Pangenome graphs and their applications in biodiversity genomics

Simona Secomandi[1,*], Guido Roberto Gallo[2,*], Riccardo Rossi[3], Carlos Rodríguez Fernandes[4,5], Erich D. Jarvis[1,6], Andrea Bonisoli-Alquati[7], Luca Gianfranceschi[2], Giulio Formenti[6,†]

* Co-first authors

† Corresponding author (e-mail address: gformenti@rockefeller.edu)

[1]Laboratory of Neurogenetics of Language, The Rockefeller University, New York, NY, USA

[2]Department of Biosciences, The University of Milan, Via Giovanni Celoria 26, Milan MI, 20133 Italy

[3]Department of Biotechnology and Biosciences, University of Milano - Bicocca, Piazza della Scienza 2, 20126 Milan, Italy

[4]CE3C - Centre for Ecology, Evolution and Environmental Changes & CHANGE - Global Change and Sustainability Institute, Departamento de Biologia Animal, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal

[5]Faculdade de Psicologia, Universidade de Lisboa, Alameda da Universidade, 1649-013 Lisboa, Portugal

[6]The Vertebrate Genome Laboratory, 1230 York Ave, New York, NY 10065 USA

[7]Department of Biological Sciences, California State Polytechnic University, Pomona. Pomona, CA

Abstract

Complete datasets of genetic variants are key to biodiversity genomics studies. Long-read sequencing technologies allow the routine assembly of highly-contiguous, haplotype-resolved reference genomes. However, even when complete, reference genomes from a single individual may bias downstream analyses and fail to adequately represent genetic diversity within a population or species. Pangenome graphs assembled from aligned collections of high-quality genomes can overcome representation bias by integrating sequence information from multiple genomes from the same population, species or genus into a single reference. Here, we review the available tools and data structures to build, visualize, and manipulate pangenome graphs, while providing practical examples and discussing their applications in biodiversity and conservation genomics across the tree of life.

Main

The decrease in DNA sequencing costs is boosting biodiversity genomics[12]. Large-scale international initiatives aim to generate highly-contiguous, haplotype-resolved reference genomes for all species[3–7], to characterize biodiversity, clarify its evolution, and help its conservation. Reference genomes are the backbones to annotate genomic variation, helping its preservation in the current biodiversity crisis[2]. Haplotype-resolved reference genomes better capture genetic variation across individuals, enabling more effective mapping of phenomes to genomes, and illuminating both the adaptive uniqueness of taxa[8] and their position in the tree of life[9]. However, even with accurate reference genomes, sequences mapped against them can be misplaced or fail to align because of divergent or missing regions in the reference (**Fig. 1a**). These 'blind spots'[10–12] introduce a reference bias[13,14], potentially misrepresenting genetic variation. A solution to reference bias lies in pangenomics, the systematic capturing of genetic variation within dedicated composite assemblies, called pangenomes[15,16]. Historically, a pangenome can be a collection of unaligned sequences[17,18], or a graph derived from raw reads[19] or assembled sequences[15] (**Fig. 1b**; **Box 1**). A pangenome graph is a graphical model storing genomic data from different individuals, together with their relationships and variability, in a single data structure (**Fig. 1c**). Data embedded in the graph can be accessed to perform bioinformatics tasks, such as read mapping, by graph indexing[15,20]. A pangenome graph accommodates multiple alternative alleles and provides a more comprehensive representation of genomic variation within a species[21,22], depending on the degree to which the haplotypes set reflects the overall diversity of the species. At higher taxonomic ranks (e.g. genus or family), super-pangenomes expand our ability to capture genetic diversity[23–27]. While cataloging genetic variability within a species is critical[28], examples of pangenome graphs in non-model species are still limited[29]. We expect this to change in the coming years, as long-read sequencing and high-quality genomes become more accessible[2,30]. Pangenome graphs improve downstream analyses, providing researchers with the flexibility to analyze genetic variation at multiple levels in a single data structure. In agriculture, pangenomes for major crop plants proved effective in identifying resistance genes and beneficial alleles for crop improvement, while in humans they show potential to improve clinical research (**Box 1**). Pangenome graphs enable the accurate detection of structural variants (SVs)[31,32], large (> 50 bp) genomic rearrangements that underlie phenotype and fitness variation[31,32] and disproportionately contribute to local adaptation[33], and even speciation[34]. Pangenomes graphs will increase the accuracy and power of resequencing projects investigating population dynamics and insights

into the genetic bases of phenotypic traits, by streamlining variant phasing, haplotype reconstruction, and genotyping through imputation. They will aid in understanding genome evolution across species and benefit studies linking genetic variation and gene expression to phenotypes. Here, we illustrate the available data structures and tools for building, visualizing, and manipulating pangenome graphs, and provide practical advice for their use in downstream analyses. We also highlight their potential future contribution to conservation and biodiversity genomics.

Pangenomes as variation graphs

Over the years, the term pangenome acquired different meanings (**Box 1**). Here, we focus on graph-based pangenomes constructed from whole-genome alignments of assembled sequences, i.e. variation graphs[15]. Variation graphs more comprehensively represent eukaryotic genomes by storing complete genomic sequences and their variation[21,35–37]. Variation graphs compress redundant sequences into bidirected networks where each node represents a sequence and edges connect nodes into complete sequences[15] (**Fig. 1c**). Linear genomes or phased haplotypes are stored as explicit paths through the graph[15], and sequence variation is represented by subgraphs called bubbles, or snarls, in which variants are defined by alternative paths connected by shared start and end nodes[38]. Here, we refer to variation graphs from whole-genome alignments simply as "pangenome graphs". They were adopted by the Human Pangenome Reference Consortium (HPRC)[39], a global initiative that generated the first human pangenome[37,40] (**Box 1**). Pangenome graphs were also assembled for the chicken (*Gallus gallus*)[21], and, among non-model organisms, for the barn swallow (*Hirundo rustica;* **Box 2**)[29]. Super-pangenome graphs are being generated for economically important species, such as tomato (*Solanum lycopersicum*)[25], grape (*Vitis sp.;* **Box 2**)[24] and cattle (*Bos taurus*)[26].

Maximizing capture of genetic diversity via sampling and sequencing

The sampling strategy is critical for successful biodiversity pangenomic studies. It should maximize genomic and biogeographic diversity within natural populations, ideally sampling through the entire geographic range, while balancing sex representation[18,24,25,37,39,41–44] (**Fig. 2a**). If a panel of variants, both SNPs and SVs, is available, estimates of heterozygosity, relatedness, and inbreeding offer insights into the sample size required to achieve a comprehensive representation. Ordination and clustering analyses can help select representative individuals for inclusion in a pangenome[24]. The ideal sample size can also be

retrospectively verified by a pangene number analysis, where a number of representative genomes are progressively added until a pangene plateau is reached, i.e., when the full set of genes is captured and adding more individuals does not recover novel genes[45]. The inclusion of more individuals augment the existing references by increasing the representation of accessory, or dispensable, sequences, which are shared only by a subset of individuals and often have functional and adaptive roles[46]. Pangenome graphs facilitate the pinpointing of functional and adaptive roles of accessory genomic regions and how they vary among geographic populations and subspecies. Accessory genomes may provide hotspots for population differentiation and speciation through divergent selection processes or via hybridization[47,48]. Investigating how much accessory regions are affected by introgression in populations and species undergoing hybridization provides insights into the dynamics of gene flow and speciation processes[49].

Among the collected samples, priority should be given to highest-quality ones, maximizing long-read sequencing throughput (**Fig. 2a**) and allowing accurate and contiguous genome assemblies[5,50]. The quality of the input genomes minimizes noise propagation in the pangenome graph. Haplotype phasing with parental sequence data or chromatin conformation data (Hi-C), is crucial to prevent haplotype false duplication and related errors[4,5,51,52]. Sequencing coverages of ~30-fold PacBio high-fidelity long reads (HiFi)[11,53] and ~60-fold Oxford Nanopore Technologies (ONT)[54] Duplex reads, in combination with ~30-fold Hi-C per haplotype and manual curation, generate reference genomes that meet the current quality standards[5,7,55]. Genome sequencing is a rapidly evolving field, and generating complete, haplotype-resolved, and near-error-free genomes (telomere-to-telomere, T2T) is now feasible[37,50,56], by complementing HiFi with ultra-long ONT reads. Incorporating high-quality genomes aids in the discovery of rare SVs[57], particularly in admixed populations and those with large effective population sizes[58,59]. Moreover, it will improve the representation of hard-to-sequence and assemble regions like centromeres, variable number tandem repeats (VNTRs)[37] and other complex repeats. Examining base-level polymorphism in VNTRs may clarify their role in shaping gene expression and complex traits[10,26,60]. Highly repetitive regions might also underlie the regulation of complex behavioral phenotypes, such as migratory behavior[61].

Overall, a pangenome graph benefits from the inclusion of all T2T-level or high-quality genomes, whereas pangenomes derived from sub-T2T genomes will limit the study of genetic diversity due to the incompleteness of challenging regions. However, acquiring multiple high-quality samples from non-model species may be difficult, especially for rare and threatened

species, and sequencing multiple individuals with different long-read technologies may currently be cost-prohibitive. We suggest that a pangenome graph should include at least one high-quality assembly as a backbone for graph construction[55], providing a robust coordinate system for downstream analyses.

Building pangenome graphs to represent complex and accessory genomic regions

Pangenome graph construction starts with the alignment of the input genomes to identify sequence similarities. Alignment can be reference-based[20,62] or involve all-versus-all comparisons[63,64], and it can be either at the base-level[20,63] or at a higher level (e.g. including only variant sites)[62]. Alignment information is embedded in the graph, which can be manipulated for downstream analyses. Two main pipelines were developed by the HPRC for graph construction: Minigraph-Cactus (MC)[20] and the PanGenome Graph Builder (PGGB)[63] (**Fig. 2b**; **Supplementary Table 1**). MC implements Minigraph[62], a sequence-to-graph aligner, as a graph constructor. In MC, a user-selected reference genome is used as the initial backbone, which is progressively augmented with structural variation from the other genomes. The resulting graph is SV-only (> 50 bp, **Fig. 2b**) and all assemblies are aligned back to the graph with a minimap2-like[65] algorithm that generates base-level alignments for each reference chromosome. MC implements a modified version of the reference-free aligner Progressive Cactus[66] to combine the alignments into base-level pangenome graphs that contain variants of all sizes (**Fig. 2b**). Chromosomal graphs are then combined and post-processed to reduce path complexity by collapsing redundant sequences[20]. In addition to the chosen reference, one may specify additional assemblies whose coordinates can serve as reference in downstream analyses[20]. In contrast to MC, PGGB[63] avoids using an initial reference, and rather employs all-to-all genome alignments with wfmash[67], a software for homology mapping that generates base-wise pairwise alignments. PGGB uses seqwish[64] as a sequence-to-graph aligner, which starts from the all-versus-all alignments to generate a complete pangenome graph, representing all variant types and sizes (**Fig. 2b**). The graph is then post-processed with a smoothing and normalization step[37,63]. PGGB can be run in parallel on each chromosome community to reduce computation time[37,63]. In PGGB graphs, every genome included in the graph can be used as reference for downstream analyses[63]. However, like in MC, only one reference can be used at a time as a coordinate system for downstream analyses such as variant calling. New computational methods and file formats other than the linear binary alignment map (BAM) and variant call format (VCF) need to be developed to overcome this limitation and represent all

the information embedded in the graph. Pangenome graphs can be combined with transcript annotations using the variation graph toolkit (vg)[68], a software for variation graph construction, handling and analysis, into splice-aware graphs, with paths through nodes (exons) and edges (splice junctions) representing the structure of mRNA transcripts. It is also possible to build pantranscriptomes by projecting a set of haplotype-specific transcripts onto a set of known haplotypes[68].

The size of a pangenome graph depends upon the genome size of the respective species, but is bound to be larger, as it incorporates accessory sequences from other individuals, and is also it is influenced by the number and diversity of the individuals contributing to a pangenome, as well as the construction pipeline (**Supplementary Table 1**). The size of the MC graph is relatively close to the genome size of the species, (~3.2 Gbp versus 3.1 Gbp for human[37], 1.2 Gbp versus 1.1 Gbp for chicken[21], ~1.6 Gbp versus 1.1 Gbp for barn swallow[29]; **Supplementary Table 1**; **Box 2**). In contrast, PGGB graph size can considerably exceed that of the genome size and MC graphs (e.g. 8.4 Gbp for human[37]). The larger size of PGGB graphs is explained by their capability to capture highly divergent satellite, centromeric and heterochromatic regions, excluded in MC graphs due to alignment issues[21,37] (**Supplementary Table 1**). The largest increase relative to true genome size was observed in grapevine[24], likely because of the inclusion of different species (**Box 2**). PGGB also has a tendency to collapse complex regions, such as copy-number polymorphic loci, into a single copy, generating loops in the graph that increase its complexity[37,63] (**Fig. 2b**). Given their greater size and complexity, PGGB graphs require more computational resources than MC graphs. To construct a graph based on 10 human haplotypes, MC currently takes ~16 hours, 154 Gb of RAM and 7 Gb of disk space, while PGGB takes 117 hours, 71 Gb of RAM and 7.6 Gb of disk space[69]. PGGB has also been experimentally shown to potentially lead to an overestimation of sequence variability. For instance, the size of the PGGB chicken pangenome was larger than expected based on the estimated variation in diverse groups of chickens[21]. A preliminary estimation of the expected species variability should be computed to detect overestimation in graph construction.

Given all these differences, a careful selection of the pangenome graph construction pipeline is of utmost importance. On one hand, PGGB graphs are based on reference-free alignments and are more complete than MC graphs. On the other hand, their complexity increases the computational resources needed for graph construction and some downstream analyses, such as variant calling after read mapping, are currently computationally infeasible[21,37]. MC graphs are easier to handle, but they omit challenging regions such as centromeres, which are hotspots

of structural variation[18] and may play a crucial role in adaptive evolution and speciation[70,71]. Moreover, MC works on single chromosome graphs during graph construction and therefore does not allow the representation of interchromosomal rearrangements[37], precluding, for example, the investigation of acrocentric chromosome evolution[72]. We suggest choosing the pipeline based on the desired analyses and the available computational resources. Both graphs can be used for graph decomposition and the identification of variants between the genomes in the graph; however, while MC should be used as a reference for resequencing projects, PGGB is particularly useful when focusing on regions of interest. Overall, pangenome graphs face conceptual and computational challenges and currently require significantly more resources for their construction, storage, and analysis than linear genomes. While these limitations are being tackled (e.g. through 'implicit' pangenome construction of only regions of interest[73]), researchers need to ensure access to sufficient computing resources, such as clusters or cloud computing infrastructures.

Improving the accessibility of biological information in pangenome graphs

Pangenome graphs enclose extensive and complex biological information, including genomic relationships and diversity among individuals. Their intricate and complex structure generates large data volumes that are challenging to navigate and interpret. Graph manipulation toolkits have been developed to improve graph accessibility to software used in downstream analyses, thereby facilitating the extraction of biological information from pangenome graphs. Graph manipulation includes tasks such as sorting, indexing, pruning and subsampling (**Fig. 2b**). Sorting optimizes the order of graph nodes to reveal the underlying latent and sparse graph structure and minimize errors in analysis and interpretation[74]. Path indexing provides faster access to specific regions of the graph, allowing software to quickly locate genes, variants, and other features of interest without scanning the entire pangenome and reducing the time required to retrieve relevant information for analyses such as read mapping and variant calling[75]. To further speed up computation, it is possible to simplify graph topology through pruning of complex or unreliable regions, or by subsampling user-defined coordinates[74,76]. Subsampling is particularly helpful to disentangle the complexity of a particular region of interest, or when computational resources are not available to query the entire graph. After graph construction and each manipulation step, it is helpful to perform diagnostic statistics, such as graph size, number of nodes and base content to get a sense of the structure of the pangenome and how each step affected the graph[74,76]. However, being a relatively new approach, pangenome graphs lack universally accepted quality metrics due to their intrinsic complexity, the multiple

construction methods, and the lack of standardized benchmarking datasets. Nonetheless, as pangenome graphs become more widely adopted, efforts will develop a unified set of metrics applicable to all pangenome graphs.

Two main software packages exist to manipulate pangenome graphs: vg[76] and the pangenome analysis toolkit ODGI[74]. Vg relies on the .vg format[76] and was the first tool to be scaled up to gigabase-scale graphs. ODGI operates on a node-centric object (.og), and was optimized for pangenome graphs with hundreds of haplotype-resolved genomes[74]. ODGI's tools work on a graph-independent universal coordinate system that remains constant among different graphs built from the same sequences[74]. This system enables coordinate translation, facilitating the lift-over of coordinate-based features between genomes and graphs[74], i.e., the accurate mapping of annotated features (such as genes, regulatory elements, or other functional elements) from one genome assembly to another[74]. CAT (Comparative Annotation Toolkit)[77] can also annotate the haplotypes in a pangenome graph by projecting the reference gene annotation to each of the genomes, which can ease within-species annotation efforts[37]. Feature annotations can also be injected in the graph[74,76] and used to interpret the functional significance of paths, nodes, and edges.

Visualizing genome diversity among individuals

Graph visualization allows the inspection of homology relationships and variation between the genomes, providing insights on the latent biological data[74]. For instance, it can disentangle variation at complex loci following haplotype paths along variation bubbles[21]. Visualization can occur at different scales, from the overall structure down to the base-level (**Fig. 2c**). 2D visualizations highlight graph structure and identify complex loci, while 1D visualizations help understanding the graph topology and the relationships between genomes, potentially providing a more immediate understanding of the complexity of a region with respect to inspecting a list of variants. Various tools exist for pangenome graph visualization. Bandage[78] and GfaViz[79], originally created to visualize assembly graphs, permit the rendering of the 2D graph layout and the interactive inspection of nodes and edges with variation that appear as bubbles in the layout. Vg viz[76] can visualize nodes, edges, paths, and the base variation among sequences. SequenceTubeMap[80] renders these elements in a 1D "tube map" model where paths representing genomes navigate through the sequence nodes of the graph, oriented from left to right. Read alignments and feature annotations injected in the graph can also be visualized. To scale to gigabase pangenomes, such as the HPRC graph[37], MoMI-G[81] combines the base-level visualization of SequenceTubeMap[80] with CIRCOS[82] plots chromosome-level connections to

efficiently browse SVs between genomes and aligned reads. ODGI[74] can render a raster image of the graph topology in either 2D or 1D. Waragraph[83], an interactive implementation of ODGI, is currently being developed to be able to inspect both 1D and 2D visualizations. When dealing with a large graph, rendering the entire graph at once can become impractical and we recommend visualizing chromosome graphs or subsampled regions of interest.

Downstream analyses and their applications

Characterizing small variants and complex SVs through graph decomposition

Variant sites (SNPs, indels, and SVs; **Fig. 3a**) in a pangenome graph can be extracted through graph decomposition[38], the process of breaking down a pangenome graph into smaller, more manageable subgraphs or components (snarls or bubbles)[38]. Graph decomposition can be performed with vg snarl[38] and gfatools bubble[84] (**Supplementary Table 2**). Vg deconstruct[37], implemented in MC[20] and PGGB pipelines[64], can process the output of vg snarls or compute snarls automatically generating a VCF with variants called from the references chosen during graph construction with MC[20] or from any genome with PGGB[64]. When working with large graphs, it is recommended to compute snarls separately before variant calling[37]. The characterization of complex SVs, which were not previously accessible using linear reference genomes and short reads, can shed light on their role in evolution[85] and in shaping phenotypic variation, often with fitness consequences[86–89], as SVs can affect fitness by altering gene expression and shaping the chromosome recombination landscape[90]. A complete representation of SVs can also help analyze synteny and collinearity within genomes. In turn, this may provide insights into chromosome evolution by encompassing the full complexity of sex chromosomes and microchromosomes, which typically are enriched in SVs and challenging to resolve due to high-repeat and GC content[91]. In addition, human pangenome graphs have also allowed the identification of recombination events between heterologous acrocentric chromosomes, especially at the breakpoints of Robertsonian translocations[72]. These translocations are the most common chromosomal rearrangement in humans, and a comprehensive pangenome graph has greatly enhanced the identification of the sequences and mechanisms involved[72].

Population genomics and alignment of transcriptomics data

A pangenome graph can be used as a reference in resequencing projects to reduce mapping bias (**Fig. 3b**; **Supplementary Table 2**). Short-reads map with greater confidence when more

genomic sequences are represented and known variation is embedded in the reference[15]. However, read mapping to a pangenome is more challenging than mapping to a single reference genome, as the search space for alignment increases due to the large number of potential paths in the graph[92]. Since canonical algorithms cannot be applied directly to pangenome graphs[15], new tools have been developed for sequence-to-graph alignment. Within the vg toolkit, the general-purpose read mapper vg map[76] is suitable for large and complex variation graphs, albeit slower than popular linear-genome aligners with comparable accuracy[92]. Vg Giraffe[92], currently being extended to support long-reads, uses a graph Burrows-Wheeler transform (GBWT)[93], an indexing strategy that supports efficient querying and retrieval of sequences and variants from the pangenome graph, to identify the paths that represent the two observed haplotypes in an individual's reference sequences and to restrict the alignment space to these regions only, avoiding biologically unlikely allele combinations. This leads to a dramatic increase in mapping speed[92]. Long reads can also be aligned with GraphAligner[94], a seed-and-extend sequence-to-graph aligner. Improved short-read mappability will benefit resequencing projects, particularly in ancient DNA (aDNA) studies, which face challenges of contamination, degradation, small amounts of endogenous DNA, shorter reads, and, therefore, lower mappability[14]. aDNA mapped against a variation graph has already been proven to mitigate reference biases by improving the allelic balance in polymorphic sites[14]. aDNA reads with non-reference alleles maps in higher proportions to a graph containing alternate alleles, with respect to a linear reference genome[14].

Vg also allows splice-aware RNA-seq mapping to splice-aware graphs, generating an alignment that can then be used to quantify haplotype-specific transcript expression[68,95]. Pantranscriptomics has the potential to efficiently quantify haplotype-specific differential gene expression by exploiting the population variation that is embedded in the pantranscriptome reference[68]. We anticipate that pantranscriptomic sequencing projects combining RNA-seq data with pangenome graph references will clarify the effects of gene flow[96], detecting adaptive genetic variation[97,98]. Chromatin accessibility analyses, such as chromatin immunoprecipitation and sequencing (ChIP-seq) or assay for transposase-accessible chromatin with sequencing (ATAC-seq), also benefit from a pangenomic approach[37]. Their combination with RNA-seq data provides a multi-omic approach that might facilitate the interpretation of regulatory events critical to a wide variety of biological processes and phenotypes[37,99]. These approaches will enable future panepigenomics studies in non-model organisms, overcoming current limitations in handling large, multi-omics data sets[37,99].

Detecting and genotyping variants in resequencing studies

Pangenome graphs can increase the accuracy of variant calling and genotyping in resequencing studies thanks to improved read mappability[21,26,29,37] (**Box 2; Fig. 3b**; **Supplementary Table 2**). Vg can be used to extract snarls from the graph and compute coverage and mapping quality of aligned reads to accurately identify known variants[100]. Larger known SVs (deletions, insertions and inversions) can be genotyped by computing read coverage for each node and edge[100]. Specifically, variable sites identified by graph decomposition are assigned the two most supported paths, representing haplotypes[100]. For *de novo* small variant calling, the graph is first "augmented" with variants identified through read alignment, after which read support is computed[100]. A potentially more accurate approach consists of projecting the graph alignment from short-read mapping back into a linear BAM file (i.e., surjection) before using traditional variant callers (e.g. Deepvariant[101], Freebayes[102] and GATK with Elprep[103,104]) to generate a VCF referenced to the genome chosen for surjection. This approach can be also used with long reads mapped by GraphAligner[94]. Given the complexity of PGGB graphs, read mapping and variant calling have been mostly performed on MC graphs so far[21,37]. Vg Giraffe mapping followed by surjection and variant calling with Deepvariant is currently the state-of-the-art approach and surjection was found as the main computational bottleneck[21,37]. For the chicken pangenome, the number of mapped and surjected reads per CPU-second dramatically decreased when mapping against a PGGB graph with respect to a MC graph (1.6 reads versus 500 reads), while the memory usage increased (>250 Gb versus 24 Gb), making mapping and surjection with a PGGB graph computationally infeasible[21]. An alternative and faster approach for known variant genotyping that does not require read mapping is implemented in PanGenie[105]. This algorithm combines long-range haplotype information embedded in the graph and *k*-mer counts from short-read data to jointly genotype SNPs, indels and SVs in the uncharacterized sample (**Fig. 3c**). Haplotypes present in the graph can support genotype assignment based on neighboring bubbles in case a given bubble is poorly covered by short-read *k*-mers[105].

In population studies, pangenome-based variant calling increases accuracy and reduces the per-sample data requirements, potentially expanding the size of assessable cohorts[37]. Determining accurate and comprehensive variant sets increases the resolution of the analysis of demographic history, linkage-disequilibrium (LD), and genome-wide selection scans. This is particularly beneficial in species with large effective population sizes, where LD is low[29,106,107]. By improving SVs genotyping, pangenome graphs can also help to integrate SVs into GWAS,

especially as long-reads gradually replace short reads in resequencing projects[24,37,105]. Performing GWAS on pangenome-based SNP and SV panels can therefore enhance our understanding of the genetic basis of complex polygenic traits, and shed light on the role of natural selection and gene-environment interactions and correlations.

Conclusions and future prospects

The last few years have seen a burgeoning of both small- and large-scale projects generating high-quality reference genomes for biodiversity studies[108], including the Vertebrate Genomes Project[5,7], the Darwin Tree of Life[109], and the European Reference Genome Atlas[110]. Most of these projects contribute to the Earth Biogenome Project[6], an ambitious proposal launched in 2020 to collectively sequence all named eukaryotic species within the next ten years. While pangenome graphs are currently available only for a handful of species, recent advances in genome and pangenome assembly potentially extend this approach to most eukaryotic species. This is a desirable goal to reduce representation bias in all analyses of biodiversity, its evolution, and conservation. Collecting, sequencing and assembling pangenomes from more than a few individuals could be impractical in many species due to costs and sample availability. In those cases, a pangenomes from single-to-few individuals would still increase representation and reduce reference bias, especially for highly heterozygous populations where a single individual may carry a high amount of allelic diversity. Pangenome graphs can benefit a broad range of applications for biodiversity, from population genomics, phylogenomics, hybridization and speciation studies, to conservation genomics, and will likely become the standard reference system for such research in the future. Many new directions are being investigated. For instance, panmitogenomes, i.e., pangenomes constructed from thousands of mitochondrial genomes, have been shown to improve haplotyping of individuals[111] and are being considered for species identification from heterogeneous samples. Another promising new direction for the field is in super-pangenome graphs, which expand the survey of variation to taxonomic ranks above species, opening new possibilities to study the molecular and evolutionary mechanisms underlying species divergence, selection and recombination processes, as well as adaptation to rapid climate changes[27]. Dense sampling and sequencing of species within a clade have proven essential for deciphering phylogenetic and phylogeographic relationships, as well as facilitating investigation of gene loss and selection events[112]. To this end, a pangenomic approach revealed complex phylogenetic relationships among bacterial

strains, allowing genetic analyses of infectious diseases to identify virulence and antimicrobial resistance genes with greater accuracy[113]. Despite the complexity of eukaryotic genomes, we envision that rapid improvements in the efficiency and scalability of pangenomic tools will soon allow such phylogenomic applications to be extended to eukaryotic species. In particular, owing to the ability of super-pangenome graphs to incorporate all types of genomic variation, they have the potential to elucidate complex evolutionary histories and phylogeographic relationships of large, panmictic, and highly recombinant wild populations[29,114], as well as to improve phylogenetic reconstructions of events, such as incomplete lineage sorting. Super-pangenomes can also assist in studying biodiversity in complex ecosystems where hybridization occurs[49,115]. Hybrid zones are a prime opportunity for pinpointing the genes responsible for phenotypic traits, as genes introgress in parallel with those traits[116]. Inclusion of both hybridizing species in a pangenome graph will mitigate biases that arise from the use of the reference genome of either species[117]. Pangenomes could also help shed light on the origin of islands of divergence, highly differentiated genomic regions that might be related to reproductive isolation and, thus, to speciation processes[118,119]. Even within the same species, a comprehensive pangenome graph that includes assemblies for all subspecies can maximize the identification of structural genetic variants that are unique to a particular subspecies (**Box 2**). We predict that species-level pangenomes will also replace linear genomes in phylogenomic comparative genomics studies, enabled by the future development of tools for aligning pangenomes of different species. Currently, the construction of a pangenome graph lacks evolutionary information across the individuals, as phylogenetic divergence is not considered in pairwise alignments, and this limitation should be taken into account when performing phylogenetic analyses.

Pangenome graphs may also effectively guide conservation strategies aimed at maximizing the preservation of genetic variation[2], by capturing a fuller spectrum of genetic diversity. Of particular interest are structural and functional genomic variation involved in adaptations and responses to environmental pressures. This will improve selection criteria for reintroducing and translocating individuals among populations of threatened and endangered species. Improved representation of structural elements, such as SVs, centromeres and telomeres, and CNVs, as well as SNPs, along with non-coding regulatory elements, can provide comprehensive conservation-relevant information regarding inbreeding, outbreeding, deleterious mutations, introgression, and local adaptation[2]. Pangenomes can also help to identify different genomic regions in cryptic species, to then develop multilocus probes that distinguish cryptic taxa and simplify conservation management[120]. Moreover, we envision that pangenomics could help

reconstruct the genomic blueprint of extinct biodiversity by improving the mappability of aDNA against a pangenome of a closely-related species. A more comprehensive comparison between the extinct species and its living relative will help identify the genetic variation underlying lost traits and ecosystem functions, which are essential information for any de-extinction and restoration efforts[14,121]. In conclusion, as methods to assemble, visualize, annotate, and analyze pangenomes graphs continue to improve, we recommend researchers in biodiversity genomics to embrace this new paradigm.

Author contributions

G.R.G., G.F., and S.S, conceived the manuscript. S.S., G.R.G., G.F., and L.G. drafted the manuscript and ideated the figures, with significant contributions from A.B-A., C.R.F., R.R., and E.D.J. All authors reviewed and approved the final text.

Acknowledgments

We are grateful to the HPRC community for the useful discussions over the years that help shape the manuscript. We would like to particularly thank Erik Garrison for his input to the manuscript. We thank Caterina Di Pietro for drawing the final figures. C.R.F. thanks the support of CE3C through an assistant researcher contract (FCiência.ID contract #366) and FCT (Fundação para a Ciência e a Tecnologia) for Portuguese National Funds attributed to CE3C within the projects UIDB/00329/2020, UIDP/00329/2020, and LA/P/0121/2020, and FPUL for a contract of invited assistant professor.

Competing interests

The authors declare no competing interests.

Boxes

Box 1: History of the pangenome concept from bacteria to the human pangenome

The origin of the pangenome concept traces back to 2005[122], when the bacterial genomes of *Streptococcus agalactiae* were first described as collections of core genes shared among strains, dispensable ('accessory') genes shared between some strains, and strain-specific (unique) genes. In *S. agalactiae*, the core genome included 80% of the genes, with the remaining 20% categorized as accessory[122]. Work has since focused on the structure and dynamics of bacterial pangenomes[123–127], proving fruitful for taxonomic identification[128], to study host-pathogen interactions[124,129] and gene families essential to pathogenicity[128] and antibiotic resistance[130]. These studies resulted in biomedical applications, with new promising candidate drug targets[130] and reverse vaccinology[131,132]. The concept was rapidly adopted by plant and animal researchers, resulting in multiple eukaryotic pangenome studies[46]. To accommodate the complexity of eukaryotic genomes, with large portions (>~50%[133]) of non-coding and yet functional sequences, the term evolved to represent a complete set of sequences found in all individuals of a population or a species[46]. In plants, the pangenome concept was first applied to the analysis of transposable elements, responsible for a large amount of variation in both genic and intergenic sequences[134]. Plant genomes are particularly dynamic as they undergo frequent polyploidization and diploidization events[135,136]. Moreover, intraspecific genetic variability is often large[137]. For these reasons, the concept of the pangenome has been rapidly extended from initial efforts in crops, such as rice[138] and tomato[139], to many other plant species[140–142], for association mapping analyses[143,144], breeding[145], crop improvement[146] and evolutionary analyses[147]. Very recently, a pangenome comprising 69 *A. thaliana* chromosome-level assemblies from a global range distribution was published[18]. Animal pangenomes have hitherto mostly focused on model species of economical value, both vertebrates and invertebrates. Among mammals, pangenome assembly and analysis projects have been conducted in the domestic pig (*Sus scrofa*)[22,148,149], cattle (*Bos taurus*)[150,151], and domestic sheep (*Ovis aries*)[152]. Among invertebrates, studies on silkworm (*Bombyx mori*)[45], Mediterranean mussel (*Mytilus galloprovincialis*)[17,153] and longwing butterflies (three *Heliconius* species)[154] pangenomes were recently published. The first step towards building a human pangenome was taken with the assembly of an Asian and an African genome and their integration within the NCBI reference human genome[155]. This work resulted in the identification of approximately 5 Mbp (mega base pairs) of novel sequences. From these early analyses, the authors estimated that a comprehensive human pangenome should contain 19-40 Mb currently missing from the state-of-art reference assembly (NCBI Build 36.3)[155]. Since then, a Danish pangenome[156] has been assembled and the GenomeDenmark project initiated[157], both utilizing family trios and revealing novel single-nucleotide and structural variants. More recently, the assembly of an African pangenome[158] included approximately 10% more DNA (296 Mbp) than the reference assembly (GRCh38), and a pangenome built from hundreds of Han Chinese individuals[159] identified 29.5 Mbp of novel genomic sequences, including at least 188 novel protein-coding genes. The generation of a comprehensive catalog of human genetic

variation has advanced through the analysis of 338 high-quality assemblies of genetically divergent populations[143], demonstrating that, for a given genome sequenced to 40-fold coverage, over 400,000 previously unmapped reads, could now be aligned. The authors also tested this new resource for more efficient mapping of previously discarded RNA-Seq reads[160]. Recently, the HPRC released the first draft human pangenome graph using three graph construction methods and 47 phased, diploid assemblies from genetically diverse individuals[37]. Soon after, another human pangenome graph was built from 116 high-quality assemblies representing 36 Chinese minority ethnic groups[161]. In parallel, Seven Bridges Genomics Company generated population-specific genome graphs for a panAfrican genome[162].

Box 2: Case studies

The barn swallow pangenome

The barn swallow (*Hirundo rustica*) is a small migratory songbird, with six subspecies that differ in body size, extent and type of secondary sexual traits, and migratory behavior[163]. Latitudinal clines exist, with partial overlap in some traits[164] and variable levels of hybridization between subspecies[165]. In a first attempt to expand the characterization of the genetic variation within this species, we generated a preliminary pangenome variation graph for the Eurasian subspecies by using 12 haplotypes combined with MC[29]. The resulting pangenome increased the reference genome by approximately 500 Mbp (1.6 Gbp versus 1.1 Gbp). We were able to use the pangenome graph to infer core and accessory genes, and tested it as a reference for read mapping and variant calling, further highlighting the potential of pangenome graphs for population genomics[29].

The grape super-pangenome

The grapevine (*Vitis vinifera*) is one of the most important fruit crops globally, with great economic and cultural value. Recently, the assembly of nine genomes from North American wild *Vitis* species aimed to comprehensively characterize the genus diversity[24]. The genomes were scaffolded at the chromosome-level and fully phased. A super-pangenome was built using PGGB[64], enabling access to intra- and inter-specific genetic variants and augmenting the genome size threefold (1.7 Gbp versus 0.5 Mbp). The decomposition of variants embedded in the graph captured valuable genetic variants, including those associated with flower sex phenotype and disease resistance, thereby shedding light on species-specific adaptations. The super-pangenome also supported a pan-genome-wide association study (panGWAS), and identified variants near gene loci that are associated with chloride exclusion, potentially influencing plant salt tolerance. These findings highlight the potential of pangenomes in studying genetic variation and uncovering the genetic basis of functional traits, as well as helping to address the challenges posed by climate change.

Figure legends

Fig. 1: Principles of pangenome graphs

a) An example of reference bias. Short reads from three different individuals (shown here is a bird as an example) aligned to a linear reference genome do not map well to a missing region or a region with divergence (shown in red). In contrast, a pangenome reference based on multiple genomes (Hap.1, Hap.2, Hap.3; only one haplotype for each individual is represented for simplicity) improves the coverage of such regions as less diverging regions are sampled, thereby facilitating variant calling and subsequent downstream analyses. b) Pangenome graphs can be constructed from unaligned raw reads (top) or from the alignment of multiple assembled sequences (bottom). c) A pangenome graph is a bidirected graph with nodes representing DNA sequences (semi-transparent squares with base pairs) connected by bidirectional edges, which define the relationships between adjacent nodes and encode for the strandedness of the sequences. The genomes walk as paths through the nodes (fine coloured lines above each node), defining their base composition. Multiple genomes can share the same node sequence or take different paths across bubbles, or snarls, which are subgraphs indicating the presence of variation in that region. The first bubble represents an insertion in Hap.1, or deletion in Hap.2 and 3. The second bubble represents a SNP where Hap.1 and Hap.2 both have a G and Hap.3 has a C.

Fig. 2: Pangenome graph construction, manipulation and visualization.

a) General representation of sample selection, sequencing and assembly of haplotype-resolved genomes (illustrated here are butterflies as an example). Samples should be collected in different geographical locations to increase variability. Then, high-quality DNA is extracted, sequenced with long-read technologies such as PacBio and Oxford Nanopore, and assembled to obtain high-quality genomes. b) Schematic illustration of the two main methods for pangenome graph construction, reference-based (MC) and reference-free (PGGB). MC starts with Minigraph, which generates a SV-only graph, i.e. it starts with Hap.1 and adds the SVs from Hap.3 (> 50 bp, highlighted in yellow). The graph construction is therefore influenced by the order of the genomes aligned. Then, Cactus adds base-level information of all sequences to the graph to also represent any SNPs. PGGB does not depend on the order of the aligned genomes and starts from all-versus-all alignments generating a graph with loops representing complex regions of the genome, such as the centromeres. After graph construction, typically performed operations include sorting, indexing, pruning and subsampling in order to correct the order of the nodes, create an index to make the pangenome elements accessible to other software, prune complex and unreliable regions, and focus on regions of interest, respectively. c) Pangenome graphs can be visualized in many ways to help interpretation. Is it possible to visualize the overall structure of the graph in 2D (high-level visualization) focusing on the relationships between nodes and edges rather than the base composition of the paths. Alternatively it is possible to visualize the paths walking through the nodes together with their

base composition in 1D (base-level visualization). The latter is a tube-like representation of the graph shown in Fig. 1c. Each coloured line is a different genomes (only one haplotype per individual is shown for simplicity). Nodes report the DNA sequence (semi-transparent squares) and variation is represented as divergence in the paths walking through the nodes.

Fig. 3: Downstream analyses. a) Identification of variants between the genomes included in the pangenome. Sites of variation (bubble or snarls) are identified through graph decomposition based on the graph topology, i.e. following the paths of the genomes across the nodes and edges. Four different bubbles are represented in a tube-like manner as in Fig. 2c (base-level visualization). By walking through the nodes of different genomes, it is possible to identify a SNP, an indel (deletion/insertion), an inversion and a duplication, present here in Hap.3 with respect to Hap.1 and Hap.2. Variants are called from the bubbles by looking at the path divergence of a genome with respect to a chosen reference. b) Use of pangenome graphs for read mapping and variant calling. Mapped reads (thin lines) after graph augmentation can be visualized together with the pangenome graph reference (thick lines). This allows to call variants already present in the graph, either a homozygous call, in which the mapped reads share a G with Hap. 3, or novel variants (e.g. second heterozygous call, in which the read has a C with respect to all the haplotypes in the pangenome graph reference). Variants are called based on the divergence between the paths of the reads and the chosen reference genome. c) Identification of structural variants and sample genotyping using PanGenie, which avoids read mapping. Raw reads of a resequenced individual are divided into *k*-mers and assigned to nodes in the pangenome graph. To assign a genotype, *k*-mer counts are compared between paths at bubbles in the graph, which usually represent known haplotypes. Genomes are represented as thick lines with different coloured outlines. In the first bubble, the raw reads of the individual lack *k*-mers that match Hap.2 in that region, but have a 1-copy *k*-mer matching Hap.1, Hap.3 and Hap.4. Hence, the genotype inferred is Hap.1/Hap.4 (note that Hap.4 and Hap.3 share the same sequence in the bubble, so here it could have also been Hap.1/Hap.3). In the second bubble, the individual has a 2-copy *k*-mer matching Hap.2 and Hap.4. Therefore the inferred genotype is Hap.2/Hap.4. In the first bubble, Hap.4 has been chosen during genotype inference over Hap.3 as a possible divergent haplotype based on the neighboring bubble. In fact Hap.3 genotype is less likely because it is not supported by any *k*-mer in the second bubble.

Glossary

Bidirected graph: graph representing both DNA strands.

Nodes: the basic unit of a pangenome graph. They represent DNA sequences included in the graph. In a syntenic region, nodes can be traversed by multiple paths. In a bubble of variation, multiple nodes represent divergent DNA sequences, and only a subset of divergent haplotypes traverse each node.

Edges: node connectors indicating their concatenation and ultimately defining the DNA sequence of each haplotype (path). In a bidirected graph, the direction of the edges indicate the strandedness.

Paths: representation of the haplotypes included in the graph. A path is defined by a concatenation of DNA sequences (nodes) connected by edges.

Bubble: region of the graph where a set of paths diverge from a common node and reconverge in the following common node. The paths walking through the bubble represent divergent haplotypes and their sequence variation.

Snarl: hierarchical generalization of a bubble. A snarl is a subgraph with a start and an and node and paths traversing the snarl can have complex interconnections, representing variation.

k-mer: substrings of nucleotides of length *k*

Haplotype false duplication: error that occurs when a single genomic region is represented twice as two distinct regions in the same assembly. This typically happens when a heterozygous region in the individual contains two highly divergent haplotypes, causing the assembler to mistakenly treat them as non-homologous regions.

Phasing: process of determining which genetic variants are inherited together on the same chromosome from each parent.

Chromatin conformation data (Hi-C): Sequencing-based molecular technique used to detect regions in the genome where physical interactions are frequent. It measures the contact frequency between all pairs of loci, offering insights into the genome's three-dimensional organization.

Hybridization: breeding between individuals from genetically different lineages

Introgression: gene flow between hybridizing populations or species through the backcrossing of hybrids with one or both of the parent populations.

Synteny: conservation (not necessarily in the same order) of blocks of genes or entire chromosomal regions across different species.

Collinearity: preservation of the linear order of genes or genetic markers along chromosomes across different species

Microchromosomes: small-sized chromosomes typically found in the genomes of various animals, including birds and reptiles. They were often found to be gene-rich and GC-rich

Seed-and-extend aligner: an aligner that begins with small, exact alignment segments, known as seeds, and then attempts to extend or merge these segments to identify larger, highly similar regions.

ChIP-seq: a method used to analyze protein-DNA interactions by combining chromatin immunoprecipitation with next-generation sequencing to identify the binding sites of DNA-associated proteins.

ATAC-seq: a technique used to assess chromatin accessibility by using Tn5 transposase to insert sequencing adapters into open chromatin regions, which are then sequenced to map these accessible sites.

-

References

1. Giani, A. M., Gallo, G. R., Gianfranceschi, L. & Formenti, G. Long walk to genomics: History and current approaches to genome sequencing and assembly. *Comput. Struct. Biotechnol. J.* **18**, 9–19 (2020).
2. Theissinger, K. *et al.* How genomics can help biodiversity conservation. *Trends Genet.* **0**, (2023).
3. Lewin, H. A. *et al.* Earth BioGenome Project: Sequencing life for the future of life. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 4325–4333 (2018).
4. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
5. Rhie, A. *et al.* Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**, 737–746 (2021).
6. Lewin, H. A. *et al.* The Earth BioGenome Project 2020: Starting the clock. *Proc. Natl. Acad. Sci. U. S. A.* **119**, (2022).
7. Larivière, D. *et al.* Scalable, accessible and reproducible reference genome assembly and evaluation in Galaxy. *Nat. Biotechnol.* **42**, 367–370 (2024).
8. Ghildiyal, K. *et al.* Genomic insights into the conservation of wild and domestic animal diversity: A review. *Gene* **886**, 147719 (2023).
9. Blaxter, M. *et al.* Why sequence all eukaryotes? *Proc. Natl. Acad. Sci. U. S. A.* **119**, (2022).
10. Chaisson, M. J. P. *et al.* Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* **10**, 1784 (2019).
11. Wenger, A. M. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019).
12. Zhao, X. *et al.* Expectations and blind spots for structural variation detection from long-read assemblies and short-read genome sequencing technologies. *Am. J. Hum. Genet.* **108**, 919–928 (2021).
13. Günther, T. & Nettelblad, C. The presence and impact of reference bias on population genomic studies of prehistoric human populations. *PLoS Genet.* **15**, e1008302 (2019).

14. Martiniano, R., Garrison, E., Jones, E. R., Manica, A. & Durbin, R. Removing reference bias and improving indel calling in ancient DNA data analysis by mapping to a sequence variation graph. *Genome Biol.* **21**, 250 (2020).
15. Eizenga, J. M. *et al.* Pangenome Graphs. *Annu. Rev. Genomics Hum. Genet.* **21**, 139–162 (2020).
16. Taylor, D. J. *et al.* Beyond the Human Genome Project: The Age of Complete Human Genome Sequences and Pangenome References. *Annu. Rev. Genomics Hum. Genet.* (2024) doi:10.1146/annurev-genom-021623-081639.
17. Gerdol, M. *et al.* Massive gene presence-absence variation shapes an open pan-genome in the Mediterranean mussel. *Genome Biol.* **21**, 275 (2020).
18. Lian, Q. *et al.* A pan-genome of 69 Arabidopsis thaliana accessions reveals a conserved genome structure throughout the global species range. *Nat. Genet.* **56**, 982–991 (2024).
19. Chin, C.-S. *et al.* A diploid assembly-based benchmark for variants in the major histocompatibility complex. *Nat. Commun.* **11**, 4794 (2020).
20. Hickey, G. *et al.* Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nat. Biotechnol.* **42**, 663–673 (2024).
21. Rice, E. S. *et al.* A pangenome graph reference of 30 chicken genomes allows genotyping of large and complex structural variants. *BMC Biol.* **21**, 267 (2023).
22. Jiang, Y.-F. *et al.* Pangenome obtained by long-read sequencing of 11 genomes reveal hidden functional structural variants in pigs. *iScience* **26**, 106119 (2023).
23. Khan, A. W. *et al.* Super-Pangenome by Integrating the Wild Side of a Species for Accelerated Crop Improvement. *Trends Plant Sci.* **25**, 148–158 (2020).
24. Cochetel, N. *et al.* A super-pangenome of the North American wild grape species. *Genome Biol.* **24**, 290 (2023).
25. Li, N. *et al.* Super-pangenome analyses highlight genomic diversity and structural variation across wild and cultivated tomato species. *Nat. Genet.* **55**, 852–860 (2023).
26. Leonard, A. S., Crysnanto, D., Mapel, X. M., Bhati, M. & Pausch, H. Graph construction method impacts variation representation and analyses in a bovine super-pangenome. *Genome Biol.* **24**, 124 (2023).

27. Shi, T. *et al.* The super-pangenome of Populus unveils genomic facets for its adaptation and diversification in widespread forest trees. *Mol. Plant* (2024) doi:10.1016/j.molp.2024.03.009.
28. Theissinger, K. *et al.* How genomics can help biodiversity conservation. *Trends Genet.* (2023) doi:10.1016/j.tig.2023.01.005.
29. Secomandi, S. *et al.* A chromosome-level reference genome and pangenome for barn swallow population genomics. *Cell Rep.* **42**, 111992 (2023).
30. Rhie, A. *et al.* Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**, 737–746 (2021).
31. Pang, A. W. *et al.* Towards a comprehensive structural variation map of an individual human genome. *Genome Biol.* **11**, R52 (2010).
32. Chakraborty, M., Emerson, J. J., Macdonald, S. J. & Long, A. D. Structural variants exhibit widespread allelic heterogeneity and shape variation in complex traits. *Nat. Commun.* **10**, 4872 (2019).
33. Tigano, A. *et al.* Chromosome-Level Assembly of the Atlantic Silverside Genome Reveals Extreme Levels of Sequence Diversity and Structural Genetic Variation. *Genome Biol. Evol.* **13**, (2021).
34. Todesco, M. *et al.* Massive haplotypes underlie ecotypic differentiation in sunflowers. *Nature* **584**, 602–607 (2020).
35. Garg, S., Balboa, R. & Kuja, J. Chromosome-scale haplotype-resolved pangenomics. *Trends Genet.* **38**, 1103–1107 (2022).
36. Wang, S., Qian, Y.-Q., Zhao, R.-P., Chen, L.-L. & Song, J.-M. Graph-based pan-genomes: increased opportunities in plant genomics. *J. Exp. Bot.* **74**, 24–39 (2023).
37. Liao, W.-W. *et al.* A draft human pangenome reference. *Nature* **617**, 312–324 (2023).
38. Paten, B. *et al.* Superbubbles, Ultrabubbles, and Cacti. *J. Comput. Biol.* **25**, 649–663 (2018).
39. Wang, T. *et al.* The Human Pangenome Project: a global resource to map genomic diversity. *Nature* **604**, 437–446 (2022).
40. Miga, K. H. & Wang, T. The Need for a Human Pangenome Reference Sequence. *Annu. Rev. Genomics Hum. Genet.* **22**, 81–102 (2021).

41. Read, B. A. *et al.* Pan genome of the phytoplankton Emiliania underpins its global distribution. *Nature* **499**, 209–213 (2013).
42. Liu, Y. *et al.* Pan-Genome of Wild and Cultivated Soybeans. *Cell* **182**, 162–176.e13 (2020).
43. Talenti, A. *et al.* A cattle graph genome incorporating global breed diversity. *Nat. Commun.* **13**, 910 (2022).
44. Bozan, I. *et al.* Pangenome analyses reveal impact of transposable elements and ploidy on the evolution of potato species. *Proc. Natl. Acad. Sci. U. S. A.* **120**, e2211117120 (2023).
45. Tong, X. *et al.* High-resolution silkworm pan-genome provides genetic insights into artificial selection and ecological adaptation. *Nat. Commun.* **13**, 5619 (2022).
46. Golicz, A. A., Bayer, P. E., Bhalla, P. L., Batley, J. & Edwards, D. Pangenomics Comes of Age: From Bacteria to Plant and Animal Applications. *Trends Genet.* **36**, 132–145 (2020).
47. Eberlein, C. *et al.* Hybridization is a recurrent evolutionary stimulus in wild yeast speciation. *Nat. Commun.* **10**, 923 (2019).
48. Mavárez, J. *et al.* Speciation by hybridization in Heliconius butterflies. *Nature* **441**, 868–871 (2021).
49. Hübner, S. *et al.* Sunflower pan-genome analysis shows that hybridization altered gene content and disease resistance. *Nat Plants* **5**, 54–62 (2019).
50. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
51. Koren, S. *et al.* De novo assembly of haplotype-resolved genomes with trio binning. *Nat. Biotechnol.* (2018) doi:10.1038/nbt.4277.
52. Cheng, H. *et al.* Haplotype-resolved assembly of diploid genomes without parental data. *Nat. Biotechnol.* **40**, 1332–1335 (2022).
53. Korlach, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Methods Enzymol.* **472**, 431–455 (2010).
54. Mikheyev, A. S. & Tin, M. M. Y. A first look at the Oxford Nanopore MinION sequencer. *Mol. Ecol. Resour.* **14**, 1097–1102 (2014).
55. Leonard, A. S. *et al.* Structural variant-based pangenome construction has low sensitivity to variability of haplotype-resolved bovine assemblies. *Nat. Commun.* **13**, 3012 (2022).

764 56. Smith, T. *et al.* The first complete T2T Assemblies of Cattle and Sheep Y-Chromosomes
765 uncover remarkable divergence in structure and gene content. *Res Sq* (2024)
766 doi:10.21203/rs.3.rs-4033388/v1.
767 57. Beyter, D. *et al.* Long-read sequencing of 3,622 Icelanders provides insight into the role of
768 structural variants in human diseases and other traits. *Nat. Genet.* **53**, 779–786 (2021).
769 58. Kim, J. *et al.* The genome landscape of indigenous African cattle. *Genome Biol.* **18**, 34 (2017).
770 59. Kim, K. *et al.* The mosaic genome of indigenous African cattle as a unique genetic resource for
771 African pastoralism. *Nat. Genet.* **52**, 1099–1110 (2020).
772 60. Bakhtiari, M. *et al.* Variable number tandem repeats mediate the expression of proximal genes.
773 *Nat. Commun.* **12**, 2075 (2021).
774 61. Caballero-López, V., Lundberg, M., Sokolovskis, K. & Bensch, S. Transposable elements mark a
775 repeat-rich region associated with migratory phenotypes of willow warblers (Phylloscopus
776 trochilus). *Mol. Ecol.* **31**, 1128–1141 (2022).
777 62. Li, H., Feng, X. & Chu, C. The design and construction of reference pangenome graphs with
778 minigraph. *Genome Biol.* **21**, 265 (2020).
779 63. Garrison, E. *et al.* Building pangenome graphs. *bioRxiv* (2023) doi:10.1101/2023.04.05.535718.
780 64. Garrison, E. & Guarracino, A. Unbiased pangenome graphs. *Bioinformatics* **39**, (2023).
781 65. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100
782 (2018).
783 66. Armstrong, J. *et al.* Progressive Cactus is a multiple-genome aligner for the thousand-genome
784 era. *Nature* **587**, 246–251 (2020).
785 67. Guarracino, A., Mwaniki, N., Marco-Sola, S. & Garrison, E. *Wfmash: Base-Accurate DNA*
786 *Sequence Alignments Using WFA and mashmap2*. (Github).
787 68. Sibbesen, J. A. *et al.* Haplotype-aware pantranscriptome analyses using spliced pangenome
788 graphs. *Nat. Methods* **20**, 239–247 (2023).
789 69. Andreace, F., Lechat, P., Dufresne, Y. & Chikhi, R. Comparing methods for constructing and
790 representing human pangenome graphs. *Genome Biol.* **24**, 274 (2023).
791 70. Nergadze, S. G. *et al.* Birth, evolution, and transmission of satellite-free mammalian centromeric

domains. *Genome Res.* **28**, 789–799 (2018).

71. Ávila Robledillo, L. *et al.* Extraordinary Sequence Diversity and Promiscuity of Centromeric Satellites in the Legume Tribe Fabeae. *Mol. Biol. Evol.* **37**, 2341–2356 (2020).
72. Guarracino, A. *et al.* Recombination between heterologous human acrocentric chromosomes. *Nature* **617**, 335–343 (2023).
73. Garrison, E., Guarracino, A. & Kille, B. *Impg: Implicit Pangenome Graph*. (Github).
74. Guarracino, A., Heumos, S., Nahnsen, S., Prins, P. & Garrison, E. ODGI: understanding pangenome graphs. *Bioinformatics* **38**, 3319–3326 (2022).
75. Sirén, J. Indexing Variation Graphs. in *2017 Proceedings of the Meeting on Algorithm Engineering and Experiments (ALENEX)* 13–27 (Society for Industrial and Applied Mathematics, 2017).
76. Garrison, E. *et al.* Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotechnol.* **36**, 875–879 (2018).
77. Fiddes, I. T. *et al.* Comparative Annotation Toolkit (CAT)-simultaneous clade and personal genome annotation. *Genome Res.* **28**, 1029–1038 (2018).
78. Wick, R. R., Schultz, M. B., Zobel, J. & Holt, K. E. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* **31**, 3350–3352 (2015).
79. Gonnella, G., Niehus, N. & Kurtz, S. GfaViz: flexible and interactive visualization of GFA sequence graphs. *Bioinformatics* **35**, 2853–2855 (2019).
80. Beyer, W. *et al.* Sequence tube maps: making graph genomes intuitive to commuters. *Bioinformatics* **35**, 5318–5320 (2019).
81. Yokoyama, T. T., Sakamoto, Y., Seki, M., Suzuki, Y. & Kasahara, M. MoMI-G: modular multi-scale integrated genome graph browser. *BMC Bioinformatics* **20**, 548 (2019).
82. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).
83. Fischer, C. *Waragraph*. (Github).
84. Li, H. gfatools: Tools for manipulating sequence graphs in the GFA and rGFA formats. https://github.com/lh3/gfatools.

85. Sedlazeck, F. J., Lee, H., Darby, C. A. & Schatz, M. C. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nat. Rev. Genet.* **19**, 329–346 (2018).
86. Noor, M. A., Grams, K. L., Bertucci, L. A. & Reiland, J. Chromosomal inversions and the reproductive isolation of species. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 12084–12088 (2001).
87. Küpper, C. *et al.* A supergene determines highly divergent male reproductive morphs in the ruff. *Nat. Genet.* **48**, 79–83 (2016).
88. Weissensteiner, M. H. *et al.* Discovery and population genomics of structural variation in a songbird genus. *Nat. Commun.* **11**, 3403 (2020).
89. Wold, J. *et al.* Expanding the conservation genomics toolbox: Incorporating structural variants to enhance genomic studies for species of conservation concern. *Mol. Ecol.* **30**, 5949–5965 (2021).
90. Wellenreuther, M. & Bernatchez, L. Eco-Evolutionary Genomics of Chromosomal Inversions. *Trends Ecol. Evol.* **33**, 427–440 (2018).
91. Peona, V. *et al.* The hidden structural variability in avian genomes. *bioRxiv* 2021.12.31.473444 (2022) doi:10.1101/2021.12.31.473444.
92. Sirén, J. *et al.* Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science* **374**, abg8871 (2021).
93. Sirén, J., Garrison, E., Novak, A. M., Paten, B. & Durbin, R. Haplotype-aware graph indexes. *Bioinformatics* **36**, 400–407 (2020).
94. Rautiainen, M. & Marschall, T. GraphAligner: rapid and versatile sequence-to-graph alignment. *Genome Biol.* **21**, 253 (2020).
95. Wu, Z. *et al.* Human pangenome analysis of sequences missing from the reference genome reveals their widespread evolutionary, phenotypic, and functional roles. *Nucleic Acids Res.* **52**, 2212–2230 (2024).
96. Tamagawa, K., Yoshida, K., Ohrui, S. & Takahashi, Y. Population transcriptomics reveals the effect of gene flow on the evolution of range limits. *Sci. Rep.* **12**, 1318 (2022).
97. DeBiasse, M. B., Kawji, Y. & Kelly, M. W. Phenotypic and transcriptomic responses to salinity stress across genetically and geographically divergent Tigriopus californicus populations. *Mol. Ecol.* **27**, 1621–1632 (2018).

98. Liu, L., Wang, Z., Su, Y. & Wang, T. Population transcriptomic sequencing reveals allopatric divergence and local adaptation in Pseudotaxus chienii (Taxaceae). *BMC Genomics* **22**, 388 (2021).
99. Ma, S. & Zhang, Y. Profiling chromatin regulatory landscape: insights into the development of ChIP-seq and ATAC-seq. *Mol Biomed* **1**, 9 (2020).
100. Hickey, G. *et al.* Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biol.* **21**, 35 (2020).
101. Poplin, R. *et al.* A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983–987 (2018).
102. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *arXiv [q-bio.GN]* (2012).
103. Van der Auwera, G. A. & O'Connor, B. D. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. ('O'Reilly Media, Inc.', 2020).
104. Herzeel, C. *et al.* Multithreaded variant calling in elPrep 5. *PLoS One* **16**, e0244471 (2021).
105. Ebler, J. *et al.* Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Nat. Genet.* **54**, 518–525 (2022).
106. Li, M.-H. & Merilä, J. Population differences in levels of linkage disequilibrium in the wild. *Mol. Ecol.* **20**, 2916–2928 (2011).
107. Charmantier, A., Garant, D. & Kruuk, L. E. B. *Quantitative Genetics in the Wild*. (Oxford University Press, 2014).
108. Formenti, G. *et al.* The era of reference genomes in conservation genomics. *Trends Ecol. Evol.* (2022) doi:10.1016/j.tree.2021.11.008.
109. Darwin Tree of Life Project Consortium. Sequence locally, think globally: The Darwin Tree of Life Project. *Proc. Natl. Acad. Sci. U. S. A.* **119**, (2022).
110. Mazzoni, C. J., Ciofi, C. & Waterhouse, R. M. Biodiversity: an atlas of European reference genomes. *Nature* **619**, 252 (2023).
111. Rubin, J. D., Vogel, N. A., Gopalakrishnan, S., Sackett, P. W. & Renaud, G. HaploCart: Human mtDNA haplogroup classification using a pangenomic reference graph. *PLoS Comput. Biol.* **19**,

876 e1011148 (2023).
877 112. Feng, S. *et al.* Dense sampling of bird diversity increases power of comparative genomics.
878 *Nature* **587**, 252–257 (2020).
879 113. Yang, Z. *et al.* Pangenome graphs in infectious disease: a comprehensive genetic variation
880 analysis of Neisseria meningitidis leveraging Oxford Nanopore long reads. *Front. Genet.* **14**,
881 1225248 (2023).
882 114. Lombardo, G. *et al.* The Mitogenome Relationships and Phylogeography of Barn Swallows
883 (Hirundo rustica). *Mol. Biol. Evol.* **39**, (2022).
884 115. Payseur, B. A. & Rieseberg, L. H. A genomic perspective on hybridization and speciation. *Mol.*
885 *Ecol.* **25**, 2337–2360 (2016).
886 116. Hewitt, G. M. Hybrid zones-natural laboratories for evolutionary studies. *Trends Ecol. Evol.* **3**,
887 158–167 (1988).
888 117. Sebastianelli, M. *et al.* A genomic basis of vocal rhythm in birds. *Nat. Commun.* **15**, 3095
889 (2024).
890 118. Irwin, D. E. *et al.* A comparison of genomic islands of differentiation across three young avian
891 species pairs. *Mol. Ecol.* **27**, 4839–4855 (2018).
892 119. Hejase, H. A. *et al.* Genomic islands of differentiation in a rapid avian radiation have been driven
893 by recent selective sweeps. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 30554–30565 (2020).
894 120. van der Sprong, J. *et al.* A novel target-enriched multilocus assay for sponges (Porifera): Red Sea
895 Haplosclerida (Demospongiae) as a test case. *Mol. Ecol. Resour.* **24**, e13891 (2024).
896 121. Novak, B. J. De-Extinction. *Genes* **9**, (2018).
897 122. Tettelin, H. *et al.* Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae:
898 implications for the microbial 'pan-genome'. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 13950–13955
899 (2005).
900 123. Obert, C. *et al.* Identification of a Candidate Streptococcus pneumoniae core genome and regions
901 of diversity correlated with invasive pneumococcal disease. *Infect. Immun.* **74**, 4766–4777
902 (2006).
903 124. Rasko, D. A. *et al.* The pangenome structure of Escherichia coli: comparative genomic analysis

of E. coli commensal and pathogenic isolates. *J. Bacteriol.* **190**, 6881–6893 (2008).

125. Donati, C. *et al.* Structure and dynamics of the pan-genome of Streptococcus pneumoniae and closely related species. *Genome Biol.* **11**, R107 (2010).

126. Pinto, M. *et al.* Insights into the population structure and pan-genome of Haemophilus influenzae. *Infect. Genet. Evol.* **67**, 126–135 (2019).

127. Aggarwal, S. K. *et al.* Pangenomics in Microbial and Crop Research: Progress, Applications, and Perspectives. *Genes* **13**, (2022).

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

129. Kiu, R., Caim, S., Alexander, S., Pachori, P. & Hall, L. J. Probing Genomic Aspects of the Multi-Host Pathogen Clostridium perfringens Reveals Significant Pangenome Diversity, and a Diverse Array of Virulence Factors. *Front. Microbiol.* **8**, 2485 (2017).

130. Poulsen, B. E. *et al.* Defining the core essential genome of Pseudomonas aeruginosa. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 10072–10080 (2019).

131. Hisham, Y. & Ashhab, Y. Identification of Cross-Protective Potential Antigens against Pathogenic Brucella spp. through Combining Pan-Genome Analysis with Reverse Vaccinology. *J Immunol Res* **2018**, 1474517 (2018).

132. Naz, K. *et al.* PanRV: Pangenome-reverse vaccinology approach for identifications of potential vaccine candidates in microbial pangenome. *BMC Bioinformatics* **20**, 123 (2019).

133. Francis, W. R. & Wörheide, G. Similar Ratios of Introns to Intergenic Sequence across Animal Genomes. *Genome Biol. Evol.* **9**, 1582–1598 (2017).

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

135. Adams, K. L. & Wendel, J. F. Polyploidy and genome evolution in plants. *Curr. Opin. Plant Biol.* **8**, 135–141 (2005).

136. Cheng, F. *et al.* Gene retention, fractionation and subgenome differences in polyploid plants. *Nat Plants* **4**, 258–268 (2018).

137. Shi, J., Tian, Z., Lai, J. & Huang, X. Plant pan-genomics and its applications. *Mol. Plant* **16**,

168–186 (2023).

138. Sun, C. *et al.* RPAN: rice pan-genome browser for ~3000 rice genomes. *Nucleic Acids Res.* **45**, 597–605 (2017).

139. Gao, L. *et al.* The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nat. Genet.* **51**, 1044–1051 (2019).

140. Jayakodi, M. *et al.* The barley pan-genome reveals the hidden legacy of mutation breeding. *Nature* **588**, 284–289 (2020).

141. Song, J.-M. *et al.* Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of Brassica napus. *Nat Plants* **6**, 34–45 (2020).

142. Li, J. *et al.* Cotton pan-genome retrieves the lost sequences and genes during domestication and selection. *Genome Biol.* **22**, 119 (2021).

143. Tao, Y., Zhao, X., Mace, E., Henry, R. & Jordan, D. Exploring and Exploiting Pan-genomics for Crop Improvement. *Mol. Plant* **12**, 156–169 (2019).

144. Qin, P. *et al.* Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell* **184**, 3542–3558.e16 (2021).

145. Zhou, Y. *et al.* Graph pangenome captures missing heritability and empowers tomato breeding. *Nature* **606**, 527–534 (2022).

146. Zanini, S. F. *et al.* Pangenomics in crop improvement-from coding structural variations to finding regulatory variants with pangenome graphs. *Plant Genome* **15**, e20177 (2022).

147. Qiao, Q. *et al.* Evolutionary history and pan-genome dynamics of strawberry (Fragaria spp.). *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).

148. Tian, X. *et al.* Building a sequence map of the pig pan-genome from multiple de novo assemblies and Hi-C data. *Sci. China Life Sci.* **63**, 750–763 (2020).

149. Li, Z. *et al.* The pig pangenome provides insights into the roles of coding structural variations in genetic diversity and adaptation. *Genome Res.* **33**, 1833–1847 (2023).

150. Jang, J. *et al.* Chromosome-level genome assembly of Korean native cattle and pangenome graph of 14 Bos taurus assemblies. *Sci Data* **10**, 560 (2023).

151. Dai, X. *et al.* A Chinese indicine pangenome reveals a wealth of novel structural variants

introgressed from other Bos species. *Genome Res.* **33**, 1284–1298 (2023).

152. Li, R. *et al.* A sheep pangenome reveals the spectrum of structural variations and their effects on tail phenotypes. *Genome Res.* **33**, 463–477 (2023).
153. Saco, A. *et al.* Gene presence/absence variation in Mytilus galloprovincialis and its implications in gene expression and adaptation. *iScience* **26**, 107827 (2023).
154. Ruggieri, A. A. *et al.* Erratum: A butterfly pan-genome reveals that a large amount of structural variation underlies the evolution of chromatin accessibility. *Genome Res.* **32**, 2145 (2022).
155. Li, R. *et al.* Building the sequence map of the human pan-genome. *Nat. Biotechnol.* **28**, 57–63 (2010).
156. Besenbacher, S. *et al.* Novel variation and de novo mutation rates in population-wide de novo assembled Danish trios. *Nat. Commun.* **6**, 5969 (2015).
157. Maretty, L. *et al.* Sequencing and de novo assembly of 150 genomes from Denmark as a population reference. *Nature* **548**, 87–91 (2017).
158. Sherman, R. M. *et al.* Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat. Genet.* **51**, 30–35 (2019).
159. Duan, Z. *et al.* HUPAN: a pan-genome analysis pipeline for human genomes. *Genome Biol.* **20**, 149 (2019).
160. Wong, K. H. Y. *et al.* Towards a reference genome that captures global genetic diversity. *Nat. Commun.* **11**, 5482 (2020).
161. Gao, Y. *et al.* A pangenome reference of 36 Chinese populations. *Nature* **619**, 112–121 (2023).
162. Tetikol, H. S. *et al.* Pan-African genome demonstrates how population-specific genome graphs improve high-throughput sequencing data analysis. *Nat. Commun.* **13**, 4384 (2022).
163. Spina, F. The EURING swallow project: a large-scale approach to the study and conservation of a long-distance migrant. *Migrating birds know no boundaries Proceedings of the*.
164. M Ller, A. P. SEXUAL SELECTION IN THE BARN SWALLOW (HIRUNDO RUSTICA). IV. PATTERNS OF FLUCTUATING ASYMMETRY AND SELECTION AGAINST ASYMMETRY. *Evolution* **48**, 658–670 (1994).
165. Scordato, E. S. C. *et al.* Genomic variation across two barn swallow hybrid zones reveals traits

associated with divergence in sympatry and allopatry. *Mol. Ecol.* **26**, 5676–5691 (2017).