Unraveling the hidden complexity of cancer through long-read sequencing

Qiuhui Li, Ayse G. Keskus, Justin Wagner, et al.

Genome Res. published online March 20, 2025
Access the most recent version at doi:10.1101/gr.280041.124

Review

Unraveling the hidden complexity of cancer through long-read sequencing

Qiuhui Li,[1] Ayse G. Keskus,[2] Justin Wagner,[3] Michal B. Izydorczyk,[4] Winston Timp,[5] Fritz J. Sedlazeck,[4,6,7] Alison P. Klein,[8] Justin M. Zook,[3] Mikhail Kolmogorov,[2] and Michael C. Schatz[1,8]

[1]*Department of Computer Science, Johns Hopkins University, Baltimore, Maryland 21218, USA;* [2]*Cancer Data Science Laboratory, Center for Cancer Research, National Cancer Institute, Bethesda, Maryland 20892, USA;* [3]*Material Measurement Laboratory, National Institute of Standards and Technology, Gaithersburg, Maryland 20899, USA;* [4]*Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas 77030, USA;* [5]*Department of Biomedical Engineering, Johns Hopkins University, Baltimore, Maryland 21218, USA;* [6]*Department of Molecular and Human Genetics, Baylor College of Medicine, Texas 77030, USA;* [7]*Department of Computer Science, Rice University, Houston, Texas 77251, USA;* [8]*Sidney Kimmel Comprehensive Cancer Center, Department of Oncology, Johns Hopkins Medicine, Baltimore, Maryland 21031, USA*

Cancer is fundamentally a disease of the genome, characterized by extensive genomic, transcriptomic, and epigenomic alterations. Most current studies predominantly use short-read sequencing, gene panels, or microarrays to explore these alterations; however, these technologies can systematically miss or misrepresent certain types of alterations, especially structural variants, complex rearrangements, and alterations within repetitive regions. Long-read sequencing is rapidly emerging as a transformative technology for cancer research by providing a comprehensive view across the genome, transcriptome, and epigenome, including the ability to detect alterations that previous technologies have overlooked. In this review, we explore the current applications of long-read sequencing for both germline and somatic cancer analysis. We provide an overview of the computational methodologies tailored to long-read data and highlight key discoveries and resources within cancer genomics that were previously inaccessible with prior technologies. We also address future opportunities and persistent challenges, including the experimental and computational requirements needed to scale to larger sample sizes, the hurdles in sequencing and analyzing complex cancer genomes, and opportunities for leveraging machine learning and artificial intelligence technologies for cancer informatics. We further discuss how the telomere-to-telomere genome and the emerging human pangenome could enhance the resolution of cancer genome analysis, potentially revolutionizing early detection and disease monitoring in patients. Finally, we outline strategies for transitioning long-read sequencing from research applications to routine clinical practice.

A hallmark of cancer is widespread genetic and epigenetic instability (Hanahan and Weinberg 2011). Cancer often originates through somatic mutations that accumulate throughout an individual's lifetime due to exposure to carcinogens, DNA replication errors, and other factors, thereby leading to the clonal evolution of cancer cells (Qing et al. 2020; Chakravarty and Solit 2021). Somatic mutations in key genes and their regulatory sequences, especially oncogenes, tumor suppressors, and DNA repair genes, are frequently found in cancers and contribute to tumor progression and therapeutic resistance (Olivier et al. 2010; Prior et al. 2012). In addition, germline pathogenic variants, while less prevalent than somatic mutations for driving cancer risk, account for at least 5%–10% of all cancers. These inherited variants originate in reproductive cells and are passed from parents to offspring (Pudjihartono et al. 2022). Clinically pathogenic variants, particularly those in high-penetrance genes such as *BRCA1/2* and *APC*, are of primary importance in cancer risk assessment and treatment (Preisler et al. 2021; Nepomuceno et al. 2024).

The rapid advancement of genome sequencing technologies over the past 30 years has revolutionized our ability to catalog cancer risk variants and understand the genomic landscape of cancer. Early studies relied on targeted sequencing and microarrays to detect genomic and transcriptomic variations enriched in cancer patients, revealing, for example, the prevalence of widespread mutations in *TP53* (Nigro et al. 1989) and the molecular basis of the subtypes of breast cancers (Perou et al. 2000). While exome-wide studies were possible with these technologies, sample sizes were modest (Wood et al. 2007; Jones et al. 2009). Nevertheless, these technologies remain clinically important to this day due to their predictable performance and low costs.

More recently, short-read sequencing has gained widespread adoption for cancer research as it enables the genome-wide identification of alterations, including single-nucleotide variants (SNVs), small insertions and deletions (indels), copy number alterations, and some structural variants (SVs) (Choo et al. 2023). This high-throughput, cost-effective technology has facilitated large-scale cancer genomics projects like The Cancer Genome Atlas (TCGA) (The Cancer Genome Atlas Research Network et al. 2013), the International Cancer Genome

Corresponding authors: mikhail.kolmogorov@nih.gov, mschatz@cs.jhu.edu

Article published online before print. Article and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.280041.124. Freely available online through the *Genome Research* Open Access option.

Consortium (ICGC) (Zhang et al. 2019), and the Pan-Cancer Analysis of Whole Genomes (PCAWG) project (The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium 2020) to characterize thousands of cancers across dozens of cancer types. Building on these population-scale efforts, the Catalogue Of Somatic Mutations In Cancer (COSMIC) has emerged as a comprehensive resource for somatic mutation data (Tate et al. 2019). In its latest release, COSMIC V100 expanded to include over 24 million gene mutations, incorporating genome-wide sequencing results from tens of thousands of tumors across diverse cancer types. These collaborative efforts have yielded invaluable insights into the molecular mechanisms underlying cancer initiation, progression, and metastasis, uncovering recurring patterns of genomic alterations, novel cancer genes, and pathways. Consequently, short-read sequencing for both tumor and germline is routinely conducted as part of patient management.

While short-read sequencing has greatly advanced our understanding of cancer susceptibility and progression, it faces several major challenges that have systematically excluded certain parts of the genome and certain types of variations from studies. Most notably, short-read sequencing is notoriously limited for germline analysis of structural variations, defined as any variant at least 50 bp in size, including insertions, deletions, inversions, duplications, and other complex variant types. While fewer in number than SNVs, because of their larger size, SVs account for a larger number of variant bases across the germline genome. SVs are also often strongly associated with alterations in gene expression (Chiang et al. 2017; Scott et al. 2021), and are emerging as an important source of pathogenic variations (Kleinert and Kircher 2022). In addition to SVs, short-reads struggle with detecting alterations in tandem repeats (TRs), and more generally, with variations found within repetitive sequences (e.g., segmental duplications, satellite sequences, transposable elements), leaving a large fraction of the genome inaccessible. Short-reads also have limited power to phase variants within the same haplotype, meaning they often cannot conclusively determine if there has been a complete loss of function in genes from compound heterozygous mutations (Sedlazeck et al. 2018). In addition to these limitations for detecting DNA variations and mutations, short-reads also face related challenges resolving alterations in transcriptomes or epigenomes and cannot simultaneously capture genomic and epigenomic variations, which further hinders the phasing of methylation and the detection of allele-specific methylation (ASM).

Addressing these challenges, long-read sequencing technologies, such as Pacific Biosciences (PacBio) Single Molecule Real Time sequencing and Oxford Nanopore Technologies (ONT) sequencing, have emerged as powerful alternatives to short-read sequencing (Fig. 1; van Dijk et al. 2023). Their extended read lengths can span difficult-to-resolve regions, improving the identification performance of structural variants, as well as improving the identification and phasing of variations in repetitive regions (Fig. 1B). The growth in adoption has mirrored the overall improvements to these platforms: while the initial release of these technologies was plagued by high error rates, high costs, and low throughput, current long-read sequencing platforms have achieved high accuracy and throughput at competitive costs (Kovaka et al. 2023) promoting a variety of new applications.

One of the most notable examples of the performance of long reads was how in 2022 long-read sequencing was used to produce the first telomere-to-telomere (T2T) assembly of a human genome (Nurk et al. 2022). Compared to previous references, the new T2T reference genome corrected thousands of structural errors and revealed >100 Mbp of previously uncharacterized regions of the human genome. Moreover, long reads have been essential to develop population-scale references of structural variations (SVs) in the human population (Ebert et al. 2021; Liao et al. 2023; Mahmoud et al. 2024), as well as to detect novel gene models and epigenetic changes (Gershman et al. 2022; Glinos et al. 2022; Kolmogorov et al. 2023; Kovaka et al. 2024; Stefansson et al. 2024). In addition to its transformative impact on basic sciences, long-read sequencing is emerging as a critical tool in clinical diagnostics. Long-read's ability to accurately detect SVs and other complex variants missed by traditional short-read sequencing has proven invaluable in identifying causal SVs in Mendelian disease (Merker et al. 2018), repeat expansions underlying neurodevelopmental disorders (Ishiura et al. 2018; Hiatt et al. 2021), and many other conditions (Marwaha et al. 2022; Damaraju et al. 2024). A particularly notable clinical application of long-read sequencing is the development of rapid (same day) sequencing workflows (Gorzynski et al. 2022), which are crucial for critically ill patients, where timely and accurate genetic diagnosis can directly influence treatment decisions, potentially improving prognosis and reducing healthcare costs.

Within cancer genomics, long-read sequencing has similarly witnessed growing adoption, especially to resolve SVs and other complex variations (Sakamoto et al. 2020a, 2021b). The earliest applications used long-read sequencing of cancer cell lines to validate known structural variations (Norris et al. 2016) and then construct whole-genome catalogs of variations for the first time (Nattestad et al. 2018). These early studies revealed that these cell lines contained tens of thousands of structural variations, most of which had been missed by short-read sequencing, as well as intricate genomic rearrangements in key oncogenes and novel gene fusions (Nattestad et al. 2018). More recent studies have expanded the scope of this work to consider larger numbers of patient samples and more complex types of variation, as well as transcriptomic and epigenomic readouts (Fig. 1B–D; Sereewattanawoot et al. 2018; Aganezov et al. 2020; Sakamoto et al. 2020b, 2022; Zheng et al. 2024a). These efforts have consistently demonstrated the superior capability of long-read sequencing technologies in constructing a more comprehensive cancer genome landscape and elucidating their functional consequences. Here, we summarize the rise of these efforts, focusing on the unique opportunities long reads offer for cancer analysis across the genome, transcriptome, and epigenome (Fig. 1A–D). We then discuss the remaining needs and challenges, and conclude with our perspective on the path toward integrating long reads as a routine component of clinical cancer diagnosis and treatment.

Long-read genome sequencing

Growth of germline population sequencing and resources

Before advancing to cancer genomes, long-read sequencing was initially applied to improve the resolution of the reference human genome and to develop comprehensive catalogs of human genetic variation in healthy donors. The first genome-wide analysis of a human genome using long reads was published by Chaisson et al. (2015). In this work, they sequenced a haploid human cell line (CHM1) with PacBio SMRT technology, and found using these data they could close or extend 55% of the remaining gaps in the GRCh37 reference genome—78% of which contained extensive degenerate short TRs, particularly in (G + C)-rich areas. They also resolved over 26,000 structural variants at base pair resolution, including previously undetectable inversions, complex insertions,

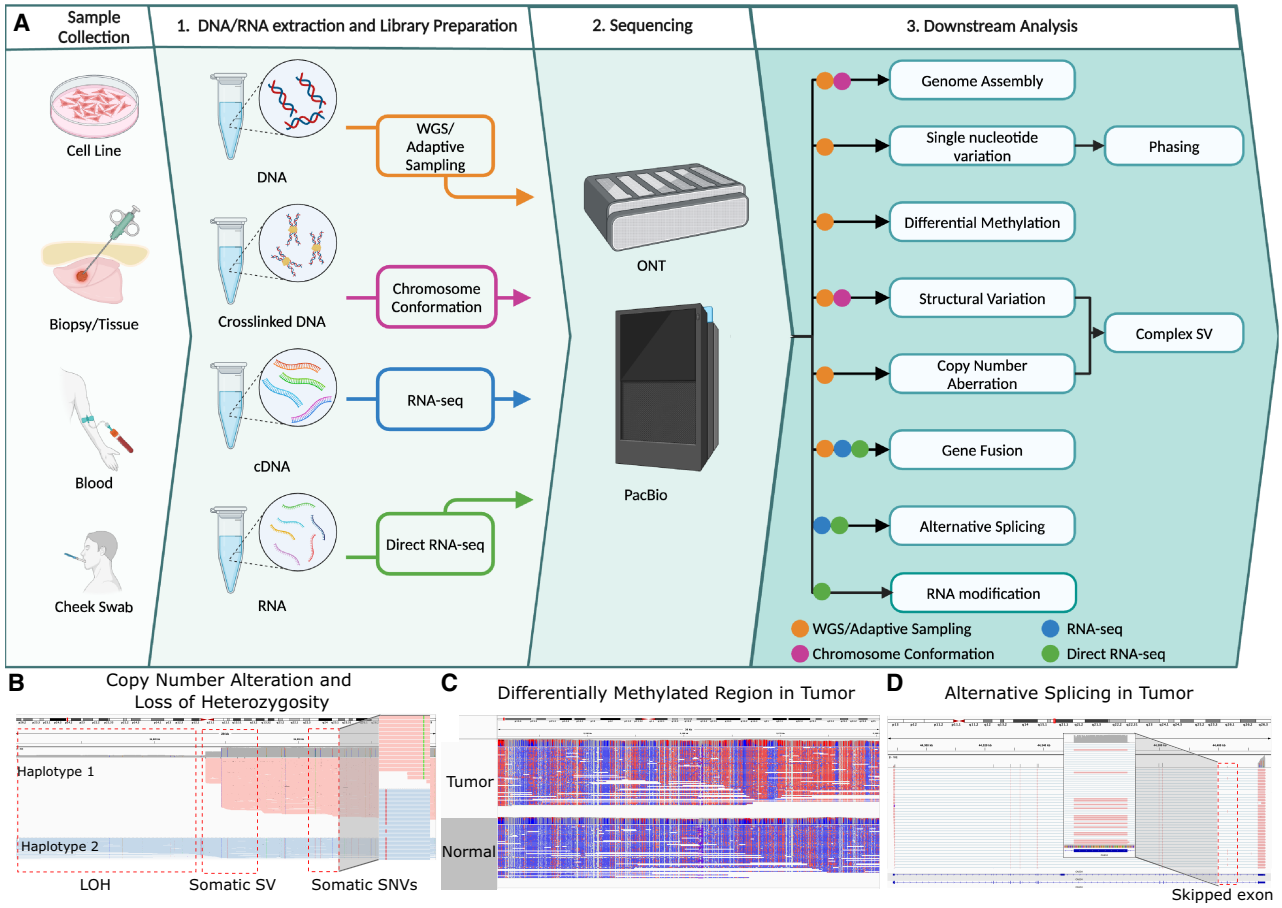

Figure 1. Overview of long-read sequencing protocols and analyses. (*A*) The overview of the long-read sequencing workflow including downstream analysis options. (*B*) Somatic SNV and a structural variation in the COLO829 cell line resulting in haplotype-specific copy number change and loss of heterozygosity (data from Keskus et al. 2024). Reads are grouped and colored with the alleles using long-read-based phasing. (*C*) Differentially methylated region between tumor and normal sample. Reads are colored with 5mC modifications (red: high methylation, blue: low methylation). (*D*) An alternative splicing in *CASC4* gene in cancer cell line SCC152 (Rodriguez et al. 2024).

and long TR tracts. This landmark study highlighted the increased complexity of the human genome, particularly in repetitive DNA regions, which could now be more completely and accurately resolved with long-read sequencing.

In the decade following, further advances in generating longer, more accurate reads enabled the T2T consortium to achieve the first complete human genome sequencing. This began in 2020 with the publication of the first gapless assembly of the human X Chromosome using ultra-long-read nanopore sequencing (Miga et al. 2020). Then in 2022, the consortium published T2T-CHM13 as the first fully gapless human genome assembled primarily from PacBio HiFi and ONT Ultralong reads. This genome comprises 3.055 billion base pairs and addresses the previously unresolved 8% of the genome, including complex heterochromatic regions (Nurk et al. 2022). In 2023, the consortium finalized the complete Y Chromosome sequence (T2T-Y) from the HG002 donor (Rhie et al. 2023). The integration of T2T-Y with the CHM13 reference, along with population variation, clinical variants, and functional genomics data, has created a comprehensive reference for all 24 human chromosomes. The use of the T2T-CHM13 genome as a reference broadly improves our ability to detect variation in human genomes. Specifically, T2T-CHM13 improves the Mendelian concordance rate among trios and eliminates tens of thousands of spurious SNVs per sample, including a reduction of false positives in 269 challenging, medically relevant genes by up to a factor of 12 (Aganezov et al. 2022). Altogether, the T2T consortium has opened previously inaccessible genomic regions, such as centromeric satellite arrays and recent segmental duplications, to variation and functional studies, bringing the routine completion of the entire human genome within reach.

Beyond establishing reference genomes, long reads have been used to probe the genetic landscape of diverse populations and enable the comprehensive characterization of genetic variation through fully reconstructed haplotypes. For example, Seo et al. (2016) used PacBio single-molecule real-time sequencing to construct a contiguous diploid human genome assembly from a Korean individual (AK1), uncovering thousands of previously unreported Asian-specific structural variants and providing high-quality haplotype information of clinically relevant alleles. More recently, in 2021, long reads from PacBio were used to analyze a diverse human panel representing 25 globally diverse populations, uncovering 15.8 million SNVs, 2.3 million indels, and over 107,000 SVs, 42% of which were previously unidentified, particularly in African genomes (Ebert et al. 2021). Following this and related efforts, in 2023 the Human Pangenome Reference Consortium introduced a first draft of the human pangenome

reference, which includes 47 phased, diploid assemblies from a genetically diverse cohort (Liao et al. 2023). These assemblies cover over 99% of the expected sequence in each genome and are more than 99% accurate at both the structural and base pair levels. This pangenome adds 119 million base pairs of euchromatic polymorphic sequences and 1115 gene duplications relative to the existing GRCh38 reference, with ~90 million of these additional base pairs originating from SVs. Furthermore, using the pangenome for analyzing short-read data reduced small variant discovery errors by 34% and increased the detection of SVs per haplotype by 104% compared to GRCh38-based workflows.

Current applications of long-read sequencing technologies are enabling even larger population studies, enhancing our ability to associate genetic variations—particularly those previously undetectable—with human diversity, disease, and other phenotypes. For example, in 2021, the deCODE genetics initiative generated a major large-scale long-read SV callset using Oxford Nanopore sequencing from 3622 Icelanders, identifying three to five times more SVs per sample than short-read data (Beyter et al. 2021). This study also uncovered SVs in strong linkage disequilibrium with disease- or trait-associated variants from the genome-wide association study (GWAS) catalog. This study effectively doubled the number of detected variants compared to short-read sequencing alone. Subsequent efforts have extended to additional diverse populations, such as the resequencing of 1000 Genomes Project samples using long reads that discovered extensive SVs missed in earlier studies (Gustafson et al. 2024; Schloissnig et al. 2024). One of the largest long-read projects to date is the *All of Us* Research Program, which has employed PacBio HiFi reads to sequence over 1000 African American samples in the phase I project (Mahmoud et al. 2024) and is progressing to over 10,000 samples in phase II to be completed by 2025. Meanwhile, the Consortium of Long Read Sequencing (CoLoRS, https://colorsdb.org/) has established an open-resource for population-wide variation cataloging. By aggregating nearly 1400 long-read genomes from various institutions, CoLoRS provides a comprehensive data set with diverse characteristics in read depth, disease focus, trio availability, and ancestry, facilitating the exploration of genetic diversity across different populations and disease states. Additionally, the NIH Center for Alzheimer's and Related Dementias (CARD) has implemented nanopore sequencing technology to analyze brain samples and cell lines (Kolmogorov et al. 2023). CARD intends to expand its analysis to thousands of samples, aiming to yield unprecedented insights into the genetic foundations of Alzheimer's and related dementias.

Benchmarking standards for germline analysis

In parallel to population-wide sequencing efforts, the Genome in a Bottle Consortium (GIAB), the Platinum Genomes Project (Eberle et al. 2017), and related efforts, have developed benchmarks for germline variant calling using a small number of extensively characterized normal cell lines. The first GIAB benchmark from 2014 included ~77% of autosomal bases in GRCh37 (Zook et al. 2014). However, this excluded challenging regions and variants, including medically relevant regions (e.g., one large study found that one in seven pathogenic variants was challenging to detect with short-reads) (Lincoln et al. 2021). Therefore, GIAB has expanded benchmarks to include increasingly difficult regions and variants, such as SVs (Zook et al. 2020), TRs (English et al. 2024), the Major Histocompatibility Complex (Chin et al. 2020), challenging medically relevant genes including those identified from COSMIC (Wagner et al. 2022), and the X and Y Chromosomes (Wagner et al. 2025). The latest draft benchmarks currently being evaluated by GIAB are based on a T2T assembly of HG002 and include 3.6 million SNVs, 950,000 indels, and 29,000 SVs in 2.74 and 2.76 Gbp of GRCh38 (2.77 and 2.82 Gbp of CHM13) for small variants and SVs, respectively (https://github.com/marbl/HG002). Long-read sequencing has demonstrated superior performance in SV detection compared to Illumina-based approaches (Fig. 2A; Kolmogorov et al. 2023). Nevertheless, the T2T assembly of CHM13v2.0 with X and Y Chromosomes is >3.1 Gbp, so substantial regions are still excluded from variant benchmarks even with near-perfect assemblies due to challenges in aligning, representing, and comparing large complex SVs around segmental duplications and satellite repeats.

Related studies such as the PrecisionFDA Challenge (Olson et al. 2022) helped show how long reads improved variant call accuracy in regions difficult to map with short-reads, though short-reads still had advantages in certain contexts like homopolymers. This has led to a virtuous cycle where benchmarks drive advancements in sequencing technologies and bioinformatics methods, and method improvements, in turn, enable expanded benchmarks. In parallel, genome stratifications help determine regions where variants are more likely identified through long-read approaches, such as 12 genes from the COSMIC Cancer Gene Census are entirely located within a segmental duplication on GRCh38 (Wagner et al. 2022; Dwarshuis et al. 2024).

Germline variation analysis in cancer patients

Germline variants influence cancer susceptibility by modifying the regulation or function of essential genes and pathways. Every human carries millions of variations in their germline compared to the reference genome, including hundreds to thousands of rare and common variants that affect disease and cancer risks (MacArthur et al. 2012). Consequently, there is interest to use long-read sequencing to explore SNVs, indels, SVs, repeat expansions, and other complex variations that were previously missed or incorrectly identified in an effort to identify new risk factors for disease. These efforts also help establish the landscape of germline variation to aid in the interpretation of somatic variation, as we discuss below.

The impact of long-read sequencing on variant detection and clinical interpretation

In one of the first studies using long-read sequencing of patient germline genomes, Aganezov et al. (2020) used multiple short- and long-read sequencing technologies to analyze the germline genomes from breast cancer patients. The study demonstrated that long-read sequencing allows for substantially more accurate and sensitive SV detection, achieving 90%–95% concordance between the PacBio and ONT long-read platforms, including in both the germline and cancer genomes. SVs detected exclusively with short-reads that occur near different types of long-read variants or within TR regions were often found to be miscalls from short-reads. Notably, the researchers identified hundreds of variants in two cancer patients' germline genomes within known cancer-related genes detectable only through long-read sequencing, including an intronic SV in *BRCA1* that was not detectable using short-reads. Other germline variations in *BRCA1* can increase the lifetime risk of developing breast cancers to more than 85% although it remains unknown if these specific variations carry disease risk. Nevertheless, these and related findings highlight the

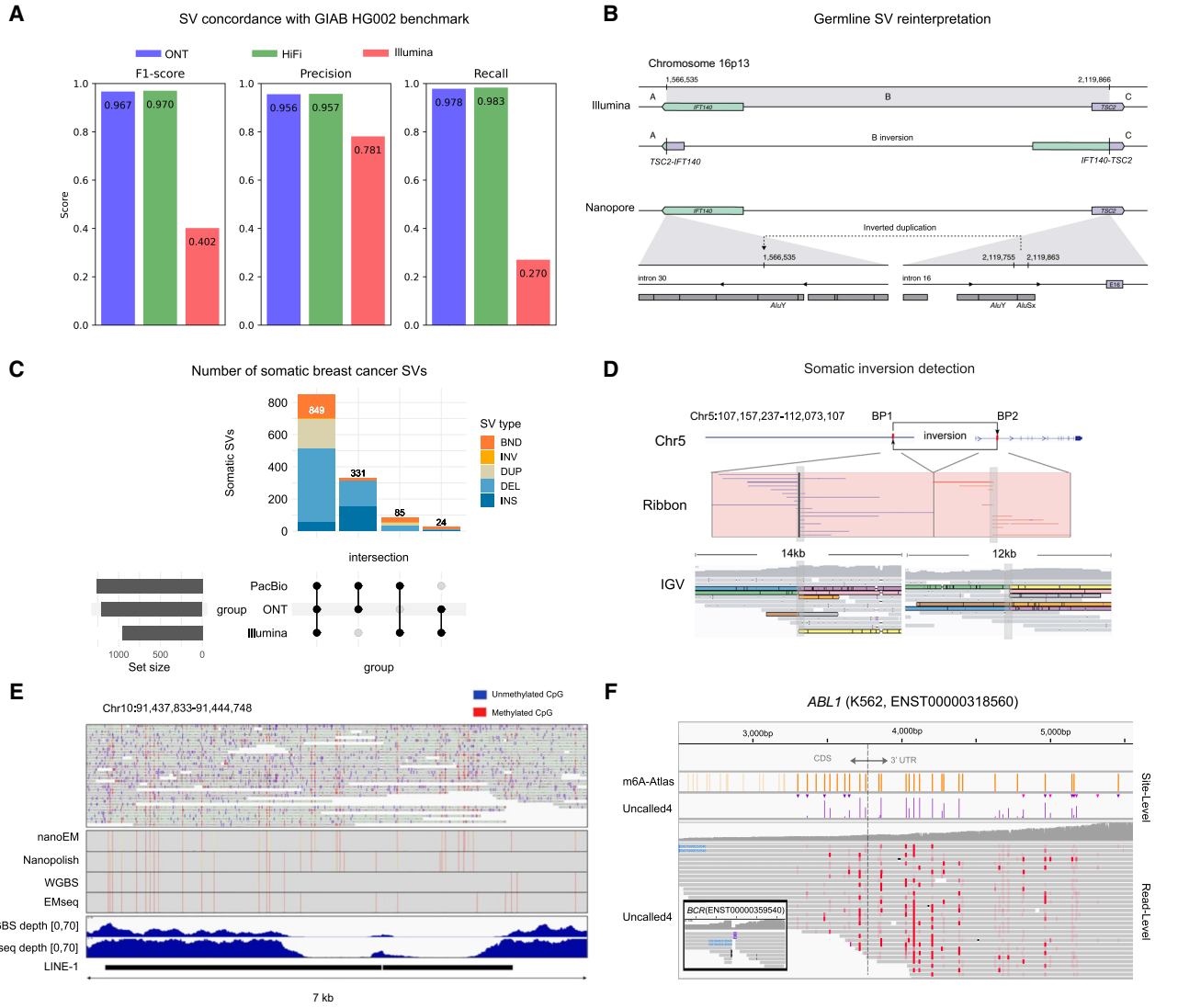

Figure 2. Comparative analysis of structural variation and epigenetics using short- and long-read sequencing. (*A*) Performance comparison of structural variant detection across sequencing technologies, evaluated against the GIAB Tier1 V0.6 benchmark (Kolmogorov et al. 2023). (*B*) Reinterpretation of a germline recurrent event using long-read sequencing. (*Upper* panel) Illumina data indicate a long-range inversion on Chromosome 16p13 with breakpoints in *IFT140* and *TSC2*. (*Lower* panel) ONT sequencing reveals the actual structure: an insertion in intron 30 of *IFT140* (Thibodeau et al. 2020). (*C*) UpSet plot of somatic structural variants identified in the breast cancer cell line using three sequencing technologies (Keskus et al. 2024). (DEL) Deletion, (BND) breakend junction, (DUP) duplication, (INV) inversion, (INS) insertion. (*D*) 4.9 Mbp somatic inversion detected by nanopore sequencing in colorectal cancer, encompassing exon 1 of the *APC* gene (Xu et al. 2023). (*E*) Methylation patterns of LINE-1 located in *HECTD2* intron. (*Upper* panel) Reads from breast cancer cell lines aligned to the region. (*Lower* panel) Read coverage of EMseq and WGBS for the same region (Sakamoto et al. 2021a). (*F*) m6A profiling of an *ABL1* transcript and *BCR-ABL1* fusion in chronic myeloid leukemia (CML) (Kovaka et al. 2024). (C–F, adapted from Thibodeau et al. 2020, Sakamoto et al. 2021a, Xu et al. 2023, and Kovaka et al. 2024).

strong potential of long-read sequencing in cancer genomics for accurately assessing genetic instability.

Long-read sequencing also facilitates the detailed characterization of diverse variants of cancer susceptibility genes, including private, recurrent, and founder variants. For example, a breast cancer study using nanopore sequencing precisely profiled fourteen variants, ranging from single-exon alterations to whole-gene changes (Dixon et al. 2023). They confirmed a 6126 bp tandem duplication in three samples, and reported a shared 1.08 Mb haplotype block by analyzing SNVs beyond the *BRCA1* founder variant boundaries. The accurate identification and phasing of SV breakpoints uncovered previously unrecognized allelic heterogeneity in key genes such as *BRCA1* and *CHEK2* and elucidated the formation mechanisms of recurrent deletions. To achieve more cost-effective identification of disease-causing variations, Nakamura et al. (2024) developed a computational workflow for target adaptive sampling long-read sequencing (see below for a discussion of adaptive sampling) and applied it to 33 suspected hereditary cancer patients. They identified 14 putative pathogenic SNVs and indels, uncovered newly identified SINE-R/VNTR/*Alu* elements affecting the *APC* gene in two familial adenomatous polyposis patients, and demonstrated the utility of off-target reads from adaptive sampling for SNP genotyping, enabling polygenic risk score calculations.

The application of long-read sequencing further reduces false-positive calls and enhances clinical interpretation. A study using nanopore sequencing reassessed SVs initially identified by Illumina in 669 cancer patients, confirming eight pathogenic or likely pathogenic SVs and resolving three additional variants that were ambiguous with short-read data (Thibodeau et al. 2020). For example, a recurrent inversion on Chromosome 16p13, initially misclassified as pathogenic, was revealed to be an inverted duplication of an *Alu* element, leading to its reclassification as likely benign (Fig. 2B). Similarly, Ban et al. (2024) combined long-read sequencing with transcript analysis to identify a novel intragenic duplication in the *PALB2* gene in a patient with triple-negative breast cancer. This *PALB2* duplication was reclassified as pathogenic, suggesting a potential causal link between this genetic alteration and the aggressive phenotype observed in the patient. Further investigation of the tandem duplication mechanism revealed microhomology regions at both proximal and distal breakpoints within *Alu* elements, which were codirectionally aligned with the transcriptional direction of *PALB2*, suggesting the role of repetitive elements in facilitating the duplication event. These advancements demonstrate the potential of long-read sequencing in resolving ambiguous variants and correcting variant classification, thereby improving the cancer screening accuracy.

Moreover, family-based studies using long-read sequencing have advanced our understanding of inherited variations in cancer. Kramer et al. (2024) performed long-read whole-genome sequencing with ONT PromethION on three families with early onset cancer probands, including two colorectal cancer trios and a quad with two siblings affected by testicular cancer, all with unaffected parents. They identified SNVs, SVs, and epigenetic profiles for each individual and applied a family-based filtering strategy to prioritize candidate variants. This approach effectively eliminated nonrelevant variants shared with unaffected parents, facilitating the identification of candidate pathogenic variants, including de novo and compound heterozygous variants (Kramer et al. 2024). In another investigation involving a family with two siblings affected by congenital atypical teratoid rhabdoid tumor, PacBio HiFi sequencing determined the position of an insertion within intron 2 of *SMARCB1*, and identified it to be a SINE–VNTR–*Alu* (SVA) retrotransposon element present in a mosaic state in the mother (Sabatella et al. 2021). A broader analysis of 120 previously unsolved families severely affected by breast, ovarian, pancreatic, or metastatic prostate cancer identified rare deep intronic variants in eight families (6%), where variants in *BRCA1*, *PALB2*, and *ATM* created intronic pseudoexons that were spliced into transcripts, resulting in premature truncations (Gulsuner et al. 2024). Overall, the elucidation of the inheritance patterns of previously hidden variations promises to enhance genetic counseling and surveillance strategies for affected families.

Advances in variant bioinformatics tools: from detection to interpretation

SNVs and indels are common genetic variants contributing to genetic diversity and disease susceptibility. While short-read methods effectively detect these variants in "high-confidence" regions, they struggle to resolve variants in complex or repetitive genomic regions (Kosugi and Terao 2024). To address these gaps, several tools using long reads have been developed. DeepVariant pioneered small variant calling using convolutional neural networks (Poplin et al. 2018). ClairS implements a deep learning model that uses summaries of adjacent genomic positions around candidate sites (Luo et al. 2020). Later, NanoCaller was developed to improve variant calling by incorporating long-range haplotype information and generating features from distant heterozygous SNPs (Ahsan et al. 2021). PEPPER-Margin-DeepVariant, a haplotype-aware genotyping pipeline, has demonstrated superior performance in identifying small variants from nanopore and PacBio HiFi data, particularly in segmental duplications and low-mappability regions where short-read methods often fail (Shafin et al. 2021). Beyond SNVs and indels, TR detection promotes the further understanding of cancer mechanisms. Straglr performs genome-wide screening for TR expansions by identifying insertions composed of repeat elements and genotyping the expanded loci (Readman et al. 2021). TRcaller enables the characterization of TR alleles from both short- and long-read sequences, achieving high accuracy and sensitivity (Wang et al. 2023).

Multiple studies have shown that long-read sequencing detects two to three times more germline SVs than short-read methods, especially for insertions where long reads capture up to 80% more variants than short-reads (Chaisson et al. 2019; Aganezov et al. 2020; Schloissnig et al. 2024). Many of these newly identified SVs are located in highly repetitive regions associated with various diseases, such as repeat expansions in medically important genes and structural variations affecting key pharmacogenes (Gustafson et al. 2024). Therefore, long-read sequencing has become instrumental in comprehensive SV detection, with algorithms broadly categorized into alignment-based and assembly-based methods. Alignment-based tools, such as Sniffles2 (Smolka et al. 2024), cuteSV (Jiang et al. 2020), and SVIM (Heller and Vingron 2019), analyze reads mapped to a reference genome, identify SV signatures, and aggregate similar signatures to produce consensus SV calls. While computationally faster than assembly-based methods, they struggle with diverse SV signatures for complex variants as well as sequencing and alignment errors. Recent advancements have achieved more precise SV profiling, particularly for large inversions and translocations in cancer genomes (Smolka et al. 2024).

Assembly-based methods reconstruct genomic sequences into longer contigs before comparing them with a reference genome. Assembly-based methods are further divided into de novo and reference-guided local assembly approaches. De novo assembly tools like phased assembly variant (PAV) (Ebert et al. 2021) and SVIM-asm (Heller and Vingron 2021) excel at detecting large-scale alterations and novel insertions. But these approaches can miss heterozygous variants if they are not resolved in the assembly and require computationally intensive genome-wide assembly before variant identification. Furthermore, assembly-based methods are often unable to capture mosaic mutations that are often prevalent in cancer tissues as they may have insufficient coverage to find these mutations. Comparatively, reference-guided local assembly methods provide effective SV identification and mitigate issues such as assembly collapse around segmental duplications. For example, DeBreak employs a local de novo assembly approach to cluster SVs and reconstruct long insertions, and it offers a specialized "tumor" mode for analyzing cancer genomes (Chen et al. 2023a).

The increasing cost-effectiveness of long-read sequencing has enabled large-scale cancer genomics studies, facilitating the construction of cohort-level SV callsets to classify rare and common variants potentially implicated in cancer predisposition or progression. Relatedly, despite major advancements in SV detection methods for individual samples, comprehensively profiling all SV types in a sample often requires integrating results from

multiple bioinformatics algorithms. To address the challenges, several specialized SV merging tools have been developed to improve the analysis of complex genomic structures at both individual and population levels. Early methods integrate structural variants based on their coordinates and user-defined parameters such as size and type (Jeffares et al. 2017). Recently developed tools have further improved the SV merging performance by considering both sequence similarity and genomic position (English et al. 2022; Kirsche et al. 2023; Zheng et al. 2024b). These approaches aim to maintain high genotype accuracy and optimize computational efficiency at both individual and population scales. With the established cohort-level variant callset/pangenome, variants in an individual of interest can be characterized through genotyping, i.e., determining an individual's genotype for each known variant in the data set (Chen et al. 2019; Ebler et al. 2022; Grytten et al. 2023). Notably, genotyping can be performed within both short-read and long-read data sets, allowing researchers to tap into the large numbers of patient genomes sequenced with short-reads, albeit with reduced performance in repetitive sequences. Nevertheless, the genotyping process not only enhances the sensitivity and specificity of variant calling but also serves as a critical step in population genetics, quantitative trait locus mapping, and genome-wide association studies.

The related problem of the interpretation of variants and their impact on phenotypes and diseases has gained increasing attention in recent years. For small variants, several tools were developed to assess pathogenicity, especially for missense variants and small in-frame indels. The Combined Annotation Dependent Depletion (CADD) method offers a standardized, genome-wide scoring system (C-score) by integrating sequence conservation and functional annotations (Kircher et al. 2014). FATHMM-MKL improves predictions of functional consequences for both coding and noncoding variants by optimally weighting diverse genomic features (Shihab et al. 2015). To streamline the annotation process, PhD-SNPg, a lightweight, sequence-based machine learning tool, enables efficient analysis of SNPs and indels across coding and noncoding regions (Capriotti and Fariselli 2023). With advancements in SV detection, tools like SVScore (Ganel et al. 2017) and AnnotSV (Geoffroy et al. 2018), integrate multiple genomic features to annotate SVs and identify potentially deleterious ones. Later, Kumar et al. (2020) developed SVFX, a machine learning-based workflow that quantifies the pathogenicity of both somatic and germline SVs. Within cancer cohorts, the predicted pathogenic SVs were enriched in known cancer genes and pathways (Kumar et al. 2020). CADD-SV enhances the SV effect prediction by integrating diverse variant annotations and employing random forest models, producing scores that correlate to known pathogenic and rare variants (Kleinert and Kircher 2022).

Relatedly, the availability of population-level SV data sets has enabled more accurate stratification of variants as common or rare in the general population, thereby improving SV annotation performance (Nicholas et al. 2022). This has led to the development of methods for analyzing rare variations. For example, a recent advancement, Watershed-SV, employs a probabilistic model integrating diverse omic signals and novel SV-related features to characterize the functional effects of rare SVs on expression (Jensen et al. 2025). When applied to the Undiagnosed Diseases Network patient data set, Watershed-SV identified more disease-relevant rare SV candidates than CADD-SV, offering valuable insights into gene disruption mechanisms. These evolving tools promote a more comprehensive and accurate evaluation of variant effects. While these tools help to identify candidate variants, functional studies are still needed before these findings can directly impact diagnostic and therapeutic strategies in clinical genomics.

Identification, analysis, and benchmarking of somatic variations

Accurate and complete detection of somatic variation is essential for cancer genomics studies aiming to identify and catalog driver mutations (Cortés-Ciriano et al. 2021), study tumor evolution (Gerstung et al. 2020), stratify cancers into clinical subtypes (Chakravarty and Solit 2021), and inform patient treatment (Sosinsky et al. 2024). Somatic variation in cancer is a result of various mutational processes (Carvalho and Lupski 2016) that could be specific to cancer types or environmental exposures (Alexandrov et al. 2020). In addition to gradually accumulating somatic mutations, cancer genomes are often characterized by genomic instability (Drews et al. 2022) and rearranged genomes (Li et al. 2020), which facilitate rapid tumor evolution (Stephens et al. 2011). As long-read sequencing continues to advance cancer genomics, it promises to unveil more complex genetic mechanisms underlying cancer predisposition and to deepen our understanding of the interplay between germline variants and somatic mutations across various cancer types.

Early studies overcoming challenges of long-read sequencing of tumors

Tumor clonality, imperfect purity (i.e., normal cells in tumor sample), tumor in normal (TIN) contamination, and loss of heterozygosity events complicate the interpretation of sequencing data, and generally require deeper sequencing to capture low-abundance variants. Extracting high molecular weight (HMW) DNA from tumor tissues (rather than cell lines or blood) is often challenging due to lower input material volumes. Either fresh or high-quality fresh frozen samples are currently required to produce long-read DNA libraries, which is a limitation as most archival tumor samples are formalin-fixed paraffin-embedded (FFPE). To distinguish between germline and somatic variants, tumor samples are often complemented by matching normal samples (The ICGA/TCGA Pan-Cancer Analysis of Whole Genomes Consortium 2020). This is highly effective in identifying somatic variants as those not observed in the paired normal samples, although it suffers from increased sequencing costs. An alternative strategy is to use panels of normals that represent the variation commonly observed in a population; such panels are produced by the population studies highlighted above. This strategy is effective for filtering common variations, although it can lead to misclassification of rare variants and variants in challenging regions that are not captured in the panel of normals.

Initial studies applied long-read sequencing to cancer cell lines and organoids. One early study from 2016 (Norris et al. 2016) used ONT sequencing to detect well-characterized SVs, including large deletions, inversions, and translocations in pancreatic cell lines. Another study from 2018 (Nattestad et al. 2018) performed PacBio sequencing to detect germline and somatic SVs in the SKBR3 breast cancer cell line, revealing highly complex rearrangements around the *ERBB2* oncogene and extensive rearrangements and other SVs genomewide. In a related study from 2020 (Aganezov et al. 2020), the authors showed high consistency and specificity of PacBio and ONT technologies for SV discovery in breast cancer patients-derived organoids.

More recent studies demonstrated the feasibility of producing long-read sequencing from various clinical tumor samples. For example, Sakamoto et al. (2020b) characterized regions of clustered

copy number changes, inversions, and deletions in 20 lung adenocarcinoma patients. This team further used long-read phasing to illustrate the uneven distribution of somatic mutations between haplotypes (Sakamoto et al. 2022). In another study, Fujimoto et al. (2021) analyzed mechanisms of somatic insertions in 11 clinical liver samples with matching normals. Recently, ONT sequencing was applied to 189 advanced tumors from the Personalized OncoGenomics (POG) program, demonstrating the potential of long-read sequencing as a scalable precision oncology tool (O'Neill et al. 2024).

Advantages of long-read sequencing for somatic variation detection

Compared to germline SVs, a higher proportion of somatic SVs are accessible to short-read sequencing because somatic SV breakpoints are less likely to coincide with repetitive elements (Carvalho and Lupski 2016). Indeed, a recent study highlighted that certain classes of somatic SVs (such as large copy number alterations) can be reliably detected with short-reads without the need for long reads (Choo et al. 2023). However, their analysis focused exclusively on copy number alternations larger than 10 kb, which represent only a small fraction of SVs (roughly 10% or less), limiting the applicability of their comparison between short- and long-read technologies. Other studies (Keskus et al. 2024) demonstrated that short-read methods have "blind spots" and miss certain SV types, such as clustered SVs or insertions (Fig. 2C; Mahmoud et al. 2019). In addition, short-read somatic SV calls do not always achieve sequence-level resolution of breakpoints and insertions and typically produce more false-positive calls (Keskus et al. 2024), which might impact the interpretation (Mahmoud et al. 2019). For example, Xu et al. (2023) identified a 4.9 Mbp inversion in a colorectal cancer sample (C546-T) that encompassed exon 1 of the *APC* gene (Fig. 2D). RNA-seq analysis revealed a substantial decrease in *APC* expression in the tumor sample (tumor fragments per kilobase of transcript per million fragments mapped (FPKM): 0.296 vs. normal FPKM: 2.262). The SV was not detected by short-read sequencing, likely because the base sequence of the inverted exon remained unchanged. Since most published cancer genomics studies are based on short-reads, our current understanding of the landscape of somatic SVs in cancer is likely to be incomplete.

Long-read sequencing has enabled the analysis of difficult-to-map repetitive DNA, which can play an important role in cancer progression (Erwin et al. 2023). For example, centromeres and telomeres play a key role in maintaining genome stability and are involved in cancer mutational processes such as breakage-fusion-bridge (BFB) amplification (Gisselsson et al. 2000). A recent study characterized various classes of rearrangements involving telomeric repeats using long-read sequencing, including neotelomeres and chromosome arm fusions (Tan et al. 2024). New targeted long-read-based technologies are also being developed to profile changes in telomere lengths (Schmidt et al. 2024).

Some cancers could be driven by viral integrations or LINE-1 retrotranspositions that are difficult to fully resolve using short-reads and, therefore, will benefit from long-read sequencing (Rodriguez-Martin et al. 2020). For example, a recent study used long-read sequencing to discover reciprocal chromosomal translocations mediated by L1 activity (Zumalave et al. 2024). Another work profiled integrations of human papillomavirus (HPV) in 16 cervical cancers using ONT sequencing, revealing different types of integrations and SVs around the integration sites (Zhou et al. 2022). Other recent applications of long-read sequencing to HPV-infected cancers revealed complex HPV-host concatemers that exist as extrachromosomal amplicons and drive complex rearrangements and intratumoral heterogeneity (Akagi et al. 2023; Rossi et al. 2023).

Long-read connectivity can also be used to better characterize the structure and mechanisms of complex rearrangement and amplification processes, such as chromothripsis, extrachromosomal DNA (ecDNA), or BFB. For example, one germline study used nanopore sequencing to identify disease-related gene losses resulting from chromothripsis in a Langer–Giedion syndrome (Lei et al. 2020). Using ONT sequencing of a childhood medulloblastoma, Rausch et al. (2023) discovered a new SV type termed templated insertion thread and assembled a chromosome copy affected by chromothripsis. Ng et al. (2024) used long-read de novo assembly to resolve complex BFB and ecDNA amplicons in nine cases of esophageal adenocarcinoma. Rodriguez et al. (2024) analyzed 19 cervical cancer cell lines to reveal recurrent BFB events that amplify *YAP1-BIRC3-BIRC2* genes, associated with 10-year-earlier age of diagnosis and three times more common in African American women.

Finally, liquid biopsy is a promising alternative to traditional tissue biopsies in cancer management, offering a minimally invasive method for analyzing circulating tumor cells, cell-free DNA (cfDNA), and other biomarkers in bodily fluids (Ignatiadis et al. 2021; Zhu et al. 2023). Recent advancements in long-read sequencing technologies have further enhanced liquid biopsy capabilities, allowing for real-time analysis of cfDNA genomic and fragmentomic signatures and the detection of previously unobserved longer cfDNA fragments in cancer patients (van der Pol et al. 2023).

New algorithms and tools for long-read somatic variation analysis

Analysis of cancer genomes presents additional computational challenges over germline analysis as most of the existing tools for long reads were not optimized to work with matching normal samples, variable tumor clonality, allelic imbalance, loss of heterozygosity, and complex somatic SV patterns. The recently developed deep learning-based method ClairS (Zheng et al. 2023) supports matching normal samples and takes advantage of long-read phasing to improve the detection of variants with low allelic fractions; DeepSomatic (Park et al. 2024) is another recently developed tool that was trained using a panel of multiple cancer cell lines with different mutational patterns (Table 1). For somatic SV analysis, GASOLINE jointly extracts SV signatures from tumor and normal samples to distinguish somatic from germline variants (Magi et al. 2023a). nanomonsv (Shiraishi et al. 2023) uses a local assembly approach to distinguish somatic and germline variants and introduced a single breakend mode to represent SVs that are partially mappable (Table 1). SAVANA (Elrick et al. 2024) employs a machine learning-based classifier as an additional filter for somatic variant calls to improve precision. Sniffles2 (Smolka et al. 2024) implemented a new function for multisample joint SV genotyping that can be used to separate germline and somatic variants. The recently developed Severus (Keskus et al. 2024) algorithm incorporates long-read phasing to produce haplotype-resolved SV calls and uses a breakpoint graph approach to detect complex rearrangements that consist of many clustered SVs (Table 1).

Long-range phasing of somatic and germline variants is another advantage of long-read technologies, facilitating the discovery of biallelic variants (O'Neill et al. 2024) and improving the accuracy of

Table 1. Summary of software tools for analyzing long reads in cancer

Category	Tool name	Description	Reference
Alignment	NGMLR	Convex gap-cost scoring model to achieve long-read alignment	Sedlazeck et al. 2018
	minimap2	Pairwise alignment method for long reads and large genomes	Li 2018
	Winnowmap	Improvements in aligning long reads in repetitive regions	Jain et al. 2020
Small variant detection	ClairS	Deep learning method designed for detecting somatic small variants, primarily for ONT long-read data	Zheng et al. 2023
	DeepSomatic	Discovery of somatic small variants across multiple sequencing platforms	Park et al. 2024
SV calling	nanomonsv	Identification of somatic SVs at single-nucleotide resolution	Shiraishi et al. 2023
	SAVANA	Somatic SV and copy number aberrations caller for long reads	Elrick et al. 2024
	SVision-pro	Neural network framework enabling somatic structural variant discovery and genotyping	Wang et al. 2024
	Sniffles2	Updated version of Sniffles; capable of identifying mosaic and population-level SVs using long reads	Smolka et al. 2024
	Severus	Improved detection and characterization of somatic SVs in tumor genomes	Keskus et al. 2024
Variant phasing	WhatsHap	Phasing of SNVs and smaller indels	Martin et al. 2023
	LongPhase	Fast chromosome-scale phasing for both small and large variations	Lin et al. 2022
Methylation identification	DeepSignal	Detection of DNA methylation states from ONT long reads	Ni et al. 2019
	Uncalled4	Toolkit for ONT signal alignment, analysis, and visualization, improving DNA and RNA modification detection	Kovaka et al. 2024
	ccsmeth	DNA 5mCpGs caller for PacBio CCS data	Ni et al. 2023
Methylation phasing	NanoMethPhase	Phasing of 5mCpGs from ONT sequencing data	Akbari et al. 2021
	MethPhaser	Utilization of ONT methylation signals to extend SNV-based phasing	Fu et al. 2024
	ccsmethphase	Nextflow pipeline for haplotype-aware methylation detection using PacBio CCS reads	Ni et al. 2023

the somatic/germline variant classification (Simpson 2024). For example, a study on nonsmall cell lung cancer generated 834 kb long-phased blocks and used the phased information to uncover regions with uneven mutation distribution between haplotypes (Sakamoto et al. 2022). A more recent study of long-read sequencing for retinoblastoma demonstrated improved sensitivity to somatic SVs, particularly insertions, relative to short-reads, as well as highlighting the ability of long reads to phase variants and show both copies of the gene are impacted (Zheng et al. 2024a). In addition, long-read phasing and methylation information can be used to distinguish the tumor clones and haplotypes (Ermini and Driguez 2024; Fu et al. 2024; Simpson 2024). A few methods for multiallelic phasing exist, such as WhatsHap for polyploid phasing (Schrinner et al. 2020) or Strainy for bacterial strain phasing (Table 1; Kazantseva et al. 2024). However, specialized methods for phasing multiple cancer clones are yet to be developed.

De novo assembly, showcased by many recent germline long-read studies, could also reveal somatic variation in cryptic and difficult-to-map regions of tumor genomes (Garg 2023; Ijaz et al. 2024). It is, however, complicated by tumor heterogeneity, allelic imbalance, and long near-perfect duplications. Some of these challenges can be addressed by the local assembly of tumor-specific regions (Le et al. 2024). Alternatively, high-quality assemblies of matching normal genomes (Xiao et al. 2022) or improved human reference assemblies (Paulin et al. 2025) can enhance the detection of somatic variants. Related approaches that analyze genome graphs can also resolve ecDNA amplicons, which may consist of repeated chromosomal fragments (Giurgiu et al. 2024; Zhu et al. 2024).

Additional long-range technologies for use in cancer genomics

To further elucidate large, repetitive, and complex genomic regions, such as those surrounding the *MYC* gene—frequently rearranged and amplified on a megabase scale in cancer—long-range technologies have been integrated. Hi-C, a genome-wide chromosome conformation capture technique, identifies large-scale rearrangements by capturing the spatial proximity of genomic regions, and has been widely used for variant phasing and genome assembly (Liao et al. 2023). A recent study used a combination of long-read and Hi-C sequencing to reveal chromosome-scale structure of germline-rearranged genomes (Schöpflin et al. 2022). In another study, chromothripsis-affected chromosomes of esophageal adenocarcinomas were haplotype-resolved and assembled using Hi-C (Ijaz et al. 2024). Further, Hi-C can be used to capture large structural rearrangements and reconstruct complete cancer karyotypes (Brunette et al. 2024), as well as validate the rearranged structure of ecDNA amplicons (Helmsauer et al. 2020). Furthermore, combination of long reads and Pore-C sequencing, which is a Hi-C-like assay but uses long-read sequencing to potentially capture several physically localized segments of DNA simultaneously, generated near-complete T2T assemblies of the human genome, resolving remaining difficult regions and providing a precise, highly continuous framework for structural genomic studies (Koren et al. 2024). Pore-C is particularly useful in studying chromothripsis; long reads allow for better resolution of these complex events by covering repetitive regions and capturing structural variations without amplification (Ulahannan et al. 2019). Combining long reads and Pore-C is particularly useful in ecDNA detection as

well as detecting translocations without the need for breakpoint-spanning reads (Hickey et al. 2024).

Another long-range technology is optical genome mapping (OGM), such as Bionano (global change) Genomics's technology. It visualizes sequence motifs on DNA molecules longer than 100 kb, to enable the detection of copy number alterations and structural variations, including chromothripsis and large inter and intrachromosomal translocations. By using fluorescent labels and restriction enzymes, OGM produces long-range genomic maps and sensitively reveals structural variants >500 bp even at low allele frequencies. While OGM does not resolve the specific sequence of a genome, it can be faster and more cost-effective than other methods (karyotyping, copy number variation (CNV) microarrays, polymerase chain reaction (PCR), and/or next-generation sequencing (NGS)) for detecting SVs, with turnaround time in 3–5 days (Levy et al. 2023). Comparison of OGM in 10 different solid tumor types showed it can be used to detect complex SVs including inversions and translocations genome-wide, requiring only as low as 6 µg of input tissue samples (Goldrich et al. 2021). A more specific study comparing optical maps of breast cancer cell line SKBR3 with long-read whole-genome sequencing (LR-WGS) found that 74% of insertions and 80% of deletions detected with Bionano can be confirmed with PacBio and ONT, but lower concordance for inversions and duplications (Savara et al. 2021). Similarly, OGM detected 77.1% of large variants not identified by WGS in cases of lung squamous cell carcinoma, including multiple variants private to primary tumor tissue (Peng et al. 2020).

Beyond its use with genome assembly or resequencing, a unique capability of nanopore sequencing is "adaptive sampling" (or "ReadUntil" sequencing) where the decision to sequence a molecule of DNA or RNA can be determined in real time. This operates by sequencing the first few hundred nucleotides of a molecule (~1 sec worth of sequencing), and ejecting molecules not of interest. This enables purely computational targeting sequencing from native DNA, enabling the detection of genetic and epigenetic changes, demonstrated in an initial study of 148 genes associated with hereditary cancer (Kovaka et al. 2021).

New benchmarking standards for cancer genomics

Multiple recent studies have produced long-read sequencing of tumor cell lines and normal cell lines from the same individual with accompanying somatic variant benchmarks (Craig et al. 2016; Fang et al. 2021; Espejo Valle-Inclan et al. 2022; Talsania et al. 2022; Keskus et al. 2024; Liu et al. 2024; Masood et al. 2024); however, the long-read data are typically from older, noisier chemistries, and the amount of high-quality data and curated variant calls is still scarce compared to long-read germline projects. In addition, tumor cell lines can have unstable genomes and/or changes in clonality (Craig et al. 2016; Paulin et al. 2025), requiring characterization of large batches of cells to enable robust benchmarking, similar to the National Institute of Standards and Technology's (NIST's) human genome reference materials for normal cell lines (Zook et al. 2016). Another challenge is a lack of explicit consent to publicly share genomic data for most cancer cell lines.

To address these limitations, the Genome in a Bottle consortium has published extensive short- and long-read sequencing data of a pancreatic tumor cell line and paired normal tissues from a patient that was explicitly consented for public sharing of genomic data and cell lines (McDaniel et al. 2024). Benchmark somatic variants for these samples and additional benchmarks for a diverse set of broadly consented cancer and normal cell line pairs with different somatic mutation signatures are in development to enable the community to develop, optimize, and demonstrate the performance of new sequencing and analysis methods.

Multiomics with long reads

Transcriptomics with long reads

Transcriptome analysis facilitates the interpretation of gene expression within dynamic cancer and normal cells at a molecular level, which is crucial for advancing precision oncology (Supplitt et al. 2021). While short-read RNA sequencing has been widely employed in transcriptomic research, it struggles to capture full-length transcripts and complex transcriptional events, particularly alternative splicing (AS) and gene fusions (Byrne et al. 2019). Long-read sequencing technologies overcome these limitations by enabling the sequencing of entire transcripts and revealing the exact sequence and structure of fusion transcripts, thereby providing deeper insights into transcriptome isoform diversity. PacBio uses its SMRT sequencing technology for full-length cDNA molecules, capturing entire transcript isoforms with high accuracy (Logsdon et al. 2020). Comparatively, ONT can sequence both native RNA and full-length cDNA, allowing for the detection of complete transcript structures (Soneson et al. 2019). For example, within non-cancer donor tissues, a recent study analyzed a large human long-read RNA-seq data set using the ONT platform from 88 samples from Genotype-Tissue Expression (GTEx) tissues and cell lines, complementing the short-read GTEx resource (Glinos et al. 2022). Through the long-read analysis, the researchers identified over 70,000 novel transcripts for annotated genes, and validated the protein expression of 10% of novel transcripts.

Characterizing transcriptomic diversity and fusion genes with long reads

Within cancer genomics, long-read transcriptome sequencing has uncovered a large range of RNA diversity inaccessible to short-reads. In 2021, the first full-length gastric cancer (GC) transcriptome database was built, covering the four major GC molecular subtypes and identifying 60,239 nonredundant transcripts—over 66% of which were novel (Huang et al. 2021). These novel isoforms, often expressed at higher levels and with greater variability, provide additional prognostic insights. Another colorectal cancer study integrated short- and long-read RNA-seq data, revealing over 50,000 (>60%) unannotated transcripts and identifying thousands of prognostically significant AS events (Sun et al. 2023), suggesting the necessity of long-read sequencing in capturing transcriptome complexity in tumors.

The accurate detection of fusion genes is necessary due to their prevalence in tumor tissues and their role as driver mutations in various cancers. Taking advantage of long-read sequencing, researchers have developed computational tools such as FusionSeeker (Chen et al. 2023b) and CTAT-LR-fusion (Qin et al. 2025) to promote comprehensive characterization of gene fusions. For example, a study used full-length transcript information to identify a novel three-gene fusion, *BMPR2-TYW5-ALS2CR11*, in a lung cancer cell line, demonstrating the capacity of long-read sequencing to identify complex fusion transcripts beyond mere breakpoint discovery (Davidson et al. 2022). Recently, an analysis of PacBio full-length RNA isoform sequencing data uncovered 23 known and 99 novel fusions in sarcoma samples, including

ASPSCR1-TFE3 fusion, a known marker of sarcomas (Volden et al. 2023).

Specific spliced isoforms play a crucial role in cancer progression, metastasis, and drug resistance, with certain AS events strongly correlated with patient survival (Stricker et al. 2017). In 2022, Veiga et al. (2022) developed a long-read RNA sequencing and annotation platform to predict the functional consequences of spliced isoforms in cancer. Their comprehensive analysis of breast cancer and normal breast samples identified 142,514 isoforms in breast tumors, 66% of which are novel, with many affecting protein-coding exons and potentially altering protein function. In addition, they identified 3059 tumor-specific splicing events, including 35 associated with patient survival and 10 enriched in specific breast cancer subtypes. The findings demonstrate the complexity and clinical significance of these isoforms and splicing events, providing a valuable resource for potential immuno-oncology therapeutic targets.

For more effective application of long-read RNA sequencing in clinical settings, several methods for transcriptome-based identification of disease variation are available. For example, the Capture and Ultradeep Long-Read RNA Sequencing (CAPLRseq) workflow evaluates the impact of various genomic alterations on mRNA structural integrity and expression, including coding and intronic SNVs, as well as structural variants like retrotransposon insertions and large genomic rearrangements (Schwenk et al. 2023). The workflow has been used for diagnosing hereditary nonpolyposis colorectal cancer (HNPCC), also known as Lynch syndrome. In the analysis of 123 hereditary cancer-related transcripts, CAPLRseq reclassified two variants of uncertain significance (VUS) in the *MSH6* and *PMS2* genes as likely pathogenic or benign and confirmed 17 cases of HNPCC/Lynch syndrome by identifying splicing defects and allele-specific expression loss in mismatch repair genes. This highlights the promise of long-read RNA sequencing to improve genetic diagnoses and the interpretation of complex hereditary cancer syndromes.

Toward long-read single-cell transcriptomics in cancer

Beyond traditional bulk RNA sequencing, long-read single-cell RNA sequencing (LR scRNA-seq) has become increasingly relevant in oncology, providing a more detailed view of the cancer landscape at single-cell resolution. A 2023 ovarian cancer study performed LR scRNA-seq on clinical samples from three patients, producing the deepest PacBio scRNA-seq data set to date (Dondi et al. 2023). They identified 152,000 isoforms, with one-third being novel, and uncovered many cell-type-specific isoforms. Notably, isoform-level analysis that accounted for noncoding isoforms demonstrated that protein-coding gene expression had been overestimated by an average of 20%, suggesting the necessity for isoform-specific quantification. Leveraging the advantages of long-read technology, the study provided evidence suggesting that cancer cells may induce epithelial–mesenchymal transition in tumor microenvironment mesothelial cells, and identified gene fusions, such as *IGF2BP2::TESPA1*, previously misclassified as high *TESPA1* expression in short-read data.

Furthermore, genetic and transcriptomic variations could cause cancer clonal heterogeneity and impact treatment outcomes. Linking genetic to transcriptomic variations is crucial for unraveling the mechanisms underlying treatment resistance in cancer (Vasan et al. 2019). LR scRNA-seq enables simultaneous detection of both genetic and transcriptomic variations. Leveraging this potential, LongSom, a computational workflow for LR scRNA-seq data, was developed to detect de novo somatic SNVs, copy number alterations, and gene fusions, further reconstructing the tumor clonal heterogeneity (Dondi et al. 2025). Application of LongSom to ovarian cancer samples identified clinically relevant somatic SNVs that were missed by short-read scRNA-seq. By integrating somatic SNVs and fusions, LongSom distinguished subclones with varying predicted treatment responses. The work demonstrates LR scRNA-seq is effective for exploring cancer evolution, clonal heterogeneity, and treatment outcome prediction.

DNA and RNA methylation analysis

Long-read methylation profiling in cancer

Abnormal DNA methylation, a hallmark of cancer development, typically manifests as genome-wide hypomethylation and site-specific CpG island promoter hypermethylation (Sonar et al. 2024). In cancer cells, hypermethylation drives malignant transformation through silencing tumor suppressor genes and disrupting essential cellular processes, such as cell–cell adhesion and apoptotic pathways. Meanwhile, global hypomethylation leads to genomic instability, activation of transposable elements, and potential oncogene upregulation (Magi et al. 2023b). Therefore, DNA methylation patterns serve as valuable biomarkers for cancer diagnosis, prognosis assessment, and therapeutic decision-making (Jones et al. 2016). Long-read sequencing technologies have enhanced epigenetic profiling by reducing GC bias, identifying CpG islands at lower read depths, and enabling direct examination of native DNA modifications (Ermini and Driguez 2024). ONT detects DNA modifications by measuring changes in the electrical signal as native DNA passes through the nanopore, while PacBio uses real-time kinetic analysis of DNA polymerase activity to infer modifications (Xu and Seki 2020). Importantly, long reads enable methylation phasing alongside genetic variations, revealing ASM patterns associated with complex diseases and cancers, which are challenging with short-read data (Gigante et al. 2019).

Recent studies have demonstrated the power of long-read sequencing in characterizing cancer methylation patterns. Using the ONT PromethION platform, Ewing et al. (2020) accurately assessed transposable element methylation in hepatocellular carcinoma (HCC) tissues, detecting pronounced demethylation of LINE-1 retrotransposons in cancer. Later, nanoEM successfully analyzed two tandem LINE-1 elements within *HECTD2* introns that traditional short-read methods like EMseq and whole-genome bisulfite sequencing (WGBS) failed to fully capture (Fig. 2E; Sakamoto et al. 2021a). Analysis of the Personalized OncoGenomics data set revealed that long-range phasing enables identification of allelically differentially methylated regions (aDMRs) in cancer genes, such as *RET* and *CDKN2A*. This study also observed *MLH1* promoter hypermethylation in Lynch syndrome, and showed promoter methylation in *BRCA1* and *RAD51C* as potential drivers of homologous recombination deficiency in cancers lacking known mutations (O'Neill et al. 2024). A medulloblastoma study using long-read sequencing revealed ASM effects, complex rearrangements with differential methylation, and distinct methylation patterns in cancer driver genes (Rausch et al. 2023). Besides, ONT-based methylation profiling enables rapid central nervous system (CNS) tumor subtype classification (Patel et al. 2022), which could be performed using real-time sequencing during surgery (Vermeulen et al. 2023). Similarly, PacBio sequencing, integrated with advanced computational methods, has shown superior performance in detecting distinctive methylation patterns with greater accuracy than short-

read technologies, offering enhanced diagnostic power in cancer studies (Choy et al. 2022; Ni et al. 2023)

Assessing cfDNA methylation in plasma holds great promise for the early, noninvasive cancer detection (Li and Zhou 2020). Traditional bisulfite or enzymatic conversion methods introduce biases, which are further complicated by limited cfDNA yields from plasma samples, making comprehensive methylome characterization challenging. Long-read sequencing technology enables direct, unbiased methylation identification, advancing liquid biopsy applications (Lau et al. 2023). Notably, the longer reads provide more CpG sites, enriching methylation information in plasma DNA. These extended methylation patterns serve as more complete "molecular barcodes" compared to shorter fragments, improving tissue-of-origin analysis specificity and supporting the development of crucial noninvasive cancer biomarkers (Choy et al. 2022). In 2022, an innovative methylation analysis used nanopore sequencing and detected cell-of-origin and cancer-specific methylation, suggesting a liquid biopsy approach would be possible with long reads (Katsman et al. 2022). Later, Lau et al. (2023) developed a single-molecule sequencing strategy to analyze cfDNA methylomes, and observed distinct methylation patterns between cancer patients and healthy individuals. By comparing the methylomes of matched tumors and immune cells, they characterized cfDNA methylation in cancer patients for longitudinal monitoring throughout treatment. Finally, an HCC study using PacBio sequencing discovered previously unreported long cfDNA fragments (>1 kb) in cancer patients. They designed the "HCC Methylation Score," a metric reflecting single-molecule methylation patterns associated with cancer, and found that long cfDNA improved discriminatory power than short cfDNA (Choy et al. 2022).

Long-read sequencing technologies have also advanced the study of RNA modifications, particularly N^6-methyladenosine (m6A) patterns. m6A influences multiple aspects of RNA metabolism, such as translation, degradation, splicing, and export, playing crucial roles in tumor regulation and development (Wang et al. 2022). ONT detects RNA modifications by sequencing native RNA molecules directly and capturing the changes in electrical signals, while PacBio indirectly infers modifications from cDNA sequencing (Xu and Seki 2020). Using native RNA sequencing, Jenjaroenpun et al. (2021) identified cell-type-specific m6A patterns in lung cancer cell line, discovering distinct modifications absent in reference cells. More recently, a liver cancer study identified 3968 potential m6A modification sites across 1396 genes and revealed these m6A-modified genes involved in multiple key cellular processes (Arzumanian et al. 2023). A recent long-read study of clear cell renal cell carcinoma (ccRCC) using nanopore identified 9644 genes with altered m6A modifications, with 5224 genes showing concurrent changes in both m6A modification and RNA expression (Li et al. 2024). Notably, this analysis found four obesity-associated genes with prognostic significance in ccRCC. These studies demonstrate the potential of long-read sequencing to detect RNA base modifications, providing valuable insights into the cancer epitranscriptome and the exploration of novel therapeutic targets.

Methylation detection, analysis, and visualization using long reads

Recent advances in computational methods have improved methylation detection from long-read sequencing data. The newest ONT basecallers enable direct detection of 5mC and 5hmC from single DNA reads (Liu et al. 2021). Advanced tools employ pre-trained models, like hidden Markov models and deep neural networks, to translate electrical current signals into specific bases (https://github.com/nanoporetech/megalodon, Ni et al. 2019). The recently developed Uncalled4 further improves the performance in detecting DNA and RNA modifications, identifying 26% more m6A modifications than Nanopolish using m6Anet in seven normal and cancerous human cell lines, including in cancer-implicated genes like *ABL1* and *JUN* (Table 1; Fig. 2F; Kovaka et al. 2024). Similar to how accurate read alignments are needed for robust detection of sequence variations, Uncalled4's improved signal alignments lead to improved detection of epigenetic changes. Using PacBio sequencing data, Tse et al. (2021) developed a convolutional neural network method named the holistic kinetic model for genome-wide 5mCpG detection. To mitigate high subread depth requirements, PacBio later introduced primrose that can achieve 85% read-level accuracy in 5mCpG detection (Ni et al. 2023). Relatedly, ccsmeth, a deep learning method for PacBio data, demonstrates 90% detection accuracy for DNA 5mCpGs at single-molecule resolution, facilitating precise, genome-wide methylation profiling (Table 1; Ni et al. 2023).

ASM, occurring throughout the human genome with frequent presence at imprinted loci, influences gene regulation and disease susceptibility (Benton et al. 2019). To identify ASM using long reads, NanoMethPhase (Akbari et al. 2021) integrates SNV and methylation signals from nanopore, showing successful ASM detection from about 10× coverage in the COLO829BL B lymphoblast cell line (Table 1). MethPhaser enhances phasing accuracy by connecting phase blocks through heterozygous methylation information, particularly improving resolution across human leukocyte antigen (HLA) and other medically relevant genes (Table 1; Fu et al. 2024). The ccsmethphase Nextflow pipeline, designed for PacBio data, enables accurate genome-wide ASM identification, even in repetitive regions (Table 1; Ni et al. 2023). Differential methylation analysis has further expanded our understanding of cancer epigenetics. PoreMeth provides high-resolution detection of DNA methylation alterations between sample pairs for both CpG islands and sparse CpG regions (Magi et al. 2023b). Its application in acute myeloid leukemia (AML) revealed drug resistance phenotypes are driven by selective epigenetic alterations in transcription factor regions. xPore enables differential analysis from direct RNA sequencing data. It identified between 800 and 2000 differentially modified sites per cancer cell line when comparing five cancer cell lines against HEK293T-KO cells, with most sites conforming to the m6A DRACH motif (Pratanwanich et al. 2021).

Efficient visualization tools promote the analysis and interpretation of methylation patterns. NanoMethViz provides a comprehensive solution for methylation data visualization by processing outputs from diverse methylation callers and implementing efficient data compression for large-scale data sets (Su et al. 2021). Its multiresolution visualization capabilities allow researchers to explore methylation patterns from broad genomic regions down to single-read resolution at specific loci of interest. Recent standardization efforts by the Global Alliance for Genomics and Health (GA4GH) have introduced two new tags (MM and ML) to the SAM/BAM file specification. Using this standardized format of files, modbamtools enables visualization, manipulation, and comparative analysis of base modifications with robust performance (Razaghi et al. 2022). The tool produces publication-quality visualizations and supports downstream analyses of methylation patterns. Additionally, WashU Epigenome Browser has added the modbed track, which displays modification details at single-

read and aggregated levels across multiple resolutions. Users can access these visualizations by uploading modbed files locally or through URLs on the WashU platform (Li et al. 2023a,b).

Chromatin accessibility and conformation

Histone modifications influence chromatin structure and gene regulation in carcinogenesis (Yang et al. 2022). For example, alterations in histone-modifying enzymes, such as *SETD2* mutations in renal cell carcinoma, lead to dysregulated gene expression (Yu et al. 2023). Chromatin conformation capture technologies, such as Hi-C or Pore-C were originally developed to study the 3D organization of the genome (Belton et al. 2012). Somatic SVs can change the spatial organization of the genome, which may be specific to different cancer types (Dubois et al. 2022). For identifying mechanisms of cancer development, Hi-C helped identify *TP53* acting as a regulator of chromatin structure (Serra et al. 2024) as well as how *TP53* loss causes whole-genome doubling that impacts chromatin segregation (Lambuta et al. 2023). A related technique, scNanoHi-C (Li et al. 2023b), is a single-cell long-read concatemer sequencing method to reveal high-order chromatin structures within individual cells. Their results suggest that extensive high-order chromatin structures exist in active chromatin regions across the genome, and multiway interactions between enhancers and their target promoters were systematically identified within individual cells.

A related technique called Fiber-seq leverages the ability of long-read sequencing to measure chromatin accessibility (Stergachis et al. 2020). This process operates by "stenciling" the structure of individual chromatin fibers onto their composite DNA templates using nonspecific DNA N^6-adenine methyltransferases. Single-molecule long-read sequencing of chromatin stencils then enables nucleotide-resolution readout of the primary architecture of multikilobase chromatin fibers. Fiber-seq exposed widespread plasticity in the linear organization of individual chromatin fibers and illuminated principles guiding regulatory DNA actuation, the coordinated actuation of neighboring regulatory elements, single-molecule nucleosome positioning, and single-molecule transcription factor occupancy.

Future opportunities and challenges in long-read cancer genomics

Scaling to large samples from diverse populations

The rapidly evolving landscape of cancer genomics and precision oncology necessitates accurate and comprehensive variation profiling across large, diverse populations. Long-read sequencing has emerged as an indispensable tool for cancer studies, but it requires suitable HMW DNA extraction protocols and size selection methods for optimal performance (Ermini and Driguez 2024). However, HMW DNA molecules are often challenging to obtain from clinical tissues compared to the input for short-read reactions. A comparative study demonstrated that cryopreserved tumor samples provide superior DNA quality, integrity, and quantity compared to FFPE tissues (Okojie et al. 2024). High-quality material benefits long-read sequencing, particularly for detecting large SVs such as chromosomal translocations, inversions, and copy number variations, common in genomically unstable tumors. The prevalent use of FFPE tissue may compromise the identification of these critical variations, underscoring the need to optimize sample selection, preservation, and DNA extraction methods (Haile et al. 2019). Relatedly, high-depth sequencing is needed to detect low allele fraction variants, especially when considering tumor heterogeneity and mixtures of tumor and normal cells (Talsania et al. 2022). It is also more costly, and thus guidelines for the depth of sequencing for clinical sequencing need to be established (Amarasinghe et al. 2020). Besides, long-read sequencing technologies face limitations in accurately identifying certain classes of small variants, especially indels, in homopolymeric and low-complexity regions (Nyaga et al. 2024). Although ONT's new R10 pore provides better resolution in homopolymer regions, systematic biases persist, particularly with certain *k*-mers varying in the distinctness of the signals they produce (Amarasinghe et al. 2020). Similarly, PacBio reads, despite their high overall accuracy, exhibit bias in homopolymer regions (Fig. 3; Amarasinghe et al. 2020). Addressing these challenges will require continued advancement in both sequencing chemistry and computational algorithms.

Another major consideration is that researchers are increasingly recognizing the importance of including diverse populations in oncology clinical trials, aiming to uncover critical disparities in cancer incidence, prevalence, and treatment outcomes (Fig. 3; Vidal et al. 2024). For instance, a study on nonsmall-cell lung cancer revealed significant differences in survival rates and *EGFR* mutation frequencies between Asian and non-Asian patients treated with gefitinib (Mok et al. 2018). Moving beyond single-cancer studies, pan-cancer projects have shed light on genomic differences across various cancer types (Fig. 3). A recent analysis of 7152 tumors sequenced with short-reads uncovered metastatic tumors typically exhibit lower intratumor heterogeneity, higher genomic instability, and more frequent SVs compared to primary tumors (Martínez-Jiménez et al. 2023). As long-read sequencing technologies improve and become more cost-effective, their applications in cross-population and pan-cancer studies promise to reduce genetic biases in complex genomic regions and elucidate the shared genetic basis across cancers, thereby revealing the global landscape of human cancer genetics.

Considering current pan-cancer studies mainly use short-read sequencing, we advocate for the utilization of long-read sequencing technologies to uncover previously missed variant characteristics and establish stronger links between variations, cancer driver genes, and tumor progression. Furthermore, multiple studies have verified the potential of T2T-CHM13 reference genome in cancer research for improved variant detection accuracy (Paulin et al. 2025). As sequencing technology advances and becomes more accessible, the possibility of complete personalized T2T genome assemblies, with both parental haplotypes phased from telomere to telomere, emerges as a potential new standard in human genetics (Fig. 3; Miga and Eichler 2023). We expect this effort would pave the way for deeper exploration of variation mechanisms and their impact on human health. Overall, expanding pan-cancer studies to encompass large-scale, high-quality sequencing data from diverse populations promise to enhance knowledge of tumor biology and accelerate the development of personalized, more effective cancer treatments in the future.

Long-read data analysis challenges

Long-read sequencing introduces several unique computational challenges over standard short-read approaches (Fig. 3). Firstly, computational requirements are often higher for base-calling raw data from long-read sequencing, such as with PacBio HiFi sequencing and Oxford Nanopore (Wenger et al. 2019; Baid et al. 2022; Pagès-Gallego and de Ridder 2023). Furthermore, sequencing

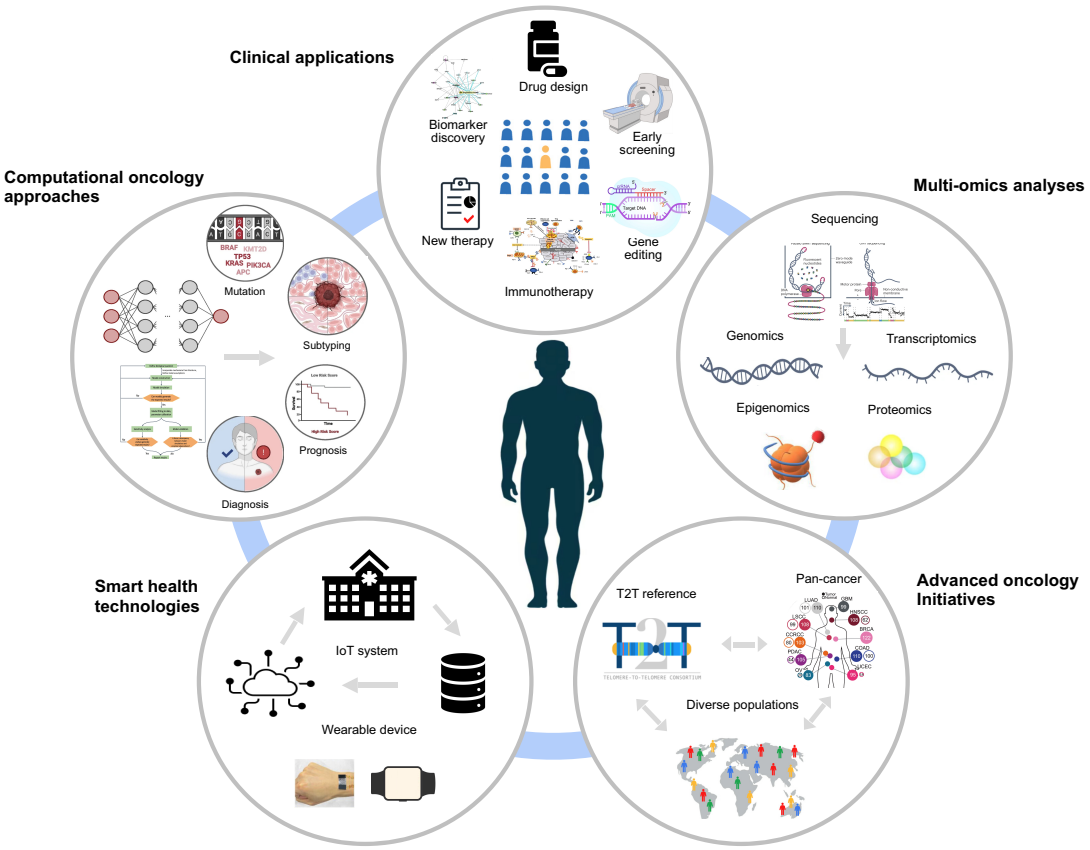

Figure 3. Advanced strategies in precision oncology. (*Top right* panel) Multiomics analyses illustrate the integration of advanced sequencing technologies with various omics approaches to elucidate the molecular foundations of cancer pathogenesis and progression. The sequencing, genomics, and transcriptomics illustrations are from Kovaka et al. (2023), the epigenetics illustration is adapted from Mehrmohamadi et al. (2021), and the proteomics illustration is from Zhang et al. (2021). (*Bottom right* panel) Advanced oncology initiatives highlight emerging trends such as T2T reference genomes and pan-cancer studies across diverse populations. The pan-cancer illustration was adapted from Savage et al. (2024). (*Top left* panel) Computational oncology approaches demonstrate the application of advanced algorithms and artificial intelligence methods in cancer research. Four potential applications and a deep learning model adapted from Unger and Kather (2024), along with a computational model flowchart from Ma and Gurkan-Cavusoglu (2024) are depicted. (*Bottom left* panel) Smart health technologies show the development of applications for efficient data collection, transmission, real-time monitoring, and treatment in cancer care, such as internet of things (IoT) systems and wearable devices. The wearable device illustration on the *left* is modified from Ma et al. (2023). (*Top center* panel) Clinical applications illustrate some key clinical directions in precision oncology. The biomarker discovery illustration is modified from Kawata-Shimamura et al. (2022). The early screening illustration is from Liao et al. (2022), the gene editing illustration from Zhang et al. (2021), and the immunotherapy illustration from Guerrouahen et al. (2020).

chemistries and base-calling algorithms have been evolving rapidly, posing challenges in validating pipelines and requiring extensive frequent benchmarking. Relatedly, long reads can have unique computer hardware requirements. For example, the algorithm advances related to genome sketching (Rowe 2019) allow for performant alignment of long reads given sufficient I/O bandwidth, along with the incorporation of deep learning architectures such as Transformers (Vaswani et al. 2017) for DeepConsensus that heavily use graphics processing units (GPUs) to process reads. The large data can make data security requirements for patient genomic data more challenging, though risk assessment and management exercises (Pulivarti 2023) can yield appropriate analysis and storage infrastructure including the use of cloud platforms that meet NIH Genomic Data Sharing (GDS) policy (https://datascience.cancer.gov/data-commons/cloud-resources, https://sharing.nih.gov/genomic-data-sharing-policy).

SVs are key drivers of cancer development, underpinning major driver mutations and somatic copy number alterations (Elrick et al. 2024). Copy number alterations (CNAs), which include large deletions and amplifications spanning genes or entire chromosomes, influence cancer progression by modulating oncogene and tumor suppressor gene expression (Muñoz-Barrera et al. 2022). While long-read sequencing has enhanced our ability to detect these variations, the effective representation and comparison of complex SVs and CNAs require new standards and tools. The distinction between SVs and CNAs often blurs due to the lack of universal size thresholds for classification. In general, SVs tend to have precise breakpoints, whereas CNAs have fuzzy boundaries due to the coverage-based detection methods. In some cases, the breakpoints of a CNA may be caused by a translocation, which is generally called an SV, but there are no standards for how to represent this relationship. New tools are needed to compare somatic SV and CNA callsets both for benchmarking and cross-individual comparisons. While more robust comparison tools for sequence-resolved SVs have recently been developed for sequence-resolved germline variants (English et al. 2022; Dunn et al. 2024), these tools do not generally accommodate somatic SVs and CNAs, which often are much larger and sometimes more complex with many breakpoints.

Beyond these considerations, data interpretation and visualization is a critical component of analyzing the complex somatic structural variation that can be discovered by long reads in cancer genomes. An array of existing tools enable data overview and summarization, sense-making, and hypothesis generation with more improvements continually being developed. Considering the canonical visualization approach of overview first, zoom and filter, then details on demand (Shneiderman 1996), Genome Ribbon (Nattestad et al. 2021), and PGR-TK (Chin et al. 2023) provide a mix of chromosome-scale views paired with read-level or gene data.

Application of deep learning in cancer research

Genomic alterations drive the development and progression of cancer, reshaping the genetic landscape as the disease advances. To address challenges in cancer genomics, such as sequencing artifacts, tumor-normal cross-contamination, and low-frequency somatic mutations, several artificial intelligence (AI) models have been developed for long reads to improve base-calling accuracy (Amarasinghe et al. 2020; Pagès-Gallego and de Ridder 2023) and enhance the variation detection (Fig. 3; Zheng et al. 2023; Park et al. 2024; Wang et al. 2024). As long-read sequencing enables more accurate identification of structural variations, distinguishing driver from passenger SVs has become crucial in cancer genome analysis. Murakami et al. (2024) developed an explainable artificial intelligence (XAI) framework that integrates knowledge graphs and deep tensors with large language models to predict pathogenicity and elucidate mechanisms of SVs involving fusion genes. Beyond SV identification and interpretation, accurate tumor characterization benefits to diagnosis, prognosis, and treatment selection (Singh et al. 2021). The Mutation-Attention (MuAt) deep neural network was developed to learn representations of diverse genetic alterations, integrating SNVs/multiple nucleotide variants (MNVs), indels, and SV breakpoints through multimodal data embeddings (Sanjaya et al. 2023). This approach enables precise identification of histological tumor types and specific entities, such as *SHH*-activated medulloblastoma and *SPOP*-associated prostate cancer, advancing the potential for precision oncology. To understand the complex nature of tumor pathogenesis, researchers have developed multiomics data fusion algorithms (Chaudhary et al. 2018; Picard et al. 2021) that integrate diverse molecular profiles, providing a more detailed view of cancer development. A recent Tumor MultiOmics pretrained Network (TMO-Net) model (Wang et al. 2024), integrates multiomics data from 32 cancer types, incorporating genomic, transcriptomic, proteomic, and metabolomic information. TMO-Net demonstrates enhanced performance in various oncology tasks, including gene mutation prediction, cancer subtype classification, drug response forecasting, and prognosis prediction. Beyond omics applications, AI has been widely adopted to advance cancer research in various fields, including biomedical image analysis, novel drug candidate identification, and the prediction of drug–drug interactions (Li et al. 2023a; Han and Tao 2024). These ongoing efforts have accelerated the drug discovery process and opened new avenues for personalized cancer treatments.

Translating long-read research into the clinic

Long-read sequencing holds immense potential for transforming cancer diagnostics and treatment. However, transitioning this technology from a powerful research tool to a routine clinical application presents several challenges, alongside promising opportunities.

One of the primary obstacles to the clinical adoption of long-read sequencing is data accessibility, especially the cost, throughput, and automation of the technology. Although technologies like Oxford Nanopore, PacBio, and potentially other long-read platforms (Zhang et al. 2024) offer unmatched resolution of structural variants and complex genomic regions, they remain more expensive compared to short-read sequencing. For widespread clinical use, it is essential to reduce these costs and increase throughput to make long-read sequencing a viable option for routine diagnostics. This is particularly important for somatic mutations because high coverage is required to detect low-frequency mutations. Relatedly, the clinical environment demands consistent and reproducible results, necessitating highly automated and standardized workflows. Presently, long-read sequencing involves multiple manual steps—from DNA extraction to data analysis—that are rapidly improving, introducing variability, and extending turnaround times. To facilitate clinical adoption, the development of fully automated sequencing workflows, including robust and stable software pipelines, is crucial (Fig. 3). These systems must be capable of efficiently processing and interpreting the vast data sets generated by long-read sequencing. Ensuring that the technology meets clinical standards for accuracy, reproducibility, and patient safety is critical.

After the experimental requirements for long-read sequencing, the largest remaining challenge is in the interpretation of the variations found. Long-read sequencing frequently uncovers a multitude of variants, including many that are of unknown significance (VUS). This presents a challenge for clinicians in making informed decisions based on these results. Developing comprehensive databases and catalogs of variants best resolved by long reads, supported by initiatives like *All of Us*, ColorsDB, and the 1000 Genomes Project, as well as advanced interpretation algorithms like CADD-SV and Watershed-SV, are essential to help clinicians accurately interpret these findings. Additionally, integrating data from these initiatives can enhance our understanding of variants across diverse populations.

Despite these challenges, we are optimistic for the growing adoption of long reads for cancer research and clinical care. Long-read sequencing excels at detecting complex genomic, transcriptomic, and epigenomic variants, including structural variants, gene fusions, and TRs, which are often missed by short-read technologies. This capability opens new avenues for identifying clinically relevant variants that could serve as biomarkers for cancer diagnosis, prognosis, and treatment selection (Fig. 3). For example, known fusions that are difficult to detect with short-reads can be more easily identified with long reads, offering crucial information for targeted therapies.

One promising opportunity to increase the accessibility of long reads is the development of targeted long-read sequencing panels focusing on specific genes or genomic regions known to be implicated in certain cancers (Iyer et al. 2024). These panels could be particularly valuable for genes located in segmental duplications or other challenging regions poorly resolved by short-reads. Such targeted panels would provide actionable insights, facilitating personalized medicine, and enabling the routine clinical use of long-read sequencing. Another opportunity is to design and conduct meaningful studies with smaller cohorts, such as family-based studies. These studies can provide valuable insights into inherited cancer susceptibility and the role of rare variants in disease. In particular, trio or quad family structures help phase variants to trace cancer predisposition

alleles within families, while facilitating the filtration of background variations unlikely to contribute to cancer risk.

As more success stories emerge from programs like Genomics England Cancer 2.0 (https://www.genomicsengland.co.uk/initiatives/cancer), the Personalized OncoGenomics program (O'Neill et al. 2024), and the programs at Children's Mercy Hospital, where long-read sequencing has led to significant clinical breakthroughs, the technology will gain greater acceptance in the clinical community. Demonstrating the real-world impact of long-read sequencing on patient outcomes will be key to driving its adoption in routine clinical practice. While there are remaining challenges to overcome, the opportunities presented by long-read sequencing in the clinical setting are substantial. By addressing cost, automation, sample requirements, and variant interpretation, and by building on successful case studies, long-read sequencing has the potential to revolutionize cancer diagnostics and personalized medicine.

Competing interest statement

F.J.S. has received research support from Pacific Biosciences and Oxford Nanopore Technologies.

Acknowledgments

We thank W. Richard McCombie, Sara Goodwin, Kim Doheny, and the teams at both PacBio and Oxford Nanopore. This work was supported, in part, by National Institutes of Health (NIH) awards U01CA253481, U24CA284167, U24HG010263, and OT2OD034190 (to M.C.S.) as well as the support of the Lustgarten Foundation 90101412 (to A.P.K.) Certain commercial equipment, instruments, or materials are identified to specify adequately experimental conditions or reported results. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the equipment, instruments, or materials identified are necessarily the best available for the purpose. M.K. and A.P.K. were supported by the Intramural Research Program of the NIH.

References

Aganezov S, Goodwin S, Sherman RM, Sedlazeck FJ, Arun G, Bhatia S, Lee I, Kirsche M, Wappel R, Kramer M, et al. 2020. Comprehensive analysis of structural variants in breast cancer genomes using single-molecule sequencing. *Genome Res* **30:** 1258–1273. doi:10.1101/gr.260497.119

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

Ahsan MU, Liu Q, Fang L, Wang K. 2021. NanoCaller for accurate detection of SNPs and indels in difficult-to-map regions from long-read sequencing by haplotype-aware deep neural networks. *Genome Biol* **22:** 261. doi:10.1186/s13059-021-02472-2

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

Akbari V, Garant J-M, O'Neill K, Pandoh P, Moore R, Marra MA, Hirst M, Jones SJM. 2021. Megabase-scale methylation phasing using nanopore long reads and NanoMethPhase. *Genome Biol* **22:** 68. doi:10.1186/s13059-021-02283-5

Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, Boot A, Covington KR, Gordenin DA, Bergstrom EN, et al. 2020. The repertoire of mutational signatures in human cancer. *Nature* **578:** 94–101. doi:10.1038/s41586-020-1943-3

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

Arzumanian VA, Kurbatov IY, Ptitsyn KG, Khmeleva SA, Kurbatov LK, Radko SP, Poverennaya EV. 2023. Identifying N6-methyladenosine sites in HepG2 cell lines using Oxford Nanopore Technology. *Int J Mol Sci* **24:** 16477. doi:10.3390/ijms242216477

Baid G, Cook DE, Shafin K, Yun T, Llinares-López F, Berthet Q, Belyaeva A, Töpfer A, Wenger AM, Rowell WJ, et al. 2022. DeepConsensus improves the accuracy of sequences with a gap-aware sequence transformer. *Nat Biotechnol* **41:** 232–238. doi:10.1038/s41587-022-01435-7

Ban IO, Chabert A, Guignard T, Puechberty J, Cabello-Aguilar S, Pujol P, Vendrell JA, Solassol J. 2024. Characterizing *PALB2* intragenic duplication breakpoints in a triple-negative breast cancer case using long-read sequencing. *Front Oncol* **14:** 1355715. doi:10.3389/fonc.2024.1355715

Belton J-M, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. 2012. Hi–C: a comprehensive technique to capture the conformation of genomes. *Methods* **58:** 268–276. doi:10.1016/j.ymeth.2012.05.001

Benton MC, Lea RA, Macartney-Coxson D, Sutherland HG, White N, Kennedy D, Mengersen K, Haupt LM, Griffiths LR. 2019. Genome-wide allele-specific methylation is enriched at gene regulatory regions in a multi-generation pedigree from the Norfolk Island isolate. *Epigenetics Chromatin* **12:** 60. doi:10.1186/s13072-019-0304-7

Beyter D, Ingimundardottir H, Oddsson A, Eggertsson HP, Bjornsson E, Jonsson H, Atlason BA, Kristmundsdottir S, Mehringer S, Hardarson MT, et al. 2021. Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nat Genet* **53:** 779–786. doi:10.1038/s41588-021-00865-4

Brunette GJ, Tourdot RW, Zong D, Pellman D, Zhang C-Z. 2024. Haplotype-resolved karyotype construction from Hi-C data using refLinker. bioRxiv doi:10.1101/2024.03.02.583108

Byrne A, Cole C, Volden R, Vollmers C. 2019. Realizing the potential of full-length transcriptome sequencing. *Philos Trans R Soc Lond B Biol Sci* **374:** 20190097. doi:10.1098/rstb.2019.0097

The Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. 2013. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* **45:** 1113–1120. doi:10.1038/ng.2764

Capriotti E, Fariselli P. 2023. PhD-SNPg: updating a webserver and lightweight tool for scoring nucleotide variants. *Nucleic Acids Res* **51:** W451–W458. doi:10.1093/nar/gkad455

Carvalho CMB, Lupski JR. 2016. Mechanisms underlying structural variant formation in genomic disorders. *Nat Rev Genet* **17:** 224–238. doi:10.1038/nrg.2015.25

Chaisson MJP, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, Antonacci F, Surti U, Sandstrom R, Boitano M, et al. 2015. Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517:** 608–611. doi:10.1038/nature13907

Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, Gardner EJ, Rodriguez OL, Guo L, Collins RL, et al. 2019. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat Commun* **10:** 1784. doi:10.1038/s41467-018-08148-z

Chakravarty D, Solit DB. 2021. Clinical cancer genomic profiling. *Nat Rev Genet* **22:** 483–501. doi:10.1038/s41576-021-00338-8

Chaudhary K, Poirion OB, Lu L, Garmire LX. 2018. Deep learning-based multi-omics integration robustly predicts survival in liver cancer. *Clin Cancer Res* **24:** 1248–1259. doi:10.1158/1078-0432.CCR-17-0853

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 Y, Wang AY, Barkley CA, Zhang Y, Zhao X, Gao M, Edmonds MD, Chong Z. 2023a. Deciphering the exact breakpoints of structural variations using long sequencing reads with DeBreak. *Nat Commun* **14:** 283. doi:10.1038/s41467-023-35996-1

Chen Y, Wang Y, Chen W, Tan Z, Song Y, Human Genome Structural Variation Consortium, Chen H, Chong Z. 2023b. Gene fusion detection and characterization in long-read cancer transcriptome sequencing data with FusionSeeker. *Cancer Res* **83:** 28–33. doi:10.1158/0008-5472.CAN-22-1628

Chiang C, Scott AJ, Davis JR, Tsang EK, Li X, Kim Y, Hadzic T, Damani FN, Ganel L, GTEx Consortium, et al. 2017. The impact of structural variation on human gene expression. *Nat Genet* **49:** 692–699. doi:10.1038/ng.3834

Chin C-S, Wagner J, Zeng Q, Garrison E, Garg S, Fungtammasan A, Rautiainen M, Aganezov S, Kirsche M, Zarate S, et al. 2020. A diploid assembly-based benchmark for variants in the major histocompatibility complex. *Nat Commun* **11:** 4794. doi:10.1038/s41467-020-18564-9

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

Choo Z-N, Behr JM, Deshpande A, Hadi K, Yao X, Tian H, Takai K, Zakusilo G, Rosiene J, Da Cruz Paula A, et al. 2023. Most large structural variants

in cancer genomes can be detected without long reads. *Nat Genet* **55:** 2139–2148. doi:10.1038/s41588-023-01540-6
Choy LYL, Peng W, Jiang P, Cheng SH, Yu SCY, Shang H, Olivia Tse OY, Wong J, Wong VWS, Wong GLH, et al. 2022. Single-molecule sequencing enables long cell-free DNA detection and direct methylation analysis for cancer patients. *Clin Chem* **68:** 1151–1163. doi:10.1093/clinchem/hvac086
Cortés-Ciriano I, Gulhan DC, Lee JJ-K, Melloni GEM, Park PJ. 2021. Computational analysis of cancer genome sequencing data. *Nat Rev Genet* **23:** 298–314. doi:10.1038/s41576-021-00431-y
Craig DW, Nasser S, Corbett R, Chan SK, Murray L, Legendre C, Tembe W, Adkins J, Kim N, Wong S, et al. 2016. A somatic reference standard for cancer genome sequencing. *Sci Rep* **6:** 24607. doi:10.1038/srep24607
Damaraju N, Miller AL, Miller DE. 2024. Long-read DNA and RNA sequencing to streamline clinical genetic testing and reduce barriers to comprehensive genetic testing. *J Appl Lab Med* **9:** 138–150. doi:10.1093/jalm/jfad107
Davidson NM, Chen Y, Sadras T, Ryland GL, Blombery P, Ekert PG, Göke J, Oshlack A. 2022. JAFFAL: detecting fusion genes with long-read transcriptome sequencing. *Genome Biol* **23:** 10. doi:10.1186/s13059-021-02588-5
Dixon K, Shen Y, O'Neill K, Mungall KL, Chan S, Bilobram S, Zhang W, Bezeau M, Sharma A, Fok A, et al. 2023. Defining the heterogeneity of unbalanced structural variation underlying breast cancer susceptibility by nanopore genome sequencing. *Eur J Hum Genet* **31:** 602–606. doi:10.1038/s41431-023-01284-1
Dondi A, Lischetti U, Jacob F, Singer F, Borgsmüller N, Coelho R, Tumor Profiler Consortium, Heinzelmann-Schwarz V, Beisel C, Beerenwinkel N. 2023. Detection of isoforms and genomic alterations by high-throughput full-length single-cell RNA sequencing in ovarian cancer. *Nat Commun* **14:** 7780. doi:10.1038/s41467-023-43387-9
Dondi A, Borgsmüller N, Ferreira PF, Haas BJ, Jacob F, Heinzelmann-Schwarz V, Tumor Profiler Consortium, Beerenwinkel N. 2025. De novo detection of somatic variants in high-quality long-read single-cell RNA sequencing data. *Genome Res* (this issue) doi:10.1101/gr.279281.124
Drews RM, Hernando B, Tarabichi M, Haase K, Lesluyes T, Smith PS, Morrill Gavarró L, Couturier D-L, Liu L, Schneider M, et al. 2022. A pan-cancer compendium of chromosomal instability. *Nature* **606:** 976–983. doi:10.1038/s41586-022-04789-9
Dubois F, Sidiropoulos N, Weischenfeldt J, Beroukhim R. 2022. Structural variations in cancer and the 3D genome. *Nat Rev Cancer* **22:** 533–546. doi:10.1038/s41568-022-00488-9
Dunn T, Zook JM, Holt JM, Narayanasamy S. 2024. Jointly benchmarking small and structural variant calls with vcfdist. *Genome Biol* **25:** 253.
Dwarshuis N, Kalra D, McDaniel J, Sanio P, Jerez PA, Jadhav B, Huang W, Mondal R, Busby B, Olson ND, et al. 2024. The GIAB genomic stratifications resource for human reference genomes. *Nat Commun* **15:** 9029.
Eberle MA, Fritzilas E, Krusche P, Källberg M, Moore BL, Bekritsky MA, Iqbal Z, Chuang H-Y, Humphray SJ, Halpern AL, et al. 2017. A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome Res* **27:** 157–164. doi:10.1101/gr.210500.116
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
Elrick H, Sauer CM, Valle-Inclan JE, Trevers K, Tanguy M, Zumalave S, De Noon S, Muyas F, Cascão 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, Menon VK, Gibbs RA, Metcalf GA, Sedlazeck FJ. 2022. Truvari: refined structural variant comparison preserves allelic diversity. *Genome Biol* **23:** 271. doi:10.1186/s13059-022-02840-6
English AC, Dolzhenko E, Ziaei Jam H, McKenzie SK, Olson ND, De Coster W, Park J, Gu B, Wagner J, Eberle MA, et al. 2024. Analysis and benchmarking of small and large genomic variants across tandem repeats. *Nat Biotechnol* doi:10.1038/s41587-024-02225-z
Ermini L, Driguez P. 2024. The application of long-read sequencing to cancer. *Cancers (Basel)* **16:** 1275. doi:10.3390/cancers16071275
Erwin GS, Gürsoy G, Al-Abri R, Suriyaprakash A, Dolzhenko E, Zhu K, Hoerner CR, White SM, Ramirez L, Vadlakonda A, et al. 2023. Recurrent repeat expansions in human cancer genomes. *Nature* **613:** 96–102. doi:10.1038/s41586-022-05515-1
Espejo Valle-Inclan J, Besselink NJM, de Bruijn E, Cameron DL, Ebler J, Kutzera J, van Lieshout S, Marschall T, Nelen M, Priestley P, et al. 2022. A multi-platform reference for somatic structural variation detection. *Cell Genomics* **2:** 100139. doi:10.1016/j.xgen.2022.100139
Ewing AD, Smits N, Sanchez-Luque FJ, Faivre J, Brennan PM, Richardson SR, Cheetham SW, Faulkner GJ. 2020. Nanopore sequencing enables comprehensive transposable element epigenomic profiling. *Mol Cell* **80:** 915–928.e5. doi:10.1016/j.molcel.2020.10.024
Fang LT, Zhu B, Zhao Y, Chen W, Yang Z, Kerrigan L, Langenbach K, de Mars M, Lu C, Idler K, et al. 2021. Establishing community reference samples, data and call sets for benchmarking cancer mutation detection using whole-genome sequencing. *Nat Biotechnol* **39:** 1151–1160. doi:10.1038/s41587-021-00993-6
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
Fujimoto A, Wong JH, Yoshii Y, Akiyama S, Tanaka A, Yagi H, Shigemizu D, Nakagawa H, Mizokami M, Shimada M. 2021. Whole-genome sequencing with long reads reveals complex structure and origin of structural variation in human genetic variations and somatic mutations in cancer. *Genome Med* **13:** 65. doi:10.1186/s13073-021-00883-1
Ganel L, Abel HJ, FinMetSeq Consortium, Hall IM. 2017. SVScore: an impact prediction tool for structural variation. *Bioinformatics* **33:** 1083–1085. doi:10.1093/bioinformatics/btw789
Garg S. 2023. Towards routine chromosome-scale haplotype-resolved reconstruction in cancer genomics. *Nat Commun* **14:** 1358. doi:10.1038/s41467-023-36689-5
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
Gerstung M, Jolly C, Leshchiner I, Dentro SC, Gonzalez S, Rosebrock D, Mitchell TJ, Rubanova Y, Anur P, Yu K, et al. 2020. The evolutionary history of 2,658 cancers. *Nature* **578:** 122–128. doi:10.1038/s41586-019-1907-7
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
Gisselsson D, Pettersson L, Höglund M, Heidenblad M, Gorunova L, Wiegant J, Mertens F, Dal Cin P, Mitelman F, Mandahl N. 2000. Chromosomal breakage-fusion-bridge events cause genetic intratumor heterogeneity. *Proc Natl Acad Sci* **97:** 5357–5362. doi:10.1073/pnas.090013497
Giurgiu M, Wittstruck N, Rodriguez-Fos E, Chamorro Gonzalez R, Brueckner L, Krienelke-Szymansky A, Helmsauer K, Hartebrodt A, Euskirchen P, Koche RP, et al. 2024. Reconstructing extrachromosomal DNA structural heterogeneity from long-read sequencing data using Decoil. *Genome Res* **34:** 1355–1364. doi:10.1101/gr.279123.124
Glinos DA, Garborcauskas G, Hoffman P, Ehsan N, Jiang L, Gokden A, Dai X, Aguet F, Brown KL, Garimella K, et al. 2022. Transcriptome variation in human tissues revealed by long-read sequencing. *Nature* **608:** 353–359. doi:10.1038/s41586-022-05035-y
Goldrich DY, LaBarge B, Chartrand S, Zhang L, Sadowski HB, Zhang Y, Pham K, Way H, Lai C-YJ, Pang AWC, et al. 2021. Identification of somatic structural variants in solid tumors by optical genome mapping. *J Pers Med* **11:** 142. doi:10.3390/jpm11020142
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
Grytten I, Rand KD, Sandve GK. 2023. KAGE 2: Fast and accurate genotyping of structural variation using pangenomes. bioRxiv doi:10.1101/2023.12.23.572333
Guerrouahen BS, Maccalli C, Cugno C, Rutella S, Akporiaye ET. 2020. Reverting immune suppression to enhance cancer immunotherapy. *Front Oncol* **9:** 1554. doi:10.3389/fonc.2019.01554
Gulsuner S, AbuRayyan A, Mandell JB, Lee MK, Bernier GV, Norquist BM, Pierce SB, King M-C, Walsh T. 2024. Long-read DNA and cDNA sequencing identify cancer-predisposing deep intronic variation in tumor-suppressor genes. *Genome Res* **34:** 1825–1831. doi:10.1101/gr.279158.124
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
Haile S, Corbett RD, Bilobram S, Bye MH, Kirk H, Pandoh P, Trinh E, MacLeod T, McDonald H, Bala M, et al. 2019. Sources of erroneous sequences and artifact chimeric reads in next generation sequencing of

genomic DNA from formalin-fixed paraffin-embedded samples. *Nucleic Acids Res* **47:** e12. doi:10.1093/nar/gky1142
Han Y, Tao J. 2024. Revolutionizing pharma: unveiling the AI and LLM trends in the pharmaceutical industry. arXiv doi:10.48550/arXiv.2401.10273
Hanahan D, Weinberg RA. 2011. Hallmarks of cancer: the next generation. *Cell* **144:** 646–674. doi:10.1016/j.cell.2011.02.013
Heller D, Vingron M. 2019. SVIM: structural variant identification using mapped long reads. *Bioinformatics* **35:** 2907–2915. doi:10.1093/bioinformatics/btz041
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
Helmsauer K, Valieva ME, Ali S, Chamorro González R, Schöpflin R, Röefzaad C, Bei Y, Dorado Garcia H, Rodriguez-Fos E, Puiggròs M, et al. 2020. Enhancer hijacking determines extrachromosomal circular *MYCN* amplicon architecture in neuroblastoma. *Nat Commun* **11:** 5823. doi:10.1038/s41467-020-19452-y
Hiatt SM, Lawlor JMJ, Handley LH, Ramaker RC, Rogers BB, Partridge EC, Boston LB, Williams M, Plott CB, Jenkins J, et al. 2021. Long-read genome sequencing for the molecular diagnosis of neurodevelopmental disorders. *HGG Adv* **2:** 100023. doi:10.1016/j.xhgg.2021.100023
Hickey S, Dai X, Aganezov S, Beaulaurier J, Harrington E, Juul S. 2024. Abstract 405: Translocation detection in cancer using low-pass pore-c sequencing. *Cancer Res* **84:** 405. doi:10.1158/1538-7445.AM2024-405
Huang KK, Huang J, Wu JKL, Lee M, Tay ST, Kumar V, Ramnarayanan K, Padmanabhan N, Xu C, Tan ALK, et al. 2021. Long-read transcriptome sequencing reveals abundant promoter diversity in distinct molecular subtypes of gastric cancer. *Genome Biol* **22:** 44. doi:10.1186/s13059-021-02261-x
The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. 2020. Pan-cancer analysis of whole genomes. *Nature* **578:** 82–93. doi:10.1038/s41586-020-1969-6
Ignatiadis M, Sledge GW, Jeffrey SS. 2021. Liquid biopsy enters the clinic - implementation issues and future challenges. *Nat Rev Clin Oncol* **18:** 297–312. doi:10.1038/s41571-020-00457-x
Ijaz J, Harry E, Raine K, Menzies A, Beal K, Quail MA, Zumalave S, Jung H, Coorens THH, Lawson ARJ, et al. 2024. Haplotype-specific assembly of shattered chromosomes in esophageal adenocarcinomas. *Cell Genomics* **4:** 100484. doi:10.1016/j.xgen.2023.100484
Ishiura H, Doi K, Mitsui J, Yoshimura J, Matsukawa MK, Fujiyama A, Toyoshima Y, Kakita A, Takahashi H, Suzuki Y, et al. 2018. Expansions of intronic TTTCA and TTTTA repeats in benign adult familial myoclonic epilepsy. *Nat Genet* **50:** 581–590. doi:10.1038/s41588-018-0067-2
Iyer SV, Goodwin S, McCombie WR. 2024. Leveraging the power of long reads for targeted sequencing. *Genome Res* **34:** 1701–1718. doi:10.1101/gr.279168.124
Jain C, Rhie A, Zhang H, Chu C, Walenz BP, Koren S, Phillippy AM. 2020. Weighted minimizer sampling improves long read mapping. *Bioinformatics* **36:** i111–i118. doi:10.1093/bioinformatics/btaa435
Jeffares DC, Jolly C, Hoti M, Speed D, Shaw L, Rallis C, Balloux F, Dessimoz C, Bähler J, Sedlazeck FJ. 2017. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat Commun* **8:** 14061. doi:10.1038/ncomms14061
Jenjaroenpun P, Wongsurawat T, Wadley TD, Wassenaar TM, Liu J, Dai Q, Wanchai V, Akel NS, Jamshidi-Parsian A, Franco AT, et al. 2021. Decoding the epitranscriptional landscape from native RNA sequences. *Nucleic Acids Res* **49:** e7. doi:10.1093/nar/gkaa620
Jensen TD, Ni B, Reuter CM, Gorzynski JE, Fazal S, Bonner D, Ungar RA, Goddard PC, Raja A, Ashley EA, et al. 2025. Integration of transcriptomics and long-read genomics prioritizes structural variants in rare disease. *Genome Res* (this issue) doi:10.1101/gr.279323.124
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
Jones PA, Issa J-PJ, Baylin S. 2016. Targeting the cancer epigenome for therapy. *Nat Rev Genet* **17:** 630–641. doi:10.1038/nrg.2016.93
Jones S, Hruban RH, Kamiyama M, Borges M, Zhang X, Parsons DW, Lin JC-H, Palmisano E, Brune K, Jaffee EM, et al. 2009. Exomic sequencing identifies *PALB2* as a pancreatic cancer susceptibility gene. *Science* **324:** 217. doi:10.1126/science.1171202
Katsman E, Orlanski S, Martignano F, Fox-Fisher I, Shemer R, Dor Y, Zick A, Eden A, Petrini I, Conticello SG, et al. 2022. Detecting cell-of-origin and cancer-specific methylation features of cell-free DNA from Nanopore sequencing. *Genome Biol* **23:** 1–25. doi:10.1186/s13059-022-02710-1
Kawata-Shimamura Y, Eguchi H, Kawabata-Iwakawa R, Nakahira M, Okazaki Y, Yoda T, Grénman R, Sugasawa M, Nishiyama M. 2022. Biomarker discovery for practice of precision medicine in hypopharyngeal cancer: a theranostic study on response prediction of the key therapeutic agents. *BMC Cancer* **22:** 779. doi:10.1186/s12885-022-09853-1
Kazantseva E, Donmez A, Frolova M, Pop M, Kolmogorov M. 2024. Strainy: phasing and assembly of strain haplotypes from long-read metagenome sequencing. *Nat Methods* **21:** 2034–2043.
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
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
Kolmogorov M, Billingsley KJ, Mastoras M, Meredith M, Monlong J, Lorig-Roach R, Asri M, Alvarez Jerez P, Malik L, Dewan R, et al. 2023. Scalable Nanopore sequencing of human genomes provides a comprehensive view of haplotype-resolved variation and methylation. *Nat Methods* **20:** 1483–1492. doi:10.1038/s41592-023-01993-x
Koren S, Bao Z, Guarracino A, Ou S, Goodwin S, Jenike KM, Lucas J, McNulty B, Park J, Rautiainen M, et al. 2024. Gapless assembly of complete human and plant chromosomes using only nanopore sequencing. *Genome Res* **34:** 1919–1930. doi:10.1101/gr.279334.124
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
Kovaka S, Fan Y, Ni B, Timp W, Schatz MC. 2021. Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED. *Nat Biotechnol* **39:** 431–441. doi:10.1038/s41587-020-0731-9
Kovaka S, Ou S, Jenike KM, Schatz MC. 2023. Approaching complete genomes, transcriptomes and epi-omes with accurate long-read sequencing. *Nat Methods* **20:** 12–16. doi:10.1038/s41592-022-01716-8
Kovaka S, Hook PW, Jenike KM, Shivakumar V, Morina LB, Razaghi R, Timp W, Schatz MC. 2024. Uncalled4 improves nanopore DNA and RNA modification detection via fast and accurate signal alignment. bioRxiv doi:10.1101/2024.03.05.583511
Kramer M, Goodwin S, Wappel R, Borio M, Offit K, Feldman DR, Stadler ZK, McCombie WR. 2024. Exploring the genetic and epigenetic underpinnings of early-onset cancers: variant prioritization for long read whole genome sequencing from family cancer pedigrees. bioRxiv doi:10.1101/2024.06.27.601096
Kumar S, Harmanci A, Vytheeswaran J, Gerstein MB. 2020. SVFX: a machine learning framework to quantify the pathogenicity of structural variants. *Genome Biol* **21:** 274. doi:10.1186/s13059-020-02178-x
Lambuta RA, Nanni L, Liu Y, Diaz-Miyar J, Iyer A, Tavernari D, Katanayeva N, Ciriello G, Oricchio E. 2023. Whole-genome doubling drives oncogenic loss of chromatin segregation. *Nature* **615:** 925–933. doi:10.1038/s41586-023-05794-2
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
Le MK, Qin Q, Li H. 2024. Long-range somatic structural variation calling from matched tumor-normal co-assembly graphs. bioRxiv doi:10.1101/2024.07.29.605160
Leger A, Amaral PP, Pandolfini L, Capitanchik C, Capraro F, Miano V, Migliori V, Toolan-Kerr P, Sideri T, Enright AJ, et al. 2021. RNA modifications detection by comparative Nanopore direct RNA sequencing. *Nat Commun* **12:** 7198. doi:10.1038/s41467-021-27393-3
Lei M, Liang D, Yang Y, Mitsuhashi S, Katoh K, Miyake N, Frith MC, Wu L, Matsumoto N. 2020. Long-read DNA sequencing fully characterized chromothripsis in a patient with Langer–Giedion syndrome and Cornelia de Lange syndrome-4. *J Hum Genet* **65:** 667–674. doi:10.1038/s10038-020-0754-6
Levy B, Baughn LB, Akkari Y, Chartrand S, LaBarge B, Claxton D, Lennon PA, Cujar C, Kolhe R, Kroeger K, et al. 2023. Optical genome mapping in acute myeloid leukemia: a multicenter evaluation. *Blood Adv* **7:** 1297–1307. doi:10.1182/bloodadvances.2022007583
Li C, Wong C, Zhang S, Usuyama N, Liu H, Yang J, Naumann T, Poon H, Gao J. 2023a. LLaVA-Med: training a large language-and-vision assistant for BioMedicine in one day. *Neural Inf Process Syst* abs/2306.00890: 28541–28564.
Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34:** 3094–3100. doi:10.1093/bioinformatics/bty191
Li H, Li C, Zhang Y, Jiang W, Zhang F, Tang X, Sun G, Xu S, Dong X, Shou J, Yang Y, et al. 2024. Comprehensive analysis of m6A methylome and

transcriptome by Nanopore sequencing in clear cell renal carcinoma. *Mol Carcinogenesis* **63:** 677–687.
Li W, Lu J, Lu P, Gao Y, Bai Y, Chen K, Su X, Li M, Liu J, Chen Y, et al. 2023b. scNanoHi-C: a single-cell long-read concatemer sequencing method to reveal high-order chromatin structures within individual cells. *Nat Methods* **20:** 1493–1505. doi:10.1038/s41592-023-01978-w
Li W, Zhou XJ. 2020. Methylation extends the reach of liquid biopsy in cancer detection. *Nat Rev Clin Oncol* **17:** 655–656. doi:10.1038/s41571-020-0420-0
Li Y, Roberts ND, Wala JA, Shapira O, Schumacher SE, Kumar K, Khurana E, Waszak S, Korbel JO, Haber JE, et al. 2020. Patterns of somatic structural variation in human cancer genomes. *Nature* **578:** 112–121. doi:10.1038/s41586-019-1913-9
Liao J, Li X, Gan Y, Han S, Rong P, Wang W, Li W, Zhou L. 2022. Artificial intelligence assists precision medicine in cancer treatment. *Front Oncol* **12:** 998222. doi:10.3389/fonc.2022.998222
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 J-H, Chen L-C, Yu S-C, Huang Y-T. 2022. LongPhase: an ultra-fast chromosome-scale phasing algorithm for small and large variants. *Bioinformatics* **38:** 1816–1822. doi:10.1093/bioinformatics/btac058
Lincoln SE, Hambuch T, Zook JM, Bristow SL, Hatchell K, Truty R, Kennemer M, Shirts BH, Fellowes A, Chowdhury S, et al. 2021. One in seven pathogenic variants can be challenging to detect by NGS: an analysis of 450,000 patients with implications for clinical sensitivity and genetic test implementation. *Genet Med* **23:** 1673–1680. doi:10.1038/s41436-021-01187-w
Liu L, Zhang J, Wood S, Newell F, Leonard C, Koufariotis LT, Nones K, Dalley AJ, Chittoory H, Bashirzadeh F, et al. 2024. Performance of somatic structural variant calling in lung cancer using Oxford Nanopore sequencing technology. *BMC Genomics* **25:** 898. doi:10.1186/s12864-024-10792-3
Liu Y, Rosikiewicz W, Pan Z, Jillette N, Wang P, Taghbalout A, Foox J, Mason C, Carroll M, Cheng A, et al. 2021. DNA methylation-calling tools for Oxford Nanopore sequencing: a survey and human epigenome-wide evaluation. *Genome Biol* **22:** 295. doi:10.1186/s13059-021-02510-z
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
Luo R, Wong C-L, Wong Y-S, Tang C-I, Liu C-M, Leung C-M, Lam T-W. 2020. Exploring the limit of using a deep neural network on pileup data for germline variant calling. *Nat Mach Intell* **2:** 220–227. doi:10.1038/s42256-020-0167-4
Ma C, Gurkan-Cavusoglu E. 2024. A comprehensive review of computational cell cycle models in guiding cancer treatment strategies. *NPJ Syst Biol Appl* **10:** 71. doi:10.1038/s41540-024-00397-7
Ma X, Wu X, Cao S, Zhao Y, Lin Y, Xu Y, Ning X, Kong D. 2023. Stretchable and skin-attachable electronic device for remotely controlled wearable cancer therapy. *Adv Sci* **10:** e2205343. doi:10.1002/advs.202205343
MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, Jostins L, Habegger L, Pickrell JK, Montgomery SB, et al. 2012. A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335:** 823–828. doi:10.1126/science.1215040
Magi A, Mattei G, Mingrino A, Caprioli C, Ronchini C, Frigè G, Semeraro R, Baragli M, Bolognini D, Colombo E, et al. 2023a. GASOLINE: detecting germline and somatic structural variants from long-reads data. *Sci Rep* **13:** 20817. doi:10.1038/s41598-023-48285-0
Magi A, Mattei G, Mingrino A, Caprioli C, Ronchini C, Frigè G, Semeraro R, Bolognini D, Rambaldi A, Candoni A, et al. 2023b. High-resolution Nanopore methylome-maps reveal random hyper-methylation at CpG-poor regions as driver of chemoresistance in leukemias. *Commun Biol* **6:** 382. doi:10.1038/s42003-023-04756-8
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, Huang Y, Garimella K, Audano PA, Wan W, Prasad N, Handsaker RE, Hall S, Pionzio A, Schatz MC, et al. 2024. Utility of long-read sequencing for all of us. *Nat Commun* **15:** 837. doi:10.1038/s41467-024-44804-3
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
McDaniel JH, Patel V, Olson ND, He H-J, He Z, Cole KD, Schmitt A, Sikkink K, Sedlazeck FJ, Doddapaneni H, et al. 2024. Development and extensive sequencing of a broadly-consented Genome in a Bottle matched tumor-normal pair. bioRxiv doi:10.1101/2024.09.18.613544
Martínez-Jiménez F, Movasati A, Brunner SR, Nguyen L, Priestley P, Cuppen E, Van Hoeck A. 2023. Pan-cancer whole-genome comparison of primary and metastatic solid tumours. *Nature* **618:** 333–341. doi:10.1038/s41586-023-06054-z
Marwaha S, Knowles JW, Ashley EA. 2022. A guide for the diagnosis of rare and undiagnosed disease: beyond the exome. *Genome Med* **14:** 23. doi:10.1186/s13073-022-01026-w
Masood D, Ren L, Nguyen C, Brundu FG, Zheng L, Zhao Y, Jaeger E, Li Y, Cha SW, Halpern A, et al. 2024. Evaluation of somatic copy number variation detection by NGS technologies and bioinformatics tools on a hyper-diploid cancer genome. *Genome Biol* **25:** 163. doi:10.1186/s13059-024-03294-8
Mehrmohamadi M, Sepehri MH, Nazer N, Norouzi MR. 2021. A comparative overview of epigenomic profiling methods. *Front Cell Dev Biol* **9:** 714687. doi:10.3389/fcell.2021.714687
Merker JD, Wenger AM, Sneddon T, Grove M, Zappala Z, Fresard L, Waggott D, Utiramerur S, Hou Y, Smith KS, et al. 2018. Long-read genome sequencing identifies causal structural variation in a Mendelian disease. *Genet Med* **20:** 159–163. doi:10.1038/gim.2017.86
Miga KH, Eichler EE. 2023. Envisioning a new era: complete genetic information from routine, telomere-to-telomere genomes. *Am J Hum Genet* **110:** 1832–1840. doi:10.1016/j.ajhg.2023.09.011
Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, Bzikadze A, Brooks S, Howe E, Porubsky D, Logsdon GA, et al. 2020. Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585:** 79–84. doi:10.1038/s41586-020-2547-7
Mok TS, Cheng Y, Zhou X, Lee KH, Nakagawa K, Niho S, Lee M, Linke R, Rosell R, Corral J, et al. 2018. Improvement in overall survival in a randomized study that compared dacomitinib with gefitinib in patients with advanced non-small-cell lung cancer and *EGFR*-activating mutations. *J Clin Oncol* **36:** 2244–2250. doi:10.1200/JCO.2018.78.7994
Muñoz-Barrera A, Rubio-Rodríguez LA, Díaz-de Usera A, Jáspez D, Lorenzo-Salazar JM, González-Montelongo R, García-Olivares V, Flores C. 2022. From samples to germline and somatic sequence variation: a focus on next-generation sequencing in melanoma research. *Life (Basel)* **12:** 1939. doi:10.3390/life12111939
Murakami K, Tago S-I, Takishita S, Morikawa H, Kojima R, Yokoyama K, Ogawa M, Fukushima H, Takamori H, Nannya Y, et al. 2024. Pathogenicity prediction of gene fusion in structural variations: a knowledge graph-infused explainable artificial intelligence (XAI) framework. *Cancers (Basel)* **16:** 1915. doi:10.3390/cancers16101915
Nakamura W, Hirata M, Oda S, Chiba K, Okada A, Mateos RN, Sugawa M, Iida N, Ushiama M, Tanabe N, et al. 2024. Assessing the efficacy of target adaptive sampling long-read sequencing through hereditary cancer patient genomes. *NPJ Genom Med* **9:** 11. doi:10.1038/s41525-024-00394-z
Nattestad M, Goodwin S, Ng K, Baslan T, Sedlazeck FJ, Rescheneder P, Garvin T, Fang H, Gurtowski J, Hutton E, et al. 2018. Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line. *Genome Res* **28:** 1126–1135. doi:10.1101/gr.231100.117
Nattestad M, Aboukhalil R, Chin C-S, Schatz MC. 2021. Ribbon: intuitive visualization for complex genomic variation. *Bioinformatics* **37:** 413–415. doi:10.1093/bioinformatics/btaa680
Nepomuceno TC, Lyra P, Zhu J, Yi F, Martin RH, Lupu D, Peterson L, Peres LC, Berry A, Iversen ES, et al. 2024. Assessment of *BRCA1* and *BRCA2* germline variant data from patients with breast cancer in a real-world data registry. *JCO Clin Cancer Inform* **8:** e2300251. doi:10.1200/CCI.23.00251
Ng AWT, McClurg DP, Wesley B, Zamani SA, Black E, Miremadi A, Giger O, Hoopen RT, Devonshire G, Redmond AM, et al. 2024. Disentangling oncogenic amplicons in esophageal adenocarcinoma. *Nat Commun* **15:** 4074. doi:10.1038/s41467-024-47619-4
Ni P, Huang N, Zhang Z, Wang D-P, Liang F, Miao Y, Xiao C-L, Luo F, Wang J. 2019. DeepSignal: detecting DNA methylation state from Nanopore sequencing reads using deep-learning. *Bioinformatics* **35:** 4586–4595. doi:10.1093/bioinformatics/btz276
Ni P, Nie F, Zhong Z, Xu J, Huang N, Zhang J, Zhao H, Zou Y, Huang Y, Li J, et al. 2023. DNA 5-methylcytosine detection and methylation phasing using PacBio circular consensus sequencing. *Nat Commun* **14:** 4054. doi:10.1038/s41467-023-39784-9
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.
Nigro JM, Baker SJ, Preisinger AC, Jessup JM, Hostetter R, Cleary K, Bigner SH, Davidson N, Baylin S, Devilee P. 1989. Mutations in the *p53* gene occur in diverse human tumour types. *Nature* **342:** 705–708. doi:10.1038/342705a0
Norris AL, Workman RE, Fan Y, Eshleman JR, Timp W. 2016. Nanopore sequencing detects structural variants in cancer. *Cancer Biol Ther* **17:** 246–253. doi:10.1080/15384047.2016.1139236
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

Nyaga DM, Tsai P, Gebbie C, Phua HH, Yap P, Le Quesne Stabej P, Farrow S, Rong J, Toldi G, Thorstensen E, et al. 2024. Benchmarking nanopore sequencing and rapid genomics feasibility: validation at a quaternary hospital in New Zealand. *NPJ Genom Med* **9:** 57. doi:10.1038/s41525-024-00445-5

Okojie J, O'Neal N, Burr M, Worley P, Packer I, Anderson D, Davis J, Kearns B, Fatema K, Dixon K, et al. 2024. DNA quantity and quality comparisons between cryopreserved and FFPE tumors from matched pan-cancer samples. *Curr Oncol* **31:** 2441–2452. doi:10.3390/curroncol31050183

Olivier M, Hollstein M, Hainaut P. 2010. TP53 mutations in human cancers: origins, consequences, and clinical use. *Cold Spring Harb Perspect Biol* **2:** a001008. doi:10.1101/cshperspect.a001008

Olson ND, Wagner J, McDaniel J, Stephens SH, Westreich ST, Prasanna AG, Johanson E, Boja E, Maier EJ, Serang O, et al. 2022. PrecisionFDA truth challenge V2: calling variants from short and long reads in difficult-to-map regions. *Cell Genom* **2:** 100129. doi:10.1016/j.xgen.2022.100129

O'Neill K, Pleasance E, Fan J, Akbari V, Chang G, Dixon K, Csizmok V, MacLennan S, Porter V, Galbraith A, et al. 2024. Long-read sequencing of an advanced cancer cohort resolves rearrangements, unravels haplotypes, and reveals methylation landscapes. *Cell Genome* **4:** 100674.

Pagès-Gallego M, de Ridder J. 2023. Comprehensive benchmark and architectural analysis of deep learning models for nanopore sequencing basecalling. *Genome Biol* **24:** 71. doi:10.1186/s13059-023-02903-2

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

Patel A, Dogan H, Payne A, Krause E, Sievers P, Schoebe N, Schrimpf D, Blume C, Stichel D, Holmes N, et al. 2022. Rapid-CNS2: rapid comprehensive adaptive nanopore-sequencing of CNS tumors, a proof-of-concept study. *Acta Neuropathol* **143:** 609–612. doi:10.1007/s00401-022-02415-6

Paulin LF, Fan J, O'Neill K, Pleasance E, Porter VL, Jones SJM, Sedlazeck FJ. 2025. Closing the gaps, and improving somatic structural variant analysis and benchmarking using CHM13-T2T. *Genome Res* (this issue) doi:10.1101/gr.279352.124

Peng Y, Yuan C, Tao X, Zhao Y, Yao X, Zhuge L, Huang J, Zheng Q, Zhang Y, Hong H, et al. 2020. Integrated analysis of optical mapping and whole-genome sequencing reveals intratumoral genetic heterogeneity in metastatic lung squamous cell carcinoma. *Transl Lung Cancer Res* **9:** 670–681. doi:10.21037/tlcr-19-401

Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, et al. 2000. Molecular portraits of human breast tumours. *Nature* **406:** 747–752. doi:10.1038/35021093

Picard M, Scott-Boyer M-P, Bodein A, Périn O, Droit A. 2021. Integration strategies of multi-omics data for machine learning analysis. *Comput Struct Biotechnol J* **19:** 3735–3746. doi:10.1016/j.csbj.2021.06.030

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

Pratanwanich PN, Yao F, Chen Y, Koh CWQ, Wan YK, Hendra C, Poon P, Goh YT, Yap PML, Chooi JY, et al. 2021. Identification of differential RNA modifications from nanopore direct RNA sequencing with xPore. *Nat Biotechnol* **39:** 1394–1402. doi:10.1038/s41587-021-00949-w

Preisler L, Habib A, Shapira G, Kuznitsov-Yanovsky L, Mayshar Y, Carmel-Gross I, Malcov M, Azem F, Shomron N, Kariv R, et al. 2021. Heterozygous APC germline mutations impart predisposition to colorectal cancer. *Sci Rep* **11:** 5113. doi:10.1038/s41598-021-84564-4

Prior IA, Lewis PD, Mattos C. 2012. A comprehensive survey of Ras mutations in cancer. *Cancer Res* **72:** 2457–2467. doi:10.1158/0008-5472.CAN-11-2612

Pudjihartono M, Perry JK, Print C, O'Sullivan JM, Schierding W. 2022. Interpretation of the role of germline and somatic non-coding mutations in cancer: expression and chromatin conformation informed analysis. *Clin Epigenetics* **14:** 120. doi:10.1186/s13148-022-01342-3

Pulivarti R. 2023. *Cybersecurity framework profile for genomic data*. National Institute of Standards and Technology, Gaithersburg, MD.

Qin Q, Popic V, Wienand K, Yu H, White E, Khorgade A, Shin A, Georgescu C, Campbell CD, Dondi A, et al. 2025. Accurate fusion transcript identification from long- and short-read isoform sequencing at bulk or single-cell resolution. *Genome Res* (this issue) doi:10.1101/gr.279200.124

Qing T, Mohsen H, Marczyk M, Ye Y, O'Meara T, Zhao H, Townsend JP, Gerstein M, Hatzis C, Kluger Y, et al. 2020. Germline variant burden in cancer genes correlates with age at diagnosis and somatic mutation burden. *Nat Commun* **11:** 2438. doi:10.1038/s41467-020-16293-7

Rausch T, Snajder R, Leger A, Simovic M, Giurgiu M, Villacorta L, Henssen AG, Fröhling S, Stegle O, Birney E, et al. 2023. Long-read sequencing of diagnosis and post-therapy medulloblastoma reveals complex rearrangement patterns and epigenetic signatures. *Cell Genomics* **3:** 100281. doi:10.1016/j.xgen.2023.100281

Razaghi R, Hook PW, Ou S, Schatz MC, Hansen KD, Jain M, Timp W. 2022. Modbamtools: Analysis of single-molecule epigenetic data for long-range profiling, heterogeneity, and clustering. bioRxiv doi:10.1101/2022.07.07.499188

Readman C, Indhu-Shree R-B, Jan MF, Inanc B. 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

Rhie A, Nurk S, Cechova M, Hoyt SJ, Taylor DJ, Altemose N, Hook PW, Koren S, Rautiainen M, Alexandrov IA, et al. 2023. The complete sequence of a human Y chromosome. *Nature* **621:** 344–354. doi:10.1038/s41586-023-06457-y

Rodriguez I, Rossi NM, Keskus AG, Xie Y, Ahmad T, Bryant A, Lou H, Paredes JG, Milano R, Rao N, et al. 2024. Insights into the mechanisms and structure of breakage-fusion-bridge cycles in cervical cancer using long-read sequencing. *Am J Hum Genet* **111:** 544–561. doi:10.1016/j.ajhg.2024.01.002

Rodriguez-Martin B, Alvarez EG, Baez-Ortega A, Zamora J, Supek F, Demeulemeester J, Santamarina M, Ju YS, Temes J, Garcia-Souto D, et al. 2020. Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition. *Nat Genet* **52:** 306–319. doi:10.1038/s41588-019-0562-0

Rossi NM, Dai J, Xie Y, Wangsa D, Heselmeyer-Haddad K, Lou H, Boland JF, Yeager M, Orozco R, Freites EA, et al. 2023. Extrachromosomal amplification of human papillomavirus episomes is a mechanism of cervical carcinogenesis. *Cancer Res* **83:** 1768–1781. doi:10.1158/0008-5472.CAN-22-3030

Rowe WPM. 2019. When the levee breaks: a practical guide to sketching algorithms for processing the flood of genomic data. *Genome Biol* **20:** 199. doi:10.1186/s13059-019-1809-x

Sabatella M, Mantere T, Waanders E, Neveling K, Mensenkamp AR, van Dijk F, Hehir-Kwa JY, Derks R, Kwint M, O'Gorman L, et al. 2021. Optical genome mapping identifies a germline retrotransposon insertion in *SMARCB1* in two siblings with atypical teratoid rhabdoid tumors. *J Pathol* **255:** 202–211. doi:10.1002/path.5755

Sakamoto Y, Sereewattanawoot S, Suzuki A. 2020a. A new era of long-read sequencing for cancer genomics. *J Hum Genet* **65:** 3–10. doi:10.1038/s10038-019-0658-5

Sakamoto Y, Xu L, Seki M, Yokoyama TT, Kasahara M, Kashima Y, Ohashi A, Shimada Y, Motoi N, Tsuchihara K, et al. 2020b. Long-read sequencing for non-small-cell lung cancer genomes. *Genome Res* **30:** 1243–1257. doi:10.1101/gr.261941.120

Sakamoto Y, Zaha S, Nagasawa S, Miyake S, Kojima Y, Suzuki A, Suzuki Y, Seki M. 2021a. Long-read whole-genome methylation patterning using enzymatic base conversion and nanopore sequencing. *Nucleic Acids Res* **49:** e81. doi:10.1093/nar/gkab397

Sakamoto Y, Zaha S, Suzuki Y, Seki M, Suzuki A. 2021b. Application of long-read sequencing to the detection of structural variants in human cancer genomes. *Comput Struct Biotechnol J* **19:** 4207–4216. doi:10.1016/j.csbj.2021.07.030

Sakamoto Y, Miyake S, Oka M, Kanai A, Kawai Y, Nagasawa S, Shiraishi Y, Tokunaga K, Kohno T, Seki M, et al. 2022. Phasing analysis of lung cancer genomes using a long read sequencer. *Nat Commun* **13:** 3464.

Sanjaya P, Maljanen K, Katainen R, Waszak SM, Genomics England Research Consortium, Aaltonen LA, Stegle O, Korbel JO, Pitkänen E. 2023. Mutation-Attention (MuAt): deep representation learning of somatic mutations for tumour typing and subtyping. *Genome Med* **15:** 47. doi:10.1186/s13073-023-01204-4

Savage SR, Yi X, Lei JT, Wen B, Zhao H, Liao Y, Jaehnig EJ, Somes LK, Shafer PW, Lee TD, et al. 2024. Pan-cancer proteogenomics expands the landscape of therapeutic targets. *Cell* **187:** 4389–4407.e15. doi:10.1016/j.cell.2024.05.039

Savara J, Novosád T, Gajdoš P, Kriegová E. 2021. Comparison of structural variants detected by optical mapping with long-read next-generation sequencing. *Bioinformatics* **37:** 3398–3404. doi:10.1093/bioinformatics/btab359

Schloissnig S, Pani S, Rodriguez-Martin B, Ebler J, Hain C, Tsapalou V, Söylev A, Hüther P, Ashraf H, Prodanov T, et al. 2024. Long-read sequencing and structural variant characterization in 1,019 samples from the 1000 Genomes Project. bioRxiv doi:10.1101/2024.04.18.590093

Schmidt TT, Tyer C, Rughani P, Haggblom C, Jones JR, Dai X, Frazer KA, Gage FH, Juul S, Hickey S, et al. 2024. High resolution long-read telomere sequencing reveals dynamic mechanisms in aging and cancer. *Nat Commun* **15:** 5149.

Schöpflin R, Melo US, Moeinzadeh H, Heller D, Laupert V, Hertzberg J, Holtgrewe M, Alavi N, Klever M-K, Jungnitsch J, et al. 2022. Integration of Hi-C with short and long-read genome sequencing reveals the structure of germline rearranged genomes. *Nat Commun* **13:** 1–15. doi:10.1038/s41467-022-34053-7

Schrinner SD, Mari RS, Ebler J, Rautiainen M, Seillier L, Reimer JJ, Usadel B, Marschall T, Klau GW. 2020. Haplotype threading: accurate polyploid

phasing from long reads. *Genome Biol* **21:** 1–22. doi:10.1186/s13059-020-02158-1
Schwenk V, Leal Silva RM, Scharf F, Knaust K, Wendlandt M, Häusser T, Pickl JMA, Steinke-Lange V, Laner A, Morak M, et al. 2023. Transcript capture and ultradeep long-read RNA sequencing (CAPLRseq) to diagnose HNPCC/Lynch syndrome. *J Med Genet* **60:** 747–759. doi:10.1136/jmg-2022-108931
Scott AJ, Chiang C, Hall IM. 2021. Structural variants are a major source of gene expression differences in humans and often affect multiple nearby genes. *Genome Res* **31:** 2249–2257. doi:10.1101/gr.275488.121
Sedlazeck FJ, Lee H, Darby CA, Schatz MC. 2018. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nat Rev Genet* **19:** 329–346. doi:10.1038/s41576-018-0003-4
Seo J-S, Rhie A, Kim J, Lee S, Sohn M-H, Kim C-U, Hastie A, Cao H, Yun J-Y, Kim J, et al. 2016. *De novo* assembly and phasing of a Korean human genome. *Nature* **538:** 243–247. doi:10.1038/nature20098
Sereewattanawoot S, Suzuki A, Seki M, Sakamoto Y, Kohno T, Sugano S, Tsuchihara K, Suzuki Y. 2018. Identification of potential regulatory mutations using multi-omics analysis and haplotyping of lung adenocarcinoma cell lines. *Sci Rep* **8:** 4926. doi:10.1038/s41598-018-23342-1
Serra F, Nieto-Aliseda A, Fanlo-Escudero L, Rovirosa L, Cabrera-Pasadas M, Lazarenkov A, Urmeneta B, Alcalde-Merino A, Nola EM, Okorokov AL, et al. 2024. P53 rapidly restructures 3D chromatin organization to trigger a transcriptional response. *Nat Commun* **15:** 2821. doi:10.1038/s41467-024-46666-1
Shafin K, Pesout T, Chang P-C, Nattestad M, Kolesnikov A, Goel S, Baid G, Kolmogorov M, Eizenga JM, Miga KH, et al. 2021. Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. *Nat Methods* **18:** 1322–1332. doi:10.1038/s41592-021-01299-w
Shihab HA, Rogers MF, Gough J, Mort M, Cooper DN, Day INM, Gaunt TR, Campbell C. 2015. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics* **31:** 1536–1543. doi:10.1093/bioinformatics/btv009
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
Shneiderman B. 1996. The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE Symposium on Visual Languages*, pp. 336–343. IEEE. Piscataway, NJ. doi:10.1109/VL.1996.545307
Simpson JT. 2024. Detecting somatic mutations without matched normal samples using long reads. bioRxiv doi:10.1101/2024.02.26.582089
Singh MP, Rai S, Pandey A, Singh NK, Srivastava S. 2021. Molecular subtypes of colorectal cancer: an emerging therapeutic opportunity for personalized medicine. *Genes Dis* **8:** 133–145. doi:10.1016/j.gendis.2019.10.013
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
Sonar S, Nyahatkar S, Kalele K, Adhikari MD. 2024. Role of DNA methylation in cancer development and its clinical applications. *Clin Transl Discov* **4:** e279. doi:10.1002/ctd2.279
Soneson C, Yao Y, Bratus-Neuenschwander A, Patrignani A, Robinson MD, Hussain S. 2019. A comprehensive examination of Nanopore native RNA sequencing for characterization of complex transcriptomes. *Nat Commun* **10:** 3359. doi:10.1038/s41467-019-11272-z
Sosinsky A, Ambrose J, Cross W, Turnbull C, Henderson S, Jones L, Hamblin A, Arumugam P, Chan G, Chubb D, et al. 2024. Insights for precision oncology from the integration of genomic and clinical data of 13,880 tumors from the 100,000 Genomes Cancer Programme. *Nat Med* **30:** 279–289. doi:10.1038/s41591-023-02682-0
Stefansson OA, Sigurpalsdottir BD, Rognvaldsson S, Halldorsson GH, Juliusson K, Sveinbjornsson G, Gunnarsson B, Beyter D, Jonsson H, Gudjonsson SA, et al. 2024. The correlation between CpG methylation and gene expression is driven by sequence variants. *Nat Genet* **56:** 1624–1631. doi:10.1038/s41588-024-01851-2
Stephens PJ, Greenman CD, Fu B, Yang F, Bignell GR, Mudie LJ, Pleasance ED, Lau KW, Beare D, Stebbings LA, et al. 2011. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144:** 27–40. doi:10.1016/j.cell.2010.11.055
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
Stricker TP, Brown CD, Bandlamudi C, McNerney M, Kittler R, Montoya V, Peterson A, Grossman R, White KP. 2017. Robust stratification of breast cancer subtypes using differential patterns of transcript isoform expression. *PLoS Genet* **13:** e1006589. doi:10.1371/journal.pgen.1006589
Su S, Gouil Q, Blewitt ME, Cook D, Hickey PF, Ritchie ME. 2021. NanoMethViz: an R/Bioconductor package for visualizing long-read methylation data. *PLoS Comput Biol* **17:** e1009524. doi:10.1371/journal.pcbi.1009524
Sun Q, Han Y, He J, Wang J, Ma X, Ning Q, Zhao Q, Jin Q, Yang L, Li S, et al. 2023. Long-read sequencing reveals the landscape of aberrant alternative splicing and novel therapeutic target in colorectal cancer. *Genome Med* **15:** 76. doi:10.1186/s13073-023-01226-y
Supplitt S, Karpinski P, Sasiadek M, Laczmanska I. 2021. Current achievements and applications of transcriptomics in personalized cancer medicine. *Int J Mol Sci* **22:** 1422. doi:10.3390/ijms22031422
Talsania K, Shen T-W, Chen X, Jaeger E, Li Z, Chen Z, Chen W, Tran B, Kusko R, Wang L, et al. 2022. Structural variant analysis of a cancer reference cell line sample using multiple sequencing technologies. *Genome Biol* **23:** 255. doi:10.1186/s13059-022-02816-6
Tan K-T, Slevin MK, Leibowitz ML, Garrity-Janger M, Shan J, Li H, Meyerson M. 2024. Neotelomeres and telomere-spanning chromosomal arm fusions in cancer genomes revealed by long-read sequencing. *Cell Genomics* **4:** 100588. doi:10.1016/j.xgen.2024.100588
Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, Boutselakis H, Cole CG, Creatore C, Dawson E, et al. 2019. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res* **47:** D941–D947. doi:10.1093/nar/gky1015
Thibodeau ML, O'Neill K, Dixon K, Reisle C, Mungall KL, Krzywinski M, Shen Y, Lim HJ, Cheng D, Tse K, et al. 2020. Improved structural variant interpretation for hereditary cancer susceptibility using long-read sequencing. *Genet Med* **22:** 1892–1897. doi:10.1038/s41436-020-0880-8
Tse OYO, Jiang P, Cheng SH, Peng W, Shang H, Wong J, Chan SL, Poon LCY, Leung TY, Chan KCA, et al. 2021. Genome-wide detection of cytosine methylation by single molecule real-time sequencing. *Proc Natl Acad Sci* **118:** e2019768118. doi:10.1073/pnas.2019768118
Ulahannan N, Pendleton M, Deshpande A, Schwenk S, Behr JM, Dai X, Tyer C, Rughani P, Kudman S, Adney E, et al. 2019. Nanopore sequencing of DNA concatemers reveals higher-order features of chromatin structure. bioRxiv doi:10.1101/833590
Unger M, Kather JN. 2024. Deep learning in cancer genomics and histopathology. *Genome Med* **16:** 44. doi:10.1186/s13073-024-01315-6
van der Pol Y, Tantyo NA, Evander N, Hentschel AE, Wever BM, Ramaker J, Bootsma S, Fransen MF, Lenos KJ, Vermeulen L, et al. 2023. Real-time analysis of the cancer genome and fragmentome from plasma and urine cell-free DNA using nanopore sequencing. *EMBO Mol Med* **15:** e17282. doi:10.15252/emmm.202217282
van Dijk EL, Naquin D, Gorrichon K, Jaszczyszyn Y, Ouazahrou R, Thermes C, Hernandez C. 2023. Genomics in the long-read sequencing era. *Trends Genet* **39:** 649–671. doi:10.1016/j.tig.2023.04.006
Vasan N, Baselga J, Hyman DM. 2019. A view on drug resistance in cancer. *Nature* **575:** 299–309. doi:10.1038/s41586-019-1730-1
Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. 2017. Attention is all you need. *Neural Inf Process Syst* **30:** 5998–6008
Veiga DFT, Nesta A, Zhao Y, Deslattes Mays A, Huynh R, Rossi R, Wu T-C, Palucka K, Anczukow O, Beck CR, et al. 2022. A comprehensive long-read isoform analysis platform and sequencing resource for breast cancer. *Sci Adv* **8:** eabg6711. doi:10.1126/sciadv.abg6711
Vermeulen C, Pagès-Gallego M, Kester L, Kranendonk MEG, Wesseling P, Verburg N, de Witt Hamer P, Kooi EJ, Dankmeijer L, van der Lugt J, et al. 2023. Ultra-fast deep-learned CNS tumour classification during surgery. *Nature* **622:** 842–849. doi:10.1038/s41586-023-06615-2
Vidal L, Dlamini Z, Qian S, Rishi P, Karmo M, Joglekar N, Abedin S, Previs RA, Orbegoso C, Joshi C, et al. 2024. Equitable inclusion of diverse populations in oncology clinical trials: deterrents and drivers. *ESMO Open* **9:** 103373. doi:10.1016/j.esmoop.2024.103373
Volden R, Kronenberg Z, Gillmor A, Verhey T, Monument M, Senger D, Dhillon H, Underwood J, Tseng E, Baker D, et al. 2023. Abstract LB078: pbfusion: detecting gene-fusion and other transcriptional abnormalities using PacBio HiFi data. *Cancer Res* **83:** LB078. doi:10.1158/1538-7445.AM2023-LB078
Wagner J, Olson ND, Harris L, McDaniel J, Cheng H, Fungtammasan A, Hwang Y-C, Gupta R, Wenger AM, Rowell WJ, et al. 2022. Curated variation benchmarks for challenging medically relevant autosomal genes. *Nat Biotechnol* **40:** 672–680. doi:10.1038/s41587-021-01158-1
Wagner J, Olson ND, McDaniel J, Harris L, Pinto BJ, Jáspez D, Muñoz-Barrera A, Rubio-Rodríguez LA, Lorenzo-Salazar JM, Flores C, et al. 2025. Small variant benchmark from a complete assembly of X and Y chromosomes. *Nat Commun* **16:** 497.
Wang F-A, Zhuang Z, Gao F, He R, Zhang S, Wang L, Liu J, Li Y. 2024. TMO-Net: an explainable pretrained multi-omics model for multi-task learning in oncology. *Genome Biol* **25:** 149. doi:10.1186/s13059-024-03293-9
Wang S, Lv W, Li T, Zhang S, Wang H, Li X, Wang L, Ma D, Zang Y, Shen J, et al. 2022. Dynamic regulation and functions of mRNA m6A modification. *Cancer Cell Int* **22:** 48. doi:10.1186/s12935-022-02452-x

Wang X, Huang M, Budowle B, Ge J. 2023. TRcaller: a novel tool for precise and ultrafast tandem repeat variant genotyping in massively parallel sequencing reads. *Front Genet* **14:** 1227176. doi:10.3389/fgene.2023.1227176

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

Wood LD, Parsons DW, Jones S, Lin J, Sjöblom T, Leary RJ, Shen D, Boca SM, Barber T, Ptak J, et al. 2007. The genomic landscapes of human breast and colorectal cancers. *Science* **318:** 1108–1113. doi:10.1126/science.1145720

Xiao C, Chen Z, Chen W, Padilla C, Colgan M, Wu W, Fang L-T, Liu T, Yang Y, Schneider V, et al. 2022. Personalized genome assembly for accurate cancer somatic mutation discovery using tumor-normal paired reference samples. *Genome Biol* **23:** 237. doi:10.1186/s13059-022-02803-x

Xu L, Seki M. 2020. Recent advances in the detection of base modifications using the Nanopore sequencer. *J Hum Genet* **65:** 25–33. doi:10.1038/s10038-019-0679-0

Xu L, Wang X, Lu X, Liang F, Liu Z, Zhang H, Li X, Tian S, Wang L, Wang Z. 2023. Long-read sequencing identifies novel structural variations in colorectal cancer. *PLoS Genet* **19:** e1010514. doi:10.1371/journal.pgen.1010514

Yang Y, Zhang M, Wang Y. 2022. The roles of histone modifications in tumorigenesis and associated inhibitors in cancer therapy. *J Natl Cancer Cent* **2:** 277–290. doi:10.1016/j.jncc.2022.09.002

Yu M, Qian K, Wang G, Xiao Y, Zhu Y, Ju L. 2023. Histone methyltransferase SETD2: an epigenetic driver in clear cell renal cell carcinoma. *Front Oncol* **13:** 1114461. doi:10.3389/fonc.2023.1114461

Zhang J, Bajari R, Andric D, Gerthoffert F, Lepsa A, Nahal-Bose H, Stein LD, Ferretti V. 2019. The International Cancer Genome Consortium data portal. *Nat Biotechnol* **37:** 367–369. doi:10.1038/s41587-019-0055-9

Zhang H, Qin C, An C, Zheng X, Wen S, Chen W, Liu X, Lv Z, Yang P, Xu W, et al. 2021. Application of the CRISPR/Cas9-based gene editing technique in basic research, diagnosis, and therapy of cancer. *Mol Cancer* **20:** 126. doi:10.1186/s12943-021-01431-6

Zhang J-Y, Zhang Y, Wang L, Guo F, Yun Q, Zeng T, Yan X, Yu L, Cheng L, Wu W, et al. 2024. A single-molecule nanopore sequencing platform. bioRxiv doi:10.1101/2024.08.19.608720

Zheng Z, Su J, Chen L, Lee Y-L, Lam T-W, Luo R. 2023. ClairS: a deep-learning method for long-read somatic small variant calling. bioRxiv doi:10.1101/2023.08.17.553778

Zheng J, Li T, Ye H, Jiang Z, Jiang W, Yang H, Wu Z, Xie Z. 2024a. Comprehensive identification of pathogenic variants in retinoblastoma by long- and short-read sequencing. *Cancer Lett* **598:** 217121. doi:10.1016/j.canlet.2024.217121

Zheng Z, Zhu M, Zhang J, Liu X, Hou L, Liu W, Yuan S, Luo C, Yao X, Liu J, et al. 2024b. A sequence-aware merger of genomic structural variations at population scale. *Nat Commun* **15:** 960. doi:10.1038/s41467-024-45244-9

Zhou L, Qiu Q, Zhou Q, Li J, Yu M, Li K, Xu L, Ke X, Xu H, Lu B, et al. 2022. Long-read sequencing unveils high-resolution HPV integration and its oncogenic progression in cervical cancer. *Nat Commun* **13:** 2563. doi:10.1038/s41467-022-30190-1

Zhu Z, Hu E, Shen H, Tan J, Zeng S. 2023. The functional and clinical roles of liquid biopsy in patient-derived models. *J Hematol Oncol* **16:** 36. doi:10.1186/s13045-023-01433-5

Zhu K, Jones MG, Luebeck J, Bu X, Yi H, Hung KL, Wong IT, Zhang S, Mischel PS, Chang HY, et al. 2024. CoRAL accurately resolves extrachromosomal DNA genome structures with long-read sequencing. *Genome Res* **34:** 1344–1354. doi:10.1101/gr.279131.124

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

Zook JM, Catoe D, McDaniel J, Vang L, Spies N, Sidow A, Weng Z, Liu Y, Mason CE, Alexander N, et al. 2016. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data* **3:** 160025. doi:10.1038/sdata.2016.25

Zook JM, Hansen NF, Olson ND, Chapman L, Mullikin JC, Xiao C, Sherry S, Koren S, Phillippy AM, Boutros PC, et al. 2020. A robust benchmark for detection of germline large deletions and insertions. *Nat Biotechnol* **38:** 1347–1355. doi:10.1038/s41587-020-0538-8

Zumalave S, Santamarina M, Espasandín NP, Garcia-Souto D, Temes J, Baker TM, Pequeño-Valtierra A, Otero I, Rodríguez-Castro J, Oitabén A, et al. 2024. Synchronous L1 retrotransposition events promote chromosomal crossover early in human tumorigenesis. bioRxiv doi:10.1101/2024.08.27.596794