Svirlpool: structural variant detection from long read sequencing by local assembly

Vinzenz May[1], Till Hartmann[1], Dieter Beule[1,*], Manuel Holtgrewe[1,2,*]

[1] Translational Bioinformatics, Institute of Health at Charité Berlin
[2] Medizinische Genetik Mainz, Limbach Genetics eGbR
* Equal contributions

Abstract

Motivation
Long-Read Sequencing (LRS) promises great improvements in the detection of structural genome variants (SVs). However, existing methods are lacking in key areas such as the reliable detection of inserted sequence, precise genotyping of variants, and reproducible calling of variants across multiple samples. Here, we present our method *Svirlpool*, that is aimed at the analysis of Oxford Nanopore Technologies (ONT) sequencing data. Svirlpool uses local assembly of candidate SV regions to obtain high-quality consensus sequences.

Results
Svirlpool obtains competitive results to the leading method Sniffles on the widely used Genome in a Bottle benchmark data sets. On trio data, however, Svirlpools shows a clear favorable performance in terms of mendelian consistency. This indicates that Svirlpool shows great promise in clinical applications and beyond benchmark datasets.

Availability and Implementation
Source code, container images, and documentation available at
https://github.com/bihealth/svirlpool

Contact
vinzenz.may@bih-charite.de

1 Introduction

Genomic structural variants (SVs) are causative for a wide array of rare diseases (Collins et al., 2020), Alzheimer's disease (Vialle et al., 2025), and play an important role in cancer biology (Aaltonen et al., 2020; Cortés-Ciriano et al., 2020; Wang et al., 2020). SVs are commonly defined as DNA sequence changes of at least 50 base pairs length. Due to the repetitive and segmentally duplicated architecture of the human genome, SV detection remains challenging (Liu et al., 2024). Longer reads can span difficult regions in the genome; carry more structural context, whereas short-read approaches still miss a substantial fraction of insertions and deletions (Chaisson et al., 2019), although more recent studies report improvements (Choo et al., 2023).

Long-read SV callers can be broadly grouped into: (1) alignment-based methods that extract SV signals from read-to-reference alignments; (2) de novo assembly approaches that first assemble contigs and then compare them to a reference; and (3) local-assembly methods that assemble subsets of reads anchored by initial mappings. De novo assembly can achieve high accuracy but typically requires very high coverage (50x to 90x), and is computationally expensive (Ahsan et al., 2023; Logsdon et al., 2020). Alignment-based methods are fast and perform best with high quality reference genomes but alignment ambiguities in challenging regions —particularly for Oxford Nanopore Technologies (ONT) reads— can introduce biases that propagate to the results. The local-assembly strategy offers a middle ground by assembling reads in targeted regions to reduce coverage requirements. A few such methods have been published and validated with Pacific Biosciences data, but, to our best knowledge, none for Nanopore data (Denti et al., 2022; Kronenberg et al., 2024).

Long reads from Pacific Biosciences (PB) and Oxford Nanopore (ONT) show different biases in their error profiles, especially given the PacBio HiFi circularized high-accuracy reads (Mikheenko et al., 2022; Park et al., 2025; Sacristán-Horcajada et al., 2021). Key factors for such sequencing errors are tandem repeats, homopolymer runs, and low sequence complexity regions.

In rare disease and cancer studies, SVs are frequently compared across individuals to identify shared variants. However, in repetitive or homologous regions, read alignments may not localize SVs unambiguously, necessitating heuristics to determine variant identity across samples. Existing strategies either offer a multi-sample mode with a built-in merging of variants across samples (Smolka et al., 2024) or refine called variants and then attempt to merge SVs (Kirsche et al., 2023). Recent benchmarks can be found in (Aydin et al., 2025; Helal et al., 2024).

To our knowledge, there are currently three SV callers that have explicit multi-sample SV calling modes: Sniffles2 (PB and ONT) is purely alignment-based and discards read sequence detail for speed, while Sawfish and SVDSS generate local assemblies, but focus on PacBio and therefore ignore the ONT specific biases (Denti et al., 2022; Saunders et al., 2025). Moreover,

post hoc merging tools such as Iris and Jasmine (Kirsche et al., 2023) do not explicitly model sequence properties but rely on general heuristics that are applied to both PB and ONT data.

We developed Svirlpool, a local-assembly SV caller that is tuned for ONT data with an explicit multi-sample mode to find shared variants across individuals from different sequencing runs. Svirlpool generates local consensus assemblies to capture sequence context, then performs SV comparison and merging using tolerance parameters informed by local sequence features, including sequence complexity, k-mer sketch similarity, and overlap with annotated tandem repeats, as well as noise patterns generated from the original read sequences. To reduce ONT-specific artefacts, Svirlpool filters SV calls driven by short homopolymer alterations and short GT-rich or AT-rich subsequences. This design aims to provide robust single- and multi-sample SV analysis in challenging regions including genotyping while maintaining the lower coverage requirements and computational efficiency of local assemblies.

2 Methods

2.1 Method Description

2.1.1 Overview

Figure 1 gives an overview of the algorithm implemented in Svirlpool. The input is the alignments of one or more samples together with the reference sequence and some annotation tracks (see Section 2.2.6 for details). The output is a list of the structural variants with their genotypes and technical/quality annotations in each sample.

As is common with complex algorithms, Svirlpool consists of a hierarchy of steps. The highest level consists of four *parts* described in the following sections. Parts 1-3 run independently for each sample whereas section 4 runs on multiple samples at once.

2.1.2 Candidate region identification (Part 1)

The algorithm scans over the read alignments of a single sample. The alignments produced by the read mapper are considered as insertions/deletions (*indels*) within the alignments (*intra-alignment signal*). Further, two local alignments of adjacent sequence in the same read may show interspersed alignments or align on opposite strands in the reference (*inter-alignment signal*). Inter-alignment signal implies break-ends of the sample's genome with respect to the reference.

Svirlpool collects intra-alignment signals (indels) using empirically determined thresholds. It then filters the signal using the unique regions and other provided annotation tracks and filters the signal using heuristics for alignment and reference artifacts (we observed empirically in ONT data). The algorithm computes depth of coverage tracks from filtered data, for normalization purposes and it computes a copy number track that is later used to guide the read clustering process to approximate the number of clusters by the expected zygosity. The copy number and depth tracks are again used in the final genotying step.

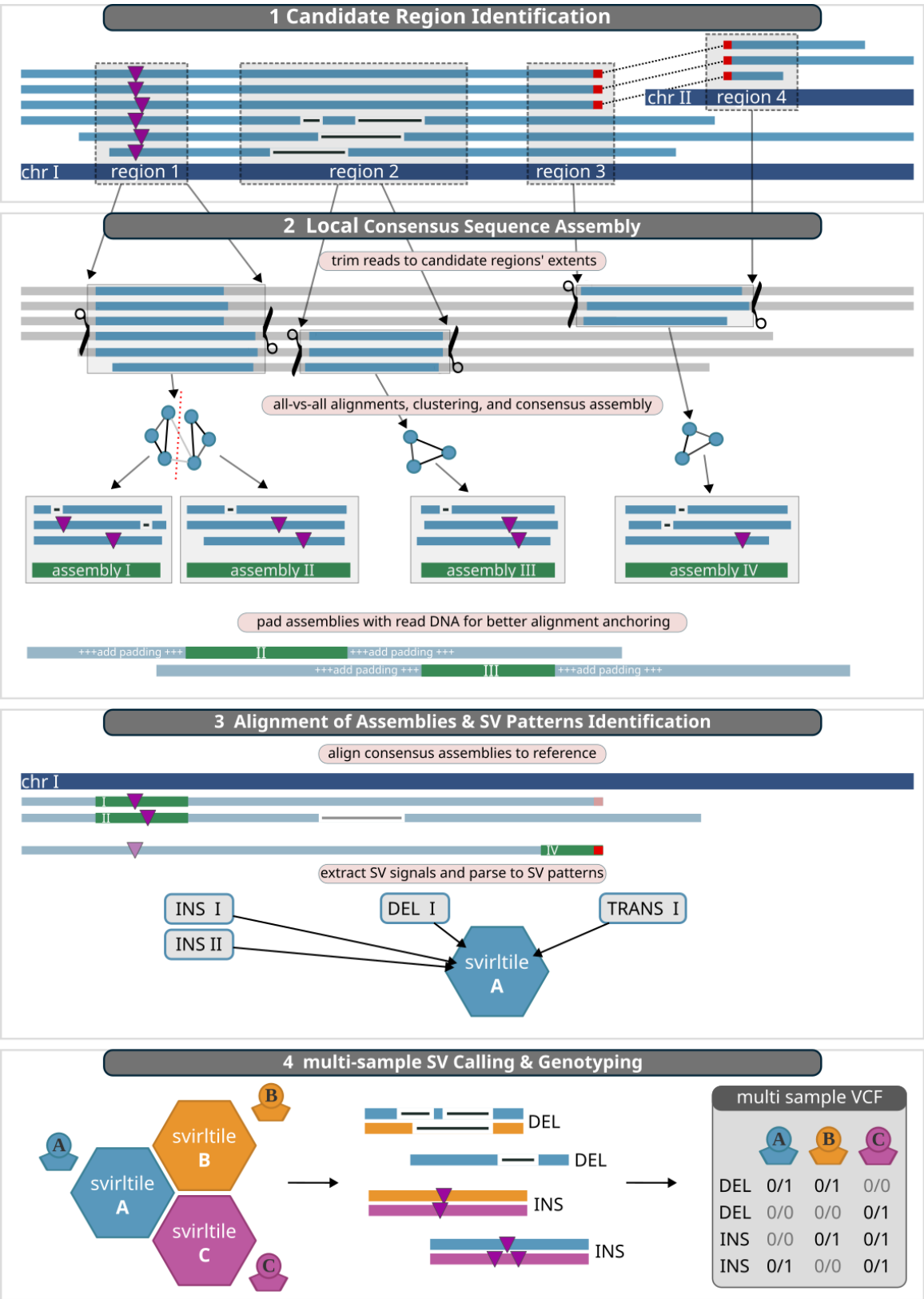

Figure 1 Svirlpool Workflow: **1**. Signals from read alignments are collected to construct candidate regions. **2**. Reads are trimmed according to the bounds of the candidate regions, respecting connected regions via break ends of underlying reads. The trimmed sequences are aligned all versus all and based on a distance measure grouped with

spectral clustering. Each cluster of trimmed reads is assembled and padded with flanking raw read sequences. **3.** All padded consensus assemblies are aligned to the reference genome. SV signals are extracted from the consensus alignments and parsed to SV pattern objects, which are stored in a database file (*svirltile*). **4.** All svirltiles are combined in a multi-sample SV processing step, where the SV patterns are merged into SV composites, which contain the signals from different consensus sequences and samples. SV calls are generated from the SV composites and genotyped and written to a multi-sample vcf file.

We assume that true SVs are expressed in the read alignments by similar, proximal signals. Thus, the task for SV identifications becomes separating the true SV signal from noise, e.g., introduced by sequencing errors, incorrect reference sequence, or alignment artifacts.

Svirlpool scores pairs of signals with a similarity score and normalizes the scores by the local depth of coverage. The algorithm then takes strong signals as seeds for candidate regions and attempts to merge signals into clusters for candidate regions supporting a putative SV based on the similarity scores. The merging takes into consideration the distance on the reference, its positioning in annotated tandem repeats, and incorporates the unique region annotation track. Svirlpool stores each candidate region together with the supporting signal, read names, and read alignment intervals. When two or more candidate regions are supported by the same read via break-ends (BNDs), we call them connected, and the further algorithm parts process them together.

2.1.3 Local read consensus Assembly (Part 2)

Svirlpool considers each connected group of candidate regions separately. For each, the algorithm first **collects all read alignments** that contributed to the signal for the candidate region. It then **trims** them to the candidate regions yielding a **pool *P* of *cut reads*** which is used as the input for the subsequent **clustering step**.

In the idealized case, a candidate region group stems from a single SV in (without loss of generality) heterozygous state and, except for sequencing errors, all reads align perfectly to one of the two distinct alleles. However, real data shows different challenges including: errors in the reference, mistakes made by the aligner, copy number variable regions, high heterogeneity of structural variants ranging from VNTRs over simple to complex structural variants. Thus, we designed our clustering algorithm to address these challenges with minimal assumptions on SV types and focusing on sequences instead.

The clustering uses all-vs-all alignments of the trimmed reads to compute distances based on a custom distance measure that is based on summing and weighting of insertion, deletion, and break-end signals with respect to local GC content and signal distributions. A dynamic cutoff is computed from all distance scores to adapt to local noise variances and to isolate outlier reads. The reads are grouped with spectral clustering based on the sparse similarity matrix that is computed from the read-to-read distances. The number of clusters is determined by the local copy number estimation, which runs the Viterbi algorithm on fixed size bins of median coverages with fixed transition probabilities of 95%, no state change and an equal distribution of state change probabilities. The bin size is scaled with the median read alignment length.

Next, Svirlpool computes **consensus assembly from each read cluster**. For this, it uses the lamassemble software (Frith et al., 2021). All trimmed reads of one cluster are then aligned to their consensus assembly to determine the interval each read covers on the consensus, and to save any occurring insertion or deletion signals to estimate the local noise distribution in the structural variant calling step.

For each consensus assembly, Svirlpool writes out the *core consensus* sequence from the cut read clusters as well as *padding sequence*. The padding is taken from the reads extending the longest beyond the core consensus using their original, non-cut sequence. This padding stabilizes global alignment and preserves flanking context but is ignored for later variant extraction.

2.1.4 From SV primitives to SV patterns (Part 3)

In this part, Svirlpool first aligns the padded consensus sequences to the target reference using minimap2. It then considers the alignments of the core consensus sequences and parses variation in this core into **SV primitives**.

The parsing works as follows. Contiguous insertions/deletions are interpreted as insertion (INS)/deletion (DEL) SVs. Large clipped or discordant split segments are interpreted as unresolved novel adjacencies (aka break-ends/BNDs). Svirlpool stores these SV primitives together with technical annotations such as identifier of sample, consensus sequence ID, and depth of coverage. It further annotates them with the tandem repeat annotation track. Also, Svirlpool performs an estimation of the local position noise/distortion by counting the insertion and deletion signals, weighted by their exponential distance.

The rationale behind the noise estimation is our observation that while some loci show highly concordant SV signals, others show a more distorted signal. Such regions are characterized by low sequence complexity, which leads to Nanopore sequencing errors and general alignment ambiguities, often expressed in many small insertion and deletion signals instead of compact single events. Most such regions are found in and around VNTRs. Svirlpool does not fully characterize VNTRs, but instead reports them as insertions or deletions, but also allows the subsequent analysis of the local consensus assemblies for more detailed characterizations of low complexity regions.

The next step is **converting the SV primitives into SV patterns.** It converts primitives from the same consensus assembly alignment into higher-level SV patterns using pre-defined topologies. Svirlpool uses a simple pattern matching algorithm without probabilistic interpretation to avoid overinterpretation. The resulting higher-level SV patterns represent SVs like insertions, deletions, inversions, and complex re-arrangements. SV patterns contain their original SV primitives and offer an interface to query properties of the determined SV types, e.g. length of inserted sequence.

A common special case are mixed sets of insertions and deletions which are frequent in alignments within tandem repeats. These sets are reduced to a single direction by the net signed length (insertions/deletions correspond to positive/negative net lengths). Such interspersed insertions and deletions inside annotated tandem repeats are merged within the consensus sequence of the same sample (*horizontal merging*) to a single insertion or deletion SV pattern.

The Svirlpool algorithm converts intra-alignment insertion/deletion SV patterns into simple insertions and deletions. It parses break ends to identify large insertions and deletions that do not occur within the local alignments generated by the read mapper. Also, translocations and inversions are identified from break ends. In the case of solitary break ends lacking adjacency, they are emitted as VCF unpaired break ends. Finally, any residual unused break ends are written out in one complex pattern. This preserves the raw structure for refinement by external downstream tools rather than forcing a classification that is most likely wrong.

2.1.5 Structural variant calling (on multiple samples) (Part 4)

This part consolidates the annotated per-sample SVs into final, genotyped cohort calls. This part scales from a single sample well to ten samples, but possibly more.

At first a per-sample **horizontal merging** handles the case that two SV patterns from the same consensus sequence express the same SV pattern, but one is constructed from intra-alignment signals (INS/DEL) and the other from break ends. This horizontal merging step is different from the horizontal merging of the SV patterns, where only INS and DEL repeats were merged. As described above, this commonly occurs in the case of tandem repeats or other challenging reference regions. The merging takes one or more SV patterns and merges them into one SV composite putatively describing the same underlying variant. Each SV composite is a container of its underlying SV patterns. Most SV composites, however, are created from a single SV pattern.

After performing merging within the signals of each single consensus sequence of one sample, Svirlpool performs the **vertical merging of SV composites** across original consensus sequences and samples into multi-sample composites. For this, it compares the distance of the signals on the reference genome, the size of the SVs, and a k-mer based similarity measure. The distance and size comparisons are scaled by the noise distortions from the aligned trimmed reads, and by the sequence complexity of the inserted or deleted sequences, if feasible.

Svirlpool genotypes each merged SV per sample by intersecting the sample's set of SV-supporting read names with a per-sample coverage tree queried over a ±500 bp window around the SV locus on the reference. This yields the number of alt-supporting reads and a number of total reads.

Svirlpool uses the most common copy number (CN) overlapping the locus. Genotype likelihoods are computed under a binomial emission model with the expected alt fraction set to (the number of alt copies)/CN, with clipping to 0.01 and 0.99 for 0 and 1, respectively. Thus, for CN=2 the

states are 0/0 (p=0.01), 0/1 (0.45), 1/1 (0.90); for CN=3 the states are 0/0/0 (0.01), 0/0/1 (0.33), 0/1/1 (0.67), 1/1/1 (0.99); and analogously for higher ploidies. Likelihoods are normalized to posteriors assuming uniform priors across the enumerated genotype states, and the maximum-posterior genotype is reported. Genotype quality (GQ) is computed as −10 log10(1 − Pmax) and capped at 60, with Pmax signifying the probability of the chosen genotype.

Records are flagged PASS if at least one sample has three or more supporting reads; otherwise, they are flagged LowQual. Breakpoint precision is approximated by repeat content: PRECISE if none of the contributing primitives overlap annotated repeats, otherwise IMPRECISE.

2.1.6 Genome Annotation Data

In addition to the samples' alignments in BAM format (Li et al., 2009), Svirlpool needs the reference in FASTA format with annotations in BED format (James Kent et al., 2002). We use the annotations by Pacific Biosciences, public with the PBSV software for **tandem repeats** (see *External Resources*). Together with our software, we provide a script to generate all genome regions of **mononucleotide runs** of length 6 and more. We create a list of all **unique regions** by subtracting the following from the whole genome: (a) tandem repeats from above, (b) mononucleotide runs from above, (c) 100mer mappability less than 1.0 (Pockrandt et al., 2020), and (d) centromere annotations (Smit, AFA, Hubley, n.d.). You can find extended information on the annotation data we used for this manuscript in section *External Resources*.

2.1.7 Implementation and Practical Notes

Svrilpool's ource code is made available as open source under the MIT license and can be found on Github together with instructions on how to run it on a sample dataset: https://github.com/bihealth/svirlpool. The software was implemented in Python 3.12 using the following libraries: biopython v.1.83, intervaltree v.3.1.0, matplotlib v.3.8.4, matplotlib-base v.3.8.4, numpy v.1.26.4, pandas v.2.2.2, plotly v.5.22.0, pysam v.0.22.1, scipy v.1.13.1. Further external software used: lamassemble v.1.7.2, samtools v.1.20, mosdepth 0.3.8, bedtools v.2.32.1, bcftools v.1.20. Between different parts of the algorithm, data is persisted as lightweight SQLite database files allowing fast downstream lookup. Svirlpool expects ONT alignments in BAM format by the Minimap2 aligner (version 2.2 and above) and creates VCF v. 4.2 files (*The Variant Call Format (VCF) Version 4.2 Specification [PDF]*, n.d.). A detailed explanation of the output can be found on the Github page referred to above.

During development and for benchmarking we used data from legacy sequencing kits on R9 flowcells. We expect read quality to increase with further versions and thus generally better results. For older chemistry versions, please contact the authors for help with parameter optimization.

2.2 Benchmark and Evaluation

2.2.1 Data

We used two publicly available real world Nanopore sequencing data sets. At first, the Ashkenazi Jewish trio, with the child HG002, the father HG003, and the mother HG004. The trio

was sequenced with two different setups: For HG002 the flowcell FLO-MIN107 and sequencing kit - SQK-RAD003 were used. For the parents, Flowcell FLO-PRO002 and kit SQK-LSK109 were used. Basecalling was done with Guppy v.3.2.4 for HG002 and v.3.2.5 for the parents' samples. The sequencing reads data are freely available in fastq format (see *External Resources*).
We aligned the already base called ultra long ONT reads to HG19 and HS1 with minimap (version 2.28-r1209) and arguments: `-x map-ont.` Second, the Platinum Pedigreee family, of which 10 samples from three generations in six trios are publicly available, were all sequenced with the same setup: flowcell FLO-PRO002 R9.4.1 and sequencing kit SQK-ULK001, then base called with Guppy v. 6.3.7. The exact sequencing specifications can be found in (Kronenberg et al., 2024). We aligned the reads to HS1 with minimap (version 2.28-r1209) and arguments: `-x map-ont.`

We subsampled all sequencing data to 30x, 20x, and 10x to approximate situations closer to our expectations in the clinical and research application based on the average depth of coverage that was calculated with mosdepth excluding secondary alignments by setting the minimum mapq to 1 (version 0.3.8: `mosdepth --no-per-base --fast-mode --mapq 1`) and samtools (version 1.20: `samtools view --subsample fraction--subsample-seed depth`), with 'fraction' chosen to approximate the target coverage and as seed we chose the target coverage, to have different subsets of reads after each sub sampling.

2.2.2 Evaluation Strategy

Sniffles (v. 2.4) offers inbuilt multi-sample variant calling that facilitates weak signals across samples to increase the sensitivity in such cases. We know of no other method that is validated on Nanoporea data with this feature, so we compared our results with Sniffles results.

We used the Genome in a Bottle SV benchmark (Zook et al., 2020) and compared the truth set with truvari version 4.3.1 (English et al., n.d.). We tested across the tier 1 (T1) high confidence regions and the canonical chromosome regions (chr1-22, X, Y) of HG19 and across subsampled alignments with average depths of coverage of 30x, 20x, and 10x using the HG002 ultra long ONT read data. We additionally ran the GiaB SV benchmark on the HG002 variant calls after doing multi-sample SV calling for the trio (HG002, HG003, HG004), after filtering the variants to represent only SV calls where HG002 has at least one alt call. We used this data to connect the GiaB benchmark results with the Mendelian consistency tests.

The Mendelian consistency tests were applied to both the HG002 trio and the six trios of the Platinum Pedigree family, which consist of ten related individuals (citation). In each trio we force-called SVs, so that genotypes like './.' were treated as '0/0'. The Mendelian consistency test checks if the genotypes can be explained by inheritance, e.g. an alt variant in the child must be found in at least one parent or in both parents in homozygous cases. The consistency test is generalized to hemizygous variants, duplications, and triplications.

We ran the multi-sample SV calling mode with all ten available family members' data from the Platinum Pedigree data set at a mean 30x, 20x, and 10x depth of coverage on the HS1 reference.

3 Results

Table 1 shows the overall summary of results of Sniffles and Svirlpool on the GiaB dataset as evaluated by truvari for both the high-confidence T1 regions and the more comprehensive canonical regions (Chr1-22, X, Y). On T1 regions, Sniffles shows better performance with Svirlpool being within about 1.5 percentage points of Sniffles' results. On the canonical regions, there is no clear difference with Sniffles showing slightly higher recall and genotype concordance and Svirlpool having higher precision and f1 score. We have run the same benchmark for the variants that were generated with both tools' multi-sample mode, after filtering them for the HG002 variants that have at least one alt allele. The results are very similar to these here and are presented in Table **S1**.

Table 1 Results of Svirlpool and Sniffles on the Tier 1 (T1) subset of the GiaB benchmark dataset and the full dataset at coverage 30x.

tool	dataset	precision	recall	f1	FP	FN	TP	GT concordance
GiaB Tier 1 Regions								
Sniffles	T1	**95.83%**	**97.61%**	**96.71%**	**410**	**230**	**9,411**	**97.72%**
Svirlpool	T1	94.33%	96.59%	95.44%	560	329	9,312	96.91%
GiaB Canonical Regions								
Sniffles	canonical	53.80%	**94.10%**	68.45%	10,294	**752**	**11,985**	**95.74%**
Svirlpool	canonical	**55.47%**	91.35%	**69.02%**	**9,342**	1102	11,635	95.01%

Figure **2** shows the Mendelian consistency results. Sub Figure **2A** shows higher consistency of Svirlpool than Sniffles across all data sets and references. The reference genome also takes a strong effect in the results, but the HG002 trio data seems to allow for better results on the HG19 reference than on HS1. The absolute numbers of reported SVs are higher for Sniffles than for Svirlpool because Sniffles merges fewer calls in the trio. Any failure to merge such a sample results in two or three separate SV calls. The exact numbers can be found in Table **S2**.

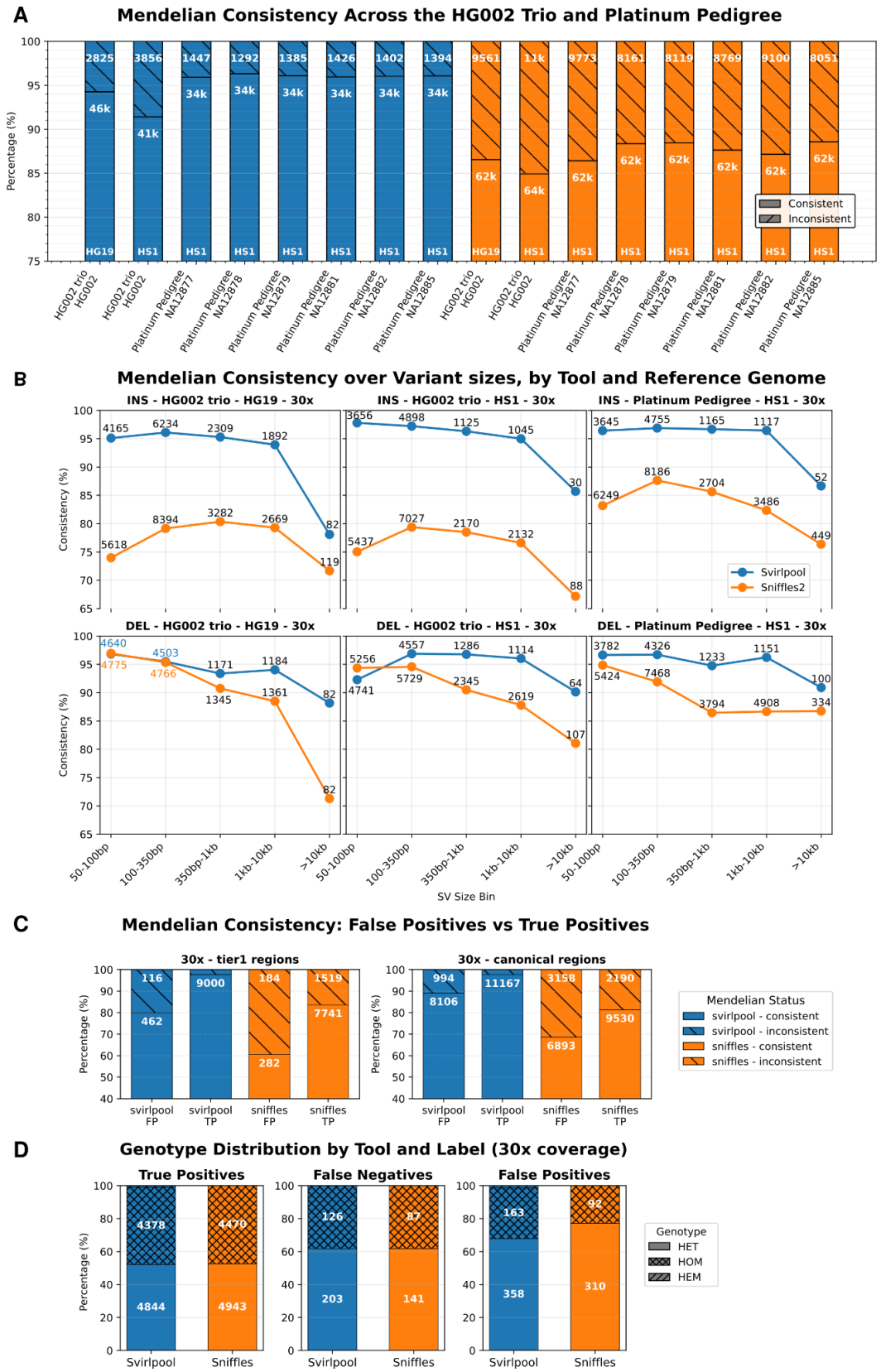

Figure 2 Mendelian consistency results of both Svirlpool and Sniffles. Throughout the figure, Sniffles results are indicated in orange and Svirlpool results are shown in blue. Sub Figure **A** shows a bar plot of the results of all SV

calls on all canonical regions on HG19 and HS1 for both tools for the HG002 trio and the Platinum Pedigree family trios. The former is shown for HG19 and HS1, and the latter for HS1 only. Sub figure **B** shows plots of the stratified SV calls by size. The top row shows plots for insertions, the bottom row for deletions. The annotated numbers show the absolute numbers, while the consistency axis scales in percentages. Sub figure **C** shows stacked bar plots of the proportions of consistent and inconsistent SV calls of HG002 given their labelling as false positive or true positive. The annotated numbers indicate the absolute counts. In sub figure **D** there are three plots showing the proportion of heterozygous (HET), homozygous (HOM), and hemizygous (HEM) variants, given a classification of the call as true positives (left sub plot), false negatives (middle plot), and false positives (right plot).

Sub figure **2B** shows the Mendelian consistency for insertions and deletions, split by sample and stratified by size of SV. Svirlpool generally performs substantially better than Sniffles. The difference is more pronounced for insertions. For deletions, the advantage of Svirlpool increases for larger sizes. We observe a generally high Mendelian consistency of more than 95% for Svirlpool for variant sizes of up to 10kb, while the insertions called by Sniffles range mostly below 80% for the HG002 trio and around 85% for the Platinum Pedigree trios. Sniffle's results show a higher bias given SV sizes than Svirlpool's results, where Sniffles seem to capture shared insertions best with sizes between 100 bp and 10 kb. In the case of deletions, Sniffles most reliably identified smaller deletions than larger ones. Although the Mendelian consistency is equal for Svirlpool and Sniffles in the case of deletions of up to 350 bp in size for the HG19 reference, there is a clear drop to below 93% for Svirlpool on the HS1 reference of the HG002 trio in 50 to 100 bp deletions. Both tools show better results on the Platinum Pedigree data than on the HG002 trio data. The results for insertions at coverages 20x and 10x can be found in the supplement in figure **S1**, and for deletions in figure **S2**.

Sub figure **2C** shows data from the HG002 trio calls that were filtered to contain only insertions and deletions that have at least one alt allele in HG002. These calls were then compared with truvari and the GiaB benchmark with parameters set to `--passonly` and `--sizemin` 50. The resulting calls connect the Mendelian consistency with the GiaB benchmark results. For all positive calls we could test the Mendelian consistency. On the T1 regions (left sub figure), in the case of Svilpool's results, we found 9000 out of 9213 (97.7%) of true positive calls Mendelian consistent and 462 out of 578 false positives (79.9%) Mendelian consistent, while the Sniffles results showed 7741 out of 9260 (84%) and 282 out of 466 (60.5%), respectively. We ran the same test for the canonical HG19 regions (chr1-22, X,Y), in the right sub figure, and found overall equal or higher consistencies, except for sniffles TPs, which are a little less Mendelian consistent than on the T1 regions. The results for 20x and 10x coverages are presented in figure **S3**.

In the final sub figure **2D**, we present the compositions of zygosity (heterozygous, homozygous, and hemizygous) of each label class (TP, FN, FP) within the same data as in sub figure **2C**. We found that Sniffles did not report any hemizygous variant calls, whereas Svirlpool generated 89. Both Svirlpool and Sniffles have a fraction of heterozygous calls of about 52% in the T1 regions among the true positives. The fraction of heterozygous calls increased to about 62% for both tools in the false positives, but then showed a difference in the false positives, where Sniffles had a fraction of 77% and Svirlpool 68%. The pattern is repeated in the canonical regions, which can be seen in Table **S3**.

We have tested to run Svirlpool on the GiaB HG002 trio (on HG19 and HS1) and the 10 freely available samples of the Platinum Pedigree (on HS1). Svirlpool processed the data (downsampled to 30x, 20x, and 10x coverage) and terminated in all cases within 24 hours on our hardware. The resource consumption varied depending on dataset and sample stage. The HG002 trio (HG002, HG003, HG004) at 30x on HG19 terminated per sample within 9, 12, and 23 hours given 32 cpu cores and consumed below 78 GB of memory. An exact table of used hardware and running times can be found in supplementary table **S5**.

4 Discussion

We developed Svirlpool as a local consensus assembly method to determine sequence properties from the underlying long read data. This procedure is computationally expensive but enables the use of consensus sequences across samples. For example, this allows comparing sequences directly when finding shared variants among individuals or downstream processing to analyse tandem repeat sequences or the contexts of more complex genomic re-arrangements.

In a direct comparison of Svirlpool with Sniffles on the widely used GiaB benchmark, we found Svirlpool to be competitive with only 1 percentage point less recall and 1.5 percentage points less precision on the tier 1 regions (T1, marked 'high confidence'). On all regions of HG19, both tools show similar results. Although this benchmark can be considered aged by the time of writing, we consider it as a useful baseline to see whether a new method can reproduce well-curated results from earlier studies.

We used consistency of inheritance ("Mendelian") tests as an additional benchmark approach that does not rely on external methods to generate a ground truth dataset and the inevitable implied bias and is purely data driven. Instead, a tested method only needs to offer multi-sample SV calling with genotyping, so that shared variants are presented as a single output record with per-sample genotypes. We force-called all variants to discriminate all results into the two classes 'consistent' and 'inconsistent'. Although this benchmark cannot directly evaluate the recall of a method, it implies important information about a tool's performance when the analysis of shared variants and the precise genotype is of importance. This is especially relevant in clinical settings, e.g., in rare disease genetics where both the reliable identification of *de novo* variants and genotyping of variants in a trio is key. Furthermore, the analysis of tumor-normal pairs is only feasible with reliable detection of shared variants. Although we have not yet optimized Svirlpool for somatic SV calling, we suspect the same advantage of local assembly methods in such cases as we see for trio variant calling.

In our benchmark across the HG002 trio and the Platinum Pedigree family, Svirlpool achieved a Mendelian Consistency (MC) of constantly over 95% at 30x coverage and SVs smaller than 10kb for both deletions and insertions. Sniffles had a generally lower MC for deletions but especially missed a lot of shared insertions and dropped below 80% on the HG002 trio data and below 88% on the platinum Pedigree data insertions. On the latter, Svirlpool's results were

consistently above 96%. Only SVs larger than 10kb still pose a challenge for Svirlpool and Sniffles, where both tools' MCs dropped significantly.

Remarkably, we observed greater differences between HG002 and his parents given the number and sites of deletion signals smaller than 100 bp. Furthermore, this trio shows to have a much lower Mendelian consistency than all the Platinum Pedigree trios with both tested methods Sniffles and Svirlool. We observed more than 3.3 times the per megabase deletions smaller than 30bp in the HG003 and HG004 aligned reads data, than in HG002, as can be seen in table **S3**. We suspect that the different flow cells and sequencing kits are causative of this difference. The bias towards more small deletions in the parent's data had a strong negative effect on the Mendelian consistency in the HG002 trio, compared to the Platinum Pedigree data, which was much more consistent. Svilpool is better suited to work with samples that were produced with different sequencing materials than Sniffles, but we would need a broader assessment to give general advice.

As discussed by (Kronenberg et al., 2024), detecting and characterizing insertions and their sequence remains a challenging task and the results of existing methods show a high discrepancy. Although insertion representations in benchmarks can be considered well-chosen on average, we assume biases in the methods used to create these data sets. We suspect this also to be true for data sets from de novo sequence assemblies, in terms of insertion location, inserted sequence, and genotype state. This motivates our choice of Mendelian Consistency for evaluating SV calling results. This allows us to be driven only by data without the influence of external tools for generating benchmark data sets or comparing tool results with expected ones, introducing their own biases. Using Mendelian Consistency allowed us to reveal the pronounced difference in insertion SV calling between tools. Further, the ability to call insertions consistently across samples justifies the computationally high cost of local assembly Svirlpool.

Overall, we introduced Svirlpool as a SV calling method with a strong focus on quality by using consensus sequence generation. Svirlpool is competitive with the leading method Sniffles on benchmark data and shows to be clearly superior when performing multi-sample variant calling, e.g., in the critically important clinical use case of analyzing trio data. Further, we expect our local assembly based method to also allow for high quality in SV calling. Finally, we see great potential for the use of the consensus sequences from local assembly in downstream applications, such as allowing sequence analysis methods for VNTRs in genes such as FMR1 (Loomis et al., 2013) or MUC1 (Vrbacká et al., 2025).

External Resources

- HG002 trio data: https://github.com/genome-in-a-bottle/giab_data_indexes (under 'Oxford Nanopore ultralong Promethion').
- Platinum Pedigree sequencing data: https://github.com/Platinum-Pedigree-Consortium/Platinum-Pedigree-Datasets
- repeat masker tracks: HS1: https://hgdownload.soe.ucsc.edu/goldenPath/hs1/bigZips/ HG38: https://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/ HG19: https://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/
- PBSV annotation tracks: https://github.com/PacificBiosciences/pbsv/tree/master/annotations

References

Aaltonen, L. A., Abascal, F., Abeshouse, A., Aburatani, H., Adams, D. J., Agrawal, N., Ahn, K. S., Ahn, S.-M., Aikata, H., Akbani, R., Akdemir, K. C., Al-Ahmadie, H., Al-Sedairy, S. T., Al-Shahrour, F., Alawi, M., Albert, M., Aldape, K., Alexandrov, L. B., Ally, A., … Consortium, T. I. P.-C. A. of W. G. (2020). Pan-cancer analysis of whole genomes. *Nature*, *578*(7793), 82–93. https://doi.org/10.1038/s41586-020-1969-6

Ahsan, M. U., Liu, Q., Perdomo, J. E., Fang, L., & Wang, K. (2023). A survey of algorithms for the detection of genomic structural variants from long-read sequencing data. *Nature Methods*, *20*(August). https://doi.org/10.1038/s41592-023-01932-w

Alena Vrbacká, Anna Přistoupilová, Kendrah O Kidd, Václav Janoušek, Martin Radina, Petr Vyleťal, Ibrahim Bitar, Viktor Stránecký, Lenka Steiner-Mrázová, Helena Trešlová, Jana Sovová, Kateřina Hodaňová, Hana Hartmannová, Dita Mušálková, Klára Svojšová, Ter, S. K. (2025). *Long-Read Sequencing of the MUC1 VNTR: Genomic Variation, Mutational Landscape, and Its Impact on ADTKD Diagnosis and Progression*. https://doi.org/doi.org/10.1101/2025.09.06.673538

Aydin, S. K., Yilmaz, K. C., & Acar, A. (2025). Benchmarking long-read structural variant calling tools and combinations for detecting somatic variants in cancer genomes. *Scientific Reports*, *15*(1), 8707. https://doi.org/10.1038/s41598-025-92750-x

Chaisson, M. J. P., Sanders, A. D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., Gardner, E. J., Rodriguez, O. L., Guo, L., Collins, R. L., Fan, X., Wen, J., Handsaker, R. E., Fairley, S., Kronenberg, Z. N., Kong, X., Hormozdiari, F., Lee, D., Wenger, A. M., … Lee, C. (2019). Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-018-08148-z

Choo, Z., Behr, J. M., Deshpande, A., Hadi, K., Yao, X., Tian, H., Takai, K., Zakusilo, G., Rosiene, J., Da, A., Paula, C., Weigelt, B., Setton, J., Riaz, N., Powell, S. N., Busam, K., Shoushtari, A. N., Ariyan, C., Reis-filho, J., … Imieliński, M. (2023). *Most large structural variants in cancer genomes can be detected without long reads*. https://doi.org/10.1038/s41588-023-01540-6

Collins, R. L., Brand, H., Karczewski, K. J., Zhao, X., Alföldi, J., Francioli, L. C., Khera, A. V., Lowther, C., Gauthier, L. D., Wang, H., Watts, N. A., Solomonson, M., O'Donnell-Luria, A., Baumann, A., Munshi, R., Walker, M., Whelan, C. W., Huang, Y., Brookings, T., … Talkowski, M. E. (2020). A structural variation reference for medical and population genetics. *Nature*, *581*(7809), 444–451. https://doi.org/10.1038/s41586-020-2287-8

Cortés-Ciriano, I., Lee, J. J. K., Xi, R., Jain, D., Jung, Y. L., Yang, L., Gordenin, D., Klimczak, L. J., Zhang, C. Z., Pellman, D. S., Akdemir, K. C., Alvarez, E. G., Baez-Ortega, A., Beroukhim, R., Boutros, P. C., Bowtell, D. D. L., Brors, B., Burns, K. H., Campbell, P. J., … von Mering, C. (2020). Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. *Nature Genetics*, *52*(3), 331–341. https://doi.org/10.1038/s41588-019-0576-7

Denti, L., Khorsand, P., Bonizzoni, P., Hormozdiari, F., & Chikhi, R. (2022). SVDSS: structural variation discovery in hard-to-call genomic regions using sample-specific strings from accurate long reads. *Nature Methods*. https://doi.org/10.1038/s41592-022-01674-1

English, A. C., Menon, V. K., Gibbs, R., Metcalf, G. A., & Sedlazeck, F. J. (n.d.). *Truvari: Refined Structural Variant Comparison Preserves Allelic Diversity*. https://doi.org/10.1101/2022.02.21.481353

Frith, M. C., Mitsuhashi, S., & Katoh, K. (2021). lamassemble: Multiple Alignment and Consensus Sequence of Long Reads. *Methods in Molecular Biology (Clifton, N.J.)*, *2231*, 135–145. https://doi.org/10.1007/978-1-0716-1036-7_9

Helal, A. A., Saad, B. T., Saad, M. T., Mosaad, G. S., & Aboshanab, K. M. (2024). Benchmarking long-read aligners and SV callers for structural variation detection in Oxford nanopore sequencing data. *Scientific Reports*, *14*(1), 6160. https://doi.org/10.1038/s41598-024-56604-2

James Kent, W., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. (2002). The human genome browser at UCSC. *Genome Research*, *12*(6), 996–1006. https://doi.org/10.1101/gr.229102

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

Kronenberg, Z., Nolan, C., Porubsky, D., Mokveld, T., Rowell, W. J., Dolzhenko, E., Chang, P., Holt, J. M., Saunders, C. T., Olson, D., Mcgee, S., Guarracino, A., Koundinya, N., Harvey, W. T., Scott, W., Munson, K. M., Hoekzema, K., Chua, K. P., Fanslow, C., … Eberle, M. A. (2024). *The Platinum Pedigree : A long-read benchmark for genetic variants*. 1–40.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Liu, Z., Xie, Z., & Li, M. (2024). Comprehensive and deep evaluation of structural variation detection pipelines with third-generation sequencing data. *Genome Biology*, *25*(1), 188. https://doi.org/10.1186/s13059-024-03324-5

Logsdon, G. A., Vollger, M. R., & Eichler, E. E. (2020). Long-read human genome sequencing and its applications. *Nature Reviews Genetics*, *21*(10), 597–614. https://doi.org/10.1038/s41576-020-0236-x

Loomis, E. W., Eid, J. S., Peluso, P., Yin, J., Hickey, L., Rank, D., McCalmon, S., Hagerman, R. J., Tassone, F., & Hagerman, P. J. (2013). Sequencing the unsequenceable: Expanded CGG-repeat alleles of the fragile X gene. *Genome Research*, *23*(1), 121–128. https://doi.org/10.1101/gr.141705.112

Mikheenko, A., Prjibelski, A. D., Joglekar, A., & Tilgner, H. U. (2022). Sequencing of individual barcoded cDNAs using Pacific Biosciences and Oxford Nanopore Technologies reveals

platform-specific error patterns. *Genome Research*, *32*(4), 726–737. https://doi.org/10.1101/gr.276405.121

Park, G., An, H., Luo, H., & Park, J. (2025). NanoMnT: an STR analysis tool for Oxford Nanopore sequencing data driven by a comprehensive analysis of error profile in STR regions. *GigaScience*, *14*, giaf013. https://doi.org/10.1093/gigascience/giaf013

Pockrandt, C., Alzamel, M., Iliopoulos, C. S., & Reinert, K. (2020). GenMap: ultra-fast computation of genome mappability. *Bioinformatics*, *April*, 1–6. https://doi.org/10.1093/bioinformatics/btaa222

Sacristán-Horcajada, E., González-De La Fuente, S., Peiró-Pastor, R., Carrasco-Ramiro, F., Amils, R., Requena, J. M., Berenguer, J., & Aguado, B. (2021). ARAMIS: From systematic errors of NGS long reads to accurate assemblies. *Briefings in Bioinformatics*, *22*(6), bbab170. https://doi.org/10.1093/bib/bbab170

Saunders, C. T., Holt, J. M., Baker, D. N., Lake, J. A., Belyeu, J. R., Kronenberg, Z., Rowell, W. J., & Eberle, M. A. (2025). Sawfish: improving long-read structural variant discovery and genotyping with local haplotype modeling. *Bioinformatics*, *41*(4), btaf136. https://doi.org/10.1093/bioinformatics/btaf136

Smit, AFA, Hubley, R. & G. (n.d.). *RepeatMasker Open-4.0* (p. http://www.repeatmasker.org).

Smolka, M., Paulin, L. F., Grochowski, C. M., Horner, D. W., Mahmoud, M., Behera, S., Kalef-Ezra, E., Gandhi, M., Hong, K., Pehlivan, D., Scholz, S. W., Carvalho, C. M. B., Proukakis, C., & Sedlazeck, F. J. (2024). Detection of mosaic and population-level structural variants with Sniffles2. *Nature Biotechnology*. https://doi.org/10.1038/s41587-023-02024-y

The Variant Call Format (VCF) Version 4.2 Specification [PDF]. (n.d.). https://samtools.github.io/hts-specs/VCFv4.2.pdf

Vialle, R. A., de Paiva Lopes, K., Li, Y., Ng, B., Schneider, J. A., Buchman, A. S., Wang, Y., Farfel, J. M., Barnes, L. L., Wingo, A. P., Wingo, T. S., Seyfried, N. T., De Jager, P. L., Gaiteri, C., Tasaki, S., & Bennett, D. A. (2025). Structural variants linked to Alzheimer's disease and other common age-related clinical and neuropathologic traits. *Genome Medicine*, *17*(1), 20. https://doi.org/10.1186/s13073-025-01444-6

Wang, W. J., Li, L. Y., & Cui, J. W. (2020). Chromosome structural variation in tumorigenesis: mechanisms of formation and carcinogenesis. *Epigenetics and Chromatin*, *13*(1), 1–17. https://doi.org/10.1186/s13072-020-00371-7

Zook, J. M., Hansen, N. F., Olson, N. D., Chapman, L., Mullikin, J. C., Xiao, C., Sherry, S., Koren, S., Phillippy, A. M., Boutros, P. C., Sahraeian, S. M. E., Huang, V., Rouette, A., Alexander, N., Mason, C. E., Hajirasouliha, I., Ricketts, C., Lee, J., Tearle, R., … Salit, M. (2020). A robust benchmark for detection of germline large deletions and insertions. *Nature Biotechnology*, *38*(11), 1347–1355. https://doi.org/10.1038/s41587-020-0538-8