This is a “preproof” accepted article for *Quantitative Plant Biology*.
This version may be subject to change during the production process.
10.1017/qpb.2025.10028

Complexity welcome: Pangenome graphs for comprehensive population genomics

Zhigui Bao[1], Detlef Weigel[1,2*]

[1]Department of Molecular Biology, Max Planck Institute for Biology Tübingen, 72076 Tübingen, Germany

[2]Institute for Bioinformatics and Medical Informatics, University of Tübingen, 72076 Tübingen, Germany

*for correspondence: weigel@tue.mpg.de

Keywords

Pangenome graphs, complex variation, plant genomics

Abstract

Pangenome graphs are revolutionizing evolutionary and population genomics by moving beyond linear reference genomes to represent the full spectrum of sequence diversity within and across species. This review traces the field’s progression from reference-augmented graphs to assembly-based, alignment-first approaches that capture complex structural variation with reduced bias. We examine key strategies for graph construction, genotyping, and implementing graph-aware tools in functional genomics, including transcriptomics and epigenomics. While much of the work to date has focused on humans, diverse and structurally complex plant genomes pose unique challenges that require further methodological innovation. Key bottlenecks—including visualization, scalability, and integration with multi-omic data—persist. By outlining trade-offs among current tools and emphasizing the need for rigorous evaluation frameworks, we argue that progress will depend on community-driven efforts to unify

graph construction, genotyping, and interpretation. Despite technical hurdles, pangenome graphs offer a powerful foundation for more inclusive evolutionary and population genomics.

Introduction

The first papers describing nearly complete genome sequences were typically entitled “The genome of …”. Implicit in these titles was the assumption that much of the genome is shared between different members of a species. Looking back today, it is clear that a great deal has been learned from the study of genes conserved in each species, the proteins they encode, their regulatory elements, and so on. Similarly, the commonalities as well as fixed differences between both closely and more distantly related species have greatly informed biology, including our understanding of evolution over different time scales.

As interesting as all this knowledge is, we cannot fully understand biology without considering the genetic differences between individuals. These are not only at the root of adaptation to specific environments but also underlie susceptibility to disease and abiotic factors. Such seemingly deleterious variants are often linked by evolutionary trade-offs, where genes and alleles that are favorable in one environment become a liability in a different environment. The original genome papers for different species, therefore, were often quickly followed by attempts to record interindividual differences at the whole-genome level. Pioneering in this regard were scientists working with the plant *Arabidopsis thaliana*. The initial genome paper already contained information on shotgun sequences from a second strain in addition to the one from which the genome sequence had been primarily generated (The Arabidopsis Genome Initiative, 2000). Soon thereafter, the first genome-wide polymorphism analyses were published, at increasingly higher resolution, relying on a series of different technologies (Clark et al., 2007; Kim et al., 2007; Nordborg et al., 2002, 2005). Some of the approaches could interrogate in principle every position in the reference genome, but were limited by the extent of sequence divergence that could be recorded, and in highly divergent or missing regions, the exact sequence remained unknown (Clark et al., 2007). This remained the case when much cheaper short-read sequencing entered the scene, although with increasing lengths of short reads, more and more of the genome became accessible to polymorphism analyses (Ossowski et al., 2008)]. Importantly, short-read sequencing provided access to sequences that were not present in the original reference genome, and some of these could even be anchored to reference positions (Cao et al., 2011; Gan et al., 2011; Long et al., 2013; Ossowski et al., 2008; Schneeberger et al., 2011). The picture for other plant species, especially the crops rice and maize, was broadly

https://doi.org/10.1017/qpb.2025.10028 Published online by Cambridge University Press

similar, albeit typically with a delay of a few years (Gao et al., 2019; Y.-H. Li et al., 2014; Zhao et al., 2018).

The use of short reads for the identification of sequence polymorphisms that distinguish the focal individual from the reference strain begins with the mapping of reads against the reference genome sequence. A limitation, therefore, is the reference bias that is caused by the degree of mismatch between a specific short read and its target. Although mismatches may still allow confident mapping, more reads can be mapped with a more closely related genome. This insight led early on to the suggestion of producing synthetic reference genome sequences that represented all possible combinations of polymorphisms across the genome, including those not yet discovered in the genomes analyzed (Schneeberger et al., 2009).

A less obvious problem has been that a specific sequence might be present only once in one genome, but multiple times in another genome. And even if a sequence is present only once, it might occur in different regions of the genome across individuals. In both cases, short-read mapping can be misleading., For example, heterozygosity may be inferred when in reality there are two closely related but not identical duplicated fragments in the short-read sequenced genome.

Rather than relying on a single reference genome, the concept of the pangenome was introduced to capture all sequence variation within a species. First applied in bacterial genomes (Tettelin et al., 2005), this framework quantified core and dispensable genes across populations, highlighting extensive diversity. As sequencing costs declined, the pangenome approach was extended to eukaryotic genomes, including humans and *Arabidopsis thaliana* (Cao et al., 2011; R. Li et al., 2010; Sudmant et al., 2010). Over time, this concept evolved into a broader model encompassing the full genomic landscape of a population, species, or clade. A somewhat unfortunate fact is that “pangenome” these days more often refers to the collection of a specific set of genomes rather than the ideal of all reasonably common variants in a population.

The evolution of building pangenome graphs

One intuitive way to overcome the limitations of a linear reference genome is through genome graphs, which offer a compact data structure where sequences are represented as node-labeled graphs with edges connecting variants across multiple genomes. Unlike traditional linear representations, genome graphs preserve original coordinates by tracing paths through the sequence graph, accommodating both shared and unique genomic regions. The interpretability

https://doi.org/10.1017/qpb.2025.10028 Published online by Cambridge University Press

of pangenome graphs and their level of detail exist on a two-dimensional spectrum. At one extreme, highly abstract graphs (Fig. 1a-b), such as those representing every nucleotide variant with numerous loops and alternative paths, may be difficult to understand and have limited practical use. At the other extreme, unaligned genomic sequences, while easy to interpret, can obscure meaningful genomic differences. Depending on the approach, fixed k-mer-based de Bruijn graphs (Fig. 1a,g) and fully aligned multi-individual genomes (Fig. 1c-h) represent different points along this continuum (Fig. 1i).

Variation-first: augmenting the reference

Early pangenome graph construction was constrained by practical realities: high-quality genome assemblies were expensive and rare, while catalogs of variation from resequencing projects were abundant. This imbalance led to reference-augmented approaches that embedded known variants into linear reference backbones. These methods typically integrate variant genotyping within the same workflow, though here we focus on the graph construction aspect.

Schneeberger and colleagues (Schneeberger et al., 2009) pioneered this concept with GenomeMapper, demonstrating that incorporating known polymorphisms could reduce mapping bias in *Arabidopsis thaliana*. The approach was later extended to the major histocompatibility complex region in humans (Dilthey et al., 2015), where high diversity makes linear references particularly inadequate. These early successes established the fundamental principle that graph representations could capture variation more faithfully than linear sequences. It was further generalized by several groups to the whole genomes of thousands of human individuals, with each group using distinct strategies (Eggertsson et al., 2017; Garrison et al., 2018; Rakocevic et al., 2019).

GraphTyper (Eggertsson et al., 2017) iteratively realigned, clipped, and unaligned short reads with an embedded genome graph for small variant calling. The VG toolkit (Variation Graph toolkit, (Garrison et al., 2018; Hickey et al., 2020) emerged as the first comprehensive open-source framework for this reference-augmented paradigm. VG constructs variation graphs by threading known variants from VCF files into reference genomes or directly from genome alignment, creating alternative paths that represent different allelic states. It supports complex structural variations, which include duplications and inversions, using a bidirectional cyclic graph. Graph Genome pipeline (Rakocevic et al., 2019) also supports SVs for genotyping with high speed, but it is limited to human genomes and is not openly distributed.

https://doi.org/10.1017/qpb.2025.10028 Published online by Cambridge University Press

Due to the flexibility of constructing graphs from precomputed VCFs, VG has become the backbone of many pipelines. By incorporating variants derived from genome assemblies or filtered long-read alignments, VG-based workflows have been successfully applied across diverse species, including humans and crops (Y. Liu et al., 2020; Qin et al., 2021; Sirén et al., 2021; Y. Zhou et al., 2022).

The alignment-first paradigm: towards unbiased representation

With decreasing costs and improving quality of long-read sequencing, generating multiple high-quality genome assemblies has become increasingly feasible, shifting the bottleneck from data generation to comparative analysis. This shift enabled a new paradigm in pangenome graph construction—moving from reference-based variant threading to graph building through whole-genome alignments directly, adopting an “alignment-first” approach. Theoretically, this approach can reduce reference bias and better capture complex structural variations, including inversion, duplications, and rearrangements that are hard to encode using VCF-based models.

Even before genome graphs were formally introduced, a multiple genome alignment (MGA) already served as an implicit representation of shared and divergent sequence features across assemblies. Multiple sequence alignments (MSAs) naturally lend themselves to representation as partially ordered sequence (POA) graphs (Lee et al., 2002), which have been extended into A-Bruijn graphs (Raphael et al., 2004), and cactus graphs (Paten et al., 2011) to better accommodate genome rearrangements and duplications. Mauve (A. C. E. Darling et al., 2004) and TBA (Threaded Blockset Aligner) (Blanchette et al., 2004) represent some of the earliest efforts to align genome regions across multiple species. Vaughn and colleagues (Vaughn et al., 2022) recently used progressiveMauve (A. E. Darling et al., 2010) to align melon genomes and convert them into a genome graph for genotyping.

To bridge traditional alignment and graph construction, several intermediate tools have been developed. REVEAL (Recursive Exact-Matching Aligner) (Linthorst et al., 2015) employs a recursive exact-matching strategy to construct alignments, while tools like NovoGraph (Biederstedt et al., 2018) and Seq-seq-pan (Jandrasits et al., 2018) utilize progressive or block-based alignment strategies to scale MGAs to a large number of genomes. ProgressiveCactus (Armstrong et al., 2020; Paten et al., 2011) dramatically improves scalability using a guide-tree-based alignment strategy. Its output can be used as alignment input to the VG toolkit, enabling the inclusion of large duplications and inversions in yeast (Garrison et al., 2018). This approach provided the first workflow for converting an MGA into a graph that can be used both to infer

https://doi.org/10.1017/qpb.2025.10028 Published online by Cambridge University Press

genotype information and to map short reads. SibeliaZ (Minkin & Medvedev, 2020) generalized these ideas based on information from a de Bruijn graph to construct improved MGAs.

The Human Pangenome Reference Consortium (HPRC) (Liao et al., 2023) has greatly advanced the field by releasing an initial pangenome draft from 47 humans, constructed using methods like Minigraph, Minigraph-Cactus, and PGGB (Pangenome Graph Builder). Minigraph (H. Li et al., 2020) extended the minimap2 chaining algorithm to progressively add large SVs (>50 bp) into the graph. Minigraph-Cactus (Hickey et al., 2024) recruits the graph from Minigraph as a backbone. It then adds base-level alignments after clipping sequences that are highly divergent from a chosen reference sequence ('clipping' is the technical term for removing the portion of a read that cannot be confidently aligned to the target genome). The details of these graphs will depend on the order of the input of sequences or the divergence between samples in the collection of genomes (Garrison & Guarracino, 2023), but it simplifies the graph structure and makes the graph suitable for downstream genotyping tasks. Similarly, ACMGA (AnchorWave-Cactus Multiple Genome Alignment) (H. Zhou et al., 2024) combines cactus with AnchorWave, which improves the alignment of long repetitive sequences in the plant genomes, for detection of large SVs (Song et al., 2022). Huijse and colleagues (Huijse et al., 2023) found that AnchorWave outperformed Minigraph-Cactus in in producing alignments in the highly divergent MHC region of human genomes. The Pangenome Graph Builder (PGGB) (Garrison et al., 2024) tries to capture all variations in the input sequences by constructing and all-to-all genome alignment by wfmash and rendering it with seqwish (Garrison & Guarracino, 2023) and gffaix, then further consensus with smoothxg. While this approach offers a comprehensive representation of variations, the computational demands of all-to-all alignments are substantial. Instead of building a whole genome graph, PGR-TK (PanGenomic Research Took Kit) (Chin et al., 2023) rapidly constructs subgraphs of specific regions using data structures designed for long-read assembly (Chin & Khalak, 2019; H. Li, 2016); it was shown to be very fast in rebuilding the complex variations in MHC haplotypes, though its use demands substantial expertise for parameter tuning and result interpretation.

Scalable alternatives to whole-genome alignments

Over the past decade, the complexity and scalability challenges of constructing and querying large genome graphs have become increasingly apparent. As a result, researchers have explored pangenomes using analyses based on specific sequence blocks—such as orthologous genes or k-mers—rather than base-resolution DNA sequences. Different strategies have been

https://doi.org/10.1017/qpb.2025.10028 Published online by Cambridge University Press

developed to make pangenome analyses more scalable, each with its own trade-offs. K-mer based approaches are computationally efficient, making them attractive for large-scale comparisons. However, they sacrifice sequence context and struggle to distinguish between repeats, particularly in complex eukaryotic genomes. In contrast, gene-based methods are more interpretable and extensible across genomes but depend heavily on good gene annotation. Annotation quality in turn is dependent on a range of factors, such as the availability of RNA and proteomics data, whether a genome is from a taxon that contains other well-annotated genomes and so on. The good news is that ever more comprehensive sampling, at the level of individuals (tissues and conditions), populations, species and higher-order groupings will undoubtedly improve gene annotations.

In bacterial pangenomics, gene presence–absence matrices generated by orthogroup clustering with OrthoMCL have been the standard (Contreras-Moreira & Vinuesa, 2013; L. Li et al., 2003; Page et al., 2015). This strategy was subsequently extended by incorporating gene graphs in tools like PPanGGOLiN (Gautreau et al., 2020) and Panaroo with partitioned and fixed annotation error (Tonkin-Hill et al., 2020). GCB (Genome Complexity Browser) visualized and quantified variability with orthogroup inference (Manolov et al., 2020). PanPA constructs graphs based on protein sequence alignments (Dabbaghie et al., 2023), and Pangene leverages rapid protein alignments to build gene graphs for eukaryotic genomes, enabling analysis of gene copy number changes and orientation—remarkably, it can build a graph from 100 human haplotypes in under one minute (H. Li et al., 2024).

Although implicit graphs constructed from fixed k-mers provide a valuable snapshot of genomic diversity, their resolution is inherently limited, and other tools have taken different routes. PanTools (Sheikhizadeh et al., 2016) detects homology groups with k-mers and builds a database for pan-proteome query, while PanKmer (Aylward et al., 2023) and Panagram (Benoit et al., 2025) decompose assembled genomes into a k-mer database with further ability to locate specific positions in assemblies. Furthermore, methods like Biforst (Holley & Melsted, 2020) and mdBG (Ekim et al., 2021) efficiently construct de Bruijn graphs for storage and rapid querying; they can be applied to genotype variable tandem repeats with short reads (Lu et al., 2021), though they fall short in accurately representing complete loci for downstream analyses (Andreace et al., 2023).

https://doi.org/10.1017/qpb.2025.10028 Published online by Cambridge University Press

Variant calling in the graph era

Once a pangenome graph is constructed, it can serve as an enhanced reference for genotyping resequenced samples—either by aligning reads or matching k-mers—capturing a broader range of sequence variation than linear references. While many current tools rely on read mapping or k-mer comparison to identify SNPs and structural variants, some have advanced to support haplotype reconstruction and the detection of novel variants—capabilities that are particularly effective with long-read resequencing.

Among these, one of the most widely adopted and versatile tools is the Variation Graph Toolkit (VG), which provides a comprehensive framework for mapping, small variant calling, and structural variant genotyping. VG has become popular since its first open-source release (Garrison et al., 2018; Hickey et al., 2020). It also reduces reference bias in ancient samples (Martiniano et al., 2020). Another VG module, Giraffe (Sirén et al., 2021), was developed as successor of VG map to accelerate the process for large-scale genotyping. PHG (Practical Haplotype Graph) utilizes established tools for mapping against linear references (e.g., GATK) for genotyping in the offspring of crops (Bradbury et al., 2022). DRAGEN (Dynamic Read Analysis for GENomics) (Behera et al., 2024) is currently the fastest for mapping and genotyping against pangenome references, exploiting hardware acceleration and tricks from machine learning, but it requires a commercial license. Apart from directly mapping to a graph, mapping reads to multiple references first, and then injecting them into a graph based on mapping coordinates is another direction; one example is Gfa2bin (Vorbrugg, Bezrukov, Bao, Xian, et al., 2024) and cosigt (Bolognini et al., 2024), which uses node coverage across multiple references for genotyping with mapping by bwa (H. Li, 2013). Such approaches benefit from the maturity of linear reference mapping and the compatibility of their outputs with downstream graph-based analyses. Mapping long reads directly to genome graphs has become increasingly viable. Graphaligner (Rautiainen & Marschall, 2020) is the first tool to achieve long-read mapping to a graph with a seed-and-extend strategy, with much higher speed than VG. Minigraph (H. Li et al., 2020) can find approximate mapping locations without base-level alignment, while Minichain (Chandra et al., 2024) introduces a recombination penalty for long reads mapping to the graph.

To sidestep the computational cost of full mapping, many tools employ k-mer comparison strategies that match sequencing reads to known variants encoded in the graph. PanGenie (Ebler et al., 2022) and KAGE (Grytten et al., 2022, 2023) compare k-mers from reads to a pangenome graph to reduce run time and mapping bias. EVG (Ensemble Variant Genotyper)

https://doi.org/10.1017/qpb.2025.10028 Published online by Cambridge University Press

(Du et al., 2024) is a framework designed to standardize the performance of various genotyping tools by accounting for the genomic features specific to plant species. Varigraph (Du et al., 2025) further optimized the k-mer-based approach with memory efficiency and extended the model for dosage estimation in autopolyploid genomes. A drawback is that these tools only genotype the known variations independently and thus cannot reconstruct the haplotypes in the population. To address this gap, Locityper (Prodanov et al., 2024) and cosigt (Bolognini et al., 2024) have been developed to utilize read alignment profiles to locate the closest haplotype in the graph.

Furthermore, structural variant (SV) calling directly from pangenome graphs remains a critical challenge. To overcome the issues, SVarp (Soylev et al., 2024) tackles this by locally assembling potential SV alleles from long-read data, while PALSS (Denti et al., 2025) augments the graph with the consensus from sample-specific long reads without mapping.

In summary, the field of pangenome graph construction is dynamic, with no single tool dominating; the optimal tool choice depends on the specific research objectives and the desired resolution. Reference-based variation graphs, for instance, facilitate population genetics analyses across extensive cohorts but may omit certain genomic variations. Tools like PGGB offer comprehensive graph representations; however, their complexity can pose challenges for downstream applications such as VG Giraffe alignment, necessitating tailored pruning strategies for effective read mapping. Notably, developing and benchmarking efforts have predominantly centered on human genomics. Given that non-human species, including plant genomes, are often much more diverse than human genomes, there is a need for expanded evaluation across diverse species of tools for the building and use of pangenomes.

Functional pangenomics: linking variation to mechanism

Reference bias not only affects variant discovery. Its shortcomings have knock-on effects in downstream functional analyses, including the comparison of chromatin accessibility, gene expression, or DNA methylation (Galli et al., 2025; Igolkina et al., 2025). Compared to the growing adoption of genome graphs for structural variation calling and genotyping, much more still needs to be done to take advantage of graph-based frameworks for functional genomics.

Grytten et.al (Grytten et al., 2019) implemented Graph Peak Caller to identify ChIP-seq peaks using a variation graph in *A. thaliana*, identifying more than twice as many base pairs absent from the linear reference than had been found with previous methods. DNA methylation studies

https://doi.org/10.1017/qpb.2025.10028 Published online by Cambridge University Press

have revealed analogous benefits—and also underscore the extent of reference bias in functional assays. In cattle, using the wrong reference genome can lead to substantial errors in methylation quantification, with up to ~2% global bias and large numbers of methylated cytosines being affected by breed-specific variation (MacPhillamy et al., 2024). In *Arabidopsis thaliana*, methylation profiling was even more sensitive to reference choice, with only ~88% sites being consistent between reference and focal strain, with one major reason being that transposable elements, which are prime targets of DNA methylation, have been much more active in this species than, for example, in humans (Igolkina et al., 2025). To address this, methylGrapher (Zhang et al., 2025) introduced the first graph-based approach for mapping bisulfite sequencing data. Compared to traditional methods such as Bismark, it uniquely identified 2.2~2.9 million ^{m}CpGs across five human samples, many of which were absent from the reference or misclassified as unmethylated before.

Reference bias also affects RNA-seq analysis. In *Arabidopsis thaliana*, expression estimates diverged for a subset of genes depending on whether reads were mapped to the reference genome or to the accession's own genome; these genes were strongly enriched for transposable elements and copy number–variable loci (Igolkina et al., 2025). Similar trends were observed in barley, but at an even higher rate, where mapping transcriptomic reads to a pan-transcriptome built with 20 genotypes improved the mapping rate by around 11% compared to a single linear reference (W. Guo et al., 2025). VG rpvg (Sibbesen et al., 2023) extends genome graph approaches to RNA-seq analysis by building spliced pangenome graphs and quantifying expression along haplotype-resolved paths (Sibbesen et al., 2023). These methods improve accuracy and enable haplotype-specific quantification, even without prior haplotype phasing, but they are ideally based on comprehensive pan-transcriptome annotation, which is absent in most species. Haplotype information in turn is immensely useful in outbred species, and perhaps even more so, in polyploid species with their complex allele ratios (Bao et al., 2022; Bird et al., 2025; Du et al., 2025).

Despite these advances, graph-based approaches to functional genomics remain in their infancy. Few tools have been developed, and most remain proof-of-concept applications limited to model species. Even where tools exist, broader adoption has been slow, partly due to the lack of comprehensive functional annotations and the complexity of graph-aware analytical workflows. Expanding these approaches across multiple omics layers—including methylation, expression, chromatin states, and chromatin accessibility—and to diverse species with more complex genomes remains a critical challenge for future research (Fig. 2).

https://doi.org/10.1017/qpb.2025.10028 Published online by Cambridge University Press

Navigating the tangled graph: visualization, comparison, and scalability

Although there are multiple strategies for graph construction, most approaches now adopt the Graphical Fragment Assembly (GFA) format to store graph information. Unfortunately, querying large-scale pangenomes remains challenging due to the inherent complexity and enormous size of these graphs. For instance, the VG toolkit offers a versatile suite of functions to construct, convert, and manipulate genome graphs, but even with VG, extracting information from Gb-scale pangenomes can be nontrivial. To overcome scalability issues, several specialized tools have been developed (Fig. 3a). ODGI (Optimized Dynamic Genome/Graph Implementation) (Guarracino et al., 2022) implements scalable algorithms to visualize graphs at multiple resolutions, to extract specific loci, and to compare path similarities. Meanwhile, tools such as Gretl (Vorbrugg, Bezrukov, Bao, & Weigel, 2024) are designed to evaluate the quality of multiple graphs by providing a range of quantitative metrics for graph description and comparison. PANCAT (Dubois et al., 2025) characterizes differences among variation graphs derived from the same sequence set using edit distance metrics.

On the visualization side, early GUI-based tools like Bandage (Wick et al., 2015) and GfaViz (Gonnella et al., 2019) provide whole-graph views of assembly graphs but are limited when it comes to base-level or Gb-scale pangenome graphs. VG view and VG viz can display sequences up to about 100 kb, whereas SequenceTubemap (Beyer et al., 2019) adopts an intuitive visualization model (inspired by public transport network maps) to display variation graphs along with read mappings at the appropriate scale. Momi-G (Yokoyama et al., 2019) extends this concept for large-scale structural variant inspection in human variation graphs, and ODGI viz further expands on the VG viz layout by exporting rasterized images suitable for chromosome-scale genome graphs.

Efforts to integrate graph layouts with functional annotation are also emerging. For example, VRPG, a visualization and interpretation framework for linear reference–projected pangenome graphs (Miao & Yue, 2025), extracts subgraphs based on reference path coordinates and annotations, while PPanG (M. Liu et al., 2024) adapts the SequenceTubemap framework to display multiple genome annotations through embedded JBrowse2 components in real time. Additionally, Gfaestus (https://github.com/chfi/gfaestus) leverages GPU frameworks to visualize full graphs from projects like HPRC, and waragraph (https://github.com/chfi/waragraph) can integrate annotation information into ODGI layouts interactively.

https://doi.org/10.1017/qpb.2025.10028 Published online by Cambridge University Press

Compared to graph construction, the visualization and comparison of pangenome graphs have lagged significantly. While multiple tools exist for assembling and processing variation graphs, there is still no comprehensive, scalable, and interactive visualization framework that can handle large-scale pangenomes efficiently and connect the functional annotation (Fig. 3b). As the pangenome expands to hundreds of individuals rapidly, it could even go beyond species, and extracting biological knowledge from the complex tangles in graphs requires better tools than what is currently available.

Conclusion and perspectives

The development of eukaryotic pangenomics has entered a transformative phase. Advances in sequencing technologies and assembly algorithms have made it feasible to generate high-quality genomes at population scale (Antipov et al., 2025; H. Cheng et al., 2024; Koren et al., 2024). As a result, pangenome references constructed from tens to hundreds of assemblies now exist for a growing number of species, including foundational species such as *Arabidopsis thaliana* (Kang et al., 2023; Lian et al., 2024; Wlodzimierz et al., 2023), key crops (L. Cheng et al., 2025; D. Guo et al., 2025; Hufford et al., 2021; Y. Liu et al., 2020; Lynch et al., 2025; Y. Zhou et al., 2022) and humans (Liao et al., 2023). Its application has shown that additional variations capture some of the heritability previously missed (Y. Zhou et al., 2022), find more associations between variations and agronomic traits (Hufford et al., 2021)), and can uncover the complex evolution history of well-studied loci (Bolognini et al., 2024).

Nevertheless, capturing the full spectrum of variation across a species in an unbiased and comprehensive way remains a challenge. While tools such as Minigraph-Cactus use iterative construction to simplify the process of graph alignment, they are sensitive to input order and tend to discard sequences that diverge too much from the reference — an issue especially problematic in high diversity species (L. Cheng et al., 2025; Garrison & Guarracino, 2023). On the other hand, all-to-all alignment approaches, such as PGGB, provide more complete graphs but require substantial computational resources, making them impractical for datasets involving hundreds of genomes (Lynch et al., 2025). Similarly, genotyping tools face scalability constraints for large graphs: VG Giraffe, for instance, typically downsamples to 64 haplotypes prior to mapping (Sirén et al., 2021).

Progress in these areas depends on the availability of high-quality benchmark datasets for validation. Yet such resources are scarce in non-human species, and even in human genomics,

https://doi.org/10.1017/qpb.2025.10028 Published online by Cambridge University Press

benchmarking is often confined to a few well-characterized individuals (Dwarshuis et al., 2024). This creates systematic bias and limits our ability to assess how well genome graphs capture rare, complex, or population-specific variation. Developing robust metrics and comparative frameworks to evaluate graph quality remains a crucial direction for the field.

Moreover, the continued reliance on biallelic SNP models restricts the development of population genetic theory capable of explaining the full complexity of pangenomic variation. However, structural variants (SVs) are generated by a wide range of distinct mutational mechanisms—including non-homologous end joining (NHEJ), non-allelic homologous recombination (NAHR), template-switching, nested transposon insertions, and tandem repeat expansion—that often result in multi-allelic loci rather than simple binary variants (Collins & Talkowski, 2025). In association studies, the assumption of biallelic variation also introduces confounding effects, particularly in the presence of genetic heterogeneity (H.-J. Liu et al., 2024). Incorporating haplotype-aware models may reveal additional associations that are otherwise missed due to the underlying complexity of structural variants and the impact of multi-allelic loci (Smith et al., 2025; Y. Zhou et al., 2022). As such, there is a growing need to develop new population genetic models.

Looking ahead, integrating pangenome graphs with evolutionary models remains a wide-open frontier. Ancient hybridization, incomplete lineage sorting, and structural rearrangements complicate cross-species alignment, increasing the difficulty of graph construction. Yet global biodiversity sequencing projects (Lewin et al., 2018) are beginning to fill the tree of life with genome assemblies. Embedding phylogenetic history directly into graph construction—rather than treating it as a downstream layer—may help generate more meaningful, interpretable graphs.

In contrast to advances in variant calling, the interpretation, visualization, and benchmarking of pangenome graphs are still in early stages. There is still no equivalent of IGV for intuitive graph exploration, and widely adopted formats for complex variant representation are lacking. While tools like ODGI offer useful summaries and visualizations, they lack interactivity and scalability for Gb-scale graphs. Even more critically, the integration of functional genomics with graph frameworks remains far behind. At present, RNA-seq, methylation, and chromatin accessibility data cannot be seamlessly analyzed in graph-aware contexts. Bridging this gap will require unified, scalable methods for aligning and interpreting multi-omic data within graph-based references for genotype-phenotype association.

https://doi.org/10.1017/qpb.2025.10028 Published online by Cambridge University Press

As we enter the next phase of pangenomic research, the field faces substantial computational and modeling hurdles. Yet the growing ecosystem of graph-based methods offers more than just an ever-expanding toolkit—it provides the foundation for a new paradigm in genomics. By embracing the full complexity of genomic variation, pangenome graphs have the potential to reshape how we conduct association studies, trace evolutionary history, and interpret regulatory landscapes. Moving beyond reference bias and linear constraints, these graphs can unify population-scale diversity, functional readouts, and comparative signals across the tree of life. Realizing this vision will require not only scalable tools and new theoretical frameworks but also sustained community efforts in benchmarking, visualization, and data integration. Still, the promise is profound: to build genomic models that reflect biological reality more accurately, and in doing so, to understand evolution in unprecedented detail.

Acknowledgement

We thank both our colleagues from the 1001 Genomes Project in the Weigel and Nordborg labs and the participants of the International Genome Graph Symposium (IGGSy'24) in Ascona 2024 for inspiration, and especially Andrea Guarracino and Erik Garrison from the University of Tennessee Health Science Center for extensive discussions.

Funding

Our work is supported by the International Max Planck Research School (IMPRS) “From Molecules to Organisms” (Z.B.), the Novozymes Prize of the Novo Nordisk Foundation (D.W.) and the Max Planck Society.

Authorship contributions

Z.B. wrote the draft of the manuscript. Z.B. and D.W. revised the manuscript.

Competing interests

D.W. holds equity in Computomics, which advises plant breeders. D.W. also consults for KWS SE, a globally active plant breeder and seed producer.

https://doi.org/10.1017/qpb.2025.10028 Published online by Cambridge University Press

References

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

Antipov, D., Rautiainen, M., Nurk, S., Walenz, B. P., Solar, S. J., Phillippy, A. M., & Koren, S. (2025). Verkko2 integrates proximity-ligation data with long-read De Bruijn graphs for efficient telomere-to-telomere genome assembly, phasing, and scaffolding. *Genome Research*, *35*(7), 1583–1594.

Armstrong, J., Hickey, G., Diekhans, M., Fiddes, I. T., Novak, A. M., Deran, A., Fang, Q., Xie, D., Feng, S., Stiller, J., Genereux, D., Johnson, J., Marinescu, V. D., Alföldi, J., Harris, R. S., Lindblad-Toh, K., Haussler, D., Karlsson, E., Jarvis, E., … Paten, B. (2020). Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature*, *587*, 246–251.

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

Bao, Z., Li, C., Li, G., Wang, P., Peng, Z., Cheng, L., Li, H., Zhang, Z., Li, Y., Huang, W., Ye, M., Dong, D., Cheng, Z., VanderZaag, P., Jacobsen, E., Bachem, C. W. B., Dong, S., Zhang, C., Huang, S., & Zhou, Q. (2022). Genome architecture and tetrasomic inheritance of autotetraploid potato. *Molecular Plant*, *15*(7), 1211–1226.

Behera, S., Catreux, S., Rossi, M., Truong, S., Huang, Z., Ruehle, M., Visvanath, A., Parnaby, G., Roddey, C., Onuchic, V., Finocchio, A., Cameron, D. L., English, A., Mehtalia, S., Han, J., Mehio, R., & Sedlazeck, F. J. (2024). Comprehensive genome analysis and variant detection at scale using DRAGEN. *Nature Biotechnology*. https://doi.org/10.1038/s41587-024-02382-1

Benoit, M., Jenike, K. M., Satterlee, J. W., Ramakrishnan, S., Gentile, I., Hendelman, A., Passalacqua, M. J., Suresh, H., Shohat, H., Robitaille, G. M., Fitzgerald, B., Alonge, M.,

https://doi.org/10.1017/qpb.2025.10028 Published online by Cambridge University Press

Wang, X., Santos, R., He, J., Ou, S., Golan, H., Green, Y., Swartwood, K., … Lippman, Z. B. (2025). Solanum pan-genetics reveals paralogues as contingencies in crop engineering. *Nature*, 1–11.

Beyer, W., Novak, A. M., Hickey, G., Chan, J., Tan, V., Paten, B., & Zerbino, D. R. (2019). Sequence tube maps: making graph genomes intuitive to commuters. *Bioinformatics (Oxford, England)*, *35*(24), 5318–5320.

Biederstedt, E., Oliver, J. C., Hansen, N. F., Jajoo, A., Dunn, N., Olson, A., Busby, B., & Dilthey, A. T. (2018). NovoGraph: Human genome graph construction from multiple long-read de novo assemblies. *F1000Research*, *7*, 1391.

Bird, K. A., Brock, J. R., Grabowski, P. P., Harder, A. M., Healy, A. L., Shu, S., Barry, K., Boston, L., Daum, C., Guo, J., Lipzen, A., Walstead, R., Grimwood, J., Schmutz, J., Lu, C., Comai, L., McKay, J. K., Pires, J. C., Edger, P. P., … Kliebenstein, D. J. (2025). Allopolyploidy expanded gene content but not pangenomic variation in the hexaploid oilseed Camelina sativa. *Genetics*, *229*(1), 1–44.

Blanchette, M., Kent, W. J., Riemer, C., Elnitski, L., Smit, A. F. A., Roskin, K. M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E. D., Haussler, D., & Miller, W. (2004). Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Research*, *14*(4), 708–715.

Bolognini, D., Halgren, A., Lou, R. N., Raveane, A., Rocha, J. L., Guarracino, A., Soranzo, N., Chin, C.-S., Garrison, E., & Sudmant, P. H. (2024). Recurrent evolution and selection shape structural diversity at the amylase locus. *Nature*, 1–9.

Bradbury, P. J., Casstevens, T., Jensen, S. E., Johnson, L. C., Miller, Z. R., Monier, B., Romay, M. C., Song, B., & Buckler, E. S. (2022). The Practical Haplotype Graph, a platform for storing and using pangenomes for imputation. *Bioinformatics (Oxford, England)*, *38*(15), 3698–3702.

Cao, J., Schneeberger, K., Ossowski, S., Günther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C.,

https://doi.org/10.1017/qpb.2025.10028 Published online by Cambridge University Press

Stegle, O., Lippert, C., Wang, X., Ott, F., Müller, J., Alonso-Blanco, C., Borgwardt, K., Schmid, K. J., & Weigel, D. (2011). Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nature Genetics*, *43*(10), 956–963.

Chandra, G., Gibney, D., & Jain, C. (2024). Haplotype-aware sequence alignment to pangenome graphs. *Genome Research*, gr.279143.124.

Cheng, H., Asri, M., Lucas, J., Koren, S., & Li, H. (2024). Scalable telomere-to-telomere assembly for diploid and polyploid genomes with double graph. *Nature Methods*, *21*(6), 967–970.

Cheng, L., Wang, N., Bao, Z., Zhou, Q., Guarracino, A., Yang, Y., Wang, P., Zhang, Z., Tang, D., Zhang, P., Wu, Y., Zhou, Y., Zheng, Y., Hu, Y., Lian, Q., Ma, Z., Lassois, L., Zhang, C., Lucas, W. J., … Huang, S. (2025). Leveraging a phased pangenome for haplotype design of hybrid potato. *Nature*, *640*(8058), 408–417.

Chin, C.-S., Behera, S., Khalak, A., Sedlazeck, F. J., Sudmant, P. H., Wagner, J., & Zook, J. M. (2023). Multiscale analysis of pangenomes enables improved representation of genomic diversity for repetitive and clinically relevant genes. *Nature Methods*. https://doi.org/10.1038/s41592-023-01914-y

Chin, C.-S., & Khalak, A. (2019). Human Genome Assembly in 100 Minutes. In *bioRxiv*. bioRxiv. https://doi.org/10.1101/705616

Clark, R. M., Schweikert, G., Toomajian, C., Ossowski, S., Zeller, G., Shinn, P., Warthmann, N., Hu, T. T., Fu, G., Hinds, D. A., Chen, H., Frazer, K. A., Huson, D. H., Schölkopf, B., Nordborg, M., Rätsch, G., Ecker, J. R., & Weigel, D. (2007). Common sequence polymorphisms shaping genetic diversity in Arabidopsis thaliana. *Science*, *317*(5836), 338–342.

Collins, R. L., & Talkowski, M. E. (2025). Diversity and consequences of structural variation in the human genome. *Nature Reviews. Genetics*, *26*(7), 443–462.

Contreras-Moreira, B., & Vinuesa, P. (2013). GET_HOMOLOGUES, a versatile software

package for scalable and robust microbial pangenome analysis. *Applied and Environmental Microbiology*, *79*(24), 7696–7701.

Dabbaghie, F., Srikakulam, S. K., Marschall, T., & Kalinina, O. V. (2023). PanPA: generation and alignment of panproteome graphs. *Bioinformatics Advances*, *3*(1), vbad167.

Darling, A. C. E., Mau, B., Blattner, F. R., & Perna, N. T. (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Research*, *14*(7), 1394–1403.

Darling, A. E., Mau, B., & Perna, N. T. (2010). progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *Plos One*, *5*(6), e11147.

Denti, L., Bonizzoni, P., Brejova, B., Chikhi, R., Krannich, T., Vinar, T., & Hormozdiari, F. (2025). Pangenome graph augmentation from unassembled long reads. In *bioRxiv*. https://doi.org/10.1101/2025.02.07.637057

Dilthey, A., Cox, C., Iqbal, Z., Nelson, M. R., & McVean, G. (2015). Improved genome inference in the MHC using a population reference graph. *Nature Genetics*, *47*(6), 682–688.

Dubois, S., Zytnicki, M., Lemaitre, C., & Faraut, T. (2025). Pairwise graph edit distance characterizes the impact of the construction method on pangenome graphs. *Bioinformatics (Oxford, England)*, *41*(6), btaf291.

Du, Z.-Z., He, J.-B., & Jiao, W.-B. (2024). A comprehensive benchmark of graph-based genetic variant genotyping algorithms on plant genomes for creating an accurate ensemble pipeline. *Genome Biology*, *25*(1), 91.

Du, Z.-Z., He, J.-B., Xiao, P.-X., Hu, J., Yang, N., & Jiao, W.-B. (2025). Varigraph: An accurate and widely applicable pangenome graph-based variant genotyper for diploid and polyploid genomes. *Molecular Plant*. https://doi.org/10.1016/j.molp.2025.08.001

Dwarshuis, N., Kalra, D., McDaniel, J., Sanio, P., Alvarez Jerez, P., Jadhav, B., Huang, W. E., Mondal, R., Busby, B., Olson, N. D., Sedlazeck, F. J., Wagner, J., Majidian, S., & Zook, J. M. (2024). The GIAB genomic stratifications resource for human reference genomes. *Nature Communications*, *15*(1), 9029.

https://doi.org/10.1017/qpb.2025.10028 Published online by Cambridge University Press

Ebler, J., Ebert, P., Clarke, W. E., Rausch, T., Audano, P. A., Houwaart, T., Mao, Y., Korbel, J. O., Eichler, E. E., Zody, M. C., Dilthey, A. T., & Marschall, T. (2022). Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Nature Genetics*, *54*(4), 518–525.

Eggertsson, H. P., Jonsson, H., Kristmundsdottir, S., Hjartarson, E., Kehr, B., Masson, G., Zink, F., Hjorleifsson, K. E., Jonasdottir, A., Jonasdottir, A., Jonsdottir, I., Gudbjartsson, D. F., Melsted, P., Stefansson, K., & Halldorsson, B. V. (2017). Graphtyper enables population-scale genotyping using pangenome graphs. *Nature Genetics*, *49*(11), 1654–1660.

Ekim, B., Berger, B., & Chikhi, R. (2021). Minimizer-space de Bruijn graphs: Whole-genome assembly of long reads in minutes on a personal computer. *Cell Systems*, *12*(10), 958–968.e6.

Galli, M., Chen, Z., Ghandour, T., Chaudhry, A., Gregory, J., Feng, F., Li, M., Schleif, N., Zhang, X., Dong, Y., Song, G., Walley, J. W., Chuck, G., Whipple, C., Kaeppler, H. F., Huang, S.-S. C., & Gallavotti, A. (2025). Transcription factor binding divergence drives transcriptional and phenotypic variation in maize. *Nature Plants*, *11*(6), 1205–1219.

Gan, X., Stegle, O., Behr, J., Steffen, J. G., Drewe, P., Hildebrand, K. L., Lyngsoe, R., Schultheiss, S. J., Osborne, E. J., Sreedharan, V. T., Kahles, A., Bohnert, R., Jean, G., Derwent, P., Kersey, P., Belfield, E. J., Harberd, N. P., Kemen, E., Toomajian, C., … Mott, R. (2011). Multiple reference genomes and transcriptomes for Arabidopsis thaliana. *Nature*, *477*(7365), 419–423.

Gao, L., Gonda, I., Sun, H., Ma, Q., Bao, K., Tieman, D. M., Burzynski-Chang, E. A., Fish, T. L., Stromberg, K. A., Sacks, G. L., Thannhauser, T. W., Foolad, M. R., Diez, M. J., Blanca, J., Canizares, J., Xu, Y., van der Knaap, E., Huang, S., Klee, H. J., … Fei, Z. (2019). The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nature Genetics*, *51*(6), 1044–1051.

Garrison, E., & Guarracino, A. (2023). Unbiased pangenome graphs. *Bioinformatics* , *39*(1), 14–

https://doi.org/10.1017/qpb.2025.10028 Published online by Cambridge University Press

19.

Garrison, E., Guarracino, A., Heumos, S., Villani, F., Bao, Z., Tattini, L., Hagmann, J., Vorbrugg, S., Marco-Sola, S., Kubica, C., Ashbrook, D. G., Thorell, K., Rusholme-Pilcher, R. L., Liti, G., Rudbeck, E., Golicz, A. A., Nahnsen, S., Yang, Z., Mwaniki, M. N., … Prins, P. (2024). Building pangenome graphs. *Nature Methods*, *21*(11), 2008–2012.

Garrison, E., Sirén, J., Novak, A. M., Hickey, G., Eizenga, J. M., Dawson, E. T., Jones, W., Garg, S., Markello, C., Lin, M. F., Paten, B., & Durbin, R. (2018). Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology*, *December 2017*. https://doi.org/10.1038/nbt.4227

Gautreau, G., Bazin, A., Gachet, M., Planel, R., Burlot, L., Dubois, M., Perrin, A., Médigue, C., Calteau, A., Cruveiller, S., Matias, C., Ambroise, C., Rocha, E. P. C., & Vallenet, D. (2020). PPanGGOLiN: Depicting microbial diversity via a partitioned pangenome graph. *PLoS Computational Biology*, *16*(3), e1007732.

Gonnella, G., Niehus, N., & Kurtz, S. (2019). GfaViz: flexible and interactive visualization of GFA sequence graphs. *Bioinformatics (Oxford, England)*, *35*(16), 2853–2855.

Grytten, I., Dagestad Rand, K., & Sandve, G. K. (2022). KAGE: fast alignment-free graph-based genotyping of SNPs and short indels. *Genome Biology*, *23*(1), 209.

Grytten, I., Rand, K. D., Nederbragt, A. J., Storvik, G. O., Glad, I. K., & Sandve, G. K. (2019). Graph Peak Caller: Calling ChIP-seq peaks on graph-based reference genomes. *PLoS Computational Biology*, *15*(2), e1006731.

Grytten, I., Rand, K. D., & Sandve, G. K. (2023). KAGE 2: Fast and accurate genotyping of structural variation using pangenomes. In *bioRxiv* (p. 2023.12.23.572333). https://doi.org/10.1101/2023.12.23.572333

Guarracino, A., Heumos, S., Nahnsen, S., Prins, P., & Garrison, E. (2022). ODGI: understanding pangenome graphs. *Bioinformatics (Oxford, England)*, *38*(13), 3319–3326.

Guo, D., Li, Y., Lu, H., Zhao, Y., Kurata, N., Wei, X., Wang, A., Wang, Y., Zhan, Q., Fan, D.,

https://doi.org/10.1017/qpb.2025.10028 Published online by Cambridge University Press

Zhou, C., Lu, Y., Tian, Q., Weng, Q., Feng, Q., Huang, T., Zhang, L., Gu, Z., Wang, C., … Han, B. (2025). A pangenome reference of wild and cultivated rice. *Nature*, *642*(8068), 662–671.

Guo, W., Schreiber, M., Marosi, V. B., Bagnaresi, P., Jørgensen, M. E., Braune, K. B., Chalmers, K., Chapman, B., Dang, V., Dockter, C., Fiebig, A., Fincher, G. B., Fricano, A., Fuller, J., Haaning, A., Haberer, G., Himmelbach, A., Jayakodi, M., Jia, Y., … Waugh, R. (2025). A barley pan-transcriptome reveals layers of genotype-dependent transcriptional complexity. *Nature Genetics*, *57*(2), 441–450.

Hickey, G., Heller, D., Monlong, J., Sibbesen, J. A., Sirén, J., Eizenga, J., Dawson, E. T., Garrison, E., Novak, A. M., & Paten, B. (2020). Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biology*, *21*(1), 35.

Hickey, G., Monlong, J., Ebler, J., Novak, A. M., Eizenga, J. M., Gao, Y., Human Pangenome Reference Consortium, Marschall, T., Li, H., & Paten, B. (2024). Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nature Biotechnology*, *42*(4), 663–673.

Holley, G., & Melsted, P. (2020). Bifrost: highly parallel construction and indexing of colored and compacted de Bruijn graphs. *Genome Biology*, *21*(1), 249.

Hufford, M. B., Seetharam, A. S., Woodhouse, M. R., Chougule, K. M., Ou, S., Liu, J., Ricci, W. A., Guo, T., Olson, A., Qiu, Y., Della Coletta, R., Tittes, S., Hudson, A. I., Marand, A. P., Wei, S., Lu, Z., Wang, B., Tello-Ruiz, M. K., Piri, R. D., … Dawe, R. K. (2021). De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science (New York, N.Y.)*, *373*(6555), 655–662.

Huijse, L., Adams, S. M., Burton, J. N., David, J. K., Julian, R. S., Meshulam-Simon, G., Mickalide, H., Tafesse, B. D., Calonga-Solís, V., Wolf, I. R., Morrison, A. J., Augusto, D. G., & Endlich, S. (2023). A pan-MHC reference graph with 246 fully contiguous phased sequences. In *bioRxiv* (p. 2023.09.01.555813). https://doi.org/10.1101/2023.09.01.555813

https://doi.org/10.1017/qpb.2025.10028 Published online by Cambridge University Press

Igolkina, A. A., Vorbrugg, S., Rabanal, F. A., Liu, H.-J., Ashkenazy, H., Kornienko, A. E., Fitz, J., Collenberg, M., Kubica, C., Mollá Morales, A., Jaegle, B., Wrightsman, T., Voloshin, V., Bezlepsky, A. D., Llaca, V., Nizhynska, V., Reichardt, I., Bezrukov, I., Lanz, C., … Nordborg, M. (2025). A comparison of 27 Arabidopsis thaliana genomes and the path toward an unbiased characterization of genetic polymorphism. *Nature Genetics*, 1–13.

Jandrasits, C., Dabrowski, P. W., Fuchs, S., & Renard, B. Y. (2018). Seq-seq-pan: Building a computational pan-genome data structure on whole genome alignment. *BMC Genomics*, *19*(1), 47.

Kang, M., Wu, H., Liu, H., Liu, W., Zhu, M., Han, Y., Liu, W., Chen, C., Song, Y., Tan, L., Yin, K., Zhao, Y., Yan, Z., Lou, S., Zan, Y., & Liu, J. (2023). The pan-genome and local adaptation of Arabidopsis thaliana. *Nature Communications*, *14*(1), 6259.

Kim, S., Plagnol, V., Hu, T. T., Toomajian, C., Clark, R. M., Ossowski, S., Ecker, J. R., Weigel, D., & Nordborg, M. (2007). Recombination and linkage disequilibrium in Arabidopsis thaliana. *Nature Genetics*, *39*(9), 1151–1155.

Koren, S., Bao, Z., Guarracino, A., Ou, S., Goodwin, S., Jenike, K. M., Lucas, J., McNulty, B., Park, J., Rautiainen, M., Rhie, A., Roelofs, D., Schneiders, H., Vrijenhoek, I., Nijbroek, K., Nordesjo, O., Nurk, S., Vella, M., Lawrence, K. R., … Phillippy, A. M. (2024). Gapless assembly of complete human and plant chromosomes using only nanopore sequencing. *Genome Research*, *34*(11), 1919–1930.

Lee, C., Grasso, C., & Sharlow, M. F. (2002). Multiple sequence alignment using partial order graphs. *Bioinformatics (Oxford, England)*, *18*(3), 452–464.

Lewin, H. A., Robinson, G. E., Kress, W. J., Baker, W. J., Coddington, J., Crandall, K. A., Durbin, R., Edwards, S. V., Forest, F., Gilbert, M. T. P., Goldstein, M. M., Grigoriev, I. V., Hackett, K. J., Haussler, D., Jarvis, E. D., Johnson, W. E., Patrinos, A., Richards, S., Castilla-Rubio, J. C., … Zhang, G. (2018). Earth BioGenome Project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences of the United States of*

https://doi.org/10.1017/qpb.2025.10028 Published online by Cambridge University Press

America, *115*(17), 4325–4333.

Lian, Q., Huettel, B., Walkemeier, B., Mayjonade, B., Lopez-Roques, C., Gil, L., Roux, F., Schneeberger, K., & Mercier, R. (2024). A pan-genome of 69 Arabidopsis thaliana accessions reveals a conserved genome structure throughout the global species range. *Nature Genetics*, *56*(5), 982–991.

Liao, W.-W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., Lu, S., Lucas, J. K., Monlong, J., Abel, H. J., Buonaiuto, S., Chang, X. H., Cheng, H., Chu, J., Colonna, V., Eizenga, J. M., Feng, X., Fischer, C., Fulton, R. S., … Paten, B. (2023). A draft human pangenome reference. *Nature*, *617*(7960), 312–324.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. In *arXiv [q-bio.GN]*. arXiv. http://arxiv.org/abs/1303.3997

Li, H. (2016). Minimap and miniasm: Fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* , *32*(14), 2103–2110.

Li, H., Feng, X., & Chu, C. (2020). The design and construction of reference pangenome graphs with minigraph. *Genome Biology*, *21*(1), 265.

Li, H., Marin, M., & Farhat, M. R. (2024). Exploring gene content with pangene graphs. *Bioinformatics (Oxford, England)*, *40*(7), btae456.

Li, L., Stoeckert, C. J., & Roos, D. S. (2003). OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Research*. https://doi.org/10.1101/gr.1224503

Linthorst, J., Hulsman, M., Holstege, H., & Reinders, M. (2015). Scalable multi whole-genome alignment using recursive exact matching. In *bioRxiv*. https://doi.org/10.1101/022715

Li, R., Li, Y., Zheng, H., Luo, R., Zhu, H., Li, Q., Qian, W., Ren, Y., Tian, G., Li, J., Zhou, G., Zhu, X., Wu, H., Qin, J., Jin, X., Li, D., Cao, H., Hu, X., Blanche, H., … Wang, J. (2010). Building the sequence map of the human pan-genome. *Nature Biotechnology*, *28*(1), 57–63.

Liu, H.-J., Swarts, K., Xu, S., Yan, J., & Nordborg, M. (2024). On the contribution of genetic

https://doi.org/10.1017/qpb.2025.10028 Published online by Cambridge University Press

heterogeneity to complex traits. In *bioRxiv* (No. biorxiv;2024.03.27.586967v1). https://www.biorxiv.org/content/10.1101/2024.03.27.586967v1

Liu, M., Zhang, F., Lu, H., Xue, H., Dong, X., Li, Z., Xu, J., Wang, W., & Wei, C. (2024). PPanG: a precision pangenome browser enabling nucleotide-level analysis of genomic variations in individual genomes and their graph-based pangenome. *BMC Genomics*, *25*(1), 405.

Liu, Y., Du, H., Li, P., Shen, Y., Peng, H., Liu, S., Zhou, G.-A., Zhang, H., Liu, Z., Shi, M., Huang, X., Li, Y., Zhang, M., Wang, Z., Zhu, B., Han, B., Liang, C., & Tian, Z. (2020). Pan-Genome of Wild and Cultivated Soybeans. *Cell*, *182*(1), 162–176.e13.

Li, Y.-H., Zhou, G., Ma, J., Jiang, W., Jin, L.-G., Zhang, Z., Guo, Y., Zhang, J., Sui, Y., Zheng, L., Zhang, S.-S., Zuo, Q., Shi, X.-H., Li, Y.-F., Zhang, W.-K., Hu, Y., Kong, G., Hong, H.-L., Tan, B., … Qiu, L.-J. (2014). De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nature Biotechnology*, *32*(10), 1045–1052.

Long, Q., Rabanal, F. A., Meng, D., Huber, C. D., Farlow, A., Platzer, A., Zhang, Q., Vilhjálmsson, B. J., Korte, A., Nizhynska, V., Voronin, V., Korte, P., Sedman, L., Mandáková, T., Lysak, M. A., Seren, U., Hellmann, I., & Nordborg, M. (2013). Massive genomic variation and strong selection in Arabidopsis thaliana lines from Sweden. *Nature Genetics*, *45*(8), 884–890.

Lu, T.-Y., Human Genome Structural Variation Consortium, & Chaisson, M. J. P. (2021). Profiling variable-number tandem repeat variation across populations using repeat-pangenome graphs. *Nature Communications*, *12*(1), 4250.

Lynch, R. C., Padgitt-Cobb, L. K., Garfinkel, A. R., Knaus, B. J., Hartwick, N. T., Allsing, N., Aylward, A., Bentz, P. C., Carey, S. B., Mamerto, A., Kitony, J. K., Colt, K., Murray, E. R., Duong, T., Chen, H. I., Trippe, A., Harkess, A., Crawford, S., Vining, K., & Michael, T. P. (2025). Domesticated cannabinoid synthases amid a wild mosaic cannabis pangenome. *Nature*, 1–10.

MacPhillamy, C., Chen, T., Hiendleder, S., Williams, J. L., Alinejad-Rokny, H., & Low, W. Y.

https://doi.org/10.1017/qpb.2025.10028 Published online by Cambridge University Press

(2024). DNA methylation analysis to differentiate reference, breed, and parent-of-origin effects in the bovine pangenome era. *GigaScience*, *13*. https://doi.org/10.1093/gigascience/giae061

Manolov, A., Konanov, D., Fedorov, D., Osmolovsky, I., Vereshchagin, R., & Ilina, E. (2020). Genome Complexity Browser: Visualization and quantification of genome variability. *PLoS Computational Biology*, *16*(10), e1008222.

Martiniano, R., Garrison, E., Jones, E. R., Manica, A., & Durbin, R. (2020). Removing reference bias and improving indel calling in ancient DNA data analysis by mapping to a sequence variation graph. *Genome Biology*, *21*(1), 250.

Miao, Z., & Yue, J.-X. (2025). Interactive visualization and interpretation of pangenome graphs by linear reference-based coordinate projection and annotation integration. *Genome Research*, *35*(2), 296–310.

Minkin, I., & Medvedev, P. (2020). Scalable multiple whole-genome alignment and locally collinear block construction with SibeliaZ. *Nature Communications*, *11*(1), 6327.

Nordborg, M., Borevitz, J. O., Bergelson, J., Berry, C. C., Chory, J., Hagenblad, J., Kreitman, M., Maloof, J. N., Noyes, T., Oefner, P. J., Stahl, E. A., & Weigel, D. (2002). The extent of linkage disequilibrium in Arabidopsis thaliana. *Nature Genetics*, *30*(2), 190–193.

Nordborg, M., Hu, T. T., Ishino, Y., Jhaveri, J., Toomajian, C., Zheng, H., Bakker, E., Calabrese, P., Gladstone, J., Goyal, R., Jakobsson, M., Kim, S., Morozov, Y., Padhukasahasram, B., Plagnol, V., Rosenberg, N. A., Shah, C., Wall, J. D., Wang, J., … Bergelson, J. (2005). The pattern of polymorphism in Arabidopsis thaliana. *PLoS Biology*, *3*(7), e196.

Ossowski, S., Schneeberger, K., Clark, R. M., Lanz, C., Warthmann, N., & Weigel, D. (2008). Sequencing of natural strains of Arabidopsis thaliana with short reads. *Genome Research*, *18*(12), 2024–2033.

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T. G., Fookes, M., Falush, D., Keane, J. A., & Parkhill, J. (2015). Roary: rapid large-scale prokaryote pan

https://doi.org/10.1017/qpb.2025.10028 Published online by Cambridge University Press

genome analysis. *Bioinformatics (Oxford, England)*, *31*(22), 3691–3693.

Paten, B., Earl, D., Nguyen, N., Diekhans, M., Zerbino, D., & Haussler, D. (2011). Cactus: Algorithms for genome multiple sequence alignment. *Genome Research*, *21*(9), 1512–1528.

Prodanov, T., Plender, E. G., Seebohm, G., Meuth, S. G., Eichler, E. E., & Marschall, T. (2024). Locityper: targeted genotyping of complex polymorphic genes. In *bioRxiv*. https://doi.org/10.1101/2024.05.03.592358

Qin, P., Lu, H., Du, H., Wang, H., Chen, W., Chen, Z., He, Q., Ou, S., Zhang, H., Li, X., Li, X., Li, Y., Liao, Y., Gao, Q., Tu, B., Yuan, H., Ma, B., Wang, Y., Qian, Y., … Li, S. (2021). Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell*, *184*(13), 3542–3558.e16.

Rakocevic, G., Semenyuk, V., Lee, W. P., Spencer, J., Browning, J., Johnson, I. J., Arsenijevic, V., Nadj, J., Ghose, K., Suciu, M. C., Ji, S. G., Demir, G., Li, L., Toptaş, B., Dolgoborodov, A., Pollex, B., Spulber, I., Glotova, I., Kómár, P., … Kural, D. (2019). Fast and accurate genomic analyses using genome graphs. *Nature Genetics*, *51*(2), 354–362.

Raphael, B., Zhi, D., Tang, H., & Pevzner, P. (2004). A novel method for multiple alignment of sequences with repeated and shuffled elements. *Genome Research*, *14*(11), 2336–2346.

Rautiainen, M., & Marschall, T. (2020). GraphAligner: rapid and versatile sequence-to-graph alignment. *Genome Biology*, *21*(1), 253.

Schneeberger, K., Hagmann, J., Ossowski, S., Warthmann, N., Gesing, S., Kohlbacher, O., & Weigel, D. (2009). Simultaneous alignment of short reads against multiple genomes. *Genome Biology*, *10*(9), R98.

Schneeberger, K., Ossowski, S., Ott, F., Klein, J. D., Wang, X., Lanz, C., Smith, L. M., Cao, J., Fitz, J., Warthmann, N., Henz, S. R., Huson, D. H., & Weigel, D. (2011). Reference-guided assembly of four diverse Arabidopsis thaliana genomes. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(25), 10249–10254.

https://doi.org/10.1017/qpb.2025.10028 Published online by Cambridge University Press

Sheikhizadeh, S., Schranz, M. E., Akdel, M., de Ridder, D., & Smit, S. (2016). PanTools: representation, storage and exploration of pan-genomic data. *Bioinformatics*, *32*(17), i487–i493.

Sibbesen, J. A., Eizenga, J. M., Novak, A. M., Sirén, J., Chang, X., Garrison, E., & Paten, B. (2023). Haplotype-aware pantranscriptome analyses using spliced pangenome graphs. *Nature Methods*, *20*(2), 239–247.

Sirén, J., Monlong, J., Chang, X., Novak, A. M., Eizenga, J. M., Markello, C., Sibbesen, J. A., Hickey, G., Chang, P.-C., Carroll, A., Gupta, N., Gabriel, S., Blackwell, T. W., Ratan, A., Taylor, K. D., Rich, S. S., Rotter, J. I., Haussler, D., Garrison, E., & Paten, B. (2021). Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science*, *374*(6574), abg8871.

Smith, C. J., Strausz, S., FinnGen, Spence, J. P., Ollila, H. M., & Pritchard, J. K. (2025). Haplotype analysis reveals pleiotropic disease associations in the HLA region. *The American Journal of Human Genetics*, *0*(0). https://doi.org/10.1016/j.ajhg.2025.06.011

Song, B., Marco-Sola, S., Moreto, M., Johnson, L., Buckler, E. S., & Stitzer, M. C. (2022). AnchorWave: Sensitive alignment of genomes with high sequence diversity, extensive structural polymorphism, and whole-genome duplication. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(1). https://doi.org/10.1073/pnas.2113075119

Soylev, A., Ebler, J., Pani, S., Rausch, T., Korbel, J., & Marschall, T. (2024). SVarp: pangenome-based structural variant discovery. In *bioRxiv* (p. 2024.02.18.580171). https://doi.org/10.1101/2024.02.18.580171

Sudmant, P. H., Kitzman, J. O., Antonacci, F., Alkan, C., Malig, M., Tsalenko, A., Sampas, N., Bruhn, L., Shendure, J., 1000 Genomes Project, & Eichler, E. E. (2010). Diversity of human copy number variation and multicopy genes. *Science (New York, N.Y.)*, *330*(6004), 641–646.

https://doi.org/10.1017/qpb.2025.10028 Published online by Cambridge University Press

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Angiuoli, S. V., Crabtree, J., Jones, A. L., Durkin, A. S., DeBoy, R. T., Davidsen, T. M., Mora, M., Scarselli, M., y Ros, I. M., Peterson, J. D., Hauser, C. R., Sundaram, J. P., Nelson, W. C., … Fraser, C. M. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial "pan-genome." *Proceedings of the National Academy of Sciences*, *102*(39), 13950–13955.

The Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, *408*(6814), 796–815.

Tonkin-Hill, G., MacAlasdair, N., Ruis, C., Weimann, A., Horesh, G., Lees, J. A., Gladstone, R. A., Lo, S., Beaudoin, C., Floto, R. A., Frost, S. D. W., Corander, J., Bentley, S. D., & Parkhill, J. (2020). Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biology*, *21*(1), 180.

Vaughn, J. N., Branham, S. E., Abernathy, B., Hulse-Kemp, A. M., Rivers, A. R., Levi, A., & Wechter, W. P. (2022). Graph-based pangenomics maximizes genotyping density and reveals structural impacts on fungal resistance in melon. *Nature Communications*, *13*(1), 7897.

Vorbrugg, S., Bezrukov, I., Bao, Z., & Weigel, D. (2024). Gretl-variation GRaph evaluation TooLkit. *Bioinformatics (Oxford, England)*, *41*(1), btae755.

Vorbrugg, S., Bezrukov, I., Bao, Z., Xian, W., & Weigel, D. (2024). Gfa2bin enables graph-based GWAS by converting genome graphs to pan-genomic genotypes. In *bioRxiv*. https://doi.org/10.1101/2024.12.05.626966

Wick, R. R., Schultz, M. B., Zobel, J., & Holt, K. E. (2015). Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics (Oxford, England)*, *31*(20), 3350–3352.

Wlodzimierz, P., Rabanal, F. A., Burns, R., Naish, M., Primetis, E., Scott, A., Mandáková, T., Gorringe, N., Tock, A. J., Holland, D., Fritschi, K., Habring, A., Lanz, C., Patel, C., Schlegel, T., Collenberg, M., Mielke, M., Nordborg, M., Roux, F., … Henderson, I. R. (2023). Cycles

of satellite and transposon evolution in Arabidopsis centromeres. *Nature*, *618*(7965), 557–565.

Yokoyama, T. T., Sakamoto, Y., Seki, M., Suzuki, Y., & Kasahara, M. (2019). MoMI-G: modular multi-scale integrated genome graph browser. *BMC Bioinformatics*, *20*(1), 548.

Zhang, W., Macias-Velasco, J. F., Zhuo, X., Belter, E. A., Jr, Tomlinson, C., Garza, J., Tekkey, N., Li, D., & Wang, T. (2025). methylGrapher: genome-graph-based processing of DNA methylation data from whole genome bisulfite sequencing. *Nucleic Acids Research*, *53*(3), gkaf028.

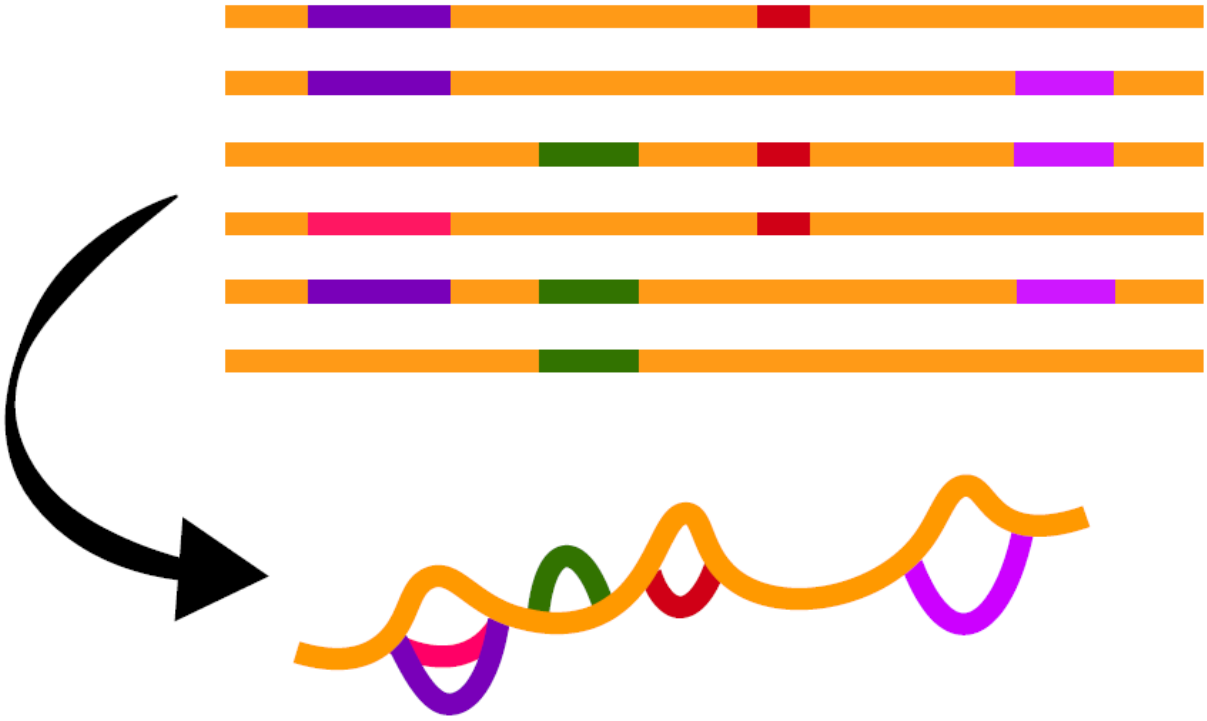
Zhao, Q., Feng, Q., Lu, H., Li, Y., Wang, A., Tian, Q., Zhan, Q., Lu, Y., Zhang, L., Huang, T., Wang, Y., Fan, D., Zhao, Y., Wang, Z., Zhou, C., Chen, J., Zhu, C., Li, W., Weng, Q., … Huang, X. (2018). Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nature Genetics*, *50*(2), 278–284.

Zhou, H., Su, X., & Song, B. (2024). ACMGA: a reference-free multiple-genome alignment pipeline for plant species. *BMC Genomics*, *25*(1), 515.

Zhou, Y., Zhang, Z., Bao, Z., Li, H., Lyu, Y., Zan, Y., Wu, Y., Cheng, L., Fang, Y., Wu, K., Zhang, J., Lyu, H., Lin, T., Gao, Q., Saha, S., Mueller, L., Fei, Z., Städler, T., Xu, S., … Huang, S. (2022). Graph pangenome captures missing heritability and empowers tomato breeding. *Nature*, *606*(7914), 527–534.

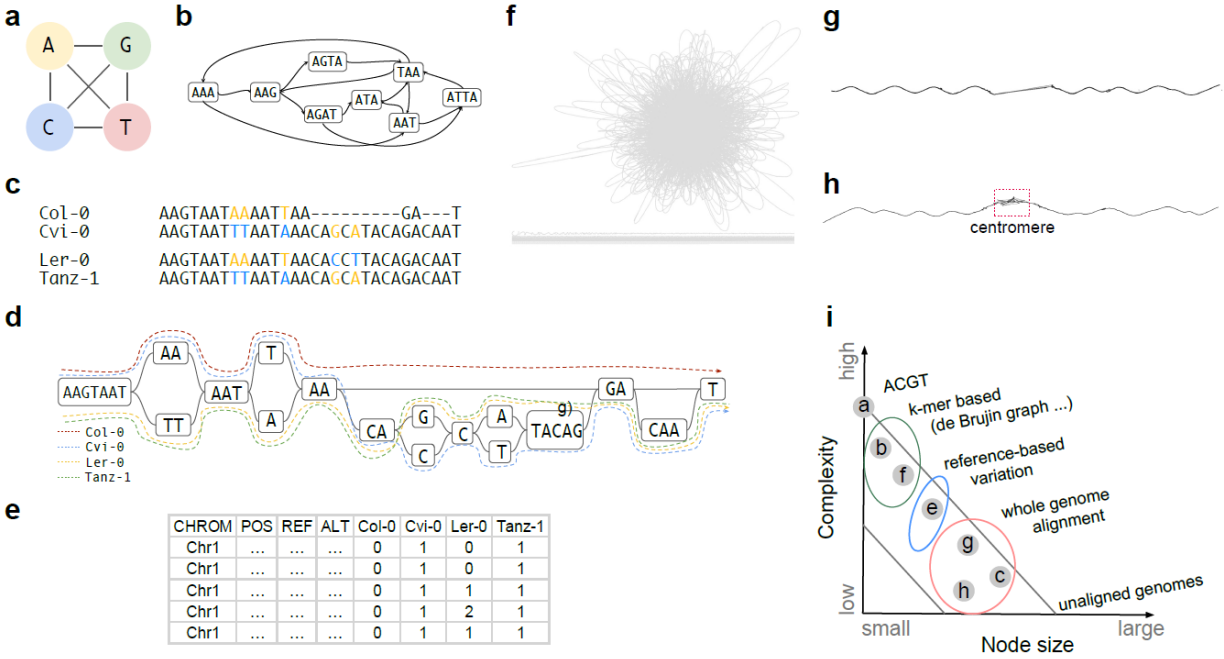
https://doi.org/10.1017/qpb.2025.10028 Published online by Cambridge University Press

Graphical Abstract



https://doi.org/10.1017/qpb.2025.10028 Published online by Cambridge University Press

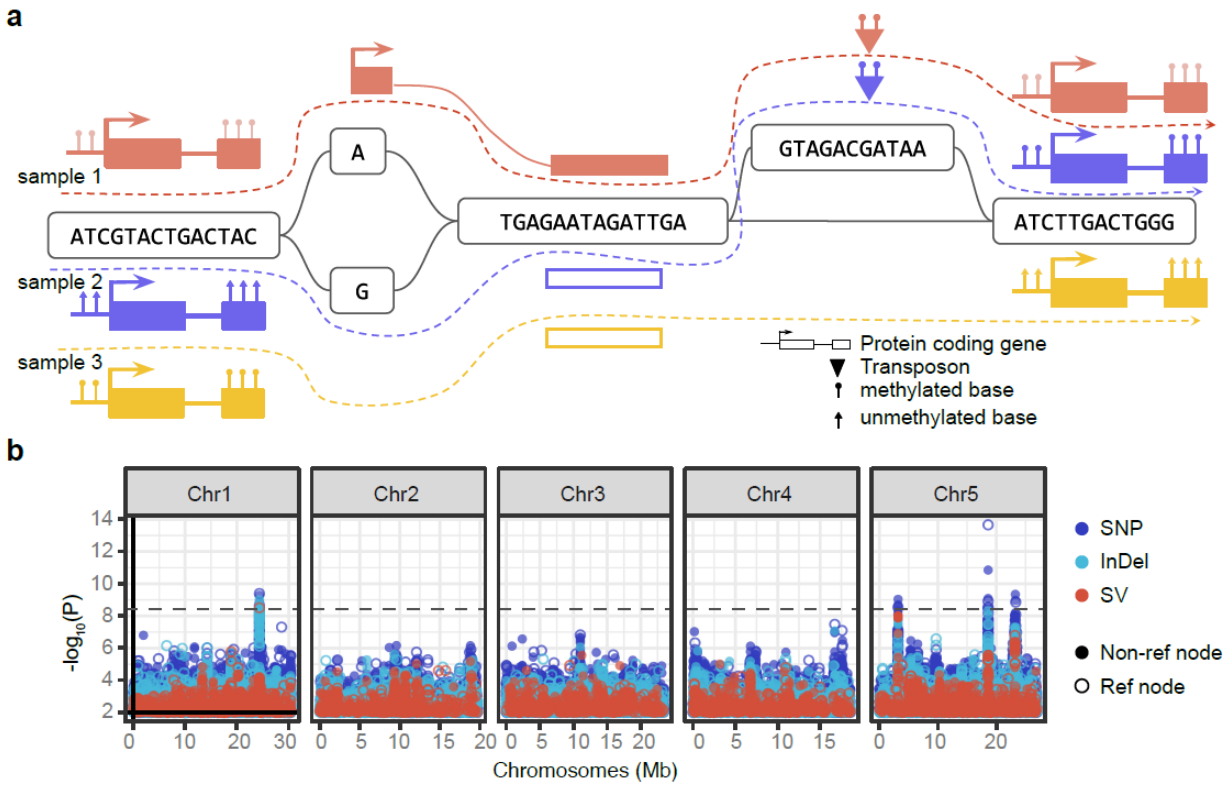
Figure 1 | Multiple approaches to building pangenome graphs. a) A graph that has only four nodes, corresponding to the four DNA bases, with all possible connections between the nodes. b) A k-mer graph based on short sequences (here, triplets). c) and d) The same sequences are combined in different representations, which highlights the equivalency of a multiple genome alignment and a genome graph. e) The same sequences shown in VCF format in a symbolic manner. f), g) and h) represent the entire chromosome 1 from five *A. thaliana* individuals. f) A Biforst 21-mer graph. g) A Minigraph-Cactus graph. h) A PGGB graph. i) Summary of the trade-offs in node size and complexity for different types of graphs. Note that even with the same method, parameter choice will result in different graphs from identical sequences.



CHROM	POS	REF	ALT	Col-0	Cvi-0	Ler-0	Tanz-1
Chr1	...	...	...	0	1	0	1
Chr1	...	...	...	0	1	0	1
Chr1	...	...	...	0	1	1	1
Chr1	...	...	...	0	1	2	1
Chr1	...	...	...	0	1	1	1

https://doi.org/10.1017/qpb.2025.10028 Published online by Cambridge University Press

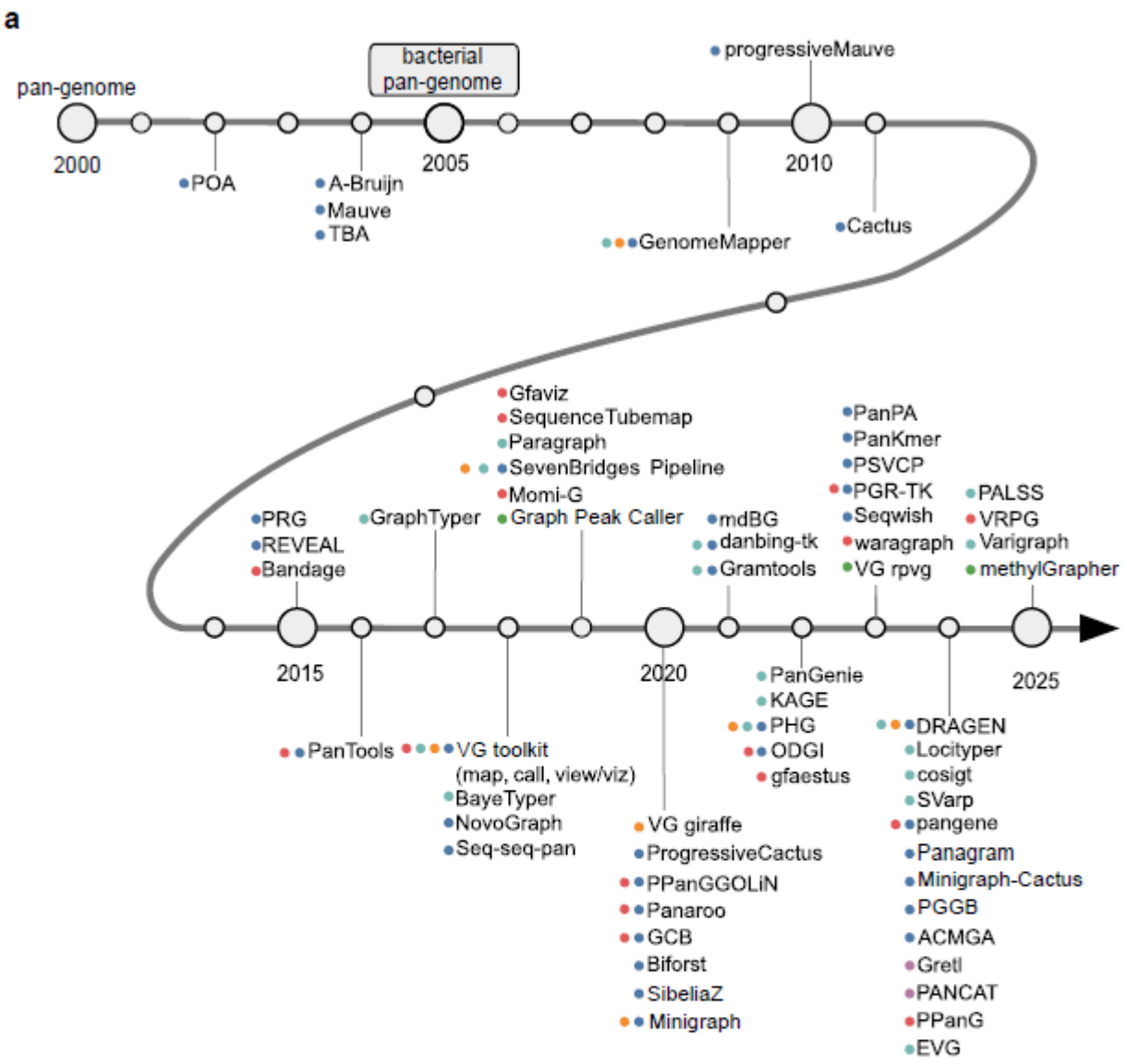
Figure 2 | Functional pangenomics. a) A pangenome graph can integrate diverse layers of functional annotations (e.g., genes, transposons, methylation level) in its reference coordinate system and serve as a unified platform for cross-genome comparison. b) Graph nodes can be used directly for genome-wide association analyses. The colors represent different length-based categories for the node, while node shapes indicate whether the sequence originates from the chosen reference genome. Broken line indicates nominal significance threshold. This figure was adapted from (Vorbrugg, Bezrukov, Bao, Xian, et al., 2024).

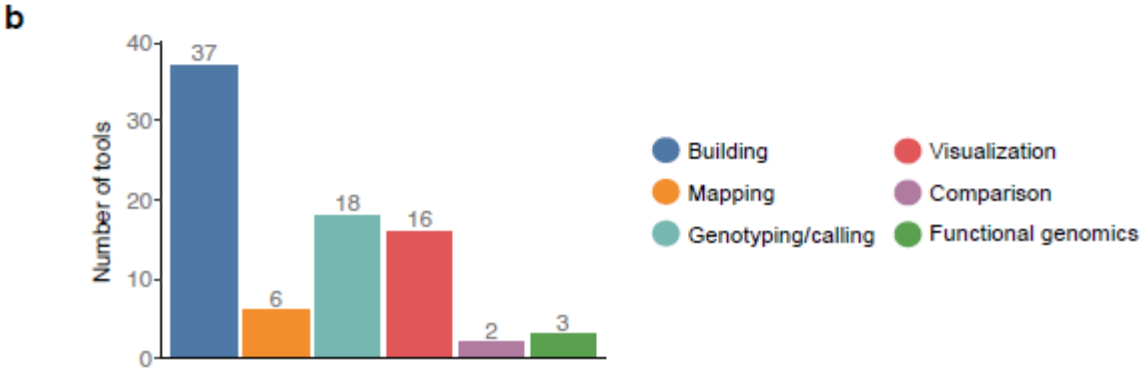


https://doi.org/10.1017/qpb.2025.10028 Published online by Cambridge University Press

Figure 3 | Timeline of pangenome graph algorithms. a) The differently colored circles indicate the main functions of tools; some workflows/tools may have multiple usages. For tools described in journals, we use the date of publication, but we note that many colleagues in this area are very generous and often release their tools long before formal publication. Given the rapid development in this area, it is perhaps not surprising that some tools only have a public GitHub repository. In the text, we provide hyperlinks as references. b) The number of graph-based tools developed for different purposes.

https://doi.org/10.1017/qpb.2025.10028 Published online by Cambridge University Press





https://doi.org/10.1017/qpb.2025.10028 Published online by Cambridge University Press