Peer Community Journal

Section: Genomics

Research article

Published
2026-09-07

Cite as
Stéphanie Bocs, Camille Carrette, Johann Confais, Siegfried Dubois, Ludovic Duvaux, Christophe Klopp, Nicolas Lapalu, Pauline Lasserre-Zuber, Fabrice Legeai, Claire Lemaitre, Benjamin Linard, Nina Marthe, Baptiste Pierre, Gautier Sarah, François Sabot, Christine Tranchant-Dubreuil and Matthias Zytnicki (2026) *Adopting graph-based pangenomics for agronomy and biodiversity studies: current resources and challenges*, Peer Community Journal, 6: e86.

Correspondence
matthias.zytnicki@inrae.fr

Peer-review
Peer reviewed and recommended by PCI Genomics, https://doi.org/10.24072/pci.genomics.100526

Adopting graph-based pangenomics for agronomy and biodiversity studies: current resources and challenges

Stéphanie Bocs[1,2,3], Camille Carrette[4], Johann Confais[5], Siegfried Dubois[6], Ludovic Duvaux[7], Christophe Klopp[8,9], Nicolas Lapalu[10], Pauline Lasserre-Zuber[11], Fabrice Legeai[6,12], Claire Lemaitre[6], Benjamin Linard[8], Nina Marthe[4], Baptiste Pierre[1], Gautier Sarah[1,3], François Sabot[5], Christine Tranchant-Dubreuil[4], and Matthias Zytnicki[8]

Volume **6** (2026), article e86

https://doi.org/10.24072/pcjournal.781

Abstract

Pangenomics is transforming the way genomics is done. By using assemblies of multiple complete genomes for the same species, it is possible to depart from a biased, reference-centric, point of view. Represented as graphs, these pangenomes open new paths for genome understanding and improvement of species of agricultural interest. They have, for instance, been used to discover new structural variants linked with traits of interest, and increased the heritability prediction. However, transitioning to a graph-based approach is not straightforward, and many tools are still needed for this shift. While numerous papers present both methodological and applied approaches on pangenomics, and many reviews present advantages of these methods, very few resources outline what remains to be done. In this paper, we would like to list methods that are still required to fully exploit pangenomes, and favor the widespread adoption of this approach.

[1]UMR AGAP Institut, Univ Montpellier, CIRAD, INRAE, Institut Agro, F-34398 Montpellier, France, [2]CIRAD, UMR AGAP Institut, F-34398 Montpellier, France, [3]South Green Bioinformatics Platform, French Institute of Bioinformatics (IFB); Bioversity, CIRAD, INRAE, IRD, F-34398 Montpellier, France, [4]UMR DIADE, University of Montpellier, CIRAD, IRD, F-34394 Montpellier, France, [5]URGI, BioinfOmics, INRAE, Université Paris-Saclay, F-78026 Versailles, France, [6]Univ Rennes, Inria, CNRS, IRISA - UMR 6074, F-35000, Rennes, France, [7]INRAE, Univ. Bordeaux, BIOGECO, F-33610 Cestas, France, [8]University of Toulouse, INRAE, UR 875 MIAT, F-32326 Castanet-Tolosan, France, [9]University of Toulouse, INRAE, BioinfOmics, GenoToul Bioinformatics Facility, F-32326 Castanet-Tolosan, France, [10]Université Paris-Saclay, INRAE, UR BIOGER, F-91120 Palaiseau, France, [11]UMR 1095 GDEC, INRAE, UCA, F-63000 Clermont-Ferrand, France, [12]IGEPP, INRAE, F-35650 Le Rheu, France

Peer Community Journal is a member of the Centre Mersenne for Open Scientific Publishing
http://www.centre-mersenne.org/

e-ISSN 2804-3871

Introduction

A wealth of pangenome models

The aim of the pangenomics is to model the genomic diversity of a population of interest. Whereas genomics studies use a unique genome as a reference, pangenomics take sets of genomes, usually from the same species, and model this diversity. The sampling intends to represent the genetic diversity under study. The notion of "pangenome" covers many models, and that it is often impossible to guess what a paper with this word in its title actually entails (Edwards, 2026). Luckily, these different models have been presented in different reviews (Eizenga et al., 2020; Golicz et al., 2020; Guerra, 2026; Hu et al., 2024; Matthews et al., 2024; Ruperao et al., 2025; Schreiber et al., 2024; Secomandi et al., 2025; Taylor et al., 2024). The term "pangenome" was first coined for prokaryotes (Tettelin et al., 2005), and they included the variations of genes, but they have been extended to eukaryotes since. Several pangenomes were also model using a genome of reference, together with an exhaustive catalog of variations: single nucleotide polymorphisms (SNPs), insertions/deletions (indels), or structural variants (SVs). SVs are usually, and arbitrarily, defined as indels of size longer than 50bp (The 1000 Genomes Project Consortium, 2010), or balanced rearrangements, which include inversions and translocations. Otherwise, the pangenome can be a collection of sequence variations, together with a reference genome.

With the advent of affordable, long, high quality sequencing technologies, such as ONT (Jain et al., 2018) and PacBio HiFi (Wenger et al., 2019) reads, it is now possible to produce accurate and telomere-to-telomere genome assemblies for a reasonable cost, and to collect several complete genomes for a given taxonomic group. These different genomes can be encoded into a graph, and at least two different types of graph are available: de Bruijn graphs and variations graphs. Both store sequences as nodes in the graph, and each genome is present in this graph, removing the principle of a reference genome. In the former, nodes are k-mers, edges are $k - 1$ overlaps, and colors indicate whether a given node belongs to a given haplotype. The latter use nodes of variable sizes, and model genomes as paths in the graph. These graphs have been discussed and compared in several reviews (Eizenga et al., 2020; Hu et al., 2024; Matthews et al., 2024; Ruperao et al., 2025). As observed by Andreace et al. (2023), de Bruijn graph scale much better, both in time and memory. However, usual operations, such as retrieving a genome from a de Bruijn-based pangenome graph, is in principle not possible. So far, variation graphs have been more widely adopted, as seen in recent literature.

Breakthroughs using variation graphs

Advantages of variation graphs have been underlined in several publications. For instance, when reads are mapped only to a reference, Sirén et al. (2021) observed that the fraction of reads supporting the reference allele was consistently greater than the fraction of reads supporting the alternative allele in human heterozygous loci (less than 30% for indels of size greater than 40bp). This bias is corrected (albeit not entirely for very large indels) using a pangenome graph. The rationale is that, most of the time, the alternative allele is present in the graph: the reads can thus be mapped without error, which simplifies the process. This finding has direct implications, for example in SNP calling, genotyping, and genome-wide association studies (GWAS). For instance, the F1 value of the genotype call increases from 0.9940 to 0.9953 when using graph (Sirén et al., 2021).

Other publications showed the advantage of including long SVs, an information that is only accessible using genome assemblies. Milia et al. (2024) used a bovine pangenome graph in order to understand the cause of a white head depigmentation, known to follow a genetic dominant pattern. They found a segmental duplication upstream of a gene of interest, affecting its expression. Of note, since the duplication was composed of transposable elements and other repeats, it was not detected using short reads. Interestingly, this duplication harbored several variants, and insertions inside this duplication were also observed. In cotton, Yang et al. (2026)

exhibited five large inversions, and one translocation, that appeared during, or before, domestication. In tomato, Zhou et al. (2022) looked for the quantitative trait locus (QTL) responsible for the expression of a gene of interest (eQTL). A SNP-only analysis points to a locus several genes away, whereas the likely causal variant is a SV which is located inside the gene, with a much stronger signal. The same analysis showed that adding indels and SVs, compared to a SNP-only model, increased the heritability from 0.28 to 0.33. Even more interestingly, when these variants were modeled in the graph and not on the reference genome, the heritability increased from 0.33 to 0.41.

SVs also sometimes complements results found for SNPs. In cucumber, Zhao et al. (2026) found that the mutation load of mildly deleterious SNPs increased during domestication. This result is expected, and confirms the cost of domestication hypothesis. However, SVs significantly decreased in the process. This would suggest that, contrary to the previous observation, SV were purged by purifying selection.

Aim of this work

The interest on pangenomics is thus growing, particularly in the biodiversity-driven fields such as agroecology (Adam et al., 2025). Several papers now use and publish pangenome graphs, and many reviews already listed several aspects of the field. Some provide an introduction of the concepts (Matthews et al., 2024). Others address the process of constructing a pangenome graph (Andreace et al., 2023; Eizenga et al., 2020), whereas other focus on technical aspects, such as mapping reads to a graph (Cui et al., 2025). Other works list the results obtained on several organisms, such as plants (Danilevicz et al., 2020; Hu et al., 2024; Wang et al., 2023; Zanini et al., 2021), animals (Gong et al., 2023), but also biodiversity genomic studies (Secomandi et al., 2025). A common conclusion is that, while pangenome graphs can be created, even in complex cases (many genomes, containing many repetitions, or being polyploid), tools exploiting them are still lacking. This is the paradox of a radical change in the pangenome model: switching from a reference genome to a graph requires all the downstream analysis tools to be re-engineered from scratch. For instance, GWAS analyses are done on a reference genome, and do not take the advantage of known variability. Bluntly speaking, whereas it is now fashionable to create pangenome graphs, it is still not clear how or why they could be useful for agronomy. Before the interest in the topic is lost, it is thus urgent to transfer tools developed for reference genomes to graphs, and to invent new methods in order to exploit the full potential of pangenome graphs. Although we are aware that transitioning from genomes to graphs poses serious computational challenges (Limasset et al., 2016), we are confident that they will be resolved soon.

The aim of this opinion paper is to list the bottlenecks that pangenome users experience once they have built their graphs. We will thus not mention the principles and difficulties of genome assembly, or graph construction. We will suppose that near telomere-to-telomere assemblies are available and used, which can be haplotype-resolved or collapsed pseudo-chromosomes. Moreover, we will focus here on multi-cellular eukaryotic species of agronomic and ecological interest, which are larger than most prokaryotes, but may also be quite different to human genomes: they can be much larger (wheat is about 15Gbp long), contain many repeated DNA, be polyploid (some sturgeons are tetraploid), and/or harbor a much higher diversity (heterozygosity is *k*-mer estimated to 1.9% in some oak species, genome size can vary up to 20% in maize). More precisely, we would like to mention here all the tools that should be developed in order to exploit these graphs. They include methods that should be adapted from so-called “linear genomes,” but also new tools, that are specifically conceived for graphs. Indeed, some questions can only be addressed using graphs: for instance, complex, nested SVs are poorly represented by a reference genome and a set of variations.

A similar review has not been addressed since 2018 in The Computational Pan-Genomics Consortium, 2018, but the landscape of pangenomics has radically changed since then. We hope here to rouse the interest of computer scientists, statisticians, data scientists, etc. so that these open questions will be solved soon. We think that these developments will be necessary to transition from genomics to pangenomics, and collect results that were previously out of reach.

Whereas this study focuses on pangenome graph, we do not want to mean that every genomic study should be performed on this data structure. Diversity analysis, association studies using SNPs and a reference genome have several advantages: they are faster, cheaper, and can be directly compared to the results accumulated so far, since they use the same methodologies.

The aim of our work is to mention all the analyses that cannot be performed, when one wants to use pangenome graphs.

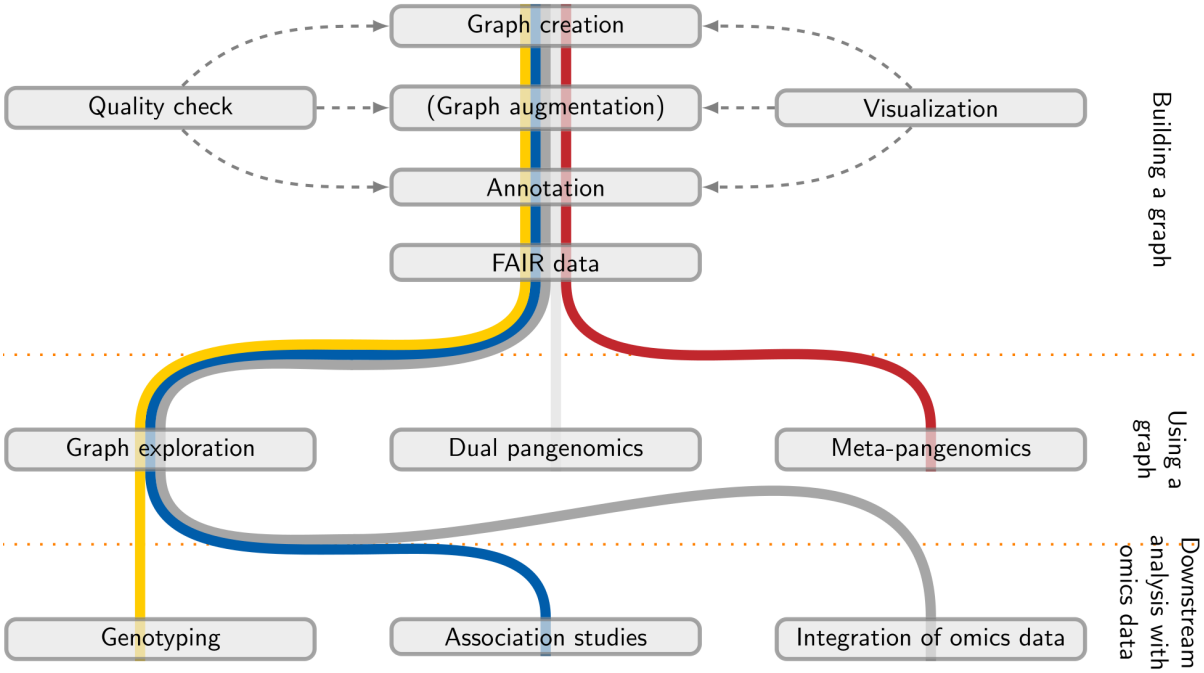

Figure 1 – A possible pipeline for the uses of a pangenome graph. This pipeline can be roughly divided into three main sections (on the right): build a graph, using it, and adding reads. A first pangenome graph is created, usually with almost complete haplotype assemblies. Since graph creating methods vary, and may be fraught with errors (unaligned regions, broken paths), assessing the quality of a graph is useful (this can be applied after each subsequent step). Similarly, it is also crucial to be able to visualize the graph, especially on particular loci of interest (can also be done throughout the pipeline). The graph can be augmented with additional data increasing allele counts, including known contig alignment or read mapping. Several augmentation iterations, with different types of data, can be considered. Since most analyses focus on genes and transposable elements, it is usually useful to add an annotation (available on one or several reference genomes), which projected to the graph. This graph, possibly extended with new variants and annotation, should be then shared with the community, following FAIR practices. Most analyses then focus on graphs with only one species, and the graph can be analyzed in order to find regions of interest. However, dual pangenomics, which studies the joint diversities of two species, can be envisioned. More generally, many species can also be modeled, for instance in the case of meta-genomics. Low coverage sequencing data of individuals can be compared or aligned to the graph in order to produce large genotyping sets enabling population genetics studies (*i.e.* coalescence, association). Other omics (*e.g.* RNA-Seq, ChIP-Seq, Hi-C, LC-MS, BS-Seq etc.) data can be added to the graph, and potentially aggregated to multi-omic layers.

Outline

In this work, we will address several practical questions that are commonly considered by pangenome graph users (Figure 1). Current analyses suggest using genome assemblies close to telomere-to-telomere completeness, preferably using haplotypes (representing each chromosome separately) for building pangenome graphs (Cheng et al., 2025; Leonard et al., 2022; Li et al.,

2024). Tools like Minigraph (Li et al., 2020), Minigraph-Cactus (Hickey et al., 2023), or PGGB (Garrison et al., 2024) create variation graphs, and de Bruijn graphs can be created by Bifrost (Holley and Melsted, 2020), for instance. Both types are pangenomes graphs. Pros and cons of each method have been described by Andreace et al. (2023). Among others, de Bruijn are much more compact, whereas it is easier to retrieve haplotypes using variation graphs.

In a pangenome graph, a node represents a sequence, and paths (in a variation graphs) or colors (in de Bruijn) represent haplotypes. In a variation graph, there is an edge between two nodes if the corresponding sequences are contiguous for at least one haplotype.

The rest of the manuscript presents several graph use cases. First, the quality of the graph should be assessed, yet no metrics have been defined to do so (Sections “Pangenome representativeness” and “Graph quality assessment”). Visualization should also be possible, although the sheer number of nodes in the graph requires innovative solutions, which are not found yet (Section “Graph visualization”). Optionally, the graph can be augmented with new variations, which are found in individuals unsampled for the graph construction step (Section “Graph augmentation”). Since many analyses will study the impact of variations on genes, it is essential to add annotations (coding, non-coding genes, or transposable elements) to the graph (Section “Graph annotation”). At this point, the graph, together with possible augmentation and annotation, can be shared by communities using the “findability, accessibility, interoperability, and reusability” (FAIR) principles (Section “Producing FAIR pangenomes”). The graph can then be analyzed and explored, in order to explore a locus of interest (Section “Manipulating and exploring data from graphs”). Alternatively, it is also possible to look for patterns inside the graph, which could be linked with population genetics, or coalescence (Section “Population genetics and coalescent theory”). The type of analysis will differ, depending on whether one or several species are included. Indeed, when two species are present, such as a host and a pathogen, one may want to analyze their relative evolutions (Section “Dual pangenomes”). This idea can be extended to several species, in the case of meta-pangenomics (Section “Pangenome graphs for metagenomics and organelle-based studies”). New genomic data, such as short or long reads can then be aligned to a graph. They can be used for genotyping individuals, or a population (Section “Variant anal-ysis and genotyping”). If contrasted populations are studied, a Genome Wide Association Study (GWAS) can be conducted (Section “Association studies”). Finally, other omics data —informing on transcription, translation, or epigenetics— can be added to the graph, in order to find links between genotype and regulation (Section “Towards multi-omics pangenomics”).

Questions

Building a graph

Pangenome representativeness. Conceptually, a pangenome is a theoretical object aiming to capture the full extent of genetic diversity within a given population, species, or species complex (hereafter referred to as the “metapopulation”). In practice, however, this goal is unattainable as any inferred pangenome graph represents only a partial and imperfect view of the metapopulation variation. Actually, no sampling design can ever be exhaustive as alleles are detected with a probability equal to their frequency in the metapopulation (*i.e.* given a sample size *N* of diploid individuals, we expect to find alleles with a frequency of *1/2N* on average). Thus, any sampling scheme must be tailored to the question at stake. For instance, detecting low frequency variants responsible for diseases in livestock species requires a much greater effort than detecting locally high-frequency adaptive variants in tree species. More importantly, the sampling design must account for the genetic structure of the metapopulation in order to capture alleles that are private to certain populations—as observed by the sudden rise of diversity in “pangenome growth” plots when a new population is first represented (Figure 3 in Liao et al. (2023)).

As mentioned in Secomandi et al. (2025), a good practice is thus to sample the population based on results obtained with chip or low-pass sequencing and tools like CoreHunter (De Beukelaer and Davenport, 2016) or SVCollector (Ranallo-Benavidez et al., 2021) in order to constitute a good diversity panel. Yet, these methods usually rely on known variations, or SNPs,

which fail to account for structural variants which add crucial information on the population diversity. Since SVs are mostly found using pangenomic data, this looks like a chicken-and-egg problem. However, a sufficient number of SNPs should be enough to get a reliable genetic structure of the population, and *k*-mer based approaches, akin to Mash distance (Ondov et al., 2016), which consider all types of variations, can be used to this end.

To summarize, it would be ideal to have a standard way to sample a population, in order to guide the user when choosing his/her panel. This procedure could include the introduction of trios, which are interesting quality controls: SV in the offspring should, in principle, also be found in the parents.

Graph quality assessment. Building a pangenome is a very complex task, especially when genomes are large, and many individuals are included. The aim of a pangenome builder is to group, in the same node, orthologous sequences. This task is particularly challenging in low-complexity or highly repetitive regions, especially where recurrent mutations can blur the true evolutionary signal. As a results, current methods rely on heuristics, and arbitrary choices, which can fail to detect sequence orthology.

Since some of the variants present in pangenome graphs may not reflect genuine genomic differences, but rather artefacts introduced by the tools and methods used, strategies to evaluate graph quality are highly needed.

The genome assembly community developed several widely used metrics. Indeed, technical limitations make it difficult to assemble telomere-to-telomere genomes, especially in low-complexity or highly repetitive regions of the genome (*e.g.* transposable element- or microsatellite-rich intervals, telomeres, centromeres), even with state-of-the-art assemblers like *hifiasm* (Cheng et al., 2021). BUSCO scores (Manni et al., 2021), N50 (Earl et al., 2011) or LAI (Ou et al., 2018), have been used to assess completeness of assembled genomes, whereas tools like *KAT* (Mapleson et al., 2016) or Merqury (Rhie et al., 2020) check whether the *k*-mers found in the reads are also present in the assembly. Recent guidelines have been published for gene-centric graphs (Heuermann et al., 2025), and it would be convenient to have metrics for graphs. So far, tools such as *vg stats* (Garrison et al., 2018) or *odgi stats* (Guarracino et al., 2022) produce a series of metrics describing the graph: number of nodes, size of the paths, number of edges per node, etc. However, it is still unclear what range of values would qualify a graph as "good." Pangenome growth analyses performed by tools such as *panacus* (Parmigiani et al., 2024) are useful for assessing whether a given sample size is sufficient to adequately represent a population: a plateau indicates a saturation of the sampling strategy. Kopalli et al. (2025) set up a benchmark, comparing major pangenome builders. Produced graphs are compared in terms of size, or variants called. They also simulated reads from these graphs, and compared the number of reads that mapped the graph. This metrics is practical and effective, yet it is not easy to determine whether the graph produced is better, or whether the mapping algorithm is particularly suited to a type graph topology. We are thus in need to get reliable indicators, which would include an assessment whether the graph faithfully represents the individual genomes (the recent *PG-SCUnK* (Cumer et al., 2026) may fill this need), and an assessment of its usability, which could include the compactness, or, on the contrary, the lack of loop.

In the absence of standard metrics, pangenome graphs can still be compared among themselves. Graph comparison is an old topic, and many methods can be devised, with the caveat that, in principle, the comparison involves graphs built with identical sequences. Methods based on ED-strings (Gabory et al., 2024) and *pancat* (Dubois et al., 2025), the latter being developed by authors of this work, can localize and quantify where and how the differences arise in graphs. These methods can be used to evaluate the impact of graph builder parameters. Interestingly, they can also be used to compare graphs produced with a "ground truth," which could be con-structed manually —for small datasets only— or simulated. This requires realistic whole genome population simulators, using controlled phylogenies, such as *MSpangenome* (Piat et al., 2026), also developed by authors of this work.

Representing variations also is a complex task. Genomic variants differentiating haplotypes are classically detected by looking for topological motifs in the graph, referred as *bubbles*, or *snarls*.

They are formed when at least two genome paths diverge in the graph due to sequence differences, and meet again further on. Several tools detect such motifs, such as *vg snarls* and *vg deconstruct* (Garrison et al., 2018; Paten et al., 2018) for *Minigraph-Cactus* and *PGGB* graphs, *gfatools bubbles* (Li et al., 2020) for *Minigraph* graphs, and *BubbleGun* (Dabbaghie et al., 2022). Complex variations can be nested, or simply co-localized, and sometimes do not fit in the bubble or snarl definitions. Once a variation has been found, it is then common to express it as a variation with respect to a reference genome (as a VCF file). Here again, there is no standard way to do so. As a result, some variations can be ignored, or mis-represented.

Moreover, there can be several ways to represent a genomic variation, and this problem has also been documented in multiple sequence alignment methods. Consider for instance a reference genome AAA, and an alternative genome AA. As is, it is impossible to guess which A has been deleted. Arbitrary decisions need to be made (*e.g.* left-normalisation), although they are not guaranteed to be evolutionary accurate, thus creating a bias for future applications (Figure 2A–C). Thus, different graph topologies do not necessarily imply different variants. Repetitions (which include sequences that occur in multiple copies throughout the genome in *cis* or in *trans*, but also copy-number variations) can either be modeled through independent sequences, or loops (Extended Figure 3 in Liao et al. (2023)). Similarly, inversions can be represented as two parallel sequences of reverse-complemented nodes rather than a bidirected node traversal (Figure 2D–E and Figure 1 in Romain et al. (2025)). As a result, graph builders may adopt different conventions, and graphs may be incomparable —although essentially equivalent— and downstream analyses will differ.

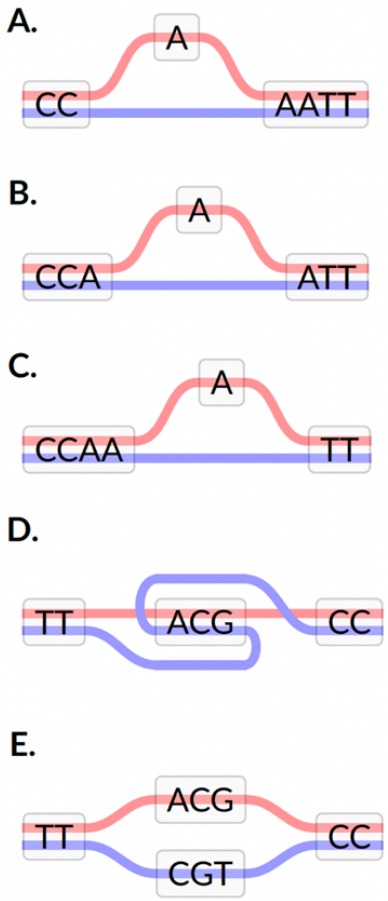

Figure 2 – A.–C.: Three equivalent ways to represent the removal of a A. **D.–E.**: Two ways of representing an inversion. We suppose here that the ancestral sequence is in red. In **A.**, **B.**, and **C.**, the first, second, and third A are removed (respectively). Deciding which nucleotide is deleted is arbitrary, yet the resulting variant call will not be identical, differing by 1 or 2 base pairs. In **D.**, the inversion is explicitly represented by introducing a reverse path that traverses the node in the opposite direction (which supposes that the reverse-complement sequence is used). In **E.**, the inversion is modeled by adding a new node containing the reverse sequence. Although the two paths yield the same sequence, their graph representations differ significantly.

Additionaly, When repetitions and inversions are modeled by loops or bidirectional node traversal, they clearly inform the reader of that particular event. Conversely, it greatly complicates the graphs, and some downstream tools cannot be used on cyclic graphs. The implementation

choices made by different graph builders affect the ease with which it can be manipulated (Liao et al., 2023).

Polyploidy is commonly found in species of agricultural interest, in plants or in animals like sturgeons. They can be divided into autopolyploids, which contain multiple sets of homologous chromosomes originating from the same species, and allopolyploids, which contain sets of chromosomes derived from different species, which are related but homoeologous rather than strictly homologous. If haplotypes are completely resolved, autopolyploids are not a problem to pangenome graph builders, since multiple homologous chromosomes can be analysed in a single graph. It is however considerably more complicated to get these phased haplotypes if they are nearly identical. When a phased pangenome is available, the graph can be used to phase new genomes, as it has been done on potato (Sun et al., 2025), and the graph may improve homeolog resolution. Downstream tools, especially genotyping tools, are not adapted to this configuration either, with the notable exception of varigraph (Du et al., 2025). Most of the time, homeologous chromosomes of allopolyploid species are handled separately. However, homeologous genes are also represented separately, and this limits the comparison. On the opposite, collapsing homeologous chromosomes into the same graph is expected to produce highly entangled topologies. Again, this is a trade-off between the quantity of information and ease of use.

As a conclusion, we advocate for the use of an explicit criterion, such as maximum likelihood, that would be optimized by graph builders. It would require an explicit molecular model of evolution, which would integrate the evolution of nucleotides, short repeats, and structural variants altogether. This would also require an improvement of the structural variants detection methods, since current ones fail to detect complex, entangled variations. Substantial work remains until such integrative models can be established, and it also requires more advanced algorithms, which would be able to exploit complex, cyclic, graphs. These graphs model more information, and should, in principle, be favored.

Graph visualization. Pangenome users need to visualize their graph, or at least a part thereof, in order to check the choices made by the pangenome builder (see previous section), or to gather information about structural variants. Admittedly, visualizing a graph with hundreds of millions of nodes is complex.

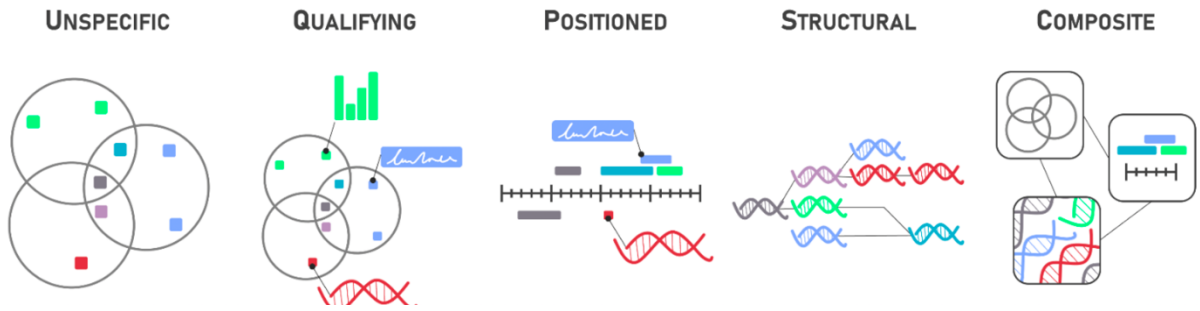

Figure 3 – Five categories of pangenome visualization tools, as presented by Durant, 2022.

In his PhD, Durant, 2022 proposed a classification of visualization systems into five categories (Figure 3). Unspecific tools use generic methods and are not dedicated to pangenome datasets, and include CytoScape (Shannon et al., 2003). Qualifying tools contain pangenome information, but no position, such as *Anvi'o* (Eren et al., 2020), *Panache* (Durant et al., 2021), *PanVA* (van den Brandt et al., 2024) using *Pantools* (Jonkheer et al., 2022), *PPanGGolin* (Gautreau et al., 2020), *VRPG* (Miao and Yue, 2025). Positioned tools anchor information on (pan)genomic coordinates. Structural tools focus on sequences and their continuity, like *SyRi* (Goel et al., 2019), *CoGe SynMap* (Haug-Baltzell et al., 2017), *SequenceTubeMap* (Beyer et al., 2019), *odgi viz* (Guarracino et al., 2022), *Bandage* (Wick et al., 2015) and Bandage-NG (https://github.com/asl/BandageNG), *Vizitig* (Degardins et al., 2024). Composite tools gather multiple categories within one tool, like Pantograph (https://www.computomics.com/home.html).

The ideal visualization tool would be able to represent the graph globally (at the chromosome scale) and locally (at base level resolution), giving the possibility to zoom in and out in an interactive, “Google Earth” way. The bird’s eye view would be able to mask small variations, and merge similar haplotypes, while retaining a global linear shape. Rules for conserving clarity and addressing the good information in the good visual channel must be the directive idea of any development. This idea has been suggested multiple times, and it is possible that some prototypes will implement this much needed feature soon. Another desirable feature is to be able to display the variations with respect to a known reference, or, on the contrary, being as agnostic as possible, and not favor any genome. The ideal tool should also be able to adapt a color pattern, in order to display a sub-population (*e.g.* cultivated vs wild individuals), or, on the contrary, display particular loci (*e.g*. genes, or exons). Hints should also mention which haplotypes belong the same individuals, in case of di- or polyploids organisms. Additional metadata should also be incorporated, and they could be reads counts, when sequences can be mapped to the genome. Moreover, it is likely that dedicated indexing data structures, or even adapted pangenome data formats, encompassing different zoom levels, should be needed to implement the ideal tool. Clearly, desirable properties for pangenome visualization are numerous, not to add that these graphs are huge. It is thus unlikely that a unique tool will fit all the needs.

Graph augmentation. Graph-based pangenomes are currently constructed using complete chromosome-level assemblies, obtained by long-read sequencing, but are often based on a limited number of genomes (Chen et al., 2023; Jayakodi et al., 2024; Liu et al., 2024). One of today’s challenges is to enrich these existing graphs to develop a more complete and accurate pangenome by integrating the diversity of a large number of additional individuals (The 3000 rice genomes project, 2014) or a small number of individuals of an exceptional ecotype (Bin Rahman and Zhang, 2013). These individuals may have been sequenced and partially assembled using long-read sequencing or resequenced using short-read sequencing technologies, thus providing access to greater inter- or intra-species variability. Currently, fragmented genome assemblies are typically discarded from the graph construction process, although they could provide valuable information on specific regions. Short-read sequencing, with its low cost and scalability, remains a practical choice for exploring diversity at the population scale or even broader levels (*e.g*., species, genus). Since the beginning of Next Generation Sequencing twenty years ago (Patrick, 2008; Reinartz et al., 2002), thousands of genomes and populations have been sequenced. This data is an incomparable resource of genomic diversity that will not be easily obtained in a few years with long reads. Despite the partial view obtained from this data, its use must be considered to add information in specific regions of graphs, and to resolve local short rearrangements.

Technically speaking, graph augmentation refers to the process of adding new genomic material to the graph, using individuals not present in the graph. The process adds paths by generating new nodes or subdivides existing ones, and inserting extra edges. This process can also enrich an existing graph with information such as the frequency of variations between individuals in a population. This can help identify rare alleles, study allele or genotype frequencies, or refine ratios of core genome (sequences present in all individuals) and dispensable genomes (sequences present in sub-populations only) in subsequent pangenomic analyses . The first step in extending a graph is typically sequencing reads mapping, but this can be a time-consuming process, especially for pangenome graphs. Short reads can be used to detect SNPs, whereas long reads can detect large variants. Of note, regions that are poorly represented in the graph (because of fragmented genomes, for instance in repetitive regions) are not expected to produce new structural variants. Nevertheless, adding information in genes or regions surrounding genes could be helpful for several biological questions, such as detecting variations in promoter regions and their impact on gene expression or mutations leading to novel coding sequences or gene alleles (Omrane et al., 2018).

SNP calling in a pangenome graph can seen counter-intuitive at a first glance: SNP calling is known to be fraught with errors, whereas variations found in pangenome graph are allegedly more accurate. This is why mapping and aligning reads to a graph is a critical step. Currently, few tools can map long reads. *Minigraph* (Li et al., 2020) clips reads into several non-connected match parts,

while *GraphAligner* (Rautiainen and Marschall, 2020) tends to generate a high number of mismatches in indels due to the use of an edit distance instead of an affine-gap cost function. *SVarp* (Söylev et al., 2024) suggests the use of a wavefront alignment algorithm (WFA) to realign the svtigs (clusters of reads) after alignment with *GraphAligner*. This realignment process was implemented in *gaftools realign* (Pani et al., 2024), requiring a two-step process to obtain a reliable alignment. Nevertheless, large insertion or deletion events lead to fragmented alignments that will need a post process treatment to chain them. A recent version of *vg giraffe* (Chang et al., 2025; Sirén et al., 2021) implements a long-read mode to map PacBio Hifi reads and ONT Nanopore sequences, and can, in principle, map read with large indels. Another tool, *Palss* (Denti et al., 2025), performs graph augmentation with an alignment-free approach. Full sequence alignment is replaced by the use of specific anchors on nodes, combined with clustering steps of positioned reads, followed by their local alignment. In the end, alignments in GAM or GAF format are used to augment the graph with *vg augment*, adding each aligned read as a new path. It is crucial to choose stringent parameters in the mapping and in the augmenting steps, in order to discard low quality SNPs. Yet, there is currently no consensus on which parameters should be chosen.

This approach does not associate reads with sequenced individuals, allowing genotype paths to be added to the graph instead of sequenced reads. Graph augmentation seems to be proposed as a way to embed read alignments into graphs to facilitate variant calling. However, it appears to us that it could also be a useful approach to add new variability to the graph with population information. With the increasing quality of long-read sequencing, low-coverage sequencing of populations with long reads will be an interesting approach to augment the graph instead of rerunning the full process of graph construction. To achieve this, it would be useful for existing tools to evolve to utilize information tags such as Read Group (RG) available in GAM/GAF formats. Similarly, incorporating all variations detected in read mapping, especially small variants (SNPs, indels < 3 bp), does not appear relevant, as these can be effectively addressed by standard variant calling methods. Allowing the choice of the type of variant to use to augment the graph is also missing in available tool implementations.

To summarize, it is necessary to improve current methods for long-read mapping, which sometimes fail at aligning long SVs. Moreover, graph augmentation methods should also be improved, especially in the filtering steps, and the node and path modification steps.

Graph annotation. A pangenome graph reveals, by construction, the structural variants between the assemblies it contains. A structural and functional annotation is therefore necessary to identify, biologically interpret, and exploit genomic variations impacting genes and possibly phenotypes. A pangenome graph annotation can be defined as a complete representation of the genomic features, which includes genes and transposable elements, specifying their positions on the different identified paths. As a by-product, this annotation also provides the determination of the genes belonging to the core, dispensable and unique genomes, allowing downstream analyses like CNV detection or identifying gene family gain and loss. One of the challenges of annotating a pangenome is to move from linear to graph-type structure annotation, and to conceive a file format of this annotation.

The annotation method can be investigated in three different strategies: i) directly *de novo* annotate the graph with an algorithm that could explore all paths, ii) linearize the graph and apply annotation tools developed for linear references, iii) project coordinates or map genomic features from linear annotated genomes to the graph. Currently, there is no direct annotation tool for pangenome graphs, although multi-genome annotation tools (Fiddes et al., 2018) exist, and could be adapted. The first annotation strategy thus requires new developments (Wang et al., 2022). Recent genome-wide studies circumvent this lack of tools by linearizing the pangenome so that conventional annotation methods can be used (Jayakodi et al., 2024; MacNish et al., 2024), which corresponds to the second strategy. By iterative alignment of genome assemblies from a diversity panel on a reference, sequences unseen so far are identified and added to the reference to form the linear pangenome. Once new genomic regions are annotated, this linear pangenome can be used for downstream analyses with conventional tools in place of linear reference genome (Wang et al., 2022). A study using the same approach goes a step further by inserting the genomic regions

presenting variations compared to the reference genome directly in the reference sequence, providing the genomic context of the structural variants on which functional analyses can be based (Wang et al., 2023). The third strategy has been more investigated by several tools such as *vg*, *odgi* and *GrAnnoT*. *Odgi position* (Guarracino et al., 2022) and *vg inject/surject* (Garrison et al., 2018) are sub-commands that can be adapted to transfer annotation by projecting coordinates between genomes of a same graph or from a genome to the graph. *Vg* annotate is another sub-command for annotation transfer on a graph. *GrAnnoT* (Marthe et al., 2025), developed by authors of this paper, also transfers a linear genome annotation to the pangenome graph and the graph annotation to other genomes embedded in the graph by projecting coordinates. A limitation of this approach is that the transferred annotation only comes from a genome taken as reference. Therefore, sequences absent from the reference genome (due to a deletion in the reference or insertion in other genomes) are not annotated on the pangenome graph. In principle, it could be possible to extract and annotate these sequences. Even though it is, in principle, as difficult as annotating a linear genome, tools are still lacking. Another issue arises when one wants to combine multiple annotations from different genomes into the same graph: currently, there is no tool available to merge and resolve conflicting annotations that may occur.

Transposable elements (TEs) have been known for decades to cause a wide range of changes in gene expression and function in plants (Lisch, 2012). Having an exhaustive representation of TE insertions at species level using pangenome graphs could enable us to better infer the evolutionary history of TE families and predict their co-option by the host genome. However, TEs are not straightforward to model in a pangenome graph: they tend to accumulate mutations, there are numerous cases of nested insertions, some elements move with a “copy-paste” strategy, and these translocations are difficult to grasp in a graph. There is, to our knowledge, no dedicated tools for TE annotation and analyses in pangenome graphs. It is however possible that graphs are not the best model to analyze TEs, due to their high evolution rates. Users can then resort to population-based tools, including *TrEMOLO* (Mohamed et al., 2023), developed by authors of this work, which compares a reference assembly with another fully assembled genome and long reads, and estimate the frequency of insertion in a population. *GraffiTE* (Groza et al., 2024) also uses long reads, but supports multiple fully assembled genomes for the same species. It produces a reference-based pangenome which includes the reference genome and the TE variations. While both tools operate with a reference, *panREPET* (Saidi et al., 2026), also developed by authors of this work, compares all the TEs of each input genome, and outputs the genomic coordinates of shared and singleton TE copies for each genome. It does not use a graph, however. Pantera (Sierra and Durbin, 2024) uses a PGGB graph to create a library of recently active TE families in a species, but it does not annotate the TE copies in the graph.

Likewise, it would be valuable to annotate pangenome graphs for other genomic distinctive features like DNA methylation, TFBS, TSS, TAD or any else informative loci. For an exhaustive annotation of a pangenome, it will still be necessary to form consortia that would invest time in manual curation of the annotation of the graphs, akin to the GENCODE project (Harrow et al., 2012), with possibly dedicated groups with expert knowledge in gene, TE families, and other types of features (methylation, TFBS, etc). Compared to the linear structure of individual genome, annotating a graph would require additional expertize on how to handle its branched nature.

Once annotation is transferred on pangenome graphs, we may want retrieve gene families to study their content and be able to find discrepancies between genomes. As seen before, finding *trans* genomic similarities (*i.e.* genomic repetitions that are not close together) raises considerable algorithmic issues, given that the graphs are sometimes built chromosome by chromosome, and that sequences may have significantly diverged (at least at the DNA level). In principle, one could imagine a refinement algorithm that would transform a given graph, so as to create bubbles and explicitly represent gene families through added edges. Polyploid genomes add a new layer of complexity, since gene families experience complex histories. Sun et al. (2025) used OrthoFinder (Emms and Kelly, 2019) on the haplotypes of the potato genome, while Huang et al. (2026) built a multiple sequence alignment of the genes of the sugarcane. Both analyses of polyploid genomes (with a mix-ploidy for sugarcan) switched from the genome-based graph to a gene-based graph.

PanTools (Jonkheer et al., 2022) adds extra information layer on the graph (like knowledge graphs), in order to model these families in metadata. This reconciles pan-gene set (Jiao et al., 2024; White et al., 2025) and pangenome graph approaches, adopting the best of both worlds. The first one excels at representing gene family expansions and contractions, whereas the second can model variations inside each gene. Either way, representing genome plasticity in the graph helps researchers to understand how species evolve, by shedding light on lesser known mechanisms, such as gene duplication and neo/sub-functionalization, TE domestication, and TE elimination.

Last, there is no convenient file format for the annotation in pangenome graphs. The Graph Alignment Format (GAF) is a text format, tab-delimited like BED files, which was proposed to represent alignments. Since they basically model paths, they have also been used to store annotations in a pangenome graph. However, its specifications does not include a field for feature types such as CDS, exon or UTR, and tools that index and query them efficiently are only appearing (Novak et al., 2024). A new standard file format, gGFF for graph Generic Feature Format, has been proposed by Llamas et al. (2021). It is a text-based, tab-separated file which generalizes the GFF3 format, and replaces genomic intervals with a subgraph (Llamas et al., 2021). An option (*--ggff*) from *vg annotate* can convert a GFF3 file from one genome into gGFF through a pangenome graph. It could be useful that existing tools that manipulate graphs integrate the gGFF format. However, this format is still not adopted, and it is still not clear how to include annotations of several individuals in a gGFF format.

To summarize, pangenomics still need methods to annotate graphs, represent these annotations in a uniform way, and to merge them from different individuals. Functional annotations, including gene or TE families, should also be adapted, with the caveat that a single graph may not be the best model for representing genome plasticity, and additional information layers may be required.

Producing FAIR pangenomes. As seen in the previous sections, many graphs (or similar data structures) have been produced and already published. However, their sharing respecting the FAIR principles is still problematic. In practice, producing and sharing FAIR pangenomes requires the adoption of standards, across communities of practices, addressing nomenclature, quality control (Section “Graph quality assessment”), data formats, and visualization (Section “Graph visualization”).

First, the names of the assemblies (genomes or haplotypes) should adopt unified nomenclatures, following Cannon et al. (2025) for genome assemblies and PanSN for path names. In case of allopolyploid genomes, a special tag should inform of the ancestry of the homeolog chromosomes. The sequences should also have an external unique identifier directing to the database where it is stored. Once the graph is finalized in its initial version, as discussed before, it could be augmented in various ways, or generated again using a new version of the building tool, or another one. In this regard, using a common and universal versioning and naming system would greatly simplify the identification and re-use of pangenome graphs. The versioning *a minima* is mandatory, as each version, related to a specific URI, will ensure a good reproducibility and finding of the same graph. A minor versioning could be performed when the dataset is not really modified but only meta-information added (such as annotation), and a major one for any change in its global structure (such as adding new paths or new nodes). In addition, even using an identical set of embedded genomes (and the order thereof), a pangenome building tool (together with the versions and the parameters) may produce different graphs, because of the stochasticity of the methods, or because multithreaded implementations order the tasks differently. It thus should be advised to provide all the metadata in the header lines in the GFA file itself, the rationale being that the user should have all the means to build the same graph, when possible. This will inform on the level of “quality” of the graph, its representativeness, its advantages and limits.

The GFA format is the current standard for variation graphs. It exists under different versions (*e.g.* rGFA/1.0, 1.1, 2.0) and if the GFAv1.1 is the most widely used, its implementation in itself varies. Some tools provide paths and other walks; some give additional tags to nodes or links, some none. *vg* defined GAM and GAF formats for aligned reads, with a clear inspiration from the

BAM and PAF formats. For annotations, gGFF format has been suggested, but it is hardly ever used. Current methods use a GAF format to describe an annotation, but it lacks the expressiveness of the GFF format (intron/exon, UTR/CDS). The VCF format can also be used to describe variants in a pangenome graph. It is not easy, however, to describe complex variants, such as nested variants, in this format. *Pantree* (Salehi Nowbandegani et al., 2026) drafted a VCF specification, with customized VCF fields that extends the standard format (while maintaining compatibility) in order to add pangenome-specific information.

Then comes the problem of sharing it in a FAIR way. Non-recommended practices are observed, such as sharing graphs through non-specialized data repositories (Zenodo), and assorted of very basic metadata, if any, except author names. It would be best to collect all the graphs in a dedicated repository, itself hosted in the INSDC databases (NCBI, EBI, DDBJ), with a standardized relationship between the graphs and the embedded sequences, the individuals, the publication, etc. A first suggestion, made for gene-centric pangenomes (Heuermann et al., 2025), could be extended to sequence-based graphs. A unified, consistent, and metadata-rich repository will ensure a long-term preservation of these datasets, improving reproducibility and scientific probity.

Finally, a FAIR pangenome graph file should be linked to derived FAIR data from the entire study, beyond the input FASTA files: annotations, variations, synteny blocks, genes families, among others. State-of-the-art methods could link these diverse objects (Soiland-Reyes et al., 2022). This would facilitate the reuse of the graph, interoperability with existing databases, and implementation of new tools (Arnoux et al., 2024). Pangenome graphs could also be part of genome portals, which include multi-species portals (*e.g.* Ensembl (Dyer et al., 2024)), and specialized ones (*e.g*. Rice Genome Annotation Project (Hamilton et al., 2024) or The Arabidopsis Information Resource (Reiser et al., 2024)). For instance, variations in the graph could be linked with the variation database, in order to assess its frequency in the sampled population, and possibly link to the phenotypes of the individuals carrying the variation of interest.

To summarize, we advocate for transparent practices, that would make any user able to reproduce the same graph using metadata. This metadata information could be stored inside the graph, using normalized keywords, or provided in an external file, in order to keep the graph file as small as possible. Moreover, it is crucial that these graphs, metadata and related information be stored in high-quality, long-term strategic repositories, encoded into community-approved formats, which support querying for a fast and intuitive retrieval and ensure open and fair data licensing.

Using a graph

Manipulating and exploring data from graphs. A pangenome graph contains in itself a wealth of information that users may want to extract. Yet, graphs are very large, and dedicated algorithms should be developed.

The most straightforward graph manipulation is to extract a locus of interest. It is already implemented in *vg chunk* (Garrison et al., 2018) or *odgi* extract (Guarracino et al., 2022). Extracting the loci that share similarity with a given sequence can also be performed through mapping the sequence beforehand. This way, it is possible to focus on a particular gene, and visualize its variations (Section “Graph visualization”). Structural variants, such as tandem repeats, inversions, or translocations, are usually more difficult to extract, and would require additional tools. Ideally, these extractions should be accessible using simple queries, so the user can add several constraints (*e.g*. find all the inversions inside the introns of gene X).

Queries should also be able to compare paths, such as the extraction of all the variations that are specific to a sub-population. For instance, one could be interested in finding all the variations that discriminate between wild and domesticated individuals. Similarly, runs of homozygosity measure inbreeding, and can help reconstructing demographic histories (Shafer and Kardos, 2025). Haplotypes of the same ancestry, in case of allopolyploid genomes, could be compared to haplotypes of another ancestry, in order to characterize the variability between them. This information is already present in the graph, and can extend previous methods to non-SNP cases.

Statistics could also be computed along the graph, and possibly, along a reference genome. They include the number of variations per kilo-base-pair, but also the minor allele frequency, or the percentage of heterozygosity. This could qualify and discriminate the different parts of the genome.

Several types of exports are possible, beyond the GFA. VCF encode variations with respect to a reference, and presence-absence matrices can also be computed with *odgi pav*. Other graphical visualisation, such as principal component analysis, or dendrograms, could also help visualizing the population structure.

Since graphs are large, dedicated index structures should be devised. *Odgi* (Guarracino et al., 2022) implements its own, and Sirén and Paten, 2022 suggested several, depending on the task. Yet, it seems that we did not reach a stable, widely accepted indexing system for the graphs.

To summarize, while toolbox such as *vg* or *odgi* provide many tools for graph manipulations, other, high-level, operations are still lacking for a full exploitation of the graphs.

Population genetics and coalescent theory. A pangenome graph, akin to a multiple sequence alignment, can be considered as an observation of an evolutionary scenario. As such, many questions on evolution could be addressed using this graph. For instance, it is tempting to find a coalescent that would match the pangenome graph, or to run an admixture analysis to find populations. It is indeed a good quality control to check whether the individuals in the graph match the expected breeding history (see for instance the cattle pangenome (Smith et al., 2023)). Discrepancies may hint incorrect or biased assemblies when, for instance, ONT-based assemblies are separated from PacBio HiFi-based assemblies. While in cultivated individuals, the evolution is known, this information could shed light to interesting discoveries in wild individuals, by tracing back the history to putative ancestors. Compared to SNP-based methods, graphs have the advantage to also include structural variants, which may add crucial information.

In principle, finding a distance between genomes, and infer the evolutionary scenario, can be easily done by exploiting the graph. Akin to a mash distance (Ondov et al., 2016), it is possible to extract a Jaccard distance between two individuals by counting the number of common (and private) nodes. Alternatively, it is also possible to compute a distance based on presence/absence of the nodes in a given path (akin to a PAV matrix, where the node replace the genes).

However, contrary to inter-species analyses, inbreeding (crossing individuals with a very recent common ancestry) are frequent inside a population, and should be considered. This is also a difference with a local analysis, spanning the length of a gene, for instance, where this effect can be neglected. It is thus not possible to rely on usual phylogeny to exploit the graph: the phylogenetic tree will differ from locus to locus.

Considering this, it is either possible to model the coalescence from the whole graph using dedicated methods such as split trees (Huson, 1998), or ancestral recombination graphs (Griffiths and Marjoram, 1996). Alternatively, it could be possible to infer local evolutionary scenarios, (partially) independently computed on sub-parts of the graph. This is a complex task since, contrary to a inter-species analysis, not only the branch length should be estimated, but also the shape of the tree (or any formalism that would model the process). Furthermore, segmentation (*i.e.* finding the boundaries where the model should be used) should also be computed, possibly based on genes.

A possible extension of this method is the search for introgression, which is the addition of genetic material from an outgroup. For instance, cultivated individuals can be crossed with wild ones, in order to increase resistance or vigor. This is a major source of genome plasticity in diploid species, even in autogamous crops (Burgarella et al., 2019). Introgression can be detected by finding unexpected patterns while constructing evolution scenarios in localized areas, or adapting *D*-statistics and f_4 statistics to the pangenome graph.

Selection is also a key issue in agronomy. Some phenotypes are highly desirable, and if they are linked with a genotype, the linked alleles tend to be under strong positive selection. This will results in loss of diversity in the regions. Many methods and tools have been developed (Hejase et al., 2020) (involving for instance the *Fst*) to address this issue, yet they primarily focus on SNPs, whereas SV should also included in the analysis. The pattern should, in principle be clearly visible in the graph, since we should observe longer nodes, and less entangled paths. On the contrary,

regions with a significantly high number of variations could be a trace of local adaptation. To date, only *Pansel*, provided by authors of this work (Zytnicki, 2024) tackles this problem. However, it is still centered on one reference genome, and uses a fixed-size sliding window, which limits the scope of the tool. To summarize, it thus crucial to develop methods and tools that could compute different metrics on the graph, which include traces of selection, or signs of introgression, to name a few.

Dual pangenomes. A prospective exploitation of pangenome graphs is querying two or more graphs together in order to search for horizontal transfers, which are an important component in agronomical research, for instance in studying plant holobionts or in insect-based biocontrol. Several methods have been developed to tackle this problem (Wijaya et al., 2025), and provide excellent result to detect transferred genetic material. Yet, using pangenome graphs could also be applied to this case, with several different advantages. A first possibility would be to look for paths or subgraphs with similar sequences which come from two different species.

Even more prospective is the search for correlated trajectories in variations contained in two different graphs. For instance, correlated variant frequency changes in subclades, geographical or ecological populations, may be signals of co-evolution between hosts and symbionts. Similarly, pathogens and hosts can be involved in an evolutionary race for which correlated variation patterns may be detected in both graphs. Developing such approaches will require new solutions to query graph variations based on metadata associated to the genome paths (sub-population, environmental factors, etc.) and to look for graph topologies correlated to these parameters in both graphs, or at least phylogenies which would correlate with an adaptation of pathogen to host. Candidate regions would then be investigated for possible changes in proteins, and would explain how interactions would take place. In the near future, we expect that pangenome graphs will be commonly used in population genomics to detect regions involved in adaptation. The pangenome graphs of *Fusarium oxysporum*, a major plant pathogen, revealed the complexity of linking accessory chromosome content and host specificity (van Westerhoven et al., 2024) when a broad range of hosts could be attacked by the same pathogen. In the case of polyploid plant genomes such as Brassica napus (*Brassica oleracea x Brassica rapa*), or bread wheat / durum wheat, some pathogens specifically target one ancestral genome. Including ancestral and hybrid genomes in a graph will help to identify sequence variability and evolution between new and ancestral chromosomes in regard to a graph of pathogen strains with specific patterns of adaptation. This kind of approach will help to understand the molecular determinants of adaptation, as studied between *L. maculans* and *B. carinata* (Noah et al., 2024). To summarize, variations in the host pangenome graph could be compared with variations in the pathogen graph to find traces of adaptation.

Pangenome graphs for metagenomics and organelle-based studies. Pangenomes graphs were recently proposed as a mean to classify metagenomic sequences from a taxonomic or functional point of view.

A first approach is to base this classification process on organelle-based graphs. While assembly graphs were already derived to study large structural rearrangements in mitochondria, such as *Master-graph* (He et al., 2022), to our knowledge the *vgan* package (https://github.com/grenaud/vgan) is currently the only attempt of organelle-based graphs for tasks of classification ranging from human sub-populations to taxonomic classification of environmental metagenomes. All *vgan* tools have in common a first step where mitochondrial query reads are mapped to a concatenate of mitochondrial variation graphs, then mapping results are exploited in probabilistic frameworks to produce a classification. Currently, this process got adapted to three categories of analyses. *Haplocart* (Rubin et al., 2023) is dedicated to human haplogroup classification: mitochondrial variations help to compute the likelihood that a sample belongs to a specific sub-population. *Euka* (Vogel et al., 2023) aims for taxonomic classification and abundance estimation for ancient environmental DNA (aeDNA) samples. A set of mitochondrial graphs is built for each targeted taxon (family, genus, species, etc.) and mappings are filtered using two estimators : their likelihood to result from aDNA-specific degradation and the likelihood of the mapping given the sequence diversity variations hold by the collection of taxon-specific graphs.

Finally, *Soilbean* (Vogel et al., 2024) extends on *Euka*'s process with a Bayesian inference to refine aeDNA classifications even for the least abundant taxa. Note that while the authors provide all necessary tools and pre-computed graph databases to classify new datasets, it is not always documented how to build a custom database for one's taxa of interest.

Going further, taxonomic classification combined to gene family functional analysis was proposed by the procaryote-centered, graph-based tool, *StrainFLAIR* (Da Silva et al., 2021). In this approach genes are extracted from a set of genomes, then families are built via sequence clustering. A variation graph is built for each family's alignment and query metagenomes can be mapped to this set. Then an algorithm relying on global coverage of mappings in all family graphs is used to produce a strain-level classification as well as relative abundance of the strains.

These applications represent a potential preliminary step toward an extended use of pan-genome graphs in typical ecosystem-integrated ("one health") projects where a agronomic ecosystem is studied in its entirety, *e.g.* from communities of domestic species to soil or aquatic ecosystems with which they interact. Strikingly, we did not find any similar approach exploiting plastid genomes, to the exception of some assembly graph based studies (Liu et al., 2023). The potential of plastid-based variation graphs to answer plant driven questions in agroecosystems remains to be explored.

Downstream analysis using omics data

Variant analysis and genotyping. Commonly, genotyping a large number of samples for representative variants requires two successive steps. First, a set of representative genomic variants is obtained from a limited number of samples by read mapping to a linear reference genome, variant discovery and variant callset merging and curation. Second, the curated variant set is genotyped in a larger number of samples, typically sequenced with shorter reads or at a lower read depth. For this second step, the benefits of representing genomic variants in a graph has already been demonstrated. Indeed, most recent variant genotypers now reduce the reference bias in the read mapping and genotype inference steps by using a graph representation of the variant alleles, making the genotyping more accurate, such as *BayesTyper* (Sibbesen et al., 2018), *GraphTyper2* (Eggertsson et al., 2019), *Paragraph* (Chen et al., 2019), *vg call* (Hickey et al., 2020), *Pangenie* (Ebler et al., 2022), *Varigraph* (Du et al., 2025) with short reads, and *SVJedi-graph* (Romain and Lemaitre, 2023), developed by authors of this work, with long reads. The benefits are even greater for structural variants compared to SNPs and indels (Hickey et al., 2020), and for variants that are close to each other, overlapping or nested in one another (Romain and Lemaitre, 2023).

These genotypers are based on the quantification of reads supporting each of the variant alleles in the graph, either by read-to-graph mapping (*GraphTyper2, Paragraph, SVJedi-graph*) or by alignment-free approaches based on *k*-mer set comparisons (*BayesTyper, Pangenie, Varigraph*). Several methods thus infer genotypes based on read or *k*-mer counts along variant paths in graphs. However these methods are not directly applicable to pangenome graphs built from whole-genome alignments, because they all build their own graph by augmenting the reference genome with a set of already characterized variants with respect to this reference. They cannot take any sequence graph as input, and the types of variants they can genotype are limited to simple types and/or those that can be described in relation to a reference genome in VCF format. For instance, to apply *Pangenie* to the variants derived from the draft human pangenome reference (HPRC) built with *Minigraph-Cactus*, an intermediate variant formatting step was required. This involved representing the variants relative to a selected reference genome and retaining only those variants compatible with *Pangenie* in the input set (Liao et al., 2023).

Pangenome graph building tools have their own pipeline to genotype variants directly in their graph, the first step of which is to detect the variants as bubble motifs in the graph (Section "Graph quality assessment"). Genotyping of these bubbles can then be performed directly on the whole graph by read-to-graph alignment methods with *vg giraffe* (Sirén et al., 2021) for *Minigraph-Cactus* and *PGGB* graphs or *Minigraph*-dedicated mapper (Li et al., 2020). However, when the number of samples to be genotyped is large, read-to-graph alignments can be very resource-intensive and

lead to scaling issues. *Pangenie* (Ebler et al., 2022) constitutes an alternative that scales better. It is based on *k*-mers, and is thus to be used on de Bruijn graphs. *Pangenie* requires a reference genome and cannot handle all possible bubbles (complex nested bubbles and cycles are not allowed, see examples in Figure 4). A more general method is thus to be developed.

One critical step in these pipelines lies in the interpretation and characterization of the obtained variants. A bubble can represent a wide range of variant types from single isolated SNPs, to indels and large structural variants. Moreover, there is no one-to-one relationship between bubbles and variants: a bubble can represent several variants combined (for example, for variants that are close to or nested within each other), and the same structural variant can be represented in several overlapping and intricate bubbles. The combination of different types of variants in the same locus, or the inclusion of small variants within larger variants, leads to a multitude of complex and intricate topological motifs in the graph. Apart from assessing the ability of bubble callers to exhaustively detect all these motifs as bubbles, the main problem lies in the fact that all these different types of variants are reported as the same generic bubble motif, with little in-formation, namely a position in the reference genome and the different walks in the graph taken by the different alleles. While isolated SNPs are easily identifiable by their distinctive allele path size of 1 bp, distinguishing and characterizing all the other variant types and their various combinations is much more difficult. Currently, tools like Bubblewave (Liao et al., 2023) are limited to insertions and deletions, and ignores other SV types, such as inversions, duplications, transpositions. As mentioned before, bubble detection tools are limited by the ability of the pangenome graph to detect these complex topologies.

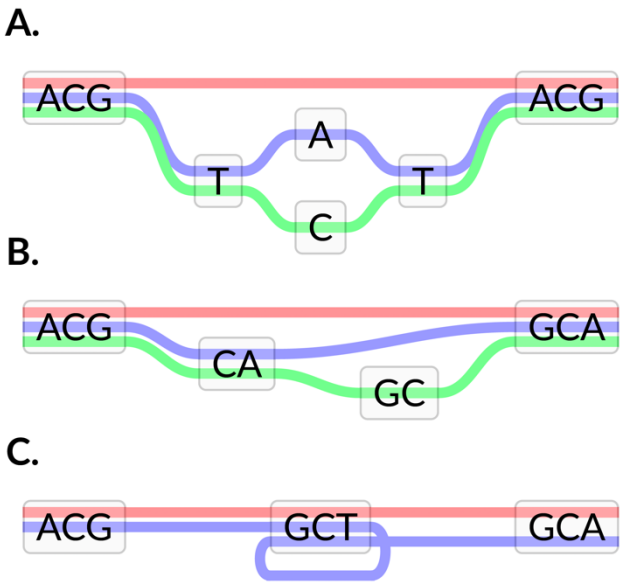

Figure 4 – Examples of nested and complex variants that are difficult to represent and genotype. In these three examples, we suppose that the reference genome is in red. In **A**., there is a SNP with two different alleles (A in the blue genome, and C in the green genome) in a region that is absent from the reference. This variant is likely to be represented as an insertion with respect to the reference genome with two possible alternative sequences (TAT and TCT). In this case, genotyping will likely genotype the different alternative sequences, but not the individual SNP themselves. In **B**., we observe two cumulative insertions, when compared to the reference genome: CA is inserted in the blue and green genome, and GC is further inserted in the green genome after CA. This forms overlapping bubbles that may be missed by vg deconstruct and cannot be genotyped for instance with *Pangenie*. In **C**., the GCT sequence is duplicated in the blue genome, forming a loop in the graph, which is not allowed in *Pangenie*.

A recent work investigating the representation of large inversions in pangenome graphs, carried out by authors of this article, showed that few of them are easily identifiable among the bubbles reported by bubble callers and that their representations vary greatly between the different pangenome pipelines (Romain et al., 2025). There is therefore a need to investigate the topological motifs generated by complex variant types, and combinations thereof, and then to provide automated tools to detect and annotate them in the graphs. Furthermore, it remains to be assessed how comprehensive and accurate are current bubble detection tools even for simple SV types, given that there are currently no published studies aiming at benchmarking pangenome graph and bubble detection tools with controlled variant sets. Finally, even without precisely characterizing the variants, current pipelines still rely on a reference genome in intermediate steps and especially in the VCF output format. Thus, representing and analyzing these variants without a reference genome remains a challenge.

Association studies. Genome-wide association studies (GWAS) are a widely used approach in genomics to identify genetic variants associated with specific traits or diseases. These studies typically begin with a large set of single nucleotide polymorphisms (SNPs) located on a linear reference genome. These SNPs are genotyped in contrasting populations, where individuals are sequenced using short reads at low coverage (Uffelmann et al., 2021). Today, thousands of studies have identified large sets of causal variants associated with heritable traits. However, these variants are often reported alongside phenotypically unrelated but physically close variants due to linkage disequilibrium (Watanabe et al., 2019). A first extension of this method consisted in including structural variants, usually detected with long reads mapped to a reference genome, and adapting available methods. This provided impressive results (Wang et al., 2024), with several structural variants discovered, but still relied on a unique reference genome.

Recently, a new GWAS approach without a reference genome uses *k*-mer count matrices (Corut and Wallace, 2023; Voichek and Weigel, 2020). This method, called *kmerGWAS*, aims to identify directly trait associated *k*-mers from a large set of resequencing samples. However, it suffers from the difficulty of repositioning the very short sequences on a reference in order to identify the underlying genomic regions. Methods for estimating the abundance of unitigs (Vicedomini et al., 2025) could also be, in principle, used. This method has several advantages, but it raises new, unanswered questions. Since the *k*-mers are considered independently, it is difficult to verify that the linkage disequilibrium cannot be leveraged in order to get stronger results. Moreover, *k*-mer counts are usually not independent: several *k*-mers encode the same variant allele. And, in case of nested variations, tests are not independent either.

The efficiency of the GWAS highly depends on the genotyping quality. Furthermore, common statistical methods, based on linear mixed models (Lipka et al., 2012; Mbatchou et al., 2021; Widmer et al., 2014; Zhou and Stephens, 2014) often do not take missing data into account, and are limited to biallelic variants, so it is common practice to reduce the variant input set to SNPs characterized in the vast majority of individuals. However, while pangenome graph approaches could make it possible to genotype a greater diversity of genetic variations, the current efficient alignment-free tools based on *k*-mers such as *Pangenie* (Ebler et al., 2022) or *Varigraph* (Du et al., 2025) fail to take into account the complexity of the graph because they need a VCF as input, which can be obtained by scanning the graph with a dedicated tool (above Section “Variant analysis and genotyping”) but consist of a simplification of the graph structure. On the other hand, another common approach (Asri et al., 2025) is to align the reads onto the graph with *vg giraffe* (Sirén et al., 2021), surject to a linear genome, and call the variants with *DeepVariant* (Poplin et al., 2018). The latter method has the advantage to consider the entire graph and should be able to identify new smaller variants not present in the pangenome graph, but it is very CPU intensive. Interestingly, Huang et al. (2026) developed a method that considers the dosage of homeolog genes, and added it into the model. Nevertheless, only a few studies (Liu et al., 2024; Yang et al., 2025; Zhou et al., 2022) used the pangenome graph in order to establish a convenient reference calling set for the association studies so far.

To leverage pangenome graphs for association studies, new tools must be created. One initial direction could investigate alignment-free variant calling directly on the graph, instead of first

exporting it as a VCF file through a reference-based simplification process. These tools could identify sub-sequences representative of specific genome paths and scan resequenced samples with a *k*-mer approach. Conversely, another strategy would be to align short sequences identified through a *kmerGWAS* approach onto a pangenome graph, and exploit its rich structure to characterize complex regions associated with a phenotype. These methods could use the full graph, or focus on the variations that have been previously exhibited in the sequencing data, and possibly added to the graph. Imputation could also be implemented in pangenome graphs.

Indeed, due to cost or technical constraints, sequencing depth is often sacrificed to increase sample size, resulting in high levels of missing data that complicate statistical analyses. Then, untyped variants are usually imputed using a phasing approach with reference individuals or populations (Browning et al., 2018). Employing the pangenome graph as a backbone for phasing facilitates the inference of missing alleles. This is the method used by Bradbury et al. (2022) in their framework, the *The Practical Haplotype Graph*. They use a treillis graph, where a hidden Markov model can be used to retrieve the haplotype. Yet, a general imputation tool still is missing.

Furthermore, considering that more and more individual genomes will be sequenced in high quality and therefore included in pangenomes, it should be interesting to directly use the graph as input of the classical GWAS. It should be possible to use the nodes and their corresponding paths to construct a presence/absence matrix and perform association tests using the phenotypes of the individuals included in the graph, and first tools have been recently suggested in this direction (Vorbrugg et al., 2024; Zhang et al., 2026). Belinchon-Moreno et al. (2025) used this method in order to find new nucleotide-binding domain leucine-rich repeat receptor genes (NLRs) associated with resistance in melon, and found new loci using graph-based approaches. Yet, the independence of the tests is not clearly solved. Alternatively, topologies representing variations, such bubbles, or snarls, could be analyzed, and paths forming these structures could be genotyped, and used for a GWAS. Currently, the number of genomes in a graph usually does not exceed a few dozens, and it is far from sufficient for an analysis.

Last, it is crucial to be able to include knowledge gathered in the last decades into a pangenome analysis. In principle, it is possible to include previously found SNPs in a pangenome graph, and perform the GWAS on the enriched graph. Previously generated sequencing data can also be used to run an analysis on the graph, instead of a linear genome. It is then expected that more QTLs, or more precise QTLs, will be found (Zhou et al., 2022). In the case of biparental or multi-parental analyses, it is even possible to trim the graph to the variations known to be present in order to reduce the complexity of the analysis (Sirén et al., 2024). Chip data are more difficult to associate to a graph, but they could be used to impute the whole genome sequence of an individual —a task that has not been addressed yet.

To summarise, although SVs from pangenome graphs have already been used on different GWAS approaches, we need new GWAS methods and tools that could directly take advantage on the graph instead of a linear genome.

Towards multi-omics pangenomics. Multi-omics integration refers to the inclusion of different sources of molecular information, such as genomics, transcriptomics, proteomics, epigenomics (methylation, histone marks, and chromosome conformation), metabolomics, and ionomics. We can foresee that, in a near future, it will be affordable to collect several omics data types together with the genomes of interest. The aim is to get a more exhaustive analysis of the cellular processes and interactions, and possibly characterize the link between genotype and phenotype. As such, pangenomics can be considered as a part of a multi-omics analysis, and several databases gather omics data for different individuals (Hu et al., 2024). However, today, most of these data types are still projected to the reference genome.

Graph-based methods can thus be used to remove this reference bias (Danilevicz et al., 2020; Hu et al., 2024; The Computational Pan-Genomics Consortium, 2018; Wang et al., 2022; Zanini et al., 2021). Genotyping, using arrays or sequencing, usually leverages SNPs in order to characterize genome variability, and can also be integrated to other omics analyses. A graph-based method is expected to increase the number of putative variants to be tested, included non-reference variants, and SVs.

Linking variations with one other omic data can be considered as an extension of the GWAS method. For instance, eQTL links genotypes with gene expression, but other analyses have been defined for proteomics, metabolomics, etc. In principle, it is thus possible to adapt methods presented in Section “Association studies” in order to tackle this question.

However, a major difficulty here is to collect information provided by sequencing data. Using a reference assembly, reads produced by RNA-Seq, ChIP-Seq, or other, usually are mapped and compared to gene locations, in order to access gene expression, histone marks, etc. Using pangenome graphs significantly complexifies this step: it requires an exhaustive annotation of the whole pangenome graph, as well as dedicated mapping and quantification tools. Some of these tools exist: *Graph Peak Caller* (Grytten et al., 2019) can be used to find ChIP-Seq peaks, and Sibbesen et al. (2023) use a combination of *vg rna*, *vg mpmap*, and *RPVG*, in order to map RNA-Seq data and quantify the expression. Interestingly, (Huang et al., 2026) showed that, compared to linear genomes, a graph-based approach improves the mapping step for autopolyploid genomes, since homeolog loci are collapsed into the same nodes; this process decreases by half the number of multi-mapped reads, which are usually discarded in downstream analyses. In the same work, the authors showed an interesting example of a structural variant modifying the chromatin accessibility, with implications on the gene expression in *cis* (assessed with ATAC-Seq and RNA-Seq respectively). However, the efficiency of these tools remains to be established, and tools for other uses (methylation, chromosome conformation, etc.) are still lacking. These data can then be aggregated using available methods, such as the mixOmics package (Rohart et al., 2017).

Another difficulty is to represent the omics data on the graph. In linear genomes, they are modeled as genomic intervals, and could be translated as sub-paths. These sub-paths should be efficiently stored and indexed, with dedicated data structures (Novak et al., 2024). Now, it is crucial to estimate distances between sub-paths, such as the distance between a promoter (found using ChIP-Seq) and a TSS (found using RNA-Seq). This distance is not clearly defined in a graph, but investigations are more than welcome here. In general, an extension to the operations included in the bedtools suite (Quinlan and Hall, 2010) to the graph would be helpful to the community.

To summarize, we have to adapt tools that use linear genomes for each omics data individually, and also need methods that would compare them. Transitioning from genomic intervals to graph-based regions is a specially complex task.

Conclusions

This work highlights several aspects of pangenomics that still cannot be satisfactorily tackled due to lack of dedicated methods and tools. To date, no need is totally covered. Pangenome builders, such as *Minigraph-Cactus* (Hickey et al., 2023) and *PGGB* (Garrison et al., 2024), are still satisfactory, but some aspects, so as the consistency of results (two identical executions do not always produce the same graphs), could be improved. Reads can be mapped using vg (Chang et al., 2025; Sirén et al., 2021), but they have problems with loops in graphs, which limits its use to *Minigraph-Cactus* graphs. Genotyping can also be done with the same tool (Hickey et al., 2020), yet results are sometimes suboptimal. Other uses are even less covered. The most crucial needs include proper quality assessment, practical visualisation tools, stable file formats and better methods to exploit graphs. Although it may look desperate at first glance, it also means that there are new avenues for developments, which will require closer collaboration between disciplines: computer scientists, statisticians, etc. on the one hand, and biologists, agronomists, etc. on the other hand. The latter group could detail and augment the wish-list presented here, whereas the first group could develop the much needed methods and tools.

Based on the previous sections, we summarize in Table 1 the list of developments that should be implemented in order to make a full use of a pangenome graph. Acknowledgedly, working on graphs with hundreds of millions nodes, as it is common in pangenomes, is hard. This is why the community also needs to develop algorithmic tools. Essential concepts for genomics, such has the Burrows–Wheeler transform (Burrows and Wheeler, 1994), seminal for read mapping on a linear genome (Li and Durbin, 2009), has been extended to graphs (Novak et al., 2017). However, other

concepts, including sketching (for example with minimizers) (Ndiaye et al., 2024) should also be extended to the non-linear case. Indeed, since many nodes can be much smaller than the *k*-mer size, indexing them is not trivial, and new indexing strategies could be devised. All the questions mentioned in this work should be accompanied with methodological developments. In particular, new integrative systems should be developed specifically for graphs and their derivatives, in order to enable user-friendly manipulation, analysis and visualisation in the light of other evidence. New artificial intelligence methods could help to make these systems faster and more accurate.

Links to other pangenomics frameworks should also be investigated further. Pangenomics has been coined twenty years ago, and already benefited from many developments. For instance, the gene-centric view highlights gene family variation/extension/reduction (Shaiber et al., 2020), an information which is not available using current tools working on pangenome graphs. De Bruijn graphs also implemented useful features, which could be adapted to variation graphs. PanTools (Jonkheer et al., 2022), for instance, includes several methods for phylogeny reconstruction. This tool also incorporates a knowledge base, which can store structural and functional annotations, gene families, etc. This idea could be generalized to other types of graphs, and support complex queries. Graphs could also be equipped with links to other resources (*i.e.* gene expression, or phenotype databases), and be explored using Web portals. An other possibility of development is an interface to metabolic networks. In principle, it could be possible to predict the impact of variations which are observed in a individual, and to compute the expected outcome in the network, paving the way to a personalized metabolism.

Of note, we chose to mention applications to agronomy. Pangenomics is obviously not restricted to this field, and could be used for studying ancient DNA, or cancer evolution, among others. Several of the questions mentioned here could be applied there.

It is also clear that the computer science community should help to perform this transition, and this work is under way. Numerous reviews explaining the concept are available (*i.e.* Matthews et al., 2024). Because of the a wealth of new methods, it is difficult to match a tool with a need. Fortunately, the community maintains a few lists, such as awesome-pagenomes (Davenport, 2026) that cover the recent developments, organized in topics. Some scientific organizations and private companies now offer courses on pangenomics for newcomers, and we can expect more in the future. Tools may also be difficult to install and use. Several pipelines have been developed in order to ease this process: an nf-core pangenome pipeline (Ewels et al., 2020), and a snakemake pipeline (made by authors of this work, Pan1C: https://forge.inrae.fr/genotoul-bioinfo/Pan1c/pan1c build a pangenome graph (together with a list of variants, and several visualizations) with minimal effort, using tools encapsulated in containers. To date, it seems that most of pangenome graph tools are not available as Galaxy plugins (Abueg et al., 2024); this could however facilitate its use in pipelines. Computational requirements can also be a limitation. One of the largest pangenome graph made so far (Liao et al., 2023) is the human one, with 47 diploid individuals. *Minigraph-Cactus* and *PGGB* took a few days (but the process can be parallelized), and a few hundreds of GB RAM, to build the graph. Using a compressed algorithm (Sirén and Paten, 2022), the graph can be stored in a 6Gb file, although it stores 94 haplotypes of size ~3Gb. Even though it is more complex to map a read to a graph compared to a linear genome, it is not necessarily more time consuming: *vg giraffe* (Sirén et al., 2021) is faster than *Bowtie2* (Langmead and Salzberg, 2012) and BWA-MEM (Li, 2013). It requires however more RAM than other tools (several tens of GB RAM). Sirén et al. (2021) also estimated that genotyping a sample took about 200 CPU-hours, and can also be efficiently parallelized. To summarize, although pangenomics still is at its infancy, and more complex than genomics, the computer science community is currently trying to build bridges in order to facilitate its adoption.

Finally, this paper only mentions the different use cases the authors have been confronted to so far. Without doubts, any improvement, tool, or implemented feature, will unlock new possibilities, and eventually new results, which will trigger new questions. We are far from having explored the wealth of pangenomics methods. It is not the intent of the authors to suggest that pangenomics will replace all previously designed methods, involving for instance SNPs and reference genome. We hope, however, that pangenomics may explain new molecular mechanisms, infer past

ancestries, or, prospectively, open new paths for improving breeding of agronomic species. Interestingly, Cheng et al. (2025) suggested a new breeding strategy in order to get as close as possible to an ideal genotype of potato, based on pangenome studies. We believe that this is the future of pangenomics.

Table 1 – Recommendation of developments and standard for an improved exploitation of pangenome graphs.

Topic	What our community needs
Diversity	A robust and standard method for assessing the diversity of a population, and suggesting individuals that would maximize it, given a budget.
Graph quality	Guidelines for uniforming SV ambiguous representations in graphs.
Graph quality	Metrics for quantifying the correctness of the graph with respect to the input genomes.
Graph quality	Metrics for quantifying the succinctness or usability of a graph.
Visualization	Indexing data structures and tools for a nucleotide-to-chromosome visualization.
Augmentation	Optimized methods for long-read mapping to graphs.
Annotation	A dedicated format for annotation in graphs.
Annotation	Methods and tools for annotating complex features such as TEs.
Annotation	Methods and tools for merging several annotations.
Annotation	Methods for representing gene families.
FAIR	Standard graph format with exhaustive information on graph construction.
FAIR	Dedicated databases for graph storing and indexing.
Manipulation	High-level, easy to use queries for sub-graph extraction.
Population genetics	Methods for phylogeny inference, with traces of domestication or introgression.
Dual	Methods for pairwise graph comparisons.
Metagenomes	Methods for exploring and comparing meta-pangenomes.
Genotyping	Improved methods for complex SV detection.
Genotyping	File format for graph-based variations.
Association studies	Methods and tools to perform GWAS on the graph.
Omics	Methods for RNA-Seq, ChIP-Seq, etc. on graphs.
Omics	Interval comparison from linear genomes to graphs.

Acknowledgements

Preprint version 6 of this article has been peer-reviewed and recommended by Peer Community In Genomics (https://doi.org/10.24072/pci.genomics.100526; Campos–Dominguez, 2026).

Fundings

Several authors of this work have been founded by the “AgroDiv” project of the Agroecology and Digital Technologies research program and received government funding managed by the Agence Nationale de la Recherche under the France 2030 program, reference ANR-22-PEAE-0005.

Conflict of interest disclosure

The authors declare that they comply with the PCI rule of having no financial conflicts of interest in relation to the content of the article. François Sabot is a recommender for PCI Genomics.

References

Abueg LAL, Afgan E, Allart O, Awan AH, Bacon WA, Baker D, Bassetti M, Batut B, Bernt M, Blankenberg D, Bombarely A, Bretaudeau A, Bromhead CJ, Burke ML, Capon PK, Čech M, Chavero-Díez M, Chilton JM, Collins TJ, Coppens F, et al. (2024). The Galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update. Nucleic Acids Research **52**, W83–W94. https://doi.org/10.1093/nar/gkae410.

Adam M, Fréville H, Alami S, Marrou H, Pot D, Ricci S, Thomas M, Guichardaz A, Billot C, Chantret N, Gouesnard B, Louafi S, Muller E, Nguepjop JR, Pradal C, Rhoné B, Sidibé-Bocs S, Tavaud M, Brocke KV, Joly H (2025). Embracing new practices in plant breeding for agroecological transition: A diversity-driven research agenda. Plants, People, Planet **8**, 1, 38–48. https://doi.org/10.1002/ppp3.70072.

Andreace F, Lechat P, Dufresne Y, Chikhi R (2023). Comparing methods for constructing and representing human pangenome graphs. Genome Biology **24**, 274. https://doi.org/10.1186/s13059-023-03098-2.

Arnoux J, Bonifati A, Calteau A, Dumbrava S, Gautreau G (2024). Integrating Complex Pangenome Graphs. In: 2024 IEEE 40th International Conference on Data Engineering Workshops (ICDEW). IEEE, pp. 350–354. https://doi.org/10.1109/icdew61823.2024.00052.

Asri M, Chang PC, Mier JC, Sirén J, Eskandar P, Kolesnikov A, Cook DE, Brambrink L, Hickey G, Novak AM, Dorfman L, Webster DR, Carroll A, Paten B, Shafin K (2025). Pangenome-aware DeepVariant. bioRxiv [Preprint]. https://doi.org/10.1101/2025.06.05.657102

Belinchon-Moreno J, Berard A, Canaguier A, Le-Clainche I, Mistral P, Leyre K, Rittener-Ruff V, Lagnel J, Hinsinger DD, Faivre-Rampant P, Boissot N (2025). Intra-specific NLR allelic diversity and genomic landscape for plant resistance association studies. bioRxiv [Preprint]. https://doi.org/10.1101/2025.08.01.668142.

Beyer W, Novak AM, Hickey G, Chan J, Tan V, Paten B, Zerbino DR (2019). Sequence tube maps: making graph genomes intuitive to commuters. Bioinformatics **35**, 24, 5318–5320. https://doi.org/10.1093/bioinformatics/btz597.

Bin Rahman AR, Zhang J (2013). Rayada specialty: the forgotten resource of elite features of rice. Rice **6**, 41. https://doi.org/10.1186/1939-8433-6-41.

Bradbury PJ, Casstevens T, Jensen SE, Johnson LC, Miller ZR, Monier B, Romay MC, Song B, Buckler ES (2022). The Practical Haplotype Graph, a platform for storing and using pangenomes for imputation. Bioinformatics **38**, 15, 3698–3702. https://doi.org/10.1093/bioinformatics/btac410.

van den Brandt A, Jonkheer EM, van Workum DJM, van de Wetering H, Smit S, Vilanova A (2024). PanVA: Pangenomic Variant Analysis. IEEE Transactions on Visualization and Computer Graphics **30**, 8, 4895–4909. https://doi.org/10.1109/tvcg.2023.3282364.

Browning BL, Zhou Y, Browning SR (2018). A One-Penny Imputed Genome from Next-Generation Reference Panels. The American Journal of Human Genetics **103**, 3, 338–348. https://doi.org/10.1016/j.ajhg.2018.07.015.

Burgarella C, Barnaud A, Kane NA, Jankowski F, Scarcelli N, Billot C, Vigouroux Y, Berthouly-Salazar C (2019). Adaptive Introgression: An Untapped Evolutionary Mechanism for Crop Adaptation. Frontiers in Plant Science **10**. https://doi.org/10.3389/fpls.2019.00004.

Burrows M, Wheeler D (1994). A block-sorting lossless data compression algorithm. Technical Report SRC-RR-**124**. Digital Equipment Corporation.

Campos Dominguez, L. (2026) Bridging the gap to graph-based plant genomics. Peer Community in Genomics, 100526. https://doi.org/10.24072/pci.genomics.100526.

Cannon EKS, Molik DC, Wright AJ, Zhang H, Honaas L, Chougule K, Dyer S (2025). Guidelines for gene and genome assembly nomenclature. Genetics **229**, 3. https://doi.org/10.1093/genetics/iyaf006.

Chang X, Novak AM, Eizenga JM, Sirén J, Monlong J, Negi S, Andreace F, Nag S, Kyriakidis K, Hickey G, Hwang S, Délot EC, Carroll A, Shafin K, Chang PC, Okamoto F, Paten B (2025). Rapid, accurate long- and short-read mapping to large pangenome graphs with vg Giraffe. bioRxiv [Preprint]. https://doi.org/10.1101/2025.09.29.678807.

Chen J, Liu Y, Liu M, Guo W, Wang Y, He Q, Chen W, Liao Y, Zhang W, Gao Y, Dong K, Ren R, Yang T, Zhang L, Qi M, Li Z, Zhao M, Wang H, Wang J, Qiao Z, et al. (2023). Pangenome analysis reveals genomic variations associated with domestication traits in broomcorn millet. Nature Genetics **55**, 12, 2243–2254. https://doi.org/10.1038/s41588-023-01571-z.

Chen S, Krusche P, Dolzhenko E, Sherman RM, Petrovski R, Schlesinger F, Kirsche M, Bentley DR, Schatz MC, Sedlazeck FJ, Eberle MA (2019). Paragraph: a graph-based structural variant genotyper for short-read sequence data. Genome biology **20**, 291. https://doi.org/10.1186/s13059-019-1909-7.

Cheng H, Concepcion GT, Feng X, Zhang H, Li H (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nature Methods **18**, 2, 170–175. https://doi.org/10.1038/s41592-020-01056-5.

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 WJ, Garrison E, et al. (2025). Leveraging a phased pangenome for haplotype design of hybrid potato. Nature **640**, 8058, 408–417. https://doi.org/10.1038/s41586-024-08476-9.

Corut AK, Wallace JG (2023). kGWASflow: a modular, flexible, and reproducible Snakemake workflow for k-mers-based GWAS. G3: Genes, Genomes, Genetics **14**, 1. https://doi.org/10.1093/g3journal/jkad246.

Cui Y, Peng C, Xia Z, Yang C, Guo Y (2025). A survey of sequence-to-graph mapping algorithms in the pangenome era. Genome Biology **26**, 1. https://doi.org/10.1186/s13059-025-03606-6.

Cumer T, Milia S, Leonard AS, Pausch H (2026). PG-SCUnK: measuring pangenome graph representativeness using single-copy and universal K-mers. BMC Bioinformatics **27**, 29. https://doi.org/10.1186/s12859-025-06355-2.

Da Silva K, Pons N, Berland M, Plaza Oñate F, Almeida M, Peterlongo P (2021). StrainFLAIR: strain-level profiling of metagenomic samples using variation graphs. PeerJ **9**, e11884. https://doi.org/https://doi.org/10.7717/peerj.11884.

Dabbaghie F, Ebler J, Marschall T (2022). BubbleGun: enumerating bubbles and superbubbles in genome graphs. Bioinformatics **38**, 17, 4217–4219. https://doi.org/10.1093/bioinformatics/btac448.

Danilevicz MF, Tay Fernandez CG, Marsh JI, Bayer PE, Edwards D (2020). Plant pangenomics: approaches, applications and advancements. Current Opinion in Plant Biology **54**, 18–25. https://doi.org/10.1016/j.pbi.2019.12.005.

Davenport C (2026). awesome-pangenomes. Zenodo [Preprint]. https://doi.org/10.5281/zenodo.18208384.

De Beukelaer H, Davenport G (2016). corehunter: Multi-Purpose Core Subset Selection. CRAN R projet. https://doi.org/10.32614/cran.package.corehunter.

Degardins B, Mouton M, Guillon B, Paperman C, Marchet C (2024). Vizitig: A visual tool for colored de Bruijn graphs exploration. Zenodo [Preprint]. https://doi.org/10.5281/zenodo.13929021.

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

Du ZZ, He JB, Xiao PX, Hu J, Yang N, Jiao WB (2025). Varigraph: An accurate and widely applicable pangenome graph-based variant genotyper for diploid and polyploid genomes. Molecular Plant **18**, 9, 1587–1601. https://doi.org/10.1016/j.molp.2025.08.001.

Dubois S, Zytnicki M, Lemaitre C, Faraut T (2025). Pairwise graph edit distance characterizes the impact of the construction method on pangenome graphs. Bioinformatics **41**, 6, btaf291. https://doi.org/10.1093/bioinformatics/btaf291.

Durant É (2022). Design of novel visual representations and tools applied to plant pangenome visualization. PhD thesis. Université de Montpellier. URL: https://theses.hal.science/tel-04135739.

Durant É, Sabot F, Conte M, Rouard M (2021). Panache: a web browser-based viewer for linearized pangenomes. Bioinformatics **37**, 23, 4556–4558. https://doi.org/10.1093/bioinformatics/btab688.

Dyer SC, Austine-Orimoloye O, Azov AG, Barba M, Barnes I, Barrera-Enriquez VP, Becker A, Bennett R, Beracochea M, Berry A, Bhai J, Bhurji SK, Boddu S, Branco Lins PR, Brooks L, Ramaraju SB, Campbell LI, Martinez MC, Charkhchi M, Cortes LA, et al. (2024). Ensembl 2025. Nucleic Acids Research **53**, D1, D948–D957. https://doi.org/10.1093/nar/gkae1071.

Earl D, Bradnam K, St. John J, Darling A, Lin D, Fass J, Yu HOK, Buffalo V, Zerbino DR, Diekhans M, Nguyen N, Ariyaratne PN, Sung WK, Ning Z, Haimel M, Simpson JT, Fonseca NA, Birol İ, Docking TR, Ho IY, et al. (2011). Assemblathon 1: A competitive assessment of de novo short read assembly methods. Genome Research **21**, 12, 2224–2241. https://doi.org/10.1101/gr.126599.111.

Ebler J, Ebert P, Clarke WE, Rausch T, Audano PA, Houwaart T, Mao Y, Korbel JO, Eichler EE, Zody MC, Dilthey AT, Marschall T (2022). Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. Nature Genetics **54**, 518–525. https://doi.org/10.1038/s41588-022-01043-w.

Edwards D (2026). On the use and misuse of pangenome and related terms. Nature Communications **17**. https://doi.org/10.1038/s41467-026-68624-9.

Eggertsson HP, Kristmundsdottir S, Beyter D, Jonsson H, Skuladottir A, Hardarson MT, Gudbjartsson DF, Stefansson K, Halldorsson BV, Melsted P (2019). GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. Nature Communications **10**. https://doi.org/10.1038/s41467-019-13341-9.

Eizenga JM, Novak AM, Sibbesen JA, Heumos S, Ghaffaari A, Hickey G, Chang X, Seaman JD, Rounthwaite R, Ebler J, Rautiainen M, Garg S, Paten B, Marschall T, Sirén J, Garrison E (2020). Pangenome Graphs. Annual Review of Genomics and Human Genetics **21**, 139–162. https://doi.org/10.1146/annurev-genom-120219-080406.

Emms DM, Kelly S (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biology **20**. https://doi.org/10.1186/s13059-019-1832-y.

Eren AM, Kiefl E, Shaiber A, Veseli I, Miller SE, Schechter MS, Fink I, Pan JN, Yousef M, Fogarty EC, Trigodet F, Watson AR, Esen ÖC, Moore RM, Clayssen Q, Lee MD, Kivenson V, Graham ED, Merrill BD, Karkman A, et al. (2020). Community-led, integrated, reproducible multi-omics with anvi'o. Nature Microbiology **6**, 3–6. https://doi.org/10.1038/s41564-020-00834-3.

Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, Garcia MU, Di Tommaso P, Nahnsen S (2020). The nf-core framework for community-curated bioinformatics pipelines. Nature Biotechnology **38**, 3, 276–278. https://doi.org/10.1038/s41587-020-0439-x.

Fiddes IT, Armstrong J, Diekhans M, Nachtweide S, Kronenberg ZN, Underwood JG, Gordon D, Earl D, Keane T, Eichler EE, Haussler D, Stanke M, Paten B (2018). Comparative Annotation Toolkit (CAT)—simultaneous clade and personal genome annotation. Genome Research **28**, 7, 1029–1038. https://doi.org/10.1101/gr.233460.117.

Gabory E, Mwaniki MN, Pisanti C, Pissis SP, Radoszewski J, Sweering M, Zuba TK (2024). Pangenome comparison via ED strings. Frontiers in Bioinformatics **4**, 1397036. https://doi.org/10.3389/fbinf.2024.1397036.

Garrison E, Guarracino A, Heumos S, Villani F, Bao Z, Tattini L, Hagmann J, Vorbrugg S, Marco-Sola S, Kubica C, Ashbrook DG, Thorell K, Rusholme-Pilcher RL, Liti G, Rudbeck E, Golicz AA, Nahnsen S, Yang Z, Mwaniki MN, Nobrega FL, et al. (2024). Building pangenome graphs. Nature Methods **21**, 11, 2008–2012. https://doi.org/10.1038/s41592-024-02430-3.

Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Dawson ET, Jones W, Garg S, Markello C, Lin MF, Paten B, Durbin R (2018). Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nature Biotechnology **36**, 9, 875–879. 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 EPC, Vallenet D (2020). PPanGGOLiN: Depicting microbial diversity via a partitioned pangenome graph. PLOS Computational Biology **16**, 3, e1007732. https://doi.org/10.1371/journal.pcbi.1007732.

Goel M, Sun H, Jiao WB, Schneeberger K (2019). SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. Genome Biology **20**. https://doi.org/10.1186/s13059-019-1911-0.

Golicz AA, Bayer PE, Bhalla PL, Batley J, Edwards D (2020). Pangenomics Comes of Age: From Bacteria to Plant and Animal Applications. Trends in Genetics **36**, 2, 132–145. https://doi.org/10.1016/j.tig.2019.11.006.

Gong Y, Li Y, Liu X, Ma Y, Jiang L (2023). A review of the pangenome: how it affects our understanding of genomic variation, selection and breeding in domestic animals? Journal of Animal Science and Biotechnology **14**. https://doi.org/10.1186/s40104-023-00860-1.

Griffiths RC, Marjoram P (1996). Ancestral inference from samples of DNA sequences with recombination. Journal of computational biology **3**, 479–502.

Groza C, Chen X, Wheeler TJ, Bourque G, Goubert C (2024). A unified framework to analyze transposable element insertion polymorphisms using graph genomes. Nature Communications **15**. https://doi.org/10.1038/s41467- 024- 53294-2.

Grytten I, Rand KD, Nederbragt AJ, Storvik GO, Glad IK, Sandve GK (2019). Graph Peak Caller: Calling ChIP-seq peaks on graph-based reference genomes. PLOS Computational Biology **15**, e1006731. https://doi.org/10.1371/journal.pcbi.1006731.

Guarracino A, Heumos S, Nahnsen S, Prins P, Garrison E (2022). ODGI: understanding pangenome graphs. Bioinformatics **38**, 13, 3319–3326. https://doi.org/10.1093/bioinformatics/btac308.

Guerra A (2026). The pangenome: a statistical model, not a fixed biological property. Bioinformatics Advances **1**, vbag069. https://doi.org/10.1093/bioadv/vbag069.

Hamilton JP, Li C, Buell CR (2024). The rice genome annotation project: an updated database for mining the rice genome. Nucleic Acids Research **53**, D1614–D1622. https://doi.org/10.1093/nar/gkae1061.

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, Barnes I, Bignell A, Boychenko V, Hunt T, Kay M, Mukherjee G, Rajan J, Despacio-Reyes G, Saunders G, Steward C, et al. (2012). GENCODE: The reference human genome annotation for The ENCODE Project. Genome Research **22**, 9, 1760–1774. https://doi.org/10.1101/gr.135350.111.

Haug-Baltzell A, Stephens SA, Davey S, Scheidegger CE, Lyons E (2017). SynMap2 and SynMap3D: web-based whole-genome synteny browsers. Bioinformatics **33**, 14, 2197–2198. https://doi.org/10.1093/bioinformatics/btx144.

He W, Xiang K, Chen C, Wang J, Wu Z (2022). Master graph: an essential integrated assembly model for the plant mitogenome based on a graph-based framework. Briefings in Bioinformatics **24**, bbac522. https://doi.org/10.1093/bib/bbac522.

Hejase HA, Dukler N, Siepel A (2020). From Summary Statistics to Gene Trees: Methods for Inferring Positive Selection. Trends in Genetics **36**, 4, 243–258. https://doi.org/10.1016/j.tig.2019.12.008.

Heuermann MC, Barros P, Beier S, Gundlach H, Alvarez-Jarreta J, Hassani-Pak K, König P, Fiebig A, Godec T, Gruden K, Nolte N, Petek M, Scholz U, Zagorščak M, Vandepoele K, Van Bel M (2025). White paper: standards for handling and analyzing plant pan-genomes. F1000Research **14**, 739. https://doi.org/10.12688/f1000research.166538.1.

Hickey G, Heller D, Monlong J, Sibbesen JA, Sirén J, Eizenga J, Dawson ET, Garrison E, Novak AM, Paten B (2020). Genotyping structural variants in pangenome graphs using the vg toolkit. Genome Biology **21**, 35. https://doi.org/10.1186/s13059-020-1941-7.

Hickey G, Monlong J, Ebler J, Novak AM, Eizenga JM, Gao Y, Abel HJ, Antonacci-Fulton LL, Asri M, Baid G, Baker CA, Belyaeva A, Billis K, Bourque G, Buonaiuto S, Carroll A, Chaisson MJP, Chang PC, Chang XH, Cheng H, et al. (2023). Pangenome graph construction from genome alignments with Minigraph-Cactus. Nature Biotechnology **42**, 4, 663–673. https://doi.org/10.1038/s41587-023-01793-w.

Holley G, Melsted P (2020). Bifrost: highly parallel construction and indexing of colored and compacted de Bruijn graphs. Genome Biology **21**, 249. https://doi.org/10.1186/s13059-020-02135-8.

Hu H, Li R, Zhao J, Batley J, Edwards D (2024). Technological Development and Advances for Constructing and Analyzing Plant Pangenomes. Genome Biology and Evolution **16,** 4, evae081. https://doi.org/10.1093/gbe/evae081.

Huang Y, Zhang Y, Zhang Q, Zhuang G, Li C, Wang B, Gao R, Xu Y, Qi Y, Hua X, Shi H, Xu Q, Yao W, Liu X, Qi Y, Chen B, Zhang M, Ming R, Tang H, Zhang J (2026). Multiscale pangenome graphs empower the genomic dissection of mixed-ploidy sugarcane species. Science **391,** 6785. https://doi.org/10.1126/science.adx1616.

Huson DH (1998). SplitsTree: analyzing and visualizing evolutionary data. Bioinformatics **14**, 68–73. https://doi.org/10.1093/bioinformatics/14.1.68.

Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT, Malla S, Marriott H, Nieto T, O'Grady J, Olsen HE, Pedersen BS, Rhie A, Richardson H, Quinlan AR, Snutch TP, et al. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. Nature Biotechnology **36**, 4, 338–345. https://doi.org/10.1038/nbt.4060.

Jayakodi M, Lu Q, Pidon H, Rabanus-Wallace MT, Bayer M, Lux T, Guo Y, Jaegle B, Badea A, Bekele W, Brar GS, Braune K, Bunk B, Chalmers KJ, Chapman B, Jørgensen ME, Feng JW, Feser M, Fiebig A, Gundlach H, et al. (2024). Structural variation in the pangenome of wild and domesticated barley. Nature **636**, 8043, 654–662. https://doi.org/10.1038/s41586-024-08187-1.

Jiao C, Xie X, Hao C, Chen L, Xie Y, Garg V, Zhao L, Wang Z, Zhang Y, Li T, Fu J, Chitikineni A, Hou J, Liu H, Dwivedi G, Liu X, Jia J, Mao L, Wang X, Appels R, et al. (2024). Pan-genome bridges wheat structural variations with habitat and breeding. Nature **637**, 8045, 384–393. https://doi.org/10.1038/s41586-024-08277-0.

Jonkheer EM, van Workum DJM, Sheikhizadeh Anari S, Brankovics B, de Haan JR, Berke L, van der Lee TAJ, de Ridder D, Smit S (2022). PanTools v3: functional annotation, classification and phylogenomics. Bioinformatics **38**, 18, 4403–4405. https://doi.org/10.1093/bioinformatics/btac506.

Kopalli V, Arslan K, Morales-Díaz N, Zanini SF, Golicz AA (2025). Toward a standardized framework for pangenome graph evaluation: assessing crop plant pangenome variation graph construction from multiple assemblies. GigaScience **14**, giaf121,https://doi.org/10.1093/gigascience/giaf121.

Langmead B, Salzberg SL (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods **9**, 4, 357–359. https://doi.org/10.1038/nmeth.1923.

Leonard AS, Crysnanto D, Fang ZH, Heaton MP, Vander Ley BL, Herrera C, Bollwein H, Bickhart DM, Kuhn KL, Smith TPL, Rosen BD, Pausch H (2022). Structural variant-based pangenome

construction has low sensitivity to variability of haplotype-resolved bovine assemblies. Nature Communication, **13,** 3012. https://doi.org/10.1038/s41467-022-30680-2.

Li H (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. bioRxiv [Preprint]. https://doi.org/10.48550/ARXIV.1303.3997.

Li H, Durbin R (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics **25**, 14, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324.

Li H, Feng X, Chu C (2020). The design and construction of reference pangenome graphs with minigraph. Genome Biology **21,** 265. https://doi.org/10.1186/s13059-020-02168-z.

Li Q, Qiao X, Li L, Gu C, Yin H, Qi K, Xie Z, Yang S, Zhao Q, Wang Z, Yang Y, Pan J, Li H, Wang J, Wang C, Rieseberg LH, Zhang S, Tao S (2024). Haplotype-resolved T2T genome assemblies and pangenome graph of pear reveal diverse patterns of allele-specific expression and the genomic basis of fruit quality traits. Plant Communications **5**, 101000. https://doi.org/10.1016/j.xplc.2024.101000.

Liao WW, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, Lu S, Lucas JK, Monlong J, Abel HJ, Buonaiuto S, Chang XH, Cheng H, Chu J, Colonna V, Eizenga JM, Feng X, Fischer C, Fulton RS, Garg S, et al. (2023). A draft human pangenome reference. Nature **617**, 7960, 312–324. https://doi.org/10.1038/s41586-023-05896-x.

Limasset A, Cazaux B, Rivals E, Peterlongo P (2016). Read mapping on de Bruijn graphs. BMC Bioinformatics **17**, 237. https://doi.org/10.1186/s12859-016-1103-9.

Lipka AE, Tian F, Wang Q, Peiffer J, Li M, Bradbury PJ, Gore MA, Buckler ES, Zhang Z (2012). GAPIT: genome association and prediction integrated tool. Bioinformatics **28**, 18, 2397–2399. https://doi.org/10.1093/bioinformatics/bts444.

Lisch D (2012). How important are transposons for plant evolution? Nature Reviews Genetics **14**, 49–61. https://doi.org/10.1038/nrg3374.

Liu J, Ni Y, Liu C (2023). Polymeric structure of the Cannabis sativa L. mitochondrial genome identified with an assembly graph model. Gene **853**, 147081. https://doi.org/10.1016/j.gene.2022.147081.

Liu Z, Wang N, Su Y, Long Q, Peng Y, Shangguan L, Zhang F, Cao S, Wang X, Ge M, Xue H, Ma Z, Liu W, Xu X, Li C, Cao X, Ahmad B, Su X, Liu Y, Huang G, et al. (2024). Grapevine pangenome facilitates trait genetics and genomic breeding. Nature Genetics **56**, 12, 2804–2814. https://doi.org/10.1038/s41588-024-01967-5.

Llamas B, Narzisi G, Schneider V, Audano PA, Biederstedt E, Blauvelt L, Bradbury P, Chang X, Chin CS, Fungtammasan A, Clarke WE, Cleary A, Ebler J, Eizenga J, Sibbesen JA, Markello CJ, Garrison E, Garg S, Hickey G, Lazo GR, et al. (2021). A strategy for building and using a human reference pangenome. F1000Research **8**, 1751. https://doi.org/10.12688/f1000research.19630.2.

MacNish TR, Al-Mamun HA, Bayer PE, McPhan C, Fernandez CGT, Upadhyaya SR, Liu S, Batley J, Parkin IAP, Sharpe AG, Edwards D (2024). Brassica Panache: A multi-species graph pangenome representing presence absence variation across forty-one Brassica genomes. The Plant Genome **18,** e20535. https://doi.org/10.1002/tpg2.20535.

Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM (2021). BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Molecular Biology and Evolution **38**, 10, 4647–4654. https://doi.org/10.1093/molbev/msab199.

Mapleson D, Garcia Accinelli G, Kettleborough G, Wright J, Clavijo BJ (2016). KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. Bioinformatics **33**, 4, 574–576. https://doi.org/10.1093/bioinformatics/btw663.

Marthe N, Zytnicki M, Sabot F (2025). GrAnnoT, a tool for efficient and reliable annotation transfer through pangenome graph. Peer Community Journal **5,** e133. https://doi.org/10.24072/pcjournal.651.

Matthews CA, Watson-Haigh NS, Burton RA, Sheppard AE (2024). A gentle introduction to pangenomics. Briefings in Bioinformatics **25,** 6, bbae588. https://doi.org/10.1093/bib/bbae588.

Mbatchou J, Barnard L, Backman J, Marcketta A, Kosmicki JA, Ziyatdinov A, Benner C, O'Dushlaine C, Barber M, Boutkov B, Habegger L, Ferreira M, Baras A, Reid J, Abecasis G, Maxwell E, Marchini J (2021). Computationally efficient whole-genome regression for quantitative and binary traits. Nature Genetics **53**, 7, 1097–1103. https://doi.org/10.1038/s41588-021-00870-7.

Miao Z, Yue JX (2025). Interactive visualization and interpretation of pangenome graphs by linear reference–based coordinate projection and annotation integration. Genome Research **35**, 2, 296–310. https://doi.org/10.1101/gr.279461.124.

Milia S, Leonard AS, Mapel XM, Bernal Ulloa SM, Drögemüller C, Pausch H (2024). Taurine pangenome uncovers a segmental duplication upstream of KIT associated with depigmentation in white-headed cattle. Genome Research **35**, 4, 1041–1052. https://doi.org/10.1101/gr.279064.124.

Mohamed M, Sabot F, Varoqui M, Mugat B, Audouin K, Pélisson A, Fiston-Lavier AS, Chambeyron S (2023). TrEMOLO: accurate transposable element allele frequency estimation using long-read sequencing data combining assembly and mapping-based approaches. Genome Biology **24**, 63. https://doi.org/10.1186/s13059-023-02911-2.

Ndiaye M, Prieto-Baños S, Fitzgerald LM, Yazdizadeh Kharrazi A, Oreshkov S, Dessimoz C, Sedlazeck FJ, Glover N, Majidian S (2024). When less is more: sketching with minimizers in genomics. Genome Biology **25**, 270. https://doi.org/10.1186/s13059-024-03414-4.

Noah JM, Gorse M, Romain CA, Gay EJ, Rouxel T, Balesdent MH, Soyer JL (2024). To be or not to be a nonhost species: A case study of the *Leptosphaeria maculans* and *Brassica carinata* interaction. Environmental Microbiology Reports **16**, 6, e70034. https://doi.org/10.1111/1758-2229.70034.

Novak AM, Chung D, Hickey G, Djebali S, Yokoyama TT, Garrison E, Narzisi G, Paten B, Monlong J (2024). Efficient indexing and querying of annotations in a pangenome graph. bioRxiv [Preprint]. https://doi.org/10.1101/2024.10.12.618009.

Novak AM, Garrison E, Paten B (2017). A graph extension of the positional Burrows–Wheeler transform and its applications. Algorithms for Molecular Biology **12**, 18. https://doi.org/10.1186/s13015-017-0109-9.

Omrane S, Audéon C, Ignace A, Duplaix C, Aouini L, Kema G, Walker AS, Fillinger S (2018). Correction for Omrane et al., "Plasticity of the *MFS1* Promoter Leads to Multidrug Resistance in the Wheat Pathogen *Zymoseptoria tritici* ". mSphere **3**, 10. https://doi.org/10.1128/msphere.00322-18.

Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM (2016). Mash: fast genome and metagenome distance estimation using MinHash. Genome Biology **17**, 132. https://doi.org/10.1186/s13059-016-0997-x.

Ou S, Chen J, Jiang N (2018). Assessing genome assembly quality using the LTR Assembly Index (LAI). Nucleic Acids Research **46**, 21, e126–e126. https://doi.org/10.1093/nar/gky730.

Pani S, Dabbaghie F, Marschall T, Soylev A (2024). gaftools: a toolkit for analyzing and manipulating pangenome alignments. bioRxiv [Preprint]. https://doi.org/10.1101/2024.12.10.627813.

Parmigiani L, Garrison E, Stoye J, Marschall T, Doerr D (2024). Panacus: fast and exact pangenome growth and core size estimation. Bioinformatics **40**, 12, btae720. https://doi.org/10.1093/bioinformatics/btae720.

Paten B, Eizenga JM, Rosen YM, Novak AM, Garrison E, Hickey G (2018). Superbubbles, Ultra-bubbles, and Cacti. Journal of Computational Biology **25**, 7, 649–663. https://doi.org/10.1089/cmb.2017.0251.

Patrick KL (2008). 454 Life Sciences: Illuminating the future of genome sequencing and personalized medicine. The Yale Journal of Biology and Medicine **80**, 191.

Piat L, Denni S, Dubois S, Linard B, Duvaux L (2026). Simulating population pangenomes under coalescent demographic models with MSpangenome. bioRxiv [Preprint]. https://doi.org/10.64898/2026.06.29.735168.

Poplin R, Chang PC, Alexander D, Schwartz S, Colthurst T, Ku A, Newburger D, Dijamco J, Nguyen N, Afshar PT, Gross SS, Dorfman L, McLean CY, DePristo MA (2018). A universal SNP and small-indel variant caller using deep neural networks. Nature Biotechnology **36**, 10, 983–987. https://doi.org/10.1038/nbt.4235.

Quinlan AR, Hall IM (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics **26**, 6, 841–842. https://doi.org/10.1093/bioinformatics/btq033.

Ranallo-Benavidez TR, Lemmon Z, Soyk S, Aganezov S, Salerno WJ, McCoy RC, Lippman ZB, Schatz MC, Sedlazeck FJ (2021). Optimized sample selection for cost-efficient long-read population sequencing. Genome Research **31**, 5, 910–918. https://doi.org/10.1101/gr.264879.120.

Rautiainen M, Marschall T (2020). GraphAligner: rapid and versatile sequence-to-graph alignment. Genome Biology **21**, 1, 253. https://doi.org/10.1186/s13059-020-02157-2.

Reinartz J, Bruyns E, Lin JZ, Burcham T, Brenner S, Bowen B, Kramer M, Woychik R (2002). Massively parallel signature sequencing (MPSS) as a tool for in-depth quantitative gene expression profiling in all organisms. Briefings in functional genomics & proteomics **1**, 95–104. https://doi.org/10.1093/BFGP/1.1.95.

Reiser L, Bakker E, Subramaniam S, Chen X, Sawant S, Khosa K, Prithvi T, Berardini TZ (2024). The Arabidopsis Information Resource in 2024. Genetics **227**, 1, iyae027 https://doi.org/10.1093/genetics/iyae027.

Rhie A, Walenz BP, Koren S, Phillippy AM (2020). Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biology **21**, 245. https://doi.org/10.1186/s13059-020-02134-9.

Rohart F, Gautier B, Singh A, Lê Cao KA (2017). mixOmics: An R package for 'omics feature selection and multiple data integration. PLOS Computational Biology **13**, 11, e1005752. https://doi.org/10.1371/journal.pcbi.1005752.

Romain S, Dubois S, Legeai F, Lemaitre C (2025). Investigating the topological motifs of inversions in pangenome graphs. bioRxiv [Preprint]. https://doi.org/10.1101/2025.03.14.643331.

Romain S, Lemaitre C (2023). SVJedi-graph: improving the genotyping of close and overlapping structural variants with long reads using a variation graph. Bioinformatics **39**, i270–i278. https://doi.org/10.1093/bioinformatics/btad237.

Rubin JD, Vogel NA, Gopalakrishnan S, Sackett PW, Renaud G (2023). HaploCart: Human mtDNA haplogroup classification using a pangenomic reference graph. PLOS Computational Biology **19**, 6, 1–27. https://doi.org/10.1371/journal.pcbi.1011148.

Ruperao P, Rangan P, Shah T, Sharma V, Rathore A, Mayes S, Pandey MK (2025). Developing pangenomes for large and complex plant genomes and their representation formats. Journal of Advanced Research **78**, 29–46. https://doi.org/10.1016/j.jare.2025.01.052.

Saidi S, Blaison M, del Pilar Rodríguez-Ordóñez M, Confais J, Quesneville H (2026). A reference-free pipeline for detecting shared transposable elements from pan-genomes to retrace their dynamics in a species. Genome Biology **27**, 117. https://doi.org/10.1186/s13059-026-03984-5.

Salehi Nowbandegani P, Zhang S, Hu H, Li H, O'Connor LJ (2026). Defining and cataloging variants in pangenome graphs. Cell Genomics, 101327. https://doi.org/10.1016/j.xgen.2026.101327.

Schreiber M, Jayakodi M, Stein N, Mascher M (2024). Plant pangenomes for crop improvement, biodiversity and evolution. Nature Reviews Genetics **25**, 8, 563–577. https://doi.org/10.1038/s41576-024-00691-4.

Secomandi S, Gallo GR, Rossi R, Rodríguez Fernandes C, Jarvis ED, Bonisoli-Alquati A, Gianfranceschi L, Formenti G (2025). Pangenome graphs and their applications in biodiversity genomics. Nature Genetics **57**, 13–26. https://doi.org/10.1038/s41588-024-02029-6.

Shafer ABA, Kardos M (2025). Runs of Homozygosity and Inferences in Wild Populations. Molecular Ecology **34**, 3, e17641. https://doi.org/10.1111/mec.17641.

Shaiber A, Willis AD, Delmont TO, Roux S, Chen LX, Schmid AC, Yousef M, Watson AR, Lolans K, Esen ÖC, Lee STM, Downey N, Morrison HG, Dewhirst FE, Mark Welch JL, Eren AM (2020).

Functional and genetic markers of niche partitioning among enigmatic members of the human oral microbiome. Genome Biology **21**, 292. https://doi.org/10.1186/s13059-020-02195-w.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003). Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Research **13**, 11, 2498–2504. https://doi.org/10.1101/gr.1239303.

Sibbesen JA, Eizenga JM, Novak AM, Sirén J, Chang X, Garrison E, Paten B (2023). Haplotype-aware pantranscriptome analyses using spliced pangenome graphs. Nature Methods **20**, 2, 239–247. https://doi.org/10.1038/s41592-022-01731-9.

Sibbesen JA, Maretty L, Krogh A (2018). Accurate genotyping across variant classes and lengths using variant graphs. Nature Genetics **50**, 7, 1054–1059. https://doi.org/10.1038/s41588-018-0145-5.

Sierra P, Durbin R (2024). Identification of transposable element families from pangenome polymorphisms. Mobile DNA **15**, 13. https://doi.org/10.1186/s13100-024-00323-y.

Sirén J, Eskandar P, Ungaro MT, Hickey G, Eizenga JM, Novak AM, Chang X, Chang PC, Kolmogorov M, Carroll A, Monlong J, Paten B (2024). Personalized pangenome references. Nature Methods **21**, 11, 2017–2023. https://doi.org/10.1038/s41592-024-02407-2.

Sirén J, Monlong J, Chang X, Novak AM, Eizenga JM, Markello C, Sibbesen JA, Hickey G, Chang PC, Carroll A, Gupta N, Gabriel S, Blackwell TW, Ratan A, Taylor KD, Rich SS, Rotter JI, Haussler D, Garrison E, Paten B (2021). Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. Science **374**, 6574. https://doi.org/10.1126/SCIENCE.ABG8871.

Sirén J, Paten B (2022). GBZ file format for pangenome graphs. Bioinformatics **38**, 22, 5012–5018. https://doi.org/10.1093/bioinformatics/btac656

Smith TPL, Bickhart DM, Boichard D, Chamberlain AJ, Djikeng A, Jiang Y, Low WY, Pausch H, Demyda-Peyrás S, Prendergast J, Schnabel RD, Rosen BD (2023). The Bovine Pangenome Consortium: democratizing production and accessibility of genome assemblies for global cattle breeds and other bovine species. Genome Biology **24**. https://doi.org/10.1186/s13059-023-02975-0.

Soiland-Reyes S, Sefton P, Crosas M, Castro LJ, Coppens F, Fernández JM, Garijo D, Grüning B, La Rosa M, Leo S, Ó Carragáin E, Portier M, Trisovic A, Community RC, Groth P, Goble C (2022). Packaging research artefacts with RO-Crate. Data Science **5**, 2, 97–138. https://doi.org/10.3233/ds-210053.

Söylev A, Ebler J, Pani S, Rausch T, Korbel JO, Marschall T (2024). SVarp: pangenome-based structural variant discovery. bioRxiv [Preprint]. https://doi.org/10.1101/2024.02.18.580171.

Sun H, Tusso S, Dent CI, Goel M, Wijfjes RY, Baus LC, Dong X, Campoy JA, Kurdadze A, Walkemeier B, Sänger C, Huettel B, Hutten RCB, van Eck HJ, Dehmer KJ, Schneeberger K (2025). The phased pan-genome of tetraploid European potato. Nature **642**, 389–397. https://doi.org/10.1038/s41586-025-08843-0.

Taylor DJ, Eizenga JM, Li Q, Das A, Jenike KM, Kenny EE, Miga KH, Monlong J, McCoy RC, Paten B, Schatz MC (2024). Beyond the Human Genome Project: The Age of Complete Human Genome Sequences and Pangenome References. Annual Review of Genomics and Human Genetics **25**, 77–104. https://doi.org/10.1146/annurev-genom-021623-081639.

Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, DeBoy RT, Davidsen TM, Mora M, Scarselli M, Margarit y Ros I, Peterson JD, Hauser CR, Sundaram JP, Nelson WC, Madupu R, et al. (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. https://doi.org/10.1073/pnas.0506758102.

The 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. Nature **467**, 7319, 1061–1073. https://doi.org/10.1038/nature09534.

The 3000 rice genomes project (2014). The 3, 000 rice genomes project. Gigascience **3**, 1, 2047–217X–3–7. https://doi.org/10.1186/2047-217x-3-7.

The Computational Pan-Genomics Consortium (2018). Computational pan-genomics: status, promises and challenges. Briefings in Bioinformatics **19**, 118–135. https://doi.org/10.1093/bib/bbw089.

Uffelmann E, Huang QQ, Munung NS, de Vries J, Okada Y, Martin AR, Martin HC, Lappalainen T, Posthuma D (2021). Genome-wide association studies. Nature Reviews Methods Primers **1**, 59. https://doi.org/10.1038/s43586-021-00056-9.

Vicedomini R, Andreace F, Dufresne Y, Chikhi R, Duitama González C (2025). MUSET: set of utilities for constructing abundance unitig matrices from sequencing data. Bioinformatics **41**, 3, btaf054. https://doi.org/10.1093/bioinformatics/btaf054.

Vogel NA, Rubin JD, Pedersen AG, Sackett PW, Pedersen MW, Renaud G (2024). soibean: High-Resolution Taxonomic Identification of Ancient Environmental DNA Using Mitochondrial Pangenome Graphs. Molecular Biology and Evolution **41**, msae203. https://doi.org/10.1093/molbev/msae203.

Vogel NA, Rubin JD, Swartz M, Vlieghe J, Sackett PW, Pedersen AG, Pedersen MW, Renaud G (2023). euka: Robust tetrapodic and arthropodic taxa detection from modern and ancient environmental DNA using pangenomic reference graphs. Methods in Ecology and Evolution **14**, 11, 2717–2727. https://doi.org/https://doi.org/10.1111/2041-210X.14214.

Voichek Y, Weigel D (2020). Identifying genetic variants underlying phenotypic variation in plants without complete genomes. Nature Genetics **52**, 5, 534–540. https://doi.org/10.1038/s41588-020-0612-7.

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

Wang J, Hu H, Jiang X, Zhang S, Yang W, Dong J, Yang T, Ma Y, Zhou L, Chen J, Nie S, Liu C, Ning Y, Zhu X, Liu B, Yang J, Zhao J (2024). Pangenome-Wide Association Study and Transcriptome Analysis Reveal a Novel QTL and Candidate Genes Controlling both Panicle and Leaf Blast Resistance in Rice. Rice **17**, 27. https://doi.org/10.1186/s12284-024-00707-x.

Wang J, Yang W, Zhang S, Hu H, Yuan Y, Dong J, Chen L, Ma Y, Yang T, Zhou L, Chen J, Liu B, Li C, Edwards D, Zhao J (2023). A pangenome analysis pipeline provides insights into functional gene identification in rice. Genome Biology **24**, 19. https://doi.org/10.1186/s13059-023-02861-9.

Wang S, Qian YQ, Zhao RP, Chen LL, Song JM (2022). Graph-based pan-genomes: increased opportunities in plant genomics. Journal of Experimental Botany **74**, 24–39. https://doi.org/10.1093/jxb/erac412.

Watanabe K, Stringer S, Frei O, Umićević Mirkov M, de Leeuw C, Polderman TJC, van der Sluis S, Andreassen OA, Neale BM, Posthuma D (2019). A global overview of pleiotropy and genetic architecture in complex traits. Nature Genetics **51**, 9, 1339–1348. https://doi.org/10.1038/s41588-019-0481-0.

Wenger AM, Peluso P, Rowell WJ, Chang PC, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND, Töpfer A, Alonge M, Mahmoud M, Qian Y, Chin CS, Phillippy AM, Schatz MC, Myers G, DePristo MA, Ruan J, et al. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nature Biotechnology **37**, 10, 1155–1162. https://doi.org/10.1038/s41587-019-0217-9.

van Westerhoven AC, Fokkens L, Wissink K, Kema G, Rep M, Author C, Seidl MF (2024). Reference-free identification and pangenome analysis of accessory chromosomes in a major fungal plant pathogen. bioRxiv [Preprint]. https://doi.org/10.1101/2024.12.12.627383.

White B, Lux T, Rusholme-Pilcher R, Juhász A, Kaithakottil G, Duncan S, Simmonds J, Rees H, Wright J, Colmer J, Ward S, Joynson R, Coombes B, Irish N, Henderson S, Barker T, Chapman H, Catchpole L, Gharbi K, Bose U, et al. (2025). De novo annotation reveals transcriptomic complexity across the hexaploid wheat pan-genome. Nature Communications **16**, 8538. https://doi.org/10.1038/s41467-025-64046-1.

Wick RR, Schultz MB, Zobel J, Holt KE (2015). Bandage: interactive visualization of de novo genome assemblies. Bioinformatics **31**, 20, 3350–3352. https://doi.org/10.1093/bioinformatics/btv383.

Widmer C, Lippert C, Weissbrod O, Fusi N, Kadie C, Davidson R, Listgarten J, Heckerman D (2014). Further Improvements to Linear Mixed Models for Genome-Wide Association Studies. Scientific Reports **4**, 6874. https://doi.org/10.1038/srep06874.

Wijaya AJ, Anžel A, Richard H, Hattab G (2025). Current state and future prospects of Horizontal Gene Transfer detection. NAR Genomics and Bioinformatics **7**, 1, lqaf005. https://doi.org/10.1093/nargab/lqaf005.

Yang L, He W, Zhu Y, Lv Y, Li Y, Zhang Q, Liu Y, Zhang Z, Wang T, Wei H, Cao X, Cui Y, Zhang B, Chen W, He H, Wang X, Chen D, Liu C, Shi C, Liu X, et al. (2025). GWAS meta-analysis using a graph-based pan-genome enhanced gene mining efficiency for agronomic traits in rice. Nature Communications **16**, 3171. https://doi.org/10.1038/s41467-025-58081-1.

Yang Z, Yang Z, Gao C, Zhang M, Hu G, Yang L, Zhang Y, Ma M, Liu R, Wang Z, Gao B, Zhang Z, Zhao H, Liu X, Ma X, Wendel JF, Ge X, Li F (2026). Graph pan-genome illuminates evolutionary trajectories and agronomic trait architecture in allotetraploid cotton. Nature Genetics **58**, 218–229. https://doi.org/10.1038/s41588-025-02462-1.

Zanini SF, Bayer PE, Wells R, Snowdon RJ, Batley J, Varshney RK, Nguyen HT, Edwards D, Golicz AA (2021). Pangenomics in crop improvement—from coding structural variations to finding regulatory variants with pangenome graphs. The Plant Genome **15**, e20177 https://doi.org/10.1002/tpg2.20177.

Zhang Y, Wang Y, Wu T, Lin Y, Tan Y, Qi Y, Wang Y, Wang B, Wang Z, Zhang Q, Zhang J, Huang Y, Tang H (2026). NodeGWAS: Leveraging Graph Pangenomes for Sensitive and Accurate Association Analysis in Diverse Diploid and Polyploid Species. Plant Communications, **7**, 101835. https://doi.org/10.1016/j.xplc.2026.101835.

Zhao X, Yu J, Zhang J, Sun H, Wu S, Zhao J, Zhou Y, Hammar SA, Lin YC, Zhang Z, Huang S, Dymerski RT, Chen F, Weng Y, Grumet R, Xu Y, Fei Z (2026). Graph-based pangenome reveals structural variation dynamics during cucumber breeding. Nature Genetics **58**, 3, 643–654. https://doi.org/10.1038/s41588-026-02506-0.

Zhou X, Stephens M (2014). Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nature Methods **11**, 4, 407–409. https://doi.org/10.1038/nmeth.2848.

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, Zhang Z, et al. (2022). Graph pangenome captures missing heritability and empowers tomato breeding. Nature **606**, 7914, 527–534. https://doi.org/10.1038/s41586-022-04808-9.

Zytnicki M (2024). Assessing genome conservation on pangenome graphs with PanSel. Bioinformatics Advances **5**, 1, vbaf018. https://doi.org/10.1093/bioadv/vbaf018.