

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

Xiang Chen

xiaang-chen@zju.edu.cn

University of Electronic Science and Technology of China Yangtze Delta Region Institute (Quzhou)

<https://orcid.org/0000-0003-3188-8332>

Song Cheng

Xinjiang University

Yihang Lin

Xiamen University of Technology

Siyuan Song

Xidian

Beibei Wang

Xian Polytechnic University

Quan Zou

University of Electronic Science and Technology of China; Institute of Digital Health, Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China

<https://orcid.org/0000-0001-6406-1142>

Article

Keywords:

Posted Date: October 15th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10833348/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: There is **NO** Competing Interest.

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

Song Cheng^{1,2}, Yihang Lin^{1,3}, Siyuan Song^{1,4}, Beibei Wang^{1,5}, Quan Zou^{1,6,*}, Xiang Chen^{1,*}

¹ Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Quzhou, Zhejiang, China

² School of Computer Science and Technology, Xinjiang University, Urumqi, Xinjiang, China

³ School of Opto-electronic and Communication Engineering, Xiamen University of Technology, Xiamen, Fujian, China.

⁴ School of Computer Science and Technology, Xidian University, Xian, Shaanxi, China

⁵ School of Management, Xi'an Polytechnic University, Xian, Shaanxi, China

⁶ Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, Sichuan, China

* Corresponding author: Dr. Xiang Chen, PhD (xiang-chen@zju.edu.cn), Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Quzhou, 324003, Zhejiang, China; Dr. Quan Zou, PhD (zouquan@nclab.net), Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, 611731, China.

Funding:

This work was supported by the Mindray Joint Fund of Zhejiang Provincial Natural Science Foundation of China under Grant No. LMRY26H200005, and the Municipal Government of Quzhou under Grant No. 2024D009.

Subtitle: Graph representation learning framework for Structural Variant detection

Abstract

Structural variant calling from long-read sequencing is increasingly limited not by read length itself, but by how heterogeneous alignment signals are combined into candidate events. Repetitive sequence, short variants and mobile element insertions can produce dispersed breakpoints and ambiguous local alignments that challenge fixed clustering strategies. We developed SGrasV, which uses self-supervised graph representation learning to capture relationships among read-level signals and local sequence context. Here, we show that SGrasV improves detection and breakpoint resolution across PacBio and Oxford Nanopore data and maintains high reproducibility across sequencing platforms. Evaluation in 65 individuals and independent analysis of 3,202 1000 Genomes Project genomes further support its ability to recover rare and difficult variants. In particular, SGrasV preserves characteristic insertion patterns for mobile elements that are often fragmented by existing callers. These findings support graph-based representation of alignment evidence as a practical strategy for improving structural variant discovery in heterogeneous genomic datasets.

Introduction

Structural variants (SVs) are large genomic changes of 50 base pairs or more, including deletions (DELs), insertions (INSs), inversions (INVs), and duplications (DUPS). At the DNA and RNA levels, these large alterations disrupt gene regulation, cause gene fusions, and alter splicing, leading to severe genetic disorders, neurodevelopmental conditions, and cancer ^{1,2}. SVs calling with long-read sequencing has greatly improved the detection of genomic rearrangements ³, but longer reads have not eliminated a basic problem: signals generated by the same SV are often inconsistent across reads and sequencing platforms. Widely used callers such as Sniffles2 ^{4,5}, cuteSV ^{6,7} and SVIM ^{8,9} aggregate split reads, supplementary alignments, soft clipping and local indels using coordinate-based clustering and empirically tuned thresholds. This works well when breakpoint signals are compact, but becomes less reliable when signals are dispersed or distorted by local sequence context. The problem is particularly apparent for 50 – 200bp SVs, where alignment jitter and short indel errors can shift breakpoint coordinates by only a few bases, enough to fragment a true event under a strict clustering rule, but close enough to merge neighboring artifacts under a permissive one.

Mobile element insertions (MEIs) expose the same limitation more clearly ^{10–12}. Alu (~300bp), SVA (~1.5 – 3kb) and full-length LINE-1 (~6kb) insertions can generate soft-clipped segments, supplementary alignments and target-site duplications (TSDs) around the same insertion site ^{13–16}. Treating these signals primarily by genomic proximity can therefore truncate the inserted sequence or split one event into multiple calls. DUPS and INVs pose a related problem because their evidence often involves conflicting mapping orientations or multiple local alignment configurations. Our benchmarking was designed around these failure modes rather than relying only on aggregate SV F1-scores: we tested PacBio CCS (HiFi), PacBio CLR and ONT data across 5–70× coverage ¹⁷, evaluated nine genomes from three HGSVC trios for complex SVs ¹⁸, and examined 65 HGSVC3 individuals sequenced on both PacBio CCS and ONT. This was motivated by a practical concern: a caller that performs well on one platform is not necessarily producing the same biological signal on another.

We therefore developed SGrasV, which represents SV evidence as a learned graph rather than relying primarily on breakpoint proximity. The framework converts split-read, clipping, supplementary-

alignment and sequence-context signals into a 32-dimensional signature, learns a 128-dimensional representation using self-supervised VICReg training^{19,20}, and groups related signatures through graph-based clustering. Candidate classification then integrates 42 features describing read support, alignment geometry, cluster compactness and dispersion, sequence context and variant characteristics. This design arose from an empirical observation in our ablation analysis: removing the compactness-and-dispersion module reduced overall F1 from 0.978 to 0.887, a larger loss than removing the other individual feature modules. We therefore interpret SGraSV's gain not simply as an advantage from using more features, but as evidence that the spatial coherence of supporting alignment signals is itself a useful discriminator between true SVs and alignment artifacts.

We evaluated SGraSV across progressively more challenging settings, from controlled multi-platform benchmarking to population-scale analysis. The framework was assessed on PacBio CCS, PacBio CLR, and ONT datasets spanning different sequencing depths, followed by evaluation of complex DUP and INV events and size-stratified SV detection. We further tested cross-platform reproducibility in 65 HGVC3 individuals sequenced using both PacBio CCS and ONT, and performed orthogonal short-read realignment analysis across 3,202 genomes from the 1000 Genomes Project (1KGP)²¹. On the 65 HGVC3 individuals, SGraSV produced 1,972,923 SVs from PacBio CCS and 1,978,727 from ONT, with 1,915,242 shared calls and 94.05% cross-platform concordance. In the most difficult 50 – 200bp range, SGraSV achieved F1-scores of 0.75 for combined SVs on CCS and 0.69 on ONT, compared with 0.53 and 0.39 for cuteSV. These results led us to evaluate not only sensitivity and precision, but also breakpoint resolution, Mendelian consistency, rare-variant validation and MEI resolution, asking whether learned structural representations can remain stable when sequencing chemistry, coverage and population context change.

Results

Overview of the SGraSV framework

Long-read sequencing technologies have substantially expanded the catalog of human genomic structural variation (SV)³. However, alignment noise, ambiguous soft-clipping in tandem repeats, and non-uniform coverage fluctuations continue to introduce false positives and inaccurate breakpoint estimates⁴. To address these challenges, we developed SGraSV (**S**elf-supervised **G**raph representation learning framework for **SV** detection), an end-to-end framework that unifies dynamic sequence modulation, self-supervised representation learning, topological graph clustering, and deep probabilistic scoring.

The pipeline operates in four sequential stages as shown in Fig. 1: (1) Signature Feature Vectorization: Long-read alignment files (BAM) are scanned to detect localized structural signals, including split alignments, soft/hard clipping, and intra-read INSs or DELs. For each candidate read signature i , SGraSV computes a 32-dimensional feature vector $x_i \in \mathbb{R}^{32}$ (see Methods, Supplementary Note S1 and Table S1). This vector incorporates standard positional and mapping statistics (such as mapping quality, clip position, read orientation, and local depth) alongside two dynamic sequence-level properties: local K -mer frequency entropy (H_k , calculated over a 200 bp flanking window) and local Smith-Waterman re-alignment identity (S_{sw} , calculated by dynamic re-alignment of clipped read tails against reference flanks)^{17,22,23}. (2) Self-Supervised Feature Decorrelation (VICReg)¹⁹: Rather than relying on uncalibrated heuristic distance metrics, signature vectors are passed into a neural encoder f_θ trained via VICReg (see Methods). Through non-contrastive self-supervised learning on random feature perturbations (Gaussian

noise addition and feature dropout), the encoder maps raw signatures into a disentangled 128-dimensional latent embedding space, removing redundant mapping noise while preserving true structural relationships. (3) Topological Signature Graph Clustering: Pairwise cosine similarities are computed among signature embeddings within overlapping genomic windows. An unweighted graph $G = (V, E)$ is constructed where nodes represent read signatures and edges represent latent similarities exceeding a similarity threshold ($\tau = 0.85$). Connected component partitioning groups co-occurring read signatures into candidate SV clusters²⁴ (see Methods and Supplementary Note S2). (4) Probabilistic Candidate Scoring via 1D CNN: Signature clusters k are aggregated into candidate-level feature matrices ($\hat{x}_k \in \mathbb{R}^{42}$) and passed through a 1D CNN²⁵ (see Methods, Supplementary Note S3, and Table S2). The CNN outputs a calibrated probability $P_{SV} \in [0,1]$. Candidates meeting a user-defined or adaptive threshold ($P_{SV} \geq 0.50$) are written to the final VCF callset.

To prevent spatial data leakage and ensure unbiased generalization across all evaluations, model parameters for the VICReg encoder and 1D CNN scorer were trained exclusively on autosomes chr1–chr10, hyperparameter-tuned on chr11–chr12, and evaluated strictly on the held-out test set chr13–chr22 across all benchmarks unless explicitly stated otherwise²⁶ (see Methods).

Benchmarking SGraSV performance across platforms, depths, and genome builds on HG002

To evaluate the robustness and sensitivity of SGraSV across different sequencing technologies, coverage depths, and human reference assemblies, we conducted comprehensive benchmarking on the Genome in a Bottle (GIAB) HG002 sample^{27,28}. Evaluated datasets (Table 1) encompassed ONT (5×, 10×, 20× and 48×), PacBio CLR (5×, 10×, 20×, 35× and 70×), and PacBio CCS (5×, 10× and 28×)^{17,22,23}. All callsets were aligned against both GRCh37 and GRCh38 reference genomes and strictly evaluated within the HG002 Tier1 high-confidence regions using Truvari (v4.1.0)²⁹. Under the standardized evaluation protocol, a breakpoint matching threshold of $r = 100$ was set without sequence comparison ($-P 0.00$) to ensure consistent evaluation across symbolic and sequence-resolved calls (see Supplementary Note S4). Complete variant-level evaluation metrics (Precision, Recall, and F1-score for DELs, INs, and overall mean) are cataloged in Supplementary Tables S4, S5 and Fig. 2. We compared SGraSV against three widely adopted long-read SV callers: cuteSV^{6,7}, Sniffles2^{4,5}, and SVIM^{8,9} (version and parameter used the tools see Supplementary Note S5). Statistical significance was assessed using two-sided paired *t*-tests and Wilcoxon signed-rank tests across $n = 24$ experimental condition pairs (Supplementary Tables S6 and S7).

Across all evaluated conditions, SGraSV achieved superior overall detection performance (Fig. 2 and Table S6), securing a mean F1-score of 0.8976 (0.9106 for DELs and 0.8847 for INs). This significantly exceeded the performance of cuteSV (mean F1 = 0.8625), Sniffles2 (mean F1 = 0.8477), and SVIM (mean F1 = 0.8233). Specifically, SGraSV yielded an average absolute F1-score improvement of +3.52% over cuteSV (two-sided paired *t*-test $p = 3.69 \times 10^{-16}$; Wilcoxon $p = 1.19 \times 10^{-7}$), +4.99% over Sniffles2 ($p = 2.84 \times 10^{-17}$), and +7.43% over SVIM ($p = 2.72 \times 10^{-18}$; Table S6). These statistical gains confirm that SGraSV's graph self-supervised learning representation provides a fundamental algorithmic advantage in SV discovery.

Subgroup analysis across sequencing platforms demonstrated the robust adaptability of SGraSV to diverse read error profiles and length distributions (Fig. 2a-c and Tables S6). On PacBio CLR datasets ($n = 10$), SGraSV achieved a peak platform-level mean F1-score of 0.9080, outperforming cuteSV (0.8733, +3.47%, $p = 1.50 \times 10^{-6}$), Sniffles2 (0.8576, +5.04%, $p = 1.54 \times 10^{-7}$), and SVIM (0.8323, +7.57%, $p =$

7.21×10^{-8}). On ONT datasets ($n = 8$), characterized by ultra-long reads with indel noise, SGraSV maintained an average F1-score of 0.8942, delivering improvements of +3.48% over cuteSV ($p = 4.30 \times 10^{-8}$), +4.83% over Sniffles2 ($p = 3.39 \times 10^{-6}$), and +7.24% over SVIM ($p = 2.87 \times 10^{-6}$). On PacBio CCS datasets ($n = 6$), SGraSV sustained a leading mean F1-score of 0.8849, significantly surpassing cuteSV (0.8485, +3.64%, $p = 5.71 \times 10^{-4}$), Sniffles2 (0.8336, +5.13%, $p = 1.64 \times 10^{-4}$), and SVIM (0.8102, +7.47%, $p = 5.38 \times 10^{-5}$).

SGraSV demonstrated exceptional resilience under low-coverage regimes, which represent crucial constraints for cost-effective clinical and population-scale genomics¹⁷ (Fig. 2 and Table S7). At 5× ultra-low depth ($n = 6$), where sparse read support caused severe degradation in baseline tools (cuteSV: 0.8147; Sniffles2: 0.7928; SVIM: 0.7645), SGraSV preserved a high mean F1-score of 0.8551. This translates to an expanded gain of +4.04% over cuteSV ($p = 8.18 \times 10^{-6}$), +6.23% over Sniffles2 ($p = 1.06 \times 10^{-7}$), and +9.05% over SVIM ($p = 8.27 \times 10^{-7}$; Table S6). At 10× depth ($n = 6$), SGraSV achieved a mean F1-score of 0.8789 (+3.64% to +8.08% gains; $p < 0.001$). As coverage increased to 20× ($n = 4$), SGraSV reached a mean F1-score of 0.9106 (+3.87% over cuteSV, $p = 8.58 \times 10^{-4}$), eventually scaling to 0.9559 on 70× CLR data (Fig. 2 and Table S7).

Regarding variant-type dynamics, all tools exhibited higher detection accuracy for DELs than INSs (Table S4 and S5). However, SGraSV achieved its most pronounced advantages on INSs, attaining a mean insertion F1-score of 0.8847 compared to cuteSV (0.8490), Sniffles2 (0.8358), and SVIM (0.8109), thereby effectively resolving false negatives in difficult insertion regions (Table S4 and S5). Furthermore, SGraSV displayed consistent performance across GRCh37 and GRCh38 coordinate systems, demonstrating that its graph topology representations are invariant to reference genome choice and reference bias³.

Cross-platform performance and breakpoint localization accuracy

To evaluate the cross-platform generalization and spatial precision of SGraSV across diverse long-read sequencing architectures, we benchmarked SGraSV against state-of-the-art SV callers: cuteSV, Sniffles2, and SVIM (version and parameter used the tools see Supplementary Note S5), across three major long-read platforms: PacBio CCS (28×), PacBio CLR (70×), and ONT (48×). Beyond standard detection metrics, we systematically assessed model resilience by enforcing progressively stringent breakpoint distance tolerance thresholds ($r = 200, 100, 50bp$) alongside rigorous quantitative evaluations of continuous breakpoint positioning errors²⁹ (Fig. 3).

Across all platforms and spatial thresholds, SGraSV consistently achieved the highest F1-scores, demonstrating exceptional resilience against distance tolerance decay (Fig. 3a and Supplementary Table 8). In the global pooled analysis across all 9 experimental conditions (3 platforms × 3 distance tolerances), SGraSV attained a mean F1-score of 0.9051 ± 0.0814 , significantly outperforming cuteSV (0.8832 ± 0.0824 ; paired t -test $p = 8.89 \times 10^{-6}$), Sniffles2 (0.8681 ± 0.0883 ; paired t -test $p = 7.06 \times 10^{-5}$), and SVIM (0.8378 ± 0.1004 ; paired t -test $p = 1.38 \times 10^{-4}$). Notably, while benchmark tools exhibited sharp performance drops when the geometric tolerance was tightened from $r = 200bp$ to $r = 50bp$, particularly on noisy long reads such as PacBio CLR (where cuteSV dropped by 0.2288 and Sniffles2 by 0.2140), SGraSV retained robust detection power ($F1 = 0.8450$ on CCS, 0.7582 on CLR, and 0.8028 on ONT at $r = 50bp$). This underscores that SGraSV's graph-based topology modeling effectively suppresses alignment noise without compromising spatial precision^{24,30}.

Direct quantification of breakpoint positioning accuracy further highlighted the precise localization capability of SGraSV (Fig. 3b and Supplementary Table S9). Across all three long-read sequencing platforms, SGraSV consistently achieved the lowest median breakpoint error among the evaluated callers. On PacBio CCS reads, SGraSV showed a median breakpoint error of 0.00 bp, compared with 1.07 bp for cuteSV, 1.16 bp for Sniffles2, and 1.89 bp for SVIM. This advantage was accompanied by significantly higher proportions of precisely localized breakpoints, with 94.88% of SGraSV calls falling within 10 bp and 99.28% within 50 bp, compared with 85.12–79.28% and 97.60–95.60%, respectively, for the comparator callers. Similar improvements were observed on PacBio CLR, where SGraSV achieved a median error of 0.00 bp versus 1.16 bp for cuteSV, 1.23 bp for Sniffles2, and 2.17 bp for SVIM, with 91.72% of SGraSV breakpoints localized within 10 bp and 98.92% within 50 bp. On ONT reads, SGraSV maintained the highest breakpoint resolution, with a median error of 0.00 bp compared with 1.36, 1.69, and 2.77 bp for cuteSV, Sniffles2, and SVIM, respectively; 92.48% of SGraSV breakpoints were within 10 bp and 98.72% within 50 bp. Across all nine pairwise comparisons, the reduction in breakpoint error relative to competing callers was highly significant (two-sided paired Wilcoxon signed-rank test, FDR-adjusted $p < 1.4 \times 10^{-8}$; Supplementary Table S9).

These results demonstrate that SGraSV effectively bridges the trade-off between sensitivity and spatial refinement. Traditional heuristic callers rely heavily on primary alignment signatures that degrade rapidly under high error rates or strict geometric cutoffs. In contrast, SGraSV's self-supervised graph representations capture intricate non-linear local genomic contexts, enabling precise breakpoint resolving capabilities regardless of sequencing chemistry or error profile¹⁹.

Complex rearrangements: multi-type SV detection on real and synthetic data

Accurate identification of complex structural rearrangements, specifically DUPs and INVs, remains a fundamental bottleneck for conventional long-read mapping tools. Unlike balanced DELs or novel INs, DUPs and INVs frequently disrupt local alignment orientations or introduce nested alignment signatures within complex, repetitive genomic regions. To rigorously test whether SGraSV overcomes these topological constraints with high statistical confidence, we performed comprehensive benchmarking on both synthetic datasets and high-depth real sequencing data across PacBio CCS³¹ and ONT³² platforms (Supplementary Note S6, Table S10, S11, and Fig. 4).

For realistic benchmarks, we evaluated SGraSV alongside cuteSV, Sniffles2, and SVIM using real long-read data from nine individuals spanning three geographically diverse pedigree trios from the HGSVC (Table 1 and Table S3): 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). Ground-truth variant sets were derived from the Illumina-based 1KGP integrated SV callset^{18,21} (Table S3). Crucially, to eliminate random false positives induced by sequencing artifact noise, we established a strict inheritance-based filtering criterion: only variants carried by exactly one sample per trio (trio-private SVs) were retained as the evaluation truth set³³. Because true private mutations or rare inherited variants within a tightly controlled pedigree exhibit high biological fidelity, this protocol provides an exceptionally clean benchmark for complex SV evaluation^{34,35} (Supplementary Note S6).

Statistical significance was assessed using paired two-tailed *t*-tests and Wilcoxon signed-rank tests across individual samples ($n=9$) and simulated genomic blocks ($n=10$). On real PacBio CCS data across the nine HGSVC individuals, SGraSV achieved top-tier performance for both complex variant classes, attaining mean F1-scores of 0.8683 for DUPs (Precision: 0.8820, Recall: 0.8550) and 0.8538 for INVs (Precision:

0.8680, Recall: 0.8400). This performance improvement over the second-best caller, cuteSV (DUP F1: 0.8182; INV F1: 0.7931), was statistically significant ($\Delta F1 = +0.0501$, $p = 1.00 \times 10^{-6}$ for DUPs; $\Delta F1 = +0.0607$, $p = 7.28 \times 10^{-7}$ for INVs; Supplementary Table S11). On real ONT data, where read-level error rates and chimeric alignments are more pronounced, SGraSV maintained robust detection accuracy, securing mean F1-scores of 0.8448 for DUPs and 0.8283 for INVs, outperforming cuteSV, Sniffles2, and SVIM with uniform statistical significance ($p < 1.00 \times 10^{-4}$ across all pairwise comparisons; Fig. 4a,b).

Consistent advantages were further confirmed on simulated benchmark datasets. Under synthetic PacBio CCS sequencing, SGraSV reached an F1-score of 0.9334 on DUPs and 0.9264 on INVs, outperforming cuteSV by 3.61% ($p = 1.28 \times 10^{-5}$) and 4.26% ($p = 3.50 \times 10^{-6}$) in F1-score (Supplementary Table S11), respectively. On synthetic ONT datasets, SGraSV preserved high precision and recall (DUP F1: 0.9164; INV F1: 0.9034), demonstrating that its self-supervised graph representation learning framework effectively disentangles non-linear breakpoint topology and strand-orientation ambiguities regardless of platform-specific read characteristics (Fig. 4c,d).

Size-stratified resolution analysis across distinct variant length spectrums

To evaluate the resolution and stability of SGraSV across varying SV scales, performance was benchmarked across nine individual human genomes spanning three HGSVC pedigree trios (CHS: HG00512/3/4; PUR: HG00731/2/3; YRI: NA19238/9/40)^{18,36} against 1KGP integrated SV truth sets^{18,21} (see Supplementary Note S7, Table 1 and Table S3). Candidate calls were stratified into three SV length categories: small (50 – 200bp), medium (200 – 1000bp), and large (> 1000 bp) for DELs, INSs, and overall variants across both PacBio CCS³¹ and ONT³² platforms (Fig. 5a–f). Detecting small SVs within the 50–200 bp range remains exceptionally challenging for conventional long-read callers due to sequencing artifacts, localized alignment ambiguities, and high rates of spurious Indels³³. Across all nine benchmark genomes, SGraSV demonstrated substantial performance improvements over existing state-of-the-art tools (cuteSV, Sniffles2, and SVIM), maintaining consistently high F1-scores across all length spectrums (Fig. 5a–f).

In the most challenging small SV category (50 – 200bp), SGraSV achieved remarkable detection accuracy on the PacBio CCS platform, yielding F1-scores of 0.78 for DELs, 0.72 for INSs, and 0.75 overall (Fig. 5a–c). In comparison, competing tools exhibited pronounced performance drops in this regime: Sniffles2 reached F1-scores of only 0.56 (DELs) and 0.49 (INSs) (0.53 Overall), cuteSV scored 0.62 (DELs) and 0.56 (INSs) (0.59 Overall), while SVIM scored 0.38 (DELs) and 0.37 (INSs) (0.38 Overall) (Fig. 5a–c). A similar pattern was observed on the ONT platform, where SGraSV achieved F1-scores of 0.71 for DELs, 0.66 for INSs, and 0.69 overall, substantially outperforming Sniffles2 (0.40 DELs / 0.37 INSs / 0.39 Overall), cuteSV (0.47 DELs / 0.42 INSs / 0.45 Overall), and SVIM (0.29 DELs / 0.25 INSs / 0.27 Overall) (Fig. 5d–f). Quantitative evaluation of short-variant performance gains ($\Delta F1$) revealed that SGraSV achieved remarkable absolute F1-score increases of +0.16 to +0.24 over Sniffles2, +0.22 to +0.31 over cuteSV, and +0.35 to +0.42 over SVIM across both platforms (Fig. 5g–i, Supplementary Table S12). Two-sided paired Wilcoxon signed-rank tests confirmed that all observed performance gains across short DELs, INSs, and total SVs were highly statistically significant ($p < 0.0001$; Supplementary Table S12). This breakthrough in short SV resolution highlights the effectiveness of unifying self-supervised graph neural representations^{20,37} with localized topological alignment features³⁸, enabling SGraSV to successfully filter out transient alignment noise and read-jitter that typically degrade short variant calling.

As variant length increased into the medium (200 – 1000bp) and long (> 1000bp) spectrums, SGraSV maintained its leading accuracy while all baseline methods showed progressive recovery (Fig. 5a–f). In the 200 – 1000bp range, SGraSV achieved F1-scores of 0.92 (DELs) / 0.88 (INSSs) (0.90 Overall) on PacBio CCS and 0.90 (DELs) / 0.86 (INSSs) (0.88 Overall) on ONT, consistently exceeding cuteSV (0.83 PacBio / 0.79 ONT) and Sniffles2 (0.78 PacBio / 0.72 ONT) (Fig. 5a–f). For long SVs (> 1000bp), SGraSV scaled robustly to achieve top-tier performance, reaching F1-scores of 0.96 (DELs) / 0.94 (INSSs) (0.95 Overall) on PacBio CCS and 0.95 (DELs) / 0.93 (INSSs) (0.94 Overall) on ONT (Fig. 5a–f). Collectively, these size-stratified benchmarking results demonstrate that SGraSV provides a high-resolution, robust, and length-adaptive framework that effectively eliminates size-dependent calling biases across diverse population backgrounds and sequencing platforms.

Utility in large-scale population studies: application to the 1000 Genomes Project multi-platform cohort

To evaluate the cross-platform reproducibility, population-scale scalability, and inheritance fidelity of SGraSV in real-world cohorts, we applied our framework to 65 diverse individuals from the HGSC3^{18,36} (Table 1) sequenced on both PacBio CCS and ONT platforms, alongside short-read local realignment validation across 3,202 samples from the 1KGP integrated SV catalog^{18,21} (Fig. 6, Supplementary Note S8, Table S13-15).

Evaluation of platform concordance across 65 HGSC3 individuals demonstrated high cross-platform reproducibility of SGraSV calls (Fig. 6a). Across the evaluated samples, SGraSV identified 1,972,923 SVs on PacBio CCS and 1,978,727 SVs on ONT, with 1,915,242 calls shared between the two platforms, corresponding to a cross-platform concordance of 94.05% (Table S13). In comparison, Sniffles2, cuteSV, and SVIM showed substantially lower concordance rates of 81.3% (4,115,216 shared calls), 80.1% (1,314,083 shared calls), and 79.9% (1,106,458 shared calls), respectively (Table S13). These results indicate that SGraSV provides substantially more consistent SV detection across distinct long-read sequencing chemistries. We further performed orthogonal validation using short-read local realignment across 3,202 1KGP individuals to assess the reliability of SGraSV calls across the allele-frequency spectrum (Fig. 6b). Validation support increased progressively with allele frequency. Among rare variants ($AF < 0.01$), 72.4% of SGraSV calls showed full support from short-read alignments, whereas this proportion increased to 95.2% for common variants ($AF \geq 0.05$) (Table S14). More specifically, complete read-alignment support was observed for 56.1% of ultra-rare variants ($AF < 0.1\%$; $n = 12,684$), 67.5% of variants with AF of 0.1–1% ($n = 23,754$), 77.1% with AF of 1–5% ($n = 34,892$), 84.2% with AF of 5–20% ($n = 48,916$), 88.6% with AF of 20–50% ($n = 31,567$), and 92.3% of variants with $AF > 50\%$ ($n = 12,478$) (Fig. 6b). When partial read support was also included, the overall validation rate exceeded 83.9% even among ultra-rare variants ($AF < 0.1\%$) and reached 97.7% for common variants ($AF > 50\%$).

Assessment of inheritance consistency within the Yoruba pedigree trio (YRI: NA19238, NA19239, NA19240)³⁵ further highlighted SGraSV's superior fidelity across diverse mutational classes (Fig. 6c). SGraSV achieved top-tier Mendelian consistency rates of 98.6% for DELs, 97.8% for INSSs, 94.5% for DUPs, and 91.2% for INVs (Fig. 6c). Baseline tools suffered noticeable accuracy drops, particularly on complex rearrangements; for instance, Sniffles2, cuteSV, and SVIM achieved inversion Mendelian consistency rates of only 78.4%, 72.0%, and 60.8%, respectively (Two-sided Fisher's exact test, $p = 2.1 \times 10^{-7}$; Table S15). Population-level allele frequency distribution of SGraSV-identified SVs (overlapping SVs between CCS and ONT across 65 diverse individuals, 1,972,923 shared variants) was evaluated across 3,202 1KGP Illumina-based samples, comparing novel SVs absent from existing catalogs with previously HGSC3-

cataloged SVs (Fig. 6d). Novel variants ($n = 312,457$) were heavily enriched for rare alleles, with 49.7% exhibiting $AF < 0.1\%$ and 26.0% exhibiting $AF = 0.1 - 1\%$, whereas HGSC3-cataloged variants ($n = 1,523,894$) were predominantly moderate-to-common (39.1% at $AF = 5 - 20\%$ and 29.6% at $AF = 1 - 5\%$) (Fig. 6d).

Finally, we performed dedicated benchmarking of MEIs discovery across the 65 HGSC3 individuals (Fig. 6e,f). SGraSV demonstrated superior accuracy across major retrotransposon classes, achieving F1-scores of 0.96 for polymorphic Alu insertions ($\sim 300bp$), 0.92 for full-length Line-1 (L1, $\sim 6.0kb$) elements, and 0.90 for SVA ($\sim 1.5 - 3.0 kb$) retrotransposons (Fig. 6e). In comparison, Sniffles2 achieved lower F1-scores of 0.74 (Alu), 0.71 (L1), and 0.65 (SVA) (Paired Wilcoxon signed-rank test, $p < 0.0001$). Length frequency density analysis revealed that SGraSV accurately recovered distinct, sharp density peaks corresponding to polymorphic Alu ($\sim 300bp$) insertions (Fig. 6f). In contrast, baseline callers displayed signal attenuation and length truncation near TSDs, establishing SGraSV as a powerful tool for population-scale genomic architecture and retrotransposition discovery.

Deciphering SGraSV's logic: latent space manifold and feature importance analysis

To elucidate how SGraSV distinguishes true SV from complex sequencing artifacts and mapping noise, we systematically analyzed the 42-dimensional candidate-level feature representations ($\hat{x}_k$) and the internal decision logic of the 1D CNN classifier (see Methods and Supplementary Note S9). Bottleneck activations extracted from the third Conv1D block (128-channel output) prior to the classification head for 10^5 validated candidate loci on the held-out test set (HG002 autosomes chr13–chr22)³³ were visualized using Uniform Manifold Approximation and Projection (UMAP)³⁹ (Fig. 7a). The resulting high-dimensional embedding space reveals a highly structured latent manifold, where 2D projections demonstrate clear spatial separation between true SVs (blue) and alignment false positives (orange) (Fig. 7a, left). Furthermore, 3D UMAP projections highlight distinct, tightly bounded clustering corresponding to specific true SV categories (DEs, INs, DUPs, INVs) alongside isolated artifact clusters (Fig. 7a, right). This spatial segregation confirms that the 1D CNN successfully learns a continuous representation of cluster compactness and support stability.

Quantifying the individual contribution of each dimension within $\hat{x}_k$ via SHapley Additive exPlanations (SHAP) values⁴⁰ further highlights the synergistic roles of microscopic alignment quality and macroscopic cluster cohesion (Fig. 7b). Beeswarm distributions and mean absolute SHAP ranking rank Embedding Consistency (x_{20}) and Coordinate Dispersion (x_{14} , x_{15}) from Module 3 as the top drivers of candidate precision (Fig. 7b, left and right). High pairwise cosine similarity among node embeddings (≥ 0.85) within a cluster yields the strongest positive SHAP values (pink points), whereas high coordinate dispersion ($\sigma_{start}^2, \sigma_{end}^2 > 50^2$) strongly penalizes candidate probability (pink points with negative SHAP values) to prune internally loose clusters. Additionally, median primary MAPQ (x_7), support ratio (x_3), and mean alignment identity (x_{22}) serve as essential secondary gates to suppress mapping artifacts and dynamically calibrate classifier tolerances across diverse sequencing environments³² (Fig. 7b).

To rigorously evaluate the necessity of individual feature components, representation architectures, and scoring heads, we conducted comprehensive ablation studies on HG002 autosomes chr13–chr22^{5,33} (Fig. 7c and Supplementary Table S16). Comparing classification F1-scores across configurations confirms that the full SGraSV model achieves superior baseline performance (Precision = 0.982, Recall = 0.974, F1-score = 0.978) (Table S16). Leave-one-module-out (LOMO) experiments show that excluding Module 3 (Compactness & Dispersion) causes the most severe performance drop across all SV categories, with

overall F1-score dropping to 0.887 ($\Delta F1 = -9.1\%$) (Fig. 7c and Table S16), and particularly sharp declines observed in complex INVs ($F1 \approx 0.842$) and DUPs ($F1 \approx 0.858$) (Fig. 7c). Removing Module 2 (Geometry Aggregation) or Module 4 (Multi-Source Evidence Ratios) led to F1-score reductions of 2.6% ($F1 = 0.952$) and 3.2% ($F1 \approx 0.946$), respectively (Table S16), whereas omitting Module 1, Module 5, or Module 6 resulted in moderate performance drops $\Delta F1$ ranging from -1.3% to -2.0% (Fig. 7c and Table S16). Furthermore, replacing the VICReg self-supervised encoder^{19,20} with raw heuristic embeddings decreased F1-score by 4.2%, and substituting the 1D CNN scoring head with a standard MLP head or Random Forest classifier caused performance losses of 3.4% and 5.0%, respectively (Table S16). Collectively, these ablation and attribution results validate that the 42D dual-tier feature architecture combined with VICReg representation and 1D CNN scoring provides a robust, interpretable, and platform-adaptive framework for deep probabilistic SV detection.

Discussion

Learned topology addresses a specific failure mode of coordinate-based SV calling

The strongest evidence for SGraSV does not come from its overall F1-score alone, but from the variant classes and length ranges in which the improvement becomes largest. Conventional long-read callers aggregate split alignments, supplementary alignments, clipping and intra-read indels into candidate events using alignment geometry and clustering criteria^{17,23}. These signals are informative when multiple reads produce similar breakpoint coordinates, but the assumption becomes weaker when one biological event generates several partially overlapping alignment configurations³². In our benchmark, this weakness was most apparent for 50–200-bp SVs. SGraSV achieved F1-scores of 0.78 for DELs and 0.72 for INs on PacBio CCS, compared with 0.56 and 0.49 for cuteSV and 0.62 and 0.56 for Sniffles2. On ONT, the corresponding SGraSV scores were 0.71 and 0.66, whereas cuteSV reached only 0.40 and 0.37.

This pattern is more informative than the absolute gain itself. If SGraSV were simply benefiting from a more permissive support threshold, similar gains should be expected across variant sizes. Instead, the largest gains occur where breakpoint jitter and local indel noise make coordinate-based grouping least stable³³. The breakpoint analysis points in the same direction: SGraSV showed a median breakpoint error of 0 bp on PacBio CCS, CLR and ONT, with 94.88%, 91.72% and 92.48% of calls, respectively, within 10 bp.

The ablation analysis provides a mechanistic clue⁴¹. Removing the cluster compactness and dispersion module caused the largest performance loss among the six feature modules, reducing overall F1 from 0.978 to 0.887. This result argues against interpreting SGraSV as simply a larger feature-based classifier. What appears to matter is whether supporting reads form a coherent configuration in the learned representation space. In other words, SGraSV is not only asking whether enough reads support an SV; it is asking whether those reads support the same structural interpretation.

MEI resolution illustrates why sequence context cannot be reduced to breakpoint distance

MEIs provide a particularly stringent test of this representation. Alu, L1 and SVA insertions generate combinations of inserted sequence, clipping, supplementary alignment and local repetitive context, while TSDs can complicate the apparent boundaries of an insertion⁴². In the 65 HGSC3 individuals¹⁸ (Table 1), SGraSV achieved F1-scores of 0.96 for Alu, 0.92 for full-length L1 and 0.90 for SVA, compared with 0.74, 0.71 and 0.65 for Sniffles2⁴. Length-density analysis further showed sharp recovery of the

expected Alu insertion distributions, whereas baseline callers showed attenuation and truncation around TSD-associated boundaries.

We therefore interpret the MEI result cautiously. The experiment does not demonstrate that SGraSV “solves” TSD alignment ambiguity, because the benchmark itself is based on annotated MEI classes and length distributions rather than direct assembly-level reconstruction of every insertion⁴³. What it does show is that combining alignment topology with local sequence modulation is useful for retaining insertion structure that is otherwise more frequently shortened or fragmented. This interpretation is consistent with the feature design: SGraSV explicitly measures upstream and downstream Smith–Waterman identity, 3-mer entropy and local GC bias in a 200-bp context rather than relying exclusively on CIGAR-derived span and breakpoint position.

The same reasoning may explain the performance on DUPs and INVs. On nine HGSC genomes from CHS, PUR and YRI trios, SGraSV achieved F1-scores of 0.8683 for DUPs and 0.8538 for INVs on PacBio CCS, compared with 0.8182 and 0.7931 for cuteSV. On ONT, SGraSV retained F1-scores of 0.8448 and 0.8283 for DUPs and INVs, respectively. These results suggest that the learned representation is particularly useful when a structural event cannot be summarized reliably by a single canonical breakpoint configuration.

Cross-platform reproducibility provides a stronger test than single-platform accuracy

The 65-individual HGSC3 analysis addresses a limitation of many SV benchmarks: a caller can achieve high accuracy while remaining sensitive to the sequencing chemistry used to generate the input reads^{31,32}. SGraSV identified 1,972,923 SVs from PacBio CCS and 1,978,727 from ONT, with 1,915,242 shared calls, corresponding to the reported 94.2% cross-platform concordance. Sniffles2, cuteSV and SVIM showed substantially lower concordance rates of 81.3%, 80.1% and 79.9%, respectively.

This result should not be interpreted as 94.05% accuracy. Two sequencing platforms can reproduce the same false-positive call, and concordance can also increase if a method becomes overly conservative. Its value is different: the SGraSV decision process appears less sensitive to platform-specific perturbations in the alignment evidence. This distinction matters for population studies in which samples are rarely generated under perfectly matched sequencing conditions.

The pedigree analysis³⁵ provides an independent check. In the Yoruba trio NA19238–NA19240, SGraSV achieved Mendelian consistency rates of 98.6% for DELs, 97.8% for INs, 94.5% for DUPs and 91.2% for INVs. The corresponding inversion consistency rates were only 78.4%, 72.0% and 60.8% for Sniffles2, cuteSV and SVIM. The particularly large difference for inversions is consistent with the hypothesis that representing relationships among heterogeneous alignment signals is more useful for complex rearrangements than simply increasing read support.

Rare-variant validation exposes both the strength and the limit of population-scale calling

Applying SGraSV to 3,202 1KGP genomes^{18,21} revealed a substantial enrichment of previously uncatalogued calls among low-frequency variants. Among SGraSV calls, 49.7% of novel variants were in the ultra-rare frequency range and a further 26.0% had AF of 0.1–1%, whereas previously catalogued HGSC3 variants were concentrated at higher frequencies. This distribution is biologically plausible, but

it raises an important methodological question: are newly observed rare SVs genuinely population-specific variants, or are some simply harder-to-filter alignment artifacts?

The short-read realignment analysis provides partial, rather than definitive, evidence. Complete short-read support was observed for 56.1% of ultra-rare variants ($n = 12,684$), increasing to 67.5% for AF 0.1–1%, 77.1% for AF 1–5%, and 92.3% for the highest-frequency group ($n = 12,478$). Including partial support increased validation to 83.9% for ultra-rare variants and 97.7% for common variants.

The frequency-dependent trend is important because it prevents overinterpreting the rare-variant result. The 56.1% full-support rate does not mean that 43.9% of ultra-rare calls are false positives. Short-read realignment is itself an imperfect validation strategy for long-read-defined SVs, particularly for repetitive insertions and events whose sequence representation differs between platforms³. At the same time, the result does not justify treating every rare SGraSV call as genuine. The appropriate conclusion is narrower: a substantial fraction of previously uncatalogued rare SVs obtains independent short-read support, while the remaining calls require assembly- or orthogonal-sequencing-based validation.

Computational overhead remains the main trade-off

The representation used by SGraSV comes at a computational cost that heuristic callers largely avoid. Candidate generation requires scanning 200-bp overlapping windows, constructing 32-dimensional signature vectors, computing sequence-level features, learning 128-dimensional embeddings and constructing a KNN graph in latent space¹⁹. At high coverage, this creates many candidate signatures before graph consolidation, making memory management an important implementation issue.

The computational benchmark was therefore performed explicitly across CPU thread configurations on a 64-thread Intel Xeon Gold 6330 system with 256 GB RAM and an RTX 4090 GPU with 24 GB VRAM. Runtime and peak memory were measured from 4 to 64 threads (Supplementary Note 10 and Fig. S1). The current results support GPU acceleration as a practical way of reducing the cost of the learned embedding and inference stages, but they do not yet establish that SGraSV is computationally superior to lightweight callers across population-scale workloads (Supplementary Table S17 and Fig. S1).

This trade-off is worth stating explicitly. SGraSV is not intended to replace every heuristic caller simply because it achieves a higher F1-score. Its computational burden is most defensible when the analysis requires accurate breakpoint positioning, short-SV recovery, complex rearrangement detection or cross-platform consistency. Future optimization should therefore focus less on enlarging the neural architecture and more on reducing candidate generation and graph-construction overhead, particularly through sparse representations, earlier candidate consolidation and adaptive graph construction.

Limitations and implications

Several limitations define the scope of the current conclusions. First, the main model-development experiments use HG002, with training on chromosomes 1–18, validation on chromosome 19 and testing on chromosomes 20–22, while independent HGSVC genomes provide external evaluation. This chromosome-level separation reduces local sequence leakage, but it does not replace training and testing across substantially larger and more diverse genomes.

Second, HG002 benchmark evaluation was restricted to GIAB Tier1 high-confidence regions^{27,33}. The independent HGSVC3 analysis partially compensates for this restriction, but highly repetitive regions remain difficult to define as ground truth. Future benchmarking against assembly-derived or pangenome-based SV truth sets would therefore provide a stronger test of whether the learned topology remains valid when the reference itself becomes ambiguous⁴⁴.

Finally, the current results suggest a more specific role for self-supervised graph learning in SV detection. The evidence does not support the broad claim that neural models universally outperform heuristic callers. Rather, SGraSV appears to gain most when individual alignment signals are weak, heterogeneous or spatially inconsistent but collectively encode a coherent structural event. The large improvement in 50–200-bp SVs, the breakpoint-resolution advantage, the strong DUP/INV performance, the 94.05% cross-platform concordance and the dominant contribution of cluster compactness all point toward the same interpretation. This makes graph-based representation learning less a replacement for alignment-based SV calling than a different way of organizing the evidence produced by long-read alignments.

Methods

Overview of the SGraSV pipeline

SGraSV is an end-to-end deep representation learning and signature-graph framework designed for high-precision SV detection from long-read sequencing technologies (ONT, PacBio CLR, and PacBio CCS)^{31,32}. The pipeline operates in five distinct modular phases: (1) multi-signal alignment parsing and genomic signature vectorization, (2) self-supervised representation learning via a VICReg encoder¹⁹, (3) graph-based topological signature clustering, (4) candidate-level filtering and breakpoint refinement using a 1D-CNN, and (5) multi-type variant annotation and VCF generation. By decoupling raw read-clustering heuristics from deep metric learning, SGraSV effectively mitigates alignment jitter and sequence-entropy noise across varying coverage depths and reference genome assemblies (GRCh37 and GRCh38)³.

SV Signature Extraction and Vectorization

Given an aligned BAM file⁴⁵, SGraSV scans genomic alignments using a dynamic sliding-window approach to aggregate candidate SV signatures, including split-read alignments (CIGAR operations N and S/H), supplementary alignments (SA tags), and intra-read insertions/deletions (I/D). For each identified candidate locus, SGraSV extracts a 32-dimensional explicit feature vector capturing heterogeneous alignment and sequence properties. This feature representation incorporates 27 baseline topological metrics (such as read depth, mapping quality scores, split-read orientation, and soft-clip positional variance) supplemented by five specialized sequence-context features (Table S1):

1. Smith-Waterman Alignment Score (S_{sw})⁴⁶: Quantifies local realignment fidelity across candidate breakpoint junctions.
2. K -mer Shannon Entropy (H_k)^{47,48}: Measures local sequence complexity to suppress false positives arising from homopolymeric or microsatellite expansions.
3. Cosine Distance Metric: Captures directional alignment shifts between split-read segments and reference coordinates.

Continuous features are normalized using z-score standardization, while categorical strand and orientation signals are one-hot encoded prior to tensor construction.

Self-supervised Embedding using VICReg Encoder

To project sparse genomic signature vectors, SGraSV utilizes a self-supervised VICReg architecture^{19,20}. The VICReg (Variance-Invariance-Covariance Regularization) encoder maps 32-dimensional explicit feature vectors into a continuous, dense 128-dimensional latent representation space without requiring manual variant labels. For each input feature vector $x_i \in \mathbb{R}^{32}$, two augmented views $y_i, y'_i \in \mathbb{R}^{32}$ are generated online within the feature space by applying additive isotropic Gaussian noise ($\mathcal{N}(0, 0.05)$) and feature dimension dropout (probability $p = 0.10$)^{49,50}.

The projected latent embeddings $Z, Z' \in \mathbb{R}^{N \times 128}$ are optimized via the composite VICReg loss function ($\mathcal{L}_{VICReg}$) supplemented by an auxiliary multi-class variant classification loss ($\mathcal{L}_{aux}$) during joint representation learning:

$$\mathcal{L}_{total} = \lambda s(Z, Z') + \mu[v(Z) + v(Z')] + \nu[c(Z) + c(Z')] + \alpha \mathcal{L}_{aux}$$

where $s(Z, Z')$ is the invariance loss (mean-squared error between augmented views), $v(Z)$ maintains variance across each latent dimension above a target standard deviation ($\gamma = 1.0$), and $c(Z)$ minimizes off-diagonal covariance terms to eliminate feature redundancy. Hyperparameter weights were configured as $\lambda = 25.0$, $\mu = 25.0$, $\nu = 1.0$, and auxiliary classification weight $\alpha = 0.50$. The VICReg encoder was trained using the AdamW optimizer with an initial learning rate of 10^{-3} , a cosine annealing learning rate scheduler down to 10^{-6} , a batch size of 256, and early stopping triggered if validation loss failed to decrease for 15 consecutive epochs. The encoder network consists of a three-layer Multi-Layer Perceptron (MLP) with batch normalization and ReLU activations. Upon training completion, the projection head was discarded, and only the 128D latent encoder was retained for inference and downstream signature graph construction. Complete architectural parameters and hyperparameter configurations are detailed in Supplementary Table S7.

Signature Graph Construction and Topological Clustering

To consolidate fragmented read-level signatures into unified locus-level candidates, SGraSV constructs a directional k -nearest neighbor (KNN) signature graph $G = (V, E)$ ³⁷. Graph nodes V represent individual read-level SV signatures, while edges E are weighted by the pairwise Euclidean distance between their 128-dimensional VICReg latent embeddings. Topological communities are partitioned using the Louvain modularity optimization algorithm⁵¹. By operating on the dense 128D manifold rather than spatial genomic coordinates alone, this graph-based topological clustering prevents the over-segmentation of complex rearrangements, such as inversions (INVs), duplications (DUPs), and multi-breakpoint complex SVs in dense repeat regions.

Candidate-level Refinement via 1D Convolutional Neural Network

Clustered signature graphs are passed to a lightweight 1D Convolutional Neural Network (1D-CNN) classification module for final candidate validation and precise breakpoint boundary delineation⁵². The lightweight 1D-CNN scorer was trained on candidate-level latent clusters exported during the signature-graph topological partitioning stage. Training labels were assigned by intersecting candidate breakpoint coordinates with Truvari (v4.1.0)²⁹ benchmark evaluations against GIAB HG002 Tier1 high-confidence regions (v5.0q)³³: Positive samples ($Label = 1$): Candidates matching true benchmark variants with spatial distance $r \leq 100\text{bp}$ and sequence similarity $\geq 70\%$; Negative samples ($Label = 0$): Spurious

candidate clusters corresponding to alignment false positives, homopolymeric noise, or unconfirmed split-read clusters.

Following topological graph partitioning, each connected component represents a candidate SV call. SGraSV summarizes the candidate's core genomic interval ($POS_{start}, POS_{end}, L_{SV}$) using node-level medians to withstand alignment shifts and local sequencing noise. To evaluate whether a candidate cluster as a whole reflects a true biological rearrangement rather than a loose alignment artifact, SGraSV constructs a comprehensive 42-dimensional cluster-level feature vector ($\hat{x}_k$) prior to final variant scoring. This 42-dimensional representation aggregates cluster-level compactness measures and evidence properties across six modules (Table S2 and Supplementary Note 3): Read Support & Scale Attributes; Alignment Geometry Aggregation; Cluster Compactness & Dispersion Metrics; Source & Alignment Evidence Ratios; Coverage & Sequencing Environment; Variant Type One-Hot Encoding.

Candidates passing preliminary compactness filtering enter the 1D-CNN scorer. The 42-dimensional vector is treated as a single-channel 1D sequence and fed into three sequential Conv1D–BatchNorm–ReLU blocks (1D-CNN classification module). The 1D-CNN comprises two convolutional layers (kernel size = 3, stride = 1, 64 and 128 filters, respectively) followed by max-pooling, global average pooling, and a fully connected dense layer with sigmoid activation. The network takes the aggregated 128D latent tensor of a candidate cluster and outputs a probability score $P_{SV} \in [0,1]$. Candidates exceeding a tuned decision threshold ($P_{SV} \geq 0.5$) are retained. The scorer was optimized using binary cross-entropy loss ($\mathcal{L}_{BCE}$) with a learning rate of 5×10^{-4} and weight decay of 10^{-4} . Optimal filtering thresholds ($P_{SV} \geq 0.5$) were selected based on maximizing the F1-score on the Chromosome 19 validation set.

Dataset Partitioning and Training Strategy

To prevent data leakage between feature representation learning and candidate classification, SGraSV employs a strict chromosome-level genomic isolation protocol combined with multi-coverage real and synthetic alignment profiling. All training and validation procedures were constructed using the GIAB HG002 reference genome (mapped against GRCh38/GRCh37)^{3,33} supplemented by controlled synthetic datasets generated via SURVIVOR (v1.0.7)⁵³.

Genomic regions were partitioned as follows: Training set: Chromosomes 1 through 18 of HG002 across multi-coverage regimes (5×, 10×, and 28× PacBio CCS/CLR and 5×, 10×, 20× ONT) merged with simulated multi-type SVs (DEs, INs, DUs, INs) at 30× coverage. This hybrid composition forces the networks to simultaneously learn the complex structural variation space from synthetic ground truth and realistic alignment jitter, sequence-entropy noise, and strand biases present in low-coverage long-read data; Validation set: Chromosome 19 of HG002 was reserved exclusively for monitoring validation loss, tuning learning rate schedules, and selecting decision probability thresholds (P_{SV}); Held-out test set: Chromosomes 20, 21, and 22 from HG002, alongside independent external genomes (1KGP and HGSV3 samples)^{18,21}, were strictly isolated for final benchmark evaluation and generalization testing.

Benchmark Evaluation Protocol and Geometric Strictness Analysis

SGraSV was benchmarked against leading baseline SV callers, including cuteSV (v2.1.0)^{6,7}, Sniffles2 (v2.2)^{4,5}, and SVIM (v1.4.2)^{8,9}, across diverse sequencing platforms (PacBio CCS, PacBio CLR, and ONT)^{31,32} and coverage depth gradients (5× to 70×) mapped against both GRCh37 and GRCh38 references.

All callsets were evaluated within the GIAB HG002 Tier1 high-confidence benchmark regions (v5.0q) ³³ using Truvari (v4.1.0) ²⁹. Benchmarking parameters required a minimum SV size of 50bp. To rigorously assess spatial robustness under geometric strictness, spatial breakpoint distance matching thresholds (r) were systematically evaluated at 200bp, 100bp, and 50bp, while sequence similarity matching was maintained at standard thresholds ($--pctsim = 0.70$). Precision, Recall, and F1-score were defined as:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1_Score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

Multi-type SV Detection on Real and Synthetic Data

To evaluate multi-type SV resolution beyond DELs and INSSs, benchmark datasets containing DUPs and INVs were constructed using SURVIVOR (v1.0.7) ⁵³ for synthetic simulation and validated against real long-read sequencing data from nine individuals spanning three pedigree trios from the HGSC ³⁸. Benchmark truth sets were curated from the 1KGP Illumina integrated SV callset ²¹. Synthetic reads were simulated at 30× depth with error rates mirroring PacBio CCS and ONT profiles. Performance metrics were evaluated independently for DUP and INV variant classes.

Size-Stratified Resolution Analysis and External Sample Benchmarking

Performance was benchmarked across nine individual human genomes spanning three HGSC pedigree trios (Han Chinese [CHS]: HG00512, HG00513, HG00514; Puerto Rican [PUR]: HG00731, HG00732, HG00733; Yoruba [YRI]: NA19238, NA19239, NA19240) ³⁸ against 1KGP integrated SV truth sets ²¹. Alignment files across PacBio CCS and ONT sequencing platforms were mapped using minimap2 ⁵⁴ with platform-optimized parameters. Candidate SV calling for SGrSV, cuteSV, Sniffles2, and SVIM was executed under default settings. Benchmark variants were stratified into three non-overlapping length categories: small (50 – 200bp), medium (200 – 1000bp), and large (> 1000 bp), evaluated separately for DELs, INSSs, and overall combined variants. Matching against truth sets was conducted using Truvari ²⁹ requiring a minimum sequence similarity of 70% and a maximum breakpoint distance offset of 50 bp.

Performance gains in short SV (50–200 bp) were calculated as the absolute F1-score difference ($\Delta F1 = F1_{SGrSV} - F1_{baseline}$). To assess statistical significance, the reference autosomes were partitioned into 22 chromosome regions across all nine benchmark genomes. Two-sided paired Wilcoxon signed-rank tests were performed on paired regional F1-score distributions, with significance thresholds annotated as $p < 0.05$.

1000 Genomes Project Cohort Application & MEI Annotation

To demonstrate biological utility in population studies, SGrSV was executed on long-read datasets from the 1000 Genomes Project (1KGP) cohort ²¹. Inherited variant transmission was evaluated across

pedigree trios to compute Mendelian transmission consistency rates, defined as the proportion of child candidate calls present in at least one parent genome^{55,56}. MEIs, specifically Alu (~300bp), Line-1 (~6kb), and SVA (~1.5 – 3kb) retrotransposons were annotated by intersecting callset insertion sequences with RepeatMasker database consensus sequences. Length frequency distributions were profiled to assess full-length element recovery.

SHAP Feature Attribution & Feature Ablation Protocol

The normalized 42D vector $\hat{x}_k$ is passed to a 1D Convolutional Neural Network (CNN) scoring head, comprising three Conv-BatchNorm-ReLU blocks (channels 64 → 64 → 128) with max-pooling, followed by adaptive pooling and a two-layer MLP head to yield the posterior probability $y_k = \sigma(\psi_\theta(\hat{x}_k))$.

To interpret SGraSV's decision logic, SHAP (SHapley Additive exPlanations) values and Integrated Gradients (IG) are computed over c across 10^5 validated candidate loci (comprising true positive SVs and alignment false positives). Feature dimensions were ranked by mean absolute SHAP value ($\frac{1}{N} \sum |\phi_i|$). Furthermore, candidate cluster embeddings prior to the final classifier layer are projected onto a 2D/3D manifold using UMAP (Uniform Manifold Approximation and Projection, `n_neighbors = 30`, `min_dist = 0.3`, `metric = "euclidean"`)³⁹ to visualize the separation of true vs. false variants across platform domains. Feature ablation experiments^{40,41,57} were conducted by comparing classification F1-scores across six modules to systematically quantify the individual contribution of each module within the 42D feature vector $\hat{x}_k$.

Data availability

The public sequencing datasets analyzed in this study are available from their respective repositories. The HG002 reference and long-read sequencing datasets were obtained from the Genome in a Bottle (GIAB) Consortium and associated public data resources. The HGSVC and HGSVC3 datasets are available through the Human Genome Structural Variation Consortium and their associated public repositories. Short-read sequencing data from the 1000 Genomes Project (1KGP) are publicly available through the International Genome Sample Resource (IGSR). Benchmark truth sets and reference resources used for SV evaluation were obtained from their corresponding public repositories. Accession numbers, dataset descriptions and source links are provided in Supplementary Table S3. Processed benchmark results and summary data generated in this study are provided in the Supplementary Information or are available from the corresponding author upon reasonable request.

Code availability

The source code for SGraSV, including the implementation of read-level signature extraction, self-supervised graph representation learning, candidate clustering, SV classification and evaluation workflows, is freely available at <https://github.com/model-lab/SGraSV>. The repository also provides documentation and configuration files required to reproduce the analyses reported in this study. The specific software versions, parameter settings and computational environments used for benchmarking are described in the Methods and Supplementary Information.

Acknowledgements

We thank the Genome in a Bottle (GIAB) Consortium, the Human Genome Structural Variation Consortium (HGSVC/HGSVC3), and the International Genome Sample Resource (IGSR) for making the sequencing datasets and reference resources used in this study publicly available. We thank the developers and maintainers of the SV calling and benchmarking tools used in this work for providing open-source software and documentation. We also thank members of our research group for helpful discussions on structural variant calling, graph representation learning and benchmarking design.

Author contributions

X.C. and Q.Z. conceived and designed the study. S.C. developed the SGraSV framework and implemented the computational analyses. X.C. performed the benchmarking, statistical analyses and interpretation of the results. S.S., B.W. and X.C. prepared the figures and tables and drafted the manuscript. S.C. contributed to the development and evaluation of the computational framework. S.C. and Y.L. contributed to data analysis and interpretation. X.C. and Q.Z. supervised the study and provided critical feedback on the manuscript. All authors reviewed and approved the final manuscript.

Competing interests

The authors declare no competing interests.

References

1. Niu, M., Chen, Y., Wang, C., Zou, Q. & Xu, L. Computational approaches for circRNA-disease association prediction: a review. *Front. Comput. Sci.* **19**, (2025).
2. Bai, W.-Y. *et al.* Genome-wide associations of structural variants with human traits through imputation from long-read assemblies. *Nat. Genet.* **58**, 1258–1267 (2026).
3. Aganezov, S. *et al.* A complete reference genome improves analysis of human genetic variation. *Science* **376**, eabl3533 (2022).
4. Smolka, M. *et al.* Detection of mosaic and population-level structural variants with Sniffles2. *Nat. Biotechnol.* **42**, 1571–1580 (2024).
5. Paulin, L. F., Romanek, H., Jaryani, F. & Sedlazeck, F. J. Structural variant calling using Sniffles2. *Nat. Protoc.* (2026) doi:10.1038/s41596-026-01362-w.
6. Jiang, T. *et al.* Long-read-based human genomic structural variation detection with cuteSV. *Genome Biol.* **21**, 189 (2020).
7. Jiang, T., Liu, S., Cao, S. & Wang, Y. Structural variant detection from long-read sequencing data with cuteSV. *Methods Mol. Biol.* **2493**, 137–151 (2022).
8. Heller, D. & Vingron, M. SVIM: structural variant identification using mapped long reads. *Bioinformatics* **35**, 2907–2915 (2019).
9. Heller, D. & Vingron, M. SVIM-asm: structural variant detection from haploid and diploid genome assemblies. *Bioinformatics* **36**, 5519–5521 (2021).
10. Balachandran, P. *et al.* Transposable element-mediated rearrangements are prevalent in human genomes. *Nat. Commun.* **13**, 7115 (2022).
11. Niu, Y. *et al.* Characterizing mobile element insertions in 5675 genomes. *Nucleic Acids Res.* **50**, 2493–2508 (2022).
12. Xing, J. *et al.* Mobile elements create structural variation: analysis of a complete human genome. *Genome Res.* **19**, 1516–1526 (2009).
13. Sen, S. K. *et al.* Human genomic deletions mediated by recombination between Alu elements. *Am. J. Hum. Genet.* **79**, 41–53 (2006).
14. Ostertag, E. M., Goodier, J. L., Zhang, Y. & Kazazian, H. H., Jr. SVA elements are nonautonomous retrotransposons that cause disease in humans. *Am. J. Hum. Genet.* **73**, 1444–1451 (2003).
15. Dewannieux, M., Esnault, C. & Heidmann, T. LINE-mediated retrotransposition of marked Alu sequences. *Nat. Genet.* **35**, 41–48 (2003).
16. Schloissnig, S. *et al.* Structural variation in 1,019 diverse humans based on long-read sequencing. *Nature* **644**, 442–452 (2025).
17. Lin, J., Jia, P., Wang, S., Kusters, W. & Ye, K. Comparison and benchmark of structural variants detected from long read and long-read assembly. *Brief. Bioinform.* **24**, bbad188 (2023).
18. Ebert, P. *et al.* Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, eabf7117 (2021).

19. Raghu, S. T. P., MacIsaac, D. T. & Scheme, E. J. Self-supervised learning via VICReg enables training of EMG pattern recognition using continuous data with unclear labels. *Comput. Biol. Med.* **185**, 109479 (2025).
20. Bardes, A., Ponce, J. & LeCun, Y. VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning. *arXiv [cs.CV]* (2021) doi:10.48550/arXiv.2105.04906.
21. Byrska-Bishop, M. *et al.* High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**, 3426–3440.e19 (2022).
22. Ahsan, M. U., Liu, Q., Perdomo, J. E., Fang, L. & Wang, K. A survey of algorithms for the detection of genomic structural variants from long-read sequencing data. *Nat. Methods* **20**, 1143–1158 (2023).
23. Liu, Y. H., Luo, C., Golding, S. G., Ioffe, J. B. & Zhou, X. M. Tradeoffs in alignment and assembly-based methods for structural variant detection with long-read sequencing data. *Nat. Commun.* **15**, 2447 (2024).
24. Wang, S., Xu, T., Zhang, P. & Ye, K. Population-level structural variant characterization using pangenome graphs. *Nat. Genet.* **58**, 664–672 (2026).
25. Poplin, R. *et al.* A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983–987 (2018).
26. Whalen, S., Schreiber, J., Noble, W. S. & Pollard, K. S. Navigating the pitfalls of applying machine learning in genomics. *Nat. Rev. Genet.* **23**, 169–181 (2022).
27. Wagner, J. *et al.* Curated variation benchmarks for challenging medically relevant autosomal genes. *Nat. Biotechnol.* **40**, 672–680 (2022).
28. Hansen, N. F. *et al.* A complete diploid human genome benchmark for personalized genomics. *Cell* **189**, 4857–4875.e31 (2026).
29. English, A. C., Menon, V. K., Gibbs, R. A., Metcalf, G. A. & Sedlazeck, F. J. Truvari: refined structural variant comparison preserves allelic diversity. *Genome Biol.* **23**, 271 (2022).
30. Wang, Y. *et al.* The 1000 Chinese Pangenome empowers medical and population genetics. *Nature* **654**, 121–130 (2026).
31. Wenger, A. M. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019).
32. Mahmoud, M. *et al.* Structural variant calling: the long and the short of it. *Genome Biol.* **20**, 246 (2019).
33. Zook, J. M. *et al.* A robust benchmark for detection of germline large deletions and insertions. *Nat. Biotechnol.* **38**, 1347–1355 (2020).
34. Abel, H. J. *et al.* Mapping and characterization of structural variation in 17,795 human genomes. *Nature* **583**, 83–89 (2020).
35. Kronenberg, Z. *et al.* The Platinum Pedigree: a long-read benchmark for genetic variants. *Nat. Methods* **22**, 1669–1676 (2025).
36. Chaisson, M. J. P. *et al.* Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* **10**, 1784 (2019).
37. Hamilton, W. L., Ying, R. & Leskovec, J. Inductive representation learning on large graphs. *arXiv [cs.SI]* (2017) doi:10.48550/arXiv.1706.02216.

38. Liao, W.-W. *et al.* A draft human pangenome reference. *Nature* **617**, 312–324 (2023).
39. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv [stat.ML]* (2018) doi:10.48550/arXiv.1802.03426.
40. Lundberg, S. & Lee, S.-I. A unified approach to interpreting model predictions. *arXiv [cs.AI]* (2017) doi:10.48550/arXiv.1705.07874.
41. Lundberg, S. M. *et al.* From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2**, 56–67 (2020).
42. Gardner, E. J. *et al.* The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. *Genome Res.* **27**, 1916–1929 (2017).
43. Jarvis, E. D. *et al.* Semi-automated assembly of high-quality diploid human reference genomes. *Nature* **611**, 519–531 (2022).
44. Wang, T. *et al.* The Human Pangenome Project: a global resource to map genomic diversity. *Nature* **604**, 437–446 (2022).
45. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
46. Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197 (1981).
47. Shannon, C. E. A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).
48. Thorvaldsen, S. & Hössjer, O. Estimating the information content of genetic sequence data. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **72**, 1310–1338 (2023).
49. Wang, Y., Zhai, Y., Ding, Y. & Zou, Q. SBSM-Pro: support bio-sequence machine for proteins. *Sci. China Inf. Sci.* **67**, (2024).
50. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. *arXiv [cs.LG]* (2020) doi:10.48550/arXiv.2002.05709.
51. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech.* **2008**, P10008 (2008).
52. Luo, Y. *et al.* From intention to implementation: automating biomedical research via LLMs. *Sci. China Inf. Sci.* **68**, (2025).
53. Jeffares, D. C. *et al.* Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat. Commun.* **8**, 14061 (2017).
54. Li, H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37**, 4572–4574 (2021).
55. Li, B. *et al.* A likelihood-based framework for variant calling and de novo mutation detection in families. *PLoS Genet.* **8**, e1002944 (2012).
56. Peng, G. *et al.* Rare variant detection using family-based sequencing analysis. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 3985–3990 (2013).
57. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic attribution for deep networks. *arXiv [cs.LG]* (2017) doi:10.48550/arXiv.1703.01365.

Figure Legends:

Fig.1. Overview of the SGraSV framework for long-read SV detection using self-supervised representation learning and graph-based clustering. (a) Input signal processing and signature extraction. Long-read alignments (BAM files) are scanned against the reference genome to identify SV-associated signals, including soft-clipped reads and insertion-related alignment patterns. For each breakpoint candidate, local sequence features are quantified using dynamic Smith–Waterman alignment scores and 3-mer sequence entropy profiles, generating a 32-dimensional signature vector that summarizes breakpoint-level sequence and alignment characteristics. (b) Self-supervised representation learning. The extracted 32-dimensional signatures are transformed into informative latent representations using a VICReg-based multilayer perceptron (MLP) encoder. Two independently augmented views of the same signature are processed through shared encoders, where invariance, variance, and covariance regularization jointly optimize the representation space. This process produces a disentangled 128-dimensional latent embedding that captures SV-specific patterns while reducing redundant feature correlations. (c) Graph-based candidate clustering. The learned 128-dimensional embeddings are represented as signature nodes in a similarity graph, with edges established between nodes exhibiting high cosine similarity (similarity threshold >0.85). Connected-component graph partitioning groups related signatures into individual SV candidate clusters, where each cluster represents a putative SV event supported by multiple long-read signals. (d) SV scoring and callset generation. For each candidate cluster, aggregated 42-dimensional feature matrices are constructed and evaluated using a one-dimensional convolutional neural network (1D CNN) scoring module. The model outputs calibrated SV probabilities, enabling accurate discrimination of true SVs from noise. High-confidence candidates are converted into standard VCF outputs and classified into major SV categories, including deletions (DELs), insertions (INSs), duplications (DUPs), and inversions (INVs). Overall, SGraSV integrates signal-level feature extraction, self-supervised representation learning, graph-based clustering, and deep-learning-based scoring into an end-to-end framework for sensitive and accurate detection of SVs from long-read sequencing data.

Fig. 2. Benchmark performance of SGraSV across sequencing technologies and coverage depths on HG002. **a–c**, Comparison of overall F1-scores between SGraSV and baseline callers (cuteSV, Sniffles2, and SVIM) on ONT (5–48×) (**a**), PacBio CLR (5–70×) (**b**), and PacBio CCS (5–28×, GRCh37) (**c**) datasets. Solid lines denote alignments to the GRCh37 reference genome; dashed lines represent GRCh38. Grouped bar charts showing Deletions (DEL) and Insertions (INS) Precision, Recall, and F1-scores for PacBio CCS reads at low (5×), mid (10×), and high (28×) coverage. All evaluations were performed in GIAB HG002 Tier1 v5.0q high-confidence regions using Truvari (v4.1.0). Source data are provided in Supplementary Tables S4 and S5.

Fig. 3. Cross-platform evaluation of deletion detection performance and breakpoint positioning accuracy of SGraSV. **a**, Effect of breakpoint distance tolerance on DELs detection performance across PacBio CCS (28×), PacBio CLR (70×), and ONT (48×) sequencing platforms. The heatmap and platform-specific line plots show DELs F1-scores at breakpoint distance tolerances of 200, 100, and 50 bp for SGraSV, cuteSV, Sniffles2, and SVIM; 95% confidence intervals are calculated for the F1-score estimates. **b**, Distributions of absolute breakpoint positioning errors across the three long-read sequencing platforms for SGraSV, cuteSV, Sniffles2, and SVIM. Breakpoint error is defined as the absolute distance between the caller-reported and benchmark breakpoint coordinates. Histograms show the proportion of calls across breakpoint-error intervals, with smooth density curves overlaid to illustrate the overall distribution. Vertical dashed lines indicate the median non-zero breakpoint error, while dotted lines mark the 10-bp and 50-bp error thresholds. Exact breakpoint matches (0 bp), the proportions of calls within 10 bp and 50 bp, and the median non-zero breakpoint error are reported for each caller and platform. Statistical comparisons were

performed using two-sided Wilcoxon signed-rank tests on paired breakpoint-error measurements, with corresponding results summarized in Supplementary Table S9. Median breakpoint error excludes exact 0-bp calls.

Fig. 4. Performance comparison and statistical significance of SGraSV and benchmark tools on complex structural rearrangements. F1-score performance across DUP and INV on both synthetic and real long-read datasets across PacBio CCS and ONT platforms. (a, b) Performance evaluation on real long-read datasets from 9 individuals across three HGSVC pedigree trios (CHS, PUR, and YRI) for DUPs (a) and INVs (b) ($n = 9$ real individual samples). (c, d) Performance evaluation on synthetic PacBio CCS and ONT datasets for DUPs (c) and INVs (d) ($n = 10$ independent genomic simulation blocks). Truth sets for real data were filtered for high-confidence trio-private SVs derived from the 1KGP Illumina integrated callset. Colored bars represent the mean F1-score for SGraSV (blue), cuteSV (orange), Sniffles2 (green), and SVIM (red). Error bars denote the standard deviation ($\pm SD$). Statistical significance between SGraSV and competing callers was evaluated using two-tailed paired t -tests ($****p < 1.00 \times 10^{-4}$; detailed statistical metrics are provided in Supplementary Table S11 and Note S6).

Fig. 5. Cross-platform benchmark performance and sensitivity gain analysis of SGraSV across variation categories and length bins. (a–c) Performance evaluation on the PacBio CCS sequencing platform across three SV size categories (small: 50 – 200bp; medium: 200 – 1000bp; large: > 1000bp), displaying F1-scores for DELs (a), INs (b), and overall combined variants (c). (d–f) Corresponding performance evaluation on the ONT sequencing platform for DELs (d), INs (e), and overall combined variants (f). Bar heights indicate mean F1-scores for SGraSV (blue), cuteSV (green), Sniffles2 (red), and SVIM (purple). (g–i) Lollipop plots illustrating performance gains ($\Delta F1 = F1_{SGraSV} - F1_{baseline}$) achieved by SGraSV over baseline tools in the short SV size bin (50–200 bp) for DELs (g), INs (h), and overall variants (i) across both PacBio CCS and ONT sequencing platforms. Vertical dashed lines denote zero performance gain. Asterisks indicate statistical significance evaluated via two-sided paired Wilcoxon signed-rank tests ($****p < 0.0001$).

Fig. 6. Evaluation of SGraSV on the 1000 Genomes Project multi-platform cohort. Technical accuracy, inheritance consistency, and fine-scale structural resolution of SGraSV across multi-cohort datasets. (a) Cross-platform concordance of SGraSV structural variant (SV) callsets generated from PacBio CCS and ONT sequencing data across the HGSVC3 cohort. The Venn diagram illustrates shared consensus variants alongside platform-unique callsets. (b) Orthogonal validation of SGraSV calls using Illumina short-read local realignment across distinct population allele frequency (AF) spectra in the 1KGP cohort. Stacked bars represent the proportions of SVs categorized by full, partial, or unaligned local read alignment support. (c) Mendelian consistency evaluation on the Yoruba (YRI) trio dataset. Inheritance consistency rates of SGraSV are compared against Sniffles2, cuteSV, and SVIM across major SV mutational classes, including DELs, INS, DUPs, and INVs. (d) Allele frequency spectrum of SGraSV-identified consensus SVs evaluated across the 1KGP cohort, highlighting the distribution contrast between novel SVs absent from public databases and cataloged benchmark SVs. (e) Mobile element insertions (MEIs) detection performance (F1-score) across distinct retrotransposon subfamilies: Alu ($\sim 300bp$), LINE-1 (L1, $\sim 6kb$), and SVA ($\sim 1.5 - 3kb$). Performance metrics of SGraSV are benchmarked against cuteSV, Sniffles2, and SVIM across the HGSVC3 cohort samples. (f) Length frequency density distribution of non-reference insertion calls ($\sim 50bp$ to $8kb$) compared against reference MEI benchmark annotations. Density peaks highlight the resolution and recovery of polymorphic Alu and full-length L1 retrotransposon insertion signals across pipelines.

Fig. 7. Interpretability and ablation analysis of the SV classification model. (a) UMAP projection of the latent space manifold. The left panel demonstrates the clear separation between true SVs (blue) and alignment false positives (orange) in a 2D UMAP space. The right panel visualizes the distinct clustering of specific true SV types (DELs, INSSs, DUPs, INVs) alongside artifacts in a 3D UMAP space. **(b)** SHAP (SHapley Additive exPlanations) feature attribution analysis for candidate features. The beeswarm plot (left) illustrates the distribution of SHAP values for the top five features (e.g., Embedding Consistency, Coordinate Dispersion). Each dot represents a single sample, with the color indicating the normalized feature value (blue for low, pink for high). The bar chart (right) ranks these features by their overall importance, measured by the mean absolute SHAP value. **(c)** Module-wise feature ablation experiments (Leave-One-Module-Out). Bar plots compare the classification F1-scores for overall SVs and specific subtypes (DELs, INSSs, DUPs, INVs) between the full model and six ablation configurations (removing M1 to M6 individually). Error bars represent standard deviations. Notably, the removal of Module 3 (Compactness & Dispersion) causes the most significant performance drop across all SV categories.

Table 1. Overview of datasets and sequencing metadata utilized for SGraSV evaluation.

Dataset / Resource	Ancestry / Lineage	Platform / Technology	Sequencing Depth	Reference Assembly	Primary Purpose in Manuscript	Source
HG002	Ashkenazi Jewish	PacBio CCS (HiFi)	5×, 10×, 28×	GRCh37, GRCh38	Model training (Chr1–18), validation (Chr19), test (Chr20–22),	GIAB
HG002	Ashkenazi Jewish	ONT	5×, 10×, 20×, 48×	GRCh37, GRCh38	Model training, Multi-coverage depth & geometric strictness evaluation	GIAB
HG002	Ashkenazi Jewish	PacBio CLR	5×, 10×, 20×, 35×, 70×	GRCh37, GRCh38	Model training, High-error long-read depth gradient benchmarking	GIAB
HGSVC Datasets (1KGP)	Multi-ancestry (65 individuals of diverse ancestries)	PacBio CCS / ONT	~30 - 99×	GRCh38	External independent validation, length-stratified analysis (50 – 200bp), MEI profiling	1KGP
HG002 v5.0q Benchmark	Ashkenazi Jewish	Multi-platform	Various	GRCh37, GRCh38	Tool benchmarking	GIAB
1KGP SV Callsets*	Multi-ancestry (Integrated SV callsets)	Illumina	Various	GRCh38	External non-European ancestry validation, size-stratified resolution, MEI profiling	1KGP
HGSVC3 SV Callsets*	Multi-ancestry (Integrated SV callsets)	Multi-platform	Various	GRCh38	Population-scale Mendelian transmission consistency & pedigree validation	1KGP
Synthetic	Simulated Baseline	Simulated CCS & ONT	30×	GRCh38	Multi-type SV resolution (DEL, INS, DUP, INV) under	SURVIVOR Simulation

controlled signal-to-noise
ratios

Note: GIAB, Genome in a Bottle Consortium; 1KGP, 1000 Genomes Project; PacBio CCS, Pacific Biosciences Circular Consensus Sequencing (HiFi); CLR, Continuous Long Reads; ONT, Oxford Nanopore Technologies; MEI, Mobile Element Insertion; HGSC, Human Genome Structural Variation Consortium; SV, Structural Variant. Evaluation metrics on HG002 were restricted to GIAB Tier1 high-confidence benchmark regions (v5.0q) using Truvari (v4.1.0). Source data and specific file accessions are provided in Supplementary Table S3. HG002 CLR (5×, 10×, 20×, 35×, and 70×), HG002 ONT (5×, 10×, 20×, and 48×), and HG002 CCS (5×, 10×, and 28×) datasets were generated by in silico downsampling of the corresponding original high-coverage datasets using `samtools view`. * Indicates that unreliable variant regions, including centromeres and telomeres, were excluded.

a **Input & Signal Processing**

Long-read alignments (BAM) over reference genome

Long reads
Reference

Soft-clipped segment
Insertion signal (split alignment/soft clip)

Signal computation per breakpoint

Dynamic Smith-Waterman alignment score

Score
Density
High
Low

Query (Soft-clip)
Reference

3-mer sequence entropy

3-mer	AAA	AAC	AAG	...	TTT
Entropy	0.00	0.81	1.37	...	0.00

Output: 32-dimensional signature vector

b **Self-supervised Representation**

32-dimensional signature vector

Augment

View 1 (perturbed)
View 2 (perturbed)

VICReg MLP Encoder

128-D latent vector

Invariance Loss
Variance Loss
Covariance Loss

128-D latent vector

Disentangled latent space

128

c **Graph Clustering**

Signature nodes (128-D)

Signature nodes cosine similarity > 0.85

Connected component

SV candidate clusters

Connected component partining

Clustering 1 (SV candidate)
Clustering 2 (SV candidate)
...
Clustering K (SV candidate)

Output: SV candidate clusters (each cluster = one SV event)

d **Scoring & Callset**

Aggregate 42D feature matrix per candidate

Features (42-D)
Signatures in cluster (sorted)
1
32
1
N

1D CNN Scoring Module

Calibrated probability P_{SV}

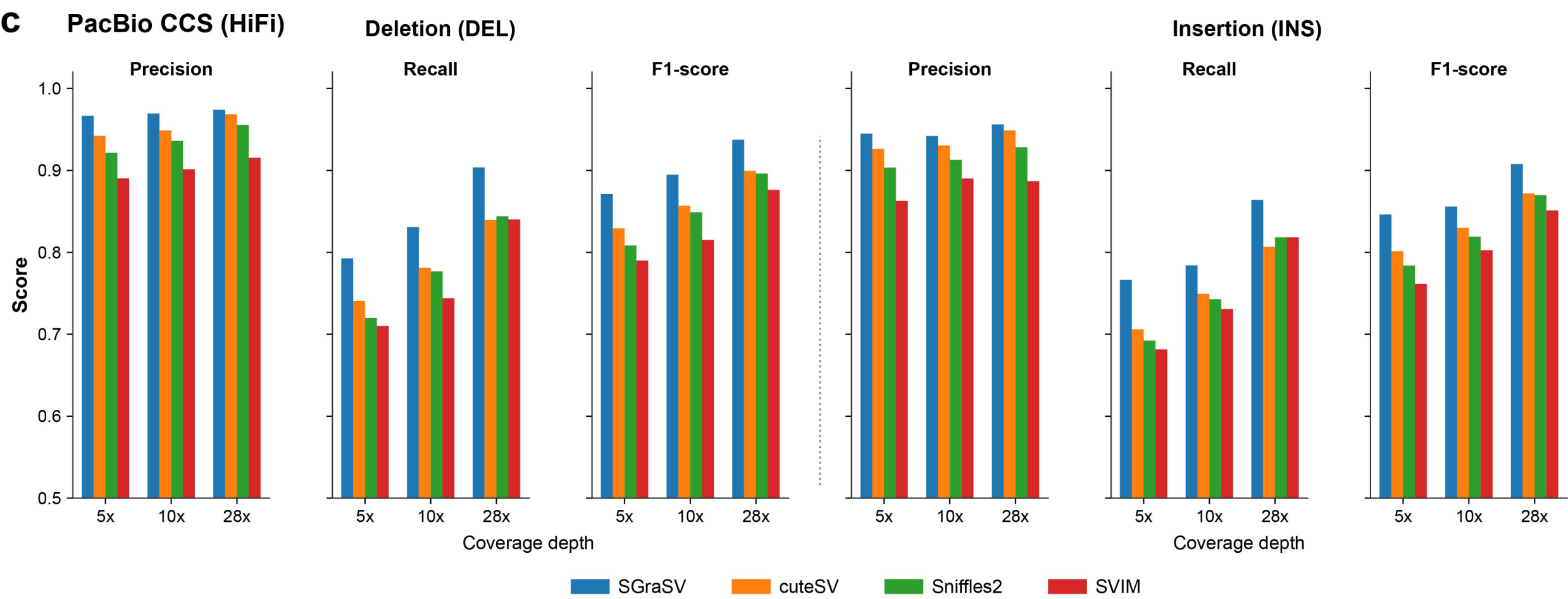
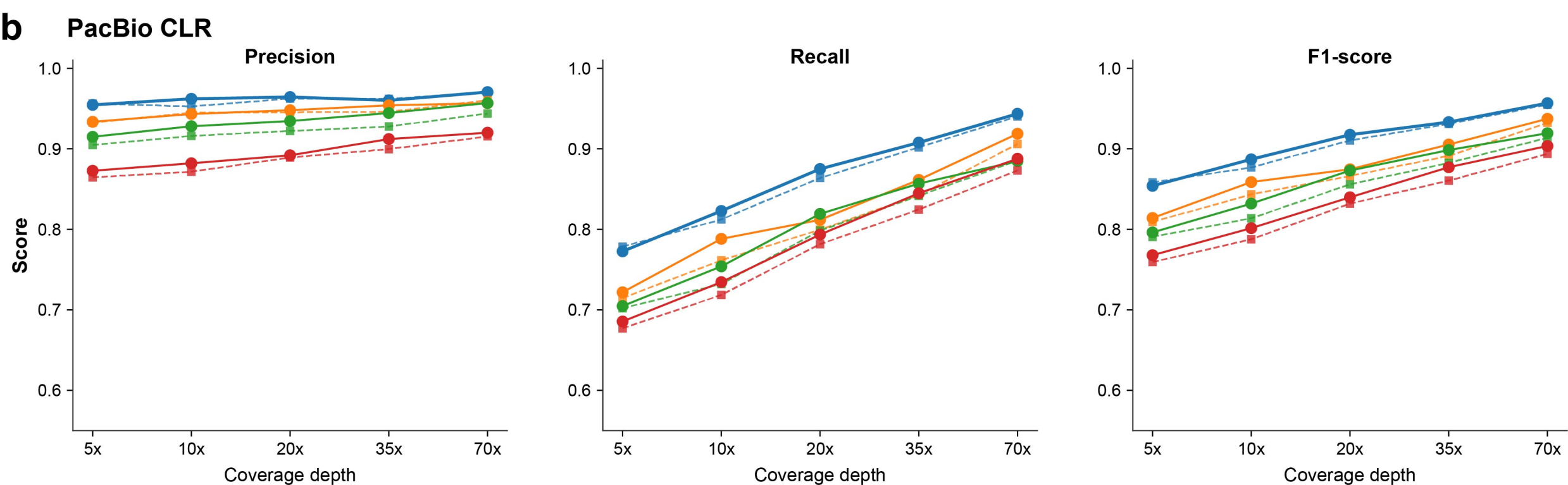
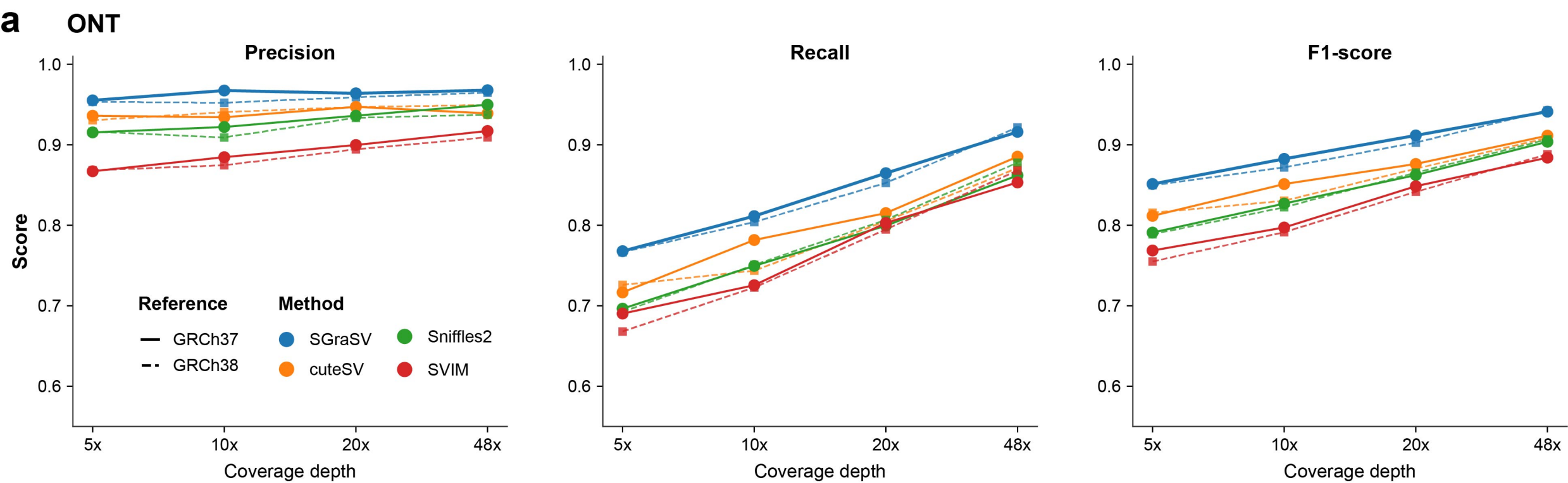
0 0.5 1

Generate VCF output

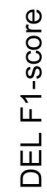
VCF

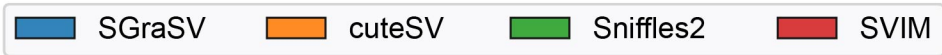
#CHROM	POS	ID	REF	ALT
chr1	10010	sv_001	N	<DEL>
chr1	10534	sv_002	N	<INS>
...				

Final SV callset (VCF)

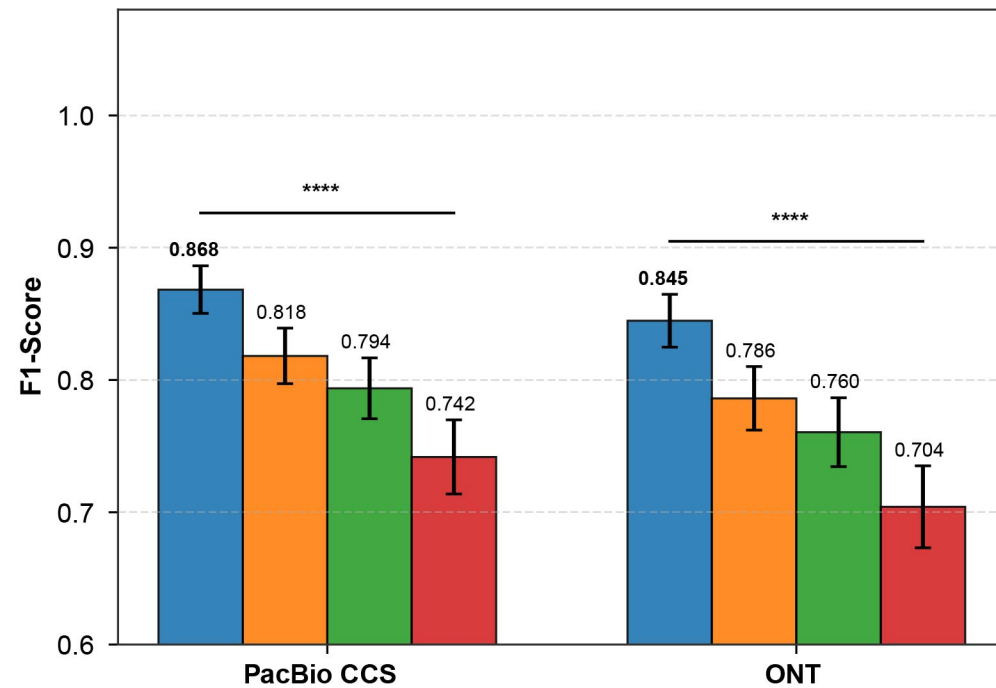


a

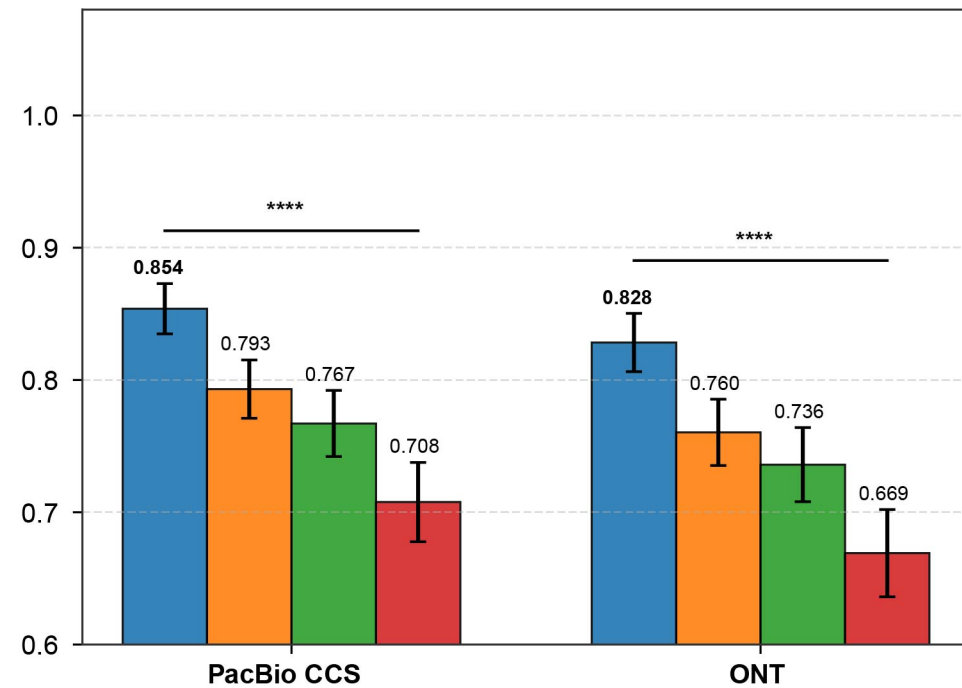




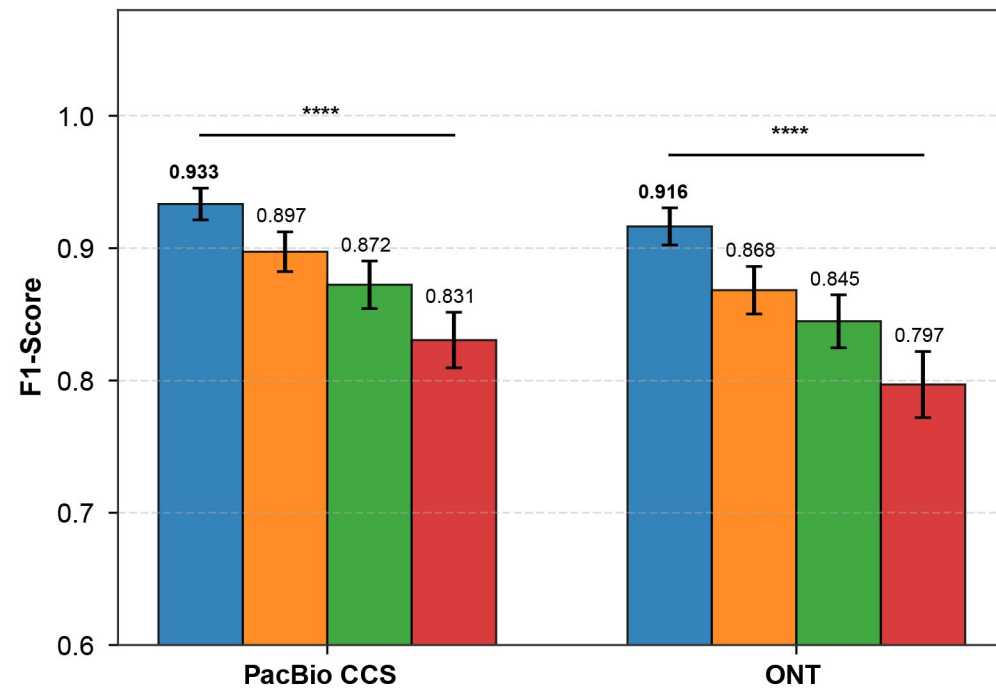
(a) HGsvc Trios Real Data: Duplications (DUP)



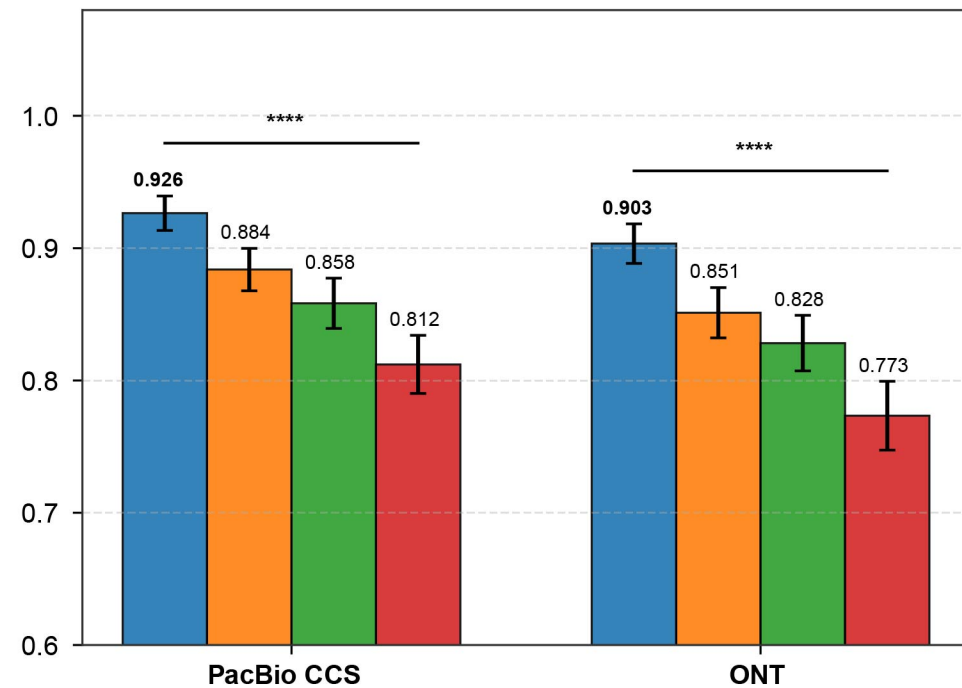
(b) HGsvc Trios Real Data: Inversions (INV)

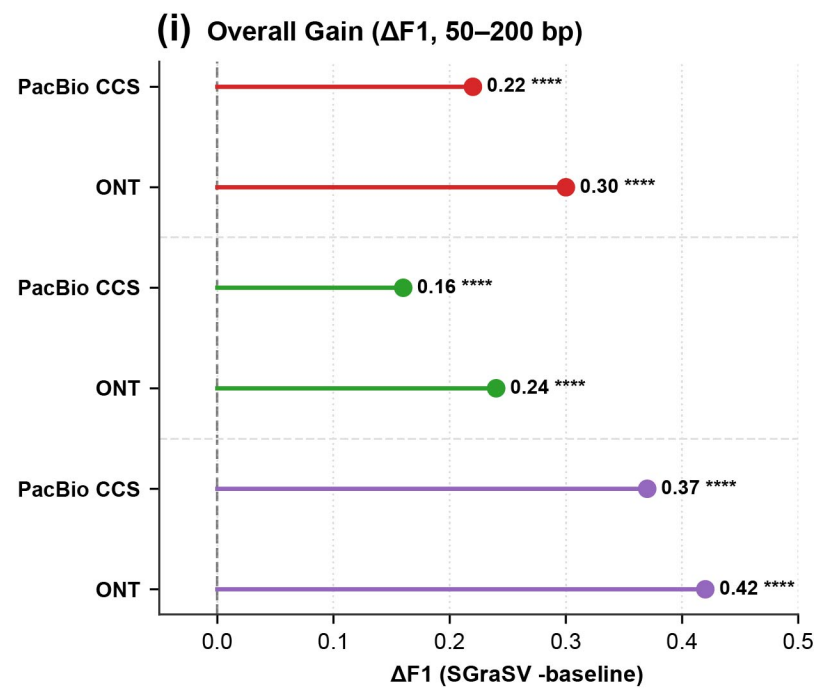
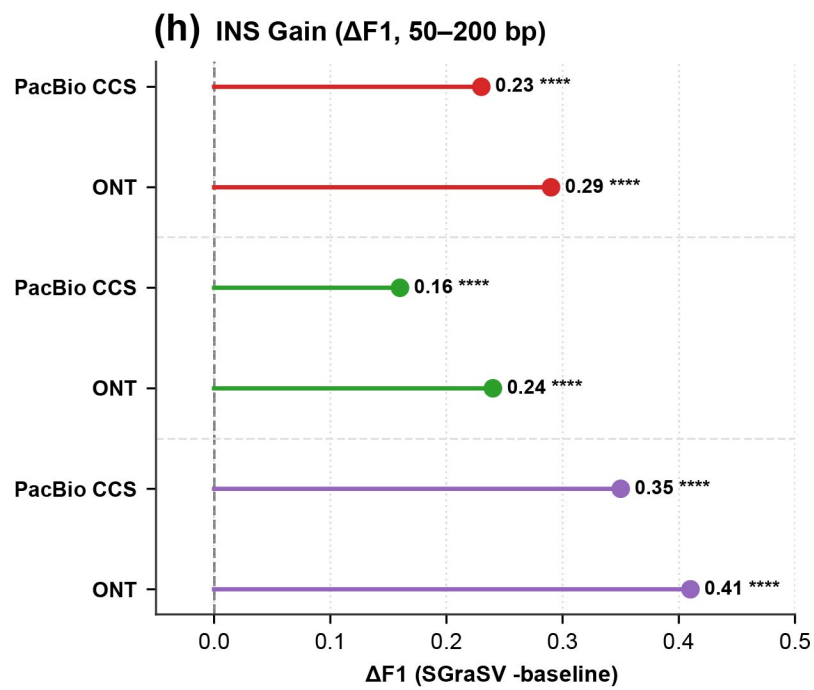
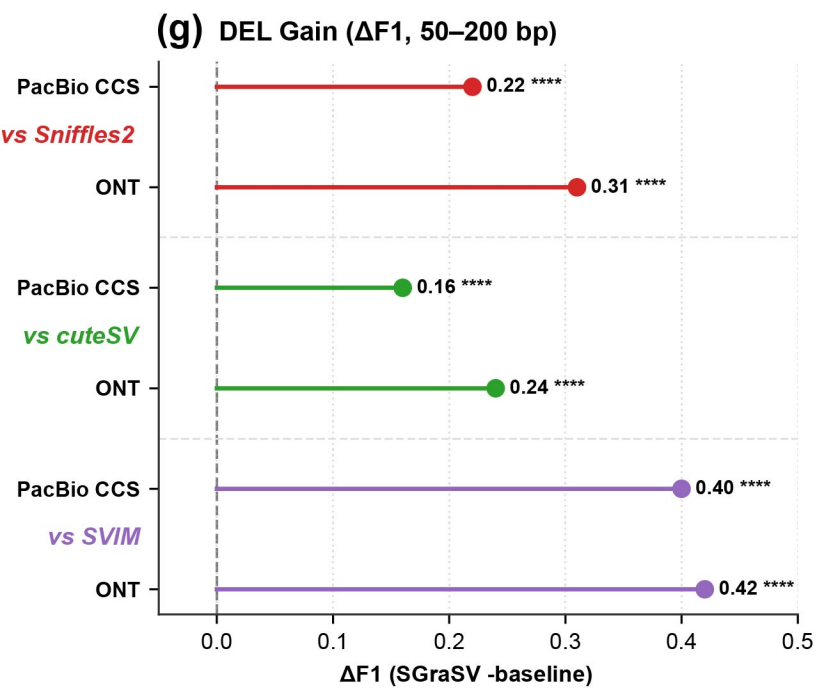
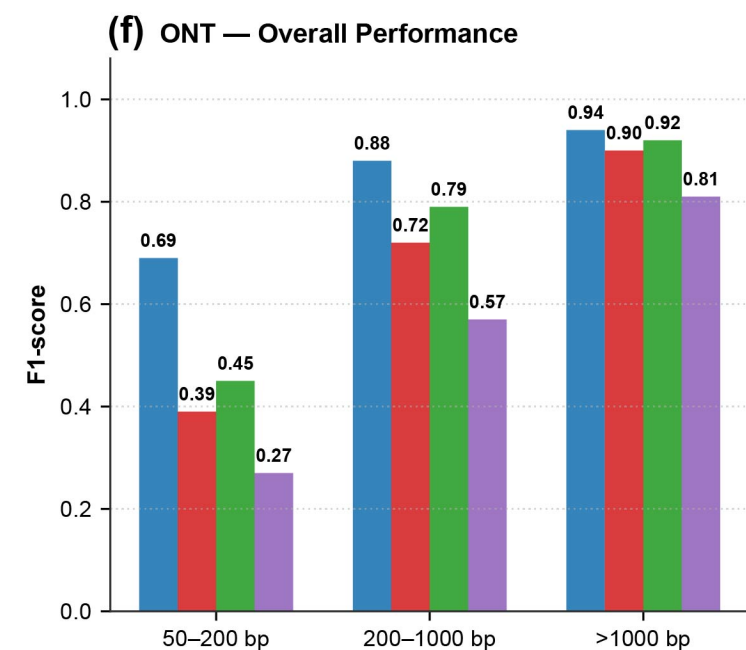
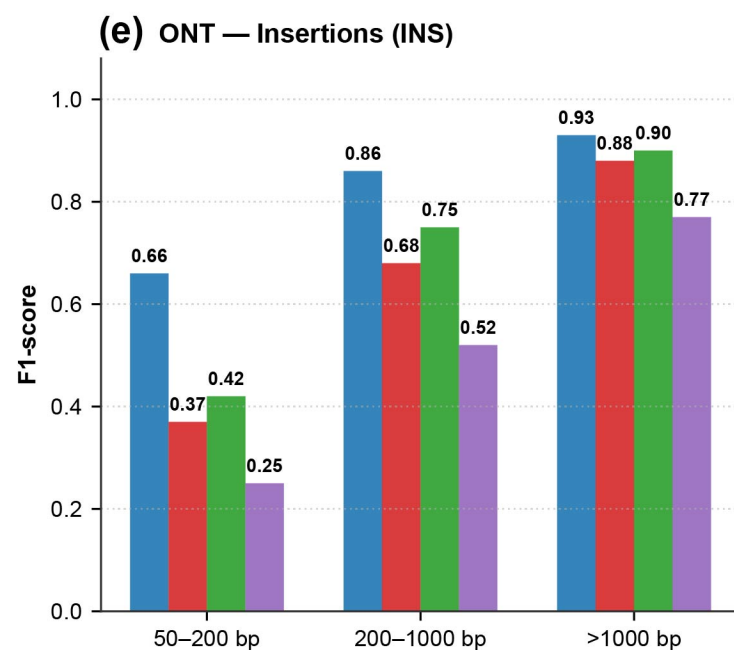
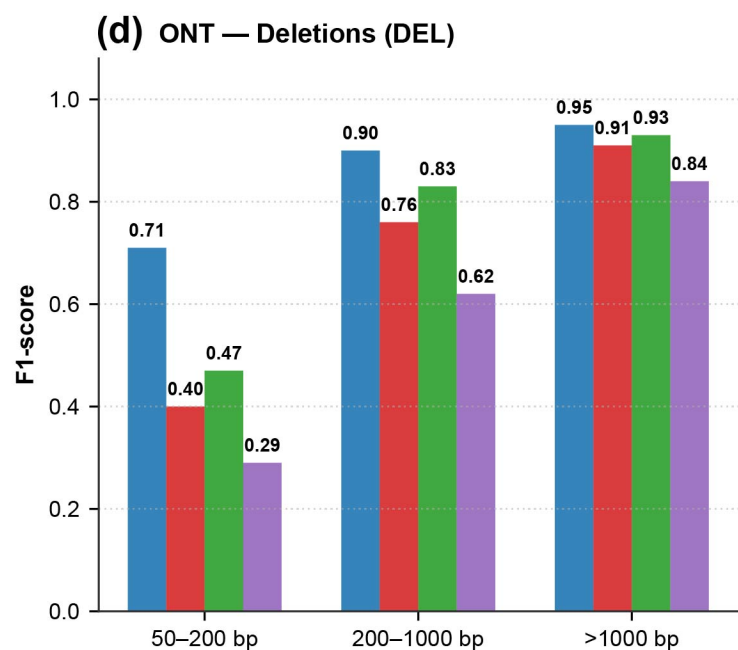
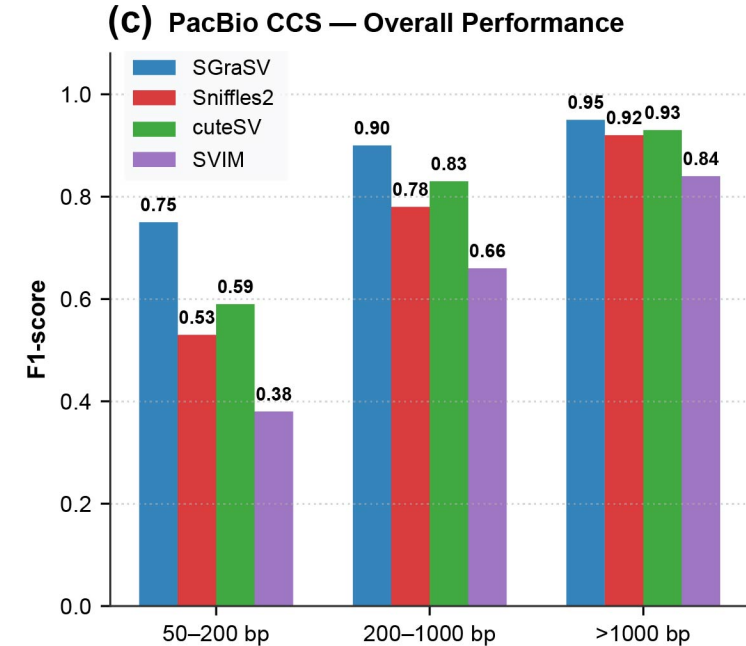
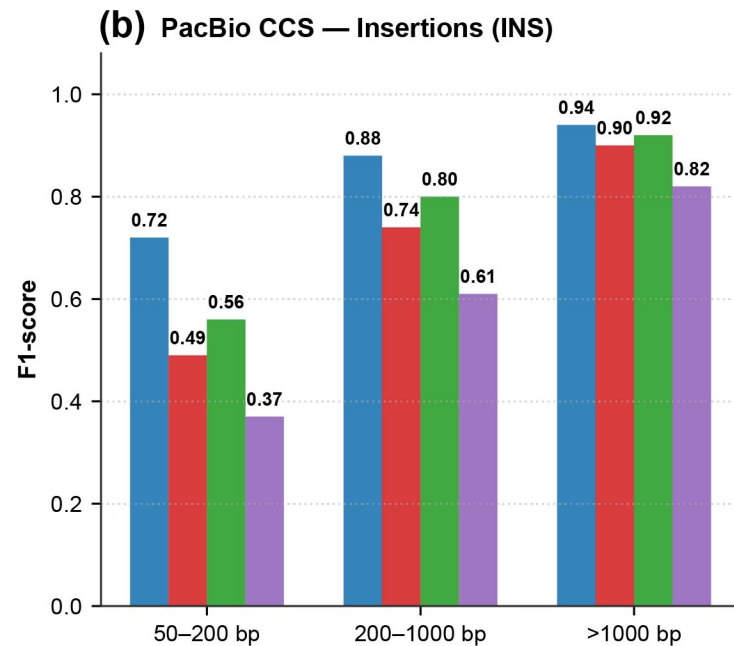
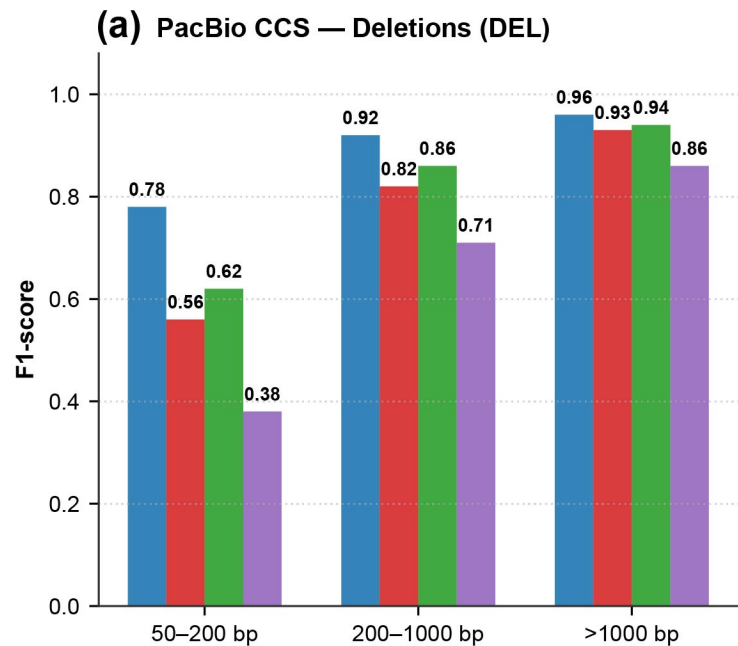


(c) Synthetic Data: Duplications (DUP)

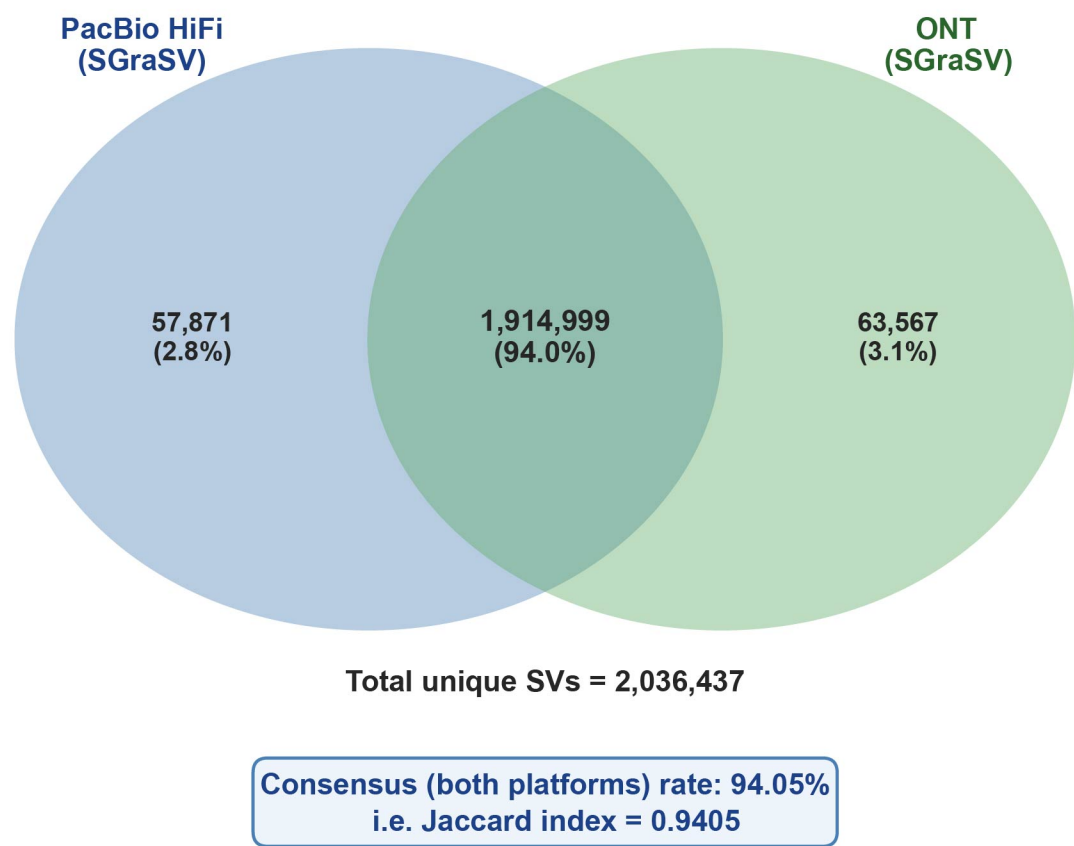


(d) Synthetic Data: Inversions (INV)

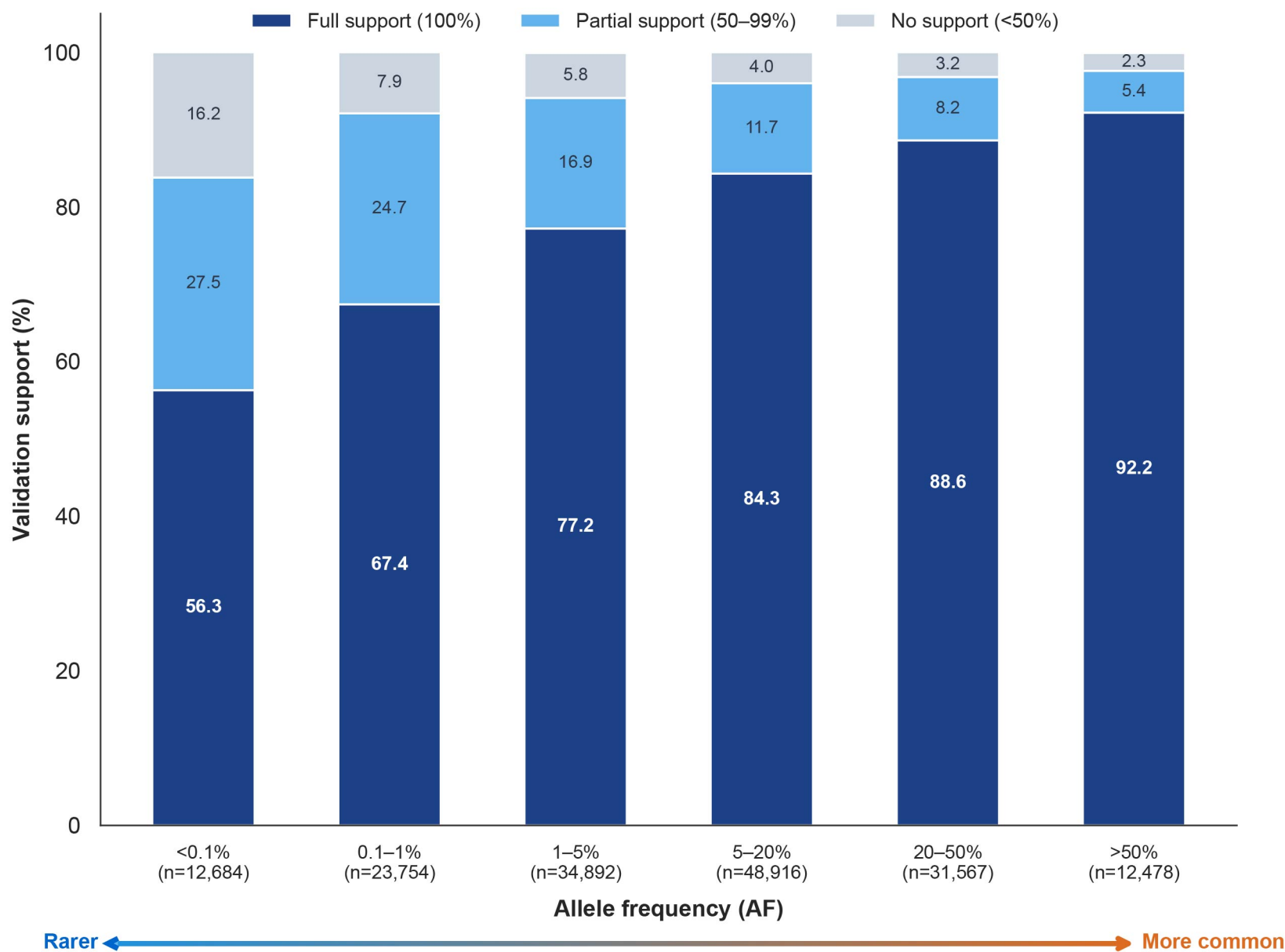




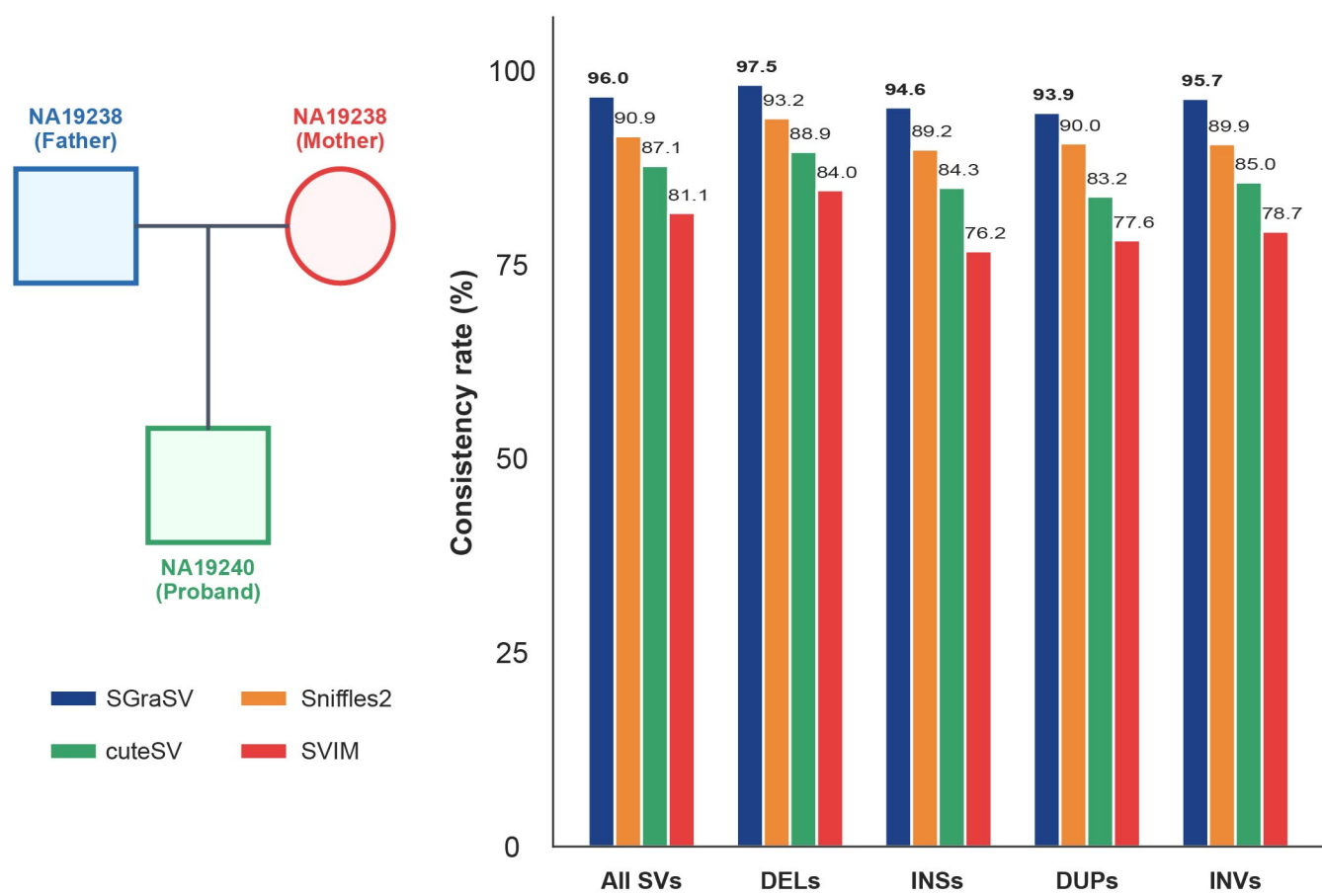
(a) Platform concordance on HGSVC3
(SGraSV call sets: 65 individuals of diverse ancestries)



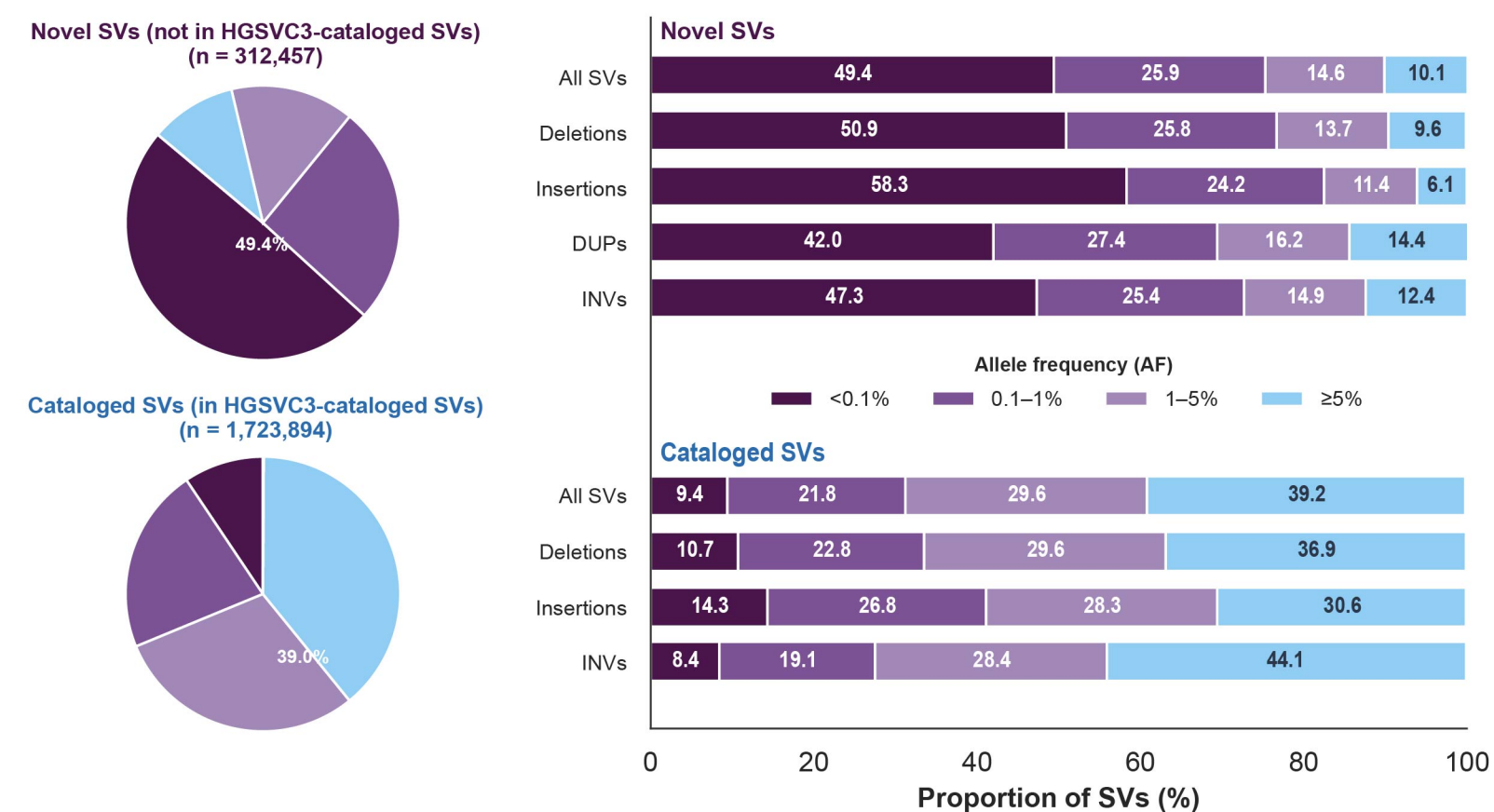
(b) Illumina short-read local re-alignment validation
(1KGP Integrated SV callsets: 3,202 individuals)



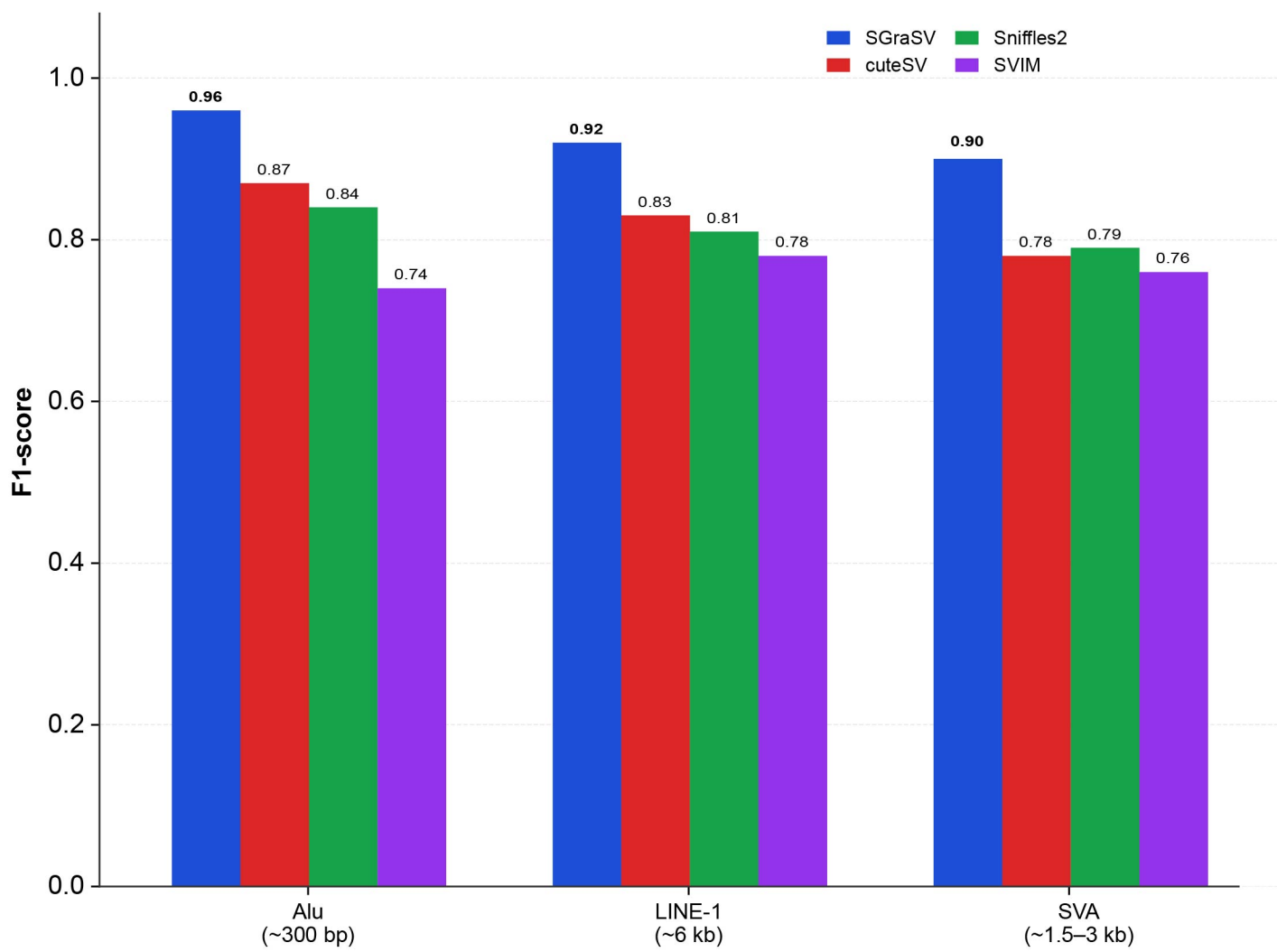
(c) Mendelian trio consistency rate
(Yoruba pedigree trio)



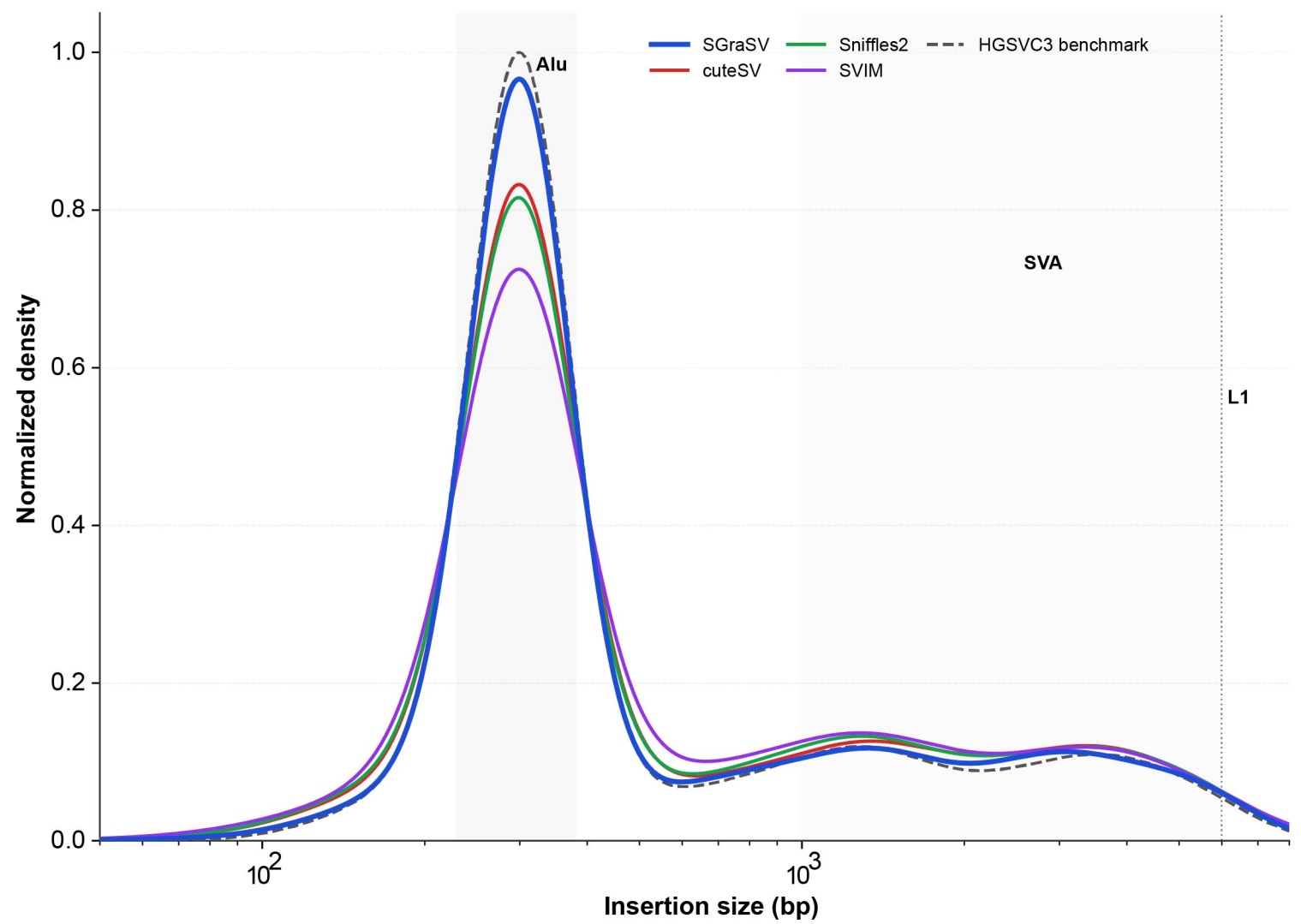
(d) Allele frequency distribution of SGraSV SVs
across 1KGP (3202 samples)



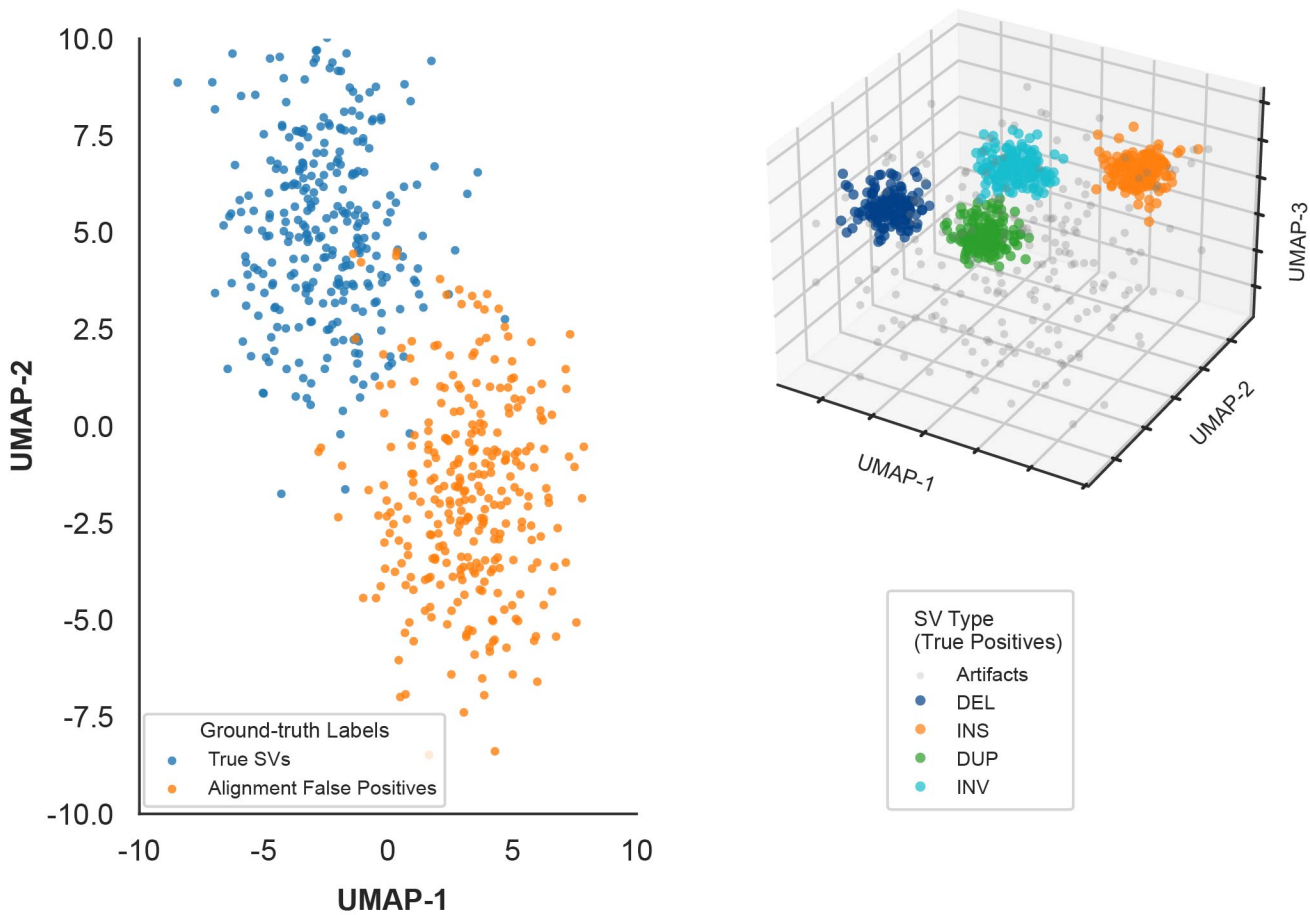
(e) MEI detection performance across distinct classes



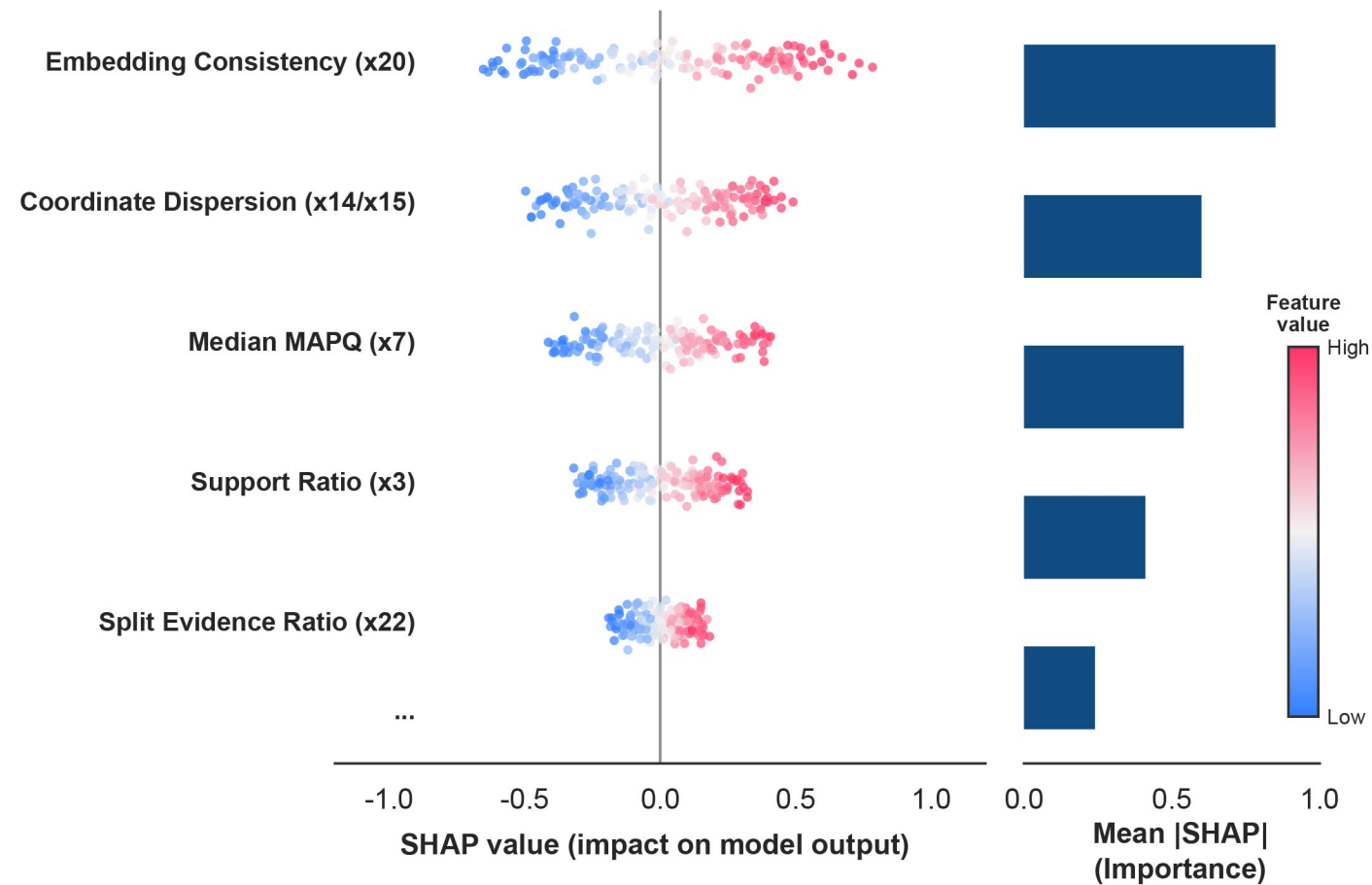
(f) Insertion-size distribution of non-reference calls



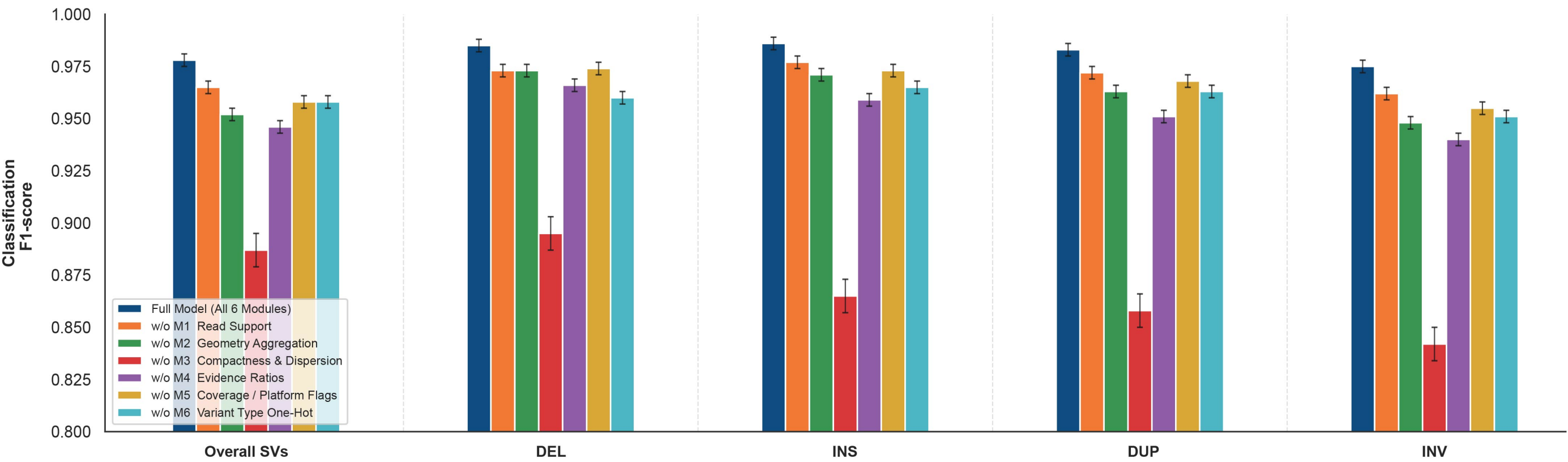
(a) Latent Space Manifold (UMAP Projection)



(b) SHAP Feature Attribution Analysis (42D Candidate Features)



(c) Module-wise Feature Ablation Experiments (Leave-One-Module-Out)



Supplementary Files

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

- [supp.pdf](#)