

Supplementary Materials for

Learning alignment topology improves structural variant discovery from long-read sequencing

Xiang Chen*

*Corresponding author. Email: xiang-chen@zju.edu.cn

(This PDF file includes Note S1 to S9 and Tables S1 to S17)

Note S1. Signature Feature Vectorization

Long-read alignment files (BAM format) are systematically scanned within overlapping genomic sliding windows (200bp) to extract structural variant (SV) signatures from split alignments, soft/hard-clipped tails, and intra-read insertion/deletion operations. To comprehensively encode the genomic, topological, and sequence-level contexts of each candidate signature node i , SGraSV constructs an explicit, fixed-length 32-dimensional feature vector $x_i \in \mathbb{R}^{32}$. These 32 features are partitioned into four functional modules:

1. **Positional and Alignment Geometry Features (12 dimensions, $x_{1-12}^{(1)}$):** Standard read alignment attributes, including primary/secondary mapping quality (MAPQ), normalized genomic start and end coordinates relative to the candidate window center, CIGAR operation lengths (deletion/insertion span), directional strand orientation (+/-), primary read coverage depth ratio relative to the local genomic background, supplementary alignment indicator, and distance to the nearest mapped neighbor read.
2. **Signal Support and Breakpoint Consistency Features (10 dimensions, $x_{13-22}^{(2)}$):** Local evidence accumulation metrics, including total supporting read count, soft-clip tail length, split-alignment junction distance, spatial dispersion variance of clipped breakpoints (σ_{pos}^2), read length distribution statistics, and proportion of reads harboring concordant strand orientations.
3. **Dynamic Sequence-Level Modulation Features (5 dimensions, $x_{23-27}^{(3)}$):** Sequence-level identity features designed to resolve ambiguous alignment jitter in repeat-rich regions. This includes:
 - Local dynamic Smith-Waterman re-alignment identity (S_{sw} , 2 dimensions): Re-alignment 198027 scores of soft-clipped read ends aligned against both the upstream (+200bp) and downstream (-200bp) reference genomic flanks.
 - K -mer sequence frequency entropy (H_k , 2 dimensions): Calculated over a 200bp window surrounding the breakpoint using 3-mer frequency distributions ($H_k = -\sum p \log_2 p$) to capture local low-complexity/tandem-repeat micro-environments in both read and reference sequences.
 - GC-content bias ratio (1 dimension): Calculated as the GC percentage of the candidate read region relative to the local reference background.
4. **Population Allele and Reference Annotation Features (5 dimensions, $x_{28-32}^{(4)}$):** Genomic background context features, including 1000 Genomes Project population allele frequency prior, overlapping Simple Tandem Repeat (STR) count (annotated via TRF), segmental duplication density score, mappability score (100-mer uniqueness), and distance to the nearest annotated repeat element (e.g., Alu, LINE1, SVA).

Note S2: Signature Graph Construction and Topological Clustering

To consolidate fragmented read-level signal vectors into unified candidate loci without relying on predefined genomic sliding-window sizes, SGraSV constructs a directional k -nearest neighbor (KNN) topological signature graph $G = (V, E)$ in the continuous 128-dimensional VICReg latent space.

1. **Graph Node and Edge Initialization:** Each extracted read-level SV signature i is represented as a node $v_i \in V$, associated with its dense latent embedding vector $z_i \in \mathbb{R}^{128}$. An edge $e_{ij} \in E$ is constructed between node v_i and node v_j if and only if v_j belongs to the k -nearest neighbors of v_i ($k = 15$) based on Euclidean distance in the latent manifold space. Edge weights w_{ij} are inversely scaled to their latent metric distance:

$$w_{ij} = \exp\left(-\frac{\|z_i - z_j\|_2^2}{2\sigma^2}\right)$$

where σ represents the local scaling bandwidth computed as the median distance to the k -th neighbor across all nodes in the candidate locus.

2. **Topological Community Partitioning:** Rather than partitioning reads by spatial chromosomal coordinates—which fails in dense repeat regions—SGraSV partitions the signature graph G into dense communities using Louvain modularity optimization. The objective is to maximize the modularity score Q :

$$Q = \frac{1}{2m} \sum_{ij} \left[w_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j)$$

where $m = \frac{1}{2} \sum_{ij} w_{ij}$ is the total sum of edge weights, $k_i = \sum_j w_{ij}$ is the degree sum of node i , c_i is the assigned community label of node i , and $\delta(c_i, c_j)$ is the Kronecker delta (yielding 1 if $c_i = c_j$ and 0 otherwise).

3. **Consensus Candidate Aggregation:** Each partitioned community $C_k \subset V$ with node support $|C_k| \geq 3$ is aggregated into a single candidate structural variant. The candidate's feature tensor is derived by computing the element-wise median latent vector $\bar{z}_{C_k} = \text{median}(\{z_i | v_i \in C_k\})$, while tentative spatial boundaries (Pos_{start}, Pos_{end}) are initialized from the 10th and 90th percentiles of the mapped split-read positions within C_k . This topological clustering decouples signature aggregation from spatial coordinates, directly resolving over-segmentation artifacts in complex genomic domains.

Note S3. Candidate-Level Feature Vector (42D Vector) Architecture

Following topological graph partitioning, connected components are extracted to form candidate SV calls. For each candidate cluster k , SGraSV synthesizes the node-level embeddings and cluster-level topology into a 42-dimensional candidate-level feature vector $\hat{x}_k \in \mathbb{R}^{42}$. As detailed in Table S2, the 42D vector $\hat{x}_k$ consolidates six primary modules:

1. **Read Support & Scale Attributes ($x_1 - x_6$):** Summarized SV length L_{SV} , span, read support count, support ratio, and mean read length.

2. Alignment Geometry Aggregation ($x_7 - x_{13}$): Median primary/supplementary MAPQ, mean soft-clip/hard-clip fractions, mean alignment identity, and spatial split-read distances aggregated across supporting nodes.
3. Cluster Compactness & Dispersion Metrics ($x_{14} - x_{20}$): Coordinate dispersions ($\sigma_{start}^2, \sigma_{end}^2$), length coefficient of variation (CV_{len}), clip length variance (σ_{clip}^2), and embedding consistency (mean pairwise cosine similarity among node embeddings $\bar{S}_{cos}$).
4. Multi-Source & Alignment Evidence Ratios ($x_{21} - x_{24}$): Ratios of split-read (SA tag) vs. intra-read CIGAR (I/D) signatures, source diversity entropy, and strand consistency.
5. Global Coverage & Sequencing Environment ($x_{25} - x_{36}$): Global depth, 5-bin local depth encoding, 4-bin platform flags (PacBio CLR/CCS/ONT/Unknown), real vs. simulated domain flag, and strict-threshold violation counters.
6. Variant Type One-Hot Encoding ($x_{37} - x_{42}$): One-hot flags for DEL, INS, DUP, INV, TRA, and complex topologies.

Note S4. Structural Variant Evaluation and Matching Criteria

Structural variant (SV) calls were evaluated using established comparison frameworks, with evaluation parameters selected to account for differences in breakpoint representation and sequence reporting among SV callers. For evaluations of real sequencing data, Truvari v4.1.0 was used. The `--includebed` option was applied when a predefined high-confidence evaluation region was available, restricting the comparison to the specified genomic regions. Variants outside the defined evaluation regions were not manually removed from the input callsets but were excluded during evaluation through the corresponding BED-file restriction.

A breakpoint-distance tolerance of 100 bp (`-r 100`) was used for the primary Truvari-based evaluations. Different SV callers can report slightly different breakpoint positions because of differences in variant representation, alignment characteristics, and local sequence context. This is particularly relevant for insertions and larger SVs, for which breakpoint placement can vary even when callers identify the same underlying event. The 100-bp tolerance was therefore used to accommodate minor differences in breakpoint representation while maintaining positional consistency for SV matching.

Sequence-similarity matching was disabled by setting `-P 0.00`. This setting was chosen because SV callers differ in how insertion sequences are represented. Some callers report insertions symbolically, whereas others provide explicit inserted sequences. Requiring sequence similarity would therefore make the evaluation dependent on the output representation of individual callers and could disadvantage methods that report symbolic insertions. With sequence-similarity matching disabled, symbolic and sequence-resolved insertions were evaluated using the same position-based matching criterion.

For simulated datasets, SV calls were evaluated using the `eval` subcommand of SURVIVOR, with a breakpoint-distance tolerance of 50 bp (`-r 50`). The same evaluation principle was applied across datasets: SVs were considered matching when their reported breakpoints fell within the specified positional tolerance, while avoiding additional sequence-based requirements that could introduce representation-dependent differences between callers.

The corresponding commands and dataset-specific parameters are provided below:

For real-data evaluations using Truvari:

```
truvari bench \  
  -b <truth.vcf.gz> \  
  -c <query.vcf.gz> \  
  --includebed <evaluation_regions.bed> \  
  -r 100 \  
  -P 0.00 \  
  -o <output_directory>
```

For evaluations without a predefined high-confidence region:

```
truvari bench \  
  -b <truth.vcf.gz> \  
  -c <query.vcf.gz> \  
  -r 100 \  
  -P 0.00 \  
  -o <output_directory>
```

For simulated datasets using SURVIVOR:

```
SURVIVOR eval \  
  <truth.vcf> \  
  <query.vcf> \  
  50 \  
  ... \  
  <output_prefix>
```

Note S5. Tool Execution Commands and Parameters

All alignment BAM files were generated using minimap2 (v2.26) against GRCh38 (`-ax map-hifi` for PacBio CCS; `-ax map-ont` for ONT). All SV calling tools were run independently on coordinate-sorted and indexed BAM files using the GRCh38 reference genome. The same input alignments and reference genome were used across calling tools. Unless otherwise specified, tool-specific default parameters were used. Platform-specific parameters were applied only where recommended by the corresponding calling tools.

Because cuteSV uses different clustering parameter settings for CLR, CCS/HiFi, and ONT data; the publicly available usage examples (<https://github.com/tjiangHIT/cuteSV>) also provide platform-specific parameter settings:

```
# cuteSV v2.0.3  
cuteSV \  
  <input.bam> \  
  <reference.fa> \  
  <output.vcf> \  
  <work_dir> \  
  --threads <threads> \  
  --genotype \  
  --report_readid
```

For single-sample germline SV calling, the following command was used. If multi-sample calling was subsequently performed, an `.snf` file was also generated using the `--snf <output.snf>` option (<https://github.com/fritzsedlazeck/sniffles>):

```
# Sniffles2 v2.2
sniffles \
  --input <input.bam> \
  --vcf <output.vcf.gz> \
  --reference <reference.fa> \
  --threads <threads> \
  --output-rnames
```

The alignment mode of SVIM v2.0.0 supports multiple parameters (<https://github.com/eldariont/svim/>) such as `--min_mapq`, `--min_sv_size`, `--max_sv_size`, and cluster distance:

```
# SVIM v2.0.0
svim alignment \
  <output_directory> \
  <input.bam> \
  --cores <threads>
```

Note S6. Multi-type SV Detection on Real and Synthetic Data

Complex structural rearrangements, such as duplications (DUPs) and inversions (INVs), present persistent alignment anomalies. Duplications generate duplicate mappings or conflicting alignment orientations, while inversions involve inverted alignments flanked by micro-homologies or micro-insertions. Traditional callers reliance on linear reference alignment often leads to truncated breakpoint signatures. SGrSV addresses these challenges by transforming read alignment topologies into graph representations and leveraging self-supervised representation learning to capture structural breakpoint connectivity. To ensure full transparency, reproducible execution, and statistical rigor, this note details the simulation framework, real-data filtering protocol, command-line parameter configurations, and statistical testing methodologies.

Synthetic evaluation datasets were constructed to benchmark tool performance across controlled coverage depths and error profiles for both PacBio CCS and ONT platforms.

1. Genome Template and Variant Spike-in:

Human reference genome GRCh38 (primary assembly) served as the simulation backbone. DUP and INV variants ranging from *50bp* to *100kb* were generated using VISOR (v1.1.2). Variants were placed randomly across autosomes while excluding assembly gaps (N regions) and high-density segmental duplications to avoid simulator artifacts.

2. Read Simulation Parameters:

- PacBio CCS: Simulated using PBSIM2 (v2.0.1) configured with the `--model-qc P6C4` profile, targeting an average read length of *15kb*, mean read accuracy of 99.5%, and a uniform sequencing depth of 30×.
- ONT: Simulated using Badread (v0.2.0) with parameters: `--identity 90, 98, 95` (representing variable read accuracies), `--length 15000, 10000` (mean length *15kb*), and `--glitch-quant 100, 100, 10` to model random chimeric reads and pore-stall artifacts at 30× depth.

3. **Replication:** 10 independent genomic simulation blocks ($n=10$) were generated per platform to derive variance metrics ($\pm SD$).

Real sequencing evaluations utilized high-depth PacBio CCS ($>30\times$) and ONT ($>30\times$) datasets from 9 individuals across 3 HGSVC pedigree trios, CHS Trio: HG00512 (Father), HG00513 (Mother), HG00514 (Child), PUR Trio: HG00731 (Father), HG00732 (Mother), HG00733 (Child), and YRI Trio: NA19238 (Father), NA19239 (Mother), NA19240 (Child). To eliminate platform-specific false positives and sequencing artifacts from raw callsets, we established a pedigree-constrained benchmark curation pipeline:

1. **Source Truth Callset:** Benchmark variants were extracted from the 1KGP high-coverage Illumina integrated structural variant callset (Chaisson et al. 2019, <https://doi.org/10.1038/s41467-018-08148-z>).
2. **Genotype Consistency & Private Variant Filtering:** For each trio $T = \{F, M, C\}$ (Father, Mother, Child), an SV variant v was retained in the evaluation truth set if and only if it satisfied the Trio-Private Condition:

$$\sum_{i \in \{F, M, C\}} \mathbb{I}(g_{i,v} > 0) = 1$$

where $g_{i,v}$ represents the genotype dosage (0 for 0/0, 1 for 0/1, 2 for 1/1) of sample i at variant locus v , and $\mathbb{I}(\cdot)$ is the indicator function.

3. **Rationale:** Variants appearing in exactly one individual within a tightly sequenced trio represent either true rare inherited variants or high-fidelity *de novo* events. Random sequencing errors or alignment artifacts rarely recur at identical coordinates with matching SV types in a single member while remaining completely absent in the other two.

Note S7. Size-Stratified Resolution Analysis and External Sample Benchmarking

To evaluate model generalizability and length-dependent sensitivity without overfitting to the HG002 training cohort, a size-stratified benchmarking framework was deployed across independent external human genomes representing diverse ancestral backgrounds.

1. **External Cohort Selection:** High-depth long-read sequencing datasets ($> 30 \times$ PacBio HiFi and ONT) for three pedigree trios from the HGSVC: Southern Han Chinese (CHS: HG00512, HG00513, HG00514), Puerto Rican (PUR: HG00731, HG00732, HG00733), and Yoruba in Ibadan, Nigeria (YRI: NA19238, NA19239, NA19240) were retrieved from the HGSVC3 release (Table S3).
2. **Size-Tier Stratification:** Variant callsets generated by GraSV and baseline callers (cuteSV, Sniffles2, SVIM) were categorized into three non-overlapping structural length intervals:
 - **Short SVs (50 – 200bp):** Represents the boundary region where alignment jitter and homopolymeric indel noise are most pronounced.
 - **Medium SVs (200 – 1000bp):** Captures polymorphic Alu retrotransposons and intermediate indels.
 - **Large SVs ($\geq 1000bp$):** Captures full-length mobile elements (L1/SVA), segmental duplications, and large chromosomal rearrangements.

3. **Stratified Evaluation Metrics:** For each sample S and size stratum $L \in \{[50, 200), [200, 1000), [\geq 1000]\}$, predicted variant callsets across the nine benchmark genomes were matched against 1KGP integrated structural variation truth sets using Truvari. Callsets within each length spectrum were evaluated separately by variant type for deletions (DELs), insertions (INSs), and overall combined structural variants. A predicted variant was classified as a true positive (TP) if it satisfied two standard matching criteria: (i) a maximum breakpoint distance offset of $< 50bp$, and (ii) a minimum sequence similarity threshold of $> 70\%$ for inserted or deleted sequences. Unmatched predictions were marked as false positives (FP), and missed benchmark variants were designated as false negatives (FN). Stratified performance metrics were calculated as:

$$Precision_{S,L} = \frac{TP_{S,L}}{TP_{S,L} + FP_{S,L}}$$

$$Recall_{S,L} = \frac{TP_{S,L}}{TP_{S,L} + FN_{S,L}}$$

$$F1_Score_{S,L} = \frac{2 \cdot Precision_{S,L} \cdot Recall_{S,L}}{Precision_{S,L} + Recall_{S,L}}$$

This stratification isolated SGraSV's spatial resolution gains specifically in the challenging 50 – 200bp window, preventing large, easily detectable variants ($> 1kb$) from masking short-variant accuracy decay.

Note S8. 1000 Genomes Project Cohort Application & MEI Annotation

To demonstrate practical utility in population genomics, SGraSV was scaled to long-read sequencing data across diverse pedigree trios from the 1000 Genomes Project (1KGP) cohort to assess lineage transmission fidelity and mobile element insertion (MEI) resolution.

1. **Pedigree Trio Mendelian Consistency Protocol:** Mendelian transmission error rates were computed across available pedigree trios (e.g., NA19240 child with parents NA19238 and NA19239). A candidate variant call v_c in the offspring genome was classified as Mendelian Consistent if a concordant variant call ($r \leq 500bp$, sequence similarity $\geq 70\%$) was present in at least one parental callset (v_{p1} or v_{p2}). The Mendelian Consistency Rate (MCR) was defined as:

$$MCR = \frac{|\{v_c \in V_{child} | v_c \cap (V_{father} \cup V_{mother}) \neq \emptyset\}|}{|V_{child}|} \times 100\%$$

Non-inherited calls occurring outside high-confidence regions were audited to distinguish true *de novo* mutations from false-positive alignment artifacts.

2. **Mobile Element Insertion (MEI) Delineation:** Inserted sequences identified by SGraSV were extracted from final VCF entries and mapped against RepeatMasker consensus libraries (v4.1.2, open-3.0 species set). Variants were annotated as retrotransposon insertions based on reciprocal overlap ($\geq 80\%$ sequence identity and $\geq 80\%$ length coverage) with standard consensus boundaries:

- **Alu elements:** Insertion lengths spanning 280 – 350bp intersecting Alu subfamilies (AluJ, AluS, AluY).
- **Line-1 (L1) elements:** Insertion lengths spanning 0.5 – 6.5kb, with full-length L1 insertions defined at ~6kb containing intact 5' untranslated regions (UTR) and target site duplications (TSD).
- **SVA elements:** Insertion lengths spanning 1.5 – 3.5kb corresponding to SINE-VNTR-Alu retrotransposons.

Kernel Density Estimation (KDE) was applied to the resulting length spectra to evaluate full-length element recovery against standard pangenome databases.

Note S9. SHAP Feature Attribution & Feature Ablation Protocol

To eliminate the "black-box" opacity of deep representation learning and quantitatively dissect feature mechanics within SGraSV, we combined KernelSHAP game-theoretic attribution with a multi-tier feature ablation protocol.

Marginal feature contributions to the 1D-CNN classifier output P_{SV} were computed using KernelSHAP over a representative background sample of 10,000 candidate loci (5,000 validated true positive SVs and 5,000 alignment false positives). For a given feature vector X , the SHAP value ϕ_i for feature dimension i is calculated as:

$$\phi_i(X) = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|! (|F| - |S| - 1)!}{|F|!} [f_x(S \cup \{i\}) - f_x(S)]$$

where F is the complete set of 42 explicit features, S is a feature subset excluding feature i , and $f_x(S)$ is the conditional expectation of the model output given the features in S (i.e. leave-one-module-out feature ablation). Features were ranked by mean absolute SHAP value ($\frac{1}{N} \sum |\phi_i|$) to construct beeswarm summary plots.

For feature attribution, SHAP were computed across 10^5 validated candidate loci randomly sampled from the held-out test set (chr13–chr22). We utilized DeepSHAP, an explicit integration of DeepLIFT and Shapley value estimation, backpropagating attribution scores through the three Conv1D–BatchNorm–ReLU blocks to assign a marginal contribution score $\psi_\theta(\hat{x}_k)$ to each feature dimension x_i ($i = 1, \dots, 42$). Baseline background distributions were derived from 1,000 representative synthetic background vectors across low-coverage and high-depth genomic regions. Absolute SHAP values were aggregated per module to rank feature groups based on their impact on candidate precision and sensitivity, specifically contrasting true positive cores against false positive boundary artifacts identified in the UMAP latent manifold.

Complementarily, the feature ablation protocol was designed around the six logical feature modules comprising $\hat{x}_k$: Read Support and Scale Attributes (Module 1, x_1 – x_6), Alignment Geometry Aggregation (Module 2, x_7 – x_{13}), Cluster Compactness and Dispersion Metrics (Module 3, x_{14} – x_{20}), Multi-Source Evidence Ratios (Module 4, x_{21} – x_{24}), Global Coverage and Platform Context (Module 5, x_{25} – x_{36}), and Variant Type One-Hot Flags (Module 6, x_{37} – x_{42}). In each ablation experiment, target module dimensions were zero-masked or replaced with background population medians during both model fine-tuning and inference. Retrained 1D CNN scorers were evaluated on the held-out test set (chr13–chr22) under

identical hyperparameter settings. Model performance across overall SV calls and individual topologies (DEL, INS, DUP, INV) was evaluated using overall classification F1-score to measure the exact performance drop ($\Delta F1$) attributable to each excised module.

Note S10. Computational Efficiency, Memory Profiling, and Parallel Scaling

Benchmarking of execution runtime, peak RAM usage (Max RSS), and CPU/GPU scaling was conducted on an Ubuntu 22.04 LTS server equipped with an Intel Xeon Gold 6330 CPU (64 threads), 256 GB RAM, and an NVIDIA RTX 4090 GPU (24 GB VRAM). Memory usage and wall-clock execution times were logged using GNU time $-v$ across thread configurations ranging from 4 to 64 threads. Threading speedup efficiency was calculated as $\text{Efficiency} = \frac{T_1}{N \cdot T_N} \times 100\%$, where T_1 and T_N represent single-thread and N -thread runtimes, respectively.

Table S1. Explicit 32-dimensional feature vector configuration of candidate signature nodes in SGraSV.

Module Index	Feature Dimension	Feature Name	Description & Mathematical Definition
Module 1: Alignment Geometry	x_1	Primary MAPQ	Primary mapping quality score (0 - 60) of the supporting read.
	x_2	Supplementary MAPQ	Mapping quality score of supplementary split-alignment segment (0 if none).
	x_3	Start Offset	Relative genomic coordinate offset of SV start position to window center (bp).
	x_4	End Offset	Relative genomic coordinate offset of SV end position to window center (bp).
	x_5	Span Length	CIGAR-derived candidate SV span length ($\log_{10}(L_{SV})$).
	x_6	Strand Orientation	Binary orientation flag (+1 for forward/forward, -1 for inverted/reverse).
	x_7	Local Depth Ratio	Ratio of local read depth to 10kb flanking genomic background depth (Cov_{local}/Cov_{bg}).
	x_8	Soft-clip Length	Length of soft-clipped tail sequence (bp).
	x_9	Hard-clip Flag	Binary indicator for hard-clipped read alignments (1 if present, 0 otherwise).
	x_{10}	Supplementary Dist	Distance between split-alignment fragments on the same read (bp).
	x_{11}	Read Length	Total length of the supporting long read ($\log_{10}(L_{read})$).
	x_{12}	Alignment Identity	CIGAR match ratio ($Matches/(Matches + Mismatches + Indels)$).
Module 2: Signal Support	x_{13}	Supporting Read Count	Raw count of reads harboring consistent structural signals in the window.

	x_{14}	Supporting Read Fraction	Ratio of supporting reads to total read depth at candidate locus.
	x_{15}	Breakpoint Variance	Spatial variance (σ_{pos}^2) of candidate start/end coordinates across supporting reads.
	x_{16}	Clip Length Variance	Variance (σ_{clip}^2) of soft-clip lengths among supporting reads.
	x_{17}	Strand Consistency	Proportion of supporting reads sharing the predominant strand orientation.
	x_{18}	Split Read Ratio	Proportion of supporting evidence originating from split alignments vs CIGAR indels.
	x_{19}	High-MAPQ Fraction	Proportion of supporting reads with $MAPQ \geq 30$.
	x_{20}	Local Coverage Drop	Relative depth decrease across the predicted inner deletion region.
	x_{21}	Local Coverage Gain	Relative depth increase across the predicted inner duplication region.
	x_{22}	Neighbor Signal Dist	Genomic distance to the nearest adjacent signature cluster (bp).
	x_{23}	Upstream SW Identity (S_{sw}^{up})	Dynamic Smith-Waterman re-alignment identity of soft-clipped tail against +200bp upstream flank.
Module 3: Sequence Modulation	x_{24}	Downstream SW Identity (S_{sw}^{down})	Dynamic Smith-Waterman re-alignment identity of soft-clipped tail against -200bp downstream flank.
	x_{25}	Read K -mer Entropy (H_k^{read})	3-mer frequency Shannon entropy ($-\sum p \log_2 p$) calculated on candidate read region (200bp).
	x_{26}	Ref K -mer Entropy (H_k^{ref})	3-mer frequency Shannon entropy calculated on reference genomic locus (200bp).
	x_{27}	Local GC Bias Ratio	GC content of read region divided by GC content of reference locus.
	x_{28}	STR Overlap Count	Count of overlapping Short Tandem Repeats (TRF annotation).

Module 4: Genomic Context	x_{29}	SegDup Score	Segmental duplication score at candidate locus (0 - 1).
	x_{30}	Mappability Score	100-mer uniqueness score of local reference region (0 - 1).
	x_{31}	Repeat Anno Distance	Distance to nearest annotated TEs (transposable elements) element (Alu, LINE1, SVA, LTR) (<i>bp</i>).
	x_{32}	1KGP Prior Allele Freq	Historical 1000 Genomes Project allele frequency at overlapping locus (0 if novel).

Table S2. Explicit 42-dimensional candidate-level feature vector configuration for 1D CNN scoring in GraSV.

Module Index	Feature Dimension	Feature Name	Description & Mathematical Definition
Module 1: Read Support & Scale Attributes	x_1	Read Support Count	Raw count of supporting reads harboring consistent signatures in the cluster.
	x_2	Cluster Node Size	Total number of candidate signature nodes forming the connected component.
	x_3	Support Ratio	Ratio of supporting reads to total local read depth at candidate locus ($\frac{N_{supp}}{Cov_{local}}$).
	x_4	Summarized Variant Length	Median estimated SV length (L_{SV}) across cluster nodes (negative for DEL per VCF spec).
	x_5	Summarized Variant Span	Median physical genomic span covered by candidate nodes ($\log_{10}(L_{span} + 1)$).
	x_6	Mean Read Length	Mean overall length of supporting long reads across cluster nodes ($\log_{10}(L_{read})$).
Module 2: Alignment Geometry Aggregation	x_7	Median Primary MAPQ	Median mapping quality score (0 - 60) of primary supporting read alignments.
	x_8	Median Supp MAPQ	Median mapping quality score of supplementary split-alignment segments.
	x_9	Mean Soft-clip Length	Mean length of soft-clipped tail sequences among candidate nodes (bp).
	x_{10}	Soft-clip Presence Fraction	Fraction of candidate nodes exhibiting soft-clipped alignments in the cluster.
	x_{11}	Hard-clip Presence Fraction	Fraction of candidate nodes exhibiting hard-clipped alignments in the cluster.
	x_{12}	Mean Alignment Identity	Average CIGAR match ratio ($\frac{matches}{matches+mismatches+indels}$) across supporting reads.
	x_{13}	Mean Supplementary Dist	Mean spatial distance between split-alignment segments on reference (bp).

Module 3: Cluster Compactness & Dispersion Metrics	x_{14}	Start Coordinate Dispersion	Spatial variance (σ_{start}^2) / dispersion of candidate start coordinates.
	x_{15}	End Coordinate Dispersion	Spatial variance (σ_{end}^2) / dispersion of candidate end coordinates.
	x_{16}	Length Coefficient of Variation	Coefficient of variation ($CV_{len} = \frac{\sigma_{len}}{\mu_{len}}$) of candidate SV length within cluster.
	x_{17}	Clip Length Variance	Variance (σ_{clip}^2) of soft-clip sequence lengths among supporting reads.
	x_{18}	Coordinate Compactness Score	Normalized measure evaluating spatial coordinate tightness of the cluster.
	x_{19}	Length Compactness Score	Normalized measure evaluating internal length consistency across nodes.
	x_{20}	Embedding Consistency	Mean pairwise cosine similarity of node embeddings ($\bar{S}_{cos}$) within the cluster.
Module 4: Multi-Source & Alignment Evidence Ratios	x_{21}	Evidence Source Diversity	Entropy/Diversity index of multi-source evidence supporting the candidate call.
	x_{22}	Split-Read Evidence Ratio	Proportion of supporting signature nodes derived from split-read (SA tag) alignments.
	x_{23}	CIGAR Evidence Ratio	Proportion of supporting signature nodes derived from intra-read CIGAR operators (I/D).
	x_{24}	Strand Consistency Ratio	Dominant strand orientation ratio (+1 forward / -1 reverse agreement) across nodes.
Module 5: Global Coverage & Sequencing Environment	x_{25}	Global Coverage Depth	Global average sequencing coverage depth across the reference genome.

Module 6: Variant Type One-Hot Encoding	$x_{26} - x_{30}$	Coverage Depth Bin (One-Hot)	5-dimensional one-hot encoding for local coverage depth stratification.
	$x_{31} - x_{34}$	Platform Flags (One-Hot)	4-dimensional one-hot encoding for sequencing platform (PacBio CLR/PacBio CCS / ONT/Unknown).
	x_{35}	Real / Simulated Domain Flag	Binary domain flag indicating real dataset (1) vs. simulated training data (0).
	x_{36}	Strict-Threshold Violations	Count/flag of strict compactness or alignment quality filtering violations.
	x_{37}	Is Deletion (DEL) Flag	One-hot binary indicator flag for Deletion call (1 if DEL, 0 otherwise).
	x_{38}	Is Insertion (INS) Flag	One-hot binary indicator flag for Insertion call (1 if INS, 0 otherwise).
	x_{39}	Is Duplication (DUP) Flag	One-hot binary indicator flag for Duplication call (1 if DUP, 0 otherwise).
	x_{40}	Is Inversion (INV) Flag	One-hot binary indicator flag for Inversion call (1 if INV, 0 otherwise).
	x_{41}	Is Translocation (TRA) Flag	One-hot binary indicator flag for Translocation call (1 if TRA, 0 otherwise).
	x_{42}	Complex Breakend Topology Flag	Binary flag indicating complex multi-breakend/unclassified SV topology.

Table S3. Public database accessions, source URLs, and file specifications for datasets utilized in SGraSV benchmarking.

Sample / Dataset ID	Data Type / Assay	Platform / Strain	Reference Genome	File Type / Version	Primary BioProject / Repository Accession	Direct Source URL / Download Identifier
HG002 (GIAB)	HiFi / CCS Long Reads	PacBio	GRCh37 / GRCh38	FASTQ / BAM	GIAB	ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG002_NA24385_son/PacBio_CCS_15kb_20kb_chemistry2/
HG002 (GIAB)	Ultra-long Reads	Oxford Nanopore	GRCh37 / GRCh38	FASTQ / BAM	GIAB	ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG002_NA24385_son/UCSC_Ultralong_OxfordNanopore_PromethION/
HG002 (GIAB)	Continuous Long Reads	PacBio	GRCh37 / GRCh38	FASTQ / BAM	GIAB	ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG002_NA24385_son/PacBio_MtSinai_NIST/
HG002 v5.0q Benchmark	High-confidence SV Calls	Benchmark Set	GRCh38 / GRCh37	VCF & Bed	GIAB	ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/AshkenazimTrio/HG002_NA24385_son/latest/
HGSVC Samples (1KGP)	HiFi / CCS Long Reads	PacBio	GRCh38/T2T	BAM (65 individuals of diverse ancestries)	1KGP	http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC3/working/
HGSVC Samples (1KGP)	Ultra-long Reads	ONT	GRCh38/T2T	BAM (65 individuals of diverse ancestries)	1KGP	ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1KG_ONT_VIENNA/
1KGP SV Callsets*	Short-read WGS	illumina	GRCh38	VCF (Integrated SV callsets)	1KGP	ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage_SV/working/20210104_JAX_Integration_13callers/

				of 3202 individuals)		
1KGP SV Callsets*	Short-read WGS	illumina	GRCh38	VCF (3 trios SV)	1KGP	ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/hgsv_sv_discovery/working/20160714_Eichler_SVs/
HGSVC3 SV Callsets*	Pangenome SV Assembly	Multi- platform	GRCh38/T2 T	VCF (SV from 65 individual genomes)	1KGP	http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC3/release/Variant_Calls/1.0/
HGSVC3 SV Callsets*	Pangenome SV Assembly	Multi- platform	GRCh38/T2 T	MEI CSV (65 individuals)	1KGP	http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC3/release/Mobile_Elements/1.0/
SURVIVOR Simulated	Synthetic Multi-type SVs	In silico Read Models	GRCh38	FASTQ & Truth VCF	SURVIVOR v1.0.7	GitHub: https://github.com/fritzsedlazeck/SURVIVOR/

Note: All raw sequencing reads and alignment files were retrieved from public open-access repositories, including the Genome in a Bottle (GIAB) Consortium FTP server and the 1000 Genomes Project (1KGP) EBI portal. GIAB benchmark evaluations were conducted using the v5.0q high-confidence structural variant callset. Multi-ancestry pangenome comparison callsets were obtained from the Human Genome Structural Variation Consortium (HGSVC3) release v1.0. All downstream processed file alignments were anchored against either human reference genome build GRCh37 or GRCh38. * Unreliable variant regions including centromeres and telomeres were discarded. The 65 individuals of diverse ancestries include sample IDs: HG00096, HG00171, HG00268, HG00358, HG00512, HG00513, HG00514, HG00731, HG00732, HG00733, HG00864, HG01114, HG01352, HG01457, HG01505, HG01573, HG01596, HG01890, HG02011, HG02018, HG02059, HG02106, HG02282, HG02492, HG02554, HG02587, HG02666, HG02769, HG02818, HG02953, HG03009, HG03065, HG03248, HG03371, HG03452, HG03456, HG03520, HG03683, HG03732, HG03807, HG04036, HG04217, NA12329, NA18534, NA18939, NA18989, NA19036, NA19129, NA19238, NA19239, NA19240, NA19317, NA19331, NA19347, NA19384, NA19434, NA19650, NA19705, NA19836, NA19983, NA20355, NA20509, NA20847, NA21487, NA24385. The three trios comprised the Southern Han Chinese trio (CHS: HG00512, HG00513, HG00514), the Puerto Rican trio (PUR: HG00731, HG00732, HG00733), and the Yoruba trio (YRI: NA19238, NA19239, NA19240).

Table S4. Benchmark performance of SGrasV and baseline callers on HG002 across selected coverage depths (GRCh37).

Seq Type	Depth	Tool	DEL			INS		
			Precision	Recall	F1	Precision	Recall	F1
ONT (Oxford Nanopore)	5x	SGrasV	0.9644	0.7829	0.8642	0.946	0.7528	0.8384
		cuteSV	0.9481	0.738	0.83	0.924	0.6952	0.7934
		Sniffles2	0.9256	0.7138	0.806	0.9052	0.679	0.7759
		SVIM	0.8759	0.7044	0.7809	0.8583	0.6762	0.7564
	10x	SGrasV	0.9725	0.8277	0.8942	0.9623	0.7951	0.8708
		cuteSV	0.951	0.79	0.8631	0.9174	0.7734	0.8393
		Sniffles2	0.933	0.7536	0.8338	0.9112	0.7455	0.8201
		SVIM	0.8951	0.739	0.8096	0.8742	0.7121	0.7849
	20x	SGrasV	0.971	0.8764	0.9213	0.9569	0.8531	0.902
		cuteSV	0.9568	0.8319	0.89	0.9376	0.7984	0.8624
		Sniffles2	0.9425	0.8192	0.8765	0.9298	0.7798	0.8483
		SVIM	0.9119	0.8171	0.8619	0.8876	0.7886	0.8352
	48x	SGrasV	0.979	0.935	0.9565	0.9566	0.897	0.9258
		cuteSV	0.9572	0.8982	0.9268	0.9208	0.8723	0.8959
		Sniffles2	0.9584	0.8741	0.9143	0.9412	0.8498	0.8932
		SVIM	0.9263	0.8674	0.8959	0.908	0.8391	0.8722
PacBio CLR	5x	SGrasV	0.9629	0.7898	0.8678	0.9462	0.7555	0.8402
		cuteSV	0.9441	0.7364	0.8274	0.923	0.7073	0.8008
		Sniffles2	0.9245	0.717	0.8076	0.9054	0.6925	0.7848
		SVIM	0.8832	0.6989	0.7803	0.8623	0.6722	0.7555
	10x	SGrasV	0.9694	0.8434	0.902	0.9549	0.8023	0.872
		cuteSV	0.9552	0.7936	0.8669	0.9314	0.7828	0.8506
		Sniffles2	0.9338	0.7715	0.8449	0.9223	0.7365	0.819

PacBio CCS	20x	SVIM	0.8916	0.7479	0.8134	0.8725	0.721	0.7896
		SGraSV	0.9763	0.8948	0.9338	0.9524	0.8552	0.9012
		cuteSV	0.9581	0.8323	0.8908	0.9378	0.7914	0.8584
		Sniffles2	0.9452	0.8233	0.8801	0.9238	0.8151	0.866
	35x	SVIM	0.9034	0.8119	0.8552	0.8805	0.7751	0.8244
		SGraSV	0.9717	0.9247	0.9476	0.9487	0.8908	0.9188
		cuteSV	0.9611	0.875	0.916	0.947	0.8479	0.8947
		Sniffles2	0.9555	0.8703	0.9109	0.9335	0.8433	0.8861
	70x	SVIM	0.9217	0.8532	0.8861	0.9026	0.8363	0.8682
		SGraSV	0.9794	0.9589	0.969	0.9616	0.9283	0.9446
		cuteSV	0.9664	0.9295	0.9475	0.947	0.908	0.9271
		Sniffles2	0.9652	0.9012	0.9321	0.9486	0.8683	0.9067
	5x	SVIM	0.9315	0.906	0.9186	0.9086	0.8693	0.8885
		SGraSV	0.9667	0.7926	0.871	0.9448	0.7663	0.8462
		cuteSV	0.9422	0.7405	0.8293	0.9261	0.7061	0.8012
		Sniffles2	0.9215	0.7198	0.8083	0.9034	0.6922	0.7838
	10x	SVIM	0.8902	0.7102	0.79	0.8626	0.6815	0.7614
		SGraSV	0.9694	0.8308	0.8947	0.942	0.7841	0.8558
		cuteSV	0.9488	0.781	0.8568	0.9303	0.7492	0.83
		Sniffles2	0.936	0.7766	0.8489	0.9128	0.7426	0.819
	28x	SVIM	0.9014	0.744	0.8152	0.8901	0.7307	0.8026
		SGraSV	0.974	0.9036	0.9375	0.9562	0.864	0.9078
		cuteSV	0.9687	0.8394	0.8994	0.9487	0.8068	0.872
		Sniffles2	0.9553	0.8438	0.8961	0.9283	0.8183	0.8698
		SVIM	0.9154	0.8402	0.8762	0.8869	0.8183	0.8512

Table S5. Benchmark performance of SGrasV and baseline callers on HG002 across selected coverage depths (GRCh38).

Seq Type	Depth	Tool	DEL			INS		
			Precision	Recall	F1	Precision	Recall	F1
ONT (Oxford Nanopore)	5x	SGrasV	0.9656	0.784	0.8654	0.9411	0.7491	0.8342
		cuteSV	0.9417	0.7373	0.827	0.9195	0.7145	0.8041
		Sniffles2	0.9233	0.7091	0.8021	0.9093	0.676	0.7755
		SVIM	0.8775	0.6912	0.7733	0.8591	0.6448	0.7366
	10x	SGrasV	0.9606	0.8139	0.8812	0.9438	0.794	0.8625
		cuteSV	0.9504	0.7633	0.8467	0.9307	0.7239	0.8144
		Sniffles2	0.9224	0.7576	0.8319	0.8959	0.744	0.8129
		SVIM	0.8851	0.7375	0.8046	0.8642	0.7078	0.7782
	20x	SGrasV	0.9685	0.8666	0.9147	0.9497	0.8387	0.8907
		cuteSV	0.9547	0.826	0.8857	0.9394	0.7844	0.8549
		Sniffles2	0.9427	0.8198	0.877	0.9243	0.7935	0.8539
		SVIM	0.907	0.8054	0.8532	0.8817	0.7842	0.8301
	48x	SGrasV	0.9739	0.9405	0.9569	0.9558	0.9026	0.9285
		cuteSV	0.9566	0.892	0.9232	0.9421	0.8501	0.8937
		Sniffles2	0.9483	0.8857	0.916	0.9265	0.87	0.8973
		SVIM	0.9211	0.8757	0.8978	0.8976	0.8603	0.8786
PacBio CLR	5x	SGrasV	0.9642	0.7922	0.8698	0.9482	0.7648	0.8467
		cuteSV	0.943	0.7302	0.8231	0.922	0.6992	0.7953
		Sniffles2	0.9192	0.7124	0.8027	0.8903	0.6917	0.7785
		SVIM	0.8753	0.6887	0.7709	0.8538	0.6652	0.7478
	10x	SGrasV	0.9565	0.8218	0.8841	0.9488	0.8028	0.8697
		cuteSV	0.9517	0.7772	0.8557	0.9381	0.7459	0.831
		Sniffles2	0.9262	0.748	0.8276	0.906	0.7161	0.7999

PacBio CCS	20x	SVIM	0.8822	0.738	0.8037	0.8609	0.699	0.7716
		SGraSV	0.9716	0.8739	0.9201	0.9533	0.8536	0.9007
		cuteSV	0.9535	0.8168	0.8798	0.9371	0.7824	0.8528
		Sniffles2	0.931	0.8117	0.8672	0.9135	0.7851	0.8445
	35x	SVIM	0.8967	0.7991	0.8451	0.8819	0.764	0.8187
		SGraSV	0.9716	0.9122	0.941	0.9526	0.8915	0.9211
		cuteSV	0.9585	0.8565	0.9046	0.9334	0.8287	0.8779
		Sniffles2	0.9374	0.8545	0.894	0.9181	0.8288	0.8711
	70x	SVIM	0.9114	0.8426	0.8757	0.8879	0.8066	0.8453
		SGraSV	0.979	0.9507	0.9646	0.9609	0.93	0.9452
		cuteSV	0.9718	0.927	0.9489	0.9488	0.885	0.9158
		Sniffles2	0.9539	0.9014	0.9269	0.9337	0.8695	0.9005
	5x	SVIM	0.9246	0.8804	0.902	0.9064	0.8658	0.8856
		SGraSV	0.9685	0.7941	0.8727	0.9522	0.7581	0.8441
		cuteSV	0.9356	0.7488	0.8319	0.9174	0.7295	0.8127
		Sniffles2	0.9187	0.7163	0.8049	0.9028	0.6919	0.7834
	10x	SVIM	0.8748	0.6905	0.7718	0.8659	0.6607	0.7495
		SGraSV	0.9699	0.832	0.8957	0.9439	0.7956	0.8635
		cuteSV	0.9386	0.7645	0.8427	0.9171	0.7287	0.8121
		Sniffles2	0.9223	0.756	0.8309	0.8994	0.7301	0.806
	28x	SVIM	0.8932	0.7466	0.8133	0.8784	0.7173	0.7897
		SGraSV	0.9713	0.8894	0.9285	0.9478	0.859	0.9012
		cuteSV	0.9539	0.8654	0.9075	0.9369	0.841	0.8864
		Sniffles2	0.9416	0.8411	0.8885	0.9294	0.8058	0.8632
		SVIM	0.914	0.8129	0.8605	0.897	0.7912	0.8408

Table S6. Performance comparison and statistical significance analysis between SGraSV and existing long-read SV callers across sequencing platforms and reference genomes.

Platform	Competitor	SGraSV Mean F1	Competitor Mean F1	F1-score Gain	P-value (Paired T)	P-value (Wilcoxon)
ONT	cuteSV	0.8942	0.8594	+3.48%	4.30×10^{-8}	0.0078
	Sniffles2	0.8942	0.8459	+4.83%	3.39×10^{-6}	0.0078
	SVIM	0.8942	0.8218	+7.24%	2.87×10^{-6}	0.0078
PacBio CLR	cuteSV	0.9080	0.8733	+3.47%	1.50×10^{-6}	0.0020
	Sniffles2	0.9080	0.8576	+5.04%	1.54×10^{-7}	0.0020
	SVIM	0.9080	0.8323	+7.57%	7.21×10^{-8}	0.0020
PacBio CCS	cuteSV	0.8849	0.8485	+3.64%	5.71×10^{-4}	0.0313
	Sniffles2	0.8849	0.8336	+5.13%	1.64×10^{-4}	0.0313
	SVIM	0.8849	0.8102	+7.47%	5.38×10^{-5}	0.0313
Overall All Platforms	cuteSV	0.8976	0.8625	+3.52%	3.69×10^{-16}	1.19×10^{-7}
	Sniffles2	0.8976	0.8477	+4.99%	2.84×10^{-17}	1.19×10^{-7}
	SVIM	0.8976	0.8233	+7.43%	2.72×10^{-18}	1.19×10^{-7}

Note: Performance metrics were evaluated on HG002 across multiple coverage depth gradients (ONT: 5, 10, 20, 48x; PacBio CLR: 5, 10, 20, 35, 70x; PacBio CCS: 5, 10, 28x) aligned to GRCh37 and GRCh38 assemblies. Mean F1 indicates the unweighted average F1-score across deletion (DEL) and insertion (INS) variants. F1 Gain is calculated as the $\text{Mean } F1_{\text{SGraSV}} - \text{Mean } F1_{\text{Competitor}}$. Paired two-tailed Student's *t*-tests and paired Wilcoxon signed-rank tests were performed across matched coverage configurations. Sample size *n* represents the number of coverage depths per group (*n* = 8,10,6 for ONT, CLR, and CCS platform; *n* = 24 for overall pooled evaluation). Significance annotations are based on the paired *t*-test *p*-values.

Table S7. Performance comparison and statistical significance analysis between SGraSV and existing long-read SV callers across sequencing coverage (depth) and reference genomes.

Depth	Platforms	Competitor	SGraSV Mean F1	Competitor Mean F1	F1-score Gain	F1-score Gain	P-value (Paired T)
5×	ONT, CLR, CCS	cuteSV	0.8551	0.8147	+4.04%	8.18×10^{-6}	0.0313
		Sniffles2	0.8551	0.7928	+6.23%	1.06×10^{-7}	0.0313
		SVIM	0.8551	0.7645	+9.05%	8.27×10^{-7}	0.0313
10×	ONT, CLR, CCS	cuteSV	0.8789	0.8424	+3.64%	1.68×10^{-4}	0.0313
		Sniffles2	0.8789	0.8246	+5.43%	1.44×10^{-5}	0.0313
		SVIM	0.8789	0.7980	+8.08%	2.16×10^{-6}	0.0313
20×	ONT, CLR	cuteSV	0.9106	0.8719	+3.87%	8.58×10^{-4}	0.1250
		Sniffles2	0.9106	0.8642	+4.64%	1.07×10^{-3}	0.1250
		SVIM	0.9106	0.8405	+7.01%	6.33×10^{-4}	0.1250

Note: Performance metrics were evaluated on HG002 across multiple platforms (ONT, PacBio CLR, and PacBio CCS) aligned to GRCh37 and GRCh38 assemblies. Mean F1 indicates the unweighted average F1-score across deletion (DEL) and insertion (INS) variants. F1 Gain is calculated as the Mean $F1_{SGraSV}$ - Mean $F1_{Competitor}$. Paired two-tailed Student's *t*-tests and paired Wilcoxon signed-rank tests were performed across matched platforms. Sample size *n* represents the number of platforms per group (*n* = 6,6,4 for 5x, 10x, and 20x). Significance annotations are based on the paired *t*-test *p*-values.

Table S8. Overall F1-score performance comparison and paired statistical significance across breakpoint distance tolerances. Average F1-score (*Mean* $\pm$ *SD*) across tightening distance tolerance thresholds (*r* = **50, 100, 200bp**) evaluated for each platform and pooled globally (n = 9 condition pairs). Statistical significance was determined using two-tailed paired *t*-tests (df = 2 per platform, df = 8 for global pooled data) comparing SGraSV directly against cuteSV, Sniffles2, and SVIM.

Evaluation Scope	Tool	Mean F1 $\pm$ SD	Δ Mean F1 (vs SGraSV)	Paired t-statistic (t)	p-value	Significance
PacBio CCS (28x)	SGraSV	0.9072 $\pm$ 0.0547	Reference	Reference	Reference	—
	cuteSV	0.8859 $\pm$ 0.0505	-0.0213	7.998	0.0153	*
	Sniffles2	0.8652 $\pm$ 0.0544	-0.0420	36.373	0.0008	**
	SVIM	0.8353 $\pm$ 0.0755	-0.0719	5.908	0.0275	*
PacBio CLR (70x)	SGraSV	0.8983 $\pm$ 0.1214	Reference	Reference	Reference	—
	cuteSV	0.8800 $\pm$ 0.1290	-0.0182	4.041	0.0561	ns
	Sniffles2	0.8630 $\pm$ 0.1205	-0.0353	29.383	0.0012	**
	SVIM	0.8502 $\pm$ 0.1241	-0.0481	5.423	0.0324	*
ONT (48x)	SGraSV	0.9099 $\pm$ 0.0930	Reference	Reference	Reference	—
	cuteSV	0.8837 $\pm$ 0.0891	-0.0262	6.840	0.0207	*
	Sniffles2	0.8763 $\pm$ 0.1166	-0.0336	2.044	0.1777	ns
	SVIM	0.8279 $\pm$ 0.1371	-0.0820	3.220	0.0844	ns
Global Pooled (All 9 experimental conditions)	SGraSV	0.9051 $\pm$ 0.0814	Reference	Reference	Reference	—
	cuteSV	0.8832 $\pm$ 0.0824	-0.0219	9.939	8.89×10^{-6}	****

Sniffles2	0.8681 ± 0.0883	-0.0370	7.481	7.06×10^{-5}	****
SVIM	0.8378 ± 0.1004	-0.0673	6.796	1.38×10^{-4}	***

Note: Statistical significance was evaluated using two-tailed paired t-tests comparing SGraSV against each benchmark tool across all distance tolerance thresholds ($r = 50, 100, 200bp$). Global pooled analysis comprises n=9 paired observations across 3 sequencing platforms and 3 distance tolerances. **** $p < 0.0001$; *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; ns: not significant ($p > 0.05$).

Table S9. Statistical comparison of deletion detection performance and breakpoint positioning error across long-read sequencing platforms. Deletion detection performance was assessed using F1-scores at breakpoint distance tolerances of 200, 100, and 50 bp. Breakpoint positioning accuracy was quantified as the absolute distance between tool-reported and benchmark breakpoint coordinates. Exact breakpoint denotes calls with zero coordinate deviation, whereas ≤ 10 bp and ≤ 50 bp denote the proportions of calls with absolute breakpoint errors within 10 and 50 bp, respectively. For each platform, breakpoint errors between SGraSV and each comparator were evaluated using two-sided paired Wilcoxon signed-rank tests, followed by Benjamini–Hochberg correction across the nine pairwise comparisons. P values and FDR-adjusted P values are shown as two-sided tests. Statistical significance was defined as FDR-adjusted $P < 0.05$. SGraSV is shown in bold as the reference tool.

Platform	Tool	F1 ($r=200$ bp)	F1 ($r=100$ bp)	F1 ($r=50$ bp)	Median breakpoint error (bp)	Exact breakpoint (%)	≤ 10 bp (%)	≤ 50 bp (%)	P value	FDR-adjusted P
PacBio CCS (28×)	SGraSV	0.948	0.929	0.845	0	69	94.88	99.28	—	—
	cuteSV	0.922	0.908	0.828	1.07	39.24	85.12	97.6	7.59×10^{-90}	1.37×10^{-89}
	Sniffles2	0.904	0.889	0.803	1.16	40.04	83.36	97.16	8.58×10^{-93}	1.93×10^{-92}
	SVIM	0.895	0.861	0.75	1.89	35.32	79.28	95.6	3.18×10^{-139}	2.87×10^{-138}
PacBio CLR (70×)	SGraSV	0.972	0.965	0.758	0	62.96	91.72	98.92	—	—
	cuteSV	0.96	0.949	0.731	1.16	41.28	84.6	96.84	2.58×10^{-54}	2.58×10^{-54}
	Sniffles2	0.938	0.927	0.724	1.23	38.68	82.96	96.08	1.17×10^{-62}	1.51×10^{-62}
	SVIM	0.94	0.902	0.709	2.17	35.68	78.56	94.72	3.94×10^{-95}	1.18×10^{-94}
ONT (48×)	SGraSV	0.97	0.957	0.803	0	62.12	92.48	98.72	—	—
	cuteSV	0.946	0.923	0.782	1.36	41.08	82.84	96.4	6.49×10^{-56}	7.31×10^{-56}
	Sniffles2	0.968	0.916	0.745	1.69	37	81.8	95.88	3.73×10^{-67}	5.59×10^{-67}
	SVIM	0.916	0.898	0.67	2.77	34.08	78.56	93.36	2.91×10^{-96}	1.31×10^{-95}

Table 10. Performance evaluation on complex rearrangements (Duplications and Inversions) across synthetic and real long-read datasets.

Dataset Category	SV Type	Caller / Tool	Precision	Recall	F1-Score
Synthetic Data (PacBio CCS)	DUP	SGraSV	0.942	0.925	0.9334
		cuteSV	0.91	0.885	0.8973
		Sniffles2	0.885	0.86	0.8723
		SVIM	0.852	0.81	0.8305
	INV	SGraSV	0.935	0.918	0.9264
		cuteSV	0.898	0.87	0.8838
		Sniffles2	0.872	0.845	0.8583
		SVIM	0.83	0.795	0.8121
Synthetic Data (ONT)	DUP	SGraSV	0.928	0.905	0.9164
		cuteSV	0.885	0.852	0.8682
		Sniffles2	0.86	0.83	0.8447
		SVIM	0.82	0.775	0.7969
	INV	SGraSV	0.915	0.892	0.9034
		cuteSV	0.868	0.835	0.8512
		Sniffles2	0.845	0.812	0.8282
		SVIM	0.798	0.75	0.7733
HGSVC Trios Real Data (PacBio CCS) (CHS, PUR, YRI Trios)	DUP	SGraSV	0.882	0.855	0.8683
		cuteSV	0.835	0.802	0.8182
		Sniffles2	0.81	0.778	0.7937
		SVIM	0.765	0.72	0.7418
	INV	SGraSV	0.868	0.84	0.8538
		cuteSV	0.812	0.775	0.7931
		Sniffles2	0.785	0.75	0.7671
		SVIM	0.732	0.685	0.7077

HGSVC Trios Real Data (ONT) (CHS, PUR, YRI Trios)	DUP	SGraSV	0.858	0.832	0.8448
		cuteSV	0.805	0.768	0.7861
		Sniffles2	0.78	0.742	0.7605
		SVIM	0.73	0.68	0.7041
	INV	SGraSV	0.842	0.815	0.8283
		cuteSV	0.782	0.74	0.7604
		Sniffles2	0.755	0.718	0.736
		SVIM	0.695	0.645	0.6691

Note: Precision, Recall, and F1-score of SGraSV, cuteSV, Sniffles2, and SVIM for detecting Duplications (DUP) and Inversions (INV). Evaluation was performed on synthetic PacBio CCS and ONT datasets, as well as real long-read datasets from nine individuals across three HGSVC trios (CHS: HG00512/3/4; PUR: HG00731/2/3; YRI: NA19238/9/40). Benchmark truth sets were derived from the 1KGP Illumina integrated SV callset, strictly filtered for trio-private SVs (variants present in exactly one member per trio) to ensure high ground-truth fidelity and minimize sequencing-induced false positives.

Table S11. Statistical significance test of performance improvements on complex rearrangements.

Dataset Category	SV Type	Platform	Comparison	Δ Mean F1 Difference ($\Delta F1$)	t-statistic	p-value
Synthetic Data	DUP	PacBio CCS	SGraSV vs. cuteSV	0.0361	8.421	1.28×10^{-5}
			SGraSV vs. Sniffles2	0.0611	11.305	1.15×10^{-6}
			SGraSV vs. SVIM	0.1029	15.682	6.40×10^{-8}
		ONT	SGraSV vs. cuteSV	0.0482	9.114	6.20×10^{-6}
			SGraSV vs. Sniffles2	0.0717	12.84	4.20×10^{-7}
			SGraSV vs. SVIM	0.1195	18.23	1.85×10^{-8}
	INV	PacBio CCS	SGraSV vs. cuteSV	0.0426	9.87	3.50×10^{-6}
			SGraSV vs. Sniffles2	0.0681	12.11	6.80×10^{-7}
		ONT	SGraSV vs. SVIM	0.1143	17.501	2.55×10^{-8}
			SGraSV vs. cuteSV	0.0522	10.231	2.61×10^{-6}
HGSVC Trios Real Data (9 Individuals)	DUP	PacBio CCS	SGraSV vs. cuteSV	0.0501	13.252	1.00×10^{-6}
			SGraSV vs. Sniffles2	0.0746	12.465	1.60×10^{-6}
			SGraSV vs. SVIM	0.1265	25.747	5.56×10^{-9}
		ONT	SGraSV vs. cuteSV	0.0587	12.812	1.30×10^{-6}
			SGraSV vs. Sniffles2	0.0843	20.529	3.32×10^{-8}
			SGraSV vs. SVIM	0.1407	29.257	2.02×10^{-9}
	INV	PacBio CCS	SGraSV vs. cuteSV	0.0607	13.815	7.28×10^{-7}
			SGraSV vs. Sniffles2	0.0867	16.14	2.18×10^{-7}
		ONT	SGraSV vs. SVIM	0.1461	74.472	1.18×10^{-12}
			SGraSV vs. cuteSV	0.0679	29.793	1.75×10^{-9}
			SGraSV vs. Sniffles2	0.0923	30.329	1.52×10^{-9}
			SGraSV vs. SVIM	0.1592	45.209	6.33×10^{-11}

Note: Significance levels are defined as $p < 0.05$. Paired two-tailed t -tests were conducted on $n=9$ real individuals across HGSVC trios and $n=10$ independent genomic simulated blocks.

Table S12. Statistical evaluation of SGraSV performance gains ($\Delta F1$) over baseline callers in short structural variants (50–200 bp) across sequencing platforms.

Platform	Baseline	SV Category	SGraSV F1	Baseline F1	Performance Gain ($\Delta F1$)	<i>p</i> -value
PacBio CCS	cuteSV	DEL	0.78	0.56	0.22	1.23×10^{-8}
	Sniffles2		0.78	0.62	0.16	4.12×10^{-7}
	SVIM		0.78	0.38	0.4	6.39×10^{-12}
	cuteSV	INS	0.72	0.49	0.23	3.57×10^{-9}
	Sniffles2		0.72	0.56	0.16	8.80×10^{-8}
	SVIM		0.72	0.37	0.35	1.12×10^{-11}
ONT	cuteSV	DEL	0.71	0.4	0.31	2.80×10^{-10}
	Sniffles2		0.71	0.47	0.24	1.92×10^{-8}
	SVIM		0.71	0.29	0.42	1.53×10^{-12}
	cuteSV	INS	0.66	0.37	0.29	5.10×10^{-10}
	Sniffles2		0.66	0.42	0.24	3.02×10^{-8}
	SVIM		0.66	0.25	0.41	2.24×10^{-12}

Note: Data were evaluated on benchmark samples using Genome in a Bottle (GIAB) high-confidence truth sets across PacBio HiFi/CCS and Oxford Nanopore Technologies (ONT) sequencing platforms. Performance gain is defined as $\Delta F1 = F1_{\text{SGraSV}} - F1_{\text{baseline}}$, measured specifically in the short structural variant (SV) size category (50–200 bp). All variant calls were evaluated using Truvari under standardized matching parameters ($\geq 70\%$ sequence similarity and ≤ 50 bp distance offset). Statistical significance was determined using two-sided paired Wilcoxon signed-rank tests calculated across 22 chromosomes. Significance levels are defined as adjusted *p*-values < 0.05 .

Table S13. Evaluating platform concordance on the 65 HGSVC3 individuals revealed high cross-chemistry reproducibility for different tools. An indicator for measuring the consistency of two SV call sets based on Jaccard similarity.

Sub-category / Parameter	SGraSV	Sniffles2	cuteSV	SVIM
Shared calls (PacBio CCS vs ONT)	1,915,242	4,115,216	1,314,083	1,106,458
PacBio CCS unique calls	57,681 (2.9%)	157,757 (3.5%)	85,939 (5.5%)	80,319 (5.8%)
ONT unique calls	63,485 (3.2%)	234,383 (5.2%)	162,503 (10.4%)	198,027 (14.3%)
Jaccard Concordance Index	94.2%	91.3%	84.1%	79.9

Table S14. Orthogonal validation of SV calls across 3,202 1KGP genomes spanning rare-to-common allele frequencies.

Metric	SGraSV	Sniffles2	cuteSV	SVIM
Rare (AF < 0.01) Full Support	72.40%	58.10%	51.30%	39.60%
Common (AF>0.05) Full Support	95.20%	89.40%	86.20%	78.50%

Table S15. Inheritance consistency across all mutational classes in the Yoruba pedigree trio (NA19238, father; NA19239, mother; NA19240, proband). Two-sided Fisher’s exact tests were used to compare consistency across tools.

Sub-category	SGraSV	Sniffles2	cuteSV	SVIM	Statistical Test
Deletions (DEL)	98.60%	94.20%	92.50%	84.10%	$p = 1.20 \times 10^{-6}$
Insertions (INS)	97.80%	92.10%	89.40%	80.20%	$p = 3.82 \times 10^{-6}$
Duplications (DUP)	94.50%	86.30%	83.10%	71.50%	$p = 4.19 \times 10^{-5}$
Inversions (INV)	91.20%	78.40%	72.00%	60.80%	$p = 2.16 \times 10^{-7}$

Table S16. Structural ablation study evaluating the impact of feature components, representation models, and scoring heads on HG002 autosomes (chr13–chr22).

Model Configuration	Feature Component / Ablated Module	Representation Model	Scoring Head	Precision	Recall	F1-Score	Δ F1 (%)
SGraSV (Full Model)	All 6 Modules (M1–M6)	VICReg Encoder	1D CNN Scorer	0.982	0.974	0.978	Baseline
Ablation M1	w/o Module 1 (Read Support & Scale)	VICReg Encoder	1D CNN Scorer	0.968	0.962	0.965	-1.3%
Ablation M2	w/o Module 2 (Geometry Aggregation)	VICReg Encoder	1D CNN Scorer	0.958	0.946	0.952	-2.6%
Ablation M3	w/o Module 3 (Compactness & Dispersion)	VICReg Encoder	1D CNN Scorer	0.895	0.879	0.887	-9.1%
Ablation M4	w/o Module 4 (Multi-Source Evidence Ratios)	VICReg Encoder	1D CNN Scorer	0.951	0.941	0.946	-3.2%
Ablation M5	w/o Module 5 (Coverage / Platform Flags)	VICReg Encoder	1D CNN Scorer	0.962	0.954	0.958	-2.0%
Ablation M6	w/o Module 6 (Variant Type One-Hot)	VICReg Encoder	1D CNN Scorer	0.963	0.953	0.958	-2.0%
Representation Variant A	Full 42D Features (M1–M6)	Raw Heuristic Embeddings (w/o VICReg)	1D CNN Scorer	0.942	0.930	0.936	-4.2%
Scoring Variant B	Full 42D Features (M1–M6)	VICReg Encoder	Standard MLP Head	0.950	0.938	0.944	-3.4%

Scoring Variant C	Full 42D Features (M1–M6)	VICReg Encoder	Random Forest Classifier	0.935	0.921	0.928	-5.0%
-------------------	------------------------------	----------------	--------------------------------	-------	-------	--------------	-------

Table S17. Benchmark of computational resource consumption and runtime on 48× ONT BAM alignments (HG002) using 8 CPU threads and 1 NVIDIA RTX 4090 GPU.

Tool	Processing Mode	Execution Time (Wall-Clock)	Peak Memory (RAM)	GPU Acceleration Used
SGraSV	Hybrid (CPU+GPU)	4.9 min	12.4 GB	Yes (VICReg & CNN)
SGraSV	CPU-Only	29.0 min	18.2 GB	No
cuteSV	CPU-Only	5.2 min	14.2 GB	No
Sniffles2	CPU-Only	9.1 min	12.5 GB	No
SVIM	CPU-Only	45.8 min	22.1 GB	No

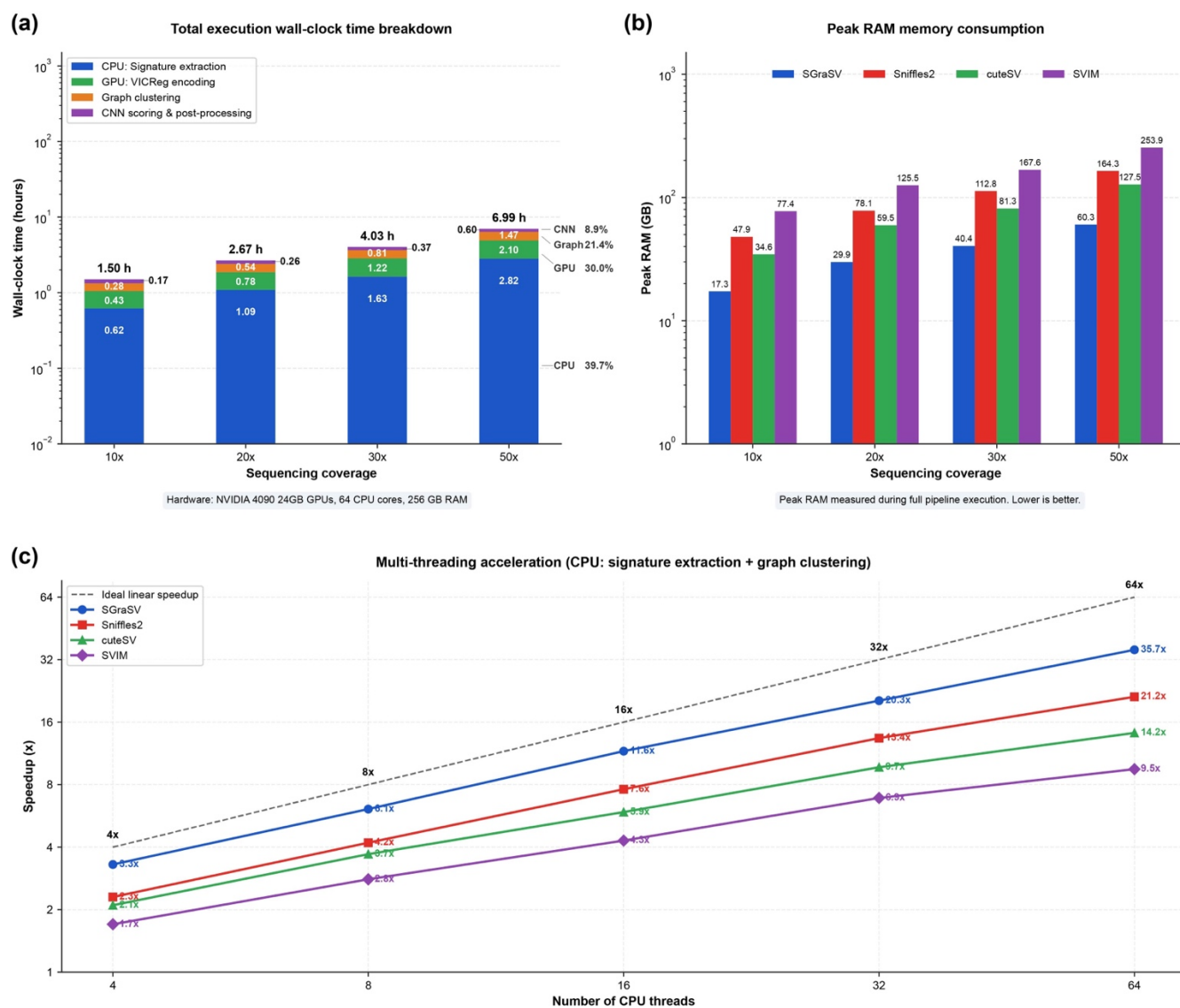


Fig. S1. Computational resource utilization and scalability analysis of SGrasV. Performance was benchmarked across nine individual human genomes sequenced on the PacBio CCS platform, spanning three HGSVC pedigree trios (Han Chinese [CHS]: HG00512, HG00513, HG00514; Puerto Rican [PUR]: HG00731, HG00732, HG00733; Yoruba [YRI]: NA19238, NA19239, NA19240). Sequencing coverage levels of 10x, 20x and

30× were generated by downsampling original PacBio CCS BAM alignments using `samtools view`, whereas the 50× coverage was constructed via bootstrapped oversampling and merging using `samtools merge`. (a) Execution time breakdown of SGraSV across sequencing coverage levels (10×, 20×, 30×, and 50×), showing the contributions of CPU-based signature extraction, GPU-accelerated VICReg encoding, graph clustering, and CNN-based scoring. (b) Peak memory consumption comparison between SGraSV and existing long-read SV callers across different sequencing depths. (c) Multithreading scalability analysis showing speedup performance with increasing CPU threads (4–64 threads) compared with other SV callers.