population facilitates cost-effective GWAS and allows the perpetual sharing of the NAM panel with global researchers.

To generate sets of RILs, NAM populations can be developed using reliable mating strategies such as diallel mating, NCD-II (North Carolina design II), eight-way cross, and single/double round robin. NAM population was first developed in maize using RILs developed from a diverse set of parents. Twenty-five diverse families in maize were used to develop 5,000 RILs that were evaluated for southern leaf blight disease resistance (Kump et al., 2011), and the wide diversity helped in the identification of 32 QTLs for the trait. A NAM population was developed using 23 different inbreds of barley in a twofold round-robin design to identify QTLs and CGs for grain morphology (Shrestha et al., 2022). NAM population has been established in both autogamous and allogamous species such as barley, rice, wheat, sorghum, and maize (Maurer et al., 2015; Bajgain et al., 2016). Several modifications of NAM, such as DH-NAM, BC-NAM, and AB-NAM, have also been used in recent times. An AB-NAM of barley consists of 796 $BC_2F_{4:6}$ lines, which were derived from 25 wild barley accessions by backcrossing to the cultivar Rasmusson (Nice et al., 2016). Using 384 SNP markers, the AB-NAM population with minimal undesirable and unadapted characteristics of the wild barley parents was genotyped and encountered 10 QTLs for grain protein content (Nice et al., 2016).

A MAGIC population is created by a group of RILs from a complex cross or a group of crosses with numerous parents. Multiple rounds of recombination occur as these populations mature, improving the accuracy of desirable recombination and desirable alleles, thereby increasing the resolution of QTL mapping. With the aid of single seed descent (SSD), highly homozygous lines will be developed to establish the MAGIC population. To develop MAGIC populations for wheat and rice that can be deployed for QTL mapping, indica and japonica lines have been adopted. Seven cycles of SSD selfing resulted in 305 F8 lines in cowpea (*Vigna unguiculata*) (Huynh et al., 2018). In the MAGIC indica rice population, 400 lines from S2 bulk were chosen on the basis of agronomic attributes and evaluated in mega-environment trials to select elite lines (Bandillo et al., 2013).

New high-throughput genotyping technologies for plants

The molecular markers are being progressively used to expedite breeding efforts in the post-genome sequencing era. Modern plant breeding is shifting from classical breeding to molecular breeding, where various genotyping technologies are being used for the discovery of molecular markers. In the last decade, a huge number of molecular markers were used for structural analysis of large germplasm populations to understand the diversity and use in GWAS. The whole-genome sequencing for most of leguminous crops has already been completed. Chromosome-level genomes are completed for most of the leguminous crops (Varshney et al., 2013). In the pre-genome sequencing era, the simple sequence repeat (SSR) markers were very powerful and potentially used for GWAS analysis. SSRs are tandem repeats highly polymorphic, abundant, co-dominant, and distributed throughout the genome. However, SSR markers are very laborious and time-consuming when compared with modern genotyping platforms such as double-digest restriction site–associated DNA sequencing (ddRAD-Seq) or specific locus amplified fragment sequencing (SLAF-Seq), whole-genome resequencing (WGRS), genotyping-by-sequencing (GBS), SNP-chip arrays, diversity array technology (DArT) array technology. With Illumina, gigabases of DNA sequencing data may be generated in a short period and cost-effectively in the NGS era (Bentley et al., 2008), Roche (Rothberg and Leamon, 2008), and AB-SOLiD (Pandey et al., 2008).

Molecular markers have become crucial components in molecular breeding over the past 2 years (Nadeem et al., 2018; Horst and Wenzel, 2007; Eathington et al., 2007). Molecular breeding has gained popularity and has been accepted by plant scientists because of its rapid and precise results for germplasm classification, back cross-breeding, and marker-assisted selection (Kumar et al., 2011; Nair and Pandey, 2021). A plethora of studies has been done using molecular markers (Kujur et al., 2015a; Wu et al., 2010; Song et al., 2013; Deokar et al., 2014; Shao et al., 2022). Different types of markers have been used for genotyping of legume, which includes rapid amplified polymorphic DNA (RAPD) (Doldi et al., 1997; Thompson et al., 1998; Iruela et al., 2002; Talebi et al., 2008), amplified fragment length polymorphism (AFLP) (Nguyen et al., 2004; Singh et al., 2008; Ude et al., 2002), inter-SSR (ISSR) (Yadav et al., 2014; Bhagyawant and Srivastava 2008; Iruela et al., 2002; Souframanien and Gopalakrishna, 2004), and SSR (Saxena et al., 2010; Choudhary et al., 2012; Zavinon et al., 2020).

However, continuous improvement of next-generation sequencing (NGS) technologies in recent years has made it cost-effective and accessible for any crop, including legumes (Poland et al., 2012). Reference genome sequencing has been completed in some legume crops like soybean, pigeon pea, groundnut, cowpea,

chickpea, and common bean (Afzal et al., 2022; Salgotra and Stewart, 2022). The currently available NGS technologies sequence each molecular or base pair of the DNA of any organism and make it feasible for us to identify the number of SNP markers with high precision and in a very short period (Liew et al., 2004). Although SNPs are biallelic and their polymorphism information is much lower compared to SSRs, they cover a significantly large part of the genome, which makes them markers to go for GWASs. In the last decade, a plethora of genotyping studies were carried out using SNPs in chickpea (Kujur et al., 2015b; Gaur et al., 2012; Hiremath et al., 2012; Deokar et al., 2014), pigeon pea (Raju et al., 2010; Singh et al., 2016; Arora et al., 2017), groundnut (Varshney, 2016; Pandey et al., 2017; Abady et al., 2021), soybean (Wu et al., 2010; Song et al., 2013; Shao et al., 2022), and other legume crops (Bohra et al., 2021; Shilpa and Lohithaswa, 2021).

PCR-based genotyping methods

Amplification of DNA segments with PCR leads to the development of multiple genotyping methods. If the primers in a PCR reaction include the variation of interest, then it is called as allele-specific PCR. Allele-specific markers are generally used during foreground selection during marker assisted selection. PCR-Restriction Fragment Length Polymorphism (RFLP) is another method of PCR-based genotyping (Saiki et al., 1985), where the genomic region of interest is PCR-amplified using the markers and then digested with restriction enzymes specifically recognize a DNA sequence, so that the digested product can produce alleles of different size, which can distinguish among the individuals. Microsatellites or short tandem repeat polymorphisms are ideal markers for PCR-based genotyping as the length of the amplified DNA fragment varies based on repeats of microsatellites in the genome (Weber and May, 1989). Before NGS technologies, a variety of DNA-based markers have been developed and used for genotyping, for instance, RAPD, SSRs (Hong et al., 2021), ISSRs, and AFLP. Among them, SSRs were most widely used in genotyping and genetic mapping studies. PCR-based genotyping methods are cheaper as compared to NGS technologies. However, the PCR-based genotyping methods are laborious and not highly efficient as NGS-based genotyping. The NGS-based genotyping includes restriction digestion of DNA and sequencing of libraries.

Double-digest restriction site–associated DNA

Although the SSRs are a potent marker system because of high reproducibility, co-dominance, and polymorphism, it is time, therefore, to generate the thousands of genome-wide SNP markers, restriction-sites associated with DNA sequencing (RADSeq) for large populations to study population genetics and genetic dissection of complex traits (Davey and Blaxter, 2011). However, in RADSeq, ~30%–50% of data were discarded because of repeated variable sites. The more reliable technique of double-digest restriction site–associated DNA sequencing (ddRAD-Seq) was developed to boost the efficiency (Peterson et al., 2012). The ddRAD-seq simultaneously uses two restriction enzymes to decrease the genome entanglement and library preparation cost by five-folds and can capture the genomic regions in hundreds of thousands for enhanced representation of the genome. It was successfully used in genetic mapping studies in peanut to map the QTLs for late leaf resistance and plant type–related traits (Zhou et al., 2014; Zhou et al., 2016). The ddRAD-seq was further advanced to reduce the repetitive DNA sequences, and the optimized version of ddRAD-Seq was developed called SLAF-Seq. The steps in SLAF-Seq are the same as in ddRAD-Seq. The DNA fragments are optimized for even distribution and to reduce the repetitive sequences. However, both technologies do not cover the whole genome (Sun et al., 2013).

Genotyping-by-sequencing

GBS is a robust genotyping technology used for SNP discovery for a multitude of applications (Elshire et al., 2011). It is a variation of ddRAD-seq, first discovered in maize and barley used for genotyping recombinant inbred line populations. In GBS, methylation-sensitive restriction enzymes play a vital role in DNA digestion that lessens the genome complexity while constructing the sequence libraries. The genomic areas that are difficult to access to contemporary sequencing techniques can be captured by GBS. The GBS was efficiently used in groundnut for trait mapping (Jadhav et al., 2021) and diversity analysis (Khera et al., 2013). Pandey and co-workers (2014) performed GWAS analysis using SSR and GBS-based SNP genotyping data to identify the SNPs associated with aflatoxin contamination and agronomic traits in groundnut. GBS was used for genotyping cultivated and wild accessions of chickpea to discover 82,489 SNPs used for diversity, population structure, and LD analysis (Bajaj et al., 2015a; Kujur et al., 2015a). A total of 3,187 SNPs were used to reveal the genetic cluster associated with black-seeded genotypes of chickpea. GBS was also used for genotyping biparental populations in trait mapping studies to identify the QTLs for sterility mosaic disease (Saxena et al., 2017a), fusarium wilt (Saxena et al., 2017b), and fertility restoration (Saxena et al., 2018) in pigeon pea. In chickpea, drought tolerance–related "*QTL-hotspot*" was discovered with 743 SNP loci (Jaganathan et al., 2015), and 3,228 SNP loci were used for mapping and identification of CGs of seed traits (Verma et al., 2015). The multiplex sequencing strategy by using adapter sequences makes GBS very inexpensive. However, it produces more missing calls, and imputations are highly recommended during quality analysis. However, GBS is also incomplete, as its sequencing covers only a limited genome (~2.5%). GBS has replaced the previous genotyping markers, i.e., RAPD, ISSR, and SSRs, as it requires less time and labor and is highly cost-effective. GBS technology has been done in legumes like chickpea and soybean (Shingote et al., 2022; Torkamaneh et al., 2021; Iquira et al., 2015; Bajaj et al., 2016; Sudheesh et al., 2021; Kujur et al., 2015b; Verma et al., 2015).

Diversity array technology

The polymorphic DNA segments called DArT markers in a genome are recognized through differential hybridization on a diversity genotyping array (Jaccoud et al., 2001). DArT is a very cost-effective whole-genome DNA fingerprinting tool for a variety of genetic analyses. It is a high-throughput sequence–independent technology that combines restricted-based hybridization and PCR. It is a very efficient marker system that can discover thousands of polymorphic sites in a very short time in any crop species. DArT is very popular in terms of high genome coverage, speed, reproducibility, and reliability (Aitken et al., 2014). Furthermore, polymorphic fragment calling does not require the reference genome. The DArT technology can be effectively used for genomic selection (GS) (Varshney et al., 2017) and marker-assisted selection (Stojaowski et al., 2011). However, the DArT markers are redundant due to clones with common sequences. Therefore, the presence of redundancy and markers with low frequencies (~41%) may affect the statistical analysis that is needed to filter out. DArT procedure includes generating a diversity panel followed by genotyping using a diversity panel. The first-ever genetic map of any legume crop was designed using DArT technology by Yang and coworkers (2011) in pigeon pea. A biparental population (F_2) was screened using 554 DArT markers. Olukolu et al. (2012) used the DArT marker technology for genetic diversity assessment of 124 accessions of groundnut representing 25 countries of Africa. Roorkiwal et al. (2014) used the DArT markets to diversify the 10 *Cicer* species, including 94 genotypes. Aldemir and coworkers (2017) used an advanced version of DArT technology, i.e., DArt sequencing (DArTseq), for the identification of QTL for iron content in lentil seeds. DArTseq is also a hybridization-based technology but combines with NGS and provides a much simpler form of sequencing than DArT (Courtois et al., 2013; Aldemir et al., 2017). Ates (2019) estimated the genetic diversity of 94 lentil landraces with DArt-based 19,383 SNPs.

SNP arrays

NGS technologies discovered an ample number of SNP markers because the demand for high-throughput genotyping has increased. The hybridization-based microarray or SNP arrays are very popular in genetic mapping, diversity analysis, and population genomics (You et al., 2018). SNP array or DNA microarray are highly polymorphic and use designed probes hybridized with fragmented DNA, which determines the alleles of all the SNP positions for hybridized DNA samples (LaFramboise, 2009). On the basis of the density, the SNP arrays can be divided into high-density (>50K), mid-density 5–10K), and low-density (>5K) SNP arrays. High-density SNP arrays can be used for high-density genetic mapping, GWAS, and population genomics studies. Mid-density assays can be used in GS because a few thousand SNPs are enough based on the genome size of the individual. However, the low-density SNP arrays can be used for foreground and background selection during marker-assisted selection and several breeding purposes. For instance, the quality control panel of rice is a low-density SNP array (25 SNPs), highly used for F1 confirmation, hybrid purity testing, and DNA fingerprinting in rice (Ndjiondjop et al., 2018). SNP arrays have been efficiently developed in several crops for genotyping, such as maize (600K SNP array) (Unterseer et al., 2014), apple (480K SNP array) (Bianco et al., 2016), and rice (700K SNP array) (McCouch et al., 2016). In leguminous crops such as peanut, the SNP Arachis array with 58K SNPs (Pandey et al., 2017) was very successful for genetic mapping (Pandey et al., 2020) and association analysis (Gangurde et al., 2020) for several traits. In pigeon pea, 56K Axiom Cajanus SNP Array and chickpea 11K Axiom Cicer SNP Array were developed (Roorkiwal et al., 2018). However, they are fixed and may not capture all recombination or diversity in an association panel, which are the limitations of SNP arrays. For instance, for genotyping a multi-parent population such as MAGIC or NAM, the whole-genome resequencing–based genotyping is helpful to capture maximum recombination regions.

Whole-genome resequencing

Advanced NGS technologies reduced per-sample sequencing cost, and WGRS-based genotyping was used for many populations to identify the presence of absence variations for genome-wide association analysis. WGRS can be carried out at high depth or low depth based on the objective of the study. For instance, in the case of genetic mapping, 0.5–1.0X coverage is sufficient; however, for GWAS, 10–15X coverage can be used. Several NGS platforms can be used for generating WGRS data, such as Illumina Hi-seq (read length of 150–250 bp), PacBio (10–25Kb), and NanoPore (read size of 500 bp to 2.3 Mb). Large LD blocks (several hundred kilo–base pairs) in plants, specially self-pollinating. Large LD blocks include several CGs. Therefore, with dense genotyping, we can have SNP variants in each of the CGs in the block and individual CGs can be identified using GWAS carried out on WGRS genotyping data. A gene for salinity tolerance *Glyma03g32900*, using sequencing data on 106 soybean diversity panels and the SNP-based KASP markers, was developed to improve salinity tolerance in soybean (Patil et al., 2016). Recently, 2,980 chickpea accessions are sequenced to discover 3.94 million SNPs, phenotyping data on 16 traits was used for GWAS analysis and identified 205 SNPs associated with 11 traits, and the associated SNPs were in the genomic regions of 79 CGs playing a role in controlling key traits like seed weight (Varshney et al., 2021).

Alleviating the phenotyping bottleneck

In the era of different omics like genomics, transcriptomics, and proteomics with the help of NGS technologies, genotyping of large germplasm at multiple locations has become feasible for plant scientists. Thus, phenotyping these large germplasms/populations with higher accuracy have become difficult. Thus, high-throughput genotyping technologies have shifted the bottleneck of plant science

from genotyping to phenotyping (Mir et al., 2019). Thus, it has become the need for time to develop high-throughput phenotyping (HTP) approaches (Mir et al., 2019). Several advanced artificial intelligence–based HTP platforms have been developed for crops like rice, maize, and Arabidopsis (Yang et al., 2020). Still, a lot of improvement is required in HTP, which can record multiple phenotypic traits in less time and manpower, which can be associated with large genotypic data of large populations (Mir et al., 2019). The major limitation in phenotyping is recording the multiple traits (agronomic traits, physiological traits, and stress-related scoring) data of large populations at multiple locations in several replications (Furbank and Tester, 2011). There are a lot of chances for error in phenotypic data when recorded manually, and less accuracy leads to false significant associations with molecular markers and wrong interpretation of alleles and genes. HTP is a non-destructive data recording method that allows the plant scientist to increase the size of the experiment by the number of genotypes or replication, or locations (Awlia et al., 2016). PHENOPSIS was one of the first automated imaging and weighing systems developed in Arabidopsis to estimate its response to water deficiency (Granier et al., 2006). However, it has its limitations. HTP platforms are of two types, i.e., HTP platforms for greenhouse or laboratory experiments and open field experiments (Shafiekhani et al., 2017). Although, HTP technologies have been used successfully for genetic dissection of agronomic traits in major field crops like rice, maize, wheat, barley, and brassica (Zhang et al., 2017a; Shi et al., 2013; Yang et al., 2014; Muraya et al., 2017; Topp et al., 2013; Tanabata et al., 2012). The use of these HTP platforms in legume crops is yet to be evaluated at the large fields, population, and multiple location levels (Zhang et al., 2021). A handful of studies has been conducted on legumes such as pea, soybean, and chickpea using a HTP approach for biotic and abiotic stress (Zhang et al., 2012; Friedli et al., 2016; Humplík et al., 2015). Zhang et al. (2021) used the quadcopter unmanned aircraft vehicle multispectral imaging data to predict the yield of chickpea and dry pea with a multivariate regression model. Humplík et al. (2015) used the automatic red blue green image analyzing software in pea to estimate the shoot biomass and photosynthetic activity for cold tolerance. Friedli et al. (2016) used the yerrestrial 3D laser scanning system in soybean for canopy-related traits.

Advanced methods and tools for GWAS

GWAS has continuously expanded in the last few decades due to advancements in sequencing technologies and the collective effort of the research community. In addition, HTP technologies have allowed us to measure many plant traits that are now frequently analyzed through GWAS tools. Recent years have seen GWAS methods solving issues of computation complexity or enhancing statistical power. It is utilized to detect new associations with traits of interest and to replicate loci detected by other different approaches. A diverse set of researchers is involved in rare-variant detection, statistical model optimization, synthetic associations, and using GWAS findings to better our knowledge of disease etiology. These methods can detect genetic variants associated with biochemical or agronomic and molecular phenotypes. In the future, this will enhance the utility of GWAS methods and their implications for plant science.

Naïve methods

In the GWAS, linear or logistic regression models are used to test for associations. The linear model is used for continuous traits such as plant height, whereas logistic regression models are used for binary traits indicating that the disease is present or absent. In addition, some covariates are included to account for confounding effects from demographic factors. However, naïve approaches often suffer from inflated false-positive rates that might be induced due to genetic relatedness among study participants (Oetjens et al., 2016). In GWAS, usually, diverse populations are selected, which often have related individuals, making subpopulations within the population. This might lead to spurious associations between SNPs that are more common in each subpopulation and phenotypes of interest if the phenotype has a higher prevalence in that subpopulation.

Mixed linear model methods

The MLM frameworks used in GWAS have drastically decreased the false-positive rates in comparison with conventional naïve approaches. Among these, the fast GWA tool is an ultra-efficient tool for MLM-based GWAS analysis of biobank-scale data (Jiang et al., 2019). MLM approaches resolve the issue of genetic relatedness among individuals following correction at two levels. These refer to population structure and kinship (Yu et al., 2006). At the first level, the population structure is inferred using genotype data with STRUCTURE tool (Pritchard et al., 2000) or through principal component analysis (Price et al., 2006). The kinship matrix is used at the second level to estimate inter-individual relatedness using the genotype data (Yu et al., 2006). In recent years, many methods have been developed to efficiently solve MLM equations. For instance, a recently available method referred to as EMMA (efficient mixed-model association) provided superior computational speed by eliminating the duplicate matrix operations at each iteration while estimating the likelihood function (Kang et al., 2008). MLM-based methods become computationally intensive for large numbers of samples. The FaST-LMM solves this issue but requires that the number of SNPs be less than the number of samples to derive kinship. The SUPER (Settlement of MLM Under Progressively Exclusive Relationship) method has been developed to extract a subset of SNPs and use them in FaST-LMM to increase the statistical power. Moreover, the compress MLM (CMLM) and enriched CMLM (ECMLM) methods are available for kinship optimization. The modified MLM method called multiple-locus linear mixed model (MLMM) incorporates multiple markers simultaneously as covariates in a stepwise MLM to partially remove the confounding between testing markers and kinship. Furthermore, a new method referred to as fixed and random model circulating

probability unification (FarmCPU) completely removes the confounding by dividing MLMM into a fixed-effect model and a random-effect model and using them iteratively. The FarmCPU can analyze the dataset with half million individuals and half million markers within 3 days. However, the random-effect model is computationally intensive in FarmCPU. The new method called Bayesian information and linkage disequilibrium iteratively nested keyway (BLINK) replaces the random-effect model with the fixed-effect model by using Bayesian information criteria. This method also replaces the bin method used in FarmCPU with LD information to eliminate the requirement that quantitative trait nucleotides be uniformly distributed throughout the genome. These all methods are summarized in Table 2.

Machine learning methods

Recent years have seen tremendous growth in the machine learning methods targeted for GWAS. The approaches used by these methods include classification, regression, ensemble-based learning, and neural networks.

Regression

Logistic regression coupled with the least absolute shrinkage and selection operator (LASSO) regularization approach is a famous method for GWAS. The penalized logistic regression method was used for the classification of patients with Crohn's disease using genotyping data at the genome-wide level. The LASSO and ridge regression are among the most frequently utilized penalized regression algorithms (Tibshirani et al., 1996; Hoerl et al., 1970). Recently, a faster and more powerful algorithm was developed by binning the closely occurring SNPs based on LD (An et al., 2020). In addition, the SNPs and phenotypes were mapped using LASSO regression in this method. This method was found to provide a reduced type 1 error rate in comparison with regular MLM and LASSO. To discover variations closely associated with the duloxetine response, some researchers used the standard genome-wide logistic regression (Maciukiewicz et al., 2018). In addition, they extracted the top predictors using LASSO regression. In another study, a preconditioned random forest regression was used to predict late genitourinary toxicity after radiotherapy. This preconditioning involved usage of logistic regression for making a continuous surrogate outcome from the original binary outcomes, which were followed by random forest regression where the surrogate outcome is utilized as a target for prediction. In this study, five-fold cross-validation was conducted for testing the model stability against existing baseline models (Lee et al., 2018). The major drawback of regression approaches is the failure to find higher-order non-linear SNP interactions involved in susceptibility to diseases. The process developed by Zhang and coworkers (2012) utilizes prior information from proteomics and biological pathways for SNP groups. To find the top predictive SNP groups, they used linear regression standardized by group sparse constraint. In the end, group LASSO was used for the regularized linear regression (Yuan et al., 2005). Thus, this approach overcomes the limitations of the regular MLM used in GWAS.

Classification

Support vector machine (SVM)–based classification methods such as COMBI have been developed for unknown phenotype prediction for a given unseen genotype (Mieth et al., 2016). In this approach, the SNPs having larger SVM weight are chosen, and the remaining SNPs are removed. Next, a chi-squared test is

TABLE 2 Advanced methods and tools for GWAS.

S.No.	Method	Description	Reference
1.	MLM	At the first level, the population structure is inferred using genotype data with STRUCTURE tool or through principal component analysis. The kinship matrix is used at the second level to estimate inter-individual relatedness using the genotype data.	Jiang et al., 2019
2.	CMLM	Clusters the individuals into groups and fits the genetic values of groups as random effects in the model that improves statistical power compared to regular MLM methods.	Zhang et al., 2010
3.	ECMLM	Calculate kinship using several different algorithms and then choose the best combination b/w kinship algorithms and grouping algorithms.	Li et al., 2014
4.	FaST-LMM	An algorithm for genome-wide association studies (GWAS) that scales linearly with cohort size in both run time and memory use. This method requires that the number of SNPs be less than the number of samples to derive kinship.	Lippert et al., 2011
5.	SUPER	Uses the associated genetic markers referred as pseudo quantitative trait nucleotides instead of all the markers, to derive kinship.	Wang et al., 2014
6.	MLMM	Include multiple markers simultaneously as covariates in a stepwise MLM to partially remove the confounding between testing markers and kinship.	Segura et al., 2012
7.	FarmCPU	Uses a bin method under the assumption that quantitative trait nucleotides are evenly distributed throughout the genome. Completely eliminates the confounding by dividing MLMM into a fixed effect model and a random effect model and using them iteratively.	Liu et al., 2016
8.	BLINK	Replaces the random effect model with the fixed effect model by using Bayesian information criteria. Uses linkage disequilibrium information to eliminate the requirement that quantitative trait nucleotides be uniformly distributed throughout the genome.	Huang et al., 2019

performed, and SNPs that exhibit a p-value below the significant criterion are taken into consideration for intensive study. The SVM method separates labeled data points into two groups with a large difference between them. Some authors proposed using SVM for genetic risk prediction (Mittag et al., 2012). This method has been used for genome-wide risk profiling for diseases such as type 1 diabetes and Parkinson's disease. In this algorithm, model training is performed using SNP data, which is followed by binary classification of the test dataset. Another researcher used the K-nearest neighbor learning algorithm for the classification of individuals into breast cancer positive and negative groups using their SNPs (Hajiloo et al., 2013). They used a leave-one-out cross validation strategy and external holdout methods for evaluating the performance of their classification algorithm.

Ensemble learning methods

These methods comprise an ensemble of decision trees. For example, random forest is an example of the ensemble learning algorithm. A bootstrapped subsample of the initial training dataset is used to create each decision tree in this instance. Some authors used gradient-boosting and random forest approaches to identify potent SNPs (Dorani et al., 2018). Nguyen et al. (2015) used a random forest method for selecting informative SNPs. They used a two-stage quality-based approach in model learning for the selection of informative SNPs. This method seems quite useful for the high-dimensional GWAS data. They also used five-fold cross-validation for assessing the potential of the model on different GWAS datasets. In addition, gradient boosting of decision trees was used for GWAS datasets. Others proposed using the XGBoost model for SNP selection (Behravan et al., 2018). This model could be used as an alternative to polygenic risk scoring. In addition, SVM classifier is used at the backend for SNP classification. Using principal component analysis and logistic regression, Oh et al. (2017) suggested a preconditioned random forest regression that converts a binary variable into a continuous variable. Later, Lee and a group of researchers (2020) used this preconditioned model for predicting the risk of breast cancer.

Neural network-based methods

Liu et al. (2019) developed a convoluted neural network (CNN) model for phenotype prediction using the SNP dataset. Moreover, they applied a saliency map for the first time to choose significant SNPs from training model. They also compared them with statistical methods such as best linear unbiased prediction and Bayesian ridge regression (BRR). In this study, association analysis was performed for quantitative traits of soybean and SNP datasets. Some authors found that increasing the hidden neuron's number does not affect the performance of the classification model for the case-control settings (Romagnoni et al., 2019). In a different study, authors compared the deep mixed model constituted of CNN and long and short-term memory with standard univariate testing and MLM (Wang et al., 2019).

Transcriptome-wide association study methods

Transcriptome-wide association study (TWAS) methods perform association analysis for gene expression variations and quantitative traits. TWAS is an approach based on genes with the ability to expand GWAS for a better understanding of functional relationships in complex traits. These methods are alternatives to variant-based association methods representing a subgroup of multi-marker association or locus-based methods. The locus-based methods have been so popular due to the larger apprehension and acceptability of the polygenic framework of the complex traits. In principle, locus-based approaches rely on multiple genetic variants to estimate the contribution of a gene or loci. TWAS uses GWAS results and transcriptome-level information to perform association testing at the gene level (Pividori et al., 2020). The ability to separate and assess the analytical procedures in TWAS simultaneously, provides several opportunities for the development of effective statistical models for the study of gene disease connections.

PheWAS methods

PheWAS methods perform unique associations in addition to utilizing known genotype–phenotype associations acquired through GWAS. These established relationships might serve as "positive controls" for additional high-throughput analysis. PheWAS methods suffer from high false-positive rates due to thousands of genotype–phenotype associations being tested in such studies (Bastarache et al., 2022). In addition, sample sizes usually also vary across studies impacting the statistical power and the replication among studies. PLATO tool is used to identify associations in PheWAS (Hall et al., 2017). DNAnexus is another tool for genomic analysis that was hosted on Amazon Web Services. This provides a distributed cluster of computers on the cloud allowing much lesser computation time for such studies. With the assistance of the DNAnexus app for PLATO, scatter-process-gather can be used on the platform to train regression models concurrently. This scatter-gather approach initiated multiple AWS virtual machines to simultaneously fit the regression models. Deep-PheWAS is another platform for PheWAS that intertwines quantitative phenotypes from primary care data, disease progression, longitudinal trajectories of quantitative measures, and drug response phenotypes with the composite phenotypes generated from clinically curated data (Packer et al., 2023). Moreover, several tools are available on this platform for efficiently analyzing the association with genetic data under different genetic models.

GWAS-assisted genomic selection

GS has been utilized as a practical genomic approach for upgrading complex traits in various crops (Thudi et al., 2016; Sandhu et al., 2022). In segregating populations, GS allows identifying lines with higher genomic estimated breeding value

(GEBV) using genome-wide marker data. A training population (TP) is used to estimate GEBV, which consists of elite breeding lines and is evaluated for multi-seasons and locations for the target phenotype. Then, a candidate population (CP) is developed by selecting parents based on the GEBVs. GS utilizes all the available genome-wide marker data irrespective of any significant effects. The GS prediction accuracies depend on several factors, including the genome size, ploidy level, interactions between gene and QTL, sample number, relatedness, number and distribution of markers, and model (Yadav et al., 2020). Several statistical methods are used for GS, including Ridge regression best linear unbiased prediction (rrBLUP) and genomic best linear unbiased predictor (gBLUP); both hypothesize a normal distribution of the SNP effects, whereas Bayesian methods like BayesA, BayesB, BayesC, and BayesR allow different variance distributions considering marker effect sizes (Heslot et al., 2012; de Los Campos et al., 2013). On the other hand, kernel approaches help predict non-additive models along with complex multi-environment/trait data (Gianola and van Kaam, 2008; Bandeira E Sousa et al., 2017).

Zhang and coworkers (2016) showed a GWAS-assisted GS with 309 soybean lines and 31,405 SNPs for seed weight using the rrBLUP approach. They showed GS prediction accuracies of 0.75–0.87, outperforming marker-assisted selection with prediction accuracies of 0.62–0.75. Ravelombola et al. (2020) performed a GS approach for soybean cyst nematode tolerance with biomass reduction using 234 soybean accessions in the greenhouse. They used five methods to compute GEBVs, including gBLUP (Zhang et al., 2007), random forest (RF) (Ogutu et al., 2011), rrBLUP (Meuwissen et al., 2001), SVMs (Maenhout et al., 2007), and Bayesian LASSO (Legarra et al., 2011). They found that the prediction accuracies were dependent on the model used, the marker set, and the size of TP. However, the accuracy of GS was higher than the SNPs from GWAS for all selection models and TP sizes.

In alfalfa, Li et al. (2015) used clonal ramets from 185 to 190 individuals for GS of biomass yield across three locations and recorded prediction accuracies of 0.43 to 0.66 for each location. Another study used 322 individual genotypes from 75 genetically diverse alfalfa populations. They tested three Bayesian models (BayesA, BayesB, and BayesC) for 25 agronomic traits, including forage quality traits, dry matter, and fall dormancy (Jia et al., 2018). They reported prediction accuracies of 0.0021 to 0.6485 with no significant differences in the three Bayesian models.

In chickpeas, Roorkiwal et al. (2016) used 320 breeding lines and six different models, including rrBLUP, RF, Bayesian LASSO, BayesB, Kinship GAUSS, and Bayes Cπ for four traits, i.e., seed yield, 100 seed weight, days to maturity, and days to 50% flowering. They reported prediction accuracies ranging between 0.138 (seed yield) to 0.192 (100 seed weight).

Li et al. (2018) showed low prediction accuracies using rrBLUP, Bayesian LASSO, and BRR for grain yield/ha, seed number per plant, 100 seed weight, and early vigor score in chickpea, which can be attributed to the small size of TP.

In common bean, cooking time (CKT), seed weight, and water absorption capacity were evaluated using 922 lines consisting of four populations (a Mesoamerican 8-parental MAGIC population, a biparental RIL, an Andean, and a Mesoamerican breeding line (MIP) panel (Diaz et al., 2020). Six models based on additive effects (BRR, BayesA, BayesB, BayesC, Bayesian Lasso, and gBLUP) and a Bayesian reproducing kernel Hilbert spaces regression (RKHS) models based on both additive and non-additive effects were used. They reported prediction accuracies for CKT ranging from MIP (0.22) to MAGIC population (0.55). A recent study showed prediction abilities ranging between 0.6 and 0.8 were shown in common bean for four agronomic traits under several environmental stresses (Keller et al., 2020)

In peanut or groundnut, 281 Kersting's groundnut lines were used for GWAS-assisted GSs for several traits, including seed traits, 100 seed weight, leaf length, days to 50% flowering, and days to maturity using 493 SNPs and rrBLUP model (Akohoue et al., 2020). They recorded prediction accuracies ranging from 0.42 to 0.79 for 100 seed weight, seed length and width, days to maturity, and days to 50% flowering. A low prediction accuracy of 0.11–0.20 was reported for traits including plant architecture traits such as height and diameter, petiole length, leaf width, number of seeds, grain yield, number of pods per plant, and number of seeds per pod.

Recently, genomic resources have been made available in some minor legume crops (Varshney et al., 2019; Bohra et al., 2020). In peas, the predictive abilities based on Bayesian LASSO model were 0.28, 0.30, 0.64, and 0.65 for lodging susceptibility, yield, seed weight, and onset of flowering, respectively (Annicchiarico et al., 2019).

Several GS methods were used for predicting GEBVs in legume crops (Figure 2), but the progress still needs to catch up compared to grain crops, including wheat, rice, and maize. However, the GS approach proved helpful and could be applied in the early stages of legume breeding programs to identify promising progenies and parents based on the predicted breeding values.

Applications in plant breeding

Since the last decade, GWAS has been successfully used in major legume crops to dissect or identify the genetic bases for various agronomic traits (Bajaj et al., 2015b; Kujur et al., 2015a; Wen et al., 2015; Zhang et al., 2015), quality traits (Hwang et al., 2014; Upadhyaya et al., 2015; Shaibu et al., 2019a), biotic (Butenhoff, 2015; Zhang et al., 2019; Banoo et al., 2020; Zhang et al., 2020; Faridi et al., 2021), and abiotic stress (Thudi et al., 2002; Thudi et al., 2014; Dhanapal et al., 2015; Assefa et al., 2020; Kohli et al., 2020). A plethora of GWASs have been conducted using different types of markers (SSR, SNPs, etc.) in some of the major legumes like soybean (Butenhoff, 2015; Zhang et al., 2015; Zhang et al., 2019; Assefa et al., 2020), chickpea (Kujur et al., 2015b; Upadhyaya et al., 2017; Kohli et al., 2020), groundnut (Zhang et al., 2017c; Shaibu et al., 2019a; Wang et al., 2019), and minor legume crops (Ali et al., 2016; Faridi et al., 2021). Brief details of GWASs conducted in legume crops are given in Table 1.

Revolutionization and rapid development in genomic techniques in the recent past have accelerated molecular studies not only in model crops but also in other crops like legumes (Varshney et al., 2005). Sequencing and availability of reference genomes have also made it feasible for the researchers to identify the alleles/QTL association with the desired trait in any germplasm. Although the GWAS approach can be used in any crop with extensive phenotyping and genotyping, it has

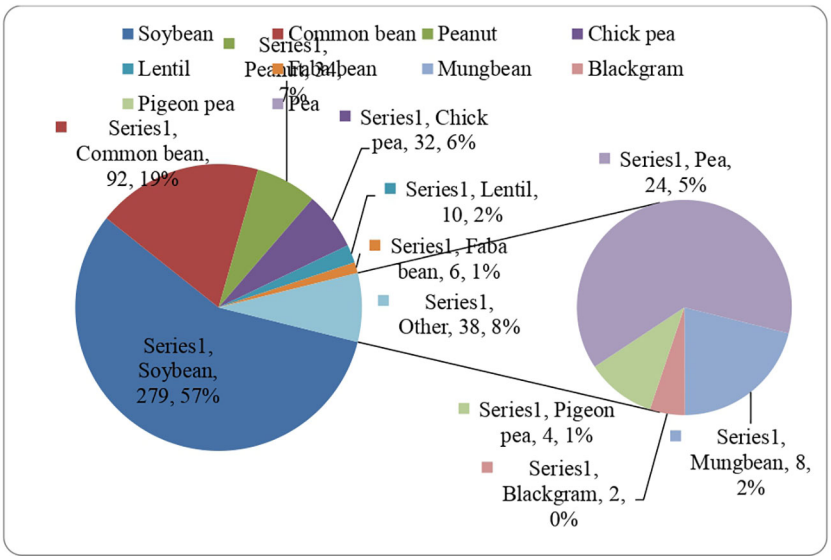


FIGURE 2
Number of GWAS studies conducted in different leguminous crops from timeline 2012 to 2023.

been used in major legume crops like soybean, chickpea, and groundnut (Mousavi-Derazmahalleh et al., 2019). These major legume crops' research community has sufficient funding for high-throughput genotyping and phenotyping. As these crops cover a significantly larger area across the globe, significantly diverse and classified germplasms are available for these crops (Mousavi-Derazmahalleh et al., 2019). However, GWAS has its limitations like false-positive association and exclusion of a significant association. All the limitations can be overcome by accurate phenotyping, large enough diverse germplasm, multilocation trials for phenotyping, and accuracy in genotyping. The use of the best suitable model, method, and bioinformatic tools also determines the accuracy of GWAS. The development of model tools for legume crops can trigger the GWAS in major and minor legume crops.

Conclusion and future perspectives

Legumes are an essential component of human nutrition and play a vital role in sustainable agriculture due to their protein-rich content, soil quality improvement, and reduced environmental impact. With the increasing global population and changing climatic conditions, there is a pressing need to develop high-yielding, disease-resistant legume cultivars that can meet the nutritional needs of the growing population. Improvement in the nutritional and production quality of legume crops with the use of conventional breeding methods is not at the required rate. Whole-genome sequencing is available only for a few major crops. As per the availability, low (RFLP) to high-throughput (SNP) markers have been used in various crops for AM, QTL mapping, or GWAS. NGS has become feasible in the model and even in non-model crops with improved efficiency and affordable sequencing methods. GWAS has been used in major legume crops to identify the genomic region linked with desired characteristics of the plant. It is yet to be exploited in minor legumes with sufficient germplasm/population. The availability of reference genomes and rapid development in genomic techniques has made it feasible for researchers to identify the alleles/QTL association with the desired trait in any germplasm. The use of suitable models, methods, and bioinformatic tools determines the accuracy of GWAS. The development of model tools for legume crops can trigger GWAS in major and minor legume crops. Authentication or precision of identified marker–trait association is required for their utilization in plant breeding programs or MAS/BAC programs. Using NGS and other high-throughput techniques for sequencing will make it possible to develop a genomic-assisted crop improvement program in legumes. Rapid development can be gained concerning agronomic traits, biotic/abiotic stress tolerance, and after-use quality improvement. Legume yield potential is meager compared to other major crops; this yield plateau can be broken in legumes for climate change problems using GWAS with multi-location phenotyping. The integration of these new and improved technologies with traditional breeding methods will help to accelerate the development of new legume cultivars with improved yield and nutritional qualities.

Author contributions

PS, PK conceptualized the review study and edited the manuscript. PY contributed to advanced methods and tools for GWAS. SS put forth experimental populations for association mapping studies in various crops. GK was associated with GWAS-assisted genomic selection. SSG and MP contributed to new high throughput-genotyping technologies in plants. VS and TT assisted in collecting data for applications of GWAS in plant

breeding and drafting the manuscript. All authors contributed to the article and approved the submitted version.

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

Abady, S., Shimelis, H., Janila, P., Yaduru, S., Shayanowako, A. I., Deshmukh, D., et al. (2021). Assessment of the genetic diversity and population structure of groundnut germplasm collections using phenotypic traits and SNP markers: implications for drought tolerance breeding. *PloS One* 16 (11), e0259883. doi: 10.1371/journal.pone.0259883

Abou-Khater, L., Maalouf, F., Jighly, A., Alsamman, A. M., Rubiales, D., Rispail, N., et al. (2022). Genomic regions associated with herbicide tolerance in a worldwide faba bean (*Vicia faba* l.) collection. *Sci. Rep.* 12, 1–13. doi: 10.1038/s41598-021-03861-0

Afzal, M., Alghamdi, S. S., Nawaz, H., Migdadi, H. H., Altaf, M., El-Harty, E., et al. (2022). Genome-wide identification, and expression analysis of CC-NB-ARC-LRR (NB-ARC) disease-resistant family members from soybean (*Glycine max* l.) reveal their response to biotic stress. *J. King Saud Univ* 34, 1758.

Agarwal, C., Chen, W., Varshney, R. K., and Vandemark, G. (2022). Linkage QTL mapping and genome-wide association study on resistance in chickpea to. *Pythium ultimum. Front. Genet.* 13, Art–945787.

Agarwal, C., Vandemark, G. J., Chen, W., and Varshney, R. K. (2019). *Genome-wide association studies and QTL mapping in chickpeas for resistance to ascochyta blight and pythium ultimum*. San Antonio, TX: International Annual Meeting: Embracing the Digital Environment.

Ahmed, S. M., Alsamman, A. M., Jighly, A., Mubarak, M. H., Al-Shamaa, K., Istanbuli, T., et al. (2021). Genome-wide association analysis of chickpea germplasms differing for salinity tolerance based on DArTseq markers. *PloS One* 16, e0260709. doi: 10.1371/journal.pone.0260709

Aitken, K. S., McNeil, M. D., Hermann, S., Bundock, P. C., Kilian, A., Heller-Uszynska, K., et al. (2014). A comprehensive genetic map of sugarcane that provides enhanced map coverage and integrates high-throughput diversity array technology (DArT) markers. *BMC Genomics* 15 (1), 1–12. doi: 10.1186/1471-2164-15-152

Akohoue, F., Achigan-Dako, E. G., Sneller, C., Van Deynze, A., and Sibiya, J. (2020). Genetic diversity, SNP-trait associations, and genomic selection accuracy in a West African collection of kersting's groundnut [*Macrotyloma geocarpum* (Harms) maréchal & baudet]. *PloS One* 15, e0234769. doi: 10.1371/journal.pone.0234769

Aldemir, S., Ateş, D., TEMEL, H. Y., Yağmur, B., Alsaleh, A., Kahriman, A., et al. (2017). QTLs for iron concentration in seeds of the cultivated lentil (*Lens culinaris* medic.) *via* genotyping by sequencing. *Turkish J. Agric. Forestry* 41 (4), 243–255. doi: 10.3906/tar-1610-33

Ali, M. B., Welna, G. C., Sallam, A., Martsch, R., Balko, C., Gebser, B., et al. (2016). Association analyses to genetically improve drought and freezing tolerance of faba bean (Vicia faba l.). *Crop Sci.* 56, 1036–1048. doi: 10.2135/cropsci2015.08.0503

An, B., Gao, X., Chang, T., Xia, J., Wang, X., Miao, J., et al. (2020). Genome-wide association studies using binned genotypes. *Heredity* 124 (2), 288–298. doi: 10.1038/s41437-019-0279-y

Annicchiarico, P., Nazzicari, N., Pecetti, L., Romani, M., and Russi, L. (2019). Pea genomic selection for Italian environments. *BMC Genomics* 20, 603. doi: 10.1186/s12864-019-5920-x

Arora, S., Mahato, A. K., Singh, S., Mandal, P., Bhutani, S., Dutta, S., et al. (2017). A high-density intraspecific SNP linkage map of pigeon pea (*Cajanus cajan* l. millsp.). *PloS One* 12 (6), e0179747.

Assefa, T., Zhang, J., Chowda-Reddy, R. V., Moran Lauter, A. N., Singh, A., O'Rourke, J. A., et al. (2020). Deconstructing the genetic architecture of iron deficiency chlorosis in soybean using genome-wide approaches. *BMC Plant Biol.* 20, 1–13. doi: 10.1186/s12870-020-2237-5

Awlia, M., Nigro, A., Fajkus, J., Schmoeckel, S. M., Negrao, S., Santelia, D., et al. (2016). High-throughput non-destructive phenotyping of traits that contribute to salinity tolerance in arabidopsis thaliana. *Front. Plant Sci.* 7, 1414. doi: 10.3389/fpls.2016.01414

Bajaj, D., Das, S., Badoni, S., Kumar, V., Singh, M., Bansal, K. C., et al. (2015a). Genome-wide high-throughput SNP discovery and genotyping for understanding natural (functional) allelic diversity and domestication patterns in wild chickpea. *Sci. Rep.* 5 (1), 1–7. doi: 10.1038/srep12468

Bajaj, D., Das, S., Upadhyaya, H. D., Ranjan, R., Badoni, S., Kumar, V., et al. (2015b). A genome-wide combinatorial strategy dissects complex genetic architecture of seed coat color in chickpea. *Front. Plant Sci.* 6, 979. doi: 10.3389/fpls.2015.00979

Bajaj, D., Upadhyaya, H. D., Das, S., Kumar, V., Gowda, C. L. L., Sharma, S., et al. (2016). Identification of candidate genes for dissecting complex branch number traits in chickpea. *Plant Sci.* 245, 61–70. doi: 10.1016/j.plantsci.2016.01.004

Bajgain, P., Rouse, M. N., Tsilo, T. J., Macharia, G. K., Bhavani, S., and Jin, Y. (2016). Nested association mapping of stem rust resistance in wheat using genotyping by sequencing. *PloS One* 11, e0155760. doi: 10.1371/journal.pone.0155760

Bandeira E Sousa, M., Cuevas, J., de Oliveira Couto, E. G., Pérez-Rodríguez, P., Jarquín, D., Fritsche- Neto, R., et al. (2017). Genomic-enabled prediction in maize using kernel models with genotype × environment interaction. *G3 (Bethesda)* 7, 1995–2014. doi: 10.1534/g3.117.042341

Bandillo, N., Raghavan, C., Muyco, P. A., Sevilla, M. A. L., Lobina, I. T., Dilla-Ermita, C. J., et al. (2013). Multi-parent advanced generation inter-cross (MAGIC) populations in rice: progress and potential for genetics research and breeding. *Rice* 6, 11. doi: 10.1186/1939-8433-6-11

Banoo, A., Nabi, A., Rasool, R. S., Shah, M. D., Ahmad, M., Sofi, P. A., et al. (2020). North-western Himalayan common beans: population structure and mapping of quantitative anthracnose resistance through genome-wide association study. *Front. Plant Sci.* 11, 571618. doi: 10.3389/fpls.2020.571618

Bastarache, L., and Denny, J. C. (2022). And roden, dPhenome-wide association studies. *JAMA* 327 (1), 75–76. doi: 10.1001/jama.2021.20356

Basu, U., Srivastava, R., Bajaj, D., Thakro, V., Daware, A., Malik, N., et al. (2018). Genome-wide generation and genotyping of informative SNPs to scan molecular signatures for seed yield in chickpea. *Sci. Rep.* 8 (1), 13240. doi: 10.1038/s41598-018-29926-1

Basu, U., Bajaj, D., Sharma, A., Malik, N., Daware, A., Narnoliya, L., et al. (2019). Genetic dissection of photosynthetic efficiency traits for enhancing seed yield in chickpea. *Plant Cell Environ.* 42 (1), 158–173. doi: 10.1111/pce.13319

Behravan, H., Hartikainen, J. M., Tengström, M., et al. (2018). Machine learning identifies interacting genetic variants contributing to breast cancer risk: A case study in finnish cases and controls. *Sci. Rep.* 8 (1), 13149. doi: 10.1038/s41598-018-31573-5

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., et al. (2008). Accurate whole human genome sequencing using reversibleterminator chemistry. *Nature* 456, 53–59. doi: 10.1038/nature07517

Bhagyawant, S. S., and Srivastava, N. (2008). Genetic fingerprinting of chickpea (Cicer arietinum l.) germplasm using ISSR markers and their relationships. *Afr. J. Biotech.* 7 (24), 4428–4431. doi: 10.5897/AJB08.973

Bianco, L., Cestaro, A., Linsmith, G., Muranty, H., Denancé, C., Théron, A., et al. (2016). Development and validation of the axiom® Apple480K SNP genotyping array. *Plant J.* 86 (1), 62–74. doi: 10.1111/tpj.13145

Bohra, A., Jha, U. C., Satheesh Naik, S. J., Mehta, S., Tiwari, A., Maurya, A. K., et al. (2021). "Genomics: Shaping legume improvement," in *Genetic enhancement in major food legumes* (Cham: Springer), 49–89. doi: 10.1007/978-3-030-64500-7_3

Bohra, A., Saxena, K. B., Varshney, R. K., and Saxena, R. K. (2020). Genomics-assisted breeding for pigeon pea improvement. *Theor. Appl. Genet.* 133, 1721–1737. doi: 10.1007/s00122-020-03563-7

Breria, C. M., Hsieh, C. H., Yen, T. B., Yen, J. Y., Noble, T. J., and Schafleitner, R. (2020). A SNPbased genome-wide association study to mine genetic loci associated to salinity tolerance in mungbean *(Vigna radiata* l.). *Genes* 11 (7), 759.

Butenhoff, K. J. (2015). *QTL mapping and gwas identify sources of iron deficiency chlorosis and canopy wilt tolerance in the fiskeby iii x mandarin (ottawa) soybean population (Doctoral dissertation, university of Minnesota)*. Available at: https://hdl.handle.net/11299/170730.

Carlson, C. H., Gouker, F. E., Crowell, C. R., Evans, L., DiFazio, S. P., Smart, C. D., et al. (2019). Joint linkage and association mapping of complex traits in shrub willow (Salix purpurea l.). *Ann. Bot.* 124 (4), 701–716. doi: 10.1093/aob/mcz047

Casa, A. M., Pressoir, G., Brown, P. J., Mitchell, S. E., Rooney, W. L., Tuinstra, M. R., et al. (2008). Community resources and strategies for association mapping in sorghum. *Crop Sci.* 48 (1), 30–40. doi: 10.2135/cropsci2007.02.0080

Chang, H. X., Lipka, A. E., Domier, L. L., and Hartman, G. L. (2016). Characterization of disease resistance loci in the USDA soybean germplasm collection using genome-wide association studies. *Phytopathology* 106 (10), 1139–1151. doi: 10.1094/PHYTO-01-16-0042-FI

Che, Z., Yan, H., Liu, H., Yang, H., Du, H., Yang, Y., et al. (2020). Genome-wide association study for soybean mosaic virus SC3 resistance in soybean. *Mol. Breed.* 40, 1–14. doi: 10.1007/s11032-020-01149-1

Chen, H., Liu, X., Zhang, H., Yuan, X., Gu, H., and Cui, X. (2018). Advances in salinity tolerance of soybean: genetic diversity, heredity, and gene identification contribute to improving salinity tolerance. *J. Integr. Agric.* 17, 2215–2221. doi: 10.1016/S2095-3119(17)61864-1

Choudhary, P., Khanna, S. M., Jain, P. K., Bharadwaj, C., Kumar, J., Lakhera, P. C., et al. (2012). Genetic structure and diversity analysis of the primary gene pool of chickpea using SSR markers. *Genet. Mol. Res.* 11 (2), 891–905. doi: 10.4238/2012.April.10.5

Contreras-Soto, R. I., Mora, F., de Oliveira, M. A. R., Higashi, W., Scapim, C. A., and Schuster, I. (2017). A genome-wide association study for agronomic traits in soybean using SNP markers and SNP-based haplotype analysis. *PloS One* 12, e0171105. doi: 10.1371/journal.pone.0171105

Courtois, B., Audebert, A., Dardou, A., Roques, S., Ghneim-Herrera, T., Droc, G., et al. (2013). Genome-wide association mapping of root traits in a japonica rice panel. *PloS One* 8 (11), e78037. doi: 10.1371/journal.pone.0078037

Cui, C., Mei, H., Liu, Y., Zhang, H., and Zheng, Y. (2017). Genetic diversity, population structure, and linkage disequilibrium of an association-mapping panel revealed by genome-wide SNP markers in sesame. *Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.01189

Das, S., Porter, L. D., Ma, Y., Coyne, C. J., Chaves-Cordoba, B., and Naidu, R. A. (2022). Resistance in lentil (Lens culinaris) genetic resources to the pea aphid (Acyrthosiphon pisum). *Entomologia Experimentalis Applicata* 170, 755– 769. doi: 10.1111/eea.13202

Davey, J., and Blaxter, M. L. (2011). RADSeq: next-generation population genetics. *Briefings Funct. Genomics* 10, 108.

de Los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D., and Calus, M. P. L. (2013). Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics* 193, 327–345. doi: 10.1534/genetics.112.143313

Deokar, A. A., Ramsay, L., Sharpe, A. G., Diapari, M., Sindhu, A., Bett, K., et al. (2014). Genome-wide SNP identification in chickpea for use in the development of a high-density genetic map and improvement of chickpea reference genome assembly. *BMC Genomics* 15 (1), 1–19. doi: 10.1186/1471-2164-15-708

Dhanapal, A. P., Ray, J. D., Singh, S. K., Hoyos-Villegas, V., Smith, J. R., Purcell, L. C., et al. (2015). Genome-wide association study (GWAS) of carbon isotope ratio (δ13C) in diverse soybean [*Glycine max* (L.) merr.] genotypes. *Theor. Appl. Genet.* 128, 73–91. doi: 10.1007/s00122-014-2413-9

Diaz, S., Ariza-Suarez, D., Ramdeen, R., Aparicio, J., Arunachalam, N., Hernandez, C., et al. (2020). Genetic architecture and genomic prediction of cooking time in common bean (*Phaseolus vulgaris* l.). *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.622213

Do, T. D., Vuong, T. D., Dunn, D., Clubb, M., Valliyodan, B., and Patil, G. (2019). Identification of new loci for salt tolerance in soybean by high-resolution genome-wide association mapping. *BMC Genomics* 20, 1–16. doi: 10.1186/s12864-019-5662-9

Doldi, M. L., Vollmann, J., and Lelley, T. (1997). Genetic diversity in soybean as determined by RAPD and microsatellite analysis. *Plant Breed.* 116 (4), pp.331–pp.335. doi: 10.1111/j.1439-0523.1997.tb01007.x

Dorani, F., Hu, T., Woods, M. O., and Zhai, G. (2018). Ensemble learning for detecting gene-gene interactions in colorectal cancer. *PeerJ* 6, e5854. doi: 10.7717/peerj.5854

Eathington, S. R., Crosbie, T. M., Edwards, M. D., Reiter, R. S., and Bull, J. K. (2007). Molecular markers in a commercial breeding program. *Crop Sci.* 47, S–154-S-163. doi: 10.2135/cropsci2007.04.0015IPBS

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., et al. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS One* 6 (5), e19379. doi: 10.1371/journal.pone.0019379

Farahani, S., Maleki, M., Ford, R., Mehrabi, R., Kanouni, H., Kema, G. H. J., et al. (2022). Genome-wide association mapping for isolate-specific resistance to ascochyta rabiei in chickpea (Cicer arietinum l.). *Physiol. Mol. Plant Pathol.* 121, 101883. doi: 10.1016/j.pmpp.2022.101883

Faridi, R., Koopman, B., Schierholt, A., Ali, M. B., Apel, S., and Link, W. (2021). Genetic study of the resistance of faba bean *(Vicia faba)* against the fungus ascochyta fabae through a genome-wide association analysis. *Plant Breed.* 140 (3), 442–452. doi: 10.1111/pbr.12918

Fayaz, H., Tyagi, S., Wani, A. A., Pandey, R., Akhtar, S., and Bhat, M. A. (2022). Genomewide association analysis to delineate high-quality SNPs for seed micronutrient density in chickpea *(Cicer arietinum* l.). *Sci. Rep.* 12 (1), 11357. doi: 10.1038/s41598-022-14487-1

Friedli, M., Kirchgessner, N., Grieder, C., Liebisch, F., Mannale, M., and Walter, A. (2016). Terrestrial 3D laser scanning to track the increase in canopy height of both monocot and dicot crop species under field conditions. *Plant Methods* 12 (1), 1–15. doi: 10.1186/s13007-016-0109-7

Furbank, R. T., and Tester, M. (2011). Phenomics–technologies to relieve the phenotyping. *Trends Plant Sci.* 16 (12), 635–644. doi: 10.1016/j.tplants.2011.09.005

Gali, K. K., Sackville, A., Tafesse, E. G., Lachagari, V. B. R., McPhee, K., Hybl, M., et al. (2019). Genome-wide association mapping for agronomic and seed quality traits of field pea (*Pisum sativum* l.). *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.01538

Gangurde, S. S., Wang, H., Yaduru, S., Pandey, M. K., Fountain, J. C., Chu, Y., et al. (2020). Nested-association mapping (NAM)-based genetic dissection uncovers candidate genes for seed and pod weights in peanut (Arachis hypogaea). *Plant Biotechnol. J.* 18 (6), 1457–1471. doi: 10.1111/pbi.13311

Gaur, R., Azam, S., Jeena, G., Khan, A. W., Choudhary, S., Jain, M., et al. (2012). High-throughput SNP discovery and genotyping for constructing a saturated linkage map of chickpea (Cicer arietinum l.). *DNA Res.* 19 (5), 357–373. doi: 10.1093/dnares/dss018

Gela, T., Ramsay, L., Haile, T. A., Vandenberg, A., and Bett, K. (2021). Identification of anthracnose race 1 resistance loci in lentil by integrating linkage mapping and genome-wide association study. *Plant Genome* 14, e20131. doi: 10.1002/tpg2.20131

Gianola, D., and van Kaam, J. B. C. H. M. (2008). Reproducing kernel hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics* 178, 2289–2303. doi: 10.1534/genetics.107.084285

Granier, C., Aguirrezabal, L., Chenu, K., Cookson, S. J., Dauzat, M., Hamard, P., et al. (2006). PHENOPSIS, an automated platform for reproducible phenotyping of plant responses to soil water deficit in arabidopsis thaliana permitted the identification of an accession with low sensitivity to soil water deficit. *New Phytol.* 169, 623–635. doi: 10.1111/j.1469-8137.2005.01609.x

Hajiloo, M., Damavandi, B., HooshSadat, M., Sangi, F., Mackey, J. R., Cass, C. E., et al (2013). "Breast cancer prediction using genome wide single nucleotide polymorphism data" *BMC Bioinf.* 14 (13), S35. doi: 10.1186/1471-2105-14-S13-S3

Hall, M. A., Wallace, J., Lucas, A., Kim, D., Basile, A. O., Verma, S. S., et al. (2017). PLATO software provides analytic framework for investigating complexity beyond genome-wide association studies. *Nat. Commun.* 8, 1167. doi: 10.1038/s41467-017-00802-2

Hasanuzzaman, M., Araújo, S., and Gill, S. S. (2020). *The plant family fabaceae: biology and physiological responses to environmental stresses* (Singapore: Springer). doi: 10.1007/978-981-15-4752-2

Heslot, N., Yang, H.-P., Sorrells, M. E., and Jannink, J.-L. (2012). Genomic selection in plant breeding: a comparison of models. *Crop Sci.* 52, 146. doi: 10.2135/cropsci2011.06.0297

Hiremath, P. J., Kumar, A., Penmetsa, R. V., Farmer, A., Schlueter, J. A., Chamarthi, S. K., et al. (2012). Large-Scale development of cost-effective SNP marker assays for diversity assessment and genetic mapping in chickpea and comparative mapping in legumes. *Plant Biotechnol. J.* 10 (6), 716–732. doi: 10.1111/j.1467-7652.2012.00710.x

Hoerl, A. E., and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12, 55–67. doi: 10.1080/00401706.1970.10488634

L. Horst and G. Wenzel (Eds.) (2007). *Molecular marker systems in plant breeding and crop improvement*.

Hang, J. (2022). *Genome-wide association study of seed protein and amino acid contents in cultivated lentils as determined by near-infrared reflectance spectroscopy (Master's thesis)*. Available at: http://hdl.handle.net/1993/36250

Hong, Y., Pandey, M. K., Lu, Q., Liu, H., Gangurde, S. S., Li, S., et al. (2021). Genetic diversity and distinctness based on morphological and SSR markers in peanut. *Agron. J.* 113 (6), 4648–4660. doi: 10.1002/agj2.20671

Hu, D., Kan, G., Hu, W., Li, Y., Hao, D., Li, X., et al. (2019a). Identification of loci and candidate genes responsible for pod dehiscence in soybean *via* genome-wide association analysis across multiple environments. *Front. Plant Sci.* 10, 811. doi: 10.3389/fpls.2019.00811

Hu, J., Maalouf, F., Zhang, Z., and Yu, L.-X. (2019b). "Genome-wide association study of the resilience to high temperature of faba bean (*Vicia faba* l.) germplasm," in *2019 ASHS Annual Conference (ASHS)*.

Huang, M., Liu, X., Zhou, Y., Summers, R. M., and Zhang, Z. (2019). BLINK: a package for the next level of genome-wide association studies with both individuals and markers in the millions. *Gigascience* 8 (2), giy154. doi: 10.1093/gigascience/giy154

Humplík, J. F., Lazár, D., Fürst, T., Husičková, A., Hýbl, M., and Spíchal, L. (2015). Automated integrative high-throughput phenotyping of plant shoots: a case study of the cold-tolerance of pea (*Pisum sativum* l.). *Plant Methods* 11 (1), 1–11. doi: 10.1186/s13007-015-0063-9

Huynh, B.-L., Ehlers, J. D., Huang, B. E., Munoz, A. M., Lonardi, S., Santos, J. R. P., et al. (2018). A multi- parent advanced generation inter-cross (MAGIC) population for genetic analysis and improvement of cowpea (*Vigna unguiculata* l. walp.). *Plant J.* 93, 1129–1142. doi: 10.1111/tpj.13827

Hwang, E.-Y., Song, Q., Jia, G., Specht, J. E., Hyten, D. L., Costa, J., et al. (2014). A genome-wide association study of seed protein and oil content in soybean. *BMC Genomics* 15, 1–12. doi: 10.1186/1471-2164-15-1

Iquira, E., Humira, S., and François, B. (2015). Association mapping of QTLs for sclerotinia stem rot resistance in a collection of soybean plant introductions using a genotyping by sequencing (GBS) approach. *BMC Plant Biol.* 15 (1), 1–12. doi: 10.1186/s12870-014-0408-y

Iruela, M., Rubio, J., Cubero, J. I., Gil, J., and Millan, T. (2002). Phylogenetic analysis in the genus cicer and cultivated chickpea using RAPD and ISSR markers. *Theor. Appl. Genet.* 104 (4), 643–651. doi: 10.1007/s001220100751

Jaccoud, D., Peng, K., Feinstein, D., and Kilian, A. (2001). Diversity arrays: a solid state technology for sequence information independent genotyping. *Nucleic Acids Res.* 29 (4), 1–7. doi: 10.1093/nar/29.4.e25

Jadhav, M. P., Gangurde, S. S., Hake, A. A., Yadawad, A., Mahadevaiah, S. S., Pattanashetti, S. K., et al. (2021). Genotyping-by-sequencing based genetic mapping identified major and consistent genomic regions for productivity and quality traits in peanut. *Front. Plant Sci.* 12, 668020. doi: 10.3389/fpls.2021.668020

Jaganathan, D., Thudi, M., Kale, S., Azam, S., Roorkiwal, M., Gaur, P. M., et al. (2015). Genotyping-by-sequencing based intra-specific genetic map refines a "QTL-hotspot" region for drought tolerance in chickpea. *Mol. Genet. Genom* 290, 559–571. doi: 10.1007/s00438-014-0932-3

Jha, U. C., Jha, R., Thakro, V., Anurag, K., Sanjeev, G., Harsh, N., et al. (2021). Discerning molecular diversity and association mapping for phenological, physiological and yield traits under high temperature stress in chickpea (Cicer arietinum l.). *J. Gen.* 100, 4. doi: 10.1007/s12041-020-01254-2

Jia, C., Zhao, F., Wang, X., Han, J., Zhao, H., Liu, G., et al. (2018). Genomic prediction for 25 agronomic and quality traits in alfalfa (Medicago sativa). *Front. Plant Sci.* 9, 1220. doi: 10.3389/fpls.2018.01220

Jiang, L., Zheng, Z., Qi, T., Kemper, K. E., Wray, N. R., Visscher, P. M., et al. (2019). A resource-efficient tool for mixed model association analysis of Large-scale data. *Nat. Genet.* 51 (12), 1749–1555. doi: 10.1038/s41588-019-0530-8

Johnson, N., Boatwright, J. L., Bridges, W., Thavarajah, P., Kumar, S., Shipe, E., et al. (2021). Genome-wide association mapping of lentil (Lens culinaris medikus) prebiotic carbohydrates toward improved human health and crop stress tolerance. *Sci. Rep.* 11, 1–12. doi: 10.1038/s41598-021-93475-3

Kalve, S., Gali, K. K., and Tar'an, B. (2022). Genome-wide association analysis of stress tolerance indices in an interspecific population of chickpea. *Front. Plant Sci.* 13, 933277. doi: 10.3389/fpls.2022.933277

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., et al. (2008). Efficient control of population structure in model organism association mapping. *Genetics* 178 (3), 1709–1235. doi: 10.1534/genetics.107.080101

Karthika, R., Jean, C. C., Ping, Z., Gopesh, S., Dorrie, M., Nurul, A., et al. (2021). Genetic diversity and GWAS of agronomic traits using an ICARDA lentil (*Lens culinaris* medik.). *Reference Plus collection. Plant Gen. Res.* 19 (4), 279–288. doi: 10.1017/S147926212100006X

Keller, B., Ariza-Suarez, D., de la Hoz, J., Aparicio, J. S., Portilla-Benavides, A. E., Buendia, H. F., et al. (2020). Genomic prediction of agronomic traits in common bean (*Phaseolus vulgaris* l.) under environmental stress. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.01001

Khan, M. A., Tong, F., Wang, W., He, J., Zhao, T., and Gai, J. (2018). Analysis of QTL–allele system conferring drought tolerance at seedling stage in a nested association mapping population of soybean [*Glycine max* (L.) merr.] using a novel GWAS procedure. *Planta* 248, 947–962. doi: 10.1007/s00425-018-2952-4

Khazaei, H., Fedoruk, M., Caron, C. T., Vandenberg, A., and Bett, K. E. (2018). Single nucleotide polymorphism markers associated with seed quality characteristics of cultivated lentil. *Plant Genome* 11, 170051. doi: 10.3835/plantgenome2017.06.0051

Khazaei, H., Podder, R., Caron, C. T., Kundu, S. S., Diapari, M., Vandenberg, A., et al. (2017). Marker–trait association analysis of iron and zinc concentration in lentil (*Lens culinaris* medik.) seeds. *Plant Genome* 10, plantgenome2017–02. doi: 10.3835/plantgenome2017.02.0007

Khera, P., Upadhyaya, H. D., Pandey, M. K., Roorkiwal, M., Sriswathi, M., Janila, P., et al. (2013). Single nucleotide polymorphism–based genetic diversity in the reference set of peanut *(Arachis* spp.) by developing and applying cost-effective kompetitive allele specific polymerase chain reaction genotyping assays. *Plant Genome* 6 (3), plantgenome2013–06. doi: 10.3835/plantgenome2013.06.0019

Kim, K. H., Kim, J.-Y., Lim, W.-J., Jeong, S., Lee, H.-Y., Cho, Y., et al. (2020). Genome-wide association and epistatic interactions of flowering time in soybean cultivar. *PloS One* 15, e0228114. doi: 10.1371/journal.pone.0228114

Kohli, P. S., Kumar Verma, P., Verma, R., Parida, S. K., Thakur, J. K., and Giri, J. (2020). Genome- wide association study for phosphate deficiency responsive root hair elongation in chickpea. *Funct. Integr. Genomics* 20, 775–786. doi: 10.1007/s10142-020-00749-6

Kujur, A., Bajaj, D., Upadhyaya, H., Das, S., Ranjan, R., Shree, T., et al. (2015a). A genome-wide SNP scan accelerates trait-regulatory genomic loci identification in chickpea. *Sci. Rep.* 5 (1), 1–20. doi: 10.1038/srep11166

Kujur, A., Upadhyaya, H. D., Shree, T., Bajaj, D., Das, S., Saxena, M. S., et al. (2015b). Ultra-high density intra-specific genetic linkage maps accelerate identification of functionally relevant molecular tags governing important agronomic traits in chickpea. *Sci. Rep.* 5 (1), 1–13. doi: 10.1038/srep09468

Kulwal, P. L., and Singh, R. (2021). "Association mapping in plants," in *Crop breeding* (New York, NY: Humana), 105–117.

Kumar, K., Anjoy, P., Sahu, S., Durgesh, K., Das, A., Tribhuvan, K. U., et al. (2022). Single trait versus principal component based association analysis for flowering related traits in pigeonpea. *Sci. Rep.* 12 (1), 10453. doi: 10.1038/s41598-022-14568-1

Kumar, J., Choudhary, A. K., Solanki, R. K., and Pratap, A. (2011). Towards marker-assisted selection in pulses: a review. *Plant Breed.* 130 (3), 297–313. doi: 10.1111/j.1439-0523.2011.01851.x

Kumar, J., Gupta, S., Biradar, R. S., Gupta, P., Dubey, S., and Singh, N. P. (2018a). Association of functional markers with flowering time in lentil. *J. Appl. Genet.* 59, 9–21. doi: 10.1007/s13353-017-0419-0

Kumar, J., Gupta, S., Gupta, D. S., and Singh, N. P. (2018b). Identification of QTLs for agronomic traits using association mapping in lentil. *Euphytica* 214, 1–15. doi: 10.1007/s10681-018-2155-x

Kumar, H., Singh, A., Dikshit, H. K., Mishra, G. P., Aski, M., Meena, M. C., et al. (2019). Genetic dissection of grain iron and zinc concentrations in lentil (Lens culinaris medik.). *J. Genet.* 98, 1–14. doi: 10.1007/s12041-019-1112-3

Kump, K., Bradbury, P., Wisser, R., Edward, S., Araby, R., Marco, A., et al. (2011). Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. *Nat. Genet.* 43, 163–168. doi: 10.1038/ng.747

LaFramboise, T. (2009). Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acids Res.* 37 (13), 4181–4193. doi: 10.1093%2Fnar%2Fgkp552

Lee, S., Kerns, S., Ostrer, H., Rosenstein, B., Deasy, J. O., and Oh, J. H. (2018). Machine learning on a genome-wide association study to predict late genitourinary toxicity after prostate radiation therapy. *Int. J. Radiat. OncologyBiologyPhysics* 101 (1), 128–355. doi: 10.1016/j.ijrobp.2018.01.054

Legarra, A., Robert-Granié, C., Croiseau, P., Guillaume, F., and Fritz, S. (2011). Improved lasso for genomic selection. *Genet. Res. (Camb)* 93, 77–87. doi: 10.1017/S0016672310000534

Li, Y., Ruperao, P., Batley, J., Edwards, D., Davidson, J., Hobson, K., et al. (2017). Genome analysis identified novel candidate genes for ascochyta blight resistance in chickpea using whole genome re-sequencing data. *Front. Plant Sci.* 8, 359. doi: 10.3389/fpls.2017.00359

Li, N., Shi, J., Wang, X., Liu, G., and Wang, H. (2014). A combined linkage and regional association mapping validation and fine mapping of two major pleiotropic QTLs for seed weight and silique length in rapeseed (Brassica napus l.). *BMC Plant Biol.* 14, 114. doi: 10.1186/1471-2229-14-114

Li, D., Zhao, X., Han, Y., Li, W., and Xie, F. (2019). Genome-wide association mapping for seed protein and oil contents using a large panel of soybean accessions. *Genomics* 111 (1), 90–95. doi: 10.1016/j.ygeno.2018.01.004

Li, Y., Ruperao, P., Batley, J., Edwards, D., Khan, T., Colmer, T. D., et al. (2018). Investigating drought tolerance in chickpea using genome-wide association mapping and genomic selection based on whole-genome resequencing data. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.00190

Li, M., Liu, X., Bradbury, P., Yu, J., Zhang, Y.-M., Todhunter, R. J., et al. (2014). Enrichment of statistical power for genome-wide association studies. *BMC Biol.* 12, 73. doi: 10.1186/s12915-014-0073-5

Li, X., Wei, Y., Acharya, A., Hansen, J. L., Crawford, J. L., Viands, D. R., et al. (2015). Genomic prediction of biomass yield in two selection cycles of a tetraploid alfalfa breeding population. *Plant Genome* 8, 0. doi: 10.3835/plantgenome2014.12.0090

Li, Y., Ruperao, P., Batley, J., Edwards, D., Khan, T., Colmer, T. D., et al. (2018). Investigating drought tolerance in chickpea using genome-wide association mapping and genomic selection based on whole-genome resequencing data. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.00190

Liew, M., Pryor, R., Palais, R., Meadows, C., Erali, M., Lyon, E., et al. (2004). Genotyping of single-nucleotide polymorphisms by high-resolution melting of small amplicons. *Clin. Chem.* 50 (7), 1156–1164. doi: 10.1373/clinchem.2004.032136

Lin, J., Lan, Z., Hou, W., Yang, C., Wang, D., Zhang, M., et al. (2020). Identification and finemapping of a genetic locus underlying soybean tolerance to SMV infections. *Plant Sci.* 292, 110367. doi: 10.1016/j.plantsci.2019.110367

Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., and Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. *Nat. Methods* 8 (10), 833–835. doi: 10.1038/nmeth.1681

Liu, X., Qin, D., Piersanti, A., Zhang, Q., Miceli, C., and Wang, P. (2020). Genome-wide association study identifies candidate genes related to oleic acid content in soybean seeds. *BMC Plant Biol.* 20 (1), 399. doi: 10.1186/s12870-020-02607-w

Liu, X., Huang, M., Fan, B., Buckler, E. S., and Zhang, Z. (2016). Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. *PloS Genet.* 12 (2), e1005767. doi: 10.1371/journal.pgen.1005767

Liu, Y., Wang, D., He, F., Wang, J., Joshi, T., and Xu, D. (2019). Phenotype prediction and genome-wide association study using deep convolutional neural network of soybean. *Front. Genet.* 10, 109110. doi: 10.3389/fgene.2019.01091

Lu, Y., Zhang, S., Shah, T., Xie, C., Hao, Z., Li, X., et al. (2010). Joint linkage–linkage disequilibrium mapping is a powerful approach to detecting quantitative trait loci underlying drought tolerance in maize. *Proc. Natl. Acad. Sci. U.S.A.* 107, 19585–19590. doi: 10.1073/pnas.1006105107

Ma, Y., Marzougui, A., Coyne, C. J., Sankaran, S., Main, D., Porter, L. D., et al. (2020). Dissecting the genetic architecture of aphanomyces root rot resistance in lentil by QTL mapping and genome-wide association study. *Int. J. Mol. Sci.* 21, 2129. doi: 10.3390/ijms21062129

Maalouf, F., Abou-Khater, L., Babiker, Z., Jighly, A., Alsamman, A. M., Hu, J., et al. (2022). Genetic dissection of heat stress tolerance in faba bean (Vicia faba l.) using GWAS. *Plants* 11, 1108. doi: 10.3390/plants11091108

Maciukiewicz, M., Marshe, V. S., Hauschild, A.-C., Foster, J. A., Rotzinger, S., Kennedy, J. L., et al. (2018). GWAS-based machine learning approach to predict duloxetine response in major depressive disorder. *J. Psychiatr. Res.* 99 (April), 62–68. doi: 10.1016/j.jpsychires.2017.12.009

Maenhout, S., De Baets, B., Haesaert, G., and Van Bockstaele, E. (2007). Support vector machine regression for the prediction of maize hybrid performance. *Theor. Appl. Genet.* 115, 1003–1013. doi: 10.1007/s00122-007-0627-9

Marinangeli, C., Curran, J., Barr, S., Slavin, J., Puri, S., Swaminathan, S., et al. (2017). Enhancing nutrition with pulses: defining a recommended serving size for adults. *Nutr. Rev.* 75, 990–1006. doi: 10.1093/nutrit/nux058

Maurer, A., Draba, V., Jiang, Y., Schnaithmann, F., Sharma, R., Schumann, E., et al. (2015). Modelling the genetic architecture of flowering time control in barley through nested association mapping. *BMC Genomics* 16 (1), 290. doi: 10.1186/s12864-015-1459-7

Meuwissen, T. H., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829. doi: 10.1093/genetics/157.4.1819

McCouch, S. R., Wright, M. H., Tung, C. W., Maron, L. G., McNally, K. L., Fitzgerald, M., et al. (2016). Open access resources for genome-wide association mapping in rice. *Nat. Commun.* 7 (1), 10532. doi: 10.1038/ncomms10532

Mieth, B., Kloft, M., Rodríguez, J. A., Sonnenburg, Sören, Vobruba, R., Morcillo-Suárez, C., et al. (2016). Combining multiple hypothesis testing with machine learning increases the statistical power of genome-wide association studies. *Sci. Rep.* 6 (1), 36671. doi: 10.1038/srep36671

Mir, R. R., Reynolds, M., Pinto, F., Khan, M. A., and Bhat, M. A. (2019). High-throughput phenotyping for crop improvement in the genomics era. *Plant Sci.* 282, 60–72. doi: 10.1016/j.plantsci.2019.01.007

Mittag, F., Büchel, F., Saad, M., Jahn, A., Schulte, C., Bochdanovits, Z., et al. (2012). Use of support vector machines for disease risk prediction in genome-wide association studies: concerns and opportunities. *Hum. Mutat.* 33 (12), 1708–1718. doi: 10.1002/humu.22161

Mousavi-Derazmahalleh, M., Bayer, P. E., Hane, J. K., Valliyodan, B., Nguyen, H. T., Nelson, M. N., et al. (2019). Adapting legume crops to climate change using genomic approaches. *Plant Cell& Environ.* 42, 6–19. doi: 10.1111/pce.13203

Muraya, M. M., Chu, J., Zhao, Y., Junker, A., Klukas, C., Reif, J. C., et al. (2017). Genetic variation of growth dynamics in maize (*Zea mays* l.) revealed through automated non-invasive phenotyping. *Plant J.* 89 (2), 366–380.

Myles, S., Peiffer, J., Brown, P. J., Ersoz, E. S., Zhang, Z., Costich, D. E., et al. (2009). Association mapping: critical considerations shift from genotyping to experimental design. *Plant Cell* 21, 2194–2202. doi: 10.1105/tpc.109.068437

Nadeem, M. A., Nawaz, M. A., Shahid, M. Q., Doğan, Y., Comertpay, G., Yıldız, M., et al. (2018). DNA Molecular markers in plant breeding: current status and recent advancements in genomic selection and genome editing. *Biotechnol. Biotechnol. Equip.* 32 (2), 261–285. doi: 10.1080/13102818.2017.1400401

Nair, R. J., and Pandey, M. K. (2021). Role of molecular markers in crop breeding: a review. *Agric. Rev.* 1, 1–8. doi: 10.18805/ag.R-2322

Ndjiondjop, M. N., Semagn, K., Zhang, J., Arnaud, C. G., Kpeki, S. B., Alphonse, G., et al. (2018). Development of species diagnostic SNP markers for quality control genotyping in four rice (Oryza l.) species. *Mol. Breed.* 38, 131. doi: 10.1007/s11032-018-0885-z

Neupane, S., Wright, D. M., Martinez, R. O., Butler, J., Weller, Ji., and Bett, K. (2022). Focusing the GWAS lens on days to flower using latent variable phenotypes derived from global multi- environment trials. *The Plant Genome* 16, e20269. doi: 10.1101/2022.03.10.483676

Nguyen, T.-T., Huang, J. Z., Wu, Q., Nguyen, T. T., and Li, M. J. (2015). Genome-wide association data classification and SNPs selection using two-stage quality-based random forests. *BMC Genomics* 16 (2), S55. doi: 10.1186/1471-2164-16-S2-S5

Nguyen, T. T., Taylor, P. W. J., Redden, R. J., and Ford, R. (2004). Genetic diversity estimates in cicer using AFLP analysis. *Plant Breed.* 123 (2), 173–179. doi: 10.1046/j.1439-0523.2003.00942.x

Nice, L. M., Steffenson, B. J., Brown-Guedira, G. L., Akhunov, E. D., Liu, C., Kono, T. J. Y., et al. (2016). Development and genetic characterization of an advanced backcross-nested association mapping (AB-NAM) population of wild cultivated barley. *Genetics* 203, 1453. doi: 10.1534/genetics.116.190736

Nkhata, W., Shimelis, H., Melis, R., Chirwa, R., Mzengeza, T., Mathew, I., et al. (2021). Genomewide association analysis of bean fly resistance and agro-morphological traits in common bean. *PloS One* 16 (4), e0250729. doi: 10.1371/journal.pone.0250729

Oetjens, M. T., Brown-Gentry, K., Goodloe, R., Dilks, H. H., and Crawford, D. C. (2016). Population stratification in the context of diverse epidemiologic surveys sans genome-wide data. *Front. Genet.* 7. doi: 10.3389/fgene.2016.00076

Ogutu, J. O., Piepho, H.-P., and Schulz-Streeck, T. (2011). A comparison of random forests, boosting and support vector machines for genomic selection. *BMC Proc.* 5 Suppl 3, S11. doi: 10.1186/1753-6561-5-S3-S11

Oh, J., Kerns, S., Ostrer, H., Simon, N. P., Barry, R., and Joseph, O. D. (2017). Computational methods using genome-wide association studies to predict radiotherapy complications and to identify correlative molecular processes. *Sci. Rep.* 7, 43381. doi: 10.1038/srep43381

Olukolu, B. A., Mayes, S., Stadler, F., Quat Ng, N., Fawole, I., Dominique, D., et al. (2012). Genetic diversity in bambara groundnut (*Vigna subterranea* (L.) verdc.) as revealed by phenotypic descriptors and DArT marker analysis. *Genet. Resour. Crop Evol.* 59 (3), 347–358. doi: 10.1007/s10722-011-9686-5

Orsak, A., Deokar, A., and Tar'an, B. (2019). 27. genome-wide association study for seed quality traits in chickpea. *FOR THE People AND THE PLANET* 40.

Packer, R. J., Williams, A. T., Hennah, W., Eisenberg, M. T., Shrine, N., Fawcett, K. A., et al. (2023). DeepPheWAS: an r package for phenotype generation and association analysis for phenome-wide association studies. *Bioinformatics* 39 (4), btad073. doi: 10.1093/bioinformatics/btad073

Pan, L., He, J., Zhao, T., Xing, G., Wang, Y., Yu, D., et al. (2018). Efficient QTL detection of flowering date in a soybean RIL population using the novel restricted two-stage multi-locus GWAS procedure. *Theor. Appl. Genet.* 131, 2581–2599. doi: 10.1007/s00122-018-3174-7

Pandey, M. K., Gangurde, S. S., Sharma, V., Pattanashetti, S. K., Naidu, G. K., Faye, I., et al. (2020). Improved genetic map identified major QTLs for drought tolerance-and iron deficiency tolerance-related traits in groundnut. *Genes* 12 (1), 37.

Pandey, M. K., Agarwal, G., Kale, S. M., Clevenger, J., Nayak, S. N., Sriswathi, M., et al. (2017). Development and evaluation of a high-density genotyping 'Axiom_Arachis' array with 58 K SNPs for accelerating genetics and breeding in groundnut. *Sci. Rep.* 7, 40577. doi: 10.1038/srep40577

Pandey, V., Nutter, R. C., and Prediger, E. (2008). "Applied biosystems SOLID system: ligation-based sequencing," in *Next generation genome sequencing: towards personalized medicine*. Ed. M. Janitz (Germany: Wiley-VCH, Weinheim), 431–444.

Parida, S. K., Kujur, A., Bajaj, D., Das, S., Srivastava, R., Badoni, S., et al. (2017). Integrative genome- wide association studies (GWAS) to understand complex genetic architecture of quantitative traits in chickpea.

Patil, G., Do, T., Vuong, T., Babu, V., Dong Lee, J., Juhi, C., et al. (2016). Genomic-assisted haplotype analysis and the development of high- throughput SNP markers for salinity tolerance in soybean. *Sci. Rep.* 6, 19199. doi: 10.1038/srep19199

Pawar, S., Pandit, E., Mohanty, I. C., Saha, D., and Pradhan, S. K. (2021). Population genetic structure and association mapping for iron toxicity tolerance in rice. *PloS One* 16 (3), e0246232. doi: 10.1371/journal.pone.0246232

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., and Hoekstra, H. E. (2012). Double digest RADseq: an inexpensive method for *De novo* SNP discovery and genotyping in model and non-model species. *PloS One* 7 (5), e37135. doi: 10.1371/journal.pone.0037135

Pividori, M., Rajagopal, P. S., Barbeira, A., Liang, Y., Melia, O., Bastarache, L., et al. (2020). PhenomeXcan: mapping the genome to the phenome through the transcriptome. *Sci. Adv.* 6 (37), eaba2083. doi: 10.1126/sciadv.aba2083

Poland, J., Endelman, J., Dawson, J., Rutkoski, J., Wu, S., Manes, Y., et al. (2012). Genomic selection in wheat breeding using genotyping-by-sequencing. *Plant Genome* 5 (3). doi: 10.3835/plantgenome2012.06.0006

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38 (8), 904–995. doi: 10.1038/ng1847

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155 (2), 945–959. doi: 10.1093/genetics/155.2.945

Puspitasari, W., Allemann, B., Angra, D., Appleyard, H., Ecke, W., Möllers, C., et al. (2022). NIRS for vicine and convicine content of faba bean seed allowed GWAS to prepare for marker- assisted adjustment of seed quality of German winter faba beans. *J. Cultivated Plants* 74 (01-02). doi: 10.5073/JfK.2022.01-02.01

Raju, N. L., Gnanesh, B. N., Lekha, P., Jayashree, B., Pande, S., Hiremath, P. J., et al. (2010). The first set of EST resource for gene discovery and marker development in pigeonpea (*Cajanus cajan* l.). *BMC Plant Biol.* 10 (1), 1–22. doi: 10.1186/1471-2229-10-45

Raman, R., Warren, A., Krysinska-Kaczmarek, M., Rohan, M., Sharma, N., Dron, N., et al. (2022). Genome-wide association analyses track genomic regions for resistance to ascochyta rabiei in Australian chickpea breeding germplasm. *Front. Plant Sci.* 1314. doi: 10.3389/fpls.2022.877266

Rasool, S., Mahajan, R., Nazir, M., Bhat, K. A., Shikari, A. B., Ali, G., et al. (2022). SSR and GBS based GWAS study for identification of QTLs associated with nutritional elemental in common bean (*Phaseolus vulgaris* l.). *Scientia Hortic.* 306, 111470. doi: 10.1016/j.scienta.2022.111470

Ravelombola, W. S., Qin, J., Shi, A., Nice, L., Bao, Y., Lorenz, A., et al. (2019). Genomewide association study and genomic selection for soybean chlorophyll content associated with soybean cyst nematode tolerance. *BMC Genomics* 20, 1–18. doi: 10.1186/s12864-019-6275-z

Ravelombola, W. S., Qin, J., Shi, A., Nice, L., Bao, Y., Lorenz, A., et al (2020). Genome-wide association study and genomic selection for tolerance of soybean biomass to soybean cyst nematode infestation. romagnoni, Alberto, Simon jégou, kristel van Steen, gilles wainrib, and Jean-Pierre hugot. 2019. "Comparative performances of machine learning methods for classifying crohn disease patients using genome-wide genotyping data" *Sci. Rep.* 9 (1), 10351. doi: 10.1038/s41598-019-46649-z

Reif, J. C., Liu, W., Gowda, M., et al. (2010). Genetic basis of agronomically important traits in sugar beet (Beta vulgaris l.) investigated with joint linkage

association mapping. *Theor. Appl. Genet.* 121, 1489–1499. doi: 10.1007/s00122-010-1405-7

Revilla, P., Rodríguez, V. M., Ordás, A., Rincent, R., Charcosset, A., Giauffret, C., et al. (2016). Association mapping for cold tolerance in two large maize inbred panels. *BMC Plant Biol.* 16 (1), 127. doi: 10.1186/s12870-016-0816-2

Rincker, K., Lipka, A. E., and Diers, B. W. (2016). Genome-wide association study of brown stem rot soybean across multiple populations. *Plant Genome* 9 (2). doi: 10.3835/plantgenome2015.08.0064

Roorkiwal, M., Rathore, A., Das, R. R., Singh, M. K., Jain, A., Srinivasan, S., et al. (2016). Genome-enabled prediction models for yield related traits in chickpea. *Front. Plant Sci.* 7, 1666. doi: 10.3389/fpls.2016.01666

Romagnoni, A., Jégou, S., Van Steen, K., Gilles, W., Hugot, J., et al. (2019). Comparative performances of machine learning methods for classifying crohn disease patients using genome-wide genotyping data. *Sci. Rep.* 9, 10351. doi: 10.1038/s41598-019-46649-z

Roorkiwal, M., Jain, A., Kale, S. M., Doddamani, D., Chitikineni, A., Thudi, M., et al. (2018). Development and evaluation of high density SNP array (Axiom ® CicerSNP array) for high resolution genetic mapping and breeding applications in chickpea. *Plant Biotechnol. J.* 16, 890–901. doi: 10.1111/pbi.12836

Roorkiwal, M., Rathore, A., Das, R. R., Singh, M. K., Jain, A., Srinivasan, S., et al. (2016). Genome-enabled prediction models for yield related traits in chickpea. *Front. Plant Sci.* 7. doi: 10.3389/fpls.2016.01666

Roorkiwal, M., Von Wettberg, E. J., Upadhyaya, H. D., Warschefsky, E., Rathore, A., and Varshney, R. K. (2014). Exploring germplasm diversity to understand the domestication process in cicer spp. using SNP and DArT markers. *PloS One* 9 (7), e102016. doi: 10.1371/journal.pone.0102016

Rothberg, J., and Leamon, J. (2008). The development and impact of 454 sequencing. *Nat. Biotechnol.* 26, 1117–1124. doi: 10.1038/nbt1485

Saiki, R. K., Gelfand, D. H., Stiffel, S., Higuchi, R. H., Horn, G. T., Mullis, K. B., et al. (1985). Primer directed enzymatic amplification of DNA with a thermostaable DNA polymerase. *Science* 239, 487–492. doi: 10.1126/science.2448875

Salgotra, R. K., and Stewart, C. N. (2022). Genetic augmentation of legume crops using genomic resources and genotyping platforms for nutritional food security. *Plants* 11 (14), 1866. doi: 10.3390/plants11141866

Sallam, A., Moursi, Y. S., Martsch, R., and Eltaher, S. (2022). Genome-wide association mapping for root traits associated with frost tolerance in faba beans using KASP-SNP markers. *Front. Genet.* 13, 907267. doi: 10.3389/fgene.2022.907267

Samineni, S., Mahendrakar, M. D., Hotti, A., Chand, U., Rathore, A., and Gaur, P. M. (2022). Impact of heat and drought stresses on grain nutrient content in chickpea: genome-wide marker-trait associations for protein, fe and zn. *Environ. Exp. Bot.* 194, 104688. doi: 10.1016/j.envexpbot.2021.104688

Sandhu, K. S., Shiv, A., Kaur, G., Meena, M. R., Raja, A. K., Vengavasi, K., et al. (2022). Integrated approach in genomic selection to accelerate genetic gain in sugarcane. *Plants* 11, 2139. doi: 10.3390/plants11162139

Saxena, R. K., Kale, S. M., Kumar, V., Parupalli, S., Joshi, S., Singh, V. K., et al. (2017a). Genotyping- by-sequencing of three mapping populations for identification of candidate genomic regions for resistance to sterility mosaic disease in pigeonpea. *Sci. Rep.* 7, 1813. doi: 10.1038/s41598-017-01535-4

Saxena, R. K., Patel, K., Kumar, C. V. S., Tyagi, K., Saxena, K. B., and Varshney, R. K. (2018). Molecular mapping and inheritance of restoration of fertility (Rf) in A4 hybrid system in pigeonpea (*Cajanus cajan* (L.) millsp.). *Theor. Appl. Genet.* 131, 1605–1614. doi: 10.1007/s00122-018-3101-y

Saxena, R. K., Prathima, C., Saxena, K. B., Hoisington, D. A., Singh, N. K., and Varshney, R. K. (2010). Novel SSR markers for polymorphism detection in pigeonpea (*Cajanus* spp.). *Plant Breed.* 129 (2), 142–148. doi: 10.1111/j.1439-0523.2009.01680.x

Saxena, R. K., Rathore, A., Bohra, A., Yadav, P., Das, R. R., Khan, A. W., et al. (2018b). Development and application of high-density axiom® cajanus SNP array with 56 K SNPs to understand the genome architecture of released cultivars and founder genotypes for redefining future pigeonpea breeding programs. *Plant Genome*. doi: 10.3835/plantgenome2018.01.0005

Saxena, R. K., Singh, V. K., Kale, S. M., Tathineni, R., Parupalli, S., Kumar, V., et al. (2017b). Construction of genotyping-by-sequencing based high-density genetic maps and QTL mapping for fusarium wilt resistance in pigeonpea. *Sci. Rep.* 7, 1911. doi: 10.1038/s41598-017-01537-2

Segura, V., Vilhjálmsson, B. J., Platt, A., Korte, A., Seren, Ü, Long, Q., et al. (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat. Genet.* 44 (7), 825–830. doi: 10.1038/ng.2314

Shafiekhani, A., Kadam, S., Fritschi, F. B., and DeSouza, G. N. (2017). Vinobot and vinoculer: two robotic platforms for high-throughput field phenotyping. *Sensors* 17 (1), 214. doi: 10.3390/s17010214

Shaibu, A. S., Motagi, B., and Ayittah, P. (2019a). Genome wide association studies for fatty acids, mineral and proximate compositions in groundnut (*Arachis h*ypogaea l.) seeds. *Pertanika J. Trop. Agric. Sci.* 42 (3), 939–955.

Shaibu, A., Sneller, C., Motagi, B., Chepkoech, J., Chepngetich, M., Miko, Z., et al. (2019b). Genome wide association studies for four physiological traits in groundnut (*Arachis hypogaea* l.) minicore collection. *Agronomy* 10, 192. doi: 10.3390/agronomy10020192

Shaibu, A. S., Sneller, C., Motagi, B. N., Chepkoech, J., Chepngetich, M., Miko, Z. L., et al. (2020). Genome-wide detection of SNP markers associated with four physiological traits in groundnut (*Arachis hypogaea* l.) mini core collection. *Agronomy* 10, 192. doi: 10.3390/agronomy10020192

Shao, Z., Shao, J., Huo, X., Li, W., Kong, Y., Du, H., et al. (2022). Identification of closely associated SNPs and candidate genes with seed size and shape *via* deep re-sequencing GWAS in soybean. *Theor. Appl. Genet.* 135 (7), 2341–2351. doi: 10.1007/s00122-022-04116-w

Shi, L., Shi, T., Broadley, M. R., White, P. J., Long, Y., Meng, J., et al. (2013). High-throughput root phenotyping screens identify genetic loci associated with root architectural traits in brassica napus under contrasting phosphate availabilities. *Ann. Bot.* 112 (2), 381–389. doi: 10.1093/aob/mcs245

Shilpa, H. B., and Lohithaswa, H. C. (2021). Discovery of SNPs in important legumes through comparative genome analysis and conversion of SNPs into PCR-based markers. *J. Genet.* 100 (2), 1–12. doi: 10.1007/s12041-021-01320-3

Shingote, P. R., Gotarkar, D. N., Kale, R. R., Limbalkar, O. M., and Wasule, D. L. (2022). Recent advances and applicability of GBS, GWAS, and GS in soybean. *Genotyping by Sequencing Crop Improvement*, 218–249. doi: 10.1002/9781119745686.ch10

Shrestha, A., Cosenza, F., van Inghelandt, D., Wu, P. Y., Li, J., Casale, F. A., et al. (2022). The double round-robin population unravels the genetic architecture of grain size in barley. *J. Exp. Bot.* 73, 7344–61. doi: 10.1093/jxb/erac369

Singh, C. M., Pratap, A., Gupta, S., Biradar, R. S., and Singh, N. P. (2020). Association mapping for mungbean yellow mosaic India virus resistance in mungbean (Vigna radiata L. Wilczek). *3 Biotech* 10 (2), 33. doi: 10.1007/s13205-019-2035-7

Singh, A., Dikshit, H. K., Mishra, G. P., Aski, M., and Kumar, S. (2019). Association mapping for grain diameter and weight in lentil using SSR markers. *Plant Gene* 20, 100204. doi: 10.1016/j.plgene.2019.100204

Singh, V. K., Khan, A. W., Saxena, R. K., Kumar, V., Kale, S. M., Sinha, P., et al. (2016). Next-generation sequencing for identification of candidate genes for fusarium wilt and sterility mosaic disease in pigeonpea (*Cajanus cajan*). *Plant Biotechnol. J.* 14 (5), 1183–1194. doi: 10.1111/pbi.12470

Singh, A., Sharma, V., Dikshit, H. K., Aski, M., Kumar, H., Thirunavukkarasu, N., et al. (2017). Association mapping unveils favorable alleles for grain iron and zinc concentrations in lentil (*Lens culinaris* subsp. culinaris). *PloS One* 12, e0188296. doi: 10.1371/journal.pone.0188296

Singh, L., Dhillon, G. S., Sarabjit, K., Sandeep, D. K., Amandeep, K., Malik, P., et al. (2022). Genome-wide association study for yield and yield-related traits in diverse blackgram panel (Vigna mungo l. hepper) reveals novel putative alleles for future breeding programs. *Front. Genet.* 13. doi: 10.3389/fgene.2022.849016

Singh, R., Singhal, V., and Randhawa, G. J. (2008). Molecular analysis of chickpea (*Cicer arietinum* l) cultivars using AFLP and STMS markers. *J. Plant Biochem. Biotechnol.* 17 (2), 167–171. doi: 10.1007/BF03263279

Song, Q., Hyten, D. L., Jia, G., Quigley, C. V., Fickus, E. W., Nelson, R. L., et al. (2013). Development and evaluation of SoySNP50K, a high-density genotyping array for soybean. *PloS One* 8 (1), e54985. doi: 10.1371/journal.pone.0054985

Souframanien, J., and Gopalakrishna, T. (2004). A comparative analysis of genetic diversity in blackgram genotypes using RAPD and ISSR markers. *Theor. Appl. Genet.* 109 (8), 1687–1693. doi: 10.1007/s00122-004-1797-3

Srungarapu, R., Mahendrakar, M. D., Mohammad, L. A., Chand, U., Jagarlamudi, V. R., Kondamudi, K. P., et al. (2022). Genome-wide association analysis reveals trait-linked markers for grain nutrient and agronomic traits in diverse set of chickpea germplasm. *Cells* 11, 2457. doi: 10.3390/cells11152457

Stagnari, F., Albino, M., Angelica, G., and Michele, P. (2017). Multiple benefits of legumes for agriculture sustainability: an overview. *Chem. Biol. Technol. Agric.* 4, 2. doi: 10.1186/s40538-016-0085-1

Stojaowski, S. A., Milczarski, P., Hanek, M., Bolibok-Bragoszewska, H., Myśków, B., Kilian, A., et al. (2011). DArT markers tightly linked with the *Rfc1* gene controlling restoration of Male fertility in the CMS-c system in cultivated rye (*Secale cereale* l.). *J. Appl. Genet.* 52 (3), 313–318. doi: 10.1007/s13353-011-0049-x

Sudheesh, S., Kahrood, H. V., Braich, S., Dron, N., Hobson, K., Cogan, N. O., et al. (2021). Application of genomics approaches for the improvement in ascochyta blight resistance in chickpea. *Agronomy* 11 (10), 1937. doi: 10.3390/agronomy11101937

Sui, M., Wang, Y., Bao, Y., Wang, X., Li, R., Lv, Y., et al. (2020). Genome-wide association analysis of sucrose concentration in soybean (Glycine max l.) seed based on high-throughput sequencing. *Plant Genome* 13, e20059. doi: 10.1002/tpg2.20059

Sul, J. H., Bilow, M., Yang, W.-Y., Kostem, E., Furlotte, N., He, D., et al. (2016). Accounting for population structure in gene-by-environment interactions in genome-wide association studies using mixed models. *PloS Genet.* 12, e1005849. doi: 10.1371/journal.pgen.1005849

Sun, X., Liu, D., Zhang, X., Li, W., Liu, H., Hong, W., et al. (2013). SLAF-seq: an efficient method of large-scale *de novo* SNP discovery and genotyping using high-throughput sequencing. *PloS One* 8 (3), e58700. doi: 10.1371/journal.pone.0058700

Tafesse, E. G., Gali, K. K., Lachagari, V. B. R., Bueckert, R., and Warkentin, T. D. (2021). Genome-wide association mapping for heat and drought adaptive traits in pea. *Genes* 12 (12), 1897. doi: 10.3390/genes12121897

Tafesse, E. G., Gali, K. K., Lachagari, V. B. R., Bueckert, R., and Warkentin, T. D. (2021). Genome-wide association mapping for heat and drought adaptive traits in pea. *Genes* 12 (12), 1897. doi: 10.3390/genes12121897

Talebi, R., Fayaz, F., Mardi, M., Pirsyedi, S. M., and Naji, A. M. (2008). Genetic relationships among chickpea (*Cicer arietinum*) elite lines based on RAPD and agronomic markers. *Int. J. Agric. Biol.* 8, 301–305.

Tanabata, T., Shibaya, T., Hori, K., Ebana, K., and Yano, M. (2012). SmartGrain: high-throughput phenotyping software for measuring seed shape through image analysis. *Plant Physiol.* 160 (4), 1871–1880. doi: 10.1104/pp.112.205120

Thompson, J. A., Nelson, R. L., and Vodkin, L. O. (1998). Identification of diverse soybean germplasm using RAPD markers. *Crop Sci.* 38 (5), pp.1348–1355. doi: 10.2135/cropsci1998.0011183X000380050033x

Thudi, M., Khan, A. W., Kumar, V., Gaur, P. M., Katta, K., Garg, V., et al. (2016). Whole genome re-sequencing reveals genome-wide variations among parental lines of 16 mapping populations in chickpea (*Cicer arietinum* l.). *BMC Plant Biol.* 16 Suppl 1, 10. doi: 10.1186/s12870-015-0690-3

Thudi, M., Upadhyaya, H. D., Rathore, A., Gaur, P. M., Krishnamurthy, U. G., Pillay, M., et al. (2002). Genetic diversity in musa acuminata colla and musa balbisiana colla and some of their natural hybrids using AFLP markers. *Theor. Appl. Genet.* 104 (8), 1246–1252. doi: 10.1007/s00122-002-0914-4

Thudi, M., Upadhyaya, H. D., Rathore, A., Gaur, P. M., Krishnamurthy, L., Roorkiwal, M., et al. (2014). Genetic dissection of drought and heat tolerance in chickpea through genome-wide and candidate gene-based association mapping approaches. *Plos One* 9 (5), e96758. doi: 10.1371/journal.pone.0096758

Tibshirani, R. (1996). Regression shrinkage and selection *via* the lasso. journal of the royal statistical society. *Ser. B (Methodological)* 58 (1), 267–288. http://www.jstor.org/stable/2346178.

Topp, C. N., Iyer-Pascuzzi, A. S., Anderson, J. T., Lee, C. R., Zurek, P. R., Symonova, O., et al. (2013). 3D phenotyping and quantitative trait locus mapping identify core regions of the rice genome controlling root architecture. *Proc. Nat. Acad. Sci.* 110 (18), E1695–E1704. doi: 10.1073/pnas.1304354110

Torkamaneh, D., Laroche, J., Boyle, B., Hyten, D. L., and Belzile, F. (2021). A bumper crop of SNPs in soybean through high- density genotyping-by-sequencing (HD- GBS). *Plant Biotechnol. J.* 19 (5), 860. doi: 10.1111%2Fpbi.13551

Tran, D. T., Steketee, C. J., Boehm, J. D., Noe, J., and Li, Z. (2019). Genome-wide association analysis pinpoints additional major genomic regions conferring resistance to soybean cyst nematode (Heterodera glycines ichinohe). *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.00401

Ude, G., Pillay, M., Nwakanma, D., and Tenkouano, A. (2002). Genetic diversity in musa acuminata colla and musa balbisiana colla and some of their natural hybrids using AFLP markers. *Theor. Appl. Genet.* 104 (8), 1246–1252. doi: 10.1007/s00122-002-0914-4

Upadhyaya, H. D., Bajaj, D., Das, S., Kumar, V., Gowda, C. L. L., Sharma, S., et al. (2016a). Genetic dissection of seed-iron and zinc concentrations in chickpea. *Sci. Rep.* 6, 1–12. doi: 10.1038/srep24050

Upadhyaya, H. D., Bajaj, D., Das, S., Saxena, M. S., Badoni, S., Kumar, V., et al. (2015). A genome- scale integrated approach aids in genetic dissection of complex flowering time trait in chickpea. *Plant Mol. Biol.* 89, 403–420. doi: 10.1007/s11103-015-0377-z

Upadhyaya, H. D., Bajaj, D., Narnoliya, L., Das, S., Kumar, V., Gowda, C. L. L., et al. (2016b). Genome-wide scans for delineation of candidate genes regulating seed-protein content in chickpea. *Front. Plant Sci.* 7, 302. doi: 10.3389/fpls.2016.00302

Upadhyaya, H. D., Bajaj, D., Srivastava, R., Daware, A., Basu, U., Tripathi, S., et al. (2017). Genetic dissection of plant growth habit in chickpea. *Funct. Integr. Genomics* 17, 711–723. doi: 10.1007/s10142-017-0566-8

Unterseer, S., Bauer, E., Haberer, G., Seidel, M., Knaak, C., Ouzunova, M., et al. (2014). A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k SNP genotyping array. *BMC Genomics* 15 (1), 1–15. doi: 10.1186/1471-2164-15-823

Varshney, R. K., Song, C., Saxena, R. K., Azam, S., Yu, S., Sharpe, A. G., et al. (2013). Draft genome sequence of chickpea (Cicer arietinum) provides a resource for trait improvement. *Nat. Biotechnol.* 31 (3), 240–246. doi: 10.1038/nbt.2491

Varshney, R. K. (2016). Exciting journey of 10 years from genomes to fields and markets: some success stories of genomics-assisted breeding in chickpea, pigeonpea and groundnut. *Plant Sci.* 242, 98–107. doi: 10.1016/j.plantsci.2015.09.009

Varshney, R. K., Graner, A., and Sorrells, M. E. (2005). Genomics-assisted breeding for crop improvement. *Trends Plant Sci.* 10, 621–630. doi: 10.1016/j.tplants.2005.10.004

Varshney, R. K., Pandey, M. K., Bohra, A., Singh, V. K., Thudi, M., and Saxena, R. K. (2019). Toward the sequence-based breeding in legumes in the post-genome sequencing era. *Theor. Appl. Genet.* 132 (3), 797–816. doi: 10.1007/s00122-018-3252-x

Varshney, R. K., Roorkiwal, M., and Sorrells, M. E. (2017). Genomic selection for crop improvement: new molecular breeding strategies for crop improvement 1–258. doi: 10.1007/978-3-319-63170-7

Varshney, R. K., Roorkiwal, M., Sun, S., Prasad, B., Annapurna, C., Mahendar, T., et al. (2021). A chickpea genetic variation map based on the sequencing of 3,366 genomes. *Nature* 599, 622–627. doi: 10.1038/s41586-021-04066-1

Varshney, R. K., Saxena, R. K., Upadhyaya, H. D., Khan, A. W., Yu, Y., Kim, C., et al. (2017). Whole-genome resequencing of 292 pigeonpea accessions identifies genomic regions associated with domestication and agronomic traits. *Nat. Genet.* 49, 1082–1088. doi: 10.1038/ng.3872

Verma, S., Gupta, S., Bandhiwal, N., Kumar, T., Bharadwaj, C., and Bhatia, S. (2015). High-density linkage map construction and mapping of seed trait QTLs in chickpea (Cicer arietinum l.) using genotyping-by-Sequencing (GBS). *Sci. Rep.* 5 (1), 17512. doi: 10.1038/srep17512

Vuong, T. D., Sonah, H., Meinhardt, C. G., Deshmukh, R., Kadam, S., Nelsonet, R. L., et al. (2015). Genetic architecture of cyst nematode resistance revealed by genome-wide association study in soybean. *BMC Genomics* 16, 593. doi: 10.1186/s12864-015-1811-y

Wu, X., Islam, A. S. M. F., Limpot, N., Mackasmiel, L., Mierzwa, J., Cortés, A. J., et al. (2020). Genome-wide SNP identification and association mapping for seed mineral concentration in mung bean (Vigna radiata l.). *Fron. Gen.* 11. doi: 10.3389/fgene.2020.00656

Wang, Q., Tian, F., Pan, Y., Buckler, E. S., and Zhang, Z. (2014). A SUPER powerful method for genome wide association study. *PloS One* 9, e107684. doi: 10.1371/journal.pone.0107684

Wang, J., Yan, C., Li, Y., Li, C., Zhao, X., Yuan, C., et al. (2019). GWAS discovery of candidate genes for yield-related traits in peanut and support from earlier QTL mapping studies. *Genes* 10, 803. doi: 10.3390/genes10100803

Wang, L., Yang, Y., Zhang, S., Che, Z., Yuan, W., and Yu, D. (2020). GWAS reveals two novel loci for photosynthesis-related traits in soybean. *Mol. Genet. Genomics* 295, 705–716. doi: 10.1007/s00438-020-01661-1

Wang, H., Yue, T., Yang, J., Wu, W., and Xing, E. P. (2019). Deep mixed model for marginal epistasis detection and population stratification correction in genome-wide association studies. *BMC Bioinf.* 20 (23), 6565. doi: 10.1186/s12859-019-3300-9

Warsame, A. O., Angra, D., and O'Sullivan, D. (2019). Identification of a candidate gene controlling hilum colour in faba bean. *For People Planet* 31, 1–44.

Weber, J. L., and May, P. E. (1989). Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am. J. Hum. Genet.* 44 (3), 388–396.

Wen, Z., Boyse, J. F., Song, Q., Cregan, P. B., and Wang, D. (2015). Genomic consequences of selection and genome-wide association mapping in soybean. *BMC Genomics* 16, 1–14. doi: 10.1186/s12864-015-1872-y

Wu, X., Ren, C., Joshi, T., Vuong, T., Xu, D., and Nguyen, H. T. (2010). SNP discovery by high- throughput sequencing in soybean. *BMC Genomics* 11 (1), 1–10. doi: 10.1186/1471-2164-11-469

Wurschum, T., Liu, W., Gowda, M., Maurer, H. P., Fischer, S., Schechert, A., et al. (2012). Comparison of biometrical models for joint linkage association mapping. *Heredity* 108, 332–340. doi: 10.1038/hdy.2011.78

Yadav, S., Jackson, P., Wei, X., Ross, E. M., Aitken, K., Deomano, E., et al. (2020). Accelerating genetic gain in sugarcane breeding using genomic selection. *Agronomy* 10, 585. doi: 10.3390/agronomy10040585

Yadav, K., Yadav, S. K., Yadav, A., Pandey, V. P., and Dwivedi, U. N. (2014). Comparative analysis of genetic diversity among cultivated pigeonpea (Cajanus cajan (L) millsp.) and its wild relatives (*C. albicans* and *C. lineatus*) using randomly amplified polymorphic DNA (RAPD) and inter simple sequence repeat (ISSR) fingerprinting. *Am. J. Plant Sci.* 5 (11), 1665. doi: 10.3390/agronomy10040585

Yan, L., Hofmann, N., Li, S., Ferreira, M. E., Song, B., Jiang, G., et al. (2017). Identification of QTL with large effect on seed weight in a selective population of soybean with genome-wide association and fixation index analyses. *BMC Genomics* 18, 1–11. doi: 10.1186/s12864-017-3922-0

Yang, W., Feng, H., Zhang, X., Zhang, J., Doonan, J. H., Batchelor, W. D., et al. (2020). Crop phenomics and high-throughput phenotyping: past decades, current challenges, and future perspectives. *Mol. Plant* 13, 187–214. doi: 10.1016/j.molp.2020.01.008

Yang, W., Guo, Z., Huang, C., Duan, L., Chen, G., Jiang, N., et al. (2014). Combining high-throughput phenotyping and genome-wide association studies to reveal natural genetic variation in rice. *Nat. Commun.* 5 (1), 1–9. doi: 10.1038/ncomms6087

Yang, S. Y., Saxena, R. K., Kulwal, P. L., Ash, G. J., Dubey, A., Harper, J. D., et al. (2011). The first genetic map of pigeon pea based on diversity arrays technology (DArT) markers. *J. Genet.* 90 (1), 103–109. doi: 10.1007/s12041-011-0050-5

Yang, Y., Wang, L., Zhang, D., Cheng, H., Wang, Q., Yang, H., et al. (2020). GWAS identifies two novel loci for photosynthetic traits related to phosphorus efficiency in soybean. *Mol. Breed.* 40, 1–14. doi: 10.1007/s11032-020-01112-0

You, Q., Yang, X., Peng, Z., Xu, L., and Wang, J. (2018). Development and applications of a high throughput genotyping tool for polyploid crops: single nucleotide polymorphism (SNP) array. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.00104

Yu, Z., Chang, F., Lv, W., Sharmin, R. A., Wang, Z., Kong, J., et al. (2019). Identification of QTN and candidate gene for seed-flooding tolerance in soybean [*Glycine max* (L.) merr.] using genome- wide association study (GWAS). *Genes* 10, 957.

Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* 38 (2), 203–208. doi: 10.1038/ng1702

Yuan, M., and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc.: Ser. B (Stat.Methodology)* 68, 49–67. doi: 10.1111/j.1467-9868.2005.00532.x

Zatybekov, A., Abugalieva, S., Didorenko, S., Gerasimova, Y., Sidorik, I., Anuarbek, S., et al. (2017). GWAS of agronomic traits in soybean collection included in breeding pool in Kazakhstan. *BMC Plant Biol.* 17 (1), 1–8. doi: 10.1186/s12870-017-1125-0

Zatybekov, A., Abugalieva, S., Didorenko, S., Rsaliyev, A., and Turuspekov, Y. (2018). GWAS of a soybean breeding collection from South East and South Kazakhstan for resistance to fungal diseases. *Vavilov Journal of Genetics and Breeding* 22 (5), 536–543. doi: 10.18699/vj18.392

Zavinon, F., Adoukonou-Sagbadja, H., Keilwagen, J., Lehnert, H., Ordon, F., and Perovic, D. (2020). Genetic diversity and population structure in beninese pigeon pea [*Cajanus cajan* (L.) huth] landraces collection revealed by SSR and genome wide SNP markers. *Genet. Resour. Crop Evol.* 67 (1), 191–208. doi: 10.1007/s10722-019-00864-9

Zeng, A., Chen, P., Korth, K., Hancock, F., Pereira, A., Brye, K., et al. (2017). Genome-wide association study (GWAS) of salt tolerance in worldwide soybean germplasm lines. *Mol. Breed.* 37, 1–14. doi: 10.1007/s11032-017-0634-8

Zhang, H., Chu, Y., Dang, P., Tang, Y., Jiang, T., Clevenger, J. P., et al. (2020). Identification of QTLs for resistance to leaf spots in cultivated peanut (*Arachis hypogaea* l.) through GWAS analysis. *Theor. Appl. Genet.* 133, 2051–2061. doi: 10.1007/s00122-020-03576-2

Zhang, Z., Ersoz, E., Lai, C. Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., et al. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* 42, 355–360. doi: 10.1038/ng.546

Zhang, X., Huang, C., Wu, D., Qiao, F., Li, W., Duan, L., et al. (2017a). High-throughput phenotyping and QTL mapping reveal the genetic architecture of maize plant growth. *Plant Physiol.* 173 (3), 1554–1564. doi: 10.1104/pp.16.01516

Zhang, C., McGee, R. J., Vandemark, G. J., and Sankaran, S. (2021). Crop performance evaluation of chickpea and dry pea breeding lines across seasons and locations using phenomics data. *Front. Plant Sci.* 12, 640259. doi: 10.3389/fpls.2021.640259

Zhang, J., Song, Q., Cregan, P. B., and Jiang, G.-L. (2016). Genome-wide association study, genomic prediction, and marker-assisted selection for seed weight in soybean (Glycine max). *Theor. Appl. Genet.* 129, 117–130. doi: 10.1007/s00122-015-2614-x

Zhang, J., Song, Q., Cregan, P. B., Nelson, R. L., Wang, X., Wu, J., et al. (2015). Genome-wide association study for flowering time, maturity dates and plant height in early maturing soybean (*Glycine max*) germplasm. *BMC Genomics* 16, 1–11. doi: 10.1186/s12864-015-1441-4

Zhang, Z., Todhunter, R. J., Buckler, E. S., and Van Vleck, L. D. (2007). Technical note: use of marker-based relationships with multiple-trait derivative-free restricted maximal likelihood. *J. Anim. Sci.* 85, 881–885. doi: 10.2527/jas.2006-656

Zhang, J., Wen, Z., Li, W., Zhang, Y., Zhang, L., Dai, H., et al. (2017b). Genome-wide association study for soybean cyst nematode resistance in Chinese elite soybean cultivars. *Mol. Breed.* 37, 1–10. doi: 10.1007/s11032-017-0665-1

Zhang, T., Wu, T., Wang, L., Jiang, B., Zhen, C., Yuan, S., et al. (2019). A combined linkage and GWAS analysis identify QTLs linked to soybean seed protein and oil content. *Int. J. Mol. Sci.* 20, 5915. doi: 10.3390/ijms20235915

Zhang, Z., Xu, Y., Liu, J., and Kwoh, C. K. (2012). "Identify predictive SNP groups in genome-wide association study: a sparse learning approach. *Procedia computer science*," in *Proceedings of the 3rd International Conference on Computational Systems-Biology and Bioinformatics*, 11 (January), 107–114. doi: 10.1016/j.procs.2012.09.012

Zhang, W., Xu, W., Zhang, H., Liu, X., Cui, X., Li, S., et al. (2021). Comparative selective signature analysis and high-resolution GWAS reveal a new candidate gene controlling seed weight in soybean. *Theor. Appl. Genet.* 134, 1329–1341. doi: 10.1007/s00122-021-03774-6

Zhang, X., Zhang, J., He, X., Wang, Y., Ma, X., and Yin, D. (2017c). Genome-wide association study of major agronomic traits related to domestication in peanut. *Front. Plant Sci.* 8, 611. doi: 10.3389/fpls.2017.01611

Zhao, J., Bayer, P. E., Ruperao, P., Saxena, R. K., Khan, A. W., Golicz, A. A., et al. (2020). Trait associations in the pangenome of pigeon pea (Cajanus cajan). *Plant Biotechnol. J.* 18 (9), 1946–1954. doi: 10.1111/pbi.13354

Zhao, X., Jiang, H., Feng, L., Qu, Y., Teng, W., Qiu, L., et al. (2019). Genome-wide association and transcriptional studies reveal novel genes for unsaturated fatty acid synthesis in a panel of soybean accessions. *BMC Genomics* 20, 1–16. doi: 10.1186/s12864-019-5449-z

Zhou, X., Xia, Y., Liao, J., Liu, K., Li, Q., Dong, Y., et al. (2016). Quantitative trait locus analysis of late leaf spot resistance and plant-type-related traits in cultivated peanut (Arachis hypogaea l.) under multi-environments. *PloS One* 11 (11), e0166873.

Zhou, X., Xia, Y., Ren, X., Chen, Y., Huang, L., Huang, S., et al. (2014). Construction of a SNP-based genetic linkage map in cultivated peanut based on large scale marker development using next-generation double-digest restriction-site-associated DNA sequencing (ddRADseq). *BMC Genomics* 15, 1–14.

Zou, K., Kang, D., Kim, K.-S., and Jun, T.-H. (2020). Screening of salinity tolerance and genome- wide association study in 249 peanut accessions (*Arachis hypogaea* l.). *Plant Breed. Biotechnol.* 8, 434–438. doi: 10.9787/PBB.2020.8.4.434

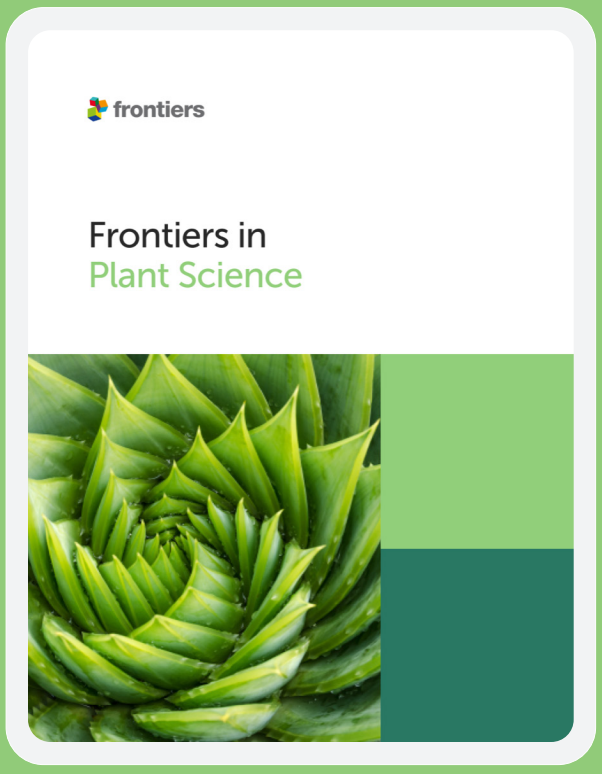
Frontiers in
Plant Science

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