nature biotechnology

Supplementary information

https://doi.org/10.1038/s41587-026-03332-9

Structural variant detection across sequencing technologies

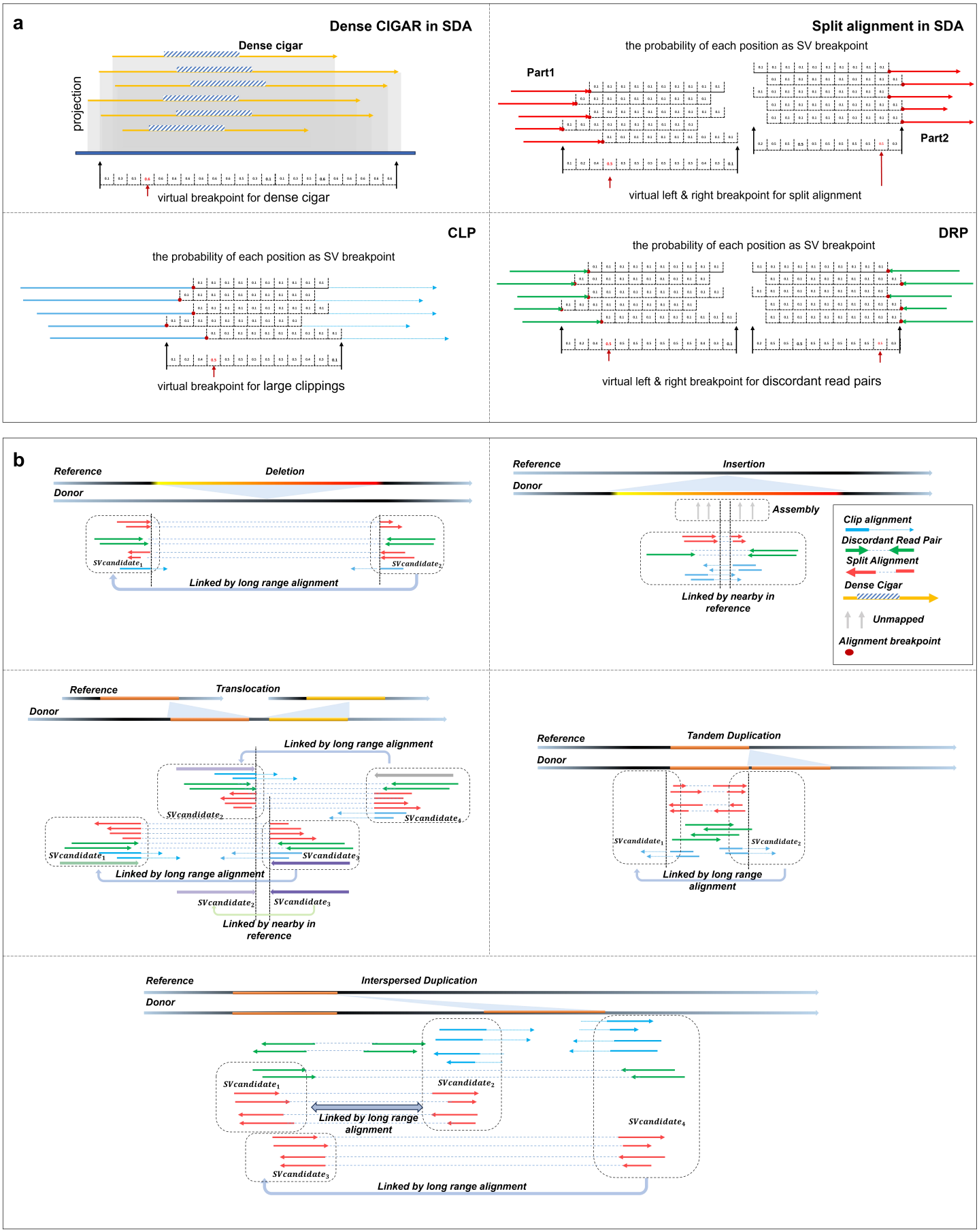
In the format provided by the authors and unedited

Structural variant detection across sequencing technologies

Supplementary Material

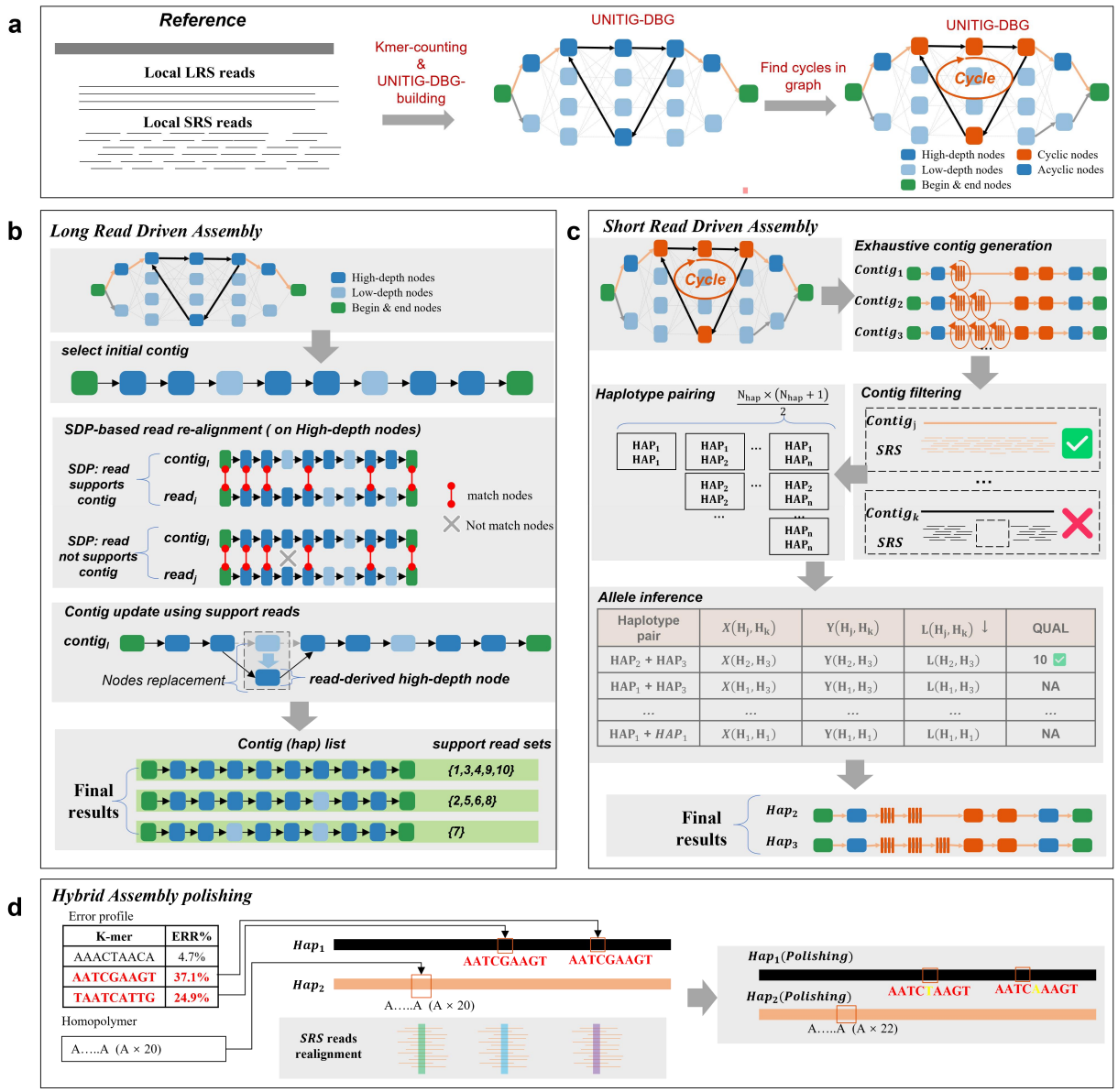
Supplementary Fig. 1. A schematic illustration on refined two-phase SV signature clustering 5
Supplementary Fig. 2. A schematic illustration of local haplotype sequence reconstruction and polishing 6
Supplementary Fig. 3. A schematic illustration on the successful of SV allele reconstruction and genotyping with evident long reads. 8
Supplementary Fig. 4. Allele counts of population-specific SVs in HitSV-callset and 1KGP-callset 10
Supplementary Fig. 5. Tandem repeat-associated SVs in the HitSV-callset 11
Supplementary Fig. 6. Phenotypic annotation of SVs in the 1KGP-callset and HitSV-callset 12
Supplementary Fig. 7. An example of CSV detection in a HG002 dataset. 13
Supplementary Fig. 8. An example of CSV detection in a family-based dataset. 14
Supplementary Fig. 9. An example of CSV detection in a HG002 dataset. 15
Supplementary Fig. 10. An example of CSV detection in a HG002 dataset. 16
Supplementary Fig. 11. An example of CSV detection in a family-based dataset. 17
Supplementary Fig. 12. An example of CSV detection in a family-based dataset. 18
Supplementary Fig. 13. An example of CSV detection in a family-based dataset. 19
Supplementary Fig. 14. An example of CSV detection in a family-based dataset. 20
Supplementary Fig. 15. Detection of complex structural variants across multiple ONT samples. 21
Supplementary Fig. 16. Chromosome-wide distribution and classification of CSV regions in HPRC datasets 22
Supplementary Fig. 17. Effects of conservative quality-control filters on CHN100K-HPRC CSV comparison. 23
Supplementary Fig. 18. Experimental validation for a subset of CHN100K CSV alleles. 32
Supplementary Fig. 19. Comparison of clustering complexity for CSV regions in the ONT-HPRC datasets. 33
Supplementary Fig. 20. Detection of TR-MEI–associated complex structural variants across multiple ONT samples. 34
Supplementary Fig. 21. Detection of TR-MEI–associated complex structural variants across multiple ONT samples. 35
Supplementary Fig. 22. Repeat unit count and length variability of tandem repeat (TR) regions. 36
Supplementary Fig. 23. Haplotype-resolved characterization of a TR-MEI array at the *TLCD5* locus. 37
Supplementary Fig. 24. A CSV reconstructed by HitSV but hindered by contig re-alignment. 38
Supplementary Note 1. The characterization of the signature distributions 39
Supplementary Note 2. Comparison between HitSV and Pangenome-based SV callers on SRS data 43
Supplementary Note 3. Runtime Environment and Command lines 45
Supplementary Note 4. The characterization of the false negatives and positives of HitSV 49
Supplementary Note 5. External support for HitSV-specific SVs in the 1KGP cohort 55
Supplementary Note 6. Analytical workflow used for the CHN100K–HPRC comparison 57

Supplementary Note 7. The SV-call comparison between the distribution-modeling- and other 1KGP samples 61
Supplementary Note 8. SV genotyping and quality control 63


a
Dense CIGAR in SDA
Dense cigar
projection
virtual breakpoint for dense cigar
Split alignment in SDA
the probability of each position as SV breakpoint
Part1
Part2
virtual left & right breakpoint for split alignment
CLP
the probability of each position as SV breakpoint
virtual breakpoint for large clippings
DRP
the probability of each position as SV breakpoint
virtual left & right breakpoint for discordant read pairs
b
Reference
Deletion
Donor
SVcandidate1
SVcandidate2
Linked by long range alignment
Reference
Insertion
Donor
Assembly
Linked by nearby in reference
Clip alignment
Discordant Read Pair
Split Alignment
Dense Cigar
Unmapped
Alignment breakpoint
Reference
Translocation
Donor
Linked by long range alignment
SVcandidate2
SVcandidate4
SVcandidate1
SVcandidate3
Linked by long range alignment
SVcandidate2
SVcandidate3
Linked by nearby in reference
Reference
Tandem Duplication
Donor
SVcandidate1
SVcandidate2
Linked by long range alignment
Reference
Interspersed Duplication
Donor
SVcandidate1
SVcandidate2
Linked by long range alignment
SVcandidate4
SVcandidate3
Linked by long range alignment

Supplementary Fig. 1. A schematic illustration on refined two-phase SV signature clustering

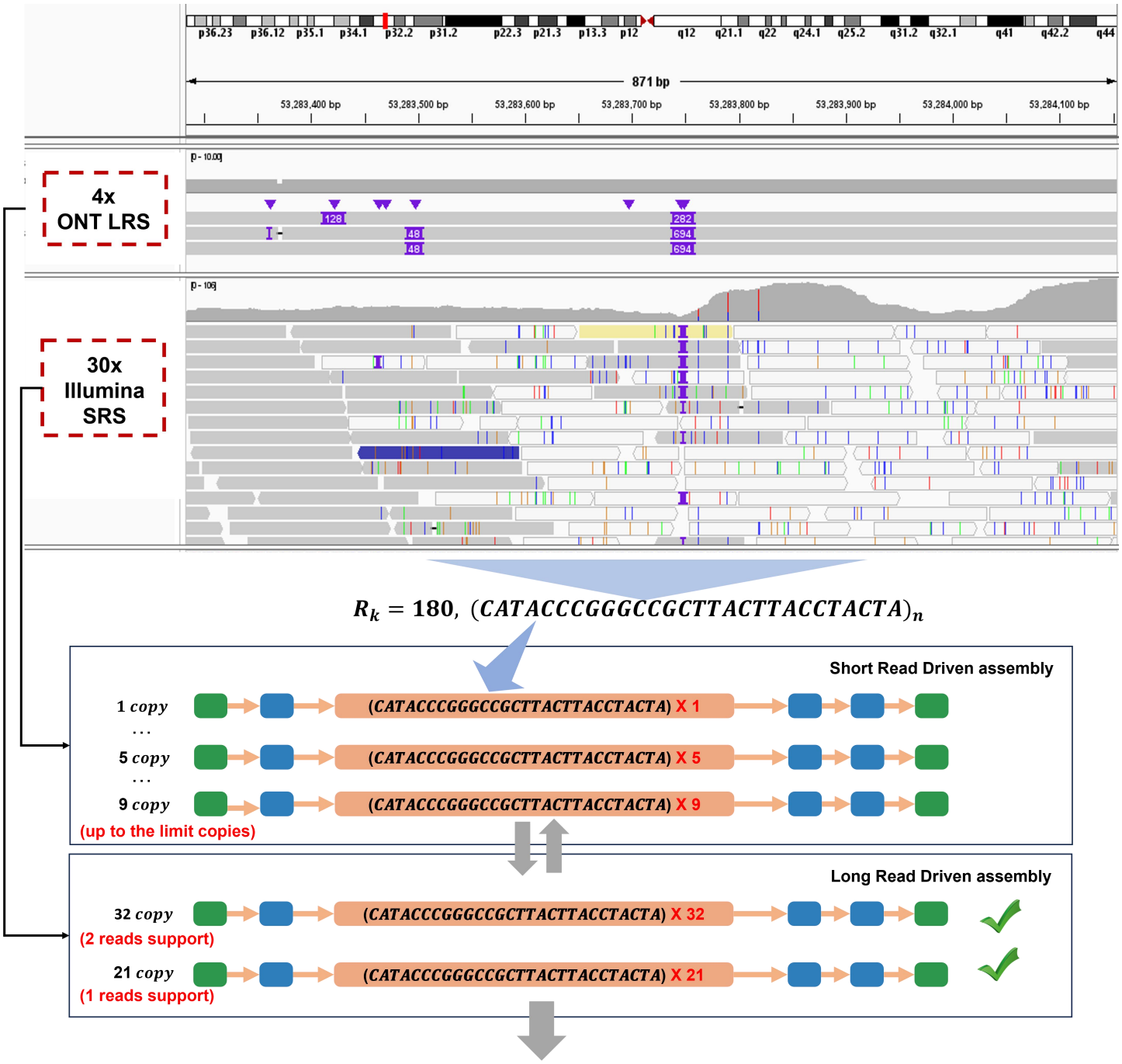
a, Phase1: virtual breakpoint-based signature clustering and refinement. HitSV estimates the virtual breakpoint implied by SDA, CLIP, DRP signature clusters in three different approaches. For SDA signature in dense cigar form (upper left panel, the dashed blocks indicate the dense cigars), HitSV projects all the bases of the reads of signature to the reference and define the distribution of SV breakpoint on the positions having at least one base being mapped to. For each of the position, a probability is assigned as the number of bases being mapped (normalized by the total number of bases), further, HitSV separately tests each of the signatures where the p-value is determined as the probability at the centroid of the signature in reference. For SDA signature in split alignment form (upper right panel), HitSV separately handles the multiple parts of the alignments. For each part, HitSV extends an L bp (default value: 10 bp) block from the split point, projects the block to reference and assigns a weight (default value: 0.1) to each of the projected positions. Further, the null distribution is defined on all the reference positions having at least one projected base, and for each position the probability is set as the sum weights of all the bases being projected there. Further, HitSV tests each of the signatures using this distribution. For CLP signature (lower left panel), HitSV extends an L bp (default value: 10 bp) block from the clipping point and implements base projection and weighting in a similar way to that of split alignment. Further, the null distribution is also defined on all the reference positions being projected for signature testing. For DRP signature (lower right panel), HitSV also handles the involved read pairs similarly. The difference is that the extended blocks are generated from the downstream and upstream alignment endpoints of the two paired reads, respectively. With the extended blocks, the null distribution is also composed based on the sum weights of the read projection positions. **b,** Phase2: the connection of refined clusters. HitSV greedily connects the signature clusters in nearby regions and/or being linked to combine the signature clusters belonging to the same SV event but having various types of signatures or being interspersed caused by multiple breakpoints and/or genomic repeats. For a deletion event (upper left panel), HitSV directly integrates the nearby clusters of SDA, CLP and DRP. Moreover, the DRP clusters can be further linked by read pairing information. For an insertion event (upper right panel), nearby CLP, SDA and DRP clusters are directly integrated similarly. Since insertion events can shorten the junctions of paired reads, in most cases the clusters can be comprehensively collected in a local region without regarding to the distant linking information. For a translocation event (middle left panel), it typically has two breakpoints and four involving clusters, nearby clusters were linked initially and then the long-range clusters are then clustered. For a tandem duplication event (middle right panel), CLP, SDA and DRP produce two breakpoints located around the boundaries of the duplication, separately. At each boundary, the nearby breakpoints from each category are first combined, and then the long-range clusters are straightforwardly merged. For an interspersed duplication event (lower panel), CLP and SDA produce four involving clusters, the split alignment signature support long-range information to combine all the clusters. Also refer to Fig. 1b for the illustration for inversion events.



Supplementary Fig. 2. A schematic illustration of local haplotype sequence reconstruction and polishing

HitSV performs a unitig-DBG–based local haplotype-resolved assembly to reconstruct SV alleles through the following steps. **a,** Read collection and graph construction. All reads including those without explicit SV signatures in the vicinity of candidate SV events are retrieved for assembly. From these reads and the local reference sequence, HitSV builds a unitig de Bruijn graph (unitig-DBG). Then unitigs are annotated by their average coverage, classified as high- or low-depth using an adaptive threshold, and labeled as cyclic or acyclic after cycle detection. **b,** Long-read–based SV detection. Each long read is represented as a path through the graph. HitSV initializes contig from the path spanning the greatest total length of high-depth nodes, and then iteratively aligns other reads to existing paths using dynamic programming on high-depth nodes. A read is considered not to support a contig if high-depth nodes at any corresponding positions diverge in sequence, or if a large gap is observed at any position; otherwise, the read is regarded as supporting the contig. Reads that support a contig refine the corresponding path, whereas reads that do not support any existing contig initiate new ones. **c,** Short-read–based SV calling. To handle cycles that confound conventional assemblers, HitSV enumerates candidate alleles by controlled graph traversal between flanking anchors, limiting node revisits to a fixed number (at most nine times by default) to prevent infinite walks. Generated contigs are filtered by realigning reads and discarding those lacking sufficient support. Surviving contigs are treated as candidate haplotypes; all diploid haplotype pairs are evaluated using a

likelihood that penalizes unmapped reads and inconsistencies between read and haplotype k-mer counts. The best pair is realigned to the reference to infer SV calls. **d,** Hybrid calling and polishing. When both data types are available, HitSV prioritizes long-read contigs as backbones, and polishes them with short reads. An empirical 9-mer error profile derived from high-confidence alignments guides identification of low-fidelity regions and homopolymers, which are locally reassembled from short reads and corrected to improve base-level accuracy.

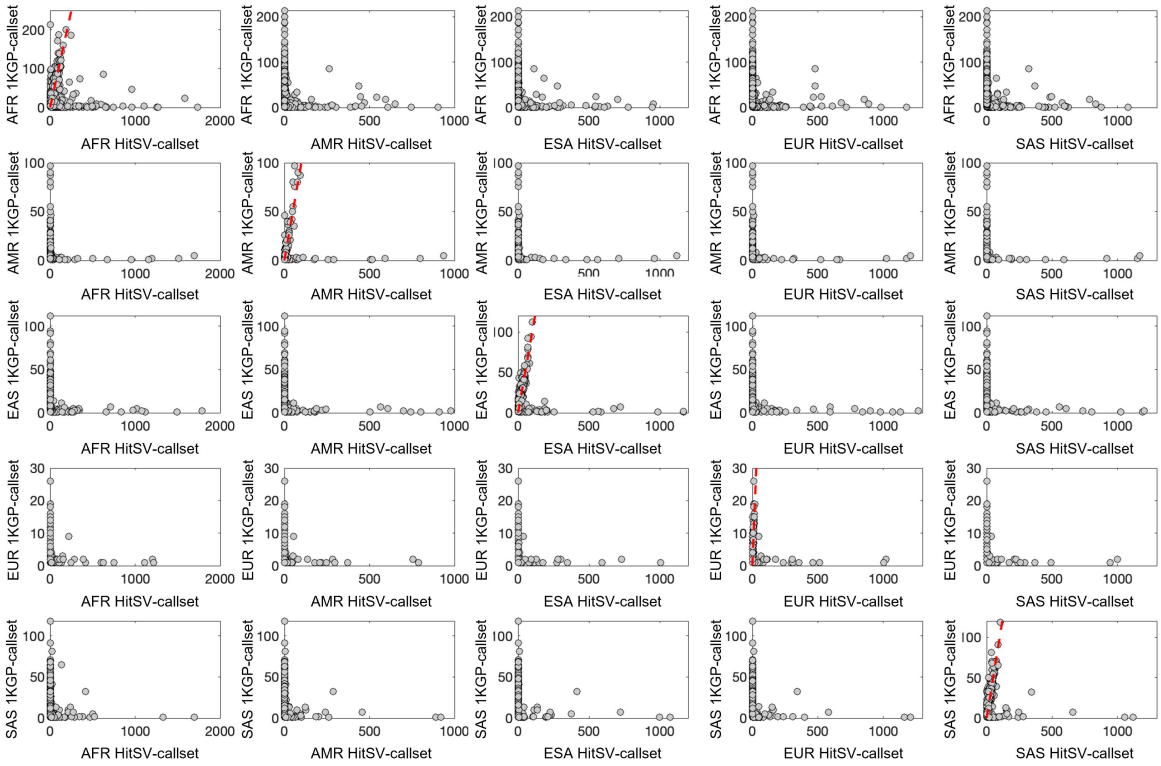


Caller	Alleles in Hap1	Alleles in Hap2	SV allele match	Zygosity match	Sequence match
HitSV	chr1:53283531 410 INS (0|1)	chr1:53283531 694 INS (1|0)	√	√	√

Supplementary Fig. 3. A schematic illustration on the successful of SV allele reconstruction and genotyping with evident long reads.

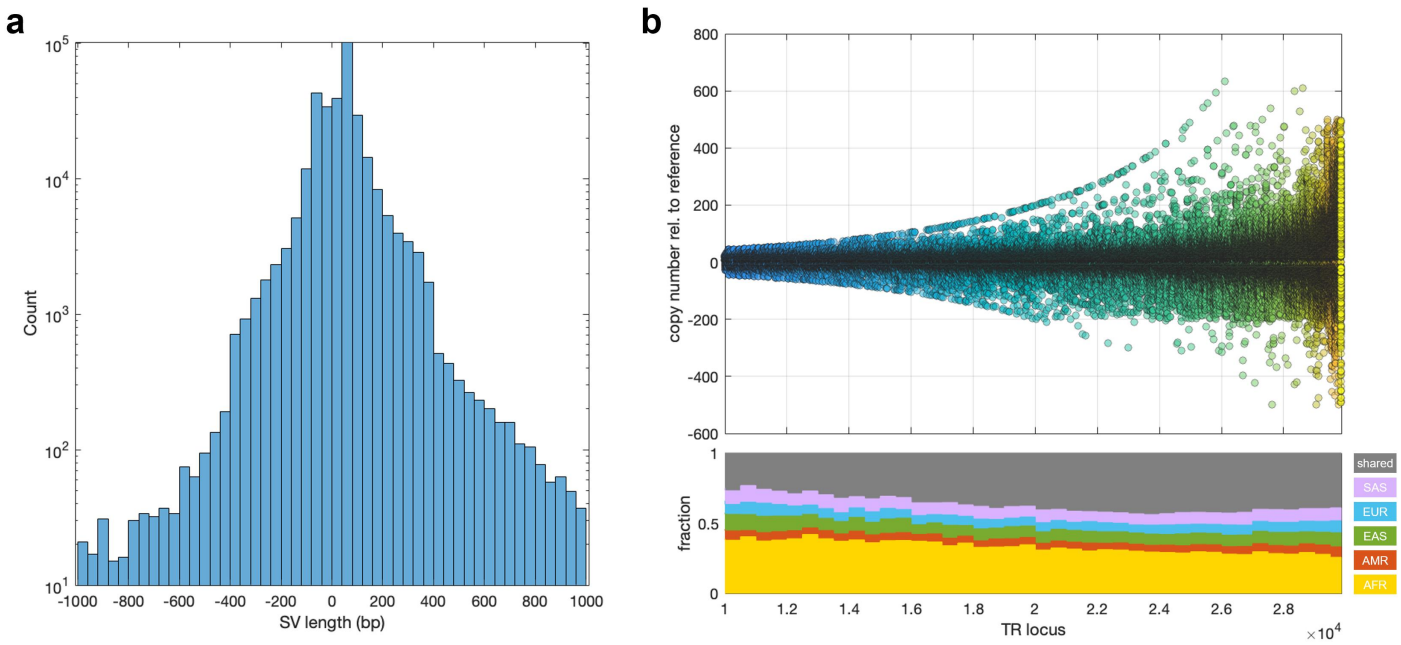
The IGV snapshot for a VNTR in a highly repetitive region (GRCh38, chr1:53,283,531-53,283,934, $R_k = 180$) highlights the improvement of evident long reads during SV allele reconstruction. The reference genome (GRCh38) contains 10 tandem copies of a 28-bp repeat unit. Two target haplotypes harbor 32 and 21 repeat units, corresponding to a 694-bp and a 410-bp insertion, respectively. (Bottom) In the 30x Illumina dataset, the high local repetitiveness prevents short reads from spanning the SV site. Despite heuristic graph walking, the reconstructed alleles are truncated (1–9 copies), failing to recover the full SV architecture. (Top) In the 4x ONT dataset, spanning long reads (one for the 410-bp and two for the

694-bp alleles) provide sufficient structural scaffolds. HitSV leverages these evidential long reads to successfully reconstruct both divergent contigs, enabling accurate SV calling and genotyping in regions recalcitrant to short-read-only methods.



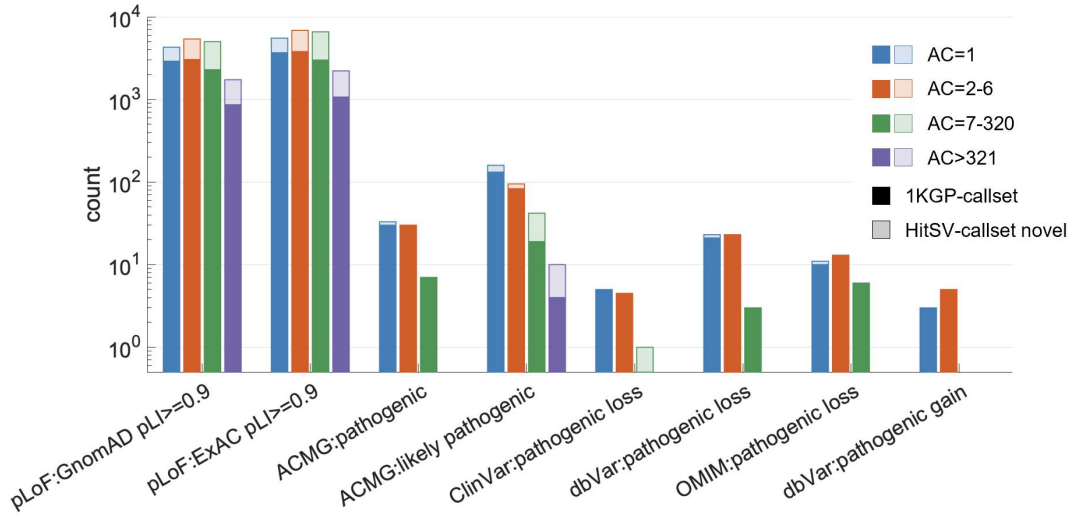
Supplementary Fig. 4. Allele counts of population-specific SVs in HitSV-callset and 1KGP-callset

Only SVs regarded unique to the superpopulation indicated on the y-axis in the 1KGP-callset were plotted. The allele count of these SVs in the same superpopulation (diagonal) or in another superpopulation, in the HitSV-callset, is shown on the x-axis. The allele count of these SVs in the 1KGP-callset is shown on the y-axis. In the diagonal subplots, the red dashed lines mark the 1:1 ratio. In the main text, SVs with strongly elevated allele frequencies are those in the diagonal subplots whose allele frequencies in HitSV-callset became at least 2-fold of their allele frequencies in 1KGP-callset, in the same superpopulation.



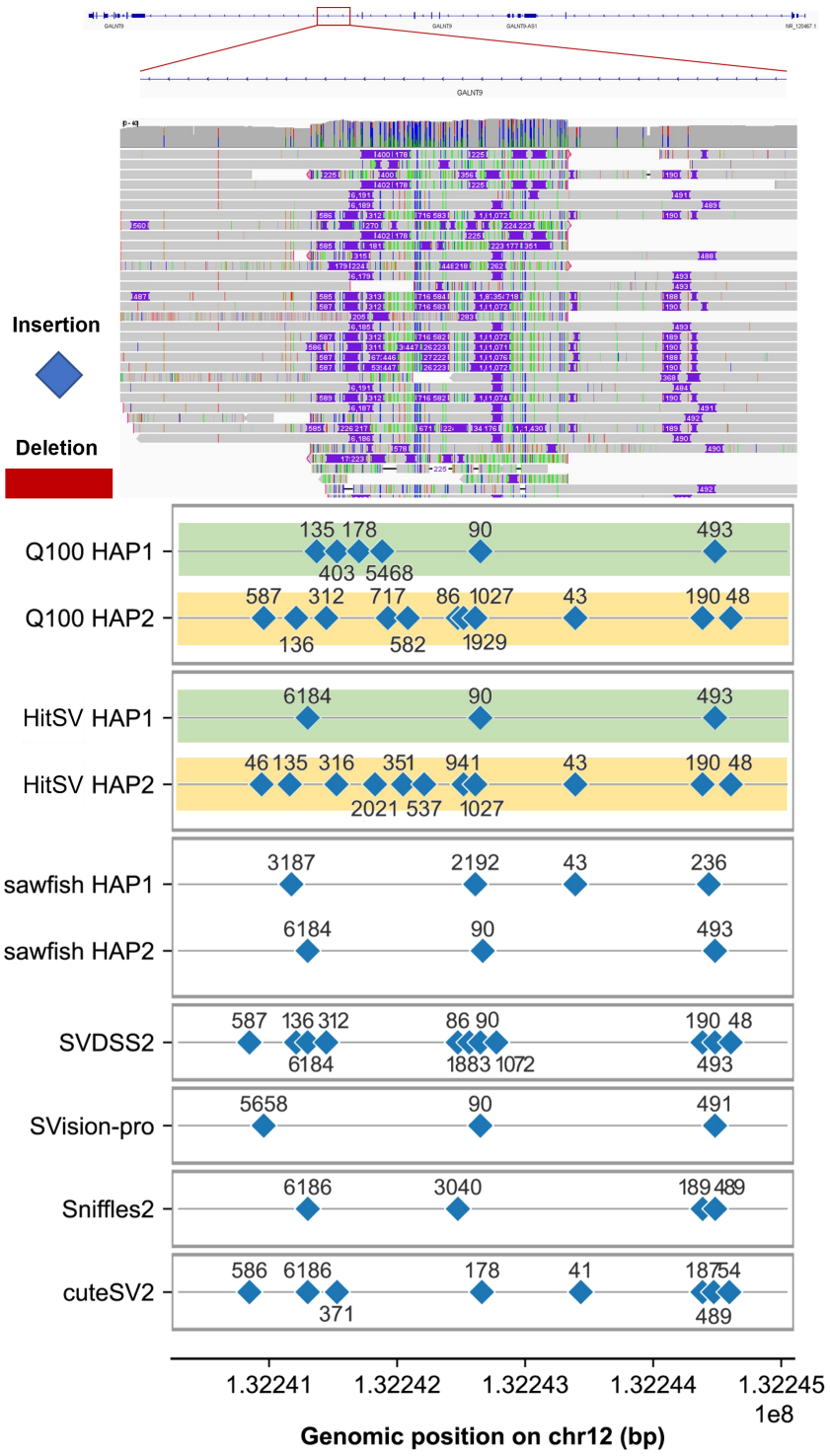
Supplementary Fig. 5. Tandem repeat-associated SVs in the HitSV-callset

a, The length distribution of tandem repeat-associated SVs in HitSV-callset, shown on the logarithmic scale. **b,** Copy number variations (upper panel) and the fractions of population unique SVs (lower panel) at each tandem repeat locus. The tandem repeat loci are sorted on the x-axis by increasing variance.



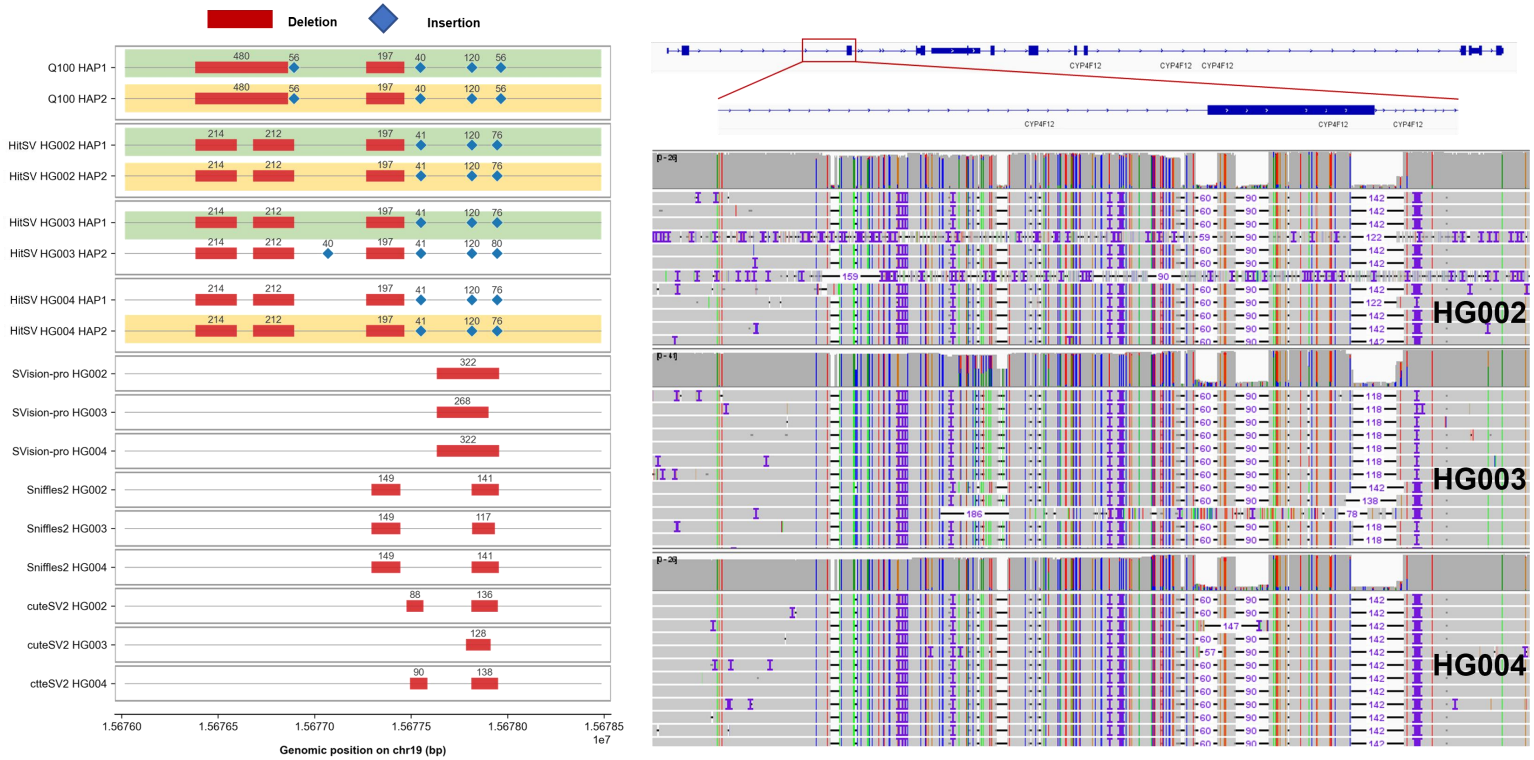
Supplementary Fig. 6. Phenotypic annotation of SVs in the 1KGP-callset and HitSV-callset

Phenotypic annotation of SVs in the 1KGP-callset (dark colors) and SVs distinctly identified in HitSV-callset (boxed light colors). For each annotation group, SVs are partitioned into four allele frequency bins.



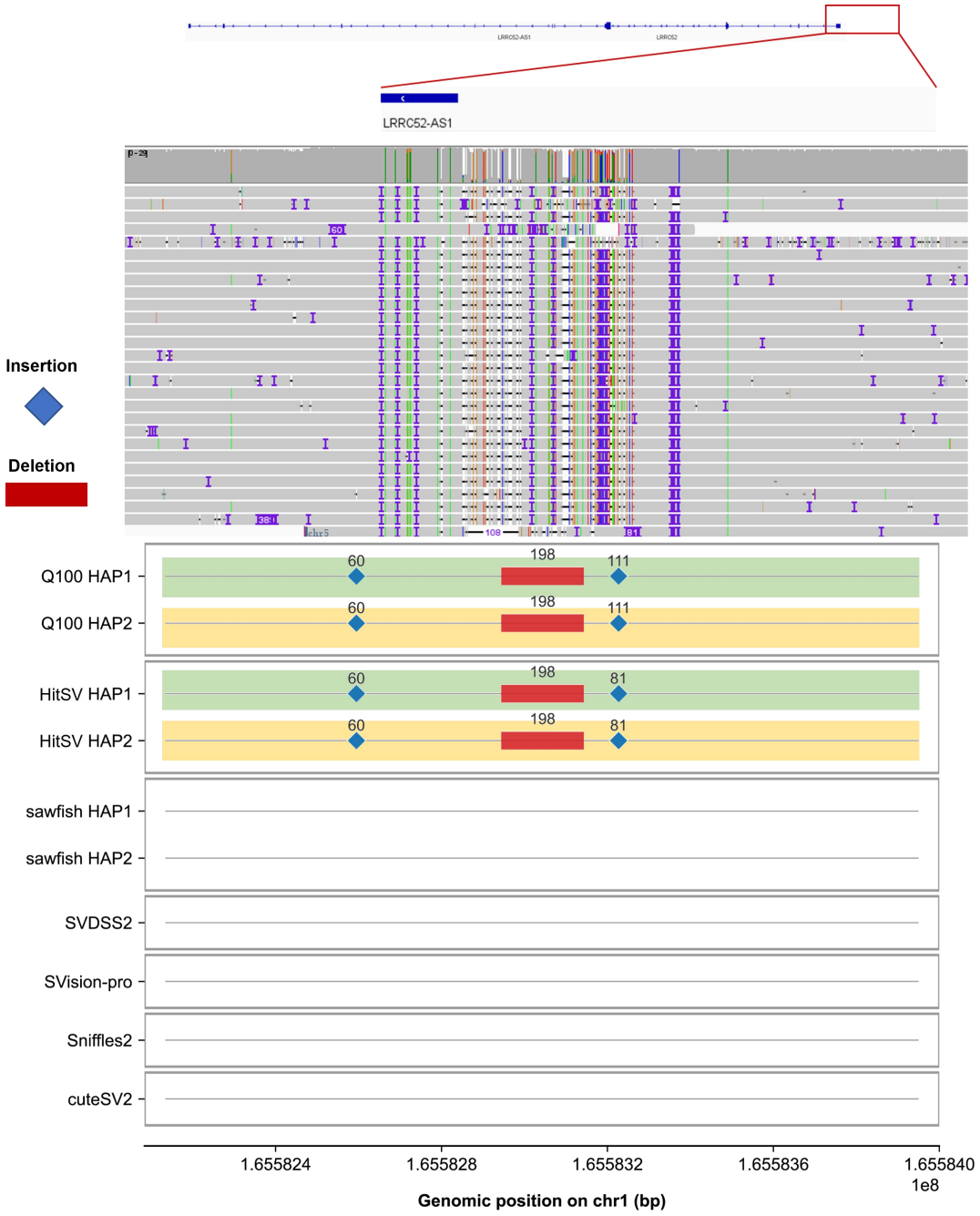
Supplementary Fig. 7. An example of CSV detection in a HG002 dataset.

An example of CSV detection in an intronic region of *GALNT9* on chromosome 12 (chr12:132,239,289–132,246,050 bp) is shown. The upper panel presents an Integrative Genomics Viewer (IGV) snapshot of this region with heterogeneous SV signatures from read alignment. The lower panel summarizes the SVs detected by different tools in this region, displaying their genomic positions and variant lengths. Q100 HAP1 and Q100 HAP2 represent the ground-truth SVs on the two haplotypes. Among the evaluated methods, HitSV and sawfish provide haplotype-resolved results and can distinguish SVs on HAP1 and HAP2, whereas the remaining tools lack phasing capability and therefore report all detected SVs without haplotype assignment. Notably, HitSV shows high concordance with the ground truth in both the genomic distribution and length of detected variants, whereas other tools exhibit substantial discrepancies.



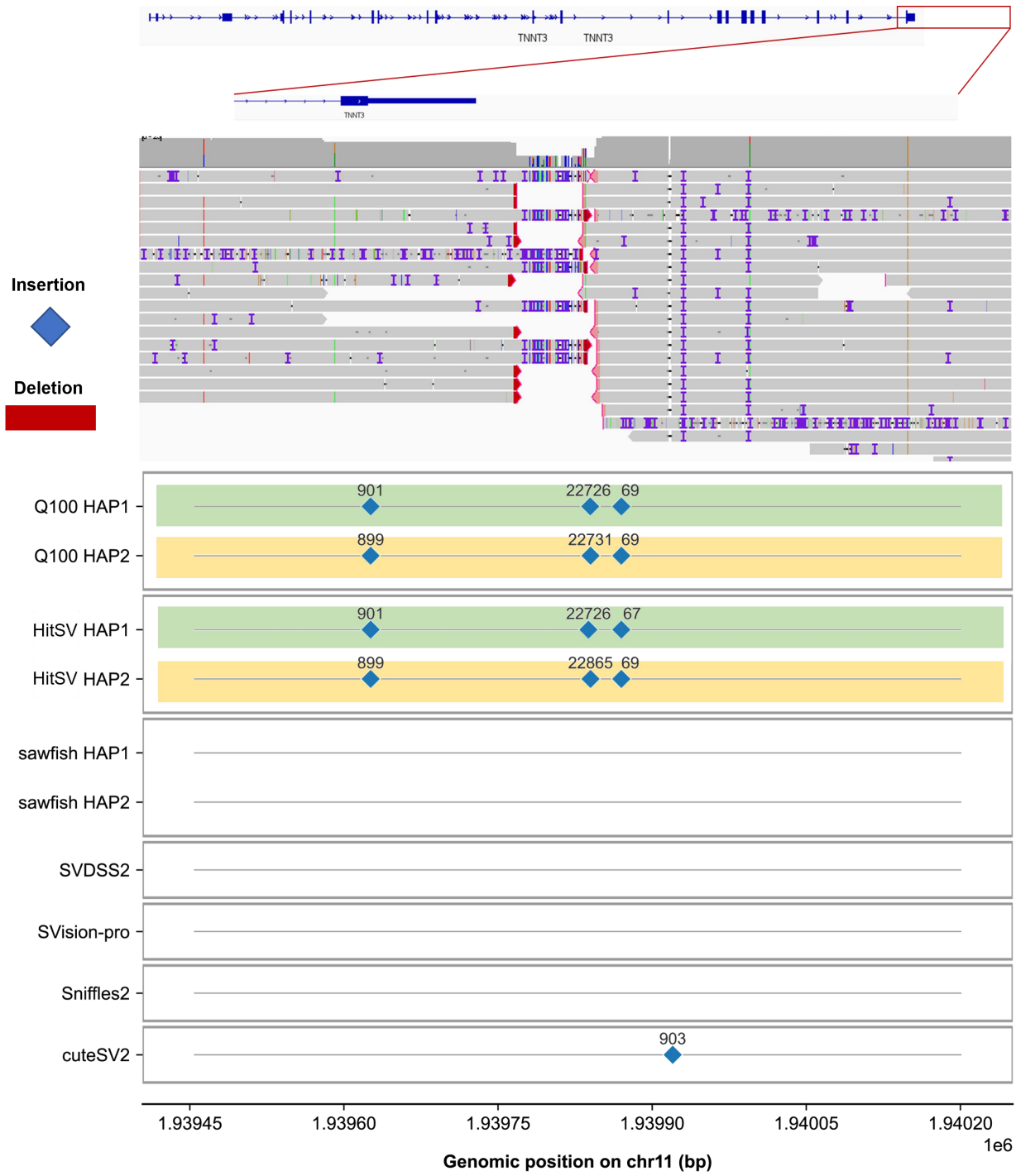
Supplementary Fig. 8. An example of CSV detection in a family-based dataset.

The analysis focuses on a local genomic region encompassing the *CYP4F12* locus on chromosome 19 (chr19:15,677,010-15,678,477 bp), illustrating CSV inheritance patterns across related individuals. The left panel summarizes the SVs detected by different tools across haplotypes and samples. Among the evaluated methods, HitSV provides haplotype-resolved calls and enables consistent tracking of CSVs across family members, facilitating the identification of shared and inherited variants, whereas other tools either lack phasing capability or fail to recover the complete set of events. This example highlights the advantage of HitSV in accurately resolving complex SV structures and their transmission patterns in pedigree-based analyses. The right panel presents IGV snapshots of the same region for HG002, HG003, and HG004, showing read alignments contain a series of concordantly mapped small divergences.



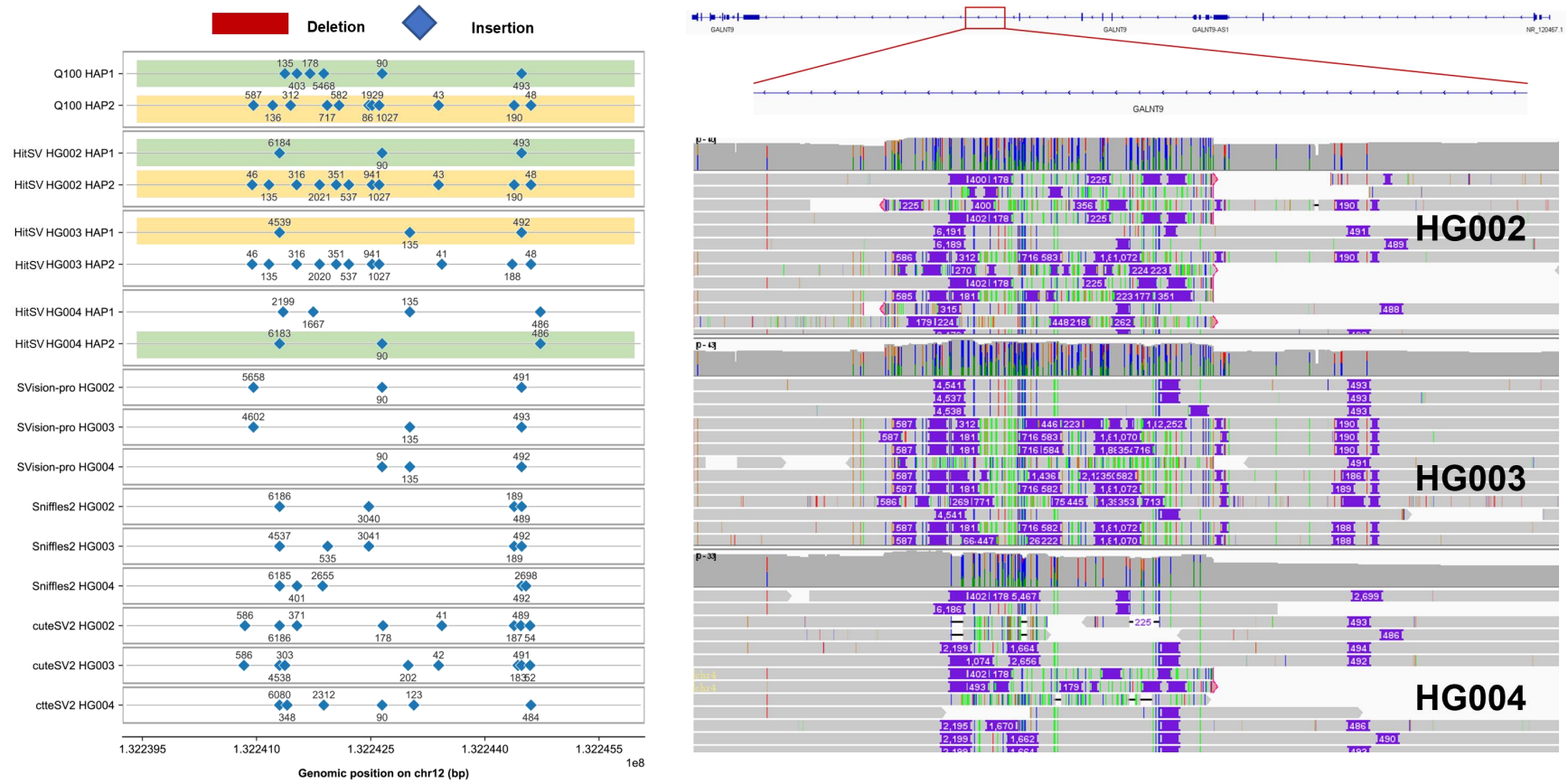
Supplementary Fig. 9. An example of CSV detection in a HG002 dataset.

An example of CSV detection in an exonic region of *LRRC52* on chromosome 1 (chr1:165,582,134-165,583,950 bp) is shown. The upper panel presents an IGV snapshot of this region, illustrating the read alignments containing a series of concordantly mapped small divergences. The lower panel summarizes the SVs detected by different tools in this region, displaying their genomic positions and variant lengths. Q100 HAP1 and Q100 HAP2 represent the ground-truth SVs on the two haplotypes. Among the evaluated methods, HitSV and sawfish provide haplotype-resolved results and can distinguish SVs on HAP1 and HAP2, whereas the remaining tools lack phasing capability and therefore report all detected SVs without haplotype assignment. Notably, HitSV achieves complete concordance with the ground truth in both variant distribution and length, while no variants are detected by the other methods.



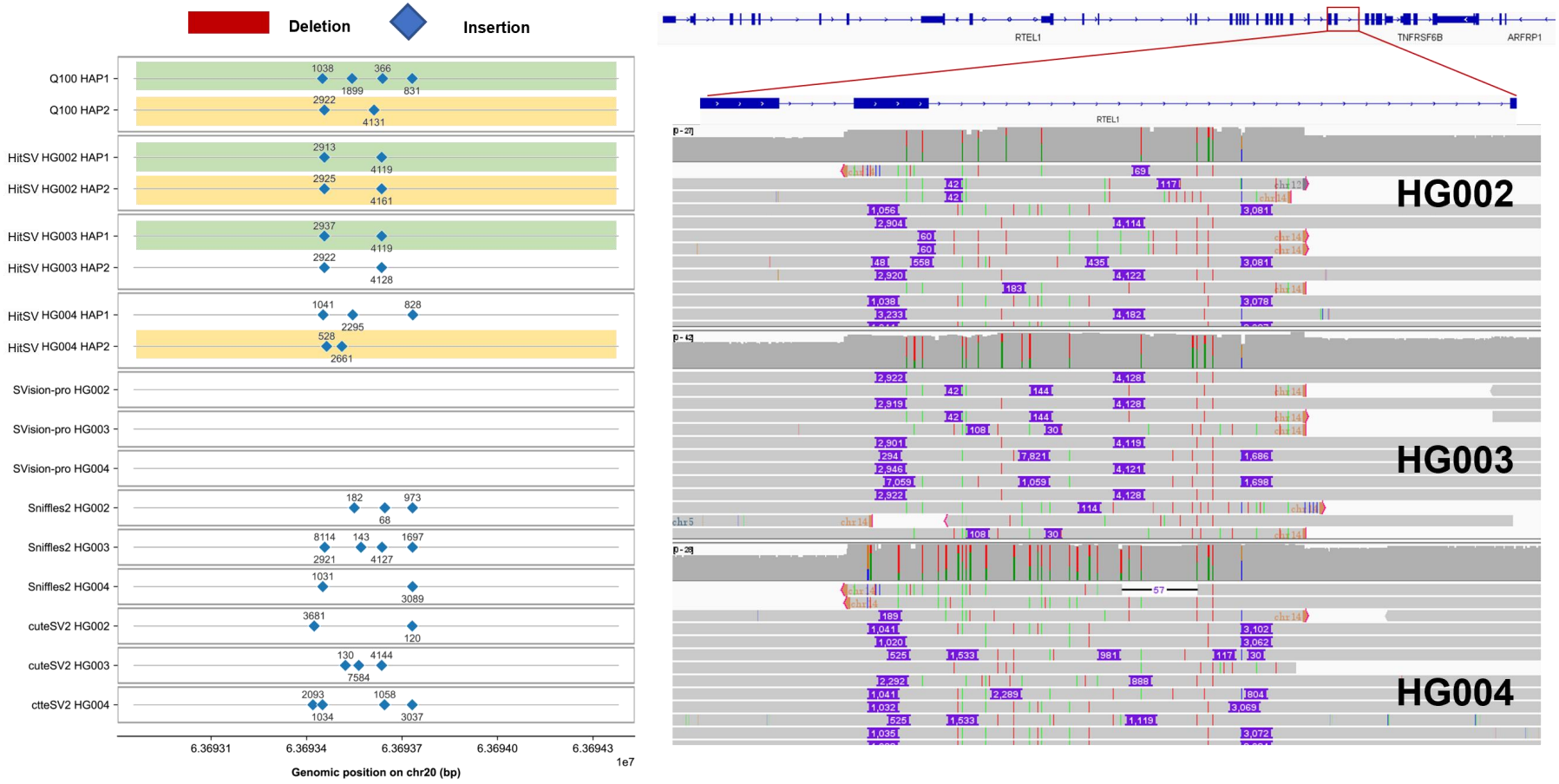
Supplementary Fig. 10. An example of CSV detection in a HG002 dataset.

An example of CSV detection in an exonic region of *TNNT3* on chromosome 11 (chr11:1,939,454-1,940,201 bp) is shown. The upper panel presents an IGV snapshot of this region, illustrating abnormal read alignments including large soft clipping, and split. The lower panel summarizes the SVs detected by different tools in this region, displaying their genomic positions and variant lengths. Q100 HAP1 and Q100 HAP2 represent the ground-truth SVs on the two haplotypes. Among the evaluated methods, HitSV and sawfish provide haplotype-resolved results and can distinguish SVs on HAP1 and HAP2, whereas the remaining tools lack phasing capability and therefore report all detected SVs without haplotype assignment. Notably, HitSV achieves complete concordance with the ground truth in both variant distribution and length, while no variants are detected by the other methods except cuteSV2 identifies a 903-bp insertion.



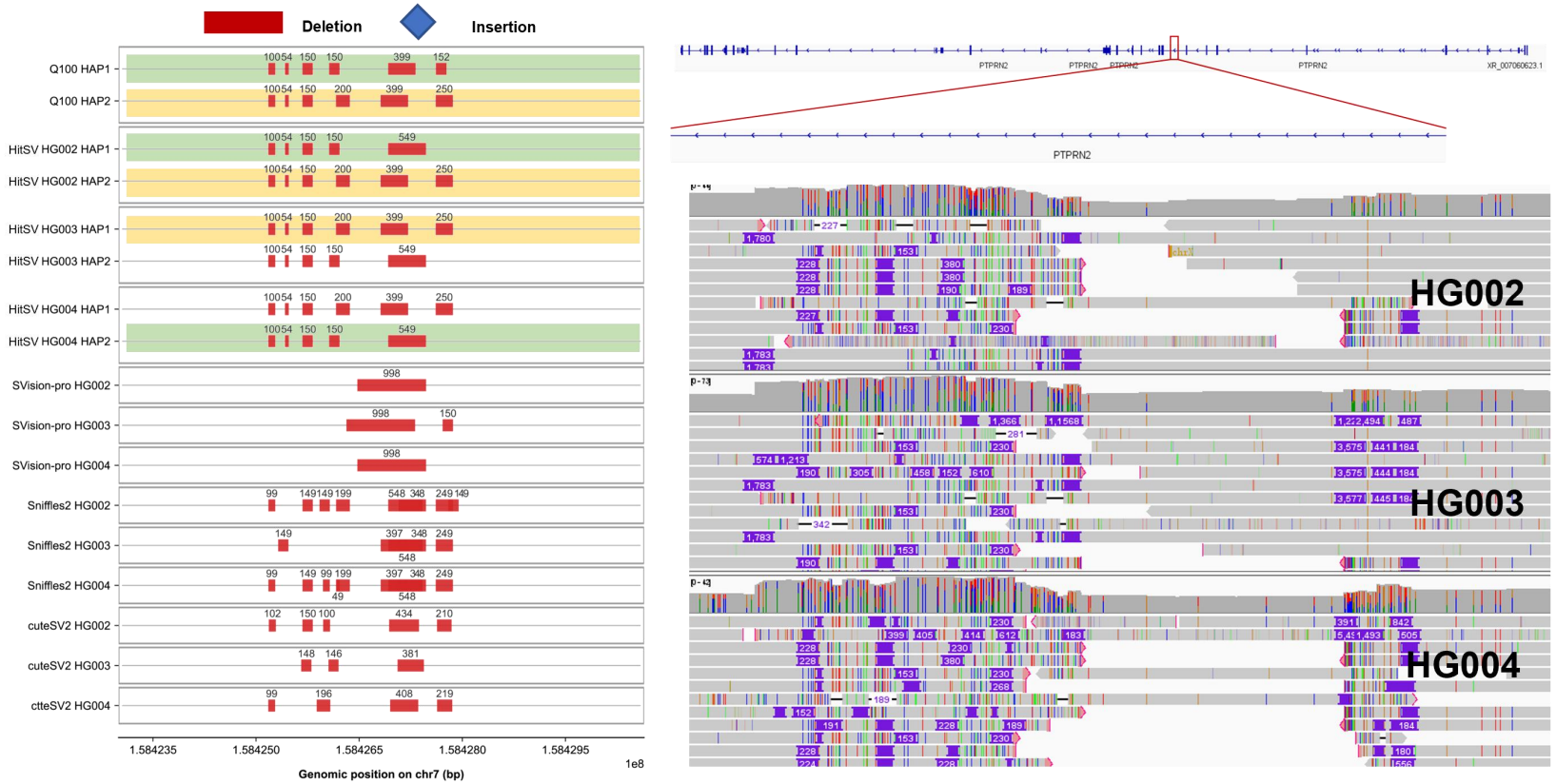
Supplementary Fig. 11. An example of CSV detection in a family-based dataset.

The analysis focuses on a local genomic region encompassing the *GALNT9* locus on chromosome 12 (chr12:132,239,289-132,246,050 bp), illustrating CSV inheritance patterns across related individuals. The left panel summarizes the SVs detected by different tools across haplotypes and samples. Among the evaluated methods, HitSV produces haplotype-resolved calls and consistently recovers concordant SV patterns across family members, enabling reliable tracking of inherited variants. In contrast, other tools either lack haplotype resolution or show fragmented, inconsistent, or missing calls. This example highlights the advantage of HitSV in accurately resolving complex SV structures and their transmission patterns in pedigree-based analyses. The right panel presents IGV snapshots of the same region for HG002, HG003, and HG004, showing read alignments with heterogeneous SV signatures.



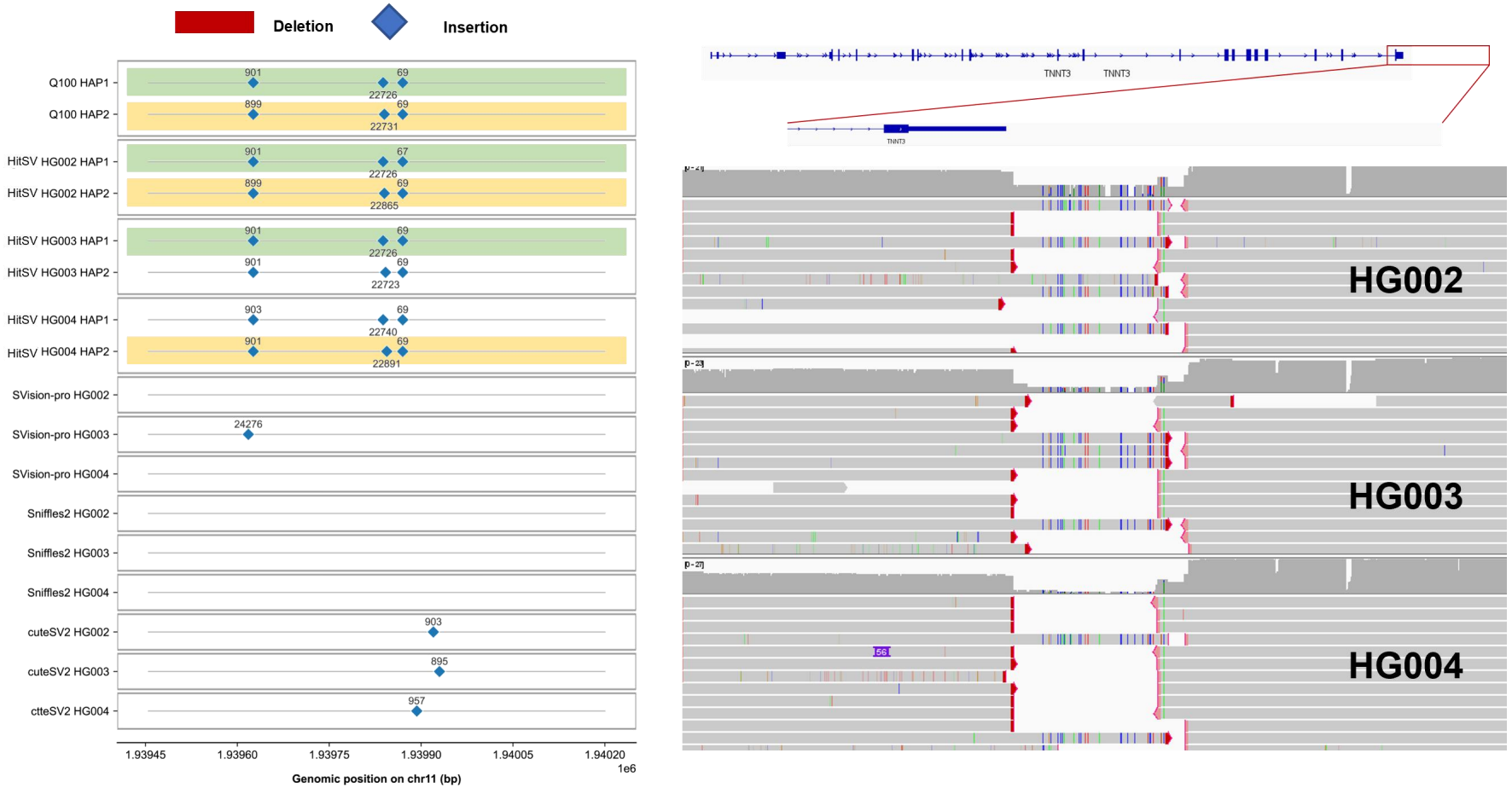
Supplementary Fig. 12. An example of CSV detection in a family-based dataset.

The analysis focuses on a local genomic region encompassing the *RTEL1* locus on chromosome 20 (chr20:63,692,856-63,694,381 bp), illustrating CSV inheritance patterns across related individuals. The left panel summarizes the SVs detected by different tools across haplotypes and samples. HitSV produces haplotype-resolved calls and consistently recovers SVs that are concordant with the ground truth and shared across family members, enabling reliable tracing of inherited CSVs. The right panel presents IGV snapshots of the same region for HG002, HG003, and HG004, showing abnormal read alignments signatures including large soft clipping, and split.



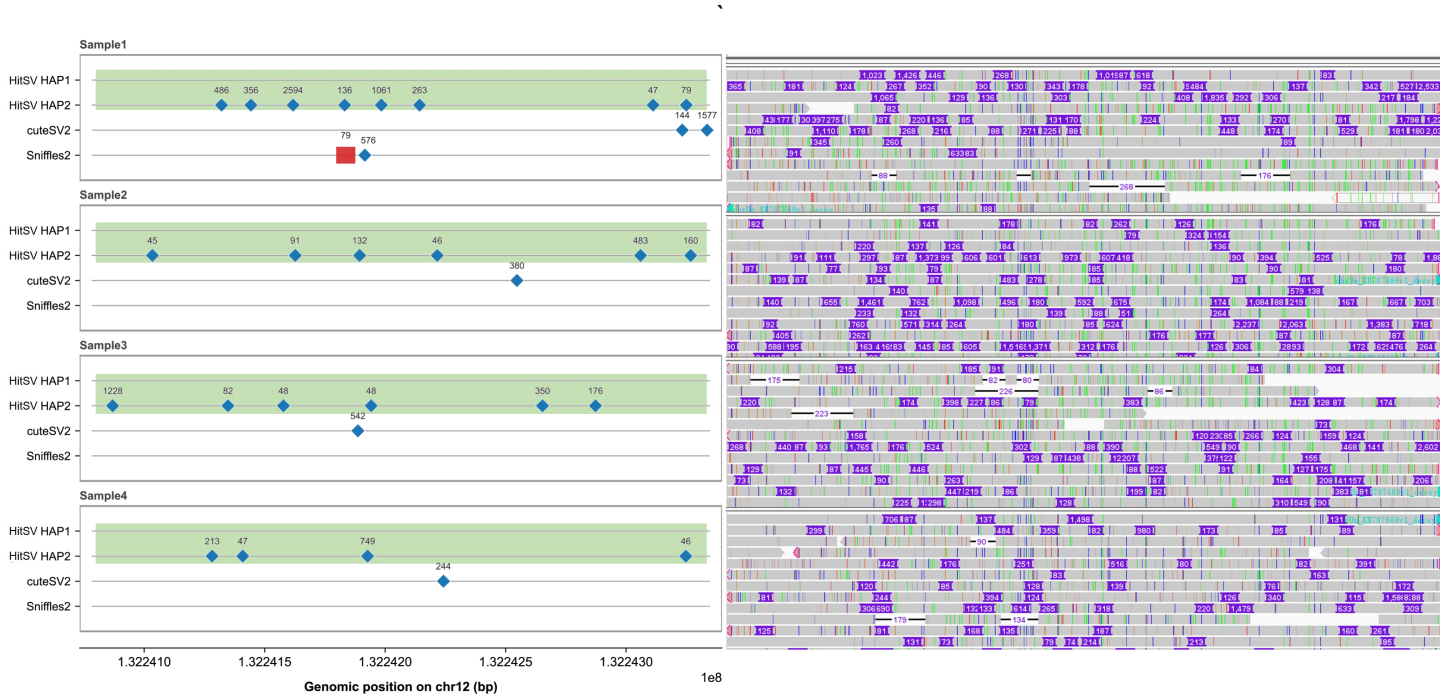
Supplementary Fig. 13. An example of CSV detection in a family-based dataset.

The analysis focuses on a local genomic region encompassing the *PTPRN2* locus on chromosome 7 (chr7:158423050-158430593 bp), illustrating CSV inheritance patterns across related individuals. The left panel summarizes the SVs detected by different tools across haplotypes and samples. HitSV produces haplotype-resolved calls and consistently recovers SVs that are concordant with the ground truth and shared across family members, enabling reliable tracing of inherited CSVs. In contrast, other tools lack haplotype resolution and often report inconsistent or missing calls. The right panel presents IGV snapshots of the same region for HG002, HG003, and HG004, showing abnormal read alignments signatures including large soft clipping, and split.



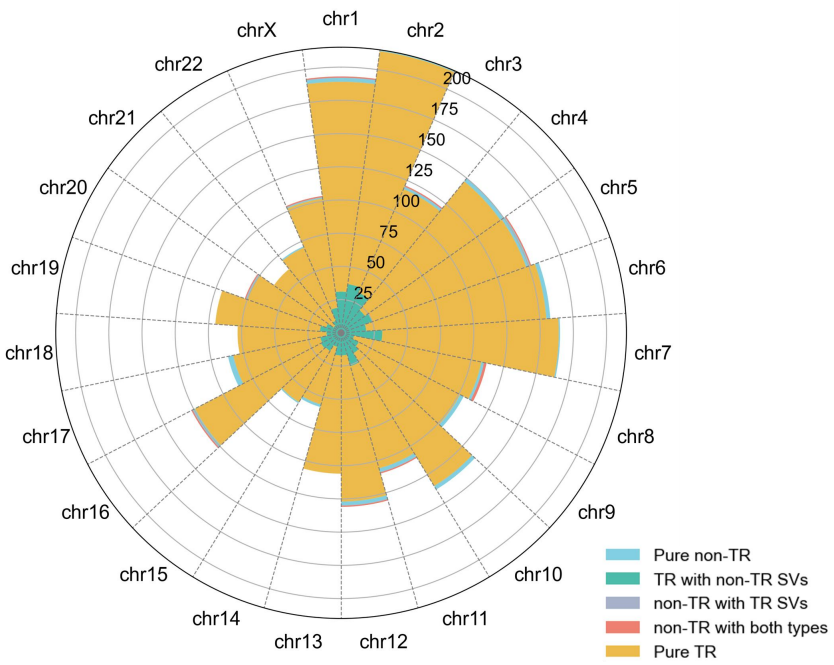
Supplementary Fig. 14. An example of CSV detection in a family-based dataset.

The analysis focuses on a local genomic region encompassing the PTPRN2 locus on chromosome 7 (chr7:158423050-158430593 bp), illustrating CSV inheritance patterns across related individuals. The left panel summarizes the SVs detected by different tools across haplotypes and samples. HitSV produces haplotype-resolved calls and consistently recovers SVs that are concordant with the ground truth and shared across family members, while other methods fail to detect these events. The right panel presents IGV snapshots of the same region for HG002, HG003, and HG004, showing abnormal read alignments signatures including large soft clipping, and split.



Supplementary Fig. 15. Detection of complex structural variants across multiple ONT samples.

Another example focuses on a common CSV region on human chromosome 12 (chr12:132240782-132243348 bp), illustrating CSV detection results across five samples from the 557-ONT dataset. The left panel summarizes the structural variants detected by different tools across samples and haplotypes. Haplotype-resolved calls produced by HitSV enable explicit assignment of CSVs to individual haplotypes, facilitating the tracing and comparison of complex variant patterns across samples. In contrast, other tools lack haplotype resolution and report aggregated SV calls. The right panel presents IGV snapshots of the same region for multiple samples, showing read alignments and supporting signals consistent with TR-MEI–associated CSV events.



Supplementary Fig. 16. Chromosome-wide distribution and classification of CSV regions in HPRC datasets

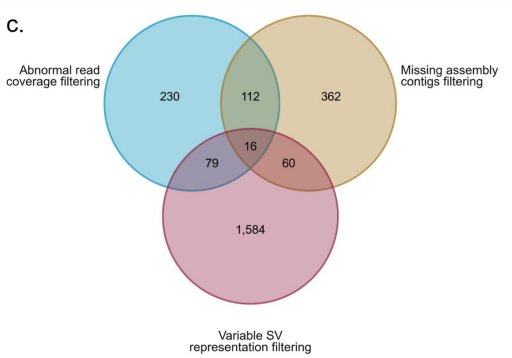
Chromosome-level classification of CSV regions across 47 Human Pangenome Reference Consortium (HPRC) datasets. Each sector corresponds to one chromosome, and radial bars indicate the number of CSV regions assigned to different structural categories.

a. Before quality filtering

Number of CSV Regions		CHN100K-557					
		Pure TR	TR with non-TR SVs	Pure non-TR	non-TR with TR SVs	non-TR with both types	No CSV
HPRC-47	Pure TR	1261	1391	0	0	0	0
	TR with non-TR SVs	33	463	0	0	0	0
	Pure non-TR	0	0	28	0	21	0
	non-TR with TR SVs	0	0	0	0	0	0
	non-TR with both types	0	0	0	0	10	0
	No CSV	729	1354	55	0	88	0
		Missing rate in HPRC cohort: 40.97%					Concordance rate:32.43%

b. After three high-confidence filters

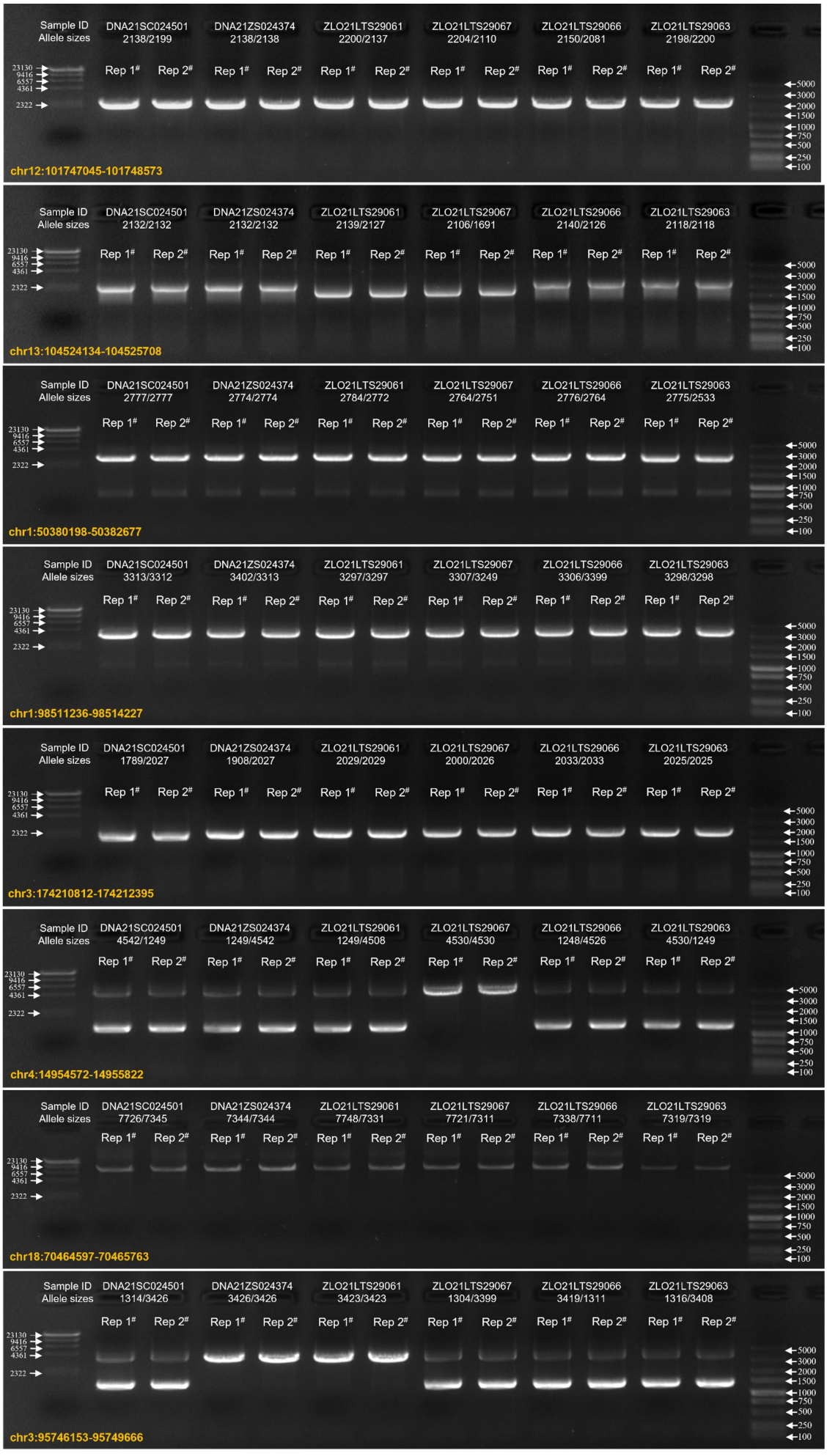
Number of CSV Regions		CHN100K-557					
		Pure TR	TR with non-TR SVs	Pure non-TR	non-TR with TR SVs	non-TR with both types	No CSV
HPRC-47	Pure TR	1219	969	0	0	0	0
	TR with non-TR SVs	14	422	0	0	0	0
	Pure non-TR	0	0	25	0	17	0
	non-TR with TR SVs	0	0	0	0	0	0
	non-TR with both types	0	0	0	0	10	0
	No CSV	71	163	27	0	53	0
		Missing rate in HPRC cohort: 10.50%					Concordance rate:56.05%



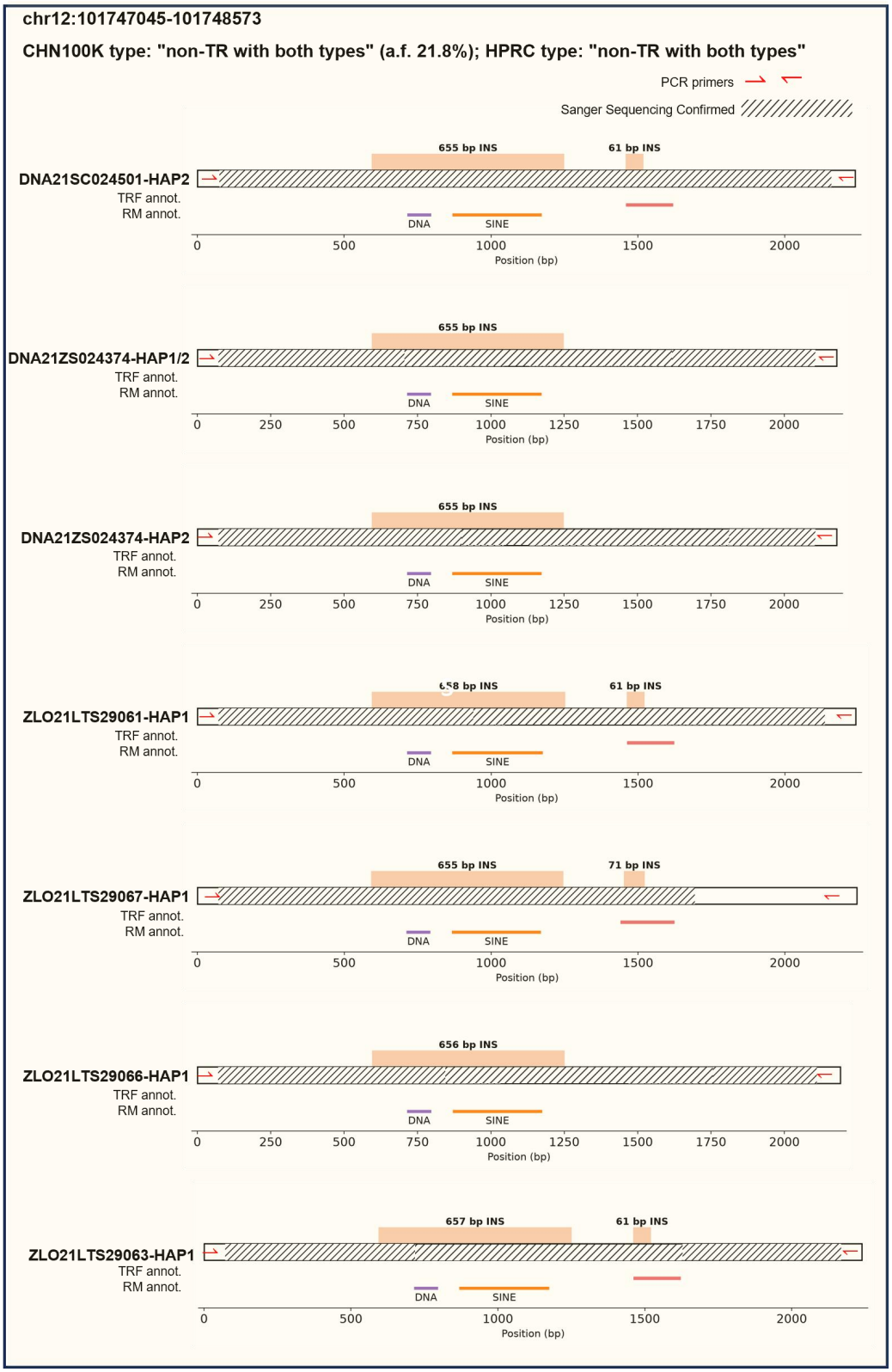
Supplementary Fig. 17. Effects of conservative quality-control filters on CHN100K-HPRC CSV comparison.

a, CHN100K-HPRC comparison for CHN100K CSV alleles with allele frequency >1% before applying quality-control filters. **b,** Conservative comparison after excluding alleles or loci affected by HPRC contig missingness, abnormal CHN100K read coverage or variable SV representation. After filtering, 2,676 CSV alleles were supported in both CHN100K and HPRC, whereas 314 remained unsupported by HPRC, corresponding to 10.50% of 2,990 high-confidence comparable CSV alleles. **c,** Number of CSV alleles or loci removed by each quality-control filter.

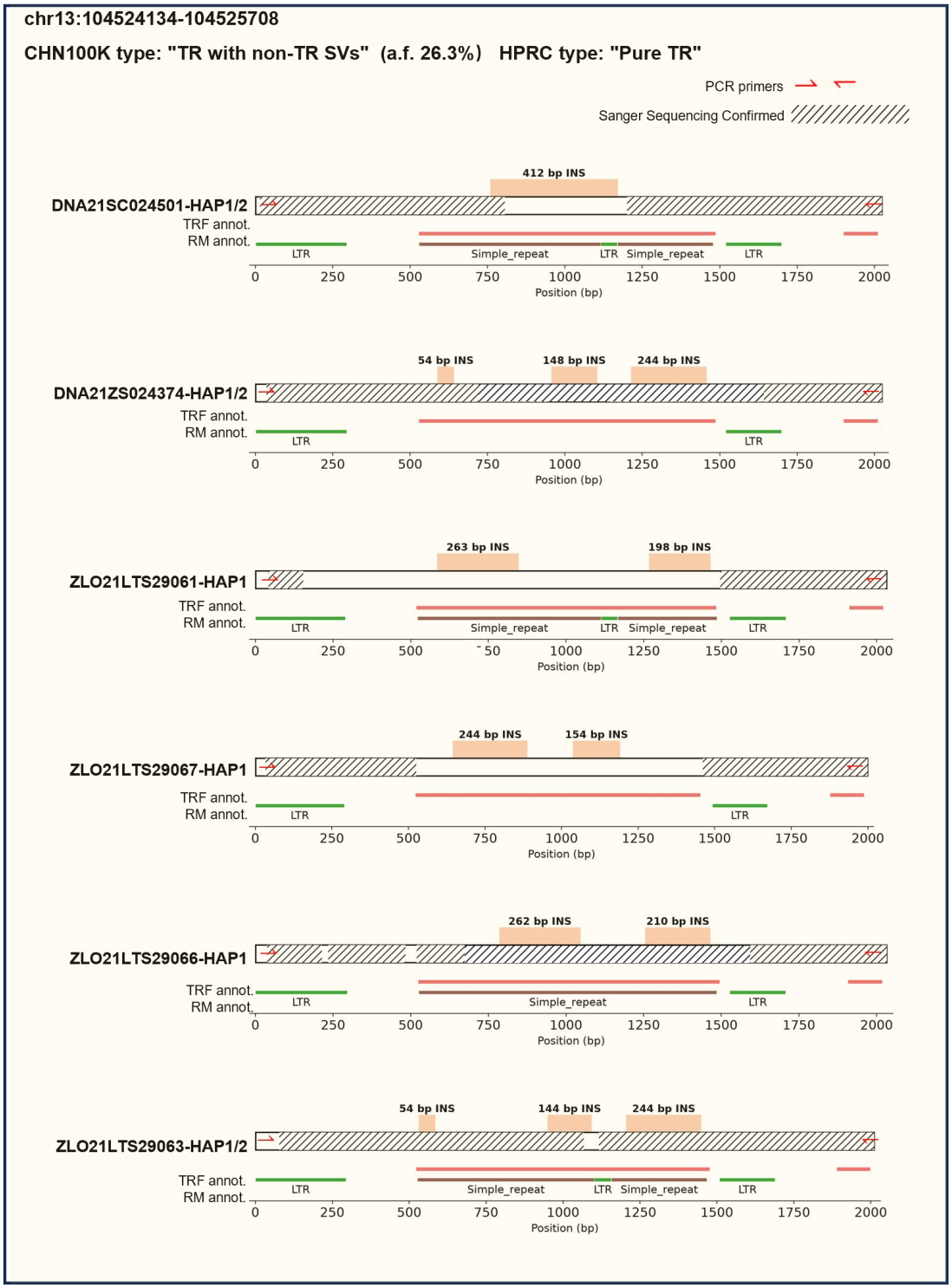
a


Sample ID
Allele sizes
DNA21SC024501 2138/2199
DNA21ZS024374 2138/2138
ZLO21LTS29061 2200/2137
ZLO21LTS29067 2204/2110
ZLO21LTS29066 2150/2081
ZLO21LTS29063 2198/2200
Rep 1# Rep 2# Rep 1# Rep 2# Rep 1# Rep 2# Rep 1# Rep 2# Rep 1# Rep 2# Rep 1# Rep 2#
23130
9416
6557
4361
2322
5000
3000
2000
1500
1000
750
500
250
100
chr12:101747045-101748573
Sample ID
Allele sizes
DNA21SC024501 2132/2132
DNA21ZS024374 2132/2132
ZLO21LTS29061 2139/2127
ZLO21LTS29067 2106/1691
ZLO21LTS29066 2140/2126
ZLO21LTS29063 2118/2118
Rep 1# Rep 2# Rep 1# Rep 2# Rep 1# Rep 2# Rep 1# Rep 2# Rep 1# Rep 2# Rep 1# Rep 2#
23130
9416
6557
4361
2322
5000
3000
2000
1500
1000
750
500
250
100
chr13:104524134-104525708
Sample ID
Allele sizes
DNA21SC024501 2777/2777
DNA21ZS024374 2774/2774
ZLO21LTS29061 2784/2772
ZLO21LTS29067 2764/2751
ZLO21LTS29066 2776/2764
ZLO21LTS29063 2775/2533
Rep 1# Rep 2# Rep 1# Rep 2# Rep 1# Rep 2# Rep 1# Rep 2# Rep 1# Rep 2# Rep 1# Rep 2#
23130
9416
6557
4361
2322
5000
3000
2000
1500
1000
750
500
250
100
chr1:50380198-50382677
Sample ID
Allele sizes
DNA21SC024501 3313/3312
DNA21ZS024374 3402/3313
ZLO21LTS29061 3297/3297
ZLO21LTS29067 3307/3249
ZLO21LTS29066 3306/3399
ZLO21LTS29063 3298/3298
Rep 1# Rep 2# Rep 1# Rep 2# Rep 1# Rep 2# Rep 1# Rep 2# Rep 1# Rep 2# Rep 1# Rep 2#
23130
9416
6557
4361
2322
5000
3000
2000
1500
1000
750
500
250
100
chr1:98511236-98514227
Sample ID
Allele sizes
DNA21SC024501 1789/2027
DNA21ZS024374 1908/2027
ZLO21LTS29061 2029/2029
ZLO21LTS29067 2000/2026
ZLO21LTS29066 2033/2033
ZLO21LTS29063 2025/2025
Rep 1# Rep 2# Rep 1# Rep 2# Rep 1# Rep 2# Rep 1# Rep 2# Rep 1# Rep 2# Rep 1# Rep 2#
23130
9416
6557
4361
2322
5000
3000
2000
1500
1000
750
500
250
100
chr3:174210812-174212395
Sample ID
Allele sizes
DNA21SC024501 4542/1249
DNA21ZS024374 1249/4542
ZLO21LTS29061 1249/4508
ZLO21LTS29067 4530/4530
ZLO21LTS29066 1248/4526
ZLO21LTS29063 4530/1249
Rep 1# Rep 2# Rep 1# Rep 2# Rep 1# Rep 2# Rep 1# Rep 2# Rep 1# Rep 2# Rep 1# Rep 2#
23130
9416
6557
4361
2322
5000
3000
2000
1500
1000
750
500
250
100
chr4:14954572-14955822
Sample ID
Allele sizes
DNA21SC024501 7726/7345
DNA21ZS024374 7344/7344
ZLO21LTS29061 7748/7331
ZLO21LTS29067 7721/7311
ZLO21LTS29066 7338/7711
ZLO21LTS29063 7319/7319
Rep 1# Rep 2# Rep 1# Rep 2# Rep 1# Rep 2# Rep 1# Rep 2# Rep 1# Rep 2# Rep 1# Rep 2#
23130
9416
6557
4361
2322
5000
3000
2000
1500
1000
750
500
250
100
chr18:70464597-70465763
Sample ID
Allele sizes
DNA21SC024501 1314/3426
DNA21ZS024374 3426/3426
ZLO21LTS29061 3423/3423
ZLO21LTS29067 1304/3399
ZLO21LTS29066 3419/1311
ZLO21LTS29063 1316/3408
Rep 1# Rep 2# Rep 1# Rep 2# Rep 1# Rep 2# Rep 1# Rep 2# Rep 1# Rep 2# Rep 1# Rep 2#
23130
9416
6557
4361
2322
5000
3000
2000
1500
1000
750
500
250
100
chr3:95746153-95749666

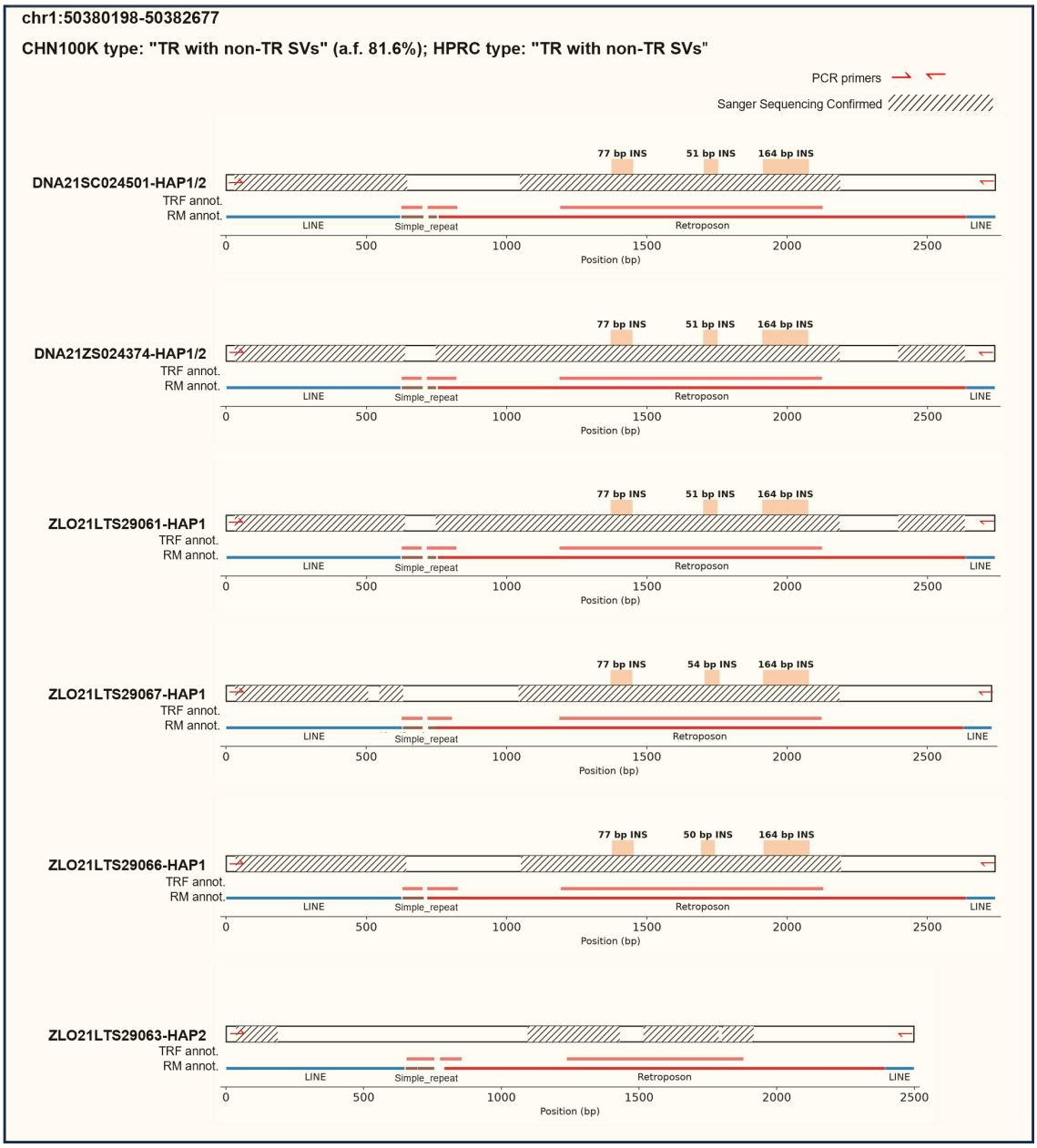
b


chr12:101747045-101748573
CHN100K type: "non-TR with both types" (a.f. 21.8%); HPRC type: "non-TR with both types"
PCR primers
Sanger Sequencing Confirmed
655 bp INS
61 bp INS
DNA21SC024501-HAP2
TRF annot.
RM annot.
DNA
SINE
0
500
1000
1500
2000
Position (bp)
655 bp INS
DNA21ZS024374-HAP1/2
TRF annot.
RM annot.
DNA
SINE
0
250
500
750
1000
1250
1500
1750
2000
Position (bp)
655 bp INS
DNA21ZS024374-HAP2
TRF annot.
RM annot.
DNA
SINE
0
250
500
750
1000
1250
1500
1750
2000
Position (bp)
658 bp INS
61 bp INS
ZLO21LTS29061-HAP1
TRF annot.
RM annot.
DNA
SINE
0
500
1000
1500
2000
Position (bp)
655 bp INS
71 bp INS
ZLO21LTS29067-HAP1
TRF annot.
RM annot.
DNA
SINE
0
500
1000
1500
2000
Position (bp)
656 bp INS
ZLO21LTS29066-HAP1
TRF annot.
RM annot.
DNA
SINE
0
250
500
750
1000
1250
1500
1750
2000
Position (bp)
657 bp INS
61 bp INS
ZLO21LTS29063-HAP1
TRF annot.
RM annot.
DNA
SINE
0
500
1000
1500
2000
Position (bp)

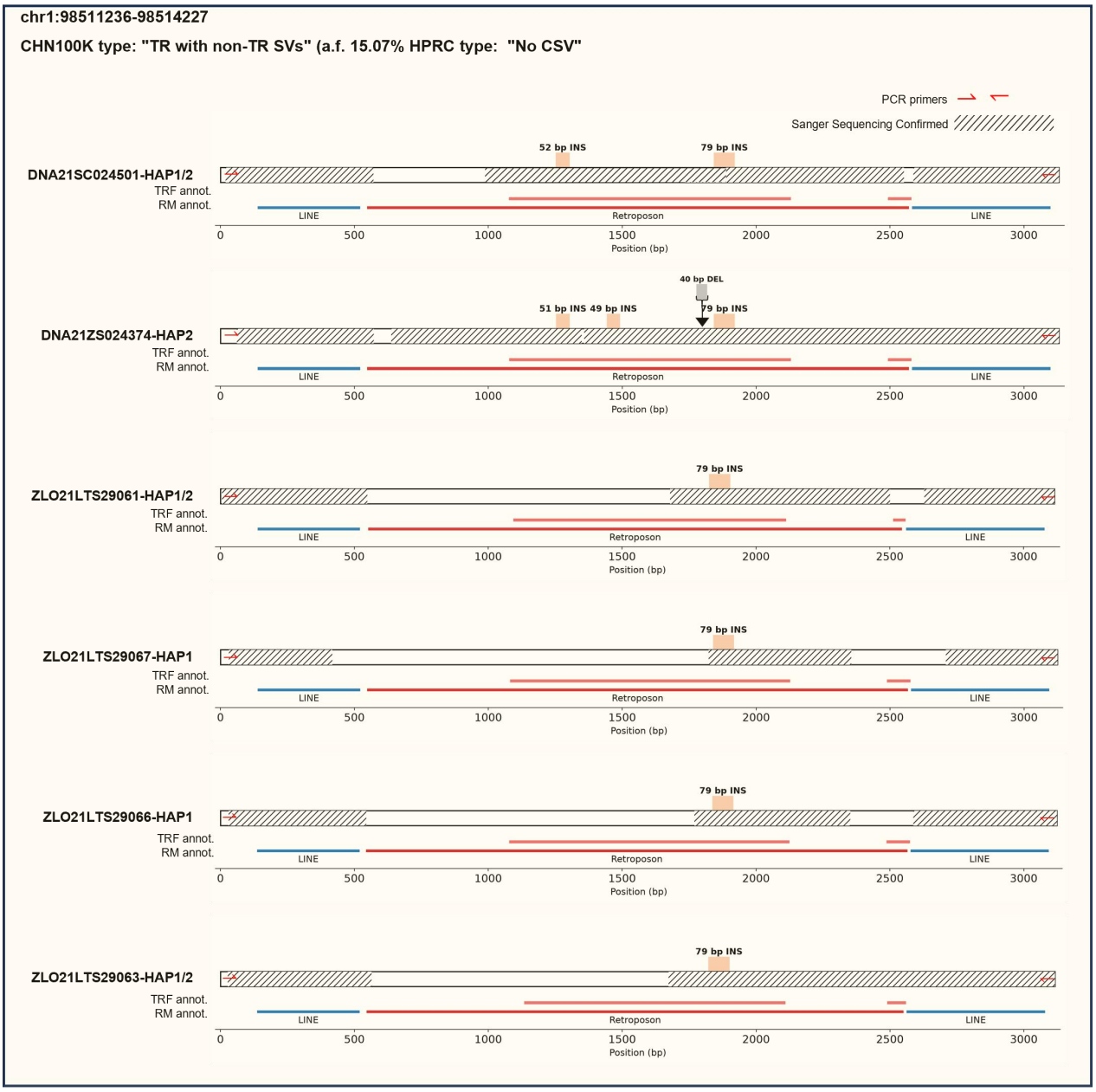
c


chr13:104524134-104525708
CHN100K type: "TR with non-TR SVs" (a.f. 26.3%) HPRC type: "Pure TR"
PCR primers
Sanger Sequencing Confirmed
412 bp INS
DNA21SC024501-HAP1/2
TRF annot.
RM annot.
LTR
Simple_repeat
LTR
Simple_repeat
LTR
0 250 500 750 1000 1250 1500 1750 2000
Position (bp)
54 bp INS
148 bp INS
244 bp INS
DNA21ZS024374-HAP1/2
TRF annot.
RM annot.
LTR
LTR
0 250 500 750 1000 1250 1500 1750 2000
Position (bp)
263 bp INS
198 bp INS
ZLO21LTS29061-HAP1
TRF annot.
RM annot.
LTR
Simple_repeat
LTR
Simple_repeat
LTR
0 250 500 750 1000 1250 1500 1750 2000
Position (bp)
244 bp INS
154 bp INS
ZLO21LTS29067-HAP1
TRF annot.
RM annot.
LTR
LTR
0 250 500 750 1000 1250 1500 1750 2000
Position (bp)
262 bp INS
210 bp INS
ZLO21LTS29066-HAP1
TRF annot.
RM annot.
LTR
Simple_repeat
LTR
0 250 500 750 1000 1250 1500 1750 2000
Position (bp)
54 bp INS
144 bp INS
244 bp INS
ZLO21LTS29063-HAP1/2
TRF annot.
RM annot.
LTR
Simple_repeat
LTR
Simple_repeat
LTR
0 250 500 750 1000 1250 1500 1750 2000
Position (bp)

d


chr1:50380198-50382677
CHN100K type: "TR with non-TR SVs" (a.f. 81.6%); HPRC type: "TR with non-TR SVs"
PCR primers
Sanger Sequencing Confirmed
DNA21SC024501-HAP1/2
DNA21ZS024374-HAP1/2
ZLO21LTS29061-HAP1
ZLO21LTS29067-HAP1
ZLO21LTS29066-HAP1
ZLO21LTS29063-HAP2
TRF annot.
RM annot.
77 bp INS
51 bp INS
54 bp INS
50 bp INS
164 bp INS
LINE
Simple_repeat
Retroposon
0
500
1000
1500
2000
2500
Position (bp)

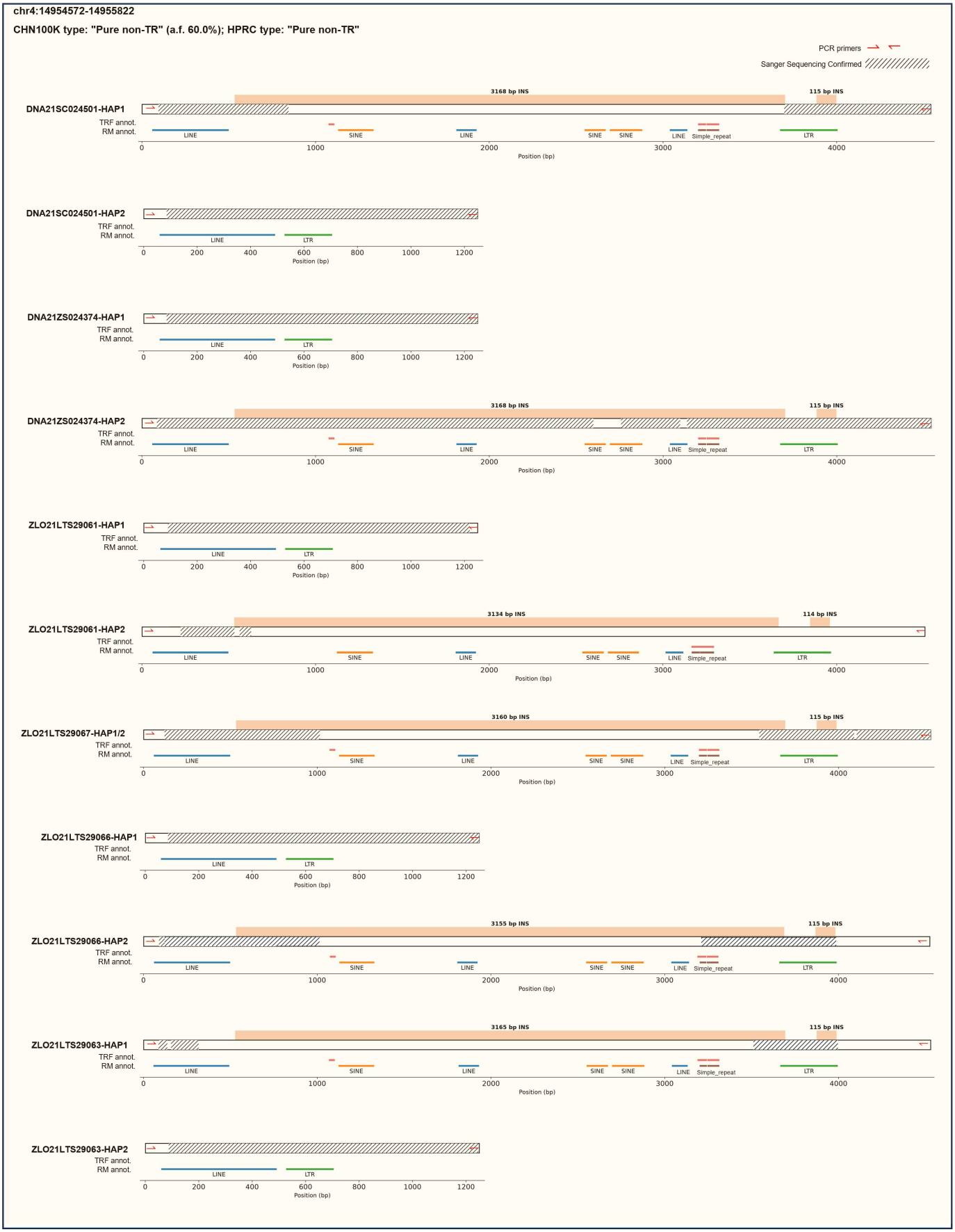
e


chr1:98511236-98514227
CHN100K type: "TR with non-TR SVs" (a.f. 15.07% HPRC type: "No CSV"
PCR primers
Sanger Sequencing Confirmed
DNA21SC024501-HAP1/2
52 bp INS
79 bp INS
DNA21ZS024374-HAP2
40 bp DEL
51 bp INS
49 bp INS
79 bp INS
ZLO21LTS29061-HAP1/2
79 bp INS
ZLO21LTS29067-HAP1
79 bp INS
ZLO21LTS29066-HAP1
79 bp INS
ZLO21LTS29063-HAP1/2
79 bp INS
TRF annot.
RM annot.
LINE
Retroposon
LINE
0
500
1000
1500
2000
2500
3000
Position (bp)

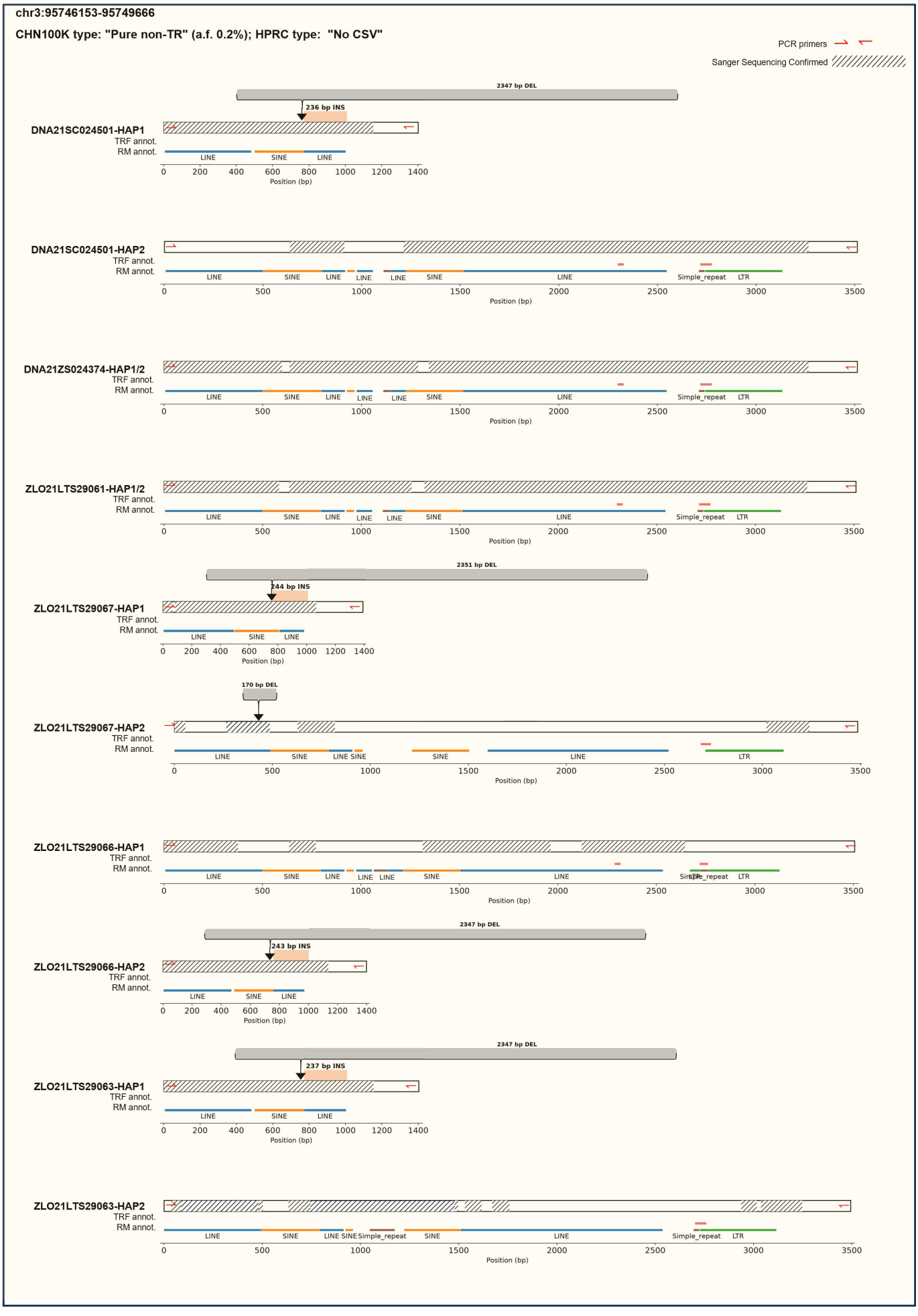
f


chr3:174210812-174212395
HN100K type: "TR with non-TR SVs" (a.f. 77.15%) HPRC type: "TR with non-TR SVs"
PCR primers
Sanger Sequencing Confirmed
DNA21SC024501-HAP1
TRF annot.
RM annot.
DNA
LTR
0 200 400 600 800 1000 1200 1400 1600
Position (bp)
DNA21ZS024374-HAP2
119 bp INS
119 bp INS
TRF annot.
RM annot.
DNA
LTR
0 250 500 750 1000 1250 1500 1750
Position (bp)
ZLO21LTS29061-HAP1/2
118 bp INS
121 bp INS
TRF annot.
RM annot.
DNA
LTR
0 250 500 750 1000 1250 1500 1750
Position (bp)
ZLO21LTS29067-HAP2
122 bp INS
116 bp INS
TRF annot.
RM annot.
DNA
LTR
0 250 500 750 1000 1250 1500 1750
Position (bp)
ZLO21LTS29066-HAP1/2
118 bp INS
119 bp INS
TRF annot.
RM annot.
DNA
LTR
0 250 500 750 1000 1250 1500 1750
Position (bp)
ZLO21LTS29063-HAP1/2
119 bp INS
117 bp INS
TRF annot.
RM annot.
DNA
LTR
0 250 500 750 1000 1250 1500 1750
Position (bp)

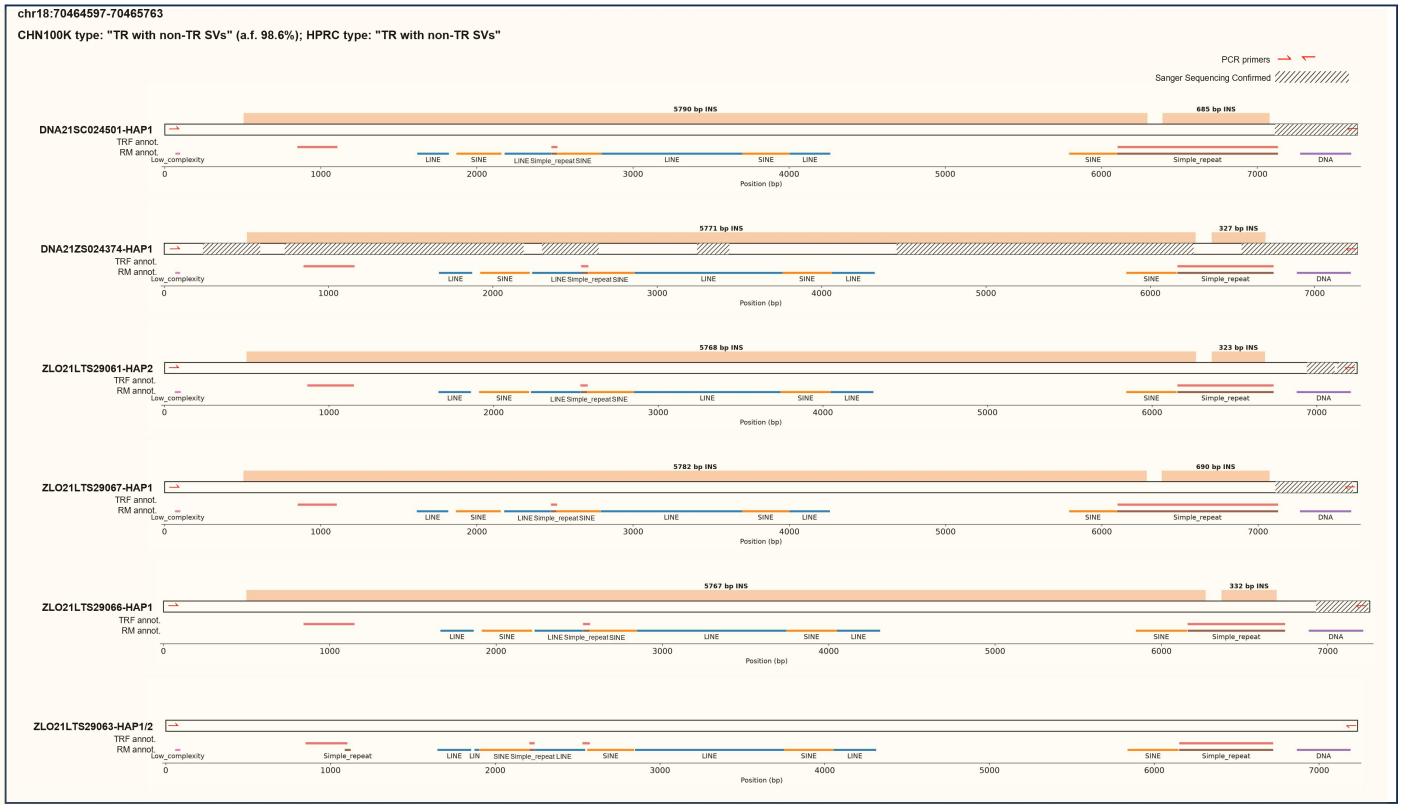
g


chr4:14954572-14955822
CHN100K type: "Pure non-TR" (a.f. 60.0%); HPRC type: "Pure non-TR"
PCR primers
Sanger Sequencing Confirmed
DNA21SC024501-HAP1
DNA21SC024501-HAP2
DNA21ZS024374-HAP1
DNA21ZS024374-HAP2
ZLO21LTS29061-HAP1
ZLO21LTS29061-HAP2
ZLO21LTS29067-HAP1/2
ZLO21LTS29066-HAP1
ZLO21LTS29066-HAP2
ZLO21LTS29063-HAP1
ZLO21LTS29063-HAP2
TRF annot.
RM annot.
3168 bp INS
115 bp INS
3134 bp INS
114 bp INS
3160 bp INS
3155 bp INS
3165 bp INS
LINE
SINE
Simple_repeat
LTR
Position (bp)

h

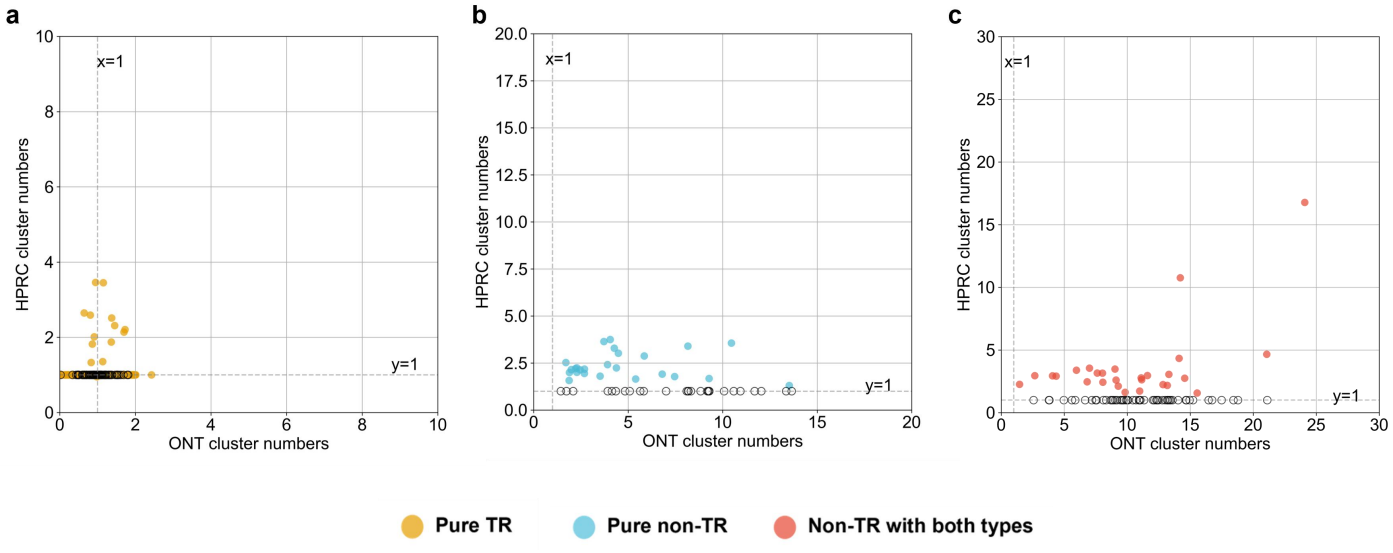

chr3:95746153-95749666
CHN100K type: "Pure non-TR" (a.f. 0.2%); HPRC type: "No CSV"
PCR primers
Sanger Sequencing Confirmed
DNA21SC024501-HAP1
TRF annot.
RM annot.
2347 bp DEL
236 bp INS
LINE
SINE
LINE
Position (bp)
DNA21SC024501-HAP2
Simple_repeat
LTR
DNA21ZS024374-HAP1/2
ZLO21LTS29061-HAP1/2
ZLO21LTS29067-HAP1
2351 bp DEL
244 bp INS
ZLO21LTS29067-HAP2
170 bp DEL
ZLO21LTS29066-HAP1
ZLO21LTS29066-HAP2
2347 bp DEL
243 bp INS
ZLO21LTS29063-HAP1
2347 bp DEL
237 bp INS
ZLO21LTS29063-HAP2

i



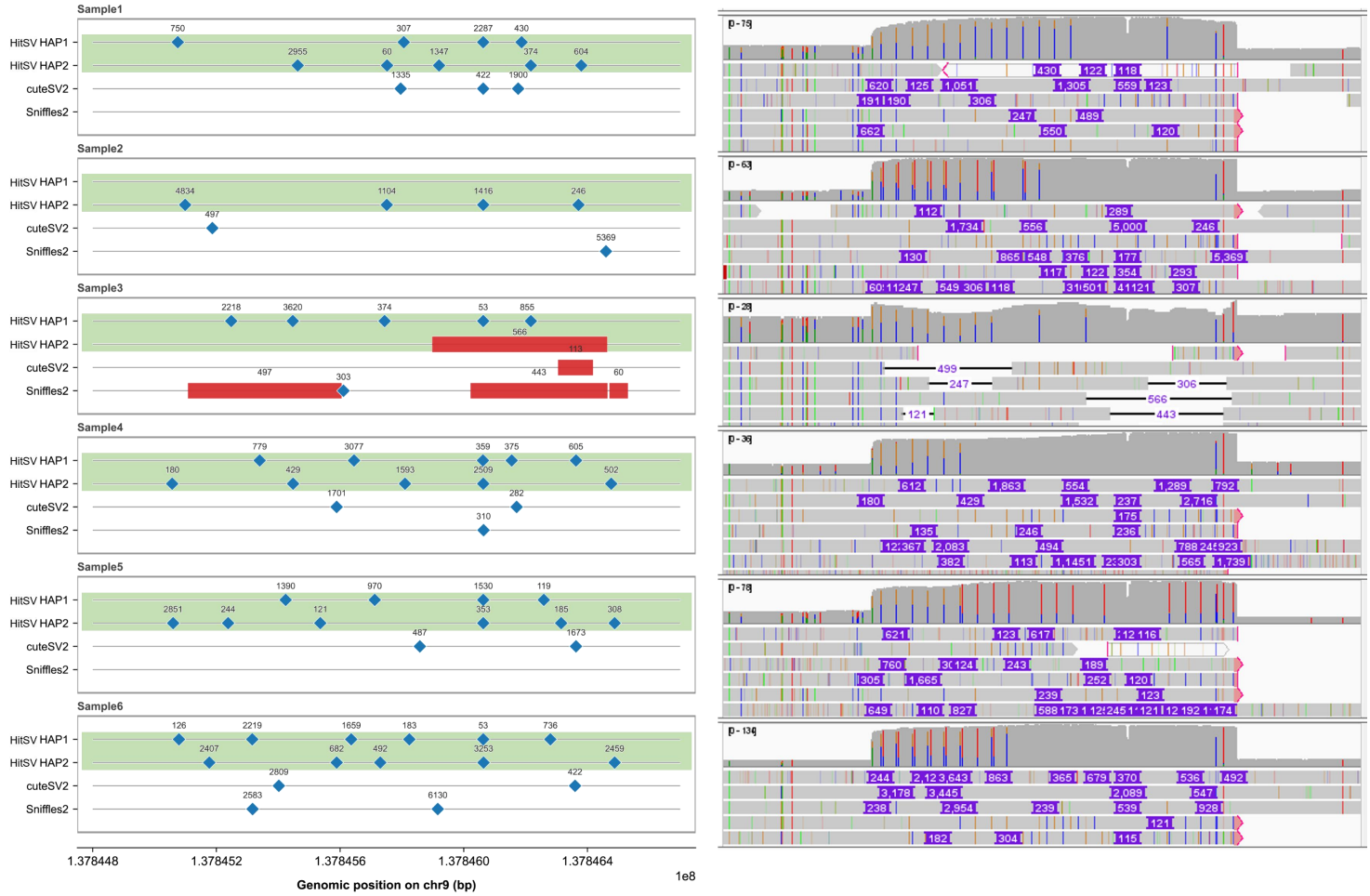
Supplementary Fig. 18. Experimental validation for a subset of CHN100K CSV alleles.

a, The polymerase chain reaction (PCR) amplification of the target intervals; **b-i,** Sanger sequencing of validated CHN100K CSV loci.



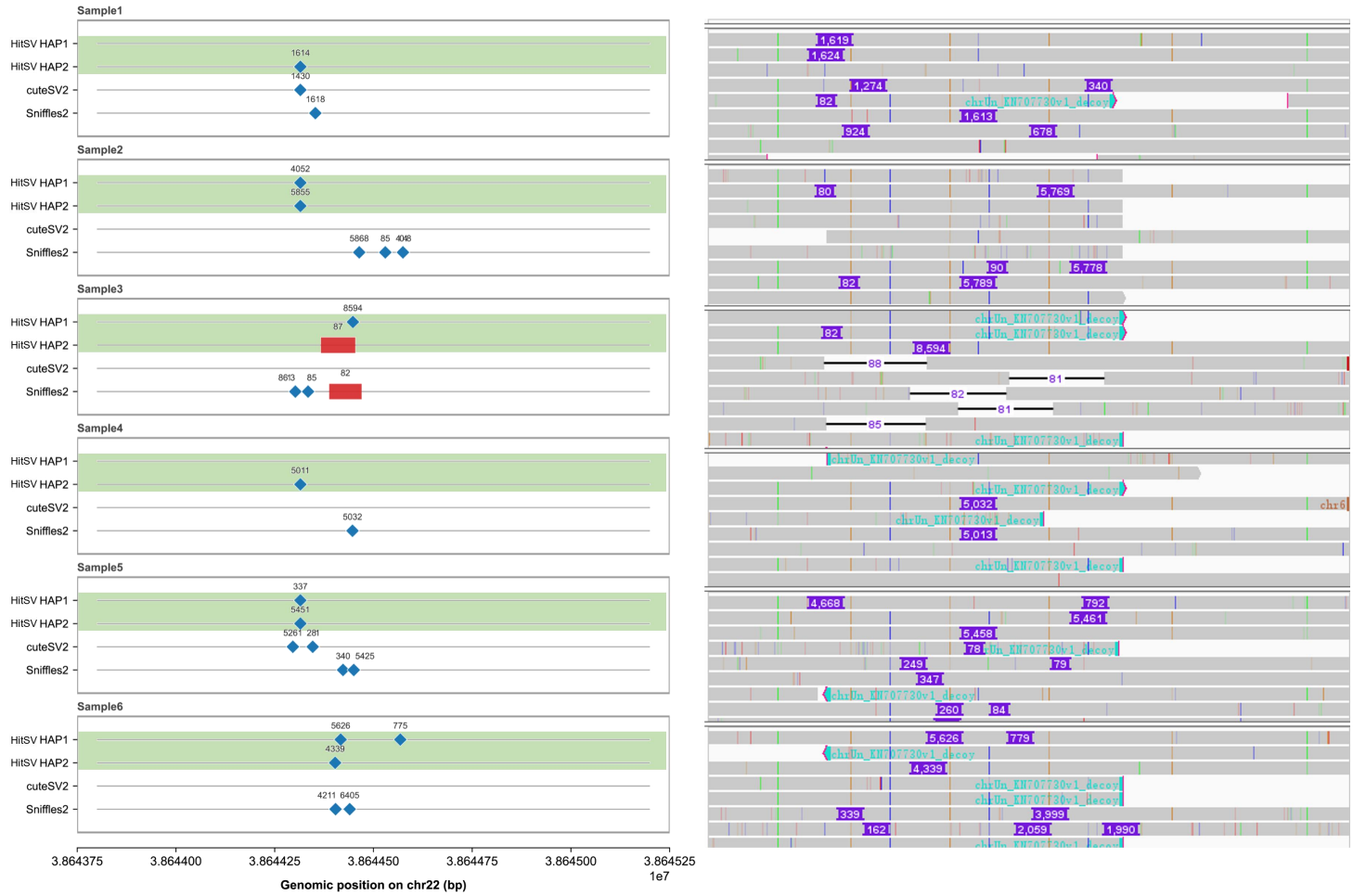
Supplementary Fig. 19. Comparison of clustering complexity for CSV regions in the ONT-HPRC datasets.

Scatter plots compare the number of clusters derived from Oxford Nanopore Technologies (ONT) data (x-axis) with the number of clusters inferred from HPRC data (y-axis) across different classes of CSV regions. Each point represents one CSV region. Dashed lines indicate the reference threshold of a single cluster (x = 1 or y = 1).
a, Pure TR regions; **b,** Pure non-TR regions; **c,** Non-TR with both types regions.



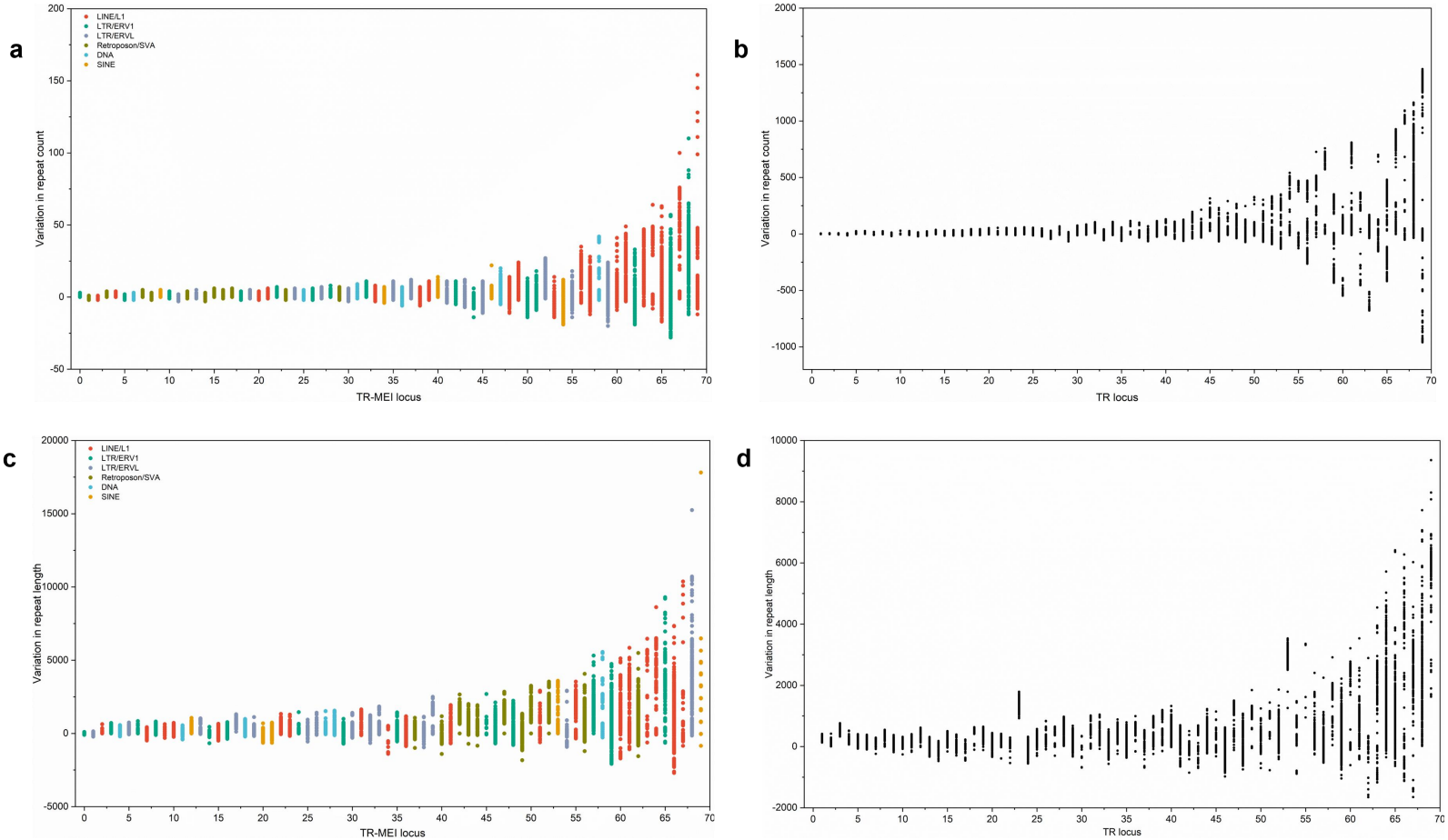
Supplementary Fig. 20. Detection of TR-MEI–associated complex structural variants across multiple ONT samples.

Another example focuses on a TR-MEI–related region on human chromosome 9 (chr9:137,844,500-137,847,000 bp), illustrating CSV detection results across six samples from the 557-ONT dataset. The left panel summarizes the structural variants detected by different tools across samples and haplotypes. Haplotype-resolved calls produced by HitSV enable explicit assignment of CSVs to individual haplotypes, facilitating the tracing and comparison of complex variant patterns across samples. In contrast, other tools lack haplotype resolution and report aggregated SV calls. The right panel presents IGV snapshots of the same region for multiple samples, showing read alignments and supporting signals consistent with TR-MEI–associated CSV events.



Supplementary Fig. 21. Detection of TR-MEI–associated complex structural variants across multiple ONT samples.

Another example focuses on a TR-MEI–related region on human chromosome 22 (chr22:38,643,500–38,645,500 bp), illustrating CSV detection results across six samples from the 557-ONT dataset. The left panel summarizes the structural variants detected by different tools across samples and haplotypes. Haplotype-resolved calls produced by HitSV enable explicit assignment of CSVs to individual haplotypes, facilitating the tracing and comparison of complex variant patterns across samples. In contrast, other tools lack haplotype resolution and report aggregated SV calls. The right panel presents IGV snapshots of the same region for multiple samples, showing read alignments and supporting signals consistent with TR-MEI–associated CSV events.



Supplementary Fig. 22. Repeat unit count and length variability of tandem repeat (TR) regions.

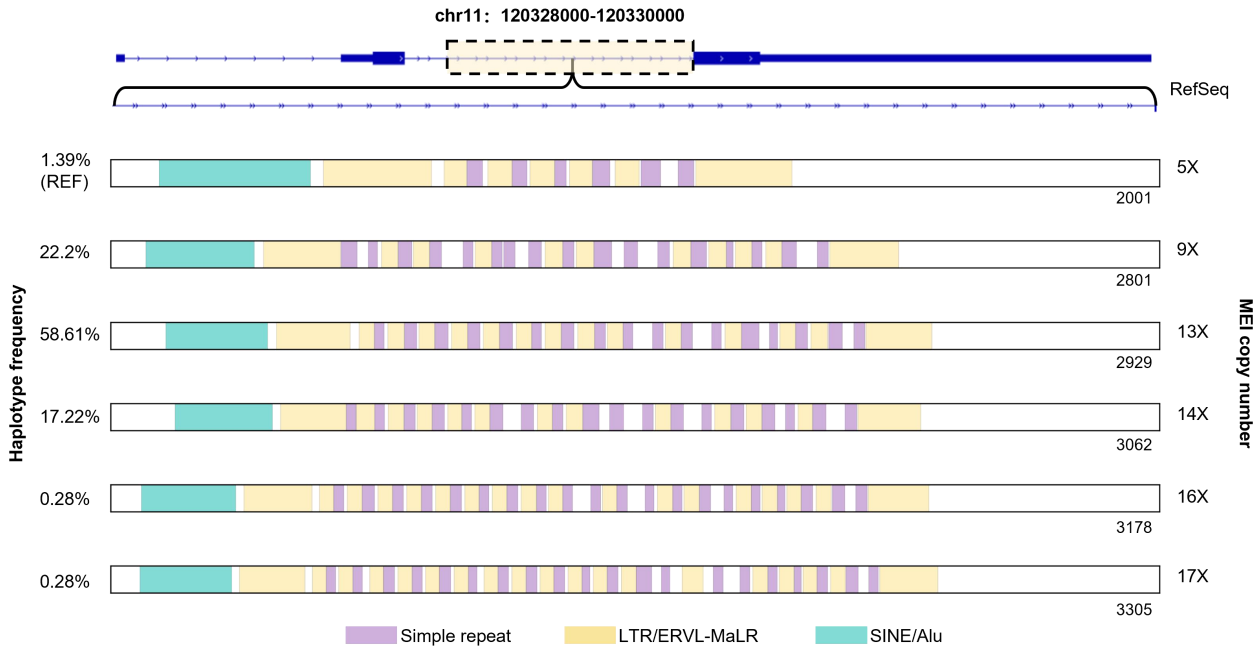
The x-axis denotes individual TR loci, and the y-axis indicates variability relative to the reference genome sequence.

a, Variability in the number of repeat units within TR-MEI arrays. Each point represents one TR-MEI interval, with colors indicating the predominant MEI type associated with the interval.

b, Variability in the number of repeat units across regular TR regions lacking MEI insertions.

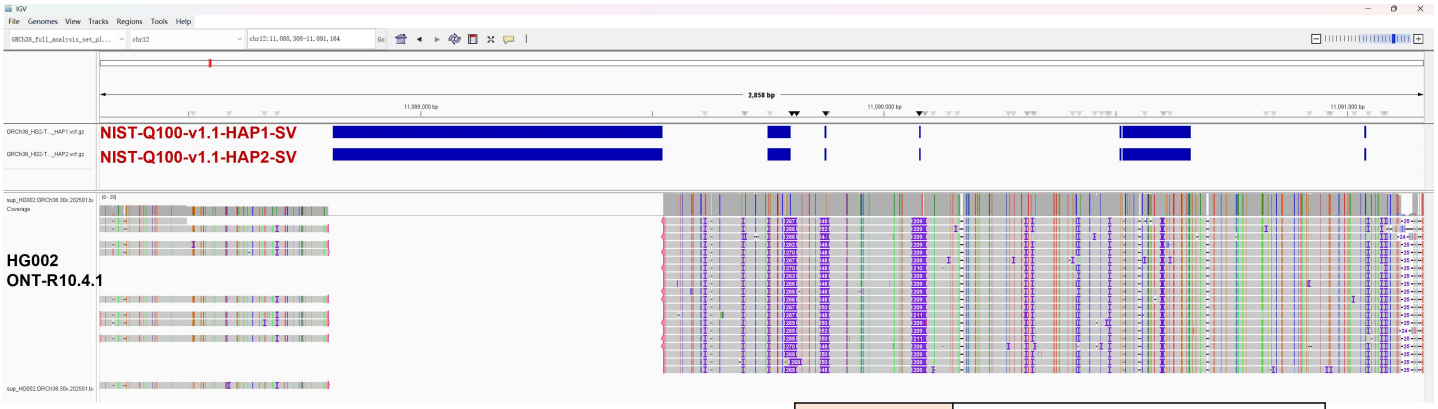
c, Length variability of TR-MEI arrays.

d, Length variability of regular TR regions without MEI insertions.



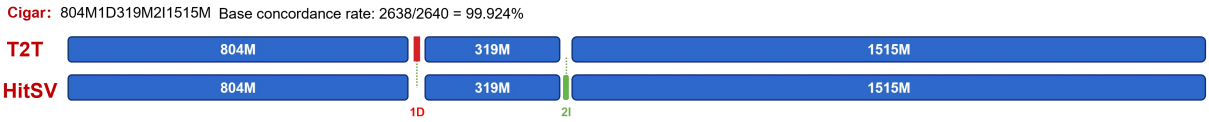
Supplementary Fig. 23. Haplotype-resolved characterization of a TR-MEI array at the *TLCD5* locus.

The analysis focuses on a TR-MEI–associated region spanning the *TLCD5* locus on human chromosome 11 (chr11:120,328,000–120,330,000 bp), illustrating structural variation across samples from the 557-ONT dataset. The upper panel shows the genomic context of the locus, highlighting the position of the TR-MEI array within an intronic region of TLCD5. The lower panel displays RepeatMasker annotations of representative locally assembled contigs from the reference genome and multiple samples, visualized as colored rectangular bars (contig lengths are indicated beneath each bar). Numbers shown on the left of each contig denote the allele frequency of the corresponding assembly across the full 557-ONT cohort, whereas numbers on the right indicate the repeat counts of MEI arrays within each contig. The TR-MEI arrays at this locus are primarily composed of LTR/ERVL elements interspersed with simple repeat sequences, illustrating haplotype-specific expansion and contraction patterns of TR-MEI structures.



T2T hap1	714D + 318I + 49D + 50I + 220I + 152I + 150D + 107I
T2T hap2	714D + 318I + 49D + 48I + 220I + 152I + 150D + 107I

CuteSV2	713D (1/1) + 315I(1/1) + 209I(1/1)
SVDSS2	266I (0/0) + 48I (0/0) + 52I (0/0) + 209I(0/0)
sawfish	715D (1/1) + 266I(1/1) + 49I(1/1)
Sniffles2	NA
Svision-PRO	NA



Supplementary Fig. 24. A CSV reconstructed by HitSV but hindered by contig re-alignment.

The upper panel shows the IGV snapshot of a HG002 CSV in the chromosome 12 (chr12:11088308-11091164). According to the ground truth, this CSV has two highly similar SV alleles consisting of a number of SV events (3 deletions and 5 insertions). cuteSV2, SVDSS2 and sawfish only recovered a part of the events while Sniffles and Svision-pro failed to detect at all. HitSV successfully reconstructed the whole SV allele which can achieve very high base concordance comparing to the ground truth (99.924%, the sequence alignment is illustrated in the lower panel). However, due to the large divergence between the SV allele and local reference, the employed contig aligner (ksw2) failed for the SV contig realignment even if highly tolerant parameters (ASM20) were used. We realized that this CSV was also successfully detected, and more robust alignment tool could be also necessary for SV discovery, especially for those highly complex ones.

Supplementary Note 1. The characterization of the signature distributions

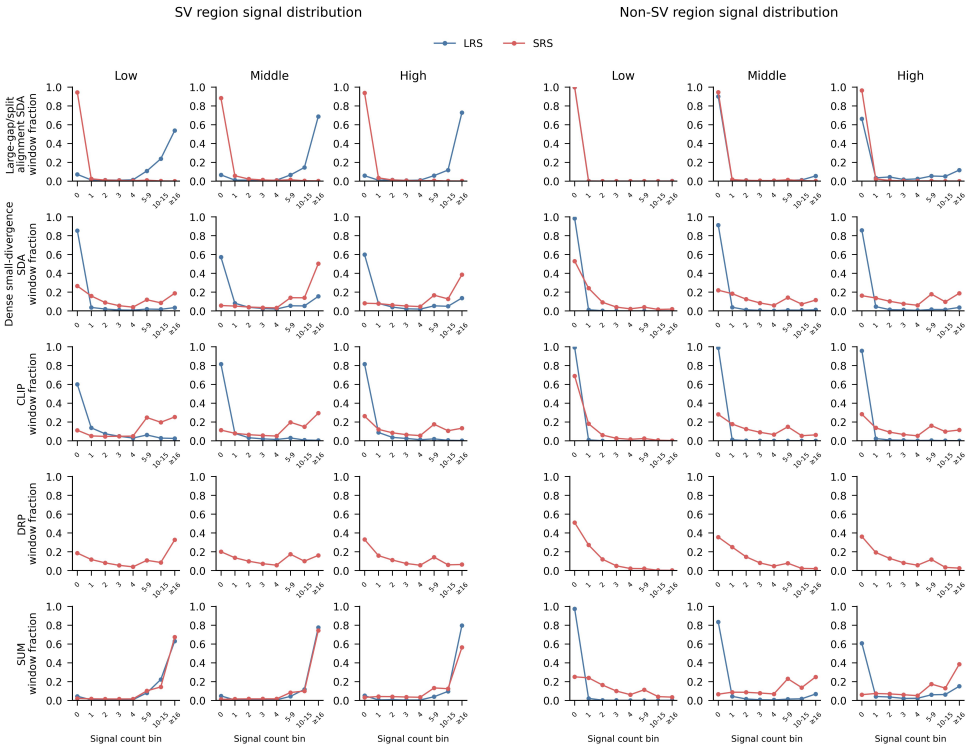
The signature distributions across sequencing (LRS/SRS) and window (SV/non-SV) categories were compared to characterize the various kinds of signatures in detail and systematically explain the observed heterogeneity of SV signatures. Four major observations are made as following.

Firstly, different signature patterns occur for LRS and SRS. The signature distributions for LRS and SRS were divided and separately compared by three ranges of local repetitiveness (i.e., low: 1≤R<40, medium: 40≤R<120 and high: R>120, Supplementary Fig. 25). Long insertion/deletion signatures are dominant in LRS data and relatively small number of spanning reads have large clippings. On the contrary, a large proportion of short SV-spanning reads have CLP signatures, meanwhile, dense mismatches and small indels are also frequently observed in the alignments instead of long insertions/deletions. Moreover, short reads also have distinctive DRP signatures, especially significant for those low R windows.

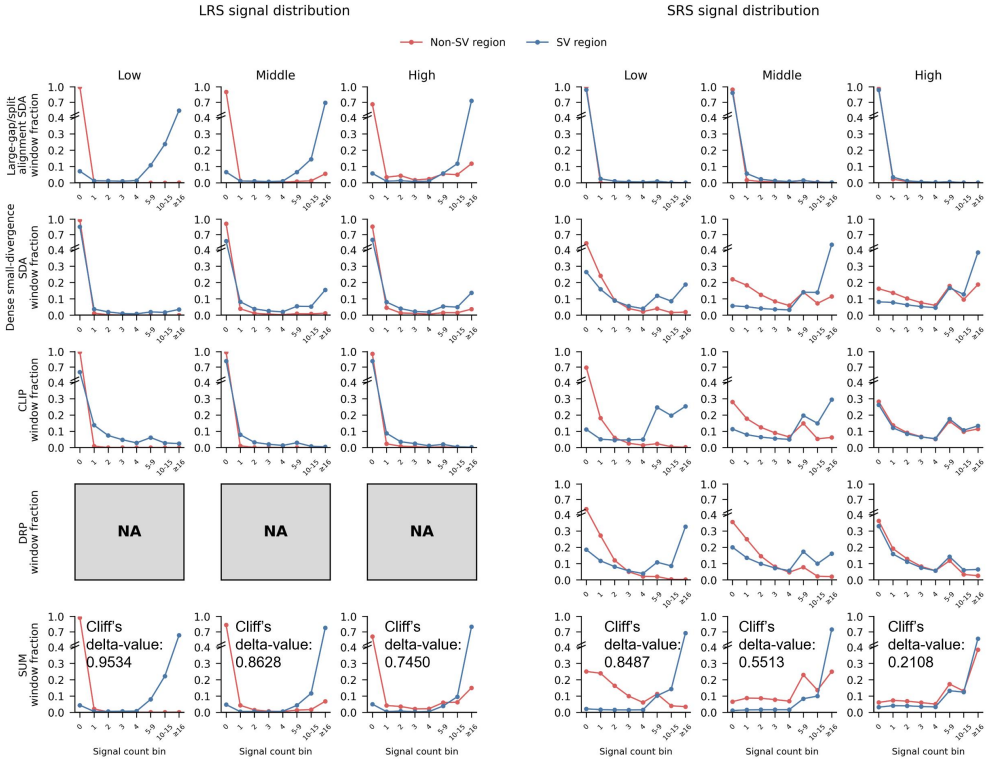
Secondly, signature evidence reduces with increasing local repetitiveness. The signature distributions for SV and non-SV windows under various local repetitiveness ranges are shown in Supplementary Fig. 26. In the windows having low R values, the difference in combined signature (SUM, last row of Supplementary Fig. 26) between SV and non-SV windows are significant (Cliff D rank test, Cliff's delta-value: 0.9534 for LRS and 0.8487 for SRS), while it obscures for high R windows. For both of LRS and SRS, the proportion of non-SV windows having SDA and CLP signatures increase in repetitive regions. This could be primarily caused by incorrect read mapping, i.e., repetitive SV-spanning reads are co-linearly aligned to wrong repeat copies. Most of such alignments have dense small divergences and also a proportion of them may have long insertion/deletions (LRS) or clippings (SRS). This is caused by the fact that the alignments can have higher scores than that of correct SV-spanning alignments since the latter ones need to afford the large penalties of non-colinear or partial alignments to span SV events. However, this issue has little impact on detection, since SUM evidence is still prevalent across most SV windows. Moreover, SVs usually span multiple adjacent windows, allowing sufficient evidence to be captured in adjacent regions. It is only except for SRS in extremely repetitive regions (R>120) since the alignments are more deteriorated and the signature discrepancy between SV- and non-SV windows is quite weak.

Thirdly, difficult regions are related to well-known repeats. We investigated the repeat and gene annotations for all the windows, using the official Repeat Masker, TRF and Ensemble Gene annotation files for GRCh38 (Supplementary Fig. 27). Overall, most of the medium to high R windows are related to well-known repeats and SV hotspots. In details, a large proportion of such windows can be annotated by Repeat Masker as transposons/retrotransposons like SINEs, LINEs and LTRs (Supplementary Fig. 27b), and tandem duplications marked as simple repeat, microsatellite and low complexity sequence (Supplementary Fig. 27c). Moreover, most of these windows are largely overlapped with TRF annotations (Supplementary Fig. 28), also indicating that they are extensively involved in tandem repeats. It is also worthnoting that the lengths of repeat units are various and not correlated with R. With these annotations, R metric can be an indicator to “SV-hard” regions and the signature distributions can also be used as a reference to explain the attenuation of signal and assess the difficulty of SV calling in regions of interest.

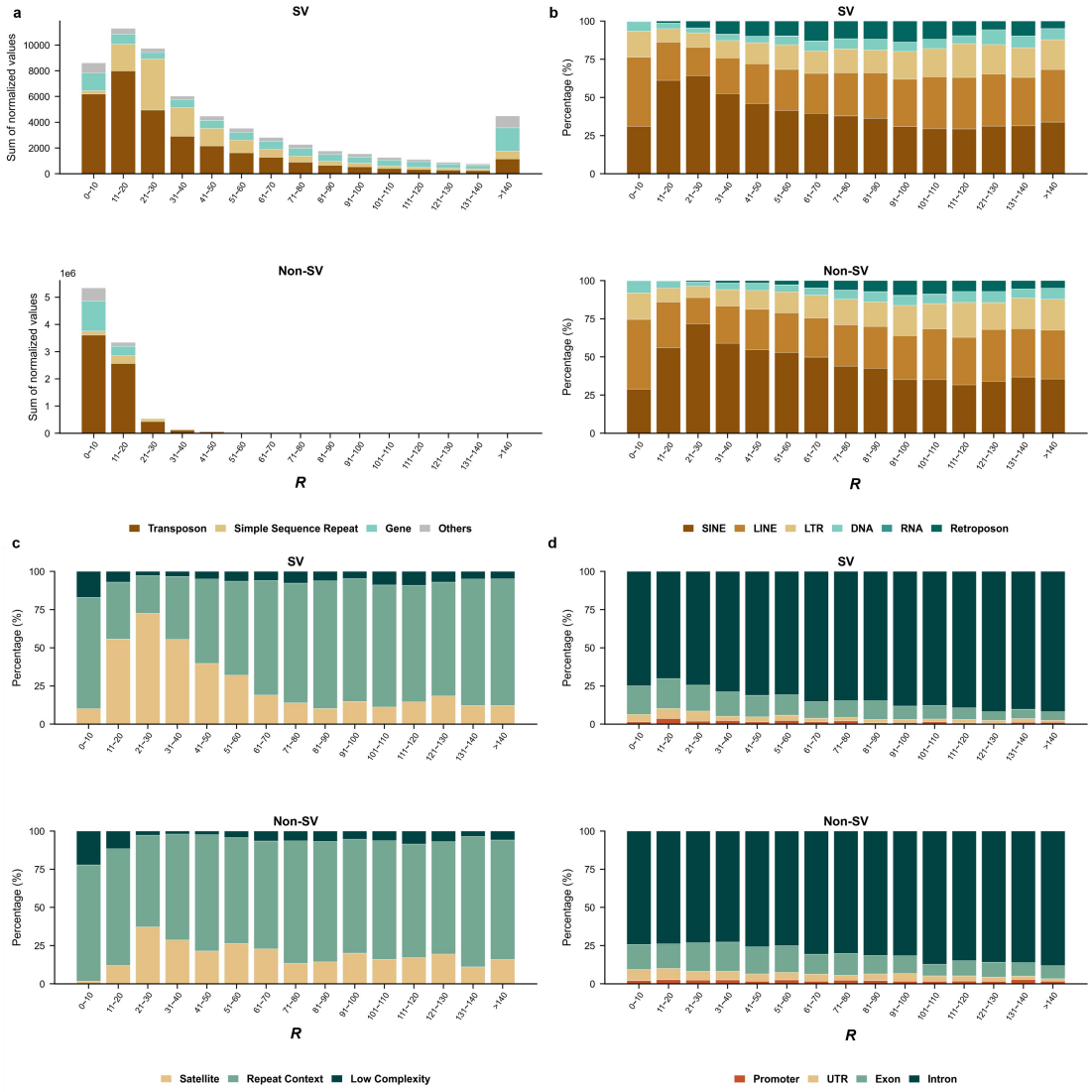
Fourthly, the signature distributions provide a quantitative approach to SV discovery. With the distributions, the yield of SV site recognition using various alignment signatures can be empirically estimated which could be applied to many studies. We further plotted the ROC curves of the sum of the signatures (SUM, Supplementary Fig. 29). The AUC values (0.975 for LRS and 0.929 for SRS) indicate that the SUM metric is useful, meanwhile, the curves can also be a reference to assess the tradeoff between the sensitivity and specificity during SV detection.



Supplementary Fig. 25. The comparison on the SV signatures between LRS and SRS data in SV and non-SV regions. Each of the plots indicate the fractions of the windows for a kind of SV signature (rows: SDA-dense, SDA-gap, CLP, DRP and SUM, respectively) in a specific range of local repetitiveness (columns: Low, 0<R<40; Medium: 40≤R<120; High: R≥120, respectively), and the blue and red lines indicate the plots for LRS and SRS respectively for head-to-head comparison. The left and right panels are for the comparison in SV and non-SV contexts, respectively. In each line, the dots indicate the fractions of the windows having a certain number (respectively 0, 1, 2, 3, 4, 5-9, 10-15 and >15) of reads containing the corresponding SV signature, which can be treated as an intensity distribution of the signature.

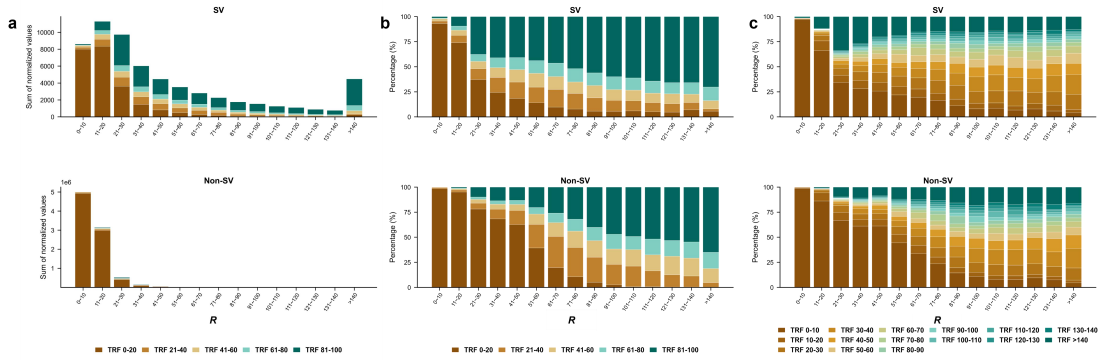


Supplementary Fig. 26. The comparison on the SV signatures between SV- and non-SV windows for LRS and SRS data. Each of the plots indicate the fractions of the windows for a kind of SV signature (rows: SDA-dense, SDA-gap, CLP, DRP and SUM, respectively) in a specific range of local repetitiveness (columns: Low, 0<R<40; Medium: 40≤R<120; High: R≥120, respectively), and the blue and red lines indicate the plots for SV and non-SV respectively for head-to-head comparison. The left and right panels are for the comparison in LRS and SRS data types, respectively. In each line, the dots indicate the fractions of the windows having a certain number (respectively 0, 1, 2, 3, 4, 5-9, 10-15 and >15) of reads containing the corresponding SV signature, which can be treated as an intensity distribution of the signature.

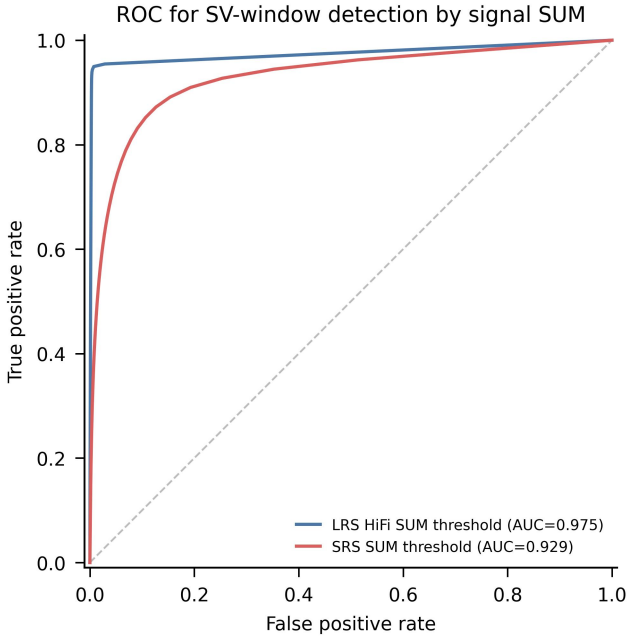


Supplementary Fig. 27. The distributions of SV and non-SV windows in various annotation terms. The genomic windows of all the distribution-modeling samples were annotated by Repeat Masker and Ensemble Gene and categorized to plot the distributions of annotated SV and non-SV windows (upper and lower sub-panels, respectively). **a,** The distributions of windows related to various genome annotations (i.e., Transposon, Simple Sequence Repeat, Gene and others). Each of the stacked bars indicate the number of the windows in absolute terms for a specific range of local repetitiveness (R metric) and each of the colored blocks indicate the number of windows belonging to a specific kind annotation. **b,** The distributions of windows related to specific categories of Transposon annotation (i.e., SINE, LINE, LTR, DNA, RNA and Retroposon). Each of the stacked bars indicate the percentage of the windows for a specific range of local repetitiveness (R metric) and each of the colored blocks indicate the percentage of the windows belonging to a specific kind annotation. **c,** The distributions of windows related to specific categories of Simple Sequence Repeat annotation (i.e., Satellite, Repeat Context and Low Complexity). Each of the stacked bars indicate the percentage of the windows for a specific range of local repetitiveness (R metric) and each of the colored blocks indicate the percentage of the windows belonging to a specific kind annotation. d. The distributions of windows related to specific categories of GENE annotation (i.e., Promoter, UTR, Exon and Intron). Herein, for a specific gene, the Intron regions was inferred by the whole region subtracting the corresponding promoter, UTR and exon

regions. Each of the stacked bars indicate the percentage of the windows for a specific range of local repetitiveness (R metric) and each of the colored blocks indicate the percentage of the windows belonging to a specific kind annotation.



Supplementary Fig. 28. The distributions of SV and non-SV windows in various tandem repeat coverages. The genomic windows of all the distribution-modeling samples were annotated by TRF to investigate the local coverages of tandem repeats. a-b. The distributions of windows having various coverage of tandem repeat coverages. Each of the stacked bars indicate the distributions of windows for a specific range of local repetitiveness (R metric). Each colored block indicates the numbers (a) and fractions (b) of the windows covered by a specific percentage of tandem repeats (i.e., 0-20%, 20-40%, 40-60%, 60-80%, 80-100%, respectively). c. The distributions of windows having various lengths of repeat unit. Each of the stacked bars indicate the distributions of windows for a specific range of local repetitiveness (R metric). Each colored block indicates the fractions of the windows having tandem repeats having specific unit lengths (grouped by every 10bp).



Supplementary Fig. 29. The ROC curves by SUM signature. The blue and red lines respectively indicate the ROC curves for LRS and SRS using various thresholds of SUM signatures for SV site recognition. The AUC values are 0.975 (LRS) and 0.929 (SRS), respectively.

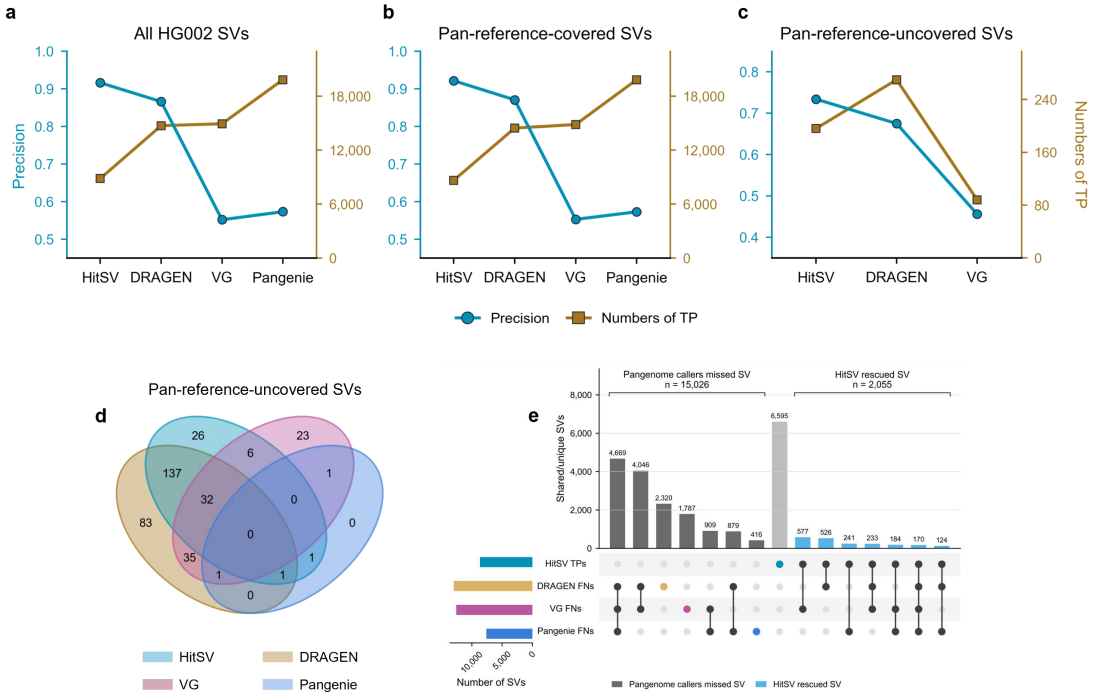
Supplementary Note 2. Comparison between HitSV and Pangenome-based SV callers on SRS data

We compared HitSV with three pangenome-based callers, i.e., PanGenie, VG and DRAGEN on HG002 by an additional assessment on pan-reference-covered and HG002 specific SVs). Specifically, we divided HG002 truth-set SVs into two categories:

(i) pan-reference-covered SVs (27424 in total), defined as HG002 truth-set SVs that are represented in the 44 HPRC-based pangenome reference (according to PanGenie's official pangenome index) used by the pangenome genotyping methods; and

(ii) pan-reference-uncovered, or HG002-specific SVs (733 in total), defined as HG002 truth-set SVs that are not represented in that pangenome reference.

The sensitivities and precisions of the callers on the two categories of SVs were separately assessed (Supplementary Fig. 30). Overall, pangenome-informed approaches can achieve high sensitivity for SVs represented in the pangenome reference, whereas HitSV shows three distinct advantages as a de novo caller.



Supplementary Fig. 30: Comparisons between HitSV and three pangenome-based callers. a-c, The precisions, number of true positives of various callers for (**a**) all SVs, (**b**) pan-reference-covered SVs, and (**c**) pan-reference-uncovered SVs. **d,** Venn diagram of pan-reference-uncovered SVs called by different callers. **e,** Upset plot between HitSV true positive calls and false negative calls from DRAGEN, VG, and Pangenie, illustrating the number of pan-reference-covered SVs rescued by HitSV.

Firstly, HitSV has outstanding overall precision (HitSV: 91.61%, PanGenie: 57.32%, VG: 55.19%, DRAGEN: 86.60%, respectively, Supplementary Fig. 30a and Supplementary Table 8). PanGenie and VG output many false positive calls on known SVs of HPRC samples which are recorded in their pangenome references (precisions on pan-reference-covered SVs, HitSV: 92.15%, PanGenie: 57.30%, VG: 55.26%, DRAGEN: 87.07%, Supplementary Fig. 30b). Considering their pangenome structure-leveraging mechanism, these mistakes could be due to the crosstalk between the known SV alleles and noisy read alignments, especially in repetitive regions. However, HitSV keeps high precision along the whole genome and leaves much fewer false positives, primarily owning to its tailored de novo discovery design which handles each of SV candidates more conservatively and asks for sufficient evidences to make the call.

Secondly, HitSV can make distinct contributions on novel SV discovery. A Venn diagram analysis (Supplementary Fig.30d) indicate that all the four callers (HitSV, PanGenie, VG and DRAGEN) can distinctively recover a set of HG002 SVs

which are not recorded by the 45 HPRC samples-based reference, although the numbers for VG and PanGenie are very few due to their genotyping-only design. Importantly, the novel SVs identified by HitSV exhibited a higher precision than DRAGEN (73.41% vs 67.50%, Supplementary Fig. 30c). We further annotated the distinct SV calls (26) of HitSV and found that, most of them are related to mobile elements insertions (LTR, SINE and LINE), simple repeats and tandem duplications, which could be still difficult to existing tools (Supplementary Table 40). At this point, the ability of HitSV is also valuable, especially considering that PanGenie and VG are mainly designed for genotyping but not de novo discovery, and such cases are not applicable for them.

Thirdly, although pangenome has been used, some of known SVs in individual genomes could still be missed by pangenome-based callers, and HitSV can rescue a part of them. We did a false negative analysis and observed that all of PanGenie, VG and DRAGEN still has high numbers (6497, 11221, 10747, respectively) of false negative calls for the HG002 SVs already being recorded in the pangenome reference (Supplementary Fig. 30e and Supplementary Table 10). HitSV can help to rescue 719, 1164 and 1053 of them, respectively. More importantly, 4839 known HG002 SVs were missed by all of the three pangenome-based callers and HitSV can recover 170 of them. At this point, HitSV has the potential to be an important participant and play its own role in future SV studies.

Supplementary Note 3. Runtime Environment and Command lines

Description about the running time and memory footprint.

HitSV showed affordable runtime for various data configurations (30x long reads: 0.83 CPU hours, 4×+30×/60× hybrid sequencing data: 3.35 and 7.94 CPU hours, 30×/60× short reads: 3.14 and 7.39 CPU hours, respectively) as well as relatively small memory footprints (30x long reads: 1.23 GB, 4×+30×/60× hybrid sequencing data: 4.76 and 8.39 GB, 30x/60x short reads: 4.24 and 8.29 GB, respectively).

Command lines.

1. The commands and parameters used for read alignment

BWA-MEM

bwa mem -t 10 -K 100000000 -Y ref.fa read1.fq read2.fq > sample.unsorted.sam

pbmm2

pbmm2 align ref.fa samle.fastq.gz sample.sort.bam -j 10 --sort --rg '@RG\tID:ID_002\tSM:HG002'

minimap2

minimap2 -x asm5 -t 10 -a ref.fa samle.fastq.gz > sample.unsorted.sam

samtools

cat sample.unsorted.sam | samtools view -@ 10 -b -o samle.unsorted.bam

samtools sort -@ 10 -m 2000000000 --output-fmt=BAM -o sample.bam sample.unsorted.bam

samtools index sample.bam

2. Down sample with samtools.

samtools

samtools view -s 0.25/0.5 -b -o output.bam input.bam

3. The commands and parameters used for LRS SV calling

HitSV

HitSV call -I input.bam -r ref.fa -o output.vcf 2> /dev/null

sniffle2 default mode

sniffles --input input.bam --vcf output.vcf --reference ref.fa

SVDSS2

SVDSS_linux_x86-64 index --reference ref.fa --index ref.fmd

SVDSS_linux_x86-64 smooth --reference ref.fa --bam input.bam --threads 10 > SMO.bam

samtools index SMO.bam

SVDSS_linux_x86-64 search --index ref.fmd --bam SMO.bam > SPE.txt

SVDSS_linux_x86-64 call --reference ref.fa --bam SMO.bam --sfs SPE.txt --threads 10 > output.vcf

cuteSV2

cuteSV --genotype input.bam ref.fa output.vcf work_dir

SVision-PRO

SVision-pro --detect_mode germline --process_num 10 --target_path sample.bam --genome_path ref.fa

--model_path model_liteunet_256_8_16_32_32_32.pth --out_path workdir --sample_name SAMPLE_NAME

Sawfish

sawfish discover --ref ref.fa --bam sample.bam --output-dir work_dir --threads 10

sawfish joint-call --sample work_dir

4. The commands and parameters used for SRS SV calling

HitSV ref stat *[only run once for all samples]*

HitSV srs_fa_stat ref.fa > ref.stat.txt

HitSV

HitSV srs_trans_reads ref.fa sample.bam TL.bam

samtools sort -@ 10 --output-fmt=BAM -o TL.sort.bam TL.bam

samtools index TL.sort.bam

HitSV call -n input.bam -L TL.sort.bam -r ref.fa -I ref.stat.txt -o output.vcf 2> /dev/null

manta

configManta.py --bam input.bam --referenceFasta ref.fa --runDir manta_work_dir

python manta_work/runWorkflow.py -m local -j 10

delly

delly call -g ref.fa input.bam > output.vcf

lumpy

samtools view -F 1294 input.bam > DIS.bam

samtools sort -@ 16 DIS.bam -o DIS.sort.bam && samtools index DIS.sort.bam

samtools view -h input.bam | ./lumpy-sv/scripts/extractSplitReads_BwaMem -i stdin | samtools view -Sb - > SIP.bam

samtools sort -@ 16 SIP.bam -o SIP.sort.bam && samtools index SIP.sort.bam

lumpyexpress -B input.bam -S SIP.sort.bam -D DIS.sort.bam -o output.vcf

5. The commands and parameters used for Hybrid SV calling

HitSV-9mer_profile [only run once for all samples]

python error_profile_analysis.py HG002.ONT.sort.bam ref.fa HG002_GRCh38_1_22_v4.2.1_benchmark.vcf.gz 9mer_profile

HitSV

HitSV srs_trans_reads ref.fa NGS.input.bam TL.bam

samtools sort --output-fmt=BAM -o TL.sort.bam TL.bam

samtools index TL.sort.bam

HitSV call -l lrs.input.bam -n srs.input.bam -L TL.sort.bam -r ref.fa -I ref.stat.txt -k 9mer_profile -o output.vcf 2> /dev/null

6 The commands and parameters used for benchmarking

bcftools & bgzip

cat tools.vcf | grep -v -E "SVTYPE=BND" | bcftools sort | bgzip -cf > output.vcf.gz

tabix output.vcf.gz

truvari bench

truvari bench --passonly -p 0 -P 0.7/0.9/1--dup-to-in -c cmp.vcf.gz -b base.vcf.gz -o ./result.dir --reference ref.fa --includebed region.bed

truvari refine

truvari bench -p 0.7/0.9/1 -P 0.7/0.9/1 --passonly --pick ac --dup-to-ins --reference ref.fa --includebed region.bed -c cmp.vcf.gz -b base.vcf.gz -o ./result.dir

truvari refine --write-phab --use-region-coords --use-original-vcfs --align mafft --reference ref.fa --regions ./result.dir/candidate.refine.bed refine_output_dir

truvari ga4gh --write-phab --input refine_output_dir --output combined_result

7 Mendelian consistency evaluation.

Run variant detection for different algorithms on different samples (HG002/3/4 and HG005/6/7), then store the VCF file paths produced by the same algorithm in the file "SUR_l.txt", and execute:

SURVIVOR merge SUR_l.txt 10 1 1 1 0 50 merge.vcf

sed -i "s/\.\/\./0\/0/g" merge.vcf

cat merge.vcf | grep -v "SVTYPE=BND" | grep -v "SVTYPE=TRA" | grep -v "SVTYPE=INV" | grep -v "HLA" | bcftools sort | bgzip -cf > merge.vcf.gz

tabix merge.vcf.gz

bcftools +mendelian2 merge.vcf.gz -p 1X:SAMPLE_2,SAMPLE_1,SAMPLE -m c

8 The commands and parameters used for 557 ont samples analysis

HitSV single sample call

HitSV call -p ERR_PRONE -l ont.sample.bam -r ref -o sample.vcf

Single-sample csv region analysis

HitSV tools csv_region_500w sample.vcf 500 50 > sample.csv.bed

cat sample.csv.bed | grep "HAP1" | grep "500W" | sort | uniq > sample.csv.hap1.bed

cat sample.csv.bed | grep "HAP2" | grep "500W" | sort | uniq > sample.csv.hap2.bed

Merge csv region of all samples and filtering

*cat */*hap*bed > ALL_single_hap.bed*

cat ALL_single_hap.bed | awk '{printf "%s:%s-%s\n",$1,$2,$3;}' | sort | uniq -c > ALL_single_hap_summary.txt

cat ALL_single_hap_summary.txt| awk '{split($2, A, ":"); split(A[2], B, "-"); if($1 >=56) printf "%s\t%s\t%s\t%s\n", A[1],B[1],B[2],$1}' > AF_0.05_CSV_AF.bed

bedtools sort -i AF_0.05_CSV_AF.bed > AF_0.05_CSV_AF.sort.bed

bedtools merge -i AF_0.05_CSV_AF.sort.bed > cohot.csv.bed

Recall local contig sequences for each sample and csv region

HitSV call -p ERR_PRONE -b cohot.csv.bed -B PURE_STR -l ont.sample.bam -r ref.fa -o sample.FC.txt

Store contigs in corresponding csv regions

*find *.FC.txt > all_sample.FC.file_name.txt*

HitSVL tools contig_file_split all_sample.FC.file_name.txt csv_result_dir GRCh38.fa

Annotation and contig clustering analysis in single csv region

For each csv region in "*csv_result_dir* ":

RepeatMasker contig_in_region.fa

trf contig_in_region.fa 2 7 7 80 10 10 500 -h -d -ngs 1> contig_in_region.fa.trf.out

HitSV tools fc_joint_ana contig_in_region.fa contig_in_region.fa.out contig_in_region.fa.trf.out 0

9 The commands and parameters used for 1KGP re-analysis

HitSV single sample SV calling

For each sample in 1KGP P4:

HitSV srs_trans_reads ref.fa sample.bam TL.bam

samtools sort --output-fmt=BAM -o TL.sort.bam TL.bam

samtools index TL.sort.bam

HitSV call -n input.bam -L TL.sort.bam -r ref.fa -I ref.stat.txt -o sample.single.vcf

SURVIVOR (SV merging)

*find *single.vcf > vcf_list.txt*

SURVIVOR merge vcf_list.txt 500 1 1 1 0 30 output.vcf

AnnotSV

AnnotSV -SVinputFile inputFile -outputDir output -outputFile sample_name.csv -genomeBuild GRCh38 -SVminSize 30

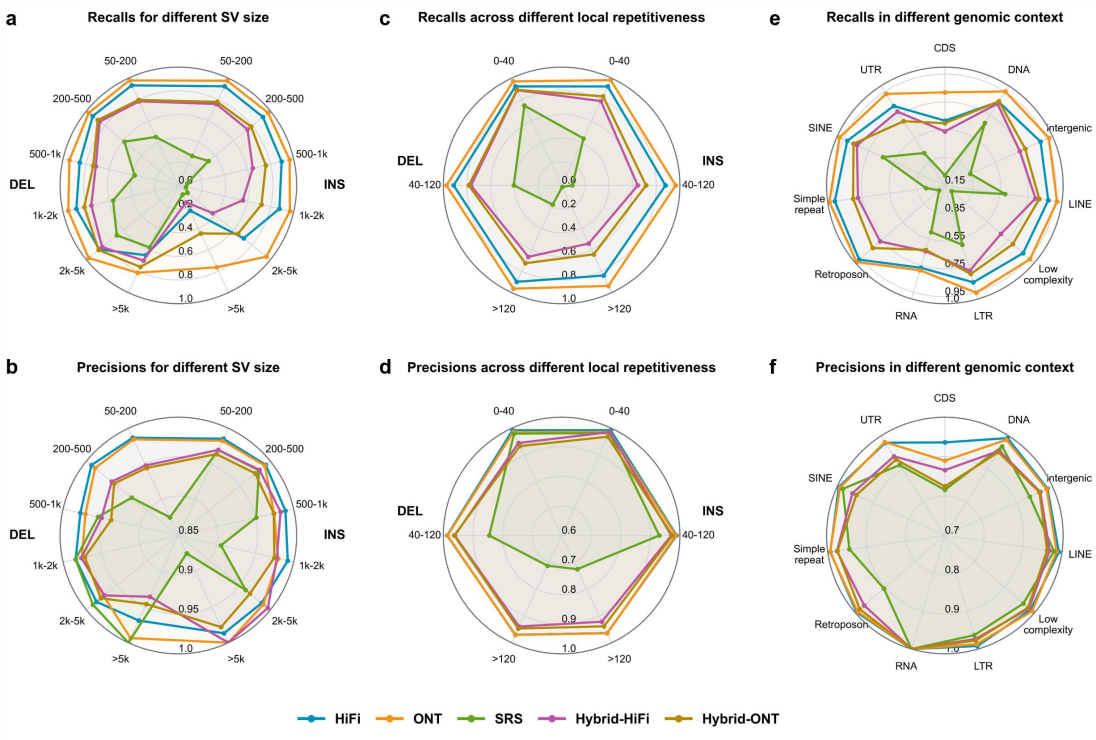
Supplementary Note 4. The characterization of the false negatives and positives of HitSV

We added a new error-profile analysis for HitSV on HG002 across LRS (both of PacBio and ONT), hybrid and SRS settings. For each setting, we quantified FP and FN calls and stratified them by SV type, SV size, local repetitiveness measured by the R metric, and genome annotations including RepeatMasker classes and gene-related annotations. The results are shown in Supplementary Fig. 31. We further extended the same analysis to the 35 HGSVC samples for LRS data (Supplementary Fig. 32). The HGSVC results showed broadly consistent error patterns, supporting that the HG002 observations are not specific to a single benchmark genome.

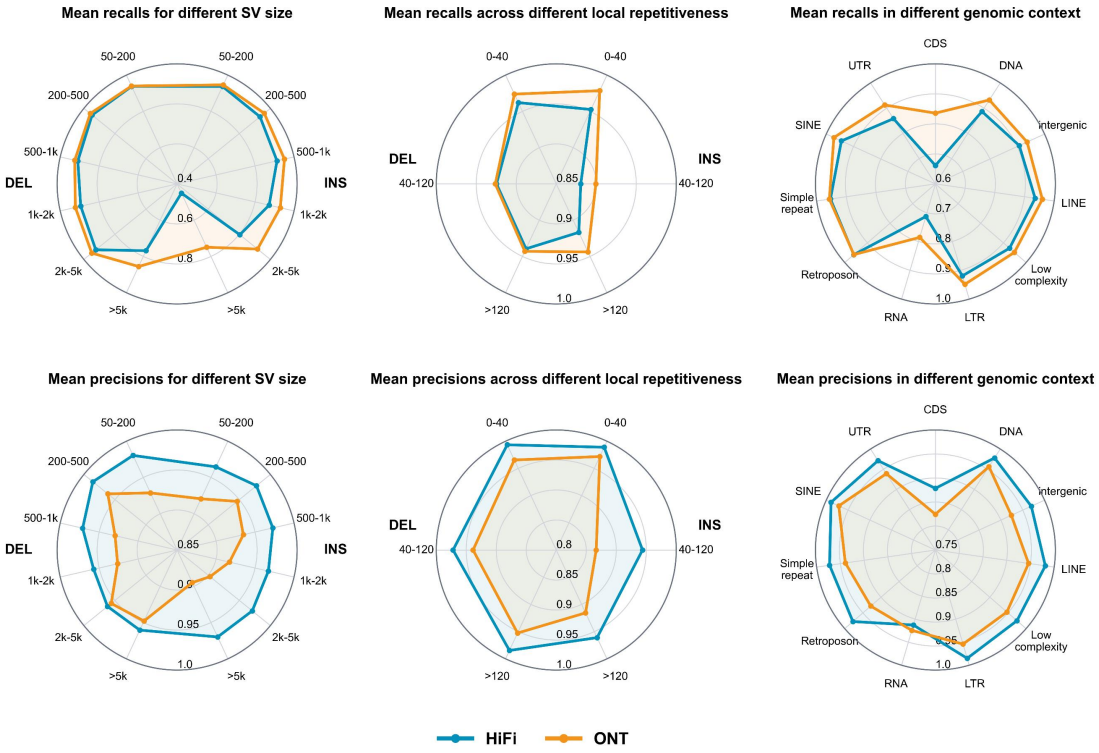
Several clear trends emerged from these analyses. First, FNs were strongly dependent on read length and SV type Supplementary Fig. 31a). HitSV produced relatively few FN deletions across LRS, hybrid and SRS settings, whereas FN insertions, especially insertions larger than 1 kb, were more frequent in SRS data. These missed insertions were enriched in medium- to high-repetitiveness regions and in regions annotated as simple repeats, low-complexity sequences, and GIAB hard regions (Supplementary Fig. 31c&e). Manual inspection indicated that many of these events lacked sufficient correctly mapped spanning reads (an example is shown in Supplementary Fig. 33) in SRS data. In some cases, candidate alleles could be accurately reconstructed by the heuristic dBG strategy, but local realignment to highly repetitive reference contexts remained ambiguous (Supplementary Fig. 34). The addition of even shallow long-read data substantially reduced this class of FNs, consistent with the ability of individual long reads to span repetitive insertion alleles and improve allele selection in the hybrid mode.

Second, FP calls were generally concentrated in highly repetitive regions, particularly regions with very high R values (Supplementary Fig. 31d). The major sources of FP calls differed by sequencing strategy. In SRS data, a recurrent source of FPs was inaccurate copy-number inference in VNTR-like regions, where base-level SV signatures were weak or absent and only subtle read-depth fluctuations remained. In such cases, local copy number may be underestimated, producing apparent small deletions calls (an example is shown in Supplementary Fig. 35). In LRS data, a subset of apparent FPs occurred in extreme repeat context where sequence-level comparison by Truvari (refine mode) was also affected by repeat ambiguity; so that small representation divergences between reconstructed alleles and benchmark alleles could be decomposed into multiple FP-SVs (an example is shown in Supplementary Fig. 36). We now describe these as apparent evaluation-level FPs where appropriate, rather than treating all such cases as biological false calls. In hybrid data, FP calls reflected a combination of these effects: low long-read coverage may provide only one local long-read allele, while the short-read component may still retain ambiguous candidates from repetitive regions to avoid sacrificing sensitivity.

These analyses clarify the remaining failure modes of HitSV. HitSV performs particularly well for deletions, mobile-element-related insertions, and SVs in low- to moderate-repetitiveness regions, while the most difficult cases remain large non-MEI insertions from SRS data and subtle copy-number changes in extremely repetitive VNTR-like regions.

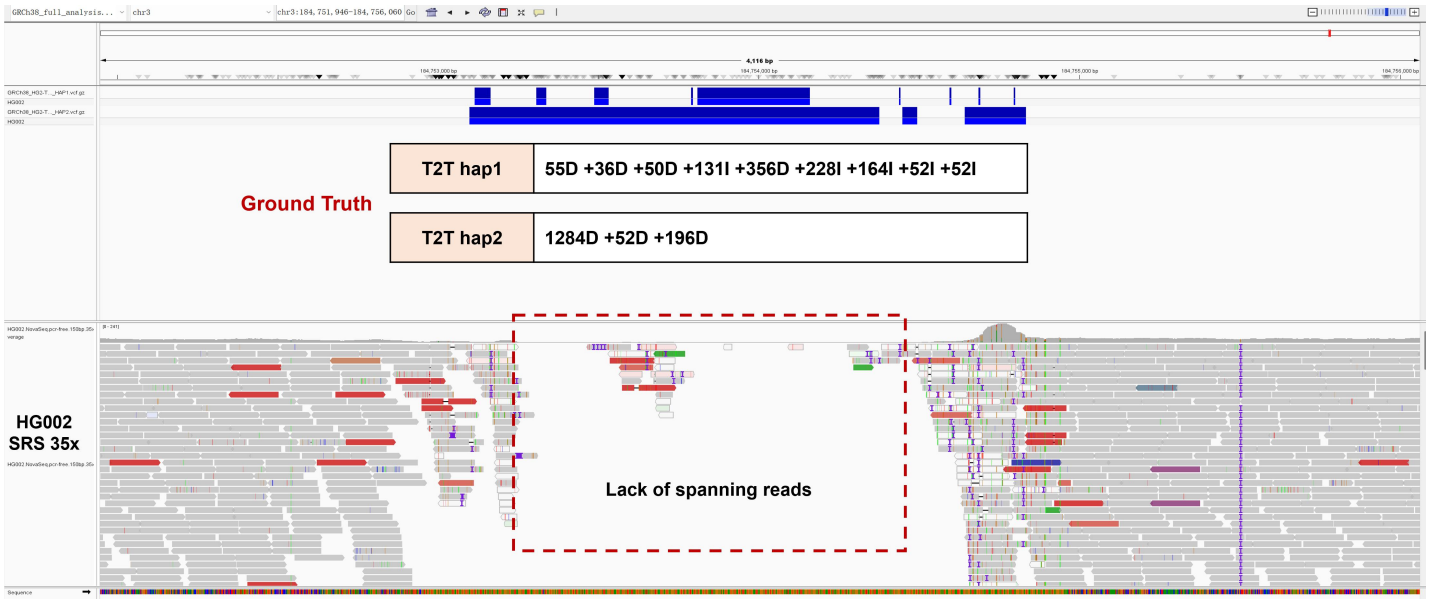


Supplementary Fig. 31: The recalls and precisions of HitSV in various contexts (HG002 benchmark). Comparing to HG002 ground truth, each radar plot represents the recall (upper panel, **a,c,e**) or precision (lower panel, **b,d,f**) of HitSV in the contexts of SV types/sizes (first column, type: deletion (left) or insertion (right), size: 50-200bp, 200-500bp, 500-1kbp, 1k-2kbp, 2k-5kbp, >5kbp), local reference repetitiveness (second column, Low, 0<R<40; Medium: 40≤R<120; High: R≥120, also grouped by deletion (left) and insertion (right) types) and Repeat Masker annotations (third column, in terms of CDS, DNA, intergenic, LINE, low complexity, LTR, RNA, simple repeat, SINE and UTR).

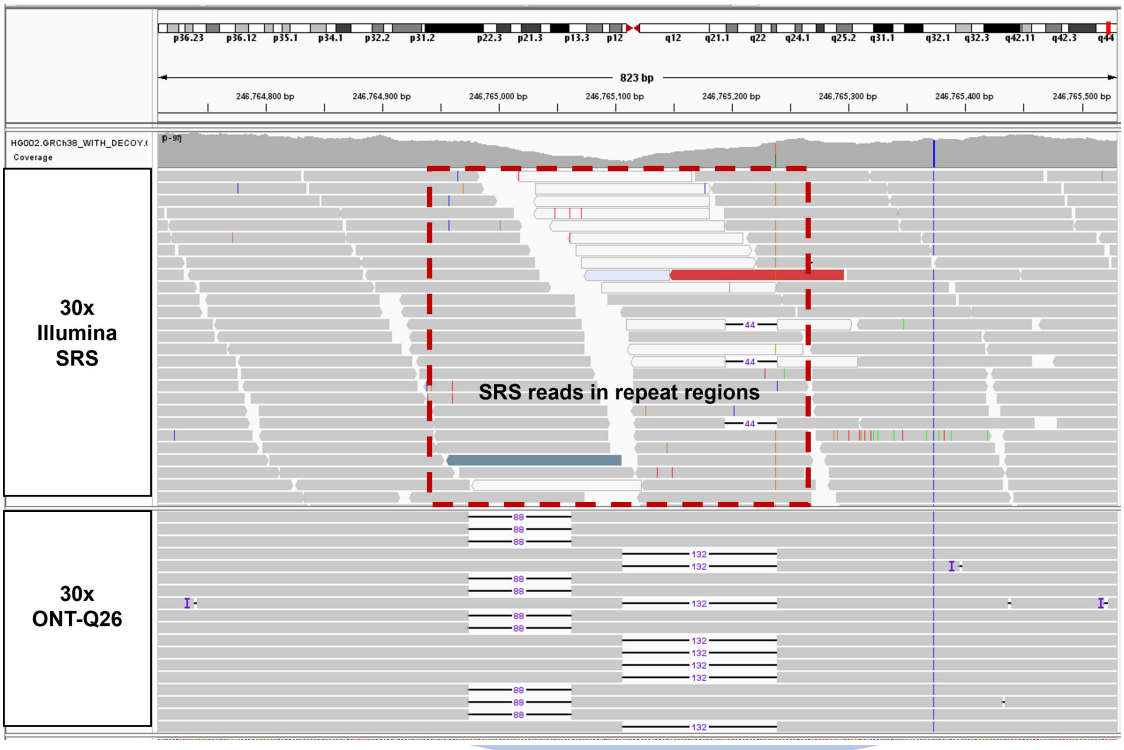


Supplementary Fig. 32: The average recalls and precisions of HitSV in various contexts for HGSVC samples. Comparing to HGSVC (35 samples) ground truth, each radar plot represents the average recall (upper) or precision (lower) of HitSV in the contexts of

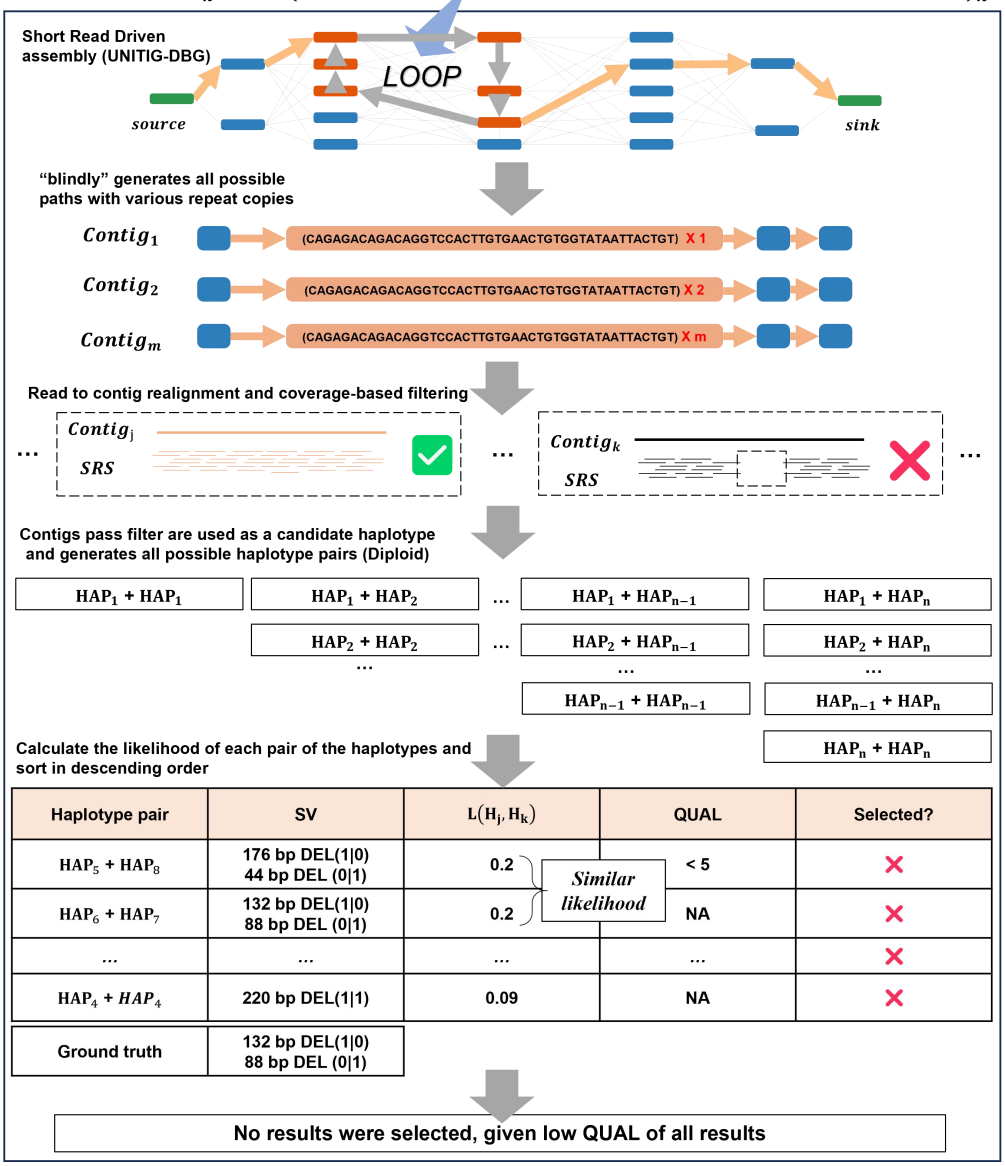
SV types/sizes (first column, type: deletion (left) or insertion (right), size: 50-200bp, 200-500bp, 500-1kbp, 1k-2kbp, 2k-5kbp, >5kbp), local reference repetitiveness (second column, Low, $0 < R < 40$; Medium: $40 \leq R < 120$; High: $R \geq 120$, also grouped by deletion (left) and insertion (right) types) and Repeat Masker annotations (third column, in terms of CDS, DNA, intergenic, LINE, low complexity, LTR, RNA, simple repeat, SINE and UTR).



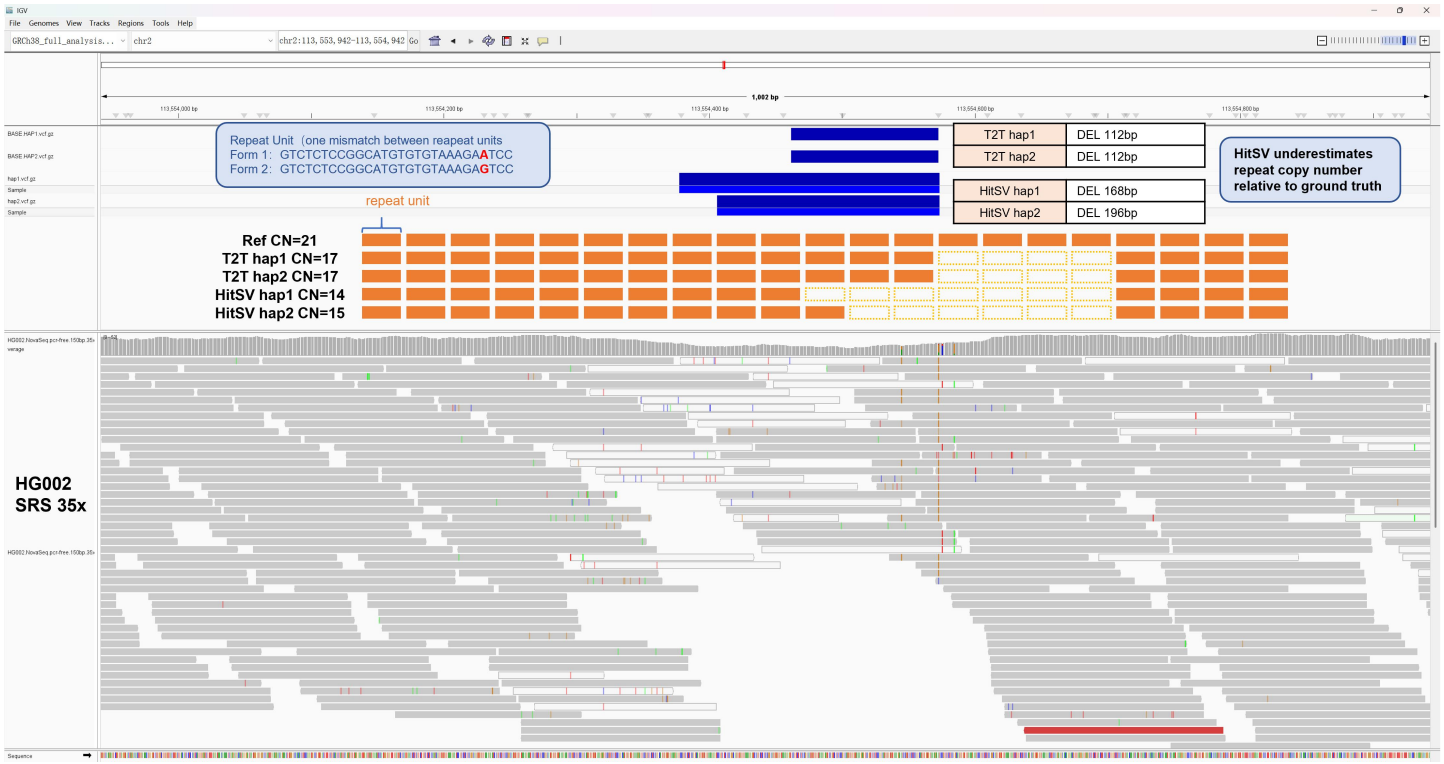
Supplementary Fig. 33. An example of false negative SV caused by lack of spanning reads. An example of SV not successfully detected by hitSV in a local region of chromosome 3 (chr3: 184,751,946-184,756,060) is shown. The upper panel represents the ground truth of the events, i.e., in each of the two SV alleles, there are a couple of insertion/ deletion events which compose a quite complex heterozygous SV. The lower panel represents the IGV snapshot of the reads being mapped to the surrounding region. It obviously shows that there are few reads mapped to the SV region with partial alignments and poor mapping quality, which are far from sufficient to reconstruct correct SV alleles. In this situation, HitSV failed to detect this event and so did other callers.



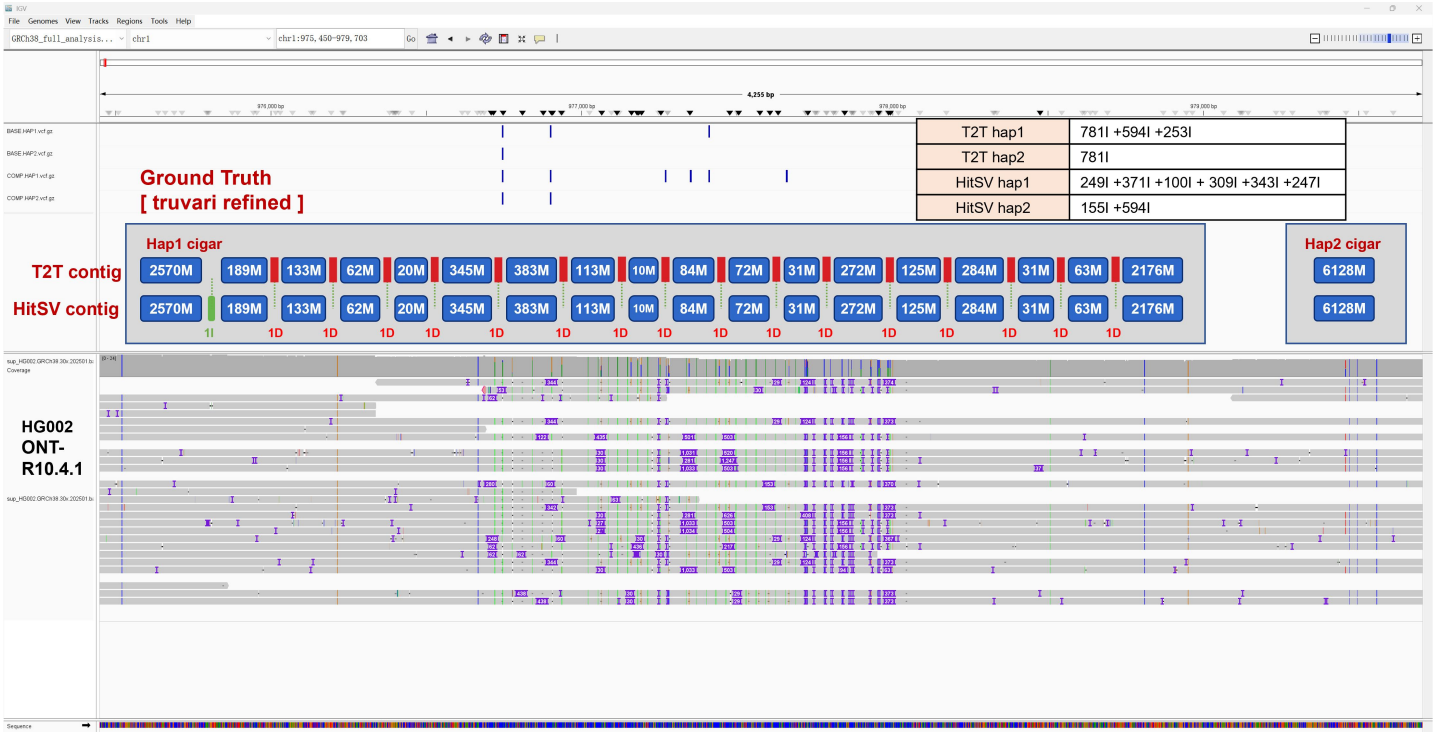
$R_k > 40,\ (CAGAGACAGACAGGTCCACTTGTGAACTGTGGTATAATTACTGT)_n$



Supplementary Fig. 34. A schematic illustration on the failure of SV detection even when the SV alleles themselves were finely reconstructed. The IGV snapshot shows a tandem-repeat locus at chr1:246,764,708-246,765,529, composed of a 44-bp repeat unit (CAGAGACAGACAGGTCCACTTGTGAACTGTGGTATAATTACTGT). In the truth set, HG002 is heterozygous, carrying two haplotypes with 6 and 7 repeat copies, corresponding to 132-bp and 88-bp deletions relative to the reference, which contains 9 repeat copies. HitSV collects all reads in the local region and constructs a localized unitig-based de Bruijn graph, then enumerates all feasible paths under node-count constraints to generate candidate contigs. Reads are aligned back to these contigs, and only those fully supported by read coverage are retained and added to the haplotype pool. All retained haplotypes are subsequently paired, and for each pair HitSV evaluates the likelihood of being the final solution by jointly considering consistency between read and contig k-mer frequencies as well as the requirement that all reads are explained by the assembly. In this locus, two combinations—5+8 and 6+7 repeat copies—receive identical composite scores, leading to indistinguishable haplotype likelihoods. As a result, HitSV cannot resolve the optimal configuration and fails to report the correct genotype for this VNTR region.



Supplementary Fig. 35. An example of false positive SV caused by the copy number underestimation of HitSV (for SRS data). An example of SV in the region of chromosome 2 (chr2: 113,553,942-113,554,942) is shown which HitSV produced a false positive call due to the underestimation of local copy number. According to the ground truth, this is a homozygous SV occurred in a VNTR locus. The reference allele indicates that the region consists of 21 copies of repeat units (shown in the upper panel, unit length is 28bp), and both of the two alleles of HG002 have 17 copies. However, with the relatively large number of background copy number and the poor mappability of short reads, the true copy number is quite difficult to estimate in theory. HitSV captured the lowered local read depth (IGV snapshot shown in the lower panel), however, its dBG-based heuristic mistakenly estimated the allelic copy numbers as 14 and 15, causing a false positive heterozygous SV.

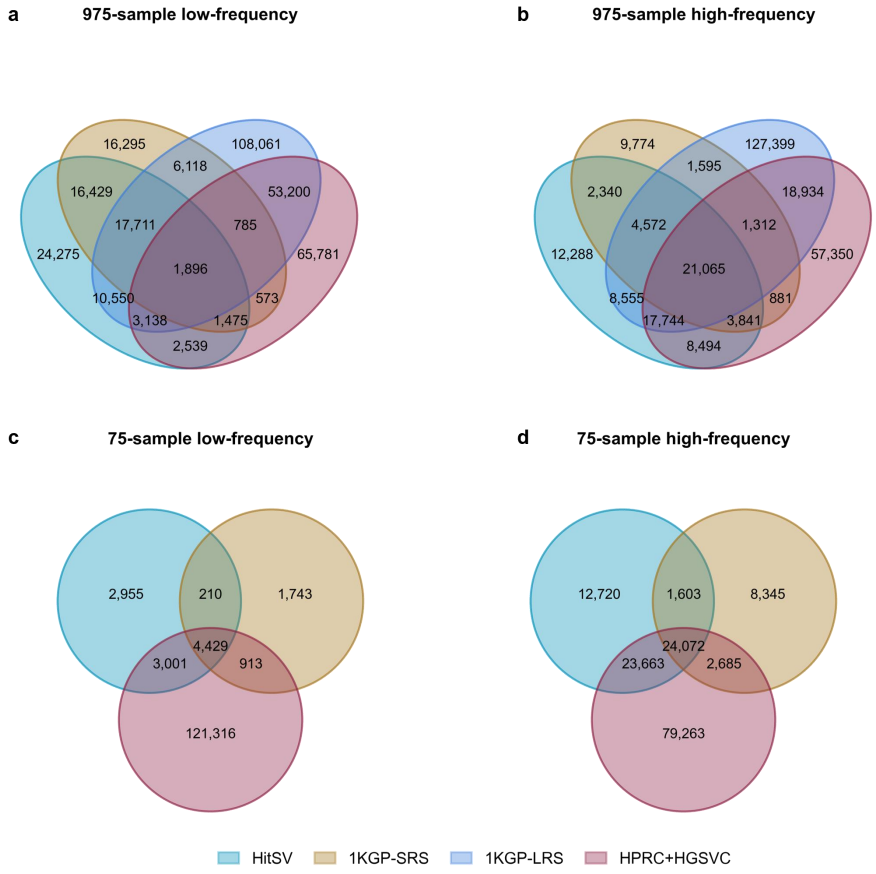


Supplementary Fig. 36. An example of false positive SV caused by the misjudgment of Truvari. An example of SV in a local region of chromosome 1 (chr1: 975,450-979,703) is shown which was misjudged during benchmark. The upper panel represents the ground truth, i.e., a complex, heterozygous SV which has two SV alleles consisting of three and one insertion events, respectively (hap1 and hap2 shown in the figure). The reads mapped there are also shown in the low panel. HitSV detected this SV site and reconstructed both of the two alleles. The outputs (HitSV hap1 and hap2) represent the forms different from the ground truths primarily due to the effects of re-alignment parameters and different local anchor sequences used during breakpoint inference (the last step in SV calling). However, the reconstructed SV alleles of HitSV have very high identity to the corresponding ground truth alleles. For example, the alignment between the local contigs of Hap1 extracted from HG002 T2T assembly and reconstructed by HitSV is represented in the middle panel. Their base identity rate is 99.75% which can be seen as a highly successful reconstruction of SV allele. For Hap2, the identity rate is exact 100% (alignment not shown). Considering the different output SV form and nearly perfect sequencing identity, we also referred to Truvari refine mode for evaluation. However, Truvari judged the result of HitSV as several false positive SVs which also lowered the overall precision. In our opinion, this could be a limitation of the evaluation tool for the handling of complex SV, but not a real false positive call.

Supplementary Note 5. External support for HitSV-specific SVs in the 1KGP cohort

We stratified HitSV-specific and official 1KGP SRS-specific SVs by allele frequencies, and separately examined the SVs in various frequencies. In the 975 samples covered by the 1KGP-LRS callset, HitSV-specific SVs had more long-read-supported calls than the official 1KGP SRS-specific set in both rare and more common frequency classes. For very rare SVs ($MAF \leq 0.2\%$), 13,688 HitSV-specific SVs were supported by the 1KGP long-read callset, compared with 6,903 official-1KGP-SRS-specific SVs (Supplementary Fig. 37a). For relatively common SVs ($MAF > 0.2\%$), the corresponding numbers were 26,299 vs 2,907, respectively, a nine-folds difference in justifiable callset-specific SVs (Supplementary Fig. 37b). A similar trend was observed in the HGSVC/HPRC assembly-based comparison, where 3,001 rare HitSV-specific SVs were supported by assembly-based callsets compared with 913 official-1KGP-SRS-specific SVs (Supplementary Fig. 37c), and 23,663 more common HitSV-specific SVs were supported compared with 2,685 official-1KGP-SRS-specific SVs (Supplementary Fig. 37d).

Besides absolute counts, the proportions of SVs being supported are also worthnoting. Across the external-resource comparison, ~74% of HitSV-specific common ($MAF > 0.2\%$) SVs were supported by either 1KGP-LRS-callset or HGSVC/HPRC callsets, while the number is only 28% for 1KGP-SRS-callset. For very rare ($MAF \leq 0.2\%$) SVs, the supported proportion was lower, as expected, but HitSV-specific SVs still showed a higher support fraction (40%) than official 1KGP-SRS-callset-specific SVs (31%). We note that rare SVs are expected to have lower external-support rates in a population callset. True SVs shared by multiple samples are more likely to be merged in into recurrent loci and supported by independent resources, whereas random false-positives are more likely to appear as rare or singleton events. Nevertheless, matched-sample evaluation against HPRC and HGSVC assembly-based callsets showed high precision for HitSV-specific calls, indicating that the additional SVs recovered by HitSV are not primarily driven by recurrent artifacts. The remaining unsupported singleton calls may reflect both single-sample calling errors and the sue of simple SV merging rather than dedicated joint calling/force calling. Because HitSV focuses on de novo single-sample SV discovery, population-level joint calling is beyond the scope of this study and is now discussed as a future improvement.



Supplementary Fig. 37: The Venn diagrams on the rare and SVs of 1KGP samples in various callsets. a-b, The Venn diagrams on

the rare (part a, MAF≤0.2%) and common (part b, MAF>0.2%) SVs specifically reported by HitSV-SRS-callset (cyan) and 1KGP-SRS-callset (orange) for the 975 samples of 1KGP-LRS-callset, the whole 1KGP-LRS-callset (blue) and the assembly-based HPRC/HGSVC callset (red, merged in advance). **c-d,** The Venn diagrams on the rare (part a, MAF≤0.2%) and common (part b, MAF>0.2%) SVs specifically reported by HitSV-SRS-callset (cyan) and 1KGP-SRS-callset (orange) for the 75 HPRC/HGSVC samples and the assembly-based HPRC/HGSVC callset (red).

Supplementary Note 6. Analytical workflow used for the CHN100K–HPRC comparison

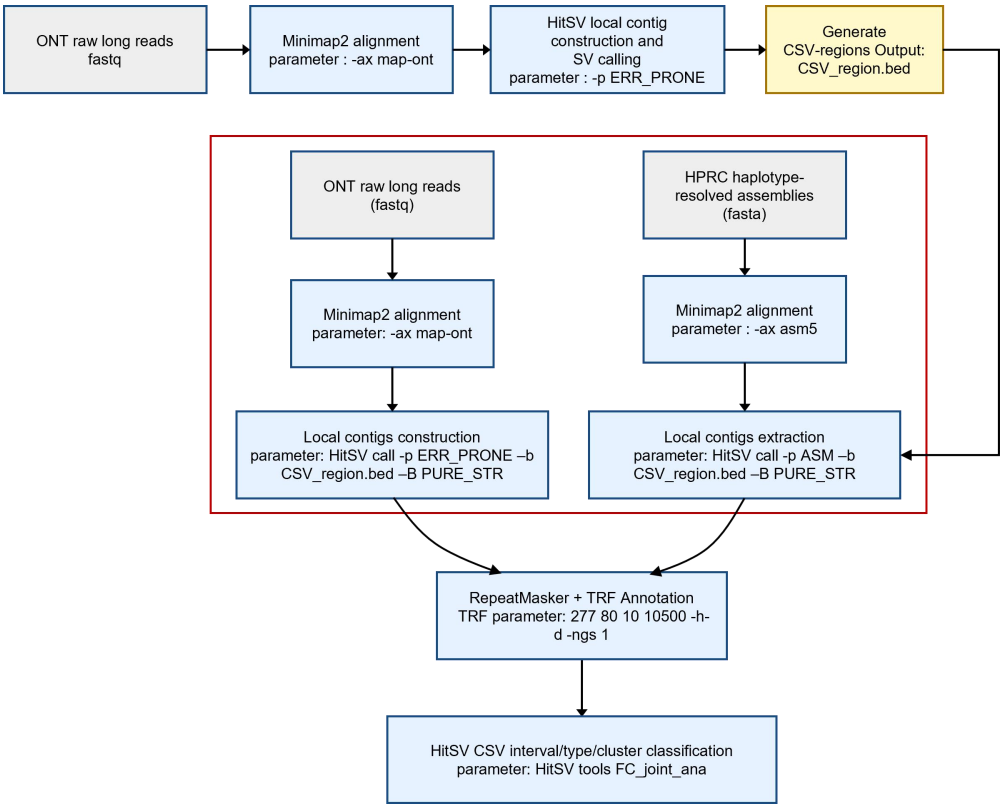
We built an analytical workflow for the comparison of SVs in the callsets of CHN100K and HPRC which consists of four major steps.

Step1: (the detailed description for the "*Population-level CSV interval from 557 ONT samples*" section of Online method) generation and annotation of CSV contigs and regions. For each of the 557 samples from CHN100K, the SVs were initially detected by HitSV based on the read alignments produced by minimap2, then the population level CSV were detected and the corresponding covered genomic regions were also collected as a BED file. Two sets of SV-related contigs were then prepared based on the BED file separately for CHN100K and HPRC as following to implement the comparison (Supplementary Fig. 38).

1) For CHN100K samples, the raw ONT long-read data were realigned to the reference genome using the `minimap2 -x map-ont` parameter, and perform local assembly within each CSV interval using HitSV to generate local contigs.

2) For HPRC samples, the haplotype-resolved assemblies were aligned by minimap2 (using ASM5 parameter) and the contigs overlapping with the BED recorded regions were extracted to compose the SV-related contig set.

All the SV-related contigs were annotated by RepeatMasker (default parameter setting) and TRF (parameter setting: 277 80 10 10500 -h -d -ngs 1) to mark the mobile elements and tandem duplications in them. For each of the SV regions, a script built in HitSV clustered all the CHN100K and HPRC SV contigs lying in it and counted their types and ranges.



Supplementary Fig. 38: The analytical workflow for CSV interval identification in CHN100K and verification in HPRC. Schematic overview of the CHN100K-HPRC comparison. For CHN100K, ONT reads from 557 individuals were aligned to GRCh38 and locally assembled by HitSV within CHN100K population CSV intervals. For HPRC, haplotype-resolved assembled contigs from 47 individuals were aligned to GRCh38 and extracted from the same intervals. CHN100K and HPRC contigs were annotated, clustered and compared using the same downstream allele-classification framework.

Step2: SV region filtration. Given an SV region, filtrations were implemented by the following three criteria for quality control.

1) Abnormal read coverage filtering: the average read depth on the contigs of all the involved samples was 50% lower than or 200% higher than the sequencing coverage of the whole dataset (computed by Mosdepth: -t 8 -x -f).

2) Missing assembly contigs filtering: there were more than 20% HPRC haplotypes missing (19 in absolute terms).

3) Variable SV representation filtering: all the CHN100K SV-contigs have high similarity to any of HPRC SV-contigs (without sequence differences of more than 50 consecutive bases), which was computed by all-to-all alignment using ksw2 (parameters: -A1 -B4 -O6,26 -E2,1).

According to these filters, we apply the CHN100K–HPRC comparison CSV loci with allele frequency >0.1%. Before quality filtering, there are 9,294 CSV alleles were supported in both CHN100K and HPRC, where 5,788 intervals were unsupported by HPRC, where the concordance rate of CSV classification is 20.61%. After filtering, 5,601 CSV loci remained and the concordance rate improved to 29.49% (Supplementary Fig. 39).

a. Before quality filtering

Number of CSV Regions		CHN100K-557					
		Pure TR	TR with non-TR SVs	Pure non-TR	non-TR with TR SVs	non-TR with both types	No CSV
HPRC-47	Pure TR	1372	1523	0	0	0	0
	TR with non-TR SVs	37	501	0	0	0	0
	Pure non-TR	0	0	30	0	30	0
	non-TR with TR SVs	0	0	0	0	0	0
	non-TR with both types	0	0	0	0	13	0
	No CSV	1451	2567	679	15	1076	0
		Missing rate in HPRC cohort: 62.27%					Concordance rate:20.61%

b. Abnormal read coverage filtering

Number of CSV Regions		CHN100K-557					
		Pure TR	TR with non-TR SVs	Pure non-TR	non-TR with TR SVs	non-TR with both types	No CSV
HPRC-47	Pure TR	1281	1395	0	0	0	0
	TR with non-TR SVs	34	469	0	0	0	0
	Pure non-TR	0	0	26	0	30	0
	non-TR with TR SVs	0	0	0	0	0	0
	non-TR with both types	0	0	0	0	13	0
	No CSV	1333	2402	658	12	1058	0
		Missing rate in HPRC cohort: 62.71%					Concordance rate:20.53%

c. Missing assembly contigs filtering

Number of CSV Regions		CHN100K-557					
		Pure TR	TR with non-TR SVs	Pure non-TR	non-TR with TR SVs	non-TR with both types	No CSV
HPRC-47	Pure TR	1143	1268	0	0	0	0
	TR with non-TR SVs	30	444	0	0	0	0
	Pure non-TR	0	0	24	0	28	0
	non-TR with TR SVs	0	0	0	0	0	0
	non-TR with both types	0	0	0	0	13	0
	No CSV	1258	2280	631	11	1026	0
		Missing rate in HPRC cohort: 63.83%					Concordance rate:19.88%

d. Variable SV representation filtering

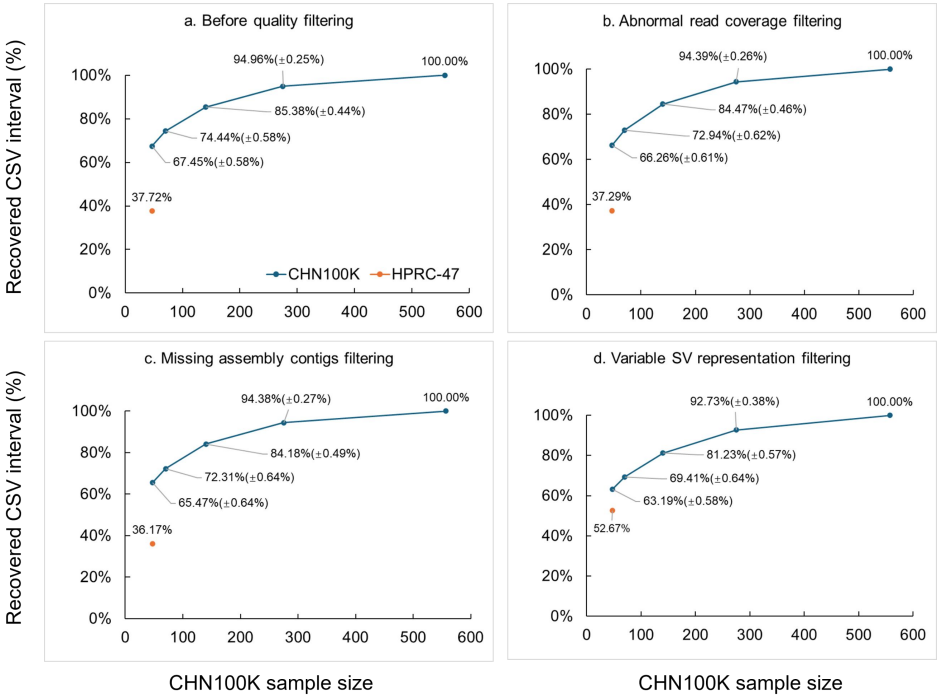
Number of CSV Regions		CHN100K-557					
		Pure TR	TR with non-TR SVs	Pure non-TR	non-TR with TR SVs	non-TR with both types	No CSV
HPRC-47	Pure TR	1143	1268	0	0	0	0
	TR with non-TR SVs	30	444	0	0	0	0
	Pure non-TR	0	0	24	0	28	0
	non-TR with TR SVs	0	0	0	0	0	0
	non-TR with both types	0	0	0	0	13	0
	No CSV	409	842	522	4	874	0
		Missing rate in HPRC cohort: 47.33%					Concordance rate:29.49%

Supplementary Fig. 39: Effects of conservative quality-control filters on CHN100K-HPRC CSV comparison with allele frequency >0.1%. a, CHN100K-HPRC comparison for CHN100K CSV alleles with allele frequency >0.1% before (**a**) applying quality-control filters. **b-c**, results passing all three high-confidence filters: (**b**) normal CHN100K read coverage, (**c**) low HPRC contig missingness, and (**e**) no evidence of variable SV representation.

Step3: down-sampling analysis of SV regions. The 557 CHN100K samples were randomly down-sampled to four subsets (respectively 47, 70, 140 and 275 samples) for evaluating the consistency of the SV region numbers with various sample sizes. Given a certain subset of samples, we assessed the proportion of the regions remaining at least one CHN100K SV-contig as an indicator. For each of the four sample sizes, bootstrapping (n=100) was conducted to estimate the mean and standard deviation of recovery rate.

To model this effect, we downsampled CHN100K to the HPRC sample size and estimated the expected fraction of CHN100K CSV loci that would be missed at 47 individuals. We first performed this analysis for CHN100K CSV loci with

allele frequency >0.1%, which captures nearly all recurrent loci while excluding singleton-only observations (Supplementary Fig. 40a). We then repeated the analysis after applying the same conservative filters used in the above HPRC reads-contig comparison, including removal of HPRC-contig-missing intervals, CHN100K abnormal-coverage intervals, and loci likely affected by variable SV representation (Supplementary Figs. 40b-d). These downsampling curves showed that when CHN100K was reduced to the HPRC sample size, the expected CSV missing rate (1-percentage of recovered CSV intervals) is 32.55%/36.81% before/after filtering, indicating that a substantial component of the raw CHN100K–HPRC difference can be explained by cohort size alone.



Supplementary Fig. 40: Downsampling analysis of CHN100K CSV allele recovery. CHN100K samples were downsampled to 47, 70, 140 and 275 individuals to estimate the effect of cohort size on CSV allele recovery. For each sample size, 100 random replicates were generated. a, Recovery curves before applying quality-control filters. b-d, Recovery curves after excluding loci affected by HPRC contig missingness, abnormal CHN100K read coverage and variable SV representation. When CHN100K was downsampled to the HPRC sample size of 47 individuals, the expected missing rate was 32.55% before filtering and 36.81% after filtering, indicating that cohort-size limitation explains a substantial component of the raw CHN100K-HPRC difference. Values in parentheses indicate standard deviations across random replicates.

Step4: Report CHN100K–HPRC comparison to use a more conservative denominator. Due to the discovery limit of 47 diploid HPRC samples is approximately 1/94 haplotypes, we restricted the reported CHN100K–HPRC comparison to CHN100K CSV loci with allele frequency >1%. Before quality filtering, 5,433 loci were supported in both CHN100K and HPRC population, the HPRC missed 40.97% of CHN100K-supported CSV loci. After passing all the three filters: low HPRC contig missingness, normal CHN100K read coverage, and no evidence of variable SV representation, there were 2,676 loci supported in both CHN100K and HPRC, whereas 314 loci remained unsupported by HPRC, corresponding to 10.50% of the high-confidence comparable loci (Supplementary Fig. 17).

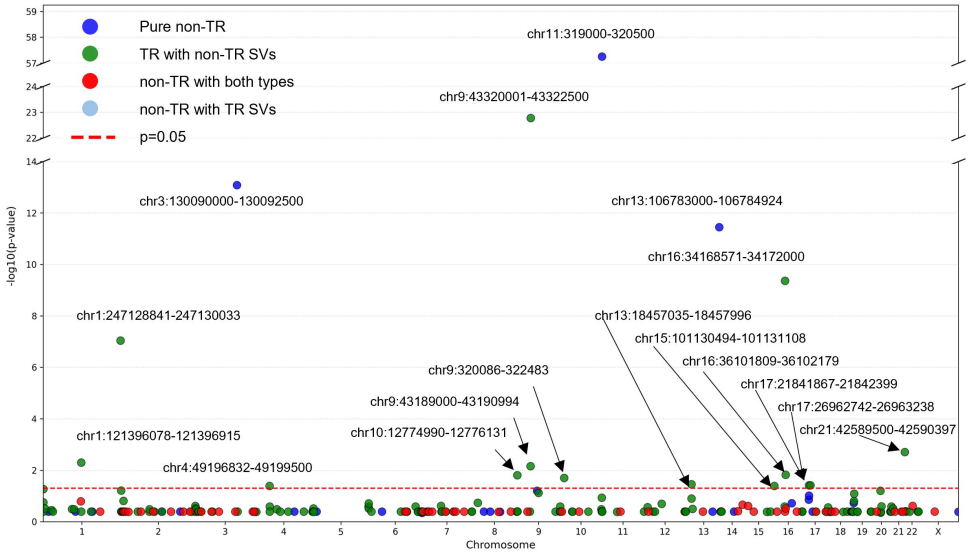
Step5: recognition of CHN100K specific SV regions. To recognize the SV enriched regions specific to CHN100K population, a further filtration was implemented each SV region by the two following criteria.

1) There were more than 1% of CHN100K samples having SV contigs in the region.

2) There was no HPRC sample having CSV contig there.

For each of the remaining regions, a Fisher's exact test was conducted based on the numbers of CHN100K CSV contigs, and the total numbers of CHN100K and HPRC haplotypes. If $p < 0.05$ (after Benjamini–Hochberg adjustment),

the alternative hypothesis was selected, i.e., the region can be seen as CHN100K SV-enriched where the SV allele frequency of CHN100K population was significantly higher than that of HPRC. Finally, we identified 17 CHN100K-enriched CSV alleles, including 14 TR–MEI composite alleles and 3 non-TR compound CSV alleles. A Manhattan plot (Supplementary Fig. 41 and Supplementary Table 41) was provided to visualize all the CHN100K SV-enriched regions among the chromosomes.



Supplementary Fig. 41: Identification of CHN100K-enriched CSV alleles. Manhattan plot showing one-sided Fisher's exact test results for high-confidence HPRC-unsupported CHN100K CSV alleles. Each point represents one candidate CSV allele. The y-axis shows -log10 adjusted P value after Benjamini-Hochberg correction, and the x-axis shows genomic position. CSV alleles with adjusted P < 0.05 and allele-frequency difference >1% were defined as CHN100K-enriched. In total, 17 CHN100K-enriched CSV alleles were identified, including 14 TR-MEI composite alleles and 3 non-TR compound CSV alleles.

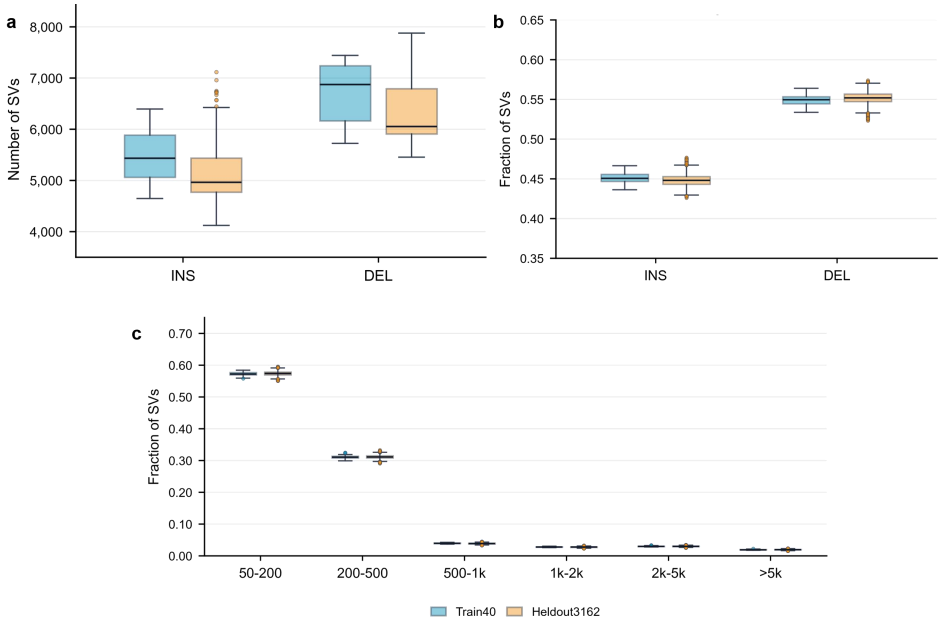
Supplementary Note 7. The SV-call comparison between the distribution-modeling- and other 1KGP samples

We divided the 1KGP re-analysis into two groups: the 40 samples used for signature-distribution modeling and the remaining 3,162 samples that were not used for this purpose. Between the two groups, we then compared per-sample SV counts, SV types proportions and SV length distributions. The 40 distribution-modeling samples showed slightly higher raw SV counts (Supplementary Fig. 42a and Supplementary Table 42). However, these 40 samples are not population-matched to the remaining cohort: 20 of them are African, only 4 are East Asian, 1 is South Asian, and none are European, while the remaining 3,162 samples have a more balanced population composition. After normalizing by SV types and length class, the distributions of the distribution-modeling and held-out groups were highly similar, indicating that the difference was not caused by the caller (Supplementary Figs. 42b&c).

We next asked whether the samples used for signature-distribution construction showed inflated calling accuracy. Using assembly-based callsets as independent truth sets, we compared HitSV SRS callsets of the 40 HPRC samples used for signature-distribution construction with those from 35 HGSVC samples that were not used for signature-distribution construction (Supplementary Figs. 43a and Supplementary Table 43). The average precisions were comparable, reaching 90.73% and 91.21% for HPRC and HGSVC samples, respectively. Thus, the distribution-modeling samples did not show artificially elevated precision relative to independent samples.

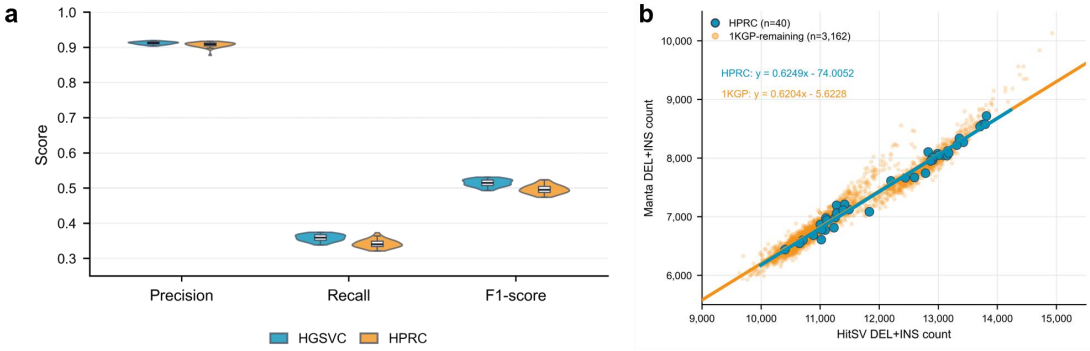
As an additional check for potential sample-group-specific inflation, we ran Manta on the 1KGP samples and compared the relationship between HitSV and Manta SV counts separately for the 40 distribution-modeling samples and the remaining 3,162 samples. Manta showed a similar raw count trend across the two groups, and the regression coefficients between HitSV and Manta SV counts were nearly identical between the distribution-modeling and held-out samples, with slopes of 0.6249 and 0.6204, respectively (Supplementary Fig. 43b). This indicates that the modest raw count difference between the two groups is not specific to HitSV, but is more consistent with differences in sample composition and genetic background.

Together, these analyses do not reveal evidence that inclusion of the 40 distribution-modeling samples caused systematic inflation of the 1KGP results. We have therefore revised the manuscript to explicitly describe the overlap between the distribution-modeling samples and the 1KGP cohort, and to report the additional overlap-control analyses. We also now describe the 1KGP reanalysis results as externally supported callset-level findings, while acknowledging that the empirical signature distributions were derived from a finite set of reference samples.



Supplementary Fig. 42: The SV counts and distributions for training and other 1KGP samples. a, The number of SV calls by HitSV on the 40 distribution-modeling samples and the other 3162 1KGP samples. **b-c,** The distribution of the types (insertion/deletion, **b**) and lengths (50-200bp, 200-500bp, 500-1kbp, 1k-2kbp, 2k-5kbp and >5kbp, **c**) of the SV calls of HitSV on the 40 training (cyan) and the

other 3162 (orange) 1KGP samples. For all box plots, the centre line indicates the median, the box bounds indicate the 25th and 75th percentiles, and the whiskers extend to the most extreme observed values within 1.5 times the interquartile range (IQR). Values beyond the whiskers are shown individually as outliers. The training and held-out groups comprise n = 40 and n = 3,162 independent individuals, respectively.



Supplementary Fig. 43: Additional check for potential sample-group-specific inflation. a, The violin plots in cyan and yellow colors respectively represent precisions (left), recalls (middle) and F1-scores (right) the SRS-based callsets of HitSV on the 35 HGSVC and 40 HPRC samples. All of the HPRC samples were used for the construction of SV signature distribution and none for the HGSVC samples. In the violin plots, the centre line indicates the median, the box bounds indicate the 25th and 75th percentiles, and the whiskers extend to the most extreme observed values within 1.5 times the interquartile range (IQR). Values beyond the whiskers are not displayed as individual box-plot outliers. **b,** Two linear regressions were separately done for the SV counts (SRS-based) of HitSV (x-axis) and Manta (y-axis) on the 40 training (HPRC, cyan line) and the remaining 3162 (orange line) 1KGP samples.

Supplementary Note 8. SV genotyping and quality control

HitSV calculates the posterior probability $P(G|D)$ of a specific genotype G given an observed set of related alignment D for genotyping. Assuming the list of the candidate alleles as $A = \{r, x_1, ..., x_n\}$, where r refers to reference allele and $x_i, i = 1, ..., n$ refers to SV-alleles, there are $|A|^2$ types of candidate genotypes, i.e., $G = \{rr, rx_1, rx_2, rx_3, ..., x_{n-1}x_n\}$ and $P(G|D)$ is formed as following:

$$P(G|D) = \frac{P(D|G) \times P(G)}{P(D)} \tag{1}$$

Further, the equation is reduced as $P(D|G) \times P(G)$ since D has been observed. The prior $P(G)$ is set as:

$$P(G) = \begin{cases} \theta/(n^2), & \mathrm{G} = \mathrm{pq} \\ 0.5\theta/n, & \mathrm{G} = \mathrm{rp} \\ 1 - 1.5\theta, & \mathrm{G} = \mathrm{rr} \end{cases} \tag{2}$$

where θ is a constant (default value: 10^{-5}) and n is the number of non-reference alleles, $\mathrm{p}, \mathrm{q} \in A$.

Further, each alignment is considered as an independent event and the conditional probability $P(D|G)$ is written as the product of the probabilities of the events:

$$P(D|G) = \prod_{d \in D} P(d|G) \tag{3}$$

where $P(d|G)$ is the probability of observing a specific abnormal alignment d given the genotype G,

$$P(d|G) = \sum_{a \in A} P(d|a) \times P(a|G) \tag{4}$$

Similarly, $P(d|a)$ represents the probability of observing a specific alignment d given a particular haplotype a. $P(a|G)$ denotes the probability of observing a specific haplotype a under the condition of a given genotype G, which is taken from $\{0, 0.5, 1.0\}$. All $P(d|a)$values form a support matrix, an $M \times N$matrix, where M is the number of reads and N is the number of contigs (two under a diploid assumption).

After realigned reads to different candidate haplotypes, HitSV recalculates the alignment scores by setting the penalty scores as 0, 3, 4, 1 for match, mismatch, gap open and gap extension, respectively. The penalty score has an upper limit (default is 5) and the total penalty score for realignment is denoted as *S*, $P(d|a)$ is set as 10^{-S} consequently.

For long-read data, a binary support matrix is generated by determining whether each long read supports a given contig in the pair (1 indicating support and 0 indicating non-support).

The genotype G maximizing $P(G|D)$ is selected and the quality score is computed by the following equation,

$$QUAL(G) = -10 \times log_{10}\big(1 - P(G|D)\big) \tag{5}$$

The results corresponding to homozygous reference genotypes are directly discarded. Additionally, results with quality scores below the threshold (default is 10) are labeled as *LOW*, while others are labeled as *PASS*.