Additional file 2

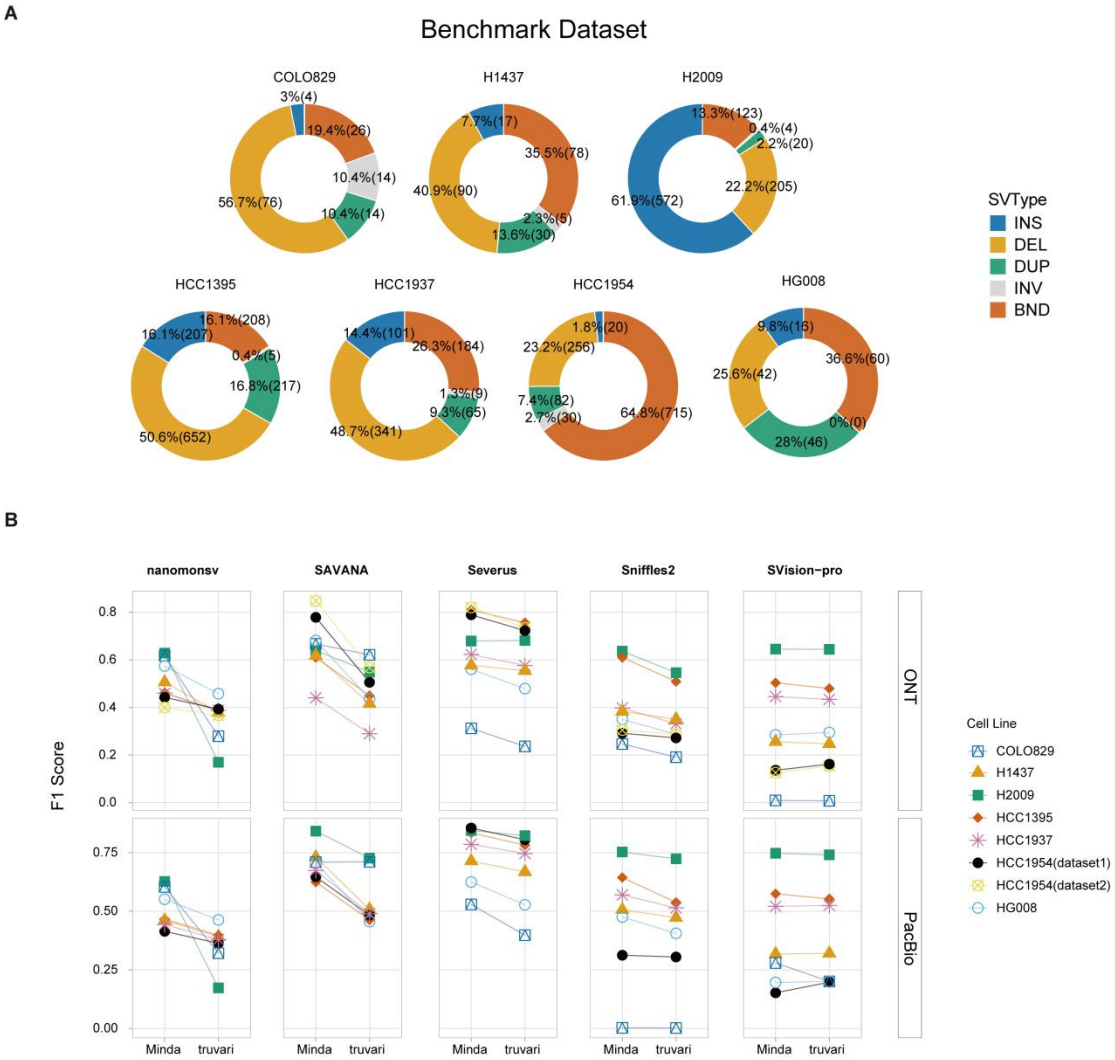

Fig. S1. Construction workflow and structural variation distribution of benchmark datasets.
A. Pie charts depicting the distribution of somatic SV types across seven cell line benchmark datasets. Data for H1437, HCC1395, HCC1937, H2009, and HCC1954 were generated through multi-omics voting consensus, while COLO829 and HG008 datasets were derived from experimentally validated results in published literature (see Methods).
B. F1 score comparison of five somatic SV detection callers across the eight benchmark datasets (COLO829, H1437, H2009, HCC1395, HCC1937, HCC1954 [dataset 1 & 2], and HG008). Performance is evaluated separately on ONT (top panel) and PacBio (bottom panel) sequencing data using two different benchmarking tools: Minda (junction-based evaluation) and Truvari (type-specific evaluation). Results are shown for nanomonsv, SAVANA, Severus, Sniffles2, and SVision-pro.

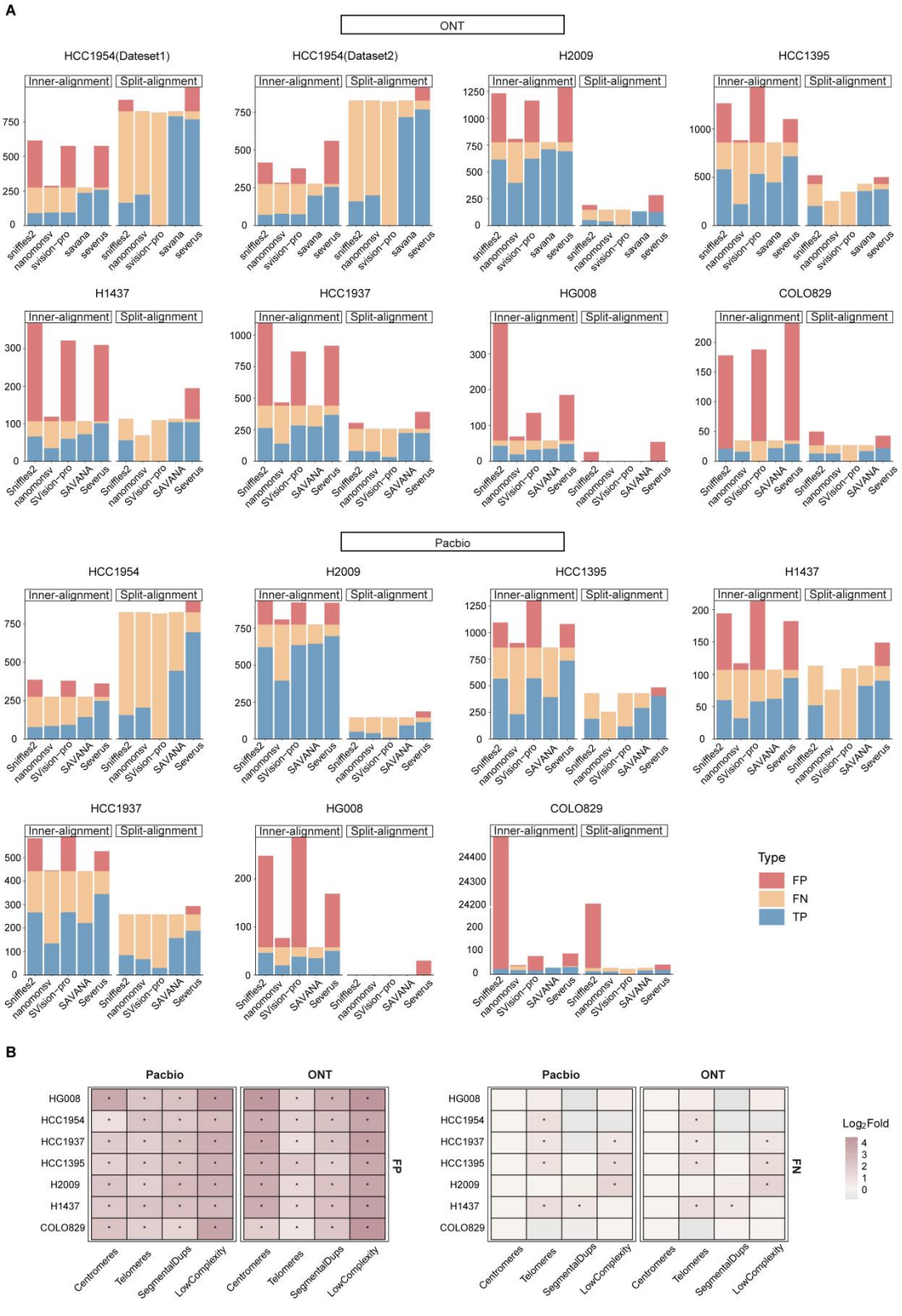
A
ONT
HCC1954(Dateset1)
HCC1954(Dataset2)
H2009
HCC1395
H1437
HCC1937
HG008
COLO829
Inner-alignment
Split-alignment
Sniffles2
nanomonsv
SVision-pro
SAVANA
Severus
Pacbio
HCC1954
H2009
HCC1395
H1437
HCC1937
HG008
COLO829
Type
FP
FN
TP
B
Pacbio
ONT
HG008
HCC1954
HCC1937
HCC1395
H2009
H1437
COLO829
FP
FN
Centromeres
Telomeres
SegmentalDups
LowComplexity
Log2Fold

Fig. S2. Performance characterization of somatic SV detection algorithms across different SV types and genomic regions.

A. Error distribution of somatic SV detection calls by SV type and sequencing platform. Bar plots show the count of True Positives (TP, blue), False Negatives (FN, yellow), and False Positives (FP, red) for five somatic SV callers (nanomonsv, Sniffles2, SAVANA, Severus, and SVision-pro) across eight benchmark datasets. Results are categorized by sequencing platform (ONT vs. PacBio) and SV type (Inner-alignment SVs vs. Split-alignment SVs).

B. Enrichment of False Positives (FP, left) and False Negatives (FN, right) across challenging genomic regions. Heatmaps illustrate the Log2 Fold enrichment of FP and FN events in four challenging genomic regions (Telomeres, Centromeres, Segmental Duplications, and Tandem Repeat & Low Complexity) for each cell line. Enrichment is calculated separately for PacBio and ONT sequencing platforms. Color intensity (darker red) corresponds to a higher Log2 Fold enrichment value.

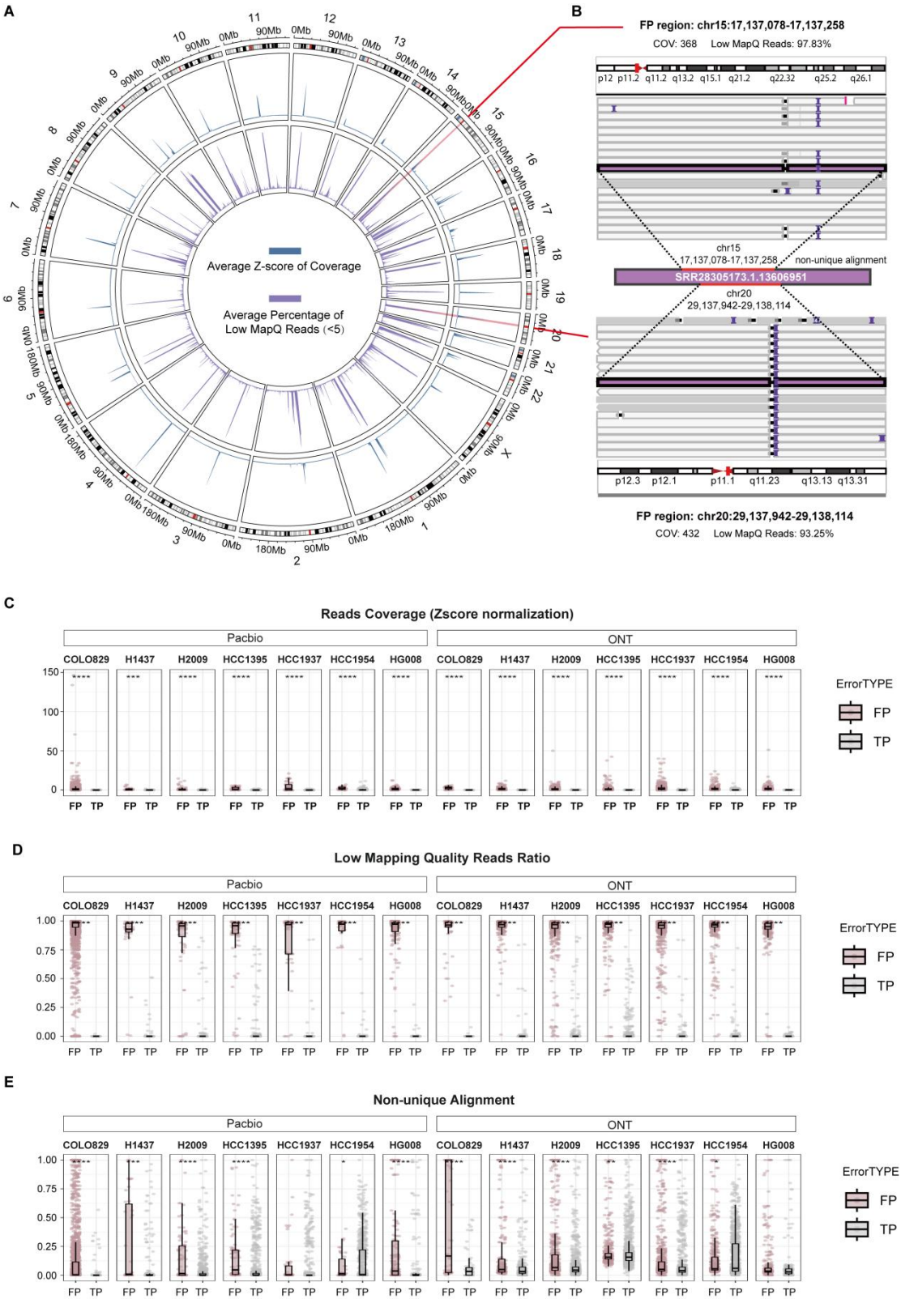
A
B
C
D
E
Average Z-score of Coverage
Average Percentage of Low MapQ Reads (<5)
FP region: chr15:17,137,078-17,137,258
COV: 368 Low MapQ Reads: 97.83%
chr15
17,137,078-17,137,258
non-unique alignment
SRR28305173.1.13606951
chr20
29,137,942-29,138,114
FP region: chr20:29,137,942-29,138,114
COV: 432 Low MapQ Reads: 93.25%
Reads Coverage (Zscore normalization)
Low Mapping Quality Reads Ratio
Non-unique Alignment
Pacbio
ONT
COLO829
H1437
H2009
HCC1395
HCC1937
HCC1954
HG008
FP
TP
ErrorTYPE

Fig. S3. Genome-wide coverage, mapping quality and alignment ambiguity performance.

A. Circle plot of chromosomes 1-22 and X displays three concentric layers: the outermost chromosomal karyotype bands, the middle blue line plot showing mean coverage Z-scores, and the innermost purple line plot indicating the mean proportion of low mapping quality (MAPQ <5) reads.

B. IGV visualization of a non-unique alignment in H2009 cell line demonstrates a read simultaneously mapped to chr15:17,137,078-17,137,258 and chr20:29,137,942-29,138,114.

C. Normalized read coverage distribution associated with False Positives (FP, red) and True Positives (TP, gray). Boxplots and scatter plots show the Z-score normalized read coverage in regions surrounding FP and TP calls across 15 benchmark datasets, separated by sequencing platform (PacBio vs. ONT).

D. Low mapping quality reads ratio associated with FP and TP. Boxplots and scatter plots show the ratio of low-mapping-quality reads within regions surrounding FP and TP calls across the 15 benchmark datasets, separated by sequencing platform.

E. Non-unique alignment ratio associated with FP and TP. Boxplots and scatter plots show the ratio of non-uniquely aligned reads within regions surrounding FP and TP calls across the 15 benchmark datasets, separated by sequencing platform. Asterisks in panels C, D and E indicate significance levels from the Mann-Whitney U Test, with thresholds defined as * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$), **** ($p < 0.0001$)

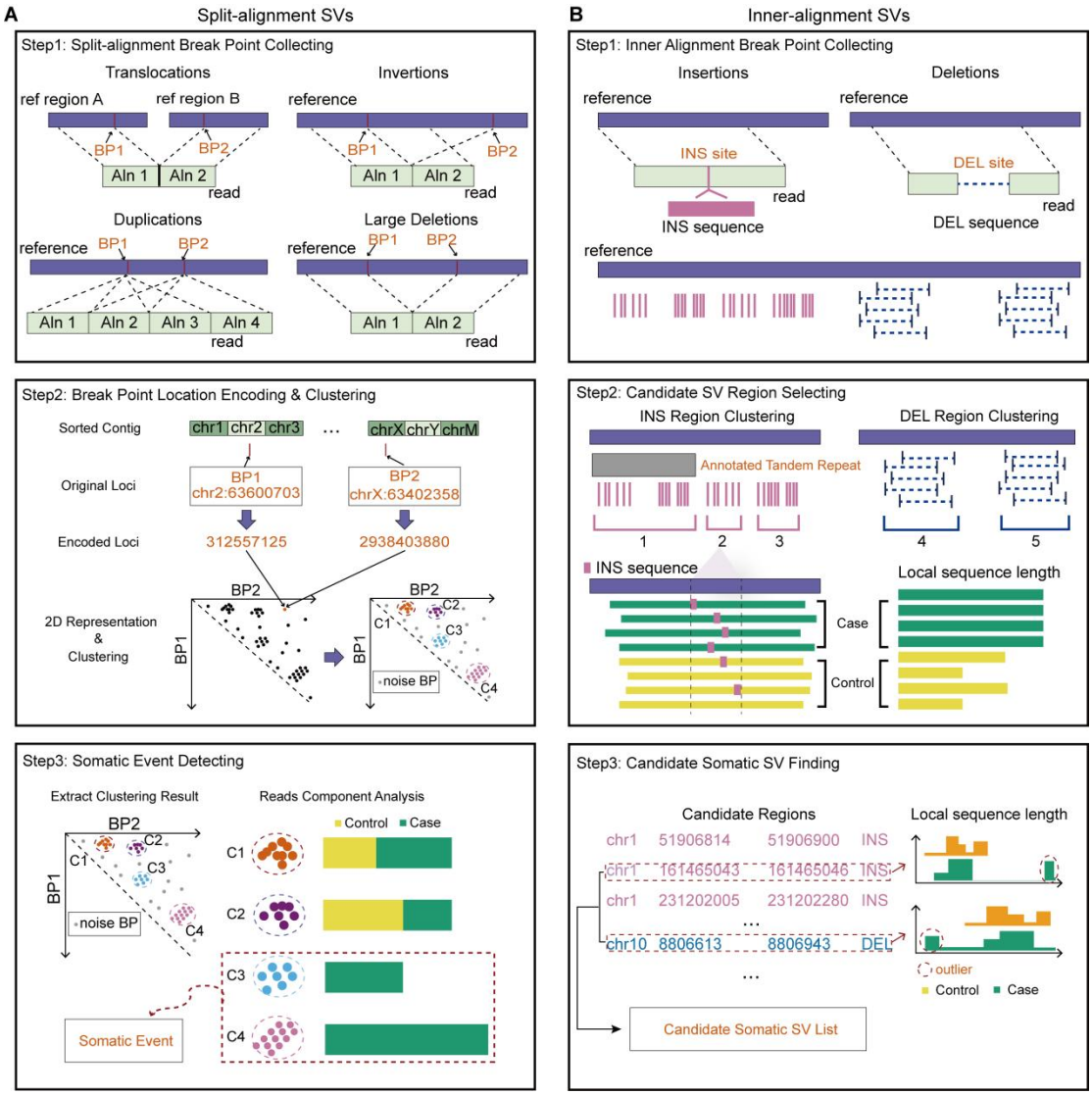

Fig. S4. Detailed Schematic of the candidate somatic SV selection workflow.

A. Split-alignment SVs are selected based on paired-breakpoint DBScan clustering methods.

B. Inner-alignment SVs are divided into insertion and deletions. Breakpoints of insertions are merged and located within the same tandem repeat regions or within the length of distance threshold (by default 200bp). Breakpoints of deletions are directly merged if containing overlap. Full-length alignment information of each read from case and control samples is extracted to filter candidate inner-alignment SVs. Intervals with outlier reads more than 3 would be considered as candidate somatic intervals.

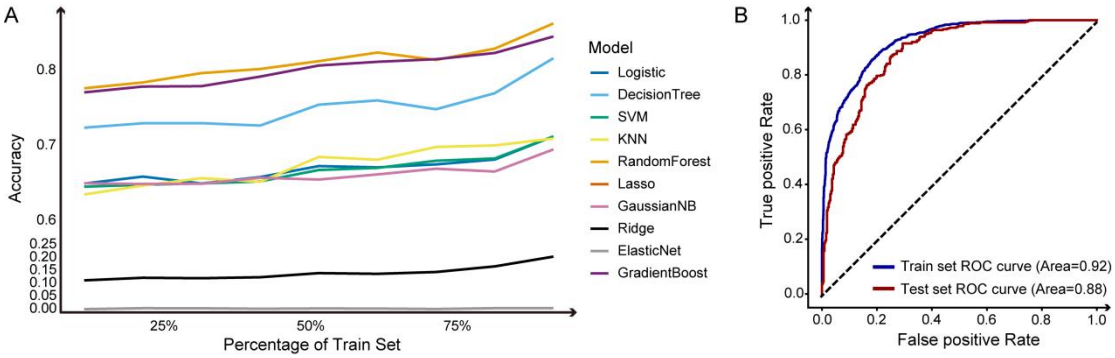

Fig. S5. Machine learning model performance evaluation.

A. Accuracy trend lines of ten machine learning methods: Logistic Regression, Decision Tree, Support Vector Machine (SVM), k-Nearest Neighbors (KNN), Random Forest, Lasso Regression, Gaussian Naive Bayes (GaussianNB), Ridge Regression, ElasticNet, and Gradient Boosting. Each method is represented by a distinct color.

B. ROC curves of the Random Forest model. Blue curve: training set performance; red curve: test set performance.

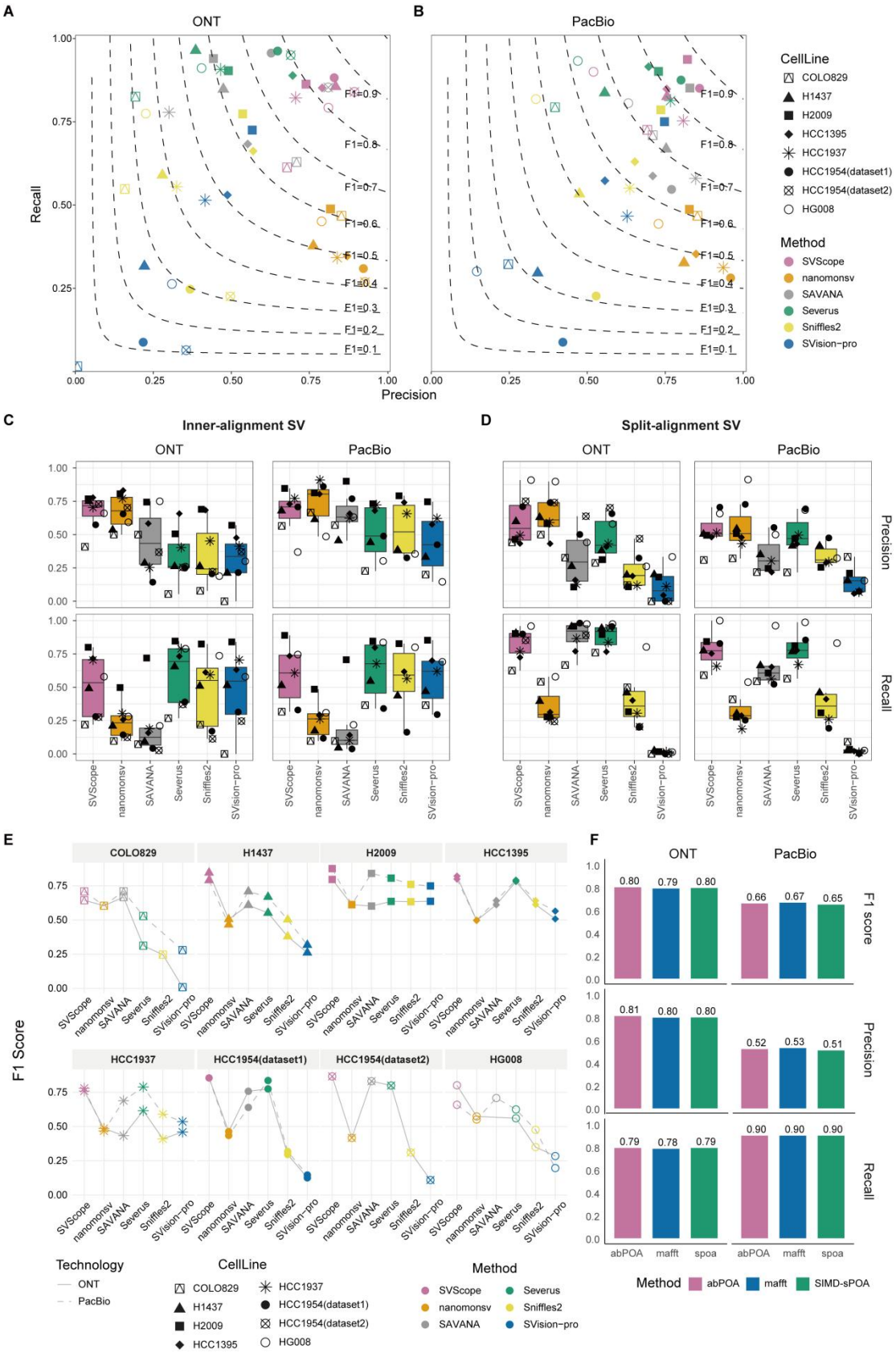
A
ONT
B
PacBio
Recall
Precision
F1=0.9
F1=0.8
F1=0.7
F1=0.6
F1=0.5
F1=0.4
F1=0.3
F1=0.2
F1=0.1
CellLine
COLO829
H1437
H2009
HCC1395
HCC1937
HCC1954(dataset1)
HCC1954(dataset2)
HG008
Method
SVScope
nanomonsv
SAVANA
Severus
Sniffles2
SVision−pro
C
Inner-alignment SV
D
Split-alignment SV
Precision
Recall
E
F1 Score
Technology
ONT
PacBio
F
F1 score
Precision
Recall
abPOA
mafft
spoa
SIMD-sPOA

Fig. S6. Performance comparison of SVScope and five other methods across seven cell lines.

A-B. Precision-Recall curves for all somatic SVs. Scatter plots showing Precision (X-axis) versus Recall (Y-axis) performance for six somatic SV callers (SVScope, nanomonsv, SAVANA, Severus, Sniffles2, and SVision-pro) across eight benchmark datasets, evaluating all somatic SVs. Results are separated by sequencing platform: ONT (A) and PacBio (B). Shapes denote cell lines, colors distinguish methods, and dashed contour bands indicate varying F1 score levels (from 0.1 to 0.9).
C-D. Precision and Recall comparison by SV types. Boxplots summarizing the distribution of Precision and Recall for inner-alignment SVs (C) and split-alignment SVs (D), separated by sequencing platform (ONT vs. PacBio). Shapes denote cell lines, and colors distinguish methods.
E. F1 score comparison across individual cell lines. Line chart displaying the F1 scores achieved by SVScope and five competitor methods across the eight benchmark datasets. Solid lines represent ONT datasets, and dashed lines represent PacBio datasets. Shapes correspond to cell lines, and colors distinguish the methods. Cell lines are categorized based on their sequencing platform (ONT vs. PacBio).
F. Impact of Multiple Sequence Alignment (MSA) algorithm on SVScope performance. Bar plots comparing the F1 score, Precision, and Recall of SVScope when utilizing different core MSA algorithms (abPOA, mafft, and SIMD-sPOA), assessed separately on ONT and PacBio sequencing platforms.

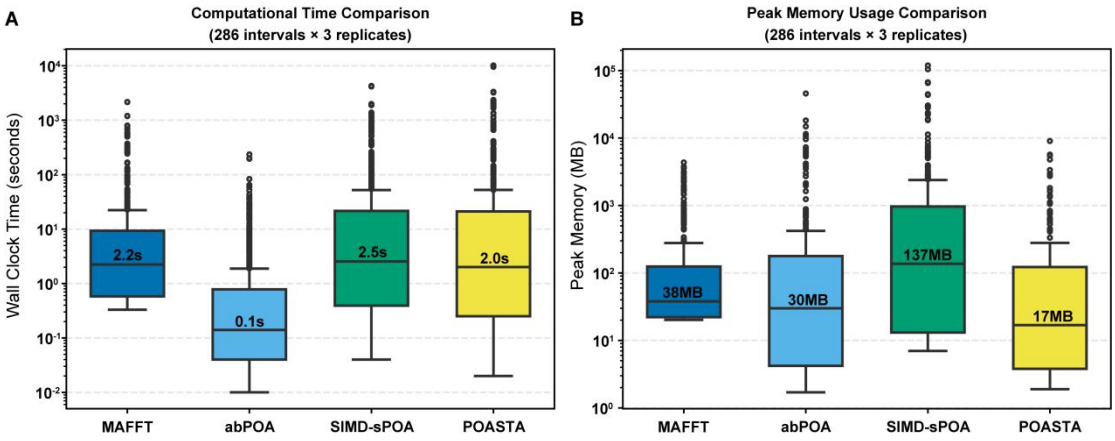

Fig. S7. Performance benchmarking of isolated Multiple Sequence Alignment (MSA) aligners. Box plots illustrating the (A) wall clock time in seconds and (B) peak memory usage in MB. The benchmarking was conducted using 286 candidate SV intervals from HG008 ONT data. Four MSA aligners—MAFFT, abPOA, SIMD-sPOA, and POASTA—were evaluated, with each SV interval processed in triplicate. Median values for wall clock time and peak memory usage are labeled within each box.

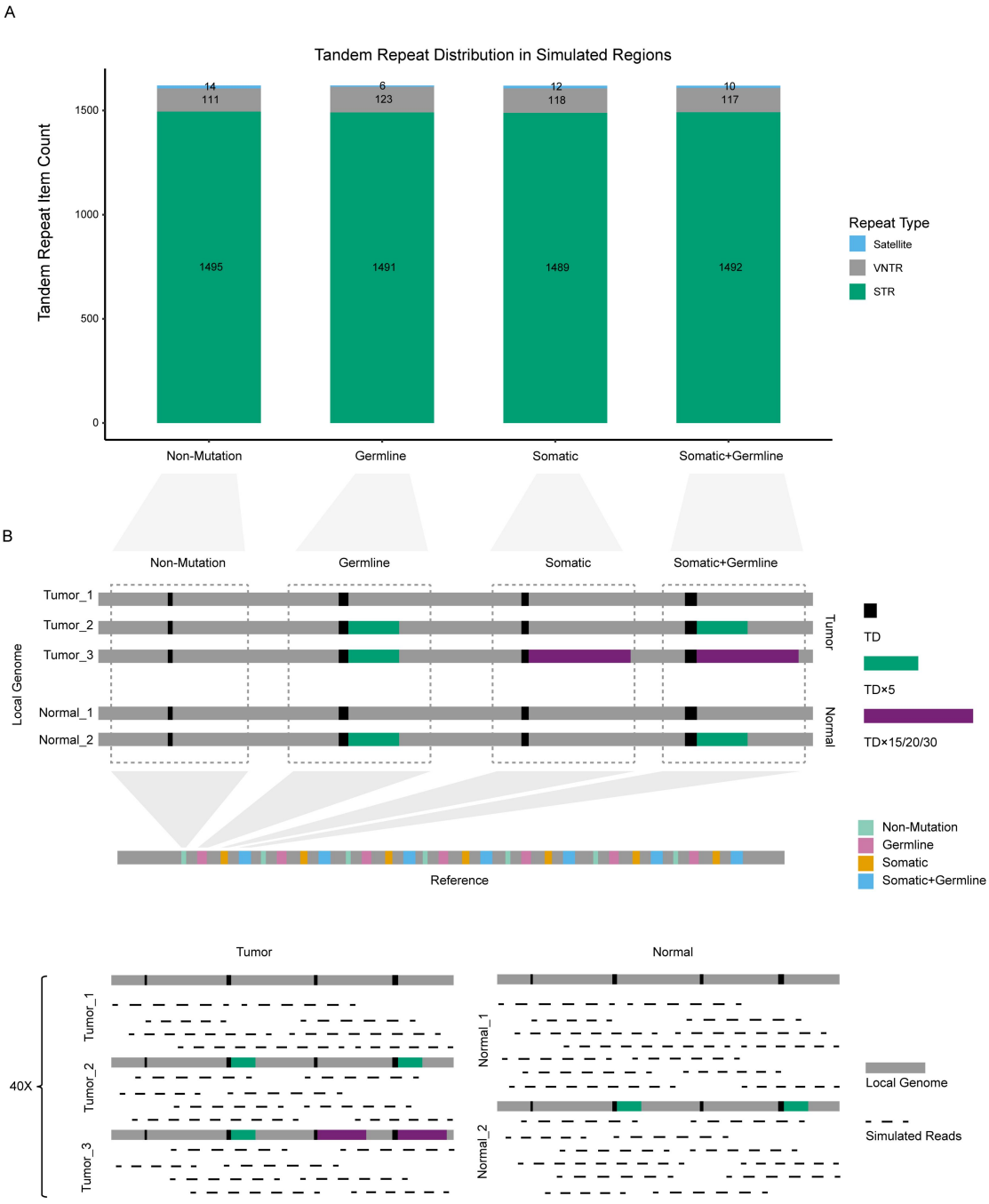
A
Tandem Repeat Distribution in Simulated Regions
Tandem Repeat Item Count
1500
1000
500
0
14
111
1495
6
123
1491
12
118
1489
10
117
1492
Non-Mutation
Germline
Somatic
Somatic+Germline
Repeat Type
Satellite
VNTR
STR
B
Non-Mutation
Germline
Somatic
Somatic+Germline
Local Genome
Tumor_1
Tumor_2
Tumor_3
Normal_1
Normal_2
Tumor
Normal
TD
TD×5
TD×15/20/30
Reference
Non-Mutation
Germline
Somatic
Somatic+Germline
Tumor
Normal
40X
Tumor_1
Tumor_2
Tumor_3
Normal_1
Normal_2
Local Genome
Simulated Reads

Fig. S8. Construction of the haplotype-aware simulated dataset for tandem repeat evaluation.

A. Tandem repeat type distribution across simulated genomic regions. Stacked bar charts showing the count and proportion of different tandem repeat types (STR, VNTR, and Satellite) within the four categories of simulated regions: Non-Mutation, Germline, Somatic, and Somatic + Germline. The total number of tandem repeat items in each category is displayed on the bars.

B. Workflow for simulating heterogeneous tumor-normal pairs and long-read data. Schematic diagram illustrating the simulation process:

Local Genome States (Top): Shows the structural status of the target tandem repeat region across five local genomes (Tumor_1, Tumor_2, Tumor_3, Normal_1, Normal_2). These states reflect the four major simulated conditions: Non-mutation, Germline (TD * 5 units), Somatic (TD * 15/20/30 units), and Somatic + Germline (Germline TD * 5 units, somatic TD * 15/20/30 units).

Reference Region Selection (Middle): The color-coded bar of the reference genome illustrates the four simulated region types: Non-mutation (green), Germline (pink), Somatic (orange), and Mixed Somatic/Germline (light blue), which were extracted from hg38 RepeatMasker annotation files to ensure a realistic genomic context for simulation.

Simulated Reads Generation (Bottom): Long reads are simulated from the respective local genomes. Tumor purity levels (20%, 30%, 40%, 50%, 100%) are achieved by controlling the sequencing depth ratio of the simulated reads derived from the tumor local genomes versus the normal local genomes, and then merging the combined reads.

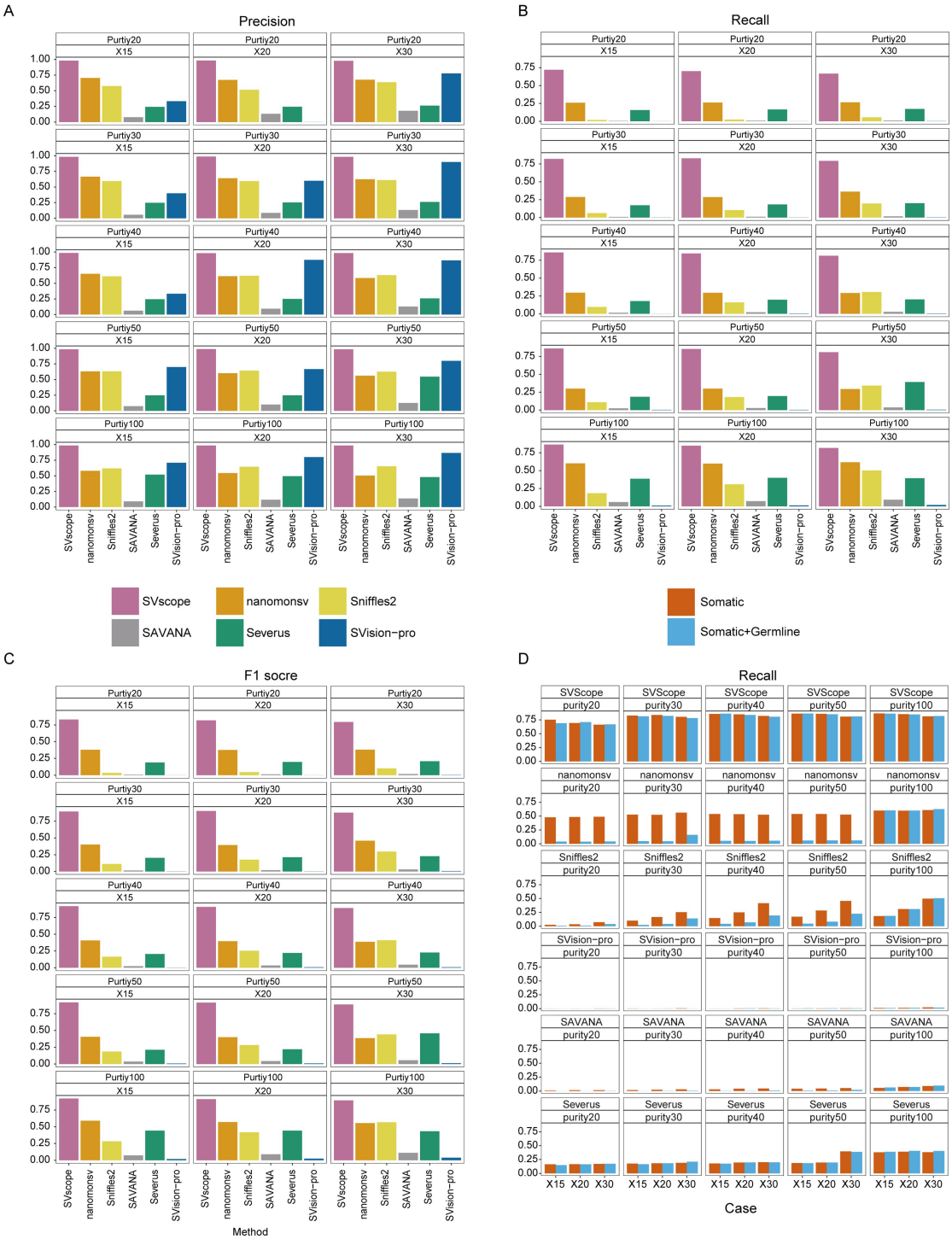

Fig. S9. Performance evaluation of SV detection methods on simulated data.

A-C. Bar plots compare Precision (A), Recall (B), and F1-score (C) across six methods: Sniffles2 (yellow), nanomonsv (orange), SVision-pro (green), SAVANA (gray), Severus (blue), and SVScope (pink). X-axis: methods; Y-axis: score percentages. Results are stratified by case (X15/X20/X30) and purity (20%, 30%, 40%, 50%, 100%).

D. Recall comparison for Somatic (red) and Mixed Somatic/Germline (blue) across methods, cases, and purity levels. X-axis: methods; Y-axis: recall.

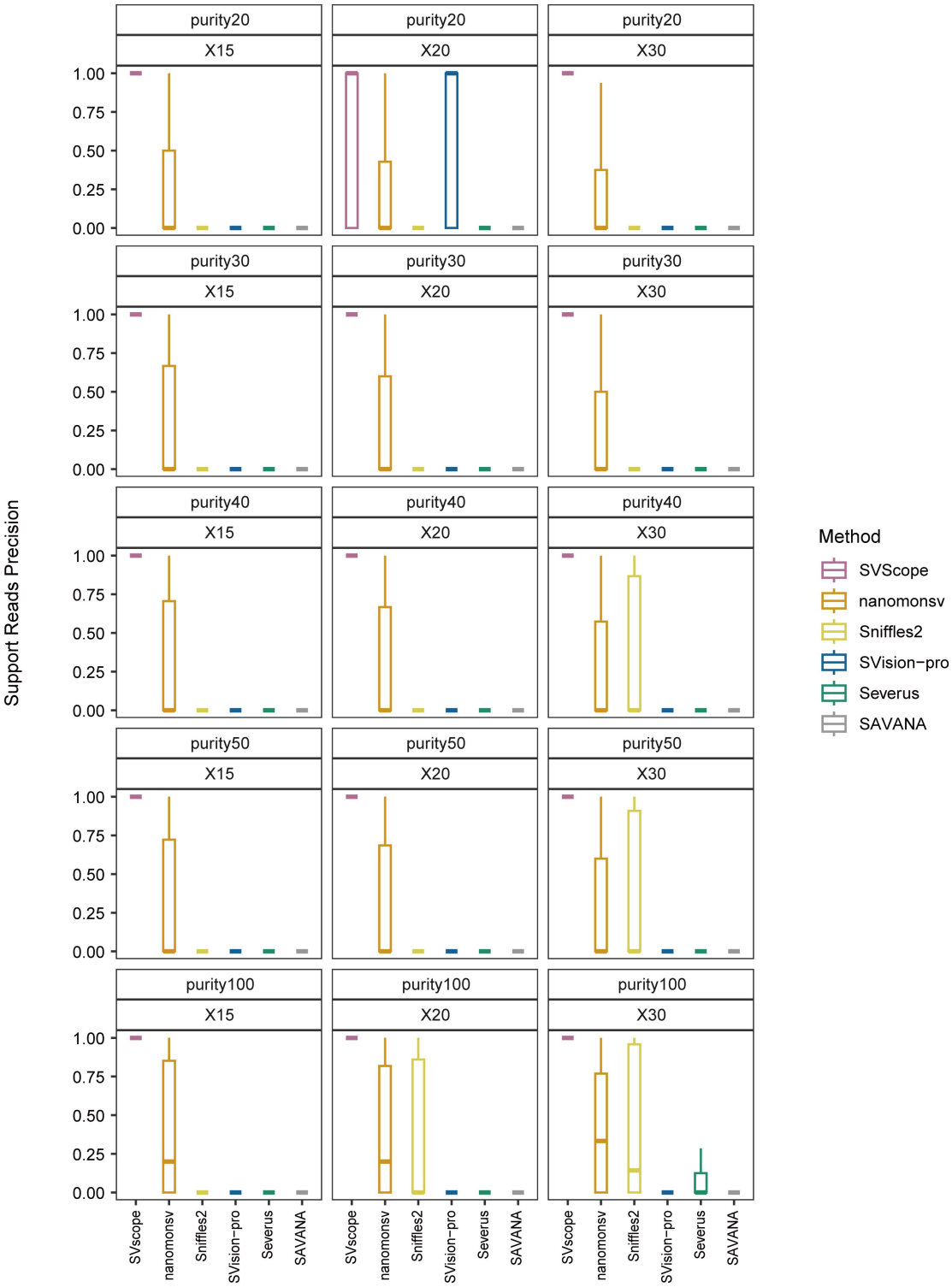

Fig. S10. Support reads precision of SV detection methods on simulated data. Boxplots compare support reads precision (Y-axis: score) for six methods: Sniffles2 (yellow), nanomonsv (orange), SVision-pro (blue), SAVANA (gray), Severus (dark green), and SVScope (pink). X-axis labels represent methods, with box distributions aggregating results across three simulated cases (X15, X20, X30) and five tumor purity levels (20%, 30%, 40%, 50%, 100%).

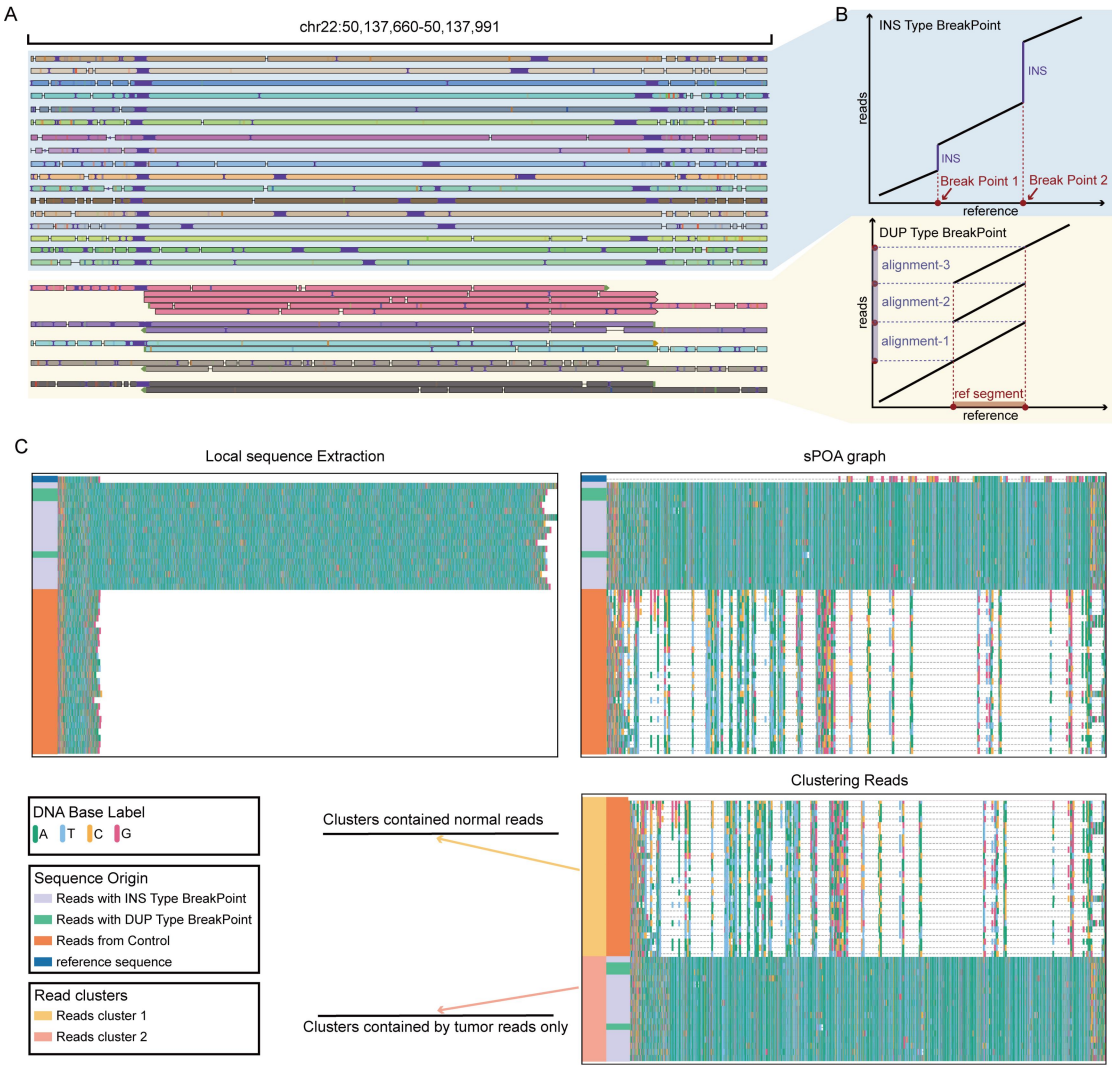

Fig. S11. Representation of SVScope in heterogenous alignment breakpoint somatic SV case on chr22:50,137,660-50,137,991.

A. IGV screen for simulated reads harboring somatic insertions in chr22:50,137,660-50,137,991. B. Line plot representing the alignment detail of INS type breakpoint (blue) and DUP type breakpoint (yellow). C. ScopeVIZ plot representing the process of local graph-genome optimization process of SVScope.

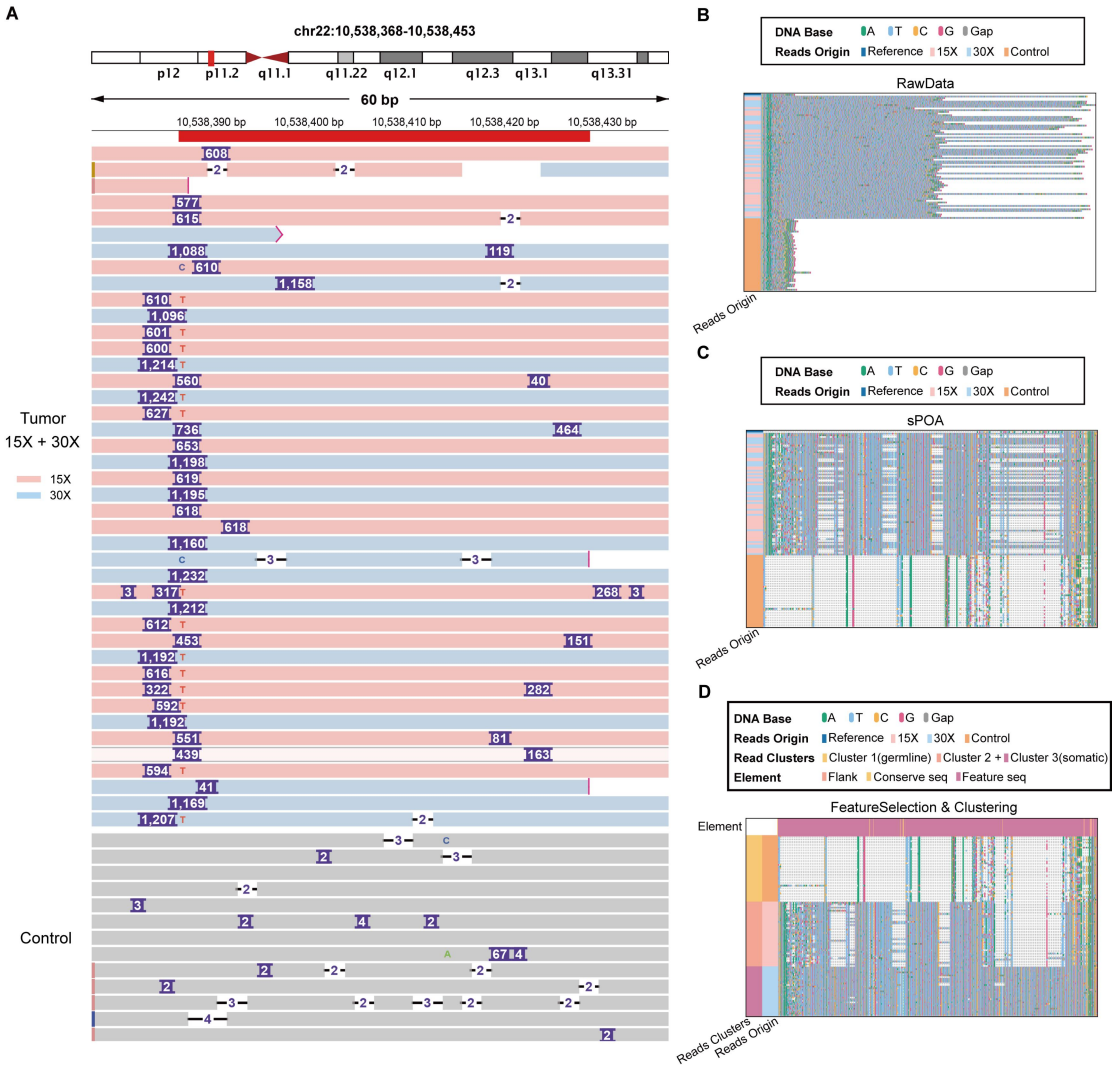

Fig. S12. Integrated visualization and ScopeVIZ analysis of simulated structural variations in the chr22:10,538,368-10,538,453 region.

A. IGV visualization of tumor (upper panel) and normal (lower panel) read alignments in the chr22:10,538,368-10,538,453 region. Tumor reads are a mixture of simulated X15 (pink) and X30 (blue) tandem duplications; normal reads (orange) lack these alterations.

B. Sequence-level view of reads: Nucleotides are color-coded (A=green, T=cyan, C=yellow, G=magenta; gaps=gray dashes). Read origins are labeled (X15=pink, X30=blue, normal=orange).

C. Partial order alignment graph generated by sparse partial order alignment (POA) from raw reads.

D. SVScope-clustered reads: Top track classifies bases as Feature (magenta), Common (yellow), or Flank (peach). Left labels indicate SVScope-defined clusters (1/2/3) and read origins (X15/X30/normal).

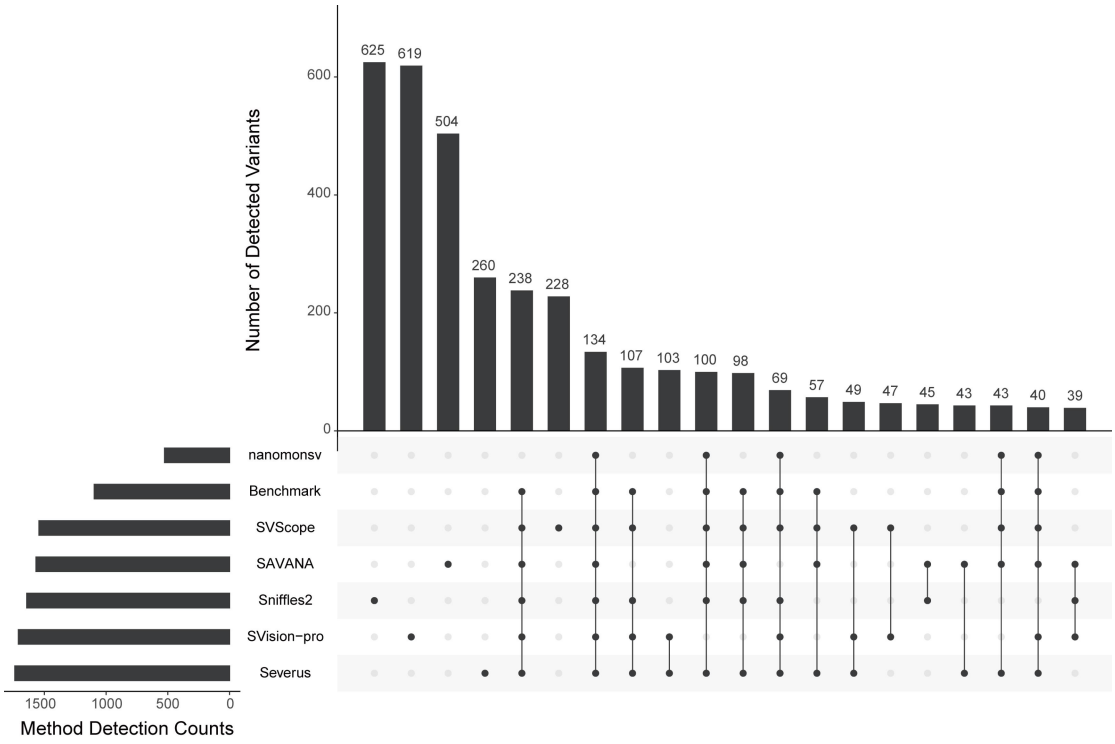

Fig. S13. Overlap of structural variants (SVs) detected in the HCC1395 cell line. The upset plot displays intersections among SV calls from SVScope, five methods (Sniffles2, nanomonsv, SVision-pro, SAVANA, Severus), and a benchmark dataset. The left bar plot shows the total SV counts per method and the benchmark dataset. The upper bar plot ranks the top 20 largest intersections by size (number of overlapping SVs). The central matrix uses dots to indicate method/dataset participation in each intersection (rows: methods/dataset; columns: intersection sets).

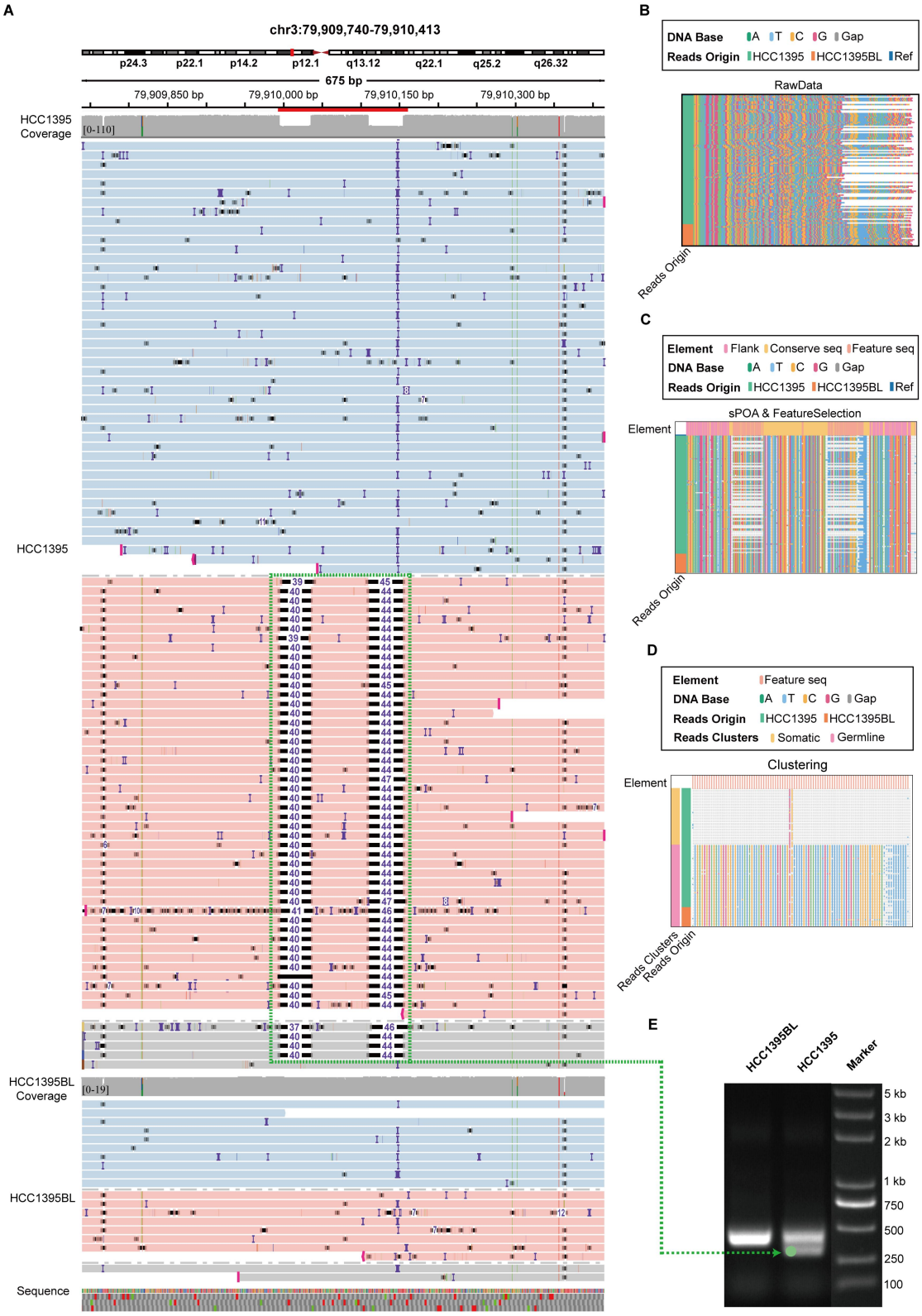
A
chr3:79,909,740-79,910,413
p24.3
p22.1
p14.2
p12.1
q13.12
q22.1
q25.2
q26.32
675 bp
79,909,850 bp
79,910,000 bp
79,910,150 bp
79,910,300 bp
HCC1395 Coverage
[0-110]
HCC1395
HCC1395BL Coverage
[0-19]
HCC1395BL
Sequence
B
DNA Base A T C G Gap
Reads Origin HCC1395 HCC1395BL Ref
RawData
Reads Origin
C
Element Flank Conserve seq Feature seq
DNA Base A T C G Gap
Reads Origin HCC1395 HCC1395BL Ref
sPOA & FeatureSelection
Element
Reads Origin
D
Element Feature seq
DNA Base A T C G Gap
Reads Origin HCC1395 HCC1395BL
Reads Clusters Somatic Germline
Clustering
Element
Reads Clusters
Reads Origin
E
HCC1395BL
HCC1395
Marker
5 kb
3 kb
2 kb
1 kb
750
500
250
100

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.

A. Genome browser plot shows the example somatic SV region chr3:79,909,993-79,910,160 in HCC1395 (top) and HCC1395BL (bottom).

B. Bar plot shows the raw sequence extracted from HCC1395BL (colored in orange) and HCC1395 reads (colored in green).

C. Bar plot shows multiple sequence alignment representation of POA graph based on read sequences from HCC1395BL (colored in orange) and HCC1395 reads (colored in green).

D. SVScope-clustered reads: Top track classifies bases as Feature (magenta), Common (yellow), or Flank (peach). Left labels indicate SVScope-defined clusters (1/2) and reads origins (tumor/normal).

E. The gel electrophoresis image illustrates the targeted PCR results for the somatic SV event. Notably, the target bands are highlighted with green dots, clearly indicating the presence of the somatic SV event in both genomes.