frontiers | Frontiers in Genetics

TYPE Review
PUBLISHED 19 September 2025
DOI 10.3389/fgene.2025.1679660

OPEN ACCESS

EDITED BY
Jared C Roach,
Institute for Systems Biology (ISB), United States

REVIEWED BY
Yosuke Kawai,
National Center for Global Health and Medicine, Japan
Zhikun Wu,
Guangzhou Medical University, China
Fawaz Dabbaghie,
Seoul National University, Republic of Korea

*CORRESPONDENCE
Denis M. Nyaga,
denis.nyaga@auckland.ac.nz
Justin M. O'Sullivan,
justin.osullivan@auckland.ac.nz

RECEIVED 04 August 2025
ACCEPTED 08 September 2025
PUBLISHED 19 September 2025

CITATION
Nyaga DM, Zaied RE, Silander OK, Black MA and O'Sullivan JM (2025) Beyond single references: pangenome graphs and the future of genomic medicine.
Front. Genet. 16:1679660.
doi: 10.3389/fgene.2025.1679660

Beyond single references: pangenome graphs and the future of genomic medicine

Denis M. Nyaga[1]*, Roan E. Zaied[1], Olin K. Silander[1], Michael A. Black[2] and Justin M. O'Sullivan[1]*

[1]Liggins Institute, University of Auckland, Auckland, New Zealand, [2]Department of Biochemistry, University of Otago, Dunedin, New Zealand

Genomic medicine relies on single reference genomes that miss crucial genetic diversity, creating diagnostic gaps that disproportionately affect underrepresented populations. Pangenome graphs, collections of diverse genomes represented as interconnected genetic paths, offer a powerful alternative to the standard reference genome approach. Pangenome-based approaches capture the spectrum of human variation, dramatically improving how we detect complex structural variants, reconstruct haplotypes, and reduce bias in genetic studies. Projects like the Human Pangenome Reference Consortium have identified hundreds of megabases of missing genetic diversity, leading to remarkable improvements in variant detection across different populations. Yet, as pangenomes grow larger and computationally complex, they become more challenging to interpret clinically, creating a trade-off between comprehensiveness and usability. This review discusses the technical and conceptual advances enabling clinical applications of pangenomes in rare disease diagnosis. Realizing the future potential of pangenome graphs in genomic medicine will require innovative implementation strategies, thorough clinical testing, and user-friendly approaches.

KEYWORDS

clinical diagnosis, genetic diversity, genome assembly, haplotypes, long-readsequencing, pangenome graphs, reference genomes

1 Introduction

Genome sequencing is transforming medicine, enabling the detection of rare genetic variants (i.e., single nucleotide variants [SNVs], structural variants [SVs], insertions and deletions [indels], copy number variants [CNVs], and short tandem repeats [STRs]) that are missed with traditional genotyping. However, standard approaches to variant discovery rely almost entirely on comparison to a single linear reference genome, which, by its nature, lacks genetic diversity and does not represent the full range of human populations (Aganezov et al., 2022; Nurk et al., 2022; Liao et al., 2023; Hickey et al., 2024; Sirén et al., 2024; Taylor et al., 2024). Over-reliance on a single reference genome is a substantial barrier to equitable, high-resolution diagnosis (Matalon et al., 2023). In this review, we argue that pangenomes (i.e., collections of genomes) are not merely an incremental improvement but, together with graph-based genome encoding (i.e., the storage of genomic data as haplotype paths), constitute a disruptive paradigm shift that will render current variant discovery pipelines in genomic medicine obsolete. Ironically, as more pangenomes are built with increasingly large collections of genetic variations, it will

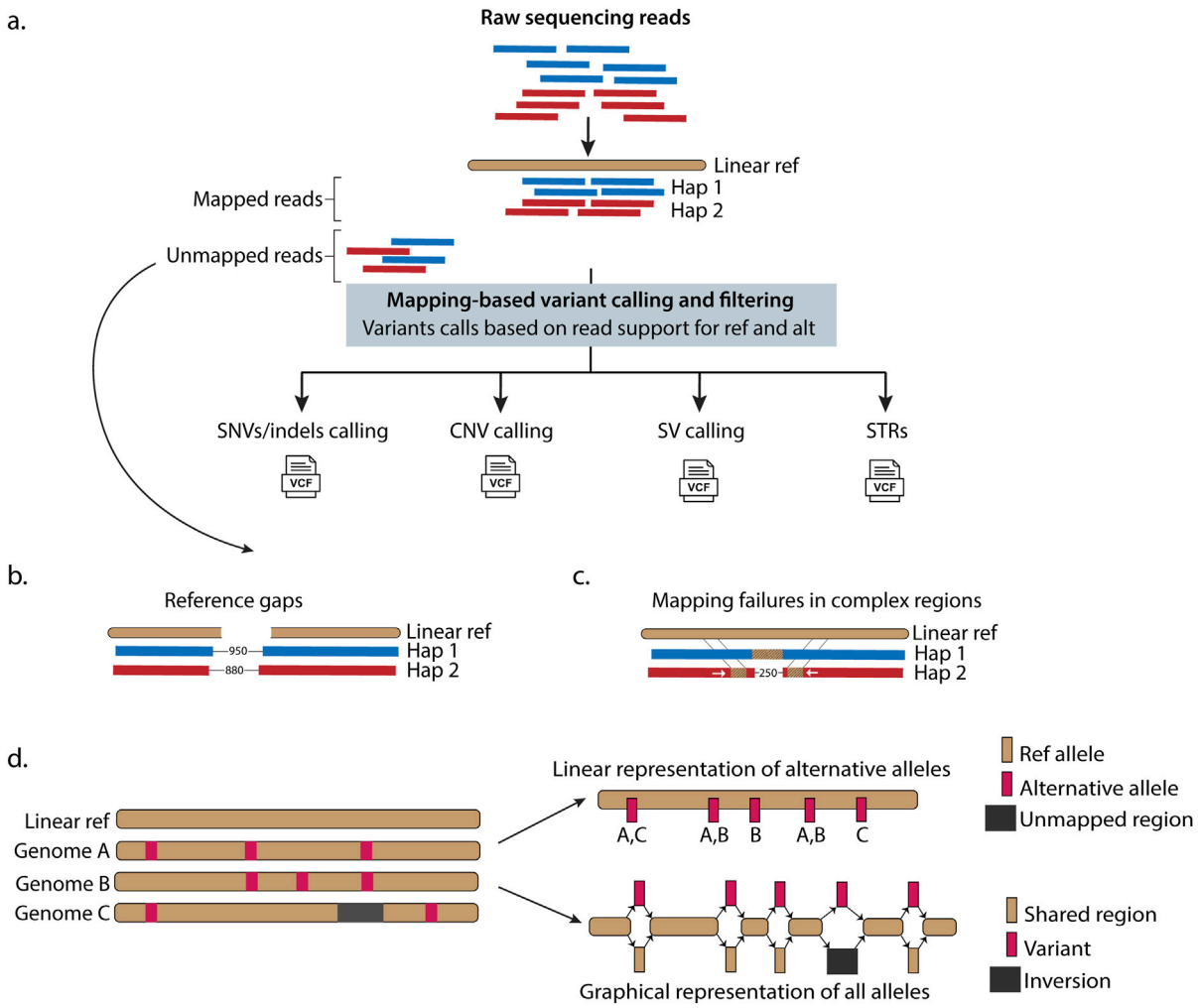

FIGURE 1
Linear reference-based genomic analysis has inherent limitations in alignment, variant detection, and representing population genetic diversity. **(a)** Raw sequencing reads from paired haplotypes are aligned to a linear reference genome, introducing inherent bias when genomic regions diverge significantly from the reference. Successfully aligned reads undergo variant detection for SNVs, small indels, SVs, CNVs, and STRs based on the read depth support for the linear reference (ref) and alternative (alt) sequence. However, variants in highly divergent regions remain undetected when reference genome sequencing reads fail to map to the linear reference due to reference genome gaps **(b)** or complex variations within genomic regions (such as an insertion on hap (haplotype) 1 and two inverted translocations with a deletion on hap (haplotype 2) **(c)**. **(d)** Linear references inadequately represent population genetic diversity. Linear models represent variations relative to the reference while graph-based models enable direct all-to-all genome comparisons, capturing complete sequence relationships. Graphs models efficiently represent SVs (such as the inversion shown in dark grey) that linear models miss. Hap - haplotypes, ref - reference, VCF - variant call format file.

become harder for clinicians and researchers to understand and use them effectively. As such, new approaches are required to enable rapid, interpretable pangenome queries.

2 The linear reference paradox

The Genome Reference Consortium (GRC) currently maintains two primary human reference assemblies: GRCh37 (hg19, released 2009) (Church et al., 2011) and its successor GRCh38 (hg38, published 2013) (Schneider et al., 2017). The GRCh38 assembly is a composite of unphased single haplotypes, with about 70% derived from a single individual, 23% from ten, and 7% from over fifty additional sources (Ballouz et al., 2019). These two reference genomes serve as critical foundations for genomic research, enabling clinical, comparative, developmental, population, and disease analyses (Lowy-Gallego et al., 2019; Abascal et al., 2020; Collins et al., 2020; GTEx Consortium, 2020; Taliun et al., 2021; Aganezov et al., 2022; Nassar et al., 2023; Reis et al., 2023; Mahmoud et al., 2024; Nyaga et al., 2024). The power that the incorporation of these references into biological studies brings has been unequivocally demonstrated, including through the identification of the genetic and molecular basis of rare diseases (Lunke et al., 2023; Nyaga et al., 2024; Sinha et al., 2025).

GRCh37 and GRCh38 are not and have never been fixed entities. Rather, the GRC has continually worked to improve these assemblies by implementing patches, fixes, and alternate scaffolds to represent allele diversity. For example, the transition from GRCh37 to GRCh38 included approximately 100 Megabases (Mb) of improvements, particularly in immune-related regions

(Schneider et al., 2017). Despite these efforts, the lack of ancestral diversity within these references remains a considerable limitation (Figure 1), particularly in clinical settings. For example, it is possible that diagnostic *de novo* pathogenic variants remain undetected because they lie outside the reference structure in the gaps (i.e., 7% or 210 Mb of its primary chromosome, unknown sequences [151 Mb], or computationally simulated regions [59 Mb] that are present in the reference structure (Figures 1A–C) (Aganezov et al., 2022; Nurk et al., 2022). Additionally, GRCh38 includes alternative contigs (i.e., continuous stretches of DNA sequence representing alternative haplotype diversity), which can lead to variant calling errors (Jia et al., 2020; Li H. et al., 2021; Aganezov et al., 2022) and biased variant interpretation, particularly towards insertions and deletions (indels) (Church et al., 2011; Schneider et al., 2017; Pan et al., 2019; Li H. et al., 2021). For example, mishandling alternative scaffold inclusion resulted in incorrect reports of genetic variation in 641 genes in UK Biobank exome data (Jia et al., 2020).

Despite its limitations, it must be acknowledged that substantial progress has been made in the continual development of the reference human genome. Since the first draft of the human genome was published in 2001, the quality (i.e., accuracy of base calling and assembly) and contiguity (i.e., the length of continuous DNA sequences without gaps) have improved substantially. Indeed, the initial human genome was incomplete and highly fragmented, consisting of more than 150,000 contigs, with just over half the genome represented in contigs greater than 50 Kilobases (Kb) (Church et al., 2011). The recent incorporation of long reads, which can span multiple Kb in a single continuous read, has allowed researchers to create a telomere-to-telomere genome assembly (T2T) (Rautiainen et al., 2023). The T2T-CHM13v2.0 human reference genome (Cai et al., 2020; Nurk et al., 2022) is a near-gapless, 'error-free' telomere-to-telomere assembly consisting of over 3.0 Gigabases (Gb) of fully resolved sequence that is contiguous across all autosomes and chromosome X (except for the highly repetitive ~9 Mb sequence from ribosomal DNA arrays) for a haploid human genome (Nurk et al., 2022).

The complete T2T genome assembly has contributed to the successfully resolved previously difficult-to-sequence regions, including the short arms of all five acrocentric chromosomes, centromeric repeats and segmental duplication (Nurk et al., 2022). This has enabled researchers to study genetic variation across these complex regions (Jeong et al., 2025). Despite representing only a single human haplotype, the T2T-CHM13v2.0 reference has already contributed to improvements in genomic variant discovery. For example, the assembly has enabled the discovery of over 2 million additional SNVs in regions missing from GRCh38 and has improved CNV detection across the 1000 Genomes Project samples (Aganezov et al., 2022; Nurk et al., 2022). In addition, researchers have attempted to genotype population-level SVs by utilizing long-read sequencing in diverse genomes and mapping these reads to the T2T assembly, identifying a large number of novel SVs (Reis et al., 2023; Gustafson et al., 2024; Logsdon et al., 2025; Schloissnig et al., 2025). However, even with these significant advances, the T2T-CHM13v2.0 assembly does not fully represent the genetic diversity of the human population, as variation can only be comprehensively studied in the context of multiple populations, not just by comparison to a single reference (Figure 1D) (Garrison et al., 2018; Eizenga et al., 2020; Sirén et al., 2021; Wang et al., 2022; Liao et al., 2023).

The standardized coordinate systems provided by reference genomes are essential for communication and coordinated analysis across the scientific and medical communities (Ballouz et al., 2019). However, no single reference genome can fully represent human diversity. This and the other inherent limitations of the GRCh38 and T2T linear reference assemblies are becoming particularly obvious as individualized approaches to medicine and rare disease increase. This is particularly true for patients of non-European ancestry, who experience substantially lower diagnostic rates. One indication of this disparity is an observed ~23% increase in the burden of variants of uncertain significance (VUS) compared to individuals of European ancestry (Dawood et al., 2024). Such missed diagnoses translate into increased morbidity, highlighting the importance of the inequalities that arise from the clinical use of a single human genome reference sequence (Green et al., 2023; Matalon et al., 2023). To address this, clinical studies must transition to use collections of genome assemblies as part of ancestrally diverse pangenome-based approaches that are backwardly compatible with published knowledge.

3 Graph-based pangenomes as next-generation references

The concept of 'pangenome' was initially introduced in 2000 by Sigaux, who applied it to describe a comprehensive database containing genomic and transcriptomic changes found in tumors, healthy cells and experimental systems (Matthews et al., 2024). However, the term has evolved to the current graph reference (as reviewed by (Marschall et al., 2018) to describe a set of whole-genome assemblies from multiple individuals that are used together as a reference or analyzed collectively (Marschall et al., 2018; Eizenga et al., 2020; Wang et al., 2022; Liao et al., 2023). These collections of multiple genomes enable the inclusion of variation from human populations, especially when they include samples from individuals underrepresented in previous genetic studies. As such, pangenomes offer a promising alternative to single linear reference assemblies for studying genetic variation (Figure 2A) (Wang et al., 2022; Gao et al., 2023; Liao et al., 2023). However, pangenomes also risk creating new inequities if not carefully implemented. If pangenomes are built predominantly from well-resourced populations or lack diverse ancestral representation, they will perpetuate existing biases despite appearing inclusive. The technical and cost implications of pangenome initiatives should not be overlooked. For example, the substantial amount of genetic data required to build ancestrally diverse pangenomes could restrict access for researchers and clinicians in resource-constrained settings (Gong et al., 2023; Liao et al., 2023). Additionally, as these pangenomes grow larger, they become computationally demanding, requiring extensive memory to process complete graphs, complex sorting algorithms, specialized visualization tools (e.g., Optimized Dynamic Genome/Graph Implementation [ODGI] (Guarracino et al., 2022)), and sophisticated indexing methods, which consequently make clinical interpretation challenging (Gong et al., 2023). This creates a trade-off between comprehensiveness and usability,

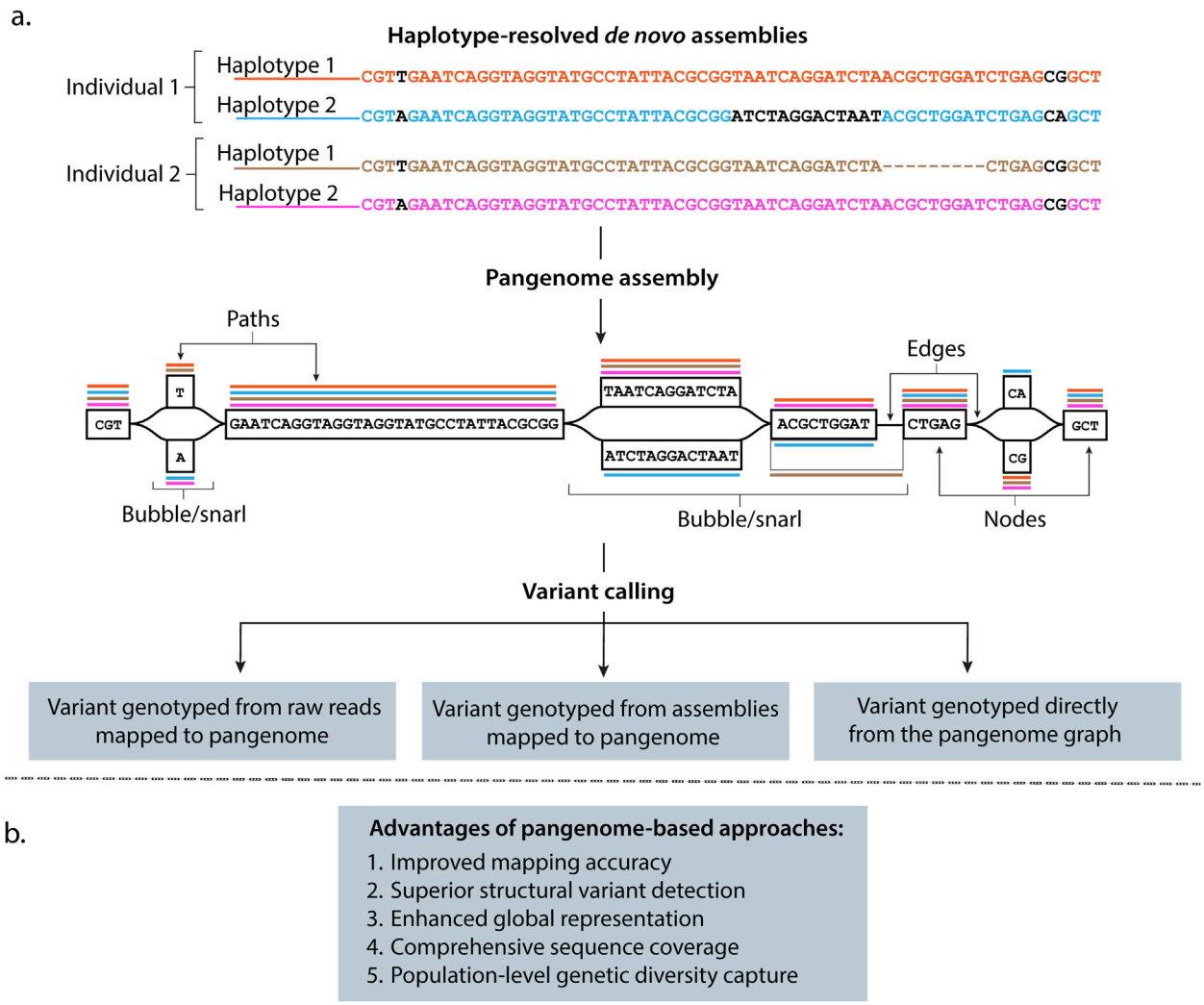

FIGURE 2
Graph-based approaches overcome limitations of linear genome representations while ensuring backward compatibility and preserving analytical continuity. **(a)** High-quality haplotype-resolved *de novo* assemblies provide the foundation for building pangenome frameworks. Pangenome graphs represent assembled genomes from individuals (i.e., Individual 1, Individual 2, . . . , Individual n) as embedded haplotype paths (colored lines) where nodes contain DNA sequences linked by directional connections, with edges illustrating interconnections between DNA sequences. Individual haplotypes follow specific paths (colored lines) through these nodes, while branching regions called "bubbles" or "snarls" capture genetic variations (nucleotides colored in black), including SVs, where genomes differ. For example, the inversion in haplotype 2 from Individual 1 and the deletion in haplotype 1 from Individual 2 are represented in the graph as bubbles or snarls with different paths. This structure preserves genetic diversity from all contributing genomes. Variant detection from pangenomes can be performed by mapping sequencing reads directly to the graph, aligning new assemblies to the pangenome reference, or analyzing variants already represented within the graph structure. **(b)** Advantages of using pangenome-based approaches instead of linear references in genomic analysis.

potentially widening the genomic equity gap between well-resourced and under-resourced communities.

Pangenome construction relies on generating high-quality, haplotype-resolved genome assemblies. Long-read sequencing technologies have enabled the generation of these haplotype-resolved genomes, where maternal and paternal chromosome segments are distinctly identified. These precisely assembled and haplotype-resolved genomes can be organized into graph-based data structures, using two dominant computational approaches: 1) sequence graphs (Garrison et al., 2018; Hickey et al., 2020; Hickey et al., 2023; Sirén et al., 2021; Groza et al., 2024) (e.g., minigraph (Li et al., 2020); and 2) de Bruijn graphs (Iqbal et al., 2012; Minkin et al., 2017; Holley and Melsted, 2020) (Box 1), that efficiently compress and index the sequence information while maintaining an intuitive coordinate system for genetic variant identification (Garrison et al., 2018; Hickey et al., 2020; Sirén et al., 2021; Liao et al., 2023).

However, representing complex structural variants in pangenomes remains challenging. While efforts exist to unify coordinates using the reference graphical fragment assembly (rGFA) format, which provides a stable coordinate system indicating the origin of segments from linear genomes, no straightforward method exists for representing complex SVs in VCF files (Li et al., 2020). This is particularly problematic for nested variants (i.e., bubbles within bubbles) or variants occurring only on alternative haplotypes (Li et al., 2020). This

complex problem is compounded by the requirement for a linear reference backbone, which reintroduces the very biases that pangenomes are designed to eliminate. Despite these challenges, the incorporation of haplotype-resolved genome assemblies into pangenomes has improved upon linear references, while the stable coordinate system ensures backward compatibility, thereby preserving analytical continuity (Garrison et al., 2018; Sirén et al., 2021; Wang et al., 2022; Liao et al., 2023; Secomandi et al., 2025).

Box 1 Sequence and de Bruijn graphs: a quick summary

Sequence graphs have clear advantages over de Bruijn graphs for clinical application due to their stable coordinate systems and support for complex SVs. This enables precise backwardly compatible connections between graph structures and biological features for accurate variant identification (Li et al., 2020; Andreace et al., 2023; Chin et al., 2023; Hickey et al., 2023; Groza et al., 2024).

Sequence graphs:

- Nodes represent variable length DNA sequences (or reverse complement, depending on traversal direction). Edges represent interconnections between DNA sequences (Figure 2a).
- "Bubbles" or "snarls" are defined as divergent paths in the graph where sequences from different individuals branch apart and then reconverge, thereby connecting common head and tail nodes and representing genetic variations (Figure 2a) (Onodera et al., 2013; Paten et al., 2018; Eizenga et al., 2020; Dabbaghie et al., 2022; Secomandi et al., 2025).
- Variation graphs consist of all possible sequences from a population and embed sample haplotype sequences as navigable paths.
- Key advantage: provides a stable coordinate system that remains consistent regardless of methodology of the graph construction (Li et al., 2020; Hickey et al., 2023; Hickey et al., 2024; Sirén et al., 2024).
- Limitation: scalability is a significant limitation for sequence/variation graphs.
- Enable precise alignment, annotation, and comparative analysis across variation graphs and linear reference genomes (Figure 2b).

De Bruijn graphs:

- Nodes represent fixed-length sequences (*k*-mers) while edges represent interconnections between DNA sequences (Iqbal et al., 2012; Andreace et al., 2023).
- Colored de Bruijn graphs enhance pangenome analysis by assigning sample- or population-specific identifiers, enabling efficient population-scale genomic studies (Iqbal et al., 2012; Holley and Melsted, 2020).
- Limitations: these graphs struggle to resolve repetitive genomic regions due to their reliance on fixed *k*-mer lengths (Bankevich et al., 2022), and cannot be effectively built from noisy sequencing reads (Nie et al., 2024)
- Computational efficiency is improved through using compact data structures and the Burrows-Wheeler transform (BWT) (Baier et al., 2016).
- Challenges persist in maintaining connections between graph structures and the original sequence coordinates, which is essential for reference pangenome applications (Li et al., 2020; Hickey et al., 2023).

Early attempts to build a human pangenome identified novel sequences absent from the standard reference genome assembly. For example, the African Pangenome Project uncovered >290 Mb of novel contigs from 910 individuals of African descent (Sherman et al., 2019) Similarly, the HUPAN initiative, which constructed the first Chinese pangenome from 275 individuals, revealed 29.5 Mb of population-specific novel sequences (Duan et al., 2019) Li et al. also utilized deep sequencing data from 486 Han Chinese individuals to build a pangenome that contained 276 Mb of sequences absent from the current human reference (Li Q. et al., 2021).

The Human Pangenome Reference Consortium (HPRC) used long-read, phased, diploid assemblies to create a more inclusive reference (Wang et al., 2022; Liao et al., 2023). The first HPRC release contained 47 phased diploid genomes representing 94 haplotypes. This has subsequently expanded to include phased haplotypes from 232 individuals in the most recent release. The genomes that were included in the HPRC pangenome were initially selected from the 1000 Genomes Project (1KGP) to represent diverse ancestries: 24% African, 30% Americas, 18% East Asian, 28% South Asian (https://humanpangenome.org/samples/) (Wang et al., 2022). To improve global representation, the HPRC is expanding beyond these initial samples to incorporate 700 haploid genomes from cohorts that include the BioMe Biobank and African American individuals to maximize diversity and create a truly representative global pangenome reference (Wang et al., 2022). This population-level approach has improved SV detection, identifying an average of over 29,000 SVs per individual compared to fewer than 16,000 SVs when using linear reference (Groza et al., 2024; Schloissnig et al., 2025). Additionally, this approach has enhanced genotyping accuracy and reduced the reference bias inherent in traditional genomic analysis methods (Hickey et al., 2020; Sirén et al., 2021; Sirén et al., 2024; Lee et al., 2022; Liao et al., 2023).

4 Building, manipulating and querying pangenome graphs

While pangenome graphs offer powerful representations of genomic diversity, their practical application depends on our ability to effectively construct, query, and analyze these complex data structures. This involves building graphs from multiple assemblies, extracting specific genomic regions, performing comparative analyses, and integrating functional annotations. However, efficient pangenome graph manipulation requires substantial computational resources and technical expertise (Garrison et al., 2024; Secomandi et al., 2025). To address these challenges, methods have been developed for *de novo* genome assembly, collecting and mapping assemblies into pangenomes, and annotating pangenome graphs (Table 1).

Pangenome graph builders such as PanGenome Graph Builder (pggb), Minigraph-Cactus, and TwoPaCo can deal with mammalian-sized (~3 Gb) assemblies. Minigraph constructs specialized pangenome graphs through iterative sequence alignment to reference templates. Human pangenome projects have utilized these tools at varying scales: Pantools (7 genomes (Sheikhizadeh et al., 2016); Minigraph-Cactus and pggb (94 single chromosomes (Hickey et al., 2023; Liao et al., 2023); TwoPaCo (100 simulated genomes (Minkin et al., 2017); and Minigraph (94 (Li et al., 2020) and 574 (Groza et al., 2024) haplotype-resolved assemblies). Recently, the HPRC released a draft human reference pangenome constructed using pggb and Minigraph-Cactus pipelines (Liao et al., 2023). As more diverse genomes are assembled *de novo*, the resulting pangenome references will progressively capture the full spectrum of human genomic

TABLE 1 Popular open-source tools for *de novo* genome assembly, construction and annotation of pangenomes.

Tool name (Year) *GitHub URL*	Description
Genome assembly	
Canu (2017) https://github.com/marbl/canu	Assembly tool for noisy long reads (i.e., PacBio CLR and ONT reads) into a graphical fragment assembly that can be integrated with complementary phasing and scaffolding methods
Flye (2019) https://github.com/fenderglass/Flye	A de novo assembler for long reads (i.e., PacBio CLR, HiFi and ONT reads) into genomes using repeat graph
GoldRush (2023) https://github.com/bcgsc/goldrush	GoldRush produces a "golden path" of long reads (i.e., PacBio CLR and ONT reads) with ~1 fold coverage, which are then polished and scaffolded into the final assembly
Hifiasm (2021) https://github.com/chhylp123/hifiasm	Constructs haplotype-resolved assemblies from accurate PacBio HiFi reads and ultralong ONT reads
La Jolla Assembler https://github.com/AntonBankevich/LJA	A tool for genome assembly from PacBio HiFi reads based on de Bruijn graphs
MECAT2 (2019) https://github.com/xiaochuanle/MECAT2	An ultra-fast and accurate mapping, error correction and de novo assembly tool for long reads (i.e., PacBio CLR, HiFi reads)
Miniasm (2016) https://github.com/lh3/miniasm	A fast OLC-based de novo assembler for noisy long reads (i.e., ONT reads) into an assembly graph in the GFA format
NECAT (2021) https://github.com/xiaochuanle/NECAT	An error correction and de novo assembly tool for noisy long reads (i.e., ONT reads)
NextDenovo (2024) https://github.com/Nextomics/NextDenovo	A string graph-based de novo assembler for long reads (i.e., PacBio CLR and ONT reads) that uses a "correct-then-assemble"
PECAT (2024) https://github.com/lemene/PECAT	A haplotype-aware correction and assembly tool for long noisy reads (i.e., PacBio CLR and ONT reads)
Peregrine-2021 (2022) https://github.com/cschin/peregrine-2021	A genome assembler designed for long reads that have good enough accuracy (i.e., PacBio CLR and ONT reads)
Raven (2021) https://github.com/lbcb-sci/raven	A de novo genome assembler for long uncorrected reads (i.e., PacBio CLR and ONT reads)
Rust-mdbg (2021) https://github.com/ekimb/rust-mdbg/	An ultra-fast minimizer-space de Bruijn graph implementation, geared towards the assembly of long and accurate PacBio HiFi reads
Shasta (2020) https://github.com/paoloshasta/shasta	A de novo assembler for long reads optimized for ONT reads
SMARTdenovo (2021) https://github.com/ruanjue/smartdenovo	A de novo assembler for long reads (i.e., PacBio CLR and ONT reads) into an assembly from all-vs-all raw read alignments without an error correction
Verkko (2023) https://github.com/marbl/verkko	A hybrid telomere-to-telomere genome assembly pipeline of accurate long reads (PacBio HiFi, ONT Duplex, and HERRO corrected ONT simplex reads) and ONT ultra-long reads
Wtdbg2 (2020) https://github.com/ruanjue/wtdbg2	A de novo sequence assembler for noisy long reads (i.e., PacBio CLR and ONT reads) into FBG
Pangenome construction	
Bifrost (2020) https://github.com/pmelsted/bifrost	Tool for parallel construction, indexing and querying of colored and compacted de Bruijn graphs
MEMO (2025) https://github.com/StephenHwang/MEMO	A pangenome indexing method based on maximal exact matches between genomes
Minigraph (2020) https://github.com/lh3/minigraph	Tool for sequence-to-graph mapping, with incrementally sequence mapping to existing graphs, and variant calling.
Minigraph-Cactus (2023) https://github.com/ComparativeGenomicsToolkit/cactus	A pangenome graph construction toolkit
Pangene (2024) https://github.com/lh3/pangene	Constructs pangenome gene graphs, with nodes representing marker genes and edges between two genes indicating their genomic adjacency on input genomes
Pangenome (2025) https://github.com/nf-core/pangenome	A bioinformatics best-practice analysis pipeline that renders a collection of sequences into a pangenome graph
Pannagram (2025) https://github.com/iganna/pannagram	A tool for constructing pan-genome alignments, analyzing SVs, and translating annotations between genomes
PanPipes (2022) https://github.com/USDA-ARS-GBRU/PanPipes	An end-to-end pipeline for pan-genomic graph construction and genetic analysis
PanTools (2025) https://git.wur.nl/bioinformatics/pantools	A toolkit for building pangenomes from genomes using de Bruijn graphs and constructing pan-proteomes from proteins

(Continued on following page)

TABLE 1 (*Continued*) Popular open-source tools for *de novo* genome assembly, construction and annotation of pangenomes.

Tool name (Year) *GitHub URL*	Description
Pggb (2025) https://github.com/pangenome/pggb	Builds pangenome variation graphs from a set of input sequences
Psvcp (2023) https://github.com/wjian8/psvcp_v1.01	A tool for pangenome construction and population structure variation genotype calling pipeline
Pangenome annotation	
GrAnnot (2025) https://forge.ird.fr/diade/dynadiv/grannot	An annotation transfer tool for pangenome graphs that transfers linear genome annotations to a pangenome graph
PanTools (2025) https://git.wur.nl/bioinformatics/pantools	Constructs and expands the annotation layer of an existing pangenome using genomic features like genes, mRNAs, proteins, tRNAs from GFF files

ONT, Oxford Nanopore Technologies; PacBio - Pacific Biosciences; CLR, continuous long reads; HiFi - high-fidelity; HERRO, Haplotype-aware ERRor cOrrection; FBG, fuzzy Bruijn graph; GFA, graphical fragment assembly text format describing a set of sequences and their overlap; OLC, Overlap-Layout-Consensus paradigm; GFF, General Feature Format; tRNA, transfer RNA; mRNA, messenger RNA; Year - represents year of publication or the year of the latest version available on GitHub.

TABLE 2 A list of popular open-source tools for pangenome-based variant genotyping.

Tool name (Year) *GitHub URL*	Description
Ctyper (2024) https://github.com/ChaissonLab/Ctyper	A pangenome allele-specific and copy number specific genotyping tool
DeepVariant (2025) https://github.com/google/deepvariant	A deep learning-based variant caller for alignments (BAM or CRAM) and pangenome graphs
Graphtyper2 (2019) https://github.com/DecodeGenetics/graphtyper	A graph-based variant caller capable of genotyping population-scale short read data sets
Minigraph (2020) https://github.com/lh3/minigraph	Tool for sequence-to-graph mapping, with incrementally sequence mapping to existing graphs, and variant calling
Minigraph-Cactus (2023) https://github.com/ComparativeGenomicsToolkit/cactus	A graph construction and variant genotyping toolkit
PanGenie (2024) https://github.com/eblerjana/pangenie	A short-read genotyper for SNPs, indels and SVs represented in a pangenome graph
Paragraph (2019) https://github.com/Illumina/paragraph	A graph-based structural variant genotyping tool for short-read sequence data
PHI (2024) https://github.com/at-cg/PHI	A pangenome-based genotyping method from low-coverage sequencing data (short-reads or long-reads)
SVarp (2024) https://github.com/asylvz/SVarp	A tool to discover haplotype resolved SVs on top of a pangenome graph reference using long sequencing reads
Varigraph (2025) https://github.com/JiaoLab2021/varigraph	A pangenome graph-based variant genotyper for diploid and polyploid genomes
Vg (2020) https://github.com/vgteam/vg	A pangenome-based SV genotyping tool

BAM, Binary Alignment Map of genome sequencing data; CRAM, Compressed Reference-oriented Alignment Map of genome sequencing data; Year - represents year of publication or the year of the latest version available on GitHub.

diversity, ultimately enhancing our ability to detect and interpret rare and clinically relevant variants in precision medicine applications.

4.1 Advantages of genome graphs for variant calling

Traditional variant calling relies on aligning reads to a single reference genome (i.e., GRCh37, GRCh38, or more recently, T2T-CHM13). However, linear reference approaches struggle with regions where individuals differ substantially from the reference, including CNVs, SVs, large indels, and highly polymorphic regions (e.g., killer immunoglobulin-like receptor [KIR] and human leucocyte antigen [HLA] loci) (Kulski et al., 2022; Olson et al., 2023). To overcome these challenges, researchers have developed methods to map reads to pangenome references using graph-based structures that incorporate variants from diverse individuals as alternative paths, with alignments often converted back to linear references for compatibility with conventional variant-calling tools (Table 2).

Recent advances by the HPRC have significantly improved variant detection through graph-based references constructed from long-read high-fidelity (HiFi) reads that provide per-base accuracy of 99.9%. Specialized tools further enhance variant calling (e.g., Giraffe-DeepVariant) and variant genotyping (e.g., PanGenie), particularly for large indels, SVs and variations in highly polymorphic regions previously problematic in GRCh38 (Table 2) (Li et al., 2020; Sirén et al., 2021; Ebler et al., 2022). The Minigraph-Cactus pangenome pipeline represents a significant computational advancement by combining fast assembly-to-graph mapping with an improved base aligner and including all SNVs and small indels in the pangenome (Hickey et al., 2023). This approach constructs nucleotide-resolution pangenome graphs through a two-stage process: first extracting SVs from each of hundreds of haplotype-resolved assemblies, then using these variants as

alignment anchors to generate a comprehensive base-level graph. The resulting graph represents variation at all resolutions (i.e., from SNVs to complex SVs such as inversions) (Hickey et al., 2023).

Minigraph-generated pangenome graphs have improved short-read and long-read mapping, variant calling, and SV genotyping (Hickey et al., 2023). For example, minigraph has recently been employed to identify 200,000 unique SVs from a pangenome graph of 574 assemblies, outperforming standard methods (Groza et al., 2024). However, despite multiple tools promising perfect recall for complex SVs, few clinical labs have validated these graph-based callers on patient cohorts. Therefore, randomized trials testing graph genotypers are required to determine if they improve the detection of clinically relevant indels and variants.

4.2 Clinical impact and research applications of pangenome graphs

Pangenomes represent a paradigm change in the conceptualization and analysis of human genetic diversity (Eggertsson et al., 2019; Hickey et al., 2020; Sirén et al., 2021; Ebler et al., 2022; Lee et al., 2022; Tetikol et al., 2022; Liao et al., 2023; Groza et al., 2024; Secomandi et al., 2025). The potential impact of the use of pangenomes is particularly notable in applications that incorporate accurate haplotype reconstruction into the diagnosis of rare disorders (Abondio et al., 2024; Groza et al., 2024) and complex SV interpretation (Hickey et al., 2020; Ebler et al., 2022; Lee et al., 2022; Tetikol et al., 2022; Groza et al., 2024). However, simply aggregating more genomes will not solve the fundamental problem of missing diversity, as bigger is not always better. Clinically, strategic sampling and generation of pangenomes from related individuals (e.g., trios–mother, father, child, or siblings) may yield more clinically actionable variants per unit (e.g., terabase) of sequencing, particularly in genomic regions that are poorly captured by standard linear reference genomes. In addition, by moving away from short-read sequencing, which suffers from a limited ability to resolve complex SVs and repetitive regions, pangenome efforts will improve the clinical utility of these genomic features (Abondio et al., 2024; Groza et al., 2024).

4.2.1 Accurate haplotype reconstruction

Many genome assembly methods collapse heterozygous alleles that are present in diploid organisms, erasing heterozygous variation and potential misrepresentation in regions of significant haplotype divergence (Dilthey et al., 2015; Li et al., 2020; Ebler et al., 2022; Chin et al., 2023; Matthews et al., 2024; Secomandi et al., 2025). This limitation is particularly problematic in genomic regions with high sequence diversity or complex SVs, such as the major histocompatibility complex (MHC) region on chromosome 6 that encodes the classical human leucocyte antigen alleles (Dilthey et al., 2015; Li et al., 2020).

However, a diversely sampled pangenome graph for the highly complex MHC region allows inference of haplotypes using only short-read sequencing data even in regions that were previously difficult to characterize accurately (Dilthey et al., 2015). Notably, approximately 1% of the human genome is poorly represented by linear references, including gene-dense loci containing the olfactory receptors and ubiquitin-specific peptidases (Dilthey et al., 2015). As such, incorporating alternative sequence variants through graph-based models significantly enhances our ability to reconstruct accurate genomic representations. Thus, the development of population reference graphs across the MHC locus highlights the broader potential impact of graph-based methods in regions of high sequence diversity.

4.2.2 Precise detection and phasing of genetic variants

The success of pangenome-based variant calling, however, depends critically on both variant characteristics and sequencing technology, with effectiveness varying significantly by variant size when using short-read sequencing data (Eizenga et al., 2020). For SNVs and small indels, pangenomes offer modest improvements in detection accuracy (Eizenga et al., 2020; Li et al., 2020; Hickey et al., 2023; Hickey et al., 2024; Secomandi et al., 2025). However, pangenome graphs demonstrate substantial improvement in genotyping SVs (e.g., over 10% increase in number of SVs detected), addressing a key limitation of traditional approaches (Hickey et al., 2020; 2024; Li et al., 2020; Sirén et al., 2021). This improvement stems from a fundamental technical constraint for short-read sequencing. Specifically, short reads can encompass small variants entirely but fail to span larger structural changes (Hickey et al., 2020; Hickey et al., 2024; Sirén et al., 2021). By contrast, pangenome graphs incorporate SVs into their framework, significantly improving variant detection, even from short-read data, compared to the currently used single-reference methods (Ebler et al., 2022; Groza et al., 2024). For example, graphs built from haplotype-resolved assemblies can harness short-read k-mer patterns to identify previously undetectable SVs (Groza et al., 2024). Additionally, some pangenome graph representations (e.g., de Bruijn) are capable of SV detection without requiring a reference genome of any type, offering a flexible alternative for variant discovery (Iqbal et al., 2012).

Pangenomes have proven valuable for population-scale variant detection (Tetikol et al., 2022; Hickey et al., 2024). For example, the pan-African (Tetikol et al., 2022) and the Chinese pangenomes (Gao et al., 2023) have substantially improved variant detection accuracy compared to traditional linear reference approaches. The effectiveness of these graphs is influenced by two key factors: nucleotide diversity within populations and the level of absolute divergence from linear reference sequences (Tetikol et al., 2022; Hickey et al., 2024). This is particularly relevant for highly diverse populations like Africans (i.e., >290 Mb of novel contigs), or groups with significant archaic admixture, such as some individuals from Australo-Melanesian populations, who may retain Denisovan haplotypes over 250 kb in length (Jacobs et al., 2019; Sherman et al., 2019; Tetikol et al., 2022).

Population-specific graphs that incorporate cohort-specific information enable the identification of functionally important variants within coding regions that are missed by standard variant calling pipelines (Ebler et al., 2022; Tetikol et al., 2022; Gao et al., 2023; Groza et al., 2024; Hickey et al., 2024). Notably, these graphs provide improvements in sensitivity and specificity typically achieved by calling variants jointly from cohorts, but without requiring simultaneous processing of all cohort samples. Thus, they represent a computationally efficient solution for large-scale genomic studies (Eggertsson et al., 2019; Ebler et al., 2022;

Tetikol et al., 2022; Gao et al., 2023; Groza et al., 2024; Hickey et al., 2024; Wu et al., 2024).

4.2.3 Exemplar application of pangenome graphs to rare disease diagnosis

Recent application of pangenome graph approaches has demonstrated their promise in rare disease diagnosis (Chin et al., 2023; Gao et al., 2023; Groza et al., 2024). Groza et al. established a practical framework for clinical implementation (Groza et al., 2024). Their graph-based analysis of 574 rare disease cases identified >200,000 unique and >500,000 shared SVs and ~1,000 rare (MAF <0.01) coding variants (Groza et al., 2024). The pangenome approach proved particularly useful in complex genomic regions where traditional methods fail, enabling the identification of a previously undetectable diagnostic variant in *KMT2E* associated with macrocephaly, hypotonia, and developmental delay. These results highlight the potential of pangenome graphs to enhance diagnostic yields through improved variant detection and prioritization of candidate SVs, while providing a scalable resource for the rare disease community (Groza et al., 2024).

5 Discussion and future perspectives

The transition from linear reference genomes to pangenome graphs represents a transformative paradigm shift in how we conceptualize and analyze human genetic diversity, addressing fundamental limitations in variant detection and population representation. Through initiatives like the HPRC and population-specific pangenome projects, we can access sophisticated graph-based frameworks that capture hundreds of megabases of previously missing genetic diversity. Tools such as Minigraph-Cactus (Hickey et al., 2023) and specialized variant callers have demonstrated remarkable improvements in structural variant detection, accurate haplotype reconstruction, and genotyping accuracy across diverse populations, improving diagnostic accuracy for rare genetic disorders and reducing reference bias in underrepresented populations. However, significant challenges remain in computational scalability, clinical validation, interpretability, and user accessibility.

Realizing the transformative potential of pangenomes requires several key advances, which are currently being developed and researched around the world. Specifically: 1) strategic selection of samples to ensure genetic diversity by prioritizing quality over quantity in genome selection, as demonstrated by the HPRC initiative (Wang et al., 2022); 2) leveraging long-read sequencing technologies and tools that facilitate T2T genome assembly, as demonstrated by (Rautiainen et al., 2023), enabling haplotypes to be added to the pangenome graphs; 3) developing efficient algorithms for building, genotyping and annotating genetic variants from pangenome graphs, with tools such as Minigraph-Cactus (Hickey et al., 2023) being continuously improved for computational efficiency and GrAnnot (Marthe et al., 2025) ensuring annotation of sequences within these graphs; 4) conducting rigorous validation for variant detection in clinical cohorts, such as benchmarking variant calling using standards set by the Global Alliance for Genomics and Health (GA4GH (Krusche et al., 2019)) and including Genome in a Bottle (GIAB (Zook et al., 2014)) samples in pangenomes; and 5) creation of simplified analytical workflows that create equitable detection of clinically relevant variants for routine genomic medicine. Additionally, the development of splice-aware population-level RNA sequencing analysis algorithms has enabled precise quantification of haplotype-specific transcript expression (Sibbesen et al., 2023; Secomandi et al., 2025). The impacts of this extend beyond transcriptomics, specifically providing deeper insights into the relationship between genetic variation and biological function in rare diseases (Grytten et al., 2019; Wang et al., 2024).

In conclusion, graph-based approaches represent a transformative shift towards truly equitable precision medicine that delivers accurate, clinically actionable insights across all populations regardless of their ancestral background or genetic diversity.

Author contributions

DN: Conceptualization, Visualization, Writing – original draft, Writing – review and editing. RZ: Writing – review and editing. OS: Writing – review and editing. MB: Writing – review and editing. JO: Conceptualization, Writing – review and editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This project was supported by donations from the Dines Family Trust, The Tautoku Trust, and The Kelliher Trust.

Acknowledgments

We gratefully acknowledge the generous donations from the Dines Family Trust, The Toutoku Trust, and The Kelliher Trust.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abascal, F., Acosta, R., Addleman, N. J., Adrian, J., Afzal, V., Aken, B., et al. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* 583, 699–710. doi:10.1038/s41586-020-2493-4

Abondio, P., Bruno, F., Passarino, G., Montesanto, A., and Luiselli, D. (2024). Pangenomics: a new era in the field of neurodegenerative diseases. *Ageing Res. Rev.* 94, 102180. doi:10.1016/j.arr.2023.102180

Aganezov, S., Yan, S. M., Soto, D. C., Kirsche, M., Zarate, S., Avdeyev, P., et al. (2022). A complete reference genome improves analysis of human genetic variation. *Science* 1979, eabl3533. doi:10.1126/science.abl3533

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

Baier, U., Beller, T., and Ohlebusch, E. (2016). Graphical pan-genome analysis with compressed suffix trees and the burrows-wheeler transform. *Bioinformatics* 32, 497–504. doi:10.1093/bioinformatics/btv603

Ballouz, S., Dobin, A., and Gillis, J. A. (2019). Is it time to change the reference genome? *Genome Biol.* 20, 159. doi:10.1186/s13059-019-1774-4

Bankevich, A., Bzikadze, A. V., Kolmogorov, M., Antipov, D., and Pevzner, P. A. (2022). Multiplex de Bruijn graphs enable genome assembly from long, high-fidelity reads. *Nat. Biotechnol.* 40, 1075–1081. doi:10.1038/s41587-022-01220-6

Cai, R., Dong, Y., Fang, M., Guo, C., and Ma, X. (2020). *De novo* genome assembly of a Han Chinese male and genome-wide detection of structural variants using Oxford Nanopore sequencing. *Mol. Genet. Genomics* 295, 871–876. doi:10.1007/s00438-020-01672-y

Chin, C. S., Behera, S., Khalak, A., Sedlazeck, F. J., Sudmant, P. H., Wagner, J., et al. (2023). Multiscale analysis of pangenomes enables improved representation of genomic diversity for repetitive and clinically relevant genes. *Nat. Methods* 20, 1213–1221. doi:10.1038/s41592-023-01914-y

Church, D. M., Schneider, V. A., Graves, T., Auger, K., Cunningham, F., Bouk, N., et al. (2011). Modernizing reference genome assemblies. *PLoS Biol.* 9, e1001091. doi:10.1371/journal.pbio.1001091

Collins, R. L., Brand, H., Karczewski, K. J., Zhao, X., Alföldi, J., Francioli, L. C., et al. (2020). A structural variation reference for medical and population genetics. *Nature* 581, 444–451. doi:10.1038/s41586-020-2287-8

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

Dawood, M., Fayer, S., Pendyala, S., Post, M., Kalra, D., Patterson, K., et al. (2024). Using multiplexed functional data to reduce variant classification inequities in underrepresented populations. *Genome Med.* 16, 143. doi:10.1186/s13073-024-01392-7

Dilthey, A., Cox, C., Iqbal, Z., Nelson, M. R., and McVean, G. (2015). Improved genome inference in the MHC using a population reference graph. *Nat. Genet.* 47, 682–688. doi:10.1038/ng.3257

Duan, Z., Qiao, Y., Lu, J., Lu, H., Zhang, W., Yan, F., et al. (2019). HUPAN: a pan-genome analysis pipeline for human genomes. *Genome Biol.* 20, 149. doi:10.1186/s13059-019-1751-y

Ebler, J., Ebert, P., Clarke, W. E., Rausch, T., Audano, P. A., Houwaart, T., et al. (2022). Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Nat. Genet.* 54, 518–525. doi:10.1038/s41588-022-01043-w

Eggertsson, H. P., Kristmundsdottir, S., Beyter, D., Jonsson, H., Skuladottir, A., Hardarson, M. T., et al. (2019). GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nat. Commun.* 10, 5402. doi:10.1038/s41467-019-13341-9

Eizenga, J. M., Novak, A. M., Sibbesen, J. A., Heumos, S., Ghaffaari, A., Hickey, G., et al. (2020). Pangenome graphs. *Annu. Rev. Genomics Hum. Genet.* 21, 139–162. doi:10.1146/annurev-genom-120219-080406

Gao, Y., Yang, X., Chen, H., Tan, X., Yang, Z., Deng, L., et al. (2023). A pangenome reference of 36 Chinese populations. *Nature* 619, 112–121. doi:10.1038/s41586-023-06173-7

Garrison, E., Sirén, J., Novak, A. M., Hickey, G., Eizenga, J. M., Dawson, E. T., et al. (2018). Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotechnol.* 36, 875–879. doi:10.1038/nbt.4227

Garrison, E., Guarracino, A., Heumos, S., Villani, F., Bao, Z., Tattini, L., et al. (2024). Building pangenome graphs. *Nat. Methods* 21, 2008–2012. doi:10.1038/s41592-024-02430-3

Gong, Y., Li, Y., Liu, X., Ma, Y., and Jiang, L. (2023). A review of the pangenome: how it affects our understanding of genomic variation, selection and breeding in domestic animals? *J. Anim. Sci. Biotechnol.* 14, 73. doi:10.1186/s40104-023-00860-1

Green, S., Prainsack, B., and Sabatello, M. (2023). Precision medicine and the problem of structural injustice. *Med. Health Care Philos.* 26, 433–450. doi:10.1007/s11019-023-10158-8

Groza, C., Schwendinger-Schreck, C., Cheung, W. A., Farrow, E. G., Thiffault, I., Lake, J., et al. (2024). Pangenome graphs improve the analysis of structural variants in rare genetic diseases. *Nat. Commun.* 15, 657. doi:10.1038/s41467-024-44980-2

Grytten, I., Rand, K. D., Nederbragt, A. J., Storvik, G. O., Glad, I. K., and Sandve, G. K. (2019). Graph peak caller: calling chip-seq peaks on graph-based reference genomes. *PLoS Comput. Biol.* 15, e1006731. doi:10.1371/journal.pcbi.1006731

GTEx Consortium (2020). The GTEx consortium atlas of genetic regulatory effects across human tissues. *Science* 369, 1318–1330. doi:10.1126/science.aaz1776

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

Gustafson, J. A., Gibson, S. B., Damaraju, N., Zalusky, M. P. G., Hoekzema, K., Twesigomwe, D., et al. (2024). High-coverage nanopore sequencing of samples from the 1000 genomes project to build a comprehensive catalog of human genetic variation. *Genome Res.* 34, 2061–2073. doi:10.1101/gr.279273.124

Hickey, G., Heller, D., Monlong, J., Sibbesen, J. A., Sirén, J., Eizenga, J., et al. (2020). Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biol.* 21, 35. doi:10.1186/s13059-020-1941-7

Hickey, G., Monlong, J., Ebler, J., Novak, A. M., Eizenga, J. M., Gao, Y., et al. (2023). Pangenome graph construction from genome alignments with minigraph-cactus. *Nat. Biotechnol.* 42, 663–673. doi:10.1038/s41587-023-01793-w

Hickey, G., Monlong, J., Ebler, J., Novak, A. M., Eizenga, J. M., Gao, Y., et al. (2024). Combining reference genomes into a pangenome graph improves accuracy and reduces bias. *Nat. Biotechnol.* 42, 580–581. doi:10.1038/s41587-023-01828-2

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

Iqbal, Z., Caccamo, M., Turner, I., Flicek, P., and McVean, G. (2012). *De novo* assembly and genotyping of variants using colored de Bruijn graphs. *Nat. Genet.* 44, 226–232. doi:10.1038/ng.1028

Jacobs, G. S., Hudjashov, G., Saag, L., Kusuma, P., Darusallam, C. C., Lawson, D. J., et al. (2019). Multiple deeply divergent denisovan ancestries in papuans. *Cell* 177, 1010–1021. doi:10.1016/j.cell.2019.02.035

Jeong, H., Dishuck, P. C., Yoo, D. A., Harvey, W. T., Munson, K. M., Lewis, A. P., et al. (2025). Structural polymorphism and diversity of human segmental duplications. *Nat. Genet.* 57, 390–401. doi:10.1038/s41588-024-02051-8

Jia, T., Munson, B., Lango Allen, H., Ideker, T., and Majithia, A. R. (2020). Thousands of missing variants in the UK biobank are recoverable by genome realignment. *Ann. Hum. Genet.* 84, 214–220. doi:10.1111/ahg.12383

Krusche, P., Trigg, L., Boutros, P. C., Mason, C. E., De La Vega, F. M., Moore, B. L., et al. (2019). Best practices for benchmarking germline small-variant calls in human genomes. *Nat. Biotechnol.* 37, 555–560. doi:10.1038/s41587-019-0054-x

Kulski, J. K., Suzuki, S., and Shiina, T. (2022). Human leukocyte antigen super-locus: nexus of genomic supergenes, SNPs, indels, transcripts, and haplotypes. Hum. Genome Var. 9 9, 49. doi:10.1038/s41439-022-00226-5

Lee, H., Greer, S. U., Pavlichin, D. S., Hughes, C. R., Zhou, B., Weissman, T., et al. (2022). *The human Pangenome's sequence conservation reveals a landscape of polymorphic structural variations*. doi:10.1101/2022.10.06.511239

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

Li, H., Dawood, M., Khayat, M. M., Farek, J. R., Jhangiani, S. N., Khan, Z. M., et al. (2021). Exome variant discrepancies due to reference-genome differences. *Am. J. Hum. Genet.* 108, 1239–1250. doi:10.1016/j.ajhg.2021.05.011

Li, Q., Tian, S., Yan, B., Liu, C. M., Lam, T. W., Li, R., et al. (2021). Building a Chinese pan-genome of 486 individuals. *Commun. Biol.* 4, 1016. doi:10.1038/s42003-021-02556-6

Liao, W. W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., et al. (2023). A draft human pangenome reference. *Nature* 617, 312–324. doi:10.1038/s41586-023-05896-x

Logsdon, G. A., Ebert, P., Audano, P. A., Loftus, M., Porubsky, D., Ebler, J., et al. (2025). Complex genetic variation in nearly complete human genomes. *Nature* 644, 430–441. doi:10.1038/s41586-025-09140-6

Lowy-Gallego, E., Fairley, S., Zheng-Bradley, X., Ruffier, M., Clarke, L., Flicek, P., et al. (2019). Variant calling on the grch38 assembly with the data from phase three of the 1000 genomes project [version 2; peer review: 1 approved, 1 not approved]. *Wellcome Open Res.* 4, 50. doi:10.12688/wellcomeopenres.15126.2

Lunke, S., Bouffler, S. E., Patel, C. V., Sandaradura, S. A., Wilson, M., Pinner, J., et al. (2023). Integrated multi-omics for rapid rare disease diagnosis on a national scale. *Nat. Med.* 29, 1681–1691. doi:10.1038/s41591-023-02401-9

Mahmoud, M., Huang, Y., Garimella, K., Audano, P. A., Wan, W., Prasad, N., et al. (2024). Utility of long-read sequencing for all of us. *Nat. Commun.* 15, 837. doi:10.1038/s41467-024-44804-3

Marschall, T., Marz, M., Abeel, T., Dijkstra, L., Dutilh, B. E., Ghaffaari, A., et al. (2018). Computational pan-genomics: status, promises and challenges. *Brief. Bioinform* 19, 118–135. doi:10.1093/bib/bbw089

Marthe, N., Zytnicki, M., and Sabot, F. (2025). *GrAnnoT, a tool for efficient and reliable annotation transfer through pangenome graph*. doi:10.1101/2025.02.26.640337

Matalon, D. R., Zepeda-Mendoza, C. J., Aarabi, M., Brown, K., Fullerton, S. M., Kaur, S., et al. (2023). Clinical, technical, and environmental biases influencing equitable access to clinical genetics/genomics testing: a points to consider statement of the American college of medical genetics and genomics (ACMG). *Genet. Med.* 25, 100812. doi:10.1016/j.gim.2023.100812

Matthews, C. A., Watson-Haigh, N. S., Burton, R. A., and Sheppard, A. E. (2024). A gentle introduction to pangenomics. *Brief. Bioinform* 25, bbae588. doi:10.1093/bib/bbae588

Minkin, I., Pham, S., and Medvedev, P. (2017). TwoPaCo: an efficient algorithm to build the compacted de Bruijn graph from many complete genomes. *Bioinformatics* 33, 4024–4032. doi:10.1093/bioinformatics/btw609

Nassar, L. R., Barber, G. P., Benet-Pagès, A., Casper, J., Clawson, H., Diekhans, M., et al. (2023). The UCSC genome browser database: 2023 update. *Nucleic Acids Res.* 51, D1188–D1195. doi:10.1093/nar/gkac1072

Nie, F., Ni, P., Huang, N., Zhang, J., Wang, Z., Xiao, C., et al. (2024). *De novo* diploid genome assembly using long noisy reads. *Nat. Commun.* 15, 2964. doi:10.1038/s41467-024-47349-7

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., et al. (2022). The complete sequence of a human genome. *Science* 376, 44–53. doi:10.1126/science.abj6987

Nyaga, D. M., Tsai, P., Gebbie, C., Phua, H. H., Yap, P., Le Quesne Stabej, P., et al. (2024). Benchmarking nanopore sequencing and rapid genomics feasibility: validation at a quaternary hospital in New Zealand. *NPJ Genom Med.* 9, 57. doi:10.1038/s41525-024-00445-5

Olson, N. D., Wagner, J., Dwarshuis, N., Miga, K. H., Sedlazeck, F. J., Salit, M., et al. (2023). Variant calling and benchmarking in an era of complete human genome sequences. *Nat. Rev. Genet.* 24, 464–483. doi:10.1038/s41576-023-00590-0

Onodera, T., Sadakane, K., and Shibuya, T. (2013). *Detecting superbubbles in assembly graphs*. 338–348. doi:10.1007/978-3-642-40453-5_26

Pan, B., Kusko, R., Xiao, W., Zheng, Y., Liu, Z., Xiao, C., et al. (2019). Similarities and differences between variants called with human reference genome HG19 or HG38. *BMC Bioinforma.* 20, 101. doi:10.1186/s12859-019-2620-0

Paten, B., Eizenga, J. M., Rosen, Y. M., Novak, A. M., Garrison, E., Hickey, G., et al. (2018). Superbubbles, ultrabubbles, and cacti. *J. Comput. Biol.*, 649–663. doi:10.1089/cmb.2017.0251

Rautiainen, M., Nurk, S., Walenz, B. P., Logsdon, G. A., Porubsky, D., Rhie, A., et al. (2023). Telomere-to-telomere assembly of diploid chromosomes with verkko. *Nat. Biotechnol.* 41, 1474–1482. doi:10.1038/s41587-023-01662-6

Reis, A. L. M., Rapadas, M., Hammond, J. M., Gamaarachchi, H., Stevanovski, I., Ayuputeri Kumaheri, M., et al. (2023). The landscape of genomic structural variation in Indigenous Australians. *Nature* 624, 602–610. doi:10.1038/s41586-023-06842-7

Schloissnig, S., Pani, S., Ebler, J., Hain, C., Tsapalou, V., Söylev, A., et al. (2025). Structural variation in 1,019 diverse humans based on long-read sequencing. *Nature* 644, 442–452. doi:10.1038/s41586-025-09290-7

Schneider, V. A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H. C., Kitts, P. A., et al. (2017). Evaluation of GRCh38 and *de novo* haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* 27, 849–864. doi:10.1101/gr.213611.116

Secomandi, S., Gallo, G. R., Rossi, R., Rodríguez Fernandes, C., Jarvis, E. D., Bonisoli-Alquati, A., et al. (2025). Pangenome graphs and their applications in biodiversity genomics. *Nat. Genet.* 57, 13–26. doi:10.1038/s41588-024-02029-6

Sheikhizadeh, S., Schranz, M. E., Akdel, M., De Ridder, D., and Smit, S. (2016). "PanTools: representation, storage and exploration of pan-genomic data," in *Bioinformatics* (Oxford University Press), i487–i493. doi:10.1093/bioinformatics/btw455

Sherman, R. M., Forman, J., Antonescu, V., Puiu, D., Daya, M., Rafaels, N., et al. (2019). Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat. Genet.* 51, 30–35. doi:10.1038/s41588-018-0273-y

Sibbesen, J. A., Eizenga, J. M., Novak, A. M., Sirén, J., Chang, X., Garrison, E., et al. (2023). Haplotype-aware pantranscriptome analyses using spliced pangenome graphs. *Nat. Methods* 20, 239–247. doi:10.1038/s41592-022-01731-9

Sinha, S., Rabea, F., Ramaswamy, S., Chekroun, I., El Naofal, M., Jain, R., et al. (2025). Long read sequencing enhances pathogenic and novel variation discovery in patients with rare diseases. *Nat. Commun.* 16, 2500. doi:10.1038/s41467-025-57695-9

Sirén, J., Monlong, J., Chang, X., Novak, A. M., Eizenga, J. M., Markello, C., et al. (2021). Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science* 1979, abg8871. doi:10.1126/science.abg8871

Sirén, J., Eskandar, P., Ungaro, M. T., Hickey, G., Eizenga, J. M., Novak, A. M., et al. (2024). Personalized pangenome references. *Nat. Methods* 21, 2017–2023. doi:10.1038/s41592-024-02407-2

Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., et al. (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed program. *Nature* 590, 290–299. doi:10.1038/s41586-021-03205-y

Taylor, D. J., Eizenga, J. M., Li, Q., Das, A., Jenike, K. M., Kenny, E. E., et al. (2024). Beyond the human genome project: the age of complete human genome sequences and pangenome references. *Annu. Rev. Genomics Hum. Genet.* 25, 77–104. doi:10.1146/annurev-genom-021623-081639

Tetikol, H. S., Turgut, D., Narci, K., Budak, G., Kalay, O., Arslan, E., et al. (2022). Pan-african genome demonstrates how population-specific genome graphs improve high-throughput sequencing data analysis. *Nat. Commun.* 13, 4384. doi:10.1038/s41467-022-31724-3

Wang, T., Antonacci-Fulton, L., Howe, K., Lawson, H. A., Lucas, J. K., Phillippy, A. M., et al. (2022). The human pangenome project: a global resource to map genomic diversity. *Nature* 604, 437–446. doi:10.1038/s41586-022-04601-8

Wang, D., Bouwmeester, R., Zheng, P., Dai, C., Sanchez, A., Shu, K., et al. (2024). Proteogenomics analysis of human tissues using pangenomes. *Biorxiv*. doi:10.1101/2024.05.24.595489

Wu, Z., Li, T., Jiang, Z., Zheng, J., Gu, Y., Liu, Y., et al. (2024). Human pangenome analysis of sequences missing from the reference genome reveals their widespread evolutionary, phenotypic, and functional roles. *Nucleic Acids Res.* 52, 2212–2230. doi:10.1093/nar/gkae086

Zook, J. M., Chapman, B., Wang, J., Mittelman, D., Hofmann, O., Hide, W., et al. (2014). Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat. Biotechnol.* 32, 246–251. doi:10.1038/nbt.2835