Genome Biology

METHODOLOGY

Open Access

SVScope improves somatic structural variations detection via graph-genome optimization

Kailing Tu[1†], Qilin Zhang[1†], Yang Li[1], Yucong Li[1], Lanfang Yuan[1], Jing Wang[1], Jie Tang[1], Lin Xia[1], Wei Huang[2*] and Dan Xie[1*]

[†]Kailing Tu and Qilin Zhang contributed equally to this work.

*Correspondence: huangw482@nenu.edu.cn; danxie@scu.edu.cn

[1] Laboratory of Omics Technology and Bioinformatics, Frontiers Science Center for Disease-Related Molecular Network, State Key Laboratory of Biotherapy, West China Hospital, Sichuan University, No. 17, Section 3, Renmin South Road, Chengdu 610041, Sichuan, China
[2] School of Mathematics and Statistics, Key Laboratory for Applied Statistics of the Ministry of Education, Northeast Normal University, 5268 Renmin Street, Changchun 130024, China

Abstract

Somatic structural variations (SVs) are critical in cancer genomes, yet their detection from long-read sequencing remains challenging due to alignment errors in repetitive regions. We develop SVScope, leveraging full-length reads and local graph-genome optimization with a random forest strategy to improve somatic SV calling. We also provide ScopeVIZ, a companion pipeline for visualizing read clustering at breakpoints. Across seven benchmark cell lines sequenced with ONT and PacBio platforms, as well as simulated datasets, SVScope consistently outperforms state-of-the-art methods, achieving up to 23.64% improvement in F1-score. Using SVScope, we validate 32 somatic SVs, expanding the ground-truth dataset by 47.06%.

Keywords: Somatic structural variation, Long-read sequencing (LRS), Local graph genome optimization

Background

Somatic structural variations (somatic SVs) analogous to de novo SVs but occurring in somatic clones can be broadly defined as large genomic alterations exceeding 50 bp in length that occur in specific somatic cell clones. Somatic SVs, which involve larger segments of base variations compared to point mutations, are predominantly derived from genomic rearrangement events [1, 2]. Existing studies have categorized SV events into five basic types based on their impact range: large-scale chromosomal rearrangements or breakend (BND) SV events, including duplications (DUP), inversions (INV), and translocations (TRA); and small-scale local genomic sequence alterations, including insertions (INS) and deletions (DEL) [3]. These SVs tend to have a greater effect on gene expression than point mutations [4, 5]. Recent studies have identified that somatic SV events mediate genomic copy number variations (CNVs) [6], gene fusions [7], and cis-regulatory alterations [8]. In addition to these mechanisms, somatic SV events

also influence the topological structure of the genome by altering CTCF binding sites, thereby participating in long-range gene regulation [9], and modulate the methylation status of the genome [10]. Accurate detection of somatic SVs is of great importance for discovering potential therapeutic targets for tumors, elucidating the regulatory networks of tumor genomes, and understanding the evolutionary patterns of cancer.

The scope of the somatic SV detection task is to identify variants unique to the somatic genome. Typically, the detection process includes reads-to-reference alignment, abnormal alignment breakpoint extraction, and somatic event detection. Historically, short-read sequencing (SRS) techniques have been constrained by the short length of sequence reads, which often fail to span entire SV regions. SRS-based algorithms predominantly relied on inter-read information arising from read populations, including alignment breakpoint [11–13] and sequencing depth [14, 15] to decipher somatic SVs. Long-read sequencing (LRS) data has revolutionized this field by providing full-length sequences that span the breakpoints of even the most complex SVs and often span entire SVs up to several tens of kilobases that were previously undetectable by SRS [3]. Leveraging the extensive sequence information from each individual read, current LRS-based algorithms utilize the intra-read information for somatic SV detection. To enhance detection accuracy, these methods focus on improving the extraction of alignment breakpoint features, with innovations including breakpoint clustering-based somatic genome polishing [16], tandem repeat sequence breakpoints merging [17], and sequence-to-image approaches [18]. Alignment breakpoint information is the key reflection of SV events in the alignment results. For large-scale chromosomal rearrangement SV events, including TRA, INV, and DUP, long-read sequences are decomposed by aligners into several independent alignment records, manifesting as a split-alignment signature [19, 20]. While these events may involve micro-insertions or deletions at the breakpoints, the large-scale split-read signal remains the dominant feature. Breakpoint coordinates information is powerful for yielding accurate results in these large-scale structural variation detection tasks.

In contrast, for small-scale somatic SVs, like local INS and DEL, breakpoint information primarily originates from gaps in the read alignment CIGAR strings, manifesting as an inner-alignment signature. Relying solely on breakpoint position information faces two challenges in the inner-alignment somatic SV detection: incomplete information and ambiguous localization. On the one hand, inner-alignment somatic SVs are fundamentally defined by local sequence alterations, such as the inserted sequence content or the span of sequence loss, where a single coordinate fails to capture all information of the SV event. On the other hand, influenced by local variable number tandem repeat sequences and alignment errors, reads supporting the same SV event may produce multiple breakpoint location features [21]. Beyond these local issues, the misalignment of reads in regions with highly repetitive sequence, including alpha-satellite DNA from centromeres [22–24] and conserved tandem repeat sequences from telomeres [25], also serves as a source of error in somatic SV detection tasks. Furthermore, the current landscape is marked by a lack of large-scale ground-truth somatic SV datasets. Past evaluations of somatic SV callers have predominantly relied on benchmark datasets derived from multi-omics data and the consensus of somatic SV callers. However, limitations in methodological design can lead to the omission of certain somatic SV events, thereby

affecting the accuracy and comprehensiveness of somatic SV assessments. Despite these challenges, long-read sequencing offers unique advantages that lie in its ability to capture the full-length sequence spanning the entire somatic SV event and its flanking regions. Therefore, a systematic approach combining alignment breakpoint and full-length sequence information is essential for precise somatic SV detection.

To address this need, we introduce SVScope, a computational framework that leverages full-length sequence information and local graph genome optimization to accurately detect somatic SVs. The framework utilizes read alignment breakpoint information from the whole-genome scale to cluster and identify split-alignment somatic SVs and candidate inner-alignment somatic SVs. To mitigate the impact of alignment errors on inner-alignment somatic SV detection, SVScope re-analyzes the alignment relationships among all full-length sequences spanning the candidate somatic SV interval using a partial order alignment (POA) graph with multi-sequence alignment (MSA) representation [26, 27] and accurately clusters reads with a sequence mixture model. To avoid read coordinate errors affected by centromeres, telomeres, and segmental duplication sequences, SVScope also implements a random forest machine learning approach based on local alignment features to filter high-confidence somatic SV events. Evaluated on Oxford Nanopore and PacBio HiFi whole-genome sequencing datasets from seven tumor-normal benchmark cell line pairs, as well as simulated datasets with high heterogeneity in tandem repeat regions, SVScope outperformed existing state-of-the-art somatic SV callers. Based on the results of SVScope, we validated 32 somatic SV events with PCR methods, expanding the size of the experimental validation dataset by 47.06%. Furthermore, SVScope provides a single-read resolution visualization tool, ScopeVIZ, to facilitate detailed analysis of somatic SVs. All code implementations are publicly available on GitHub (https://github.com/Goatofmountain/SVScope).

Results

Error analysis of somatic SV detection in cell lines

To elucidate the challenges in somatic structural variant (SV) detection, we conducted an in-depth analysis of existing multiplatform whole-genome sequencing datasets from seven tumor-normal paired cell lines (Additional file 1: Table S1) derived from both ONT and PacBio platforms. Among these, only COLO829 and HG008T have gold-standard somatic SV call sets established through extensive validation [28, 29]. For the remaining five tumor-normal cell line pairs (H1437, H2009, HCC1395, HCC1937, HCC1954), which lack such experimentally validated gold-standard datasets, we utilized the CASTLE benchmark dataset [17]. This dataset was constructed by identifying recurrent somatic SV events across 3 sequencing platforms and 8 somatic SV callers. Our analysis revealed distinct patterns: COLO829, HCC1395, H2009, and HCC1937 predominantly exhibited INS and DEL, which are categorized as inner-alignment somatic SVs, while HCC1954, HG008T and H1437 showed a higher prevalence of DUP, INV, and BND, which are classified as split-alignment somatic SVs (Additional file 2: Fig. S1A).

Four out of five existing somatic SV callers exhibit higher detection accuracy for split-alignment somatic SVs compared to inner-alignment somatic SVs in both ONT and PacBio datasets (Fig. 1A, Additional file 2: Fig. S2A). To ensure a rigorous and fair evaluation of these callers, we employed Minda, a junction-based somatic SV benchmarking

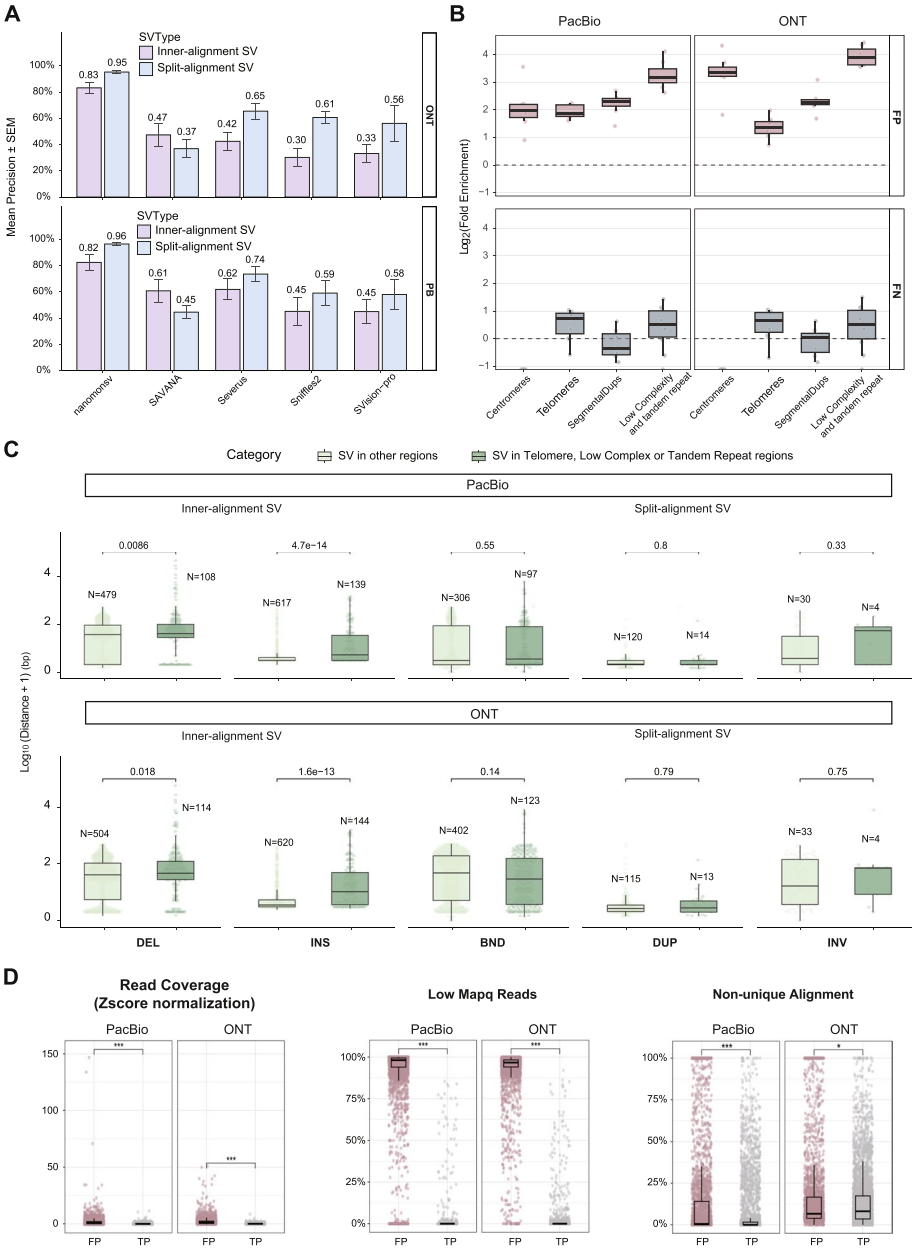

Fig. 1 Error summary of somatic SV calling task. **A** Barplot shows the mean precision of 5 existing somatic SV callers in detecting inner-alignment SVs (light purple) and split-alignment SVs (light blue) tasks for ONT (top panel) and PacBio (bottom panel) datasets. **B** Boxplot shows the log-scaled odds ratio of all false positive (light red) and all false negative (light blue) somatic SVs in four types of genomic regions. These regions include centromeres, telomeres, segmental duplication regions, and low-complexity and tandem repeat sequence regions annotated by RepeatMasker. **C** Boxplot shows log-scaled breakpoint deviation of somatic SV located within telomere, low-complexity or tandem regions (dark green) and other regions (light green). The number of somatic SV events (N) in each category is annotated above the corresponding box. **D** Boxplots show the difference of Zscore normalized local read coverage (left), percentage of low-mapping-quality reads (middle), and rate of non-unique alignment reads (right) in false positive (light red) and true positive (light gray) somatic SV intervals with statistical significance (Mann–Whitney U Test) annotated as * ($p < 0.05$), ** ($p < 0.01$), and *** ($p < 0.001$)

tool [17], rather than the coordinate-centric matching approach of Truvari [30]. While Truvari is a robust standard originally optimized for high-consistency germline SVs, we observed that different matching philosophies can lead to variations in performance metrics when evaluating the heterogeneous outputs of somatic SV callers (Additional file 2: Fig. S1B, Methods). Minda evaluates accuracy based on junction coordinate overlap regardless of the specific SV type assigned by each caller. This junction-centric approach provides a stable assessment of the physical detection capabilities of somatic SV algorithms.

To further investigate the challenges in somatic SV detection, we analyzed the genomic distribution of all false positive (FP) and false negative (FN) calls across seven cell lines on both ONT and PacBio platforms, using five LRS-based somatic SV callers. Our findings indicated that telomeres, low-complexity and tandem repeat sequences are hotspots for errors in both ONT and PacBio data (Fig. 1B, Additional file 2: Fig. S2B). Previous studies have attributed this to the limitations of the minimizer-and-chain alignment algorithm [31], which often fails to precisely anchor reads within repetitive sequences, leading to ambiguous breakpoint positioning [17, 21]. Consequently, our analysis reveals that inner-alignment somatic SVs (DEL and INS) within these repetitive regions exhibit significantly larger deviations between the detected breakpoints and the true SV center positions compared to those in non-repetitive regions (ONT DEL: p-value = 0.018, ONT INS: $p = 1.6e\text{-}13$, PacBio DEL: $p = 0.0086$, PacBio INS: $p = 4.7e\text{-}14$, Mann–Whitney U Test, Fig. 1C). In contrast, large-scale split-alignment SVs, such as BND and INV, exhibit considerable breakpoint dispersion regardless of whether they are located in repetitive or non-repetitive regions ($p > 0.05$, Fig. 1C). This indicates that their high breakpoint deviation is intrinsic to the complexity of large-scale rearrangements rather than being specifically driven by local sequence homology. Since existing callers maintain high detection accuracy for these types (Fig. 1A), it suggests that current algorithms are robust to this intrinsic, global deviation. However, the context-specific breakpoint instability induced by repetitive sequences remains a critical and unresolved failure mode specifically for inner-alignment SVs.

Beyond telomeres, low-complexity and tandem repeat sequences, we also observed an enrichment of false positive somatic SV events in centromeres and segmental duplications (SD) (Fig. 1B). These regions, characterized by their high sequence homology [32], pose significant challenges for accurate SV detection. For instance, centromeres are primarily composed of alpha-satellite DNA, which consists of arrays of ~ 171 base-pair units arranged into higher-order arrays throughout the centromere of each chromosome [22–24]. The repetitive nature of these sequences leads to alignment ambiguities. Reads that fail to span these regions often align non-uniquely to multiple genomic intervals with similar sequences, resulting in artifacts such as high local coverage depth and poor mapping quality (Additional file 2: Fig. S3A-B). Consequently, compared to true positive somatic SVs, false positive SVs in these regions exhibited higher local read counts (both ONT and PacBio $p < 0.001$, Mann–Whitney U Test Fig. 1D, Additional file 2: Fig. S3C), a higher proportion of low mapQ reads (both ONT and PacBio $p < 0.001$, Mann–Whitney U Test, Fig. 1D, Additional file 2: Fig. S3D), and a higher proportion of non-uniquely aligned reads (PacBio $p < 0.001$, ONT $p < 0.05$, Mann–Whitney U Test, Fig. 1D, Additional file 2: Fig. S3E). These characteristics reflect the alignment inaccuracies caused

by long-read sequences that cannot effectively span these regions, thereby significantly impacting the accuracy of somatic SV detection. In summary, our analysis highlights the complexities and challenges in somatic SV detection, particularly in regions with low complexity, tandem repeats, centromeres, telomeres and segmental duplication regions. The limitations of current SV callers in accurately detecting somatic SVs in these regions underscore the need for a more robust and accurate detection approach.

Design of SVScope

To address the challenge of accurate somatic SV detection, refinement, and visualization, we introduce SVScope, which consists of four essential modules. First, the candidate somatic SV selection module is designed to analyze read alignment results and identify candidate somatic SV intervals. It categorizes potential somatic SVs into split-alignment SVs and inner-alignment SVs based on breakpoint characteristics (Fig. 2A). Split-alignment SVs are characterized by reads with different segments aligning to multiple positions in the reference genome, and are identified through a paired-breakpoint

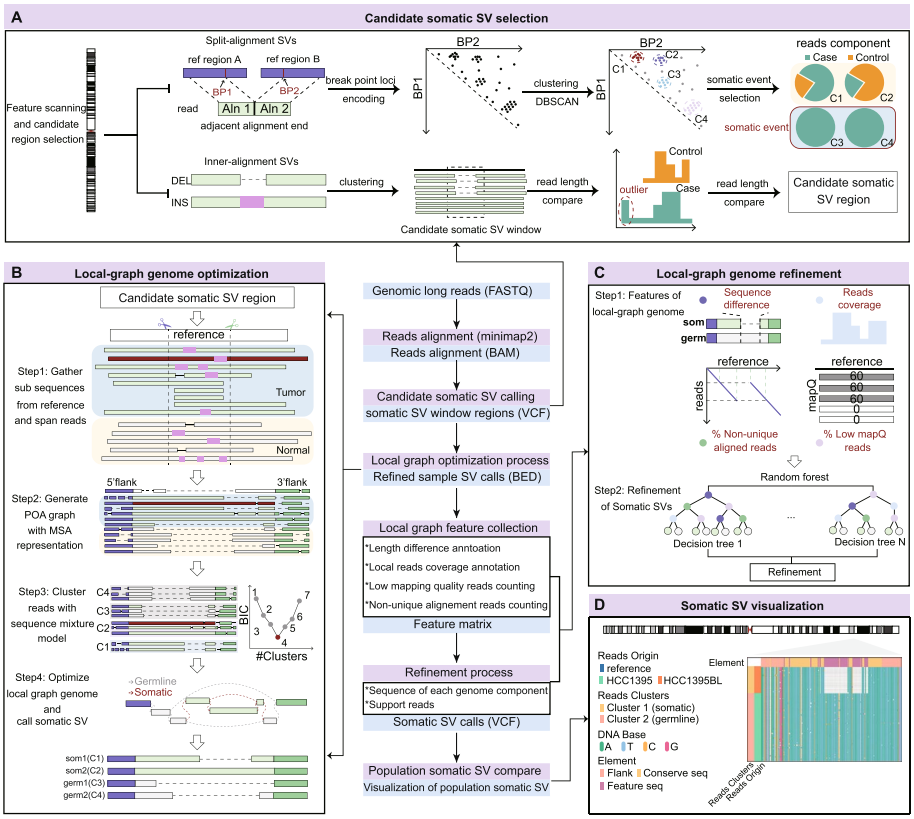

Fig. 2 Schematic of the SVScope Workflow. This schematic illustrates the workflow of the SVScope algorithm for detecting somatic structural variations (SVs) from FASTQ data and comparing somatic SV sequences across population samples. The detection process consists of four modules: **A** Candidate somatic SV selection module selects candidate somatic SV intervals through paired-breakpoint clustering (split-alignment SVs) and full-length sequence length comparison (inner-alignment SVs). **B** Local-graph genome optimization utilizes the full-length sequence information of span-reads to optimize local graph-genome and find case-reads-specific paths as somatic SVs. **C** Local-graph genome refinement module evaluates the confidence of the local-graph genome with a random forest model based on local alignment features. **D** Somatic SV visualization module, visualizes read contributions in SV detection

coordinate, re-encoding, clustering and component analysis process (Methods, Additional file 2: Fig. S4A). Inner-alignment SVs, featuring internal breakpoints with sequence insertions or deletions, are processed by clustering adjacent breakpoints to form candidate regions. The module then compares the lengths of case and control reads within these regions using full-length information from long reads to identify candidate somatic SV intervals (Methods, Additional file 2: Fig. S4B). Second, the local-graph genome optimization module targets all inner-alignment somatic SV intervals. It constructs a local graph genome based on the complete sequences of span reads from case and control samples using the POA algorithm [26, 27], and optimizes it with a sequence mixture model (Fig. 2B, Methods). Candidate somatic SVs are selected from case-read-specific paths in the local graph genome. Third, to mitigate read misalignment issues caused by highly repetitive regions, the local-graph genome refinement module employs a random forest model in conjunction with read alignment features to assess the accuracy of all previously obtained local-graph genomes, thereby filtering out low-confidence somatic SVs (Methods, Additional file 2: Fig. S5A, B). Finally, to facilitate a clear comparison of read contributions in the SV detection process, SVScope includes a somatic SV visualization module that enables users to inspect the clustering of all span reads (Fig. 2D).

Accurate detection of somatic SVs with SVScope

To comprehensively evaluate the performance of SVScope, we conducted a systematic benchmark using 15 whole-genome sequencing datasets derived from seven tumor-normal cell line pairs across both ONT and PacBio HiFi platforms (Additional file 1: Table S1). The compiled datasets capture a wide range of sequencing characteristics representative of current long-read studies. For ONT data, samples were sequenced using both R9.4.1 and R10.4.1 flow cells, with sequencing coverage ranging from 14 to 86X and read length N50 ranging from 11.8 kb to 47.5 kb. The PacBio HiFi datasets provided coverage between 29 and 79X, with N50 read lengths consistently ranging from 10 to 21 kb (detailed metrics in Additional file 1: Table S1). For ground truth validation, we utilized experimentally validated gold-standard somatic SV call sets for the COLO829 and HG008 cell lines. For the remaining five cell line pairs (H1437, H2009, HCC1395, HCC1937, and HCC1954), which lack external gold-standard datasets, we constructed a robust consensus benchmark. Following the methodology of the Severus study [17], we performed a voting consensus using results from eight somatic SV callers across three platforms, with the crucial addition of SVScope's own calls to the voting pool to ensure comprehensiveness (Methods).

Across this diverse multi-platform dataset, SVScope demonstrated superior detection accuracy. It achieved the highest average F1 scores on both ONT and PacBio platforms compared to all other somatic SV callers (Fig. 3A, Additional file 2: Fig. S6A, B) and maintained exceptional performance in detecting both inner-alignment SVs and split-alignment SVs (Fig. 3B, Additional file 2: Fig. S6C, D). Notably, SVScope was the top-performing algorithm in 11 out of the 15 datasets (comprising 7 ONT and 4 PacBio datasets, Additional file 2: Fig. S6E). In these cases, SVScope significantly outperformed the second-best methods, achieving a maximum F1 score improvement of

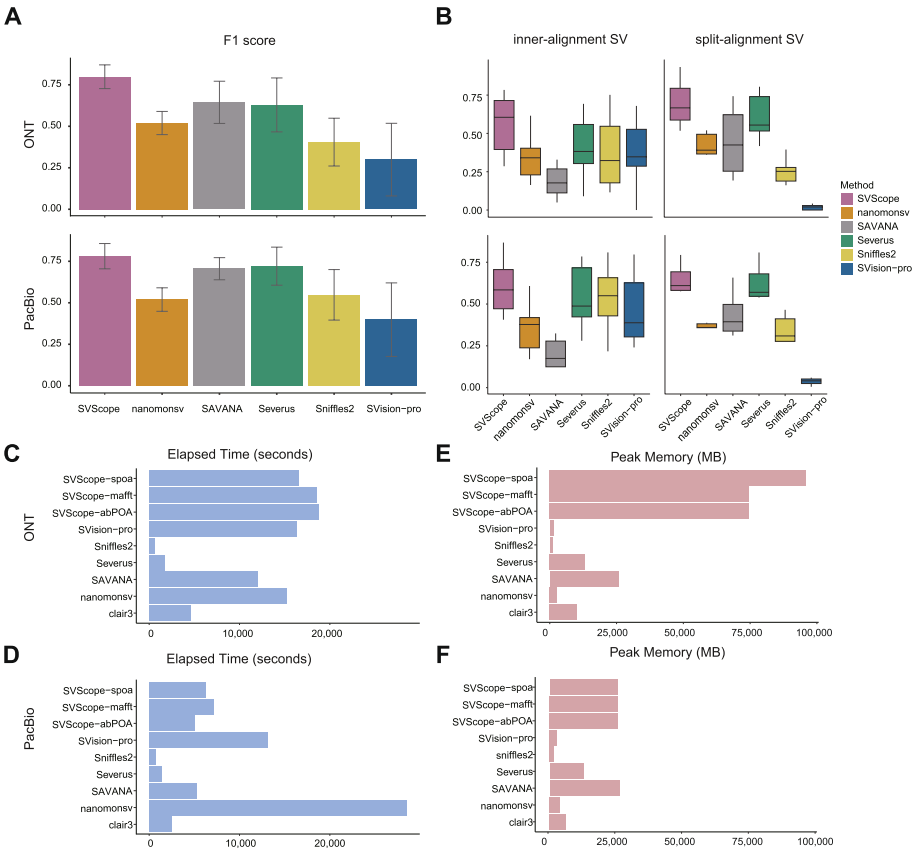

Fig. 3 Performance of somatic SV callers across benchmark cell lines on ONT and PacBio platforms. **A** Bar plot showing the mean F1 scores for detecting all types of somatic SVs aggregated across seven tumor-normal benchmark cell line pairs. The evaluation includes datasets derived from both Oxford Nanopore Technologies (ONT, encompassing R9.4.1 and R10.4.1 flow cells) and PacBio HiFi platforms. Error bars represent the standard error of the mean (SEM). **B** Box plots summarizing the distribution of F1 scores stratified by SV classification: inner-alignment somatic SVs (e.g., insertions, deletions) and split-alignment somatic SVs (e.g., inversions, translocations, breakends). Detailed sequencing metrics for all datasets are provided in Additional file 1: Table S1. **C**-**F** Computational performance benchmarks conducted using the GIAB HG008 dataset. All tools were executed on the same unloaded high-performance computing node configured with 64 CPU threads. Total wall-clock runtime (in seconds) on ONT data (**C**) and PacBio HiFi data (**D**). Peak memory usage (in MB) on ONT data (**E**) and PacBio HiFi data (**F**). Note: For Severus, the reported runtime includes the mandatory pre-processing step of SNP calling performed by Clair3

23.64% (H1437 cell line, ONT platform). In the remaining four datasets (COLO829 ONT and PacBio, HCC1937 PacBio and HG008 PacBio) where SVScope did not rank first, it consistently ranked second, with performance metrics very close to the top caller (maximum F1 score gap of only 4.9% in the HG008 PacBio dataset).

Computational performance of SVScope and other somatic SV callers

Beyond detection accuracy, we comprehensively evaluated the computational efficiency of SVScope and other state-of-the-art somatic SV callers using the Genome in a Bottle (GIAB) HG008 dataset on both ONT (HG008T 45X, HG008-N-P 41X) and PacBio HiFi (HG008T 79X, HG008-N-P 35X) platforms (Additional file 1: Table S1). To ensure a fair comparison, all benchmarks were conducted on the same unloaded

high-performance computing node, with all tools and their requisite pre-processing steps uniformly configured to utilize 64 CPU threads (Methods).

The benchmark results reveal distinct performance profiles driven by data characteristics. Notably, Severus demonstrated the highest computational efficiency among specialized somatic SV callers. Even when accounting for the runtime of its mandatory SNP calling step via Clair3 [33], the Severus pipeline remained the fastest solution on both platforms. In comparison, SVScope's performance varied by platform. On the PacBio HiFi dataset, characterized by high sequencing depth and low read error rates, SVScope demonstrated high efficiency regardless of the MSA kernel used. Its computational speed and peak memory usage were comparable to SAVANA [34] (Fig. 3C, D) and significantly faster than nanomonsv [16] and SVision-pro [18] (Fig. 3C).

In contrast, when processing ONT data, which typically features lower sequencing depth but higher read error rates, SVScope exhibited increased resource demands. While its runtime remains within the same order of magnitude as SAVANA, SVision-pro, and nanomonsv, SVScope generally required longer processing times and recorded the highest peak memory usage among the tested tools (Fig. 3E-F). We attribute this to the noisy nature of ONT reads, which generate a significantly higher frequency of ambiguous alignment breakpoints. This data characteristic forces SVScope to perform resource-intensive Multiple Sequence Alignment (MSA) and unsupervised local sequence clustering on a much larger number of candidate inner-alignment intervals compared to HiFi data.

Under these high-load sequencing conditions, the architectural advantage of the adaptive banded POA (abPOA) [27] became evident. Utilizing abPOA mitigates the peak memory usage and consistently yielded the most robust detection F1-score on the HG008 dataset compared to the other two MSA algorithms, SIMD-sPOA [26] and MAFFT [35] (Additional file 2: Fig. S6F). To further investigate the specific contributions of the MSA step to the overall pipeline efficiency, we conducted an isolated benchmarking of four MSA algorithms, abPOA, SIMD-sPOA, MAFFT, and POASTA [36], using 286 candidate SV intervals extracted from the HG008 ONT dataset. As shown in Additional file 2: Fig. S7A, abPOA exihibited a significant advantage in CPU runtime, with a median of 0.1 s compared to the 2.0 s-2.5 s required by other algorithms. Regarding memory efficiency, POASTA demostrated the lowest peak memory consumption (median 17 MB), where as SIMD-sPOA required substantially more memory (median 137 MB) to process the same intervals (Additional file 2: Fig. S7B).

Subsequently, we re-evaluated the impact of these MSA algorithms on the final SV detection performance within the SVScope pipeline using the HG008 ONT dataset. In addition to the default settings, we implemented a high-accuracy MAFFT strategy (–maxiterate 1000 –globalpair, see Methods) to assess whether increased alignment precision would yield better calling results. Our evaluation (detailed in Additional file 1: Table S2) revealed that the high-accuracy MAFFT strategy produced results nearly identical to those of the default MAFFT setting, suggesting that SVScope's overall performance is robust and not heavily dependent on extreme MSA precision. While POASTA offered memory advantages, it resulted in the lowest detection accuracy among all tested methods (Additional file 1: Table S2). Consequently, abPOA

ensures that SVScope maintains a feasible balance between computational cost and superior detection accuracy.

Robust detection and read clustering on simulated somatic SVs in tandem repeat regions

To further assess SVScope's detection capability for somatic SVs in tandem repeat regions, we constructed a simulated ONT dataset anchored to 3238 tandem repeat regions on chromosome 22 annotated by RepeatMasker. This diverse dataset comprises 2,981 Short Tandem Repeats (STRs), 235 Variable Number Tandem Repeats (VNTRs), and 22 Satellites. To mimic the "diploid haplotype interference" encountered in real biological samples, we modeled the diploid background of paired normal samples to include germline polymorphisms. Specifically, in 50% of the simulated loci, the two normal alleles were designed to be heterozygous, with one allele matching the reference and the other exhibiting a germline expansion (Additional file 2: Fig. S8A). For the somatic tumor genome, we introduced a further expansion of the repeat sequence (X15, X20, or X30 units) on top of the target allele (Additional file 2: Fig. S8B, Methods). Finally, by downsampling and mixing reads from these simulated genomes, we generated datasets with varying tumor purities ranging from 20 to 100% (Methods). Under varying tumor purity and SV length conditions, SVScope consistently outperformed other somatic SV callers (Fig. 4A, Additional file 2: Fig. S9A-C) and maintained high-precision results across different tumor purity levels (Fig. 4B). Notably, this precision advantage persists even when facing the added complexity of diploid haplotype interference, a critical challenge in distinguishing somatic variants from germline structural polymorphisms in repetitive regions. This demonstrates SVScope's robustness against the influence of multiple-allele haplotypes compared to other alignment-based somatic SV callers (Additional file 2: Fig. S9D). Beyond somatic SV event identification, SVScope also exhibited superior performance in the calculation of somatic SV support reads (Fig. 4C) and provided robust estimations across different tumor purities (Additional file 2: Fig. S10).

The employment of full-length sequence based local-graph optimization strategy is critical in somatic SV support read detection tasks. For instance, a long tandem repeat somatic insertion interval on chr22:50,137,660–50,137,991 illustrates how reads harboring somatic SVs are prone to alignment breakpoint errors due to the variable choice of minimizers in the nearby flanking sequences, leading to different alignment breakpoints for reads originating from the same somatic genome (Additional file 2: Fig. S11A). In addition to the high-dispersion INS breakpoints, such long tandem repeat insertions can even introduce duplication-like breakpoints primarily manifested as supplementary alignments (Additional file 2: Fig. S11B). By leveraging full-length span-read sequences and POA graph-based multiple sequence alignment strategies, SVScope is unaffected by such read-reference alignment errors, ensuring proper read alignment (Additional file 2: Fig. S11C). Based on this, the local graph genome optimization process employing the sequence mixture model can accurately estimate the genomic information of different cellular components, thereby enabling precise parsing of heterogeneous genomes (Additional file 2: Fig. S11D).

To further test SVScope's ability to distinguish read origins in heterogeneous genomes, we mixed two simulated somatic genomes with different tandem repeat lengths (S2: X15, S3: X30) and a control genome (S1). We then used SVScope to perform read source

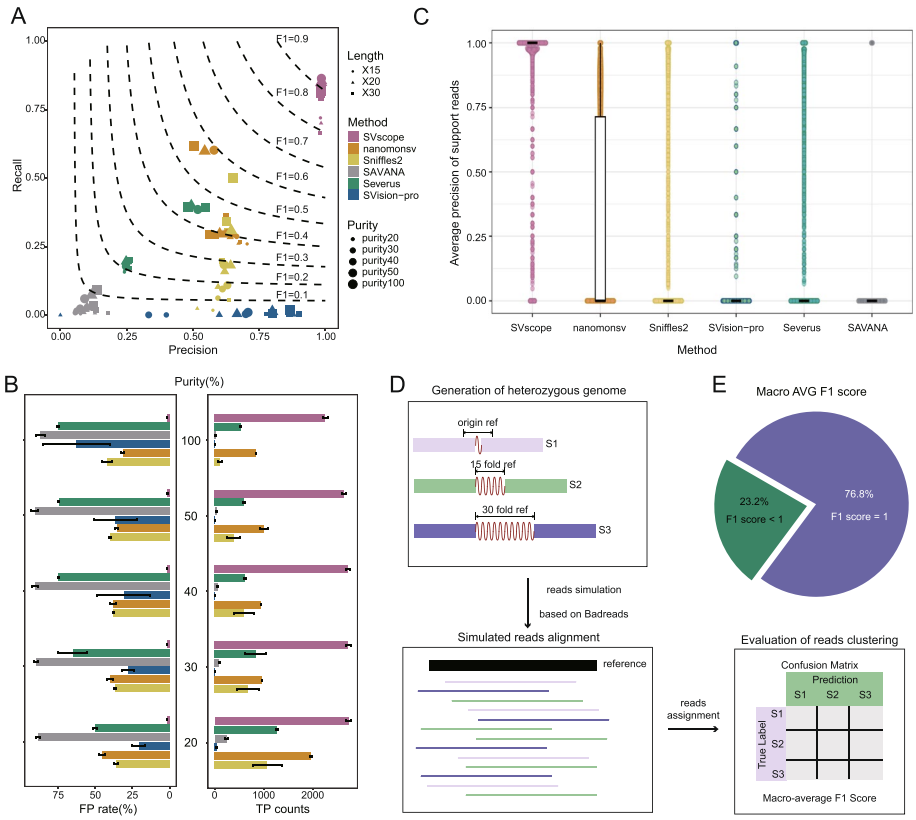

Fig. 4 Performance of somatic SV callers in simulated somatic tandem repeat datasets. **A** Scatter plot with contour lines illustrating the precision and recall of SVScope and other state-of-the-art somatic SV callers across all simulated datasets with 5 tumor purity thresholds and 3 tandem repeat extension lengths. The parameters of datasets and somatic SV callers are denoted on the right. Dashed lines represent various gradients of F1 scores. **B** Bar plot shows the count of true positive somatic SVs and the false positive rate of each somatic SV caller under different tumor purity. **C** Scatter plot shows the precision of somatic SV supporting reads in each somatic SV caller under different tumor purity. Tumor purity and SVcaller name in Fig. 4B-C are color-coded as in Fig. 4A. **D** Schematic of simulated data for the heterogeneous genome. **E** Pie chart shows the percentage of simulated heterogeneous regions with perfect read source assignment performance using SVScope (F1 score = 1)

assignment on these mixtures and evaluated the accuracy of the resulting clusters (Fig. 4D). In this task, SVScope achieved completely accurate read clustering in 2359 out of 3225 (76.8%) simulated mixed heterogeneous loci (Fig. 4E). Taking the heterogeneous simulated region chr22:10,538,368–10,538,453 as an example (Additional file 2: Fig. S12A), SVScope effectively mitigates the impact of local breakpoint errors, accurately clustering reads by leveraging full-length sequence information (Additional file 2: Fig. S12B) and local graph genome optimization process (Additional file 2: Fig. S12C, D). This performance underscores SVScope's robustness in accurately clustering reads from different genomic origins, highlighting its potential for resolving multi-allelic heterogeneity in regions where traditional alignment or SNP-based phasing is challenging.

Application of SVScope in somatic SV detection and validation

Given SVScope's superior performance in somatic SV detection tasks across both cell lines and simulated datasets, we further optimized the gold-standard somatic SV dataset

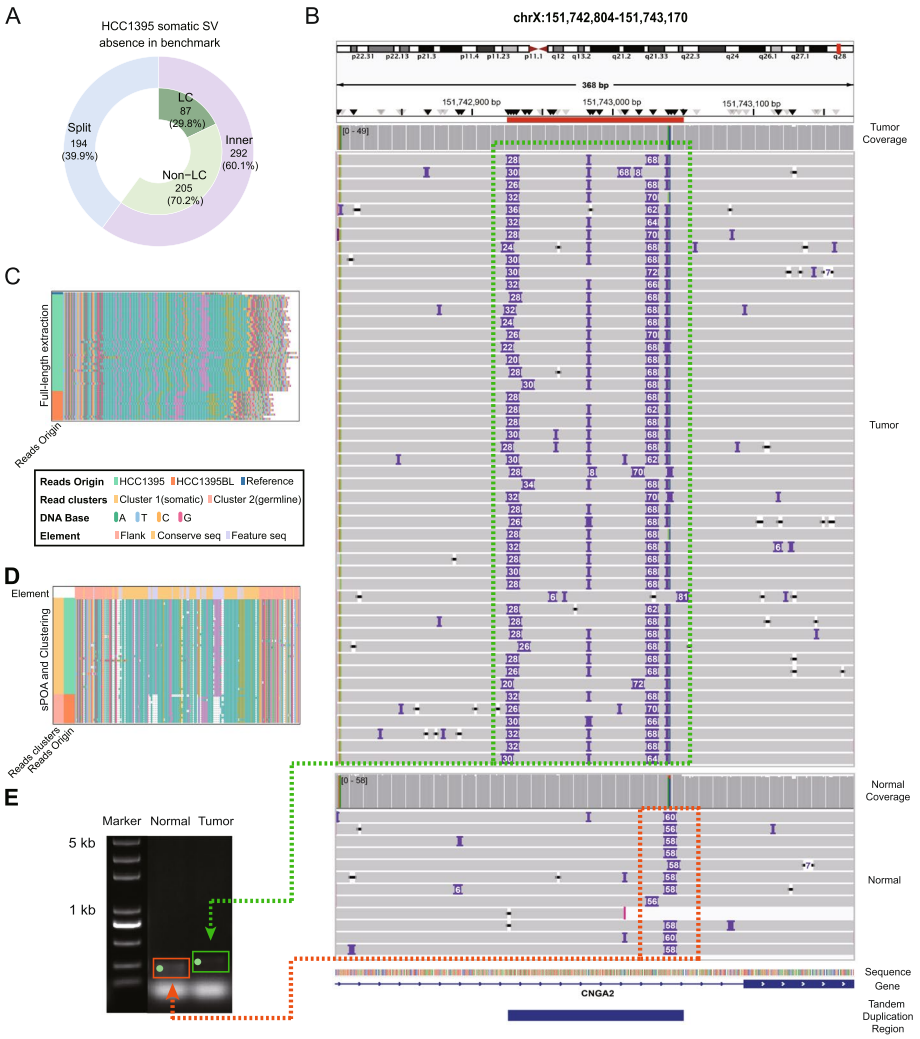

Fig. 5 Detailed characterization and validation of somatic SV events. **A** The outer pie chart shows the percentage of inner-alignment (purple) and split-alignment (light blue) somatic SVs identified by SVScope from HCC1395 cell line that are absent from the benchmark SV sets. The inner pie chart shows the percentage of low-complexity or tandem repeat sequence associated (light green) and non-associated (dark green) somatic SV events. **B** Genome browser plot shows the example somatic SV region chrX:151,742,804–151,743,170 (within the *CNGA2* gene) in HCC1395 (top) and HCC1395BL (bottom). **C** Bar plot shows the raw sequence extracted from HCC1395 (colored in green) and HCC1395BL reads (colored in orange). **D** Bar plot shows the result of read clustering based on SVScope. **E** The gel electrophoresis image illustrates the targeted PCR result for the target somatic SV event in the genomes of HCC1395BL (left panel, Normal) and HCC1395 (right panel, Tumor). The target bands are highlighted with green dots

for the HCC1395 cell line. Based on SVScope, we identified 486 somatic SV events not present in the existing benchmark datasets from the HCC1395 cell line, comprising 292 inner-alignment SVs and 194 split-alignment SVs (Fig. 5A). Among these, 228 out of 486 somatic SV events were uniquely detected by SVScope (Additional file 2: Fig. S13). These SVs exhibit complex breakpoint distributions. For example, a somatic tandem repeat insertion event within the intron of the *CNGA2* gene located at chrX:151,742,802–151,743,170 manifests multiple alignment breakpoints in the read alignments of HCC1395, with longer insertion fragments ranging from 62 to 81 bp and shorter fragments ranging from 8 to 34 bp (Fig. 5B). These short fragment breakpoints are often

overlooked by other somatic SV callers. In contrast, HCC1395BL shows a single alignment breakpoint with insertion fragment lengths ranging from 56 to 60 bp (Fig. 5B). Visualization results from ScopeVIZ indicate that reads originating from HCC1395 are longer than those from HCC1395BL (Fig. 5C). Through the local graph genome optimization process, we reconstructed the genomic sequences of HCC1395 and HCC1395BL in this region (Fig. 5D). PCR experimental gel electrophoresis also reflects bands consistent with SVScope's computational results (Fig. 5E). Similarly, in the case of a somatic deletion SV located at chr3:79,909,993–79,910,160, SVScope demonstrates robust detection and read clustering capabilities (Additional file 2: Fig. S14). The breakpoints of these SVs exhibit highly dispersed distributions in the read alignments relative to the reference genome. These features interfere with other alignment-based algorithms, causing them to fail to accurately identify such SVs. In total, based on SVScope's computational results, we validated 32 somatic SV events through a rigorous combination of PCR and haplotype-phased read visualization (Methods, Additional file 1: Table S3). This comprehensive analysis explicitly confirmed the somatic origin of these variants and resolved complex allelic contexts, including loss of heterozygosity (LOH) events, which were corroborated by both gel electrophoresis patterns and phased sequencing data. Among these validated events, 22 were uniquely detected by SVScope. These validated events significantly enhance the existing somatic SV dataset for the HCC1395 cell line, providing a robust foundation for future research.

Discussion

SVScope introduces a novel computational approach that combines local graph genome optimization and random forest machine learning strategies, representing a significant advancement in the field of somatic structural variant (SV) detection. Specifically, SVScope leverages full-length sequence information from spanning reads and integrates it with a local graph genome optimization process. This approach effectively mitigates alignment errors introduced by read aligners, thereby ensuring the accuracy of the algorithm. Additionally, the employment of a machine learning strategy based on local alignment features reduces false positives caused by alignment errors in highly repetitive regions, such as centromeres and telomeres. These innovations collectively enhance the robustness and precision of somatic SV detection and read source assignment tasks. SVScope provides a more accurate and reliable tool for analyzing complex and heterogeneous genomes, paving the way for future advancements in genomic research and clinical applications.

The development of SVScope highlights the potential of leveraging full-length information to improve the accuracy and reliability of somatic genome analysis. This approach not only enhances the detection of somatic SVs but also provides valuable insights into the heterogeneity of complex genomes. Beyond the nucleotide sequence, the integration of extensive epigenetic data including methylation [37, 38] and chromatin accessibility [39–42] derived from single-molecule sequencing offers a more comprehensive view of the genome, which is crucial for understanding the underlying mechanisms of genetic disease and cancer. Future research should focus on optimizing algorithms to further explore the potential functions of somatic SVs, leveraging the multi-dimensional information provided by long-read single-molecule sequencing technologies.

The local graph genome optimization step is the core component of the SVScope algorithm, relying on spanning reads to provide full-length sequence for the target genomic intervals. However, due to the unavailability of span reads limited by the read length of current long-read sequencing technology, somatic SVs with ultra-long insertions cannot always be precisely analyzed through SVScope. Advances in ultra-long read sequencing technologies [43] are expected to mitigate these challenges in the future. Recent developments in DNA large language models, such as Evo [44], Nucleotide Transformer [45], and DNABERT [46], have demonstrated the power of adapting BERT-like self-supervised masked prediction strategies [47] to learn genomic syntax. These models have achieved remarkable performance in downstream classification tasks, including enhancer prediction [45, 46], CRISPR sgRNA generation [44], and species classification [45]. However, their application to interpreting the complex functional impact of somatic structural variations in cancer remains limited. A primary bottleneck is the scarcity of high-precision sequence labels required to fine-tune these models on heterogeneous tumor data. Future cancer genomics will increasingly rely on single-molecule long-read sequencing to simultaneously capture sequence and multi-dimensional epigenetic information [40, 41]. In this context, SVScope's capability to perform high-precision read clustering and source assignment becomes critical. By accurately deconvoluting somatic variants from germline backgrounds and linking them to their specific epigenetic states, SVScope can facilitate the automated generation of large-scale, high-quality labeled datasets. These structured datasets will serve as essential ground truth for training the next generation of LLMs to predict the regulatory and functional consequences of somatic rearrangements.

Conclusions

In summary, SVScope has demonstrated superior performance in somatic SV detection, SV support reads calculation, and read source assignment tasks. It has identified new somatic SV events in paired tumor-normal cell lines, which were validated through PCR experiments and phased read analysis, thereby contributing to the refinement and expansion of the gold-standard somatic SV dataset for the HCC1395-HCC1395BL benchmark cell line pair. Furthermore, the development of the ScopeVIZ visualization system allows users to visually inspect the contribution of sequencing reads to somatic SV detection, facilitating detailed confirmation of each event. SVScope thus provides a new paradigm for leveraging single-molecule long-read information, offering enhanced accuracy and interpretability in the analysis of somatic SV events.

Methods

Establishment of somatic SV benchmark dataset for cell lines via multiomics detection voting strategy

To create a benchmark dataset of somatic SVs for tumor-normal B lymphocyte pairs HCC1395-HCC1395BL, HCC1937-HCC1937BL, HCC1954-HCC1954BL, H1437-BL1437, and H2009-BL2009, we gathered high-depth whole-genome sequencing datasets obtained from Illumina short-read sequencing platforms and long-read sequencing platforms, including PacBio and Oxford Nanopore from published databases. For short-read sequencing data, we employed three SV callers—SvABA, GRIPSS,

and Manta—to detect somatic SVs. For long-read sequencing data, six SV callers—SVScope, Sniffles2, SAVANA, Severus, SVision-pro, and nanomonsv—were used for somatic SV detection. The detection results were integrated using the Minda Ensemble module [17], and somatic SVs supported by at least five SV callers and data from at least two sequencing platforms were selected. For the COLO829 and HG008 tumor-normal pair, we directly adopted the somatic SV list provided in previous studies [28, 29], which was validated through extensive experimental evidence, as the gold-standard datasets. Large-scale somatic SV events including INV, DUP, and BND are labeled as split-alignment somatic SVs. Small-scaled somatic SV events including INS and DEL are labeled as inner-alignment somatic SVs.

Evaluating cell line structural variants detection.

To benchmark the performance of SV detection tools, we downloaded SV calls, ONT and PacBio HiFi reads of seven tumor and matched normal cell lines: HCC1954-HCC1954BL, H2009-BL2009, HCC1937-HCC1937BL, H1437-BL1437, HCC1395-HCC1395BL, COLO829-COLO829BL, and HG008T-HG008-N-P. All sequencing reads were aligned to the UCSC human reference genome (hg38). For Oxford Nanopore Technologies data, raw reads were aligned using minimap2 (v2.22-r1101) with the parameters -ax map-ont –MD. For PacBio HiFi data, datasets originally obtained in BAM format were first converted to FASTQ format using samtools (v1.21). Subsequently, these converted reads, along with other raw PacBio FASTQ datasets, were aligned using minimap2 (v2.22-r1101) with the parameters -ax map-pb –MD. To enable haplotype-aware analysis, we performed Single Nucleotide Polymorphism (SNP) calling on all normal cell line datasets using Clair3 (v1.0.4). Then LongPhase was utilized to generate phased haplotypes. The resulting phased variants served as input for the Severus somatic SV caller and facilitated the haplotype-grouped visualization of somatic SV candidates in IGV. The somatic calling tools nanomonsv(v0.7.0), SAVANA(v1.2.1), Svision-pro(v1.8) and Severus(v1.6) all follow the recommended workflows and default parameters. Although Sniffles2(v2.0.7) was not designed for a paired tumor-normal analysis, we utilized the recommended sniffles2-merge process and its mosaic mode to subtract the normal sample SV calls from the tumor sample calls. For all tools, the minimum supporting read count was set to 3, and the minimum SV length was established at 50 bp. For the benchmarking of somatic SVs, we employed Truvari (v5.0) [30] with the –typeignore argument enabled to accommodate heterogeneous SV type assignments. For evaluation using Minda (v0.0.1) [17], we utilized the following parameters: –min_size 50, –tolerance 500, and –multimatch. All other parameters for both tools were maintained at their default settings to ensure a standard and reproducible evaluation.

Generation of simulated data

To benchmark the performance of SV detection tools, we simulated ONT sequencing data with a depth of 40 × based on the human reference genome hg38 on chromosome 22. We designed five allele frequency levels: 0.20, 0.30, 0.40, 0.50 and 1. Additionally, we included a control group without insertion variants and three new sequence insertion variants of different lengths: tandem repeat (TR) sequence X15, X20 and X30. The aim was to evaluate the accuracy, recall, and F1 scores of TR detection by Somatic SV

callers under varying tumor purities, different new sequence insertion lengths, and the presence or absence of germline mutations. From the human repeat segment annotation database RepeatMasker, we retained TRs on chromosome 22 that are 1000 bp or shorter and ensured that there are no other TRs within 200 bp of each TR. This filtering resulted in a total of 6,478 candidate TR windows, with TR types including STR, VNTR and Satellite. These were divided into four files: Group A(Reference): Homozygous reference sequences without variation; Group B (Germline Heterozygosity): Diploid genomes where one allele matches the reference and the other harbors a germline expansion, creating a heterozygous background; Group C (Somatic Only): Homozygous reference background with a somatic expansion introduced; and Group D (Somatic + Germline Interference): The most complex scenario, where a somatic expansion is introduced into the heterozygous germline background. The positions and lengths of TRs in files C and D were recorded as a gold-standard set for subsequent evaluations. First, using the simulation sequencing software VISOR (v1.1.2), we took chromosome 22 of the hg38 genome as the reference genome and used the recorded mutation files as input to generate corresponding variant genomes for different mutation types. Next, we employed the simulation sequencing software badread (v0.4.0) [48] (https://github.com/rrwick/Badread) to simulate sequencing at a depth of 40 × for each mutation type, using the variant genomes generated by VISOR [49] as the reference. This process produced a dataset of ONT simulated sequencing fragments with three lengths of new sequence insertions across five different tumor purity levels. The parameters for the badread simulation software were as follows: -quantity 40x –glitches 1000,20,20 -junk_reads 0.1 -random_reads 0.1 –chimeras 0.1 –identity 80,90,6 –length 20000,5000. The error model used was nanopore2023.

Evaluating simulated structural variants detection.

The ONT reads are aligned to the hg38 reference genome using minimap2 (v2.22-r1101) with default settings [50]. The somatic calling tools nanomonsv, SAVANA, Svision-pro, Sniffles2 and Severus all follow the recommended workflows and default parameters. The minimum supporting read count was set to 3 for all tools, and the minimum SV length was set to 50 bp. To examine the correctness of detected somatic SVs, we employed region matching, which requires accurate detection of SV sites. For example, assume a particular benchmark somatic SV [b.chrom, b.start, b.end], and a prediction [p.chrom, p.start, p.end]; then a correct region-match should at least overlap by 1 bp. Any called somatic SVs without a matched prediction were counted as false negatives. Based on the numbers of true positives (TP) and false negatives (FN), we computed the recall, precision and F1-score with the following equations, respectively.

$$Precision = \frac{N_{Correct}}{N_{Method}}$$

$$Recall = \frac{N_{Correct}}{N_{Golden\_Somatic}}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

Each caller was run with different AF and insertion length of somatic SV, and the performance of detecting simulated SVs was assessed correspondingly.

Validation of somatic SVs

DNA was extracted from HCC1395 and HCC1395BL using the QIAamp DNA Mini Kit (QIAGEN) according to the manufacturer's instructions. The forward and reverse primers were designed according to the 2000 bp upstream and downstream sequences of the target somatic SV genomic intervals and synthesized by Sangon Biotech (Additional file 1: Table S3 Primers). Extracted DNA was subjected to PCR using matched primers. Each 50-μL reaction consisted of 2 × Premix Taq (Takara) 25 μL, 2 μL 10 μM forward primer, 2 μL 10 μM reverse primer, 50–150 ng DNA, and add ddH2O to 50 μL. PCR was performed as follows: an initial step of 5 min at 95 °C, 30 cycles of 15 s at 95 °C, 15 s at 56 °C and 60–300 s at 72 °C (size of PCR products between 200 bp and 6 kb), with a final extension at 72 °C for 5 min. To address non-specific amplification observed in specific candidates (SV 17, 22, 24, 27, and 28), we optimized the validation assay by utilizing the 2 × Taq Master Mix (Vazyme) and repeating the PCR experiments to ensure specific target amplification. Finally, combining the PCR results with haplotype-phased read visualization, we confirmed and retained 32 high-confidence somatic SV events for the final validated dataset.

Computational performance evaluation

To benchmark the computational efficiency of SVScope and other somatic SV callers, we executed all tools on the same unloaded high-performance computing node to avoid resource contention. Each tool, along with its requisite pre-processing steps (e.g., alignment or SNP calling), was configured to utilize 64 CPU threads to ensure a consistent parallelization environment. We measured the total wall-clock runtime and peak memory usage for each command using the GNU time utility (specifically/usr/bin/time -v).

The raw output logs from the time command were parsed using custom Python scripts to extract precise performance metrics. In detail, the elapsed wall-clock time was extracted using the regular expression 'Elapsed \(wall clock\) time \(h:mm:ss or m:ss\): (?:(\d+):)?(\d+):(\d+)'. The matched groups (hours, minutes, seconds) were then converted into total seconds for comparison. The maximum resident set size was extracted using the regular expression 'Maximum resident set size \(kbytes\): (\d+)'. The captured value (in kilobytes) was subsequently converted to Megabytes (MB) for reporting.

Isolated benchmarking of MSA algorithms

To evaluate the specific computational overhead and impact of different MSA algorithms on SVScope's performance, we performed an isolated benchmarking on 286 candidate SV intervals extracted from the HG008 ONT dataset. These intervals were selected from the SVScope pipeline where MSA and sequence mixture modeling are required for resolution. We compared four MSA aligners: abPOA(version 1.5.5) [27], SIMD-sPOA(4.1.5) [26], MAFFT(v7.525) [35], and POASTA (0.1.0) [36]. The benchmarking focused on two

primary metrics: wall clock time (seconds) and peak memory usage (MB). To ensure statistical stability, each aligner was executed in triplicate for every SV interval. In addition to the default settings, we implemented a high-accuracy MAFFT strategy using the parameters –maxiterate 1000 –globalpair to assess the tradeoff between alignment precision and computational cost. To further investigate how these MSA results influence variant calling accuracy, the outputs from each aligner were reintegrated into the SVScope pipeline for downstream analysis. The final SV detection performance, including F1-score, Precision, and Recall, was evaluated using Minda against the GIAB somatic SV gold standard for the HG008 cell line. All benchmarking experiments were conducted in a single-threaded environment to ensure a fair comparison of intrinsic algorithmic efficiency.

SVScope methodology

Alignment breakpoint encoding

The candidate somatic selection module of SVScope is the initial step in identifying candidate somatic SV loci. This process involves clustering alignment breakpoint information. SVScope collects all read alignment information from both case and control samples. Subsequently, it extracts alignment breakpoint pairs from two sources: CIGAR values (for inner-alignment breakpoints) and supplementary alignment records (for split-alignment breakpoints). Each breakpoint pair i is encoded as below,

$$BP_i = \left\{ D_i^Z, P_i^Z, S_i^Z, d_i, p_i^Z, L \right\}$$

where status $Z \in \{0,1\}$ represents the order of alignment breakpoint end with $Z=0$ for 5' end, and $Z=1$ for 3' end. The first 3 elements represent the location of each breakpoint end in reference genome with D for contig ID, P for chromosome location, and S for the DNA strand. The last 3 elements represent for the location of each breakpoint end in read with d for readID, p for read location, and L for read resource.

Processing of split-alignment somatic SVs (INV, TRA, BND)

For split-alignment somatic SVs, SVScope designs a systematic location encoder to further analyze their location relationships. Specifically, the chromosomes are sorted according to their IDs. Breakpoint positions on chromosomes with higher IDs are transformed into a unified coordinate system by summing the lengths of all preceding chromosomes. This approach ensures that breakpoint coordinates are consistently measured across different chromosomes, facilitating more accurate and systematic analysis of split-alignment somatic SVs. Paired split-alignment breakpoints which are close in each read (by default located within 100 bp in a single read) but at different coordination in the reference genome were re-encoded through this systematic location encoder and clustered through the DBSCAN algorithm with default settings (distance = 500 eps = 3). Finally, only breakpoint clusters sourced from the case sample are identified as somatic SV events. This step ensures that the identified SV events are specific to the case sample and not present in the control sample, thereby improving the accuracy of somatic SV detection.

Screening of candidate inner-alignment somatic SVs (INS, DEL)

For deletion type inner-alignment breakpoints, SVScope merged all breakpoints from case reads to generate a candidate deletion set. For insertion type inner-alignment breakpoints, case read breakpoints located within Repeatmasker annotated tandem repeat regions were firstly merged into the target tandem repeat genome interval. The others would be merged if the distance is lower than a preset threshold (by default 200 bp). To select candidate inner-alignment somatic SVs with significant sequence loss or gain, all sub-read sequences from both case and control samples spanning each candidate inner-alignment somatic SV were extracted with pysam packages. The candidate inner-alignment somatic SV intervals would be processed further if case reads have a larger (insertion case) or smaller (deletion case) length of span sequence over a preset threshold (by default 50 bp).

Local graph genome construction

The candidate somatic SV selection module provides a set of candidate somatic SV genomic intervals. In the local graph genome construction module, the SVScope algorithm targets all somatic structural variation (SV) candidate intervals and their flanking regions (by default captures a 50 bp range upstream and downstream of the target somatic SV interval). Utilizing the pysam toolkit, it extracts sequence sets from the reference genome and reads spanning these intervals in both tumor and paired normal samples. For every sequencing read length within the candidate somatic SV interval, all alignment records are extracted based on their alignment coordinates. Alignments on the negative strand are converted to positive strand sequences using the pysam toolkit. Subsequently, SVScope constructs a local graph genome for each candidate somatic SV interval sequence set using the POA-graph algorithm, supporting abPOA [27] and SIMD-sPOA [26], and extracts the output results of its multiple sequence alignment representation. Highly heterogeneous alignment sites are filtered out to serve as features for subsequent multi-class sequence mixture models (second largest sequence frequency > 3 or frequency > = 5%). For the case of N reads and m highly heterogeneous alignment sites, we encoded data as an observed matrix,

$$R = \begin{bmatrix} R_{1,1} & \cdots & R_{1,m} \\ \vdots & \ddots & \vdots \\ R_{N,1} & \cdots & R_{N,m} \end{bmatrix}$$

where one-hot vector $R_{i,j} = [R^0_{i,j}, R^1_{i,j}, R^2_{i,j}, R^3_{i,j}, R^4_{i,j}]$ represents the observed value of the j-th feature in the i-th read, which is corresponding to the 4 DNA bases A,T,C,G and the alignment gap in the multisequence alignment representation of POA graph. The matrix encapsulates the presence or absence of specific nucleotide bases or gaps at each heterogeneous feature site across all reads.

Multi-category sequence mixture model

We consider each candidate somatic structural variation (SV) interval as a local graph genome assembly constituted by K types of latent linear genomic sequences

$G = \{G_1, ..., G_k\}$. The composition of each genome in the whole cellular population is represented by the parameter $\pi = [\pi_1, ..., \pi_k]$. Each linear genome G_r is represented as a set of m features, and all features constituting the local graph genome can be represented by the parameter matrix,

$$\Theta = \begin{bmatrix} \theta_{1,1} & \cdots & \theta_{1,m} \\ \vdots & \ddots & \vdots \\ \theta_{K,1} & \cdots & \theta_{K,m} \end{bmatrix}$$

where each element $\theta_{i,j}$ within Θ is composed of 5 probabilities $[\theta_{ij}^0, \theta_{ij}^1, \theta_{ij}^2, \theta_{ij}^3, \theta_{ij}^4]$, corresponding to the four DNA bases A, T, C, G, and the alignment gaps.

For each sequencing read in the multisequence alignment representation of POA graph, $R_r = [R_{r,1}, ..., R_{r,m}]$ we employ a multi-class sequence mixture model for modeling, denoted as,

$$P(R_r|(\pi, \Theta))$$

$$= \sum_{G_r} P(R_r|G_r, (\pi, \Theta))P(G_r|(\pi, \Theta))$$

Then, the complete likelihood function of (R, G) is,

$$L((\pi, \Theta); G, R) = P(R, G|(\pi, \Theta)) = \prod_{r=1}^{N} P(R_r, G_r|(\pi, \Theta))$$

$$= \prod_{r=1}^{N}\prod_{k=1}^{K} \pi_k^{I(G_r=k)}[\prod_{j=1}^{m} {\theta_{kj}^0}^{R_{r,j}^0} \times {\theta_{kj}^1}^{R_{r,j}^1} \times {\theta_{kj}^2}^{R_{r,j}^2} \times {\theta_{kj}^3}^{R_{r,j}^3} \times {\theta_{kj}^4}^{R_{r,j}^4}]^{I(G_r=k)}$$

Therefore, the log-likelihood function of the complete data is,

$$l((\pi, \Theta); G, R) = \sum_{r=1}^{N}\sum_{k=1}^{K} I(G_{r=k})\pi_k + \sum_{r=1}^{N}\sum_{k=1}^{K} I(G_{r=k})\left\{\sum_{j=1}^{m}\sum_{i=0}^{4}\left[R_{r,j}^i \log(\theta_{kj}^i)\right]\right\}$$

Parameter estimation. The SVScope algorithm initiates the parameter estimation process by employing hierarchical clustering to approximate the initial model parameters. Specially, SVScope leverages the scipy package (version 1.7.3) to perform hierarchical clustering on the reads with the edit distance matrix,

$$D = \begin{bmatrix} D_{1,1} & \cdots & D_{1,N} \\ \vdots & \ddots & \vdots \\ D_{N,1} & \cdots & D_{N,N} \end{bmatrix}$$

where $D_{i,j} = \frac{|R_i - R_j|}{M}$, represents the edit distance between read i and read j. The outcome of this clustering, denoted as $N \times k$ one-hot matrix,

$$\gamma^{(0)} = \begin{bmatrix} \gamma^0_{1,1} & \cdots & \gamma^0_{1,k} \\ \vdots & \ddots & \vdots \\ \gamma^0_{N,1} & \cdots & \gamma^0_{N,k} \end{bmatrix}$$

serves as the foundation for initial model parameters $(\pi^{(0)}, \Theta^{(0)})$.

$$\pi^{(0)} = \frac{1}{N} \times [\sum_{r=1}^{N} \gamma^{(0)}_{r,1}, ..., \sum_{r=1}^{N} \gamma^{(0)}_{r,k}]$$

$$\Theta^{(0)} = (\gamma^{(0)})^T R / \mathrm{RowSum}((\gamma^{(0)})^T)$$

Next, the SVScope algorithm commences the parameter estimation phase by employing the Expectation–Maximization (EM) algorithm to refine the parameters of the multi-category sequence mixture model (π, Θ). The Q function induced by the complete log-likelihood function is,

$$Q((\pi, \Theta), (\pi^{(t)}, \Theta^{(t)})) = E(l(\pi, \Theta; G, R) | \pi^{(t)}, \Theta^{(t)})$$

$$= \sum_{r=1}^{N} \sum_{k=1}^{K} \gamma^{(t)}_{r,k} \pi_k + \sum_{r=1}^{N} \sum_{k=1}^{K} \gamma^{(t)}_{r,k} \left\{ \sum_{j=1}^{m} \sum_{i=0}^{4} \left[R^i_{r,j} \log(\theta^i_{kj}) \right] \right\}$$

where in the Expectation step,

$$E(I(G_r = k) | \pi^{(t)}, \Theta^{(t)}) = \gamma^{(t)}_{r,k} = \frac{\pi^{(t)}_k \prod_{j=1}^{m} \prod_{i=1}^{m} R^i_{r,j} \log(\theta^i_{kj})}{\sum_{k=1}^{K} \pi^{(t)}_k \prod_{j=1}^{m} \prod_{i=1}^{m} R^i_{r,j} \log(\theta^i_{kj})}$$

In the Maximization step,

$$\pi^{(t+1)}_k = \frac{\sum_{r=1}^{N} \gamma^{(t)}_{r,k}}{N}$$

$$\Theta^{(t+1)} = \frac{\sum_{r=1}^{N} \gamma^{(t)}_{r,k} R_r}{\sum_{r=1}^{N} \gamma^{(t)}_{r,k}}$$

SVScope systematically estimates the model parameters for a spectrum of mixture model sizes, ranging from $k = 1$ to $k = 10$, and ascertains the optimal number of genome components K with minimal Bayesian information criterion (BIC),

$$\mathrm{argmin}(BIC) = \mathrm{argmin}[-2\log(\widehat{L}) + k\log(N)]$$

Local graph genome refinement

To evaluate the confidence of optimized local graph genome and avoid the false positives caused by highly-repetitive regions or segmental duplications, we trained a Random Forest (RF) classifier using Scikit-Learn (version 1.0.2). To this aim, we selected 3,452

labeled candidate somatic SV intervals from five cell lines (HCC1395, H2009, H1437, HCC1954, HCC1937) using the local graph genome optimization module. These intervals were used as the candidate training set. We then summarized the local alignment features of these intervals as interval features. Specifically, we calculated the absolute sequence difference length between the tumor genome and the paired normal genome, marked as ABSMisScore. Additionally, we computed the proportion of reads spanning the interval in the case and control samples, the proportion of low mapping quality (< 5) reads, the proportion of non-uniquely aligned reads, and the normalized coverage score based on the average and standard deviation of read counts in 10-kb windows across the genome. All data were randomly split into training and testing sets at a ratio of 4:1. The optimal values for the hyperparameters of the Random Forest (RF) model including Number of trees (n_estimators), Maximum tree depth (max_depth), Feature selection (max_features), Bootstrap sampling (bootstrap), Split criterion (criterion), Minimum samples required to split (min_samples_split), and Minimum samples per leaf (min_samples_leaf) were determined using cross-validation and grid search, with the number of decision trees ranging from 5 to 30 and the maximum depth ranging from 2 to 64. To detect somatic SVs in each cell line, we trained an RF model using the labeled SVs from all other cell lines to ensure that no data from the cell line to be analyzed was used during model training.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-026-04076-0.

Additional file 1: Table S1. Source of cell line long-read WGS data. Table S2. Minda evaluation of SVScope performance on HG008-T vs HG008-N-P ONT dataset using different MSA algorithms. Table S3. Validation information for somatic SVs in HCC1395-HCC1395BL.

Additional file 2: Fig. S1. Construction workflow and structural variation distribution of benchmark datasets. Fig. S2. Performance characterization of somatic SV detection algorithms across different SV types and genomic regions. Fig. S3. Genome-wide coverage, mapping quality and alignment ambiguity performance. Fig. S4. Detailed Schematic of the candidate somatic SV selection workflow. Fig. S5. Machine learning model performance evaluation. Fig. S6. Performance comparison of SVScope and five other methods across seven cell lines. Fig. S7. Performance benchmarking of isolated Multiple Sequence Alignmentaligners. Fig. S8. Construction of the haplotype-aware simulated dataset for tandem repeat evaluation. Fig. S9. Performance evaluation of SV detection methods on simulated data. Fig. S10. Support reads precision of SV detection methods on simulated data. Fig. S11. Representation of SVScope in heterogenous alignment breakpoint somatic SV case on chr22:50,137,660-50,137,991. Fig. S12. Integrated visualization and ScopeVIZ analysis of simulated structural variations in the chr22:10,538,368-10,538,453 region. Fig. S13. Overlap of structural variantsdetected in the HCC1395 cell line. Fig. S14. Detailed characterization of somatic SV and validation of somatic tandem repeat SV events of HCC1395 cell line in chr3:79,909,993-79,910,160.

Acknowledgements

We thank Xuyan Liu and Jing Yang, PhD students at Sichuan University, and Dr. Jiaxun Zhang from The Third People's Hospital of Chengdu for their valuable discussions regarding the algorithm design and pipeline testing.

Peer review information

Andrew Cosgrove and Wenjing She were the primary editors of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. The peer-review history is available in the online version of this article.

Authors' contributions

K.T. and Q.Z. contributed equally to this work. K.T., W.H., and D.X. were responsible for the study design. W.H. developed the original theoretical model, K.T. derived the extended model for somatic SV detection and implemented the associated algorithms. Q.Z. reviewed the code and performed the evaluation of the algorithms. Y.L. and J.T. designed the experiments, J.W. executed the experiments. Q.Z., L.F.Y. and Y.C.L. performed the simulation and visualization of data. L.X. analyzed the application scenarios of the algorithms and contributed to the discussion of their implications. All authors reviewed and approved the final manuscript.

Funding

This work was supported by grants from the National Natural Science Foundation of China (82173383 to D. X., 11671073 to W.H., 32300522 to K.T., 32200508 to L. X., 82072582 to J.T.), the 1·3·5 project for disciplines of excellence, West China

Hospital, Sichuan University (ZYYC23024 to D. X.). This work is also supported by the Science and Technology Department of Sichuan Province, China (2025ZNSFSC1098 to K.T., 2024NSFSC1187 to Y.L.), and the Fundamental Research Funds for the Central Universities, China to W.H.

Data availability

The sequencing data in this study is downloaded from SRA database under accession number SRP494767 (BioProject: PRJNA1086849) [17, 51], the DNA Data Bank of Japan (DDBJ) DRP006799 (BioProject: PRJDB10898) [16, 52], and HG008 tumor-normal celline pairs from the GIAB repository [28, 53]. Detail accession number for each cell lines were listed in Additional file 1: Table S1. The original source code for SVScope and ScopeVIZ is available on GitHub at https://github.com/Goatofmountain/SVScope [54] under the Apache License version 2.0. A persistent version of the SVScope source code, the code used in this paper, and demo ONT/PacBio datasets are archived in Zenodo at https://doi.org/10.5281/zenodo.18681800 [55]. Benchmark somatic structural variations (SVs), including PCR validation labels and phased IGV images for the HCC1395 cell line, are available in Zenodo at https://doi.org/10.5281/zenodo.18688405 [56].

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing Interests

The authors declare no competing interests.

Received: 27 May 2025 Accepted: 7 April 2026
Published online: 23 April 2026

References

1. Stephens PJ, McBride DJ, Lin ML, Varela I, Pleasance ED, Simpson JT, et al. Complex landscapes of somatic rearrangement in human breast cancer genomes. Nature. 2009;462:1005–10.
2. Li Y, Roberts ND, Wala JA, Shapira O, Schumacher SE, Kumar K, et al. Patterns of somatic structural variation in human cancer genomes. Nature. 2020;578:112–21.
3. Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. Structural variant calling: the long and the short of it. Genome Biol. 2019;20:246.
4. Chiang C, Scott AJ, Davis JR, Tsang EK, Li X, Kim Y, et al. The impact of structural variation on human gene expression. Nat Genet. 2017;49:692–9.
5. Scott AJ, Chiang C, Hall IM. Structural variants are a major source of gene expression differences in humans and often affect multiple nearby genes. Genome Res. 2021;31:2249–57.
6. Wang Z, Tu K, Xia L, Luo K, Luo W, Tang J, et al. The open chromatin landscape of non-small cell lung carcinoma. Cancer Res. 2019;79:4840–54.
7. Xu L, Wang X, Lu X, Liang F, Liu Z, Zhang H, et al. Long-read sequencing identifies novel structural variations in colorectal cancer. PLoS Genet. 2023;19:e1010514.
8. Zhang Y, Chen F, Fonseca NA, He Y, Fujita M, Nakagawa H, et al. High-coverage whole-genome analysis of 1220 cancers reveals hundreds of genes deregulated by rearrangement-mediated cis-regulatory alterations. Nat Commun. 2020;11:736.
9. Dubois F, Sidiropoulos N, Weischenfeldt J, Beroukhim R. Structural variations in cancer and the 3D genome. Nat Rev Cancer. 2022;22:533–46.
10. Zhang Y, Yang L, Kucherlapati M, Hadjipanayis A, Pantazi A, Bristow CA, et al. Global impact of somatic structural variation on the DNA methylome of human cancers. Genome Biol. 2019;20:209.
11. Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods. 2009;6:677–81.
12. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics. 2009;25:2865–71.
13. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 2012;28:i333–9.
14. Boeva V, Popova T, Bleakley K, Chiche P, Cappo J, Schleiermacher G, et al. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. Bioinformatics. 2012;28:423–5.
15. Dolzhenko E, Deshpande V, Schlesinger F, Krusche P, Petrovski R, Chen S, et al. ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. Bioinformatics. 2019;35:4754–6.
16. Shiraishi Y, Koya J, Chiba K, Okada A, Arai Y, Saito Y, et al. Precise characterization of somatic complex structural variations from tumor/control paired long-read sequencing data with nanomonsv. Nucleic Acids Res. 2023;51:e74.
17. Keskus AG, Bryant A, Ahmad T, Yoo B, Aganezov S, Goretsky A, et al. Severus detects somatic structural variation and complex rearrangements in cancer genomes using long-read sequencing. Nat Biotechnol. 2025. https://doi.org/10.1038/s41587-025-02618-8.
18. Wang S, Lin J, Jia P, Xu T, Li X, Liu Y, et al. De novo and somatic structural variant discovery with SVision-pro. Nat Biotechnol. 2024;43:181–5.

19. Jiang T, Liu Y, Jiang Y, Li J, Gao Y, Cui Z, et al. Long-read-based human genomic structural variation detection with cuteSV. Genome Biol. 2020;21:189.
20. Smolka M, Paulin LF, Grochowski CM, Horner DW, Mahmoud M, Behera S, et al. Detection of mosaic and population-level structural variants with Sniffles2. Nat Biotechnol. 2024;42:1571–80.
21. Kolmogorov M, Billingsley KJ, Mastoras M, Meredith M, Monlong J, Lorig-Roach R, et al. Scalable nanopore sequencing of human genomes provides a comprehensive view of haplotype-resolved variation and methylation. Nat Methods. 2023;20:1483–92.
22. Waye JS, Willard HF. Chromosome-specific alpha satellite DNA: nucleotide sequence analysis of the 2.0 kilobasepair repeat from the human X chromosome. Nucleic Acids Res. 1985;13:2731–43.
23. Sen Gupta A, Seidel C, Tsuchiya D, McKinney S, Yu Z, Smith SE, et al. Defining a core configuration for human centromeres during mitosis. Nat Commun. 2023;14:7947.
24. Logsdon GA, Rozanski AN, Ryabov F, Potapova T, Shepelev VA, Catacchio CR, et al. The variation and evolution of complete human centromeres. Nature. 2024;629:136–45.
25. Moyzis RK, Buckingham JM, Cram LS, Dani M, Deaven LL, Jones MD, et al. A highly conserved repetitive DNA sequence, (TTAGGG)n, present at the telomeres of human chromosomes. Proc Natl Acad Sci U S A. 1988;85:6622–6.
26. Lee C, Grasso C, Sharlow MF. Multiple sequence alignment using partial order graphs. Bioinformatics. 2002;18:452–64.
27. Gao Y, Liu Y, Ma Y, Liu B, Wang Y, Xing Y. abpoa: an SIMD-based C library for fast partial order alignment using adaptive band. Bioinformatics. 2021;37:2209–11.
28. McDaniel JH, Patel V, Olson ND, He HJ, He Z, Cole KD, et al. Development and extensive sequencing of a broadly-consented Genome in a Bottle matched tumor-normal pair. Sci Data. 2025;12:1195.
29. Espejo Valle-Inclan J, Besselink NJM, de Bruijn E, Cameron DL, Ebler J, Kutzera J, et al. A multi-platform reference for somatic structural variation detection. Cell Genom. 2022;2:100139.
30. English AC, Menon VK, Gibbs RA, Metcalf GA, Sedlazeck FJ. Truvari: refined structural variant comparison preserves allelic diversity. Genome Biol. 2022;23:271.
31. Sahlin K, Baudeau T, Cazaux B, Marchet C. A survey of mapping algorithms in the long-reads era. Genome Biol. 2023;24:133.
32. Vollger MR, Guitart X, Dishuck PC, Mercuri L, Harvey WT, Gershman A, et al. Segmental duplications and their variation in a complete human genome. Science. 2022;376:eabj6965.
33. Zheng Z, Li S, Su J, Leung AW, Lam TW, Luo R. Symphonizing pileup and full-alignment for deep learning-based long-read variant calling. Nat Comput Sci. 2022;2:797–803.
34. Elrick H, Sauer CM, Espejo Valle-Inclan J, Trevers K, Tanguy M, Zumalave S, et al. SAVANA: reliable analysis of somatic structural variants and copy number aberrations using long-read sequencing. Nat Methods. 2025;22:1436–46.
35. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002;30:3059–66.
36. van Dijk LR, Manson AL, Earl AM, Garimella KV, Abeel T. Fast and exact gap-affine partial order alignment with POASTA. Bioinformatics. 2024. https://doi.org/10.1093/bioinformatics/btae757.
37. Akbari V, Garant JM, O'Neill K, Pandoh P, Moore R, Marra MA, et al. Megabase-scale methylation phasing using nanopore long reads and NanoMethPhase. Genome Biol. 2021;22:68.
38. Yuen ZW, Srivastava A, Daniel R, McNevin D, Jack C, Eyras E. Systematic benchmarking of tools for CpG methylation detection from nanopore sequencing. Nat Commun. 2021;12:3438.
39. Lee I, Razaghi R, Gilpatrick T, Molnar M, Gershman A, Sadowski N, et al. Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing. Nat Methods. 2020;17:1191–9.
40. Altemose N, Maslan A, Smith OK, Sundararajan K, Brown RR, Mishra R, et al. DiMeLo-seq: a long-read, single-molecule method for mapping protein-DNA interactions genome wide. Nat Methods. 2022;19:711–23.
41. Battaglia S, Dong K, Wu J, Chen Z, Najm FJ, Zhang Y, et al. Long-range phasing of dynamic, tissue-specific and allele-specific regulatory elements. Nat Genet. 2022;54:1504–13.
42. Tu K, Li X, Zhang Q, Huang W, Xie D. A data-adaptive methods in detecting exogenous methyltransferase accessible chromatin in human genome using nanopore sequencing. Bioinformatics. 2024. https://doi.org/10.1093/bioinformatics/btae206.
43. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol. 2018;36:338–45.
44. Nguyen E, Poli M, Durrant MG, Kang B, Katrekar D, Li DB, et al. Sequence modeling and design from molecular to genome scale with Evo. Science. 2024;386:eado9336.
45. Dalla-Torre H, Gonzalez L, Mendoza-Revilla J, Lopez Carranza N, Grzywaczewski AH, Oteri F, et al. Nucleotide transformer: building and evaluating robust foundation models for human genomics. Nat Methods. 2025;22:287–97.
46. Ji Y, Zhou Z, Liu H, Davuluri RV. DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. Bioinformatics. 2021;37:2112–20.
47. Devlin J, Chang M-W, Lee K, Toutanova K: Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). 2019: 4171–4186.
48. Wick RR. Badread: simulation of error-prone long reads. J Open Source Softw. 2019;4:1316.
49. Bolognini D, Sanders A, Korbel JO, Magi A, Benes V, Rausch T. VISOR: a versatile haplotype-aware structural variant simulator for short- and long-read sequencing. Bioinformatics. 2020;36:1267–9.
50. Li H. New strategies to improve minimap2 alignment accuracy. Bioinformatics. 2021;37:4572–4.
51. Keskus AG, Bryant A, Ahmad T, Yoo B, Aganezov S, Goretsky A, Donmez A, Lansdon LA, Rodriguez I, Park J, et al: Severus detects somatic structural variation and complex rearrangements in cancer genomes using long-read sequencing. Sequence Read Archive. https://www.ncbi.nlm.nih.gov/sra/?term=SRP494767 (2025 of deposition).
52. Shiraishi Y, Koya J, Chiba K, Okada A, Arai Y, Saito Y, Shibata T, Kataoka K: Precise characterization of somatic complex structural variations from tumor/control paired long-read sequencing data with nanomonsv. DNA Data Bank of Japan (DDBJ). https://ddbj.nig.ac.jp/search/entry/sra-study/DRP006799 (2023 of deposition).

53. McDaniel JH, Patel V, Olson ND, He HJ, He Z, Cole KD, Gooden AA, Schmitt A, Sikkink K, Sedlazeck FJ, et al: Development and extensive sequencing of a broadly-consented Genome in a Bottle matched tumor-normal pair the Genome in a Bottle (GIAB). https://ftp.ncbi.nlm.nih.gov/ReferenceSamples/giab/data_somatic/HG008/ (2025 of deposition).
54. Kailing T, Qilin Z, Yang L, Yucong L, Lanfang Y, Jing W, Jie T, Lin X, Wei H, Dan X: SVScope improves somatic structural variations detection via graph-genome optimization. Github. https://github.com/Goatofmountain/SVScope (2026 of deposition).
55. Kailing T, Qilin Z, Yang L, Yucong L, Lanfang Y, Jing W, Jie T, Lin X, Wei H, Dan X: SVScope improves somatic structural variations detection via graph-genome optimization (Source Code). Zenodo. https://doi.org/10.5281/zenodo.18681800 (2026 of deposition).
56. Kailing T, Qilin Z, Yang L, Yucong L, Lanfang Y, Jing W, Jie T, Lin X, Wei H, Dan X: SVScope improves somatic structural variations detection via graph-genome optimization (Somatic SV validations). Zenodo. https://doi.org/10.5281/zenodo.18688405 (2026 of deposition).

Publisher's Note