Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Current Structural Variant Calling Biases Compromise Clinical Genome Diagnostics

Blaž Vrhovšek
University of Ljubljana
Ana Markež Vrhovšek
University of Ljubljana
Doroteja Vujinović
University of Ljubljana
Jernej Kovač
University of Ljubljana
Maruša Debeljak
University of Ljubljana
Tadej Battelino
University of Ljubljana
Barbara Jenko Bizjan
barbara.jenko.bizjan@kclj.si
University of Ljubljana

Article

Keywords:

Posted Date: April 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9179453/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Current Structural Variant Calling Biases Compromise Clinical Genome Diagnostics

Authors:

Main author: Blaž Vrhovšek; 1,2; 0009-0006-5533-742X
Ana Markež Vrhovšek; 1,2; 0009-0005-2758-8128
Doroteja Vujinović; 2,4; 0009-0009-6211-4947
Jernej Kovač; 1,2; 0000-0003-3080-4838
Maruša Debeljak; 1,2; 0000-0003-0446-5465
Tadej Battelino; 1,3; 0000-0002-0273-4732
Corresponding author: Barbara Jenko Bizjan; 1,2; 0000-0003-3452-7006

1. **Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia**
2. **Clinical Institute of Special Laboratory Diagnostics, University Children's Hospital, University Medical Centre Ljubljana, Ljubljana, Slovenia**
3. **Department of Endocrinology, Diabetes and Metabolic Diseases, University Children's Hospital, University Medical Centre Ljubljana, Ljubljana, Slovenia**
4. **Biotechnical Faculty, University of Ljubljana, Ljubljana, Slovenia**

Corresponding author mail: barbara.jenko.bizjan@kclj.si

Abstract

Structural variants (SVs) account for a substantial proportion of pathogenic mutations in rare diseases, yet their detection remains challenging in clinically relevant coding regions. We systematically benchmarked SV detection across Illumina WGS, Illumina WES, ONT WGS, and PacBio WGS using the HG002 GIAB truth set on GRCh37 and GRCh38. Performance was assessed within three nested interval sets: high-confidence intervals (HCI), a pediatric disorders gene panel (GP), and exons plus UTRs (EX+UTR). To isolate genomic context effects from interval size, we developed a novel simulation framework generating exon-like target sets outside coding regions. Long-read platforms outperformed short-read sequencing across all intervals, with PacBio Pbsv achieving F1=0.94 in HCI versus 0.62 for Illumina WGS. SV detection performance consistently dropped in EX+UTR regions compared to simulated counterparts, indicating systematic bias beyond interval size effects. Analysis on GRCh38 revealed that region-dependent performance gaps persisted, although they were less pronounced and predominantly affecting precision. These results demonstrate that SV detection accuracy is both

technology- and region-dependent, underscoring the need for interval-specific benchmarking in clinical genomics.

Introduction

Structural variants (SVs) represent a major source of human genetic variation, impacting more base pairs in the genome than single nucleotide variants (SNVs). They are significantly enriched among pathogenic variants in rare diseases and are known contributors to cancer development and progression [1,2]. They include deletions, insertions, duplications, inversions, and translocations, and can disrupt gene function, alter dosage, or perturb regulatory networks. SVs are responsible for a significant fraction of pathogenic variants associated with Mendelian disorders, neurodevelopmental conditions, and cancers [3–5]. Given their disproportionate impact on phenotypic variation and disease risk, accurate SV detection is essential for precision medicine and clinical diagnostics [1,6].

Despite major advances in sequencing technology, SV detection remains less reliable than single-nucleotide variant (SNV) calling. Challenges include the wide size spectrum of SVs, their frequent occurrence in repetitive or low-mappability regions, and biases in variant distribution across the genome. Approximately 90% of SVs are noncoding deletions, and SVs are depleted from coding regions and 3′ UTRs, where functional constraint is highest [7,8]. Small polymorphisms within homologous regions are an additional source of detected SV displacement [7]. These distributional biases have important implications for benchmarking, as performance estimates from genome-wide analyses may overstate accuracy in clinically relevant intervals.

Sequencing technology further modulates detection performance. Short-read WGS (e.g., Illumina) offers high base accuracy and low cost but struggles with large or complex SVs [6]. Long-read platforms (PacBio, ONT) better resolve SV breakpoints and repetitive regions but are costlier [9,10]. While several benchmarking studies exist [11–14], none have examined performance differences between coding-rich and noncoding intervals, or compared short- and long-read technologies directly in a clinical diagnostic context.

Here, we present a systematic benchmarking of SV detection across four sequencing platforms: Illumina WGS, Illumina WES, ONT WGS, and PacBio WGS. Using the GIAB HG002 truth set on both GRCh37 and GRCh38 reference genomes, we evaluated detection performance in three nested genomic interval categories: high-confidence intervals (HCI), a pediatric disorders gene panel (GP), and exons plus untranslated regions (EX+UTR). To our knowledge, this is the first study to introduce a simulated-interval framework that generates exon-like regions outside coding space to isolate the effect of genomic location from interval size. By comparing real and simulated target sets, we quantify the intrinsic bias in SV detection between coding and noncoding regions across technologies and callers. Our results provide actionable guidance for platform and pipeline selection in rare disease diagnostics, emphasizing the importance of

interval-specific benchmarking and transparent reporting of regional performance differences to appropriately inform clinical decision-making.

Methods

Data Acquisition and Alignment

We utilized Genome in a Bottle (GIAB) data for the HG002 sample, aligned to both GRCh37 and GRCh38 reference genomes to assess reference-dependent effects on SV detection [15]. The inclusion of GRCh37 as the primary reference reflects current clinical reality: despite the availability of GRCh38, the majority of clinical diagnostic laboratories worldwide continue to perform routine genetic analysis on GRCh37 due to established validation pipelines, regulatory considerations, and the extensive annotation resources developed for this assembly [16]. Understanding SV detection performance on GRCh37 therefore remains directly relevant to contemporary diagnostic practice. The dual-reference approach was employed because the transition from GRCh37 to GRCh38 represents a major shift in clinical genomics, and understanding how this affects SV detection performance is essential for laboratories planning reference migrations.

For the GRCh37 analysis, pre-aligned BAM files were obtained from the GIAB FTP repository representing near-optimal detection conditions: Illumina WES (~180x mean target coverage), Illumina WGS (~60x coverage), PacBio HiFi WGS (~48x coverage), and Nanopore WGS (~47x coverage). These coverage levels exceed typical clinical thresholds (30x for short-read WGS [17], 20x-10x for long-read WGS, and even lower [18]). As the ground truth for GRCh37, we used the GIAB HG002 SV Tier1 v0.6 passed variants, restricting our analysis to the provided high-confidence intervals (HCI).

For the GRCh38 analysis, we selected datasets matching the sequencing technologies and coverage depths of the GRCh37 counterparts to ensure comparability. We obtained the Illumina WGS (downsampled to ~60x) and PacBio HiFi Revio (~48x) datasets from the GIAB Ashkenazim Trio repository, alongside the corresponding ONT Ultralong (~47x) dataset. For the GRCh38 ground truth, we utilized the NIST HG002 Draft Benchmark (v0.012-20231107, *GRCh38_HG002-T2TQ100-V1.0_stvar.vcf.gz*) and its corresponding high-confidence intervals.

GRCh38 BAM files, specifically Illumina and PacBio were aligned to references with differing non-canonical contigs. As the HCI does not include any of those contigs we generated new BAM headers containing only the primary assembly chromosomes (1–22, X, Y, and M) and filtered reads to exclude those mapping to non-canonical contigs. For Illumina data, we further filtered to retain only properly paired reads. This pre-processing step was critical to standardize the genomic search space and ensure that differences in performance were attributable to the sequencing technology and reference assembly rather than alignment artifacts in decoy regions.

Structural Variant Calling

SV calling was performed using different tools depending on the sequencing technology. The selection of SV callers was guided by several criteria: (1) widespread adoption in clinical and research laboratories, ensuring reproducibility and comparability with existing literature; (2)

established performance benchmarks in prior studies; (3) active maintenance and community support; and (4) technology-specific optimization. While newer SV callers continue to emerge, we deliberately focused on well-validated tools that represent the current standard of care in clinical genomics, as novel algorithms require extensive validation before clinical deployment.

For Illumina WES and WGS, Manta (v1.6.0) [19] was selected as it remains one of the most widely used and clinically validated SV callers for short-read data, offering robust performance across diverse SV types and sizes while maintaining computational efficiency suitable for clinical throughput requirements. Given Manta's established dominance in clinical short-read SV calling and the relative maturity of the Illumina SV calling landscape, a single caller was deemed sufficient for short-read evaluation. WES analysis was restricted to target regions defined by the Agilent SureSelect Human All Exon V5 BED file.

For long-read platforms, we deliberately employed two callers per technology to enable both cross-platform comparison and technology-specific optimization assessment. On Nanopore PromethION WGS, SVs were called using CuteSV (v2.0.2) [20] and Sniffles (v2.4) [21]. CuteSV was chosen as a platform-agnostic caller enabling direct comparison between ONT and PacBio results, while Sniffles represents the flagship ONT-optimized caller with demonstrated superior performance in long-read benchmarks. With Sniffles we provided a tandem repeats BED file to improve accuracy in repetitive regions. For PacBio WGS, SV calling was performed using CuteSV (v2.1.0) and Pbsv (v2.11.0) (https://github.com/PacificBiosciences/pbsv). Pbsv was included as the vendor-recommended, PacBio-optimized caller, while CuteSV enables cross-platform comparison with ONT results. This dual-caller strategy for long-read data allows us to assess both the performance ceiling achievable with technology-specific tools and the consistency of results when using a common caller across platforms. Default parameters were used unless otherwise specified, as parameter optimization for specific genomic contexts was beyond the scope of this benchmarking study, and default settings represent the most common clinical implementation scenario. We acknowledge that default parameters are typically optimized for genome-wide performance; region-specific parameter tuning may partially mitigate the coding-region biases we observe, though this remains to be systematically evaluated.

Benchmarking and Interval-Based Analysis

To evaluate performance, we analyzed variants overlapping three nested genomic interval categories, representing different diagnostic target ranges. The first level consists of high-confidence intervals (HCI), as defined by GIAB, covering approximately 86% of the genome and containing most genes and intergenic regions (see Figure 1). The second level comprises the Pediatric Disorders Gene Panel (GP); Ensembl gene IDs were mapped using biomaRt to obtain genome coordinates. This interval covers approximately 14% of the HCI, while the HCI itself covers 95% of the GP. It contains 3886 clinically important genes, excluding intergenic regions. The third level consists of exonic regions combined with UTRs (EX+UTR) extracted from GENCODE v19. This set covers approximately 4% of the HCI, with 90% of target regions overlapping the HCI. For both GP and EX+UTR only the regions overlapping HCI were used. Refer to Figure 1 for visual representation of the three genomic intervals.

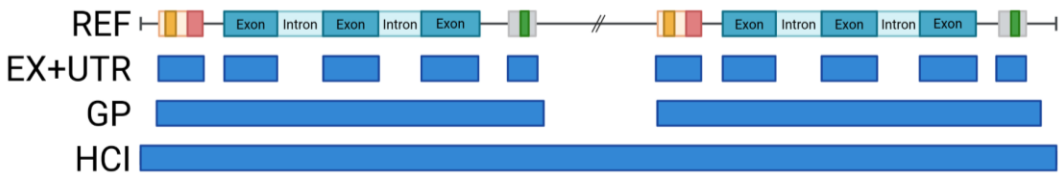


Figure 1. *Schematic overview of genomic interval hierarchies used in variant analysis. The reference gene model (REF) is shown with annotated exons, introns, UTRs, and intergenic regions. Three hierarchical target regions are illustrated below the gene model: EX+UTR (exons and UTRs), GP (gene panel), and HCI (high-confidence intervals). The diagram highlights the nested nature of these interval sets, with EX+UTR representing the most specific and HCI the most inclusive range used in benchmarking.*

Variant Comparison and Accuracy Metrics

To compare called variants, we utilized Truvari (v5.0.0) [22], obtaining recall, precision, and F1-score for each dataset. Given our benchmarking requirements, we modified the Truvari code to allow comparison of variants that partially overlap target intervals (minimum 1 bp overlap), where the original Truvari disregards all variants not fully contained within target intervals. This modification was essential because clinically relevant SVs frequently span exon-intron boundaries, and strict containment filtering would systematically exclude such variants from evaluation, potentially biasing performance estimates. This modified version is available on GitHub: https://github.com/CISLD/truvari.

For benchmarking only passed variants were considered. All SVs were classified as insertions or deletions to enable consistent comparison across callers. This simplification was necessary because different callers use varying nomenclature for complex SV types (e.g., duplications may be reported as insertions by some tools), and harmonizing to two fundamental categories eliminates classification discrepancies that would otherwise confound performance comparisons. For WES calls, the minimal allele size similarity threshold was set to 0, as most variants extend beyond exonic regions and strict size matching would penalize correct calls with slightly different breakpoint resolution. Additionally, to avoid distinguishing duplications from insertions, the minimum percentage allele sequence similarity was set to 0, ensuring that sequence-level differences in SV representation do not artificially inflate false positive or negative counts.

Simulated Exon Intervals

To test the hypothesis that performance differences in coding regions reflect genomic location rather than interval size alone, we developed a novel simulation framework generating exon-like target sets outside coding regions. This approach is critical because coding regions differ from noncoding regions not only in size distribution but also in sequence composition, GC content, repetitive element density, and evolutionary constraint; all factors that could independently affect

SV detection. By generating size-matched intervals in noncoding space, we isolate the effect of genomic context from the confounding influence of target size.

First, we calculated the median length of each decile of original exon interval lengths to preserve the size distribution of real exons. The probability of an exon occurring on each chromosome was determined based on its frequency in the genome, ensuring simulated intervals reflect the chromosomal distribution of coding regions. Non-exonic regions within HCI were split into decile median lengths and randomly selected according to chromosome-specific probabilities. Importantly, simulated intervals were constrained to lie within GIAB high-confidence regions, which excludes known problematic areas such as centromeres, telomeres, and large segmental duplications. To avoid repetition and ensure independence between simulations, selected intervals were excluded from HCI in subsequent iterations. This process was repeated 500 times, generating 500 simulated exome target BED files. The large number of iterations provides robust statistical power to detect systematic differences between real and simulated regions. Simulations were not performed for the Illumina WES technology, as it inherently provides insights only into true exonic regions due to capture-based enrichment. Figure 2 provides a visual representation of the simulation workflow.

Simulations were performed on both the GRCh37 and GRCh38 references, in accordance with the obtained reference data sets.

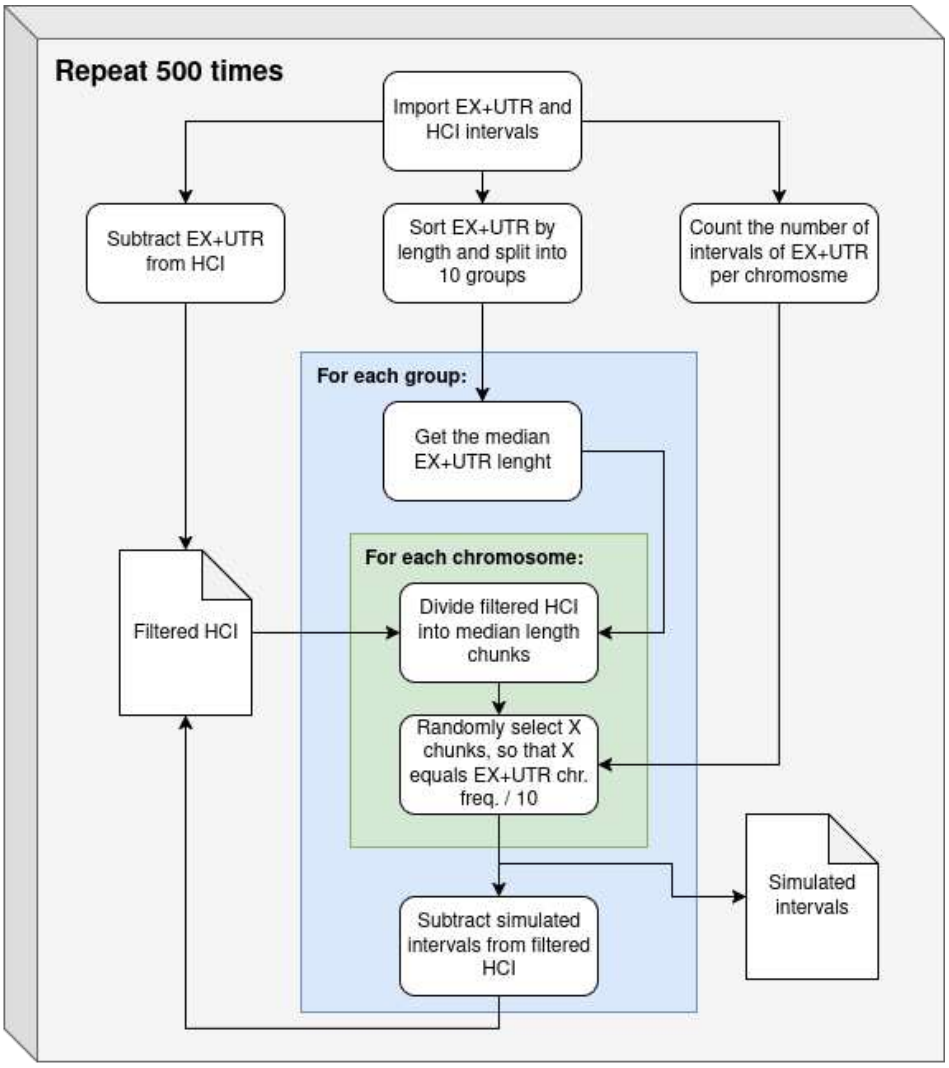


Figure 2. *Diagram highlighting the pipeline for simulating EX+UTR like intervals. The process is optimized to be run in parallel and the final output are 500 simulated intervals bed files.*

Statistics and reproducibility

Benchmarking metrics (precision, recall, F1-score) were calculated using Truvari across all interval categories. To isolate the effect of genomic context, we generated a null distribution using 500 independent simulation iterations with a provided seed. We visualized the resulting performance distributions using Kernel Density Estimation (KDE) and determined statistical significance by computing the empirical percentile rank of each real EX+UTR metric within the simulated distribution. Using a one-tailed framework ($\alpha = 0.05$), real metrics falling below the 5th percentile of the simulated values were considered significantly reduced relative to expectations. All analyses are deterministic and reproducible using the provided code repository.

Code availability

All scripts and pipelines used in this study are available on GitHub: https://github.com/CISLD/SV_coding_regions_benchmark_nextflow. The analysis workflow was organized into a bash script for data download and preparation, and a Nextflow pipeline for running simulations and benchmarks. Custom code modifications to Truvari are available at https://github.com/CISLD/truvari.

Results

Benchmarking variants across genome ranges on GRCh37

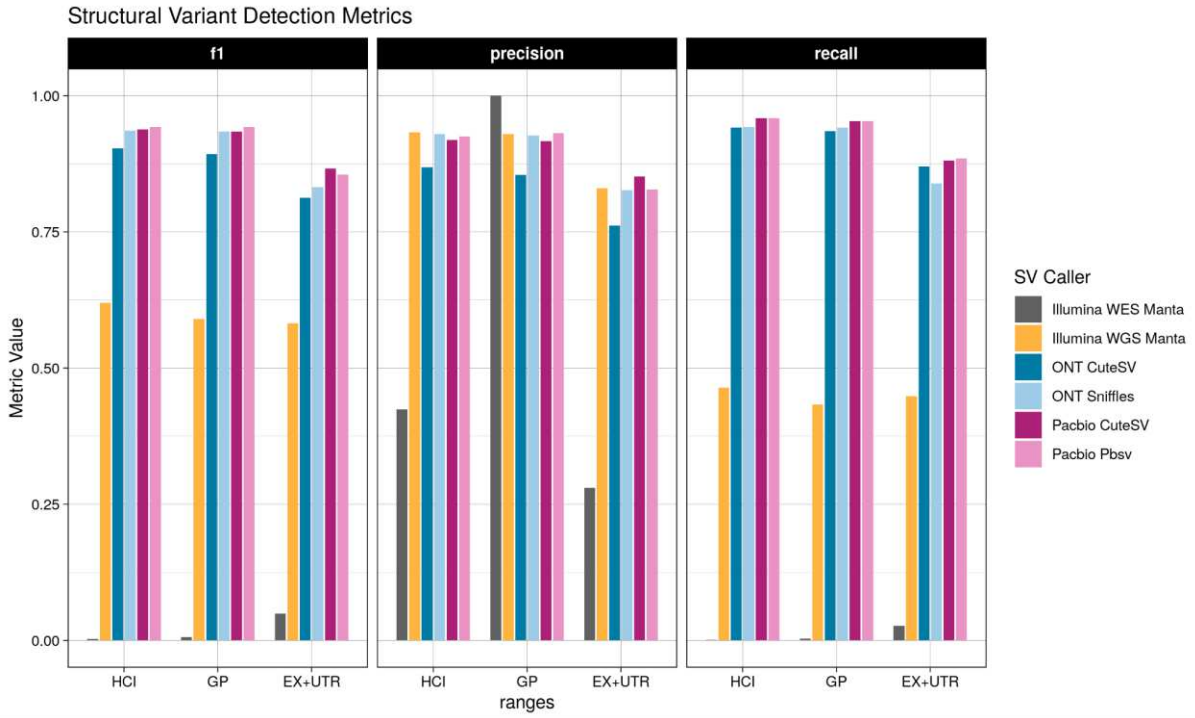


Figure 3. *Structural Variant Detection Metrics across Sequencing Platforms and Genomic Ranges. The bar plots show F1-score, precision, and recall for six structural variant (SV) calling pipelines evaluated over three genomic interval categories: HCI, GP and EX+UTR. Precision (middle panel) reflects the proportion of called SVs that are true positives, meaning how many of the calls are actually correct. Recall (right panel) indicates the proportion of all true SVs in the genome that were successfully detected, meaning how many real variants were actually captured. F1-score (left panel) is the harmonic mean of precision and recall, providing a balanced measure of overall performance. PacBio-based callers (CuteSV and Pbsv) demonstrate consistently high accuracy across all metrics and regions, while Illumina-based methods, particularly WES, underperform, especially in broader interval categories. Oxford Nanopore (ONT) methods show strong recall, though slightly lower precision compared to PacBio.*

Long-read sequencing consistently outperforms short-read approaches in SV detection across all genomic intervals, with the largest differences observed in recall. Figure 3 summarizes F1 scores, precision, and recall for each platform-caller combination across high-confidence intervals (HCI), the pediatric disorders gene panel (GP), and exons plus untranslated regions (EX+UTR). PacBio Pbsv achieved the highest overall accuracy in HCI (F1 = 0.94, precision = 0.93, recall = 0.96), closely followed by PacBio CuteSV (F1 = 0.94) and ONT Sniffles (F1 = 0.94). In contrast, Illumina WGS (Manta) maintained high precision (0.93) but had substantially reduced recall (0.46), while Illumina WES exhibited extremely poor recall (<0.01) despite moderate-to-high precision, underscoring its limited utility for genome-wide SV detection.

Performance within GP intervals closely mirrored HCI results, with only minor drops in all metrics (<1% absolute change for most callers), suggesting that clinically relevant gene panels maintain similar detection characteristics to broader high-confidence regions. However, EX+UTR regions consistently showed the lowest accuracy, with an average reduction of 9.6% in precision, 7% in recall, and 7.9% in F1 compared to HCI. This drop was most pronounced for ONT Sniffles (F1 decrease: 10.3%) and least for Illumina WGS (3.7%).

Illumina WGS exhibited the smallest relative performance change from HCI to EX+UTR (F1 decrease: 3.7%), reflecting its already limited recall across all regions. In contrast, long-read callers, while maintaining high absolute accuracy, experienced more pronounced drops when restricted to coding-rich intervals, suggesting that part of their advantage lies in better performance in noncoding or repetitive regions where most SVs occur.

Detailed per-interval metrics are provided in Supplementary Table 2.

Simulated-interval analysis confirms intrinsic coding–noncoding detection bias

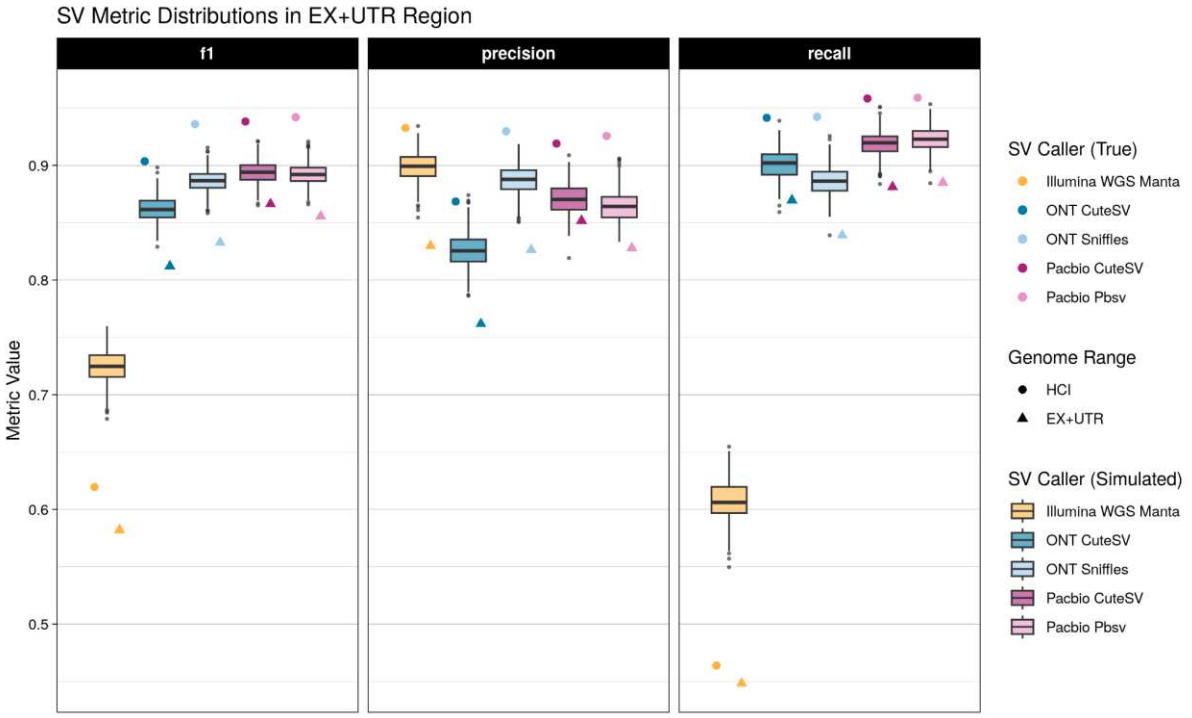


Figure 4: *Structural Variant Detection Performance Across Simulated EX+UTR Regions. Performance metrics for SVs within simulated EX+UTR-like regions are shown as boxplots, corresponding metrics for true EX+UTR regions are represented by colored dots, while colored triangles represent true HCI regions. Notably, the real values are consistently lower than those of the simulated regions, indicating a systematic bias in SV detection between coding and noncoding regions.*

To determine whether EX+UTR performance losses stem from smaller target sizes or from genomic location, we generated 500 random exon-like interval sets located outside coding regions. Across all technologies, simulated EX+UTR-like regions yielded consistently higher median precision, recall, and F1 than the real EX+UTR targets (Figure 4). Real EX+UTR metrics fell below the 5th percentile of the simulated distribution for all caller–platform combinations.

The largest real-simulated performance gaps were seen in Illumina WGS (median Δprecision = +0.11, Δrecall = +0.16, ΔF1 = +0.16 for simulated over real). PacBio CuteSV showed the smallest differences (Δprecision = +0.02, Δrecall = +0.04), indicating greater robustness to genomic context. Simulation summary statistics are presented in *Supplementary Table 4b* and Percentile and KDE analyses confirming coding-region bias are reported in *Supplementary Table 4a*. Individual simulation metrics are available in *Supplementary Data 1*. The differences are visualised in *Supplementary Figure 2*.

These results demonstrate a systematic SV detection bias against coding-rich regions that cannot be explained by target size alone. The effect is technology-dependent, with short reads

suffering both from lower baseline recall and from larger relative losses in coding contexts, while long reads, despite higher overall accuracy, still exhibit measurable context-specific declines.

Reference genome comparison: GRCh37 versus GRCh38

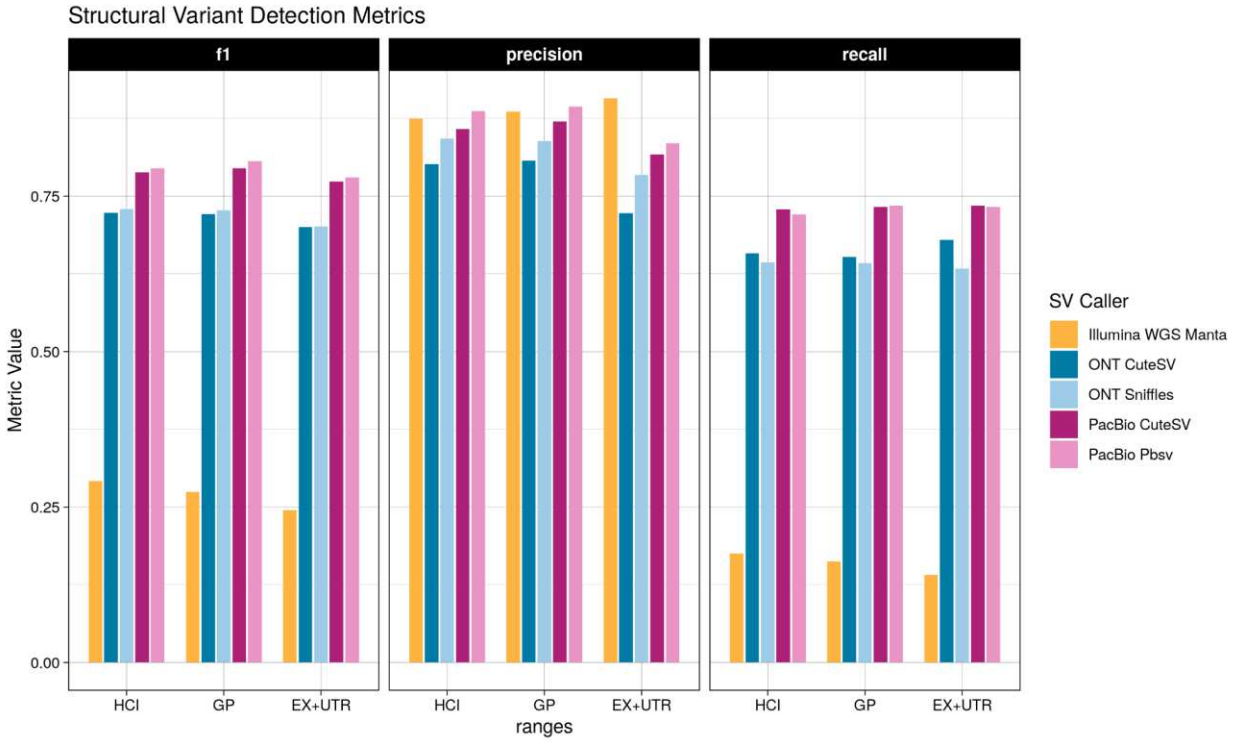


Figure 5: *Barplot showing the Structural Variant Detection Metrics across Sequencing Platforms and Genomic Ranges for the GRCh38 reference. Note the mirrored trend with technology and callers. The only noticeable drop in performance between intervals is with precision on EX+UTR.*

To assess reference-dependent effects on SV detection, we repeated all analyses using GRCh38. Overall, metrics declined substantially across all platforms and callers when transitioning from GRCh37 to GRCh38 (Figure 5). In HCI regions, PacBio Pbsv F1 dropped from 0.94 (GRCh37) to 0.79 (GRCh38), driven primarily by reduced recall (0.96 to 0.72) with a smaller precision decline (0.93 to 0.89). ONT Sniffles showed a similar pattern, with F1 declining from 0.94 to 0.73, recall from approximately 0.94 to 0.64, and precision from approximately 0.93 to 0.85. Illumina WGS (Manta) exhibited the most dramatic recall reduction (0.46 to 0.17), though precision remained relatively stable (0.93 to 0.87).

Detailed per-interval metrics are provided in *Supplementary Table 1*.

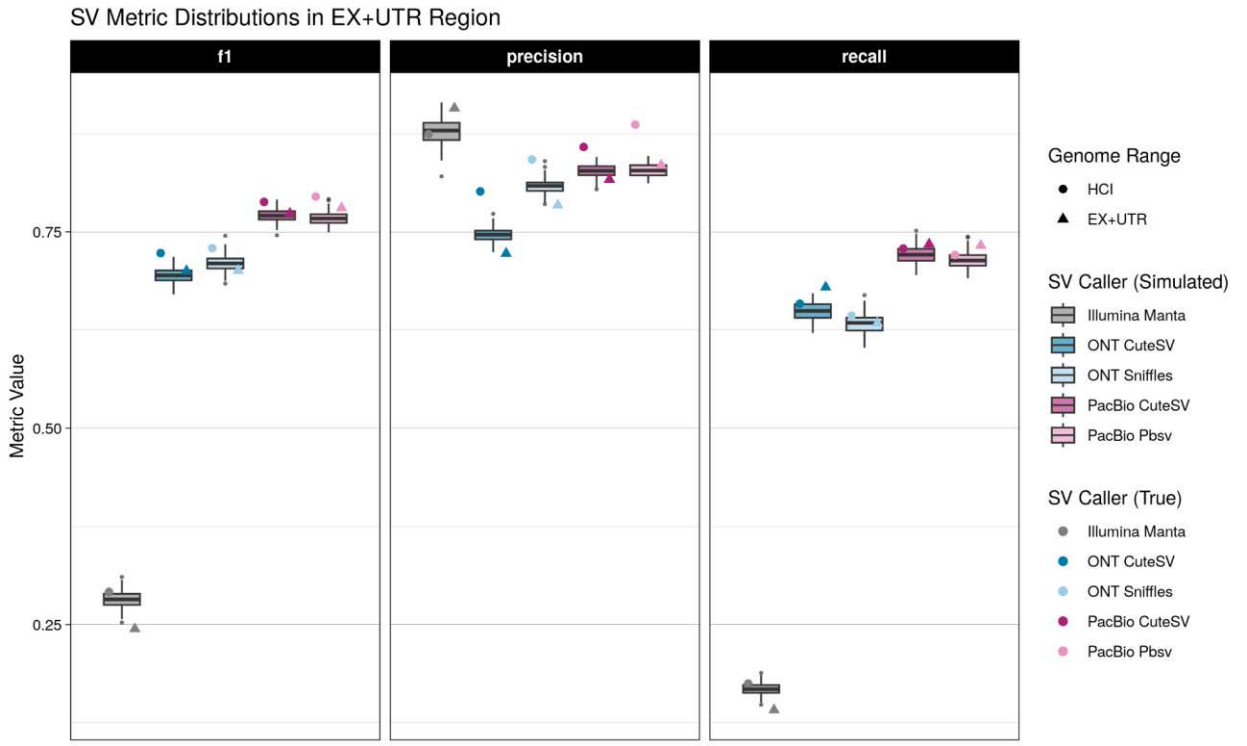


Figure 6: *Boxplot showing performance metrics for SVs within simulated EX+UTR-like regions. Notably, the real values are consistently lower only partly with precision, indicating that the systematic bias in SV detection between coding and noncoding regions has been measurably corrected with the use of a newer reference genome.*

The region-dependent performance differences observed on GRCh37 partly persisted on GRCh38. EX+UTR regions continued to show the lowest metrics, with precision remaining disproportionately affected in coding regions across both reference assemblies. Notably, PacBio demonstrated greater robustness to reference changes compared to ONT: PacBio Pbsv retained higher absolute F1 scores (0.79 in HCI) than ONT Sniffles (0.73), and PacBio CuteSV (F1 = 0.79) outperformed ONT CuteSV (F1 = 0.73) on GRCh38 (Figure 6). This platform-specific difference suggests that PacBio-based pipelines may require less recalibration when migrating between reference assemblies. Simulation summary statistics are presented in *Supplementary Table 3b* and Percentile and KDE analyses confirming coding-region bias are reported in *Supplementary Table 3a*. Individual simulation metrics are available in *Supplementary Data 1*. The differences are visualised in *Supplementary Figure 1*.

These metric declines on GRCh38 may be partially attributable to the use of a draft GIAB HG002 truth set, which lacks the mature curation of the established GRCh37 benchmark. Nevertheless, the persistence of relative region-dependent biases across both references underscores that coding-region performance deficits represent a fundamental challenge for SV detection.

Discussion

Our benchmarking between sequencing technologies is consistent with previous studies [23–25] and confirms that Illumina WGS, while delivering high precision (~93%) on the GP, suffers from low recall, missing more than half of true SVs. In contrast, long-read platforms such as ONT and PacBio maintain similar precision but achieve much higher recall (~94%), making them better suited for comprehensive SV detection in rare disease contexts. However, SV detection accuracy is not uniform across the genome. Performance varied measurably between GP, HCI, and EX+UTR regions. GP intervals showed only minor drops compared to HCI, but EX+UTR regions consistently yielded the lowest metrics, with precision disproportionately affected. These differences likely arise from distinct SV distributions in coding-rich regions, which are especially pronounced within the GRCh37 reference.

While several comprehensive SV benchmarking studies have been published [11–14,23–25], our work addresses critical gaps that limit the clinical applicability of prior findings. First, existing benchmarks predominantly report genome-wide performance metrics, which systematically overestimates accuracy in the coding regions most relevant to clinical diagnosis. By stratifying performance across HCI, GP, and EX+UTR intervals, we provide the first granular assessment of region-dependent SV detection biases across different human references with direct implications for diagnostic sensitivity.

Second, our novel simulation framework represents a methodological advance not employed in previous studies. By generating 500 size-matched exon-like interval sets in noncoding regions, we definitively isolate genomic context effects from interval size confounding, a distinction that genome-wide benchmarks cannot make.

Third, we explicitly benchmark on both GRCh37 and GRCh38 reference genomes. Despite the availability of GRCh38, the majority of clinical laboratories worldwide continue to use GRCh37 for routine diagnostics due to established validation pipelines and regulatory requirements. Prior studies focusing exclusively on GRCh38 provide limited guidance for these laboratories. Our dual-reference analysis demonstrates that reference choice substantially affects performance and that GRCh37-derived metrics cannot be directly extrapolated to GRCh38 pipelines.

Finally, our caller selection strategy, where we employ both a cross-platform caller and technology-optimized callers, enables direct platform comparison while also assessing the performance ceiling achievable with specialized tools. This design provides more nuanced guidance for clinical implementation than studies using only platform-specific callers.

Implications for Clinical Genome Diagnostics

The observed region-dependent performance biases have profound implications for clinical genome diagnostics, where accurate SV detection in coding regions is paramount. Pathogenic SVs causing Mendelian disorders frequently disrupt exonic sequences, and our finding that all

tested platforms exhibit reduced performance specifically in these clinically critical EX+UTR regions reveals a systematic blind spot in current diagnostic pipelines, especially those utilizing the older GRCh37 reference. This bias is particularly concerning because it operates silently as laboratories relying on genome-wide performance metrics may be unaware that their diagnostic sensitivity for exonic SVs is substantially lower than headline figures suggest.

The clinical consequence is a measurable increase in diagnostic false negatives. To quantify this impact, the GIAB HG002 truth set on the GRCh37 contains 9699 SVs within HCI and approximately 261 SVs overlapping EX+UTR regions. With the average reduction of 9.6% in precision and 6.8% in recall, approximately 25 SVs may be wrongly identified and approximately 18 SVs missed, compared to genome-wide expectations. Extrapolating to rare disease cohorts where approximately 5–10% of diagnoses involve coding SVs, this bias could contribute to diagnostic delays in a meaningful fraction of unresolved cases. For rare disease patients, this translates directly to delayed or missed diagnoses, prolonged diagnostic odysseys, and potentially inappropriate clinical management. Our simulation-based analysis demonstrates that this performance degradation cannot be attributed to the smaller size of exonic targets but reflects intrinsic properties of coding sequences that challenge current SV detection algorithms. These properties likely include higher GC content, distinct repeat architectures, and the presence of segmental duplications and processed pseudogenes that complicate read mapping and variant calling in coding space [26].

Technology Selection for Clinical Implementation

Our results provide actionable guidance for technology selection in clinical genomics laboratories. Long-read platforms (PacBio and ONT) substantially outperform short-read WGS for SV detection across all genomic contexts, with the performance advantage most pronounced in recall. Critically, our benchmarking demonstrates that Illumina WES offers negligible utility for SV detection, with recall rates frequently falling below 5%, rendering it clinically insufficient and practically useless for structural variant screening. For clinical laboratories where comprehensive SV detection is essential, particularly in pediatric rare disease, neurodevelopmental disorders, and cancer predisposition syndromes, long-read WGS offers a clear advantage. Notably, PacBio demonstrated more robust performance than ONT.

Reference Genome Considerations

The choice of reference genome emerged as an underappreciated variable affecting SV detection performance. When analyses were repeated using GRCh38, overall metrics declined across all platforms and callers. We attribute that to the GRCh38 SV truth set as it contains 26702 SVs within the HCI, which is approximately 3x times more than contained within the GRCh37 truth set. Nonetheless, the differential performance between genomic regions partly persisted on GRCh38, with precision remaining disproportionately lower in coding regions. This finding suggests that reference-specific factors, such as gap closures, centromeric sequence additions, and improved representation of segmental duplications in GRCh38, may partially confound the coding-noncoding performance differential. The persistence of precision deficits in real exonic regions on both references suggests that precision, rather than recall, remains the primary bottleneck for exonic SV detection across reference assemblies.

These reference-dependent effects carry important implications for clinical laboratories. The transition from GRCh37 to GRCh38 is not performance-neutral, and validation studies conducted on one reference may not directly translate to the other. Laboratories should conduct reference-specific benchmarking when implementing or updating SV detection pipelines, particularly for coding-focused diagnostic applications.

Caller-Specific Considerations

Caller choice further influenced regional performance. For long reads, flagship tools (Pbsv for PacBio, Sniffles for ONT) achieved the highest absolute F1 scores, but also exhibited the largest drops from HCI to EX+UTR, suggesting they may be tuned for optimal performance in noncoding or repetitive regions where most SVs occur. In contrast, CuteSV for PacBio showed slightly lower peak accuracy but was less affected by genomic context, hinting at more balanced parameterization.

Recommendations for Clinical Practice

Based on our findings, we propose several recommendations for clinical genome diagnostics:

1. **Interval-specific benchmarking**: Diagnostic pipelines should be validated on the specific interval sets they will be applied to, particularly for coding-rich regions. Genome-wide performance metrics overestimate accuracy in clinically relevant targets.

2. **Technology matching to diagnostic goals**: For comprehensive SV detection in coding regions, long-read platforms should be preferred where available. Laboratories limited to short-read WGS should acknowledge and communicate the reduced sensitivity for exonic SVs.

3. **Reference genome validation**: Performance validation should be repeated when changing reference assemblies, as reference-dependent effects on SV detection are substantial and technology-specific.

As sequencing technologies evolve and new SV callers are introduced, addressing this context-specific bias will be essential for achieving accurate, unbiased SV detection in precision medicine. This includes tailoring caller parameters to the target genomic space, incorporating benchmarking intervals representative of diagnostic applications, and transparently reporting regional performance differences alongside genome-wide metrics.

Limitations

This benchmarking was performed on a single reference sample (HG002) and may not capture population-level variation or performance differences across genomes with distinct repeat landscapes or ancestry-specific SV profiles. HG002 is of Ashkenazi Jewish ancestry, and SV landscapes are known to vary substantially across global populations, with African-ancestry genomes exhibiting higher SV diversity. While GIAB HG002 represents the gold standard for SV benchmarking due to its extensive validation, the generalizability of our findings to ancestry-diverse clinical populations requires verification. Notably, other GIAB samples (HG001, HG005, HG006, HG007), which include individuals of Han Chinese and Puerto Rican ancestry, were not evaluated as they lack comprehensive truth sets. Extending this analysis to these samples would strengthen the generalizability of our conclusions.

Additionally, the GIAB HG002 truth set represents a curated subset of "callable" SVs within high-confidence regions. Complex SVs, nested events, and variants in segmental duplications are underrepresented. This curation may bias our results toward SVs that are inherently easier to detect, and performance for challenging variant classes may be lower than reported here.

The sequencing coverage levels used in this study (60x Illumina WGS, 48x PacBio, 47x ONT) exceed typical clinical thresholds. SV detection is coverage-dependent, particularly for long-read technologies, and real-world performance at lower coverage depths (e.g., 15–20x long-read) may be reduced compared to our findings.

While our GRCh38 analysis provides important insights, these results must be interpreted with caution. When analyses were repeated using GRCh38, overall metrics declined across all platforms and callers; however, this performance shift may not represent genuine degradation but rather reflect the inherent maturation gap between truth sets. The GRCh38 draft benchmark (defrabbV0.012-20231107) contains 26,702 SVs within the HCI, approximately 3x more than the GRCh37 truth set, and while this potentially captures a more comprehensive range of complex variants, it has not yet reached the same level of exhaustive curation and validation. Until a GRCh38 truth set achieves equivalent maturation and curation stringency as the established GRCh37 benchmark, a direct, quantitative comparison of performance metrics across reference assemblies remains inherently confounded.

Our selection of SV callers focused on established, clinically validated tools rather than the newest algorithms. While this ensures immediate clinical relevance, it means our findings may not reflect the performance of recently published callers. Newer tools specifically designed for

coding-region SV detection may partially address the biases we identify, though they would require independent validation before clinical deployment.

The simulation framework, while controlling for interval size, cannot fully reproduce the complex sequence contexts of coding regions, including specific repeat architectures, GC content gradients, and evolutionarily constrained sequences. The inability to detect significant differences between real and simulated exon regions on GRCh38 suggests that reference-specific factors may influence the detectability of coding-noncoding performance differentials.

Our analysis was restricted to deletions and insertions as classified by Truvari. Complex SVs, inversions, and translocations were not specifically evaluated, though these variant classes may exhibit different region-dependent biases. Furthermore, we did not assess phasing accuracy or genotype concordance beyond basic metrics, which are important considerations for compound heterozygote identification in recessive disorders.

This study focused exclusively on germline SV detection. Somatic SV detection, which is critical for cancer genomics, may exhibit different region-dependent biases due to distinct variant characteristics including lower allele frequencies, tumor heterogeneity, and different mutational processes. The applicability of our findings to somatic variant calling pipelines remains to be established.

Future work should extend benchmarking to additional truth sets, clinical genomes with confirmed pathogenic SVs, and population-diverse samples to comprehensively characterize regional performance biases across the full spectrum of human genetic variation.

Data availability

All datasets analyzed during the current study are publicly available. The HG002 GIAB benchmark set was obtained from the NIST repository. All data can be automatically downloaded with scripts from the official experiment repository https://github.com/CISLD/SV_coding_regions_benchmark_nextflow.

Author contributions

B.V. conceived, coded and deployed the study and wrote the manuscript with input from all authors. A.M.V. performed the statistical analysis for benchmarking SV on real and simulated genome intervals. D.V. designed the plots and result reporting. J.K. provided mentoring and guidance on study design and on manuscript writing. M.D. and T.B. provided guidance and support within the research program. B.J.B. designed the study and provided mentoring on advanced bioinformatics concepts. All authors reviewed the manuscript.

Acknowledgements

This study was supported by the Slovenian research programs:

- Etiology, early detection and treatment of diseases in children and adolescents (P3-0343).
- Slovenian Reference Genome Project - SRGP (J7-60116).
- Identification of altered regulatory pathways, epigenetic changes, and innate genetic variants in pediatric multisystem inflammatory syndrome associated with COVID-19 (J3-50115)
- Genomic screening of newborns (J3-50122)
- Development of genomic infrastructure for the implementation of the Slovenian reference pangenome (TP20250221)

The funder had no role in study design, data collection, or manuscript preparation

Competing interests

The authors declare no competing financial or non-financial interest.

References

1. Liu, Z. *et al.* Towards accurate and reliable resolution of structural variants for clinical diagnosis. *Genome Biol.* **23**, 68 (2022).

2. Ho, S. S., Urban, A. E. & Mills, R. E. Structural variation in the sequencing era. *Nat. Rev. Genet.* **21**, 171–189 (2020).

3. Pande, S., Dawood, M. & Grochowski, C. M. Structural Variants: Mechanisms, Mapping, and Interpretation in Human Genetics. *Genes* **16**, 905 (2025).

4. D'haene, E. & Vergult, S. Interpreting the impact of noncoding structural variation in neurodevelopmental disorders. *Genet. Med.* **23**, 34–46 (2021).

5. Sanchis-Juan, A. *et al.* Complex structural variants in Mendelian disorders: identification and breakpoint resolution using short- and long-read genome sequencing. *Genome Med.* **10**, 95 (2018).

6. Chaisson, M. J. P. *et al.* Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* **10**, 1784 (2019).

7. Audano, P. A. *et al.* Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell* **176**, 663-675.e19 (2019).

8. Abel, H. J. *et al.* Mapping and characterization of structural variation in 17,795 human genomes. *Nature* **583**, 83–89 (2020).

9. Espinosa, E., Bautista, R., Larrosa, R. & Plata, O. Advancements in long-read genome sequencing technologies and algorithms. *Genomics* **116**, 110842 (2024).

10. Mahmoud, M. *et al.* Structural variant calling: the long and the short of it. *Genome Biol.* **20**, 246 (2019).

11. Dierckxsens, N., Li, T., Vermeesch, J. R. & Xie, Z. A benchmark of structural variation detection by long reads through a realistic simulated model. *Genome Biol.* **22**, 342 (2021).

12. De La Vega, F. M. *et al.* Benchmarking of germline copy number variant callers from whole genome sequencing data for clinical applications. *Bioinforma. Adv.* **5**, vbaf071 (2025).

13. Joe, S. *et al.* Comparison of structural variant callers for massive whole-genome sequence data. *BMC Genomics* **25**, 318 (2024).

14. Liu, Z., Xie, Z. & Li, M. Comprehensive and deep evaluation of structural variation detection pipelines with third-generation sequencing data. *Genome Biol.* **25**, 188 (2024).

15. Zook, J. M. *et al.* A robust benchmark for detection of germline large deletions and insertions. *Nat. Biotechnol.* **38**, 1347–1355 (2020).

16. Lansdon, L. A. *et al.* Factors Affecting Migration to GRCh38 in Laboratories Performing Clinical Next-Generation Sequencing. *J. Mol. Diagn.* **23**, 651–657 (2021).

17. Schobers, G. *et al.* Genome sequencing as a generic diagnostic strategy for rare disease. *Genome Med.* **16**, 32 (2024).

18. Höps, W. *et al.* HiFi long-read genomes for difficult-to-detect, clinically relevant variants. *Am. J. Hum. Genet.* **112**, 450–456 (2025).

19. Chen, X. *et al.* Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinforma. Oxf. Engl.* **32**, 1220–1222 (2016).

20. Jiang, T. *et al.* Long-read-based human genomic structural variation detection with cuteSV. *Genome Biol.* **21**, 189 (2020).

21. Smolka, M. *et al.* Detection of mosaic and population-level structural variants with Sniffles2. *Nat. Biotechnol.* **42**, 1571–1580 (2024).

22. English, A. C., Menon, V. K., Gibbs, R. A., Metcalf, G. A. & Sedlazeck, F. J. Truvari: refined structural variant comparison preserves allelic diversity. *Genome Biol.* **23**, 271 (2022).

23. Mahmoud, M., Agustinho, D. P. & Sedlazeck, F. J. A Hitchhiker's Guide to long-read genomic analysis. *Genome Res.* **35**, 545–558 (2025).

24. Kosugi, S. & Terao, C. Comparative evaluation of SNVs, indels, and structural variations detected with short- and long-read sequencing data. *Hum. Genome Var.* **11**, 18 (2024).

25. Rao, J. *et al.* Performance evaluation of structural variation detection using DNBSEQ whole-genome sequencing. *BMC Genomics* **26**, 299 (2025).

26. Ebbert, M. T. W. *et al.* Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight. *Genome Biol.* **20**, 97 (2019).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supp.zip