A Hitchhiker's Guide to long-read genomic analysis

Medhat Mahmoud, Daniel P. Agustinho and Fritz J. Sedlazeck

Genome Res. 2025 35: 545-558
Access the most recent version at doi:10.1101/gr.279975.124

Mini-Review

A Hitchhiker's Guide to long-read genomic analysis

Medhat Mahmoud,[1,4] Daniel P. Agustinho,[1,4] and Fritz J. Sedlazeck[1,2,3]

[1]*Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas 77030, USA;* [2]*Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas 77030, USA;* [3]*Department of Computer Science, Rice University, Houston, Texas 77005, USA*

Over the past decade, long-read sequencing has evolved into a pivotal technology for uncovering the hidden and complex regions of the genome. Significant cost efficiency, scalability, and accuracy advancements have driven this evolution. Concurrently, novel analytical methods have emerged to harness the full potential of long reads. These advancements have enabled milestones such as the first fully completed human genome, enhanced identification and understanding of complex genomic variants, and deeper insights into the interplay between epigenetics and genomic variation. This mini-review provides a comprehensive overview of the latest developments in long-read DNA sequencing analysis, encompassing reference-based and de novo assembly approaches. We explore the entire workflow, from initial data processing to variant calling and annotation, focusing on how these methods improve our ability to interpret a wide array of genomic variants. Additionally, we discuss the current challenges, limitations, and future directions in the field, offering a detailed examination of the state-of-the-art bioinformatics methods for long-read sequencing.

Short-read sequencing revolutionized genomics by providing a fast and cost-effective method for sequencing entire genomes, establishing it as a cornerstone of modern genomic research (Heather and Chain 2016; Foox et al. 2021). The emergence of long-read sequencing, producing reads of ~10 kbp–4 Mbp, has enabled unprecedented insights into previously inaccessible genome regions, such as repetitive sequences (Sulovari et al. 2019; Nurk et al. 2022; Chaisson et al. 2023; Olson et al. 2023; Mahmoud et al. 2024b). In addition, long-read sequencing enabled the simultaneous assessment of genomic and epigenomic changes within complex regions (Logsdon et al. 2020; Mahmoud et al. 2021; Vollger et al. 2025). Nevertheless, long-read sequencing requires specialized analysis techniques to unlock its full potential, often requiring in-depth knowledge of rapidly evolving bioinformatic methods.

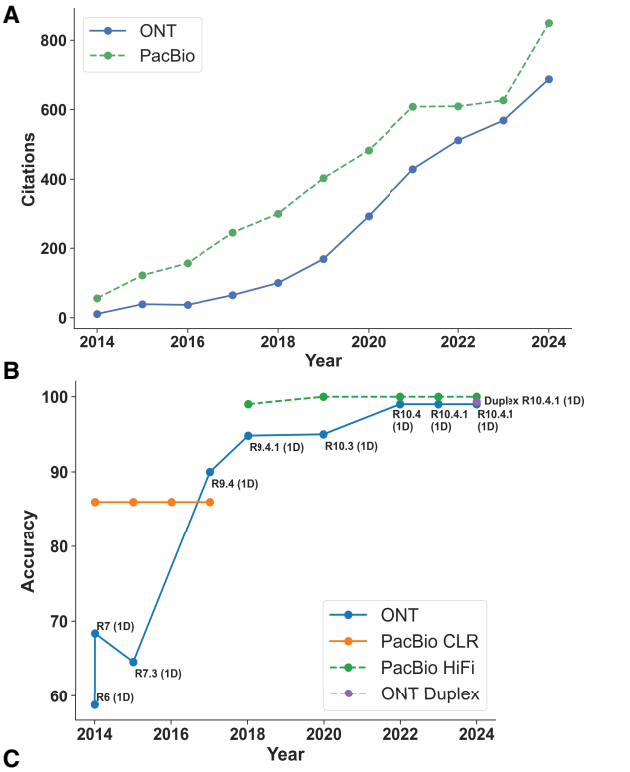
Two leading long-read sequencing technologies currently dominate the market and have significantly impacted the genomics field (Fig. 1A): Pacific Biosciences (PacBio) HiFi and Oxford Nanopore Technologies (ONT) sequencing. While both technologies produce continuous long reads, they present significant differences in accuracy, price point, read-length profiles, sample requirements, and the amount of DNA, where ONT requires less DNA than PacBio HiFi. DNA quality is crucial for the success of a long-read sequencing run (Oehler et al. 2023), and both PacBio and ONT have multiple extraction and preparation protocols for different organisms. The optimal choice of technology is contingent upon the specific application, with factors such as the complexity of variant calling or the desired level of assembly influencing the decision.

Initially, the adoption of long-read sequencing technologies suffered from both high error rates, at times up to 30%, and high costs, which have so far been significantly reduced over time (Fig. 1B). Currently, PacBio HiFi and ONT generate highly accurate long reads, exceeding 99% accuracy, expanding the applicability of long-read sequencing across diverse genomic studies (Wenger et al. 2019; Koren et al. 2024). Both platforms are capable of DNA and cDNA sequencing and detecting DNA methylation. At the same time, ONT offers additional functionalities such as adaptive sampling and direct RNA-seq (including epigenetic modifications). We recommend that readers interested in more details about the two platforms read the review by Logsdon et al. (2020).

Both PacBio and ONT excel in resolving repetitive elements and identifying complex genomic variants, including structural variants (SVs), which have historically posed challenges for short-read approaches (see Fig. 1C; Cameron et al. 2019; English et al. 2024b; Mahmoud et al. 2024b). SVs, defined as genomic alterations of 50 bp or more, encompass deletions, duplications, insertions, inversions, translocations, or a combination thereof (Escaramís et al. 2015; Collins et al. 2020). Furthermore, long-read utility expanded to include haplotype phasing, enabling the study of how genetic variants are inherited together. In addition, long-read sequencing has revolutionized methylation analysis, especially in repetitive regions. Long-read sequencing allows for the simultaneous determination of methylation levels and haplotypes (Gigante et al. 2019), a limited capability in short-read sequencing. This advancement facilitates the identification of differentially methylated regions and enhances our understanding of their potential impact on gene regulation and epigenetic mechanisms. Thus, long read enabled solving epigenetic-related diseases (Xie et al. 2021; Lucas and Novoa 2023), achieving a complete human genome assembly (Nurk et al. 2022), novel insights into repetitive regions (English et al. 2024b), accelerated diagnosis in various diseases (Gorzynski et al. 2022; Akagi et al. 2023; Lau et al. 2023), and comprehensive methylation maps (Gershman et al. 2022).

This mini-review delves into the analysis of long-read sequencing technologies, highlighting how these insights can be replicated and produced for samples of interest. We provide insight into some of the most prevalent long-read analysis approaches and assist in choosing individual analysis tools. This spans different aspects of analysis focusing on DNA, from reference-based

[4]**These authors contributed equally to this work.**
Corresponding author: fritz.sedlazeck@bcm.edu
Article and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.279975.124.



C

	SNVs	Indels	TRs	SVs	CNVs	Methylation	Phasing	Assembly
Short-Reads	+++	+++	++	+	++	+++**	+	+
Long-reads	+++	++	+++	+++	+++	+++	+++	+++

\+ Minimum performance +++ Maximum performance

Figure 1. Long-read accuracy, citation trends over time, and comparison to short reads. (*A*) Citations of PacBio and Oxford Nanopore Technologies (ONT) long-read sequencing publications from 2014 to the present demonstrate their growing impact in the field. We collected citations from PubMed and excluded review articles. (*B*) This figure presents the evolution of long-read sequencing accuracy over time for ONT (Ashton et al. 2015; Goodwin et al. 2015; Laver et al. 2015; Suzuki et al. 2017; Ferguson et al. 2022; Ni et al. 2023b; Sanderson et al. 2024) and PacBio (Wenger et al. 2019; Amarasinghe et al. 2020; Logsdon et al. 2020; Oxford Nanopore Technologies 2020), illustrating their progress toward achieving >99% accuracy. For ONT, the analysis focuses exclusively on the 1D technology, with the 2014 R7 (1D) and the Duplex data points representing the median value, while the remaining points represent the mean values. We excluded ONT's 2D and $1D^2$ technologies because they ceased production in 2016. The plot distinguishes between PacBio's continuous long read (CLR) and high fidelity (HiFi) technologies. (*C*) Comparison between short reads and long reads in variant calling accuracy, methylation calling, and genome assembly (Oehler et al. 2023; Ni et al. 2023a; Dolzhenko et al. 2024; Espinosa et al. 2024; Kosugi and Terao 2024; Höps et al. 2025), where one plus represents the minimum performance and three pluses represent the maximum performance. ** Short reads required biochemical treatment and were used as the benchmark for methylation.

mapping to de novo assembly. Although this mini-review mainly focuses on the human genome, many methods described here also apply to other species, while microbiomes have been recently reviewed elsewhere (Agustinho et al. 2024). For readers interested in long-read transcriptomics analyses, we refer them to other works (Calvo-Roitberg et al. 2024; Pardo-Palacios et al. 2024). We cover the alignment-based and de novo assembly-based variant detection, phasing, and annotation range. A comprehensive understanding of these analysis methods is crucial for unlocking the full potential of long-read sequencing in diverse research areas.

Considerations for long-read-based analysis

Long-read sequencing technologies, such as ONT and PacBio, generate raw electrical or optical signals (squiggle) that require specialized basecalling algorithms to convert into nucleotide sequences. ONT's basecaller is continuously updated, with Dorado being the latest development (Table 1). These updates improve basecalling accuracy, now approaching 99% (Fig. 1)—however, frequent updates could present challenges in clinical and multisample projects. Clinical workflows require reproducibility, consistency, and regulatory validation, making frequent revalidation necessary whenever the basecalling software changes. This can delay the integration of advancements into clinical pipelines and affect standardization across laboratories. PacBio's basecaller is integrated into the sequencing machine directly and is not publicly available. PacBio's accuracy is mainly driven by the generation of HiFi reads, where the DNA polymerase reads both forward and reverse strands of the same DNA molecule multiple times in a continuous loop, allowing the software to create a highly accurate consensus sequence from these multiple passes (Wenger et al. 2019). PacBio utilizes the circular consensus sequencing (CCS) method for PacBio platforms to collapse reads and improve their quality. In addition, Google DeepConsensus (Baid et al. 2023) is also available (Table 1). The choice between these approaches depends on the specific project requirements, such as the need for cutting-edge accuracy versus reproducibility and standardization. To ensure data quality, rigorous quality control (QC) is essential. Software like LongQC (Fukasawa et al. 2020) and NanoPack (De Coster et al. 2018) assess read length distribution, base quality, and other metrics, providing crucial insights for downstream analyses.

Following sequencing with either PacBio or ONT, researchers are confronted with a critical decision: using reference-based mapping or de novo assembly (Fig. 2). Reference-based mapping relies on a high-quality reference genome, while de novo assembly often demands longer reads and/or higher coverage (Coster et al. 2021). The choice of either analysis approach often impacts the experimental design itself (Harvey et al. 2023). Other factors influencing this decision include computational resources, DNA quality and quantity, sequencing depth, and the ability of assembly-based approaches to improve existing reference genomes (e.g., for nonmodel organisms). For instance, if only a fragmented and incomplete genome is available as a reference or if the goal is to analyze, e.g., segmental duplications, then a de novo assembly approach might be preferred. Conversely, a reference-based mapping is often more appropriate for projects aiming to interpret and compare variants.

Below, we discuss and provide insights and methods for the reference-guided alignment and de novo assembly approaches (Fig. 2). In addition, see Table 1 for a complete list of suggested tools for each analysis step.

Reference-guided analysis

Aligning reads to a reference genome (i.e., mapping) identifies the likely origin of a sequence read within the reference and helps identify sequence variations relative to a reference genome. These genomic variations encompass a spectrum of sizes, from single-nucleotide variants (SNVs) and short insertions or deletions (indels, <50 bp) to SVs (≥50 bp). Thus, identifying the proper reference genome for long-read alignment is critical

Table 1. List of methods for long-read analysis and its function

Function	Tool	Technology	GitHub	Citation
Basecalling (A)	CCS	PacBio	https://github.com/PacificBiosciences/ccs	
	Dorado	ONT	https://github.com/nanoporetech/dorado	
	Google Deep Consensus	PacBio	https://github.com/google/deepconsensus	Baid et al. 2023
Reads QC (B)	LongReadSum	ONT/PacBio	https://github.com/WGLab/LongReadSum	
	LongQC	ONT/PacBio	https://github.com/yfukasawa/LongQC	Fukasawa et al. 2020
	NanoPack	ONT/PacBio	https://github.com/wdecoster/nanopack	De Coster et al. 2018
Alignment (C)	LRA	ONT/PacBio	https://github.com/ChaissonLab/LRA	Ren and Chaisson 2021
	minimap2	ONT/PacBio	https://github.com/lh3/minimap2	Li 2018, 2021
	NGMLR	ONT/PacBio	https://github.com/philres/ngmlr	Sedlazeck et al. 2018b
	pbmm2	PacBio	https://github.com/PacificBiosciences/pbmm2	
	VACmap	ONT/PacBio	https://github.com/micahvista/VACmap	Ding et al. 2024
	Vulcan	ONT/PacBio	https://gitlab.com/treangenlab/vulcan	Fu et al. 2021
	Winnowmap2	ONT/PacBio	https://github.com/marbl/Winnowmap	Jain et al. 2022
Alignment QC (D)	NanoPack	ONT/PacBio	https://github.com/wdecoster/nanopack	De Coster et al. 2018
	Sambamba	ONT/PacBio	https://github.com/biod/sambamba	Tarasov et al. 2015
	SAMtools	ONT/PacBio	https://github.com/samtools/samtools	Li et al. 2009
SNV & indels calling (E)	Clair3	ONT/PacBio	https://github.com/HKU-BAL/Clair3	Zheng et al. 2022
	DeepSomatic	Illumina/ONT/PacBio	https://github.com/google/deepsomatic	Park et al. 2024
	DeepVariant	ONT/PacBio	https://github.com/google/deepvariant	Poplin et al. 2018
	Longshot	ONT/PacBio	https://github.com/pjedge/longshot	Edge and Bansal 2019
SV calling (F)	cuteSV	ONT/PacBio	https://github.com/tjiangHIT/cuteSV	Jiang et al. 2020
	DELLY	ONT/PacBio/Illumina	https://github.com/dellytools/delly?tab=readme-ov-file	Rausch et al. 2012
	NanomonSV	ONT/PacBio	https://github.com/friend1ws/nanomonsv	Shiraishi et al. 2023
	pbsv	PacBio	https://github.com/PacificBiosciences/pbsv	
	Sawfish	PacBio	https://github.com/PacificBiosciences/sawfish	Saunders et al. 2024
	Severus	ONT/PacBio	https://github.com/KolmogorovLab/Severus	Keskus et al. 2024
	Sniffles2	ONT/PacBio	https://github.com/fritzsedlazeck/Sniffles	Smolka et al. 2024
	SAVANA	ONT/PacBio	https://github.com/cortes-ciriano-lab/savana	Elrick et al. 2024
	SVision	ONT/PacBio	https://github.com/xjtu-omics/SVision	Lin et al. 2022b
Copy number variants (G)	HiFiCNV	PacBio	https://github.com/PacificBiosciences/HiFiCNV	
	Spectre	ONT/PacBio	https://github.com/fritzsedlazeck/Spectre	
Tandem repeat (TRs) (H)	Medaka	ONT	https://github.com/nanoporetech/medaka	Lee et al. 2021
	pathSTR	ONT	https://github.com/wdecoster/pathSTR	De Coster et al. 2024
	Straglr	ONT	https://github.com/bcgsc/straglr	Chiu et al. 2021
	StrSpy	ONT	https://github.com/unique379r/strspy	Hall et al. 2022
	TRGT	PacBio	https://github.com/PacificBiosciences/trgt	Dolzhenko et al. 2024
Genotyping (I)	kanpig	ONT/PacBio	https://github.com/ACEnglish/kanpig	English et al. 2024a
	SVJedi	ONT/PacBio	https://github.com/llecompte/SVJedi	Romain and Lemaitre 2023
Downstream analysis—phasing (J)	HapCUT2	ONT/PacBio	https://github.com/vibansal/HapCUT2	Edge et al. 2017
	HiPhase	PacBio	https://github.com/PacificBiosciences/HiPhase	Holt et al. 2024
	WhatsHap	ONT/PacBio	https://github.com/whatshap/whatshap	Martin et al. 2023

(*continued*)

Table 1. *Continued*

Function	Tool	Technology	GitHub	Citation
Downstream analysis—variant annotation (K)	AnnotSV	NA	https://github.com/lgmgeo/AnnotSV	Geoffroy et al. 2018
	ANNOVAR	NA	http://annovar.openbioinformatics.org/	Wang et al. 2010
	CADD	NA	https://cadd.gs.washington.edu/	Kircher et al. 2014
	CADD-SV	NA	https://github.com/kircherlab/CADD-SV	Kleinert and Kircher 2022
	DANN	NA	https://cbcl.ics.uci.edu/public_data/DANN/	Quang et al. 2015
	PolyPhen-2	NA	http://genetics.bwh.harvard.edu/pph2/	Adzhubei et al. 2013
	SIFT	NA	https://sift.bii.a-star.edu.sg/	Sim et al. 2012
	SnpEff	NA	https://pcingola.github.io/SnpEff/	Cingolani et al. 2012
	SvAnna	ONT/PacBio	https://github.com/monarch-initiative/SvAnna	Danis et al. 2022
	SVAFotate	NA	https://github.com/fakedrtom/SVAFotate	Nicholas et al. 2022
	VEP	NA	https://github.com/Ensembl/ensembl-vep	Hunt et al. 2022
Downstream analysis—methylation calling (L)	Dorado	ONT	https://github.com/nanoporetech/dorado	
	fibertools-rs/ft	PacBio	https://github.com/fiberseq/fibertools-rs	Jha et al. 2024
	Jasmine	PacBio	https://github.com/PacificBiosciences/jasmine	
	Modkit	ONT	https://github.com/nanoporetech/modkit	
Assembly (M)	Canu	ONT/PacBio	https://github.com/marbl/canu	Koren et al. 2017
	Flye	ONT	https://github.com/mikolmogorov/Flye	Konings et al. 2024
	hifiasm	ONT/PacBio	https://github.com/chhylp123/hifiasm	Cheng et al. 2021
	Shasta	ONT	https://github.com/chanzuckerberg/shasta	Shafin et al. 2020
	Verkko	ONT + PacBio + Illumina	https://github.com/marbl/verkko	Rautiainen et al. 2023
Assembly QC (N)	AssemblyQC	NA	https://github.com/Plant-Food-Research-Open/assemblyqc	Rashid et al. 2024
	BUSCO	NA	https://gitlab.com/ezlab/busco	Manni et al. 2021
	Flagger	ONT/PacBio	https://github.com/mobinasri/flagger	Liao et al. 2023
	Merqury	Illumina	https://github.com/marbl/merqury	Rhie et al. 2020
	NucFreq	ONT/PacBio	https://github.com/mrvollger/NucFreq	Vollger et al. 2019
	QUAST-LG	ONT/PacBio	https://quast.sourceforge.net/quast-lg.html	Mikheenko et al. 2018
Variant calling from assembly (O)	Dipcall	NA	https://github.com/lh3/dipcall	Li et al. 2018
	PAV	NA	https://github.com/EichlerLab/pav	Ebert et al. 2021
Graph genome tools (P)	Giraffe	NA	https://github.com/vgteam/vg	Sirén et al. 2021
	Graph Genome pipeline	NA	https://github.com/graph-genome/pipeline	Guarracino et al. 2022
	Minigraph-Cactus	NA	https://github.com/ComparativeGenomicsToolkit/cactus	Hickey et al. 2024
	PanGenie	NA	https://github.com/eblerjana/pangenie	Ebler et al. 2022
	PGGB	NA	https://github.com/pangenome/pggb	Garrison et al. 2024
	PGR-TK	NA	https://github.com/GeneDx/pgr-tk	Chin et al. 2023
	VG	NA	https://github.com/vgteam/vg	Garrison et al. 2018

Letters in parentheses correspond to Figure 2 workflow steps.

as it might impact downstream analyses (Majidian et al. 2023). For example, while GRCh37 and GRCh38 provide more annotation, T2T-CHM13 (newest release v2.0) has been shown to reduce artifacts in the analysis (Aganezov et al. 2022), and a fixed GRCh38 reduces false positive calls in collapsed or falsely duplicated regions of the genome (Behera et al. 2023; Mahmoud et al. 2024a). Moreover, including alternative contigs within reference genomes has been shown to impact alignment precision negatively, necessitating careful consideration during the reference selection process (https://lh3.github.io/2017/11/13/which-human-reference-genome-to-use).

Interestingly, the increased read length and different error characteristics of long-read sequencing pose significant challenges for accurate alignment (Sedlazeck et al. 2018b). Novel computational approaches are continuously developed to improve the efficiency and accuracy of mapping long reads to reference genomes. Furthermore, mapping is an essential step since it is the basis of all subsequent analyses, and hence, an error in this stage will impact

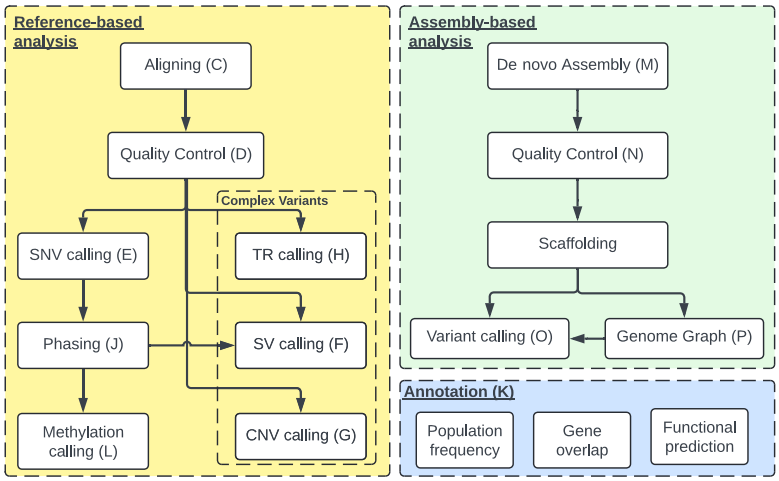


Figure 2. Schematic workflow for long-read-based genomic analysis. The workflow outlines the two approaches to analyzing long-read sequencing data. It details the different routes a researcher can take, either a reference-based approach or an assembly-based approach, and concludes with the types of annotations that can be generated. The letters in parentheses next to each step correspond to their detailed list of tools provided in Table 1.

the overall results substantially. In the past, multiple aligners have been developed for long reads, such as BLASR (Chaisson and Tesler 2012) and NGMLR (Sedlazeck et al. 2018b), which have been so far overtaken by minimap2 (Table 1; Li 2018). Although minimap2 has become the most widely adopted aligner for long reads, it still struggles to map reads accurately within repetitive genomic regions (5%–10% of the human genome) and regions affected by rearrangements (e.g., inversions) (Ding et al. 2024). Newer aligners, including Winnowmap2 (Jain et al. 2022) and VACmap (Ding et al. 2024), demonstrated improved accuracy in challenging genomic regions compared to minimap2 (Table 1). It has been recently shown that the combination of aligners can achieve improvements over single aligners in speed and accuracy. Vulcan was developed by aligning reads quickly with minimap2 and then reprocessing the ones with abnormally high edit distances using different aligners (Fu et al. 2021).

Several proposed aligners utilize the raw base signal instead of relying on the base caller. For example, cwDTW (Han et al. 2018) employs a dynamic time-warping (DTW) algorithm to measure DNA similarity but is computationally intensive and sensitive to noise (Shih et al. 2022). While tools like Sigmap (Zhang et al. 2021) aim to address these challenges, their performance can suffer when aligning large genomes (Firtina et al. 2024), and the difficulty of processing large FAST5/POD5 files limits their broader adoption. The alignment process typically produces a tab-separated output file in SAM or compressed BAM formats, which includes detailed information (e.g., mapping location and alignment differences) on a per-alignment fragment basis (Li et al. 2009). It is important to note that a read can be aligned in one or multiple fragments (i.e., split reads), while each part of the read is typically aligned only once. Additional information, such as edit distance or methylation tags, is stored with the alignments.

After alignment, rigorous QC is essential to evaluate its performance. Metrics such as percentage of aligned reads, alignment identity, and average base quality are commonly investigated. These assessments are facilitated by tools like SAMtools (Danecek et al. 2021) and NanoPlot (De Coster and Rademakers 2023). In this context, SAMtools can be used to calculate alignment metrics such as the percentage of mapped reads, the presence of split-read alignments, or high numbers of soft-clipped bases. For a typical human genome realignment, one might expect >90% of reads to align successfully, with only a small proportion (e.g., ~10%) showing split-read alignments. These metrics help identify potential issues, such as low-quality data or misalignments, ensuring reliable downstream analyses.

Identification of genetic and epigenetic alterations from long-read sequencing data

After an alignment has passed the quality assessment, variant calling is the next step in the analysis. Detecting genomic alterations fosters our understanding of genomic differences between individuals and thus may also give insights into diseases or other important phenotypes. In general, genomic alterations can be classified into four different groups based on their type and size:

i) SNVs and insertions/deletions (indels) are often defined as smaller than 50 bp. SNVs and indels are the most studied variant type in genetics studies due to their abundance in coding regions and clear relationship to protein changes, with ~4–5 million variants expected per human genome (The 1000 Genomes Project Consortium 2015). Tools that detect SNVs and indels are broadly categorized into two major approaches: traditional statistical methods and machine learning-based techniques. An example of a statistical method is Longshot, which utilizes the Pair-Hidden Markov Model (Edge and Bansal 2019). While statistics-based tools offer faster runtimes, machine learning methods like DeepVariant (Poplin et al. 2018) and Clair (Zheng et al. 2022) are now preferred for both short- and long-read variant calling due to their higher accuracy. However, these machine learning models require careful selection as they are technology-specific, optimized for particular flow cell generations, and primarily trained on human genome data.

ii) Tandem repeats (TRs) consist of consecutive copies of DNA motifs, ranging from 1 bp or larger, with variable copy numbers. TRs span around 8% of the human genomes, range from homopolymers to large segmental duplications, and are highly variable between individuals. So far, over 60 diseases have been linked to TRs, with the majority of them being neurodegenerative (English et al. 2024b). The repetitiveness of the TR sequence often requires specialized methods to overcome alignment artifacts. To account for these, multiple methods have been introduced, such as the Gaussian mixed model (Ummat and Bashir 2014; Liu et al. 2017; Dolzhenko et al. 2024), deep learning (De Roeck et al. 2019; Giesselmann et al. 2019; Fang et al. 2022), and a network-based approach (Guo et al. 2018). In addition, specific methods such as StrSpy leverage TRs' high variability between individuals for forensic applications by targeting only a small set of TRs (Hall et al. 2022). A significant challenge in TR detection lies in the inconsistent definition of repeat units across different tools. While some methods focus on shorter repeats, others target longer, more

complex structures, hindering direct comparisons and meta-analyses. Recent tools from PacBio (TRGT [Dolzhenko et al. 2024]) and ONT (Medaka and pathSTR [De Coster et al. 2024]) have been improved and have shown great performance across the entire genome (English et al. 2024b).

iii) Structural variants (SVs) are genome alterations longer than 50 bp, typically numbering between 23,000 and 27,000 variants per healthy human genome (Mahmoud et al. 2019). SV encompasses deletions, duplications, insertions, inversions, and translocations, often occurring in repetitive regions. SVs and TRs are inherently interconnected, as many SVs involve alterations in TR regions, such as expansions or contractions of repeat units. Recognizing this overlap is crucial for a comprehensive understanding of genomic variation. By analyzing SVs and TRs together, we can capture both large-scale structural changes and finer-scale repeat dynamics, offering complementary insights into the genome's complexity (Jensen et al. 2024).

Various methods have been developed to call SVs from mapped reads, including Sniffles2 (Smolka et al. 2024), cuteSV (Jiang et al. 2020), and PBSV (see Table 1). These methods operate on the shared principle of detecting discordant mappings and inferring the SV type by consensus or localized assembly. Additionally, AI-based methods like SVision (Lin et al. 2022b), BreakNet (deletions only) (Luo et al. 2021), and MAMnet (Ding and Luo 2022) have been developed. Most SV callers perform similarly, so differences are often based on runtime or additional features. For example, SVision Pro can detect more complex SVs, while Sniffles2 allows rapid comparison of multiple samples. Of note, there are also cancer-specific variant callers that streamline the comparison of tumor and normal samples, such as NanomonSV (Shiraishi et al. 2023), SAVANA (Elrick et al. 2024), and Severus (Keskus et al. 2024). Additionally, using long reads, cohort analysis tools such as SVJedi (Romain and Lemaitre 2023) and Kanpig (Table 1; English et al. 2024a) were developed to genotype SVs.

iv) Copy number variants (CNVs) are typically large alterations that span multiple megabases. While smaller CNVs (50 bp–1 Mbp) are often also detected by SV callers, larger CNVs or even chromosome-size alterations are frequently missed. Due to the undefined boundaries between CNVs and SVs and the predominant focus on SVs within the long-read sequencing community, only a few CNV callers for long reads have been developed. The most prominent CNV callers are HiFiCNV (https://github.com/PacificBiosciences/HiFiCNV) for HiFi reads and Spectre (https://github.com/fritzsedlazeck/Spectre), which works on both ONT and HiFi to identify large CNVs (>100 kb). It is essential to highlight that neither tool reports precise breakpoints, making comparison to other variants typically challenging.

v) Epigenetic alterations can be detected directly from long-read sequencing data without prior biochemical alterations of the DNA (such as bisulfite sequencing). The presence of 5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC, ONT only) at CpG sites is recorded during the basecalling process. Other modifications, such as N^6-methyladenine (6mA), can also be detected using specific contexts or methods (Kong et al. 2023; Agustinho et al. 2024). Methylation has been linked to the regulation of promoters, playing a crucial role in tissue-specific gene expression and the regulation of oncogenes in cancer (Wang et al. 2022; Bhootra et al. 2023). Postprocessing of methylation based on raw reads is typically carried out using tools like Modkit and Jasmine for PacBio or the basecaller Dorado for ONT (see Table 1). Fiber-seq (Stergachis et al. 2020), which requires prior handling of methyltransferases to alter the sample, allows the detection of open chromatin to provide additional insights.

Long-read sequencing has been instrumental in recent breakthroughs in understanding complex genomic regions and epigenetic regulation. The Telomere-to-Telomere (T2T) consortium (Nurk et al. 2022) demonstrated the power of these technologies by resolving previously intractable regions of the human genome. In particular, long reads enabled the first complete characterization of human centromeres, revealing their complex satellite DNA organization and epigenetic states, and provided unprecedented resolution of segmental duplications, which had been resistant to accurate assembly using short reads due to their highly repetitive nature (Nurk et al. 2022). These advances have transformed our understanding of genome architecture, showing how segmental duplications contribute to evolutionary innovation and genomic diversity, while also illuminating the structural organization of centromeres and their role in chromosome segregation (Altemose et al. 2022a). This milestone was accompanied by innovative methodologies leveraging long-read capabilities for epigenetic profiling. Novel approaches such as NanoNOMe (Lee et al. 2020), DiMeLo-seq (Altemose et al. 2022b), and Fiber-seq (Stergachis et al. 2020) have further expanded our ability to profile DNA modifications and chromatin.

Another feature of long-read sequencing is its ability to enable the phasing of variants. Phasing refers to determining whether or not two or more variants co-occur on the same DNA molecule (i.e., haplotype). Long-read-based phasing can detect rare or de novo alleles that population-based phasing methods often miss. Additionally, phasing can reveal inheritance patterns and identify carriers of potentially disease-causing mutations. Phasing is typically conducted by analyzing if variants co-occur on a single read, which is then extended by overlapping reads and statistical clustering of variants, often assuming a diploid model and focusing on heterozygous variants. The primary focus is on SNVs since their frequency across the genome allows them to be phased within a read length. Phased SNVs are then reported in phase blocks (i.e., regions where phasing is consistent) and haplotypes (e.g., HP1, HP2).

There are two main methods for phasing: WhatsHap (Martin et al. 2023) and HapCUT2 (Edge et al. 2017). Other SNV-based phasing methods include LongPhase (Lin et al. 2022a) and HiPhase (Holt et al. 2024), which can incorporate SVs. A novel method, MethPhaser, extends the SNV-based phasing concept by leveraging haplotype-specific methylation signals, which span regions of homozygosity (Fu et al. 2024). Notably, many variant calling methods can leverage phasing information across different variant types (e.g., SNV and SV by Sniffles2). Additionally, when parental data is available, it can serve as a powerful validation tool to assess phasing accuracy by comparing the inferred haplotypes with the expected inheritance patterns. To enable SV phasing, we first need to phase SNVs and then tag the BAM file (e.g., using WhatsHap) before proceeding with additional SV calling.

Another common task after variant identification and phasing (optional) is the comparison of variants across samples. For SNVs, the best practice is to use a genomic variant call format file (gVCF), especially if more than one sample is studied. A gVCF file allows the merging of two or more samples and their variants into a fully genotyped VCF file. This avoids erroneous interpretations of variants whose genotypes are unknown per sample (i.e., ./.). gVCF is a particular type of VCF that contains records for every position or interval in the genome (e.g., read depth),

regardless of whether it contains variants. GLnexus (Lin et al. 2018) or BCFtools (Danecek et al. 2021) merge gVCF files across samples. Tools like Sniffles2, Jasmine (Kirsche et al. 2023), and Truvari (English et al. 2024b) are commonly used for multiple sample SV comparisons. Sniffles2 relies on the binary file it creates for each sample to merge SVs. Jasmine uses a minimum spanning forest algorithm to merge SVs but does not call SVs. Truvari can merge and benchmark SVs, including a module named "phab," which is specific for TR calling (English et al. 2024b).

Variant identification using long-read sequencing has significantly improved over the years and continues to evolve rapidly. This advancement often results in a more comprehensive understanding of the genome. However, annotations are typically required to infer their functional impact on various phenotypes, including diseases, to utilize these variants effectively.

Annotation of variants from long-read sequencing

The next crucial step is variant annotation, which has two primary goals: estimating the potential functional impact of variants and population frequency annotation. Variant annotation does not distinguish between variants identified from long or short reads. However, certain aspects need to be considered. Researchers utilize biological databases and annotation files to predict functional annotation of the identified variant calls. These resources provide crucial information about the variants' locations within genes, their predicted effects on protein sequence and function, and any known disease associations. Tools such as ANNOVAR (Wang et al. 2010), SnpEff (Cingolani et al. 2012), and Ensembl Variant Effect Predictor (VEP; Hunt et al. 2022) are commonly used, enabling researchers to systematically annotate variants with known functional effects or potential impacts on genes and regulatory elements. ANNOVAR, in particular, queries many databases, including ClinVar (Landrum et al. 2014) for disease associations, dbSNP (Sherry et al. 1999) for known variants, and OMIM (Hamosh et al. 2005) for disease genes. Beyond empirical annotation, VEP also facilitates access to other in silico pathogenicity prediction tools such as PolyPhen-2 (Adzhubei et al. 2013), SIFT (Sim et al. 2012), and CADD (Kircher et al. 2014). These tools generate scores indicative of a variant's potential deleteriousness, ranging from benign to probably damaging. However, their predictions require cautious interpretation as protein context and individual genetics can influence the actual effects.

While functional annotation is crucial, it cannot be interpreted without considering the allele frequency (AF) of the variant in the population. Population AF is a key factor in ranking variants for their likelihood of being pathogenic (Kobayashi et al. 2017; Gudmundsson et al. 2022). Generally, the likelihood of pathogenicity is positively correlated with the rarity of variants in the population, with few exceptions (Kobayashi et al. 2017). Databases such as gnomAD (Chen et al. 2024) contain annotations about population AF. SVAFotate (Nicholas et al. 2022), AnnotSV (Geoffroy et al. 2018), CADD-SV (Kleinert and Kircher 2022), and gnomAD also enable the annotation of SV and prediction of their deleteriousness. However, these databases are based on short-read sequencing data and often lack comprehensive genome-wide SV annotations (Mahmoud et al. 2024b). A recent study reported that only ~35% of SV from the HG002 GIAB benchmark could be annotated using gnomAD, while another long read-based annotation achieves 95% (Zheng et al. 2024). Initiatives like the HGSVC, HPRC, and the All of Us Research Program projects are currently developing SV catalogs based on long-read sequencing data of large population groups for AF annotation to address this gap (Gustafson et al. 2024; Mahmoud et al. 2024b).

In conclusion, variant annotation in long-read sequencing workflows is essential for understanding the functional significance of genetic variations, contributing to a more comprehensive understanding of their roles in health and disease.

De novo assembly of long-read data

De novo assembly, the process of reconstructing complete genome sequences from raw sequencing reads without a reference genome, has historically been a significant challenge for short-read sequencing technologies. Due to their inherent length limitations, short reads struggle to span repetitive regions and complex rearrangements in many genomes (Logsdon et al. 2020). These limitations often result in fragmented assemblies riddled with gaps and misassemblies. Long-read sequencing technologies have revolutionized de novo assembly, offering significant advantages over traditional short-read approaches (Espinosa et al. 2024). Several long-read assembly tools have emerged, such as Canu (Koren et al. 2017), Flye (Kolmogorov et al. 2019), Hifiasm (Cheng et al. 2021), and Shasta (Shafin et al. 2020). Hifiasm (Cheng et al. 2021) is currently the most widely used method working on PacBio and ONT, and it provides phased assemblies. Verkko (Rautiainen et al. 2023) is a hybrid assembly pipeline that uses PacBio and/or ONT data with a graph-based approach to producing highly accurate assemblies. Additionally, Verkko could utilize parental Illumina short-read data for phasing. For more details on near-complete genome assembly and assembly algorithms, we encourage the reader to follow the review of Li and Durbin (2024).

In general, genome assemblers first produce continuous sequences built upon overlapping individual reads, effectively tiling them together to create longer contiguous sequences (i.e., contigs) (Sedlazeck et al. 2018a). With modern long-read technologies, particularly PacBio HiFi reads, many assemblers can now produce highly contiguous assemblies directly from the assembly graph, often achieving chromosome-scale contigs without additional scaffolding steps, especially for human genomes (Cheng et al. 2021; Nurk et al. 2022). When needed, scaffolding methods utilize additional information to combine contigs, sometimes with an unresolved sequence in between (i.e., gaps represented with NN). To help create scaffolds, Hi-C, Bionano, or ONT ultra-long reads can provide long-range information extending beyond large repeats (e.g., segmental duplications) and thus joining different contigs into scaffolds. However, scaffolding outside the assembly graph may introduce errors (Nurk et al. 2022) and should be considered carefully based on specific project needs. Given the linking of two or more contigs (i.e., scaffolding), the sequence spanning is often undefined as Hi-C or Bionano does not provide sequence context. Thus, gap-filling strategies can replace these undefined sequences by utilizing unmapped parts of reads originating from the region (Xu et al. 2020; Schmeing and Robinson 2023).

Phasing is another important aspect of genome assembly. Multiple assemblers provide phased assemblies directly from long-read data. These can be extended by Hi-C or Pore-C (Deshpande et al. 2022). Their integration typically leads to larger phase blocks (Garg et al. 2021; Li and Durbin 2024). Alternatively, parental data alongside the proband is another effective approach for whole-genome assembly phasing (Koren et al. 2018).

After generating a preliminary assembly, multiple QC metrics are needed to assess the assembly's completeness and accuracy. Metrics like N50 (i.e., defined as the length of the shortest contig such that the sum of this contig and all larger contigs covers at least 50% of the total assembly length) and the total number and size distribution of contigs provide insights into the assembly continuity. However, these metrics can sometimes be misleading, thus, auN (also known as E-Size) was introduced as a more comprehensive measure (https://lh3.github.io/2020/04/08/a-new-metric-on-assembly-contiguity), calculated as the weighted average of contig lengths across all cumulative coverage thresholds (e.g., N10, N20, N30, …, N100) (Salzberg et al. 2012).

That said, these metrics alone fail to provide deeper insights into the accuracy or completeness of the assembly itself. To address that, benchmarking universal single-copy orthologs (BUSCO) (Manni et al. 2021) analysis is a commonly used approach to evaluate the assembly's content by searching for a set of highly conserved, single-copy orthologous genes. BUSCO reports results in three categories: "Complete" (single-copy or duplicated), "Fragmented," and "Missing," providing a quantitative measure of genome assembly quality. In addition to BUSCO, other methods are available for assessing assembly quality. For example, HMM-Flagger (Liao et al. 2023) uses a hidden Markov model to detect misassemblies by analyzing patterns in read mapping, coverage depth, and other alignment signals. Similarly, *k*-mer-based tools such as Merqury (Rhie et al. 2020) evaluate the assembly by comparing the *k*-mers from the assembled genome against those derived from unassembled high-accuracy raw sequencing reads, providing insights into the completeness and accuracy of the assembly.

Once the quality of an assembly is established, variant calling can retrieve information concerning sequence differences between the new assembly and another assembly or reference genome. The most commonly used methods for variant calling in assemblies currently include Dipcall (Li et al. 2018), SVIM-asm (Heller and Vingron 2021), and phased assembly variant (PAV) (Ebert et al. 2021). Dipcall can infer SNVs, insertions and deletions (indels), and SVs, though it is limited to detecting only insertions and deletions. In contrast, PAV extends its capabilities by also identifying inversions. SVIM-asm, on the other hand, specializes in detecting a broader range of SVs, including deletions, insertions, tandem and interspersed duplications, and inversions. Variant calling using assembled genomes has advantages over reference-based methods. Assembly-based approaches avoid biases introduced by incomplete or biased reference genomes (Behera et al. 2023). Additionally, assembly-based methods enable the detection of larger insertions, which can still be challenging using long-read mapping-based methods.

Finally, while variant calling can utilize existing reference annotations (e.g., from T2T-CHM13v2 or GRCh), direct functional annotation of a genome assembly can provide additional insights about genes or other functional elements. This is particularly valuable when studying novel sequences or SVs that might affect gene structure or regulation. This process is still highly complicated (Salzberg 2019), often including either a de novo approach over RNA-seq or a liftover approach of a close relative available genome. Multiple methods have been suggested (e.g., liftoff [Shumate and Salzberg 2021]), but most of them require manual curation on top of automated pipelines such as Apollo (Dunn et al. 2019).

In conclusion, long-read sequencing technologies have significantly improved the continuity and accuracy of de novo assembly, paving the way for more accurate and complete genome reconstructions.

Graph genomes for long-read analysis

Reference-based methods rely on mapping reads to a reference genome, assuming that it accurately represents the sample's genetic makeup. However, this single-reference approach is limited by the reference's completeness and accuracy. It often overlooks structural variations and polymorphisms in complex regions such as the human leukocyte antigen (Lai et al. 2024; Zhou et al. 2024), *LPA* (Behera et al. 2024a), and major histocompatibility complex (Liao et al. 2023) in the human genome. Moreover, a single-reference genome can introduce bias, particularly in nonmodel organisms or genetically diverse populations (Gong et al. 2023; Secomandi et al. 2023; Sun et al. 2025). Graph genome (GG) approaches, or pangenome graphs, address these limitations by representing multiple genomes as graphs (Garrison et al. 2018; Miga and Wang 2021). These graphs provide a more comprehensive representation of genetic diversity and capture variations like insertions, deletions, and SVs as nodes and branches. GGs have proven valuable for studying complex genomes (Paten et al. 2017) and understanding how mobile element insertions impact the epigenome (Groza et al. 2023). Hence, these data structures hold substantial promise for multiple applications, including cancer research and population analysis (Sherman and Salzberg 2020). However, this comes at the cost of increased computational complexity, and variant calls still need to be projected onto a linear reference within the graph for downstream analysis.

GG workflows construct a graph by integrating existing reference genomes with known variants (Garrison et al. 2018). Several methods capitalize on the availability of highly accurate and comprehensive de novo genome assemblies. These approaches build graphs that capture complete genomic variation, enabling accurate population-level analyses (Eizenga et al. 2020; Nurk et al. 2022). For instance, Minigraph (Li et al. 2020) and Minigraph-Cactus (Hickey et al. 2024) leverage de novo assemblies to construct graph-based genome representations and align reads to existing graphs. Other tools like vg, Giraffe (Sirén et al. 2021), HISAT2 (Kim et al. 2019), or DRAGEN (Behera et al. 2024b) align DNA and RNA short reads to GG that are constructed either based on assemblies or previously identified variants. Alternatively, custom graphs with specific alleles enable the visualization of more complex regions across a population set (Chin et al. 2022, 2023). Furthermore, some methods in SV genotyping are already utilizing regional graphs to identify if variants are present in a BAM file (Chen et al. 2019; Ebler et al. 2022; English et al. 2024a)

Finally, while the graph approach holds great promise in genome analysis, further research is essential to fully evaluate its scalability and effectiveness in identifying variants across both genic and intergenic regions, ensuring their utility in diverse genomic analyses; different papers have discussed this in more depth (Sherman and Salzberg 2020; Abondio et al. 2023, 2024; Liao et al. 2023; Rocha et al. 2024).

Discussion

Long-read sequencing is advancing rapidly, with continuous improvements in read length and accuracy, revolutionizing genomic research (Fig. 1A). Advances in basecalling algorithms and sequencing chemistry have significantly enhanced accuracy, making long-read data more reliable and precise, as evidenced in Figure 1B. These improvements have enabled more accurate detection of SVs, including large insertions, deletions, duplications, and complex rearrangements, which are challenging or impossible to

identify with short-read sequencing alone. Additionally, long-read sequencing has dramatically improved de novo assembly capabilities, particularly in regions with complex repetitive elements, where short-read technologies have fallen short. Beyond genome assembly, long-read technologies have expanded their applications to transcriptomics, epigenomics, and metagenomics, providing deeper insights into gene regulation and the genetic diversity within populations. Both PacBio and ONT sequencing technologies support the detection of DNA modifications, such as methylation, while ONT uniquely enables the detection of RNA modifications. This capability provides valuable insights for understanding epigenetic state, chromatin structure, and RNA modifications. Furthermore, developing portable, real-time sequencing devices has opened new possibilities for immediate data analysis, with potential applications in clinical and field settings (Jain et al. 2016; King et al. 2020; Wasswa et al. 2022). These advancements collectively contribute to a more comprehensive understanding of genomic structure and function, with broad implications for personalized medicine (Wojcik et al. 2023) and evolutionary biology (Stergachis et al. 2020).

Despite its revolutionary potential, long-read sequencing still faces several challenges. Compared to short-read technologies, it requires larger DNA input quantities and involves higher per-base sequencing costs. The complexity of data analysis presents ongoing challenges, particularly in accurate alignment, GG creation, and variant calling within complex or duplicated genomic regions. While bioinformatic innovations are helping manage the large data sets generated by this technology, the generation of high-quality libraries and the operation of complex higher-throughput instruments still require specialized expertise and infrastructure. These current limitations notwithstanding, long-read sequencing continues to expand its role in genomics research and clinical applications, particularly in areas where comprehensive structural variation detection and complete genome assembly are crucial.

Nevertheless, efforts are progressing to simplify these processes and reduce DNA input requirements (Heavens et al. 2021). Finally, there is an annotation gap, as current databases predominantly rely on short-read data, which misses around 50% of SVs (Ebert et al. 2021). Integrating long-read data into existing databases or developing new ones tailored for long-read-derived variants, particularly SVs, is essential to fully harness the power of long-read sequencing. This integration will ultimately enhance our understanding of genetic variation and provide a more comprehensive view of genomes. Despite these challenges, the continuous improvements in long-read sequencing technologies promise to unlock new possibilities in genomics research. The future of long-read sequencing is poised for significant advancements to broaden its impact across various research fields (Conesa et al. 2024).

Integrating long-read sequencing with other "omics" technologies, such as proteomics and metabolomics, promises a more comprehensive understanding of biological systems. Additionally, novel single-cell and spatial genomics applications are emerging alongside real-time and in-field sequencing capabilities with platforms like ONT (Izydorczyk et al. 2024), which will enhance field-based genomics, clinical diagnostics, and environmental monitoring. Furthermore, the development of specialized analysis tools, particularly those leveraging machine learning approaches, promises to make long-read data interpretation more efficient and accessible to a broader scientific community (Poplin et al. 2018; Mastoras et al. 2024). Graph genomes will also play a crucial role in addressing reference bias and enabling the exploration of the diversity from complex genomic regions (Miga and Wang 2021). This is particularly valuable for detecting links between genetic markers and diseases, facilitating the genetic study of more prevalent pathologies in different populations. Combining long-read sequencing with CRISPR for targeted sequencing and expanding direct RNA-seq technologies will offer deeper insights into genomic and transcriptomic complexities. The integration of long-read sequencing into clinical diagnostics represents a crucial frontier, with several sequencing centers beginning to validate these platforms for clinical use. Although the rapid evolution of long-read technologies poses challenges for clinical validation, successful diagnostic cases in research settings have demonstrated their potential to resolve previously unsolved cases, particularly those involving complex SVs or repeat expansions (Goenka et al. 2022; Gorzynski et al. 2022).

Long-read sequencing has emerged as a transformative technology in genomics, fundamentally changing our ability to explore complex genomic landscapes. By overcoming key limitations of short-read sequencing, it provides unprecedented insights into genome structure and variation, particularly in challenging regions that have long remained inaccessible. As technical challenges continue to be addressed and new applications emerge, long-read sequencing is poised to revolutionize both basic research and clinical diagnostics, especially in cases where traditional approaches have proven insufficient. The continuous evolution of this technology, coupled with advances in bioinformatics and clinical validation, promises to deepen our understanding of human genetics and accelerate the path toward more precise and personalized medicine.

Competing interest statement

F.J.S. receives research support from Illumina, PacBio, and Oxford Nanopore. The other authors declare no competing interests.

Acknowledgments

The authors would like to thank Androo Markham for constructive discussions. This research was supported in part by the National Institutes of Health (NIH) grants (1U01HG011758-01 and 1UG3NS132105-01), the National Institute of Child Health and Human Development (NICHD) (R01HD106056), and the National Institute of Allergy and Infectious Diseases (1U19AI144297).

Author contributions: All authors contributed to the research and writing of the manuscript.

References

The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature* **526:** 68–74. doi:10.1038/nature15393

Abondio P, Cilli E, Luiselli D. 2023. Human pangenomics: promises and challenges of a distributed genomic reference. *Life* **13:** 1360. doi:10.3390/life13061360

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

Adzhubei I, Jordan DM, Sunyaev SR. 2013. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* **Chapter 7:** Unit7.20. doi:10.1002/0471142905.hg0720s76

Aganezov S, Yan SM, Soto DC, Kirsche M, Zarate S, Avdeyev P, Taylor DJ, Shafin K, Shumate A, Xiao C, et al. 2022. A complete reference genome improves analysis of human genetic variation. *Science* **376:** eabl3533. doi:10.1126/science.abl3533

Agustinho DP, Fu Y, Menon VK, Metcalf GA, Treangen TJ, Sedlazeck FJ. 2024. Unveiling microbial diversity: harnessing long-read

sequencing technology. *Nat Methods* **21:** 954–966. doi:10.1038/s41592-024-02262-1

Akagi K, Symer DE, Mahmoud M, Jiang B, Goodwin S, Wangsa D, Li Z, Xiao W, Dunn JD, Ried T, et al. 2023. Intratumoral heterogeneity and clonal evolution induced by HPV integration. *Cancer Discov* **13:** 910–927. doi:10.1158/2159-8290.CD-22-0900

Altemose N, Logsdon GA, Bzikadze AV, Sidhwani P, Langley SA, Caldas GV, Hoyt SJ, Uralsky L, Ryabov FD, Shew CJ, et al. 2022a. Complete genomic and epigenetic maps of human centromeres. *Science* **376:** eabl4178. doi:10.1126/science.abl4178

Altemose N, Maslan A, Smith OK, Sundararajan K, Brown RR, Mishra R, Detweiler AM, Neff N, Miga KH, Straight AF, et al. 2022b. DiMeLo-seq: a long-read, single-molecule method for mapping protein–DNA interactions genome wide. *Nat Methods* **19:** 711–723. doi:10.1038/s41592-022-01475-6

Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. 2020. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol* **21:** 30. doi:10.1186/s13059-020-1935-5

Ashton PM, Nair S, Dallman T, Rubino S, Rabsch W, Mwaigwisya S, Wain J, O'Grady J. 2015. MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nat Biotechnol* **33:** 296–300. doi:10.1038/nbt.3103

Baid G, Cook DE, Shafin K, Yun T, Llinares-López F, Berthet Q, Belyaeva A, Töpfer A, Wenger AM, Rowell WJ, et al. 2023. DeepConsensus improves the accuracy of sequences with a gap-aware sequence transformer. *Nat Biotechnol* **41:** 232–238. doi:10.1038/s41587-022-01435-7

Behera S, LeFaive J, Orchard P, Mahmoud M, Paulin LF, Farek J, Soto DC, Parker SCJ, Smith AV, Dennis MY, et al. 2023. FixItFelix improving genomic analysis by fixing reference errors. *Genome Biol* **24:** 31. doi:10.1186/s13059-023-02863-7

Behera S, Belyeu JR, Chen X, Paulin LF, Nguyen NQH, Newman E, Mahmoud M, Menon VK, Qi Q, Joshi P, et al. 2024a. Identification of allele-specific KIV-2 repeats and impact on Lp(a) measurements for cardiovascular disease risk. *BMC Med Genomics* **17:** 255. doi:10.1186/s12920-024-02024-0

Behera S, Catreux S, Rossi M, Truong S, Huang Z, Ruehle M, Visvanath A, Parnaby G, Roddey C, Onuchic V, et al. 2024b. Comprehensive genome analysis and variant detection at scale using DRAGEN. *Nat Biotechnol* doi:10.1038/s41587-024-02382-1

Bhootra S, Jill N, Shanmugam G, Rakshit S, Sarkar K. 2023. DNA methylation and cancer: transcriptional regulation, prognostic, and therapeutic perspective. *Med Oncol* **40:** 71. doi:10.1007/s12032-022-01943-1

Calvo-Roitberg E, Daniels RF, Pai AA. 2024. Challenges in identifying mRNA transcript starts and ends from long-read sequencing data. *Genome Res* **34:** 1719–1734. doi:10.1101/gr.279559.124

Cameron DL, Di Stefano L, Papenfuss AT. 2019. Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. *Nat Commun* **10:** 3240. doi:10.1038/s41467-019-11146-4

Chaisson MJ, Tesler G. 2012. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* **13:** 238. doi:10.1186/1471-2105-13-238

Chaisson MJP, Sulovari A, Valdmanis PN, Miller DE, Eichler EE. 2023. Advances in the discovery and analyses of human tandem repeats. *Emerg Top Life Sci* **7:** 361–381. doi:10.1042/ETLS20230074

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

Chen S, Francioli LC, Goodrich JK, Collins RL, Kanai M, Wang Q, Alföldi J, Watts NA, Vittal C, Gauthier LD, et al. 2024. A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* **625:** 92–100. doi:10.1038/s41586-023-06045-0

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

Chin C-S, Behera S, Metcalf GA, Gibbs RA, Boerwinkle E, Sedlazeck FJ. 2022. A pan-genome approach to decipher variants in the highly complex tandem repeat of *LPA*. bioRxiv doi:10.1101/2022.06.08.495395

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

Chiu R, Rajan-Babu I-S, Friedman JM, Birol I. 2021. Straglr: discovering and genotyping tandem repeat expansions using whole genome long-read sequences. *Genome Biol* **22:** 224. doi:10.1186/s13059-021-02447-3

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. 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)* **6:** 80–92. doi:10.4161/fly.19695

Collins RL, Brand H, Karczewski KJ, Zhao X, Alföldi J, Francioli LC, Khera AV, Lowther C, Gauthier LD, Wang H, et al. 2020. A structural variation reference for medical and population genetics. *Nature* **581:** 444–451. doi:10.1038/s41586-020-2287-8

Conesa A, Hoischen A, Sedlazeck FJ. 2024. Revolutionizing genomics and medicine—one long molecule at a time. *Genome Res* **34:** ix–xi. doi:10.1101/gr.280179.124

Coster WD, De Coster W, Weissensteiner MH, Sedlazeck FJ. 2021. Towards population-scale long-read sequencing. *Nat Rev Genet* **22:** 572–587. doi:10.1038/s41576-021-00367-3

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. 2021. Twelve years of SAMtools and BCFtools. *Gigascience* **10:** giab008. doi:10.1093/gigascience/giab008

Danis D, Jacobsen JOB, Balachandran P, Zhu Q, Yilmaz F, Reese J, Haimel M, Lyon GJ, Helbig I, Mungall CJ, et al. 2022. SvAnna: efficient and accurate pathogenicity prediction of coding and regulatory structural variants in long-read genome sequencing. *Genome Med* **14:** 44. doi:10.1186/s13073-022-01046-6

De Coster W, Rademakers R. 2023. NanoPack2: population-scale evaluation of long-read sequencing data. *Bioinformatics* **39:** btad311. doi:10.1093/bioinformatics/btad311

De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. 2018. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* **34:** 2666–2669. doi:10.1093/bioinformatics/bty149

De Coster W, Höijer I, Bruggeman I, D'Hert S, Melin M, Ameur A, Rademakers R. 2024. Visualization and analysis of medically relevant tandem repeats in nanopore sequencing of control cohorts with pathSTR. *Genome Res* **34:** 2074–2080. doi:10.1101/gr.279265.124

De Roeck A, De Coster W, Bossaerts L, Cacace R, De Pooter T, Van Dongen J, D'Hert S, De Rijk P, Strazisar M, Van Broeckhoven C, et al. 2019. NanoSatellite: accurate characterization of expanded tandem repeat length and sequence through whole genome long-read sequencing on PromethION. *Genome Biol* **20:** 239. doi:10.1186/s13059-019-1856-3

Deshpande AS, Ulahannan N, Pendleton M, Dai X, Ly L, Behr JM, Schwenk S, Liao W, Augello MA, Tyer C, et al. 2022. Identifying synergistic high-order 3D chromatin conformations from genome-scale nanopore concatemer sequencing. *Nat Biotechnol* **40:** 1488–1499. doi:10.1038/s41587-022-01289-z

Ding H, Luo J. 2022. MAMnet: detecting and genotyping deletions and insertions based on long reads and a deep learning approach. *Brief Bioinform* **23:** bbac195. doi:10.1093/bib/bbac195

Ding H, Sedlazeck FJ, Liao Z, Pu L, Zhu S. 2024. VACmap: an accurate long-read aligner for unraveling complex genomic rearrangements. bioRxiv doi:10.1101/2023.08.03.551566

Dolzhenko E, English A, Dashnow H, De Sena Brandine G, Mokveld T, Rowell WJ, Karniski C, Kronenberg Z, Danzi MC, Cheung WA, et al. 2024. Characterization and visualization of tandem repeats at genome scale. *Nat Biotechnol* **42:** 1606–1614. doi:10.1038/s41587-023-02057-3

Dunn NA, Unni DR, Diesh C, Munoz-Torres M, Harris NL, Yao E, Rasche H, Holmes IH, Elsik CG, Lewis SE. 2019. Apollo: democratizing genome annotation. *PLoS Comput Biol* **15:** e1006790. doi:10.1371/journal.pcbi.1006790

Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, Sulovari A, Ebler J, Zhou W, Serra Mari R, et al. 2021. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372:** eabf7117. doi:10.1126/science.abf7117

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

Edge P, Bansal V. 2019. Longshot enables accurate variant calling in diploid genomes from single-molecule long read sequencing. *Nat Commun* **10:** 4660. doi:10.1038/s41467-019-12493-y

Edge P, Bafna V, Bansal V. 2017. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Res* **27:** 801–812. doi:10.1101/gr.213462.116

Eizenga JM, Novak AM, Sibbesen JA, Heumos S, Ghaffaari A, Hickey G, Chang X, Seaman JD, Rounthwaite R, Ebler J, et al. 2020. Pangenome graphs. *Annu Rev Genomics Hum Genet* **21:** 139–162. doi:10.1146/annurev-genom-120219-080406

Elrick H, Sauer CM, Espejo Valle-Inclan J, Trevers K, Tanguy M, Zumalave S, De Noon S, Muyas F, Cascao R, Afonso A, et al. 2024. SAVANA: reliable analysis of somatic structural variants and copy number aberrations in clinical samples using long-read sequencing. bioRxiv doi:10.1101/2024.07.25.604944

English AC, Cunial F, Metcalf GA, Gibbs RA, Sedlazeck FJ. 2024a. *k*-mer analysis of long-read alignment pileups for structural variant genotyping. bioRxiv doi:10.1101/2024.10.22.619642
English AC, Dolzhenko E, Ziaei Jam H, McKenzie SK, Olson ND, De Coster W, Park J, Gu B, Wagner J, Eberle MA, et al. 2024b. Analysis and benchmarking of small and large genomic variants across tandem repeats. *Nat Biotechnol* doi:10.1038/s41587-024-02225-z
Escaramís G, Docampo E, Rabionet R. 2015. A decade of structural variants: description, history and methods to detect structural variation. *Brief Funct Genomics* **14:** 305–314. doi:10.1093/bfgp/elv014
Espinosa E, Bautista R, Larrosa R, Plata O. 2024. Advancements in long-read genome sequencing technologies and algorithms. *Genomics* **116:** 110842. doi:10.1016/j.ygeno.2024.110842
Fang L, Liu Q, Monteys AM, Gonzalez-Alegre P, Davidson BL, Wang K. 2022. DeepRepeat: direct quantification of short tandem repeats on signal data from nanopore sequencing. *Genome Biol* **23:** 108. doi:10.1186/s13059-022-02670-6
Ferguson S, McLay T, Andrew RL, Bruhl JJ, Schwessinger B, Borevitz J, Jones A. 2022. Species-specific basecallers improve actual accuracy of nanopore sequencing in plants. *Plant Methods* **18:** 137. doi:10.1186/s13007-022-00971-2
Firtina C, Soysal M, Lindegger J, Mutlu O. 2024. RawHash2: mapping raw nanopore signals using hash-based seeding and adaptive quantization. *Bioinformatics* **40:** btae478. doi:10.1093/bioinformatics/btae478
Foox J, Tighe SW, Nicolet CM, Zook JM, Byrska-Bishop M, Clarke WE, Khayat MM, Mahmoud M, Laaguiby PK, Herbert ZT, et al. 2021. Performance assessment of DNA sequencing platforms in the ABRF Next-Generation Sequencing Study. *Nat Biotechnol* **39:** 1129–1140. doi:10.1038/s41587-021-01049-5
Fu Y, Mahmoud M, Muraliraman VV, Sedlazeck FJ, Treangen TJ. 2021. Vulcan: improved long-read mapping and structural variant calling via dual-mode alignment. *Gigascience* **10:** giab063. doi:10.1093/gigascience/giab063
Fu Y, Aganezov S, Mahmoud M, Beaulaurier J, Juul S, Treangen TJ, Sedlazeck FJ. 2024. MethPhaser: methylation-based long-read haplotype phasing of human genomes. *Nat Commun* **15:** 5327. doi:10.1038/s41467-024-49588-0
Fukasawa Y, Ermini L, Wang H, Carty K, Cheung M-S. 2020. LongQC: a quality control tool for third generation sequencing long read data. *G3 (Bethesda)* **10:** 1193–1196. doi:10.1534/g3.119.400864
Garg S, Fungtammasan A, Carroll A, Chou M, Schmitt A, Zhou X, Mac S, Peluso P, Hatas E, Ghurye J, et al. 2021. Chromosome-scale, haplotype-resolved assembly of human genomes. *Nat Biotechnol* **39:** 309–312. doi:10.1038/s41587-020-0711-0
Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Dawson ET, Jones W, Garg S, Markello C, Lin MF, et al. 2018. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat Biotechnol* **36:** 875–879. doi:10.1038/nbt.4227
Garrison E, Guarracino A, Heumos S, Villani F, Bao Z, Tattini L, Hagmann J, Vorbrugg S, Marco-Sola S, Kubica C, et al. 2024. Building pangenome graphs. *Nat Methods* **21:** 2008–2012. doi:10.1038/s41592-024-02430-3
Geoffroy V, Herenger Y, Kress A, Stoetzel C, Piton A, Dollfus H, Muller J. 2018. AnnotSV: an integrated tool for structural variations annotation. *Bioinformatics* **34:** 3572–3574. doi:10.1093/bioinformatics/bty304
Gershman A, Sauria MEG, Guitart X, Vollger MR, Hook PW, Hoyt SJ, Jain M, Shumate A, Razaghi R, Koren S, et al. 2022. Epigenetic patterns in a complete human genome. *Science* **376:** eabj5089. doi:10.1126/science.abj5089
Giesselmann P, Brändl B, Raimondeau E, Bowen R, Rohrandt C, Tandon R, Kretzmer H, Assum G, Galonska C, Siebert R, et al. 2019. Analysis of short tandem repeat expansions and their methylation state with nanopore sequencing. *Nat Biotechnol* **37:** 1478–1481. doi:10.1038/s41587-019-0293-x
Gigante S, Gouil Q, Lucattini A, Keniry A, Beck T, Tinning M, Gordon L, Woodruff C, Speed TP, Blewitt ME, et al. 2019. Using long-read sequencing to detect imprinted DNA methylation. *Nucleic Acids Res* **47:** e46. doi:10.1093/nar/gkz107
Goenka SD, Gorzynski JE, Shafin K, Fisk DG, Pesout T, Jensen TD, Monlong J, Chang P-C, Baid G, Bernstein JA, et al. 2022. Accelerated identification of disease-causing variants with ultra-rapid nanopore genome sequencing. *Nat Biotechnol* **40:** 1035–1041. doi:10.1038/s41587-022-01221-5
Gong Y, Li Y, Liu X, Ma Y, Jiang L. 2023. A review of the pangenome: how it affects our understanding of genomic variation, selection and breeding in domestic animals? *J Anim Sci Biotechnol* **14:** 73. doi:10.1186/s40104-023-00860-1
Goodwin S, Gurtowski J, Ethe-Sayers S, Deshpande P, Schatz MC, McCombie WR. 2015. Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome Res* **25:** 1750–1756. doi:10.1101/gr.191395.115
Gorzynski JE, Goenka SD, Shafin K, Jensen TD, Fisk DG, Grove ME, Spiteri E, Pesout T, Monlong J, Baid G, et al. 2022. Ultrarapid Nanopore genome sequencing in a critical care setting. *N Engl J Med* **386:** 700–702. doi:10.1056/NEJMc2112090
Groza C, Chen X, Pacis A, Simon M-M, Pramatarova A, Aracena KA, Pastinen T, Barreiro LB, Bourque G. 2023. Genome graphs detect human polymorphisms in active epigenomic state during influenza infection. *Cell Genom* **3:** 100294. doi:10.1016/j.xgen.2023.100294
Guarracino A, Heumos S, Nahnsen S, Prins P, Garrison E. 2022. ODGI: understanding pangenome graphs. *Bioinformatics* **38:** 3319–3326. doi:10.1093/bioinformatics/btac308
Gudmundsson S, Singer-Berk M, Watts NA, Phu W, Goodrich JK, Solomonson M, Genome Aggregation Database Consortium, Rehm HL, MacArthur DG, O'Donnell-Luria A. 2022. Variant interpretation using population databases: lessons from gnomAD. *Hum Mutat* **43:** 1012–1030. doi:10.1002/humu.24309
Guo R, Li Y-R, He S, Ou-Yang L, Sun Y, Zhu Z. 2018. RepLong: de novo repeat identification using long read sequencing data. *Bioinformatics* **34:** 1099–1107. doi:10.1093/bioinformatics/btx717
Gustafson JA, Gibson SB, Damaraju N, Zalusky MP, Hoekzema K, Twesigomwe D, Yang L, Snead AA, Richmond PA, De Coster W, et al. 2024. Nanopore sequencing of 1000 Genomes Project samples to build a comprehensive catalog of human genetic variation. medRxiv doi:10.1101/2024.03.05.24303792
Hall CL, Kesharwani RK, Phillips NR, Planz JV, Sedlazeck FJ, Zascavage RR. 2022. Accurate profiling of forensic autosomal STRs using the Oxford Nanopore Technologies MinION device. *Forensic Sci Int Genet* **56:** 102629. doi:10.1016/j.fsigen.2021.102629
Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. 2005. Online Mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* **33:** D514–D517. doi:10.1093/nar/gki033
Han R, Li Y, Gao X, Wang S. 2018. An accurate and rapid continuous wavelet dynamic time warping algorithm for end-to-end mapping in ultra-long nanopore sequencing. *Bioinformatics* **34:** i722–i731. doi:10.1093/bioinformatics/bty555
Harvey WT, Ebert P, Ebler J, Audano PA, Munson KM, Hoekzema K, Porubsky D, Beck CR, Marschall T, Garimella K, et al. 2023. Whole-genome long-read sequencing downsampling and its effect on variant-calling precision and recall. *Genome Res* **33:** 2029–2040. doi:10.1101/gr.278070.123
Heather JM, Chain B. 2016. The sequence of sequencers: the history of sequencing DNA. *Genomics* **107:** 1–8. doi:10.1016/j.ygeno.2015.11.003
Heavens D, Chooneea D, Giolai M, Cuber P, Aanstad P, Martin S, Alston M, Misra R, Clark MD, Leggett RM. 2021. How low can you go? Driving down the DNA input requirements for nanopore sequencing. bioRxiv doi:10.1101/2021.10.15.464554
Heller D, Vingron M. 2021. SVIM-asm: structural variant detection from haploid and diploid genome assemblies. *Bioinformatics* **36:** 5519–5521. doi:10.1093/bioinformatics/btaa1034
Hickey G, Monlong J, Ebler J, Novak AM, Eizenga JM, Gao Y, Human Pangenome Reference Consortium, Marschall T, Li H, Paten B. 2024. Pangenome graph construction from genome alignments with minigraph-cactus. *Nat Biotechnol* **42:** 663–673. doi:10.1038/s41587-023-01793-w
Holt JM, Saunders CT, Rowell WJ, Kronenberg Z, Wenger AM, Eberle M. 2024. HiPhase: jointly phasing small, structural, and tandem repeat variants from HiFi sequencing. *Bioinformatics* **40:** btae042. doi:10.1093/bioinformatics/btae042
Höps W, Weiss MM, Derks R, Galbany JC, den Ouden A, van den Heuvel S, Timmermans R, Smits J, Mokveld T, Dolzhenko E, et al. 2025. HiFi long-read genomes for difficult-to-detect, clinically relevant variants. *Am J Hum Genet* **112:** 450–456. doi:10.1016/j.ajhg.2024.12.013
Hunt SE, Moore B, Amode RM, Armean IM, Lemos D, Mushtaq A, Parton A, Schuilenburg H, Szpak M, Thormann A, et al. 2022. Annotating and prioritizing genomic variants using the ensembl variant effect predictor—a tutorial. *Hum Mutat* **43:** 986–997. doi:10.1002/humu.24298
Izydorczyk MB, Kalef-Ezra E, Horner DW, Zheng X, Holmes N, Toffoli M, Sahin ZG, Han Y, Mehta HH, Muzny DM, et al. 2024. Single cell long read whole genome sequencing reveals somatic transposon activity in human brain. medRxiv doi:10.1101/2024.11.11.24317113
Jain M, Olsen HE, Paten B, Akeson M. 2016. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol* **17:** 239. doi:10.1186/s13059-016-1103-0
Jain C, Rhie A, Hansen NF, Koren S, Phillippy AM. 2022. Long-read mapping to repetitive reference sequences using Winnowmap2. *Nat Methods* **19:** 705–710. doi:10.1038/s41592-022-01457-8
Jensen TD, Ni B, Reuter CM, Gorzynski JE, Fazal S, Bonner D, Ungar RA, Goddard PC, Raja A, Ashley EA, et al. 2024. Integration of transcriptomics and long-read genomics prioritizes structural variants in rare disease. medRxiv doi:10.1101/2024.03.22.24304565
Jha A, Bohaczuk SC, Mao Y, Ranchalis J, Mallory BJ, Min AT, Hamm MO, Swanson E, Dubocanin D, Finkbeiner C, et al. 2024. DNA-m6A calling

and integrated long-read epigenetic and genetic analysis with *fibertools*. *Genome Res* **34:** 1976–1986. doi:10.1101/gr.279095.124

Jiang T, Liu Y, Jiang Y, Li J, Gao Y, Cui Z, Liu Y, Liu B, Wang Y. 2020. Long-read-based human genomic structural variation detection with cuteSV. *Genome Biol* **21:** 189. doi:10.1186/s13059-020-02107-y

Keskus A, Bryant A, Ahmad T, Yoo B, Aganezov S, Goretsky A, Donmez A, Lansdon LA, Rodriguez I, Park J, et al. 2024. Severus: accurate detection and characterization of somatic structural variation in tumor genomes using long reads. medRxiv doi:10.1101/2024.03.22.24304756

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37:** 907–915. doi:10.1038/s41587-019-0201-4

King J, Harder T, Beer M, Pohlmann A. 2020. Rapid multiplex MinION nanopore sequencing workflow for influenza A viruses. *BMC Infect Dis* **20:** 648. doi:10.1186/s12879-020-05367-y

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46:** 310–315. doi:10.1038/ng.2892

Kirsche M, Prabhu G, Sherman R, Ni B, Battle A, Aganezov S, Schatz MC. 2023. Jasmine and Iris: population-scale structural variant comparison and analysis. *Nat Methods* **20:** 408–417. doi:10.1038/s41592-022-01753-3

Kleinert P, Kircher M. 2022. A framework to score the effects of structural variants in health and disease. *Genome Res* **32:** 766–777. doi:10.1101/gr.275995.121

Kobayashi Y, Yang S, Nykamp K, Garcia J, Lincoln SE, Topper SE. 2017. Pathogenic variant burden in the ExAC database: an empirical approach to evaluating population data for clinical variant interpretation. *Genome Med* **9:** 13. doi:10.1186/s13073-017-0403-7

Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* **37:** 540–546. doi:10.1038/s41587-019-0072-8

Kong Y, Mead EA, Fang G. 2023. Navigating the pitfalls of mapping DNA and RNA modifications. *Nat Rev Genet* **24:** 363–381. doi:10.1038/s41576-022-00559-5

Konings M, Gerrits van den Ende B, Raats MWJ, Fahal AH, van de Sande WWJ, Hagen F. 2024. Complete genome sequence of the itraconazole decreased susceptible *Madurella fahalii* type-strain CBS 129176. *Mycopathologia* **189:** 6. doi:10.1007/s11046-023-00807-0

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Res* **27:** 722–736. doi:10.1101/gr.215087.116

Koren S, Rhie A, Walenz BP, Dilthey AT, Bickhart DM, Kingan SB, Hiendleder S, Williams JL, Smith TPL, Phillippy AM. 2018. De novo assembly of haplotype-resolved genomes with trio binning. *Nat Biotechnol* **36:** 1174–1182. doi:10.1038/nbt.4277

Koren MJ, Rodriguez F, East C, Toth PP, Watwe V, Abbas CA, Sarwat S, Kleeman K, Kumar B, Ali Y, et al. 2024. An "inclisiran first" strategy vs usual care in patients with atherosclerotic cardiovascular disease. *J Am Coll Cardiol* **83:** 1939–1952. doi:10.1016/j.jacc.2024.03.382

Kosugi S, Terao C. 2024. Comparative evaluation of SNVs, indels, and structural variations detected with short- and long-read sequencing data. *Hum Genome Var* **11:** 18. doi:10.1038/s41439-024-00276-x

Lai S-K, Luo AC, Chiu I-H, Chuang H-W, Chou T-H, Hung T-K, Hsu JS, Chen C-Y, Yang W-S, Yang Y-C, et al. 2024. A novel framework for human leukocyte antigen (HLA) genotyping using probe capture-based targeted next-generation sequencing and computational analysis. *Comput Struct Biotechnol J* **23:** 1562–1571. doi:10.1016/j.csbj.2024.03.030

Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR. 2014. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* **42:** D980–D985. doi:10.1093/nar/gkt1113

Lau BT, Almeda A, Schauer M, McNamara M, Bai X, Meng Q, Partha M, Grimes SM, Lee H, Heestand GM, et al. 2023. Single-molecule methylation profiles of cell-free DNA in cancer with nanopore sequencing. *Genome Med* **15:** 33. doi:10.1186/s13073-023-01178-3

Laver T, Harrison J, O'Neill PA, Moore K, Farbos A, Paszkiewicz K, Studholme DJ. 2015. Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomol Detect Quantif* **3:** 1–8. doi:10.1016/j.bdq.2015.02.001

Lee I, Razaghi R, Gilpatrick T, Molnar M, Gershman A, Sadowski N, Sedlazeck FJ, Hansen KD, Simpson JT, Timp W. 2020. Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing. *Nat Methods* **17:** 1191–1199. doi:10.1038/s41592-020-01000-7

Lee JY, Kong M, Oh J, Lim J, Chung SH, Kim J-M, Kim J-S, Kim K-H, Yoo J-C, Kwak W. 2021. Comparative evaluation of nanopore polishing tools for microbial genome assembly and polishing strategies for downstream analysis. *Sci Rep* **11:** 20740. doi:10.1038/s41598-021-00178-w

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34:** 3094–3100. doi:10.1093/bioinformatics/bty191

Li H. 2021. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37:** 4572–4574. doi:10.1093/bioinformatics/btab705

Li H, Durbin R. 2024. Genome assembly in the telomere-to-telomere era. *Nat Rev Genet* **25:** 658–670. doi:10.1038/s41576-024-00718-w

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/Map format and SAMtools. *Bioinformatics* **25:** 2078–2079. doi:10.1093/bioinformatics/btp352

Li H, Bloom JM, Farjoun Y, Fleharty M, Gauthier L, Neale B, MacArthur D. 2018. A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nat Methods* **15:** 595–597. doi:10.1038/s41592-018-0054-7

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

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

Lin MF, Rodeh O, Penn J, Bai X, Reid JG, Krasheninina O, Salerno WJ. 2018. GLnexus: joint variant calling for large cohort sequencing. bioRxiv doi:10.1101/343970

Lin J-H, Chen L-C, Yu S-C, Huang Y-T. 2022a. Longphase: an ultra-fast chromosome-scale phasing algorithm for small and large variants. *Bioinformatics* **38:** 1816–1822. doi:10.1093/bioinformatics/btac058

Lin J, Wang S, Audano PA, Meng D, Flores JI, Kosters W, Yang X, Jia P, Marschall T, Beck CR, et al. 2022b. SVision: a deep learning approach to resolve complex structural variants. *Nat Methods* **19:** 1230–1233. doi:10.1038/s41592-022-01609-w

Liu Q, Zhang P, Wang D, Gu W, Wang K. 2017. Interrogating the "unsequenceable" genomic trinucleotide repeat disorders by long-read sequencing. *Genome Med* **9:** 65. doi:10.1186/s13073-017-0456-7

Logsdon GA, Vollger MR, Eichler EE. 2020. Long-read human genome sequencing and its applications. *Nat Rev Genet* **21:** 597–614. doi:10.1038/s41576-020-0236-x

Lucas MC, Novoa EM. 2023. Long-read sequencing in the era of epigenomics and epitranscriptomics. *Nat Methods* **20:** 25–29. doi:10.1038/s41592-022-01724-8

Luo J, Ding H, Shen J, Zhai H, Wu Z, Yan C, Luo H. 2021. BreakNet: detecting deletions using long reads and a deep learning approach. *BMC Bioinformatics* **22:** 577. doi:10.1186/s12859-021-04499-5

Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. 2019. Structural variant calling: the long and the short of it. *Genome Biol* **20:** 246. doi:10.1186/s13059-019-1828-7

Mahmoud M, Doddapaneni H, Timp W, Sedlazeck FJ. 2021. PRINCESS: comprehensive detection of haplotype resolved SNVs, SVs, and methylation. *Genome Biol* **22:** 268. doi:10.1186/s13059-021-02486-w

Mahmoud M, Harting J, Corbitt H, Chen X, Jhangiani SN, Doddapaneni H, Meng Q, Han T, Lambert C, Zhang S, et al. 2024a. Closing the gap: solving complex medically relevant genes at scale. medRxiv doi:10.1101/2024.03.14.24304179

Mahmoud M, Huang Y, Garimella K, Audano PA, Wan W, Prasad N, Handsaker RE, Hall S, Pionzio A, Schatz MC, et al. 2024b. Utility of long-read sequencing for all of us. *Nat Commun* **15:** 837. doi:10.1038/s41467-024-44804-3

Majidian S, Agustinho DP, Chin C-S, Sedlazeck FJ, Mahmoud M. 2023. Genomic variant benchmark: if you cannot measure it, you cannot improve it. *Genome Biol* **24:** 221. doi:10.1186/s13059-023-03061-1

Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 2021. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol Biol Evol* **38:** 4647–4654. doi:10.1093/molbev/msab199

Martin M, Ebert P, Marschall T. 2023. Read-based phasing and analysis of phased variants with WhatsHap. *Methods Mol Biol* **2590:** 127–138. doi:10.1007/978-1-0716-2819-5_8

Mastoras M, Asri M, Brambrink L, Hebbar P, Kolesnikov A, Cook DE, Nattestad M, Lucas J, Won TS, Chang P-C, et al. 2024. Highly accurate assembly polishing with DeepPolisher. bioRxiv doi:10.1101/2024.09.17.613505

Miga KH, Wang T. 2021. The need for a human pangenome reference sequence. *Annu Rev Genomics Hum Genet* **22:** 81–102. doi:10.1146/annurev-genom-120120-081921

Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A. 2018. Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics* **34:** i142–i150. doi:10.1093/bioinformatics/bty266

Ni P, Nie F, Zhong Z, Xu J, Huang N, Zhang J, Zhao H, Zou Y, Huang Y, Li J, et al. 2023a. DNA 5-methylcytosine detection and methylation phasing using PacBio circular consensus sequencing. *Nat Commun* **14:** 4054. doi:10.1038/s41467-023-39784-9

Ni Y, Liu X, Simeneh ZM, Yang M, Li R. 2023b. Benchmarking of nanopore R10.4 and R9.4.1 flow cells in single-cell whole-genome amplification and whole-genome shotgun sequencing. *Comput Struct Biotechnol J* **21:** 2352–2364. doi:10.1016/j.csbj.2023.03.038

Nicholas TJ, Cormier MJ, Quinlan AR. 2022. Annotation of structural variants with reported allele frequencies and related metrics from multiple datasets using SVAFotate. *BMC Bioinformatics* **23:** 490. doi:10.1186/s12859-022-05008-y

Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, et al. 2022. The complete sequence of a human genome. *Science* **376:** 44–53. doi:10.1126/science.abj6987

Oehler JB, Wright H, Stark Z, Mallett AJ, Schmitz U. 2023. The application of long-read sequencing in clinical settings. *Hum Genomics* **17:** 73. doi:10.1186/s40246-023-00522-3

Olson ND, Wagner J, Dwarshuis N, Miga KH, Sedlazeck FJ, Salit M, Zook JM. 2023. Variant calling and benchmarking in an era of complete human genome sequences. *Nat Rev Genet* **24:** 464–483. doi:10.1038/s41576-023-00590-0

Oxford Nanopore Technologies 2020. R10.3: the newest nanopore for high accuracy nanopore sequencing—now available in store. https://nanoporetech.com/news/news-r103-newest-nanopore-high-accuracy-nanopore-sequencing-now-available-store [accessed August 19, 2024].

Pardo-Palacios FJ, Wang D, Reese F, Diekhans M, Carbonell-Sala S, Williams B, Loveland JE, De María M, Adams MS, Balderrama-Gutierrez G, et al. 2024. Systematic assessment of long-read RNA-seq methods for transcript identification and quantification. *Nat Methods* **21:** 1349–1363. doi:10.1038/s41592-024-02298-3

Park J, Cook DE, Chang P-C, Kolesnikov A, Brambrink L, Mier JC, Gardner J, McNulty B, Sacco S, Keskus A, et al. 2024. DeepSomatic: accurate somatic small variant discovery for multiple sequencing technologies. bioRxiv doi:10.1101/2024.08.16.608331

Paten B, Novak AM, Eizenga JM, Garrison E. 2017. Genome graphs and the evolution of genome inference. *Genome Res* **27:** 665–676. doi:10.1101/gr.214155.116

Poplin R, Chang P-C, Alexander D, Schwartz S, Colthurst T, Ku A, Newburger D, Dijamco J, Nguyen N, Afshar PT, et al. 2018. A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol* **36:** 983–987. doi:10.1038/nbt.4235

Quang D, Chen Y, Xie X. 2015. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* **31:** 761–763. doi:10.1093/bioinformatics/btu703

Rashid U, Wu C, Shiller J, Smith K, Crowhurst R, Davy M, Chen T-H, Carvajal I, Bailey S, Thomson S, et al. 2024. AssemblyQC: a nextflow pipeline for reproducible reporting of assembly quality. *Bioinformatics* **40:** btae477. doi:10.1093/bioinformatics/btae477

Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. 2012. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28:** i333–i339. doi:10.1093/bioinformatics/bts378

Rautiainen M, Nurk S, Walenz BP, Logsdon GA, Porubsky D, Rhie A, Eichler EE, Phillippy AM, Koren S. 2023. Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nat Biotechnol* **41:** 1474–1482. doi:10.1038/s41587-023-01662-6

Ren J, Chaisson MJP. 2021. lra: a long read aligner for sequences and contigs. *PLoS Comput Biol* **17:** e1009078. doi:10.1371/journal.pcbi.1009078

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

Rocha JL, Lou RN, Sudmant PH. 2024. Structural variation in humans and our primate kin in the era of telomere-to-telomere genomes and pangenomics. *Curr Opin Genet Dev* **87:** 102233. doi:10.1016/j.gde.2024.102233

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

Salzberg SL. 2019. Next-generation genome annotation: we still struggle to get it right. *Genome Biol* **20:** 92. doi:10.1186/s13059-019-1715-2

Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, Koren S, Treangen TJ, Schatz MC, Delcher AL, Roberts M, et al. 2012. GAGE: a critical evaluation of genome assemblies and assembly algorithms. *Genome Res* **22:** 557–567. doi:10.1101/gr.131383.111

Sanderson ND, Hopkins KMV, Colpus M, Parker M, Lipworth S, Crook D, Stoesser N. 2024. Evaluation of the accuracy of bacterial genome reconstruction with Oxford nanopore R10.4.1 long-read-only sequencing. *Microb Genom* **10:** 001246. doi:10.1099/mgen.0.001246

Saunders CT, Holt JM, Baker DN, Lake JA, Belyeu JR, Kronenberg Z, Rowell WJ, Eberle MA. 2024. Sawfish: improving long-read structural variant discovery and genotyping with local haplotype modeling. bioRxiv doi:10.1101/2024.08.19.608674

Schmeing S, Robinson MD. 2023. Gapless provides combined scaffolding, gap filling, and assembly correction with long reads. *Life Science Alliance* **6:** e202201471. doi:10.26508/lsa.202201471

Secomandi S, Gallo GR, Sozzoni M, Iannucci A, Galati E, Abueg L, Balacco J, Caprioli M, Chow W, Ciofi C, et al. 2023. A chromosome-level reference genome and pangenome for barn swallow population genomics. *Cell Rep* **42:** 111992. doi:10.1016/j.celrep.2023.111992

Sedlazeck FJ, Lee H, Darby CA, Schatz MC. 2018a. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nat Rev Genet* **19:** 329–346. doi:10.1038/s41576-018-0003-4

Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, Schatz MC. 2018b. Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods* **15:** 461–468. doi:10.1038/s41592-018-0001-7

Shafin K, Pesout T, Lorig-Roach R, Haukness M, Olsen HE, Bosworth C, Armstrong J, Tigyi K, Maurer N, Koren S, et al. 2020. Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nat Biotechnol* **38:** 1044–1053. doi:10.1038/s41587-020-0503-6

Sherman RM, Salzberg SL. 2020. Pan-genomics in the human genome era. *Nat Rev Genet* **21:** 243–254. doi:10.1038/s41576-020-0210-7

Sherry ST, Ward M, Sirotkin K. 1999. dbSNP—database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Res* **9:** 677–679. doi:10.1101/gr.9.8.677

Shih PJ, Saadat H, Parameswaran S, Gamaarachchi H. 2022. Efficient real-time selective genome sequencing on resource-constrained devices. *Gigascience* **12:** giad046. doi:10.1093/gigascience/giad046

Shiraishi Y, Koya J, Chiba K, Okada A, Arai Y, Saito Y, Shibata T, Kataoka K. 2023. Precise characterization of somatic complex structural variations from tumor/control paired long-read sequencing data with nanomonsv. *Nucleic Acids Res* **51:** e74. doi:10.1093/nar/gkad526

Shumate A, Salzberg SL. 2021. Liftoff: accurate mapping of gene annotations. *Bioinformatics* **37:** 1639–1643. doi:10.1093/bioinformatics/btaa1016

Sim N-L, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. 2012. SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res* **40:** W452–W457. doi:10.1093/nar/gks539

Sirén J, Monlong J, Chang X, Novak AM, Eizenga JM, Markello C, Sibbesen JA, Hickey G, Chang P-C, Carroll A, et al. 2021. Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science* **374:** abg8871. doi:10.1126/science.abg8871

Smolka M, Paulin LF, Grochowski CM, Horner DW, Mahmoud M, Behera S, Kalef-Ezra E, Gandhi M, Hong K, Pehlivan D, et al. 2024. Detection of mosaic and population-level structural variants with Sniffles2. *Nat Biotechnol* **42:** 1571–1580. doi:10.1038/s41587-023-02024-y

Stergachis AB, Debo BM, Haugen E, Churchman LS, Stamatoyannopoulos JA. 2020. Single-molecule regulatory architectures captured by chromatin fiber sequencing. *Science* **368:** 1449–1454. doi:10.1126/science.aaz1646

Sulovari A, Li R, Audano PA, Porubsky D, Vollger MR, Logsdon GA, Human Genome Structural Variation Consortium, Warren WC, Pollen AA, Chaisson MJP, et al. 2019. Human-specific tandem repeat expansion and differential gene expression during primate evolution. *Proc Natl Acad Sci* **116:** 23243–23253. doi:10.1073/pnas.1912175116

Sun B, Pashkova L, Pieters PA, Harke AS, Mohite OS, Santos A, Zielinski DC, Palsson BO, Phaneuf PV. 2025. PanKB: an interactive microbial pangenome knowledgebase for research, biotechnological innovation, and knowledge mining. *Nucleic Acids Res* **53:** D806–D818. doi:10.1093/nar/gkae1042

Suzuki A, Suzuki M, Mizushima-Sugano J, Frith MC, Makalowski W, Kohno T, Sugano S, Tsuchihara K, Suzuki Y. 2017. Sequencing and phasing cancer mutations in lung cancers using a long-read portable sequencer. *DNA Res* **24:** 585–596. doi:10.1093/dnares/dsx027

Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. 2015. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **31:** 2032–2034. doi:10.1093/bioinformatics/btv098

Ummat A, Bashir A. 2014. Resolving complex tandem repeats with long reads. *Bioinformatics* **30:** 3491–3498. doi:10.1093/bioinformatics/btu437

Vollger MR, Dishuck PC, Sorensen M, Welch AE, Dang V, Dougherty ML, Graves-Lindsay TA, Wilson RK, Chaisson MJP, Eichler EE. 2019. Long-read sequence and assembly of segmental duplications. *Nat Methods* **16:** 88–94. doi:10.1038/s41592-018-0236-3

Vollger MR, Korlach J, Eldred KC, Swanson E, Underwood JG, Bohaczuk SC, Mao Y, Cheng Y-HH, Ranchalis J, Blue EE, et al. 2025. Synchronized long-read genome, methylome, epigenome, and transcriptome profiling resolve a Mendelian condition. *Nat Genet* **57:** 469479. doi:10.1038/s41588-024-02067-0

Wang K, Li M, Hakonarson H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38:** e164. doi:10.1093/nar/gkq603

Wang Q, Xiong F, Wu G, Liu W, Chen J, Wang B, Chen Y. 2022. Gene body methylation in cancer: molecular mechanisms and clinical applications. *Clin Epigenetics* **14:** 154. doi:10.1186/s13148-022-01382-9

Wasswa FB, Kassaza K, Nielsen K, Bazira J. 2022. MinION whole-genome sequencing in resource-limited settings: challenges and opportunities. *Curr Clin Microbiol Rep* **9:** 52–59. doi:10.1007/s40588-022-00183-1

Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND, et al. 2019. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol* **37:** 1155–1162. doi:10.1038/s41587-019-0217-9

Wojcik MH, Reuter CM, Marwaha S, Mahmoud M, Duyzend MH, Barseghyan H, Yuan B, Boone PM, Groopman EE, Délot EC, et al. 2023. Beyond the exome: what's next in diagnostic testing for Mendelian conditions. *Am J Hum Genet* **110:** 1229–1248. doi:10.1016/j.ajhg.2023.06.009

Xie S, Leung AW-S, Zheng Z, Zhang D, Xiao C, Luo R, Luo M, Zhang S. 2021. Applications and potentials of nanopore sequencing in the (epi)genome and (epi)transcriptome era. *Innovation (Camb)* **2:** 100153. doi:10.1016/j.xinn.2021.100153

Xu M, Guo L, Gu S, Wang O, Zhang R, Peters BA, Fan G, Liu X, Xu X, Deng L, et al. 2020. TGS-GapCloser: a fast and accurate gap closer for large genomes with low coverage of error-prone long reads. *Gigascience* **9:** giaa094. doi:10.1093/gigascience/giaa094

Zhang H, Li H, Jain C, Cheng H, Au KF, Li H, Aluru S. 2021. Real-time mapping of nanopore raw signals. *Bioinformatics* **37:** i477–i483. doi:10.1093/bioinformatics/btab264

Zheng Z, Li S, Su J, Leung AW-S, Lam T-W, Luo R. 2022. Symphonizing pileup and full-alignment for deep learning-based long-read variant calling. *Nat Comput Sci* **2:** 797–803. doi:10.1038/s43588-022-00387-x

Zheng X, Chowdhury M, Mirpochoev B, Clauset A, Layer RM, Sedlazeck FJ. 2024. STIX: long-reads based accurate structural variation annotation at population scale. bioRxiv doi:10.1101/2024.09.30.615931

Zhou Y, Song L, Li H. 2024. Full-resolution HLA and KIR gene annotations for human genome assemblies. *Genome Res* **34:** 1931–1941. doi:10.1101/gr.278985.124