Received: 14 April 2024 | Revised: 13 May 2024 | Accepted: 4 June 2024 | Published online: 13 June 2024

RESEARCH ARTICLE

Plant Pan-genomics: Opportunities, Advances, and Challenges

Journal of Data Science and Intelligent Systems
2024, Vol. 00(00) 1–10
DOI: 10.47852/bonviewJDSIS42023107

Mingjun Hou[1*], Shaozi Pang[1]
[1]*College of Life Science, Northeast Forestry University, China,* 18766957953@nefu.edu.cn

***Corresponding author:** Mingjun Hou, College of Life Science, Northeast Forestry University, China. Email: 18766957953@nefu.edu.cn

Abstract: The pan-genome concept encompasses all individual genome sequences within a species, comprising both core and variable genes. In recent years, it has emerged as a promising avenue for advancing plant genetic evolution and cultivating desirable traits. This review provides a comprehensive overview of pan-genome representation, detailing both linear pan-genome and graph-based pan-genome. It further presents a thorough analysis of the assembly methods for these pan-genomic structures and discusses the widely-used software tools and processes for their construction, including VG, Minigraph and PGGB. The paper concurrently delineates the fundamental factors that influence pan-genome quality, which include genome assembly quality, accurate annotation, identification of homologous genes, and strategic sample selection. Furthermore, we investigate the pan-genome's applications in plant genomics and provide a summary of recent research findings related to the plant pan-genomes, particularly in elucidating genetic transformations during domestication, uncovering genetic variations linked to agronomic traits, and offering insights to guide breeding endeavors.

Keywords: pan-genomics, crop breeding, core genes, variable genes

1. Introduction

The ever-changing natural environment prompts organisms to adapt for survival, reproduction, and prolonged fitness. Adaptations manifest in many forms, encompassing behavioral shifts like habitat changes or migration to more favorable locales, as well as physiological adjustments such as modifying body temperature and metabolic processes. Across extended periods, species undergo genetic adaptations to environmental shifts, yielding genetic variations among individuals of the same species [1]. For instance, hybridization studies on various isolates of bacterial species like *Helicobacter pylori*, *Escherichia coli*, and *Staphylococcus aureus* reveal substantial genetic diversity, with approximately 20%-35% of genes being unique to each strain [2-4]. This genetic distinctiveness suggests a broader adaptive capacity, facilitating better adjustment to diverse environmental conditions.

However, individual genetic uniqueness presents challenges in data analysis. Traditional genetics and epigenetics analyses rely on a reference genome typically constructed from DNA sequences of specific individuals or a small cohort, failing to capture genetic variations among other members of the species [5]. This reliance on a reference genome can yield misleading outcomes, as it may not accurately reflect genetic variations among other individuals within the species, potentially hindering a comprehensive grasp of genetic diversity and adaptability. The construction of the GRCh38 [6] reference genome primarily relies on genomic data from European and Asian populations. However, it does not fully account for the genetic diversity of certain indigenous groups (such as Australian Aborigines and Native Americans) or specific Asian subraces (like South Asians and Southeast Asians). Consequently, when utilizing GRCh38 as a reference for genomic studies, some genetic variations may be overlooked. Yet, research methods centered on a single reference genome fail to encompass all genetic diversity, thus curtailing our ability to fully comprehend the nature of these genetic variations.

As genomic research progresses, the limitations of traditional single reference genome study models have become

increasingly evident. To tackle this challenge, researchers introduced the pan-genome concept, which encompasses the non-redundant collection of genome sequences of an entire species [7]. The first pan-genomes were constructed for microbes due to their smaller genomes, mostly single-stranded DNA, and relatively simple variations. *Streptococcus agalactiae* (group B Streptococcus or GBS), a bacterium known for its high genetic diversity and adaptability, causing various human infections like neonatal sepsis and highly invasive infections in the elderly, served as a notable example [8]. In GBS, each strain's genome sequence provides novel sequence information. Mathematical models based on these genomes predict that even after sequencing hundreds or thousands of genomes, new genes will continually be discovered from additional individual strains.

In 2007, the concept of the pan-genome was first applied to plants through the description of short variable regions within rice and maize genomes [9].

Subsequently, in 2014, the inaugural plant pan-genome study was published. This study focused on the establishment and analysis of a pan-genome for cultivated soybean wild relatives [10]. The approach involved sequencing and de novo assembly of genomic material from seven phylogenetically and geographically representative accessions. Notably, approximately 80% of the genome is shared across all samples, constituting the core genome. However, the remaining 20% exhibits significant variability and contributes to greater genetic diversity. This variable portion of the genome may influence various agricultural traits, including biotic resistance, seed composition, flowering and maturation time, organ size, and final biomass.

Advancements in sequencing technology have led to increasingly high-quality constructed genomes. Simultaneously, the quality of pan-genomes has also improved [11].

The increasing applications of pan-genomes in the study of plant diversity and variation will gradually establish them as new reference genomes [12], offering broader insights into evolution, selection, and particularly genome functions. Additionally, graph-based pan-genome construction enhances storage and visualization methods for pan-genomes while further uncovering their functional attributes.

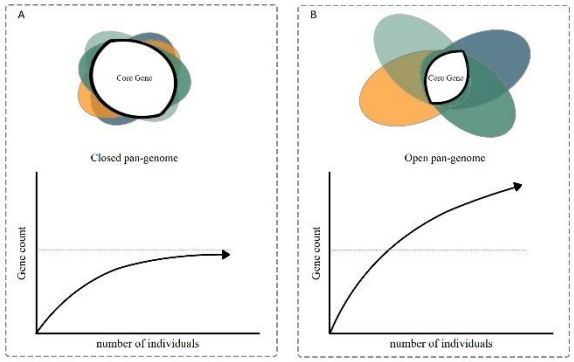
2. Basic Concepts of the Pan-genome

The pan-genome is primarily composed of the core gene and the variable gene. "Core gene" are genes present in all individuals of a species, whereas "variable gene" exist only in certain individuals. [8]. In the pan-genome of GBS, the core gene accounts for approximately 80% of the strain's genes, with the variable gene constituting approximately 20%. Approximately 18% of the variable genes are shared among two or more strains, while 1.5% are unique to each strain. Occasionally, these strain-specific genes are delineated separately and termed private genes.

The core gene encompasses all genes that define the fundamental characteristics of a species. For instance, in the pan-genome of Streptococcus agalactiae, the core gene predominantly includes genes related to essential cellular functions, regulatory functions, and transport and binding proteins [8]. This suggests that the core gene typically governs the organism's essential functions and metabolic processes, encompassing aspects like cell structure, energy conversion, and DNA replication [12]. The extent of the core gene reflects the genetic conservation of a species, representing genes that remain largely stable or undergo minimal evolution over time. These genes usually form most of the pan-genome but diminish proportionately as more individual genomes are incorporated into the analysis. Conversely, the variable gene comprises genes are often associated with specific traits and environmental adaptations of the organism, such as disease resistance, cold tolerance, and plant flower color [13]. The size of the variable gene signifies the genetic diversity within a species, encompassing genes gained or lost during evolution, which constitute only a small fraction of the pan-genome and increase as more individual genomes are examined.

The concept of whether the number of genes in the pan-genome will stabilize with an increase in individual genomes distinguishes pan-genomes into open and closed types [14]. In species with an open pan-genome, the gene count continues to rise as new individual genomes are added, resulting in an unpredictable pan-genome size (Figure 1). In species with a closed pan-genome, the gene count remains relatively constant, allowing for size predictions (Figure 1). The differentiation between open and closed pan-genomes underscores the genetic diversity and the prevalence of horizontal gene transfer within a species [15]. Typically, species with open pan-genomes demonstrate larger population sizes, higher rates of horizontal gene transfer, and greater ecological adaptability. Conversely, species with closed pan-genomes tend to exhibit smaller population sizes, reduced rates of horizontal gene transfer, and diminished environmental adaptability. Pan-genomes of animals and plants primarily belong to the closed category. This provides valuable insights and research opportunities for studying eukaryotic pan-genomes.

Figure 1
Closed and Open Pan-genomes



3. Construction and Presentation of the Pan-genome

3.1. Construction of the pan-genome

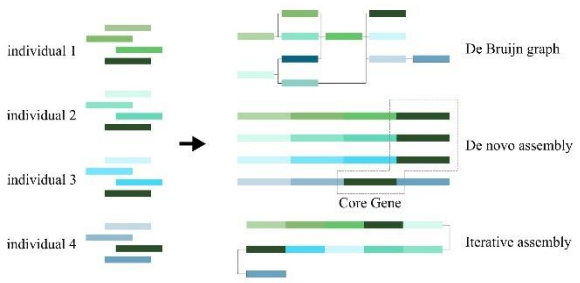
Presently, there are three primary methods for assembling the pan-genome: k-mer-based assembly, iterative assembly, and *de novo* assembly (comparative *de novo* approach) (Figure 2). The k-mer-based method [16] utilizes a

de Bruijn graph technique, where each sequence is fragmented into numerous sub-sequences of length k, referred to as k-mers. These k-mers are visually depicted, with each acting as a node; nodes that overlap by k-1 characters are connected by an edge. This graphical representation may contain multiple edges linking the same nodes and can even form loops. Genomes are reconstructed by analyzing the connections among these k-mer nodes [17]. When examining multiple genomes, it is essential to preserve information on the origin of the nodes, achieved through node coloring. Each node derived from different samples is assigned a unique color, facilitating straightforward tracking. This method allows the entire pan-genome to be visualized as a colored de Bruijn graph, aiding in identifying shared sequences across genomes and unique sequences from individual organisms [18, 19]. Commonly used software for k-mer analysis includes KmerGenie, GenomeScope, and KAT.

Iterative assembly begins by choosing a fully sequenced genome as the initial reference genome [20]. Sequencing data from other individuals' genomes are sequentially aligned to this reference, with unaligned sequences extracted for further assembly. The reference genome is continuously updated with the unaligned sequences, and additional reads from individual genomes are aligned to the updated reference. This iterative procedure of updating and aligning enables the creation of a non-redundant pan-genome. For instance, if a genome fragment is identified in two individual genomes, it is assembled and integrated only once using the reads of the first individual genome. This approach conserves computational resources and is beneficial for larger genomes or when computational capacity is limited. Software tools such as Minimus2, QUAST, and CISA integrate assembled sequences with the reference genome.

De novo assembly stands out as the most prevalent method for pan-genome assembly [21]. Through *de novo* assembly, each individual undergoes assembly given adequate genome sequencing coverage. Comparisons of whole genomes subsequently unveil common or variable regions across different individuals. A key advantage of *de novo* assembly lies in its ability to detect more structural variations (SVs) compared to iterative assembly. However, it necessitates substantial computational power and deep sequencing, rendering it less practical for species with expansive genomes or extensive population studies. Furthermore, assembled genomes can undergo annotation to pinpoint homologous gene clusters, alongside core and variable gene families.

Figure 2
Methods for Pan-genome Assembly



MUMmer4 [22] and Minimap2 [23] are widely employed tools for genome comparisons, particularly in the context of identifying presence-absence variations (PAVs) and constructing linear pan-genomes.

3.2. Graph-based pan-genome

To effectively depict and reference the pan-genome, a comprehensive understanding of the sequence components and their location within the variable genome is paramount. However, a linear pan-genome solely exhibits the sequence components of the variable genome, lacking chromosomal location information and leading to notable data loss. To address this limitation, two strategies are employed in pan-genome construction: one integrates location information into the sequence fragments of the linear pan-genome. In contrast, the other employs nodes to represent sequence information and edges to illustrate connections between sequences. This approach effectively merges the reference genome with genetic variations and stores the pan-genome in a graphical format [7, 24]. graph-based pan-genomes offer the advantage of expanding the coordinate system of linear reference genomes to encompass broader genetic diversity, thereby enhancing the precision and efficiency of genome annotation, alignment, and variant detection [25]. However, this method necessitates substantial computational resources and a more intricate analysis process, constrained by the lack of standardized protocols and tools. While the utilization of graph-based pan-genomes in plants is currently limited, emerging studies underscore its potential in genomic research, breeding, and evolutionary studies.

The construction methodologies of graph-based pan-genomes are primarily classified into three types: the variation graph-based pan-genome, sequence graph-based pan-genome, and component graph-based pan-genome [24].

The variation graph-based pan-genome utilizes the reference genome as a framework, aligning sequences from other individual genomes and branching off at points of divergence. It evolves and expands as new sequences are incorporated, forming a complex graphical structure that encapsulates all variation information in the genome. This method efficiently captures diverse variations, including single nucleotide polymorphisms (SNPs), small fragment insertions and deletions (indel), and structural variations (SV), providing a visual representation of the genome's variability. However, it encounters challenges in managing large-scale datasets and complex recombination events [26].

The sequence graph-based pan-genome integrates all genetic variation information of a species into the reference genome via *de novo* assembly. It is depicted as a graphic illustrating alternative pathway, facilitating the construction of a graph-based pan-genome [25]. This approach utilizes high-quality genomes from *de novo* assembly to effectively capture large segment structural variations (LSV), such as copy number variations (CNVs), transposable elements (TEs), and repetitive sequences (REPs), thus covering the majority of individual differences and providing extensive genetic information for species research. However, it demands substantial computational resources and storage capacity.

The component graph-based pan-genome employs genomic components (e.g., genes, transcripts, exons) as nodes in the graphical structure. It compares and aligns genome components of different individuals or species to construct a graph-based pan-genome containing comprehensive component information. This method efficiently captures data on the presence or absence (P/A), organizational structure (OS), and functional annotation (FA) of genome sequences, catering to various research objectives and analytical requirements [27]. However, it necessitates precise definition and annotation of genome components.

The construction of graph-based pan-genome currently relies on several popular methods, including Minigraph [28-30], and VG [31]. PGGB utilizes ODGI to precisely extract and explore variant graphs. Notably, it does not require constructing a graph based on a reference sequence. Minigraph constructs a pangenome graph based on a reference sequence. It is particularly suitable for capturing complex variations in the genome. Leveraging the Cactus aligner, Minigraph-Cactus introduces small variations into the graph without relying on a reference sequence. VG is a versatile toolkit for building and manipulating graphical pangenomes. It supports both long sequences (VG map) and short sequences (VG giraffe) for alignment on the graph. Additionally, VG includes a set of annotation tools (RPVG) that facilitate comparative analyses and aid in identifying pathways of variation within the pangenome.

The computational resources required for constructing pangenomes are a critical consideration. Current algorithms and processes consume significant time and computational resources during execution, which limits the practical application scope of pangenomes. Furthermore, the data structures resulting from the completion of graphical pangenome construction are static and unmodifiable [32]. Therefore, enhancing the efficiency of pangenome construction and utilization, as well as updating existing graphical pangenomes, become key challenges in the development of pangenomics.

4. Factors Influencing Pan-genome Quality

4.1. Genome assembly quality

The quality of genome assembly, a crucial factor directly impacting pan-genome analysis, can be assessed quantitatively using methods like CEGMA [33], BUSCO [34], and LAI [35].

In plant genomes, there are numerous repetitive sequences, high heterozygosity, and polyploidy. Moreover, the genomes of certain gymnosperms, amphibians, and reptiles exceed 5 Gb in size. These factors significantly complicate genome sequencing and assembly.

The advent of third-generation sequencing technologies, such as PacBio [36] and Oxford Nanopore [37], has substantially increased read lengths. Moreover, the introduction of high-fidelity (HiFi) [38] sequencing technologies and haplotype-resolved assembly software has dramatically simplified the exploration of polyploid and highly heterozygous plant genomes.

By leveraging HiFi and Hi-C data, researchers have successfully constructed autopolyploid and heterozygous genomes of cultivated alfalfa [39], potato [40], sugarcane [41], and tea [42].

New assembly algorithms, such as canu [43] and flye [44], can also be developed to accommodate the characteristics of third-generation sequencing data, thereby improving both assembly speed and quality.

4.2. Annotation quality

Gene annotation is the process of identifying the positions and functions of genes within a genome. Two primary methods dominate gene prediction: de novo prediction and evidence-based prediction. De novo prediction involves analyzing genomic DNA sequences to systematically identify features indicative of protein-coding potential. In contrast, evidence-based prediction relies on existing expression evidence. This includes data from expressed sequence tags, RNA sequencing, and protein sequences. By comparing these data with reference sequences, scientists infer gene locations and structures

The complexity of large genomes, characterized by high GC content and a significant proportion of repetitive sequences, poses challenges for gene annotation. To address this, a combined approach, leveraging both de novo and evidence-based methods, can be employed. However, it's essential to recognize that this hybrid approach requires substantial computational resources due to the vast amount of data involved.

Discrepancies in gene annotation may stem from differences in platforms or software, which may apply varying criteria for gene start codons, stop codons, and intron splice sites. Additionally, annotation quality depends on annotation databases, which cover gene function annotation, homology comparison, and gene ontology. The comprehensiveness of these databases directly impacts the accuracy and comprehensiveness of annotation. A notable annotation software is MAKER [45], which integrates both *de novo* and evidence-based prediction methods to produce highly reliable gene models.

4.3. Homologous gene detection

Homologous gene detection plays a crucial role in pan-genome research by identifying genes with similar or identical functions and sequences across different species or individuals within the same species. This process aids in the identification of core and variable gene families [46], thus reflecting the composition and diversity of the pan-genome. Currently, two primary methods are utilized for homologous gene detection: sequence alignment and phylogenetic approaches. The sequence alignment method employs computer software to align functional unknown sequences from an individual's genome with a known gene database, such as using BLAST [47], to identify the most similar sequences. This method determines homology based on similarity and conservation, necessitating efficient and precise sequence alignment algorithms and extensive gene databases.

In contrast, the phylogenetic method, rooted in molecular evolution theory, constructs a phylogenetic tree based on sequence disparities between functional unknown se-

quences from an individual's genome and known genes, illustrating their phylogenetic relationships and evolutionary history. This method relies on appropriate evolutionary models and tree construction algorithms, such as MEGA [48], which integrates several unique phylogenetic analysis functions.

4.4. Choosing the right samples

Thoughtful selection of the sample range for pan-genome analysis significantly influences the size and diversity of the pan-genome [49]. Sample selection for the pan-genome should be guided by the following considerations: (1) Representation of genetic variation and evolutionary history within the species, guaranteeing the pan-genome comprehensively reflects the genomic structure and function of the species. This encompasses wild types, cultivated varieties, and individuals from various geographical areas and ecological environments. (2) Coverage of primary phylogenetic relationships and genotypic groups within the species, limiting bias towards certain related or homologous individuals, minimizing redundant sequences and false-positive variations in the pan-genome. (3) Reflection of key traits and functional genes within the species, including stress resistance, disease resistance, yield, and quality, enhancing the practical use and research significance of the pan-genome. (4) Satisfaction of technical requirements for pan-genome construction, encompassing sequencing depth, data quality, and assembly efficiency, to ensure the accuracy and dependability of the pan-genome representation.

5. Applications of Plant Pan-genomes

Plant pan-genomes involve extensive sequencing and analysis of genomes from multiple individuals within a species, providing comprehensive genetic insights for enhancement. They aid in tracking gene retention and loss during domestication and breeding, with the potential to reintroduce lost genes into modern varieties and restore genetic diversity [50]. In the domestication process of cotton, scientists have identified numerous genes that exist in wild varieties but are lost in cultivated species [51]. Through pan-genome analysis, researchers have pinpointed these missing genes and explored their potential impact on cotton fiber quality and yield. Some of these lost genes may be associated with disease resistance, drought tolerance, or other crucial agronomic traits in cotton. Consequently, modern gene-editing techniques can reintroduce these beneficial lost genes into contemporary cotton varieties, thereby restoring genetic diversity and enhancing crop adaptability and productivity. Gao et al. utilized 725 tomato (*Solanum lycopersicum*) varieties with diverse traits and geographic locations to identify 4873 genes absent in the reference genome. They observed substantial gene loss or negative selection during domestication and improvement, several of which were genes necessary for disease resistance [49]. The production of graph-structured pan-genomes has steadily advanced, integrating more functional elements and sequence space.

The utilization of the graphical pan-genome as a reference genome for analyzing individual genomes significantly enhances our comprehension of the intricate genetic makeup within both individual organisms and entire species. In 2021, Qin et al. produced a high-quality graph-based pan-genome of rice (*Oryza sativa* and *Oryza glaberrima*) [27]. Their study offered genetic resources for rice genome variation and domestication, examined the distribution and processes of SV development, and examined the influence of SVs on gene expression. An insertion of a long terminal repeat (LTR), spanning 987 base pairs (bp), was identified using the rice graph genome. This LTR insertion is considered a probable causal variant associated with leaf senescence. It remained undetectable when common single nucleotide polymorphisms (SNPs) were called against the single linear reference genome. This study highlighted instances in which SVs directly impacted environmental adaptability and agronomic traits, underscoring the role of high-quality genome assembly and graph-based pan-genomes in plant and functional genomics. In 2020, a graph-based pan-genome was developed from the *de novo* assembly of genomes from 26 soybean lines and three previously documented genomes. This was supplemented with resequencing data derived from 2,898 different lines, revealing numerous variations undetectable in a single reference genome. This endeavor offered a more comprehensive genomic background for soybean evolution and functional genomics investigation. By examining whole-genome duplication regions and structural variations (SVs), it was demonstrated that genome duplication exerts a function in SV evolution [25].

Modern pan-genome research transcends genes, with numerous studies suggesting that vital agronomic traits can be shaped by variations in non-coding regions [12, 52, 53]. For instance, by examining the tomato pan-genome, researchers identified sequence variations linked to tomato flavor within the *TomLoxC* promoter, a selection made during domestication. This understanding of non-coding regions paves the way for leveraging heterosis to bolster trait enhancement in production [49].

Table 1
Recent Advances in Pan-genome Research

Year	Species	Genome Size	Number of Individuals	Reference
2019	*Helianthus annuus* L.	3 Gb	493	[54](Hübner et al., 2019)
2019	*Sesamum indicum* L.	275 Mb	5	[55] (Yu et al., 2019)
2019	*Solanum lycopersicum* L.	827 Mb	725	[49](Gao et al., 2019)
2020	*Cajanus cajan*	590 Mb	89	[56](Zhao et al., 2020)
2020	*Glycine max*	978 Mb	26	[25](Liu et al., 2020)
2020	*Prunus persica*	227 Mb	4	[57](Cao et al., 2020)

2021	*Carya illinoinensis*	674 Mb	4	[58](Lovell et al., 2021)
2021	*Fragaria spp.*	794 Mb	5	[59](Qiao et al., 2021)
2021	*Brassica napus*	1 Gb	18	[60](Cai et al., 2021)
2021	*Gossypium hirsutum* L.	2.3 Gb	1913	[51](Li et al., 2021)
2021	*Oryza sativa* L.	374 Mb	33	[27](Qin et al., 2021)
2021	*Solanum melongena* L.	833 Mb	23	[61](Barchi et al., 2021)
2021	*Sorghum bicolor* L.	709 Mb	13	[62](Tao et al., 2021)
2022	*Cucumis sativus* L.	225 Mb	12	[63](Li et al. 2022)
2023	*Camellia sinensis*	3.1 Gb	22	[64](Chen et al., 2023)
2023	*Citrus*	299 Mb	12	[65](Huang et al., 2023)
2023	*Cucumis melo ssp.*	438 Mb	9	[66](Lyu et al., 2023)
2023	*Malus domestica*	703 Mb	13	[67](Wang et al., 2023)
2023	*Setaria italica*	406 Mb	110	[68](He et al., 2023)
2023	*Zea mays* L.	2.2 Gb	12	[69](Wang et al., 2023)
2024	*Brassica oleracea*	489 Mb	8	[70](Guo et al., 2024)

6. Future Perspectives

Advancing research on plant pan-genomes promises to unveil the genetic diversity and evolutionary lineage of species. It facilitates the discovery and mapping of genes and markers linked to vital agronomic traits, while also identifying and harnessing large segment variations like presence or absence variants, copy number variants, and structural variants. Furthermore, it illuminates gene flow and transfer between species, aiding breeders in selecting suitable parent candidates, enhancing compatibility and heterosis in hybrid combinations, and boosting breeding efficiency and precision. This comprehensive approach broadens the genetic foundation of breeding, providing guidance for intergeneric hybridization and gene transfer, and producing novel germplasm resources for breeding.

Currently, the development of graph-based pan-genomes is in its infancy, facing challenges in its application to species characterized by high complexity and large genome sizes. Additionally, there is a lack of established standards for evaluating graph-based pan-genomes. Most bioinformatics tools are geared towards linear reference genomes, necessitating the creation of more algorithms and tools for downstream examination of graph-based pan-genomes. Beyond uncovering new SVs and associated phenotypic traits, it is crucial to develop graph-based pan-genome applications that can seamlessly integrate with multi-omics data for multidimensional association analysis to locate candidate loci. For example, integrating DNA methylation and other relevant information into graph-based pan-genomes could enable the comparison of various phenotypes resulting from the methylation of distinct alleles. graph-based pan-genomes, operating as reference genomes, signify a future trajectory with massive potential to revolutionize crop breeding and make substantial contributions to global food security and sustainable agricultural development.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

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

Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

References

[1] Nannan, L., Huamiao, L., Yan, J., Xingan, L., Yang, L., Tianjiao, W., ..., & Xiumei, X. (2022). Geometric morphology and population genomics provide insights into the adaptive evolution of Apis cerana in Changbai Mountain. *BMC Genomics, 23*(1), 64.

[2] Dorrell, N., Mangan, J. A., Laing, K. G., Hinds, J., Linton, D., Al-Ghusein, H., ..., & Karlyshev, A. V. (2001). Whole genome comparison of *Campylobacter jejuni* human isolates using a low-cost microarray reveals extensive genetic diversity. *Genome research, 11*(10), 1706-1715.

[3] Fitzgerald, J. R., Sturdevant, D. E., Mackie, S. M., Gill, S. R., & Musser, J. M. (2001). Evolutionary genomics of *Staphylococcus aureus:* Insights into the origin of methicillin-resistant strains and the toxic shock syndrome epidemic. *Proceedings of the National Academy of Sciences of the United States of America, 98*(15), 8821-8826.

[4] Fukiya, S., Mizoguchi, H., & Tobe, T. (2004). Extensive genomic diversity in pathogenic *Escherichia coli* and *Shigella* Strains revealed by comparative genomic hybridization microarray. *Journal of Bacteriology, 186*(12), 3911-

3921.

[5] Lee, J. S., & Kim, N. S. (2022). Genomic perspectives on epigenetics. *Genes & Genomics, 44*(3), 247-249.

[6] Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., ..., & Pääbo, S. (2012) A draft sequence of the Neandertal genome. *Science*, *328*(5979), 710-722.

[7] Bian, P. P., Zhang, Y., & Jiang, Y. (2021). Pan-genome: Setting a new standard for high-quality reference genomes. *Yi Chuan= Hereditas, 43*(11), 1023-1037.

[8] Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., ..., & Fraser, C. M. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". In *Proceedings of the National Academy of Sciences of the United States of America, 102*(39), 13950-13955.

[9] Morgante, M., De Paoli, E., & Radovic, S. (2007). Transposable elements and the plant pan-genomes. *Current Opinion in Plant Biology*, *10*(2), 149–155.

[10] Li, Y. H., Zhou, G., Ma, J., Jiang, W., Jin, L. G., Zhang, Z., ..., & Qiu, L. J. (2014). *De novo* assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nature biotechnology, 32*(10), 1045-1052.

[11] Torkamaneh, D., Lemay, M. A., & Belzile, F. (2021). The pan-genome of the cultivated soybean (PanSoy) reveals an extraordinarily conserved gene content. *Plant Biotechnology Journal*, *19*(9), 1852–1862.

[12] Bayer, P. E., Golicz, A. A., Scheben, A., Batley, J., & Edwards, D. (2020). Plant pan-genomes are the new reference. *Nature Plants, 6*(8), 914-920.

[13] Lei, L., Goltsman, E., Goodstein, D., Wu, G. A., Rokhsar, D. S., & Vogel, J. P. (2021). Plant pan-genomics comes of age. *Annual Review of Plant Biology, 72*, 411-435.

[14] Lefébure, T., Bitar, P. D., Suzuki, H., & Stanhope, M. J. (2010). Evolutionary dynamics of complete *Campylobacter* pan-genomes and the bacterial species concept. *Genome Biology and Evolution, 2*, 646-655.

[15] Rouli, L., Merhej, V., Fournier, P. E., & Raoult, D. (2015). The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New Microbes New Infect, 7*, 72-85.

[16] Jayakodi, M., Padmarasu, S., Haberer, G., Bonthala, V. S., Gundlach, H., Monat, C., ..., & Stein, N. (2020). The barley pan-genome reveals the hidden legacy of mutation breeding. *Nature, 588*(7837), 284-289.

[17] Aylward, A. J., Petrus, S., Mamerto, A., Hartwick, N. T., & Michael, T. P. (2023). PanKmer: *k*-mer-based and reference-free pangenome analysis. *Bioinformatics, 39*(10), btad621.

[18] Iqbal, Z., Caccamo, M., Turner, I., Flicek, P., & McVean, G. (2012). *De novo* assembly and genotyping of variants using colored de Bruijn graphs. *Nature genetics, 44*(2), 226-232.

[19] Marcus, S., Lee, H., & Schatz, M. C. (2014). SplitMEM: A graphical algorithm for pan-genome analysis with suffix skips. *Bioinformatics, 30*(24), 3476-3483.

[20] Montenegro, J. D., Golicz, A. A., Bayer, P. E., Hurgobin, B., Lee, H., Chan, C. K., ..., & Batley, J. (2017). The pan-genome of hexaploid bread wheat. *The Plant journal, 90*(5), 1007-1013.

[21] Li, R., Li, Y., Zheng, H., Luo, R., Zhu, H., Li, Q., ..., & Wang, J. (2010). Building the sequence map of the human pan-genome. *Nature Biotechnology, 28*(1), 57-63.

[22] Marçais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L., & Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system. *PLoS Computational Biology*, *14*(1), e1005944.

[23] Li H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics, 34*(18), 3094–3100.

[24] Wang, S., Qian, Y. Q., Zhao, R. P., Chen, L. L., & Song, J. M. (2023). Graph-based pan-genomes: Increased opportunities in plant genomics. *Journal of Experimental Botany, 74*(1), 24-39.

[25] Liu, Y., Du, H., Li, P., Shen, Y., Peng, H., Liu, S., ..., & Tian, Z. (2020). Pan-genome of wild and cultivated soybeans. *Cell, 182*(1), 162-176.

[26] Tetikol, H. S., Turgut, D., Narci, K., Budak, G., Kalay, O., Arslan, E., ..., & Davis-Dusenbery, B. N. (2022). Pan-African genome demonstrates how population-specific genome graphs improve high-throughput sequencing data analysis. *Nature Communications, 13*(1), 4384.

[27] Qin, P., Lu, H., Du, H., Wang, H., Chen, W., Chen, Z., ..., & Li, S. (2021). Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell, 184*(13), 3542-3558.

[28] Li, H., Feng, X., & Chu, C. (2020). The design and construction of reference pangenome graphs with minigraph. *Genome Biology*, *21*, 1-19.

[29] Hickey, G., Monlong, J., Ebler, J., Novak, A. M., Eizenga, J. M., Gao, Y., ..., & Paten, B. (2024). Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nature biotechnology*, 1-11.

[30] Garrison, E., Guarracino, A., Heumos, S., Villani, F., Bao, Z., Tattini, L., ..., & Prins, P. (2023). Building pangenome graphs. *bioRxiv Preprint: 2023.04.05.535718*.

[31] Hickey, G., Heller, D., Monlong, J., Sibbesen, J. A., Sirén, J., Eizenga, J., ..., & Paten, B. (2020). Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome biology*, *21*, 1-17.

[32] Andreace, F., Lechat, P., Dufresne, Y., & Chikhi, R. (2023). Comparing methods for constructing and representing human pangenome graphs. *Genome biology*, *24*(1), 274.

[33] Parra, G., Bradnam, K., & Korf, I. (2007). CEGMA: A pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics, 23*(9), 1061-1067.

[34] Simao, F. A., Waterhouse, R. M., Ioannidis, P., & Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics, 31*(19), 3210-3212.

[35] Ou, S., Chen, J., & Jiang, N. (2018). Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Research, 46*(21), e126.

[36] Rhoads, A., & Au, K. F. (2015). PacBio sequencing and its applications. *Genomics, Proteomics & Bioinformatics*, *13*(5), 278–289.

[37] Lu, H., Giordano, F. & Ning, Z. (2016). Oxford Nanopore MinION sequencing and genome assembly. *Genomics, Proteomics & Bioinformatics*, *14*(5), 265–279.

[38] Feng, X., Cheng, H., Portik, D., & Li, H. (2022). Metagenome assembly of high-fidelity long reads with hifiasm-meta. *Nature methods*, *19*(6), 671–674.

[39] Chen, H., Zeng, Y., Yang, Y., Huang, L., Tang, B., Zhang, H., ..., & Qiu, Q. (2020). Allele-aware chromosome-level genome assembly and efficient transgene-free genome editing for the autotetraploid cultivated alfalfa. *Nature communications, 11*(1), 2494.

[40] Wang, L., Li, Z., Liu, Y., Chen, S., Li, L., Duan, P., ..., & Tian, Y. (2022). A chromosome-level genome assembly of the potato grouper (Epinephelus tukula). *Genomics, 114*(5), 110473.

[41] Shearman, J. R., Pootakham, W., Sonthirod, C., Naktang, C., Yoocha, T., Sangsrakru, D., ..., & Tangphatsornruang, S. (2022). A draft chromosome-scale genome assembly of a commercial sugarcane. *Scientific reports, 12*(1), 20474.

[42] Liu, H., Zhang, R., Zhou, B. F., Shen, Z., Chen, X. Y., Gao, J., & Wang, B. (2023). Chromosome-scale genome assembly of sweet tea (Lithocarpus polystachyus Rehder). *Scientific Data, 10*(1), 873.

[43] 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 Research*, *27*(5), 722–736.

[44] Freire, B., Ladra, S., & Parama, J. R. (2022). Memory-efficient assembly using Flye. In *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *19*(6), 3564–3577.

[45] Cantarel, B. L., Korf, I., Robb, S. M., Parra, G., Ross, E., Moore, B., ..., & Yandell, M. (2008). MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome research, 18*(1), 188-196.

[46] Lin, K., Zhang, N., Severing, E. I., Nijveen, H., Cheng, F., Visser, R. G., ..., & Bonnema, G. (2014). Beyond genomic variation--Comparison and functional annotation of three *Brassica rapagenomes*: A turnip, a rapid cycling and a Chinese cabbage. *BMC Genomics, 15*, 1-17.

[47] Mount, D. W. (2007). Using the basic local alignment search tool (BLAST). *Cold Spring Harbor Protocols, 2007*(7), pdb.top17.

[48] Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Molecular biology and evolution, 35*(6), 1547.

[49] Gao, L., Gonda, I., Sun, H., Ma, Q., Bao, K., Tieman, D. M., ..., & Fei, Z. (2019). The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nature Genetics, 51*(6), 1044-1051.

[50] Song, J. M., Guan, Z., Hu, J., Guo, C., Yang, Z., Wang, S., ..., & Guo, L. (2020). Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of Brassica napus. *Nature plants, 6*(1), 34-45.

[51] Li, J., Yuan, D., Wang, P., Wang, Q., Sun, M., Liu, Z., ..., & Wang, M. (2021). Cotton pan-genome retrieves the lost sequences and genes during domestication and selection. *Genome Biology, 22*, 1-26.

[52] Alonge, M., Wang, X., Benoit, M., Soyk, S., Pereira, L., Zhang, L., ..., & Lippman, Z. B. (2020). Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell, 182*(1), 145-161.

[53] Liu, Y. & Tian, Z. (2020a). From one linear genome to a graph-based pan-genome: A new era for genomics. *Science China. Life sciences, 63*(12), 1938-1941.

[54] Hübner, S., Bercovich, N., Todesco, M., Mandel, J. R., Odenheimer, J., Ziegler, E., ..., & Rieseberg, L. H. (2019). Sunflower pan-genome analysis shows that hybridization altered gene content and disease resistance. *Nature plants, 5*(1), 54-62.

[55] Yu, J., Golicz, A. A., Lu, K., Dossa, K., Zhang, Y., Chen, J., ..., & Zhang, X. (2019). Insight into the evolution and functional characteristics of the pan-genome assembly

from sesame landraces and modern cultivars. *Plant Biotechnology Journal, 17*(5), 881-892.

[56] Zhao, J., Bayer, P. E., Ruperao, P., Saxena, R. K., Khan, A. W., Golicz, A. A., ..., & Varshney, R. K. (2020). Trait associations in the pangenome of pigeon pea (*Cajanus cajan*). *Plant Biotechnology Journal, 18*(9), 1946-1954.

[57] Cao, K., Peng, Z., Zhao, X., Li, Y., Liu, K., Arus, P., ..., & Wang, L. (2020). Pan-genome analyses of peach and its wild relatives provide insights into the genetics of disease resistance and species adaptation. *bioRxiv Preprint: 2020.07.13.200204*.

[58] Lovell, J. T., Bentley, N. B., Bhattarai, G., Jenkins, J. W., Sreedasyam, A., Alarcon, Y., ..., & Randall, J. J. (2021). Four chromosome scale genomes and a pan-genome annotation to accelerate pecan tree breeding. *Nature communications, 12*(1), 4125.

[59] Qiao, Q., Edger, P. P., Xue, L., Qiong, L., Lu, J., Zhang, Y., ..., & Zhang, T. (2021). Evolutionary history and pan-genome dynamics of strawberry (*Fragaria spp*.). In *Proceedings of the National Academy of Sciences of the United States of America, 118*(45), e2105431118.

[60] Cai, X., Chang, L., Zhang, T., Chen, H., Zhang, L., Lin, R., Liang, J., Wu, J., Freeling, M., & Wang, X. (2021). Impacts of allopolyploidization and structural variation on intraspecific diversification in Brassica rapa. *Genome biology, 22*(1), 166.

[61] Barchi, L., Rabanus-Wallace, M. T., Prohens, J., Toppino, L., Padmarasu, S., Portis, E., ..., & Lanteri, S. (2021). Improved genome assembly and pan-genome provide key insights into eggplant domestication and breeding. *The Plant journal, 107*(2), 579-596.

[62] Tao, Y., Luo, H., Xu, J., Cruickshank, A., Zhao, X., Teng, F., ..., & Mace, E. (2021). Extensive variation within the pan-genome of cultivated and wild sorghum. *Nat Plants, 7*(6), 766-773.

[63] Li, H., Wang, S., Chai, S., Yang, Z., Zhang, Q., Xin, H. ..., & Zhang, Z. (2022). Graph-based pan-genome reveals structural and sequence variations related to agronomic traits and domestication in cucumber. *Nature communications, 13*(1), 682.

[64] Chen, S., Wang, P., Kong, W., Chai, K., Zhang, S., Yu, J., ..., & Zhang, X. (2023). Gene mining and genomics-assisted breeding empowered by the pangenome of tea plant Camellia sinensis. *Nature plants, 9*(12), 1986-1999.

[65] Huang, Y., He, J., Xu, Y., Zheng, W., Wang, S., Chen, P., ..., & Xu, Q. (2023). Pangenome analysis provides insight into the evolution of the orange subfamily and a key gene for citric acid accumulation in citrus fruits. *Nature genetics, 55*(11), 1964-1975.

[66] Lyu, X., Xia, Y., Wang, C., Zhang, K., Deng, G., Shen, Q., ..., & Zhang, M. (2023) Pan-genome analysis sheds light on structural variation-based dissection of agronomic traits in melon crops. *Plant Physiology, 193*(2), 1330-1348.

[67] Wang, T., Duan, S., Xu, C., Wang, Y., Zhang, X., Xu, X., ..., & Wu, T. (2023a). Pan-genome analysis of 13 *Malus* accessions reveals structural and sequence variations associated with fruit traits. *Nature Communications, 14*(1), 7377.

[68] He, Q., Tang, S., Zhi, H., Chen, J., Zhang, J., Liang, H., ..., & Jia, G. (2023). A graph-based genome and pan-genome variation of the model plant *Setaria*. *Nature genetics, 55*(7), 1232-1242.

[69] Wang, B., Hou, M., Shi, J., Ku, L., Song, W., Li, C., ..., & Wang, H. (2023b). De novo genome assembly and analyses of 12 founder inbred lines provide insights into maize heterosis. *Nature Genetics, 55*(2), 312-323.

[70] Guo, N., Wang, S., Wang, T., Duan, M., Zong, M., Miao, L., Han, S., ..., & Liu, F. (2024). Graph-based pan-genome of *Brassica oleracea* provides new insights into its domestication and morphotype diversification. *Plant Commun*, *5*(2), 100791.