REVIEW

How to Cite: *Aitzhanova A, Bektayev R, Daniyarov A, Karimov A, Makhmetov S, Ordabayeva A, Temirkhanova K (2026) Methods and Bioinformatics Tools for Genomic Analysis. Contig. 1(4):202622. DOI: 10.66273/3134-6359.2026.1.4.004*

Received: 20/10/2026 | Accepted: 17/11/2026 | Published online: 18/11/2026

Methods and Bioinformatics Tools for Genomic Analysis

Aizhamal Aitzhanova[1,†,*], Rakhimbek Bektayev[2,†,*], Asset Daniyarov[2,†,*], Aibek Karimov[2,†,*], Sultan Makhmetov[3,†,*], Anel Ordabayeva[2,†,*], Kamila Temirkhanova[4,†,*]

[1] Université de Montréal, Montréal, Canada

[2] Laboratory of Bioinformatics and Systems Biology, Center for Life Sciences, National Laboratory Astana, Nazarbayev University, Astana, Kazakhstan

[3] University of Pavia, Pavia, Italy

[4] Institute for Clinical Chemistry and Laboratory Medicine, TU Dresden, Dresden, Germany

† These authors contributed equally to this work and are listed in alphabetical order.

* Corresponding authors: aizhamal.aitzhanova@umontreal.ca (A.A.); rakhimbek.bektayev@nu.edu.kz (R.B.); asset.daniyarov@nu.edu.kz (A.D.); aibek.karimov@nu.edu.kz (A.K.); sultan.makhmetov01@universitadipavia.it (S.M.); anel.ordabayeva@nu.edu.kz (A.O.); kamila.temirkhanova@ukdd.de (K.T.)

ORCID: Rakhimbek Bektayev, 0000-0002-7658-5952; Asset Daniyarov, 0000-0003-3886-718X; Aibek Karimov, 0000-0001-5186-8561; Sultan Makhmetov, 0009-0007-5203-4834; Anel Ordabayeva, 0009-0009-8857-3199; Kamila Temirkhanova, 0009-0001-7885-8758

Abstract

Complete, phased genome assemblies and population-scale sequencing have changed how human genomes are analysed. This review surveys current methods and bioinformatics tools in four connected areas, with emphasis on the human genome and its evolution. First, we describe genome assembly, from overlap and de Bruijn graph algorithms to telomere-to-telomere assemblers for high-fidelity and ultra-long reads, together with phasing, scaffolding, polishing and quality assessment. Second, we cover the analysis of variation: read mapping, calling of small and structural variants, phasing and imputation, annotation and effect prediction, rare-variant association testing, and population-genetic inference of demography, genealogies and selection. Third, we discuss comparative genomics and the pangenome, including whole-genome alignment, evolutionary constraint, complete primate genomes, pangenome graph construction and graph-based genotyping. Fourth, we review the study of repeats: transposable elements, centromeric satellites, tandem repeats and segmental duplications, which together make up more than half of the genome and were largely inaccessible until recently. For each area we list representative tools and note practical limitations. We conclude that assembly-based and pangenome-based analysis is replacing the single linear reference. The main open problems are cost at cohort scale, standards for graph-based data, functional interpretation of complex variants, and the under-representation of many populations, including those of Central Asia.

Keywords: *genome assembly; telomere-to-telomere; long-read sequencing; structural variation; rare variants; pangenome; comparative genomics; repetitive DNA; human evolution*

Word count: *main text 5152 words; abstract 203 words; figures 1; tables 4*

1. Introduction

The first drafts of the human genome [1, 2] gave genetics a shared coordinate system, and successive reference builds up to GRCh38 served as the basis of read mapping, variant calling and annotation for two decades [3]. That reference was nevertheless a mosaic of a few donors, and about 8% of the genome was missing or misassembled. The unresolved fraction consisted mostly of centromeric satellite arrays, recent segmental duplications and the short arms of the acrocentric chromosomes [4].

Three developments have changed this situation. First, long-read sequencing matured into two complementary data types: PacBio high-fidelity (HiFi) reads of about 10–25 kb with accuracy of 99.8% or higher [5], and Oxford Nanopore Technologies (ONT) ultra-long reads that exceed 100 kb [6–8]. Second, assembly algorithms designed for these data produced the first gapless human genome, T2T-CHM13, which spans 3.055 Gb and adds nearly 200 Mb of sequence absent from GRCh38 [4], followed by a complete Y chromosome of 62.46 Mb [9]. Third, the field began to replace the single linear reference with a pangenome that represents many haplotypes at once. The Human Pangenome Reference Consortium (HPRC) released a draft built from 47 phased diploid assemblies [10, 11], and its second release (HPRC2) comprises 460 haplotypes that capture more than 99% of the common variation observed in the All of Us cohort [12].

These advances arrived alongside population-scale sequencing of hundreds of thousands of genomes [13, 14] and telomere-to-telomere (T2T) assemblies of ape genomes [15, 16]. As a result, questions that were previously out of reach can now be addressed directly: how centromeres and duplicated genes vary among people, how often tandem repeats mutate, and which structural changes distinguish the human lineage from other primates.

The bioinformatics toolkit has changed as fast as the data. This review surveys methods and software in four connected areas, with the human genome and its evolution as the main focus: (i) genome assembly; (ii) analysis of variation, including polymorphism and rare variants; (iii) comparative genomics and the pangenome; and (iv) the study of repeats. For each area we summarize the underlying algorithmic ideas, list widely used tools, and point out limitations that matter in practice. Figure 1 shows how the four areas connect.

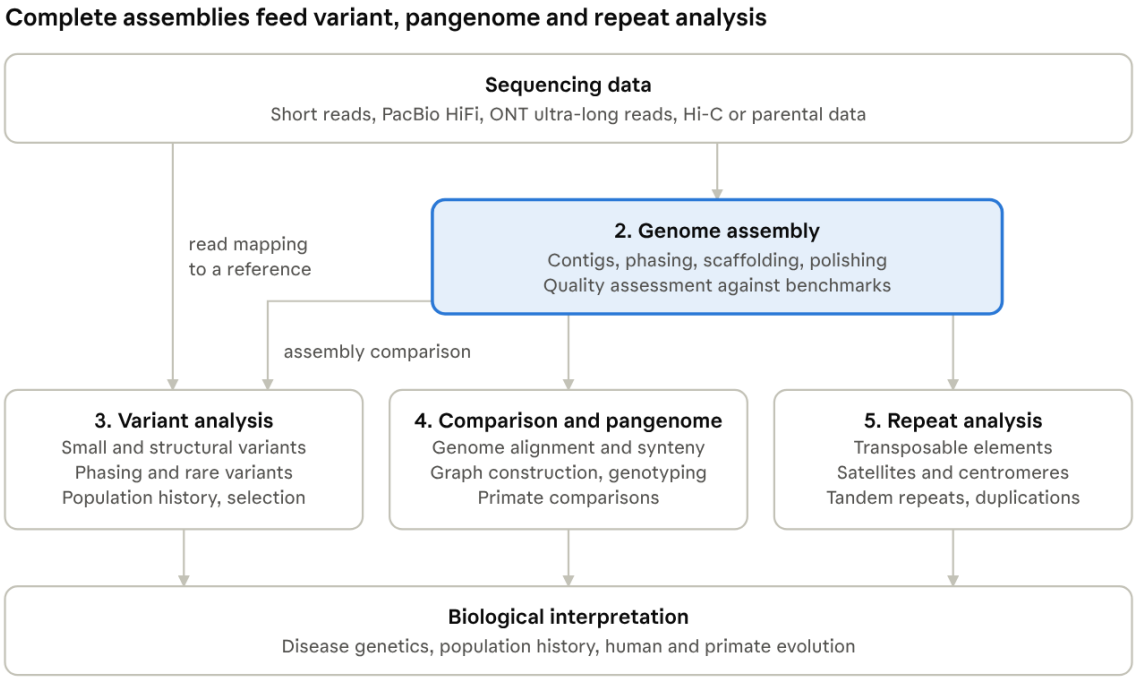


Fig 1. Overview of the analyses covered in this review. Sequencing reads can be mapped to a reference directly (left), but complete assemblies (Section 2) increasingly serve as the common input for variant

analysis (Section 3), comparative and pangenome analysis (Section 4) and repeat analysis (Section 5). Pangenome graphs in turn act as references for genotyping.

2. Genome assembly methods

2.1 Algorithmic foundations

De novo assembly reconstructs chromosomes from reads by detecting overlaps and resolving repeats. Three graph formulations underlie almost all assemblers: overlap–layout–consensus, in which reads are nodes and overlaps are edges; de Bruijn graphs, in which reads are broken into k-mers [17]; and string graphs, which remove redundant overlaps to give a compact representation of the genome [18]. De Bruijn graph assemblers such as SPAdes [19] remain standard for short reads and small genomes. In a human genome, however, repeats longer than the read or insert length break short-read assemblies into thousands of contigs and collapse duplicated sequence.

A repeat can be resolved only by reads that span it or that contain differences between its copies. Read length and read accuracy, more than coverage, therefore set the limit on contiguity [20].

2.2 Assembly of noisy long reads

Early PacBio continuous long reads and ONT reads carried error rates of 5–15%, so assemblers had to tolerate or remove errors before building a graph. Canu corrects and trims reads and then assembles them, using adaptive k-mer weighting to down-weight repetitive overlaps [21]. FALCON-Unzip introduced diploid-aware assembly that separates haplotypes into phased blocks [22]. Miniasm showed that the correction step can be skipped altogether [23], Flye builds repeat graphs from error-prone disjointigs [24], and wtdbg2 uses fuzzy de Bruijn graphs to cut running time [25]. Shasta relies on run-length encoding and a marker representation of reads to assemble a human genome from nanopore data within hours [26].

All of these methods need consensus polishing, for example with Racon [27], to reach useful base accuracy. They produce contigs of tens of megabases but usually collapse segmental duplications and satellite arrays.

2.3 High-fidelity reads and telomere-to-telomere assembly

HiFi reads made it possible to tell apart near-identical repeat copies and the two haplotypes of a diploid genome. HiCanu adapted Canu to HiFi data through homopolymer compression and strict overlap filtering [28]. Hifiasm builds a phased assembly graph that keeps both haplotypes instead of collapsing them, and is now the most widely used assembler for HiFi data [29]. The La Jolla Assembler applies multiplex de Bruijn graphs with very large k to the same problem [30].

The T2T-CHM13 genome combined a HiFi string graph with ultra-long ONT reads that were used to untangle the remaining repeats, and it benefited from the effectively haploid CHM13 cell line [4, 31]. Verkko automated this strategy: it builds a HiFi-based graph, aligns ONT reads to it, and uses trio or Hi-C information to trace each haplotype through the graph [32]. Verkko2 improves repeat resolution and gap closing and adds proximity-ligation phasing and scaffolding [33]. Hifiasm follows a comparable route with a "double graph" that integrates ultra-long reads [34].

The need for two sequencing platforms raises cost and DNA input, and several groups have worked to remove it. ONT duplex reads combined with Pore-C produced gapless human and plant chromosomes from nanopore data alone [35]. HERRO applies haplotype-aware deep-learning correction to ultra-long simplex reads so that they can be assembled with HiFi-oriented tools [36]. Hifiasm (ONT) reaches near-T2T contiguity from standard simplex reads without ultra-long libraries [37].

2.4 Haplotype phasing and scaffolding

A diploid assembly can be collapsed, pseudo-haplotype (a primary assembly plus alternate contigs, cleaned with tools such as purge_dups [38]), or fully phased. Full phasing requires long-range information. Trio binning partitions reads or graph paths with parent-specific k-mers [39]. When parents are unavailable, Hi-C contacts [40, 41] or single-cell Strand-seq [42] give chromosome-scale haplotypes.

Hi-C is also the usual source of scaffolding information. 3D-DNA [43], SALSA2 [44] and YaHS [45] order and orient contigs, and the result is normally inspected and corrected by hand in Juicebox [46]. The Vertebrate Genomes Project and the HPRC have published end-to-end pipelines that combine these steps and document their error modes [47, 48].

2.5 Polishing and quality assessment

Even T2T assemblies contain residual errors, most of them small indels in homopolymers and low-complexity sequence. The T2T Consortium described a conservative polishing strategy that avoids introducing errors in repeats [49]. Merfin filters candidate corrections by k-mer consistency with the reads [50], and DeepPolisher uses a transformer model on phased HiFi alignments [51].

Assembly quality has several independent dimensions, and no single statistic captures them:

- Contiguity: contig and scaffold N50 or NG50, and the number of chromosomes assembled telomere to telomere.
- Completeness: conserved single-copy genes with BUSCO [52] or compleasm [53], and k-mer completeness.
- Base accuracy: reference-free consensus quality values from Merqury [54].
- Structural correctness: reference-based comparison with QUAST-LG [55], and read-based detection of misassemblies with Inspector [56] or Flagger [11].
- Phasing: switch and Hamming error rates from parental k-mers [54].

A complete diploid assembly of the HG002 genome now provides a benchmark with near-perfect accuracy across 99.4% of the genome, against which reads, variant calls and assemblies can be scored [57].

2.6 Current state and open problems

Near-complete diploid human assemblies are now routine. The Human Genome Structural Variation Consortium assembled 130 haplotypes from 65 individuals with a median contiguity of 130 Mb, closed 92% of earlier assembly gaps and finished 39% of chromosomes telomere to telomere [58]. Population-specific T2T references, such as T2T-YAO for Han Chinese [59], are also appearing.

Several problems remain open. Ribosomal DNA arrays and the largest satellite arrays are still not assembled automatically. Ultra-long libraries need large amounts of high-molecular-weight DNA, which is rarely available from clinical samples. Manual curation remains a bottleneck, and methods for aneuploid, polyploid and tumour genomes lag behind those for normal diploid samples. Table 1 lists representative tools.

Table 1. Representative tools for genome assembly and assembly evaluation.

Stage	Tool	Input data	Key idea	Reference
Contig assembly	Canu / HiCanu	PacBio CLR, ONT, HiFi	Correction, trimming and overlap-based assembly; homopolymer compression for HiFi	[21, 28]
Contig assembly	Flye	PacBio CLR, ONT, HiFi	Repeat graphs built from disjointigs	[24]
Contig assembly	Shasta	ONT	Run-length encoding and marker-based overlap; very fast	[26]

Stage	Tool	Input data	Key idea	Reference
Contig assembly	wtdbg2	PacBio CLR, ONT	Fuzzy de Bruijn graph	[25]
Contig assembly	La Jolla Assembler	HiFi	Multiplex de Bruijn graph	[30]
Phased assembly	hifiasm	HiFi, with trio, Hi-C or ultra-long ONT	Phased string graph; haplotype-resolved contigs	[29, 34, 41]
Phased assembly	hifiasm (ONT)	ONT simplex	Near-T2T assembly without ultra-long reads	[37]
T2T assembly	Verkko / Verkko2	HiFi or corrected ONT, ultra-long ONT, trio or Hi-C	HiFi graph resolved with ONT alignments and haplotype walks	[32, 33]
Read correction	HERRO	ONT ultra-long simplex	Haplotype-aware deep-learning correction	[36]
Haplotig purging	purge_dups	Primary assembly and reads	Removes haplotypic duplication by depth and self-alignment	[38]
Scaffolding	3D-DNA, SALSA2, YaHS	Contigs and Hi-C	Ordering and orientation from contact frequency	[43–45]
Polishing	Merfin, DeepPolisher	Assembly and reads	k-mer validation of edits; transformer-based correction	[50, 51]
Evaluation	Merqury	Assembly and read k-mers	Reference-free QV, completeness and phasing	[54]
Evaluation	BUSCO, compleasm	Assembly	Conserved single-copy orthologues	[52, 53]
Evaluation	QUAST-LG, Inspector	Assembly, reference or reads	Misassembly and structural error detection	[55, 56]

3. Analysis of genomes, polymorphism and rare variants

3.1 From reads to small variants

Most human genomes are still analysed by resequencing: reads are aligned to a reference, variants are called and filtered, and the calls are annotated. BWA and BWA-MEM are the standard aligners for short reads [60, 61]. Minimap2 fills the same role for long reads and assemblies [62, 63], and Winnowmap2 improves placement in repetitive sequence [64]. SAMtools and BCFtools handle alignment and variant files [65], and the Genome Analysis Toolkit supplies a widely used framework for duplicate marking, recalibration and calling [66, 67]. DRAGEN combines hardware acceleration, multigenome mapping and machine-learning-based calling of all variant classes in one pipeline [68].

For single-nucleotide variants (SNVs) and short indels, haplotype-based callers have been joined by deep-learning methods. DeepVariant classifies candidate sites from pileup images with a convolutional network [69] and scales to cohorts through joint genotyping with GLnexus [70]. Clair3 [71] and PEPPER-Margin-DeepVariant [72] bring comparable accuracy to long reads by using haplotype information during calling.

Benchmarking relies on the Genome in a Bottle samples and standard comparison tools [73]. The precisionFDA Truth Challenge V2 and the benchmark for challenging medically relevant genes showed that most remaining errors fall in segmental duplications, low-complexity sequence and other difficult regions [74, 75]. The complete diploid HG002 benchmark extends evaluation to the whole genome [57].

The reference itself affects accuracy. Mapping 3,202 samples of the 1000 Genomes Project to T2T-CHM13 instead of GRCh38 removed tens of thousands of spurious variants per sample and reduced false positives in 269 medically relevant genes by up to 12-fold [76]. Many disease-relevant genes lie in regions that are dark or camouflaged for short reads [77]; long reads and pangenome references (Section 4) are the main routes to recovering them.

3.2 Structural variants

Structural variants (SVs; 50 bp or longer) affect more bases per genome than SNVs. A multi-platform study found about 27,600 SVs per genome, three to seven times more than standard short-read analyses detect [78].

Short-read callers combine discordant read pairs, split reads, read depth and local assembly. DELLY [79], LUMPY [80] and Manta [81] are common choices. Their sensitivity is limited for insertions and for variants inside repeats, and results differ considerably between tools [82–84]. Short-read population references such as the 1000 Genomes Project SV map [85] nevertheless remain valuable for filtering. gnomAD-SV catalogued 433,371 SVs in 14,891 genomes and estimated that SVs cause 25–29% of rare protein-truncating events per genome [86].

Long reads resolve most SVs directly. Alignment-based callers include Sniffles and Sniffles2, which also handles mosaic and population-level calling [87, 88], cuteSV [89] and SVIM [90]. Assembly-based discovery with SMRT-SV [91] or PAV compares phased assemblies to the reference; PAV identified 107,590 SVs in 64 haplotypes, 68% of which had not been found with short reads [92]. Jasmine [93] and Truvari [94] merge and compare call sets, and a Genome in a Bottle benchmark supports evaluation of large insertions and deletions [95].

Long-read sequencing of population panels is now under way. Nanopore sequencing of 1000 Genomes Project samples has yielded more than 100,000 sequence-resolved biallelic SVs and genotypes for 300,000 multiallelic tandem repeats in 1,019 individuals from 26 populations [96, 97]. Scalable protocols also make nanopore-only analysis of single genomes practical [98].

3.3 Phasing, imputation and reference panels

Haplotype phase is needed to interpret compound heterozygosity, to impute missing genotypes and to infer genealogies. Read-based phasing with WhatsHap [99] or HiPhase, which phases small variants, SVs and tandem repeats jointly from HiFi data [100], yields blocks limited by read length and heterozygosity. Statistical phasing with SHAPEIT4 [101], Beagle [102] or SHAPEIT5 uses haplotype sharing across a cohort. SHAPEIT5 keeps switch error rates below 5% even for variants seen once in 100,000 samples, which makes rare compound heterozygous events tractable [103]. GLIMPSE imputes genotypes from low-coverage sequencing [104].

These methods depend on reference panels. Widely used open panels include the 1000 Genomes Project [105] and its high-coverage extension to 3,202 samples with 602 trios [106], the Human Genome Diversity Project [107], the Simons Genome Diversity Project [108] and TOPMed [109].

3.4 Rare variants: discovery, annotation and association

Most human variants are rare. Among more than 400 million variants found in 53,831 TOPMed genomes, 97% had frequencies below 1% and 46% were singletons [109]. The first 150,119 UK Biobank genomes contained 585 million SNVs, about 7% of all possible single-nucleotide changes [110]; the resource now includes 490,640 genomes [14]. The All of Us programme reported more than 1 billion variants in 245,388 genomes, over 275 million of them previously unreported [13]. Aggregated allele frequencies from ExAC and gnomAD are the standard filter in rare-disease diagnostics [111–113].

Functional annotation starts with consequence prediction by the Ensembl Variant Effect Predictor [114], SnpEff [115] or ANNOVAR [116], and with clinical assertions from ClinVar [117]. Predictive scores add evidence for variants that have never been observed in patients: CADD integrates many annotations

[118], SpliceAI predicts splicing effects from sequence [119], AlphaMissense scores missense changes [120] and AlphaGenome targets regulatory variants [121]. Constraint metrics derived from the depletion of variation in large cohorts indicate which genes and non-coding regions are intolerant to change [112, 113]. Clinical classification combines these lines of evidence under the ACMG/AMP guidelines [122].

Single-variant association tests have little power at low allele counts, so rare variants are tested jointly within genes or regions. Burden tests collapse variants into one score, the sequence kernel association test (SKAT) allows opposite effect directions, and SKAT-O combines both [123–125]. At biobank scale, SAIGE-GENE and SAIGE-GENE+ [126, 127] and REGENIE [128] control for relatedness and case–control imbalance, and they have been applied to hundreds of thousands of exomes [129]. STAAR and STAARpipeline incorporate functional annotations and extend testing to non-coding regions [130, 131].

De novo mutations are the rarest class of variation and are identified in trios or larger pedigrees. Early work established the strong effect of paternal age on the number of de novo SNVs [132], and family-based sequencing has since measured de novo SV rates [133]. Near-complete assembly of a four-generation, 28-member family raised the estimate to 98–206 de novo events per transmission once tandem repeats, centromeres and other difficult regions were included [134].

3.5 Population-genetic and evolutionary inference

Variant call sets feed several families of evolutionary analysis:

- Population structure and admixture: principal component analysis [135], ADMIXTURE [136], PLINK 2 [137], ANGSD for low-coverage data [138], RFMix for local ancestry [139] and f-statistics in ADMIXTOOLS [140].
- Demographic history: sequentially Markovian coalescent methods such as PSMC [141], MSMC [142] and SMC++ [143], and site-frequency-spectrum methods such as dadi [144] and fastsimcoal2 [145].
- Genome-wide genealogies: ARGweaver [146], Relate [147], tsinfer [148], ARG-Needle [149] and SINGER [150] infer ancestral recombination graphs at increasing scale, and a unified genealogy now links modern and ancient genomes [151].
- Positive selection: haplotype-based statistics such as iHS [152] and XP-EHH [153], implemented in selscan [154], and the singleton density score for very recent adaptation [155].
- Archaic introgression: comparison with Neanderthal genomes [156, 157] and reference-free detection of introgressed segments [158, 159].

All of these analyses inherit the biases of the underlying data. Many populations are still poorly represented in reference panels and frequency databases. For Central Asia, for example, published whole-genome data remain limited to small studies such as five high-coverage Kazakh genomes [160, 161], which constrains imputation accuracy and frequency-based filtering for patients from the region. Table 2 summarizes the main tools.

Table 2. Representative tools for variant discovery, annotation and population analysis.

Task	Tools	Notes	References
Short-read alignment	BWA, BWA-MEM, DRAGEN	Burrows–Wheeler indexing; DRAGEN adds multigenome mapping	[60, 61, 68]
Long-read alignment	minimap2, Winnowmap2	Minimizer seeding and chaining; weighted minimizers for repeats	[62, 64]
SNV and indel calling	GATK, DeepVariant with GLnexus, Clair3, PEPPER-Margin-DeepVariant	Haplotype-based and deep-learning callers for short and long reads	[67, 69–72]
SV calling, short reads	DELLY, LUMPY, Manta	Paired-end, split-read and assembly evidence	[79–81]

Task	Tools	Notes	References
SV calling, long reads	Sniffles2, cuteSV, SVIM, PAV	Alignment-based and assembly-based discovery	[88–90, 92]
SV merging and comparison	Jasmine, Truvari	Population-scale merging; benchmarking	[93, 94]
Phasing and imputation	WhatsHap, HiPhase, SHAPEIT5, Beagle, GLIMPSE	Read-based and statistical phasing; low-coverage imputation	[99, 100, 102–104]
Annotation	VEP, SnpEff, ANNOVAR	Consequence prediction against gene models	[114–116]
Effect prediction	CADD, SpliceAI, AlphaMissense, AlphaGenome	Integrative, splicing, missense and regulatory scores	[118–121]
Rare-variant association	SKAT and SKAT-O, SAIGE-GENE+, REGENIE, STAAR	Set-based tests; mixed models for biobanks; annotation-informed tests	[123, 124, 127, 128, 130]
Demography	PSMC, MSMC, SMC++, dadi, fastsimcoal2	Coalescent and frequency-spectrum inference	[141–145]
Genealogy inference	Relate, tsinfer, ARG-Needle, SINGER	Ancestral recombination graphs	[147–150]
Selection scans	selscan (iHS, XP-EHH), SDS	Haplotype homozygosity; singleton density	[154, 155]

4. Comparative genomics and the pangenome

4.1 Whole-genome alignment, synteny and orthology

Comparative analysis begins with alignment of whole genomes. Pairwise pipelines based on chained and netted local alignments [162], LAST with adaptive seeds [163] and MUMmer4 [164] are long-established, and minimap2 is now widely used for assembly-to-assembly alignment [62]. The wavefront algorithm made exact gap-affine alignment fast enough for long, similar sequences and underlies recent all-to-all aligners [165]. For many species, Progressive Cactus produces reference-free multiple alignments stored in the HAL format [166, 167].

Alignments are then interpreted at several levels. SyRI identifies syntenic regions, inversions, translocations and duplications between assemblies [168]. MCScanX [169] and GENESPACE [170] work with gene order. OrthoFinder infers orthogroups from protein sequences [171], TOGA combines gene annotation with orthology inference from genome alignments [172], and Liftoff transfers annotations between assemblies of the same or related species [173].

4.2 Constraint, acceleration and selection across species

Sequence conservation across species is one of the most reliable indicators of function. PhastCons and phyloP estimate conserved elements and per-base rates of substitution from multiple alignments [174, 175]. In the Zoonomia alignment of 240 placental mammals, at least 332 Mb of the human genome (about 10.7%) is constrained [176], and constrained positions are enriched for variants that explain the heritability of common diseases [177]. Whole-genome data from 233 primate species add resolution at shorter evolutionary distances [178].

The opposite signal, lineage-specific acceleration, marks candidate regions of adaptive change such as the human accelerated regions [179]. At the protein level, codon models in PAML [180] and HyPhy

[181] test for positive selection, and CAFE 5 models the expansion and contraction of gene families [182].

Incomplete lineage sorting complicates all of these analyses in primates. It affects up to 64% of the genome at individual ancestral nodes, so local genealogies often disagree with the species tree and must be modelled explicitly [183].

4.3 Complete primate genomes and human-specific change

Until recently, comparisons between humans and other apes excluded the most dynamic parts of the genome because they were missing from the assemblies. Long-read assemblies of bonobo and other primates first showed how much structural change had been overlooked [184]. An analysis of eight primate lineages estimated that SVs have affected about 27% of the genome during primate evolution and identified 1,607 structurally divergent regions where genes are recurrently lost or gained [185].

T2T assemblies of ape sex chromosomes [15] and of six ape genomes with 215 gapless chromosomes [16] now allow segmental duplications, centromeres and immune loci to be compared without mapping reads to the human reference. This matters for human evolution because several human-specific genes implicated in cortical development, including SRGAP2 and NOTCH2NL paralogues, arose by segmental duplication in regions that were misassembled in earlier references [186, 187].

4.4 The pangenome and graph construction

The term pangenome was introduced for bacteria to denote the full gene repertoire of a species [188]. In humans the concept addresses reference bias: sequence absent from the reference cannot be mapped to or genotyped. Short-read data from 910 individuals of African descent already contained 296.5 Mb of sequence missing from GRCh38 [189].

A pangenome graph stores sequences as nodes, adjacencies as edges and haplotypes as paths [190]. Construction methods differ in what they align and what they keep:

- The vg toolkit builds variation graphs from a reference and a variant call set [191].
- Minigraph adds assemblies to a reference backbone one at a time and records SVs of 50 bp or more [192].
- Minigraph-Cactus adds base-level alignment to the minigraph structure and scaled to 90 human haplotypes [193].
- The PanGenome Graph Builder (PGGB) uses all-to-all alignment without a privileged reference [194], and ODGI provides manipulation and visualization of the resulting graphs [195].
- PGR-TK targets complex, clinically relevant loci at multiple scales [196], and pangene graphs represent gene order and copy number [197].

A systematic comparison found that these methods differ in the variation they represent, their scalability and the downstream tools they support, so the choice depends on the application [198].

The HPRC draft pangenome of 47 individuals added 119 Mb of euchromatic polymorphic sequence and 1,115 gene duplications relative to GRCh38 [11]. Regional efforts show that much variation is still unsampled. The Chinese Pangenome Consortium contributed 116 assemblies from 36 minority ethnic groups and 189 Mb of additional sequence [199], and a pangenome of 53 individuals from the United Arab Emirates reported about 112 Mb of euchromatic sequence absent from existing references [200].

4.5 Using pangenomes for mapping, genotyping and evolutionary analysis

Giraffe maps short reads to a haplotype-aware graph at a speed comparable to linear mappers [201], and GraphAligner aligns long reads to graphs [202]. Sampling a personalized subgraph that matches the k-mer content of each sample further reduces mapping errors caused by irrelevant rare alleles [203]. In the HPRC analysis, pangenome-based mapping reduced small-variant errors by 34% compared with GRCh38-based workflows [11].

Genotyping benefits even more. Known SVs can be genotyped from read alignments to the graph [204], and PanGenie infers genotypes from k-mer counts and panel haplotypes without alignment [205]. The draft pangenome increased the number of SVs detected per haplotype from short reads by 104% [11]. Locityper uses pangenome haplotypes to genotype 256 challenging medically relevant loci from short or long reads [206], and graph-based analysis has improved SV interpretation in rare-disease cohorts [207].

Pangenomes also support evolutionary inference in regions that were previously inaccessible. All-to-all comparison of HPRC assemblies provided evidence of recombination between heterologous acrocentric short arms [208]. Unambiguous alignment of segmental duplications across 102 haplotypes showed that SNV density is 60% higher there than in unique sequence, with interlocus gene conversion accounting for at least 23% of the excess [209].

4.6 Limitations

Pangenome analysis is not yet a drop-in replacement for linear references. Annotations, clinical databases and most pipelines still use linear coordinates, so results must be projected back with some loss. Centromeres and other satellite-rich regions are often excluded from graphs because they cannot be aligned reliably. Computational cost grows with the number of haplotypes, and file formats and tools are still converging. Finally, a pangenome is only as representative as its sample set, and broad, ethically governed sampling remains a prerequisite [10]. Table 3 lists the main tools.

Table 3. Representative tools for comparative genomics and pangenome analysis.

Task	Tools	Notes	References
Pairwise genome alignment	LAST, MUMmer4, minimap2	Seed-and-extend or minimizer chaining	[62, 163, 164]
Multiple genome alignment	Progressive Cactus, HAL tools	Reference-free alignment of hundreds of genomes	[166, 167]
Synteny and rearrangements	SyRI, MCScanX, GENESPACE	Sequence-level and gene-level collinearity	[168–170]
Orthology and annotation transfer	OrthoFinder, TOGA, Liftoff	Orthogroups; alignment-based orthologous loci; annotation lift-over	[171–173]
Conservation and acceleration	phastCons, phyloP	Conserved elements and per-base rate tests	[174, 175]
Selection and gene families	PAML, HyPhy, CAFE 5	Codon models; gene family size evolution	[180–182]
Graph construction	vg, minigraph, Minigraph-Cactus, PGGB	From variants, SV-only, reference-anchored base-level, all-to-all	[191–194]
Graph manipulation and local graphs	ODGI, PGR-TK, pangene	Visualization, statistics, complex loci, gene graphs	[195–197]
Read-to-graph alignment	Giraffe, GraphAligner	Short and long reads; personalized subgraphs	[201–203]
Graph-based genotyping	vg call, PanGenie, Locityper	Alignment-based, k-mer-based, targeted complex loci	[204–206]

5. Analysis of repeats

5.1 Repeat classes in the human genome

Repeats make up more than half of the human genome. The initial draft attributed about 45% of the sequence to transposable elements [1], and annotation of T2T-CHM13 raised the total repeat content to about 54% [4, 210]. Four classes are usually distinguished because they differ in origin, mutational behaviour and the tools needed to study them: interspersed transposable elements (TEs), satellite DNA, short and variable-number tandem repeats, and segmental duplications. Repeats were long masked and ignored in genome analysis; complete assemblies now make them directly measurable.

5.2 Transposable elements

TEs are classified hierarchically by transposition mechanism and structure [211]. Reference annotation normally runs RepeatMasker [212] with curated consensus sequences or profile hidden Markov models from Repbase and Dfam [213, 214]. For new assemblies and non-model species, families are first discovered de novo with RepeatModeler2 [215], EDTA [216] or HiTE [217]. Library-free tools that report interspersed and tandem repeats together are described in Section 5.6. Manual curation of the resulting libraries is still advisable, and the available software has been reviewed in detail [218].

In humans, only L1, Alu and SVA elements remain active. Their insertions are a recurrent cause of genetic disease [219, 220], and older copies have been co-opted as regulatory elements [221]. Polymorphic insertions absent from the reference are detected from short reads with MELT [222] or MEGAnE [223], from long reads with PALMER [224], and from combined data with xTea [225]. Nanopore reads additionally report the methylation state of individual elements [226], and graph genomes offer a unified way to genotype insertion polymorphisms [227].

Population studies show that these variants matter. Phased assemblies identified 130 of the most active source elements [92], long-read sequencing of 1,019 individuals traced L1- and SVA-mediated transductions [97], and mobile element variants are associated with changes in gene expression more often than SNVs and can be tested directly in genome-wide association studies [223]. In T2T-CHM13, repeat annotation was combined with transcription and CpG methylation data to define which elements are active, including those in centromeres [210, 228].

5.3 Satellite DNA and centromeres

Pericentromeric and centromeric satellites cover 6.2% of T2T-CHM13 (189.9 Mb) and were almost entirely absent from earlier references [229]. Human centromeres consist of α-satellite monomers of about 171 bp arranged in higher-order repeats (HORs) that form arrays of several megabases [230].

Assembly of these arrays required dedicated methods such as centroFlye [231]. Annotation tools decompose arrays into monomers and HORs: StringDecomposer [232], HORmon [233], Satellite Repeat Finder [234] and TRASH [235]. Sequence identity heatmaps from StainedGlass [236] and ModDotPlot [237] reveal the layered structure of arrays, and a local dip in CpG methylation marks the probable site of kinetochore assembly [228].

Centromeres are among the fastest-evolving regions of the genome. Comparison of two complete sets of human centromeres showed at least a 4.1-fold excess of single-nucleotide variation relative to unique flanking sequence, up to 3-fold differences in array size, and 45.8% of centromeric sequence that could not be aligned reliably between the two genomes [238]. Across 1,246 assembled centromeres from diverse individuals, α-satellite HOR arrays vary in length by up to 30-fold [58]. Standard alignment-based variant calling is therefore unsuitable for these regions, and analysis relies on assembly-to-assembly comparison of HOR structure.

5.4 Short tandem repeats and variable-number tandem repeats

Tandem repeats mutate by changes in copy number at rates far above those of point mutation, and they contribute to both normal variation and disease [239–241]. Expansions cause more than 50 disorders, most of them neurological [242]. A genome-wide catalogue of tandem repeat regions covers 8.1% of the genome but contains about 24.9% of the variants found in an individual [243].

Tandem Repeats Finder remains the basis of most repeat catalogues [244]. Genotyping from short reads developed through lobSTR [245] and HipSTR [246] for alleles shorter than the read, and through ExpansionHunter [247], GangSTR [248], STRetch [249] and STRling [250] for expansions; STRling also detects expansions at loci absent from the reference. Genome sequencing combined with linkage analysis identified the intronic RFC1 expansion as a frequent cause of late-onset ataxia [251], and whole-genome sequencing has been evaluated against PCR-based testing for the diagnosis of neurological repeat expansion disorders [252]. Population resources include genotypes at more than 1.7 million loci in 3,550 individuals, with a phased panel for imputing tandem repeats from nearby SNPs [253] and TR-gnomAD, with 0.86 million loci genotyped in 338,963 genomes [254].

Short reads cannot size long or interrupted alleles. Long-read tools read the full allele sequence, its motif composition and, where available, its methylation: TRGT for HiFi data [255], Straglr [256] and LongTR [257]. Vamos annotates variable-number tandem repeats with efficient motif sets [258], and danbing-tk genotypes them from short reads with repeat-pangenome graphs [259]. A benchmark built from 86 haplotype-resolved assemblies and the HG002 sample now allows these tools to be compared [243].

Pedigree data confirm that tandem repeats dominate the mutational input per generation. In a four-generation family, about 65 of the de novo indels and SVs per transmission arose in tandem repeats, compared with about 75 de novo SNVs, and 32 loci mutated recurrently [134].

5.5 Segmental duplications

Segmental duplications (SDs) are blocks longer than 1 kb with at least 90% identity. Early whole-genome analysis linked them to hotspots of genomic instability and disease [260]. With the complete genome, the estimate of SD content rose from 5.4% to 7.0% (218 Mb), and SDs accounted for nearly one-third of the sequence added by T2T-CHM13 [261]. SEDEF [262] and BISER [263] detect SDs in one or many assemblies.

SDs are central to human evolution and to medical genetics. They gave rise to human-specific genes (Section 4.3), they mediate recurrent microdeletions and microduplications, and they are subject to frequent interlocus gene conversion [209]. Short reads rarely map uniquely to paralogues, so genotyping needs specialized tools. Paraphase resolves SMN1 and SMN2 haplotypes from HiFi reads [264], and Locityper selects the best-fitting pair of pangenome haplotypes for each locus [206]. Phased assemblies now span loci such as SMN1/SMN2 and AMY1/AMY2 in full [58]. The ribosomal DNA arrays on the acrocentric short arms remain the largest repeats without routine assembly [4, 208].

5.6 Library-free *de novo* identification of repeats across classes

The tools described above are specialized: each targets one repeat class, and many depend on curated libraries or on a prior model of repeat structure. A complementary strategy detects every repeated sequence in an assembly directly, without a library or reference annotation, and groups the copies into families afterwards. This is useful for newly assembled genomes, for regions such as satellite arrays where libraries are incomplete, and as an independent check on library-based masking.

Repeater implements this strategy as a comparison of a sequence with itself [265]. In the current version, all k-mers on the forward and reverse-complement strands are indexed and those that occur only once are discarded. Each pair of positions that share a seed is extended into an ungapped local alignment that tolerates isolated mismatches, overlapping alignments are merged into repeat copies, and copies that align to one another are clustered into families. Direct and inverted repeats are therefore found in a single pass. The same procedure reports interspersed elements, tandem repeats and higher-order arrays such as telomeric and satellite sequences, and a separate linear-time mode reports microsatellites and other low-complexity loci.

Repeater writes a coordinate table of repeat copies with their cluster assignments, a soft-masked sequence, and a repeat-landscape figure that distinguishes forward from reverse-complement copies. It is a Java command-line program released under the GPL-3.0 licence, and the seed length (18 by

default) is the main parameter controlling sensitivity and speed. Sequences in a multi-record file are analysed one at a time, so copies located on different chromosomes are not linked into one family.

TotalRepeats applies an alignment-free design to whole genomes and to sets of genomes [266]. Short tandem and low-complexity repeats are identified first. All k-mers are then counted on both strands, bases covered by k-mers that occur at least twice are marked, and covered runs are merged into repeat blocks. Blocks are grouped into families by the cosine similarity of their 4-mer composition profiles, with a threshold calibrated on the data of each run, or alternatively by k-mer containment. Families can optionally be named against Repbase, Dfam or a custom library, which connects de novo discovery to established classifications.

In addition to annotation tables, soft-masked sequence and landscape plots, TotalRepeats has comparative modes for collections of genomes. These classify repeat families as core, soft-core, shell or cloud, estimate the openness of the repeat pangenome, and derive distances between genomes from shared repeat content. The program can also import RepeatMasker masks and compare them with its own, so the two approaches can be used together. TotalRepeats is currently documented in its software repository and has no peer-reviewed description yet; the run times reported there are for sequences of 1.4–15 Mb.

Both tools are best regarded as complementary to library-based annotation. Similarity-based families are not a transposable element classification, and for human assemblies their output should be compared with RepeatMasker and Dfam annotation before biological conclusions are drawn. Table 4 lists representative tools.

Table 4. Representative tools for the analysis of repeats.

Repeat class	Task	Tools	References
Transposable elements	Reference annotation	RepeatMasker with Dfam or Repbase	[212–214]
Transposable elements	De novo family discovery	RepeatModeler2, EDTA, HiTE	[215–217]
Transposable elements	Polymorphic insertions, short reads	MELT, MEGAnE, xTea	[222, 223, 225]
Transposable elements	Polymorphic insertions, long reads	PALMER, xTea, TLDR	[224–226]
Satellites and centromeres	Assembly	centroFlye, Verkko, hifiasm	[32, 34, 231]
Satellites and centromeres	Monomer and HOR annotation	StringDecomposer, HORmon, Satellite Repeat Finder, TRASH	[232–235]
Satellites and centromeres	Visualization	StainedGlass, ModDotPlot	[236, 237]
Tandem repeats	Detection in sequence	Tandem Repeats Finder	[244]
Tandem repeats	Genotyping, short reads	HipSTR, ExpansionHunter, GangSTR, STRetch, STRling	[246–250]
Tandem repeats	Genotyping, long reads	TRGT, Straglr, LongTR, vamos	[255–258]
Segmental duplications	Detection	SEDEF, BISER	[262, 263]
Segmental duplications	Paralogue-aware genotyping	Paraphase, Locityper	[206, 264]
All classes	Library-free *de novo* detection, masking and visualization	Repeater, TotalRepeats	[265, 266]

6. Conclusions and future perspectives

Genomic analysis is shifting from a model centred on one incomplete reference and short reads to one centred on complete, phased assemblies and collections of such assemblies. Three consequences follow from the work reviewed here.

First, assembly has become a routine starting point for variant discovery. In the complete HG002 benchmark, de novo assembly recovered more of the genome and was more accurate than mapping-based variant calling [57]. Second, classes of variation that were previously set aside can now be measured: structural variants, tandem repeats, segmental duplications and centromeric satellites together account for most of the bases that differ between two genomes and for a large share of new mutations [58, 134]. Third, pangenomes and complete primate genomes allow comparisons within and between species that do not depend on what a single reference happens to contain [11, 16].

Several challenges will shape the next few years:

- Scale and cost. Long-read sequencing of biobank-sized cohorts is still expensive. In the meantime, pangenome-based genotyping and imputation are the practical way to recover complex variation from existing short-read data.
- Standards. Coordinate systems, graph formats, annotation of pangenomes and the representation of variants inside repeats are not yet settled, which hinders exchange between tools and databases.
- Interpretation. Predicting the functional effect of non-coding, repeat-associated and structural variants lags behind their discovery. Deep-learning predictors need independent validation, particularly for ancestries that are scarce in training data.
- Representation. Reference panels, frequency databases and pangenomes still under-sample much of the world, including Central Asia. Regional sequencing and assembly projects are needed for equitable use of genomic medicine.
- Remaining technical gaps. Ribosomal DNA arrays, the largest satellite arrays, somatic mosaicism and fully automated curation of assemblies remain open problems.

For practitioners, the choice of tools should follow the data type and the class of variation under study, results should be validated against the available benchmarks, and software versions and parameters should be reported so that analyses can be reproduced.

Data Availability Statement

No new data were generated for this review. Bibliographic details and DOIs of the cited articles were checked against PubMed records.

Author Contributions

All authors contributed equally to this work: conceptualization, literature search, writing of the original draft, and review and editing. The authors are listed in alphabetical order. All authors read and approved the final manuscript.

Ethics Approval and Consent to Participate

Not applicable.

Conflicts of Interest

The authors declare no conflicts of interest.

Declaration of AI and AI-Assisted Technologies

During the preparation of this work the authors used Claude (model Claude Fable 5.2; Anthropic) in order to search and screen the literature, to draft and structure the text of the manuscript, to prepare the tables and the workflow figure, and to compile and format the reference list. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. *Nature*. 2001; 409(6822): 860-921. https://doi.org/10.1038/35057062
2. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. *Science*. 2001; 291(5507): 1304-51. https://doi.org/10.1126/science.1058040
3. Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen HC, Kitts PA, et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res*. 2017; 27(5): 849-864. https://doi.org/10.1101/gr.213611.116
4. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete sequence of a human genome. *Science*. 2022; 376(6588): 44-53. https://doi.org/10.1126/science.abj6987
5. Wenger AM, Peluso P, Rowell WJ, Chang PC, Hall RJ, Concepcion GT, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol*. 2019; 37(10): 1155-1162. https://doi.org/10.1038/s41587-019-0217-9
6. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol*. 2018; 36(4): 338-345. https://doi.org/10.1038/nbt.4060
7. Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and its applications. *Nat Rev Genet*. 2020; 21(10): 597-614. https://doi.org/10.1038/s41576-020-0236-x
8. Wang Y, Zhao Y, Bollas A, Wang Y, Au KF. Nanopore sequencing technology, bioinformatics and applications. *Nat Biotechnol*. 2021; 39(11): 1348-1365. https://doi.org/10.1038/s41587-021-01108-x
9. Rhie A, Nurk S, Cechova M, Hoyt SJ, Taylor DJ, Altemose N, et al. The complete sequence of a human Y chromosome. *Nature*. 2023; 621(7978): 344-354. https://doi.org/10.1038/s41586-023-06457-y
10. Wang T, Antonacci-Fulton L, Howe K, Lawson HA, Lucas JK, Phillippy AM, et al. The Human Pangenome Project: a global resource to map genomic diversity. *Nature*. 2022; 604(7906): 437-446. https://doi.org/10.1038/s41586-022-04601-8
11. Liao WW, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, et al. A draft human pangenome reference. *Nature*. 2023; 617(7960): 312-324. https://doi.org/10.1038/s41586-023-05896-x
12. Lucas JK, Hebbar P, Liao WW, Macias-Velasco JF, Novak AM, Asri M, et al. HPRC2: A human pangenome reference with near-complete coverage of common genetic variation. *bioRxiv* [Preprint], version 1, posted 22 July 2026; not peer reviewed. 2026. https://doi.org/10.64898/2026.07.21.739710. PMID: 42539208; PMCID: PMC13419748
13. All of Us Research Program Genomics Investigators. Genomic data in the All of Us Research Program. *Nature*. 2024; 627(8003): 340-346. https://doi.org/10.1038/s41586-023-06957-x
14. UK Biobank Whole-Genome Sequencing Consortium. Whole-genome sequencing of 490,640 UK Biobank participants. *Nature*. 2025; 645(8081): 692-701. https://doi.org/10.1038/s41586-025-09272-9

15. Makova KD, Pickett BD, Harris RS, Hartley GA, Cechova M, Pal K, et al. The complete sequence and comparative analysis of ape sex chromosomes. *Nature*. 2024; 630(8016): 401-411. https://doi.org/10.1038/s41586-024-07473-2
16. Yoo D, Rhie A, Hebbar P, Antonacci F, Logsdon GA, Solar SJ, et al. Complete sequencing of ape genomes. *Nature*. 2025; 641(8062): 401-418. https://doi.org/10.1038/s41586-025-08816-3
17. Pevzner PA, Tang H, Waterman MS. An Eulerian path approach to DNA fragment assembly. *Proc Natl Acad Sci U S A*. 2001; 98(17): 9748-53. https://doi.org/10.1073/pnas.171285098
18. Myers EW. The fragment assembly string graph. *Bioinformatics*. 2005; 21 Suppl 2: ii79-85. https://doi.org/10.1093/bioinformatics/bti1114
19. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol*. 2012; 19(5): 455-77. https://doi.org/10.1089/cmb.2012.0021
20. Li H, Durbin R. Genome assembly in the telomere-to-telomere era. *Nat Rev Genet*. 2024; 25(9): 658-670. https://doi.org/10.1038/s41576-024-00718-w
21. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res*. 2017; 27(5): 722-736. https://doi.org/10.1101/gr.215087.116
22. Chin CS, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods*. 2016; 13(12): 1050-1054. https://doi.org/10.1038/nmeth.4035
23. Li H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics*. 2016; 32(14): 2103-10. https://doi.org/10.1093/bioinformatics/btw152
24. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol*. 2019; 37(5): 540-546. https://doi.org/10.1038/s41587-019-0072-8
25. Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. *Nat Methods*. 2020; 17(2): 155-158. https://doi.org/10.1038/s41592-019-0669-3
26. Shafin K, Pesout T, Lorig-Roach R, Haukness M, Olsen HE, Bosworth C, et al. Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nat Biotechnol*. 2020; 38(9): 1044-1053. https://doi.org/10.1038/s41587-020-0503-6
27. Vaser R, Sović I, Nagarajan N, Šikić M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res*. 2017; 27(5): 737-746. https://doi.org/10.1101/gr.214270.116
28. Nurk S, Walenz BP, Rhie A, Vollger MR, Logsdon GA, Grothe R, et al. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res*. 2020; 30(9): 1291-1305. https://doi.org/10.1101/gr.263566.120
29. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods*. 2021; 18(2): 170-175. https://doi.org/10.1038/s41592-020-01056-5
30. Bankevich A, Bzikadze AV, Kolmogorov M, Antipov D, Pevzner PA. Multiplex de Bruijn graphs enable genome assembly from long, high-fidelity reads. *Nat Biotechnol*. 2022; 40(7): 1075-1081. https://doi.org/10.1038/s41587-022-01220-6
31. Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, Bzikadze A, et al. Telomere-to-telomere assembly of a complete human X chromosome. *Nature*. 2020; 585(7823): 79-84. https://doi.org/10.1038/s41586-020-2547-7
32. Rautiainen M, Nurk S, Walenz BP, Logsdon GA, Porubsky D, Rhie A, et al. Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nat Biotechnol*. 2023; 41(10): 1474-1482. https://doi.org/10.1038/s41587-023-01662-6
33. Antipov D, Rautiainen M, Nurk S, Walenz BP, Solar SJ, Phillippy AM, et al. Verkko2 integrates proximity-ligation data with long-read De Bruijn graphs for efficient telomere-to-telomere genome assembly, phasing, and scaffolding. *Genome Res*. 2025; 35(7): 1583-1594. https://doi.org/10.1101/gr.280383.124

34. Cheng H, Asri M, Lucas J, Koren S, Li H. Scalable telomere-to-telomere assembly for diploid and polyploid genomes with double graph. *Nat Methods*. 2024; 21(6): 967-970. https://doi.org/10.1038/s41592-024-02269-8
35. Koren S, Bao Z, Guarracino A, Ou S, Goodwin S, Jenike KM, et al. Gapless assembly of complete human and plant chromosomes using only nanopore sequencing. *Genome Res*. 2024; 34(11): 1919-1930. https://doi.org/10.1101/gr.279334.124
36. Stanojević D, Lin D, Nurk S, de Sessions PF, Šikić M. Telomere-to-telomere assembly using HERRO-corrected Nanopore Simplex reads. *Nature*. 2026; 655(8121): 158-165. https://doi.org/10.1038/s41586-026-10563-y
37. Cheng H, Qu H, McKenzie S, Lawrence KR, Windsor R, Vella M, et al. Efficient near-telomere-to-telomere assembly of nanopore simplex reads. *Nature*. 2026; 655(8121): 166-173. https://doi.org/10.1038/s41586-026-10105-6
38. Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics*. 2020; 36(9): 2896-2898. https://doi.org/10.1093/bioinformatics/btaa025
39. Koren S, Rhie A, Walenz BP, Dilthey AT, Bickhart DM, Kingan SB, et al. De novo assembly of haplotype-resolved genomes with trio binning. *Nat Biotechnol*. 2018; 36(12): 1174-1182. https://doi.org/10.1038/nbt.4277
40. Garg S, Fungtammasan A, Carroll A, Chou M, Schmitt A, Zhou X, et al. Chromosome-scale, haplotype-resolved assembly of human genomes. *Nat Biotechnol*. 2021; 39(3): 309-312. https://doi.org/10.1038/s41587-020-0711-0
41. Cheng H, Jarvis ED, Fedrigo O, Koepfli KP, Urban L, Gemmell NJ, et al. Haplotype-resolved assembly of diploid genomes without parental data. *Nat Biotechnol*. 2022; 40(9): 1332-1335. https://doi.org/10.1038/s41587-022-01261-x
42. Porubsky D, Ebert P, Audano PA, Vollger MR, Harvey WT, Marijon P, et al. Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads. *Nat Biotechnol*. 2021; 39(3): 302-308. https://doi.org/10.1038/s41587-020-0719-5
43. Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, et al. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science*. 2017; 356(6333): 92-95. https://doi.org/10.1126/science.aal3327
44. Ghurye J, Rhie A, Walenz BP, Schmitt A, Selvaraj S, Pop M, et al. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLoS Comput Biol*. 2019; 15(8): e1007273. https://doi.org/10.1371/journal.pcbi.1007273
45. Zhou C, McCarthy SA, Durbin R. YaHS: yet another Hi-C scaffolding tool. *Bioinformatics*. 2023; 39(1). https://doi.org/10.1093/bioinformatics/btac808
46. Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, et al. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst*. 2016; 3(1): 99-101. https://doi.org/10.1016/j.cels.2015.07.012
47. Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, et al. Towards complete and error-free genome assemblies of all vertebrate species. *Nature*. 2021; 592(7856): 737-746. https://doi.org/10.1038/s41586-021-03451-0
48. Jarvis ED, Formenti G, Rhie A, Guarracino A, Yang C, Wood J, et al. Semi-automated assembly of high-quality diploid human reference genomes. *Nature*. 2022; 611(7936): 519-531. https://doi.org/10.1038/s41586-022-05325-5
49. Mc Cartney AM, Shafin K, Alonge M, Bzikadze AV, Formenti G, Fungtammasan A, et al. Chasing perfection: validation and polishing strategies for telomere-to-telomere genome assemblies. *Nat Methods*. 2022; 19(6): 687-695. https://doi.org/10.1038/s41592-022-01440-3
50. Formenti G, Rhie A, Walenz BP, Thibaud-Nissen F, Shafin K, Koren S, et al. Merfin: improved variant filtering, assembly evaluation and polishing via k-mer validation. *Nat Methods*. 2022; 19(6): 696-704. https://doi.org/10.1038/s41592-022-01445-y

51. Mastoras M, Asri M, Brambrink L, Hebbar P, Kolesnikov A, Cook DE, et al. Highly accurate assembly polishing with DeepPolisher. *Genome Res*. 2025; 35(7): 1595-1608. https://doi.org/10.1101/gr.280149.124
52. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol Biol Evol*. 2021; 38(10): 4647-4654. https://doi.org/10.1093/molbev/msab199
53. Huang N, Li H. compleasm: a faster and more accurate reimplementation of BUSCO. *Bioinformatics*. 2023; 39(10). https://doi.org/10.1093/bioinformatics/btad595
54. Rhie A, Walenz BP, Koren S, Phillippy AM. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol*. 2020; 21(1): 245. https://doi.org/10.1186/s13059-020-02134-9
55. Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A. Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics*. 2018; 34(13): i142-i150. https://doi.org/10.1093/bioinformatics/bty266
56. Chen Y, Zhang Y, Wang AY, Gao M, Chong Z. Accurate long-read de novo assembly evaluation with Inspector. *Genome Biol*. 2021; 22(1): 312. https://doi.org/10.1186/s13059-021-02527-4
57. Hansen NF, Dwarshuis N, Ji HJ, Rhie A, Loucks H, Logsdon GA, et al. A complete diploid human genome benchmark for personalized genomics. *Cell*. 2026; 189(16): 4857-4875.e31. https://doi.org/10.1016/j.cell.2026.06.016
58. Logsdon GA, Ebert P, Audano PA, Loftus M, Porubsky D, Ebler J, et al. Complex genetic variation in nearly complete human genomes. *Nature*. 2025; 644(8076): 430-441. https://doi.org/10.1038/s41586-025-09140-6
59. He Y, Chu Y, Guo S, Hu J, Li R, Zheng Y, et al. T2T-YAO: A Telomere-to-telomere Assembled Diploid Reference Genome for Han Chinese. *Genomics Proteomics Bioinformatics*. 2023; 21(6): 1085-1100. https://doi.org/10.1016/j.gpb.2023.08.001
60. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009; 25(14): 1754-60. https://doi.org/10.1093/bioinformatics/btp324
61. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*. 2013; 1303.3997. https://doi.org/10.48550/arXiv.1303.3997
62. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. 2018; 34(18): 3094-3100. https://doi.org/10.1093/bioinformatics/bty191
63. Li H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics*. 2021; 37(23): 4572-4574. https://doi.org/10.1093/bioinformatics/btab705
64. Jain C, Rhie A, Hansen NF, Koren S, Phillippy AM. Long-read mapping to repetitive reference sequences using Winnowmap2. *Nat Methods*. 2022; 19(6): 705-710. https://doi.org/10.1038/s41592-022-01457-8
65. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. *Gigascience*. 2021; 10(2). https://doi.org/10.1093/gigascience/giab008
66. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010; 20(9): 1297-303. https://doi.org/10.1101/gr.107524.110
67. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*. 2011; 43(5): 491-8. https://doi.org/10.1038/ng.806
68. Behera S, Catreux S, Rossi M, Truong S, Huang Z, Ruehle M, et al. Comprehensive genome analysis and variant detection at scale using DRAGEN. *Nat Biotechnol*. 2025; 43(7): 1177-1191. https://doi.org/10.1038/s41587-024-02382-1
69. Poplin R, Chang PC, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol*. 2018; 36(10): 983-987. https://doi.org/10.1038/nbt.4235

70. Yun T, Li H, Chang PC, Lin MF, Carroll A, McLean CY. Accurate, scalable cohort variant calls using DeepVariant and GLnexus. *Bioinformatics*. 2021; 36(24): 5582-5589. https://doi.org/10.1093/bioinformatics/btaa1081
71. Zheng Z, Li S, Su J, Leung AWS, Lam TW, Luo R. Symphonizing pileup and full-alignment for deep learning-based long-read variant calling. *Nat Comput Sci*. 2022; 2(12): 797-803. https://doi.org/10.1038/s43588-022-00387-x
72. Shafin K, Pesout T, Chang PC, Nattestad M, Kolesnikov A, Goel S, et al. Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. *Nat Methods*. 2021; 18(11): 1322-1332. https://doi.org/10.1038/s41592-021-01299-w
73. Krusche P, Trigg L, Boutros PC, Mason CE, De La Vega FM, Moore BL, et al. Best practices for benchmarking germline small-variant calls in human genomes. *Nat Biotechnol*. 2019; 37(5): 555-560. https://doi.org/10.1038/s41587-019-0054-x
74. Olson ND, Wagner J, McDaniel J, Stephens SH, Westreich ST, Prasanna AG, et al. PrecisionFDA Truth Challenge V2: Calling variants from short and long reads in difficult-to-map regions. *Cell Genom*. 2022; 2(5): 100129. https://doi.org/10.1016/j.xgen.2022.100129
75. Wagner J, Olson ND, Harris L, McDaniel J, Cheng H, Fungtammasan A, et al. Curated variation benchmarks for challenging medically relevant autosomal genes. *Nat Biotechnol*. 2022; 40(5): 672-680. https://doi.org/10.1038/s41587-021-01158-1
76. Aganezov S, Yan SM, Soto DC, Kirsche M, Zarate S, Avdeyev P, et al. A complete reference genome improves analysis of human genetic variation. *Science*. 2022; 376(6588): eabl3533. https://doi.org/10.1126/science.abl3533
77. Ebbert MTW, Jensen TD, Jansen-West K, Sens JP, Reddy JS, Ridge PG, et al. Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight. *Genome Biol*. 2019; 20(1): 97. https://doi.org/10.1186/s13059-019-1707-2
78. Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat Commun*. 2019; 10(1): 1784. https://doi.org/10.1038/s41467-018-08148-z
79. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*. 2012; 28(18): i333-i339. https://doi.org/10.1093/bioinformatics/bts378
80. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol*. 2014; 15(6): R84. https://doi.org/10.1186/gb-2014-15-6-r84
81. Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*. 2016; 32(8): 1220-2. https://doi.org/10.1093/bioinformatics/btv710
82. Kosugi S, Momozawa Y, Liu X, Terao C, Kubo M, Kamatani Y. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol*. 2019; 20(1): 117. https://doi.org/10.1186/s13059-019-1720-5
83. Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. Structural variant calling: the long and the short of it. *Genome Biol*. 2019; 20(1): 246. https://doi.org/10.1186/s13059-019-1828-7
84. Ho SS, Urban AE, Mills RE. Structural variation in the sequencing era. *Nat Rev Genet*. 2020; 21(3): 171-189. https://doi.org/10.1038/s41576-019-0180-9
85. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. *Nature*. 2015; 526(7571): 75-81. https://doi.org/10.1038/nature15394
86. Collins RL, Brand H, Karczewski KJ, Zhao X, Alföldi J, Francioli LC, et al. A structural variation reference for medical and population genetics. *Nature*. 2020; 581(7809): 444-451. https://doi.org/10.1038/s41586-020-2287-8
87. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, et al. Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods*. 2018; 15(6): 461-468. https://doi.org/10.1038/s41592-018-0001-7

88. Smolka M, Paulin LF, Grochowski CM, Horner DW, Mahmoud M, Behera S, et al. Detection of mosaic and population-level structural variants with Sniffles2. *Nat Biotechnol*. 2024; 42(10): 1571-1580. https://doi.org/10.1038/s41587-023-02024-y
89. Jiang T, Liu Y, Jiang Y, Li J, Gao Y, Cui Z, et al. Long-read-based human genomic structural variation detection with cuteSV. *Genome Biol*. 2020; 21(1): 189. https://doi.org/10.1186/s13059-020-02107-y
90. Heller D, Vingron M. SVIM: structural variant identification using mapped long reads. *Bioinformatics*. 2019; 35(17): 2907-2915. https://doi.org/10.1093/bioinformatics/btz041
91. Huddleston J, Chaisson MJP, Steinberg KM, Warren W, Hoekzema K, Gordon D, et al. Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res*. 2017; 27(5): 677-685. https://doi.org/10.1101/gr.214007.116
92. Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*. 2021; 372(6537): eabf7117. https://doi.org/10.1126/science.abf7117
93. Kirsche M, Prabhu G, Sherman R, Ni B, Battle A, Aganezov S, et al. Jasmine and Iris: population-scale structural variant comparison and analysis. *Nat Methods*. 2023; 20(3): 408-417. https://doi.org/10.1038/s41592-022-01753-3
94. English AC, Menon VK, Gibbs RA, Metcalf GA, Sedlazeck FJ. Truvari: refined structural variant comparison preserves allelic diversity. *Genome Biol*. 2022; 23(1): 271. https://doi.org/10.1186/s13059-022-02840-6
95. Zook JM, Hansen NF, Olson ND, Chapman L, Mullikin JC, Xiao C, et al. A robust benchmark for detection of germline large deletions and insertions. *Nat Biotechnol*. 2020; 38(11): 1347-1355. https://doi.org/10.1038/s41587-020-0538-8
96. Gustafson JA, Gibson SB, Damaraju N, Zalusky MPG, Hoekzema K, Twesigomwe D, et al. High-coverage nanopore sequencing of samples from the 1000 Genomes Project to build a comprehensive catalog of human genetic variation. *Genome Res*. 2024; 34(11): 2061-2073. https://doi.org/10.1101/gr.279273.124
97. Schloissnig S, Pani S, Ebler J, Hain C, Tsapalou V, Söylev A, et al. Structural variation in 1,019 diverse humans based on long-read sequencing. *Nature*. 2025; 644(8076): 442-452. https://doi.org/10.1038/s41586-025-09290-7
98. Kolmogorov M, Billingsley KJ, Mastoras M, Meredith M, Monlong J, Lorig-Roach R, et al. Scalable Nanopore sequencing of human genomes provides a comprehensive view of haplotype-resolved variation and methylation. *Nat Methods*. 2023; 20(10): 1483-1492. https://doi.org/10.1038/s41592-023-01993-x
99. Patterson M, Marschall T, Pisanti N, van Iersel L, Stougie L, Klau GW, et al. WhatsHap: Weighted Haplotype Assembly for Future-Generation Sequencing Reads. *J Comput Biol*. 2015; 22(6): 498-509. https://doi.org/10.1089/cmb.2014.0157
100. Holt JM, Saunders CT, Rowell WJ, Kronenberg Z, Wenger AM, Eberle M. HiPhase: jointly phasing small, structural, and tandem repeat variants from HiFi sequencing. *Bioinformatics*. 2024; 40(2). https://doi.org/10.1093/bioinformatics/btae042
101. Delaneau O, Zagury JF, Robinson MR, Marchini JL, Dermitzakis ET. Accurate, scalable and integrative haplotype estimation. *Nat Commun*. 2019; 10(1): 5436. https://doi.org/10.1038/s41467-019-13225-y
102. Browning BL, Tian X, Zhou Y, Browning SR. Fast two-stage phasing of large-scale sequence data. *Am J Hum Genet*. 2021; 108(10): 1880-1890. https://doi.org/10.1016/j.ajhg.2021.08.005
103. Hofmeister RJ, Ribeiro DM, Rubinacci S, Delaneau O. Accurate rare variant phasing of whole-genome and whole-exome sequencing data in the UK Biobank. *Nat Genet*. 2023; 55(7): 1243-1249. https://doi.org/10.1038/s41588-023-01415-w
104. Rubinacci S, Ribeiro DM, Hofmeister RJ, Delaneau O. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. *Nat Genet*. 2021; 53(1): 120-126. https://doi.org/10.1038/s41588-020-00756-0
105. 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*. 2015; 526(7571): 68-74. https://doi.org/10.1038/nature15393

106. Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, et al. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell*. 2022; 185(18): 3426-3440.e19. https://doi.org/10.1016/j.cell.2022.08.004
107. Bergström A, McCarthy SA, Hui R, Almarri MA, Ayub Q, Danecek P, et al. Insights into human genetic variation and population history from 929 diverse genomes. *Science*. 2020; 367(6484): eaay5012. https://doi.org/10.1126/science.aay5012
108. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*. 2016; 538(7624): 201-206. https://doi.org/10.1038/nature18964
109. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*. 2021; 590(7845): 290-299. https://doi.org/10.1038/s41586-021-03205-y
110. Halldorsson BV, Eggertsson HP, Moore KHS, Hauswedell H, Eiriksson O, Ulfarsson MO, et al. The sequences of 150,119 genomes in the UK Biobank. *Nature*. 2022; 607(7920): 732-740. https://doi.org/10.1038/s41586-022-04965-x
111. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. 2016; 536(7616): 285-91. https://doi.org/10.1038/nature19057
112. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 2020; 581(7809): 434-443. https://doi.org/10.1038/s41586-020-2308-7
113. Chen S, Francioli LC, Goodrich JK, Collins RL, Kanai M, Wang Q, et al. A genomic mutational constraint map using variation in 76,156 human genomes. *Nature*. 2024; 625(7993): 92-100. https://doi.org/10.1038/s41586-023-06045-0
114. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. *Genome Biol*. 2016; 17(1): 122. https://doi.org/10.1186/s13059-016-0974-4
115. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)*. 2012; 6(2): 80-92. https://doi.org/10.4161/fly.19695
116. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010; 38(16): e164. https://doi.org/10.1093/nar/gkq603
117. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res*. 2018; 46(D1): D1062-D1067. https://doi.org/10.1093/nar/gkx1153
118. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res*. 2019; 47(D1): D886-D894. https://doi.org/10.1093/nar/gky1016
119. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li YI, et al. Predicting Splicing from Primary Sequence with Deep Learning. *Cell*. 2019; 176(3): 535-548.e24. https://doi.org/10.1016/j.cell.2018.12.015
120. Cheng J, Novati G, Pan J, Bycroft C, Žemgulytė A, Applebaum T, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science*. 2023; 381(6664): eadg7492. https://doi.org/10.1126/science.adg7492
121. Avsec Ž, Latysheva N, Cheng J, Novati G, Taylor KR, Ward T, et al. Advancing regulatory variant effect prediction with AlphaGenome. *Nature*. 2026; 649(8099): 1206-1218. https://doi.org/10.1038/s41586-025-10014-0
122. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*. 2015; 17(5): 405-24. https://doi.org/10.1038/gim.2015.30

123. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet*. 2011; 89(1): 82-93. https://doi.org/10.1016/j.ajhg.2011.05.029
124. Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet*. 2012; 91(2): 224-37. https://doi.org/10.1016/j.ajhg.2012.06.007
125. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. *Am J Hum Genet*. 2014; 95(1): 5-23. https://doi.org/10.1016/j.ajhg.2014.06.009
126. Zhou W, Zhao Z, Nielsen JB, Fritsche LG, LeFaive J, Gagliano Taliun SA, et al. Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts. *Nat Genet*. 2020; 52(6): 634-639. https://doi.org/10.1038/s41588-020-0621-6
127. Zhou W, Bi W, Zhao Z, Dey KK, Jagadeesh KA, Karczewski KJ, et al. SAIGE-GENE+ improves the efficiency and accuracy of set-based rare variant association tests. *Nat Genet*. 2022; 54(10): 1466-1469. https://doi.org/10.1038/s41588-022-01178-w
128. Mbatchou J, Barnard L, Backman J, Marcketta A, Kosmicki JA, Ziyatdinov A, et al. Computationally efficient whole-genome regression for quantitative and binary traits. *Nat Genet*. 2021; 53(7): 1097-1103. https://doi.org/10.1038/s41588-021-00870-7
129. Backman JD, Li AH, Marcketta A, Sun D, Mbatchou J, Kessler MD, et al. Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature*. 2021; 599(7886): 628-634. https://doi.org/10.1038/s41586-021-04103-z
130. Li X, Li Z, Zhou H, Gaynor SM, Liu Y, Chen H, et al. Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nat Genet*. 2020; 52(9): 969-983. https://doi.org/10.1038/s41588-020-0676-4
131. Li Z, Li X, Zhou H, Gaynor SM, Selvaraj MS, Arapoglou T, et al. A framework for detecting noncoding rare-variant associations of large-scale whole-genome sequencing studies. *Nat Methods*. 2022; 19(12): 1599-1611. https://doi.org/10.1038/s41592-022-01640-x
132. Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, et al. Rate of de novo mutations and the importance of father's age to disease risk. *Nature*. 2012; 488(7412): 471-5. https://doi.org/10.1038/nature11396
133. Belyeu JR, Brand H, Wang H, Zhao X, Pedersen BS, Feusier J, et al. De novo structural mutation rates and gamete-of-origin biases revealed through genome sequencing of 2,396 families. *Am J Hum Genet*. 2021; 108(4): 597-607. https://doi.org/10.1016/j.ajhg.2021.02.012
134. Porubsky D, Dashnow H, Sasani TA, Logsdon GA, Hallast P, Noyes MD, et al. Human de novo mutation rates from a four-generation pedigree reference. *Nature*. 2025; 643(8071): 427-436. https://doi.org/10.1038/s41586-025-08922-2
135. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*. 2006; 38(8): 904-9. https://doi.org/10.1038/ng1847
136. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*. 2009; 19(9): 1655-64. https://doi.org/10.1101/gr.094052.109
137. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*. 2015; 4: 7. https://doi.org/10.1186/s13742-015-0047-8
138. Korneliussen TS, Albrechtsen A, Nielsen R. ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics*. 2014; 15(1): 356. https://doi.org/10.1186/s12859-014-0356-4
139. Maples BK, Gravel S, Kenny EE, Bustamante CD. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am J Hum Genet*. 2013; 93(2): 278-88. https://doi.org/10.1016/j.ajhg.2013.06.020
140. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, et al. Ancient admixture in human history. *Genetics*. 2012; 192(3): 1065-93. https://doi.org/10.1534/genetics.112.145037

141. Li H, Durbin R. Inference of human population history from individual whole-genome sequences. *Nature*. 2011; 475(7357): 493-6. https://doi.org/10.1038/nature10231
142. Schiffels S, Durbin R. Inferring human population size and separation history from multiple genome sequences. *Nat Genet*. 2014; 46(8): 919-25. https://doi.org/10.1038/ng.3015
143. Terhorst J, Kamm JA, Song YS. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nat Genet*. 2017; 49(2): 303-309. https://doi.org/10.1038/ng.3748
144. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet*. 2009; 5(10): e1000695. https://doi.org/10.1371/journal.pgen.1000695
145. Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. Robust demographic inference from genomic and SNP data. *PLoS Genet*. 2013; 9(10): e1003905. https://doi.org/10.1371/journal.pgen.1003905
146. Rasmussen MD, Hubisz MJ, Gronau I, Siepel A. Genome-wide inference of ancestral recombination graphs. *PLoS Genet*. 2014; 10(5): e1004342. https://doi.org/10.1371/journal.pgen.1004342
147. Speidel L, Forest M, Shi S, Myers SR. A method for genome-wide genealogy estimation for thousands of samples. *Nat Genet*. 2019; 51(9): 1321-1329. https://doi.org/10.1038/s41588-019-0484-x
148. Kelleher J, Wong Y, Wohns AW, Fadil C, Albers PK, McVean G. Inferring whole-genome histories in large population datasets. *Nat Genet*. 2019; 51(9): 1330-1338. https://doi.org/10.1038/s41588-019-0483-y
149. Zhang BC, Biddanda A, Gunnarsson ÁF, Cooper F, Palamara PF. Biobank-scale inference of ancestral recombination graphs enables genealogical analysis of complex traits. *Nat Genet*. 2023; 55(5): 768-776. https://doi.org/10.1038/s41588-023-01379-x
150. Deng Y, Nielsen R, Song YS. Robust and accurate Bayesian inference of genome-wide genealogies for hundreds of genomes. *Nat Genet*. 2025; 57(9): 2124-2135. https://doi.org/10.1038/s41588-025-02317-9
151. Wohns AW, Wong Y, Jeffery B, Akbari A, Mallick S, Pinhasi R, et al. A unified genealogy of modern and ancient genomes. *Science*. 2022; 375(6583): eabi8264. https://doi.org/10.1126/science.abi8264
152. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. *PLoS Biol*. 2006; 4(3): e72. https://doi.org/10.1371/journal.pbio.0040072
153. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, et al. Genome-wide detection and characterization of positive selection in human populations. *Nature*. 2007; 449(7164): 913-8. https://doi.org/10.1038/nature06250
154. Szpiech ZA, Hernandez RD. selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. *Mol Biol Evol*. 2014; 31(10): 2824-7. https://doi.org/10.1093/molbev/msu211
155. Field Y, Boyle EA, Telis N, Gao Z, Gaulton KJ, Golan D, et al. Detection of human adaptation during the past 2000 years. *Science*. 2016; 354(6313): 760-764. https://doi.org/10.1126/science.aag0776
156. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A draft sequence of the Neandertal genome. *Science*. 2010; 328(5979): 710-722. https://doi.org/10.1126/science.1188021
157. Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, et al. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*. 2014; 505(7481): 43-9. https://doi.org/10.1038/nature12886
158. Browning SR, Browning BL, Zhou Y, Tucci S, Akey JM. Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture. *Cell*. 2018; 173(1): 53-61.e9. https://doi.org/10.1016/j.cell.2018.02.031
159. Skov L, Hui R, Shchur V, Hobolth A, Scally A, Schierup MH, et al. Detecting archaic introgression using an unadmixed outgroup. *PLoS Genet*. 2018; 14(9): e1007641. https://doi.org/10.1371/journal.pgen.1007641

160. Kairov U, Molkenov A, Rakhimova S, Kozhamkulov U, Sharip A, Karabayev D, et al. Whole-genome sequencing data of Kazakh individuals. *BMC Res Notes*. 2021; 14(1): 45. https://doi.org/10.1186/s13104-021-05464-4
161. Kairov U, Molkenov A, Sharip A, Rakhimova S, Seidualy M, Rhie A, et al. Whole-Genome Sequencing and Genomic Variant Analysis of Kazakh Individuals. *Front Genet*. 2022; 13: 902804. https://doi.org/10.3389/fgene.2022.902804
162. Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc Natl Acad Sci U S A*. 2003; 100(20): 11484-9. https://doi.org/10.1073/pnas.1932072100
163. Kiełbasa SM, Wan R, Sato K, Horton P, Frith MC. Adaptive seeds tame genomic sequence comparison. *Genome Res*. 2011; 21(3): 487-93. https://doi.org/10.1101/gr.113985.110
164. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: A fast and versatile genome alignment system. *PLoS Comput Biol*. 2018; 14(1): e1005944. https://doi.org/10.1371/journal.pcbi.1005944
165. Marco-Sola S, Moure JC, Moreto M, Espinosa A. Fast gap-affine pairwise alignment using the wavefront algorithm. *Bioinformatics*. 2021; 37(4): 456-463. https://doi.org/10.1093/bioinformatics/btaa777
166. Hickey G, Paten B, Earl D, Zerbino D, Haussler D. HAL: a hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics*. 2013; 29(10): 1341-2. https://doi.org/10.1093/bioinformatics/btt128
167. Armstrong J, Hickey G, Diekhans M, Fiddes IT, Novak AM, Deran A, et al. Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature*. 2020; 587(7833): 246-251. https://doi.org/10.1038/s41586-020-2871-y
168. Goel M, Sun H, Jiao WB, Schneeberger K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol*. 2019; 20(1): 277. https://doi.org/10.1186/s13059-019-1911-0
169. Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res*. 2012; 40(7): e49. https://doi.org/10.1093/nar/gkr1293
170. Lovell JT, Sreedasyam A, Schranz ME, Wilson M, Carlson JW, Harkess A, et al. GENESPACE tracks regions of interest and gene copy number variation across multiple genomes. *Elife*. 2022; 11: e78526. https://doi.org/10.7554/eLife.78526
171. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol*. 2019; 20(1): 238. https://doi.org/10.1186/s13059-019-1832-y
172. Kirilenko BM, Munegowda C, Osipova E, Jebb D, Sharma V, Blumer M, et al. Integrating gene annotation with orthology inference at scale. *Science*. 2023; 380(6643): eabn3107. https://doi.org/10.1126/science.abn3107
173. Shumate A, Salzberg SL. Liftoff: accurate mapping of gene annotations. *Bioinformatics*. 2021; 37(12): 1639-1643. https://doi.org/10.1093/bioinformatics/btaa1016
174. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*. 2005; 15(8): 1034-50. https://doi.org/10.1101/gr.3715005
175. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res*. 2010; 20(1): 110-21. https://doi.org/10.1101/gr.097857.109
176. Christmas MJ, Kaplow IM, Genereux DP, Dong MX, Hughes GM, Li X, et al. Evolutionary constraint and innovation across hundreds of placental mammals. *Science*. 2023; 380(6643): eabn3943. https://doi.org/10.1126/science.abn3943
177. Sullivan PF, Meadows JRS, Gazal S, Phan BN, Li X, Genereux DP, et al. Leveraging base-pair mammalian constraint to understand genetic variation and human disease. *Science*. 2023; 380(6643): eabn2937. https://doi.org/10.1126/science.abn2937

178. Kuderna LFK, Gao H, Janiak MC, Kuhlwilm M, Orkin JD, Bataillon T, et al. A global catalog of whole-genome diversity from 233 primate species. *Science*. 2023; 380(6648): 906-913. https://doi.org/10.1126/science.abn7829
179. Pollard KS, Salama SR, Lambert N, Lambot MA, Coppens S, Pedersen JS, et al. An RNA gene expressed during cortical development evolved rapidly in humans. *Nature*. 2006; 443(7108): 167-72. https://doi.org/10.1038/nature05113
180. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 2007; 24(8): 1586-91. https://doi.org/10.1093/molbev/msm088
181. Kosakovsky Pond SL, Poon AFY, Velazquez R, Weaver S, Hepler NL, Murrell B, et al. HyPhy 2.5: A Customizable Platform for Evolutionary Hypothesis Testing Using Phylogenies. *Mol Biol Evol*. 2020; 37(1): 295-299. https://doi.org/10.1093/molbev/msz197
182. Mendes FK, Vanderpool D, Fulton B, Hahn MW. CAFE 5 models variation in evolutionary rates among gene families. *Bioinformatics*. 2021; 36(22-23): 5516-5518. https://doi.org/10.1093/bioinformatics/btaa1022
183. Rivas-González I, Rousselle M, Li F, Zhou L, Dutheil JY, Munch K, et al. Pervasive incomplete lineage sorting illuminates speciation and selection in primates. *Science*. 2023; 380(6648): eabn4409. https://doi.org/10.1126/science.abn4409
184. Mao Y, Catacchio CR, Hillier LW, Porubsky D, Li R, Sulovari A, et al. A high-quality bonobo genome refines the analysis of hominid evolution. *Nature*. 2021; 594(7861): 77-81. https://doi.org/10.1038/s41586-021-03519-x
185. Mao Y, Harvey WT, Porubsky D, Munson KM, Hoekzema K, Lewis AP, et al. Structurally divergent and recurrently mutated regions of primate genomes. *Cell*. 2024; 187(6): 1547-1562.e13. https://doi.org/10.1016/j.cell.2024.01.052
186. Dennis MY, Nuttle X, Sudmant PH, Antonacci F, Graves TA, Nefedov M, et al. Evolution of human-specific neural SRGAP2 genes by incomplete segmental duplication. *Cell*. 2012; 149(4): 912-22. https://doi.org/10.1016/j.cell.2012.03.033
187. Fiddes IT, Lodewijk GA, Mooring M, Bosworth CM, Ewing AD, Mantalas GL, et al. Human-Specific NOTCH2NL Genes Affect Notch Signaling and Cortical Neurogenesis. *Cell*. 2018; 173(6): 1356-1369.e22. https://doi.org/10.1016/j.cell.2018.03.051
188. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, et al. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". *Proc Natl Acad Sci U S A*. 2005; 102(39): 13950-5. https://doi.org/10.1073/pnas.0506758102
189. Sherman RM, Forman J, Antonescu V, Puiu D, Daya M, Rafaels N, et al. Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat Genet*. 2019; 51(1): 30-35. https://doi.org/10.1038/s41588-018-0273-y
190. Eizenga JM, Novak AM, Sibbesen JA, Heumos S, Ghaffaari A, Hickey G, et al. Pangenome Graphs. *Annu Rev Genomics Hum Genet*. 2020; 21: 139-162. https://doi.org/10.1146/annurev-genom-120219-080406
191. Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Dawson ET, et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat Biotechnol*. 2018; 36(9): 875-879. https://doi.org/10.1038/nbt.4227
192. Li H, Feng X, Chu C. The design and construction of reference pangenome graphs with minigraph. *Genome Biol*. 2020; 21(1): 265. https://doi.org/10.1186/s13059-020-02168-z
193. Hickey G, Monlong J, Ebler J, Novak AM, Eizenga JM, Gao Y, et al. Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nat Biotechnol*. 2024; 42(4): 663-673. https://doi.org/10.1038/s41587-023-01793-w
194. Garrison E, Guarracino A, Heumos S, Villani F, Bao Z, Tattini L, et al. Building pangenome graphs. *Nat Methods*. 2024; 21(11): 2008-2012. https://doi.org/10.1038/s41592-024-02430-3
195. Guarracino A, Heumos S, Nahnsen S, Prins P, Garrison E. ODGI: understanding pangenome graphs. *Bioinformatics*. 2022; 38(13): 3319-3326. https://doi.org/10.1093/bioinformatics/btac308

196. Chin CS, Behera S, Khalak A, Sedlazeck FJ, Sudmant PH, Wagner J, et al. Multiscale analysis of pangenomes enables improved representation of genomic diversity for repetitive and clinically relevant genes. *Nat Methods*. 2023; 20(8): 1213-1221. https://doi.org/10.1038/s41592-023-01914-y
197. Li H, Marin M, Farhat MR. Exploring gene content with pangene graphs. *Bioinformatics*. 2024; 40(7). https://doi.org/10.1093/bioinformatics/btae456
198. Andreace F, Lechat P, Dufresne Y, Chikhi R. Comparing methods for constructing and representing human pangenome graphs. *Genome Biol*. 2023; 24(1): 274. https://doi.org/10.1186/s13059-023-03098-2
199. Gao Y, Yang X, Chen H, Tan X, Yang Z, Deng L, et al. A pangenome reference of 36 Chinese populations. *Nature*. 2023; 619(7968): 112-121. https://doi.org/10.1038/s41586-023-06173-7
200. Nassir N, Almarri MA, Kumail M, Mohamed N, Balan B, Hanif S, et al. A draft UAE-based Arab pangenome reference. *Nat Commun*. 2025; 16(1): 6747. https://doi.org/10.1038/s41467-025-61645-w
201. Sirén J, Monlong J, Chang X, Novak AM, Eizenga JM, Markello C, et al. Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science*. 2021; 374(6574): eabg8871. https://doi.org/10.1126/science.abg8871
202. Rautiainen M, Marschall T. GraphAligner: rapid and versatile sequence-to-graph alignment. *Genome Biol*. 2020; 21(1): 253. https://doi.org/10.1186/s13059-020-02157-2
203. Sirén J, Eskandar P, Ungaro MT, Hickey G, Eizenga JM, Novak AM, et al. Personalized pangenome references. *Nat Methods*. 2024; 21(11): 2017-2023. https://doi.org/10.1038/s41592-024-02407-2
204. Hickey G, Heller D, Monlong J, Sibbesen JA, Sirén J, Eizenga J, et al. Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biol*. 2020; 21(1): 35. https://doi.org/10.1186/s13059-020-1941-7
205. Ebler J, Ebert P, Clarke WE, Rausch T, Audano PA, Houwaart T, et al. Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Nat Genet*. 2022; 54(4): 518-525. https://doi.org/10.1038/s41588-022-01043-w
206. Prodanov T, Plender EG, Seebohm G, Meuth SG, Eichler EE, Marschall T. Locityper enables targeted genotyping of complex polymorphic genes. *Nat Genet*. 2025; 57(11): 2901-2908. https://doi.org/10.1038/s41588-025-02362-4
207. Groza C, Schwendinger-Schreck C, Cheung WA, Farrow EG, Thiffault I, Lake J, et al. Pangenome graphs improve the analysis of structural variants in rare genetic diseases. *Nat Commun*. 2024; 15(1): 657. https://doi.org/10.1038/s41467-024-44980-2
208. Guarracino A, Buonaiuto S, de Lima LG, Potapova T, Rhie A, Koren S, et al. Recombination between heterologous human acrocentric chromosomes. *Nature*. 2023; 617(7960): 335-343. https://doi.org/10.1038/s41586-023-05976-y
209. Vollger MR, Dishuck PC, Harvey WT, DeWitt WS, Guitart X, Goldberg ME, et al. Increased mutation and gene conversion within human segmental duplications. *Nature*. 2023; 617(7960): 325-334. https://doi.org/10.1038/s41586-023-05895-y
210. Hoyt SJ, Storer JM, Hartley GA, Grady PGS, Gershman A, de Lima LG, et al. From telomere to telomere: The transcriptional and epigenetic state of human repeat elements. *Science*. 2022; 376(6588): eabk3112. https://doi.org/10.1126/science.abk3112
211. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet*. 2007; 8(12): 973-82. https://doi.org/10.1038/nrg2165
212. Smit AFA, Hubley R, Green P. RepeatMasker Open-4.0. 2015. Available at: http://www.repeatmasker.org
213. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA*. 2015; 6: 11. https://doi.org/10.1186/s13100-015-0041-9
214. Storer J, Hubley R, Rosen J, Wheeler TJ, Smit AF. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mob DNA*. 2021; 12(1): 2. https://doi.org/10.1186/s13100-020-00230-y

215. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci U S A*. 2020; 117(17): 9451-9457. https://doi.org/10.1073/pnas.1921046117
216. Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol*. 2019; 20(1): 275. https://doi.org/10.1186/s13059-019-1905-y
217. Hu K, Ni P, Xu M, Zou Y, Chang J, Gao X, et al. HiTE: a fast and accurate dynamic boundary adjustment approach for full-length transposable element detection and annotation. *Nat Commun*. 2024; 15(1): 5573. https://doi.org/10.1038/s41467-024-49912-8
218. Goerner-Potvin P, Bourque G. Computational tools to unmask transposable elements. *Nat Rev Genet*. 2018; 19(11): 688-704. https://doi.org/10.1038/s41576-018-0050-x
219. Kazazian HH, Moran JV. Mobile DNA in Health and Disease. *N Engl J Med*. 2017; 377(4): 361-370. https://doi.org/10.1056/NEJMra1510092
220. Payer LM, Burns KH. Transposable elements in human genetic disease. *Nat Rev Genet*. 2019; 20(12): 760-772. https://doi.org/10.1038/s41576-019-0165-8
221. Chuong EB, Elde NC, Feschotte C. Regulatory activities of transposable elements: from conflicts to benefits. *Nat Rev Genet*. 2017; 18(2): 71-86. https://doi.org/10.1038/nrg.2016.139
222. Gardner EJ, Lam VK, Harris DN, Chuang NT, Scott EC, Pittard WS, et al. The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. *Genome Res*. 2017; 27(11): 1916-1929. https://doi.org/10.1101/gr.218032.116
223. Kojima S, Koyama S, Ka M, Saito Y, Parrish EH, Endo M, et al. Mobile element variation contributes to population-specific genome diversification, gene regulation and disease risk. *Nat Genet*. 2023; 55(6): 939-951. https://doi.org/10.1038/s41588-023-01390-2
224. Zhou W, Emery SB, Flasch DA, Wang Y, Kwan KY, Kidd JM, et al. Identification and characterization of occult human-specific LINE-1 insertions using long-read sequencing technology. *Nucleic Acids Res*. 2020; 48(3): 1146-1163. https://doi.org/10.1093/nar/gkz1173
225. Chu C, Borges-Monroy R, Viswanadham VV, Lee S, Li H, Lee EA, et al. Comprehensive identification of transposable element insertions using multiple sequencing technologies. *Nat Commun*. 2021; 12(1): 3836. https://doi.org/10.1038/s41467-021-24041-8
226. Ewing AD, Smits N, Sanchez-Luque FJ, Faivre J, Brennan PM, Richardson SR, et al. Nanopore Sequencing Enables Comprehensive Transposable Element Epigenomic Profiling. *Mol Cell*. 2020; 80(5): 915-928.e5. https://doi.org/10.1016/j.molcel.2020.10.024
227. Groza C, Chen X, Wheeler TJ, Bourque G, Goubert C. A unified framework to analyze transposable element insertion polymorphisms using graph genomes. *Nat Commun*. 2024; 15(1): 8915. https://doi.org/10.1038/s41467-024-53294-2
228. Gershman A, Sauria MEG, Guitart X, Vollger MR, Hook PW, Hoyt SJ, et al. Epigenetic patterns in a complete human genome. *Science*. 2022; 376(6588): eabj5089. https://doi.org/10.1126/science.abj5089
229. Altemose N, Logsdon GA, Bzikadze AV, Sidhwani P, Langley SA, Caldas GV, et al. Complete genomic and epigenetic maps of human centromeres. *Science*. 2022; 376(6588): eabl4178. https://doi.org/10.1126/science.abl4178
230. Miga KH, Alexandrov IA. Variation and Evolution of Human Centromeres: A Field Guide and Perspective. *Annu Rev Genet*. 2021; 55: 583-602. https://doi.org/10.1146/annurev-genet-071719-020519
231. Bzikadze AV, Pevzner PA. Automated assembly of centromeres from ultra-long error-prone reads. *Nat Biotechnol*. 2020; 38(11): 1309-1316. https://doi.org/10.1038/s41587-020-0582-4
232. Dvorkina T, Bzikadze AV, Pevzner PA. The string decomposition problem and its applications to centromere analysis and assembly. *Bioinformatics*. 2020; 36(Suppl_1): i93-i101. https://doi.org/10.1093/bioinformatics/btaa454
233. Kunyavskaya O, Dvorkina T, Bzikadze AV, Alexandrov IA, Pevzner PA. Automated annotation of human centromeres with HORmon. *Genome Res*. 2022; 32(6): 1137-1151. https://doi.org/10.1101/gr.276362.121

234. Zhang Y, Chu J, Cheng H, Li H. De novo reconstruction of satellite repeat units from sequence data. *Genome Res*. 2023; 33(11): 1994-2001. https://doi.org/10.1101/gr.278005.123
235. Wlodzimierz P, Hong M, Henderson IR. TRASH: Tandem Repeat Annotation and Structural Hierarchy. *Bioinformatics*. 2023; 39(5). https://doi.org/10.1093/bioinformatics/btad308
236. Vollger MR, Kerpedjiev P, Phillippy AM, Eichler EE. StainedGlass: interactive visualization of massive tandem repeat structures with identity heatmaps. *Bioinformatics*. 2022; 38(7): 2049-2051. https://doi.org/10.1093/bioinformatics/btac018
237. Sweeten AP, Schatz MC, Phillippy AM. ModDotPlot: rapid and interactive visualization of tandem repeats. *Bioinformatics*. 2024; 40(8). https://doi.org/10.1093/bioinformatics/btae493
238. Logsdon GA, Rozanski AN, Ryabov F, Potapova T, Shepelev VA, Catacchio CR, et al. The variation and evolution of complete human centromeres. *Nature*. 2024; 629(8010): 136-145. https://doi.org/10.1038/s41586-024-07278-3
239. Hannan AJ. Tandem repeats mediating genetic plasticity in health and disease. *Nat Rev Genet*. 2018; 19(5): 286-298. https://doi.org/10.1038/nrg.2017.115
240. Chaisson MJP, Sulovari A, Valdmanis PN, Miller DE, Eichler EE. Advances in the discovery and analyses of human tandem repeats. *Emerg Top Life Sci*. 2023; 7(3): 361-381. https://doi.org/10.1042/ETLS20230074
241. Tanudisastro HA, Deveson IW, Dashnow H, MacArthur DG. Sequencing and characterizing short tandem repeats in the human genome. *Nat Rev Genet*. 2024; 25(7): 460-475. https://doi.org/10.1038/s41576-024-00692-3
242. Depienne C, Mandel JL. 30 years of repeat expansion disorders: What have we learned and what are the remaining challenges? *Am J Hum Genet*. 2021; 108(5): 764-785. https://doi.org/10.1016/j.ajhg.2021.03.011
243. English AC, Dolzhenko E, Ziaei Jam H, McKenzie SK, Olson ND, De Coster W, et al. Analysis and benchmarking of small and large genomic variants across tandem repeats. *Nat Biotechnol*. 2025; 43(3): 431-442. https://doi.org/10.1038/s41587-024-02225-z
244. Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res*. 1999; 27(2): 573-80. https://doi.org/10.1093/nar/27.2.573
245. Gymrek M, Golan D, Rosset S, Erlich Y. lobSTR: A short tandem repeat profiler for personal genomes. *Genome Res*. 2012; 22(6): 1154-62. https://doi.org/10.1101/gr.135780.111
246. Willems T, Zielinski D, Yuan J, Gordon A, Gymrek M, Erlich Y. Genome-wide profiling of heritable and de novo STR variations. *Nat Methods*. 2017; 14(6): 590-592. https://doi.org/10.1038/nmeth.4267
247. Dolzhenko E, Deshpande V, Schlesinger F, Krusche P, Petrovski R, Chen S, et al. ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. *Bioinformatics*. 2019; 35(22): 4754-4756. https://doi.org/10.1093/bioinformatics/btz431
248. Mousavi N, Shleizer-Burko S, Yanicky R, Gymrek M. Profiling the genome-wide landscape of tandem repeat expansions. *Nucleic Acids Res*. 2019; 47(15): e90. https://doi.org/10.1093/nar/gkz501
249. Dashnow H, Lek M, Phipson B, Halman A, Sadedin S, Lonsdale A, et al. STRetch: detecting and discovering pathogenic short tandem repeat expansions. *Genome Biol*. 2018; 19(1): 121. https://doi.org/10.1186/s13059-018-1505-2
250. Dashnow H, Pedersen BS, Hiatt L, Brown J, Beecroft SJ, Ravenscroft G, et al. STRling: a k-mer counting approach that detects short tandem repeat expansions at known and novel loci. *Genome Biol*. 2022; 23(1): 257. https://doi.org/10.1186/s13059-022-02826-4
251. Cortese A, Simone R, Sullivan R, Vandrovcova J, Tariq H, Yau WY, et al. Biallelic expansion of an intronic repeat in RFC1 is a common cause of late-onset ataxia. *Nat Genet*. 2019; 51(4): 649-658. https://doi.org/10.1038/s41588-019-0372-4
252. Ibañez K, Polke J, Hagelstrom RT, Dolzhenko E, Pasko D, Thomas ERA, et al. Whole genome sequencing for the diagnosis of neurological repeat expansion disorders in the UK: a retrospective diagnostic accuracy and prospective clinical validation study. *Lancet Neurol*. 2022; 21(3): 234-245. https://doi.org/10.1016/S1474-4422%2821%2900462-2

253. Ziaei Jam H, Li Y, DeVito R, Mousavi N, Ma N, Lujumba I, et al. A deep population reference panel of tandem repeat variation. *Nat Commun*. 2023; 14(1): 6711. https://doi.org/10.1038/s41467-023-42278-3
254. Cui Y, Ye W, Li JS, Li JJ, Vilain E, Sallam T, et al. A genome-wide spectrum of tandem repeat expansions in 338,963 humans. *Cell*. 2024; 187(9): 2336-2341.e5. https://doi.org/10.1016/j.cell.2024.03.004
255. Dolzhenko E, English A, Dashnow H, De Sena Brandine G, Mokveld T, Rowell WJ, et al. Characterization and visualization of tandem repeats at genome scale. *Nat Biotechnol*. 2024; 42(10): 1606-1614. https://doi.org/10.1038/s41587-023-02057-3
256. Chiu R, Rajan-Babu IS, Friedman JM, Birol I. Straglr: discovering and genotyping tandem repeat expansions using whole genome long-read sequences. *Genome Biol*. 2021; 22(1): 224. https://doi.org/10.1186/s13059-021-02447-3
257. Ziaei Jam H, Zook JM, Javadzadeh S, Park J, Sehgal A, Gymrek M. LongTR: genome-wide profiling of genetic variation at tandem repeats from long reads. *Genome Biol*. 2024; 25(1): 176. https://doi.org/10.1186/s13059-024-03319-2
258. Ren J, Gu B, Chaisson MJP. vamos: variable-number tandem repeats annotation using efficient motif sets. *Genome Biol*. 2023; 24(1): 175. https://doi.org/10.1186/s13059-023-03010-y
259. Lu TY, Human Genome Structural Variation Consortium, Chaisson MJP. Profiling variable-number tandem repeat variation across populations using repeat-pangenome graphs. *Nat Commun*. 2021; 12(1): 4250. https://doi.org/10.1038/s41467-021-24378-0
260. Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, et al. Recent segmental duplications in the human genome. *Science*. 2002; 297(5583): 1003-7. https://doi.org/10.1126/science.1072047
261. Vollger MR, Guitart X, Dishuck PC, Mercuri L, Harvey WT, Gershman A, et al. Segmental duplications and their variation in a complete human genome. *Science*. 2022; 376(6588): eabj6965. https://doi.org/10.1126/science.abj6965
262. Numanagić I, Gökkaya AS, Zhang L, Berger B, Alkan C, Hach F. Fast characterization of segmental duplications in genome assemblies. *Bioinformatics*. 2018; 34(17): i706-i714. https://doi.org/10.1093/bioinformatics/bty586
263. Išerić H, Alkan C, Hach F, Numanagić I. Fast characterization of segmental duplication structure in multiple genome assemblies. *Algorithms Mol Biol*. 2022; 17(1): 4. https://doi.org/10.1186/s13015-022-00210-2
264. Chen X, Harting J, Farrow E, Thiffault I, Kasperaviciute D, Hoischen A, et al. Comprehensive SMN1 and SMN2 profiling for spinal muscular atrophy analysis using long-read PacBio HiFi sequencing. *Am J Hum Genet*. 2023; 110(2): 240-250. https://doi.org/10.1016/j.ajhg.2023.01.001
265. Kalendar R, Kairov U. Genome-Wide Tool for Sensitive de novo Identification and Visualisation of Interspersed and Tandem Repeats. *Bioinform Biol Insights*. 2024; 18: 11779322241306391. https://doi.org/10.1177/11779322241306391
266. Kalendar R. TotalRepeats: self-calibrated, alignment-free discovery and pangenome comparison of repeat families. Version 4.1.3 [computer software]. *Zenodo*. 2026. https://doi.org/10.5281/zenodo.13735264. Source code: https://github.com/rkalendar/TotalRepeats

Author Information

Full Name	Aizhamal Aitzhanova
Department	https://www.jacquemont-lab.org/team/aizhamal-aitzhanova
Institution	Université de Montréal, Montréal, Canada
Email	aizhamal.aitzhanova@umontreal.ca
ORCID	
Scopus ID	

Full Name	Rakhimbek Bektayev
Department	Laboratory of Bioinformatics and Systems Biology, Center for Life Sciences, National Laboratory Astana
Institution	Nazarbayev University, Astana, Kazakhstan
Email	rakhimbek.bektayev@nu.edu.kz
ORCID	0000-0002-7658-5952
Scopus ID	59408517300

Full Name	Asset Daniyarov
Department	Laboratory of Bioinformatics and Systems Biology, Center for Life Sciences, National Laboratory Astana
Institution	Nazarbayev University, Astana, Kazakhstan
Email	asset.daniyarov@nu.edu.kz
ORCID	0000-0003-3886-718X
Scopus ID	57204420341

Full Name	Aibek Karimov
Department	Laboratory of Bioinformatics and Systems Biology, Center for Life Sciences, National Laboratory Astana
Institution	Nazarbayev University, Astana, Kazakhstan
Email	aibek.karimov@nu.edu.kz
ORCID	0000-0001-5186-8561
Scopus ID	57216751025

Full Name	Sultan Makhmetov
Department	
Institution	University of Pavia, Pavia, Italy
Email	sultan.makhmetov01@universitadipavia.it
ORCID	0009-0007-5203-4834
Scopus ID	58749802600

Full Name	Anel Ordabayeva
Department	Laboratory of Bioinformatics and Systems Biology, Center for Life Sciences, National Laboratory Astana
Institution	Nazarbayev University, Astana, Kazakhstan
Email	anel.ordabayeva@nu.edu.kz
ORCID	0009-0009-8857-3199
Scopus ID	60387390600

Full Name	Kamila Temirkhanova
Department	Institute for Clinical Chemistry and Laboratory Medicine
Institution	TU Dresden, Dresden, Germany
Email	kamila.temirkhanova@ukdd.de
ORCID	0009-0001-7885-8758
Scopus ID	57366748300