Genome Biology

RESEARCH

Open Access

Benchmarking of sequencing technologies defines optimal strategies for genetic variants detection in a human genome

Robert J. M. Eveleigh[1,3], Sarah J. Reiling[2,3], Jose Hector Galvez[1,3], Mathieu Bourgey[1,2,4], Jiannis Ragoussis[2,3*] and Guillaume Bourque[1,2,3,4*]

*Correspondence: ioannis.ragoussis@mcgill.ca; guil.bourque@mcgill.ca

[1] Canadian Centre for Computational Genomics, McGill University, Montreal, QC H3A 1A4, Canada
[2] Department of Human Genetics, McGill University, Montreal, QC H3A 0C7, Canada
[3] McGill Genome Centre, Victor Phillip Dahdaleh Institute of Genomic Medicine, Montreal, QC H3A 1A4, Canada
[4] DNA to RNA - Health Data Science Platform, McGill University, Montreal, QC H3A 2R7, Canada

Abstract

Background: Advances in sequencing technologies continue to improve the resolution and completeness with which human genetic variation can be characterized. Short-read sequencing remains widely used due to its high base accuracy, throughput, and cost efficiency; however, its limited ability to resolve repetitive and structurally complex regions has accelerated adoption of long-read sequencing platforms, including those from Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT).

Results: We systematically compared sequencing technologies and variant calling pipelines for small variants and structural variants across diverse genomic contexts and sequencing depths. Short-read sequencing combined with DRAGEN achieved high accuracy for single-nucleotide variants (SNVs) and indels in well-mapped and moderately complex regions but showed reduced sensitivity and completeness for structural variant detection. In contrast, long-read sequencing platforms demonstrated clear advantages in detecting structural variants and resolving small variants in difficult genomic regions, although challenges remain in specific indel-prone sequence contexts. Among long-read pipelines, PacBio Revio with DeepVariant achieved the highest SNV and indel accuracy genome-wide, while ONT R10 with DeepVariant performed particularly well in clinically relevant loci. Structural variant detection was dominated by long-read optimized callers, with SVIM and Sawfish performing best for PacBio, and Sniffles2 and CuteSV2 for ONT, consistently outperforming short-read-based methods across variant classes and sizes. Coverage analyses indicated that long-read sequencing reached accuracy saturation between 20× and 45×, whereas short-read sequencing required more than 60× coverage to approach maximal genome completeness.

Conclusions: These results provide practical guidance for platform and pipeline selection. Long-read sequencing enables more comprehensive detection and resolution of structural variants and variation in complex genomic regions, while short-read sequencing remains a cost-effective and scalable solution for high-throughput genotyping and clinically focused applications.

Keywords: Third generation sequencing, Genome in a bottle, T2T, CMRG, Variant calling, Structural variants, LRS, SRS, Illumina, MGI, PacBio, Nanopore, WGS, Human reference genome, ILMN, MGI, PB, ONT

Background

The evolution of sequencing technologies is generally categorized into three generations, each marked by groundbreaking innovations that have progressively improved the speed, cost-efficiency, and accuracy in the detection of various forms of genomic variation. First-generation Sanger sequencing [1], while highly accurate, was laborious and expensive, restricting its application to short DNA fragments and hindering large-scale genomic studies. The advent of second-generation, or Next-Generation Sequencing (NGS)-like Illumina's sequencing-by-synthesis [1] (SBS) and, more recently, MGI's DNA nanoball technology [2]- collectively known as short-read sequencing (SRS), dramatically increased throughput and reduced costs, enabling genome-wide analyses at an unprecedented scale. However, NGS's reliance on short reads presented challenges, particularly in repetitive genomic regions [3, 4], leading to potential sequencing errors and gaps in genome coverage. To address these limitations, synthetic and third-generation, or linked- and long-read sequencing (LRS) technologies, including 10X Genomics' linked-read sequencing technology [5], PacBio's Single Molecule, Real-Time (SMRT) [6] sequencing and Oxford Nanopore Technologies' [7], emerged. These platforms introduced the ability to reconstruct long-range genomic information or sequence long stretches of DNA in a single read providing comprehensive insights into complex genomic architectures, including repetitive sequences and large structural variants. Furthermore, LRS offer functionalities like real-time sequencing and epigenetic analysis. While ongoing efforts focus on minimizing error rates and enhancing bioinformatics algorithms to improve data accuracy and utility, third-generation sequencing represents a significant leap forward.

The past decade was marked by advancements in small (SNP and Indels) and large (structural; SV) germline variant detection, driven by improvements in sequencing technology, bioinformatics and computational infrastructure. Early tools like GATK [8] HaplotypeCaller laid the foundation for detecting single nucleotide variants (SNVs) and small insertions or deletions (indels) from SRS. More recently, the advent of machine learning (ML)-powered tools such as Clair3 [9], DeepVariant [10] and Pepper-Margin-DeepVariant [11] has been coupled with advancements and utilization of hardware acceleration with tools like DRAGEN [11], Parabricks [12], MegaBolt [13] to enable rapid, large-scale variant detection while significantly enhancing sensitivity and accuracy for both SRS and LRS.

Structural variant (SV)-large (typically greater than 50 bp) deletions, insertions, duplications, inversions, and translocations-detection has similarly evolved, transitioning from limited resolution with early sequencing methodologies and tools to more accurate detection methods through integration of multiple callers like MetaSV [14] or modern single-call approaches, such as FPGA-accelerated DRAGEN [11] or the assembly-based and ML filtering Dysgu [15]. Long-range and long-read

technologies have further pushed these boundaries, enabling accurate identification of larger and more complex SVs with novel complementary tools like LongRanger [5], CuteSV [16], Nanocaller [17], pbsv [18], Sawfish [19], Sniffles2 [20], and SVIM [21].

As sequencing technologies and variant calling algorithms evolve, robust validation and benchmarking methodologies have become essential for assessing the accuracy, precision, and reliability of genomic data [22, 23]. Organisations like the Genome in a Bottle [24] (GIAB) consortium, SEQC2 [25] and the Global Alliance for Genomics and Health [22] (GA4GH) have played pivotal roles in developing standards for benchmarking, and providing high-confidence reference datasets that allow for the detection and quantification of errors, biases, and inconsistencies in genomic analyses. GIAB has established itself as a leader in creating well-characterized reference genomes with comprehensive, high-confidence variant calls, including small variants [26] (SNVs and indels), structural variants [27, 28] (SV), tandem repeats [29], and preliminary large and small variants derived from near telomere-to-telomere (T2T) assemblies.

Before the inclusion of T2T assemblies, previous GIAB small variant benchmarks, such as v4.2.1 for the HG002 genome, improved coverage using accurate linked and long reads to include challenging regions like segmental duplications and low-mappability regions, ultimately covering 92.2% of the autosomal GRCh38 assembly, up from 85% in the previously released benchmark (v3.3.2). As a result, approximately 300,000 additional SNVs and 50,000 indels were identified, largely due to the inclusion of these challenging genomic regions that were historically difficult to resolve with SRS. This new benchmark also included 16% more exon variants related to challenging, clinically relevant genes providing a more robust resource focusing on curating medically relevant genomic regions [30] (CMRG), which include loci frequently implicated in genetic disorders, pharmacogenomics, and cancer. This clinical emphasis ensures that the benchmarking dataset is particularly valuable for validating diagnostic assays targeting these areas. The GIAB reference genome panel includes widely used samples such as HG001, the Ashkenazim trio (HG002, HG003, and HG004), and the Han Chinese Trio (HG005, HG006, and HG007). This population-diverse genome panel promotes equity while providing key standards for high-confidence variant calling, and is instrumental in validating Mendelian inheritance patterns and de novo variant calls.

However, this foundational work, based on a variant-centric framework, was limited in scope, as it excluded both sex chromosomes and 12% of the autosomes [31]. To overcome these issues and avoid the inherent biases of mapping-based approaches, the T2T Consortium introduced the concept of a T2T "genome benchmark". The T2T-HG002 genome benchmark (v1.1), derived from the same Ashkenazim sample (HG002), uses the complete diploid sequence as the ground truth, enabling a fundamentally different and more comprehensive evaluation known as genome inference. The T2T-HG002 benchmark achieves near-perfect accuracy [31], covering sequences that were entirely absent from prior benchmarks. Specifically, it adds 701.4 Mb of autosomal sequence and both sex chromosomes (216.8 Mb), totaling 15.3% of the genome that was missing from v4.2.1. Excluding rDNA arrays, 99.9% of the diploid genome is covered by high-confidence T2T-HG002 benchmark regions, compared to the 85.2% covered by the GIAB v4.2.1 variant benchmark. The T2T assembly resolves regions previously inaccessible

using regular sequence-based analysis, such as all centromeric satellite arrays, recent segmental duplications, and the short arms of all five acrocentric chromosomes.

Complementing GIAB's and the T2T Consortium's efforts, GA4GH [32], have contributed by stratifying genomic regions to enhance benchmarking assessments. These stratified regions delineate areas of the genome with specific characteristics, such as coding regions, low mappability regions, high GC content regions, various types of repetitive regions, or medical relevance, enabling researchers to evaluate sequencing technologies and algorithm performance under diverse genomic contexts. To support this, small variant benchmarking tools such as vcfeval [33], hap.py [22], and large variant (SV) tools like truvari [34] are widely employed. These tools not only provide robust methodologies for comparing variant calls against high-confidence reference datasets to generate metrics like precision, recall, F1-score and genotype concordance but also enable stratification of variants to facilitate more nuances assessments. Their integration into benchmarking workflows ensures a rigorous and standardized evaluation of variant calling performance, allowing researchers to gain insights into the accuracy and limitations of sequencing platforms and bioinformatics pipelines.

In light of rapid technological advances, there has been a growing number of studies assessing the strengths and limitations of sequencing platforms [26, 31] and bioinformatic tools [17]. Prior work has provided valuable insights into variant detection accuracy, notably through previous versions of GIAB benchmark sets for small variants [26, 35, 36] and SVs [30, 37, 38], as well as comprehensive evaluations focused on specific technologies [39] or variant types and their associated callers. For instance, prior benchmarks have been critical to accelerate the development and optimization of technologies and bioinformatics approaches, highly accurate long reads [40], deep-learning- [10] and graph-based [41] variant callers, and de novo assemblies [30, 42, 43]. However, a comprehensive, head-to-head evaluation across the full spectrum of variants: SNVs, indels, and SVs, remains limited, particularly when considering modern LRS platforms paired with recent bioinformatics approaches, when varying by sequencing coverage, and complex genomic contexts. Here, we leverage the most recent LRS (PacBio Revio; ONT R10) alongside older chemistries (Sequel and R9), coupled with state-of-the-art small and large variant callers to provide a comprehensive benchmarking framework. Using the latest high-quality truth sets from the T2T and CMRG datasets for HG002, our study systematically evaluates variant calling performance across diverse scenarios, including variant type, genomic context, zygosity, and sequencing depth. By integrating these analyses, we delineate technology- and caller-specific strengths and weaknesses and provide practical recommendations for the optimal combination of sequencing platform, coverage, and variant caller, tailored to both research-oriented and clinical applications. This work extends previous studies by unifying cutting-edge sequencing platforms, variant callers, and benchmarking datasets, offering a definitive resource for variant detection best practices in complex genomes.

Results

SRS variant calling algorithms prioritize precision and sensitivity differently

To evaluate the accuracy of SRS, LRS and synthetic platforms, we assessed various libraries using the HG002 cell line (Methods). Specifically, we generated libraries and sequenced an equivalent of one lane using four different platforms: two SRS, (Illumina

Xten and MGI G400), one synthetic 10X Genomics (10X), and two LRS technologies, Oxford Nanopore Technologies with two chemistries (R9.4: ONT1 and R10.4.1: ONT3) and Pacific Biosciences Revio (PacBio; PB3). Additionally, we obtained one PacBio Sequel II LRS dataset (PB1) from GIAB directly. The average depth of coverage obtained for each platform, as well as the average physical molecule length are summarized in Additional file 2: Table S1. These libraries allowed us to test the performance of read-length on the detection of single nucleotide variants (SNVs), indels and large structural variants using the latest T2T, and challenging, medically-relevant genes (CMRG) validation sets to generate precision, recall and F1-score metrics (Fig. 1, Table 1).

The inclusion of two benchmark sets, T2T and CMRG, was motivated by their complementary scope: T2T provides a genome-wide representation (including both sex chromosomes) of HG002 suited for research applications, whereas CMRG emphasizes clinically relevant regions enriched for challenging variant contexts. This distinction is reflected in both the scale and distribution of variant contexts. T2T includes several million small variants (e.g., ~ 2.9 M SNVs and ~ 0.95 M indels across easy and difficult regions), whereas CMRG contains only a few tens of thousands in total (~ 18 K variants),

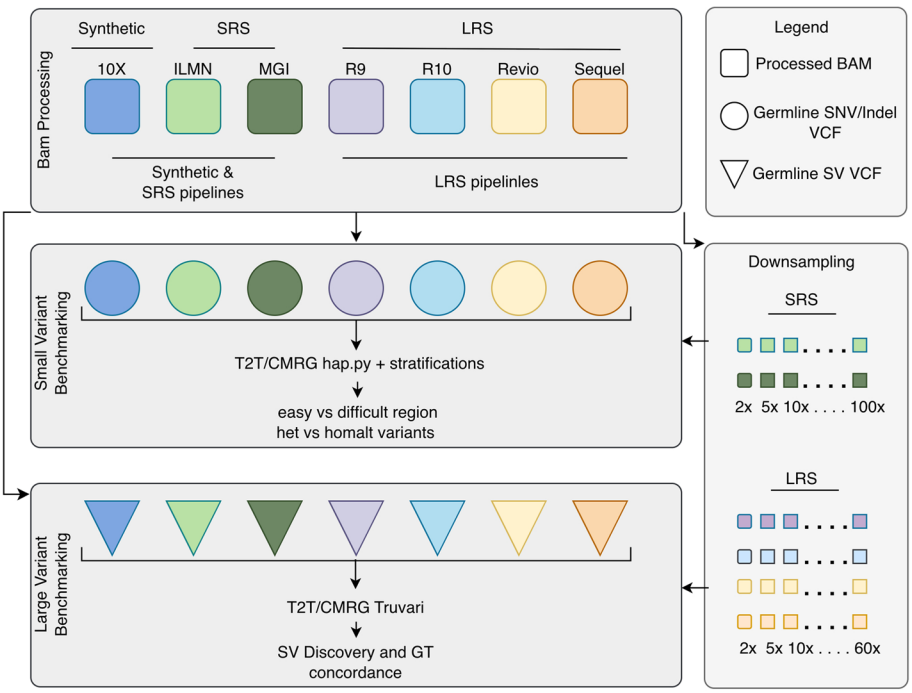

Fig. 1 Overview of the analysis workflow. This workflow diagram outlines the step-by-step rationale of the study. (Top left panel). Data from seven sequencing platforms—one synthetic (10X), two short-read platforms (Illumina and MGI), and four long-read platforms (ONT R9, ONT R10, PacBio Revio, and PacBio Sequel)—were processed using platform-specific pipelines to generate final BAM files (shown as squares). Specifically, 10X data were processed with Longranger; short-read data with GenPipes, DRAGEN, Parabricks, and MegaBOLT; and long-read data with platform-appropriate aligners (see Table 1). (Middle left panel) Small-variant benchmarking was performed using hap.py on VCFs generated by eight variant callers. Metrics including precision, recall, and F1-score were stratified by genomic region (easy vs. difficult) and zygosity. (Bottom left panel) Structural variant (SV) benchmarking was conducted using Trivari, on nine sv callers evaluating both T2T and CMRG benchmark sets to assess discovery and genotyping (GT) F1-scores. (Bottom right panel) All sequencing datasets were downsampled and reprocessed through both small-variant and SV benchmarking workflows to assess the impact of coverage

Table 1 Bioinformatic algorithms used in this study

Step/Category	Tool	Version	Reference
SRS Pipeline	GenPipes SNV/Indel/SV	4.5.0	Bourgey et al. [44]
	Parabrick SNV/Indel	2.4.0-rc3	O'Connell et al. [45]
	DRAGEN SNV/Indel/SV	4.3.13	Behera et al. [46]
	MegaBolt SNV/Indel	2.2.2.1	MGI[a]
	10X Longranger	2.2.2	Marks et al. [5]
LRS Pipeline	Minimap2	2.26	Li et al. [47]
	pbmm2	1.13.1	PacBio[b]
Germline Variant Calling	Clair3	1.2.0	Zheng et al. [9]
	DeepVariant	1.9.0	Poplin et al. [10]
	GATK HaplotypeCaller	3.8/4.1.2.0	McKenna et al. [8]
	PEPPER[c]	0.8	Shafin et al. [48]
	Nanocaller	2.0.0	Ahsan et al. [17]
Structural Variant Calling	MetaSV	0.5.4	Mohiyuddin et al. [14]
	Dysgu	1.6.1	Cleal and Baird [15]
	CuteSV	2.0.3	Jiang et al. [16]
	Sniffles	2.2.2	Smolka et al. [20]
	SVIM	2.0.0	Heller and Vingron [21]
	pbsv	2.9.0	PacBio[d]
	Sawfish	2.0.5	Saunders et al. [19]
Variant Assessment (SNV/Indels)	hap.py	0.3.15	Illumina[e]
Variant Assessment (SV)	truvari	4.3.1	English et al. [34]
Genome Stratification	GIAB/GA4GH	3.6	Dwarshuis et al. [32]
Benchmark datasets	T2T	1.1	Hansen et al. [31]
	CMRG	1.0	Wagner et al. [30]

[a] https://en.mgi-tech.com/products/software_info/6/

[b] https://github.com/PacificBiosciences/pbmm2

[c] PEPPER: Pepper-Margin-DeepVariant

[d] https://github.com/PacificBiosciences/pbsv

[e] https://github.com/Illumina/hap.py

representing a > 100 × reduction in absolute variant counts (Additional file 2: Table S2). Despite this dramatic reduction in size, CMRG is proportionally enriched for variants in difficult-to-map regions. For SNVs, 35% of CMRG calls fall in difficult regions compared to only 21% in T2T, a ~ 1.7 × increase in representation. For indels, both truth sets are dominated by difficult sites (> 75%), but CMRG shows a modest increase in the fraction found in easier regions (23% vs. 19%). Taken together, these differences highlight how CMRG prioritizes the challenging genomic regions most relevant to clinical interpretation, while T2T provides a comprehensive genome-wide benchmark.

Using these two validation sets, we first tested the following variant detection tools on the two SRS technologies using four CPU-based callers: GATK3, GATK4 haplotype caller, DeepVariant, Clair3, two graphics processing units (GPU)-based: Parabricks and MegaBOLT, and one field-programmable gate array technology (FPGA)-based: DRAGEN using default settings in most cases (see Methods). To evaluate small variant performance, we identified all easy- and difficult-to-map variants within each validation set to generate benchmarking metrics. All algorithms achieved high F1-scores across both validation sets: over 98.5% (T2T) and 95.3% (CMRG) for SNVs, and over 91.7% (T2T)

and 91.0% (CMRG) for indels (Additional file 2: Table S3). DRAGEN ranked the highest across both variant types and validation sets, based on F1-score, followed closely by DeepVariant, then typically by Clair3, GATK4, GATK3, Parabrick, and Megabolt.

Although F1 scores varied only slightly between technologies, MGI consistently exhibited less type I errors (false discovery rates or FDR), whereas Illumina (ILMN) showed less type II errors (false negative rates or FNR) across all validation sets. Despite these systematic differences, the relative ranking of variant calling algorithms remained largely consistent between SRS platforms (Fig. 2a). Consequently, we focused subsequent comparisons on ILMN data to better resolve the subtle differences in algorithmic performance. When comparing FDR and FNR across variant callers, we observed modest but consistent trade-offs between these two statistical error metrics. DRAGEN was the most sensitive algorithm, achieving the lowest FNRs across all truth sets (1.17%, and 3.76% for T2T, and CMRG, respectively), followed closely by DeepVariant, which exhibited slightly higher FNRs (2.41%, and 5.81%). DeepVariant, however, was generally the most precise caller, attaining the lowest FDRs for CMRG (0.98%) and ranking second

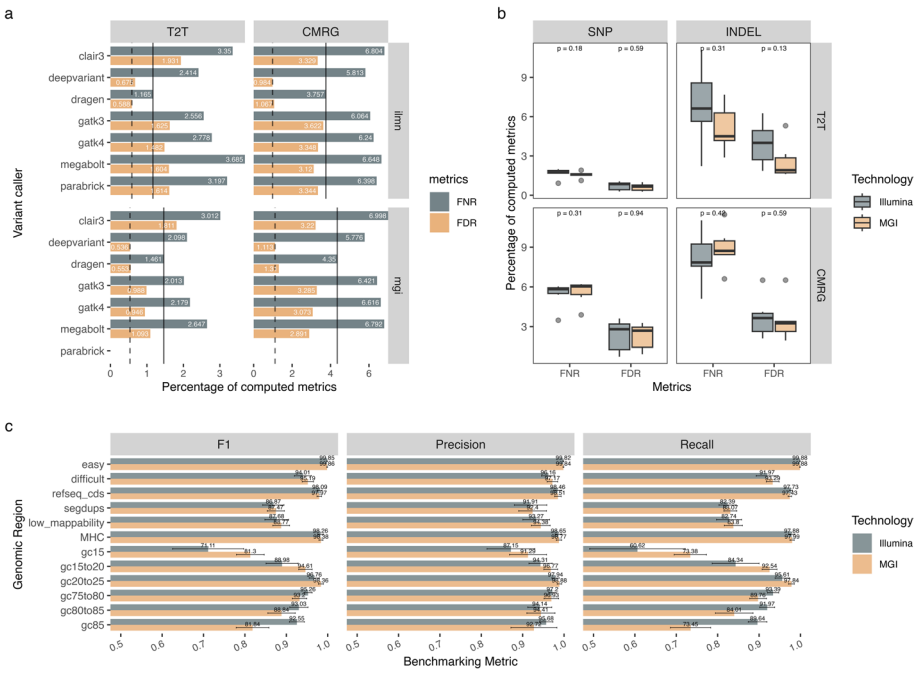

Fig. 2 SRS Variant Caller Priorities: Precision vs Sensitivity. **a** False Negative and False Discovery Rates Across Callers and Technologies Comparison of false negative rate (FNR) and false discovery rate (FDR) percentages for diploid germline variant callers using one lane of 41 × Illumina HiSeq X Ten and 45 × MGI G400 sequencing data, evaluated against the T2T and CMRG validation sets. Each point label corresponds to either FNR or FDR, with vertical dashed or solid lines indicating the minimal values, representing the top-performing caller in each category. **b** FNR and FDR Distributions Across Variant Callers and Technologies. Distribution of FNR and FDR for Illumina and MGI technologies across multiple germline variant callers (excluding Parabricks), stratified by SNPs, indels, and both T2T and CMRG validation datasets. Two-sided Wilcoxon rank-sum tests were performed, revealing no significant differences ($p > 0.05$) in FNR or FDR distributions between technologies for any variant type or validation set. **c** Performance Differences Across Genomic Contexts in T2T. Bar plots showing differences in F1 score, recall, and precision between Illumina and MGI technologies across genomic contexts within the T2T truth set. Error bars indicate variability across multiple variant callers (excluding Parabricks). MGI demonstrates higher F1 scores in difficult-to-map regions, including the MHC, segmental duplications, low-mappability regions, and high GC-content regions (> 70%). Conversely, Illumina outperforms in AT-rich regions (> 70% AT content)

only to DRAGEN for T2T (0.68% vs. 0.59%). Clair3, the other machine-learning–based caller evaluated, demonstrated high precision for CMRG (FDR: 3.33%), outperforming all GATK-based pipelines including GATK4 (3.35%), MegaBOLT (3.12%), Parabricks (3.34%), and GATK3 (3.62%). Its sensitivity, however, was intermediate; FNRs for Clair3 (6.80%) placed the caller between CPU-based GATK callers (GATK3: 5.96%; GATK4: 6.14%) and GPU-accelerated implementations (MegaBOLT: 6.55%; Parabricks: 6.65%), while Clair3 exhibited the highest FNR among all tools for CMRG. On the T2T dataset, Clair3 performed relatively poorly for both FDR (1.93%) and FNR (3.35%), suggesting limited training on or generalization to this truth set. Among GATK-based pipelines, CPU-based GATK3 and GATK4 slightly outperformed their GPU-accelerated counterparts. GATK4 achieved lower FDR than GATK3 (0.61% vs. 0.70%), whereas GATK3 had marginally lower FNR (0.85% vs. 0.88%). MegaBOLT and Parabricks, both based on GATK4, yielded performance metrics similar to the CPU implementation. Between these two GPU pipelines, MegaBOLT exhibited slightly higher FNR (1.08% vs. 0.95%), whereas Parabricks had slightly higher FDR for CMRG (3.34% vs. 3.12%).

Although the ranking of the variant callers were similar between both SRS technologies (Fig. 2a), we further stratified the SRS data by variant type, validation sets and genomic context according to GIAB/GA4GH genome stratifcation regions to identify inter-technology differences. First, we analysed the FNR and FDR distributions across all callers (excluding Parabricks) by variant type and validation set using a two-sided Wilcoxon test (significance threshold: $p = 0.05$). It was observed that for both SNV and indels, the distributions were not significantly different across all validation sets. (Fig. 2b). Next, we stratified the small variant calls made by all algorithms according to GIAB/GA4GH annotated genomic regions to assess inter-technology differences in specific genomic contexts. MGI demonstrated marginally better F1 scores in difficult-to-map regions, driven by a slight increase in sensitivity and precision in segmental duplications, low-mappability regions, and high AT-rich regions (GC < 15%). Conversely, Illumina outperformed MGI in high GC-rich regions (> 75%) (Fig. 2c). It should be noted for AT-rich region with either DRAGEN or DeepVariant, ILMN can achieve comparable precision, recall and F1 score to MGI, but the inverse cannot be said for MGI in high GC-rich regions. Together, these findings suggest that MGI and ILMN are broadly interchangeable for small variant detection, with platform-specific differences becoming apparent only in challenging genomic contexts. MGI shows a slight advantage in low-mappability and AT-rich regions, whereas ILMN demonstrates superior performance in GC-rich regions, indicating that technology choice may be most relevant for studies focused on extreme GC-content regions.

In addition to overall performance in small variant detection, we also compared the computational efficiency, including walltime (i.e. total running time), for each of the six short-read variant callers. As anticipated, callers utilizing hardware acceleration, such as FPGA or GPU, as is the case with DRAGEN, MegaBOLT, and Parabricks, were orders of magnitude faster than CPU-based callers (Additional file 2: Table S4). Our assessment of total walltimes across SRS technologies using CPU-based callers showed that MGI data generally took longer to process than ILMN (approximately-ILMN: 33.5 h ± 2 h and MGI:46 h ± 4 h), which is likely due to higher coverage. Additionally, the GATK4 haplotype caller took almost twice as long (7h05min versus 4h32min) to

run than GATK3 for MGI data, but the inverse was observed in the Illumina dataset. This suggests that GATK4 code could be internally optimized for ILMN data.

LRS outperforms SRS for most small variants, but SRS remains superior in specific genomic contexts

To benchmark small variant calling across sequencing technologies, we evaluated SRS, LRS, and synthetic platforms using standardized GA4GH/GIAB practices against two comprehensive validation sets: T2T and CMRG. Considering the top-performing callers for each platform (see Methods), most technologies exceeded an overall F1-score of 95%, with the exception of R9 and 10X (Additional file 2: Table S4). Among all combinations, Revio with DeepVariant achieved the highest accuracy (99.3%), followed by SRS analysed with Dragen (ILMN: 99.1%, MGI: 99.0%). Sequel (Clair3: 95.6%) and R10 (DeepVariant: 95.0%) also surpassed the 95% threshold, while 10X (LongRanger: 93.8%) and R9 (Clair3: 87.68%) lagged behind. Other R9 callers performed poorly (Nanocaller: 63.0%, Pepper: 77.8%). GATK4 showed strong performance on SRS (MGI: 97.87%, ILMN: 98.43%) but substantially lower accuracy on PacBio data (87.27%). Performance trends were consistent with CMRG, where Revio again led (98.9%), followed by SRS with Dragen (ILMN: 97.6%, MGI: 97.1%). Interestingly, Sequel favoured DeepVariant (96.3%) over Clair3 (95.8%) in CMRG, reversing the trend seen with T2T. R10 with DeepVariant (96.1%), 10X (92.8%), and R9 with Clair3 (88.3%) performed less well overall.

When stratified by variant type (Fig. 3), F1-scores for SNVs exceeded 98% across all platforms except R9. Revio (DeepVariant: 99.72%) and Sequel (99.52%) led, slightly outperforming SRS with Dragen (ILMN: 99.40%, MGI: 99.29%), followed by R10 (99.20%), 10X (98.04%), and R9 (97.78%). Indel detection was more variable. Revio (97.73%) remained competitive with SRS (ILMN: 97.96%, MGI: 97.75%), but older platforms showed marked reduction in F1 (Sequel: 79.94%, 10X: 73.78%, R9: 39.9%). R10 (76.62%) demonstrated substantial improvement over R9. In CMRG, Revio retained strong performance for both SNVs (99.21%) and indels (97.16%), while SRS indel accuracy decreased modestly (ILMN: 96.27%, MGI: 95.43%). Older LRS and synthetic technologies were more strongly affected, with greater drops in indel performance. Overall, Revio consistently achieved the best combined SNV and indel accuracy, while SRS retained a slight edge for indels in T2T.

To assess the impact of different genomic regions on variant calling performance, we compared the top-performing variant callers across all technologies, stratifying each validation set across T2T and CMRG genomic regions (see Methods, Fig. 3, Additional file 1: Fig. S1). As expected, F1 scores for easy-to-map and coding regions were highly consistent across most technologies, variant types, and validation sets, with the exclusion of indels for 10X and R9 platforms. For SNV, Revio (T2T: 99.98%) achieved near-perfect accuracy in easy regions, marginally surpassing both SRS platforms (ILMN: 99.95% and MGI: 99.94%) followed by Sequel (99.91%), R10 (99.91%), 10X (99.5%) and R9 (99.36%). In contrast, indel performance revealed a slightly different hierarchy: Revio (99.41%) marginally underperformed relative to SRS (MGI: 99.90%, ILMN: 99.89%), followed by same ordering as observed for SNVs (Sequel: 99.61%, R10: 98.45%, 10X: 97.49% and R9: 71.81%). Within coding regions specifically, LRS technologies demonstrated a notable advantage for SNVs, with Revio (99.50%), R10 (99.45%), and Sequel (99.41%)

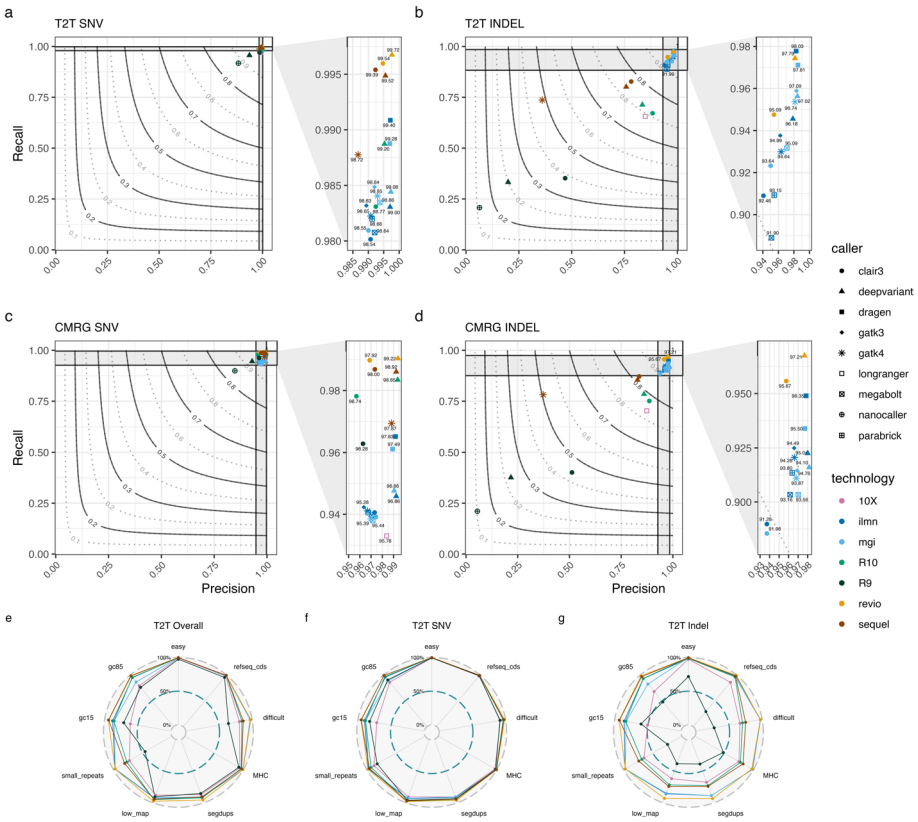

Fig. 3 Comparative Performance of SRS and LRS for Small Variant Detection. Precision–Recall plots with F1-score contours for SNVs and indels using two validation sets: T2T (panels **a**, **b**) and CMRG (panels **c**, **d**). Results are shown for five sequencing technologies: 10X Genomics (dark blue), Illumina (light red), MGI (green), Oxford Nanopore Technologies (R10: blue, R9: purple), and PacBio (Revio: yellow, Sequel II: dark red). Grey-shaded regions indicate magnified areas of the plots to improve visibility and discrimination of closely clustered values. F1 scores of top-performing variant callers across all sequencing technologies short-read sequencing (SRS: Illumina and MGI using Dragen), long-read sequencing (LRS: PacBio using DeepVariant, ONT R9 using Clair3, and ONT R10 using DeepVariant), and synthetic long reads (10X Genomics using Long Ranger)-were stratified using the GIAB/GA4GH genome stratifications. Results are presented for: (**e**) GIAB combined SNPs and indels, (**f**) GIAB SNPs only, and (**g**) GIAB indels only

outperforming SRS platforms (ILMN: 99.01%; MGI: 98.79%), with R9 (98.45%) and 10X (98.19%) trailing behind. However, ILMN (98.32%) showed improved performance for indels, placing between the two PacBio technologies (Revio: 99.41% and Sequel: 97.67%), followed by MGI (96.41%), R10 (96.33%), 10X (85.82%) and R9 (29.19%).

In difficult regions, LRS platforms outperformed SRS for SNVs (Revio: 98.34%, Sequel: 97.39% vs ILMN: 97.28%, MGI: 96.73%), while the inverse was true for indels (SRS: ILMN: 97.45%, MGI: 97.18% vs Revio: 97.17%) with remaining LRS and synthetic platforms: Sequel (75.02%), R10 (70.41%), 10X (66.53%), and R9 (27.64%) showed sharper drops in indel accuracy. The advantage of LRS was most pronounced in mapping-challenging regions: in segmental duplications, PacBio (Revio: 96.17%, Sequel: 94.31%) and R10 (95.88%) outperformed SRS (ILMN: 91.71%, MGI: 91.19%) and 10X (90.83%) for SNVs with similar trends observed in low-mappability regions. For indels, only Revio (93.48%, 93.21%) exceeded SRS (ILMN: 88.65%, 85.85%; MGI: 88.64%, 85.09%) for both segmental duplications and low-mappability regions, respectively, while other platforms

performed poorly. In the MHC, PacBio Revio (99.92%) and Sequel (99.75%) slightly surpassed SRS (ILMN: 99.55%, MGI: 99.38%) and ONT (R10: 99.43%, R9: 98.46%) for SNVs, but SRS (ILMN: 98.89% and MGI: 98.63%) outperformed all LRS and synthetic platforms for indels (Revio: 97.48%, Sequel: 82.48%, R10: 77.97%, 10X: 75.45%, and R9: 48.64%). In small repeats, Revio led for SNVs (98.27%), followed by SRS (ILMN: 97.63%, MGI: 96.72%), while indel performance favoured SRS (ILMN: 97.46% and MGI: 97.18%), with Revio (97.08%) close behind. ONT showed marked improvements in these repeat regions with R10 (SNVs: 92.9%, indels: 69.0%) compared with R9 (SNVs: 82.4%, indels: 24.7%), underscoring gains in newer chemistries. Finally, PacBio maintained consistently high SNV accuracy across most GC contexts, including the extremes, whereas SRS performance declined in at either low- or high-GC contexts (e.g., MGI GC > 85%: T2T 88.99%).

CMRG largely mirrored the platform rankings observed in T2T, with PacBio (especially Revio) providing the most robust performance across both SNVs and indels, SRS remaining competitive in easy and coding regions. Interestingly R10 showed SNV-specific strength outperforming both PacBio platforms (Revio: coding, difficult mapping regions and Sequel in all difficult regions) but had variable indel performance. 10X performing worst in challenging genomic contexts.

LRS outperform SRS for comprehensive variant detection across coverage and genomic complexity

To determine the minimum depth of coverage needed to capture the majority of SNVs and indels, samples from the Ashkenazim son (HG002/GM24385) was downsampled to various mean depths (ranging from 2 × to 120 × depending on the technology; see Methods) for both SRS (ILMN and MGI) and LRS (R9: ONT2, R10: ONT4, Sequel: PB2 and Revio: PB4) platforms then analysed using the cross platform Clair3 variant caller. Performance was assessed using both T2T and CMRG truth sets, with results categorized into GA4GH easy- and difficult-to-map regions, and then by zygosity (heterozygous: het or homozygous alternative: homalt) within those regions.

It was observed that variant discovery for both SNVs and indels showed consistent depth-dependent improvements, characterized by increasing true positives and decreasing false positives, which together enhanced recall and precision. The coverage required to reach a given performance threshold or plateau varied across region types, variant classes, and zygosity, and to a lesser degree across validation sets. Beyond these points, additional sequencing depth provided only marginal gains (Fig. 4a, b). More specifically, within the easy regions of the T2T validation set, recall and precision for SNVs plateaued at ≥ 99% around ~ 15 × coverage for Revio, Sequel, and R10, with the two PacBio platforms requiring only 10 × to achieve ≥ 99% precision. SRS platforms reached this threshold at ~ 20 ×, although MGI required more sequencing (30–35 ×) to exceed 99% precision, while R9 required higher depth (~ 30 ×) to reach comparable performance. For easy indels, similar patterns were observed, though only the PacBio platforms and R10 surpassed the ≥ 99% threshold for both metrics with Sequel at 15 ×, Revio at 15–20 ×, and R10 at 60 ×. In contrast, R9 failed to exceed 80% recall or precision even at

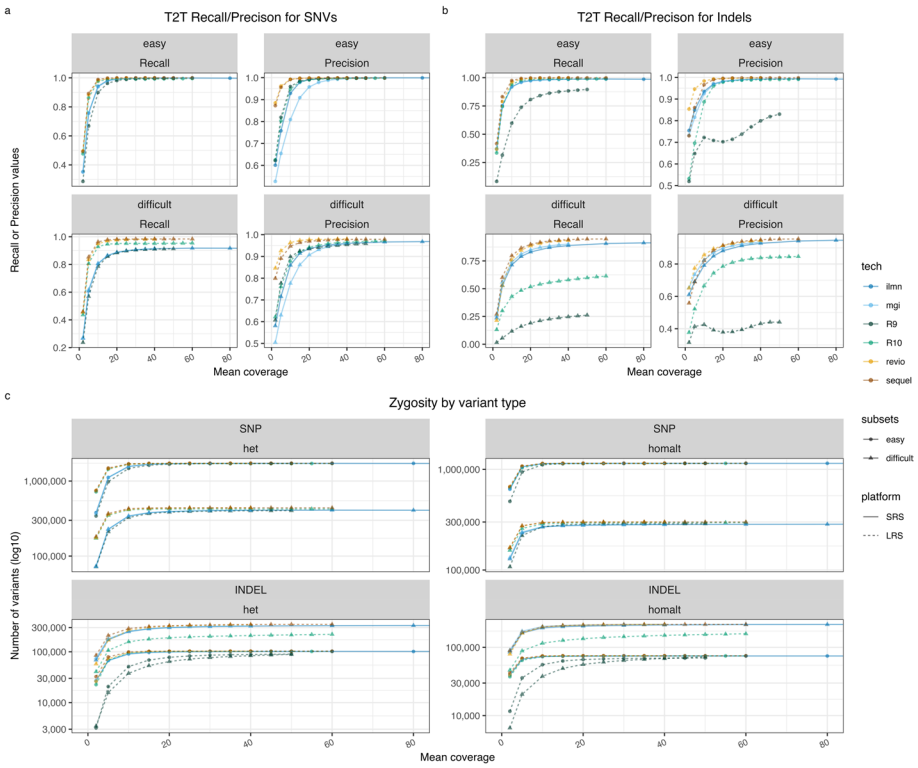

Fig. 4 Coverage Titration Analysis of HG002 Across Sequencing Technologies. Coverage titration was performed on the HG002 sample using four sequencing technologies: ILMN (downsampled from 2 × to 120 ×), MGI (2 × to 45 ×), ONT (2 × to 50 ×), and PacBio (2 × to 60 ×). Precision and Recall values were calculated for **a** SNPs and **b** indels in both easy-to-map and difficult-to-map regions using the T2T validation set. **c** Proportion of True Positive SNP and Indel Heterozygous and Homozygous Calls Across Coverage Levels for ILMN HG002 sample Line plots display the proportion of true positive SNP and indel heterozygous calls (left panel) and homozygous calls (right panel) detected at varying mean depths of coverage stratified by genomic region (easy-to-map vs. difficult-to-map) using T2T validation set

the highest coverage tested and for SRS platforms, recall stabilized at ~ 98% around 30 ×, whereas achieving 99% precision required higher coverage (~ 40 ×).

In the clinical benchmark, the relative ranking of sequencing technologies was largely consistent with the T2T results but generally required higher coverage to achieve comparable performance. For easy SNVs, all LRS and SRS platforms reached ≥ 99% recall at 20–30 × coverage, but only SRS and ONT R9 achieved ≥ 99% precision within this range. PacBio Sequel (98.97% at 45 ×) and Revio (98.27% at 25 ×) approached this threshold, while ONT R10 plateaued at ~ 97% precision near 25 ×. For indels, only the PacBio platforms achieved ≥ 99% recall (Sequel at 15 ×; Revio at 25 ×), with Sequel and SRS technologies also attaining ≥ 99% precision (Sequel at 45 ×; MGI at 30 ×; ILMN at 40 ×). Both ONT platforms underperformed relative to others, with R10 saturating near 98% recall and 96% precision at 25 ×, and R9 reaching only 89% recall and 81% precision at 40 × coverage.

Within the difficult regions of the T2T benchmark, PacBio platforms consistently required less coverage than other technologies to reach high recall and precision. For SNVs, Sequel achieved ~ 98% recall at 20 × and ~ 97% precision at 25 ×, while Revio required deeper coverage (~ 35 ×) to attain comparable recall but reached similar

precision at lower depths (15 ×). ONT R10 followed a similar trajectory, reaching 95% recall at 20 × but requiring 60 × to match PacBio's precision. SRS platforms plateaued at ~ 91% recall, requiring 35 × for MGI and 40 × for ILMN, with ILMN needing > 80 × to approach the precision of PacBio and R10. R9 resembled SRS performance, plateauing at ~ 91% recall and ~ 95% precision at 35–40 × coverage. For difficult indels, PacBio outperformed other platforms, plateauing at 35–40 × to achieve ~ 94% recall and ~ 95% precision. SRS platforms required ≥ 80 × to reach 91% recall and 94% precision, whereas ONT platforms underperformed markedly, capturing only ~ 61% (R10) and ~ 26% (R9) of indels at their highest tested coverages, with corresponding precision values of 84% and 44%, respectively.

In the CMRG, the relative ranking of technologies in difficult regions was largely consistent with the T2T results, though higher coverage was required to achieve slightly lower overall performance. For SNVs, both PacBio platforms plateaued around 35 ×, reaching ~ 97% recall and ~ 94% precision, while ONT and SRS technologies required deeper sequencing to approach similar levels. Among ONT platforms, R10 achieved higher recall than R9 (96.4% at 60 × vs. 93.2% at 50 ×) but lower precision (92.5% vs. 96.3%). SRS platforms captured fewer SNVs overall (ILMN: 84.3% recall at 60 ×; MGI: 83.5% at 40 ×) yet maintained comparable precision (ILMN: 93.7%; MGI: 92.3%). For indels, Sequel demonstrated the strongest performance in difficult regions, surpassing Revio at 35 × (recall: 93.8%, precision: 94.8%) and plateauing at 45 × with 96.2% recall and 97.0% precision, outperforming all other platforms. SRS platforms achieved similar precision (ILMN: 94.4% at 80 ×; MGI: 92.6% at 40 ×) but lower recall (ILMN: 88.2%; MGI: 85.8%) compared to Revio. ONT platforms showed modest improvements relative to T2T results, with R10 capturing ~ 71% of indels (precision: 86%) and R9 capturing ~ 31% (precision: 51%) at their highest tested coverages.

Next we investigated the impact of varying input depth on zygosity detection, observing that heterozygous (het) calls were identified more frequently in both truth sets, constituting approximately 60% of all true variants (2.82 M in T2T: ~ 61% and 12.2 k in CMRG: ~ 58%) with T2T containing > 230 × more het variants than CMRG and 194 × more homozygous alternative (homalt; 1.76 M in T2T: ~ 39% and 9.01 k in CMRG: ~ 42%; Additional file 2: Table S1). In the easy regions of the T2T truth set, ILMN and both PacBio technologies reached > 99% recall for homalt SNVs at only ~ 10 × coverage, while MGI and ONT R10 required ~ 15 × and R9 slightly more (~ 20 ×, Fig. 4c). For het SNVs, all platforms showed the expected 1.5–2 × increase in required depth to achieve comparable recall, with PacBio reaching 99% by 15 ×, and ONT and SRS technologies requiring 25–30 × (R10 and MGI at ~ 25 ×; R9 and ILMN at ~ 30 ×). For indels, PacBio outperformed other platforms with Revio and Sequel achieving 99% recall for homalts at 10–15 × coverage, while R10 required 25 × and SRS platforms plateaued around 30 ×. R9 underperformed, recovering only ~ 92% even at high coverage (60 ×). For het indels, only PacBio reached saturation (Sequel at 15 ×; Revio at 20 ×), whereas other platforms approached but did not surpass 99% recall even at high depths (R10: 98.8% at 60 ×; ILMN: 98.8% at 60 ×; MGI: 98.3% at 40 ×; R9: 87% at 50 ×).

These same coverage and zygosity trends held in the CMRG easy regions, underscoring the robustness of platform-specific performance. Interestingly, R10 closely matched PacBio for SNVs, reaching 99% recall for homoalt and het variants

at ~ 10 × and ~ 20 ×, respectively. SRS technologies and R9 required deeper coverage (ILMN and R9 at ~ 15 ×; MGI at ~ 20 × for homalts; 30–35 × for hets). For indels, PacBio led with Sequel and Revio attaining 99% of homoalts by 15–20 × (capturing all CMRG homalts at 20 × and 35 ×, for Revio and Sequel) and required up to twice as much coverage to identify over 99% of hets (Sequel: 20 ×, Revio: 30 ×). ONT R10 performed comparably for homalts (99% at 20 ×) but required 30 × to reach 98% of hets. R9 lagged substantially, requiring 35 × to reach 95% for homalts and plateauing near 86% for hets at 50 ×. Both SRS platforms exhibited similar saturation trends, achieving 99% recall for homalts at modest depth (ILMN 10 ×; MGI 20 ×) but plateauing around 97% for hets by 30–40 ×.

In the difficult regions, SNV recall plateaued at lower depths for PacBio technologies compared to ONT and SRS platforms (Fig. 4c). Both Revio and Sequel saturated near 98% recall between 15–20 × coverage for both T2T homalt and het variants, while R10 required ~ 60 × to approach similar levels (97.4% for homalt; 95.3% for hets). Among SRS platforms, ILMN showed minimal improvement beyond 60 ×, reaching 93.6% (homalt) and 91.3% (het), whereas MGI achieved comparable results at 40 ×. For indels, Revio and Sequel plateaued at ~ 94% for homalt variants, with Sequel showing a modest increase to 96.5% for hets at 60 ×. SRS platforms reached similar homalt recall (~ 94.6%) at higher depths (ILMN: 80 ×; MGI: 40 ×), but lagged behind PacBio for hets, with ILMN attaining 92% at 120 × and MGI 89% at 40 ×. ONT exhibited the lowest indel performance, consistent with known error profiles, reaching only 69.1% (homalt) and 61.6% (het) for R10 and less than half those values for R9 (31.7% and 24.8%, respectively).

In the CMRG difficult regions, PacBio platforms continued to outperform ONT and SRS technologies, with recall plateauing at lower depths but required more depth than observed in T2T (Fig. 4c). For SNVs, Revio reached 98% recall for homalt and 95% for het variants by ~ 35 ×, while Sequel required higher coverage (~ 50 ×) to achieve a similar 99% for homalts but attained 95% for hets at only 20 ×. ONT R10 performed competitively for SNVs, achieving 98% and 94% recall for homalt and hets, respectively, at just 15 ×-a notable improvement over R9, which required 20 × and 45 × to reach 96% and 91%. In contrast, SRS platforms plateaued at considerably lower values, with Illumina reaching 82% (homalt) and 85% (het) at 40–60 ×, and MGI showing comparable results of 81% and 85% at around 35 ×. For indels, Revio maintained consistent recall across genotypes (~ 94% at 35 ×), whereas Sequel achieved similar performance for homalts (94% at 30 ×) but continued to improve for hets, reaching 96% at 45 ×. SRS technologies performed comparably for homalts, with Illumina and MGI each attaining ~ 94% at 60 × and 30 ×, respectively, though both trailed PacBio for hets (91% and 89%). ONT platforms showed a pronounced drop in indel accuracy, consistent with their error modes: R10 reached only 69% and 61% (homalt and het, respectively) at 60 ×, while R9 performed substantially worse (31% and 25%). Overall, PacBio platforms displayed the most efficient and consistent performance across variant types and zygosities, whereas ONT showed major improvements from R9 to R10 but remained limited for indel detection.

Long-read technologies outperform short-read technologies in structural variant detection

Next, we evaluated the performance of the various sequencing technologies in detecting structural variants (SVs). The performance of different SV calling tools was assessed by using truvari to generate standardized metrics across the T2T and CMRG validation sets. The T2T dataset includes 20,767 deletions (DEL) and 29,640 insertions (INS) over 30 bp (excluding 13,579 DEL and 13,992 INS below 30 bp) from the HG002 sample. The CMRG subset focuses on 89 high-confidence, sequence-resolved DEL and 105 INS 50 bp or larger near 273 clinically relevant genes (Additional file 2: Table S5).

A comparative analysis of SV calling performance across sequencing technologies was conducted using 27 structural variant call sets generated from multiple technology–caller combinations. For SRS platforms, DRAGEN, dysgu, and metasv were evaluated, while long-read sequencing LRS platforms included six callers: for ONT, cutesv2, dysgu, sniffles, and svim; for PacBio, pbsv and sawfish were added. SV caller performance was assessed using two metrics. Discovery F1 reflects the balance of precision and recall in correctly identifying the presence and approximate location of structural variants, regardless of genotype. Genotyping (GT) F1 measures the accuracy of both detecting and correctly genotyping variants, providing a more stringent evaluation of caller performance. Assessment using these metrics revealed that LRS consistently outperformed

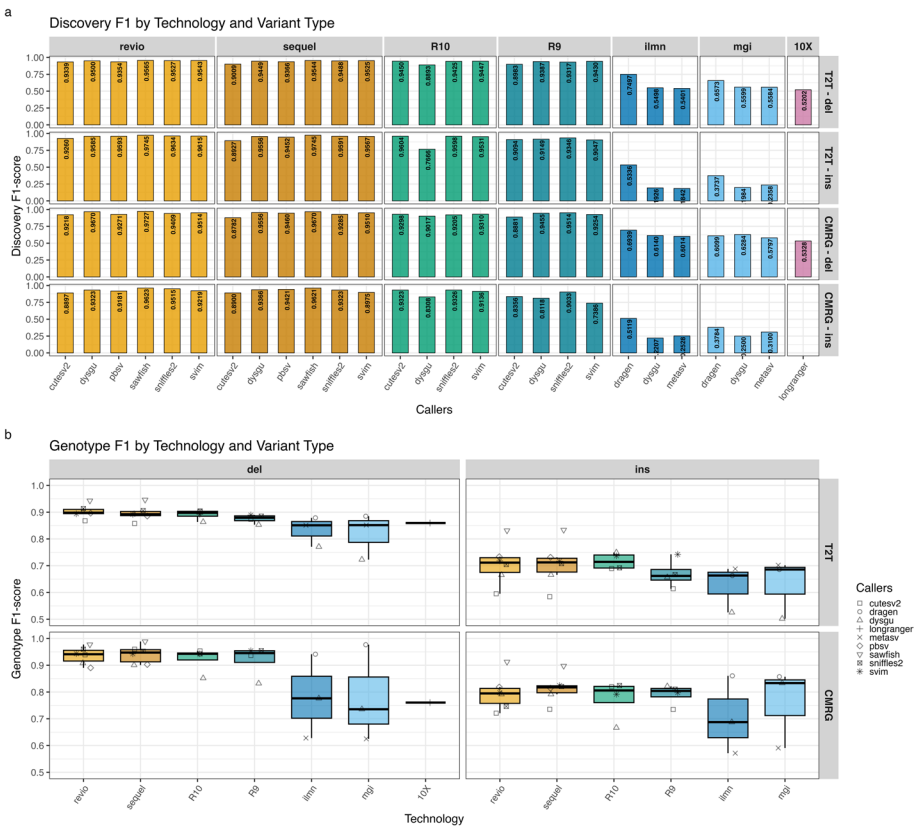

Fig. 5 Discovery and Genotyping F1 score Across Complex Validation Sets. Discovery (panel **a**) and genotyping (panel **b**) performance was evaluated using two benchmark datasets of structural variants (SVs) across seven sequencing technologies and nine SV callers. T2T : ~ 15,000 high-confidence deletions and insertions > 30 bp. CMRG : 89 deletions and 105 insertions > 50 bp across 273 medically relevant genes

SRS and synthetic technologies across all validation sets and variant types, as indicated by higher discovery F1 and GT F1 scores (Fig. 5). Specifically, for DEL, LRS achieved an average discovery F1 scores of 85.0% and 93.5% in the T2T and CMRG datasets, respectively-approximately 1.5 times higher than SRS and synthetic technologies (55.9% and 60.9%). The difference was even more pronounced for INS, where LRS attained an average discovery F1 scores of 80.8% and 90.2%, outperforming SRS by roughly three-fold (25.9% and 32.1%), with 10X failing to detect any INS variants (Additional file 2: Table S8). Similarly, LRS demonstrated substantially improved genotyping accuracy. For DEL, LRS achieved GT F1 scores of 89.3% and 93.2% for T2T and CMRG, respectively. The difference between LRS and SRS was less pronounced for DEL, with SRS scoring only 1.1–1.2 times lower (83.1% and 77.8%), particularly lower in the clinical dataset. For INS, LRS average GT F1 scores were 70.2% and 79.7%, exceeding the SRS averages by a modest 1.1-fold (62.8% and 73.3%).

Structural variant calling performance vary depending on technology, variant type and validation set

Within LRS platforms, discovery and GT F1 performance patterns varied by technology, SV type or validation set. Overall, PacBio platforms consistently outperformed ONT for both DEL and INS in the T2T and CMRG benchmarks (Fig. 5). Among platforms, newer systems (Revio and R10) generally outperformed their predecessors (Sequel and R9). For DEL, SVIM achieved the highest discovery F1 across most PacBio and ONT platforms in the T2T dataset-Revio (89.10%), Sequel (88.72%), and R9 (86.33%)-while cutesv2 ranked highest for R10 (86.65%). In contrast, in the CMRG benchmark, PacBio's sawfish outperformed all other callers, reaching 97.27% (Revio) and 96.70% (Sequel), compared to ONT's top performers sniffles2 on R9 (95.14%) and SVIM on R10 (93.10%). Interestingly, Revio paired with sawfish or SVIM, and ONT R9/R10 paired with sniffles2 or SVIM all achieved identical maximum recall (94.68%), with sawfish showing perfect precision (100%), followed by sniffles2 (95.60%) and SVIM (91.58%). GT F1 results similarly favoured PacBio. Sawfish achieved the highest DEL GT F1 in both T2T (Sequel 94.62%, Revio 94.31%) and CMRG (Sequel 98.86%, Revio 97.75%). For ONT, sniffles2 on R10 (94.44%) slightly exceeded SVIM on R9 (88.94%) in T2T, while R9 sniffles2 (95.45%) marginally surpassed R10 cutesv2 (95.35%) in CMRG.

For INS, performance patterns largely mirrored those for DEL. In T2T, SVIM achieved the highest discovery F1 for PacBio platforms (Revio: 86.66%, Sequel: 85.94%) and cutesv2 (85.07%) led for R10, outperforming R9 sniffles2 (79.35%). In CMRG, sawfish again showed superior discovery performance (Revio 96.32%, Sequel 96.21%), exceeding ONT sniffles2 (R10 93.26%, R9 90.33%). Similar recall patterns were observed across platforms, with Revio and Sequel using sawfish or sniffles2, and R10 using cutesv2, all achieving 97.25% recall. Precision was highest for sawfish (Revio 95.24%, Sequel 95.19%), followed by Revio sniffles2 (93.14%) and R10 cutesv2 (89.52%). GT F1 for INS reinforced PacBio's dominance. Sawfish led in both T2T (Sequel 83.35%, Revio 83.17%) and CMRG (Revio 91.26%, Sequel 89.75%), outperforming ONT's top performers: R10 dysgu (74.94%) and R9 SVIM (74.25%) in T2T, and R10 sniffles2 (82.41%) and R9 dysgu (82.14%) in CMRG.

As observed for LRS platforms, performance among SRS and synthetic technologies varied across SV types and validation sets. Overall, ILMN generally outperformed MGI in DEL and INS discovery for both benchmarks, whereas the opposite trend was observed for genotyping performance, with MGI surpassing ILMN across SV types and datasets, except for INS in the CMRG benchmark. For DEL discovery, DRAGEN was the top performer for both technologies in T2T (ILMN: 67.52%; MGI: 61.65%) and in CMRG (ILMN: 69.39%), while dysgu (62.84%) marginally outperformed DRAGEN (60.99%) for MGI. In contrast, GT F1 favoured MGI across both validation sets. DRAGEN achieved the highest GT F1 for DEL in T2T (MGI: 88.48%; ILMN: 87.87%) and in CMRG (MGI: 97.67%; ILMN: 94.12%).

For INS, DRAGEN again emerged as the top-performing caller for both SRS technologies, with ILMN showing markedly higher discovery F1 scores than MGI in both T2T (ILMN: 44.72%; MGI: 32.08%) and CMRG (ILMN: 51.12%; MGI: 37.84%) benchmarks. In contrast, GT F1 for INS favoured MGI in the T2T dataset, where MetaSV achieved the highest performance (MGI: 70.19%; ILMN: 68.80%). For the CMRG benchmark, however, ILMN DRAGEN led in genotyping accuracy (ILMN: 51.20%; MGI: 37.84%). Notably, the synthetic 10X LongRanger platform consistently underperformed relative to both LRS and SRS technologies. It achieved low discovery and GT F1 scores for deletions (discovery: T2T 39.69%, CMRG 58.27%; GT F1: CMRG 37.50%) and failed to detect any insertions, underscoring its limited utility for comprehensive structural variant detection.

Size-stratified performance of SV callers varies across sequencing technologies

To gain a more granular understanding of each technology's performance given various SV callers, we stratified the caller results into bins of varying, non-overlapping sizes, considering known peaks associated with Alu elements (±300 bp) and full-length LINE1 elements [28] (±6000 bp). Specifically, we created bins for SV sizes ranging from 30–50 bp (exclusively in the T2T dataset, for the purpose of genome-wide completeness), 51–250 bp, 251–500 bp, 501–6000 bp, and greater than 6000 bp. Stratification of both validation datasets revealed a general trend: the abundance of both deletions (DEL) and insertions (INS) decreased as SV size increased. Notably, the 51–250 bp size range exhibited the highest number of SV events, with a subsequent decline in event counts observed in larger size bins (Fig. 6). This pattern was consistently observed in the high-confidence datasets, T2T and CMRG, with CMRG lacking events greater than 6000 bp. However, T2T presented a slight deviation, with the 30–50 bp size range showing the second highest abundance of SV events (Additional file 2: Table S5) and number of INS events is between 1.3 to 3 times as abundant as DEL events for a given SV bin.

Applying our binning strategy using truvari across the two validation sets, SV types, and technologies reinforced a key observation: LRS platforms consistently detected a greater number of both DEL and INS events compared to SRS and synthetic data. This analysis highlighted a higher abundance of LRS-detected INS events overall and within the smaller and larger size ranges of both SV types. More specifically, LRS technologies identified approximately 4.65 times more DEL events in the smallest T2T range (30–50 bp), and approximately 2.4 (T2T) to 3.1 (CMRG)-fold more DELs in the 51–250 bp range, depending on the validation set (Additional file 2: Table S6). Similarly, for small

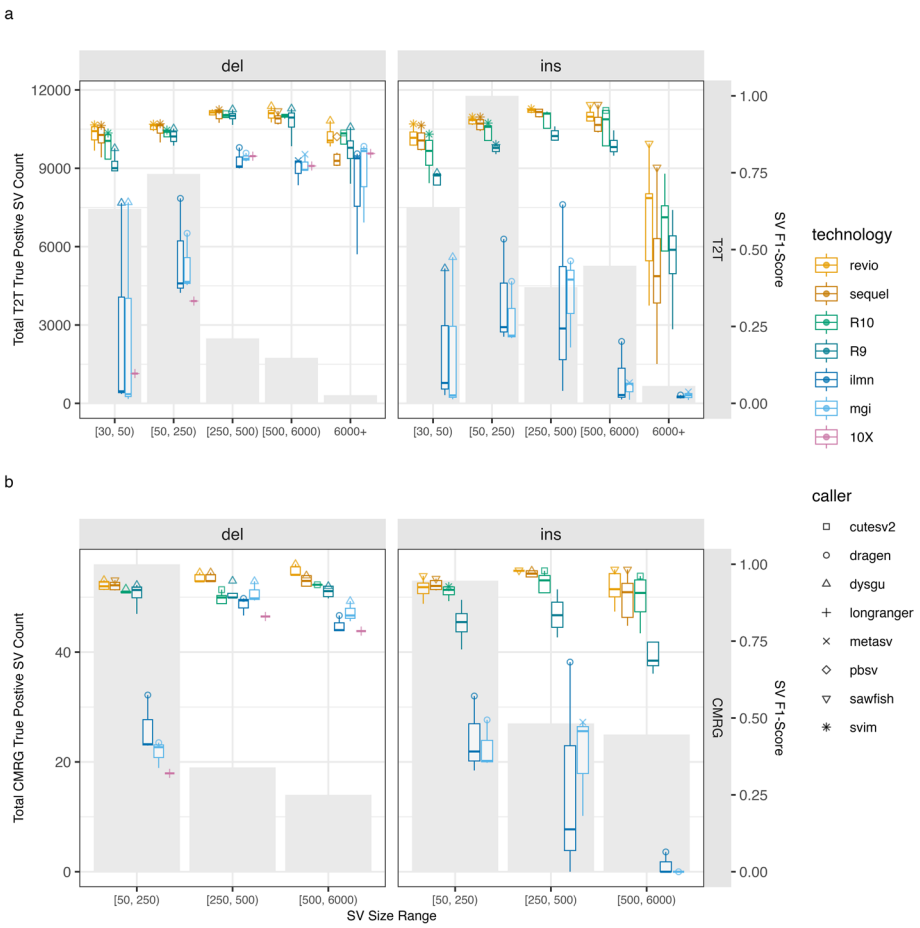

Fig. 6 F1-Score Performance by Structural Variant Size and Technology. F1-scores were evaluated for long-read sequencing (LRS), short-read sequencing (SRS), and synthetic technologies, stratified by structural variant (SV) type-deletions (DEL) and insertions (INS)-and SV size. Size categories included 30–50 bp (T2T only), 51–250 bp, 251–500 bp, 501–6,000 bp, and > 6 kb (T2T only). Boxplots show the distribution of F1-scores across SV callers within each technology, with the top-performing caller indicated for: **a** the T2T truth set and (**b**) the CMRG truth set

INS, LRS identified 6.12-fold more T2T 30–50 bp INS events, and between 4.04- (T2T) and 2.95 (CMRG)-fold more INS in the 51–250 bp range. At the opposite end of the size spectrum, LRS platforms detected between 28.7- (T2T) and 144.3- (CMRG) fold INS events in the 501–6000 bp range, and 35.7-fold more T2T events larger than 6000 bp. The synthetic 10X data typically underperformed relative to both LRS and SRS for smaller DEL and INS events, but detected approximately 1.5-fold more large 6000 + bp events than either LRS and SRS in in T2T (Fig. 6).

Across LRS platforms consistent high F1 scores were observed for both DEL and INS across all size ranges, with PacBio platforms (Revio and Sequel) consistently achieving the highest overall F1 scores across datasets. In the T2T dataset, DEL performance showed strong size-dependent trends. Revio achieved top F1 scores for small (30–50 bp, SVIM: 90.56%), mid-sized (251–500 bp, sniffles2: 95.59%), large (501–6000 bp, dysgu: 96.61%), and very large deletions (> 6000 bp, dysgu: 91.90%), while Sequel performed best in the 51–250 bp range (SVIM: 90.93%). R10 showed

comparable accuracy for mid-to-large deletions (251–500 bp; cutesv2: 94.18%, 501–6000 bp; sniffles2: 94.0%), but slightly reduced F1 for smaller (< 250 bp; SVIM:87.92 and 88.87%) and very large (> 6 kb) variants (cutesv2: 87.93%). However, with the exception of the small events (30–50 bp, dysgu: 82.94%), R9 exhibited higher accuracy compared to R10, with dysgu F1 scores of 89.36%, 95.59%, 95.84%, and 89.8% for ranges 51–250, 251–500, 501–6000 and > 6 kb bp, respectively.

For INS in T2T, Revio and Sequel again led performance, with SVIM dominating smaller and sawfish larger events. Both platforms maintained high F1 across all size ranges, with Revio achieving 90.87% F1 at 30–50 bp and Sequel slightly higher at 93.12% for 51–250 bp, only marginally outperforming Revio at 93.09% for the same range. For mid-to-large INS, Revio again led at 251–500 bp (SVIM: 95.81%), whereas Sequel performed best in the 501–6000 bp range (sawfish: 97.15%). Revio demonstrated superior robustness for large insertions, achieving 97.12% F1 at 501–6000 bp and maintaining 84.49% beyond 6 kb—an improvement of nearly 8% over Sequel (76.7%). R10 followed similar trends, with strong performance for mid-sized insertions (251–500 bp; sniffles2: 94.88%, 501–6000 bp: cutesv2: 95.3%) but lower F1 for smaller (< 250 bp: SVIM—30–50 bp: 87.54%; 51–250 bp: 91.14%) and very large variants (> 6 kb, sniffles:74.7%). R9, showed the inverse to the DEL results with lower accuracy across all ranges, particularly for small and very large insertions, with F1 values of 74.94% for 30–50 bp (dysgu). 84.15% for 51–250 bp (SVIM), 90.07% for 251–500 (sniffles2), 88.87% for 501–6000 bp (sniffles2) and 62.88% for 6000 + bp (sniffles2).

In the CMRG dataset, certain LRS platforms achieved near-saturation performance across variant sizes with both PacBio instruments performed equally well for deletions, particularly in the mid- to large-size ranges. For 50–250 bp and 250–500 bp deletions, Revio and Sequel achieved identical F1 scores of 94.83% and 97.30%, respectively, with top-performing callers being dysgu or sawfish. Revio paired with dysgu or sawfish achieved a perfect 100% F1 for large deletions (500–6000 bp). R10 achieved similarly high F1 values, though with greater caller variability depending on size. For the smallest SVs, dysgu (50–250 bp: 92.04%) outperformed sniffles2 and SVIM due to higher precision but sniffles2 and SVIM had higher recall (90.16%), mid-size (250–500 bp) cutesv2 (91.74%) and larger variant cutesv, sniffles and svim were tied at 93.33%.

For INS, Revio consistently outperformed all other LRS platforms across all size intervals, achieving 94.83%, 98.04%, and 98.31% F1, all with sawfish as the consistent leading caller. Sequel followed closely but with slightly more caller variability depending on size, sawfish was the top performer for both small (95.33%) and large (98.31%) events with dysgu, pbsv and SVIM tied for the mid-range event at 97.87%. R10 performed competitively for medium events, tying Sequel for 250–500 bp (cutesv2: 97.87%) but showed a modest drop in small (SVIM: 92.73%) and large (cuteSV2: 96.16%) insertions. R9 again produced stable yet slightly lower F1 (50–250 bp: 88.42%, 250–500 bp, 91.83, 500–6000 bp 90.45%) similar to the observations from T2T INS, but with sniffles2 being the top performer.

Across SRS technologies, both ILMN and MGI demonstrated lower F1 scores compared to long-read platforms, with the most pronounced deficits observed for INS. DEL performance was comparatively stable, though accuracy varied by size range and

technology. In the T2T benchmark, ILMN achieved its strongest performance for medium sized DELs, with DRAGEN performing best at 51–250 bp (66.65%) and 251–500 bp (83.18%). For the smallest deletions (30–50 bp), dysgu provided comparable results across platforms (ILMN: 65.19%; MGI: 65.35%), while MGI slightly surpassed ILMN for large (500–6000 bp, metasv: 80.99% vs. 78.94%) and very large (> 6 kb, DRAGEN: 83.58% vs. 81.16%) DELs. In contrast, INS detection exhibited sharp declines in F1 with increasing variant size. For ILMN, DRAGEN performed best in the 51–250 bp range (53.42%) and 251–500 bp (64.64%), but performance dropped precipitously for INS beyond 500 bp (20.13%) and > 6 kb (2.72%). MGI displayed a similar trend, with slightly higher accuracy for smaller insertions (30–50 bp, dysgu: 47.53%) but very low F1 for larger events (> 500 bp, metasv: 6.73% and 3.74%).

Performance patterns in the CMRG dataset mirrored these trends, with DRAGEN dominating most size ranges for ILMN deletions (51–250 bp: 57.47%; 251–500 bp: 88.89%; 500–6000 bp: 83.33%), whereas MGI achieved superior performance for medium-to-large deletions using dysgu (251–500 bp: 94.44%; 500–6000 bp: 88.00%). Insertion detection again lagged behind, with ILMN reaching modest F1 values for 251–500 bp (68.18%) but fell sharply for larger INS (6.54%), while MGI demonstrated limited success across all size ranges ($\leq$ 49.38%; 51–250 bp, DRAGEN: 49.38%, and 251–500 bp, metasv: 48.85%) and failed to detect INS > 500 bp. Overall, these results underscore the continued challenges of SRS-based INS detection and highlight modest but notable MGI gains in large deletion recovery compared to ILMN.

Minimum coverage requirements for SV discovery vary by technology, validation set and SV type

To determine the minimum sequencing coverage necessary for comprehensive SV discovery, we analysed multiple downsampled datasets from LRS, and SRS (excluding synthetic) platforms using truvari against T2T and CMRG datasets (see Methods). The sequencing depth required for each platform to reach 99% of its maximal F1-score was evaluated, defining this value as the coverage saturation point. This threshold represents the depth beyond which additional sequencing yields negligible improvements in SV detection accuracy.

Consistent with small variant detection, SV discovery improves with increased sequencing depth (Fig. 7). Across both datasets, LRS platforms exhibited early and consistent convergence toward their maximal mean F1-scores, saturating at approximately 30-fold coverage for both INS and DEL. Revio showed the fastest and most stable approach to saturation, achieving > 94% and > 0% mean F1 for both T2T/CMRG DEL and INS. Sequel displayed a similar trajectory, reaching the 99% saturation threshold within 30- to 35-fold depth range, confirming the high reproducibility of PacBio's HiFi-based platforms. R10 achieved comparable performance but required higher coverage (30- to 35-fold) to reach the same level of stability, while R9 lagged requiring approximately 40-fold coverage across both SV types, reflecting basecalling and alignment improvements realized in later chemistries.

In contrast, SRS platforms, ILMN and MGI, demonstrated a slower and more coverage-dependent trajectory toward performance saturation. Both required 2- to three-fold additional (40–80x) coverage to reach comparable plateaus, with DEL consistently

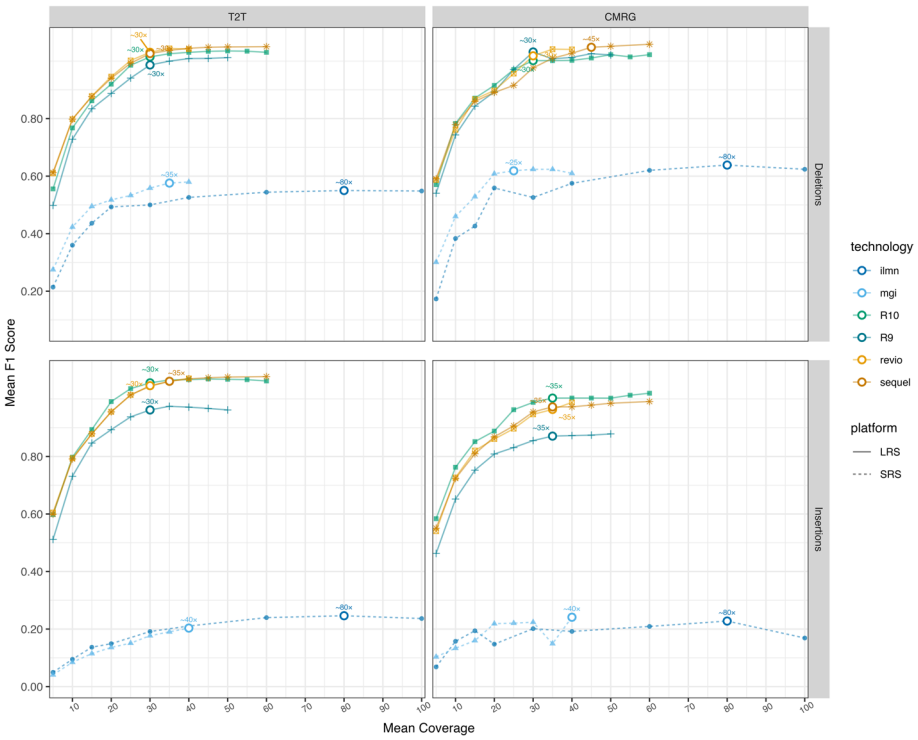

Fig. 7 SV discovery: Coverage Varies by Technology, Validation, and Type. Coverage titration of short- and long-read sequencing technologies. ILMN (5 × −100 ×) and MGI (5 × −40 ×) represent short-read sequencing (SRS), while ONT2 (2 × −50 ×), ONT4 (5 × −60 ×), PB2 (5 × −60 ×), and PB4 (5 × −40 ×) represent long-read sequencing (LRS). Mean F1 scores were computed for T2T and CMRG DEL and INS using three SRS callers (DRAGEN, dysgu, and metasv) and six LRS callers (cutesv2, dysgu, pbsv, sawfish, sniffles2, and SVIM). Circles indicate the 99% saturation threshold for each technology

outperforming INS. ILMN achieved its saturation point beyond 60-fold for DEL, while MGI approached saturation closer to 40–50-fold, suggesting marginal improvements in coverage uniformity and GC bias correction. However, even at the highest coverages tested, INS mean F1-scores for SRS datasets remained well below 30%, underscoring persistent challenges in resolving INS.

Collectively, these findings indicate that LRS technologies, particularly Revio, Sequel, and R10. reach near-maximal variant detection accuracy by ~ 30–35 × coverage, providing a efficient balance between sequencing depth and performance. Conversely, SRS platforms require substantially greater coverage (> 70 ×) to achieve comparable saturation, reinforcing the efficiency and scalability of modern long-read sequencing for high-confidence SV detection in both genome-wide research and clinical scenarios.

Discussion

In this study, the performance of various sequencing technologies and variant calling algorithms across a range of variant types, including SNVs, indels, and SVs was assessed. Moreover, by utilizing recent complex (T2T) and clinical (CMRG) validation datasets across these variant types and by further subsetting using different genomic and zygosity contexts, we highlighted distinct performance patterns between LRS, SRS, and synthetic platforms, as well as between different variant calling algorithms and their performance

Table 2 Performance ranking of sequencing technologies across variant types and genomic contexts

a

Region/Validation set	Subregion	Top 1 (T2T/CMRG)	Top 2 (T2T/CMRG)	Top 3 (T2T/CMRG)
Easy-to-map	SNV	Revio/Revio	Sequel/Sequel	ILMN/R10
	INDEL	MGI/ILMN	ILMN/Revio	Revio/MGI
Difficult-to-map	SNV	Revio/Revio	Sequel/Sequel	ILMN/R10
	INDEL	ILMN/Revio	MGI/ILMN	Revio/MGI
Easy Heterozygous	SNP	Revio/Revio	ILMN/ILMN	MGI/MGI
	INDEL	MGI/Revio	ILMN/ILMN	Revio/MGI
Easy Homozygous Alt	SNV	Revio/Revio	Sequel/Sequel	ILMN/R10
	INDEL	ILMN/MGI	MGI/ILMN	Revio/Revio
Difficult Heterozygous	SNV	Revio/Revio	Sequel/R10	ILMN/ILMN
	INDEL	ILMN/Revio	MGI/ILMN	Revio/MGI
Difficult Homozygous Alt	SNV	Revio/Revio	Sequel/Sequel	ILMN/R10
	INDEL	Revio/Revio	ILMN/ILMN	MGI/MGI

b

Region/Validation set	Subregion	Top 1	Top 2	Top 3
T2T SV	DEL	Revio	Sequel	R10
	INS	Revio	Sequel	R10
CMRG SV	DEL	Revio	Sequel	R9
	INS	Revio	Sequel	R10

Performance of LRS, SRS, and synthetic technologies was evaluated for both (a) small variants (SNVs and indels) and (b) large structural variants. The top three performing technologies were identified within each genomic region and validation set, enabling comparison of accuracy across variant types and benchmarking datasets

at varying coverage depths. These findings provide a nuanced understanding of how technology and methodology choices influence variant detection accuracy, offering insights for optimizing variant detection under certain scenarios (Table 2).

Small variant detection

Across T2T and CMRG benchmarking frameworks, the top-performing technologies for small variant detection (SNVs and indels) were led by LRS platforms, particularly Revio, followed by Sequel, with SRS platforms, ILMN and MGI, performing strongly (Fig. 2). Using the T2T benchmark, which provides a genome-wide evaluation of variant detection accuracy, Revio consistently ranked first across nearly all variant types, genomic contexts and zygosity types, highlighting its reliability for research-oriented studies that require comprehensive genome coverage [19, 32]. In contrast, within the clinically focused CMRG benchmark, Revio again achieved top performance, but was more closely followed by SRS platforms, and R10, each showing competitive results depending on variant type and zygosity. The strong performance of R10 in clinically relevant regions, alongside Revio, illustrates the growing clinical potential of LRS data, particularly for resolving complex or difficult-to-map loci (CMRG) where SRS technologies face inherent limitations [42]. Nonetheless, despite this strong overall performance, as observed in other studies [49], PacBio and more strikingly with ONT platforms continue to show limitations in indel detection, particularly for smaller variants in homopolymeric or low-complexity regions [50]. These errors often stem from

systematic basecalling [51] or alignment challenges inherent to long-read data [39], where signal noise and local sequence context can influence indel length estimation and boundary precision. Although recent chemistry and model improvements [4] have substantially reduced these errors, SRS technologies still tend to outperform LRS in detecting short indels, especially in easily mappable regions [29]. Further stratifying results by indel size and their relative abundance in each validation set would help clarify platform-specific strengths and weaknesses. Thus, while Revio provides unparalleled genome-wide completeness and strong SNV performance, integrating complementary SRS data with hybrid ML-models [52] (e.g. DeepVariant hybrid PacBio and ILMN model) or using caller ensembles could be advantageous for achieving optimal indel accuracy in both research and clinical settings.

Algorithmic performance was highly consistent across SRS and LRS platforms in easy-to-map and coding regions, with all top-performing variant callers achieving near-identical accuracy (F1 > 99% for both SNVs and indels). Recent LRS platform, Revio and R10, performed on par with ILMN and MGI in these regions with Revio reaching SNV and indel F1-scores of 99.98% and 99.41%, respectively, and R10 achieving 99.91% and 98.45%. Similarly, ILMN and MGI each exceeded 99.4% F1 for both variant types, underscoring that current sequencing and variant-calling technologies have effectively converged on maximal accuracy in well-behaved genomic contexts. However, distinctions between technologies became more evident in complex or repetitive regions [36, 38], where sequencing chemistry, read length, and algorithmic design play a larger role.

Within SRS data, subtle but reproducible platform-specific biases were observed: MGI performed slightly better in low-mappability and AT-rich regions, whereas ILMN showed superior accuracy in GC-rich loci and within the major histocompatibility complex (MHC), reflecting differences in library preparation and amplification chemistry. Among SRS pipelines, DRAGEN emerged as the top performer, achieving the highest overall F1-scores across both T2T and CMRG (SNVs ~ 99.9%, indels ~ 97.9%) and demonstrating exceptional precision-recall balance. DRAGEN's advantage stems from its FPGA-accelerated alignment engine, dynamic Smith-Waterman rescoring, and adaptive local realignment, which collectively improve indel placement and variant resolution in GC-rich or repetitive regions to outperform conventional software-based approaches such as GATK4 which use traditional aligners [13, 46].

By contrast, LRS technologies displayed more variability in difficult genomic contexts, more specifically in the ONT platform in regions associated with homopolymers and other repeat regions [36, 42]. DeepVariant, when paired with newer LRS platforms such as Revio and R10, delivered the highest, by platform, accuracy across both variant types, maintaining strong F1-scores not only in easy regions but for difficult regions such as in low-mappability, segmentally duplicated, and high-GC regions (Revio SNVs: 96–99%, indels: 93–97%). For earlier LRS chemistries, such as Sequel and R9, Clair3 was the top-performing caller (Sequel SNVs ~ 97%, indels 75–80%; R9 SNVs ~ 88%, indels < 40%), though performance remained below that of SRS or modern LRS platforms. Collectively, these findings indicate that while SRS remains a gold standard for precision and reproducibility, recent advances in LRS chemistry and neural network-based calling have narrowed, and in many challenging regions, surpassed the historical accuracy gap between SRS and LRS technologies [49].

The superior performance of LRS in complex genomic regions can be attributed to several interrelated factors. Foremost, the extended read lengths of PacBio and ONT technologies reduce mapping ambiguity by spanning repetitive elements [5], structural polymorphisms, and segmental duplications that confound short-read aligners [53]. This increased contiguity enables more accurate haplotype reconstruction and variant phasing, particularly within low-mappability and high-GC contexts where short reads often fail to anchor uniquely. Moreover, although out of scope for this paper, recent advances in haplotype-aware variant callers such as DeepVariant [10, 48], trained using the Human Pangenome Reference Consortium (HPRC) assemblies to utilize linear or pangenome alignments, have further enhanced accuracy in these regions by leveraging graph-based representations that better capture population-level allelic diversity which could conceivably further improve the results shown here.

In parallel, short-read variant calling has also evolved toward a more haplotype-aware framework. The latest DRAGEN versions now implement a multigenome mapping strategy, which aligns reads against multiple reference backbones simultaneously, effectively mitigating reference bias and improving variant resolution in polymorphic or duplicated regions such as the MHC and segmental duplications [46]. This innovation narrows the traditional disadvantage of SRS in these contexts, particularly for clinically relevant regions with complex allelic diversity. Nonetheless, even with these advances, LRS technologies maintain a distinct advantage due to their ability to directly sequence through complex haplotypes and resolve complex indels without reliance on reference-based inference. Together, the convergence of long-read data quality, haplotype-aware algorithms, and pangenome-informed models signals a shift toward more complete and unbiased variant discovery across the full spectrum of genomic complexity.

Structural variant detection

The data demonstrates that LRS platforms markedly outperform SRS and synthetic approaches in SV detection, both in discovery and genotyping accuracy. Across independent benchmarks, LRS achieved substantially higher F1-scores for DEL and INS, with the largest gains observed for INS, where discovery performance was roughly threefold greater than SRS. This outcome aligns with prior observations that short-read approaches miss a large proportion of SVs, particularly insertions and complex events [54–56]. These findings reinforce the growing consensus that the extended read lengths, improved contiguity, and reduced mapping ambiguity of LRS provide a more comprehensive and accurate representation of the structural landscape of the human genome.

Size-stratified benchmarking further revealed that the advantage of LRS extends across the full spectrum of SV sizes. Most detected events fell within the 50 to 500 bp range, consistent with the expected abundance of Alu and small retrotransposon-derived variants, while detection frequency declined with increasing variant length. LRS maintained high and stable F1-scores across all size bins, from small (30 bp) to very large (145 kb +) events, reflecting a uniform ability to resolve both localized and large-scale genomic rearrangements. The relative advantage of LRS became increasingly pronounced with variant size, particularly for large insertions (> 500 bp), where LRS detected up to 140-fold more events than SRS in CMRG. These results emphasize that read length and

long-range contiguity remain critical determinants of SV detection sensitivity across genomic contexts.

Within LRS platforms, consistent performance gains were associated with both chemistry improvements and algorithmic innovations. PacBio platforms, particularly Revio, outperformed ONT across variant types and datasets, underscoring the advantage of highly accurate HiFi reads for both variant discovery and genotyping. Nonetheless, the newest R10 chemistry substantially closed the gap with PacBio, reflecting the cumulative impact of refined basecalling, pore design, and signal modelling. These improvements translated into more confident variant detection even in clinically relevant regions (CMRG), suggesting that LRS is now reaching a level of robustness suitable for clinical and translational applications. Moreover, the consistent performance of Revio and R10 across SV size ranges highlights the growing maturity of LRS pipelines, with recent chemistries narrowing the gap between discovery and genotyping accuracy. Caller performance remained dependent on variant type, sequencing technology, and genomic context. Among PacBio callers, SVIM and Sawfish consistently achieved the highest F1-scores, whereas sniffles2 and cutesv2 were the top-performing callers for ONT data. Performance patterns also shifted by SV size: for large insertions and deletions, graph-based and local haplotype-aware callers such as sawfish provided measurable accuracy gains, underscoring the value of haplotype phasing and graph alignment in resolving complex events [19].

Although LRS technologies clearly dominated in SV discovery, SRS pipelines demonstrated competitive genotyping performance, particularly for deletions. Among SRS callers, DRAGEN and dysgu performed best in variant discovery, while DRAGEN and MetaSV achieved the highest genotyping accuracy. These results suggest that SRS data, while limited in capturing the full diversity of SVs, remain valuable for high-throughput genotyping once variant catalogs are established. Hybrid approaches that integrate LRS-based discovery with SRS-based population genotyping could therefore provide a cost efficient framework for population-scale studies [57, 58].

The inclusion of synthetic 10X Genomics data further highlighted the limitations of linked-read technologies for comprehensive SV detection. LongRanger recovered a small subset of large (>6 kb) DEL, likely reflecting barcode-linked scaffolding rather than genuine sensitivity improvements, while performance for insertions and small events remained well below that of both SRS and LRS. These results confirm that synthetic technologies have limited utility for current SV discovery applications and have largely been superseded by LRS.

Finally, the size-stratified analysis underscores the importance of reporting SV performance across variant length classes rather than relying on global F1-summaries, which can obscure technology- or algorithm-specific biases. Most contemporary LRS pipelines now achieve near-saturation accuracy (from 30 bp to 145 kb +) spanning the majority of human structural diversity whereas SRS and synthetic approaches exhibit pronounced blind spots beyond a few hundred base pairs. Continued reporting of size-resolved benchmarking metrics will be essential for accurately evaluating progress and guiding technology adoption in both research and clinical genomics. Additionally, no single algorithm was universally optimal, indicating that each caller may leverage technology-specific error profiles or alignment properties differently. The differences observed between

the T2T and CMRG datasets also underscore the influence of benchmark composition, such as repeat content, GC bias, and local sequence complexity, on caller behavior.

Optimal depth for variant detection

The downsampling analysis revealed distinct platform- and variant-specific trends in coverage requirements and detection performance across small and structural variants. For SNVs and indels, recall and precision improved predictably with increasing depth, reaching platform-specific saturation points, with heterozygous variants generally requiring 1.5–2 × higher coverage than homozygous counterparts, similar to previous studies [49, 59, 60]. In easy-to-map regions, PacBio platforms and R10 achieved near-complete SNV detection (≥ 99% recall and precision) at modest depths of ~ 10–20 ×, while SRS platforms and R9 required higher coverage (~ 20–35 ×) to reach comparable accuracy. For easy indels, PacBio consistently outperformed all other technologies, attaining high recall and precision at moderate depths (~ 15–25 ×), with SRS platforms requiring higher coverage (30–40 ×) than SNVs and R10 required up to 60 × and R9 remained limited in both metrics. These differences were amplified in difficult-to-map regions, where LRS demonstrated clear advantages in recall, precision, and coverage efficiency. PacBio maintained near-maximal accuracy for both SNVs and indels at 20–30 × coverage, while ONT required 35–60 × and SRS exceeded 60 × to approach similar recall. Across benchmarks, the CMRG dataset, enriched for clinically relevant, complex loci, required approximately 1.5–2 × higher coverage than T2T to achieve comparable recall and precision, highlighting the impact of genomic context on platform-specific saturation.

Structural variant discovery followed parallel patterns. Long-read sequencing achieved near-maximal F1 for DEL and INS at lower coverage than SRS. PacBio platforms showed a sharp increase in accuracy 5 × to 15 × [49] and saturated at ~ 30–35 × in T2T, with Revio reaching stable > 94% mean F1 faster than Sequel. R10 required similar coverage, whereas R9 lagged at ~ 40 ×. Short reads required up to 80 × to approach plateaus, with INS remaining poorly captured (< 30% mean F1). Across CMRG, coverage requirements increased: PacBio reached maximal F1 at 35–45 ×, while SRS often exceeded 70 ×.

Collectively, these results demonstrate that LRS platforms provide efficient, robust detection of both small and structural variants across genomic contexts, with predictable saturation and minimal sensitivity to benchmark complexity. SRS technologies, in contrast, require substantially higher coverage to approach comparable performance and remain constrained in complex genomic regions. These findings underscore the advantages of long contiguity, haplotype-aware alignment, and improved mappability in achieving high-confidence variant discovery for both research and clinical applications.

Clinical and research implications

For genomic studies, optimal platform and algorithm selection depends on the application context (Table 2). For research-focused, genome-wide small variant discovery, PacBio Revio paired with DeepVariant or hybrid LRS/SRS models provides maximal coverage and accuracy across both easy and difficult regions, leveraging long-read contiguity and haplotype-aware calling. R10 is a strong alternative for complex loci, particularly for large indels and low-mappability regions. For clinical applications, SRS platforms (ILMN, MGI)

remain highly competitive for coding and moderately complex loci, with DRAGEN offering superior precision, reproducibility, and speed while LRS platforms (Revio, R10) are increasingly viable for challenging clinically relevant regions such as the MHC or segmental duplications. In structural variant detection, LRS dominates discovery across the full size spectrum, with PacBio callers (SVIM, Sawfish) and ONT callers (Sniffles2, CuteSV2) optimized for large or complex events. SRS remains suitable for targeted genotyping of known SVs but cannot capture the full structural landscape. Coverage requirements reflect these distinctions, with LRS achieving saturation at moderate depths (20–45 ×), whereas SRS often requires substantially higher coverage (> 60 ×) to approach comparable performance. Together, these findings emphasize a scenario-driven strategy: leverage LRS for comprehensive discovery and complex regions, and SRS for high-throughput genotyping or coding-focused clinical applications.

Conclusion

This study provides a comprehensive evaluation of sequencing technologies and variant calling algorithms for accurately detecting SNVs, indels, and SVs. Our findings delineate the strengths and limitations of each approach, guiding optimal technology and algorithm choices for distinct genomic contexts and research objectives. While SRS excels in identifying small variants within well-mapped regions, LRS, particularly PacBio, delivers superior accuracy for SV detection and complex or repetitive regions.

Future efforts should expand benchmarking efforts to include phased variant accuracy, leveraging the phased T2T HG002 assembly and tools such as vcfdist [61] to systematically assess haplotype fidelity across LRS and SRS datasets for both small variants and SVs. A critical next step is reducing reference bias [31], which disproportionately affects repetitive and polymorphic regions when alignments and variant calls are constrained to a single linear reference genome. This bias can obscure population-specific alleles and misrepresent complex loci, leading to under-calling of insertions, paralogous variants, or divergent haplotypes.

Pangenome-based frameworks offer a promising solution by incorporating multiple haplotypes and structural configurations into graph or multigenome references. Approaches such as DRAGEN's multigenome mapping and DeepVariant's integration of the HPRC pangenome provide examples of how graph-aware or multi-reference alignment can mitigate reference bias, improve read placement, and enhance variant detection across diverse ancestries and structurally complex regions. Continued improvements in ONT R10.4.1 and PacBio Revio SPRQ chemistries, combined with the development of more accurate basecallers, indel- and haplotype-aware variant callers, will further elevate precision in both small variant and SV discovery. Together, phasing-aware benchmarking, reduced reference bias via pangenome models, and next-generation sequencing and algorithmic advances will more accurately capture the full spectrum of human genetic variation, deepening insights into genome biology, disease mechanisms, and evolutionary diversity.

Data availability

All raw alignment files generated in this study are available through European Nucleotide Archive (ENA) under project accession PRJEB109257 [62]. SNV/indel and structural variant (SV) VCF files, along with their corresponding benchmarking outputs

generated using hap.py and truvari, are available on Zenodo at: https://zenodo.org/records/18868532 under the MIT license [63].

In addition, the processed small variant benchmarking results are publicly accessible through an interactive dashboard at: https://varbench-dashboard.c3g-app.sd4h.ca. This resource will be periodically updated as new software versions and additional variant callers become available, with future development planned to incorporate structural variant (SV) benchmarking results.

Materials and methods

Library preparation and sequencing

A large batch of cells expanded from one original vial directly obtained from the Coriell Institute for Medical Research cell line repository was used for DNA isolation and subsequent genomic analysis using four different technology platforms: Illumina HiSeq Xten (ILMN), MGI DNBSEQ-400 (MGI), Oxford Nanopore Technologies R9 (ONT1) and R10 (ONT3 and ONT4), Pacific Biosciences (PacBio) Revio (PB3) and 10 × genomics (sequenced on Illumina HiSeq Xten; 10X).

Additional LRS data was downloaded directly from Genome in a Bottle (GIAB) for additional analyses. Two PacBio HiFi datasets: PacBio CCS 15 kb (PB1) and PacBio CCS 15 kb and 20 kb chemistry 2 (PB2) and one Oxford Nanopore Technologies R9 (ONT2).

Cell line culture

GM24385 (HG002), GM24149 (HG003), and GM24143 (HG004) cells (Coriell Institute) were obtained directly from Coriell Institute and cultured in RPMI medium supplemented with 2 mM L-alanyl-L-glutamine, 15% fetal bovine serum, 50 U/ml penicillin, and 50 µg/ml streptomycin (all from ThermoFisher Scientific, Nepean, ON, Canada) at 37 °C and 5% CO_2, according to the Coriell protocol. Cells were seeded at a density of 500,000 cells/ml and passaged or harvested when they reached 1 M cells/ml.

DNA extraction

Cells were extracted on the SageHLS High Molecular Weight DNA Library System (Sage Science, Inc., Beverly, Massachusetts, United States) which aims to minimize pipetting steps that shear very HMW DNA by doing extraction and fractionation of DNA based in the same cassette. Cells in suspension are lysed directly in the cassette and an enzyme is added to digest the DNA just enough so that it can be electrophoresed and separated by size. This digestion step is crucial, as overshearing will greatly reduce the yield of HMW fragments. A second electrophoresis is then done perpendicular to the first, in order to collect the DNA in 6 different elution wells.

PCRfree-whole genome sequencing library preparation

Illumina

The Ashkenazi trio genomic DNA (gDNA) quality was assessed prior to library preparation to ensure high-quality input material. DNA integrity was evaluated using either agarose gel electrophoresis or the Agilent Tapestation, while quantification was performed using the Qubit DNA High Sensitivity (HS) assay. DNA purity was assessed by measuring the A260/280 ratio, with acceptable values ranging between 1.8 and 2.0. Whole-genome

sequencing (WGS) libraries were prepared using the Illumina TruSeq PCR-Free Library Prep Kit, with 500 ng of high-quality gDNA as input. DNA was fragmented to a target size of 300 bp using the Covaris LE220 focused ultrasonicator, followed by size selection with Ampure XP beads (Beckman Coulter). Library quality control (QC) was performed to ensure appropriate fragment size and concentration before sequencing. Fragment size distribution was assessed using the Agilent Bioanalyzer DNA High Sensitivity assay, while library quantification was performed using the KAPA qPCR Library Quantification Kit. Sequencing was carried out on the Illumina HiSeq X platform with paired-end 150 bp reads (2 × 150), on one lane with 1% PhiX as an internal control.

MGI

The samples were done using the WGS Lucigen (NxSeq AmpFREE low DNA Library Kit) with the starting input of 1000 ng.The Genomic DNA was mechanically sheared using COVARIS at 300 bp. Libraries were quantified using the KAPA Library Quant Kit (Illumina/Universal qPCR) from Roche for qPCR quantification. Average size fragment was determined using a LabChip GXII (PerkinElmer) instrument. The libraries were converted to MGI using the MGIEasy Universal Library Conversion Kit (App-A) using 50 ng of each sample with 5-cycles of amplification. The converted libraries were quantified using Qubit™ ssDNA Assay Kit and pooled at equimolar. The pool was loaded at 60 fmol on a DNBSEQ-G400 as per the manufacturer's recommendations. The run was performed for 2 × 150 cycles (paired-end mode).

10X Genomics

The DNA extractions for HG002 and HG004 were prepared following the SageHLS High Molecular Weight DNA Extraction protocol using a 1/800 dilution of NEBNext® dsDNA Fragmentase® (cat# M0348) for a half hour digestion. DNA from HG003 was prep testing an alternative protocol which uses a Cas9 system (Cas9 Nuclease, *S. pyogenes*, New England BioLabs, Inc., cat# M0386) with two simple guide RNAs (a 10 CA repeat guide and a 10 AT repeat guide, Alt-R® CRISPR-Cas9 crRNA, Integrated DNA Technologies, Inc., Coralville, Iowa, United States). DNA yield from each elution well was measured by Qubit™ dsDNA BR Assay Kit (ThermoFisher Scientific, cat# Q32853) and molecule length was assessed by Femto Pulse (Genomic DNA 165 kb Kit, 3 h run, Agilent Technologies, Inc., Santa Clara, California, United States, cat# FP-1002–0275).

Nanopore

R9

Genomic DNA was manually purified from whole blood using the Qiagen protocol for high-molecular-weight (HMW) DNA (Qiagen, Toronto, ON, Canada). Briefly, cells were pelleted by centrifuging at 500 × g for 10 min at 4 °C, and the supernatant was removed, leaving approximately 200 µL with the pellet. The pellet was then resuspended in a digestion mixture containing 20 µL proteinase K, 4 µL RNase A, and 150 µL Buffer AL, vortexed, and incubated at room temperature for 30 min. Following a quick spin, DNA was bound to 15 µL MagAttract Suspension G by adding 280 µL Buffer MB and

incubating for 3 min at 1,400 rpm at room temperature in a thermoshaker. After a quick spin, the supernatant was removed after a 1-min incubation on a magnetic rack. The magnetic beads were then washed twice with 700 µL Buffer MW1 and twice with 700 µL Buffer PE, each wash involving a 1-min incubation at 1,400 rpm in the thermoshaker followed by 1 min on the magnetic rack and supernatant removal. The beads were further rinsed twice with 700 µL of water while on the magnetic rack, removing the supernatant after 1 min each time. Finally, the DNA was eluted by adding 100 µL Buffer AE to the beads (removed from the magnetic rack), incubating at room temperature for 3 min at 1,400 rpm, placing tubes on the magnetic rack, and recovering the supernatant into a new tube after 1 min. Purified DNA was stored at 4 °C, and its quality was assessed using a NanoDrop 8000 Spectrophotometer (ThermoFisher Scientific, Nepean, ON, Canada).

The Qiagen protocol for manual purification of high-molecular-weight (HMW) genomic DNA from whole blood was used as recommended by the manufacturer (Qiagen, Toronto, ON, Canada).

ONT1 library preparation was performed using 1.5 µg genomic DNA as starting material for the ligation sequencing kit (SQK-LSK109) as recommended by the manufacturer (Oxford Nanopore Technologies, Oxford, UK). The library was loaded onto a FLO-PRO002 R9.4.1 flow cell with 6052 available pores at the start of the sequencing run. Flow cell was refueled with 150 µl priming mix after ~ 45 h of sequencing. PromethION-Beta Release 19.05.1 was used with live high accuracy basecalling (Guppy v3.0.3).

Sequencing was primarily performed on the MinION using a mix of SQK-RAD003 and SQK-RAD004 library prep kits on the FLO-MIN106. Additional sequencing was performed using the GridION and PromethION. Reads were base-called with guppy (version 2.3.5), then combined into a single fastq file. The high accuracy basecalling algorithm was used for all but one flowcells (FAH03763), as the model config file was not available for that particular flowcell/sequencing kit combination (flowcell—FLO-MIN107 and kit—SQK-RAD003).

The ONT2 “ultralong" (Jain, et al. 2018) libraries were generated on GIAB Ashkenazi son cell line HG002 (GM24385) to generate a library with 52 × total coverage with approximately 15 × coverage generated by reads larger than 100 kb. DNA was prepared using various modified versions of Josh Quick’s protocol (https://doi.org/10.17504/protocols.io.mrxc57n). A manuscript with a detailed protocol and methods development is in prep. Sequencing was primarily performed on the MinION using a mix of SQK-RAD003 and SQK-RAD004 library prep kits on the FLO-MIN106 with additional sequencing performed on GridION and PromethION. Reads were base-called with guppy (version 3.4.5), then combined into a single fastq file. The hac algorithm was used for all but one flowcells (FAH03763), as the model config file was not available for that particular flowcell/sequencing kit combination (flowcell—FLO-MIN107 and kit—SQK-RAD003).

R10

High-molecular-weight genomic DNA was extracted using the Qiagen MagAttract HMW DNA kit following the manufacturer’s protocol. DNA was sheared using a Megaruptor system (Diagenode) with a target concentration of 40 ng/µL in a 100 µL volume at speed setting 31. Size selection was performed using the SRE XS 10 kb protocol to

enrich for fragments ≥ 10 kb. Libraries were prepared with the Oxford Nanopore Technologies (ONT) Ligation Sequencing Kit SQK-LSK114 and loaded onto a single R10.4.1 flow cell (FLO-PRO114M) on a PromethION 48 (PRO-SEQ048) instrument. Sequencing was performed using MinKNOW v24.06.10 with at least 6,500 active pores at the time of loading. Raw signal data were basecalled using Dorado v7.4.12 with the super-accurate (SUP) basecalling model v4.3.0 (400 bps), enabling modified base detection for 5-hydroxymethylcytosine (5hmC) and 5-methylcytosine (5mC) in all sequence contexts.

PacBio

Sequel

High-fidelity 15 kb long read dataset of HG002 (PB1) was generated by the GIAB consortium using using Binding/Sequencing chemistry 3.0 and SMRT Link 6.0. One shotgun library was prepared from a DNA sample extracted from a large homogenized growth of B-lymphoblastoid cell lines from the Coriell Institute for Medical Research. Size selection was performed, targeting a narrow size band ~ 15 kb using the sageELF DNA size-selection system from SAGE Science. Circular consensus sequences were calculated using default parameters and a predicted read accuracy filter of Q20 (99%) using SMRT Link 6.0. The library was sequenced to approximately 28 × coverage.

High-fidelity combined 15 kb and 20 kb long read dataset of HG002 (PB2) was generated by the GIAB consortium using using Sequel II System with 2.0 chemistry and SMRT Link 9.0. Two shotgun library were prepared from a DNA sample extracted from a large homogenized growth of B-lymphoblastoid cell lines from the Coriell Institute for Medical Research the first with TPK 1.0 library prep and SageELF Fraction 4 (target: 15 kb) size selection and the second with SMRTbell Express 2.0 + Enzyme Clean Up and SageELF Fraction 2 (target: 20 kb) size selection. Circular consensus sequences were calculated using default parameters and a predicted read accuracy filter of Q20 (99%) using SMRT Link 9.0. The 15 kb library consisting of four readsets was sequenced to approximately 36 × coverage and the 20 kb library with two readset totaling 16 × coverage.

Revio

High-molecular-weight genomic DNA (5 µg) was sheared using a Megaruptor system (Diagenode) at speed setting 31 in a final volume of 130 µL (38.46 ng/µL). Library preparation was performed following the PacBio protocol "Preparing Whole Genome and Metagenome Libraries Using the SMRTbell® Prep Kit 3.0" (PN 102–166–600, REV03, March 2024) using the SMRTbell Prep Kit 3.0. Size selection was conducted with the BluePippin system (Sage Science) using 0.75% agarose gel cassettes and Marker S1 (BLF7510), configured for high-pass selection between 10 and 50 kb (Cassette Definition: 0.75% DF Marker S1 High-Pass 6–10 kb v3). Sequencing was carried out on the PacBio Revio platform, and basecalling was performed using SMRT Link v25.1.0.257715.

Bioinformatics methods

Experimental design

This study aimed to evaluate the performance of various sequencing technologies and variant calling algorithms in detecting germline SNPs, InDels, and SVs in human genomes. SRS platforms, including Illumina and MGI, LRS technologies such as PacBio:

Revio, Sequel II and ONT: R10 and R9, and synthetic 10X genomics, were used to generate to compare a full, representative sequencing outputs from each platforms. All reads were linearly aligned to the human genome build GRCh37 decoy (hs37d5) to ensure consistency across all truth datasets. A range of variant callers were applied, some optimized for specific sequencing technologies, to assess their ability to accurately detect variants. Downsampling experiments were conducted to further evaluate the effects of sequencing depth on variant detection.

Data processing

SRS data processing

SRS data from Illumina and MGI platforms were processed using GenPipes version 4.5.0, a bioinformatics pipeline based on GATK Best Practices. Quality control of raw reads was performed using Skewer (0.2.2), followed by alignment to the GRCh37 decoy (hs37d5) reference genome using the BWA-MEM algorithm (7.15). Post-alignment processing included marking duplicate reads using Picard (2.9.0) and performing base quality score recalibration with GATK (v3.8). Parallel alignments were generated for comparison using Parabrick (2.4.0-rc3), DRAGEN (07.031.732.4.3.13; linear model), and MegaBolt (2.2.2.1), with each tool run using default settings.

LRS data processing

LRS data generated from PacBio Sequel II (PB1), PacBio Revio (PB2), ONT R9.4 (ONT1) and R10.4 (ONT2) platforms were aligned to the GRCh37 decoy (hs37d5) genome using technology-specific aligners. ONT reads were aligned using Minimap2 (2.26) with parameters optimized for contiguity (-a -z 600,200 -x map-ont), while PacBio long reads were aligned using pbmm2 (1.13.1) with the circular consensus sequencing (CCS) preset for Sequel II samples and HiFi preset for Revio samples. Datasets generated both in-house and from public repositories were processed with the same protocols to ensure consistency across samples.

Downsampling

To evaluate the effects of sequencing depth on variant detection, downsampling experiments were performed on both SRS and LRS datasets. Illumina data for the Ashkenazim son (GM24385/HG002) were combined to achieve mean coverage levels of 126x, and then downsampled to 11 individual coverage levels (ranging from 120 × to 2x) using a custom Perl script to randomly select read pairs. The MGI HG002 sample was similarly downsampled into nine coverage levels (from 40 × to 2x). For LRS datasets, PacBio Sequel II (PB2), PacBio Revio (PB4), ONT R9.4 (ONT2) and ONT R10.4.1 (ONT4, 61x) data were downsampled into 12, 8, 11 and 13 samples, respectively, using rasusa (v2.1.0) which accounted for read-length as part of its downsampling calculation (Table S1).

Data analysis

Germline variant calling

Germline variants, including SNVs and Indel, were called across the SRS datasets using seven variant calling algorithms: Clair3 (v1.2.0), DeepVariant (DV; v1.9.0; with small model disabled), DRAGEN, GATK HaplotypeCaller (v3.8 and v4.1.2.0), Parabrick, and MegaBolt. These tools were run with default settings to ensure a consistent comparison across platforms. For the LRS datasets, germline variants were identified using Clair3, DeepVariant, and GATK (v4.1.2.0) for PacBio data, and Clair3 (R9.4: ont model and R10.4.1: dna_r10.4.1_e8.2_400bps_sup@v4.3.0), Pepper-Margin-DeepVariant (v0.8 with DV v1.2.0; noted as DeepVariant in results figures) for R9.4, DeepVariant for R10.4.1, and Nanocaller (v2.0.0) for ONT data. All variant callers were executed with default parameters unless otherwise specified.

Structural variant detection

SVs were identified using distinct workflows for SRS and LRS data. For the SRS datasets, structural variants were detected using six SV callers—delly, lumpy, manta, WHAM, breakseq2, and cnvkit—integrated through the MetaSV (v0.5.4) ensemble approach. MetaSV was configured to enhance insertion detection with the –boost_ins option and improve breakpoint resolution. Further filtering was applied using DupHold (v0.2.1), with a fold-change threshold (DHFFC) of 0.7 to minimize false positives for deletions and duplications. Additionally, DRAGEN's SV pipeline (v4.3.13) and Dysgu (v1.6.1) was applied to the SRS datasets. For PacBio and ONT LRS datasets, four SV callers were applied to both pathforms-CuteSV (v2.0.3), Dysgu, Sniffles (v2.2.2), and SVIM (v2.0.0) with pbsv (2.9.0), Sawfish (v2.0.5) applied to only PacBio with a minimum SV size of 30 base pairs.

Germline variant assessment

The precision, sensitivity, and F1-scores of germline SNP and InDel calls were evaluated using the hap.py (v0.3.15) algorithm and RTG vcfeval (v3.12.1). Variant calls from each sequencing technology and caller were compared against the Telomere-2-Telomere (T2T-Q100 v1.1), and Clinically Relevant Molecular Genetics (CMRG) v1.0.0 truth sets for HG002. Default parameters were used for CMRG, with T2T having the additional argument of –gender male set to correctly call chromosome X as suggested by GIAB authors. This analysis provided a comprehensive assessment of variant detection accuracy across technologies, including comparisons of individual variant callers.

Genome stratification

Genome stratification was performed using hap.py (v0.3.15) and RTG vcfeval (v3.12.1) with the –stratification option enabled. We applied the pre-defined GIAB/GA4GH genome stratifications (v3.6), which include: easy, difficult, RefSeq coding (refseq_cds), non-coding (not_in_refseq_cds), MHC, low-mappability, non–low-mappability, segmental duplications (segdups), non-segmental duplications, tandem repeats

and homopolymers (all_tr_and_homopolymers), homopolymer length bins (4–6 bp, 7–11 bp, ≥ 12 bp, ≥ 21 bp), satellites, non-satellites, and GC-content bins (≤ 15%, 15–20%, 20–25%, 25–30%, 30–55%, 55–60%, 60–65%, 65–70%, 70–75%, 75–80%, 80–85%, ≥ 85%, ≤ 25% or ≥ 65%, and ≤ 30% or ≥ 55%). Stratification regions were subset against the high-confidence regions of each validation truth set using bedtools (v2.30.0) intersect.

Structural variant benchmarking

The performance of SV detection was benchmarked using the truvari (v4.3.1) on the two truth sets: the truth set for genome-wide assessment T2T and the clinical SVs from the CMRG dataset. The T2T truth set included 20,767 deletions and 29,640 insertions, all over 30 bp in length and the CMRG subset encompassed 89 high-quality deletions and 105 insertions located near 273 known clinical genes. Benchmarking metrics, including precision, recall, and F1-scores, were generated for each variant file from various sequencing technology, which were processed by bcftools (v2.30) to retain SVs with the caller specific PASS FILTER with alternative contigs removed. Truvari assessed the filtered variant using the bench arguments "–sizemin 50 –sizefilt 30 –sizemax 150 k -r 2000 -C 5000 –pctseq 0 –pick ac –dup-to-ins" with the refine arguments of "–recount –use-region-coords –use-original-vcfs –align mafft" applied to only the T2T set. SV size stratification was applied by setting the corresponding –sizemin and –sizefilt arguments to the lower bound and –sizemax to the upper bound. SVs were categorized into the following size ranges: 30–50 bp (T2T-only), 51–250 bp, 251–500 bp, 501–6000 bp, and larger than 6001 bp, based on known repeat sizes from GIAB.

Genotype performance

Genotype accuracy was assessed by calculating genotype F1-scores for each caller and technology. Additional stratification of results based on SV size allowed for detailed performance comparisons, enabling insight into the genotype calling ability across various types and sizes of SVs.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-026-04048-4.

Additional file 1: Fig. S1. F1-score performance across technologies and genomic contexts.

Additional file 2: Table S1. Coverage, Insert Size, and Read Length Metrics for HG002 Table S2. Variant Counts in T2T and CMRG Validation Sets. Table S3. Precision, Recall, and F1-Score Performance Across Sequencing Technologies. Table S4. Computational Resource Usage for Short-Read Sequencing Technologies. Table S5. Stratification of Structural Variant Counts by Size. Table S6. Discovery Metrics for Structural Variants Across Technologies. Table S7. Genotyping Metrics for Structural Variants Across Technologies. Table S8. Comparison of Mean Discovery and Genotyping F1 Scores Across Technologies. Table S9. Comparison of True Positive Counts Across Sequencing Technologies.

Acknowledgements

We would like to thank Dr. Albena Pramatarova for providing the cells. We would also like to thank Dr. Pierre Berube, Haig H.V. Djambazian, and Geneviève Geneau for the library preparations, and Shu-Huang Chen and Julio Serna Vasquez for excellent technical assistance.

Peer review information

Fritz Sedlazeck, Andrew Cosgrove and Wenjing She served as the primary editors for this article and oversaw its editorial process and peer review in collaboration with the editorial team. Peer reviewer reports are available in the online version of this article.

Authors' contributions
RE, MB, JR, and GB conceived and designed the study. SR and JR collected and sequenced samples. RE and GB wrote the main manuscript text with contributions from all authors. RE prepared all figures. All authors reviewed the manuscript.

Funding
This work was supported by a Canada Institute of Health Research (CIHR) project grant (PJT-191707). G.B. is supported by a Canada Research Chair Tier 1 award and a FRQ-S Distinguished Research Scholar award. The Canadian Centre for Computational Genomics (C3G) was supported by a Genome Canada Genome Technology Platform grant for JR. This research was also enabled in part by support provided by Calcul Québec and the Digital Research Alliance of Canada.

Data availability
All raw alignment files generated in this study are available through European Nucleotide Archive (ENA) under project accession PRJEB109257 [62]. SNV/indel and structural variant (SV) VCF files, along with their corresponding benchmarking outputs generated using hap.py and truvari, are available on Zenodo at: (https://zenodo.org/records/18868532) under the MIT license [63]. In addition, the processed small variant benchmarking results are publicly accessible through an interactive dashboard at: (https://varbench-dashboard.c3g-app.sd4h.ca). This resource will be periodically updated as new software versions and additional variant callers become available, with future development planned to incorporate structural variant (SV) benchmarking results.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 10 June 2025 Accepted: 17 March 2026
Published online: 08 April 2026

References

1. Rodriguez R, Krishnan Y. Genesis of next-generation sequencing. Nat Biotechnol. 2023;41:1709. https://doi.org/10.1038/s41587-023-01986-3.
2. Porreca GJ. Genome sequencing on nanoballs. Nat Biotechnol. 2010;28:43–4. https://doi.org/10.1038/nbt0110-43.
3. Tørresen OK, Star B, Mier P, Andrade-Navarro MA, Bateman A, Jarnot P, et al. Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. Nucleic Acids Res. 2019;47:10994. https://doi.org/10.1093/nar/gkz841.
4. Espinosa E, Bautista R, Larrosa R, Plata O. Advancements in long-read genome sequencing technologies and algorithms. Genomics. 2024;116:110842. https://doi.org/10.1016/j.ygeno.2024.110842.
5. Marks P, Garcia S, Barrio AM, Belhocine K, Bernate J, Bharadwaj R, et al. Resolving the full spectrum of human genome variation using Linked-Reads. Genome Res. 2019;29:635–45. https://doi.org/10.1101/gr.234443.118.
6. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-time DNA sequencing from single polymerase molecules. Science. 2009;323:133–8. https://doi.org/10.1126/science.1162986.
7. Kasianowicz JJ, Brandin E, Branton D, Deamer DW. Characterization of individual polynucleotide molecules using a membrane channel. Proc Natl Acad Sci USA. 1996;93:13770–3. https://doi.org/10.1073/pnas.93.24.13770.
8. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297. https://doi.org/10.1101/gr.107524.110.
9. Zheng Z, Li S, Su J, Leung AW-S, Lam T-W, Luo R. Symphonizing pileup and full-alignment for deep learning-based long-read variant calling. Preprint at bioRxiv. https://doi.org/10.1101/2021.12.29.474431. 2022.
10. Poplin R, Chang P-C, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol. 2018;36:983–7. https://doi.org/10.1038/nbt.4235.
11. Miller NA, Farrow EG, Gibson M, Willig LK, Twist G, Yoo B, et al. A 26-hour system of highly sensitive whole genome sequencing for emergency management of genetic diseases. Genome Med. 2015;7:100. https://doi.org/10.1186/s13073-015-0221-8.
12. Carpi G, Gorenstein L, Harkins TT, Samadi M, Vats P. A GPU-accelerated compute framework for pathogen genomic variant identification to aid genomic epidemiology of infectious disease: a malaria case study. Brief Bioinform. 2022;23:bbac314. https://doi.org/10.1093/bib/bbac314.
13. Li Z, Xie Y, Zeng W, Huang Y, Gu S, Gao Y, et al. An efficient large-scale whole-genome sequencing analyses practice with an average daily analysis of 100Tbp: zbolt. Clin Transl Discov. 2023;3:e252. https://doi.org/10.1002/ctd2.252.
14. Mohiyuddin M, Mu JC, Li J, Bani Asadi N, Gerstein MB, Abyzov A, et al. MetaSV: an accurate and integrative structural-variant caller for next generation sequencing. Bioinformatics. 2015;31:2741–4. https://doi.org/10.1093/bioinformatics/btv204.
15. Cleal K, Baird DM. Dysgu: efficient structural variant calling using short or long reads. Nucleic Acids Res. 2022;50:e53. https://doi.org/10.1093/nar/gkac039.

16. Jiang T, Liu Y, Jiang Y, Li J, Gao Y, Cui Z, et al. Long-read-based human genomic structural variation detection with cuteSV. Genome Biol. 2020;21:189. https://doi.org/10.1186/s13059-020-02107-y.
17. Ahsan MU, Liu Q, Fang L, Wang K. 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. 2021;22:261. https://doi.org/10.1186/s13059-021-02472-2.
18. Huddleston J, Chaisson MJP, Steinberg KM, Warren W, Hoekzema K, Gordon D, et al. Discovery and genotyping of structural variation from long-read haploid genome sequence data. Genome Res. 2017;27:677–85. https://doi.org/10.1101/gr.214007.116.
19. Saunders CT, Holt JM, Baker DN, Lake JA, Belyeu JR, Kronenberg Z, et al. Sawfish: improving long-read structural variant discovery and genotyping with local haplotype modeling. Bioinformatics. 2025;41:btaf136. https://doi.org/10.1093/bioinformatics/btaf136.
20. Smolka M, Paulin LF, Grochowski CM, Horner DW, Mahmoud M, Behera S, et al. Detection of mosaic and population-level structural variants with sniffles2. Nat Biotechnol. 2024;42:1571–80. https://doi.org/10.1038/s41587-023-02024-y.
21. Heller D, Vingron M. SVIM: structural variant identification using mapped long reads. Bioinformatics. 2019;35:2907–15. https://doi.org/10.1093/bioinformatics/btz041.
22. Krusche P, Trigg L, Boutros PC, Mason CE, De La Vega FM, Moore BL, et al. Best practices for benchmarking germline small-variant calls in human genomes. Nat Biotechnol. 2019;37:555–60. https://doi.org/10.1038/s41587-019-0054-x.
23. Roy S, Coldren C, Karunamurthy A, Kip NS, Klee EW, Lincoln SE, et al. Standards and guidelines for validating next-generation sequencing bioinformatics pipelines: a joint recommendation of the association for molecular pathology and the College of American Pathologists. J Mol Diagn. 2018;20:4–27. https://doi.org/10.1016/j.jmoldx.2017.11.003.
24. Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W, et al. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nat Biotechnol. 2014;32:246–51. https://doi.org/10.1038/nbt.2835.
25. Fang LT, Zhu B, Zhao Y, Chen W, Yang Z, Kerrigan L, et al. Establishing community reference samples, data and call sets for benchmarking cancer mutation detection using whole-genome sequencing. Nat Biotechnol. 2021;39:1151–60. https://doi.org/10.1038/s41587-021-00993-6.
26. Olson ND, Wagner J, McDaniel J, Stephens SH, Westreich ST, Prasanna AG, et al. PrecisionFDA Truth Challenge V2: Calling variants from short and long reads in difficult-to-map regions. Cell Genomics. 2022. https://doi.org/10.1016/j.xgen.2022.100129.
27. Chapman LM, Spies N, Pai P, Lim CS, Carroll A, Narzisi G, et al. A crowdsourced set of curated structural variants for the human genome. PLoS Comput Biol. 2020;16:e1007933. https://doi.org/10.1371/journal.pcbi.1007933.
28. Zook JM, Hansen NF, Olson ND, Chapman L, Mullikin JC, Xiao C, et al. A robust benchmark for detection of germline large deletions and insertions. Nat Biotechnol. 2020;38:1347–55. https://doi.org/10.1038/s41587-020-0538-8.
29. English AC, Dolzhenko E, Ziaei Jam H, McKenzie SK, Olson ND, De Coster W, et al. Analysis and benchmarking of small and large genomic variants across tandem repeats. Nat Biotechnol. 2025;43:431–42. https://doi.org/10.1038/s41587-024-02225-z.
30. Wagner J, Olson ND, Harris L, McDaniel J, Cheng H, Fungtammasan A, et al. Curated variation benchmarks for challenging medically relevant autosomal genes. Nat Biotechnol. 2022;40:672–80. https://doi.org/10.1038/s41587-021-01158-1.
31. Hansen NF, Dwarshuis N, Ji HJ, Rhie A, Loucks H, Logsdon GA, Vollger MR, Storer JM, Kim J, Adam E, et al. A complete diploid human genome benchmark for personalized genomics. Preprint at bioRxiv. https://doi.org/10.1101/2025.09.21.677443. 2025.
32. Dwarshuis N, Kalra D, McDaniel J, Sanio P, Alvarez Jerez P, Jadhav B, et al. The GIAB genomic stratifications resource for human reference genomes. Nat Commun. 2024;15:9029. https://doi.org/10.1038/s41467-024-53260-y.
33. Cleary JG, Braithwaite R, Gaastra K, Hilbush BS, Inglis S, Irvine SA, Jackson A, Littin R, Rathod M, Ware D, et al. Comparing Variant Call Files for Performance Benchmarking of Next-Generation Sequencing Variant Calling Pipelines. Preprint at bioRxiv. https://doi.org/10.1101/023754. 2015.
34. English AC, Menon VK, Gibbs RA, Metcalf GA, Sedlazeck FJ. Truvari: refined structural variant comparison preserves allelic diversity. Genome Biol. 2022;23:271. https://doi.org/10.1186/s13059-022-02840-6.
35. Zook JM, Catoe D, McDaniel J, Vang L, Spies N, Sidow A, et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. Sci Data. 2016;3:160025. https://doi.org/10.1038/sdata.2016.25.
36. Wagner J, Olson ND, Harris L, Khan Z, Farek J, Mahmoud M, et al. Benchmarking challenging small variants with linked and long reads. Cell Genomics. 2022. https://doi.org/10.1016/j.xgen.2022.100128.
37. Kosugi S, Terao C. Comparative evaluation of SNVs, indels, and structural variations detected with short- and long-read sequencing data. Hum Genome Var. 2024;11:18. https://doi.org/10.1038/s41439-024-00276-x.
38. Nardone GG, Andrioletti V, Santin A, Morgan A, Spedicati B, Concas MP, et al. A hitchhiker guide to structural variant calling: a comprehensive benchmark through different sequencing technologies. Biomedicines. 2025;13:1949. https://doi.org/10.3390/biomedicines13081949.
39. Santos R, Lee H, Williams A, Baffour-Kyei A, Lee S-H, Troakes C, et al. Investigating the performance of Oxford Nanopore long-read sequencing with respect to Illumina microarrays and short-read sequencing. Int J Mol Sci. 2025;26:4492. https://doi.org/10.3390/ijms26104492.
40. Sanderson ND, Kapel N, Rodger G, Webster H, Lipworth S, Street TL, et al. Comparison of R9.4.1/Kit10 and R10/Kit12 Oxford Nanopore flowcells and chemistries in bacterial genome reconstruction. Microb Genom. 2023;9:000910. https://doi.org/10.1099/mgen.0.000910.
41. Sirén J, Monlong J, Chang X, Novak AM, Eizenga JM, Markello C, et al. Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. Science. 2021;374:abg8871. https://doi.org/10.1126/science.abg8871.

42. Majidian S, Agustinho DP, Chin C-S, Sedlazeck FJ, Mahmoud M. Genomic variant benchmark: if you cannot measure it, you cannot improve it. Genome Biol. 2023;24:221. https://doi.org/10.1186/s13059-023-03061-1.
43. Rautiainen M, Nurk S, Walenz BP, Logsdon GA, Porubsky D, Rhie A, et al. Telomere-to-telomere assembly of diploid chromosomes with verkko. Nat Biotechnol. 2023;41:1474–82. https://doi.org/10.1038/s41587-023-01662-6.
44. Bourgey M, Dali R, Eveleigh R, Chen KC, Letourneau L, Fillon J, et al. Genpipes: an open-source framework for distributed and scalable genomic analyses. Gigascience. 2019;8:giz037. https://doi.org/10.1093/gigascience/giz037.
45. O'Connell KA, Yosufzai ZB, Campbell RA, Lobb CJ, Engelken HT, Gorrell LM, et al. Accelerating genomic workflows using NVIDIA parabricks. BMC Bioinformatics. 2023;24:221. https://doi.org/10.1186/s12859-023-05292-2.
46. Behera S, Catreux S, Rossi M, Truong S, Huang Z, Ruehle M, et al. Comprehensive genome analysis and variant detection at scale using DRAGEN. Nat Biotechnol. 2025;43:1177–91. https://doi.org/10.1038/s41587-024-02382-1.
47. Li H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics. 2016;32:2103–10. https://doi.org/10.1093/bioinformatics/btw152.
48. Shafin K, Pesout T, Chang P-C, Nattestad M, Kolesnikov A, Goel S, et al. Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. Nat Methods. 2021;18:1322–32. https://doi.org/10.1038/s41592-021-01299-w.
49. Mahmoud M, Huang Y, Garimella K, Audano PA, Wan W, Prasad N, et al. Utility of long-read sequencing for All of Us. Nat Commun. 2024;15:837. https://doi.org/10.1038/s41467-024-44804-3.
50. Harvey WT, Ebert P, Ebler J, Audano PA, Munson KM, Hoekzema K, et al. Whole-genome long-read sequencing downsampling and its effect on variant-calling precision and recall. Genome Res. 2023;33:2029–40. https://doi.org/10.1101/gr.278070.123.
51. Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. Opportunities and challenges in long-read sequencing data analysis. Genome Biol. 2020;21:30. https://doi.org/10.1186/s13059-020-1935-5.
52. Gambardella G. Joint processing of long- and short-read sequencing data with deep learning improves variant calling. Cell Reports Methods. 2025. https://doi.org/10.1016/j.crmeth.2025.101107.
53. Chaisson MJP, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, et al. Resolving the complexity of the human genome using single-molecule sequencing. Nature. 2015;517:608–11. https://doi.org/10.1038/nature13907.
54. Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nat Commun. 2019;10:1784. https://doi.org/10.1038/s41467-018-08148-z.
55. Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. Science. 2021;372:eabf7117. https://doi.org/10.1126/science.abf7117.
56. Zhao X, Collins RL, Lee W-P, Weber AM, Jun Y, Zhu Q, et al. Expectations and blind spots for structural variation detection from long-read assemblies and short-read genome sequencing technologies. Am J Hum Genet. 2021;108:919–28. https://doi.org/10.1016/j.ajhg.2021.03.014.
57. Chen S, Krusche P, Dolzhenko E, Sherman RM, Petrovski R, Schlesinger F, et al. Paragraph: a graph-based structural variant genotyper for short-read sequence data. Genome Biol. 2019;20:291. https://doi.org/10.1186/s13059-019-1909-7.
58. Quan C, Lu H, Lu Y, Zhou G. Population-scale genotyping of structural variation in the era of long-read sequencing. Comput Struct Biotechnol J. 2022;20:2639–47. https://doi.org/10.1016/j.csbj.2022.05.047.
59. Parks M, Lambert D. Impacts of low coverage depths and post-mortem DNA damage on variant calling: a simulation study. BMC Genomics. 2015;16:19. https://doi.org/10.1186/s12864-015-1219-8.
60. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008;456:53–9. https://doi.org/10.1038/nature07517.
61. Dunn T, Zook JM, Holt JM, Narayanasamy S. Jointly benchmarking small and structural variant calls with vcfdist. Genome Biol. 2024;25:253. https://doi.org/10.1186/s13059-024-03394-5.
62. Robert JM, Eveleigh Sarah J, Reiling J, Hector Galvez, Mathieu Bourgey, Jiannis Ragoussis, Guillaume Bourque. Benchmarking of sequencing technologies defines optimal strategies for genetic variants detection in a human genome. European Nucleotide Archive (ENA). 2026. https://www.ebi.ac.uk/ena/browser/view/PRJEB109257.
63. Robert JM, Eveleigh Sarah J, Reiling J, Hector Galvez, Mathieu Bourgey, Jiannis Ragoussis, Guillaume Bourque. Benchmarking of sequencing technologies defines optimal strategies for genetic variants detection in a human genome. Zenodo. 2026. Version v1. https://doi.org/10.5281/zenodo.18868532.

Publisher's Note