Journal of Genetics and Genomics 49 (2022) 833–846

Contents lists available at ScienceDirect

Journal of Genetics and Genomics

Journal homepage: www.journals.elsevier.com/journal-of-genetics-and-genomics/

Review

Plant pan-genomics: recent advances, new challenges, and roads ahead

Wei Li [1], Jianan Liu [1], Hongyu Zhang, Ze Liu, Yu Wang, Longsheng Xing, Qiang He, Huilong Du*

School of Life Sciences, Institute of Life Sciences and Green Development, Hebei University, Baoding, Hebei 071000, China

ARTICLE INFO

Article history:
Received 21 April 2022
Received in revised form
9 June 2022
Accepted 10 June 2022
Available online 21 June 2022

Keywords:
Plant pan-genome
Genomic evolution
Crop domestication
Graph-based genome

ABSTRACT

Pan-genomics can encompass most of the genetic diversity of a species or population and has proved to be a powerful tool for studying genomic evolution and the origin and domestication of species, and for providing information for plant improvement. Plant genomics has greatly progressed because of improvements in sequencing technologies and the rapid reduction of sequencing costs. Nevertheless, pan-genomics still presents many challenges, including computationally intensive assembly methods, high costs with large numbers of samples, ineffective integration of big data, and difficulty in applying it to downstream multi-omics analysis and breeding research. In this review, we summarize the definition and recent achievements of plant pan-genomics, computational technologies used for pan-genome construction, and the applications of pan-genomes in plant genomics and molecular breeding. We also discuss challenges and perspectives for future pan-genomics studies and provide a detailed pipeline for sample selection, genome assembly and annotation, structural variation identification, and construction and application of graph-based pan-genomes. The aim is to provide important guidance for plant pan-genome research and a better understanding of the genetic basis of genome evolution, crop domestication, and phenotypic diversity for future studies.

Introduction

The plant kingdom has amazing diversity and importantly provides a variety of resources and food energy intake for humans (Food and Agriculture Organization of the United Nations, 1995). The estimated number of land plant species is approximately 391,000, and their genomes are unusually diverse and complicated with genome sizes that vary dramatically from approximately 60 Mb to 150 Gb (Pellicer et al., 2010; Fleischmann et al., 2014; Kuroiwa et al., 2016; Willis, 2017). Polyploidization events and variations in the amounts of repetitive DNA have played important roles in influencing the different sizes of plant genomes, which are vital to plant speciation and evolution (Paterson et al., 2010). The dynamics of transposable elements (Jumper et al., 2021), along with self-incompatibility, have long been recognized as significant evolutionary forces that contribute to plant genome changes (Takayama and Isogai, 2005; Igic et al., 2008; Ambrožová et al., 2011; Ibarra-Laclette et al., 2013; Casacuberta et al., 2016). All of these properties, high repetitive DNA content, high degree of heterozygosity, and polyploidy, make it technically challenging and time-consuming to generate high-quality plant genome assemblies.

High-quality reference genome sequences are the prerequisite and basis for promoting fundamental and applied research in plants and animals. Triggered by developments in computing power, sequencing technologies, and assembly methods, the genomes of more than 700 plants species, from non-vascular to flowering, have been released in the past 20 years (Sun et al., 2021). Third-generation sequencing technologies, such as those that use the PacBio and Oxford Nanopore platforms, can generate reads with significantly increased lengths and they have been widely applied along with well-established assembly algorithms to construct large and complicated plant genomes at unprecedented high resolution (Koren et al., 2017; Cheng et al., 2021; Niu et al., 2022). The 25.4 Gb high-quality genome of Chinese pine, which is the largest gymnosperm genome released so far, was constructed by combining long-read PacBio and Hi-C sequencing technologies (Niu et al., 2022). A chromosome-scale

* Corresponding author.
E-mail address: huilongdu@hbu.edu.cn (H. Du).

[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.jgg.2022.06.004

genome assembly of bread wheat that is 14.66-Gb long with a contig N50 length of 30.22 Mb has also been reported (Athiyannan et al., 2022). The construction of high-quality assemblies of these very large and complicated plant genomes indicates the significant progress that has been made in giga-genome assembly. Furthermore, the emergence of high-fidelity (HiFi) sequencing technologies along with haplotype-resolved assembly software have greatly facilitated the exploration of polyploid and highly heterozygous plant genomes (Cheng et al., 2021). Allele-aware autopolyploid and heterozygous genomes of cultivated alfalfa, potato, sugarcane, and tea have been constructed by integrating HiFi and Hi-C data (Zhang et al., 2018; Chen et al., 2020; Zhang et al., 2021b; Sun et al., 2022). The advances in computing power and sequencing and assembly technologies have promoted the construction of almost complete genomes, even gap-free genomes, which has provided a solid foundation for comparative genomics analysis among different plant accessions and helped to minimize the negative effects caused by incomplete genome assembly.

The dynamics of plant genomes and processes such as the amplification of transposable elements, gene tandem duplication, genome rearrangements, and mutations can lead to a continuum of changes from single-nucleotide polymorphisms (SNPs), gene presence/absence variations (PAVs), to structural variations (SVs) that provide the raw material for natural selection, phenotypic diversity, and adaptation (McClintock, 1956; Gabur et al., 2019; Tao et al., 2019). The availability of high-quality genomes of more and more species has led to the realization that a single genome may not be enough to reflect the landscape of a species because of the large numbers of variations between accessions. Therefore, the "pan-genome" concept was conceived to represent all the genetic information of a species, including core genes that are present in all strains and dispensable genes that are present only in a subset of

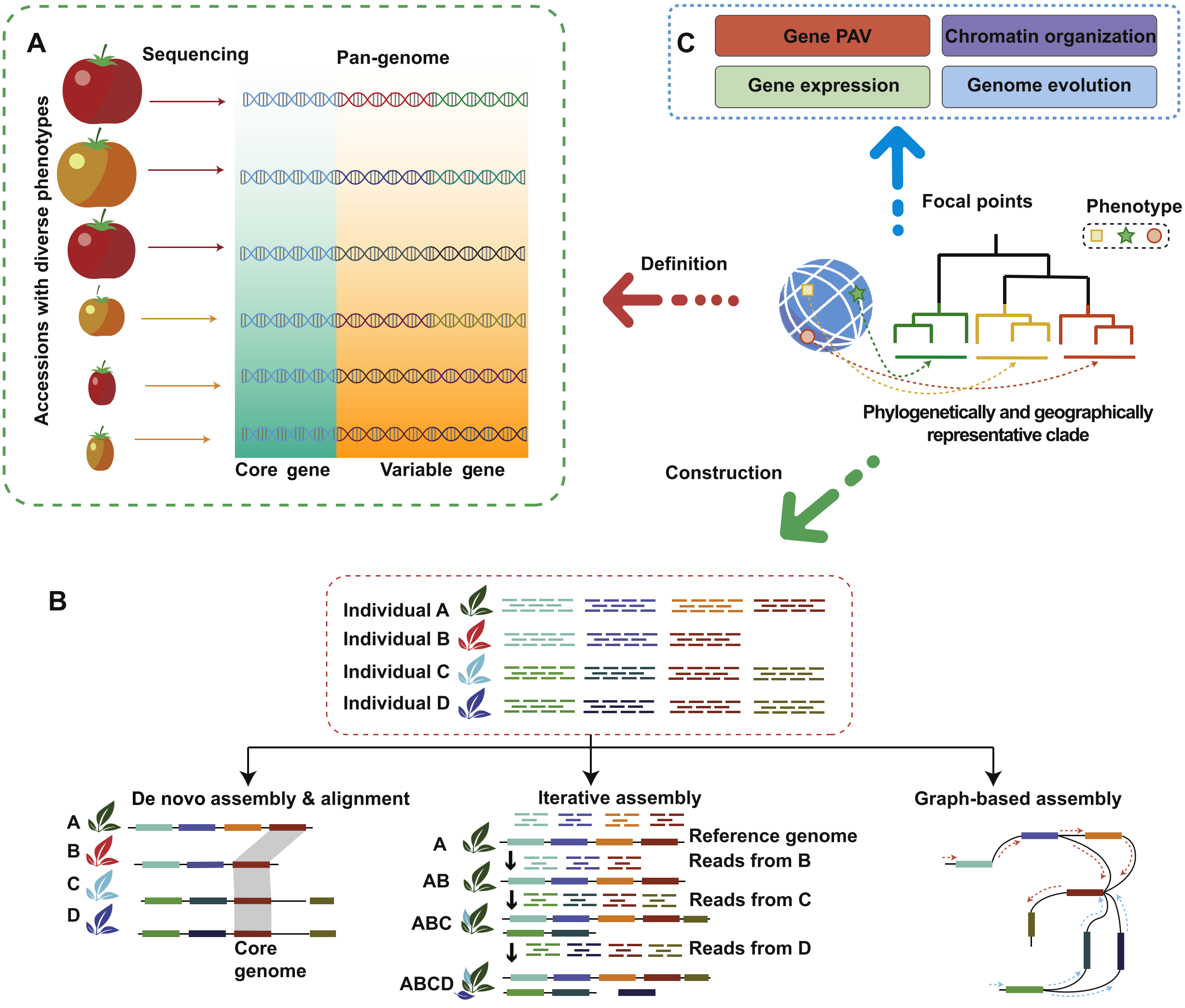

Fig. 1. Schematic diagram showing the concept, construction methods, and research focus of pan-genomics. **A**: Definition and components of a representative pan-genome. The phylogenetically or geographically representative accessions are selected to construct the pan-genome. The pan-genome can be broken down into a "core gene" that contains genes present in all individuals, and a "variable gene" that contains genes present in a subset of strains. **B**: Pan-genome assembly approaches, including de novo assembly, iterative assembly, and graph-based assembly. The sequencing reads from different samples are used for the assembly of pan-genome. Reads from shared genomic segments are indicated with the same colors. For de novo assembly, the core regions are highlighted using grey color. For iterative approach, individual A's genome is used as a reference and then the sequencing reads from other individuals are sequentially mapped to the reference genome. The unmapped are then assembled to construct non-redundant pan-genome. A pan-genome is constructed from graph assembly and the relationships between each node can be traced by following the paths of the graph. **C**: Focus of current pan-genome studies.

strains (Tettelin et al., 2005) (Fig. 1A). The definitions and objectives of pan-genome were then modified and developed since it was proposed (Rasko et al., 2008; Snipen et al., 2009; Alcaraz et al., 2010; Plissonneau et al., 2018), and the pan-genome can be either sequence-based or gene-based (Golicz et al., 2020). Compared with the gene-based pan-genome, a sequence-based pan-genome could captures genic as well as nongenic sequences, such as TEs and noncoding RNAs (ncRNAs), which play fundamental roles in the structural organization and function in plant genomes (Tahir Ul Qamar et al., 2020). Until now, pan-genomics studies in plants, including rice, soybean, maize, wheat, cucumber, chickpea, and tomato, have focused mainly on crop breeding, adaptation, and evolution (Hirsch et al., 2014; Li et al., 2014; Schatz et al., 2014; Montenegro et al., 2017; Wang et al., 2018; Zhao et al., 2018; Gao et al., 2019; Alonge et al., 2020; Liu et al., 2020; Walkowiak et al., 2020; Hufford et al., 2021; Li et al., 2022a). In this review, we briefly summarize the recent major achievements and pipelines for pan-genome research, and discuss potential challenges and perspectives for future pan-genomics studies to provide a basis for applications related to crop improvement.

Approaches for pan-genome construction

Recent advances in sequencing technologies have enabled the assembly of high-quality reference genomes for a large number of plants concurrently. However, how to integrate multiple genomes from a subset of accessions and make the integrated genetic information easily accessible to biologists remain challenging (Li et al., 2020). Methods that have been used to construct pan-genomes include de novo assembly, iterative assembly, and graph-based assembly (Li et al., 2014; Schatz et al., 2014; Golicz et al., 2016a; Danilevicz et al., 2020; Liu and Tian, 2020; Qin et al., 2021) (Fig. 1B).

The most straightforward way to construct a pan-genome is by de novo assembly of the genomes of multiple samples, followed by comparative analyses to detect all variant types and characterize the identified genes as core or dispensable (Mahmoud et al., 2019). The progress in long-read sequencing technologies and complementary approaches such as the construction of Hi-C and BioNano maps have made it feasible to obtain high-quality plant genomes at the chromosome level, including telomere-to-telomere genome assemblies (Miga et al., 2020). The de novo assembly strategy copes well with repeat regions, but it requires a high depth of sequencing reads to build highly contiguous and accurate genome assemblies, which is costly for large plant genomes and hundreds of reference genomes for one species (Hurgobin and Edwards, 2017).

Unlike the de novo assembly strategy, the iterative assembly strategy starts with the construction of a single reference genome, and then the reads from other samples are sequentially mapped to the reference genome. Unmapped reads are assembled and added to the reference genome to construct a pan-genome of non-redundant sequences (Golicz et al., 2016b). This method costs less than the de novo assembly method because each sample can be sequenced with low sequencing depth, which allows the pooling of hundreds of samples. However, because there is no assembly process, the iterative assembly method struggles to handle genomes that contain a large number of repeat regions and it cannot detect large SVs that are not spanned by single short reads (Jiao and Schneeberger, 2017).

The graph-based assembly strategy for pan-genome construction uses a graph to represent diversity and variations relative to a reference genome. The compacted de Bruijn graph is the one most commonly used to integrate genetic information from different accessions of one species (Chikhi et al., 2015, 2016; Li et al., 2020). The bi-directed variation graph has been used to integrate genetic variations across a population and label their possible locations on a reference genome. Graph-based pan-genomes show significant improvements in mitigating reference bias compared with traditional linear genomes (Garrison et al., 2018). Currently, the construction and application of graph-based pan-genomes are limited by the complexity of plant genomes, such as the high repeat content and polyploidy, and the lack of tools for common downstream analyses and visualization of the graph. However, graph-based genomes have been shown to have immense advantages over other methods, implying that graph-based assembly strategies may have extensive applications and promising prospects in the future.

Major achievements in plant pan-genomics

Pan-genomes for major crops, such as maize, rice, wheat, and soybean, have been constructed based on high-quality genomes of multiple samples, which has led to great progress in studies into the evolution of plant genomes and the identification of key genes associated with important agronomic traits (Zhao et al., 2018; Liu et al., 2020; Hufford et al., 2021; Qin et al., 2021) (Figs. 1C and 2). These studies have shown that the construction of a pan-genome can eliminate deviations from a single reference genome as much as possible and can present a nearly full view of the diversity within a species (Khan et al., 2020).

Grain species

Major grain species include rice, wheat, maize, soybean, millet, barley, oats, and sorghum, which are indispensable sources of energy in the human diet (Bansal et al., 2016). Because of their importance, a major focus of plant pan-genome research has been to obtain a full view of the genetic variations within each of these grain species. The first plant pan-genome based on high-quality genomes was released in 2014 for wild soybean, which provided a potentially rich resource for improving the genetic diversity of cultivated soybean that was lost during domestication (Li et al., 2014). In their study, seven phylogenetically and geographically representative accessions of wild soybean were de novo assembled and abundant variations associated with agronomic traits were identified, including biotic resistance, seed composition, flowering, and maturity time. This study confirmed for the first time that a single genome did not adequately represent the diversity within a species, and showed that a large number of SVs and genes associated with important agronomic traits were lost in the domestication process. The first graph-based pan-genome was constructed based on the high-quality genomes of 26 representative wild and cultivated soybean samples selected from 2898 deeply sequenced accessions (Liu et al., 2020). Numerous genetic variations that could not be detected by short read mapping were identified, and the influence of these variations on genome evolution, key agronomic traits, and generation of new genes was explored (Liu et al., 2020). This study broke through the storage form of the traditional linear genome and the obtained graph-based pan-genome provided an almost full view of the genetic variations within soybean. This work was a milestone in plant genomics. A more recently published pan-genome of 204 representative cultivated soybean was constructed based on a cladogram of 1007 soybean accessions from the GmHapMap data set, and 108 Mb of novel sequences that contained 3621 protein-coding genes that were absent from the reference genome, were detected (Torkamaneh et al., 2021). Although this pan-genome may be unrepresentative and incomplete to some extent, especially for large SVs, because of the de novo assembly of short reads, it still provided a relatively comprehensive gene pool in cultivated soybean.

A rice pan-genome dataset of the *Oryza sativa–O. rufipogon* species complex was constructed by de novo assembly of 66 divergent accessions, and 23 million sequence variants were

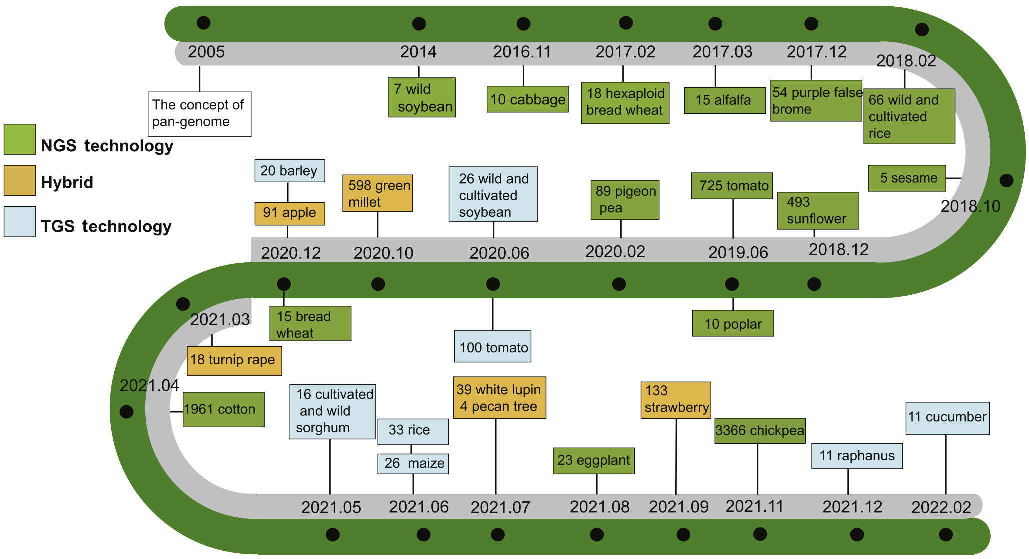

Fig. 2. Timeline and basic information for the released plant pan-genomes. The different sequencing technologies used to construct the pan-genomes are indicated using different colors. Solid black circles indicate past events in plant pan-genomics. The technologies are indicated using colored rectangular boxes: light green, next-generation sequencing; dark orange, hybrid sequencing; light blue, long-read sequencing. The sample size and species are indicated in the colored rectangular boxes.

identified by inter-genomic comparisons with the Nipponbare reference genome (Zhao et al., 2018). Although this pan-genome was constructed based on short reads rather than high-quality genomes, it was still possible to trace the evolution history of important quantitative trait loci (QTLs) of different rice accessions, including traits associated with flowering time, cold tolerance, grain weight, tiller angle, and plant height. The most complete rice graph-based pan-genome constructed so far is based on the high-quality genomes of 33 genetically diverse rice accessions (Qin et al., 2021). This is a classic study that not only comprehensively detected genomic variations and their formation mechanisms, but also systematically inferred their impacts on genome evolution, gene expression, crop domestication, and adaptability to environment for the first time. A pan-genome of maize was constructed using 26 inbred lines, and 103,033 pan-genes and 791,101 SVs were identified (Hufford et al., 2021). This was the first high-quality pan-genome of maize and the first to identify SVs associated with the DNA methylation rate, which may contribute to phenotypic variation.

Genomic studies of wheat and barley, two of the most important crop species worldwide, have been hindered by their large and highly-repetitive genomes. The bread wheat cv. Chinese Spring genome assembly was updated and the genetic diversity among 18 wheat cultivars was explored (Montenegro et al., 2017). Then, the pan-genome of wheat was constructed by iterative assembly and 350 Mb of newly assembled sequence was added to the reference genome; the variable genes were enriched mainly in the response to environmental stress and defense. Although this pan-genome was constructed based on short reads and may be incomplete to some extent, it still provided the first variation map of a representative wheat species and important guidance for the construction of high-quality pan-genomes of plant species that have very large and complex genomes. The first wheat pan-genome based on high-quality genomes was constructed by integrating the genetic information of 15 representative wheat varieties from global wheat breeding resources to explore the genomic diversity of wheat. Comparative analysis identified extensive structural rearrangements, introgressions from wild relatives, and differences in gene content that resulted from complex breeding histories, which provided a basis for functional gene discovery and breeding (Walkowiak et al., 2020). This study creatively overcame the challenges of plant genome complexity and triggered the construction of high-quality pan-genomes in plants with very large and complex genomes. A comprehensive barely pan-genome was constructed using 20 varieties of cultivated and wild barley accessions, and gene PAVs were found to be frequently associated with resistance gene homologues (Jayakodi et al., 2020). The prevalence of a large inversion was also identified in current elite germplasm, which may have far-reaching implications for the use of barley germplasm resources, understanding of the molecular mechanisms underlying the formation of important agronomic traits, as well as for breeding high-quality, high-yielding, and stress-tolerant superior varieties.

Vegetable species

Vegetable crops are major nutrient sources in the human diet and have been cultivated for thousands of years. The pan-genomes of many agronomically important vegetable crops, including tomato, cucumber, eggplant, and rapeseed, have been released, which have provided a basis for future biological studies and breeding programs (Golicz et al., 2016b; Gao et al., 2019; Alonge et al., 2020; Song et al., 2020; Li et al., 2022a).

Modern crops have narrow genetic diversity because of domestication, so it is very important for breeders to obtain as much genetic variation information as possible for crop improvement. The first tomato pan-genome was constructed based on the re-sequencing data of 725 representative accessions using a map-to-pan strategy, and gene PAV analyses detected substantial gene loss and intense negative selection of genes related to disease resistance during the domestication and improvement process (Gao et al., 2019). They found that a rare allele in the *TomLoxC* promoter was selected during domestication and further analysis showed that *TomLoxC* played a key role in apocarotenoid production. Furthermore, long-read Oxford Nanopore sequencing captured 238,490 SVs in 100 diverse tomato lines and hundreds of SV gene pairs were found to exhibit subtle and significant changes in gene expression that could broadly influence variations in quantitative traits, such as fruit flavor, size, and production (Alonge et al., 2020). These findings systematically highlighted the underexplored roles of SVs in genotype-to-phenotype relationships and their widespread importance and utility in tomato improvement.

Eight oilseed rape lines were sequenced using PacBio sequencing technology and used to construct a pan-genome (Song et al., 2020). Millions of small variants and 77.2 Mb–149.6 Mb gene PAVs were identified and a genome-wide association study (GWAS) of the gene PAVs was performed to screen candidate genes for silique length, seed weight, and flowering time. This study not only generated the first allotetraploid pan-genome and resources to support a better understanding of the genome architecture and accelerate the genetic improvement of rapeseed, but also demonstrated that PAV-GWAS was significantly complementary to SNP-GWAS in identifying associations of gene PAVs to traits. The first cucumber pan-genome was constructed from the genome information of 12 representative accessions using PacBio sequencing technology and a graph-based assembly strategy (Li et al., 2022a). Large segmental inversions were detected in some wild accessions and the graph-based pan-genome was used in a GWAS analysis of female flowering rate on a primary branch, fruit spine/wart density, and branch number. This study clarified the cucumber karyotype evolution in the domestication process and identified a number of potentially important genes related to agronomic traits, which provided a basis for mining key genes, breeding, and improvement of cucumber.

Other plant species

The large number of pan-genome studies of grain and vegetable crops has enabled gene PAVs to be tracked in domestication and breeding processes, and the potentially rich resources that were obtained have been used to improve the genetic diversity that was lost because of these processes. Pan-genomes of other species such as fruits and species that are closely to crops have also been constructed and analyzed to find agronomic traits that are affected by SVs. The pan-genome of the model grass *Brachypodium distachyon*, which was constructed by de novo assembly of 54 inbred lines, was found to contain nearly twice the number of genes found in any individual genome. The core genes were enriched mainly in essential cellular processes, whereas the shell and softcore genes were enriched in functions that may be advantageous in specific environments (Gordon et al., 2017). This study demonstrated that gene PAVs contributed substantially to phenotypic variation and that transposable elements played key roles in genome evolution, which is consistent with results for other plant pan-genomes.

Cotton is an important economic crop that is cultivated worldwide, and breeding plants with high-quality fiber that are high yielding and disease resistant has long been a goal. The first cotton pan-genome was constructed based on the re-sequencing data of 1961 cotton accessions using a map-to-pan strategy, and this variation repertoire indicated that genomic divergence during cotton domestication and improvement had informed the characterization of favorable gene alleles for improved breeding practice (Li et al., 2021). This cotton pan-genome contained the most abundant variations of cotton so far and provided the genomic basis of cotton domestication and new ideas for the precise improvement of important cotton traits. To characterize the genetic diversity in sunflower and to quantify the contributions from wild relatives, 287 cultivated lines, 17 Native American landraces, and 189 wild accessions were used to constructed a pan-genome of cultivated lines using a map-to-pan strategy (Hübner et al., 2019). The results indicated that approximately 10% of the cultivated sunflower pan-genome was derived by introgression of regions from wild sunflower species, and most of these regions contained genes related to biotic resistance. Importantly, this study showed that the comparatively extensive wild genetic resources made it possible to comprehensively analyze the introgression regions in the pan-genome of cultivated sunflower and their impacts on important agronomic traits.

Strawberry (*Fragaria* spp.) is a model system for fundamental and applied research because it has remarkable nutritional composition, different mating systems, and complicated ploidy variations (Johnson et al., 2014; Qiao et al., 2021). The pan-genome of five diploid *Fragaria* species was constructed and 128 individuals spanning 10 diploid species were resequenced. The subsequent analysis showed the genetic diversity, demographic history, and natural selection of strawberry species, and multiple independent single base mutations were detected in the *MYB10* gene that were associated with white pigmented fruit. These findings provided new insights into the evolution and resource utilization of strawberry, including the first pan-genome, phylogeny and genetic differentiation, the evolutionary dynamics of important gene families, and large-scale genome resources for further research. The haplotype-resolved genomes of cultivated apple and its two major wild progenitors (*Malus sieversii* and *M. sylvestris*) were assembled and the pan-genome was constructed and analyzed by combining them with 91 varieties that had been deeply resequenced. Thousands of new genes were discovered and hundreds of them were selected from one of the progenitors and largely fixed in cultivated apples, showing that introgression of new genes/alleles was a hallmark of apple domestication through hybridization (Sun et al., 2020). The major breakthroughs in this study were the high-quality genome assembly of highly heterozygous species and the first haplotype-resolved apple pan-genome, which indicated the potential of species domestication and improvement based on haplotype-resolved pan-genomes.

The availability of multiple reference genome within a species has provided unprecedented opportunities to identify SVs in a non-reference-biased manner. Many crop species are characterized with large and complex genomes, which make the genome assembly cost-prohibitive. Currently, the pan-genome with numerous samples were mostly constructed using "map-to-genome" method with short reads, such as cotton (Li et al., 2021), chickpea (Varshney et al., 2021), which enable the construction of the pan-genome with a relatively lower sequencing depth and cost. However, the identification of SVs has been troubled by the highly repetitive nature of crop genomes as the short reads are inefficient and unreliable in these regions (Della Coletta et al., 2021). The recently published pan-genomes constructed from new approaches have the benefits of providing the physical position of genes and other genomic features, which provide the most comprehensive characterization of SV to date (Alonge et al., 2020; Qin et al., 2021; Li et al., 2022a, 2022b), although they were relatively costly.

Current plant pan-genome studies are mainly focused on the knowledge of TEs and SVs (Morgante et al., 2007). The exhaustiveness of SV discovery is largely affected by the joint effect of sample size and sequencing depth (Torkamaneh et al., 2021). As the sample size increases, the percentage of the pan-genome increases and the percentage of core genes decreases. A rice pan-genome constructed from three accessions revealed that about 92.17% of all genes are core genes, and only 7.83% genes are dispensable genes. A pan-genome constructed from 3010 diverse Asian cultivars showed a higher percentage of variable genome (~41%) compared with previous study based on three rice accessions (~8%), indicating that the sample size will reach a saturation point where any further increase would not lead to a further expansion of the pan-genome size. Besides, the genetic properties of the selected sample, such as genome size, mode of reproduction, bottlenecks during domestications, and ploidy level, may also influence the efficiency and completeness of a pan-genome study (Tao et al., 2019). The core genes tend to have lower SNP density and/or indel density compared with dispensable gene as observed in *B. distachyon* (Gordon et al., 2017), rice (Wang et al., 2018), and soybean (Li et al., 2014), and they are functionally involved in basic cellular functions. The dispensable genes tend to have higher SNP density, and are likely involved in environment adaptability, organ size, flowering time, and gene regulation (Gordon et al., 2017).

Applications of plant pan-genomics

Pan-genomes for functional gene mapping

Genomics has been widely used for gene mapping in crop species; however, a single reference genome cannot fully represent the germplasm within a species. A pan-genome is a comprehensive representation of the genomic variation in a population or species that enables the identification of genomic regions associated with diverse traits of interest. The most common approaches used to identify genetic variations associated with a desired phenotype are QTL analysis and GWAS. QTL analysis can link phenotypes to molecular markers such as SNPs, simple sequence repeats, and restriction fragment length polymorphisms (Myles and Wayne, 2008), whereas GWAS is used to detect associations between small genetic variants and traits in a population. Both of these approaches are influenced by reference bias (Gage et al., 2019). For example, the *Xa21* gene that confers resistance to *Xanthomonas oryzae* pv. *oryzae* race 6 is absent in some cultivated rice accessions, which poses a great challenge for cloning using a single reference genome (Song et al., 1995). The availability of gene PAV information in a pan-genome complemented with data on small SVs and SNPs can overcome the problem of reference bias and facilitate marker development during the mapping process, even for large-scale SVs (Tao et al., 2019). GWASs of large SVs identified by pan-genome analysis can not only map the regions and genes related to important agronomic traits, but can also provide information about what kind of SVs are associated with the different phenotypes.

Pan-genomes for domestication and evolution studies

Understanding the processes that facilitate the origin of phenotypic diversity in plants and the domestication of important crops has long been a goal of researchers and breeders. Plant genomes are highly dynamic and vastly divergent, so a phylogenomic framework is needed to clarify the relationships between species or accessions. Pan-genomes provide an unprecedented opportunity to investigate the origin, evolution, domestication, and gene flow in plants by obtaining the genetic variations among cultivated and wild accessions (Qiao et al., 2021). Whole genomic comparisons between cultivated and wild accessions or interspecies comparisons can characterize the core and dispensable regions of genomes, which provides valuable information for gene evolutionary studies (Krasileva, 2019). Pan-genomes have been used to study the evolutionary history of genes associated with important agronomic traits in crop species, such as the identification of mutations and evolution of *MYB10* related to white fruit in wild strawberries (Qiao et al., 2021). The seed coat color of soybeans is distinct in wild and cultivated accessions; nearly all wild soybeans have black seed coats and most cultivars have yellow seed coats (Zhou et al., 2015). SVs in the genomic regions related to seed coat color were identified based on the pan-genome that was constructed based on 29 diverse accessions of cultivated and wild soybeans, and the origin and evolution history of these variants were reconstructed by phylogenetic analysis and genetic distance estimation (Liu et al., 2020). Pan-genome analysis of 33 genetically diverse rice accessions also greatly promoted rice evolution and domestication studies. For example, independent deletions in the *OsWAK112d* gene, a known negative regulator of blast resistance, in some indica and japonica genomes, suggested that these deletions may have increased rice blast resistance in the affected plants (Qin et al., 2021).

Pan-genomes for genome evolution studies

The well-established plant pan-genomes have provided a deeper understanding of the molecular mechanisms that underlie genome evolution within a species, such as the origin of SVs and the impacts of SVs on expression patterns and chromatin organization. Transposable element activity, polyploidy, and outcrossing were shown to be major driving forces for the generation of SVs (Panchy et al., 2016; Dunning et al., 2019; Zhang et al., 2019). Pan-genomes can provide a full view of the mechanisms of SV formation, which can help explain genome evolution and the complex architecture of phenotypic traits of agricultural relevance. In rice, tomato, and soybean, SVs have been shown to influence the expression of nearby genes by changing gene sequences or by affecting regulatory sequences (Alonge et al., 2020; Liu et al., 2020; Qin et al., 2021). Plant pan-genomes have also been used to explore the influence of SVs on chromatin three-dimensional organization. In both diploid and tetraploid cottons, many SVs were found in topologically associating domain boundary regions and had a large effect on disrupting TAD organization, and the SVs together with TAD disruption led to expression differences of orthologous genes (Long et al., 2021).

Polyploidy is common in angiosperm plants and proved to be tremendous source of raw material for gene genesis (Jaillon et al., 2007; Huang et al., 2013; Salman-Minkov et al., 2016). A soybean pan-genome showed that the nucleotide diversity in the WGD regions was significantly lower than that in the non-WGD regions. Compared with non-WGD regions, the WGD regions contained a more core and softcore genes and less SVs, indicating that genome duplication may be an important genetic force to shape the evolution of SVs (Liu et al., 2020). A significant higher ratio of core genes (average ~45.27%) were generated from a WGD in the sesame pan-genome, and only ~10.22% dispensable genes are influenced by WGD event (Yu et al., 2019).

Pan-genomes for genotype database construction

A comprehensive functional genomic platform that integrates the variants obtained by pan-genome analysis, the diverse phenotypes among different accessions, and other multi-omics data will provide hugely valuable resources for genetic studies and crop breeding. Genotype databases have been used to search and visualize variation types of interest in the pan-genome context. Several databases that contain large amounts of genetic and phenotypic information have been constructed to facilitate ready access to pan-genome resources in many important crops. The Molecular Breeding Knowledgebase of rice contains two reference genomes and other multi-omics data, including nearly 7000 global rice resequencing data sets, more than 4 million phenotypes, and 13,000 functional annotations of known genes, which has allowed complex functions such as germplasm screening based on genotype, individual comparisons, mutation analysis, and online annotation of genotypes to be conducted (Peng et al., 2020). ZEAMAP, a comprehensive database of maize genera, contains integrated multi-omics data, including transcriptomes, phenotypic groups, and genetic variations, which has greatly promoted the understanding of the relationship between phenotypes and genotypes in maize (Gui et al., 2020). A comprehensive functional genomic platform for 1689 rapeseed accessions that contains genome sequences, phylogenetic relationships, gene PAV information, and common multi-omics tools was established to integrate quick searches and visualization of the pan-genome data (Song et al., 2021). This platform has provided resources that form a solid foundation for genetic breeding and improvement of rapeseed.

Pan-genomes for molecular breeding in crops

Natural and occasional variants can be identified by analyzing high-quality pan-genomes, and many variants associated with important agronomic traits, such as abiotic and biotic stress resistance, flower time, fruit flavor, and production, have been detected (Yu et al., 2014; Golicz et al., 2016b; Gao et al., 2019; Tao et al., 2019). Many crop species have lost substantial genetic diversity through successive bottlenecks during domestication and selection (Li et al., 2014; Gao et al., 2019; Alonge et al., 2020). Pan-genomes enable comparisons between crop species and their wild relatives, which can help to identify genes that have been lost under intensive human cultivation. Enhanced understanding of the dispensable genome is of great importance to select suitable materials for a breeding crops (Tao et al., 2019). The CRISPR-Cas9 technology has been widely used in transformable plants to characterize gene function and improve traits; however, the editing efficiency was influenced by the genotype and target site selected (Yu et al., 2017). The availability of high-quality pan-genomes, along with phenotypic information, can help to identify variant alleles and delimit CRISPR-Cas9 target sites, which can improve the editing efficiency (Tay

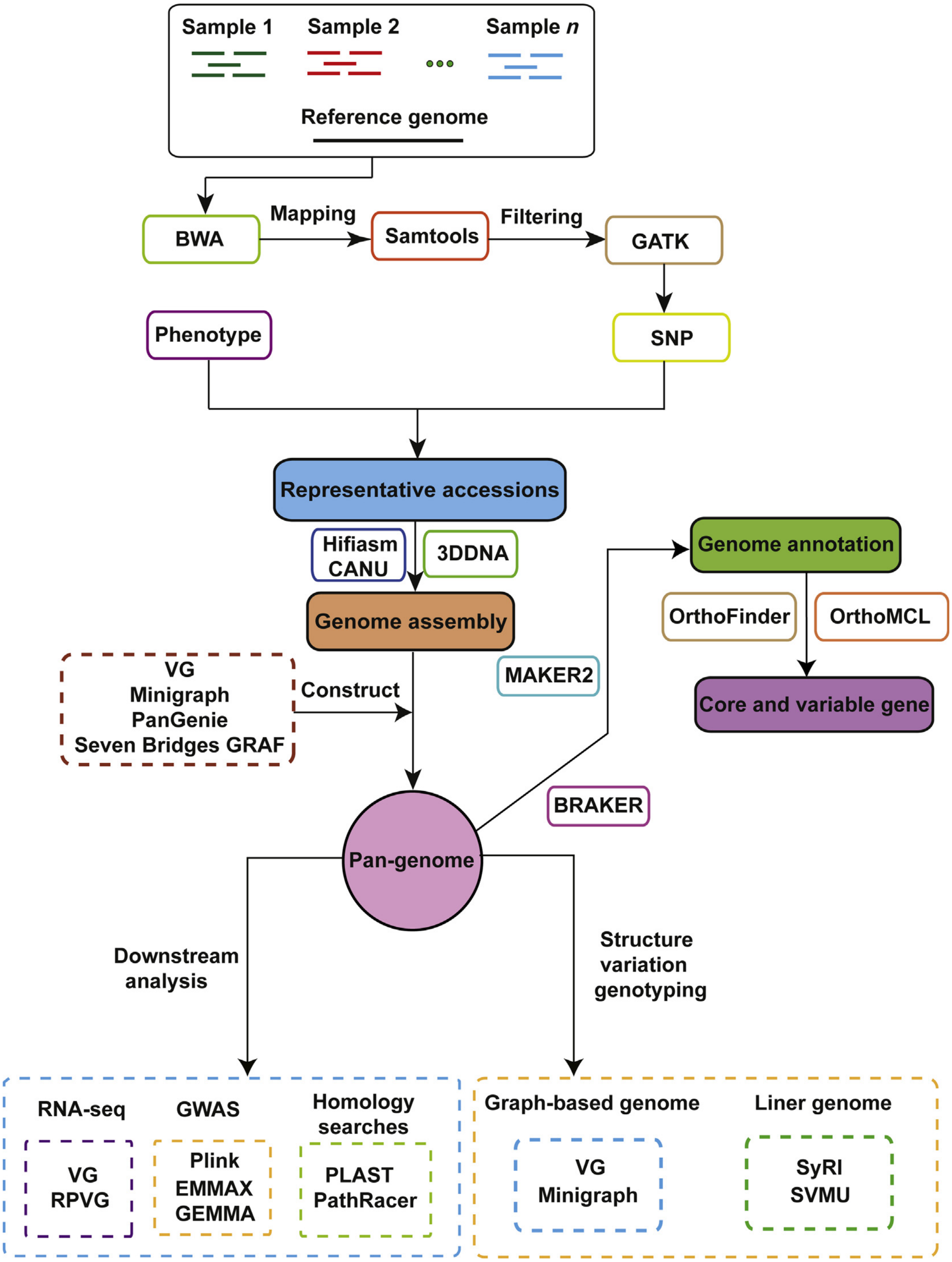

Fig. 3. The workflow for current pan-genome studies.

Fernandez et al., 2022). Hence, pan-genome analysis focused mainly on crop diversity and improvement will facilitate crop breeding programs.

Exploring the genetic basis that underlies agronomical traits is critical for crop improvement and genomics-based breeding methods have been applied successfully to many crop species. Access to important genetic resources by pan-genome research and their correlation with phenotypes are crucial for crop improvement. A genome design approach, which applied genome analyses to the creation of F_1 hybrids, was used to develop pure and fertile potato lines (Zhang et al., 2021a). Self-compatible diploid clones with low heterozygosity and few deleterious mutations were selected as the starting materials. Segregation distortion regions and important genetic loci were identified by population genetics analysis, and then highly homozygous inbred lines were developed by continuous selfing and genome-assisted selection. Finally, the genomes of the inbred lines were sequenced, and those with high genome complementarity were crossed and the resultant F_1 hybrids were evaluated for performance. The de novo domestication of wild allotetraploid rice also provided new insights into the design of ideal crop species by establishing an efficient transformation and genome editing system based on a high-quality reference genome (Yu et al., 2021).

Plant pan-genome analysis tools

Here, we briefly summarize the available tools and pipelines used for pan-genome analyses, including the selection of representative accessions, genome assembly, genome annotation, SNP identification, orthologous group identification, phylogenetic construction, SV detection, construction of a pan-genome, and GWAS mapping using the pan-genome (Fig. 3; Table 1).

Selection of representative accessions

The first step for setting up a pan-genome is the selection of diverse individuals. Adequate and representative accessions that capture the genetic diversity within a species should be selected. The most important factors to consider are phenotypic diversity and phylogenetic relationship. Construction of a phylogenetic tree of the samples based on the identified SNPs can facilitate the selection process. The used software packages are summarized in Table 1.

High-quality genome assembly

Genome assembly is conducted to decipher the base composition that is fundamental for pan-genome analysis. Assembly methods for long-read sequences from PacBio platforms and complementary approaches for chromosome construction are listed here. Canu and Hifiasm can both be used to assemble HiFi data. The quality at the single-base level and completeness, including heterozygous regions, of genomes constructed by Canu are always much higher than those for genomes constructed by Hifiasm. However, the continuity of a Canu assembly is usually much smaller than that of a Hifiasm assembly, especially for highly heterozygous species, and the amount of computation required for Canu is much more than the computation required for Hifiasm. Besides, Hifiasm assemblies can sometimes lose large fragments and contain the chimeras of heterozygous sequences. The used software packages are summarized in Table 1.

Table 1
Overview of key tools and pipelines used for pan-genome analyses.

	Application	Software	Link
Selection of representative accessions	Short-read mapping	BWA	https://github.com/lh3/bwa
	Reads filtering	SAMTools	https://github.com/Blue-Matter/SAMtool
	Variant discovery	Genome Analysis Toolkit	https://github.com/broadinstitute/gatk
	Phylogenetic tree construction	FastTree	https://github.com/PavelTorgashov/FastTree
High-quality genome assembly	Contigs construction	CANU	https://github.com/marbl/canu
		Hifiasm	https://github.com/chhylp123/hifiasm
	Analyzing kilobase resolution Hi-C data	Juicer	https://github.com/aidenlab/juicer
	Chromosome construction	3DDNA	https://github.com/aidenlab/3d-dna
Genome annotation	Screening interspersed repeats	RepeatMasker	https://github.com/rmhubley/RepeatMasker
	Gene prediction	BRAKER	https://github.com/Gaius-Augustus/BRAKER
		MAKER2	https://www.yandell-lab.org/software/maker.html
	Functional annotation	InterProScan	https://github.com/biocorecrg/interproscan_docker
Identification of core and variable genes	Genes clustering	OrthoMCL	https://github.com/stajichlab/OrthoMCL
		OrthoFinder	https://github.com/davidemms/OrthoFinder
Construction of graph-based pan-genomes	Graph-based pan-genome construction	Variation graph (vg) toolkit	https://github.com/vgteam/vg
		Minigraph toolkit	https://github.com/lh3/minigraph
		PanGenie	https://github.com/eblerjana/pangenie
Structural variation genotyping	Structural variation genotyping for linear genome	SyRI	https://schneebergerlab.github.io/syri/pipeline.html
		SVMU	https://github.com/mahulchak/svmu
		NGMLR	https://github.com/philres/ngmlr
		Sniffles	https://github.com/fritzsedlazeck/Sniffles
	Structural variation genotyping for graph-based pan-genomes	Variation graph (vg) toolkit	https://github.com/vgteam/vg
		Minigraph toolkit	https://github.com/lh3/minigraph
		GraphTyper2	https://github.com/DecodeGenetics/graphtyper
Genome wide association studies		Plink	https://www.cog-genomics.org/plink/
		GEMMA	https://github.com/genetics-statistics/GEMMA
		EMMAX	http://genetics.cs.ucla.edu/emmax/
Homology searches		PLAST	https://mesihk.github.io/plast
		PathRacer	http://cab.spbu.ru/software/pathracer/
Mapping and quantification of RNA-seq data		Variation graph toolkit	https://github.com/vgteam/vg
		RPVG	https://github.com/jonassibbesen/rpvg

Genome annotation

Genome annotation methods are used to predict gene functions and structures. The software or pipeline for pan-genome annotation is still not available. The current strategy for pan-genome annotation is to perform independent annotation of individual genomes. The automatic annotation pipelines such as Maker2 (https://www.yandell-lab.org/software/maker.html) and Braker2 (https://github.com/Gaius-Augustus/BRAKER) are generally effective at detecting protein-coding regions. The BRAKER2 pipeline generates and integrates spliced alignments of homologous proteins, which are then used for the training and gene prediction by GeneMark-EP+ (https://github.com/gatech-genemark/GeneMark-EP-plus) and AUGUSTUS (https://bioinf.uni-greifswald.de/augustus/). MAKER identifies repeats, aligns ESTs and proteins to the reference genome, performs de novo gene predictions and automatically integrate the results into consensus gene set. Besides, users need to run MAKER for multiple rounds to improve annotation. The used software packages for genome annotations are summarized in Table 1.

Identification of core and variable genes

Core and variable genes can be identified by clustering the genes from the genomes of different samples. OrthoMCL (https://github.com/stajichlab/OrthoMCL) mainly looks for lineal homologous genes between relatively complete genomes, and firstly it creates databases and builds tables. Blast is used to perform all-vs-all comparison of the protein sequences, and MCL is used to cluster gene pairs into in-paralog groups and co-ortholog groups based on sequence similarity. Unlike OrthoMCL, OrthoFinder (https://github.com/davidemms/OrthoFinder) can use DIAMOND software (https://github.com/bbuchfink/diamond) to perform all-vs-all sequence alignment, which greatly improve the blast speed, and offers an option for fast tree building.

Construction of graph-based pan-genomes

Graph-based genomes have been used effectively to integrate genetic variations within a species. Several tools, including the variation graph toolkit (Garrison et al., 2018), minigraph toolkit (Li et al., 2020), Seven Bridges GRAF pipeline (Patron et al., 2019), and PanGenie (Ebler et al., 2020), have been developed for graph-based pan-genome construction. The variation graph toolkit provides data storage, interchange formats, alignment, genotyping, and variant calling methods and is the most widely used software for pan-genome construction with relatively integrated functions. VG uses paths to project graphics-related data into a reference-related coordinate space. Paths provide stable coordinates for graphs that are constructed in different ways from the same input sequence. The minigraph toolkit uses the reference Graphical Fragment Assembly format to model reference pan-genome graphs, which made the visualization and application of graph-based pan-genomes much more convenient than they are using the variation graph toolkit. Minigraph aligns the query sequence with the sequence graph and increases the existing graph with long query subsequences that diverge from the graph. For a graph consists of many short segments, minigraph will fail to map query sequences, and the alignment is slow for highly diverse species. Mapping readings to the reference genome will introduce reference bias and computational burden. The PanGenie short-read genotyper algorithm can efficiently leverage the increasing numbers of haplotype-resolved assemblies to unravel the functional impact of previously inaccessible variants and is faster than the variation graph and minigraph toolkits. The used software packages are summarized in Table 1.

Structural variation genotyping for linear and graph-based pan-genomes

The variety of repetitive sequences in genomes make it technically challenging and time-consuming to generate an accurate SVs set for pan-genome construction. Several methods have been developed to resolve this problem. SyRI (Synteny and Rearrangement Identifier) and SVMU (SVs from MUMmer) were designed to identify SVs based on comparisons between assembled genomes. SVMU can detect PAV and CNV using lastZ or MUMmer results as input files, but it has not been extensively tested on large genomes. SyRI mainly detects chromosomal variation based on MUMmer results, and it starts by identifying the longest collinear regions. It can identify all collinear regions and local variations within rearrangement regions, including SNP, INDEL, PAV, INV, TRANS, and so on. After many attempts and comparisons, we found that SyRI identified many more types of SVs than SVMU, and although smaller numbers of SVs were identified by SVMU, its accuracy was slightly higher than that of SyRI. The NGMLR long-read mapper and Sniffles caller were designed to identify SVs in long-read sequencing data, which may lose some genetic variations, especially for large SVs. The used software packages are summarized in Table 1.

Genome-wide association studies

The SVs detected in a graph-based pan-genome can be analyzed with standard linear genome tools, including Plink (Purcell et al., 2007), EMMAX (Kang et al., 2010), and GEMMA (Parker et al., 2016). Plink association analysis is mainly aimed at case/control analysis, including standard chi-square test, logistic regression, simple linear regression, Fisher's test and so on. Complex models cannot be realized, so animal and plant data can be filtered by this software, but association analysis is not recommended. It is generally used for human association analysis (Chang et al., 2015). EMMAX mainly implements EMMA model and genotype files are usually in tped/tfam format, and it is worth noticing all chromosome names are changed to numerical type. EMMAX can only run one trait at a time, which consumes less memory (Kang et al., 2010). GEMMA needs four main input files, including genotype data, phenotype data, correlation matrix, and covariate data, which incorporated LM, MLM, MLMM, and BSLMM models. The implemented multivariate linear-mixed model offers improved computation speed and power, which can deal with more than two phenotypes. A unique Bayesian sparse linear mixing model is also used for prediction, multi-marker modeling and phenotypic prediction (Zhou and Stephens, 2014). The used software packages are summarized in Table 1.

Homology searches

Algorithms for sequence to linear alignment have been available for a long time, such as BLAST, BLAT (Kent, 2002), and HMMER (Finn et al., 2011). PLAST (Schulz et al., 2021) and PathRacer (Shlemov and Korobeynikov, 2019) were designed to perform homology searches with graph-based pan-genomes. PLAST performs a local alignment search between a DNA query sequence and a graph-based pan-genome. A new heuristic method is used to find maximum scoring local alignments, which use the assembled genomes or reads as input. This method scales sublinearly in running time and memory usage with respect to the number of genomes used. PathRacer aligns a profile hidden Markov model (HMM) directly to the assembly graph. The most probable paths traversed through the whole assembly graph are inferred by this tool, it does not matter whether interested sequence is located within the single contig or scattered across several edges. The used software packages are summarized in Table 1.

Mapping and quantification of RNA sequencing (RNA-seq) data

A graph-based pan-genome can improve the accuracy of RNA-seq analysis and can represent splice junctions with little modification (Sibbesen et al., 2021). The vg rna tool and vg mpmap are used to perform spliced graph construction and RNA-seq mapping, respectively. And then RPVG is used to quantify the haplotype-specific transcript expression. The software packages for RNA-seq mapping and quantification are summarized in Table 1.

Challenges, prospects, and future directions

Challenges for the construction and application of plant pan-genomes

Along with the development of new sequencing technologies, the construction of plant pan-genomes is becoming increasingly attractive. However, many challenges remain to be addressed for the construction and application of plant pan-genomes. Until now, most plant pan-genomes have been built using a "map-to-genome" method with short reads. However, newly obtained sequences cannot be mapped to specific positions in the pan-genome, which greatly hinders the downstream analysis and application in breeding programs, such as gene positional cloning, and massive individual genetic information from the non-reference lines is missed, particularly for larger SVs (Liu and Tian, 2020). The large number of repeats in plant genomes, such as transposable elements, which are major drivers that influence genome evolution and crop phenotypes, are the main reason why plant genomes have historically been poorly assembled, resulting in highly fragmented and incomplete pan-genomes (Alonge et al., 2020; Liu et al., 2020). Long-read sequencing technologies have made high-quality reference genome assembly practicable and a number of high-quality plant pan-genomes have been published, including those for rice, soybean, rapeseed, tomato, and wheat; however, it is still costly to construct a pan-genome with hundreds of samples. Computational time is another major challenge for pan-genome construction. Pan-genomic data can be considered "Big Data" in volume, variety, velocity, and veracity, and finding a way to store datasets from dozens of samples, especially for species such as wheat with large genomes, is essential (Consortium, 2018). Until now, now pan-genome studies have focused mainly on important agronomic crops and efforts are still needed to include the understudied plant species.

The dispensable genome, mainly driven by SVs, is the key element that contributes the phenotypic variations between accessions (Xu et al., 2012; Mace et al., 2013; Tao et al., 2019). However, current pan-genome analyses have focused mainly on the identification of SVs and gene PAVs, and have largely ignored comprehensive functional methods that cannot be fully implemented by analyzing population re-sequencing data because of the incomplete genetic information, particularly for the large SVs (Gao et al., 2019; Qin et al., 2021; Li et al., 2022a). Currently, pan-genomes contain only basic information, such as the allele frequencies, newly discovered genes, and SVs. In a previous analysis of RNA-seq data for 29 diverse rice sample types from the R527 accession, including multiple tissues at different stages as well as tissues for plants grown under different abiotic stress conditions, we found that the SVs had broadly shaped gene expression profiles (Qin et al., 2021). However, whether these SVs affect the local three-dimensional chromatin conformation and epigenetic modifications, and how these SVs regulate gene expression have still not been systematically studied and remain largely unresolved. Therefore, more omics and phenotypic data are needed to explore the potential mechanisms and regular patterns of SVs.

Currently, the tools for pan-genome analysis still lag far behind the developments in sequencing technology. The number of samples used in pan-genomic studies, sequencing depth, strategies for constructing pan-genomes, sequence annotation methods, and definition of core and dispensable genes vary greatly (Li et al., 2014; Yu et al., 2019; Liu and Tian, 2020; Qin et al., 2021). Graph-based pan-genomes can store all types of genetic variations (Paten et al., 2017), but how to effectively integrate, visualize, and use the pan-genome data remains challenging. Previous studies have tried to address these problems (Garrison et al., 2018; Li et al., 2020; Rabbani et al., 2020), but there is still no standard format for storing pan-genome graphs. Furthermore, the lack of algorithms and methods for downstream analysis, including the annotation of a variable genome and multi-omics analysis, has become a serious obstacle for the development of plant pan-genome studies. There is clearly a need to develop tools for storage and downstream analysis, not only to realize the effective integration of huge amounts of genetic information but also to realize the efficient use of the information, so as to really provide a useful pan-genome.

Prospects and future directions for plant pan-genomes

The availability of pan-genomes offers substantial new knowledge and unprecedented resources for unlocking the full genetic potential of plant species. The immense advances in sequencing and computer technologies, such as multi-omics, artificial intelligence, and gene editing, have made feasible the integration and downstream application of the extensive variants within species (Fig. 4). Phenomics technologies, which are characterized by intelligence, and high-throughput and dynamic nondestructive measurement, have developed rapidly thanks to developments in remote sensing, robotics, imaging technologies, and artificial intelligence that have made it possible to detect multi-temporal and multi-scale phenotypes, thereby enabling the dynamic and accurate identification of phenotypes in the whole growth period of crops (Furbank and Tester, 2011; McCoy, 2011; Fiorani and Schurr, 2013). The phenotypes of 368 maize materials were tested continuously at multiple growth stages under normal watering and drought stress, and combined with GWAS analysis, a large number of candidate genes and QTLs related to drought stress were identified and a genotype-to-phenotype association network was constructed (Wu et al., 2021). Therefore, combining high-throughput phenotyping data and the variants detected in pan-genome studies is a novel and effective approach to dissect the genetic architecture of complex traits and clone genes associated with agronomic traits.

Multi-omics data generated from dozens of samples, including pan-genome, phenome, transcriptome, epigenetic modification, metabolome, and proteome data, even spatial and single-cell transcriptome data, enable a deeper understanding of how SVs in a pan-genome influence the complex architecture of diverse traits of agricultural relevance. Analyses of multi-omics data can provide information about the dynamics of genomic variants, gene expression and regulation, as well as substance synthesis and metabolism that can be used to study complex biological processes and regulatory networks holistically and systematically. For example, the metabolome is the complete set of metabolites in a cell, tissue, or organism, which is complementary to phenomics and has been widely used in plant species such as tomato, rice and wheat (Chen et al., 2016, 2020; Zhu et al., 2018). Untargeted metabolomic analyses were performed in 136 representative tea accessions from China, and different phylogenetic subgroups were found to contain different signature metabolites. In particular, the accessions of *Camellia sinensis* var. *assamica* were characterized by high accumulation of diverse classes of flavonoid compounds, which may influence the flavors of these accessions (Yu et al., 2020).

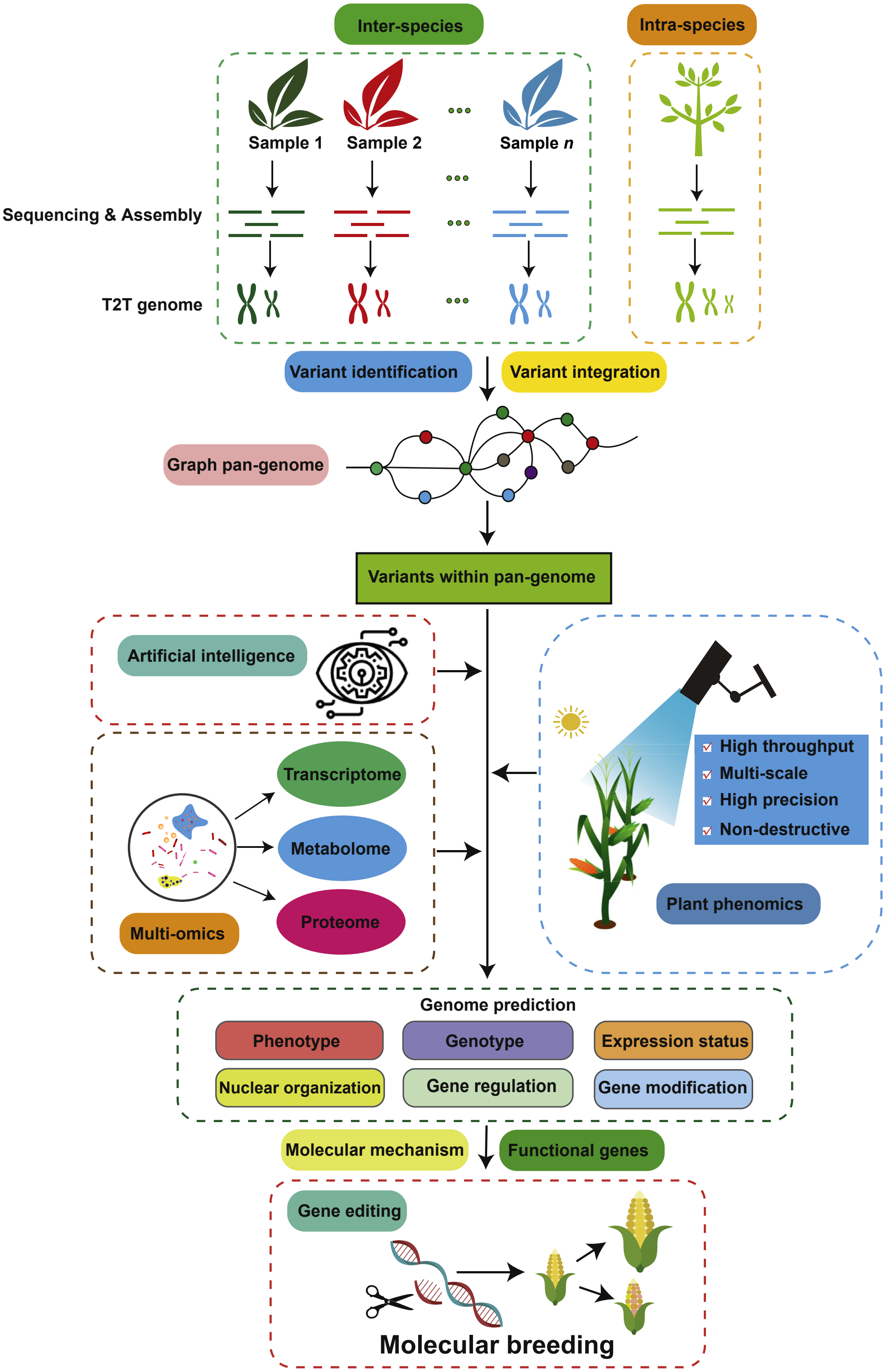

Fig. 4. Prospects and future directions for plant pan-genomics. The advances of multi-omics data, along with the high-throughput phenotyping data and artificial intelligence together, make it possible to perform precise genome prediction. It becomes possible for accurate prediction of genotype, expression status, phenotype, etc. By taking advantage of the well-established pan-genome and gene editing, we will be able to achieve the goal of genome design in crop species.

The immense technological advances, including gene editing and artificial intelligence, have made feasible the application of genetic resources in crop breeding. The variants identified from pan-genomes can support genome editing approaches that provide functional information on gene sequences and new target sites with increased efficiency (Tay Fernandez et al., 2022). The wheat susceptibility gene *MLO* was precisely manipulated using gene editing techniques, and the new germplasm showed broad spectrum resistance to powdery mildew, and had high yield and good quality. This finding showed that complex plant genomes can be edited and opened the prospect of genome editing in modern agricultural production (Li et al., 2022b). Artificial intelligence has also been widely applied in multi-omics research and a number of breakthroughs have been made. For example, AlphaFold2 is a novel machine learning approach that incorporates physical and biological knowledge about protein structure based on neural network-based model (Jumper et al., 2021).

Genomic prediction methods that use phenotypic data and the increasingly available genotypic data as the training set to construct a statistical model for predicting phenotypes can be applied to accelerate molecular plant breeding (Meuwissen et al., 2001; Wong and Bernardo, 2008; Hu et al., 2019; Keller et al., 2020). The composition and size of the training sets are critical for the accuracy of the predictions, and adequate amounts of different data from individuals are necessary (Isidro et al., 2015; Hu et al., 2019). Pan-genome and multi-omics data together with gene editing and artificial intelligence make precise genome prediction possible. Hybrid breeding is an efficient way to increase production in crop breeding, and genomic prediction methods have been shown to have competitive advantages because of their ability to predict and selecting superior hybrid descendants based on the genotypes of inbred parents (Xu et al., 2014). The successful development of a genomic prediction model will greatly improve the accuracy of breeding value prediction and dramatically reduce generation intervals (Desta and Ortiz, 2014). The intermediate data generated by the advanced high-throughput multi-omics platforms act as bridges between genotypes from pan-genomes and phenotypes from phenomics technologies (Hu et al., 2019). Artificial intelligence methods, including machine learning and deep learning, can be used to construct a genomic prediction model by integrating multi-dimensional phenotypic datasets and multi-omics data. An innovative genomic prediction framework will make it possible to accurately predict genotypes, expression status, phenotypes, and gene modification and regulation, as well as help to identify key functional genes. Once key functional genes are identified in plant species, gene editing technologies can be used to enhance target traits such as yield, quality, tolerance to biotic and abiotic stresses, and nutritional value, which will greatly accelerate the molecular breeding of various plants. The effective integration of the information generated by pan-genome analyses will not only improve the prediction accuracy of functional genes related to important agronomic traits, but also help in the construction of genotype-to-phenotype association networks. By taking advantage of the well-established pan-genomes, a series of agronomically important genes could be edited, which will help achieve the goal of genome design in crop species.

Conflict of interest

The authors declared that they have no conflicts of interest to this work.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (32100500), the Natural Science Foundation of Hebei Province (C2021201048) and Interdisciplinary Research Program of Natural Science of Hebei University. We thank Margaret Biswas, PhD, from Liwen Bianji (Edanz) (www.liwenbianji.cn/) for editing the English text of a draft of this manuscript.

References

Alcaraz, L.D., Moreno-Hagelsieb, G., Eguiarte, L.E., Souza, V., Herrera-Estrella, L., Olmedo, G., 2010. Understanding the evolutionary relationships and major traits of Bacillus through comparative genomics. BMC Genom. 11, 332.

Alonge, M., Wang, X., Benoit, M., Soyk, S., Pereira, L., Zhang, L., Suresh, H., Ramakrishnan, S., Maumus, F., Ciren, D., 2020. Major impacts of widespread structural variation on gene expression and crop improvement in tomato. Cell 182, 145–161.

Ambrožová, K., Mandáková, T., Bureš, P., Neumann, P., Leitch, I.J., Koblížková, A., Macas, J., Lysak, M.A., 2011. Diverse retrotransposon families and an AT-rich satellite DNA revealed in giant genomes of Fritillaria lilies. Ann. Bot. 107, 255–268.

Athiyannan, N., Abrouk, M., Boshoff, W.H., Cauet, S., Rodde, N., Kudrna, D., Mohammed, N., Bettgenhaeuser, J., Botha, K.S., Derman, S.S., 2022. Long-read genome sequencing of bread wheat facilitates disease resistance gene cloning. Nat. Genet. 54, 227–231.

Bansal, M., Sharma, M., Kanwar, P., Goyal, A., 2016. Recent advances in proteomics of cereals. Biotechnol. Genet. Eng. Rev. 32, 1–17.

Casacuberta, J.M., Jackson, S., Panaud, O., Purugganan, M., Wendel, J., 2016. Evolution of plant phenotypes, from genomes to traits. Genetics 6, 775–778.

Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., Lee, J.J., 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience 4, s13742, 015.

Chen, W., Wang, W., Peng, M., Gong, L., Gao, Y., Wan, J., Wang, S., Shi, L., Zhou, B., Li, Z., 2016. Comparative and parallel genome-wide association studies for metabolic and agronomic traits in cereals. Nat. Commun. 7, 1–10.

Chen, H., Zeng, Y., Yang, Y., Huang, L., Tang, B., Zhang, H., Hao, F., Liu, W., Li, Y., Liu, Y., 2020. Allele-aware chromosome-level genome assembly and efficient transgene-free genome editing for the autotetraploid cultivated alfalfa. Nat. Commun. 11, 1–11.

Cheng, H., Concepcion, G.T., Feng, X., Zhang, H., Li, H., 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat. Methods 18, 170–175.

Chikhi, R., Limasset, A., Jackman, S., Simpson, J.T., Medvedev, P., 2015. On the representation of de Bruijn graphs. J. Comput. Biol. 22, 336–352.

Chikhi, R., Limasset, A., Medvedev, P., 2016. Compacting de Bruijn graphs from sequencing data quickly and in low memory. Bioinformatics 32, i201–i208.

Danilevicz, M.F., Fernandez, C.G.T., Marsh, J.I., Bayer, P.E., Edwards, D., 2020. Plant pangenomics: approaches, applications and advancements. Curr. Opin. Plant Biol. 54, 18–25.

Della Coletta, R., Qiu, Y., Ou, S., Hufford, M.B., Hirsch, C.N., 2021. How the pan-genome is changing crop genomics and improvement. Genome Biol. 22, 1–19.

Desta, Z.A., Ortiz, R., 2014. Genomic selection: genome-wide prediction in plant improvement. Trends Plant Sci. 19, 592–601.

Dunning, L.T., Olofsson, J.K., Parisod, C., Choudhury, R.R., Moreno-Villena, J.J., Yang, Y., Dionora, J., Quick, W.P., Park, M., Bennetzen, J.L., 2019. Lateral transfers of large DNA fragments spread functional genes among grasses. Proc. Natl. Acad. Sci. U. S. A. 116, 4416–4425.

Ebler, J., Clarke, W.E., Rausch, T., Audano, P.A., Houwaart, T., Korbel, J., Eichler, E.E., Zody, M.C., Dilthey, A.T., Marschall, T., 2020. Pangenome-based genome inference. BioRxiv. https://doi.org/10.1101/2020.11.11.378133.

Finn, R.D., Clements, J., Eddy, S.R., 2011. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 39, W29–W37.

Fiorani, F., Schurr, U., 2013. Future scenarios for plant phenotyping. Annu. Rev. Plant Biol. 64, 267–291.

Fleischmann, A., Michael, T.P., Rivadavia, F., Sousa, A., Wang, W., Temsch, E.M., Greilhuber, J., Müller, K.F., Heubl, G., 2014. Evolution of genome size and chromosome number in the carnivorous plant genus *Genlisea* (Lentibulariaceae), with a new estimate of the minimum genome size in angiosperms. Ann. Bot. 114, 1651–1663.

Food and Agriculture Organization of the United Nations,1995. Staple foods: What do people eat? https://www.fao.org/3/u8480e/u8480e07.htm.

Furbank, R.T., Tester, M., 2011. Phenomics-technologies to relieve the phenotyping bottleneck. Trends Plant Sci. 16, 635–644.

Gabur, I., Chawla, H.S., Snowdon, R.J., Parkin, I.A.P., 2019. Connecting genome structural variation with complex traits in crop plants. Theor. Appl. Genet. 132, 733–750.

Gage, J.L., Vaillancourt, B., Hamilton, J.P., Manrique-Carpintero, N.C., Gustafson, T.J., Barry, K., Lipzen, A., Tracy, W.F., Mikel, M.A., Kaeppler, S.M., 2019. Multiple maize reference genomes impact the identification of variants by genome-wide association study in a diverse inbred panel. Plant Genome 12. https://doi.org/10.3835/plantgenome2018.09.0069.

Gao, L., Gonda, I., Sun, H., Ma, Q., Bao, K., Tieman, D.M., Burzynski-Chang, E.A., Fish, T.L., Stromberg, K.A., Sacks, G.L., 2019. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. Nat. Genet. 51, 1044–1051.

Garrison, E., Sirén, J., Novak, A.M., Hickey, G., Eizenga, J.M., Dawson, E.T., Jones, W., Garg, S., Markello, C., Lin, M.F., 2018. Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nat. Biotechnol. 36, 875–879.

Golicz, A.A., Batley, J., Edwards, D., 2016a. Towards plant pangenomics. Plant Biotechnol. J. 14, 1099–1105.

Golicz, A.A., Bayer, P.E., Barker, G.C., Edger, P.P., Kim, H., Martinez, P.A., Chan, C.K.K., Severn-Ellis, A., McCombie, W.R., Parkin, I.A., 2016b. The pan-genome of an agronomically important crop plant Brassica oleracea. Nat. Commun. 7, 1–8.

Golicz, A.A., Bayer, P.E., Bhalla, P.L., Batley, J., Edwards, D., 2020. Pangenomics comes of age: from bacteria to plant and animal applications. Trends Genet. 36, 132–145.

Gordon, S.P., Contreras-Moreira, B., Woods, D.P., Des Marais, D.L., Burgess, D., Shu, S., Stritt, C., Roulin, A.C., Schackwitz, W., Tyler, L., 2017. Extensive gene content variation in the Brachypodium distachyon pan-genome correlates with population structure. Nat. Commun. 8, 1–13.

Gui, S., Yang, L., Li, J., Luo, J., Xu, X., Yuan, J., Chen, L., Li, W., Yang, X., Wu, S., 2020. ZEAMAP, a comprehensive database adapted to the maize multi-omics era. iScience 23, 101241.

Hirsch, C.N., Foerster, J.M., Johnson, J.M., Sekhon, R.S., Muttoni, G., Vaillancourt, B., Peñagaricano, F., Lindquist, E., Pedraza, M.A., Barry, K., 2014. Insights into the maize pan-genome and pan-transcriptome. Plant Cell 26, 121–135.

Hu, X., Xie, W., Wu, C., Xu, S., 2019. A directed learning strategy integrating multiple omic data improves genomic prediction. Plant Biotechnol. J. 17, 2011–2020.

Huang, S., Ding, J., Deng, D., Tang, W., Sun, H., Liu, D., Zhang, L., Niu, X., Zhang, X., Meng, M., et al., 2013. Draft genome of the kiwifruit *Actinidia chinensis*. Nat. Commun. 4, 1–9.

Hübner, S., Bercovich, N., Todesco, M., Mandel, J.R., Odenheimer, J., Ziegler, E., Lee, J.S., Baute, G.J., Owens, G.L., Grassa, C.J., 2019. Sunflower pan-genome analysis shows that hybridization altered gene content and disease resistance. Native Plants 5, 54–62.

Hufford, M.B., Seetharam, A.S., Woodhouse, M.R., Chougule, K.M., Ou, S., Liu, J., Ricci, W.A., Guo, T., Olson, A., Qiu, Y., 2021. De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. Science 373, 655–662.

Hurgobin, B., Edwards, D., 2017. SNP discovery using a pangenome: has the single reference approach become obsolete? Biology 6, 21.

Ibarra-Laclette, E., Lyons, E., Hernández-Guzmán, G., Pérez-Torres, C.A., Carretero-Paulet, L., Chang, T.-H., Lan, T., Welch, A.J., Juárez, M.J.A., Simpson, J., 2013. Architecture and evolution of a minute plant genome. Nature 498, 94–98.

Igic, B., Lande, R., Kohn, J.R., 2008. Loss of self-incompatibility and its evolutionary consequences. Int. J. Plant Sci. 169, 93–104.

Isidro, J., Jannink, J.L., Akdemir, D., Poland, J., Heslot, N., Sorrells, M.E., 2015. Training set optimization under population structure in genomic selection. Theor. Appl. Genet. 128, 145–158.

Jaillon, O., Aury, J.M., Noel, B., Policriti, A., Clepet, C., Cassagrande, A., Choisne, N., Aubourg, S., Vitulo, N., Jubin, C., 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449, 463–467.

Jayakodi, M., Padmarasu, S., Haberer, G., Bonthala, V.S., Gundlach, H., Monat, C., Lux, T., Kamal, N., Lang, D., Himmelbach, A., 2020. The barley pan-genome reveals the hidden legacy of mutation breeding. Nature 588, 284–289.

Jiao, W.-B., Schneeberger, K., 2017. The impact of third generation genomic technologies on plant genome assembly. Curr. Opin. Plant Biol. 36, 64–70.

Johnson, A.L., Govindarajulu, R., Ashman, T.L., 2014. Bioclimatic evaluation of geographical range in *Fragaria* (Rosaceae): consequences of variation in breeding system, ploidy and species age. Bot. J. Linn. Soc. 176, 99–114.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., 2021. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589.

Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S.Y., Freimer, N.B., Sabatti, C., Eskin, E., 2010. Variance component model to account for sample structure in genome-wide association studies. Nat. Genet. 42, 348–354.

Keller, B., Ariza-Suarez, D., De la Hoz, J., Aparicio, J.S., Portilla-Benavides, A.E., Buendia, H.F., Mayor, V.M., Studer, B., Raatz, B., 2020. Genomic prediction of agronomic traits in common bean (*Phaseolus vulgaris* L.) under environmental stress. Front. Plant Sci. 11, 1001.

Kent, W.J., 2002. BLAT—the BLAST-like alignment tool. Genome Res. 12, 656–664.

Khan, A.W., Garg, V., Roorkiwal, M., Golicz, A.A., Edwards, D., Varshney, R.K., 2020. Super-pangenome by integrating the wild side of a species for accelerated crop improvement. Trends Plant Sci. 25, 148–158.

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

Krasileva, K.V., 2019. The role of transposable elements and DNA damage repair mechanisms in gene duplications and gene fusions in plant genomes. Curr. Opin. Plant Biol. 48, 18–25.

Kuroiwa, T., Ohnuma, M., Imoto, Y., Misumi, O., Nagata, N., Miyakawa, I., Fujishima, M., Yagisawa, F., Kuroiwa, H., 2016. Genome size of the ultrasmall unicellular freshwater green alga, *Medakamo hakoo* 311, as determined by staining with 4′, 6-diamidino-2-phenylindole after microwave oven treatments: II. Comparison with *Cyanidioschyzon merolae*, *Saccharomyces cerevisiae* (*n*, *2n*), and *Chlorella variabilis*. Cytologia 81, 69–76.

Li, Y.-h., Zhou, G., Ma, J., Jiang, W., Jin, L.-g., Zhang, Z., Guo, Y., Zhang, J., Sui, Y., Zheng, L., 2014. De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. Nat. Biotechnol. 32, 1045–1052.

Li, H., Feng, X., Chu, C., 2020. The design and construction of reference pangenome graphs with minigraph. Genome Biol. 21, 1–19.

Li, J., Yuan, D., Wang, P., Wang, Q., Sun, M., Liu, Z., Si, H., Xu, Z., Ma, Y., Zhang, B., et al., 2021. Cotton pan-genome retrieves the lost sequences and genes during domestication and selection. Genome Biol. 22, 119.

Li, H., Wang, S., Chai, S., Yang, Z., Zhang, Q., Xin, H., Xu, Y., Lin, S., Chen, X., Yao, Z., 2022a. Graph-based pan-genome reveals structural and sequence variations related to agronomic traits and domestication in cucumber. Nat. Commun. 13, 1–14.

Li, S., Lin, D., Zhang, Y., Deng, M., Chen, Y., Lv, B., Li, B., Lei, Y., Wang, Y., Zhao, L., 2022b. Genome-edited powdery mildew resistance in wheat without growth penalties. Nature 602, 455–460.

Liu, Y., Tian, Z., 2020. From one linear genome to a graph-based pan-genome: a new era for genomics. Sci. China Life Sci. 63, 1938–1941.

Liu, Y., Du, H., Li, P., Shen, Y., Peng, H., Liu, S., Zhou, G.-A., Zhang, H., Liu, Z., Shi, M., 2020. Pan-genome of wild and cultivated soybeans. Cell 182, 162–176.

Long, Y., Liu, Z., Wang, P., Yang, H., Wang, Y., Zhang, S., Zhang, X., Wang, M., 2021. Disruption of topologically associating domains by structural variations in tetraploid cottons. Genomics 113, 3405–3414.

Mace, E.S., Tai, S., Gilding, E.K., Li, Y., Prentis, P.J., Bian, L., Campbell, B.C., Hu, W., Innes, D.J., Han, X., 2013. Whole-genome sequencing reveals untapped genetic potential in Africa's indigenous cereal crop sorghum. Nat. Commun. 4, 1–9.

Mahmoud, M., Gobet, N., Cruz-Dávalos, D.I., Mounier, N., Dessimoz, C., Sedlazeck, F.J., 2019. Structural variant calling: the long and the short of it. Genome Biol. 20, 1–14.

McClintock, B., 1956. Controlling elements and the gene. Cold Spring Harb. Symp. Quant. Biol. 21, 197–216.

McCoy, J.P., 2011. High-content screening: getting more from less. Nat. Methods 8, 390–391.

Meuwissen, T.H., Hayes, B.J., Goddard, M., 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157, 1819–1829.

Miga, K.H., Koren, S., Rhie, A., Vollger, M.R., Gershman, A., Bzikadze, A., Brooks, S., Howe, E., Porubsky, D., Logsdon, G.A., 2020. Telomere-to-telomere assembly of a complete human X chromosome. Nature 585, 79–84.

Montenegro, J.D., Golicz, A.A., Bayer, P.E., Hurgobin, B., Lee, H., Chan, C.K.K., Visendi, P., Lai, K., Doležel, J., Batley, J., 2017. The pangenome of hexaploid bread wheat. Plant J. 90, 1007–1013.

Morgante, M., De Paoli, E., Radovic, S., 2007. Transposable elements and the plant pan-genomes. Curr. Opin. Plant Biol. 10, 149–155.

Myles, C., Wayne, M., 2008. Quantitative trait locus (QTL) analysis. Nat. Educ. 1, 208.

Niu, S., Li, J., Bo, W., Yang, W., Zuccolo, A., Giacomello, S., Chen, X., Han, F., Yang, J., Song, Y., 2022. The Chinese pine genome and methylome unveil key features of conifer evolution. Cell 185, 204–217.

Panchy, N., Lehti-Shiu, M., Shiu, S.-H., 2016. Evolution of gene duplication in plants. Plant Physiol. 171, 2294–2316.

Parker, C.C., Gopalakrishnan, S., Carbonetto, P., Gonzales, N.M., Leung, E., Park, Y.J., Aryee, E., Davis, J., Blizard, D.A., Ackert-Bicknell, C.L., 2016. Genome-wide association study of behavioral, physiological and gene expression traits in outbred CFW mice. Nat. Genet. 48, 919–926.

Paten, B., Novak, A.M., Eizenga, J.M., Garrison, E., 2017. Genome graphs and the evolution of genome inference. Genome Res. 27, 665–676.

Paterson, A.H., Freeling, M., Tang, H., Wang, X., 2010. Insights from the comparison of plant genome sequences. Annu. Rev. Plant Biol. 61, 349–372.

Patron, J., Serra-Cayuela, A., Han, B., Li, C., Wishart, D.S., 2019. Assessing the performance of genome-wide association studies for predicting disease risk. PLoS ONE 14, e0220215.

Pellicer, J., Fay, M.F., Leitch, I.J., 2010. The largest eukaryotic genome of them all? Bot. J. Linn. Soc. 164, 10–15.

Peng, H., Wang, K., Chen, Z., Cao, Y., Gao, Q., Li, Y., Li, X., Lu, H., Du, H., Lu, M., 2020. MBKbase for rice: an integrated omics knowledgebase for molecular breeding in rice. Nucleic Acids Res. 48, D1085–D1092.

Plissonneau, C., Hartmann, F.E., Croll, D., 2018. Pangenome analyses of the wheat pathogen *Zymoseptoria tritici* reveal the structural basis of a highly plastic eukaryotic genome. BMC Biol. 16, 5.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., De Bakker, P.I., Daly, M.J., 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559–575.

Qiao, Q., Edger, P.P., Xue, L., Qiong, L., Lu, J., Zhang, Y., Cao, Q., Yocca, A.E., Platts, A.E., Knapp, S.J., 2021. Evolutionary history and pan-genome dynamics of strawberry (*Fragaria* spp.). Proc. Natl. Acad. Sci. U. S. A. 118, e2105431118.

Qin, P., Lu, H., Du, H., Wang, H., Chen, W., Chen, Z., He, Q., Ou, S., Zhang, H., Li, X., 2021. Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. Cell 184, 3542–3558.

Rabbani, L., Müller, J., Weigel, D., 2020. An algorithm to build a multi-genome reference. BioRxiv. https://doi.org/10.1101/2020.04.11.036871.

Rasko, D.A., Rosovitz, M.J., Myers, G.S., Mongodin, E.F., Fricke, W.F., Gajer, P., Crabtree, J., Sebaihia, M., Thomson, N.R., Chaudhuri, R., et al., 2008. The pangenome structure of Escherichia coli: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. J. Bacteriol. 190, 6881–6893.

Salman-Minkov, A., Sabath, N., Mayrose, I., 2016. Whole-genome duplication as a key factor in crop domestication. Nat. Plants 2, 16115.

Schatz, M.C., Maron, L.G., Stein, J.C., Wences, A.H., Gurtowski, J., Biggers, E., Lee, H., Kramer, M., Antoniou, E., Ghiban, E., 2014. Whole genome de novo assemblies of three divergent strains of rice, *Oryza sativa*, document novel gene space of aus and indica. Genome Biol. 15, 506.

Schulz, T., Wittler, R., Rahmann, S., Hach, F., Stoye, J., 2021. Detecting high-scoring local alignments in pangenome graphs. Bioinformatics 37, 2266–2274.

Shlemov, A., Korobeynikov, A., 2019. PATHRACER: racing profile HMM paths on assembly graph. In: Holmes, I., Martín-Vide, C., Vega-Rodríguez, M. (Eds.), International conference on algorithms for computational biology. AlCoB 2019. Lecture Notes in Computer Science. Springer, Cham. https://doi.org/10.1007/978-3-030-18174-1_6.

Sibbesen, J.A., Eizenga, J.M., Novak, A.M., Sirén, J., Chang, X., Garrison, E., Paten, B., 2021. Haplotype-aware pantranscriptome analyses using spliced pangenome graphs. BioRxiv. https://doi.org/10.1101/2021.03.26.437240.

Snipen, L., Almøy, T., Ussery, D.W., 2009. Microbial comparative pan-genomics using binomial mixture models. BMC Genom. 10, 385.

Song, W.-Y., Wang, G.-L., Chen, L.-L., Kim, H.-S., Pi, L.-Y., Holsten, T., Gardner, J., Wang, B., Zhai, W.-X., Zhu, L.-H., 1995. A receptor kinase-like protein encoded by the rice disease resistance gene, Xa21. Science 270, 1804–1806.

Song, J.-M., Guan, Z., Hu, J., Guo, C., Yang, Z., Wang, S., Liu, D., Wang, B., Lu, S., Zhou, R., 2020. Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of Brassica napus. Nat. Plants 6, 34–45.

Song, J.M., Liu, D.X., Xie, W.Z., Yang, Z., Guo, L., Liu, K., Yang, Q.Y., Chen, L.L., 2021. BnPIR: Brassica napus pan-genome information resource for 1689 accessions. Plant Biotechnol. J. 19, 412.

Sun, X., Jiao, C., Schwaninger, H., Chao, C.T., Ma, Y., Duan, N., Khan, A., Ban, S., Xu, K., Cheng, L., et al., 2020. Phased diploid genome assemblies and pan-genomes provide insights into the genetic history of apple domestication. Nat. Genet. 52, 1423–1432.

Sun, Y., Shang, L., Zhu, Q.-H., Fan, L., Guo, L., 2021. Twenty years of plant genome sequencing: achievements and challenges. Trends Plant Sci. 27, 391–401.

Sun, H., Jiao, W.B., Krause, K., Campoy, J.A., Goel, M., Folz-Donahue, K., Kukat, C., Huettel, B., Schneeberger, K., 2022. Chromosome-scale and haplotype-resolved genome assembly of a tetraploid potato cultivar. Nat. Genet. 54, 342–348.

Tahir Ul Qamar, M., Zhu, X., Khan, M.S., Xing, F., Chen, L.L., 2020. Pan-genome: a promising resource for noncoding RNA discovery in plants. Plant Genome 13, e20046.

Takayama, S., Isogai, A., 2005. Self-incompatibility in plants. Annu. Rev. Plant Biol. 56, 467–489.

Tao, Y., Zhao, X., Mace, E., Henry, R., Jordan, D., 2019. Exploring and exploiting pan-genomics for crop improvement. Mol. Plant 12, 156–169.

Tay Fernandez, C.G., Nestor, B.J., Danilevicz, M.F., Marsh, J.I., Petereit, J., Bayer, P.E., Batley, J., Edwards, D., 2022. Expanding gene-editing potential in crop improvement with pangenomes. Int. J. Mol. Sci. 23, 2276.

Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., Angiuoli, S.V., Crabtree, J., Jones, A.L., Durkin, A.S., 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". Proc. Natl. Acad. Sci. U. S. A. 102, 13950–13955.

Torkamaneh, D., Lemay, M.A., Belzile, F., 2021. The pan-genome of the cultivated soybean (PanSoy) reveals an extraordinarily conserved gene content. Plant Biotechnol. J. 19, 1852.

Varshney, R.K., Roorkiwal, M., Sun, S., Bajaj, P., Chitikineni, A., Thudi, M., Singh, N.P., Du, X., Upadhyaya, H.D., Khan, A.W., et al., 2021. A chickpea genetic variation map based on the sequencing of 3,366 genomes. Nature 599, 622–627.

Walkowiak, S., Gao, L., Monat, C., Haberer, G., Kassa, M.T., Brinton, J., Ramirez-Gonzalez, R.H., Kolodziej, M.C., Delorean, E., Thambugala, D., 2020. Multiple wheat genomes reveal global variation in modern breeding. Nature 588, 277–283.

Wang, W., Mauleon, R., Hu, Z., Chebotarov, D., Tai, S., Wu, Z., Li, M., Zheng, T., Fuentes, R.R., Zhang, F., 2018. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature 557, 43–49.

Willis, K., 2017. State of the World's Plants 2017. Royal Botanics Gardens Kew, London.

Wong, C., Bernardo, R., 2008. Genomewide selection in oil palm: increasing selection gain per unit time and cost with small populations. Theor. Appl. Genet. 116, 815–824.

Wu, X., Feng, H., Wu, D., Yan, S., Zhang, P., Wang, W., Zhang, J., Ye, J., Dai, G., Fan, Y., et al., 2021. Using high-throughput multiple optical phenotyping to decipher the genetic architecture of maize drought tolerance. Genome Biol. 22, 185.

Xu, X., Liu, X., Ge, S., Jensen, J.D., Hu, F., Li, X., Dong, Y., Gutenkunst, R.N., Fang, L., Huang, L., 2012. Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. Nat. Biotechnol. 30, 105–111.

Xu, S., Zhu, D., Zhang, Q., 2014. Predicting hybrid performance in rice using genomic best linear unbiased prediction. Proc. Natl. Acad. Sci. U. S. A. 111, 12456–12461.

Yu, J., Tehrim, S., Zhang, F., Tong, C., Huang, J., Cheng, X., Dong, C., Zhou, Y., Qin, R., Hua, W., 2014. Genome-wide comparative analysis of NBS-encoding genes between Brassica species and *Arabidopsis thaliana*. BMC Genom. 15, 1–18.

Yu, Q.-h., Wang, B., Li, N., Tang, Y., Yang, S., Yang, T., Xu, J., Guo, C., Yan, P., Wang, Q., 2017. CRISPR/Cas9-induced targeted mutagenesis and gene replacement to generate long-shelf life tomato lines. Sci. Rep. 7, 1–9.

Yu, J., Golicz, A.A., Lu, K., Dossa, K., Zhang, Y., Chen, J., Wang, L., You, J., Fan, D., Edwards, D., 2019. Insight into the evolution and functional characteristics of the pan-genome assembly from sesame landraces and modern cultivars. Plant Biotechnol. J. 17, 881–892.

Yu, X., Xiao, J., Chen, S., Yu, Y., Ma, J., Lin, Y., Li, R., Lin, J., Fu, Z., Zhou, Q., 2020. Metabolite signatures of diverse *Camellia sinensis* tea populations. Nat. Commun. 11, 1–14.

Yu, H., Lin, T., Meng, X., Du, H., Zhang, J., Liu, G., Chen, M., Jing, Y., Kou, L., Li, X., 2021. A route to de novo domestication of wild allotetraploid rice. Cell 184, 1156–1170.

Zhang, J., Zhang, X., Tang, H., Zhang, Q., Hua, X., Ma, X., Zhu, F., Jones, T., Zhu, X., Bowers, J., et al., 2018. Allele-defined genome of the autopolyploid sugarcane *Saccharum spontaneum* L. Nat. Genet. 50, 1565–1573.

Zhang, L., Ren, Y., Yang, T., Li, G., Chen, J., Gschwend, A.R., Yu, Y., Hou, G., Zi, J., Zhou, R., 2019. Rapid evolution of protein diversity by de novo origination in Oryza. Nat. Ecol. Evol. 3, 679–690.

Zhang, C., Yang, Z., Tang, D., Zhu, Y., Wang, P., Li, D., Zhu, G., Xiong, X., Shang, Y., Li, C., 2021a. Genome design of hybrid potato. Cell 184, 3873–3883.

Zhang, Y., Shen, Q., Leng, L., Zhang, D., Chen, S., Shi, Y., Ning, Z., Chen, S., 2021b. Incipient diploidization of the medicinal plant Perilla within 10,000 years. Nat. Commun. 12, 1–13.

Zhao, Q., Feng, Q., Lu, H., Li, Y., Wang, A., Tian, Q., Zhan, Q., Lu, Y., Zhang, L., Huang, T., 2018. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. Nat. Genet. 50, 278–284.

Zhou, X., Stephens, M., 2014. Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nat. Methods 11, 407–409.

Zhou, Z., Jiang, Y., Wang, Z., Gou, Z., Lyu, J., Li, W., Yu, Y., Shu, L., Zhao, Y., Ma, Y., 2015. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nat. Biotechnol. 33, 408–414.

Zhu, G., Wang, S., Huang, Z., Zhang, S., Liao, Q., Zhang, C., Lin, T., Qin, M., Peng, M., Yang, C., 2018. Rewiring of the fruit metabolome in tomato breeding. Cell 172, 249–261.