

HitSV: Maximizing discovery of structural variants across sequencing technologies

Yadong Wang

ydwang@hit.edu.cn

Harbin Institute of Technology <https://orcid.org/0000-0002-0673-8503>

Gaoyang Li

Harbin Institute of Technology

Yadong Liu

Harbin Institute of Technology

Bo Liu

Harbin Institute of Technology

Long Qian

Peking University <https://orcid.org/0000-0003-3728-7917>

Article

Keywords:

Posted Date: February 20th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-8913458/v1>

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

[Read Full License](#)

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

Version of Record: A version of this preprint was published at Nature Biotechnology on October 15th, 2026. See the published version at <https://doi.org/10.1038/s41587-026-03332-9>.

HitSV: Maximizing discovery of structural variants across sequencing technologies

Gaoyang Li^{1,2†}, Yadong Liu^{1,2†}, Bo Liu^{1,2*}, Long Qian^{3*} and Yadong Wang^{1,2*}

¹Center for Bioinformatics, Faculty of Computing, Harbin Institute of Technology, Harbin, Heilongjiang, 150001, China

²Key Laboratory of Biological Bigdata, Ministry of Education, Harbin Institute of Technology, Harbin, Heilongjiang 150001, China

³Center for Quantitative Biology, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China

*Correspondence should be addressed to bo.liu@hit.edu.cn (B.L.), long.qian@pku.edu.cn (L. Q.), and ydwang@hit.edu.cn (Y. W.)

† These authors contributed equally to this work

Abstract

Structural variants (SVs) are a major source of genomic diversity, yet their discovery remains challenging due to repetitive genomic contexts, alignment ambiguity, and the trade-off between sequencing cost and read length. Here we introduce HitSV, which substantially improves SV discovery by implementing repetitiveness and signature density aware breakpoint recognition coupled with precise haplotype-resolved local assembly, thereby enabling base-resolution SV reconstruction and genotyping across various sequencing technologies. HitSV is 12-68% (long-read), 3%-36% (short-read) and 13% (hybrid-sequencing), respectively, more accurate than state-of-the-art SV callers across different coverages. Applying HitSV to the 1KGP Phase 4 cohort, we identified 31.5% more SVs, substantially reshaping allele-frequency landscapes. Notably, analysis of a large Chinese long-read cohort uncovers tandem repeat–mobile element composite arrays as a prevalent and multi-allelic class of complex SVs, highlighting composite repeat architectures as a fundamental hallmark of human genomes.

Introduction

Structural variants (SVs), typically defined as genomic alternations larger than 50 bp, arise from distinct genome-wide mutational processes^{1, 2} and contribute substantially to genetic and epigenetic variation³⁻⁶. At the phenotypic level, SVs are strongly associated with disease susceptibility and represent major drivers of oncogenesis and inherited genetic disorders⁷⁻⁹. Population-scale analyses further demonstrate that SVs account for a disproportionate fraction of inter-individual genomic differences and are frequently enriched in repetitive or structurally unstable regions of the genome. Accurate and comprehensive SV detection is therefore essential for both population genomics and medical genetics.

Reliable SV discovery remains challenging. Short-read sequencing (SRS) continues to dominate large-scale cohort studies due to its low cost and high throughput, yet its limited read length hampers the detection of many SVs, particularly those embedded in repetitive regions or involving insertions. Long-read sequencing (LRS) substantially improves mappability and breakpoint resolution, enabling more complete SV discovery, but its widespread application in population studies is constrained by sequencing cost and data requirements. As a result, existing SV callsets derived from either SRS or LRS alone are often incomplete, biased toward specific SV classes, or difficult to integrate across studies.

Most state-of-the-art SV calling approaches infer SVs from non-colinear read alignments to the reference¹⁰⁻²⁰, using signals such as split alignments, discordant read-pairs, large clippings and local read depth changes²¹⁻²³.

Within this paradigm, many widely used tools rely on signature clustering and statistical modeling to infer SV types and breakpoints¹⁵⁻¹⁷, whereas others incorporate local assembly and realignment to improve allele reconstruction^{10, 19}. Owing to their flexibility across read lengths, depths and qualities, alignment-based approaches are widely adopted in large-scale genomics studies²⁴⁻²⁶ and have been further extended to handle specific SV types or genomic regions²⁷⁻²⁹, as well as integrated into multiple-tools pipelines^{30, 31}. Besides that, efforts have also been made to assembly-based SV calling^{32, 33}, which leverage long-read contiguity to resolve large and complex rearrangements with reduced reference bias. However, their application in population-scale analyses remains constrained by high sequencing coverage requirement³⁴⁻³⁶, susceptibility to mis-assemblies³⁷⁻³⁹ and incomplete haplotype resolution⁴⁰, which inevitably introduce false positives and compromise SV genotyping. Accordingly, alignment-derived evidence continues to provide an indispensable foundation for scalable SV discovery and genotyping across diverse sequencing strategies. Nevertheless, such approaches are strongly influenced by alignment complexity, particularly in repetitive regions and for complex SVs, where SV-associated signals become weak, heterogeneous and systematically distorted.

Here, we present HitSV, a read length-agnostic SV calling approach that enables comprehensive, accurate and cost-effective SV discovery. HitSV explicitly models the impact of local genomic context—particularly repetitive sequences—on alignment signature density for SV site recognition, and integrates progressive signature clustering with haplotype-resolved local assembly and refinement for robust SV reconstruction and genotyping. These strategies enable HitSV to achieve high sensitivity and precision across long-read, short-read, and hybrid sequencing technologies, while substantially reducing the sequencing depth of long-read data required for reliable SV detection, as well as maximizing SV discovery from short-read data. Importantly, by improving SV reconstruction precision in repeat-rich and structurally ambiguous regions, HitSV significantly enhances the resolving complex forms of SVs. We further applied HitSV to scan the large-scale Chinese cohort and identified composite repeat architectures featured by tandem-mobile element composite array, underscoring composite repeat organization as a pervasive and previously underrecognized structural feature of human genomes

Results

Overview of HitSV

HitSV is an alignment-based approach for structural variant (SV) detection that supports long-read, short-read, and hybrid sequencing data. It addresses a long-standing limitation of existing SV callers: SV-associated alignment signals are systematically distorted in repetitive and complex genomic regions, leading to missed, fragmented, or misrepresented variants. To overcome this, HitSV redesigns the core stages of alignment-based SV discovery—SV site recognition, SV signature clustering, and SV reconstruction and genotyping—by explicitly modeling the interaction between genomic repetitiveness and signature intensity (Fig. 1; Online Methods).

SV site recognition. HitSV adopts a repetitiveness and signature density aware breakpoint recognition strategy that models how SV signatures density under different degrees of local repetitiveness, using population-scale reference data from long-read and short-read sequencing (40 shared samples in HPRC⁴¹ and 1KGP⁶). HitSV constructs empirical distributions that jointly capture repeat context and multiple alignment signals, including sequence divergences in alignments, clipping, and discordant read pairs. These distributions inform rule-based, data-type-specific criteria for identifying candidate SV sites, enabling sensitive detection in repeat-rich regions where conventional split-read, paired-end, or consecutive indel signals are attenuated or obscured (Supplementary Fig. 1).

SV signature clustering. SV-spanning reads frequently exhibit confounded breakpoint signals that remain intractable to existing clustering strategies, particularly in large, heterozygous, or repeat-associated rearrangements. HitSV addresses these challenges through a progressive clustering strategy that treats SV breakpoints as latent variables inferred from SV signatures (Supplementary Fig. 2). By iteratively refining breakpoint probability

distributions and greedily integrating compatible clusters, HitSV aggregates fragmented evidence from disperse signals across distant windows and disentangles overlapping signatures from distinct SV alleles. This enables accurate recovery of complex events associated with large insertions, deletions, inversions, and translocations.

SV reconstruction and genotyping. To reconstruct SV alleles with high fidelity, HitSV performs haplotype-resolved local assembly using a unitig-based de Bruijn graph (unitig-DBG) that integrates both variant-supporting and non-variant reads within SV-affected regions (Supplementary Fig. 3). For short-read data, coverage-informed graph traversal generates candidate contigs beyond intrinsic read-length limits. For long-read data, HitSV constructs candidate contigs by clustering read-derived unitig-DBG paths: it begins by selecting a dominant path and iteratively refines contig structure via SDP-based read realignment and path updating. With hybrid data, HitSV uses long-read contigs as structural backbones and polishes them with short-read consensus. Candidate alleles are evaluated based on k-mer statistical modeling that realigns constituent reads to identify the most probable contig pairs for SV calling and genotyping. By decoupling SV reconstruction from fixed read-length constraints and integrating complementary sequencing signals, HitSV mitigates the trade-off between conservative yet truncated allele inference and aggressive but error-prone reconstruction, enabling high-resolution and haplotype-aware characterization of complex SVs.

GIAB HG002 Benchmark illustrates the high yields and versatility of HitSV

Using the Genome in a Bottle (GIAB) HG002 benchmark with the NIST Q100 truth set (V1.1^{42, 43}), we systematically evaluated the performance of HitSV across long-read (30× coverage, PacBio HiFi and ONT R10.4.1), short-read (Illumina, 60× and 35× coverages), and hybrid sequencing (mixed PacBio/ONT and Illumina) configurations (Supplementary Table 1). Evaluations were conducted using Truvari⁴⁴ with both standard and refinement-aware matching criteria, enabling assessment of breakpoint-level concordance as well as sequence-level reconstruction accuracy. Eleven representative state-of-the-art SV callers (LRS: Sniffles2⁴⁵, cuteSV2¹⁷, SVDSS2¹⁹, sawfish⁴⁶, SVision-pro⁴⁷; hybrid: Blend-seq⁴⁸; SRS: DRAGEN³⁰, GATK-SV³, Manta¹⁰, Lumpy¹¹, Delly¹²) were included for comparison under identical evaluation settings (Supplementary Table 2).

In all scenarios, HitSV consistently achieved the highest genotype F1-scores (F1-GT) for SV discovery. The F1-GT of 81.3%/83.7% (30× PacBio/ONT), 63.5% (4× LRS + 30× SRS) and 39.5%/42.1% (35×/60× SRS) surpassed existing SOTA tools by 11.5-67.9%, 12.5% and 2.9-35.5% for LRS, hybrid and SRS configuration, respectively, demonstrating its outstanding ability and versatility. The advantage of HitSV becomes larger with more stringent Truvari parameters (Fig. 2a and Supplementary Tables 3-6). These results highlight the ability of HitSV to generate accurate genotyped SV across sequencing technologies.

Notably, a substantial fraction of HitSV's false-positive/negative calls were attributable to inconsistencies in variant representation relative to the ground truth (e.g., Supplementary Fig. 4). Therefore, in addition to breakpoint-level agreement, we evaluated sequence-level reconstruction accuracy using Truvari 'refine' mode. Under all stringency thresholds, HitSV outperformed sawfish for both PacBio and ONT data (Fig. 2b). Specifically, HitSV achieved F1-GT of 96.5% at as high as 95% sequence similarity (-p 0.95), whereas sawfish attained an F1-GT score of 91.4%. For perfect reconstruction (-p 1), HitSV had a Fi-GT of 85.3%, higher than sawfish by 12.7% (Fig. 2b and Supplementary Table 7). This marked precision in SV sequence reconstruction originates from the utilization of all SV-spanning reads for unitig-DBG and allele selection, rather than relying solely on read segments proximal to alignment-inferred breakpoints.

The advantages of HitSV were even more pronounced in challenging genomic regions and for challenging SV types. SVs were robustly detected across highly repetitive classes including simple repeat, SINE, and satellite-enriched regions, as well as in GIAB challenging medical genes (CMRG)⁴⁹ (Fig. 2c, and Supplementary Table 8). Further, in contrast to other callers which decay rapidly for long insertions/duplications (INS), HitSV's yield remained high for INSs >1kb and produced considerable yield for INS >5kbp, generating balanced yields for

deletions (DEL) and INS in NIST-Q100 callset (Fig. 2d, Supplementary Fig. 5 and Supplementary Table 9). These results demonstrate the effectiveness of HitSV's repeat-aware modeling capability of SV signals, and its breakpoint clustering strategy which mitigate signal fragmentation and misalignment effects to produce coherent SV representations.

To assess robustness under limited sequencing depth, we down-sampled LRS datasets to find that HitSV with 6× data surpassed major callers with 30× data; and by 10× data, HitSV arrived at the saturating yield of sawfish (Fig. 2e and Supplementary Table 10). Notably, hybrid configurations combining low-coverage LRS (2–4×) with standard-depth SRS achieved performance comparable to, and in several cases exceeding, that of high-coverage LRS alone (Fig. 2e). In direct comparison with Blend-Seq, which uses LRS for SV detection and SRS for genotyping independently, HitSV's integrated SV calling yielded around 12.5% higher F1-GT, as evaluated on 4× PacBio HiFi + 30× SRS data (Supplementary Table 5). These results underscore HitSV's effectiveness in handling low-depth LRS data and the benefit of jointly modeling long- and short-read signals, providing powerful and versatile solutions for cost-effective SV calling.

Benchmark performance across increasingly sparse long-read regimes further clarified the complementary roles of LRS and SRS data in HitSV. At extremely low LRS coverage (1–2×), SRS data contributed a significant portion of the overall SV yield (Fig. 2e). Under SRS-only conditions, HitSV achieved the highest F1-GT (41.4%) among standalone tools and integrated pipelines that combine multiple tools or pan-genome references (Fig. 2a, Supplementary Table 6). This advantage is evidenced by its ability to substantially improve recall while maintaining competitive precision. In particular, HitSV was the only method recovering a considerable amount of large insertions (>1kbp) from SRS data (Supplementary Fig. 6). Stratification by alignment difficulty⁵⁰ showed that HitSV achieved 8.7–16.9% F1-GT gains over DRAGEN and GATK-SV in alignment “easy” regions (Fig. 2f and Supplementary Table 11), while exhibiting comparable performance in “hard” regions, where read realignment to complex repeats remains challenging even when SV alleles are accurately reconstructed (Supplementary Fig. 7). This issue was partially alleviated by the inclusion of LRS data in the hybrid framework, as even one SV-spanning long read was often sufficient to resolve allele selection ambiguities in SRS-based calling (Supplementary Fig. 8). Conversely, SRS data provide dense local evidence for breakpoint identification in regions insufficiently covered by long reads (Methods). As a result, at 1–2× LRS, 22.7–43.7% SVs were uniquely called with hybrid data (Fig. 2g, Supplementary Table 12). These results suggest that our framework could maximize the power of hybrid sequencing in detecting SVs. In practice, configurations such as 2× LRS+60× SRS offers a highly cost-effective solution for large-scale studies, while 4–5× LRS+60× SRS provide higher yields approaching those of deep long-read sequencing (Fig. 2e).

Finally, HitSV demonstrated practical computational efficiency across sequencing technologies and on large cohort of 1KGP phase 4 (Supplementary Note 1, Supplementary Table 13), supporting the applicability of HitSV to population-scale studies without prohibitive computational costs. Overall, HitSV represents the first SV caller that is able to handle and achieve high performance across long-read, short-read, and benchmarked hybrid sequencing data, while substantially reducing the long-read coverage required for reliable SV discovery.

The application of HitSV to large-scale genomics datasets

To evaluate the performance of HitSV at population scale using short-read sequencing, we re-analyzed the 1000 Genomes Project (1KGP) Phase 4 cohort, comprising 3,202 samples sequenced at 30× coverage (Supplementary Table 13). The HitSV-callset was compared to the official 1KGP-callset generated by a combined pipeline of GATK-SV, SVTools and Absinthe⁶. Across the cohort, HitSV identified 228,025 SVs, a 31.5% increase relative to the 1KGP-callset (173,366 SVs). Among these, 93,242 SVs were not reported in 1KGP-callset. Moreover, on average, 18.3% more SVs (11,405 vs 9,639) were identified per genome, consistently across populations (Fig. 3a).

We examined the allele frequency landscape of SVs, as it establishes foundations for population genetic

studies. Compared to 1KGP-callset, HitSV-callset produced a smoother and faster decaying allele count distribution for all superpopulations, with a clearer separation between rare and common variants shown in bi-modal forms (Fig. 3b&c). Notably, upon detailed inspection, a substantial fraction of SVs previously classified as population-specific or ultra-rare in 1KGP-callset were re-assigned as shared variants across multiple superpopulations (Fig. 3d). For example, 0.29% of African-specific SVs in 1KGP-callset were found to be common ($AF \geq 0.05$) in at least one other superpopulation, and 0.17% were found to be commonly shared in all superpopulations. In addition, 1.52-2.84% of population-specific SVs had their frequencies strongly elevated (Fig. 3e, Supplementary Fig. 9). Moreover, 36.6% of newly identified SVs were population-specific, increasing population-specific SVs by 23.5–36.4% per superpopulation (Supplementary Fig. 10). Although over 80% of these unique SVs were extremely rare ($AF < 0.002$, Supplementary Fig. 10 inset), given HitSV's high precision, most of them could be plausible.

Stratification by genomic context revealed that the largest gains achieved by HitSV occurred in repetitive regions hard to resolve by conventional callers, particularly tandem repeats (TRs) (Fig. 3f). While short TRs can be relatively reliably called from SRS data^{51, 52}, HitSV was able to discover multi-kilobase TR expansions, extending TR unit variations to $> \pm 500$ copies relative to the reference genome and revealing stable fractions of population-unique TR alleles (Supplementary Fig. 11). Lastly, many HitSV-unique SVs were found in coding and other regulatory regions (Fig. 3f). To assess functional relevance, we annotated HitSV-unique variants using seven public open resources^{6,53} (Supplementary Fig. 12). More than 7,300 HitSV-unique SVs were predicted to cause loss-of-function effects; three large deletions (10–40 kb) unique to EUR and AFR were related to retinal dystrophy⁵⁴, spastic paraplegia and congenital hypothyroidism⁵⁵, respectively; and a 50 bp common insertion was related to primary ciliary dyskinesia, with functional annotations provided by AnnotSV⁵⁶ (Supplementary Table 14).

Together, by recovering SVs missed by existing pipelines and reshaping population-level allele frequency distributions, HitSV highlights the continued value of existing short-read resources for population genetics and disease-associated studies, particularly when coupled with methods explicitly designed to handle repetitive and structurally complex genomic regions.

HitSV resolves genomic complex structural variants

Complex structural variants (CSVs), characterized by clustered breakpoints and multi-event architectures, represent a major challenge for alignment-based SV calling approaches. To assess the ability of HitSV to resolve such variants, we evaluated its performance using a curated HG002 CSV benchmark as well as two GIAB trio datasets, focusing on both detection accuracy and sequence-level reconstruction (Methods).

We first constructed a CSV benchmark for HG002 by including regions containing multiple SV events within 500 bp proximity⁵ in the NIST Q100 benchmark. Based on this benchmark and using HG002 LRS data, HitSV outperformed other SV callers (Fig. 4a&b, Supplementary Table 15). Importantly, the performance advantage of HitSV became more pronounced as CSV complexity increased. In extremely complex regions (local repetitiveness $R > 120$ and > 6 SVs), HitSV still achieved F1-GT of 53.3%, whereas other tools consistently fell below 0.4. Further, we assessed the sequence reconstruction accuracy under 'refine' mode ($-p 0.95$) in CSV regions for HitSV and sawfish, both supporting haplotype-level SV reconstruction. HitSV achieved 86.7-100% F1 whereas sawfish lagged by 8.9-36.2% for ONT data (Fig. 4c and Supplementary Table 16), highlighting the substantial improvement of HitSV under stringent evaluation criteria.

To better understand the sources of difficulty in CSV detection, we examined alignment characteristics in CSV regions: (i) heterogeneous SV signatures: CSV loci exhibited increased breakpoint heterogeneity relative to simple SVs, quantified by elevated cumulative Kullback–Leibler divergence between alignment-derived and ground-truth breakpoint distributions (Methods). Increasing heterogeneity corresponded to larger performance gaps between HitSV and signature-clustering methods such as Sniffles2 (Fig. 4d; Supplementary Table 17), indicating improved

resilience of HitSV to breakpoint crowding and signal interference. For example, within the intron of *GALNT9*, HitSV correctly phased and resolved clustered SV signals, whereas other methods showed substantial discrepancies (Fig. 4e; Supplementary Fig. 13); (ii) alignment distortion: CSV regions frequently showed reduced gap-compressed identity (Fig. 4f; Supplementary Table 18), where concordant small mismatches and indels replaced single large events. In the exon of *CYP4F12*, HitSV recovered integral insertions and deletions with near-complete variant length, whereas other methods failed to detect insertion events (Fig. 4g; Supplementary Fig. 14); (iii) abnormal local alignments: CSV loci often display low coverage or extensive clipping without reliable split alignments; nevertheless, HitSV retained sensitivity and reconstructed SVs from such limited evidence (Supplementary Figs. 19, 22–23).

We next applied HitSV on two GIAB trio-samples' LRS datasets to assess its generalizability. Across both trios, HitSV achieved higher Mendelian consistency rates than other methods, indicating improved accuracy of SV detection and genotyping in a familial context (Fig. 4h and Supplementary Table 19). The advantage of HitSV was most pronounced for CSVs exhibiting multi-allelic heterozygosity. In such cases, HitSV clustered read-derived unitig-DBG paths according to their discrepancies in the unitig-DBG and reconstructed allelic sequences independently. This capability reduces the risk of collapsing signals from multiple alleles into a single, incorrect SV representation, a common failure mode of signature clustering-based approaches. A representative case at the *GSE1* gene, comprising an insertion-rich paternal allele and an insertion-deletion mixed maternal allele, is shown in Fig. 4i and Supplementary Fig. 15. HitSV uniquely reconstructed both alleles in HG002 and concordantly the parental haplotypes (HG003 & HG004), whereas other tools largely missed SVs or produced coarse merged calls.

Collectively, these results demonstrate that HitSV extends accurate SV detection beyond simple events to highly complex rearrangements through the integration of repetitiveness aware site recognition, probabilistic breakpoint clustering, and haplotype-resolved local allele reconstruction. More examples can be found in Supplementary Figs. 16–23.

A comprehensive panel of CSVs in 557 Chinese individuals

To interrogate the CSV landscape in human populations, we applied HitSV on the large-scale ONT sequencing resource from the China 100K Genomes Project (CHN100K)⁵⁷ (Methods), generating a comprehensive callset of CSVs in 557 Chinese individuals. This callset enabled systematic characterization of CSVs across a large and relatively under-represented population.

Across the cohort, HitSV identified a total of 4,807 CSV loci (Fig. 5a, Supplementary Table 20). 65.1% CSV loci were associated with genomic tandem repeats (TR), where canonical alignment-based callers fail to resolve (Supplementary Fig. 24). To evaluate the novelty and generalizability of the identified CSVs, we further utilized 94 haplotype-resolved T2T contigs from 47 individuals in the Human Pangenome Reference Consortium (HPRC) (Supplementary Fig. 25). A substantial fraction (37.1%) of CSV loci identified in CHN100K were absent from the HPRC dataset, indicating that many complex variants are either population-enriched or require larger cohorts for reliable detection. Likewise, among shared loci, CHN100K samples exhibited broader allele distributions and more extensive structural diversity (Fig. 5b, Supplementary Fig. 26 and Supplementary Table 21–22).

TR-rich CSV loci showed remarkable accumulation of mobile-element-associated sequences (44.1%, 1380/3130), consistent with previous observations that mobile elements are frequently retro-transpose within or near tandem repeats⁵⁸. However, our analysis further revealed the widespread occurrence of coordinated expansion or contraction of TRs and mobile element insertions (MEIs), forming TR-MEI composite arrays at 89 loci (Methods, Supplementary Fig. 27). Strikingly, 95.5% of such arrays were common alleles (AF>5%) within the cohort, rather than rare or singleton events, suggesting that these structures represent stable and recurrent features of human genomic variation (Supplementary Fig. 28a and Supplementary Table 23).

Accurate resolution of TR–MEI composite arrays require reconstruction of long, multi-allelic structures. By aggregating heterogeneous signals and graph-based allele reconstruction, HitSV captures both repeat length variation and embedded MEI sequences (Supplementary Figs. 29-30). Although multiple transposable element families were involved, copy number and length contribution were decoupled: while L1, ERV1, ERVL, DNA transposons, and SINEs were the most frequently observed components (Supplementary Fig. 28b), only ERV1- and SVA-derived sequences were overrepresented by total length relative to genomic background, whereas DNA transposons were depleted (Fig. 5c, Supplementary Table 24). This decoupling underscores the structural heterogeneity of TR–MEI arrays. Although Alu and L1 exhibited higher copy numbers within TR–MEI arrays, they are typically present as short internal fragments. In contrast, SVA elements, despite being less frequent, contributed disproportionately to array length due to their closer-to-full-length sizes (Supplementary Fig. 31). Given the non-autonomous nature of SVA retrotransposition, together with its composite sequence architecture, SVA insertions may be more likely to persist as extended structural segments once captured within TR contexts, rather than being repeatedly fragmented and amplified. Lastly, despite their modest copy numbers, TR–MEI composite arrays can generate levels of genomic length variation similar to canonical VNTR (Supplementary Fig. 32, Supplementary Table 25), highlighting their potential contribution to structural polymorphism and downstream phenotypic effects.

Supporting this regulatory potential, TR-MEI composite arrays were frequently located in gene or gene-proximal regions (Fig. 5d, Supplementary Table 26). Representative loci illustrate that segments within composite arrays can span putative exonic and promoter-associated regions, or contain Alu-derived sequences overlapping predicted enhancer elements. At these loci, moderate expansion leads to ~3 copies at high allele frequencies (19.7%), with extreme expansions reaching ~24 copies in rare cases (~0.5%), substantially amplifying regulatory sequence dosage (Fig. 5e,f; Supplementary Fig. 33).

In addition to repeat-associated CSVs, 34.9% of CSV loci identified in CHN100K were not anchored in repetitive contexts. These loci feature compact breakpoint clusters interleaved with successive insertions and deletions within short genomic intervals (~2 kb), forming highly structured multi-allelic configurations (Fig. 5g). Extending recent long-read population studies of long-range complex structural polymorphism at functional loci^{59, 60}, these findings capture microscale compound rearrangements, characterized by high allelic diversity and moderate to high allele frequencies across the cohort. These features suggest that local sequence organization, chromatin context, or selective retention may facilitate repeated remodeling of the same genomic interval yet of much shorter spans.

Collectively, the CHN100K-based analysis using HitSV provides a population-scale view of CSV at a resolution previously inaccessible. This reveals TR–MEI composite arrays and microscale compound rearrangements as two widespread but previously underappreciated modes of human genome variation.

Discussion

Comprehensive characterization of SVs has long been a cornerstone of genomic research, yet continues to be constrained by the interplay between read length, the intrinsic complexity of the human genome and sequencing cost. The results presented here demonstrate that systematic distortion of alignment signals in repetitive and structurally ambiguous regions represents a primary bottleneck for SV discovery across technologies. HitSV establishes a new paradigm enabling platform-agnostic, generalizable SV detection, with unprecedented performance.

The remarkably high sensitivity achieved by HitSV can be primarily attributed to the improved signal aggregation, by implementing repetitiveness and signature density aware site recognition coupled with probabilistic breakpoint clustering. By reconciling weak or conflicting alignment evidence into coherent SV hypotheses, HitSV

achieved an improvement of 2.9-35.5% over existing short-read callers, demonstrating that effective modeling of alignment uncertainty can partially compensate for limited data length.

Accurate reconstruction of SV alleles is critical for genotyping and downstream functional interpretation, yet is frequently compromised when variant representation relies on breakpoint-proximal alignments alone. HitSV achieved haplotype-resolved local assembly by transforming all the local reads into unitig-DBG paths and implementing path discrepancies-based clustering and merging, to fully reconstruct allelic-specific sequences. HitSV outperforms the state-of-the-art long-read callers by 11.5-67.9% in F1-GT scores. Moreover, the refined F1-GT scores of HitSV reached 96.5% at the 95% sequence similarity ($p < 0.05$), highlighting its outstanding capability of near-complete SV reconstruction for an individual genome.

A distinctive strength of HitSV lies in its unified modeling of long- and short-read data, leveraging their complementary advantages of both platforms. Low coverage (2-4 $\times$) long-read ensures contig continuity and haplotype separation, whereas short-read data (30-60 $\times$) provides refined alleles and statistical evidence, resulting in a 13% F1-GT score improvement than state-of-the-art hybrid pipeline. Such a native integration allows to achieve equal or higher yields than deep long-read sequencing, offering a practical and cost-effective alternative for scaling up population genomic studies.

Improved sensitivity on short-read data alone also enables meaningful re-analysis of existing large cohorts. Application to the 1KGP cohort uncovered a substantial number of previously missed SVs and reshaped allele-frequency distributions across populations. Notably, variants previously regarded as population-specific or ultra-rare were frequently reassigned as shared at appreciable frequencies, indicating that technical limitations in earlier pipelines contributed to biases in population stratification. Although short-read data remain insufficient for resolving complex architectures, these results highlight that methodological advances can substantially extend the value of legacy datasets.

Beyond generic SV discovery, HitSV enables systematic exploration of complex SVs when sufficient long-read information is available. Benchmark and trio-based evaluations show improved resolution of clustered, multi-allelic, and haplotype-specific rearrangements relative to alignment-driven approaches lacking robust clustering and reconstruction. Extending these analyses to a large Chinese cohort reveals thousands of CSV loci and uncovers structural patterns beyond canonical SV categories, including TR-MEI composite arrays and microscale compound rearrangements. The prevalence, allelic diversity, and moderate-to-high population frequencies of these structures suggest that coordinated repeat instability, retrotransposition, and localized sequence remodeling represent stable modes of genomic variation rather than rare mutational events.

Extremely large or highly entangled rearrangements continue to pose challenges for local assembly-based reconstruction, while the ambiguous representation of complex alleles remains a significant hurdle for downstream annotation. Furthermore, while hybrid sequencing has been evaluated within controlled benchmarks, its performance across large-scale, real-world cohorts remains to be established as such datasets emerge. Collectively, by integrating repetitiveness and signature density aware breakpoint recognition coupled with precise haplotype-resolved local assembly across sequencing technologies, HitSV significantly expands the observable spectrum of structural variation, providing a scalable and valuable foundation for population genomics and the characterization of complex genetic architectures.

Online methods

SV site recognition

Local repetitiveness and genome windows. The reference genome was divided into 600 bp size windows with 300bp overlaps. For each of the windows, a measure of local repetitiveness (termed as R) was defined as the maximum length of repeated K -mers in the window:

$$R = f(s, [l, r]) = \max \{K | \exists i, j \in [l, r], i \neq j, s_i s_{i+1} \dots s_{i+K-1} = s_j s_{j+1} \dots s_{j+K-1}\} \quad (1)$$

where, s is the reference and $[l, r]$ is the coordinates of the window. R rises with local sequence repetitiveness.

SV signature collection. SV signatures were identified in each genome window. Multiple kinds of colinear and non-colinear SV-associated signatures were collected, including sequence divergences in alignments (SDAs), large clippings (CLPs) and discordant read pairs (DRPs, only for short reads). Herein, SDA is defined by both of dense sequence divergences (mismatches and indels) shown by CIGAR information and split alignment.

Given short and long-read alignment as input, HitSV extracts SDA signatures i.e., dense sequence divergences in CIGARs (single long insertions/deletions $\geq \text{min_SV_length}$ or dense clusters of small variants, defined as more than N (10 for LRS and 3 for SRS) small-variant events occurring within a 200-bp region.) and split alignments, and CLP signatures, i.e., large clippings in CIGAR, by traveling the alignments. For short-read, DRP signatures are derived by checking the inserted size and gap between the reference and read. All the signatures are represented as 4-tuples:

$$\text{Sig} = (R, N_{\text{SDA}}, N_{\text{CLP}}, N_{\text{DRP}}) \quad (2)$$

Where, Sig denotes the signal intensity within the window, N_{SDA} , N_{CLP} , and N_{DRP} represent the numbers of SDA, CLP, and DRP signatures within the same window, respectively.

Signature distributions: SV signals manifest heterogeneously across sequencing technologies and genomic contexts, particularly under varying levels of local repetitiveness. Through extensive empirical inspection of long- and short-read alignments, we observed several recurring yet poorly quantified phenomena: (i) different SV signatures dominate in long- versus short-read data; (ii) the strength of these signatures varies systematically with local repeat complexity; and (iii) SV-like signals occasionally appear in regions without SV events, especially within highly repetitive sequences. HitSV designs a model to describe the correlation between SV signatures and local repetitiveness, furthermore to distinguish true SV-supporting patterns from background noise in a principled manner.

HitSV identifies candidate SV loci using pre-built signature distributions that quantify local signature intensity across repetitiveness levels. Separate distributions were constructed for long- and short-read data using sequencing datasets from 40 samples shared in Human Pangenome Reference Consortium and 1000 Genomes Project, respectively. Assembly-based SV callsets from HPRC were used to label SV-overlapping windows (SV-windows), within which SDA, CLP, and DRP signals were quantified. Two-dimensional empirical distributions jointly modeling repetitiveness and signature intensity were then constructed for SV-windows, alongside matched background distributions from non-SV windows (Supplementary Fig. 1), where the x-axis denotes local repetitiveness and the y-axis denotes signature count intensity.

Analysis of these distributions revealed four key empirical patterns that systematically explain the observed heterogeneity of SV signatures:

1) SDA exhibits sequencing technology-specific behavior. In long reads, SDA signals are strong and largely independent of repeat complexity, reflecting direct encoding of most SVs as long indels or split alignments. In short reads, SDA intensity decreases with increasing repetitiveness.

2) CLP signals decay with repeat complexity in both data types. This reflects increased sequence similarity between SV alleles and local repeats, enabling near-complete read alignment. CLPs are rare but informative in long reads, whereas they constitute the dominant SV signal in short reads due to the short parts involving in SV sites usually cannot be confidently mapped.

3) Combined multi-signature evidence is prevalent across most SV windows. Though a small fraction of

windows—mainly in long-read data—lack detectable signatures, typically near large SV breakpoints, where reads align colinearly with minor mismatches. This has little impact on detection, as such SVs span multiple adjacent windows, allowing sufficient evidence to be captured in adjacent regions.

4) Background SV-like signals occur sporadically in non-SV windows. These arise primarily from outlier reads with abnormal errors or insert sizes and rarely exhibit high intensity across multiple signatures simultaneously, leading the background noise can be largely mitigated by jointly considering multiple signatures. Such noise is more frequent in highly repetitive regions, where attenuation of true SV signals reduces separability. To mitigate this, HitSV explicitly flags highly repetitive regions (e.g., $R > 40$) as candidate SV loci; these regions represent only 1.46% of the reference genome yet are enriched for SV hotspots.

SV candidate site recognition. Guided by the empirical signature distributions, HitSV adopts a rule-based model to identify candidate SV windows, with separate recognition criteria tailored to long- and short-read sequencing.

For long-read data, SV signals are dominated by SDA and CLP signatures. A genome window is therefore classified as an SV candidate site if more than T_{LR} reads within it exhibit either indels longer than 30 bp or clipped segments exceeding 300 bp, as recorded in their CIGAR strings. For short-read data, where SDA, CLP, and DRP signatures jointly contribute and signal attenuation is prominent in repetitive regions, a window is recognized as an SV candidate site if any of the following conditions is satisfied: 1) the local repetitiveness score R is higher than 40; 2) the number of SDA reads is higher than $\text{MAX}(T_{SDA}, 2)$; 3) the number of CLP reads is higher than $\text{MAX}(T_{CLP}, 2)$; 4) the number of DRP reads is higher than $\text{MAX}(T_{DRP}, 2)$.

Here, the default values of the thresholds T_{LR} , T_{SDA} , T_{CLP} and T_{DRP} are set to 10% of local read depth, ensuring robustness across variable coverage while suppressing spurious signals from individual outlier reads. Using these distribution-informed rules, HitSV achieves high sensitivity in SV site recognition on the HG002 benchmark, recovering 98.1% of SV sites (19,234 sites) for long-read data and 89.2% (17,478 sites) for short-read data, demonstrating effective recall of true SV loci prior to downstream refinement.

SV signature clustering

It is difficult to comprehensively collect all SV-associated windows and obtain a purified read set for downstream assembly, which arises from i) SVs may span multiple windows owing to multiple breakpoints or dispersed repeats and ii) signals from distinct SV events may become entangled due to heterozygosity or complex rearrangements. To solve this problem, HitSV employs a progressive approach with two phases for refined SV signature clustering and candidate event inference.

The first phase infers SV breakpoints by modeling them as latent variables and constructing probability distributions from SV signatures, enabling the removal of unlikely signals and yielding purified clusters (Supplementary Fig. 2a). For SDA clustering based on CIGAR strings and split alignments, individual signals are assigned inferred breakpoint probability distributions: dense small-variant signals are uniformly distributed across their affected intervals, while isolated large signals are uniformly distributed within pre-defined length windows (default: 600 bp) centered on their indicated start and end positions. These signal distributions are then aggregated to identify a global probability peak, which defines the final inferred breakpoint. Subsequently, HitSV filters signatures by assessing whether their centroids are sufficiently close to this inferred breakpoint to ensure clustering accuracy. For CLP clustering, individual signals are assigned a uniform breakpoint probability distribution centered around their clipping bases. For DRP clustering, the breakpoint probability distribution of each signal is calculated by integrating the precomputed insert size (INSIZE) distribution with the orientation and distance of the corresponding paired read-ends. For CLP and DRP clusters, the final inferred breakpoint is obtained through a process similar to that of SDA clustering.

In the second phase, HitSV scans and links all the clusters by nearby regions or large signals (SDA/DRP) through a greedy approach (Supplementary Fig. 2b), enabling comprehensive window collection for downstream SV allele reconstruction and genotyping. This strategy handles SVs with multiple or ambiguous breakpoints and integrates windows affected by the same event. For example, inversions—typically involving two breakpoints and four windows—are grouped in a single step (Fig. 1c). This benefit is especially pronounced for dispersed repeats like mobile element insertions (MEIs; Supplementary Fig. 2). In such cases, reads align ambiguously and are hard to collect; however, proximal or linked candidate windows allow the greedy strategy to aggregate them transitively.

Haplotype-resolved local assembly for high-fidelity SV reconstruction

Read collection for local assembly. After clustering all signals of the same SV event, HitSV rescans the BAM within the event-affected region—extended upstream and downstream by a fixed length (default 500 bp)—to collect reads for assembly. To avoid missing variants and to enhance assembly ability and genotyping accuracy, HitSV includes all reads in this region, regardless of whether they contain any SV signatures or not, enabling an optimal local assembly.

unitig-DBG construction from reads. HitSV builds a unitig de Bruijn graph (unitig-DBG)⁶¹ based on k-mers from all the collected reads and the local reference sequence. Each unitig of the unitig-DBG records average reads coverage and is classified as high- or low-depth using a threshold depending on pre-estimated sequencing depth (default $\geq 20\%$ of expected depth, minimum 2). Each unitig is labeled as cyclic or acyclic after detecting cycles in the unitig-DBG. Acyclic unitigs shared with the reference are treated as anchors. Finally, all reads are decomposed into paths through the unitig-DBG for subsequent read-to-path realignment.

Short-read based SV calling. When cycles occur in the assembly graph, many current methods avoid traversing them, often preventing the recovery of valid allele sequences and leading to incomplete or incorrect SV allele reconstructions. HitSV addresses this challenge by constructing candidate SV alleles through graph walking under coverage constraints (Supplementary Fig. 3). **(i) Exhaustive (“blind”) contig generation.** The upstream and downstream anchor nodes closest to the SV are selected as the source and sink. HitSV performs depth-first search on high-depth nodes in the DBG, assigning each node a maximum visit count (default 9). A path is accepted only if it connects source to sink and the visit counts of all traversed nodes remain below the maximum threshold. This operation enables exhaustive contig enumeration at affordable cost through implicit coverage control. **(ii) Contig filtering.** HitSV realigns all reads to the generated contigs above, and reject contig without sufficient reads supporting within any base of it. **(iii) Haplotype pairing and allele inference.** Remaining contigs are treated as candidate haplotypes, and assuming diploidy, all $N_{HAP} \times (N_{HAP} + 1)/2$ haplotype pairs are evaluated. For each pair (H_1, H_2) , HitSV computes a likelihood by the following equation:

$$L(H_1, H_2) = -\frac{(2X(H_1, H_2) + Y(H_1, H_2))}{D_R} \quad (2)$$

where D_R is pre-estimated sequencing depth, $X(H_1, H_2)$ is the number of unmapped reads, and $Y(H_1, H_2)$ represent the concordance between read and haplotype k-mer counts, which is computed as following:

$$Y(H_1, H_2) = \sum_{kmer \in kmer(H_1, H_2)} |R(kmer_i, (H_1, H_2), RS) - \bar{R}((H_1, H_2), RS)|, \forall kmer_i \in kmer(H_1, H_2) \quad (3)$$

$$R(kmer_i, (H_1, H_2), RS) = \frac{count(kmer_i, (H_1, H_2))}{count(kmer_i, RS)} \quad (4)$$

where $kmer(H_1, H_2)$ indicates the k-mer set of (H_1, H_2) , RS indicates the read set, $count(kmer_i, (H_1, H_2))$ and $count(kmer_i, RS)$ indicate the counts of a specific k-mer ($kmer_i$) on (H_1, H_2) and RS , respectively. The haplotype-pair having highest likelihood is selected and realigned to reference (with the graph walking information)

using ksw2⁶² to infer SV calls.

Long-read based SV detection. After unitig-DBG construction, each long read is represented as a path through the graph. HitSV selects the read traversing the largest total length of high-depth nodes as initial path, and then assigns remaining reads to existing paths or builds new ones. Through iterative alignment, clustering, and updating, HitSV produces final path set. (i) **SDP-based read re-alignment.** For each read, its unitig-DBG-path is aligned against those of all current paths by SDP. A read supports a contig if no structural discrepancy is detected, defined as follows: if, at any SDP-anchored node (i.e., flanked by SDP-matched nodes), either (i) the gap length in two adjacent match nodes is more than 20 bp, or (ii) both allele nodes are high-depth but differ in sequence. (ii) **Path updating.** Reads aligned successfully refine the corresponding path by replacing low-depth nodes at SDP-anchored positions with read-derived high-depth nodes, whereas reads are rejected by all existing contigs initiate new paths. After all reads are processed, each path is converted into a contig.

Hybrid-based SV calling and polishing. When both short- and long-read data are available, HitSV generates candidate contigs using unitig-DBG built from long and short reads, respectively. If long-read based contigs exist, HitSV prioritizes them as backbones, and uses high-accuracy short-reads to improve the base-level accuracy of backbones. In detail, i) HitSV builds an empirical 9-mer error profile by calculating the accuracy probability of all genomic 9-mers from alignment of HG002-ONT data (by inhouse script `error_profile_analysis.py`), using only primary alignments, and all the k-mers harboring known variants from GIAB-V4.2.1 were filtered out. ii) HitSV first identifies the positions of all low-fidelity 9-mers (error rate > 0.2) based on sequencing error profiles and homopolymer contexts in the backbone sequence, and then extracts all short reads overlapping these positions to construct multiple local consensus sequences. A representative consensus is selected based on majority short reads support. Homopolymer lengths and 9-mer sequences are subsequently corrected by comparing the selected consensus with the backbone sequence, thereby refining local sequence accuracy for downstream analysis.

Then, HitSV reassigns all short reads to polished contigs to augment support estimation by re-alignment. When long-read contigs are unavailable, short-read-based contigs are used as the final assemblies, and long reads are subsequently incorporated to refine genotype accuracy (see below).

SV genotyping and quality control

HitSV infers SV genotypes by genotyping candidate contig pair with a probabilistic model, in which the probability is computed through a support matrix derived from read evidence in the local assembly region. The support matrix is an $M \times N$ matrix, where M is the number of reads and N is the number of contigs (2 for a diploid assumption). Each matrix element represents the probability of observing a specific read given a particular contig in the pair. For short-read data, each read is aligned to both contigs in the candidate pair, and the resulting alignment scores are transformed to construct a likelihood-based support matrix using an exponential function. 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 support matrix is used to estimate the probability for each genotype of the contig pair, and all SVs called from the contig pair are genotyped with the highest probability correspondingly. A Phred-scaled quality score is also computed by the probability. Homozygous-reference calls are discarded, variants having quality lower than a predefined threshold (default: 10) are labeled LOW, and the remaining calls are marked PASS.

Complex structural variation

Definition of CSV: In this study, CSVs are defined as groups of SVs that occur on the same haplotype and whose breakpoint positions lie within 500 bp of each other. Under this definition, any pair of SVs with breakpoint proximity <500 bp is merged into a larger CSV event.

HG002 CSVs benchmark: For HG002, CSV benchmark intervals were constructed as follows. All variants in the NIST HG002 Q100 v1.1 truth set were first partitioned into two haplotype-specific SV sets based on their genotypes. Within each set, SVs were clustered using a maximum breakpoint distance of 500 bp and a minimum SV length of 50 bp, yielding breakpoint groups corresponding to candidate CSVs. To ensure robust downstream evaluation, each breakpoint was padded by ± 500 bp to account for potential uncertainty in the reported variant boundaries, and all resulting intervals were stored in BED format. These preliminary CSV intervals were then sorted and merged using bedtools, and intersected with the high-confidence benchmark regions (GRCh38_HG2-T2TQ100-V1.1_stvar.benchmark.bed) to retain only regions suitable for reliable assessment.

In total, 2,274 non-overlapping CSV intervals were obtained for HG002, spanning approximately 3.77 Mb. Each CSV interval was further classified along two dimensions for evaluation the ability for CSV identification: i) CSV complexity, defined by the number of constituent haplotype-resolved SVs (1-2, 3-6, or >6 events), and ii) local repeat complexity, measured by the local repetitiveness, R ($R < 40$, $40 \leq R < 119$, or $R \geq 120$). This yielded nine mutually exclusive CSV categories used throughout the benchmarking analysis.

Kullback–Leibler divergence: For the HG002 sample, we computed the Kullback-Leibler divergence (KLD) within each CSV interval to quantify the discrepancy between the observed read-alignment pattern and the expected genotype-derived breakpoint distribution. KLD was calculated in a sliding-window manner using 10 bp windows across each CSV region. For a given window, the local KLD is defined as $p \cdot \log\left(\frac{p}{p_a}\right) + q \cdot \log\left(\frac{q}{q_a}\right)$, where p denotes the fraction of reads in the window that support the true breakpoint, and $q=1-p$. The expected probabilities p_a and q_a are determined from the ground-truth genotype: i) Homozygous SV: $p_a = 1$, $q_a = 0$; ii) Heterozygous SV: $p_a = 0.5$, $q_a = 0.5$; iii) Homozygous reference: $p_a = 0$, $q_a = 1$; To avoid zero terms, a smoothing constant of 0.01 was added to p , q , p_a , q_a computation. The regional KLD was obtained by summing the local KLD values across all windows in the CSV interval.

Panel of CSVs in population

Population-level CSV interval from 557 ONT samples: HitSV was applied to 557 ONT samples to generate per-sample VCFs, retaining variants supported by at least 2 reads. For each sample, adjacent SVs with breakpoint distance < 500 bp were grouped as CSVs, and the reference genome was partitioned into non-overlapping 500 bp windows to record sample-specific CSV intervals and allele counts (Supplementary Note 2). Population-level allele frequencies were computed across all samples, and windows with AF $\geq 5\%$ were retained. After merging adjacent intervals and excluding telomeric and centromeric regions, 2,857 population CSV intervals were obtained (refer to https://github.com/hitbc/HitSV/blob/main/cohort_csv_analysis.md).

CSV-region contigs generation and annotation: For all 557 ONT samples, local contigs within the 2,857 population CSV intervals were assembled directly from BAM files. Contigs were filtered to retain those supported by ≥ 2 reads. For comparison, haplotype-resolved contigs from the 47 HPRC samples (94 haplotypes) were obtained using the same procedure. All contigs and corresponding reference sequences from the 47 HPRC samples and 557 ONT samples were annotated using RepeatMasker (default parameters) and TRF (*trf input.fa 2 7 80 10 10 500 -h -d -ngs 1*). Based on the RepeatMasker annotations, mobile element insertion (MEI) categories (“SINE”, “LINE”, “Retroposon”, “DNA”, “RC”, or “LTR”) were parsed for every contig (Supplementary Note 2).

MEI-array identification and annotation: For each contig, elements belonging to the dominant MEI type were sorted by genomic order, and performs DBSCAN clustering based on MEI length and the distances between adjacent MEIs. A contig was classified as containing an MEI array if it met all the following criteria: i) ≥ 3 MEI elements in the dominant cluster; ii) the standard deviation of element lengths $\leq 20\%$ of their mean; iii) The standard deviation of the distances is required to be less than 20% of the mean or less than 15 bp. If a CSV interval contains at least one MEI-array contig, the interval is defined as an MEI-array interval. Using this criterion, we

identified a total of 89 CSV intervals containing MEI arrays. These CSV intervals were annotated with the GENCODE (v44), and regulatory elements were annotated using the Ensembl regulatory feature database (GRCh38, v115) (Supplementary Note 2).

Contig clustering within CSV intervals: To further characterize sequence diversity within CSV regions (Supplementary Note 2), each interval was stratified into TR-enriched and non-TR subregions based on reference sequence annotations. After splitting, the subregions of the 2,857 population CSV intervals collectively contained 4,807 CSV loci. TRF-annotated repeats and RepeatMasker categories “simple_repeat” and “Satellite” were considered as TR-enriched, whereas all other annotated segments (including MEIs) were treated as non-TR. Contigs and the corresponding reference sequences were partitioned along these boundaries to generate region-specific contig sets. Within each contig set, all contigs were compared pairwise using *ksw2* (*ksw_extd2_sse function*) to identify structural differences ≥ 50 bp.

Variants were classified as TR-derived (insertions attributable to tandem-repeat expansion within TR-enriched regions) or non-TR (insertions involving non-TR sequence or occurring outside TR regions). Two contigs were assigned to the same cluster if no structural differences were detected or if differences were limited to TR-derived events; otherwise, they were assigned to distinct clusters. For each CSV interval, the cluster number was defined as the total number of distinct contig clusters observed.

Classification of CSV interval: Based on the aggregated variant profiles above, each subregion was assigned to one of the following categories (Supplementary Note 2). 1) No SV: no structural differences detected among contigs; 2) Pure TR: all differences attributable solely to TR-derived insertions; 3) TR with non-TR SVs: non-TR insertions detected within TR-enriched subregions; 4) non-TR with TR SVs: TR-type insertions detected within non-TR subregions; 5) Pure non-TR: only non-TR insertions present within non-TR subregions; 6) non-TR with both types: both TR and non-TR insertions observed within non-TR subregions. These classifications were used to quantify the structural complexity of contigs within each CSV interval.

Benchmarking methodology

Computing environment. All experiments were conducted on a server equipped with two Intel Xeon Gold 6240 CPUs (2.60 GHz) and 256 GB of memory. During sequence alignment and variant calling, ten CPU cores were allocated per job. Workloads were managed by the Slurm scheduler, and actual memory usage and CPU time were recorded using the Slurm utility *seff*.

Datasets and down-sampling. Four real HG002 sequencing datasets were obtained from previous studies: 30× ONT long reads, 30× PacBio HiFi long reads, and 60× and 35× Illumina short reads (Supplementary Table 1). All reads were pre-aligned to the human reference genome GRCh38 with all decoy sequences using minimap2 (ONT), pbmm2 (PacBio) or BWA-MEM (short reads), and converted to BAM format, sorted and indexed with SAMtools (Supplementary Table 2). The long-read datasets were down-sampled to 1–10×, 15×, 20× and 25×, and the 60× short-read dataset to 15× and 30× using SAMtools.

Truth set and evaluation. The NIST-Q100 v1.1 callset (Supplementary Table 1) from the Genome in a Bottle Consortium (GIAB) was used as ground truth. Performance metrics, including F1 score, sensitivity, precision and genotype concordance (GT-F1 was calculated as the product of F1 score and genotype concordance), were computed with Truvari in single mode, considering only PASS variants, treating duplications as insertions (*--dup-to-in*), and setting *--sizemax* to 50,000. In non-refined mode, *pctseq* (*-p*) was set to 0 and *pctsize* (*-P*) to 0.7, 0.9 or 1.0 for benchmarking; in Truvari’s refined mode, *pctseq* was set to 0.7, 0.9 or 1.0 to assess the accuracy of SV-allele reconstruction. All evaluations were restricted to NIST high-confidence regions. Furthermore, all SVs in those regions are classified by SV type (insertions and deletions), SV size and repeat context using RepeatMasker and Challenging Medical Relevant Genes (CMRG) annotations. For short reads, HARD and EASY regions from

GIAB were also evaluated. Before benchmarking, all VCF files were sorted, compressed and indexed using bcftools.

Benchmarking configuration and tools. For all tools, ten threads were used whenever multi-threading was supported, and default parameters were applied unless otherwise specified. On down-sampled ONT and PacBio datasets, HitSV was compared with SVision-pro, sawfish, Sniffles2, cuteSV2 and SVDSS2. cuteSV2 was run with the "--genotype" option and SVision-pro with "--detect_mode germline". On down-sampled Illumina datasets, HitSV, Manta, Delly and Lumpy were evaluated at 15×, 30× and 60× coverage. To match previous DRAGEN benchmarks, HitSV was additionally run on the 35× dataset used in that study. Hybrid datasets only for HitSV, are generated by combining 1×–10× long reads (ONT and PacBio) with 30× or 60× short reads, respectively. The metrics of Blend-seq under 4× long reads (ONT and PacBio) and 30× short reads⁴⁸ were used directly for comparison.

Mendelian consistency evaluation. Mendelian consistency was assessed in two trios (HG002/3/4 and HG005/6/7, Supplementary Table 19) using ONT data. HitSV, Sniffles2 and cuteSV2 were evaluated. For each tool, trio callsets in a family were merged with SURVIVOR (parameters: 10 1 1 1 0 50), ignoring BND events and converting missing genotypes to homozygous reference. After sorting and indexing the resulting VCF files, Mendelian inheritance was examined using the bcftools mendelian2 plugin.

1000 Genomes Phase 4 dataset re-analysis. We use HitSV to call SVs for 3,202 samples (Supplementary Table 1) to obtain individual VCFs. Individual VCFs were merged using SURVIVOR (parameters: 500 1 1 1 0 50), retaining only non-BND PASS variants. SV quality scores were re-estimated by summing across supporting samples, and variants below a threshold (<50) were removed. For each SV, AC and AF were computed using bcftools and annotated with AnnotSV. The merged VCF was compared with the official 1000 Genomes high-coverage release using Truvari to identify novel SVs.

Command lines and detailed parameter settings are provided in the Supplementary Notes 2.

Data availability

All datasets used in this study are listed in Supplementary Table 1. The versions and download URLs of all tools employed are provided in Supplementary Table 2. Ten randomly selected individual VCF files across samples, together with the population VCF generated by reanalyzing the 1KGP4 high-coverage cohort (3,202 samples) using HitSV, as well as HitSV call sets produced from the HG002 ONT (30×), PacBio (30×), Illumina (60× and 35×), and hybrid datasets (4× ONT/PacBio plus 30× Illumina), and the results obtained for HG002/3/4 and HG005/6/7, are publicly available at https://github.com/hitbc/HitSV_call_results.

Code availability

The HitSV program is implemented in C++, and its source code (v2.0), auxiliary scripts, and precomputed resources for commonly used genomes and datasets (such as local sequence repetitiveness file for GRCh38 and 9-mer error profile files derived from ONT datasets) used to generate all results in this study are available on GitHub (<https://github.com/hitbc/HitSV>) under the GNU General Public License v3.0.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 62331012, No. 62572146 and No. 62402140) and the Ministry of Science and Technology of China (2024YFF1206300, 2024YFF1206200, 2023YFF1205600 and 2017YFC0907500).

Author contributions

Y. W. designed and supervised research; G. L. and B. L. developed the algorithm and software; B. L., L. Q., and Y. L. contributed to data interpretation; G. L. performed algorithm benchmarking; Q. L, Y. L. and G. L. contributed to the 1KGP4 re-analysis and CHN100K analyses. L. Q., B. L., and Y. L. wrote the paper with input from all other authors. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

References

1. Middlebrooks, C.D. et al. Association of germline variants in the APOBEC3 region with cancer risk and enrichment with APOBEC-signature mutations in tumors. *Nature Genetics* **48**, 1330-1338 (2016).
2. Hardwick, R.J. et al. Haptoglobin (HP) and Haptoglobin-related protein (HPR) copy number variation, natural selection, and trypanosomiasis. *Human Genetics* **133**, 69-83 (2014).
3. Collins, R.L. et al. A structural variation reference for medical and population genetics. *Nature* **581**, 444-451 (2020).
4. Sudmant, P.H. et al. Global diversity, population stratification, and selection of human copy-number variation. *Science* **349**, aab3761 (2015).
5. Sudmant, P.H. et al. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75-81 (2015).
6. Byrsk-Bishop, M. et al. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**, 3426-3440.e3419 (2022).
7. Weischenfeldt, J., Symmons, O., Spitz, F. & Korbel, J.O. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nature Reviews Genetics* **14**, 125-138 (2013).
8. Chiang, C. et al. The impact of structural variation on human gene expression. *Nature Genetics* **49**, 692-699 (2017).
9. Li, Y. et al. Pan-cancer proteogenomics connects oncogenic drivers to functional states. *Cell* **186**, 3921-3944.e3925 (2023).
10. Chen, X. et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220-1222 (2016).
11. Layer, R.M., Chiang, C., Quinlan, A.R. & Hall, I.M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biology* **15**, R84 (2014).
12. Rausch, T. et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333-i339 (2012).
13. Ye, K., Schulz, M.H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865-2871 (2009).
14. Fan, X., Abbott, T.E., Larson, D. & Chen, K. BreakDancer: Identification of Genomic Structural Variation from Paired-End Read Mapping. *Curr Protoc Bioinformatics* **45**, 15.16.11-11 (2014).
15. Sedlazeck, F.J. et al. Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods* **15**, 461-468 (2018).
16. Heller, D. & Vingron, M. SVIM: structural variant identification using mapped long reads. *Bioinformatics* **35**, 2907-2915 (2019).
17. Jiang, T. et al. Long-read-based human genomic structural variation detection with cuteSV. *Genome Biology* **21**, 189 (2020).
18. Chen, Y. et al. Deciphering the exact breakpoints of structural variations using long sequencing reads with DeBreak. *Nature Communications* **14**, 283 (2023).
19. Denti, L., Khorsand, P., Bonizzoni, P., Hormozdiari, F. & Chikhi, R. SVDSS: structural variation discovery in hard-to-call genomic regions using sample-specific strings from accurate long reads. *Nature Methods* **20**, 550-558 (2023).
20. Lin, J. et al. SVision: a deep learning approach to resolve complex structural variants. *Nature Methods* **19**, 1230-1233 (2022).
21. Quenez, O. et al. Detection of copy-number variations from NGS data using read depth information: a diagnostic performance evaluation. *European Journal of Human Genetics* **29**, 99-109 (2021).
22. Helal, A.A., Saad, B.T., Saad, M.T., Mosaad, G.S. & Aboshanab, K.M. Benchmarking long-read aligners and SV callers for structural variation detection in Oxford nanopore sequencing data. *Scientific Reports* **14**, 6160 (2024).

23. Ahsan, M.U., Liu, Q., Perdomo, J.E., Fang, L. & Wang, K. A survey of algorithms for the detection of genomic structural variants from long-read sequencing data. *Nat Methods* **20**, 1143-1158 (2023).
24. Wu, D. et al. Large-Scale Whole-Genome Sequencing of Three Diverse Asian Populations in Singapore. *Cell* **179**, 736-749.e715 (2019).
25. Walter, K. et al. The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82-90 (2015).
26. Gudbjartsson, D.F. et al. Large-scale whole-genome sequencing of the Icelandic population. *Nature Genetics* **47**, 435-444 (2015).
27. Zhang, S. et al. Genome-wide investigation of VNTR motif polymorphisms in 8,222 genomes: Implications for biological regulation and human traits. *Cell Genomics*.
28. Kidd, J.M. Mobile element insertions affect human pigmentation and skin cancer risk. *Nat Genet* **56**, 1546-1547 (2024).
29. Tanudisastro, H.A., Deveson, I.W., Dashnow, H. & MacArthur, D.G. Sequencing and characterizing short tandem repeats in the human genome. *Nat Rev Genet* **25**, 460-475 (2024).
30. Behera, S. et al. Comprehensive genome analysis and variant detection at scale using DRAGEN. *Nature Biotechnology* (2024).
31. Larson, D.E. et al. svtools: population-scale analysis of structural variation. *Bioinformatics* **35**, 4782-4787 (2019).
32. Li, H. et al. A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nature Methods* **15**, 595-597 (2018).
33. Heller, D. & Vingron, M. SVIM-asm: structural variant detection from haploid and diploid genome assemblies. *Bioinformatics* **36**, 5519-5521 (2020).
34. Nurk, S. et al. The complete sequence of a human genome. *Science* **376**, 44-53 (2022).
35. Bilgrav Saether, K. et al. Leveraging the T2T assembly to resolve rare and pathogenic inversions in reference genome gaps. *Genome Res* **34**, 1785-1797 (2024).
36. Garg, V. et al. Unlocking plant genetics with telomere-to-telomere genome assemblies. *Nature Genetics* **56**, 1788-1799 (2024).
37. Phillippy, A.M., Schatz, M.C. & Pop, M. Genome assembly forensics: finding the elusive mis-assembly. *Genome Biology* **9**, R55 (2008).
38. Rahman, A. & Medvedev, P. Assembler artifacts include misassembly because of unsafe unitigs and underassembly because of bidirected graphs. *Genome Res* **32**, 1746-1753 (2022).
39. Lai, S. et al. metaMIC: reference-free misassembly identification and correction of de novo metagenomic assemblies. *Genome Biology* **23**, 242 (2022).
40. Chen, Y., Zhang, Y., Wang, A.Y., Gao, M. & Chong, Z. Accurate long-read de novo assembly evaluation with Inspector. *Genome Biology* **22**, 312 (2021).
41. Liao, W.W. et al. A draft human pangenome reference. *Nature* **617**, 312-324 (2023).
42. Dwarshuis, N. et al. The GIAB genomic stratifications resource for human reference genomes. *Nat Commun* **15**, 9029 (2024).
43. Rautiainen, M. et al. Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nat Biotechnol* **41**, 1474-1482 (2023).
44. English, A.C., Menon, V.K., Gibbs, R.A., Metcalf, G.A. & Sedlazeck, F.J. Truvari: refined structural variant comparison preserves allelic diversity. *Genome Biology* **23**, 271 (2022).
45. Smolka, M. et al. Detection of mosaic and population-level structural variants with Sniffles2. *Nature Biotechnology*, 1-10 (2024).
46. Saunders, C.T. et al. Sawfish: improving long-read structural variant discovery and genotyping with local haplotype modeling. *Bioinformatics* **41** (2025).
47. Wang, S. et al. De novo and somatic structural variant discovery with SVision-pro. *Nature Biotechnology* **43**, 181-185 (2025).
48. Magnier, R. et al. Blended Length Genome Sequencing (blend-seq): Combining Short Reads with Low-Coverage Long Reads to Maximize Variant Discovery. *bioRxiv*, 2024.2011.2001.621515 (2025).
49. Wagner, J. et al. Curated variation benchmarks for challenging medically relevant autosomal genes. *Nat Biotechnol* **40**, 672-680 (2022).
50. Krusche, P. et al. Best practices for benchmarking germline small-variant calls in human genomes. *Nature Biotechnology* **37**, 555-560 (2019).
51. Cui, Y. et al. A genome-wide spectrum of tandem repeat expansions in 338,963 humans. *Cell* **187**, 6411-6412 (2024).
52. Zhang, S. et al. Genome-wide investigation of VNTR motif polymorphisms in 8,222 genomes: Implications for biological regulation and human traits. *Cell Genomics* **4** (2024).
53. Chen, S. et al. A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* **625**, 92-100 (2024).
54. Bandah-Rozenfeld, D. et al. Mutations in IMPG2, encoding interphotoreceptor matrix proteoglycan 2, cause

- autosomal-recessive retinitis pigmentosa. *The American Journal of Human Genetics* **87**, 199-208 (2010).
55. Heinen, C.A. et al. Mutations in IRS4 are associated with central hypothyroidism. *Journal of Medical Genetics* **55**, 693-700 (2018).
56. Geoffroy, V. et al. AnnotSV: an integrated tool for structural variations annotation. *Bioinformatics* **34**, 3572-3574 (2018).
57. Jiang, T. et al. A comprehensive genetic variant reference for the Chinese population. *Science Bulletin* **69**, 3820-3825 (2024).
58. Cordaux, R. & Batzer, M.A. The impact of retrotransposons on human genome evolution. *Nature reviews genetics* **10**, 691-703 (2009).
59. Schloissnig, S. et al. Structural variation in 1,019 diverse humans based on long-read sequencing. *Nature* **644**, 442-452 (2025).
60. Logsdon, G.A. et al. Complex genetic variation in nearly complete human genomes. *Nature* **644**, 430-441 (2025).
61. Zerbino, D.R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research* **18**, 821-829 (2008).
62. Suzuki, H. & Kasahara, M. Introducing difference recurrence relations for faster semi-global alignment of long sequences. *BMC bioinformatics* **19**, 45 (2018).

Figures

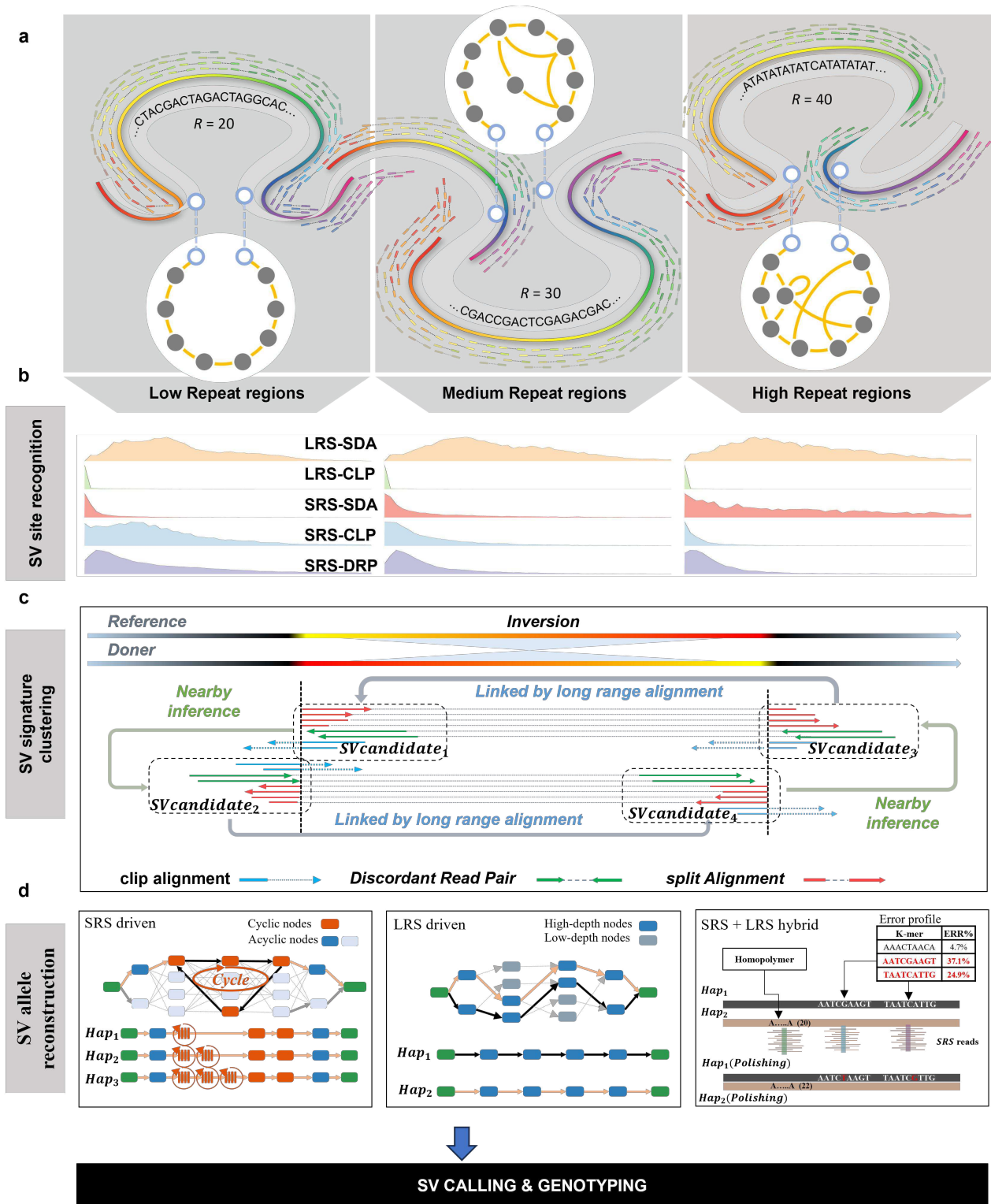


Fig. 1 | A schematic illustration of HitSV. a, Schematic of HitSV, an alignment-based SV calling approach for both LRS and SRS data. **b**, Repetitiveness and signature density-aware SV site recognition. HitSV learns the distributions of all types of alignment signals, including sequence divergences in alignments (SDA, for LRS and SRS), large clippings (CLP, for LRS and SRS) and discordant read pairs (DRP, only for SR), within SV spanning reads under various degrees of local repetitiveness. **c**, Progressive SV signature clustering. HitSV constructs probabilistic models to infer virtual SV breakpoints using alignments signals of the clustered reads, and then

connects the clusters around various breakpoints in contexts of genomic positions or long- range information of read pairs or split alignments. Inversion is shown as an example. **d**, Haplotype-resolved local assembly. HitSV performs haplotype-resolved local assembly using a unitig-based de Bruijn graph integrating variant and non-variant reads. For short reads, coverage-guided traversal generates contigs beyond read-length limits. For long reads, candidate contigs are derived by clustering unitig-DBG paths and iteratively refined via SDP-based realignment. For hybrid data, long-read contigs serve as backbones and are polished with short-read consensus.

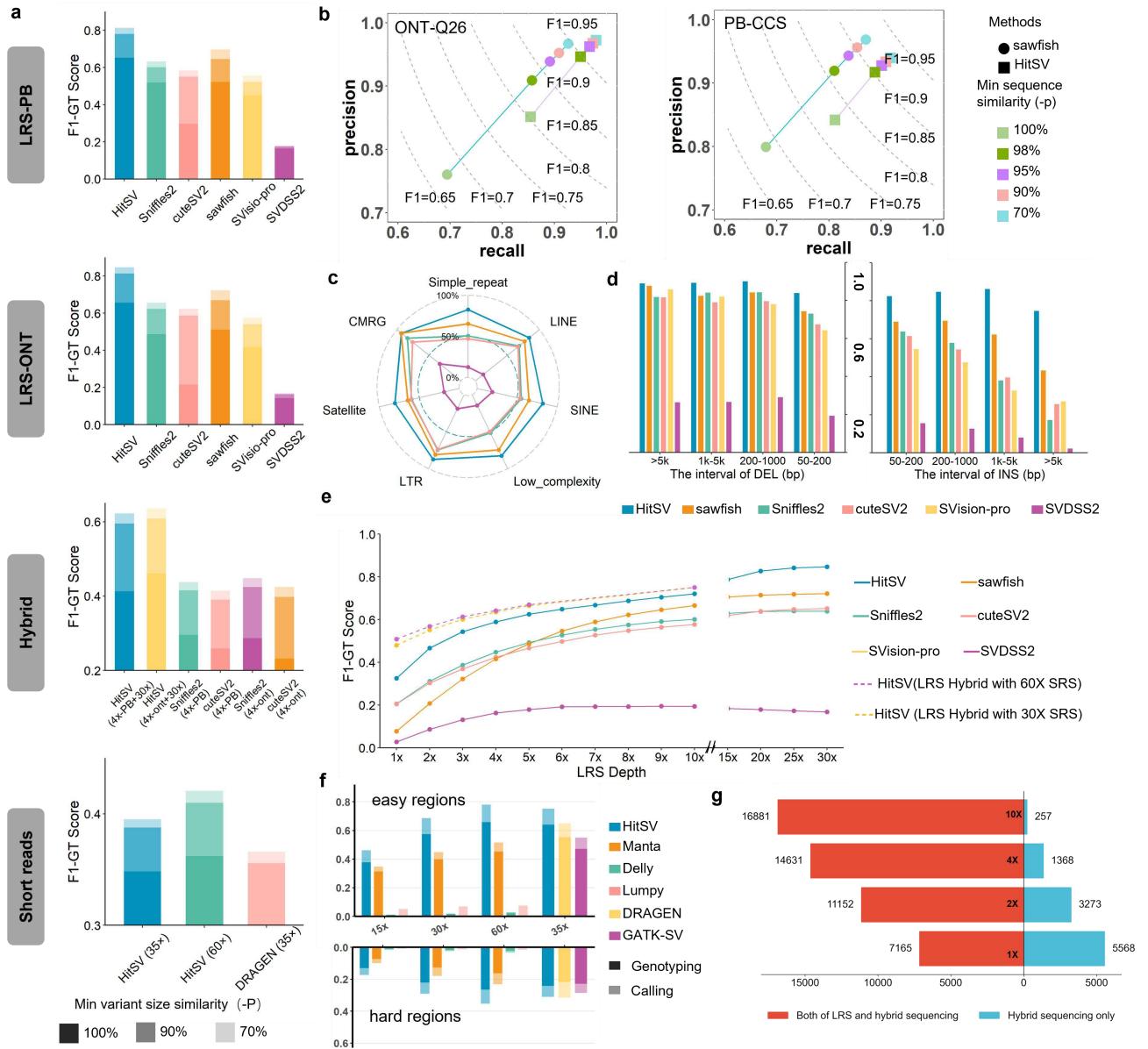


Fig. 2 | Overall performance evaluation of HitSV on GIAB HG002 sample. a, The benchmark evaluations of various SV callers on F1-GT scores for long read (PacBio and ONT), hybrid and short read sequencing data, respectively. The dark, medium and light bars indicate different configurations for the -P parameter of Truvari (-P 1, 0.9 and 0.7 for dark, medium and light bars, respectively). **b**, Base-level precision of SV reconstruction of HitSV and Sawfish by using the 'refine' mode of Truvari with different stringencies (-P, 0.7, 0.9, 0.95, 0.98, 1) and read types (PacBio and ONT reads). **c**, The GT-F1 scores of the callers in various kinds of genomic regions (according to the annotations of RepeatMasker and CMRG). **d**, The long read-based F1-GT scores of various SV callers for various types and sizes of SVs. Deletions (DEL, left) and insertions (INS, right) are binned by SV length as indicated on the x-axis. **e**, The F1-GT scores of long read-based callers for various read coverages (1-30x). The dashed lines indicate the hybrid module of HitSV with corresponding long reads plus various amounts of short reads (pink: 60x, yellow: 30x). **f**, The yields of the callers (HitSV, Manta, Delly, Lumpy, DRAGEN and GATK-SV) for various (hard, easy) genomic regions using various coverages of short reads (15x, 30x, 60x, 35x). The "hard" and "easy" regions are regions that hard and easy for short reads to be aligned, respectively, as designated in previous studies⁵⁰. **g**, The comparison of HitSV calls using hybrid sequencing data and long reads only. The SV calls of various coverages (1x, 2x, 4x and 10x) of long reads with and without 60x short reads are compared in

pairs. The red (left) and cyan (right) bars indicate numbers of true positive calls produced by both data configurations and by hybrid sequencing data only, respectively.

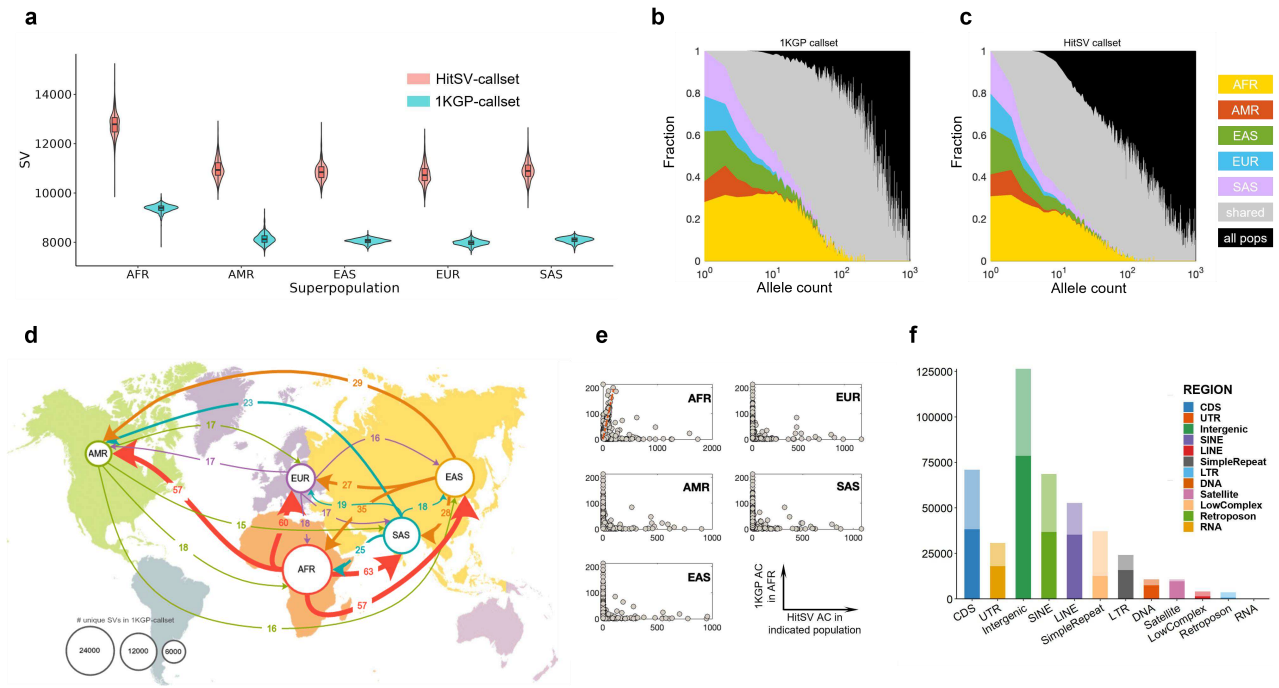


Fig. 3 | The re-analysis of 1KGP datasets by HitSV. **a**, Distribution of population-wise per-sample SV numbers in HitSV-callset and 1KGP-callset. The abbreviations for superpopulations are: AFR – African, AMR – American, EAS – East Asian, EUR – European, SAS – South Asian. **b&c**, Distributions of population unique and shared SV alleles. Shared represents alleles shared by at least two but not all superpopulations in 1KGP-callset (b) and HitSV-callset (c). **d**, Shifts in SV allele distributions between superpopulations. Arrows indicate that alleles deemed unique to the source superpopulation by 1KGP-callset are regarded common (frequency > 0.05) in the destination superpopulation by HitSV-callset. The circle size scales with the number of alleles unique to each superpopulation in HitSV-callset. The arrow width and the number marked on each arrow indicate the number of alleles re-discovered as common in the destination superpopulation. **e**, Allele counts of AFR unique SVs in 1KGP-callset vs. those of the same SVs in HitSV-callset in the same superpopulation or in another superpopulation, as indicated in each panel. Red dashed line in the first panel shows 1:1 ratio. **f**, Functional/genome region annotation of SVs in the 1KGP-callset (dark colors) and SVs distinctly identified in HitSV-callset (light colors).

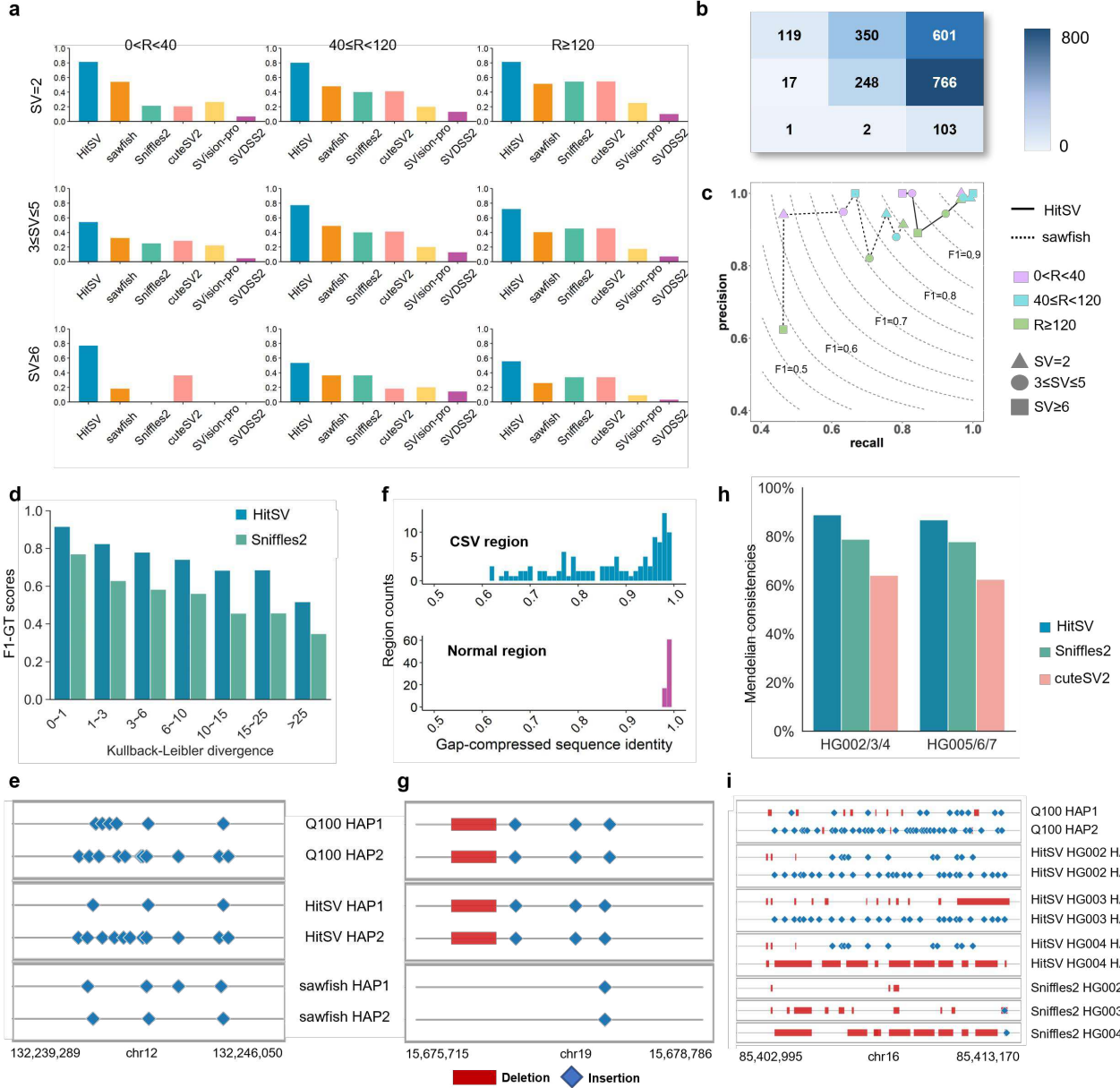


Fig. 4 | The ability of HitSV to solve complex SVs on ONT dataset. **a**, The event-level F1-GT scores of various SV callers under various scenarios and difficulties. the regions are further categorized by their complexity (represented by numbers of containing SV events, 3 stages: 2, 3-5, ≥ 6) and local reference repetitiveness (represented by R values, 3 stages: 1-39, 40-119, ≥ 120). **b**, The counts of CSV events under various scenarios and difficulties. **c**, Precision comparison of SV-related sequence between HitSV and Sawfish with 95% sequence similarity (by Truvari refine, -P, 0.95) under various scenarios and difficulties. **d**, Comparison on various Kullback-Leibler divergence (KLD) ranges between HitSV and Sniffles. The KLD were calculated by the distributions of SV breakpoints and ground truth to measures the heterogeneity of local read alignments, which used to re-categorize the HG002 CSV regions by various KLD ranges. **e**, An example of CSV detection in an intronic region of *GALNT9* on chromosome 12 (chr12:132,239,289–132,246,050 bp), HitSV shows high concordance with the ground truth in both the genomic distribution and length of detected variants. **f**, Distributions of gap-compressed sequence identities of the CSV and non-CSV regions. The identities measured the occurrence of a series of small sequence divergences such as mismatches and indels instead of whole large events, which were caused by scoring bias of read alignments and the similarity of flanking references. **g**, An example of CSV

detection on a local genomic region encompassing the CYP4F12 locus on chromosome 19 (chr19:15,677,010-15,678,477 bp), HitSV accurately resolves SVs with near-complete concordance in total variant length. **h**, Mendelian consistencies of HitSV, Sniffles2 and cuteSV2 on two GIAB trio-samples (HG002/3/4 and HG005/6/7) LRS datasets. **i**, A typical example consisting of an INS-rich paternal and an INS-DEL mixed maternal allele in the *GSE1* gene, HitSV accurately resolving complex SV structures and their transmission patterns in pedigree-based analyses.

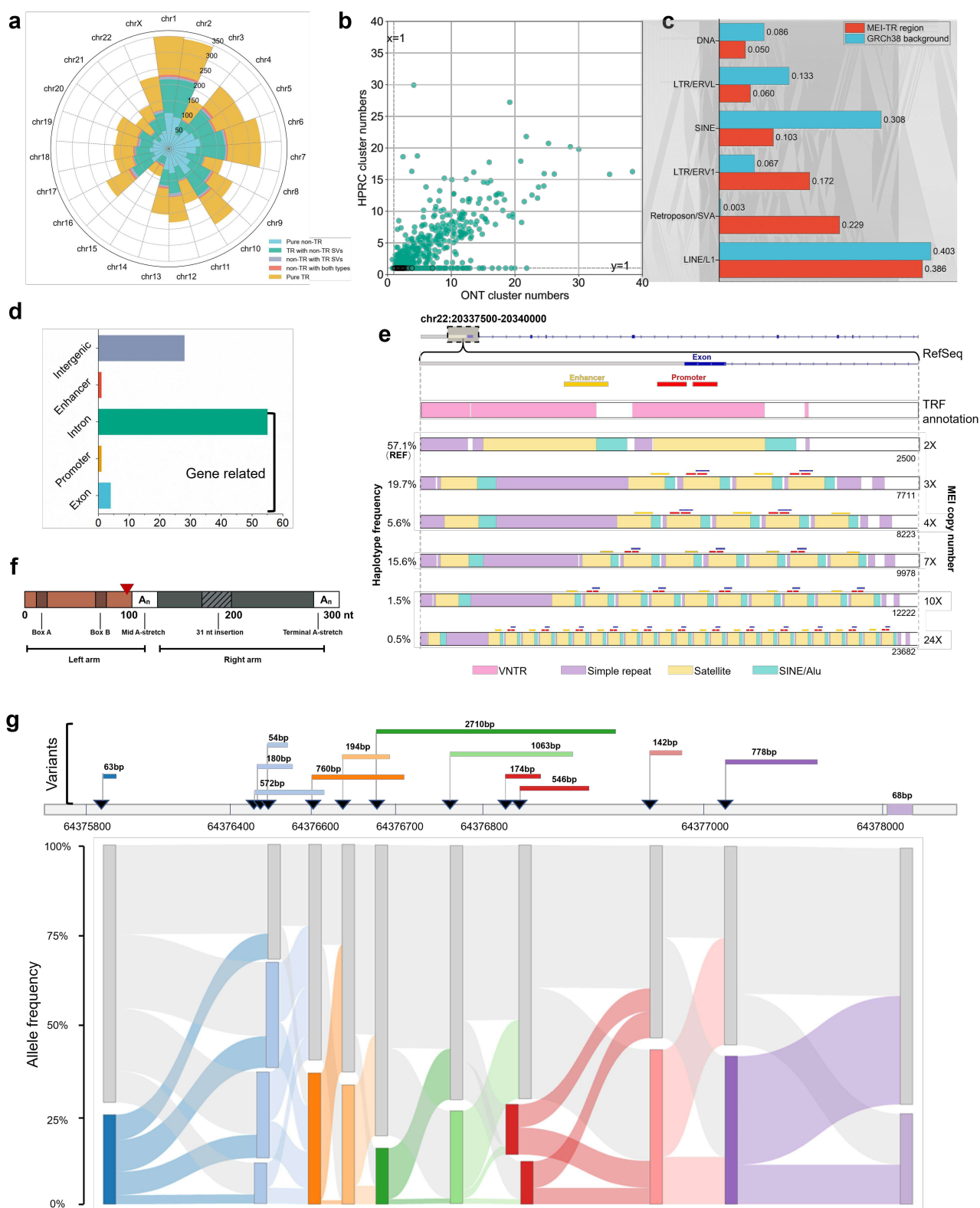


Fig. 5 | Comprehensive analysis of CSVs in 557 Chinese samples. a, Chromosome-level classification of CSV regions across 557 Chinese samples' ONT datasets. CSV regions are classified based on the genomic context and composition of detected insertions, distinguishing tandem repeat (TR)-derived events from non-TR structural variants. **b**, Scatter plots compare the number of clusters derived from ONT data (x-axis) with the number of clusters inferred from HPRC data (y-axis) for TR with non-TR SVs CSV regions. Each point represents one CSV

region. Dashed lines indicate the reference threshold of a single cluster ($x = 1$ or $y = 1$). **c**, The relative length contributions of major transposable element (TE) families within TR-MEI regions and across the whole genome based on the GRCh38 reference background. Bars indicate the proportion of total TE length contributed by each TE family, including LINE/L1, SINE, LTR/ERV1, LTR/ERVL, DNA transposons, and retroposon/SVA elements. **d**, The distribution of TR-MEI intervals across different genomic annotation categories, including intergenic, intronic, exonic, promoter, and enhancer regions. **e**, An example of haplotype-resolved structural variation at a TR-MEI-associated locus on FAM230G gene of human chromosome 22 (chr22:20,337,500–20,340,000 bp) associated transcriptional regulation. The upper panel shows the genomic context based on RefSeq annotation, highlighting the position of the locus relative to gene structure, including exon, promoter, and enhancer elements. Tandem repeat finder (TRF) annotations are shown below, indicating repeat-enriched regions overlapping the TR-MEI array. The lower panels display representative haplotypes stratified by allele frequency across the ONT cohort, with each track corresponding to a distinct MEI copy-number state ($2\times$ to $24\times$). Colored segments denote different repeat classes, including TR, simple repeats, satellite repeats, and SINE/Alu elements. Percentages on the left indicate haplotype frequencies in the population, while numbers on the right denote total MEI copy numbers. **f**, The canonical structure of the AluY mobile element. The red triangle marks the position on the AluY consensus sequence corresponding to the most frequent midpoint (97.3%) observed among repeated AluY copies within TR-MEI regions. **g**, A Sankey-style diagram summarizing the haplotype-level distribution of structural variants within a CSV region on chromosome 2 (chr2:64,375,000–64,377,000 bp) across 557 Chinese samples. A total of 13 variants are observed in this region, including 12 INS and 1 DEL. The top track indicates the genomic positions and lengths of individual variants. The leftmost column represents the allele frequency of each variant across the cohort, while flows illustrate how variants co-occur and segregate across different haplotypes. Gray blocks and flows denote the wild-type state.

Supplementary Files

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

- [HitSVSupplementary.pdf](#)