Accurate Somatic SV detection via sequence graph model-based local pan-genome optimization

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

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, Chengdu, Sichuan, 610041, China
2. School of Mathematics and Statistics, Key Laboratory for Applied Statistics of the Ministry of Education, Northeast Normal University, Changchun 130024, China

* Corresponding authors:
Laboratory of Omics Technology and Bioinformatics, Frontiers Science Center for Disease-related Molecular Network, State Key Laboratory of Biotherapy, West China Hospital, Sichuan University, Chengdu, Sichuan, 610041, China. No. 17, Section 3, Renmin South Road, Chengdu, Sichuan 610041, China. E-mail: danxie@scu.edu.cn (D.X.)
School of Mathematics and Statistics, Key Laboratory for Applied Statistics of the Ministry of Education, Northeast Normal University, 5268 Renmin Street, Changchun, Jilin Province 130024, China. E-mail: huangw482@nenu.edu.cn (W.H.)

Abstract

Somatic structural variations (Somatic SVs) are critical genomic alterations with significant implications in cancer genomics. Although long-read sequencing (LRS) theoretically provides optimal resolution for detecting these variants due to its ability to span large genomic segments, current LRS - based methods, which are derived from short - read - based somatic SV detection algorithms, mainly rely on split - read information. The high error rate of long - read sequencing and the errors introduced by the seed-and-chaining strategy of mainstream alignment algorithms affect the accuracy of these split-reads, making precise detection of somatic SVs still a challenge. To address this issue, we propose the TDScope algorithm, which uses the complete sequence information of local genomic regions provided by long-read sequencing to construct a local graph genome and combines random forest technology to achieve precise detection of somatic structural variations. TDScope outperforms state-of-the-art somatic SV detection methods on paired long-read whole-genome sequencing (WGS) benchmark cell lines, with an average F1-score improvement of 20%. It also demonstrates superior performance in detecting somatic SVs and

resolving heterogeneous genomes in tandem repeat-like simulated somatic SV datasets. We also provide the ScopeVIZ tool to offer users visualization evidence of local graph genomes and somatic SV sequences. All code implementations are publicly available on GitHub (https://github.com/Goatofmountain/TDScope).

Introduction

Somatic structural variations (somatic SVs) or de novo SVs can be broadly defined as large genomic alterations exceeding 50 bp in length that occur in specific somatic cell clones. Somatic SVs originate from genomic rearrangements and involve larger segments of base variations compared to point mutations(1,2). They participate in more complex genomic regulatory processes. Recent studies have identified that somatic SV events mediate genomic copy number variations (CNVs)(3), gene fusions(4), and cis-regulatory alterations(5). 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(6), and modulate the methylation status of genome(7). 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.

Over the past decade, second-generation sequencing techniques relied on alignment breakpoints (8-10) and local alignment depth information(11,12) to decipher somatic SVs. On one hand, the high sequencing accurcy of short-read technology ensures minimal alignment errors, thereby providing reliable breakpoint information. On the other hand, the limited read lengths preclude the direct capture of the complete sequence landscape of SVs, compelling the use of breakpoint-based strategies. Third-generation single-molecule sequencing technologies produce long-read segments that can carry complete SV information severes as a better method for somatic SV detection. Several studies have demonstrated that long-read alignment or assembly can far exceed short-read sequencing data in precisely describing SVs (13). Current somatic SV detection algorithms based on long-read sequencing still rely on breakpoint features from read alignment to detect somatic SVs, which can be seen as an extension of short-read somatic SV detection algorithms. These alignment-based algorithms are computationally efficient and can quickly identify somatic SV events. Unlike short-read sequencing, long-read sequencing technologies struggle to provide high-precision breakpoint information. In addition to the high error rate of single-base sequences, the seed-and-chain strategy of long-read alignment algorithms is also affected by high-frequency short-segment sequences, introducing local sequence alignment errors in low-complexity or tandem repeat sequences that affect the high error rate of alignment breakpoints. To address the errors introduced by alignment algorithms, past studies have proposed re-alignment strategies(14), tandem repeat sequence breakpoints merging strategies(15), and individual genome polishing strategies(16) to improve the accuracy of breakpoint features. To mitigate errors from feature selection, previous somatic SV callers have incorporated additional phasing features, such as Severus(15), and introduced machine learning frameworks, such as SAVANA(17), or deep learning frameworks, such as SVision-Pro(18), to enhance the detection capabilities of somatic SV events. However, the reliance on local alignment breakpoints limits the effectiveness of current algorithms, as the seed-and-chaining strategy can cause inaccuracies in breakpoint positioning,

particularly in low-complexity regions like tandem repeats(19). Crucially, the strength of long-read sequencing lies in its ability to capture full-length sequence information, not merely breakpoints. To fully leverage this advantage and address the heterogeneity of tumor genomes, a systematic approach based on complete sequence information is essential for precise somatic SV detection.

Different from the existing somatic SV callers, we introduce TDScope, a novel approach that frames somatic SV identification as an optimization problem of the local pangenome. Instead of relying predominantly on alignment breakpoints, TDScope integrates complete sequences from reads spanning candidate somatic SV regions, thereby minimizing feature loss. By implementing a sequence Partial Order Alignment (sPOA) graph with multi-sequence alignment representation, TDScope significantly reduces the impact of sequence alignment bias (20). Additionally, it achieves local-graph genome optimization and precise somatic SV detection through graph-based sequence mixture models combined with global alignment feature-based machine learning techniques. TDScope demonstrates superior performance in detecting somatic SVs in high-quality whole-genome Oxford Nanopore Technologies (ONT) sequencing datasets. Evaluated on 5 pairs of tumor-normal benchmark cell lines and simulated datasets with varying tumor purity and genomic heterogeneity, TDScope outperforms state-of-the-art somatic SV callers. Furthermore, its ability to parse heterogeneous genome components is exceptional, as confirmed by simulation data, where it achieved high detection accuracy for different genomic elements and read attribution.

Result

Optimize somatic SV detection, refinement and population analysis with TDScope

Addressing the need for accurate somatic SV detection, refinment and population analysis, we propose TDScope, comprising two key modules. The first module, local-graph genome optimization, establishes a local graph genome based on complete sequence of span-reads from two samples with sPOA algorithm (20), and optimized by a sequence mixture model (Fig.1 Methods). Candidate somatic SVs are selected from tumor-read specific path in the local graph genome. The local-graph genome optimization process addresses the break point bias originating from alignment algorthms, particularly in low complex sequences, and is effective for parsing genome sequences of various cell components under the conditions of high intratumoral heterogeneity. The second module, refinement of local-graph genome, utilizes a random forest model to accesses the confidence of the local-graph genome through reads alignment signatures including germline-somatic differences, local coverage, alignment quality and non-uniqueness read frequency (Fig. 1, Methods). The refinement process mitigates global errors introduced by inaccuracies in the reference genome and limitations in sequencing read length. It effectively reduces false positives associated with segmental duplications and low-quality reference genome sequences, particularly in regions such as centromeres and telomeres. Through the computations of these two modules, the TDScope algorithm provides sequence information of local graph genome components containing somatic structural variants, as well as supporting reads.

Additionally, the local-graph optimization module is also avaliable for population-scale somatic sequence comparison, enabling the identification of sample clusters with similar somatic variation patterns and facilitating the detection of both shared and cluster-specific variant features among the samples (Fig. 1). Full pipeline of TDScope is available as stand-alone package within github.

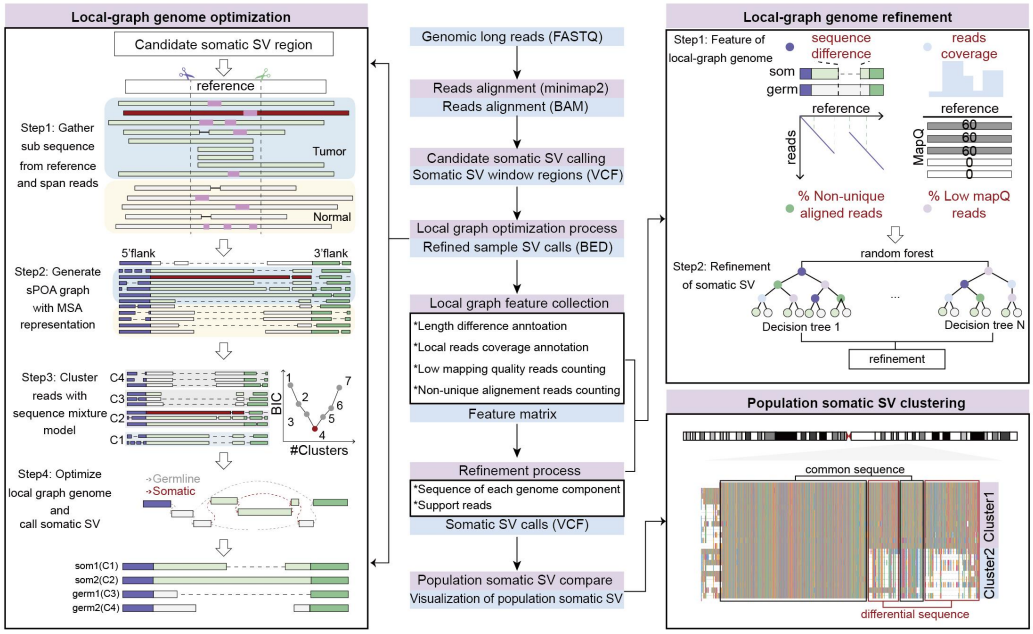

Figure 1. Schematic of the TDScope Workflow. This schematic illustrates the workflow of the TDScope algorithm for detecting somatic structural variations (SVs) from FASTQ data and comparing somatic SV sequences across population samples. The detection process is divided into two main steps: local pangenome optimization and pangenome confidence assessment. In the local pangenome optimization step, the algorithm utilizes the complete sequence information of reads to parse the heterogeneous genome and identify components containing somatic SVs. The pangenome confidence assessment step constructs a random forest model using sequence alignment features to refine the detection of somatic SVs. Additionally, TDScope can assess the similarities and differences in somatic sequences between samples within a population, facilitating comparative genomic analyses.

Accurate detection of somatic SVs by resolving local biases

We benchmarked the performance of TDScope and other approaches using publicly available dastasets from 5 pairs of matched normal-tumor cell lines (Methods, Supplementary Table S1). TDScope demostrated a significant improvement in the accuracy and specificity in somatic SV detection compared to other published methods (Fig. 2A), with an average increase of 20% (3.5%~39.2%) in F1 score accross all data sets (Fig 2B). As a supplement to the existing somatic SV benchmark set, we conducted target PCR (Supplementary Fig. S1) and Sanger sequencing validations (Extended Data Fig. 1-4) for the 15 false positive somatic SVs identified by TDScope in HCC1395-HCC1395BL cell lines (Supplementary Table S2). Notably, the detection accruacy of somatic SVs within low-complexity seqeunces, including low-complex and tandem repeat

sequence, presents a significant challenge. Algorithms reliant on alignment breakpoint information, such as sniffles2, severus, and SAVANA, incurred over 50% of false positive or false negatives stemming from these SV events (Fig 2C). Furthermore, local genomic heterogeneity also impacts the detetion of somatic SVs, with SVision-pro, NanomonSV, and SAVANA algorithms being prone to generating false positives in proximity to germline SVs (Fig 2D).

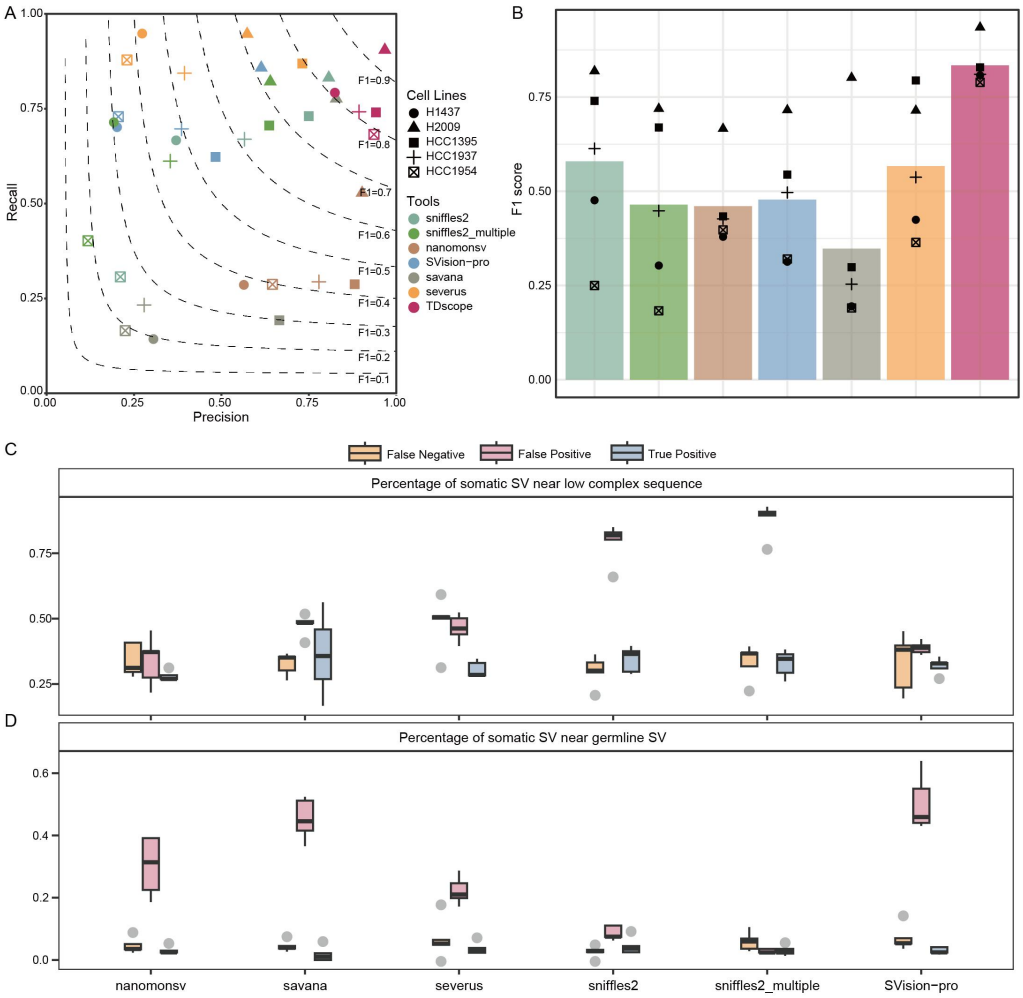

Figure 2. Performance of somatic SV caller in benchmark cell lines. A. Scatter plot with contour lines illustrating the precision and recall of TDScope and other state-of-the-art somatic SV callers across five benchmark cell lines in detecting somatic SVs. Different cell lines and callers are denoted as per the legend on the right. Dashed lines represent various gradients of F1 scores. B. Bar plot showing the F1 scores for different cell lines, comparing the performance of TDScope with other callers. C. Box plot displaying the percentage of somatic SVs detected near low-complexity sequences by various callers. The box represents the interquartile range (IQR), with the line inside the box showing the median. Whiskers extend to the most extreme data points not considered outliers, and outliers are plotted as individual points. D. Box plot illustrating the percentage of somatic SVs detected near germline SVs by different callers. The representation is consistent with panel (C), providing insights into the callers' ability to identify SVs in regions with high germline SV density.

To further assess TDScope's performance under these confoundings, we constructed a simulated dataset centered around tandem repeat sequences in the human genome (Methods). Under varying tumor purity and SV length conditions, TDScope outperforms other somatic SV callers, including Straglr (21), a specific tandem repeat expansion caller (Fig. 3A). It is challenging for the mainstream long-read aligner based on minimize-and-chainnig strategy, like minimap2, to generate highly specific minimizers in low complex sequence. Reads harboring somatic SVs are prone to alignment breakpoint errors due to the variations choice of minimizers in the nearby flanking sequences, leading to different alignment breakpoints of reads origin from the same genome (Supplementary Fig. S2A). These heterogenous breakpoint distributions cause high error rate for other somatic SV callers (Supplementary Fig. S2B, Methods). Due to the application of complete read sequences and sPOA graph based multiple sequence alignment strategies, TDScope is unaffected by such read-reference alignment errors, ensuring proper reads alignment (Supplementary Fig. S2C). Building on this foundation, the local graph genome optimization process based on sequence mixture model can accurately estimate the genomic information of different cellular components, thereby enabling precise parsing of heterogeneous genomes (Supplementary Fig. S2D). This feature allows TDScope to detect somatic SVs without being influenced by heterogenous genome sequences, achieving optimal computational performance in somatic SV event detection under different tumor purities (Fig. 3B).

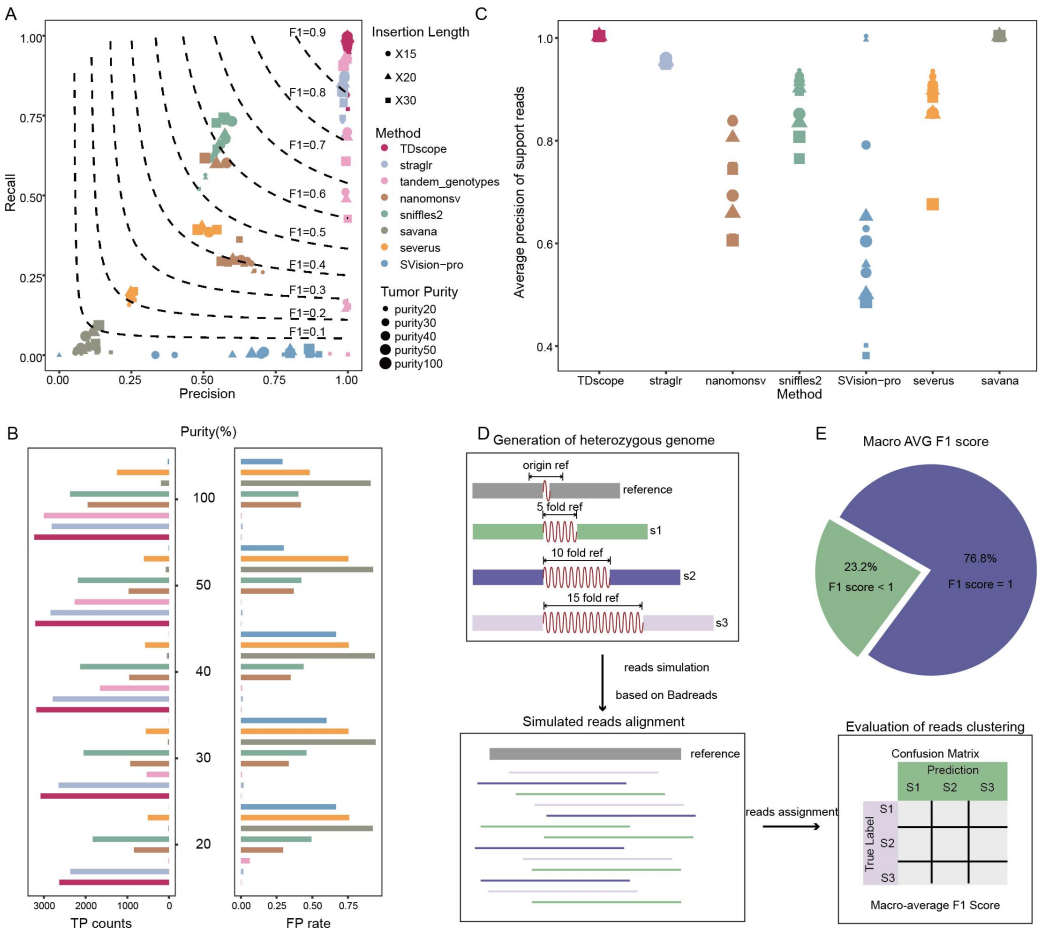

Fig.3. Performance of somatic SV callers in simulated somatic tandem repeat dataset. A. Scatter plot with contour lines illustrating the precision and recall of TDScope and other state-of-the-art somatic SV callers across all simulated dataset with 5 tumor purity threshold and 3 tandem repeat extension scale. The parameters of datasets and somatic SV callers are denoted on the right. Dashed lines represent various gradients of F1 scores. B. Barplot shows the count of true positive somatic SVs and false positive rate of each somatic SV callers under different tumor purity. C. Scatter plot shows the accuracy of somatic SV support reads in each somatic SV callers under different tumor purity. tumor purity and SVcaller name in Fig.3B-C are color-coded in Fig.3A. D. Schematic of simulated data for heterogenous genome. C. Pie chart shows the percentage of simulated heterogenous regions with perfect read assignment performance using TDScope (F1 score=1).

Detailed characteration of different genome component

In addition to its ability to accurately detect somatic structural variants, TDScope demonstrated the highest precision in identifying somatic SV support reads across all simulated datasets (Fig.3C). Notably, the presence or absence of germline SV events did not influence TDScope's precision in detecting somatic SV support reads, a finding corroborated in Supplementary Figure S3.To further evaluate TDScope's capability in resolving heterogeneous genomic components, we constructed a dataset comprising 1,864 sets of heterogeneous genomes mixed with tandem repeat sequences of three different lengths (Fig 3D). TDScope exhibited a macro average F1 score of 1 for read length discrimination in 76.8% of these sets (Fig 3E). Leveraging its advantages, TDScope is capable of identifying complex alignment breakpoints and somatic SV events within highly heterogeneous genomic regions in real data, such as the somatic deletion event in the HCC1395-HCC1395BL cell line pair at chr6:598,996-604990 (Fig. 4A). TDScope preciesly parses the sequence information of different genomic components within this local region by utilizing the complete read sequences in this interval (Fig. 4B), and through sPOA graph and sequence mixture model-based read clustering, it can accurately detect somatic SVs (Fig.4C-E). Furthermore, we have designed a somatic SV detection workflow specifically for tandem repeat regions (Methods). In the HCC1395-HCC1395BL cell lines, we identified 192 somatic SV events located within tandem repeat sequences (Fig. 4F). Among these, 22.40% of the tandem repeat somatic SVs were undetectable by other somatic SV callers (Fig. 4G). We validated 40 of these using targeted PCR, and all showed distinct band differences (Fig. 4H). For these somatic SVs, we provided the germline and somatic genomic sequences, as well as PCR primer information (Supplementary Table S2).

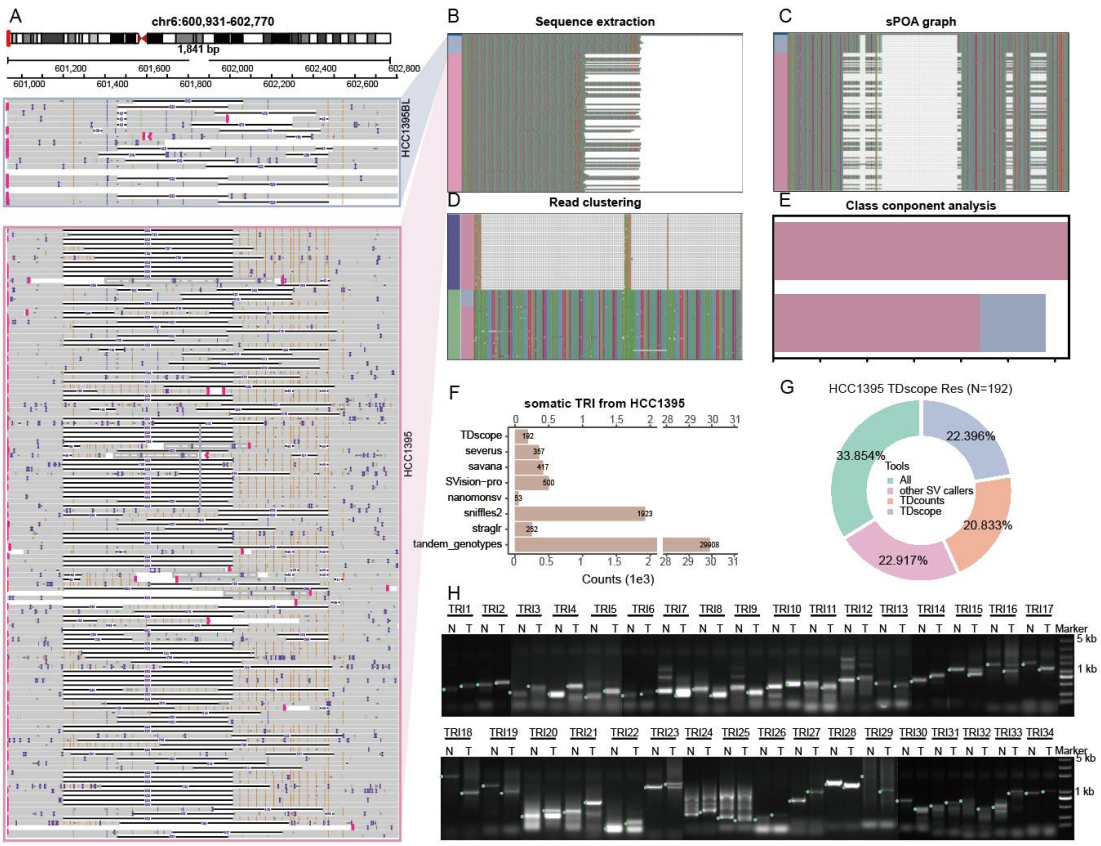

Figure 4. Detailed characterization of somatic SV and validation of somatic tandem repeat SV events. A. Genome browser plot shows the example somatic SV region chr6:600,931-602,770 in HCC1395BL(top) and HCC1395(bottom). B.Bar plot shows the raw sequence extracted from HCC1395BL (colored in blue) and HCC1395 reads (colored in red). C. Bar plot shows multiple sequence alignment representation of sPOA graph based on read sequences from HCC1395BL (colored in blue) and HCC1395 reads (colored in red). D. Bar plot shows the result of reads clustering based on TDScope. E. Bar plot shows the reads components of each cluster, with reads from HCC1395 colored in red and HCC1395BL in blue. F. Bar plot shows the count of somatic tandem repeat insertions (TRIs) detected by TDScope and other state-of-art somatic SV callers. G. Pie chart shows the percentage of somatic TRIs detected by various somatic SV callers. H. The gel electrophoresis image illustrates the targeted PCR results for 40 somatic TRI events exclusively detected by the TDScope algorithm in the genomes of HCC1395BL (left panel) and HCC1395 (right panel). Notably, the target bands are highlighted with green dots, clearly indicating the presence of somatic TRI events in both genomes.

Discussion

Here we introduced TDScope, an algorithm that employs a two-step computational approach to locally optimize heterogeneous genomic sequences and accurately detect somatic SVs. Through evaluation using 5 pairs of tumor-normal benchmark cell line datasets and simulated data, we demostrate that TDScope pipeline outperforms other state-of-the-art methods in characterizing heterogeneous genomes and detecting somatic SVs. To enhance the applicability of TDScope, we

further provide a visualization tool, ScopeVIZ, which allows users to visualize read-level evidence supporting each genomic component of the local graph genome, including somatic SVs. Additionally, ScopeVIZ enables the visualization of differences in somatic SV seuqneces among individuals at the population sample level. These companion methods will assist researchers and clinical laboratories in annotating and visually inspecting the somatic SV sequences detected by TDScope.

The local graph genome optimization step is the core component of the TDScope algorithm, relying on span-reads to provide complete information for the target intervals. However, due to limitations in reference genome accuracy and read length, large-scale somatic SVs such as translocations and large inversions cannot always be precisely analyzed due to the unavailability of span-reads. Advances in ultra-long read sequencing technologies are expected to mitigate these challenges in the future. The computational resource consumption is another significant issue with the current algorithm. This is primarily attributed to the sPOA-graph algorithm and the extensive sequence feature extraction required for the Expectation-Maximization (EM) algorithm in the sequence mixture model. Optimizing feature extraction processes is a direct solution to address this problem. Leveraging the feature learning capabilities of deep learning to distill sequence features is a promising future direction. With the development of genomic large language models, we anticipate designing more efficient somatic SV detection algorithms in future research.

In conclusion, TDScope leverages the complete sequence of reads and extract features based on the multisequence alignment representation of sPOA-graph to avoid sequence information loss and bias introduced by reads aligner. Building on this foundation, TDScope effectively parses different sequence components of heterogeneous genomes and accurately identifies the genotype information of each single-molecule read through the use of a sequence mixture model. With the development of long-read single-molecule sequencing technologies, especially ultra-long sequencing techniques, and the widespread application of multi-modal information such as single-molecule methylation and chromatin accessibility in long-read sequencing, single-molecule somatic structural variation information will play an increasingly important role in phasing and tumor evolution studies. This information will also serve as the primary basis for our future predictions of somatic structural variation function.

Methods

Evaluating Cell Line structural variants detection.

To benchmark the performance of SV detection tools, we downloaded SV calls and ONT reads from five tumor cell lines: HCC1954, H2009, HCC1937, H1437, and HCC1395. The high-confidence calls for insertions, deletions, and duplications published by previous study served as our ground truth (15). We aligned the ONT reads to hg38 using minimap2(v2.26-r1175) with the default setting for SV detection. The somatic calling tools Nanomonsv(v0.7.0), Savana(v1.2.1), and Severus(v0.1.2) 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. To examine the correctness of detected SVs, Minda(v0.0.1)(15) was employed to calculate precision, recall and F1-score between the ground truth and the call set, which is agnostic to SV types and is better suited for analysis of somatic SVs represented as junctions.

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 frequencies levels: 0.20, 0.30, 0.40,0.50and 1. Additionally, we included a control group without insertion variants and three new sequence insertion variants of different lengths: TD sequence X15, X20 and X30. The aim was to evaluate the accuracy, recall, and F1 scores of TD 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 TDs on chromosome 22 that are 1000 bp or shorter and ensured that there are no other TDs within 200 bp of each TD. This filtering resulted in a total of 6,478 candidate TD windows, with TD types including Simple_repeat and Satellite. These were divided into four files: A contains TD windows without any variations; B contains TD windows with germline variations; C contains TD windows with somatic variations; and D contains both somatic and germline variations. The positions and lengths of TDs 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) (22) (https://github.com/rrwick/Badread) to simulate sequencing at a depth of 40× for each mutation type, using the variant genomes generated by VISOR(23) 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 (v2.26-r1175) with default settings(24). The somatic calling tools Nanomonsv, Savana, 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. Sniffles utilizes the recommended sniffles2-merge process from our previous study (25), while Straglr(v1.4.1) (21) adds the -Y parameter during alignment as required by the software, followed by the same merging operations

as Sniffles. The merged BAM files are then used for Straglr with the following parameters: --genotype_in_size --min_support 3 --max_str_len 100 --max_num_clusters 9 --min_ins_size 50. The filtering process retains only the records supported by tumor reads. Tandem-genotypes(v1.9.0)(26) runs according to the software's default workflow and parameters. The results for case and control are analyzed using the Wilcoxon test and FDR correction, identifying records with at least 3 supporting reads and significant q-values as somatic.

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 1bp. 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.

TDScope methodology

Overall workflow of TDScope. TDScope initiates with identifying candidate somatic SV loci with minda workflow by merging the output result of known SV caller. After which a set of SV associated sub-sequences are extracted from reference genome and each reads of both tumor and matched normal samples that span the candidate somatic SV loci. Then, a partial-alignment graph for all sub-sequences are built with SIMID-spoa algorthm as represent of multi-sequence alignment matrix. High heterozygous alignment columns are extracted as feature for next parameter estimation of multi-category model, which clustering all sub-sequences and TDscope finally selects the clusters of somatic SV event.

Candidate somatic SV locus searching. Candidate somatic SV locus are identified by 5 SV callers, including mergeSV, nanomonSV, SVision-pro, SAVANA, and Severus. All candidate locus were merged with minda workflow with default parameters.

Local graph genome construction and feature selection. The TDscope algorithm targets all somatic structural variation (SV) candidate intervals and their flanking regions (default settings capture a 50bp 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, TDScope constructs a local graph genome for each candidate somatic SV

interval sequence set using the SIMID-sPOA algorithm 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 sPOA 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 linear genome sequences $G = \{G_1, \ldots, G_k\}$. The composition of each genome in the whole cellular population is represented by the parameter $\pi = [\pi_1, \ldots, \pi_k]$. 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 $[\theta^0_{ij}, \theta^1_{ij}, \theta^2_{ij}, \theta^3_{ij}, \theta^4_{ij}]$, same as the observed data encode, corresponding to the four DNA bases A, T, C, G, and the alignment gaps.

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

$$\begin{aligned} & P(R_r|G) \\ &= \sum_{G_r} P(R_r|G_r)P(G_r) \\ &= R_r \times (\Theta)^T \times (\pi)^T \end{aligned}$$

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

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

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

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

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

Parameter estimation. The TDscope algorithm initiates the parameter estimation process by employing hierarchical clustering to approximate the initial model parameters. Specially, TDscope

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 $(\pi^{(0)}, \Theta^{(0)})$.

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

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

Next, TDscope 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,

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

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

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,

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

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

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

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

Data and code availability

The sequencing data in this study is downloaded from SRA data base under accession number SRP494767 (BioProject: PRJNA1086849). Detail accession number for each cell lines were listed in Supplementary Table S1. Original code of TDScope and ScopVIZ are available from github at https://github.com/Goatofmountain/TDScope.

Acknowledgments

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.

Author 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 implement 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.

Competing interests

Dan Xie is a co-founder and shareholder of Qitan Tech.

Reference

1. Li, Y., Roberts, N.D., Wala, J.A., Shapira, O., Schumacher, S.E., Kumar, K., Khurana, E., Waszak, S., Korbel, J.O., Haber, J.E. *et al.* (2020) Patterns of somatic structural variation in human cancer genomes. *Nature*, **578**, 112-121.
2. Stephens, P.J., McBride, D.J., Lin, M.L., Varela, I., Pleasance, E.D., Simpson, J.T., Stebbings, L.A., Leroy, C., Edkins, S., Mudie, L.J. *et al.* (2009) Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature*, **462**, 1005-1010.
3. Wang, Z., Tu, K., Xia, L., Luo, K., Luo, W., Tang, J., Lu, K., Hu, X., He, Y., Qiao, W. *et al.* (2019) The Open Chromatin Landscape of Non-Small Cell Lung Carcinoma. *Cancer research*, **79**, 4840-4854.
4. Xu, L., Wang, X., Lu, X., Liang, F., Liu, Z., Zhang, H., Li, X., Tian, S., Wang, L. and Wang, Z. (2023) Long-read sequencing identifies novel structural variations in colorectal cancer. *PLoS Genet*, **19**, e1010514.

5. Zhang, Y., Chen, F., Fonseca, N.A., He, Y., Fujita, M., Nakagawa, H., Zhang, Z., Brazma, A., Group, P.T.W., Group, P.S.V.W. *et al.* (2020) High-coverage whole-genome analysis of 1220 cancers reveals hundreds of genes deregulated by rearrangement-mediated cis-regulatory alterations. *Nature communications*, **11**, 736.
6. Dubois, F., Sidiropoulos, N., Weischenfeldt, J. and Beroukhim, R. (2022) Structural variations in cancer and the 3D genome. *Nature reviews. Cancer*, **22**, 533-546.
7. Zhang, Y., Yang, L., Kucherlapati, M., Hadjipanayis, A., Pantazi, A., Bristow, C.A., Lee, E.A., Mahadeshwar, H.S., Tang, J., Zhang, J. *et al.* (2019) Global impact of somatic structural variation on the DNA methylome of human cancers. *Genome biology*, **20**, 209.
8. Rausch, T., Zichner, T., Schlattl, A., Stutz, A.M., Benes, V. and Korbel, J.O. (2012) DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, **28**, i333-i339.
9. Ye, K., Schulz, M.H., Long, Q., Apweiler, R. and Ning, Z. (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, **25**, 2865-2871.
10. Chen, K., Wallis, J.W., McLellan, M.D., Larson, D.E., Kalicki, J.M., Pohl, C.S., McGrath, S.D., Wendl, M.C., Zhang, Q., Locke, D.P. *et al.* (2009) BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature methods*, **6**, 677-681.
11. Dolzhenko, E., Deshpande, V., Schlesinger, F., Krusche, P., Petrovski, R., Chen, S., Emig-Agius, D., Gross, A., Narzisi, G., Bowman, B. *et al.* (2019) ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. *Bioinformatics*, **35**, 4754-4756.
12. Boeva, V., Popova, T., Bleakley, K., Chiche, P., Cappo, J., Schleiermacher, G., Janoueix-Lerosey, I., Delattre, O. and Barillot, E. (2012) Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics*, **28**, 423-425.
13. Mahmoud, M., Gobet, N., Cruz-Davalos, D.I., Mounier, N., Dessimoz, C. and Sedlazeck, F.J. (2019) Structural variant calling: the long and the short of it. *Genome biology*, **20**, 246.
14. Smolka, M., Paulin, L.F., Grochowski, C.M., Horner, D.W., Mahmoud, M., Behera, S., Kalef-Ezra, E., Gandhi, M., Hong, K., Pehlivan, D. *et al.* (2024) Detection of mosaic and population-level structural variants with Sniffles2. *Nature biotechnology*.
15. Keskus, A., Bryant, A., Ahmad, T., Yoo, B., Aganezov, S., Goretsky, A., Donmez, A., Lansdon, L.A., Rodriguez, I., Park, J. *et al.* (2024) Severus: accurate detection and characterization of somatic structural variation in tumor genomes using long reads. *medRxiv*.
16. Shiraishi, Y., Koya, J., Chiba, K., Okada, A., Arai, Y., Saito, Y., Shibata, T. and Kataoka, K. (2023) Precise characterization of somatic complex structural variations from tumor/control paired long-read sequencing data with nanomonsv. *Nucleic acids research*, **51**, e74.
17. Isidro Cortés-Ciriano, H.E., Carolin Sauer, Jose Espejo Valle-Inclan, Katherine Trevers, Melanie Tanguy, Sonia Zumalave, Solange De Noon, Francesc Muyas, Rita Cascao, Angela Afonso, Fernanda Amary, Roberto Tirabosco, Adam Giess, Timothy Freeman, Alona Sosinsky, Katherine Piculell, David Miller, Claudia Faria, Greg Elgar, Adrienne Flanagan. (2024), Research Square.
18. Wang, S., Lin, J., Jia, P., Xu, T., Li, X., Liu, Y., Xu, D., Bush, S.J., Meng, D. and Ye, K. (2024) De novo and somatic structural variant discovery with SVision-pro. *Nature biotechnology*.
19. Sahlin, K., Baudeau, T., Cazaux, B. and Marchet, C. (2023) A survey of mapping algorithms in the long-reads era. *Genome biology*, **24**, 133.

20. Lee, C., Grasso, C. and Sharlow, M.F. (2002) Multiple sequence alignment using partial order graphs. *Bioinformatics*, **18**, 452-464.
21. Chiu, R., Rajan-Babu, I.S., Friedman, J.M. and Birol, I. (2021) Straglr: discovering and genotyping tandem repeat expansions using whole genome long-read sequences. *Genome biology*, **22**, 224.
22. Wick, R.R. (2019) Badread: simulation of error-prone long reads. *Journal of Open Source Software*, **4**, 1316.
23. Bolognini, D., Sanders, A., Korbel, J.O., Magi, A., Benes, V. and Rausch, T. (2020) VISOR: a versatile haplotype-aware structural variant simulator for short- and long-read sequencing. *Bioinformatics*, **36**, 1267-1269.
24. Li, H. (2021) New strategies to improve minimap2 alignment accuracy. *Bioinformatics*, **37**, 4572-4574.
25. Dan Xie, L.X., Zhoufeng Wang, Tianfu Zeng, Xuenan Pi, Huan Wang, Guonian Zhu, Xinyue Wu, Yangqian Li, Yan Deng, Yawen Qi, Xuyan liu, Fengmei Zhang, Weimin Li. (2023), bioRxiv.
26. Mitsuhashi, S., Frith, M.C., Mizuguchi, T., Miyatake, S., Toyota, T., Adachi, H., Oma, Y., Kino, Y., Mitsuhashi, H. and Matsumoto, N. (2019) Tandem-genotypes: robust detection of tandem repeat expansions from long DNA reads. *Genome biology*, **20**, 58.