Journal of Advanced Research xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of Advanced Research

journal homepage: www.elsevier.com/locate/jare

Review Article

Developing pangenomes for large and complex plant genomes and their representation formats

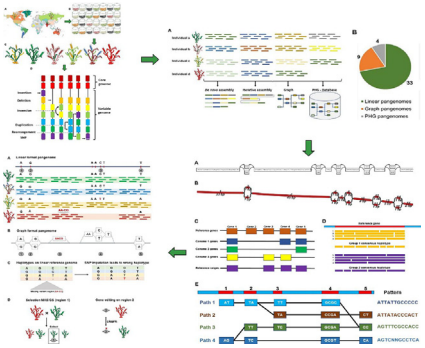
Pradeep Ruperao [a,*], Parimalan Rangan [b,c], Trushar Shah [d], Vinay Sharma [a], Abhishek Rathore [e], Sean Mayes [a], Manish K. Pandey [a,*]

[a] Center of Excellence in Genomics and Systems Biology (CEGSB) and Center for Pre-Breeding Research (CPBR), International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad, India
[b] ICAR-National Bureau of Plant Genetic Resources (NBPGR), New Delhi, India
[c] Queensland Alliance for Agriculture and Food Innovation, The University of Queensland, St Lucia, Australia
[d] International Institute of Tropical Agriculture (IITA), Nairobi, Kenya
[e] International Maize and Wheat Improvement Center (CIMMYT), Nairobi, Kenya

HIGHLIGHTS

- Recent NGS progress allows sequencing multiple genotypes per species, revolutionizing genomic analysis.
- Pan-genomes capture diverse genetic variations for comprehensive comparative analysis, especially in dioecious plants.
- Large plant genomes pose sequencing and computational challenges, addressed by methods like skim-sequencing and RNA-seq
- Emergence of specialized software tools aids in constructing pan-genomes, enhancing research efficiency in plant genomics.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 25 March 2024
Revised 27 January 2025
Accepted 27 January 2025
Available online xxxx

Keywords:
Pangenome
Graph format
PHG
Haplotype graph
Genome viewer
Graph viewer

ABSTRACT

Background: The development of pangenomes has revolutionized genomic studies by capturing the complete genetic diversity within a species. Pangenome assembly integrates data from multiple individuals to construct a comprehensive genomic landscape, revealing both core and accessory genomic elements. This approach enables the identification of novel genes, structural variations, and gene presence-absence variations, providing insights into species evolution, adaptation, and trait variation. Representing pangenomes requires innovative visualization formats that effectively convey the complex genomic structures and variations.

Aim: This review delves into contemporary methodologies and recent advancements in constructing pangenomes, particularly in plant genomes. It examines the structure of pangenome representation, including format comparison, conversion, visualization techniques, and their implications for enhancing crop improvement strategies.

Key scientific concepts of review: Earlier comparative studies have illuminated novel gene sequences, copy number variations, and presence-absence variations across diverse crop species. The concept of a pan-genome, which captures multiple genetic variations from a broad spectrum of genotypes, offers a holistic perspective of a species' genetic makeup. However, constructing a pan-genome for plants with larger genomes poses challenges, including managing vast genome sequence data and comprehending

* Corresponding authors.
E-mail addresses: pradeep.ruperao@icrisat.org (P. Ruperao), Manish.Pandey@icrisat.org (M.K. Pandey).

https://doi.org/10.1016/j.jare.2025.01.052
2090-1232/

Please cite this article as: P. Ruperao, P. Rangan, T. Shah et al., Developing pangenomes for large and complex plant genomes and their representation formats, Journal of Advanced Research, https://doi.org/10.1016/j.jare.2025.01.052

the genetic variations within the germplasm. To address these challenges, researchers have explored cost-effective alternatives to encapsulate species diversity in a single assembly known as a pangenome. This involves reducing the volume of genome sequences while focusing on genetic variations. With the growing prominence of the pan-genome concept in plant genomics, several software tools have emerged to facilitate pangenome construction.

This review sheds light on developing and utilizing software tools tailored for constructing pan-genomes in plants. It also discusses representation formats suitable for downstream analyses, offering valuable insights into the genetic landscape and evolutionary dynamics of plant species. In summary, this review underscores the significance of pan-genome construction and representation formats in resolving the genetic architecture of plants, particularly those with complex genomes. It provides a comprehensive overview of recent advancements, aiding in exploring and understanding plant genetic diversity.

Introduction

The single most important technology that has transformed genomics in the past decades is high-throughput DNA sequencing enabling cheap, fast and comprehensive generation of very large datasets on genomes. The availability of such huge volume of sequence data has been instrumental in the advancement of genomic studies for a wide range species including economically important crop.

Advanced genome sequencing systems incorporating Oxford Nanopore Technologies (ONT) and PacBio platforms alongside optical mapping and Hi-C technology have fundamentally transformed research in plant pangenomics. Further improvements in assembly technologies provide advanced tools for resolving complex plant genomes that bring together large amounts of repetitive DNA sequences and heterogenous structural variations that previous sequencing approaches could not handle. While initial efforts focused on generating high-quality reference genome assemblies for various plant species, such as maize [1], sorghum [2], soybean [3], potato [4], barley [5], chickpea [6], pigeonpea [7] and huge hexaploidy genome of bread wheat [8,9]. Recent studies have revealed the limitations of relying solely on a single reference genome to capture the extensive genetic diversity present within a species [10]. The systems developed by ONT and PacBio constitute leading technologies that generate long-read sequence information. ONT provides live sequencing operations through its system to generate DNA reads measuring between several kilobases. The direct measurement of DNA molecules using this technology simplifies sample preparation requirements while expanding genomic analysis capabilities for both large and complex genomes. PacBio's premier long-read sequencing performance using SMRT (Single Molecule Real-Time) technology allows precise identification of ambiguous genomic regions including repetitive elements together with structural variants. Highly accurate reads from this technology are fundamental for reliable genome assembly and pangenome research outputs. These technologies have transformed plant pangenomic research through their ability to sequence complete genetic information from several different species members. The sequencing genomes provide information about both genetic diversity along with structural components and uncommon gene variants which help understand plant survival abilities. Our understanding of plant evolution together with environmental adaptation benefits from comprehensive genome variation detection. The integration of advanced sequencing technologies such as PacBio, HiFi, Oxford Nanopore, along with specialized assemblers like HiCanu [11], Falcon [12], and Hifiasm [13], has significantly impacted plant pangenomics research.

The technique of optical genome mapping alongside long-read sequencing serves as an essential tool for visualizing plant genome spatial arrangements. The technique creates genomic schematic maps with high-definition detail through visual observation of solitary DNA strands. This method maintains exceptional value when it functions to resolve gaps that exist in assembled genomes when sequencing data does not provide clear information. The Hi-C technology demonstrates extraordinary usefulness by giving scientists insights into chromosomal layer structures beyond previous knowledge. The Hi-C technique detects genomic loci spatial relationships to develop long-range genome interaction maps. Hi-C delivers significant value for identifying complex genomic regions containing repetitive elements along with large-scale structure variants. By capturing chromatin conformation Hi-C has delivered crucial information that helped reveal the functional characteristics of plant genomes alongside their pangenomic structures.Using these technologies, researchers have been able to dissect the complex genetic architectures of plant species containing brand-new genomic structures and evolved evolutionary processes driving fine-scale diversity in plants. In the last few years these new such sequencing technologies have been deployed in plant pangenome studies and played a major role driving discoveries with regards to important biological processes as well agricultural research providing tools for understanding both crop traits and mechanisams of adaptation.

The concept of a pan-genome, first introduced by [14], offers a complete representation of the entire genomic collection of a given species. A pan-genome is defined as the non-redundant collection of all DNA sequences present across all individuals within a species. This holistic approach to genomic representation has gained significant power in plant genomics, with pan-genomes being constructed for numerous crop species, including maize [15], soybean [16], rice [17], sorghum [18,19], chickpea [20,21], and wheat [22,23].

Pan-genome analyses provide insights into the core genome, comprising the global genes across all individuals, and the dispensable or accessory genome, comprising genes specific to a subset of individuals. This distinction is crucial, as core genes are typically associated with essential biological functions and phenotypic traits, while accessory genes often contribute to specific adaptations and environmental responses [24]. By capturing this extensive genetic diversity, pan-genomes offer a powerful resource for understanding the evolutionary, functional, and phenotypic consequences of genomic variations within a species.

Moreover, pan-genomes have emerged as vital tools for crop improvement efforts, enabling the identification of genetic variants associated with desirable agronomic traits, such as yield, drought tolerance, and disease resistance. By leveraging the comprehensive genomic information provided by pan-genomes, researchers can focus on specific variants and develop molecular markers for marker-assisted selection or genomic prediction models, accelerating the breeding process and enhancing crop productivity.

This review provides a comprehensive overview of pan-genome construction approaches, data representation formats, and visualization tools, highlighting their applications in plant genomics

and crop improvement. Additionally, it explores the potential of pan-genomes to revolutionize genomics-assisted breeding strategies, ultimately contributing to the development of improved crop varieties to address global food security challenges.

Plant pan-genome analysis

Next-generation sequencing technologies have transformed how researchers study genetic variations across crop species through developments in NGS technology (Fig. 1). The ongoing development of sequencing methods has resolved previous challenges with short read lengths and high error rates and non-uniform data coverage to facilitate precise genome investigation [25]. The pan-genome data structure enables storage of crop species or population genomic sequences which function as a reference framework to describe genomic collections across the pan-genome. Pan-genome models let researchers study complete species-wide genetic diversity through their ability to detect genomic variations that exist between individual specimens.

Constructing a pan-genome necessitates the availability of a complete set of haplotype-resolved genetic variations. Significant progress has been made in this regard through various HapMap projects, which aim to capture linkage information for species such as cassava [26], maize [27], rice (https://www.ncgr.ac.cn/RiceHap3/), and *Cajanus*spp. [28]. However, despite these advancements, sequencing reads often lack sufficient length to assemble all repeat structures, necessitating the integration of complementary technologies like array comparative genomic hybridization (aCGH), synthetic long reads, and high-throughput optical genome mapping to detect larger-scale structural variations (SVs) [1,29,30].

Advanced sequencing methods present opportunities to integrate additional dimensional data including transcriptome profiling and DNA-protein interactions with epigenetic data for pan-genome investigations. Today's resequencing studies use whole-genome sequencing technologies to dive into genomic variations among different genotypes including single nucleotide polymorphisms (SNPs), insertions/deletions (indels) and large chromosomal structural variations. A comprehensive understand-

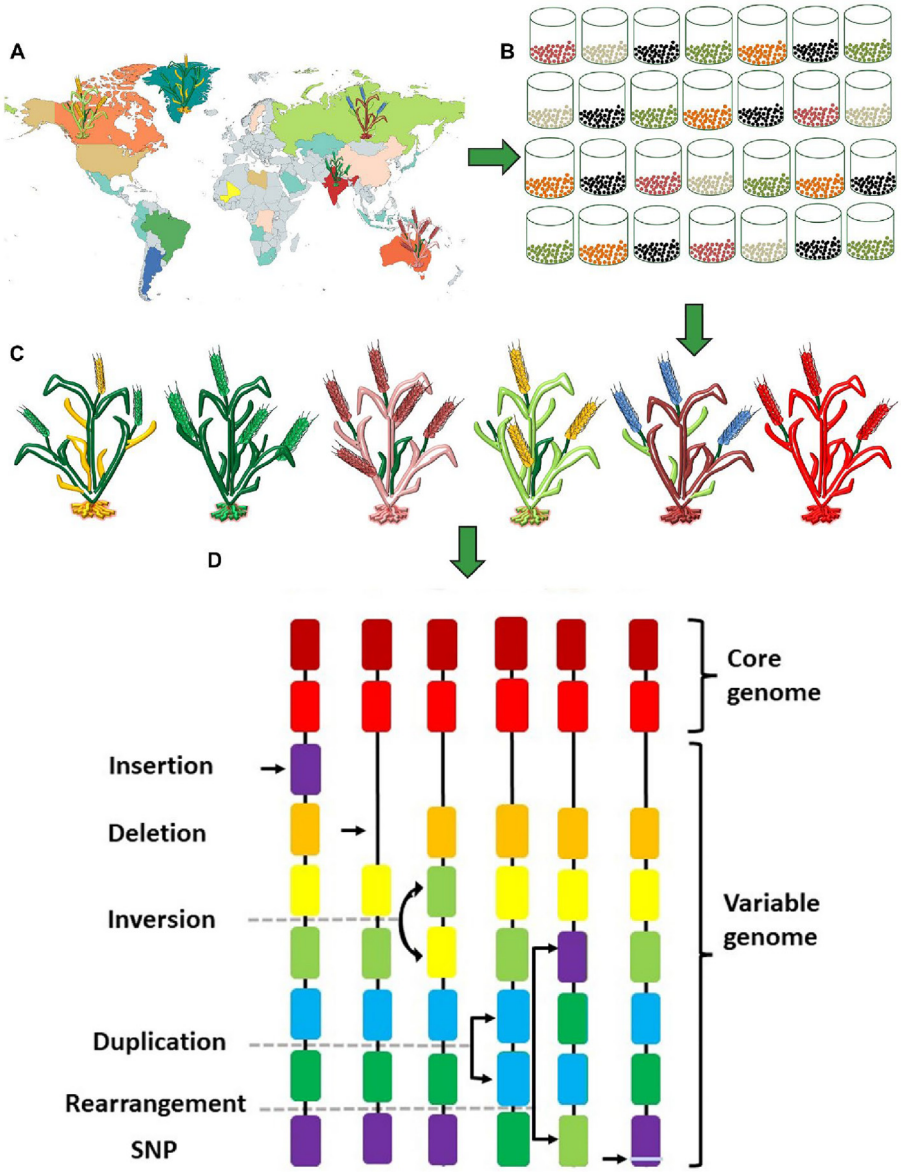


Fig. 1. Constructing pangenome from diverse global accessions: A) Representative genotypes are chosen from genetically diverse global accessions; B) the accessions are preserved; C) cultivated to extract the genetic material, and; D) analysed to construct the pangenome through assembling the high-quality genomes and identifying the variations including both core and variable sequences of a species.

ing of these genetic and genomic variants helps discover hazardous mutations and reveal plant domestication patterns and agricultural enhancement approaches [31].

The characterization of species becomes possible through pan-genome analyses which partition natural populations into core genome elements found across all individuals and overall pan-genome dimensions that include all genes or gene families represented within each population. Genomic and gene family analyses function at various levels according to the research design [32]. Essential biological processes along with major phenotypic traits constitute the core genome which exists alongside the accessory genome that includes genes able to adapt to environmental conditions and influence trait variability [33].

The research shows that core genes represent significant gene proportions across plant species where wheat contains 64 % and rice contains 89 % of total genes. When more genomes are studied variable genes linked to environmental adaptations become more prevalent throughout the core genome components.

Research on Brassica napus and wheat alongside sorghum and chickpea along with soybean has demonstrated that the pan-genome grows with additional genome sequencing, thus confirming open pan-genome organization is the dominant structural pattern in living plant species. The detection of additional genetic elements highlights the need for pan-genome methodologies which effectively capture a species' maximal genetic diversity so researchers can study phenotype-genotype interactions more thoroughly and accelerate crop enhancement programs.

How is a pan-genome assembled?

The assembly of a pan-genome can be approached through either supervised or unsupervised (*de novo*) methods, depending on the availability of a reference genome. In the supervised approach, sequence reads from each cultivar are mapped to the reference genome, and only the unmapped reads are iteratively assembled. Conversely, the unsupervised approach does not require a reference genome, and the assembly process proceeds entirely *de novo* (Fig. 2) [32,34].

Several bioinformatics tools have been developed to facilitate pan-genome construction from sequence datasets or assembled genomes. PANSEQ [35] and PGAP [36] enable the identification of novel genomic regions and the determination of additional features, such as single nucleotide polymorphisms (SNPs), core genome sequences, and accessory genome sequences. PANTOOLS [37], designed to handle large and complex genomes, can detect and annotate homologous regions within pan-genomic data.

In addition to collecting unique sequences, pan-genome assembly tools also consider minor variants between individuals within a crop species. These variants, ranging from SNPs and indels to larger structural variations, are represented as "bubbles" within a graph structure ordered by reference genome coordinates. Tools like GENOMEMAPPER [38], PANVC [39], GRAMTOOLS [40], GraphGenome Pipeline [41], PGGB [42], cactus [43], and VG [44] employ graph-based approaches to build such pan-genome representations.

The concept of the small variant graph extends further to genome assembly-level graphs. *De Bruijn* graph-based assemblers, such as Cortex [45], SplitMEM [46], TwoPaCo [47], and Bifrost [48], can adapt to pan-genome construction by assigning colours (representing specific biosamples) to nodes or unitigs. These coloured *de Bruijn* graphs enable population-scale analyses and facilitate the identification of sample-specific variations.

Furthermore, graph-based pan-genome data structures can be indexed to support efficient random access to elements and features within the graph, enhancing the utility of these resources for downstream analyses [49].

It is important to note that pan-genome assembly approaches can also incorporate transcriptomic data, providing an additional layer of information complementary to whole-genome sequences. Pan-transcriptome analyses capture partial genome information by cataloging the gene-level sequences present in each individual of a species. This approach enables the exploration of presence/absence variations (PAVs) and gene expression patterns, which can be integrated with genome-based pan-genome assemblies to enhance the resolution and accuracy of genetic variation detection.

Overall, the assembly of pan-genomes involves a diverse array of computational tools and methodologies, tailored to accommo-

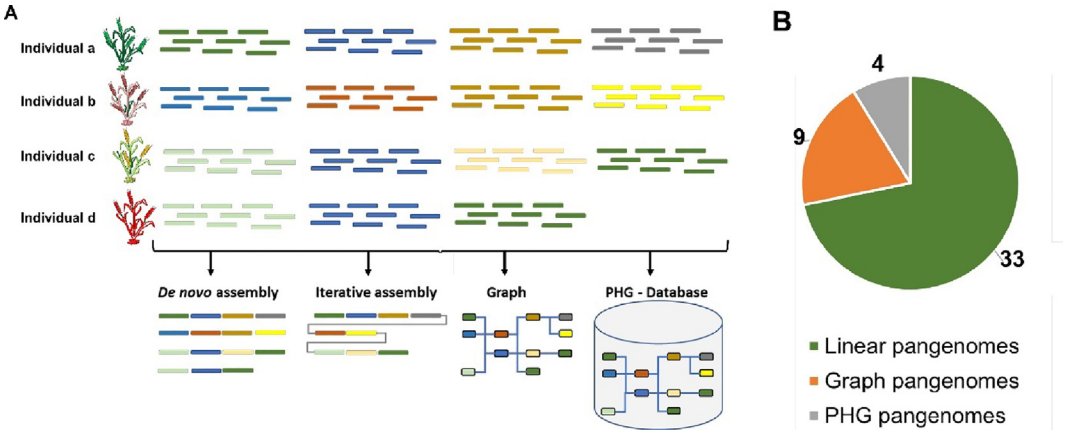


Fig. 2. Pangenome assembly methods and format: A) Schema showing pangenome assembly methods. Sequence reads from individual genomes assembled using the *de novo* method (each color indicates gene for each individual) and compared to define the core and variable regions. In the iterative assembly, one of the genomes de novo was assembled and used as a reference for assembling the remaining genomes' unique sequences. Graph pangenome assembly represents the genes/sequences as interconnected nodes and each path represents the genome. In the PHG (practical haplotype graph), the nodes represent the reference ranges sequences and the graph stored in the database. B) The number of plant' pangenomes assembled in each format. The linear format constitutes the major proportion of overall pangenome assemblies, followed by the graph and PHG format of assemblies.

date the inherent complexities of genomic data and the specific requirements of the target species or population under investigation. These approaches collectively aim to comprehensively capture and represent the genomic diversity within a species, laying the foundation for more advanced analyses and applications in crop improvement and breeding programs.

How is a pan-genome represented?

The representation of a pan-genome can take various forms, each with its own strengths and limitations. The three primary formats for pan-genome data are: classical linear draft sequence format, assembly graphs, and practical haplotype graph database (PHG). Traditionally, the pangenomes are stored in the linear structure as collections of sequences in FASTA format. The variants in this linear format are stored in VCF format (small/structural variants). Whereas for graphical format pangenomes are stored in Graphical Fragment Assembly format, (GFAv1) (https://github.com/GFA-spec/GFA-spec/blob/master/GFA1.md) or Graph Alignment Format (GAF) (https://github.com/lh3/gfatools/blob/master/doc/rGFA.md#the-graph-alignment-format-gaf). The genomic features on the pangenome assembly are represented in a linear format (expecting the common co-ordinate system between physical and genetic position) compared to the other two forms (graph and PHG) (Fig. 2B). The read alignments on the graph pangenomes are stored in GAM format [44] and supported by tools such as VG, and GraphAligner [50].

Linear sequence format pangenomes

A linear string of reference sequence base after the base in the pan-genome is a classical representation format, a standard FASTA format allowing visualization in two-dimensional genome browsers. In this representation, a novel sequence is identified from the individual genome sequence and either appended to the end of an existing reference sequence or inserted between prior known sequences. This format maintains the rearranged linear sequence bases with unique co-ordinate positions. The genetic variations between the cultivars, which could be as small as SNPs, insertions and deletions, or copy number variations, to as large as chromosomal rearrangements (deletions, duplications, inversions, and translocations), can report only one version of variation following the co-ordinate system (Fig. 1). To represent such a genome in a single linear format necessarily removes variations or finds an additional way of reporting such variations, and the co-ordinates system represents one of the genomic reference individual / parent. However, this format faces challenges in capturing the species-wide large population's specific features, like novel sequence, variations, similarity or functional content, so there is a need to address these challenges to represent species-wide information with proper genomic co-ordinate systems.

A FASTA format is a standard text-based format for representing nucleotide sequences in a linear format. The first line of each sequence starts with the '>' symbol, followed by the sequence identifier (id), and the second line contains the actual series of sequence base characters. Many such pangenome assemblies have been developed recently for small genome crop *A. thaliana* to complex genomic structures like wheat (Table 1).

Eg:

>scaffold1

CGACGAACA

>NC_003071.7:19472573-19474387 *Arabidopsis thaliana* chromosome 2 sequence

TGATTTTCTAAAAGTAGAAGAAAATAAGTG CAGTCCATAAAATAAAA TCCTATAAAAATGTTAAAACTAGATTCTTTTTTAAAAAACTAAAATTT GCT GCAGACATCTAAAATTTTCGAAAATGATTG GGTGGCTAAGA

Graph format pangenomes

A graph-based pan-genome is an alternative format addressing the above issues with the linear format. The sequence graph serves to collapse the similar sequence into a single unique data structure that is still representative of the full set [99]. A *de Bruijn* graphs-based genome assembly is a popular graph representation in which each node represents a *k-mer,* and the edge represents an overlap of *k-1* bases between from and to nodes. A direct walk following the node labels can be interpreted as a DNA sequence [100]. A graph is bidirectional when it represents both strands of DNA and the inversions between them. A graph could also represent a pan-genome of multiple individuals capable of capturing all sequences and variations between individuals [99]. A genome graph representing the genome of a species (represents the whole genome relationship) will grow with genome information as more data on that species becomes available. Such a graph imposed with the linear co-ordinate system by constructing a linear ordering of the nodes can describe the pan-genome [101]. Based on the topological relationships between each individual graph, it is possible to construct a compressed graph format as implemented in a few software like splitMEM and VG, to construct bacterial and human pan-genome [69]. This is similar to the earlier demonstrated compact representation (splicing graph) for a collection of splicing variants [102] and its application was also adapted to transcriptome data [103], highlighting the importance of graph. The tools available to construct such representations (like cortex) and calling variants through assembly (like platypus and vg: Variant Graph) and newer ones are upcoming. A graph can be a solution for a single diploid genome, addressing the above-mentioned linear graph limitations. It can also be used to represent the genomes of multiple individuals, capturing all variation between them [104].

Additionally, a graph could be a coloured path representing a specific individual, where the path of the graph is annotated. Such a *de Brujin* graph has been implemented in software like cortex and platypus [45]. For graph representation, currently, three file formats (FASTG, GBZ, and GFA) have been developed and also implemented by a few assemblers, like ALLPATHS_LG and SPAdes, which produce fastg format, and ABYSS produces GFA format. A recently developed GBZ file format is a path-based format in which the sequences are the objects connected with the edges. It is a compressed format with a specialized C++ library developed for creating and reading the compressed graph file [105], although, the GBZ was not designed for assembly graphs. The available graph representation of a genome would allow mapping reads corresponding to variants available in the graph. The binary format graph, the 'vg' graph, has been developed to store sequence and variant information and interchange the graph format [44].

FastG

FastG was the first format introduced (as FASTG) in 2012, which is an extension of the fasta format. The format mainly differed by representing edges as sequences, complicating the data operations (https://lh3.github.io/2014/07/19/a-proposal-of-the-grapical-fragment-assembly-format). A fastg format requires 'begin' and 'end' lines with each scaffold line starting with '>' symbol. The below assembly example has two scaffolds named 'scaffold1′ and 'scaffold2′ in the fastg format.

Eg:

Fastg graph format

#FASTG:begin

#FASTG:version = 1.0:assembly_name='example';

>scaffold1:scaffold1;

ACGANNNNN[5:gap:size=(5,4..6)]CATGGC

Table 1
The plant pangenomes published in linear format, graph and PHG format.

Species	Domestication status	Ploidy	Number of accessions	Reference
Linear format				
Arabidopsis thaliana	Crop	Diploid	69	[58]
Brachypodium distachyon	Wild	Diploid	54	[90]
Brassica napus	Crop	Tetraploid	53	[52]
B. napus	Crop	Tetraploid	50	[93]
B. napus	Crop	Tetraploid	9	[68]
B. oleracea	Crop	Diploid	10	[32]
Banana (*Musa and Ensete*)	Crop, hybrids	Triploid	15	[72]
B.napus, rapa, oleracea	Crop	Diploid, diploid, amphidiploid	87, 77 and 79	[97]
B.rapa	Crop	Diploid	3	[56]
Pepper (*Capsicum*)	Crop	Diploid	383	[62]
Chickpea (*Cicer arietinum*)	Crop	Diploid	3,366	[20]
Cowpea (*Vigna unguiculata*)	Crop	Diploid	6	https://doi.org/10.1101/2022.08.22.504811
Cowpea (Vigna unguiculata)	Crop	Diploid	6	[57]
Eggplant (*Solanum melongena*)	Crop	Diploid	23	[86]
Soybean (*Glycine soja*)	Wild	Tetraploid	7	[16]
Sunflower (*Helianthus annuus*)	Crop	Diploid	493	[51]
Walnut (*Juglan ssp.*)	Wild	Diploid	6	[77]
Medicago truncatula	Wild	Diploid	15	[88]
Melon (*Cucumis melo*)	Wild, landrace	Diploid	2	[63]
Mung bean (*Vigna radiata*)	Crop	Diploid	217	[54]
Oryza sativa	Crop	Diploid	3	[17]
O. sativa (indica/japonica)	Crop	Diploid	1,483	[30]
O. sativa	Crop	Diploid	3010	[85]
O. sativa/ O. rufipogon	Crop	Diploid	67	[79]
Pea (*Pisum sativum*)	Wild, Crop	Diploid	118	[84]
Pecan (*Carya illinoinensis*)	Tree	Diploid	4	[64]
Pigeon pea (*Cajanus cajan*)	Crop	Diploid	89	[80]
Populus	Tree	Diploid	19	[71]
Populus	Wild	Diploid	7	[74]
Potato (*Solanum tuberosum*)	Wild, Crop	Diploid	44	[67]
Sesame (*Sesamum indicum*)	Crop	Diploid	5	[82]
Tomato (*Solanum lycopersicum*)	Crop	Diploid	725	[91]
Sorghum (*Sorghum bicolor*)	Crop	Diploid	354	[19]
Bread wheat (*Triticum aestivum*)	Crop	Hexaploid	19	[22]
White lupin (*Lupinus albus*)	Crop	Diploid	39	[53]
Maize (*Zea mays*)	Crop	Tetraploid	503	[15]
Maize (*Zea mays*)	Crop	Diploid	721	[89]
Graph format				
Arabidopsis thaliana	Crop	Diploid	32	[61]
Broomcorn millet (*Panicum miliaceum*)	Wild	Diploid	32	[96]
Chickpea (*Cicer arietinum*)	Wild	Diploid	8	[21]
Cucumber (*Cucumis sativus*)	Crop	Diploid	11	[59]
Soybean (*G.* max)	Crop	Diploid	29	[55]
Grapevine *(Vitis vinifera)*	Wild	Diploid	9	[94]
Melon (*Cucumis melo*)	Crop	Diploid	3	[76]
Rice (*O. sativa*)	Crop	Diploid	33	[73]
Pepper (*Capsicum*)	Crop	Diploid	3	[62]
Radish (*Raphanus sativus*)	Crop	Diploid	11	[81]
Rice (*O. sativa*)	Crop	Diploid	251	[54]
Rice (*O. sativa*)	Crop	Diploid	12	[69]
Sorghum (*Sorghum bicolor*)	Crop/Wild	Diploid	16	[78]
Tea (*Camellia sinensis*)	Elit cultivars	Diploid	22	[95]
Tomato (*Solanum lycopersicon*)	Wild/Cultivated	Diploid	11	[60]
Tomato (*Solanum lycospersicum*)	Crop	Diploid	838	[87]
Grapevine (*Vitis vinifera* ssp. *Vinifera*)	Crop	Diploid	29	[66]
Brassica genomes	Crop	Tetraploid	41	[70]
Lattuce (*Lactuca sativa*)	Crop	Diploid	474	[83]
Barley (*Hordeum vulgare*)	Wild/Cultivated	Diploid	76	[92]
PHG format				
Cassava (*Manihot esculenta*)	Crop	Diploid	241	[65]
Maize (*Zea mays*)	Crop	Diploid	27	[75]
Sorghum (*Sorghum bicolor*)	Crop	Diploid	398	[18]
Wheat (*Triticum aestivum*)	Crop	Hexaploid	65	[98]

```
>scaffold2;
CGA[1:alt:allele|A,T]CGATCA
#FASTG:end;
Linear format (fasta)

>scaffold1
ACGANNNNNCATGGC

>scaffold2
CGACGATCA
```

Graphical format assembly (GFA)

Alternative to FastG, gfa is another format of a graph that is represented as a tab-delimited field like header (H), segment (S), link (L), containment (C) and path (P). GFA format was intro-

duced in 2014, compatible with *de Bruijn* and *string* graphs. More specifications of this format are available at https://github.com/GFA-spec/GFA-spec, and the tools and API listed are available at the same link. Fig. 5. A and B below are the simple gfa format graph assembly with a string in reverse complement and a base mismatch.

Eg:
Fastg graph format
#FASTG:begin
#FASTG:version=1.0:assembly_name='example';
>scaffold1:scaffold1;
CGACGA[1:alt:allele|A,T]CA
>scaffold2
ACGANNNNN[5:gap:size=(5,4..6)]CATGGC
#FASTG:end;
Linear format (fasta)
>scaffold1
CGACGA**A**CA
>scaffold2
ACGANNNNNCATGGC

Practical haplotype graph (PHG)

Compared to the genome assembly graphs, the haplotype graph is a collection of nodes and edges for the sequence within the organism inherited from a single parent. The PHG is built from a subset of sequences (conserved sequences with genetic variations) called reference ranges. Such sequence ranges are represented as graph node, and the nodes are connected with edges, which do not contain the sequence range but indicate the two haplotypes were together in a particular individual [106]. The PHG represents the sequence of haplotypes instead of the complete nucleotide sequences and stores the data in the relational database format. For example, the existing genomic resources of the breeding program founder line (whole genome sequence data or whole genome assemblies) are loaded into a graph database. Such a database supports genomic analysis such as imputation of low sequence coverage (as low as 0.01x coverage) of individuals in the breeding population achieved based on consensus haplotypes derived from the graph database (https://bitbucket.org/bucklerlab/rphg/wiki/Home). The input sequence can be a whole genome sequence, a reduced representation sequence, or SNPs called from population data. The PHG database also stores an additional layer of genomic features with genic and intergenic haplotypes, assisting in annotating the haplotypes. The data is stored in the compact format of haplotypes in the form of an imputed path through the graph, resulting in a very compact storage of the graph path list of haplotypes for many genotypes in a relational database. Thus, the organized pangenome is finally formed by storing the node and edge relationship as the path for each individual.

The first step of the PHG database is to assign the reference ranges in user-defined groups (e.g., gene and non-gene co-ordinates) followed by uploading to the database with haplotypes from other individuals [107] (Fig. 3). The database can be updated with either consensus haplotypes built from aligned genome assemblies or variants from WGS/reduced representation (GBS) data [107]. The PHG database has been implemented in sorghum [18], maize [75], wheat [98], and cassava [65] using the SNPs from diverse accessions WGS data and imputed with GBS/skim sequence data from inbred lines (Table 1). The PHG is deployed as a Docker image and available at https://hub.docker.com/r/maizegenetics/phg. Alternatively, a statistical programming language R package for PHG is available at https://bitbucket.org/bucklerlab/rphg/wiki/Home. HaploCart, working on the Bayesian inference principle is available in command-line and web interfaces [108].

Formats comparison (linear vs graph)

The choice between linear and graph-based representations of pan-genomes has significant implications for capturing and analyzing genomic diversity within a species. Each format presents distinct advantages and limitations, shaping the types of analyses and applications that can be effectively performed.

The linear sequence format, typically represented as FASTA files, has been the classical approach for representing genomic sequences. The first line of each sequence starts with the '>' symbol, followed by the sequence identifier (id), and the second line contains the actual series of sequence base characters (Fig. 4). Many such pangenome assemblies have been developed recently for small genome crop *A. thaliana* (1001 genome project in Arabidopsis) to complex genomic structures like wheat [23] (Table 1).

Linear format offers several advantages, such as maintaining a straightforward coordinate system, enabling easy mapping of genomic features such as annotations, variants, and structural variations. The position of each base and the distance between bases are readily interpretable. Numerous bioinformatics tools and pipelines have been developed over decades to operate on linear sequence data, ensuring widespread compatibility and ease of integration with existing workflows (Compatibility). FASTA files are human-readable and require minimal computational resources, making them accessible and easy to manipulate.

However, linear representations also face significant limitations when it comes to capturing the full extent of genomic complexity within a species. The strictly linear nature of sequence representation forces assemblers to make arbitrary choices when encountering ambiguities, such as uncertain bases, single nucleotide polymorphisms (SNPs), or tandem repeats. This can lead to the loss of genetic information or the introduction of errors (Loss of information). Linear formats struggle to accurately represent structural variations, inversions, and complex haplotype relationships, as they can only accommodate a single representation of variations at a given coordinate (Inability to represent variations). As more genomes are added to a pan-genome, the linear representation becomes increasingly fragmented, reducing its utility and complicating downstream analyses (Limited scalability).

Graph-based representations, such as assembly graphs and variation graphs, offer an alternative approach that addresses many of the limitations of linear formats. Plant graph construction tools play a crucial role in capturing the complex genetic variations present in plant genomes, leading to the development of plant pangenomes. Recent advancements in graph construction tools have enabled researchers to construct comprehensive plant graph pangenomes, offering a more nuanced understanding of genetic diversity and evolutionary relationships within plant species.

Tools such as Cortex [45] and SplitMEM [46] have traditionally been used for graph construction in genomic studies, but their applicability in plant genomics may be limited due to the unique complexities of plant genomes. However, newer tools and methodologies tailored for pangenomes, such as VG (Variation Graph Toolkit) [44], PGGB [42], ODGI [109], cactus [43], and GraphAligner [50], have emerged to address these challenges more effectively. The tools like PGGB, cactus, and Minigraph-Cactus are alignment based graph generating tools applied for vertebrate and human pangenome studies [110,111].

Graph structures can faithfully represent the non-linear complexities of genomes, including ambiguities, repeats, inversions, and structural variations, without the need for arbitrary decisions or loss of information (Preservation of complexity). Graph representations can accommodate variations across multiple individuals within a population, enabling population-scale analyses and the

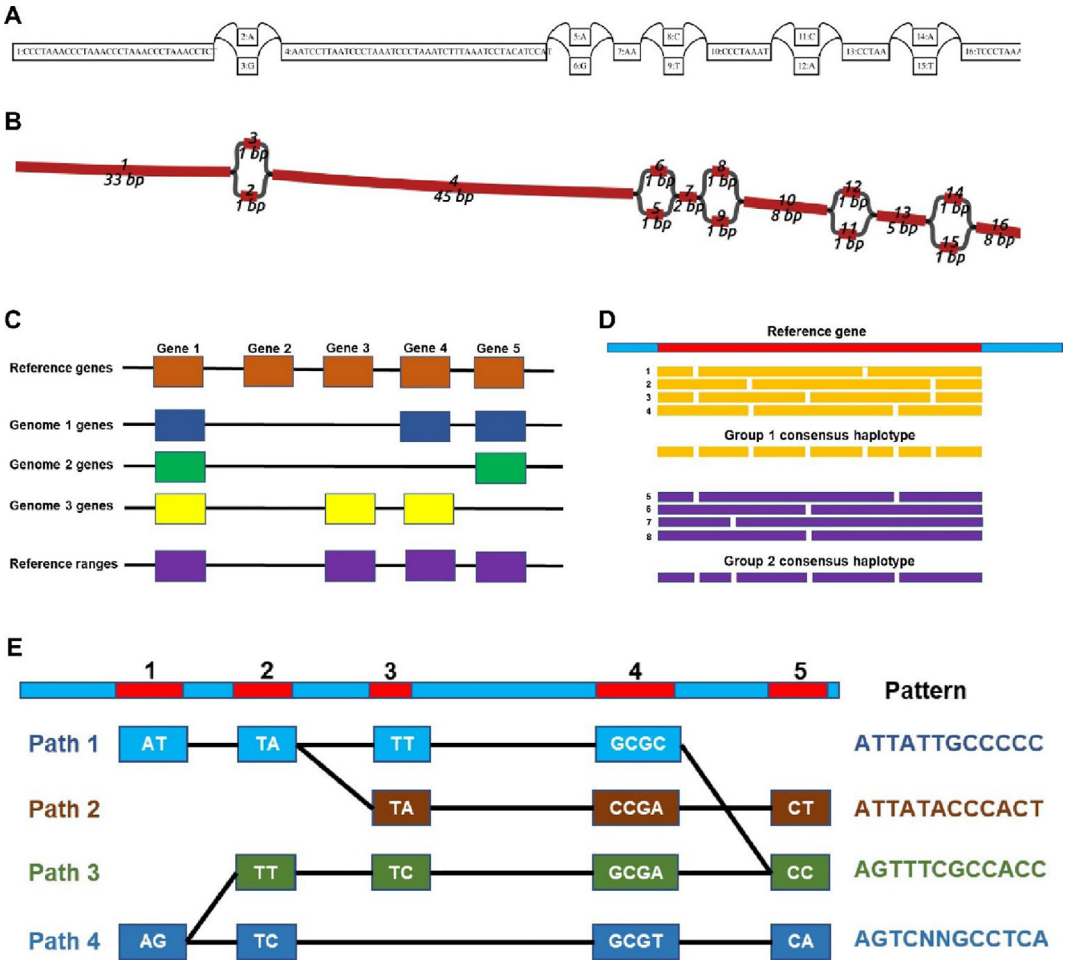


Fig. 3. A graph-based visualization in: A) dot format viewer; B) graph format in Bandage The PHG database construction includes; C) identification of reference ranges/ intervals sequences (conserved regions); D) identify the haplotypes and calling consensus for each group (of a population) and storing in the database; E) map the sequence read of a query individual of a population and follow the path to find the haplotypes from the database.

identification of sample-specific variations (Population-scale analysis). As new genomes are added to the pan-genome, graph structures can dynamically incorporate and represent the additional variations, providing a scalable framework for capturing genomic diversity (Scalability).

However, graph-based representations also face challenges. Unlike linear sequences, graph representations often lack a straightforward coordinate system for mapping genomic features, complicating analyses and requiring the development of specialized tools and methodologies (lack of coordinate system). Graph structures can be computationally intensive to construct, manipulate, and analyze, particularly for large and complex genomes (computational complexity). Effectively visualizing and interpreting the intricate patterns and relationships within graph-based pan-genomes can be challenging, requiring the development of specialized visualization tools (visualization challenges) listed in the Table 2.

Future developments in plant graph pangenomes may involve incorporating diverse genomic and epigenomic data (integration of multi-omics data) into plant graph pangenomes can provide a more holistic view of plant genomes. Advancements in graph algorithms and tools tailored for plant genomics (development of efficient graph algorithms) can enhance the accuracy and scalability of plant graph pangenome construction. Utilizing plant graph pangenomes for marker-assisted breeding and trait mapping can accelerate genetic improvement efforts in crops (application in breeding and crop improvement).

As pan-genome analyses continue to evolve, the choice between linear and graph-based representations will depend on the specific research objectives, the complexity of the target species, and the desired balance between comprehensiveness, computational efficiency, and interpretability.

Is it possible to toggle between the formats?

The linear sequence and graph-based representations of pan-genomes are not mutually exclusive, but rather complementary approaches that can be leveraged in a coordinated manner. As Iain MacCallum and David B. Jaffe (from Broad Institute of MIT and Harvard, Cambridge) indicated, while each format has its unique strengths and limitations , it is possible to transition between them, capitalizing on their respective advantages for different stages of analysis or specific applications (Fig. 4).

From linear to graph

Genome assembly tools, such as cloudSPAdes [147], ALLPATHS-LG [148], Cuttlefish 2 [149], and Minigraph-Cactus [111], typically employ a graph-based approach during the initial assembly process. These tools build assembly graphs by identifying overlapping sequence reads and representing them as nodes and edges. Subsequently, the optimal path through the graph is

Table 2
Plant pangenome visualization tools for linear and graph format assemblies.

Software	Available site	Reference
Linear format		
ABrowse (genome browser)	https://www.abrowse.org/	[116]
BasePlayer	https://github.com/rkataine/BasePlayer	[142]
Biodalliance	https://github.com/dasmoth/dalliance	[128]
Ensembl genome browser	https://useast.ensembl.org/Homo_sapiens/Location/View?r = 17:63992802-64038237	[143]
GBrowse 2	https://github.com/GMOD/GBrowse	[117]
GeneViTo	https://athina.biol.uoa.gr/bioinformatics/GENEVITO/	[126]
GenomeMaps	https://github.com/opencb/genome-maps	[125]
Gosling	https://gosling.js.org/	[113]
HiGlass	https://github.com/higlass/higlass	[145]
IGB	https://bioviz.org/	[123]
IGV	https://github.com/igvteam/igv	[121]
IGV.js	https://github.com/igvteam/igv.js/	[144]
JBrowse 2	https://jbrowse.org/jb2	[129]
Kero-BROWSE	https://kero.hgc.jp/examples/CLCL/hg38/index.html	[114]
NCBI Genome Data Viewer	https://www.ncbi.nlm.nih.gov/genome/gdv/	[140]
Nucleome browser	https://vis.nucleome.org/v1/main.html	[135]
pyGenomeTracks	https://github.com/deeptools/pyGenomeTracks	[127]
Tablet	https://ics.hutton.ac.uk/tablet/	[131]
Trackplot (python)	https://github.com/ygidtu/trackplot	[137]
Trackster	https://galaxyproject.org/learn/visualization/	[132]
UCSC genome browser	https://genome.ucsc.edu/	[146]
UTGB	https://utgenome.org/	[141]
Zenbu	https://fantom.gsc.riken.jp/zenbu/	[112]
Graph format		
AbySS-Explorer	https://github.com/bcgsc/ABySS-explorer	[134]
Assembly Graph Browser	https://www.github.com/almiheenko/AGB	[124]
Bandage	https://github.com/rrwick/Bandage	[119]
GfaViz	https://github.com/ggonnella/gfaviz	[133]
Icarus	https://bioinf.spbau.ru/icarus	[130]
IGV	https://igv.org/	[121]
MoMI-G	https://github.com/MoMI-G/MoMI-G	[139]
Panache	github.com/SouthGreenPlatform/panache	[122]
PanGraphViewer	https://github.com/TF-Chan-Lab/panGraphViewer	[136]
PGGB	https://github.com/pangenome/pggb	[42]
Ray Cloud Browser	https://deNovoAssembler.sf.Net/	[120]
SGTK	https://github.com/olga24912/SGTK.	[118]
VAG	https://ricegenomichjx.xiaomy.net/VAG/sequenceextraction.php	[115]
viralFlye	https://github.com/Dmitry-Antipov/viralFlye	[138]

selected to generate the final non-branching assembly in a linear contig sequence format.

As more genome sequences become available for a species, the linear representation can be extended to accommodate variations from additional individuals by introducing "bubbles" or branches within the graph structure. This process effectively transitions from a linear format to a graph-based representation, enabling the capture of population-level variations and structural complexities.

From graph to linear

Conversely, graph-based pan-genome representations can be linearized by exporting specific paths or haplotypes as linear sequences. This approach is particularly useful for integrating graph-based pan-genomes with existing bioinformatics pipelines and tools that operate on linear sequences.

For instance, in the construction of the wheat graph pan-genome, the gfatools gfa2bed utility was employed to linearize the graph representation, allowing the integration of genomic features and annotations from the linear coordinate system (https://doi.org/10.5281/zenodo.6085239).

In the case of Practical Haplotype Graphs (PHGs), the graph database can be queried with aligned sequence reads (e.g., from whole-genome sequencing or reduced representation sequencing) to extract linear haplotype sequences corresponding to specific individuals or accessions.

Hybrid approaches

In many cases, a hybrid approach that leverages the strengths of both linear and graph-based formats may be advantageous. Linear representations can serve as a familiar coordinate system for mapping genomic features, annotations, and small-scale variations, while graph structures can capture the broader genomic diversity, including structural variations, inversions, and complex haplotype relationships.

This hybrid approach allows researchers to seamlessly transition between formats, utilizing linear sequences for downstream analyses and feature mapping, while leveraging graph structures for comprehensive representation of pan-genomic diversity and population-scale analyses.

Ongoing bioinformatics developments create tools that enable researchers to switch between linear and graph-based analytical formats through data structures which bridge the two methods. Pan-genomic data analysis becomes more comprehensive through ongoing tool improvements which allow researchers to leverage both formats' therapeutic possibilities [150]. Effective format interoperability techniques will unlock the complete potential of pan-genome analyses to push forward crop breeding research

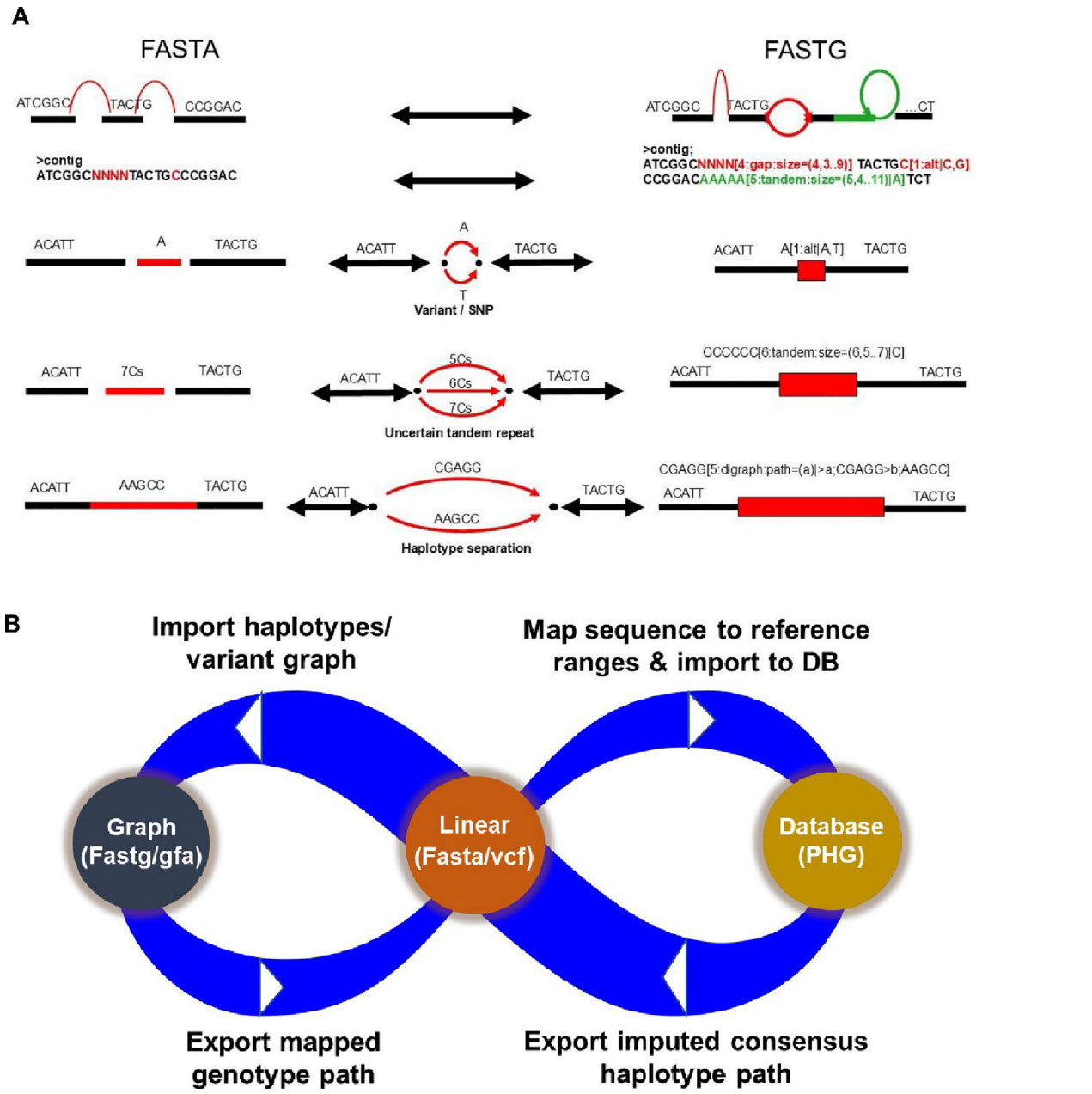


Fig. 4. Comparison of linear and graph formats in pangenome representation: A) The linear and graph format sequence comparison. The linear format (FASTA) force to choose a random base (in case of SNP variant), a single path for uncertain repeat and haplotype patterns at a sequence position, whereas graph format (FASTG) encodes and store the genome complexity; B) A pangenome can be represented in a linear format, graph and PHG can interchange with few additional steps. A linear format can be converted to a graph with identified haplotypes/variants and export genotypes into a linear format. Similarly, a list of haplotypes called on reference ranges which are based on the linear format can be imported into PHG database in graph format and can export the imputed consensus haplotype path back to a linear format.

and agricultural development through paramount insights into species genetics.

Visualization

Effective visualization is pivotal for interpreting and understanding the intricate relationships and patterns within pan-genomic data. While most visualization tools initially focused on linear reference genome structures, the increasing adoption of graph-based representations has necessitated the development of novel visualization approaches to capture the complexities inherent in these non-linear data structures (Table 2).

Linear genome visualizers: Adapting to pan-genomic representations

Traditional linear genome visualizers, such as GBrowse, JBrowse2, and Circos, have been adapted to accommodate linear pan-genome representations (Fig. 5). These tools have been employed in various studies, including the visualization of pan-genomes for species like *Brassica napus*, *Brassica oleracea*, and wheat. While effective for linear sequences, these tools may strug-

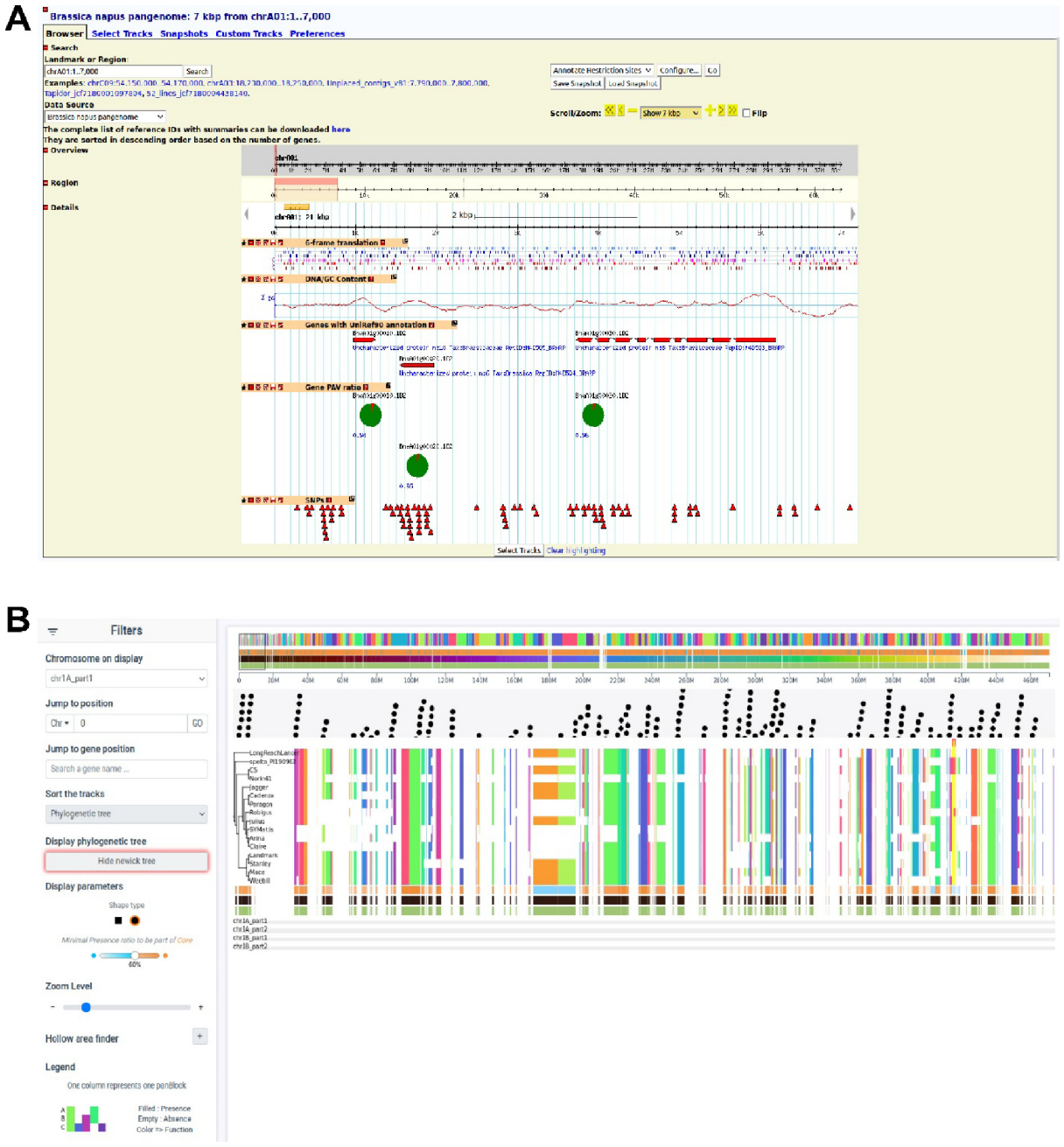


Fig. 5. Pangenome visualization of A) a linear format pangenome in Gbrowse (*Brassica napus* pangenome); B) Panache screenshot of wheat pangenome.

gle to accurately represent the intricate details and complexities present in graph-based pan-genomes.

Graph-based genome visualization: Capturing non-linear complexities

To address the challenges of visualizing graph-based pan-genomes, several specialized tools have been developed for assembly graphs and variant graphs. Bandage is a visualization tool designed specifically for assembly graphs, capable of displaying connections and patterns within the graph structure. ODGI (Open, Decentralized Genomic Research) is a command-line tool for visualizing and analyzing assembly graphs. The tool for visualizing and exploring genome assembly graphs in the Graphical Fragment Assembly (GFA) format is possible with GfaViz.

Additionally, tools like vg view (part of the Variant Graph tool suite) and the Sequence Tube Map have been developed to visualize variation graphs at different scales, ranging from individual variations to larger structural variations. More tools are listed in Table 2.

Network analysis and heatmap visualizations

Network analysis packages, such as igraph (available in Python, R, and C/C++), provide tools for visualizing and analyzing graph-based pan-genome data structures.

Moreover, heatmap visualizations have emerged as a powerful technique for representing shared genomic regions among individuals within a species. Tools like Panache can generate interactive web-based heatmaps, highlighting regions of similarity and divergence across different accessions or individuals.

Visualizing practical haplotype graphs (PHGs)

For representations like Practical Haplotype Graphs (PHGs), which focus on capturing haplotype variations within a pan-genome, specialized visualization approaches are required. While tools exist for visualizing the linear components of PHGs (e.g., conserved sequence ranges), visualizing the haplotype connectivity and relationships within the graph structure remains an active area of development.

Integrating multiple visualization approaches

As pan-genome analyses continue to advance, the development of effective visualization tools will be crucial for interpreting the intricate patterns of genomic diversity within species. By integrating multiple visualization approaches, ranging from linear genome browsers to graph-based representations and heatmaps, researchers will gain a comprehensive view of the genetic landscape, enabling deeper insights into the evolution, adaptation, and functional implications of genomic variations.

Pangenomes towards the crop improvement

The introduction of pangenomic approaches may transform crop improvement by revolutionizing how we understand and utilize genetic diversity within species. Unlike traditional methods that depend on a single reference genome, pangenomics captures the full range of genetic variations, including both core genes found in all individuals and accessory genes present in only some. This comprehensive view of a species' gene pool provides new opportunities for identifying genetic variants linked to valuable agronomic traits, making crop improvement more precise and effective.

Pangenomics for trait discovery

Pangenomics represents a powerful crop improvement method because it enables researchers to study the actual genetic variants which determine traits including yield productivity mixed with drought resilience and disease immunity. Pan-genome-wide single nucleotide polymorphisms (SNPs) along with presence/absence variations (PAVs) provide high-density molecular markers which enable researchers to run powerful genome-wide association studies (GWAS) and quantitative trait locus (QTL) mapping analyses (Fig. 6). Through these methods scientists can discover quantitative trait nucleotides (QTNs) that correspond with desired phenotypic features to develop useful markers and genomic prediction models [66].

Overcoming biases and capturing comprehensive genetic diversity

Pangenomics addresses the biases and limitations of using just one reference genome for genetic studies. It allows scientists to capture and study genetic variants that might be missing or underrepresented in a single reference genome. This leads to a more accurate and complete identification of genes linked to important traits. Additionally, pangenomics helps explore large genetic changes, such as copy number variations (CNVs) and gene presence/absence variations (PAVs), which significantly affect traits like leaf growth and disease resistance in crops like maize.

Unleashing the potential of evolutionary dynamics and functional implications

Pangenomics allows scientists to study the differences in gene content across various plant types, including wild relatives. This helps them understand how these differences evolved and their impact on plant functions. By pinpointing gene families or specific genes linked to beneficial traits, researchers can use this information to introduce or edit these genes, speeding up the creation of better crop varieties with improved performance and desirable characteristics.

Integrating pangenomics with advanced statistical models and machine learning

Combining pangenomic data with advanced statistical models and machine learning can greatly improve the accuracy of predicting genetic traits. By including structural variations, CNVs, and PAVs along with traditional SNP data in their models, researchers can get a fuller picture of the genetic makeup behind complex traits. This makes it easier to select and breed crops more accurately and efficiently, leading to better crop improvement programs.

Fueling future advancements in crop improvement

The advancement of sequencing technology and wider availability of pangenomic crop species resources will allow more extensive utilization of pangenomic information for crop improvement programs. Pangenomic data integration alongside modern techniques including genome editing combined with genomic selection and gene introgression enables scientists to develop climate-adapted crop varieties exceeding the current yield limits and containing essential nutritional components that serve food security and environmentally-friendly farming systems [31,151].

The future of modern agriculture depends heavily on pangenomic intervention to address genetic diversity needs and identify traits and create breeding precision methods while investigating genome-environment relationships and discovering new genetic material [150]. The wide-ranging genetic composition of crop species becomes accessible through pangenomics so it reshapes plant breeding approaches while fostering sustainable crop development suitable for evolving global agricultural requirements.

Conclusion and future perspectives

Recent technological advancements have revolutionized the field of pangenomics, enabling the comprehensive representation of genetic variation within species through pangenome assemblies. These innovative approaches encompass both linear and graphical models, supporting sophisticated algorithms for sequence read mapping, visualization, and association studies. While graph-based pangenomes exhibit the capability to effectively relate multiple sequences, the debate persists regarding whether they will supplant the traditional linear reference genomes. Linear references offer the advantage of maintaining coordinate systems, enhancing their utility across diverse applications.

Identifying dispensable genes or sequences throughout a species' complete germplasm is a vital component of pan-genome research. However, the field faces notable challenges stemming from the limitations of existing technologies and computational programs. These limitations encompass issues such as the accuracy of gene annotations, the complexity of analyzing large-scale genomic data, the computational resources required for comprehensive pan-genome studies, and the need for standardized methodologies

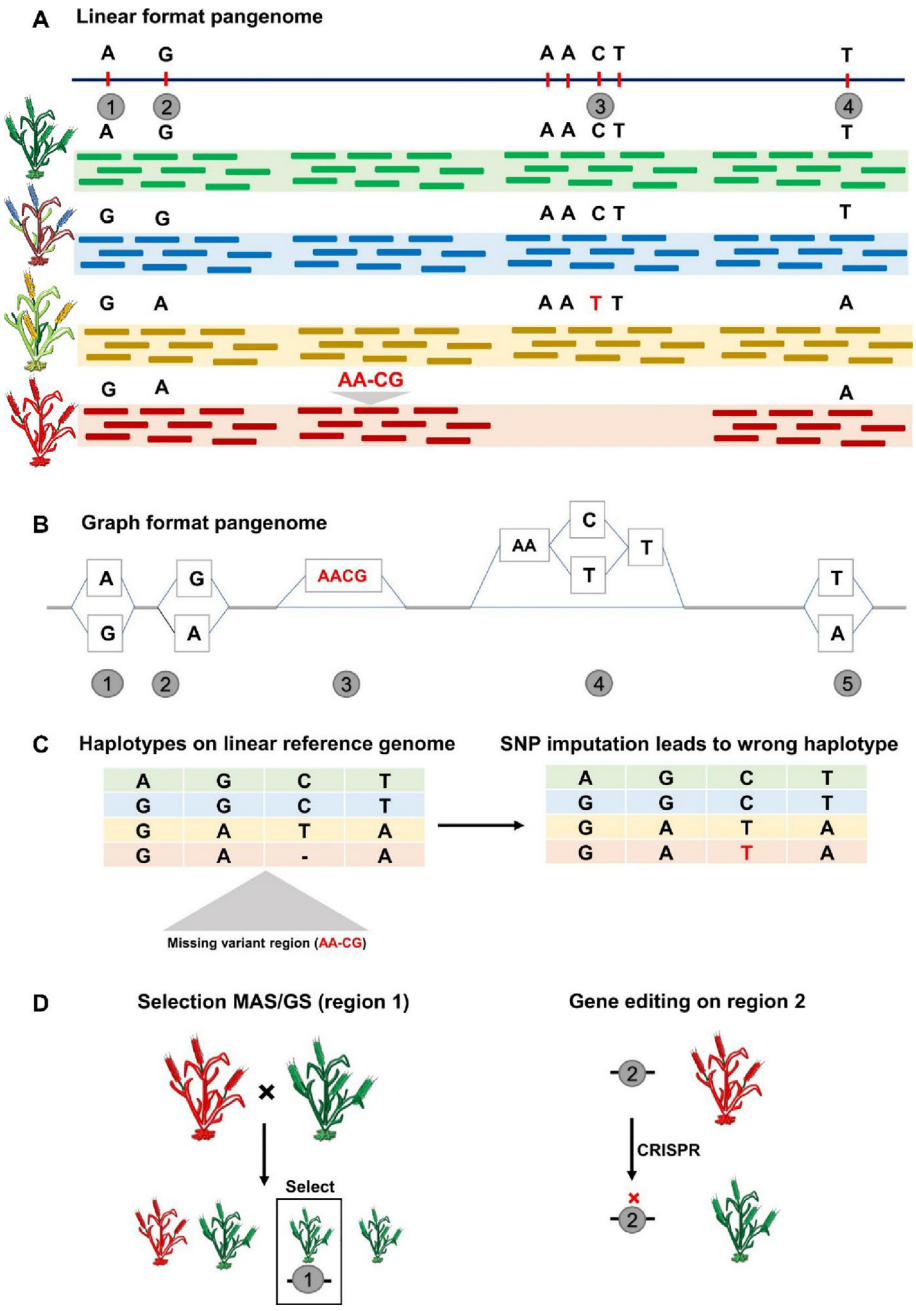


Fig. 6. The comparison of genetic variants in pangenome formats: A) The variation was ignored between the 2nd and 3rd variation; B) whereas in graph format the same missing variation was captured; C) at the downstream analysis the haplotypes were missing leading to a wrong pattern of haplotypes and; D) With more accurate genetic information, breeders can utilize it to identify variants involved in MAS/GS and make alterations to the genome through genome editing.

to ensure reproducibility and comparability across different studies. Overcoming these hurdles necessitates a detailed examination and refinement of current approaches to enhance the robustness and reliability of dispensable gene identification in pan-genome research.

Pangenome analysis offers valuable insights and tools that can significantly enhance crop breeding by facilitating the recovery of favorable genes lost in elite lines and integrating genome editing to guide future breeding strategies. Through pan-genome analysis, breeders can identify dispensable genes or sequences that are not present in all accessions but may confer beneficial traits under specific conditions. By investigating the functional roles of these dispensable genes, breeders can strategically reintroduce them into elite lines to enhance agronomic performance and resilience. The comprehensive understanding of dispensable genes provided by pan-genome analysis guides breeders in selecting and introgressing valuable genetic variants for trait improvement.

The genome editing method CRISPR-Cas9 creates precise and focused modification techniques to manipulate specific genes located in crop genomes. The utilization of pan-genome information helps breeders choose target genes for desired traits which they can modify using genome editing tools for novel alleles and deleterious mutation correction and gene expression level optimization in elite genetic lines. Through the use of this approach researchers can develop improved crop varieties with tailored benefits by accelerating breeding production and development cycles [151].

A critical challenge in pangenomics lies in addressing heterozygosity issues, where the presence of alternative alleles complicates

variant identification. Strategies must be developed to differentiate true single nucleotide polymorphisms (SNPs) from variants arising exclusively due to heterozygosity during pangenome construction. Furthermore, there is a growing interest in generating taxonomically stratified pangenomes to elucidate variable genomic regions distinguishing taxa at species or family levels. Concurrently, conserved genomic regions hold promise for marker development to classify species taxonomically. Looking ahead, the prospect of creating pangenomes at higher taxonomic levels, such as the genus or family, or even a unified pangenome for viridiplantae, emerges as a fascinating avenue for future research endeavors with profound implications for evolutionary studies and biodiversity conservation efforts.

Machine learning and artificial intelligence (ML/AI) promise to enhance formatting and haplotype graphing operation (HG) within plant pangenomes through forthcoming studies that show predictive capabilities for genomic analysis techniques. Genomic research benefits increasingly from ML/AI technologies which create streamlined data analysis solutions for enhanced annotation accuracy alongside genome assembly results. In the realm of pangenome analysis, ML/AI algorithms can be leveraged to enhance the formatting of complex genomic data and improve the construction of haplotype graphs. By developing ML models that can recognize patterns in genomic sequences and structural variations, researchers can optimize the representation of pangenome graphs and accurately capture genetic diversity within plant species.

Moreover, the application of ML/AI in pan-genomic research extends to transcriptome assembly and annotation. By incorporating ML algorithms trained on small RNAs/microRNAs data, researchers can improve the efficiency and accuracy of pan-transcriptome assembly and annotation processes.

Future genomic endeavors will benefit from adding ML/AI methodologies to pangenome and pan-transcriptome analysis which provides enhanced accurate annotations combined with more advanced comparative genomics capabilities to reveal functional plant population variation.Embracing these technologies in future research endeavors can pave the way for innovative discoveries and transformative insights into plant genomic diversity and evolution.

CRediT authorship contribution statement

Pradeep Ruperao: Conceptualization, Supervision, Writing – original draft, Visualization, Writing – review & editing. **Parimalan Rangan:** Writing – review & editing; Trushar Shah: Writing – review & editing. **Vinay Sharma:** Writing–original draft, Visualization, Writing–review & editing. **Abhishek Rathore:** Writing – review & editing. **Sean Mayes:** Writing – review & editing. **Manish K. Pandey:** Conceptualization, Supervision, Funding acquisition, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research is funded by Bill & Melinda Gates Foundation (BMGF) through Tropical Legumes III (TL III) project. The authors duly acknowledged the Indian Council of Agricultural Research (ICAR), India and Department of Biotechnology (DBT) (Grant number: 16113200037-1012166), Government of India, Global initiative project VACS (Vision for Adapted Crops and Soils) for financial support. V.S also acknowledges Council of Scientific and Industrial Research (CSIR), Government of India, for awarding the SRF-Direct fellowship (File No: 09/0800(18433)/2024-EMR-I) for PhD.

Compliance with ethics requirements

The authors declare that they have no Compliance with Ethics Requirements in this paper.

References

[1] Jiao Y, Peluso P, Shi J, et al. Improved maize reference genome with single-molecule technologies Available at:. Nature 2017;546:524–7. http://www.nature.com/doifinder/10.1038/nature22971.

[2] McCormick, R.F., Truong, S.K., Sreedasyam, A., et al. (2017) The Sorghum bicolor reference genome: Improved assembly and annotations, a transcriptome atlas, and signatures of genome organization. bioRxiv. Available at: http://scholar.google.com/scholar?q=The Sorghum bicolor reference genome: improved assembly and annotations, a transcriptome atlas, and signatures of genome organization&btnG=&hl=en&num=20&as_sdt=0%2C22.

[3] Schmutz J, Cannon SB, Schlueter J, et al. Erratum: Genome sequence of the palaeopolyploid soybean (Nature (2010) 463 (178-183)) Available at:. Nature 2010;465:120. http://www.nature.com/doifinder/10.1038/nature08957.

[4] Xu X, Pan S, Cheng S, et al. Genome sequence and analysis of the tuber crop potato Available at:. Nature 2011;475:189–95. http://www.nature.com/doifinder/10.1038/nature10158.

[5] Beier S, Himmelbach A, Colmsee C, et al. Construction of a map-based reference genome sequence for barley Available at:. Hordeum vulgare L Sci Data 2017;4:170044. http://www.ncbi.nlm.nih.gov/pubmed/28448065.

[6] Varshney RK, Song C, Saxena RK, et al. Draft genome sequence of chickpea (Cicer arietinum) provides a resource for trait improvement Available at:. Nat Biotechnol 2013;31:240–6. http://www.nature.com/doifinder/10.1038/nbt.2491.

[7] Varshney RK, Chen W, Li Y, et al. Draft genome sequence of pigeonpea (Cajanus cajan), an orphan legume crop of resource-poor farmers Available at:. Nat Biotechnol 2012;30:83–9. http://www.nature.com/doifinder/10.1038/nbt.2022.

[8] Lukaszewski AJ, Alberti A, Sharpe A, et al. A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome Available at:. Science (80-) 2014;345:1251788. http://www.sciencemag.org/content/345/6194/1250092.abstract.

[9] Alaux M, Rogers J, Letellier T, et al. Linking the International Wheat Genome Sequencing Consortium bread wheat reference genome sequence to wheat genetic and phenomic data. Genome Biol 2018;19.

[10] Bayer PE, Golicz AA, Scheben A, Batley J, Edwards D. Plant pan-genomes are the new reference. Nat Plants 2020;6:914–20.

[11] Nurk S, Walenz BP, Rhie A, et al. HiCanu: Accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. Genome Res 2020;30.

[12] Chin CS, Peluso P, Sedlazeck FJ, et al. Phased diploid genome assembly with single-molecule real-time sequencing. Methods: Nat; 2016.

[13] Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Methods: Nat; 2021. p. 18.

[14] Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, et al. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial "pan-genome. Proc Natl Acad Sci U S A 2005;102:13950–5.

[15] Hirsch CN, Foerster JM, Johnson JM, et al. Insights into the maize pan-genome and pan-transcriptome Available at:. Plant Cell 2014;26:121–35. http://www.plantcell.org/cgi/doi/10.1105/tpc.113.119982.

[16] Li Y-H, Zhou G, Ma J, et al. De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits Available at:. Nat Biotechnol 2014;32:1045–52. http://www.nature.com/doifinder/10.1038/nbt.2979.

[17] Schatz MC, Maron LG, Stein JC, et al. Whole genome de novo assemblies of three divergent strains of rice, Oryza sativa, document novel gene space of Aus and Indica. Genome Biol 2014;15:506.

[18] Jensen SE, Charles JR, Muleta K, et al. A sorghum practical haplotype graph facilitates genome-wide imputation and cost-effective genomic prediction Available at:. Plant Genome 2020;13:e20009. http://www.ncbi.nlm.nih.gov/pubmed/33016627.

[19] **Ruperao, P., Thirunavukkarasu, N., Gandham, P., et al.** (2021) Sorghum Pan-Genome Explores the Functional Utility for Genomic-Assisted Breeding to Accelerate the Genetic Gain. Front. Plant Sci., **12**. Available at: https://pubmed.ncbi.nlm.nih.gov/34140962/ [Accessed December 23, 2021].

[20] Varshney RK, Roorkiwal M, Sun S, et al. A chickpea genetic variation map based on the sequencing of 3,366 genomes. Nature 2021;599.
[21] Khan, A.W., Garg, V., Sun, S., et al. (2024) Cicer super-pangenome provides insights into species evolution and agronomic trait loci for crop improvement in chickpea. Nat. Genet., 56, 1225–1234. Available at: doi: 10.1038/s41588-024-01760-4.
[22] Montenegro JD, Golicz AA, Bayer PE, et al. The pangenome of hexaploid bread wheat. Plant J 2017;90:1007–13.
[23] Walkowiak S, Gao L, Monat C, et al. Multiple wheat genomes reveal global variation in modern breeding. Nature 2020;588.
[24] Segerman B. The genetic integrity of bacterial species: the core genome and the accessory genome, two different stories Available at:. Front Cell Infect Microbiol 2012;2:116. http://journal.frontiersin.org/article/10.3389/fcimb.2012.00116/abstract.
[25] Ruperao P, Rangan P, Shah T, Thakur V, Kalia S, Mayes S, et al. The Progression in Developing Genomic Resources for Crop Improvement. Life 2023;13.
[26] Ramu P, Esuma W, Kawuki R, et al. Cassava haplotype map highlights fixation of deleterious mutations during clonal propagation. Nat Genet 2017;49:959–63.
[27] Chia JM, Song C, Bradbury PJ, et al. Maize HapMap2 identifies extant variation from a genome in flux. Nat Genet 2012;44:803–7.
[28] Kumar V, Khan AW, Saxena RK, Garg V, Varshney RK. First-generation HapMap in Cajanus spp. reveals untapped variations in parental lines of mapping populations. Plant Biotechnol J 2016;14:1673–81.
[29] Weigel D, Mott R. The 1001 genomes project for Arabidopsis thaliana Available at:. Genome Biol 2009;10:107. http://genomebiology.biomedcentral.com/articles/10.1186/gb-2009-10-5-107.
[30] Yao W, Li G, Zhao H, Wang G, Lian X, Xie W. Exploring the rice dispensable genome using a metagenome-like assembly strategy Available at:. Genome Biol 2015;16:187. http://genomebiology.com/2015/16/1/187.
[31] Hu, H., Zhao, J., Thomas, W.J.W., Batley, J. and Edwards, D. (2025) The role of pangenomics in orphan crop improvement. Nat. Commun., 16, 118. Available at: doi: 10.1038/s41467-024-55260-4.
[32] Golicz AA, Bayer PE, Barker GC, et al. The pangenome of an agronomically important crop plant Brassica oleracea Available at:. Nat Commun 2016;7:13390. http://www.nature.com/doifinder/10.1038/ncomms13390.
[33] Vernikos G, Medini D, Riley DR, Tettelin H. Ten years of pan-genome analyses. Curr Opin Microbiol 2015;23:148–54.
[34] Shi J, Tian Z, Lai J, Huang X. Plant pan-genomics and its applications. Plant: Mol; 2023. p. 16.
[35] Laing C, Buchanan C, Taboada EN, Zhang Y, Kropinski A, Villegas A, et al. Pan-genome sequence analysis using Panseq: An online tool for the rapid analysis of core and accessory genomic regions. BMC Bioinf 2010;11.
[36] Zhao Y, Wu J, Yang J, Sun S, Xiao J, Yu J. PGAP: Pan-genomes analysis pipeline. Bioinformatics 2012;28:416–8.
[37] Sheikhizadeh Anari S, de Ridder D, Schranz ME, Smit S. Efficient inference of homologs in large eukaryotic pan-proteomes. BMC Bioinf 2018;19.
[38] Schneeberger K, Hagmann J, Ossowski S, Warthmann N, Gesing S, Kohlbacher O, et al. Simultaneous alignment of short reads against multiple genomes. Genome Biol 2009;10.
[39] Valenzuela D, Norri T, Välimäki N, Pitkänen E, Mäkinen V. Towards pan-genome read alignment to improve variation calling. BMC Genomics 2018;19.
[40] Maciuca S, Elias CDO, McVean G, Iqbal Z. A natural encoding of genetic variation in a burrows-wheeler transform to enable mapping and genome inference. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2016.
[41] Rakocevic G, Semenyuk V, Lee WP, et al. Fast and accurate genomic analyses using genome graphs. Genet.: Nat; 2019. p. 51.
[42] **Garrison, E., Guarracino, A., Heumos, S., et al.** (2023) Building pangenome graphs. bioRxiv.
[43] Paten B, Eizenga JM, Rosen YM, Novak AM, Garrison E, Hickey G. Superbubbles, Ultrabubbles, and Cacti. In. J Comput Biol 2018.
[44] Garrison E, Sirén J, Novak AM, et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nat Biotechnol 2018;36:875–81.
[45] Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G. De novo assembly and genotyping of variants using colored de Bruijn graphs. Nat Genet 2012;44:226–32.
[46] Marcus S, Lee H, Schatz MC. SplitMEM: A graphical algorithm for pan-genome analysis with suffix skips. Bioinformatics 2014;30:3476–83.
[47] Minkin I, Pham S, Medvedev P. TwoPaCo: an efficient algorithm to build the compacted de Bruijn graph from many complete genomes. Bioinformatics 2017;33.
[48] Holley G, Melsted P. Bifrost: Highly parallel construction and indexing of colored and compacted de Bruijn graphs. Genome Biol 2020;21.
[49] Eizenga JM, Novak AM, Sibbesen JA, et al. Pangenome Graphs. Rev. Genomics Hum. Genet.: Annu; 2020. p. 21.
[50] Rautiainen M, Marschall T. GraphAligner: rapid and versatile sequence-to-graph alignment. Genome Biol 2020;21.
[51] Hübner S, Bercovich N, Todesco M, Mandel JR, Odenheimer J, Ziegler E, et al. Sunflower pan-genome analysis shows that hybridization altered gene content and disease resistance. Nat Plants 2019;5.
[52] Hurgobin B, Golicz AA, Bayer PE, Chan CKK, Tirnaz S, Dolatabadian A, et al. Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid Brassica napus. Plant Biotechnol J 2018;16:1265–74.
[53] Hufnagel B, Soriano A, Taylor J, Divol F, Kroc M, Sanders H, et al. Pangenome of white lupin provides insights into the diversity of the species. Plant Biotechnol J 2021:19.
[54] Liu Y, Tian Z. Super graph-based pan-genome: Bringing rice functional genomic study into a new dawn. Mol. Plant 2022;15:1409–11.
[55] Liu Y, Du H, Li P, Shen Y, Peng H, Liu S, et al. Pan-Genome of Wild and Cultivated Soybeans. Cell ;2020:182.
[56] Lin K, Zhang N, Severing EI, Nijveen H, Cheng F, Visser RGF, et al. Beyond genomic variation - comparison and functional annotation of three *Brassica rapa* genomes: A turnip, a rapid cycling and a Chinese cabbage. BMC Genomics 2014;15:250.
[57] Liang Q, Muñoz-Amatriaín M, Shu S, Lo S, Wu X, Carlson JW, et al. A view of the pan-genome of domesticated Cowpea (*Vigna unguiculata* [L.] Walp.). Plant Genome 2024:17.
[58] Lian Q, Huettel B, Walkemeier B, Mayjonade B, Lopez-Roques C, Gil L, et al. A pan-genome of 69 *Arabidopsis thaliana* accessions reveals a conserved genome structure throughout the global species range. Nat Genet 2024;56:982–91.
[59] Li H, Wang S, Chai S, Yang Z, Zhang Q, Xin H, et al. Graph-based pan-genome reveals structural and sequence variations related to agronomic traits and domestication in cucumber. Nat Commun 2022;13.
[60] Li N, He Q, Wang J, Wang B, Zhao J, Huang S, et al. Super-pangenome analyses highlight genomic diversity and structural variation across wild and cultivated tomato species. Nat Genet 2023;55:852–60.
[61] Kang M, Wu H, Liu H, Liu W, Zhu M, Han Y, et al. The pan-genome and local adaptation of *Arabidopsis thaliana*. Nat Commun 2023:14.
[62] Ou L, Li D, Lv J, Chen W, Zhang Z, Li X, et al. Pan-genome of cultivated pepper (*Capsicum*) and its use in gene presence–absence variation analyses. New Phytol 2018:220.
[63] Lyu X, Xia Y, Wang C, Zhang K, Deng G, Shen Q, et al. Pan-genome analysis sheds light on structural variation-based dissection of agronomic traits in melon crops. Plant Physiol 2023:193.
[64] Lovell JT, Bentley NB, Bhattarai G, Jenkins JW, Sreedasyam A, Alarcon Y, et al. Four chromosome scale genomes and a pan-genome annotation to accelerate pecan tree breeding. Nat Commun 2021:12.
[65] Long, E.M., Bradbury, P.J., Cinta Romay, M., Buckler, E.S. and Robbins, K.R. (2022) Genome-wide imputation using the practical haplotype graph in the heterozygous crop cassava. G3 Genes, Genomes, Genet., 12.
[66] Liu, Zhongjie, Wang, N., Su, Y., et al. (2024) Grapevine pangenome facilitates trait genetics and genomic breeding. Nat. Genet., 56, 2804–2814. Available at: doi: 10.1038/s41588-024-01967-5.
[67] Tang D, Jia Y, Zhang J, Li H, Cheng L, Wang P, et al. Genome evolution and diversity of wild and cultivated potatoes. Nature 2022;606:535–41.
[68] Song JM, Guan Z, Hu J, Guo C, Yang Z, Wang S, et al. Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of Brassica napus. Nat Plants 2020:6.
[69] Wang S, Qian YQ, Zhao RP, Chen LL, Song JM. Graph-based pan-genomes: Increased opportunities in plant genomics. J Exp Bot 2023;74.
[70] MacNish, T.R., Al-Mamun, H.A., Bayer, P.E., et al. (2025) Brassica Panache: A multi-species graph pangenome representing presence absence variation across forty-one Brassica genomes. Plant Genome, 18, e20535. Available at: doi: 10.1002/tpg2.20535.
[71] Shi T, Zhang X, Hou Y, Jia C, Dan X, Zhang Y, et al. The super-pangenome of Populus unveils genomic facets for its adaptation and diversification in widespread forest trees. Mol Plant 2024;17:725–46.
[72] Rijzaani H, Bayer PE, Rouard M, Doležel J, Batley J, Edwards D. The pangenome of banana highlights differences between genera and genomes. Plant Genome 2022:15.
[73] Qin P, Lu H, Du H, Wang H, Chen W, Chen Z, et al. Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. Cell ;2021:184.
[74] Pinosio S, Giacomello S, Faivre-Rampant P, Taylor G, Jorge V, Le Paslier MC, et al. Characterization of the Poplar pan-genome by genome-wide identification of structural variation. Mol Biol Evol 2016:33.
[75] Franco JAV, Gage JL, Bradbury PJ, Johnson LC, Miller ZR, Buckler ES, et al. A Maize Practical Haplotype Graph Leverages Diverse NAM Assemblies. bioRxiv 2020:2.
[76] Vaughn JN, Branham SE, Abernathy B, Hulse-Kemp AM, Rivers AR, Levi A, et al. Graph-based pangenomics maximizes genotyping density and reveals structural impacts on fungal resistance in melon. Nat Commun 2022;13:7897.
[77] Trouern-Trend AJ, Falk T, Zaman S, Caballero M, Neale DB, Langley CH, et al. Comparative genomics of six Juglans species reveals disease-associated gene family contractions. Plant J ;2020:102.
[78] Tao Y, Luo H, Xu J, Cruickshank A, Zhao X, Teng F, et al. Extensive variation within the pan-genome of cultivated and wild sorghum. Nat Plants 2021 76 2021;7:766–73.
[79] Zhao Q, Feng Q, Lu H, Li Y, Wang A, Tian Q, et al. Erratum to: Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. Nat. Genet. 2018;50:278–84.

[80] Zhao J, Bayer PE, Ruperao P, Saxena RK, Khan AW, Golicz AA, et al. Trait associations in the pangenome of pigeon pea (*Cajanus cajan*). Plant Biotechnol J 2020.
[81] Zhang X, Liu T, Wang J, Wang P, Qiu Y, Zhao W, et al. Pan-genome of Raphanus highlights genetic variation and introgression among domesticated, wild, and weedy radishes. Mol. Plant 2021;14:2032–55.
[82] Yu J, Golicz AA, Lu K, Dossa K, Zhang Y, Chen J, et al. Insight into the evolution and functional characteristics of the pan-genome assembly from sesame landraces and modern cultivars. Plant Biotechnol J 2019:17.
[83] Workum, D.-J.M. van, Mehrem, S.L., Snoek, B.L., et al. (2024) Lactuca super-pangenome reduces bias towards reference genes in lettuce research. BMC Plant Biol., 24, 1019. Available at: doi: 10.1186/s12870-024-05712-2.
[84] Yang T, Liu R, Luo Y, Hu S, Wang D, Wang C, et al. Improved pea reference genome and pan-genome highlight genomic features and evolutionary characteristics. Nat Genet 2022;54:1553–63.
[85] Wang Y, Fu L, Ren J, Yu Z, Chen T, Sun F. Identifying Group-Specific sequences for microbial communities using Long k-mer sequence signatures. Front Microbiol 2018;9:872.
[86] Barchi L, Rabanus-Wallace MT, Prohens J, Toppino L, Padmarasu S, Portis E, et al. Improved genome assembly and pan-genome provide key insights into eggplant domestication and breeding. Plant J ;2021:107.
[87] Zhou Y, Zhang Z, Bao Z, Li H, Lyu Y, Zan Y, et al. Graph pangenome captures missing heritability and empowers tomato breeding. Nature 2022;606:527–34.
[88] Zhou P, Silverstein KAT, Ramaraj T, Guhlin J, Denny R, Liu J, et al. Exploring structural variation and gene family architecture with De Novo assemblies of 15 Medicago genomes. BMC Genom 2017;18:261.
[89] Gui S, Wei W, Jiang C, Luo J, Chen L, Wu S, et al. A pan-Zea genome map for enhancing maize improvement. Genome Biol 2022;23:178.
[90] Gordon SP, Contreras-Moreira B, Woods DP, Des Marais DL, Burgess D, et al. Extensive gene content variation in the *Brachypodium distachyon* pan-genome correlates with population structure. Nat. Commun. 2017;8:2184.
[91] Gao L, Gonda I, Sun H, Ma Q, Bao K, Tieman DM, et al. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. Nat Genet 2019;51:1044–51.
[92] Jayakodi, M., Lu, Q., Pidon, H., et al. (2024) Structural variation in the pangenome of wild and domesticated barley. Nature, 636, 654–662. Available at: doi: 10.1038/s41586-024-08187-1.
[93] Dolatabadian A, Bayer PE, Tirnaz S, Hurgobin B, Edwards D, Batley J. Characterization of disease resistance genes in the *Brassica napus* pangenome reveals significant structural variation. Plant Biotechnol J 2020;18:969–82.
[94] Cochetel N, Minio A, Guarracino A, Garcia JF, Figueroa-Balderas R, Massonnet M, et al. A super-pangenome of the North American wild grape species. Genome Biol ;2023:24.
[95] Chen S, Wang P, Kong W, Chai K, Zhang S, Yu J, et al. Gene mining and genomics-assisted breeding empowered by the pangenome of tea plant *Camellia sinensis*. Nat. Plant. 2023b;9, 1986-1999.
[96] Chen J, Liu Y, Liu M, Guo W, Wang Y, He Q, et al. Pangenome analysis reveals genomic variations associated with domestication traits in broomcorn millet. Nat Genet 2023a;55.
[97] Bayer PE, Scheben A, Golicz AA, Yuan Y, Faure S, Lee HT, et al. Modelling of gene loss propensity in the pangenomes of three *Brassica* species suggests different mechanisms between polyploids and diploids. Plant Biotechnol J 2021:19.
[98] Jordan, K.W., Bradbury, P.J., Miller, Z.R., et al. (2022) Development of the Wheat Practical Haplotype Graph database as a resource for genotyping data storage and genotype imputation. G3 Genes, Genomes, Genet., 12.
[99] Hein J. A new method that simultaneously aligns and reconstructs ancestral sequences for any number of homologous sequences, when the phylogeny is given. Biol. Evol.: Mol; 1989. p. 6.
[100] Paten B, Novak AM, Eizenga JM, Garrison E. Genome graphs and the evolution of genome inference. Genome Res 2017;27:665–76.
[101] Nguyen N, Hickey G, Zerbino DR, Raney B, Earl D, Armstrong J, et al. Building a pangenome reference for a population. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 2014:207–21.
[102] Heber S, Alekseyev M, Sze S-H, Tang H, Pevzner PA. Splicing graphs and EST assembly problem Available at:. Bioinformatics 2002;18(Suppl 1):S181–8. http://www.ncbi.nlm.nih.gov/pubmed/12169546.
[103] Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat Protoc 2016;11:1650–67.
[104] Garg S, Rautiainen M, Novak AM, Garrison E, Durbin R, Marschall T. A graph-based approach to diploid genome assembly. Bioinformatics 2018:i105–14.
[105] Sirén J, Paten B. GBZ file format for pangenome graphs. Bioinformatics 2022;38.
[106] Ruperao P, Gandham P, Rathore A. Construction of Practical Haplotype Graph (PHG) with the Whole-Genome Sequence Data. In Methods in Molecular Biology, 2022.
[107] Bradbury PJ, Casstevens T, Jensen SE, Johnson LC, Miller ZR, Monier B, et al. The Practical Haplotype Graph, a platform for storing and using pangenomes for imputation. Bioinformatics 2022;38.
[108] Rubin JD, Vogel NA, Gopalakrishnan S, Sackett PW, Renaud G. HaploCart: Human mtDNA haplogroup classification using a pangenomic reference graph. PLoS Comput Biol 2023;19.
[109] Guarracino A, Heumos S, Nahnsen S, Prins P, Garrison E. ODGI: Understanding pangenome graphs. Bioinformatics 2022;38.
[110] Leonard AS, Crysnanto D, Mapel XM, Bhati M, Pausch H. Graph construction method impacts variation representation and analyses in a bovine super-pangenome. Genome Biol 2023;24.
[111] Hickey G, Monlong J, Ebler J, et al. Pangenome graph construction from genome alignments with Minigraph-Cactus. Nat Biotechnol 2023.
[112] Severin J, Lizio M, Harshbarger J, et al. Interactive visualization and analysis of large-scale sequencing datasets using ZENBU. Nat Biotechnol 2014;32.
[113] L'Yi S, Wang Q, Lekschas F, Gehlenborg N. Gosling: A Grammar-based Toolkit for Scalable and Interactive Genomics Data Visualization. Vis. Comput. Graph.: IEEE Trans; 2022. p. 28.
[114] Suzuki A, Kawano S, Mitsuyama T, et al. DBTSS/DBKERO for integrated analysis of transcriptional regulation. Nucleic Acids Res 2018;46.
[115] Li F. Visualization and review of reads alignment on the graphical pan-genome with VAG Available at: https://www.biorxiv.org/content/10.1101/2023.01.20.524849v1 2023.
[116] Kong L, Wang J, Zhao S, Gu X, Luo J, Gao G. ABrowse - A customizable next-generation genome browser framework. BMC Bioinf 2012;13.
[117] Stein LD. Using GBrowse 2.0 to visualize and share next-generation sequence data. Brief Bioinform 2013;14.
[118] Kunyavskaya O, Prjibelski AD. SGTK: A toolkit for visualization and assessment of scaffold graphs. Bioinformatics 2019;35.
[119] Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: Interactive visualization of de novo genome assemblies. Bioinformatics 2015;31:3350–2.
[120] Boisvert S, Laviolette F, Corbeil J. Ray: Simultaneous assembly of reads from a mix of high-throughput sequencing technologies. J Comput Biol 2010;17.
[121] Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. Brief Bioinform 2013;14:178–92.
[122] Durant É, Sabot F, Conte M, Rouard M. Panache: A web browser-based viewer for linearized pangenomes. Bioinformatics 2021;37.
[123] Freese NH, Norris DC, Loraine AE. Integrated genome browser: Visual analytics platform for genomics. Bioinformatics 2016;32.
[124] Mikheenko A, Kolmogorov M. Assembly Graph Browser: Interactive visualization of assembly graphs. Bioinformatics 2019;35.
[125] Medina I, Salavert F, Sanchez R, de Maria A, Alonso R, Escobar P, et al. Genome Maps, a new generation genome browser. Nucleic Acids Res 2013;41.
[126] Vernikos GS, Gkogkas CG, Promponas VJ, Hamodrakas SJ. GeneViTo: Visualizing gene-product functional and structural features in genomic datasets. BMC Bioinf 2003;4.
[127] Lopez-Delisle L, Rabbani L, Wolff J, Bhardwaj V, Backofen R, Grüning B, et al. pyGenomeTracks: reproducible plots for multivariate genomic datasets. Bioinformatics 2021;37.
[128] Down TA, Piipari M, Hubbard TJP. Dalliance: Interactive genome viewing on the web. Bioinformatics 2011;27.
[129] Diesh C, Stevens GJ, Xie P, et al. JBrowse 2: a modular genome browser with views of synteny and structural variation. Genome Biol 2023;24.
[130] Mikheenko A, Valin G, Prjibelski A, Saveliev V, Gurevich A. Icarus: Visualizer for de novo assembly evaluation. Bioinformatics 2016;32.
[131] Milne I, Stephen G, Bayer M, Cock PJA, Pritchard L, Cardle L, et al. Using tablet for visual exploration of second-generation sequencing data. Brief Bioinform 2013;14:193–202.
[132] Goecks J, Coraor N, Nekrutenko A, Taylor J. NGS analyses by visualization with Trackster. Nat Biotechnol 2012;30.
[133] Gonnella G, Niehus N, Kurtz S. GfaViz: Flexible and interactive visualization of GFA sequence graphs. Bioinformatics 2019;35.
[134] Nielsen CB, Jackman SD, Birol I, Jones SJM. ABySS-explorer: visualizing genome sequence assemblies. IEEE Transactions on Visualization and Computer Graphics, 2009.
[135] Zhu X, Zhang Y, Wang Y, Tian D, Belmont AS, Swedlow JR, et al. Nucleome Browser: an integrative and multimodal data navigation platform for 4D Nucleome. Methods: Nat; 2022. p. 19.
[136] Yuan, Y. (2023) PanGraphViewer: A Versatile Tool to Visualize Pangenome Graphs. Available at: https://www.biorxiv.org/content/10.1101/2023.03.30.534931v1.
[137] Zhang Y, Zhou R, Liu L, Chen L, Wang Y. Trackplot: A flexible toolkit for combinatorial analysis of genomic data. PLoS Comput Biol 2023;19.
[138] Antipov D, Rayko M, Kolmogorov M, Pevzner PA. viralFlye: assembling viruses and identifying their hosts from long-read metagenomics data. Genome Biol 2022;23.
[139] Yokoyama TT, Sakamoto Y, Seki M, Suzuki Y, Kasahara M. MoMI-G: Modular multi-scale integrated genome graph browser. BMC Bioinf 2019;20.
[140] Rangwala SH, Kuznetsov A, Ananiev V, et al. Accessing NCBI data using the NCBI sequence viewer and genome data viewer (GDV). Genome Res 2021;31.

[141] Saito TL, Yoshimura J, Ahsan B, Sasaki A, Kurosh R, Morishita S. UTGB Toolkit for Personalized Genome Browsers. Tag-Based Next Generation Sequencing 2012.
[142] Katainen R, Donner I, Cajuso T, Kaasinen E, Palin K, Mäkinen V, et al. Discovery of potential causative mutations in human coding and noncoding genome with the interactive software BasePlayer. Protoc.: Nat; 2018. p. 13.
[143] Hubbard T, Barker D, Birney E, et al. The Ensembl genome database project. Nucleic Acids Res 2002;30.
[144] Robinson JT, Thorvaldsdottir H, Turner D, Mesirov JP. igv.js: an embeddable JavaScript implementation of the Integrative Genomics Viewer (IGV). Bioinformatics 2023;39.
[145] Kerpedjiev P, Abdennur N, Lekschas F, et al. HiGlass: Web-based visual exploration and analysis of genome interaction maps. Genome Biol 2018;19.
[146] Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The Human Genome Browser at UCSC. Genome Res 2002;12.
[147] Tolstoganov I, Bankevich A, Chen Z, Pevzner PA. CloudSPAdes: Assembly of synthetic long reads using de Bruijn graphs. In. Bioinformatics 2019.
[148] Gnerre S, MacCallum I, Przybylski D, et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Sci: Proc. Natl. Acad; 2011.
[149] Khan J, Kokot M, Deorowicz S, Patro R. Scalable, ultra-fast, and low-memory construction of compacted de Bruijn graphs with Cuttlefish 2. Genome Biol 2022;23.
[150] Hu H, Wang J, Nie S, Zhao J, Batley J, Edwards D. Plant pangenomics, current practice and future direction Available at:. Agric Commun 2024;2:100039. https://www.sciencedirect.com/science/article/pii/S2949798124000152.
[151] Pandey S, Divakar S, Singh A. Genome editing prospects for heat stress tolerance in cereal crops Available at:. Plant Physiol Biochem 2024;215:108989. https://www.sciencedirect.com/science/article/pii/S0981942824006570.

Dr. Pradeep Ruperao, an experienced Research Associate (Bioinformatics) at the Center of Excellence in Genomics and Systems Biology at ICRISAT, India, possesses over 15 years of versatile experience in Research and Development. He is a proficient genome sequencing and high-throughput genotyping data analyst with a remarkable track record in leading-edge scientific domains. His expertise encompasses various aspects of genomics analysis, including pan-genome assemblies, variant prediction, population genetics, quantitative genetics, and more. With a rich background, Pradeep has made significant contributions to genome analysis in a spectrum of crops, such as Chickpea, Pigeonpea, Wheat, Brassica, Cassava, and Fungal pathogens at the University of Queensland, University of Western Australia (Australia), NIAB (Cambridge, UK) and ICRISAT (India). Leveraging his comprehensive skill set, he actively manages and contributes to research projects dealing with extensive sequencing data for genome assemblies, pangenomics, germplasm characterization, and pre-breeding initiatives. His overarching goal is to spearhead accelerated crop improvement across diverse crops. Furthermore, Dr. Ruperao is at the forefront of incorporating Artificial Intelligence (AI) and Machine Learning (ML) concepts into his research endeavors. By integrating these advanced technologies, he aims to enhance the precision and efficiency of crop improvement initiatives, ensuring a cutting-edge approach to genomics and bioinformatics in agriculture. His commitment to innovation positions him as a valuable asset in advancing scientific frontiers and shaping the future of crop research.

Dr. Parimalan Rangan is a scientist in Agricultural Biotechnology at ICAR-National Bureau of Plant Genetic Resources, and an honorary senior fellow at the Queensland Alliance for Agriculture and Food Innovation, The University of Queensland, Australia. His expertise in transcriptomics and functional genomics had led to various research contributions in cereal and oilseed crop groups, especially rice, wheat, and sesame. Novel biological insights gained through the transcriptome-based RNA-seq experiments had helped propose for an early evolution (than the presently known) of the C4 photosynthesis pathway in the BEP clade with minimal, but functional, requirements. He is proficient in handling genome-scale data to derive biological meaning using appropriate computational tools. He also ensures the services with special reference to DNA fingerprinting profiling for sesame. He has expertise in plant tissue culture and transformation including gene cloning. Serving at the ICAR-NBPGR for more than 15 years, he had gained expertise in the plant genetic resources management, both genotyping and phenotyping, through large-scale experiments. He is involved in collaborative research projects at national and international level, dealing on a large set of plant genetic resources for efficient management and utilization through the effective use of 'omics' tools. He has more than 50 research contributions.

Dr. Trushar Shah is currently a Bioinformatician and the Country Representative-Kenya at the International Institute of Tropical Agriculture(IITA). He has a post-graduate degree with distinction in Molecular modelling and Bioinformatics from the University of London. He has successfully applied bioinformatics to data integration, comparative genomics, precision genetics and allele mining approaches for abiotic and biotic stress across crops. He also leads the development and implementation of breeding informatics tools. He has co-authored papers in high-impact journals, including Science, Nature, and PNAS, and has received several awards. He has been actively involved in building capacity and has been a resource person of the BixCOP (Bioinformatics Community of Practice), the Integrated Breeding Platform multi-year course on molecular breeding and the African Plant Breeding Academy including both the Breeder's course and the CRISPR academy.

Mr. Vinay Sharma is a PhD Research Scholar at the International Crops Research Institute for the Semi-Arid Tropics (ICRISAT) in Hyderabad, India, concurrently serving as a Visiting Research Fellow at the University of Leeds, U.K. He is also a CLIFF-GRADS awardee, supported by the New Zealand Government's Ministry for Primary Industries and the CGIAR Trust Fund Donors. His profound expertise in molecular breeding, genomics, and biotechnology, with significant contributions to research in cereal and oilseed crop groups, particularly rice and peanut. His skill set extends to plant tissue culture and transformation, encompassing gene cloning in Arabidopsis. With a prolific academic career, Vinay boasts over 45 research contributions, demonstrating his unwavering commitment to advancing agricultural science and addressing global food security challenges.

Dr. Abhishek Rathore leads the Statistical, Biometric, Bioinformatics, and Breeding Informatics initiatives for the Dryland Crops Program at International Maize and Wheat Improvement Center (CIMMYT). In this role, he ensures that crop breeding pipelines are not only meticulously designed but also optimized and adhere to the modern quantitative genetics principles. He has designed breeding pipelines aiming for higher genetic gains by implementing genomic selection and other forward breeding methodologies. In his current capacity, he also ensures that all research trials conform to appropriate and modern experimental designs and data is recorded by following standard crop ontologies and relevant scales. He oversees the analysis of data through modern statistical methodologies, guaranteeing that decisions regarding crop improvement are founded on solid statistical evidence. He has demonstrated skills in designing visualization tools, dashboards and developing visual analytics for crop breeding. He is also responsible for the digitalization of crop breeding programs and implementation of breeding data management systems, such as BMS and EBS, across the crop improvement network. Furthermore, he is at the forefront of developing capacity-building programs in biometrics, quantitative genetics, and data management for partners in CGIAR and NARS, enhancing their capabilities and expertise in these critical areas.

Prof. Sean Mayes serves as the Director of the Global Research Program - Accelerated Crop Improvement at ICRISAT, India, overseeing diverse sub-research groups such as seed systems, crop breeding, genomics pre-breeding and bioinformatics, genebank, crop physiology, cell molecular biology, and genetic engineering. With a focus on genetic diversity and germplasm characterization in various crops, Prof. Mayes excels in marker-assisted selection for genetic enhancement. His extensive involvement spans numerous projects encompassing wheat, oil palm, African rice, Bambara groundnut, winged bean, foxtail and proso millets, moth bean, amaranths, quinoa, and more. Prior to his role at ICRISAT, Dr. Mayes held key positions as an Associate Professor of Crop Genetics at the University of Nottingham, UK, and as the Theme Director for crop genetics at the Crops for the Future Research Centre, Malaysia. Additionally, he served as the Principal Investigator at the Department of Genetics, University of Cambridge, UK. In his capacity at ICRISAT, Prof. Mayes plays a pivotal role in advancing research initiatives to revolutionize agriculture, expanding the array of crops available globally. His contributions aim to foster agricultural sustainability and improve human nutrition by diversifying crop options for farmers worldwide.

Dr. Manish K. Pandey is currently leading the Groundnut and Pigeonpea Genomics, Prebreeding & Bioinformatics group at International Crops Research Institute for the SemiArid Tropics (ICRISAT), Hyderabad, India. His key contributions include reference genome sequence for diploid progenitor and both the subspecies of cultivated tetraploid groundnut, gene expression atlas, low to high density SNP genotyping assays, diagnostic markers for key traits, quality control panel, marker-assisted pyramiding and genomic selection in groundnut. Through genomics-based breeding, his team developed three high oleic groundnut varieties which are released for cultivation in six states of India. He has published more than 225 scientific articles (>10000 citations) in various journals of international repute including Nature Genetics, PNAS-USA and Molecular Plant. His efforts have been recognized at international level as he has been invited to deliver presentations in several international conferences, review the proposals for international funding agencies and he is collaborating with a large number of scientists at international level. He is also Adjunct Associate Professor in University of Southern Queensland (USQ), Australia & Distinguished Professor, in Institute of Crop Genetic Resources of Shandong Academy of Agricultural Sciences (SAAS), China. He is now being inducted in many scientific academies of India namely as Fellow in National Academy of Agricultural Sciences (NAAS)-New Delhi, and Telangana Academy of Agricultural Sciences (TAAS)-Telangana, while Associate Fellow/Member in Andhra Pradesh Akademi of Agricultural Sciences (APAS)-Andhra Pradesh, and Member of the National Academy of Sciences-India (NASI). He was also among the top 2% highly cited researcher in 2019-2022 compiled by Stanford University, USA.