BRIEF COMMUNICATION
https://doi.org/10.1038/s41592-022-01609-w

nature methods

SVision: a deep learning approach to resolve complex structural variants

Jiadong Lin[1,2,3,4,13], Songbo Wang[1,2,3,13], Peter A. Audano[5], Deyu Meng[1,6,7], Jacob I. Flores[5], Walter Kosters[4], Xiaofei Yang[1,8], Peng Jia[1,2,3], Tobias Marschall[9], Christine R. Beck[5,10] and Kai Ye[1,2,3,11,12] ✉

Complex structural variants (CSVs) encompass multiple breakpoints and are often missed or misinterpreted. We developed SVision, a deep-learning-based multi-object-recognition framework, to automatically detect and characterize CSVs from long-read sequencing data. SVision outperforms current callers at identifying the internal structure of complex events and has revealed 80 high-quality CSVs with 25 distinct structures from an individual genome. SVision directly detects CSVs without matching known structures, allowing sensitive detection of both common and previously uncharacterized complex rearrangements.

CSVs contain multiple breakpoints and may delete, duplicate, and/or invert multiple segments in both healthy[1] and diseased genomes[2,3], creating events that are more likely to be deleterious than simple structural variants (SVs)[4,5]. Previous short-read-sequencing-based studies detected CSVs through intensive breakpoint analysis and subsequent manual inspection was required to determine CSV internal structures[1,6], hindering large-scale study of CSVs. Although long-read sequencing has greatly facilitated the detection of phased SVs[7] and somatic SVs[4,8], three major issues have impeded their use in CSV detection. First, the model-based inference approach, initially designed for simple SV discovery from short reads[9], requires construction of each SV model for fitting aberrant alignment patterns (Extended Data Fig. 1), which is not applicable to largely unexplored CSV structures[10,11]. Second, ambiguous alignments at repetitive regions lead to false calls or missing events[6]. Last, the current definition of different CSV types is based on predefined models lacking a unified and computer-interpretable framework, thereby limiting cross-study comparison.

We developed an automated CSV detection and characterization method: SVision. It introduces a sequence-to-image coding schema, adapting variant detection to a problem that is amenable to deep-learning frameworks. SVision contains three core components: an encoder that represents the differences and similarities between a variant-supporting read and its corresponding segment on the reference genome as a denoised image, a targeted multi-object recognition (tMOR) framework that detects and characterizes CSVs through a convolutional neural network (CNN) in the denoised image, and an illustrator that creates and unifies each detected CSV as a graph representation from the denoised image (Fig. 1a and Methods). Specifically, the encoder first collects aberrant long-read alignments, the 'variant feature sequence' (VAR), and its aligned segment on the reference genome, referred to as the reference sequence (REF). For a VAR and REF pair, the encoder identifies matched and unmatched bases to create VAR-to-REF and REF-to-REF images (Fig. 1b). Because repetitive sequences might be present in both VAR and REF, the variant signature can be isolated and accentuated when the background noise is removed. Accordingly, a denoised image is created for each VAR by subtracting REF-to-REF image from its corresponding VAR-to-REF image. In the tMOR step, because a denoised image might contain more than one SV, SVision uses a two-step image-segmentation process to first obtain one-variant images, containing the full structure of an event. Then, SVision defines each location surrounding a breakpoint in the one-variant image as a segment of interest (SOI), and SOIs collected from a one-variant image are recognized as a single CSV through a pre-trained CNN. For all one-variant images supporting one event, SVision integrates the CNN prediction probability of each one-variant image and the similarity of one-variant images to measure the quality of a discovery. Finally, the illustrator adopts a graph-based approach to depict different CSV structures in graphical fragment assembly (GFA) format. Moreover, a given CSV graph structure and its topologically equivalent events are combined through detection of isomorphic graphs.

We examined how the sequence-to-image coding and the CNN model perform across HiFi (high-fidelity reads produced by PacBio circular consensus sequencing) and ONT (reads produced by Oxford Nanopore Technology) data for canonical SV detection, by benchmarking SVision, cuteSV[12], pbsv, SVIM[13], and Sniffles[10] against the HG002 genome[14] (Methods). SVision outperforms other callers at all different coverages (Supplementary Table 1): the *F* score of SVision ranged from 0.83 to 0.90 for HiFi and from 0.76 to 0.92 for ONT (Extended Data Fig. 2). Furthermore, their performances were assessed on a genome harboring 3,000 simulated CSVs of 10 types (Extended Data Fig. 3a–c, Supplementary Table 2, and Methods). Similar to the evaluation metric employed by

[1]MOE Key Lab for Intelligent Networks & Networks Security, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, China. [2]School of Automation Science and Engineering, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, China. [3]Genome Institute, the First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, China. [4]Leiden Institute of Advanced Computer Science, Faculty of Science, Leiden University, Leiden, the Netherlands. [5]The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA. [6]School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, China. [7]Macau Institute of Systems Engineering, Macau University of Science and Technology, Taipa, Macau. [8]School of Computer Science and Technology, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, China. [9]Heinrich Heine University, Medical Faculty, Institute for Medical Biometry and Bioinformatics, Dusseldorf, Germany. [10]Department of Genetics and Genome Sciences, Institute for Systems Genomics, University of Connecticut Health Center, Farmington, CT, USA. [11]The School of Life Science and Technology, Xi'an Jiaotong University, Xi'an, China. [12]Faculty of Science, Leiden University, Leident, the Netherlands. [13]These authors contributed equally: Jiadong Lin, Songbo Wang ✉e-mail: kaiye@xjtu.edu.cn

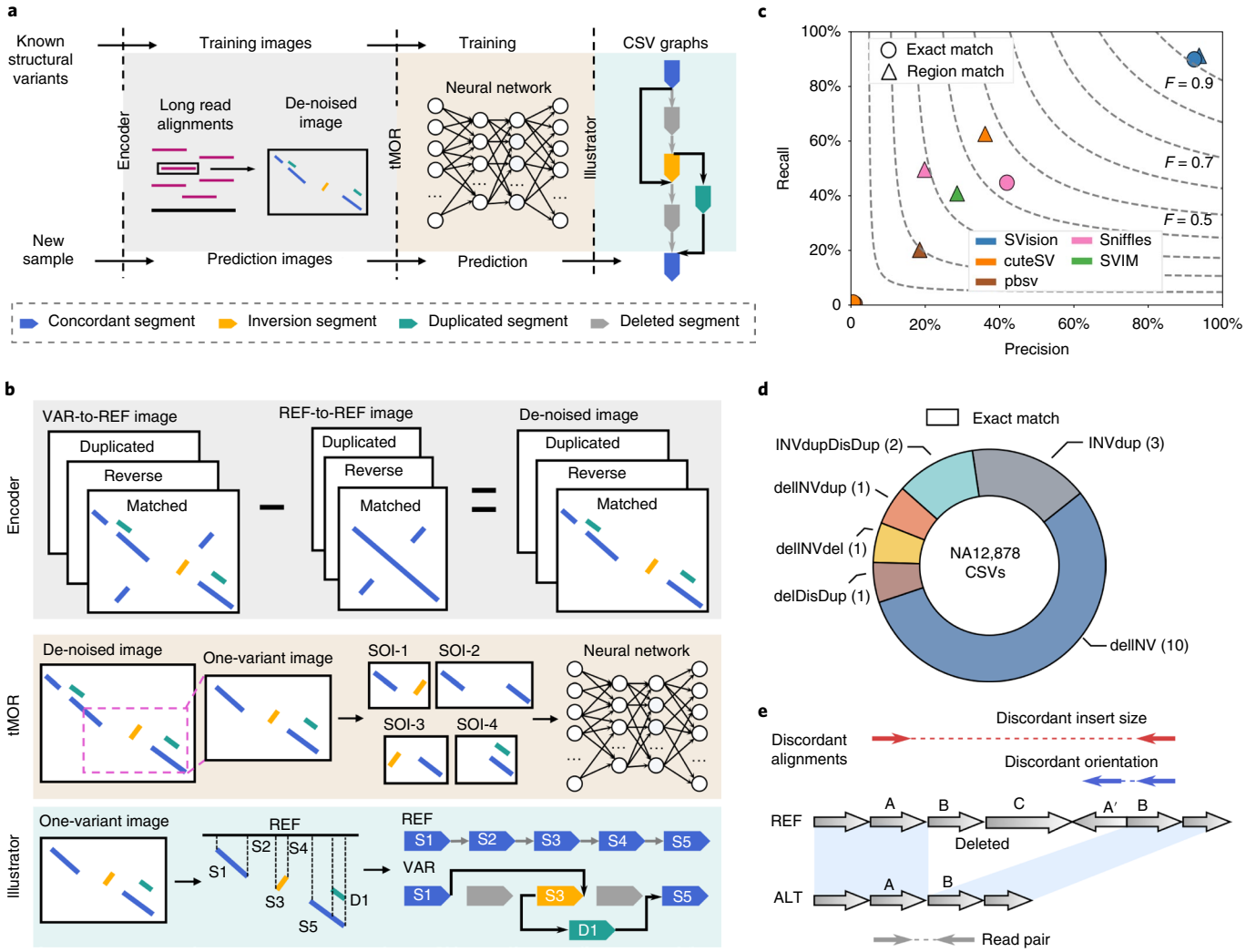


Fig. 1 | Workflow and evaluation of SVision's detection of CSVs. a, Overview of three modules in SVision. **b**, Workflow for each module. **c**, Performance for calling simulated CSVs was evaluated by recall (*y* axis), precision (*x* axis) and *F* score (*F*, dashed line). **d**, SVision detected all previously reported and manually curated NA12878 CSVs, all of which passed exact-match criteria. **e**, A misinterpreted complex event from short-read data, which is correctly detected by SVision as simple deletion. This event is surrounded by repeats that introduce two distinct patterns of discordant paired-end mapping, misleading short-read callers to report a CSV.

Sniffles[10], we introduced region-match (that is, correct detection of a CSV site) and exact-match (that is, correct detection of a CSV site and its subcomponents) for performance evaluation (Methods). For region-match, the recall and the precision of SVision were 91% and 93%; those for the second-best tool, cuteSV, were 62% and 36%, respectively (Fig. 1c and Supplementary Table 3). The low recall and precision of cuteSV and others could be largely attributed to partial CSV detection because the observed signatures were beyond existing models (Extended Data Fig. 3d and Supplementary Table 4). For exact-match, SVision detected 89% of the CSVs, more than double the percentage detected by Sniffles, and other callers were not able to characterize any CSVs (Fig. 1c and Supplementary Table 3). Additionally, we manually curated 62 complex deletion and 251 complex inversion sites in NA12878 reported by a short-read study[1] (Methods). As a result, 18 CSVs of six unique structures, including one unclassified novel structure, were verified; the remaining events were either simple SVs (64) or false discoveries (231) (Supplementary Table 5 and Supplementary File 1). SVision automatically and correctly characterized the internal structure of all 18 CSVs (Fig. 1d and Supplementary Table 6), including two CSVs that were unclassified by short-read[1], that is, a deletion replaced by an inverted segment and a duplicated segment, and a novel complex insertion structure consisting of an inverted-duplication and two dispersed-duplications (Extended Data Fig. 4). Moreover, SVision resolved a simple deletion at a region flanked by duplicates (inverted and dispersed), which was mistakenly reported as a CSV in the short-read study[1] (Fig. 1e, Extended Data Fig. 5, and Supplementary Table 5). The above results suggest that SVision can detect both simple and complex SVs with high sensitivity and specificity.

To explore novel CSV loci and structures, we applied SVision to the HG00733 genome[7], in which CSVs were not well characterized. SVision detected 80 high-quality CSVs of 25 unique structures, 20 of which were novel, accounting for half of the events, and the remaining five CSV graph structures matched frequently reported CSV types[1,2] (Extended Data Fig. 6, Supplementary Table 7, and Supplementary File 2). We then applied both computational and experimental approaches to validate the structure and breakpoint junctions of those 80 CSVs (Methods). Firstly, GraphAligner[15] was used to assess the internal structure and breakpoints of CSVs by aligning ONT reads[16] to CSV graphs. The graph

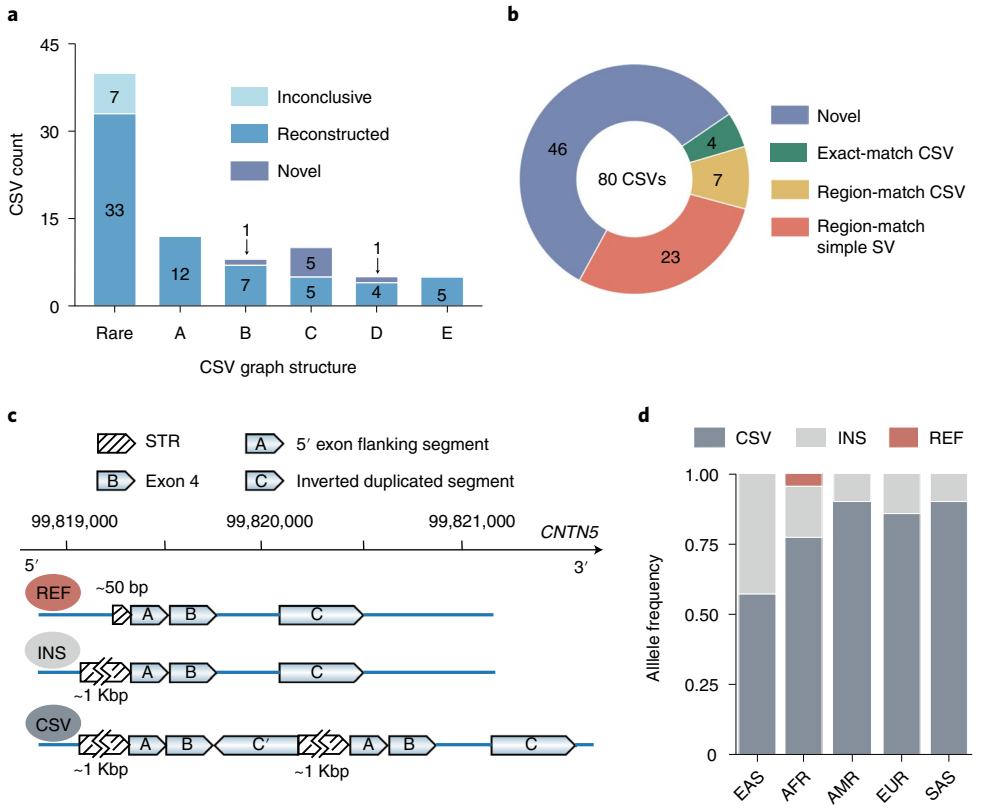


Fig. 2 | Application of SVision on HG00733 HiFi data. a, The SVision CSVs overlapping with PAV calls and reconstructed with phased HiFi contigs. The rare type represented multiple CSV graph structures, each of which contains fewer than five complex events. A: inverted duplication. B: deletion associated with inversion. C: deletion associated with inverted duplication. D: multiple deletion with spacer. E: deletion associated with duplication. Inconclusive: unable to characterize visually. Reconstructed: SVision CSV structure validated through contig-based manual curation. Novel: no overlapping PAV call. **b**, Compared with 80 CSVs in HG00733 detected by SVision, the short-read callset correctly reported the full structure of four CSVs (exact-match), partially called seven (region-match), misinterpreted 23 events as simple (region-match), and completely missed 46 CSVs. **c**, Three distinct alleles were found for the *CNTN5* locus, including a contracted tandem repeat (REF), an expanded tandem repeat (INS), and an expanded tandem repeat with a complex duplication containing *CNTN5* exon 4. **d**, The frequencies of three *CNTN5* alleles differ among populations. EAS, East Asians; AFR, Africans; AMR, Americans; EUR, Europeans; SAS, South Asians.

alignments showed that single reads cover the entire paths of 79 CSV graphs, and one CSV graph path was covered by two different reads (Supplementary Table 8). Secondly, 73 CSVs overlapped phased assembly variant (PAV)[7] calls, and 90% of them were successfully reconstructed with haplotype contigs (Fig. 2a), while others were challenging to characterize visually but were verified via GraphAligner (Supplementary Table 8 and Supplementary File 3). Furthermore, 20 CSVs were selected for manual curation and experimental validation after excluding SVs in highly repetitive regions. Of these 20 CSVs, manual inspection confirmed that 18 events matched SVision's reports, and two loci contained expansions of short tandem repeats that were collapsed in the reference (Supplementary Table 8). As for the experiment, eight CSVs failed PCR owing to repetitive sequence or high GC content, and the other 12 events were successfully confirmed by PCR and Sanger sequencing (Supplementary Table 8). The above validations indicate that SVision can detect and characterize CSVs reliably from long-read data. Compared with 80 CSVs detected by SVision, short-read misinterpreted 23 as simple events and completely missed 46 CSVs, while four and seven were fully and partially interpreted, respectively (Fig. 2b and Supplementary Table 9).

Of the 80 HG00733 CSVs detected by SVision, 19 overlapped 18 different genes (Supplementary Table 7). A complex duplication in *CNTN5*, an important neural-development gene, is composed of a direct duplication of *CNTN5* exon 4 and an inverted intronic duplication, both of which inserted in tandem within a *CNTN5* intronic tandem repeat proximal to exon 4 (Fig. 2c and Extended Data Fig. 7). This event was missed by short-read data, and PAV called only a simple insertion, leaving the duplicated exon unannotated[7]. Using contigs of 35 samples from Human Genome Structural Variant Consortium and the SVision reported structure, three distinct alleles were noted for this site, a contracted tandem repeat, an expanded tandem repeat, and an expanded tandem repeat containing the complex exon duplication (Fig. 2d, Supplementary Table 10, and Supplementary File 4). We observed the duplicated exon signature in the RNA-seq data for the human primary visual cortex and precuneus[17] (Extended Data Fig. 8, Supplementary Table 11, and Supplementary File 5). Additionally, SVision identified an insertion-inversion-insertion event, which was detected as a 1,737-bp insertion by PAV but missed in previous studies[16,18] (Extended Data Fig. 9a). This event was also re-genotyped by PanGenie[19], and it has 80% allele frequency among 2,504 unrelated samples in 1000 Genomes Project cohort[7]. The inserted sequence of this CSV was further identified in chimpanzee and gorilla genomes (Extended Data Fig. 9b), indicating the insertion state was ancestral and the reference was derived through deletion and inversion.

Long-read sequencing and associated tools have revolutionized SV detection[7,11], but it is still hard to correctly characterize multi-breakpoint events, leaving CSVs either uncalled or misinterpreted as simple SVs. Inspired by existing deep-learning-based

variant-detection methods[20,21], SVision fills this gap by applying a multi-object recognition framework to denoised images to detect both simple and complex SVs, and autonomously identifies their structures. Note that CSV structure resolution depends on knowledge-oriented image denoise and segmentation, which need further optimization with unsupervised approaches. SVision is a valuable tool to facilitate the study of complicated and novel CSVs, paving the way for the analysis of complex genomic events at the population scale.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-022-01609-w.

Received: 18 January 2022; Accepted: 11 August 2022;
Published online: 16 September 2022

References

1. Sudmant, P. H. et al. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
2. Collins, R. L. et al. Defining the diverse spectrum of inversions, complex structural variation, and chromothripsis in the morbid human genome. *Genome Biol.* **18**, 36 (2017).
3. Li, Y. et al. Patterns of somatic structural variation in human cancer genomes. *Nature* **578**, 112–121 (2020).
4. Fujimoto, A. et al. Whole-genome sequencing with long reads reveals complex structure and origin of structural variation in human genetic variations and somatic mutations in cancer. *Genome Med.* **13**, 65 (2021).
5. Baca, S. C. et al. Punctuated evolution of prostate cancer genomes. *Cell* **153**, 666–677 (2013).
6. Quinlan, A. R. & Hall, I. M. Characterizing complex structural variation in germline and somatic genomes. *Trends Genet.* **28**, 43–53 (2012).
7. Ebert, P. et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, eabf7117 (2021).
8. Aganezov, S. et al. Comprehensive analysis of structural variants in breast cancer genomes using single-molecule sequencing. *Genome Res.* **30**, 1258–1273 (2020).
9. Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* **12**, 363–376 (2011).
10. Sedlazeck, F. J. et al. Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).
11. Ho, S. S., Urban, A. E. & Mills, R. E. Structural variation in the sequencing era. *Nat. Rev. Genet.* **21**, 171–189 (2020).
12. Jiang, T. et al. Long-read-based human genomic structural variation detection with cuteSV. *Genome Biol.* **21**, 189 (2020).
13. Heller, D. & Vingron, M. SVIM: structural variant identification using mapped long reads. *Bioinformatics* **35**, 2907–2915 (2019).
14. Zook, J. M. et al. A robust benchmark for detection of germline large deletions and insertions. *Nat. Biotechnol.* **38**, 1347–1355 (2020).
15. Rautiainen, M. & Marschall, T. GraphAligner: rapid and versatile sequence-to-graph alignment. *Genome Biol.* **21**, 253 (2020).
16. Chaisson, M. J. P. et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* **10**, 1784 (2019).
17. Guennewig, B. et al. Defining early changes in Alzheimer's disease from RNA sequencing of brain regions differentially affected by pathology. *Sci. Rep.* **11**, 4865 (2021).
18. Audano, P. A. et al. Characterizing the major structural variant alleles of the human genome. *Cell* **176**, 663–675(2019).
19. Ebler, J. et al. Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Nat. Genet.* **54**, 518–525 (2022).
20. Poplin, R. et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983–987 (2018).
21. Cai, L., Wu, Y. & Gao, J. DeepSV: accurate calling of genomic deletions from high-throughput sequencing data using deep convolutional neural network. *BMC Bioinf.* **20**, 665 (2019).

Methods

Evaluating simple structural variants detection with HG002. To benchmark SV callers on HG002, we followed the procedure introduced by Genome-In-A-Bottle (GIAB) and detailed steps adopted by cuteSV. Briefly, the high-confidence insertion and deletion calls and high-confidence regions published by the GIAB consortium were used as the ground truth, and the genotype accuracy was not considered in our evaluation. The HiFi reads were aligned to reference hg19 by pbmm2 (https://github.com/PacificBiosciences/pbmm2, v1.4.0) with parameter '–preset CCS', and ONT reads were aligned with pbmm2 default settings. The 5× and 10× coverage of HiFi and ONT data were further obtained with the SAMtools[22] '-s' option. Sniffles (v1.0.12), cuteSV (v1.0.10), pbsv (v2.2.2), SVision (v1.3.6), and SVIM (v1.4.0) were applied to the pbmm2 aligned file with default parameters. The minimum supporting read was two and three for 5× and 10× data, and 10 was used for the original coverage.

Simulating complex structural variants. Ten CSV types were simulated, according to frequently reported types introduced by 1000 Genomes Project (1KGP)[1] and a cohort study of autism spectrum disorder[2] (Supplementary Note). A CSV was essentially a combination of breakpoints from simple structural variants (SSVs). Therefore, a four-step simulation process was developed as follows. VISOR[23] was first used to simulate and to randomly implant five SSV types (that is, deletion, inverted-dispersed-duplication, inverted-tandem-duplication, tandem-duplication, and dispersed-duplication) on reference genome GRCh38. Second, we followed the procedure introduced by SURVIVOR[24] to simulate CSVs, where SSVs of the above five types were randomly added adjacent to the existing SSVs on the genome. In particular, 3,000 SSVs of the five types were created by VISOR with parameters '-n 3000 –r 20:20:20:20:20 –l 500 –s 150'. Third, we added extra SSVs required in predefined CSV structures to existing SSVs by following the order of types, that is, deletion, inverted-dispersed duplication, inverted-tandem duplication, tandem-duplication and dispersed-duplication. For instance, we first used implanted deletions as seeds to create all CSV instances that involved deletions, and then turned to instances of the next type. Finally, the variation genome with 3,000 CSVs was used as input for the VISOR LASoR module to simulate 30× HiFi reads for subsequent alignment by ngmlr[10] (v0.2.7) with the default setting. Note that VISOR was used only to simulate variants at one haplotype.

Evaluating detection of simulated complex structural variants. To examine the correctness of detected CSVs, we used closeness and size similarity to assess whether two events are identical, according to Truvari (https://github.com/spiralgenetics/truvari/), developed by GIAB (Supplementary Note). The closeness, *bpDist*, and size similarity, *sim*, between prediction and benchmark were 500 bp and 0.7, respectively. For example, assume a benchmark CSV (start at *b.start*, end at *b.end* and the size is*b.size*), and a prediction (start at *p.start*, end at *p.end* and the size is *p.size*); then, a correct region-match should satisfy the following equations:

$$\max\left(|b.start - p.start|, |b.end - p.end|\right) \leq bpDist$$

$$b.size \times sim \leq p.size \leq b.size \times (2 - sim)$$

Comparably, the exact-match not only required region-match, but also required the correct detection of all CSV subcomponents, including the subcomponent breakpoint type. Therefore, for a deletion-inversion that contained two subcomponents, that is, inversion and deletion, the exact-match became a three-step evaluation:

1. Region-match between a predicted CSV and a benchmark deletion-inversion event.
2. For each subcomponent, we examined the breakpoint closeness and event size, as well as the correctness of detected type.
3. The correct exact-match detection should pass conditions (1) and (2).

Currently, we considered only insertion, deletion, duplication, and inversion as subcomponent types. Any called CSVs without a matched prediction were counted as false negatives. On the basis of the numbers of true positives (*TP*) and false negatives (*FN*), we computed the recall, precision, and *F* score with the following equations:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F - score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Each caller was run with a different number of variant-supporting reads (that is, 1, 3, 5, and 10), and the performance of simulated-CSV detection was assessed accordingly (Supplementary Note).

Examining complex structural variant detection in NA12878. The published NA12878 CSV set was obtained from Supplementary Tables 12 and 15 of a study conducted by the 1KGP[1], containing 62 deletion- and 251 inversion-associated CSV sites in hg19 coordinates. We aligned the HiFi reads of NA12878 released by Human Genome Structural Variants Consortium (HGSVC)[7] using ngmlr (v0.2.7) with the default setting for manual inspection and CSV detection. For manual curation, SAMtools was used to extract HiFi reads spanning the CSV loci, and Gepard[25] was used to create the Dotplots between HiFi reads and their corresponding reference sequences. We then manually inspected all Dotplots associated with a reported CVS locus (Supplementary File 1). SVision was run with default parameters on the ngmlr aligned file for CSV detection. Then, we compared SVision's discoveries with the curated CSV loci and examined whether the internal structures matched that reported by SVision.

Three-channel coding of feature sequence. The encoder consisted of two major steps, that is, variant feature sequence selection and sequence coding (Supplementary Note). Variant feature sequences (VAR) are directly identified from long-read aberrant alignments containing SV signatures, such as inter-read and intra-read alignments. Intra-read alignments are derived from reads spanning the entire SV locus, whereas inter-read alignments are obtained from reads that are aligned to larger SV event, resulting in supplementary alignments. SVision identifies additional SV signatures by applying a *k*-mer-based realignment approach for an unmapped segment in VAR, such as 'I's from CIGAR string and gap sequence obtained from inter-read alignments. Then, matched and unmatched segments between VAR and its mapped segment on the reference genome (that is, REF) are coded as an image. The image contains three channels, including a blue channel (0, 0, 255), a green channel (0, 255, 0), and a red channel (255, 0, 0), to code the matched, the duplicated, and the inverted segments, respectively.

To efficiently implement three-channel image coding, matched segments obtained from CIGAR string and supplementary alignments, originating from aligner's outputs, are directly used for VAR-to-REF image coding, and realignment results are further added to complete image coding. The REF-to-REF image is created by *k*-mer-based realignment. The denoised image is obtained by subtracting the REF-to-REF image from the VAR-to-REF image. Because the repetitive background noise originates from REF, the encoder subtracts the segments of two images on the basis of the REF sequence coordinates. Specifically, if segments from two images overlap on the reference dimension and their difference is larger than 50 bp (minimum SV report size), the encoder keeps the non-overlapping part of the segment in the similarity image, where its coordinates are determined by VAR-to-REF image (Supplementary Note).

Detecting complex structural variants from denoised images using targeted multi-object recognition. In principle, for each denoised image, the regions where VAR and REF are identical must be a straight line, whereas SVs introduce discontinuous segments. These discontinuous segments surrounding SV breakpoints are considered as a breakpoint object and further defined as SOI. Since long reads are likely to span more than one SVs in the denoised image, the tMOR contains a two-step image-segmentation process for further SV recognition (Supplementary Note). Firstly, the tMOR obtains a one-variant image from the denoised image, on the basis of the following steps.

1. Sorting and tagging. We sort all segments in the denoised image by their positions on the read in ascending order. Then, the major segment is defined according to the matched segments derived from CIGAR operations, and the minor segment should meet one of the following conditions:
   Condition 1: the segment is derived from the *k*-mer-based realignment.
   Condition 2: the segment is inverted compared with the reference genome.
   Condition 3: the segment is totally covered by another one.
2. Creating the one-variant image. SVision partitions the denoised image into several one-variant images through sequential combination of the major segments. Specifically, each major segment and its neighboring major segment along with the minor segments (if they exist) between them are used to create a one-variant image.

Afterwards, SVision clusters similar one-variant images by measuring the distance of segment signatures between one-variant images. Thus, one-variant images in a cluster support the same variant, and the size of a cluster is termed as the number of variant supporting image. Secondly, SVision collects SOIs from each one-variant image. Unlike traditional multi-object recognition that uses complex algorithms to select regions of interest, the discontinuous segment signatures in the one-variant image enable efficient SOI identification by sequentially combining both major and minor segments. Then, SOIs are used as input for CNN prediction, and the interpreted SV types are given by the labels involved in the training set, including deletion (DEL), inversion (INV), insertion (INS), duplication (DUP), and tandem-duplication (tDUP). Finally, the CNN assigns the probability score to assess the existence of the five SV classes in the one-variant image (Supplementary Note).

Creating complex structural variant graphs from one-variant images. SVision uses a graph to unify the definition of different CSV types and provides a computational method to compare different CSV graph structures. To create

a CSV graph $G = (V, E)$, SVision first collects the node set $V = V_s \cup V_I \cup V_D$ of G. Specifically, skeleton node set $V_s = \{S_1, S_2, \ldots, S_n\}$, insertion node set $V_I = \{I_1, I_2, \ldots, I_m\}$ and duplication node set $V_D = \{V_1, V_2, \ldots, V_k\}$ contain n, m and k skeleton nodes, insertion nodes and duplication nodes in the graph, respectively. Skeleton nodes are derived from major segments in a one-variant image and sequence between discontinuous major segments on REF (that is, concordant segments between VAR and REF). Insertion nodes consist of minor segments in the one-variant image, while insertion nodes with known origins are defined as duplication nodes. Moreover, each node $v_i \in V$ is represented as a tuple v_i = (Seq, Pos, Strand), corresponding to a segment in the one-variant image. The Seq indicates the segment sequence, Pos is the position of the segment on VAR, and Strand represents the forward or reverse strand of the segment aligned on the reference genome. The edges in G are collected by $E = E_{ad} \cup E_{dp}$. E_{ad} represents a set of adjacency edge $e^k_{ad} = (v_k, v_{k+1})$, connecting two adjacent nodes v_k and v_{k+1}, and E_{dp} represents a set of duplication edge e_{dp}, connecting the duplicated node with its known origin. For each CSV, its breakpoint and graph structure information are kept in the 'BKPS', 'GraphID' and 'GFA_FILE_PREFIX' column, and the CSV graph is saved in GFA format (Supplementary Note). Given a CSV graph G, a CSV could be interpreted by visiting each node through the E_{ad} edges. For example (Extended Data Fig. 10a), the CSV path is interpreted as 'S1+S3–S3–S4+', where '+' or '–' indicates the direction (that is, node Strand) of visiting a specific node. Specifically, nodes S1 and S4 are visited in forward direction (+), while S3 is visited in reverse direction (–), so that the path should be 'S1+S1+S3–S3–S4+S4+'. But for simplicity, only the intermediate nodes, such as S3, are kept twice, whereas the start node (S1) and the end node (S4) are used once in the path.

The comparison of two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ is a non-deterministic polynomial (NP)-hard problem, but ordering the nodes on the basis of the reference coordinate system simplifies this problem (Supplementary Note). SVision first compares the numbers of edges and nodes between two graphs G_1 and G_2, which are considered different if either number is different. However, if graphs G_1 and G_2 have a topologically identical path in addition to the same numbers of nodes and edges, they are termed as isomorphic CSV graphs, that is, $G_1 = G_2$. If graphs G_1 and G_2 have the same numbers of nodes and edges but differ in paths, we further examine whether G_1 and G_2 share symmetric topology (Extended Data Fig. 10b), since a variant might be identified on either forward or minus strand, that is, from 5′ to 3′ or from 3′ to 5′. In particular, we create a mirror graph G'_1 of the original graph G_1, and obtain a new path from G'_1. Similarly, we also create G'_2 from G_2 Then, we cross compare whether the paths between G'_1 and G_2 as well as between G'_2 and G_1 are topologically identical. We consider G_1 and G_2 isomorphic if both comparisons are equal. SVision keeps isomorphic graphs and symmetric graphs in two separate files, enabling search of CSV events of the same structure.

Training data. The CNN model in SVision is trained with both real and simulated simple SVs of DEL, INV, INS, DUP, and tDUP, to avoid usually unbalanced numbers of SV types in real data. We obtained real SVs from NA19240 (4,282) and HG00514 (3,682) by selecting calls supported by both PacBio CLR reads and Illumina reads[16]. In this integrated real SV set, we labeled SVs with the above-mentioned five SSV types (that is, INS, DEL, INV, DUP, and tDUP). Because INS and DEL dominate SVs from real samples, we further used VISOR with the parameters '-n 4000 -r 20:20:20:20:20 –l 1000 –s 500' to create more INV, DUP, and tDUP for training. For all training SVs, their one-variant images and SOIs were created as we described in the above sections, leading to 75,000 SOIs (15,000 per type) in total, where 50% SOIs are from real events. All SOIs were used for further CNN model training (Supplementary Note).

Convoluted neural network model training. SVision adopts AlexNet[26] to classify sequence differences in similarity images. The AlexNet architecture consists of five convolutional layers and three fully connected layers. The first convolution layer loads images of size 224 × 224 × 3, and it uses the 11 × 11 × 3 convolution kernel with stride 4. The last three layers are fully connected and contain a five-class SoftMax layer for classification. In the end, the input SOIs are detected as either INS, DEL, INV, DUP, tDUP, or mixed types for CSVs. We applied the idea of transfer learning to train the CNN with 75,000 SOIs. First, the parameters of all layers in the CNN were initialized to the best parameter set that was achieved on the ImageNet competition. Next, we fine-tuned the parameters of the last three fully connected layers on our training data using back propagation and gradient descent optimization with a learning rate of 0.001. The loss function was defined as the cross entropy between predicted probability and the true class labels. To evaluate the trained AlexNet model, we applied tenfold cross-validation and examine the loss and accuracy of each model on the training set and used an independent set of 7,500 SOIs to measure the accuracy. We also assessed the AlexNet accuracy and robustness with different initialization parameters (random initialization) and different network structures (InceptionV3). Moreover, we examined the interpretability of features extracted by AlexNet during training, and these features could also be used by classic machine-learning methods (for example, SVM and logistic regression) for accurate classification (Supplementary Note).

Quality score of discoveries. SVision measures the quality of each discovery on the basis of consistency and prediction reliability derived from one-variant image clusters that support an event.

1. One-variant image consistency. Intuitively, the non-linear segments in a given one-variant image indicate potential differences between REF and VAR. We thus first compute the non-linear score for all images that support each event, that is, one-variant images originated from a cluster of VARs supporting the same event. The non-linear score of a one-variant image is calculated by its segment coordinates and lengths. Specifically, for a one-variant image with segments:

$$\text{Nonlinear } score_i = \frac{\sum_k \left(\left|\text{k}.ref_{mid} - \text{k}.read_{mid}\right|\right) \times \text{k}.length}{\text{Refpan}}$$

where the summation is over all segments k in image i, and k.ref$_{mid}$ and k.read$_{mid}$ are the center of segment on reference and read, respectively. Then, we normalize the summation by dividing *RefSpan*, which denotes the distance between the leftmost and rightmost coordinates of the similarity image. Finally, for a SV of M supporting images, we calculate the consistency score with the following equation:

$$\text{Consistency} = \frac{\text{Std}\left(\{\text{nonlinear } score_1, \ldots, \text{nonlinear } score_M\}\right)}{\text{M}}$$

Accordingly, we expect a smaller consistency value for high-quality SV predictions.

2. Prediction reliability. This part evaluates the deep-learning prediction quality. The last layer in the CNN architecture is a SoftMax layer, which outputs the probability of the prediction results. Therefore, we use the average probability of all SOIs as the CNN reliability:

$$\text{Reliability} = \frac{\sum_s \text{s}.softmax \times 100}{\#\text{SOIs}}$$

where the summation is over all SOIs in a one-variant image. The reliability will range from 0 to 100 because the SoftMax probabilities always range from 0 to 1. We expect higher reliability values for accurate SVs.

Finally, we summed up the two features and normalized it to range from 0 to 100:

$$\text{qual} = \text{Consistency} + (1 - \text{Reliability})$$

$$\text{Normalized } score = \left(1 - \frac{sum(\text{Scores}) - min(\text{Scores})}{max(\text{Scores}) - min(\text{Scores})}\right) \times 100^{\text{where}}$$

Scores = {qual$_1$, …, qual$_M$}, and M is the total number of images supporting this variant.

Analysis of complex structural variants detected from HG00733. The HiFi reads of HG00733 were aligned to reference GRC h38 by ngmlr (v0.2.7) with the default setting. Then, SVision was run under the default setting, except with parameters '-s 5–graph–qname'.

First, the events detected by SVsion at low-mapping-quality regions, centromeres, genome gap regions and so on were excluded in analysis. These regions were obtained from https://github.com/mills-lab/svelter/tree/master/Support/GRCh38 and the UCSC genome centromere for reference GRCh38. Then, we applied the following steps to filter CSVs from the raw callset. (1) Filtering CSVs of length larger than 100 kbp; (2) filtering CSVs without complete graph representation, where the path ends with other node types instead of 'S'; and (3) for multiple CSVs at one site, we kept only the one with the greatest number of supporting reads. SVision revealed two special complex structures, that is, a structure consists of nodes 'S:2,I:2,D:1' and path 'S1+I1+I1+I2+I2+S2+' as well as another structure consists of nodes 'S:2,I:1,D:1' and path 'S1+I1+I1+S2+', which were visually confirmed as local targeted-site-duplication (Extended Data Fig. 10c) and tandem-duplication (Extended Data Fig. 10d). Events of these two structures were also filtered because they were considered as simple events from a biological perspective. Afterwards, we used RepeatMasker and tandem repeat finder (TRF) annotated files from UCSC genome browser to annotate the CSVs passed the filters through BEDtools[27] intersect option. The repeat type was assigned if the CSV region overlaps with the repeat element, while the size or percentage of overlaps was not required. For CSVs with multiple repeat types, the one with the largest overlapping region with the CSV was chosen. Meanwhile, CSV was annotated as STR if the repeat unit length <7 bp; otherwise, it was annotated as VNTR. Finally, we termed all CSVs outside of VNTR/STR regions as high-quality CSVs, which were validated and used for further analysis. The PAV and short-read data matched CSV loci were obtained through BEDtools without requiring overlap size. For the short-read data, a matched CSV locus was considered as completely reconstructed if both breakpoint positions and types matched what SVision reported, otherwise as partially reconstructed events if either breakpoints or types agreed with SVision's prediction. The related analysis of CSV on *CNTN5* among 35 samples and the insertion-inversion-insertion event are described in Supplementary Note.

Validation of high-quality complex structural variants detected from HG00733. We validated 80 CSVs detected by SVision in HG00733 via (1) graph-based

alignment; (2) contig-based visual confirmation; and (3) PCR and Sanger sequencing (Supplementary Note).

Graph-based alignment. For each CSV graph in rGFA format, we extracted the CSV locus-spanning reads with SAMtools and aligned these reads to each CSV graph using GraphAligner (v1.0.12) with the default settings. A CSV was successfully validated if a single ONT read could be aligned to the corresponding variant path specified in the rGFA file. We then counted the number of long reads covering the entire VAR path as the number of supports for this CSV event.

Contig-based visual confirmation. To examine the internal structure of CSVs, the phased-assembly specified in the PAV (v1.1.2, TIG_REGION column) at the reported variant region was used for further analysis. We first extracted the contig sequence harboring a variant based on the coordinates provided in the 'PAV_TIG_REGION' (Supplementary Table 8). For example, a sequence-containing variant was extracted from the h1 assembled genome for '1|1' and '1|0' genotype and from the h2 assembled genome for '0|1.' In order to validate CSV-structure-containing complex insertion, we extended 5 kbp both upstream and downstream of the CSV region to extract the reference genome via BEDtools, from which the origin of the inserted sequence could be identified. Then, Gepard was used to create the Dotplot of contig sequence (*y* axis in the Dotplot) and reference sequence (*x* axis in the Dotplot) for each CSV locus. On the basis of each contig Dotplot, the manual validation contained two tiers of metrics: (1) whether the reported region contains a variant; and (2) whether the SVision reported structure is identical to what revealed by Dotplot. A CSV was considered completely reconstructed if both (1) and (2) were satisfied, and others were considered inconclusive events.

PCR and Sanger sequencing. We first determined that about half of the 80 CSVs (39/80) were unusable for PCR owing to their location within segmental duplications, the size of the amplicon needed to validate the rearrangement, or the simple repeat nature of the rearrangement. We then randomly selected 20 of the remaining rearrangements and performed BLAT on the local region from the HG0733 assembly data. We next attempted to subject each of the 20 CSVs to PCR. Briefly, we designed primers flanking the CSV or flanking breakpoints within the CSV for each of the 20 events (Supplemental Table 12). Next, we attempted to amplify each region using Takara LA taq. We obtained the predicted band size for 12 of the 20 variant loci. The remaining eight regions did not amplify in three attempts with alterations of the PCR conditions and template amounts. All PCR products underwent Sanger sequencing and were validated as on target, and contained the correct amplicon with the breakpoint from the assembly and SVision call.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

HG002 ONT and HiFi data were downloaded from ftp://ftp.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG002_NA24385_son/UCSC_Ultralong_OxfordNanopore_Promethion/ and https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG002_NA24385_son/PacBio_CCS_15kb/, respectively. The NA12878 HiFi data was downloaded from http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/release/v1.0/assemblies/20200628_HHU_assembly-results_CCS_v12/haploid_reads.
The HG00733 HiFi and ONT data were downloaded from http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/working/20190925_PUR_PacBio_HiFi/ and http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/hgsv_sv_discovery/working/20181210_ONT_rebasecalled/, respectively. The HG00733 assembly was download from http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/working/20200417_Marschall-Eichler_NBT_hap-assm/.
The human reference genome hg19 was downloaded from ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/reference/phase2_reference_assembly_sequence/hs37d5.fa.gz. The human reference genome GRCh38 was downloaded from http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/GRCh38_reference_genome/.
The HG00733 PAV callset was downloaded from http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/working/20210806_PAV_VCF/. The merged PAV callset of 35 samples was downloaded from http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/release/v2.0/integrated_callset/.
The RNA-seq data was downloaded from Sequence Read Archive of project ID PRJNA720779.
All results generated by this study are available in Supplementary Note from the article.

Code availability

The SVision program (v1.3.6) and trained model are provided at GitHub (https://github.com/xjtu-omics/SVision), which is available under GNU General Public License v3.0. SVision is free for non-commercial use by academic, government and non-profit/not-for-profit institutions. Please contact the corresponding author for more information about commercial usage. A Code Ocean capsule of the package is provided (https://doi.org/10.24433/CO.8937098.v1).

References

22. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
23. Bolognini, D. et al. VISOR: a versatile haplotype-aware structural variant simulator for short- and long-read sequencing. *Bioinformatics* **36**, 1267–1269 (2020).
24. Jeffares, D. C. et al. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat. Commun.* **8**, 14061 (2017).
25. Krumsiek, J., Arnold, R. & Rattei, T. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics* **23**, 1026–1028 (2007).
26. Krizhevsky, A., Sutskever, I. & Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems* **25**, 1097–1105 (2012).
27. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

Acknowledgements

We thank X. Zhao, P. Balachandran, A. Wenger, and other members of the Human Genome Structural Variation Consortium for helpful discussions on methods development and structural variants analysis. K. Y. and X. Y. are supported by National Science Foundation of China (32125009, 32070663 and 62172325), the Key Construction Program of the National '985' Project, the World-Class Universities (Disciplines), the Fundamental Research Funds for the Central Universities, and the Characteristic Development Guidance Funds for the Central Universities. C. R. B., P. A. A., and J. I. F. are supported by the National Institutes of Health R35GM133600 through the NIGMS and pilot funding from the Jackson Laboratory Cancer Center (P30 CA034196). D. M. is supported by the National Science Foundation of China (61721002) and the Macao Science and Technology Development Fund under Grant (061/2020/A2).

Author contributions

K. Y. designed and supervised research; J. L. and S. W. developed the algorithm and software; D. M. contributed to the assessment and analysis of the deep-learning model; W. K., T. M. and P. A. provided constructive suggestions for the algorithm; J. L. performed the algorithm benchmarking on real data and CSV analysis; S. W. performed algorithm benchmarking on the simulated data. P. A. A., J. I. F., and C. R. B. contributed to the analysis and experimental validation of complex structural variants; P. J. and X. Y. contributed to the sequencing data processing; J. L., W. K., P. A. A., C. R. B., and K. Y. 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.

Additional information

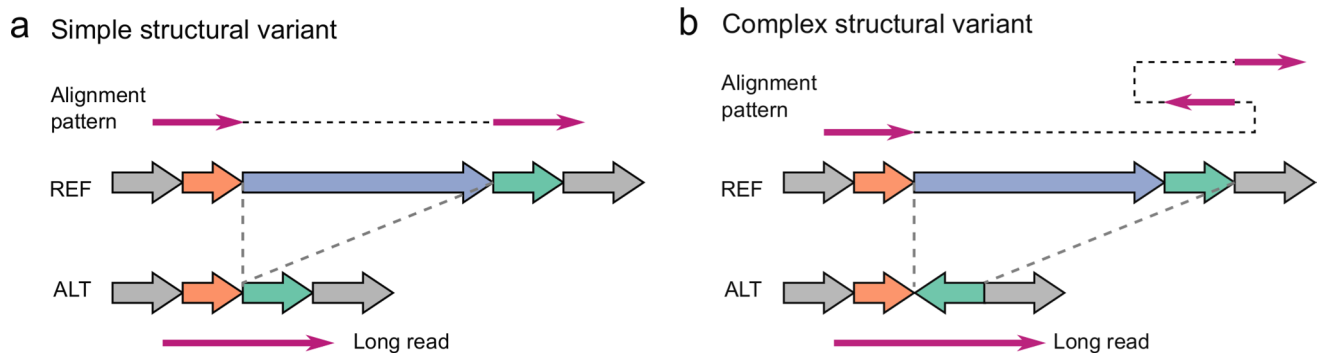
Extended data is available for this paper at https://doi.org/10.1038/s41592-022-01609-w.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-022-01609-w.

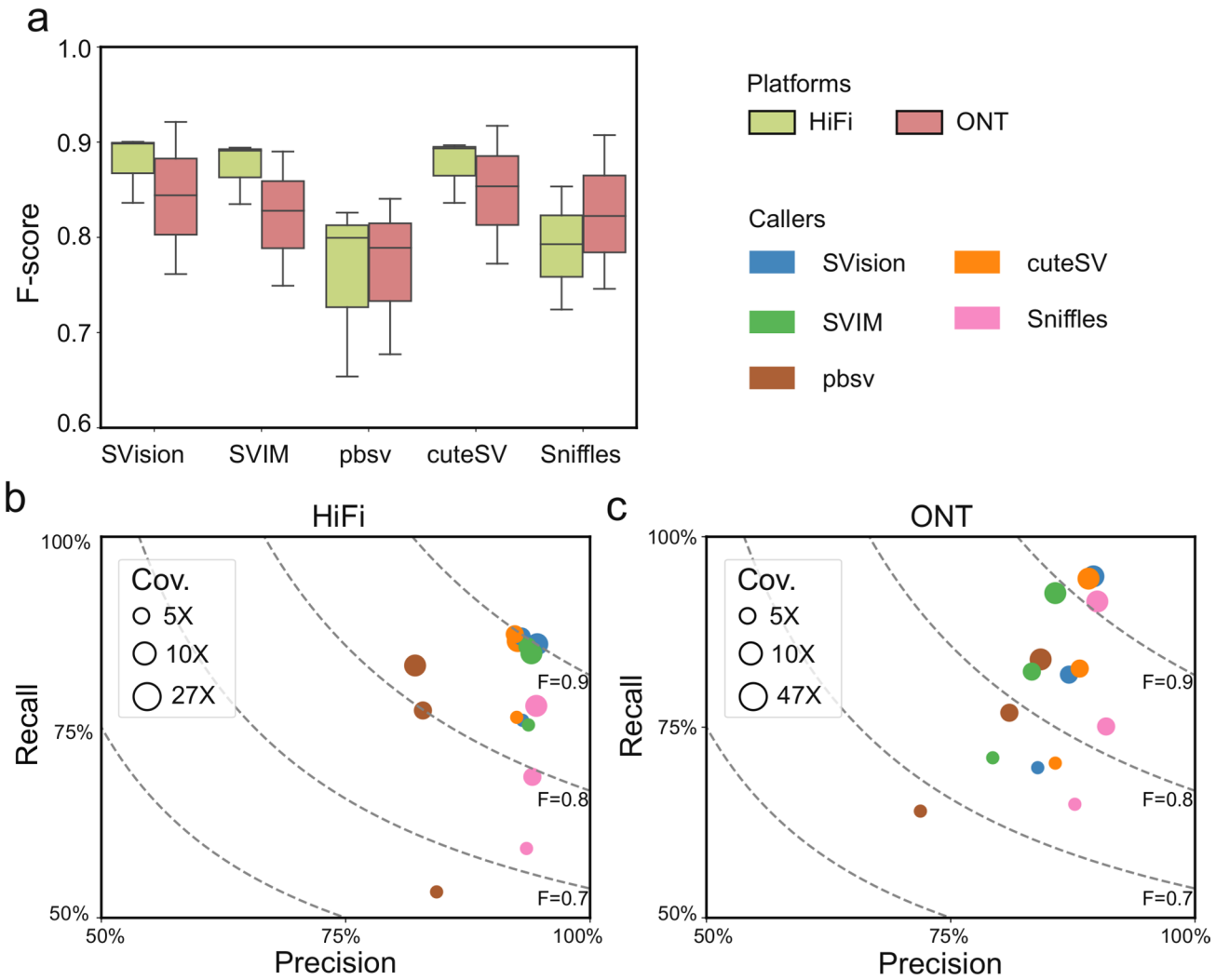
Correspondence and requests for materials should be addressed to Kai Ye.

Peer review information *Nature Methods* thanks Ryan Layer and the other, anonymous, reviewer for their contribution to the peer review of this work. Peer reviewer reports are available. Primary Handling editor: Lin Tang, in collaboration with the Nature Methods team.

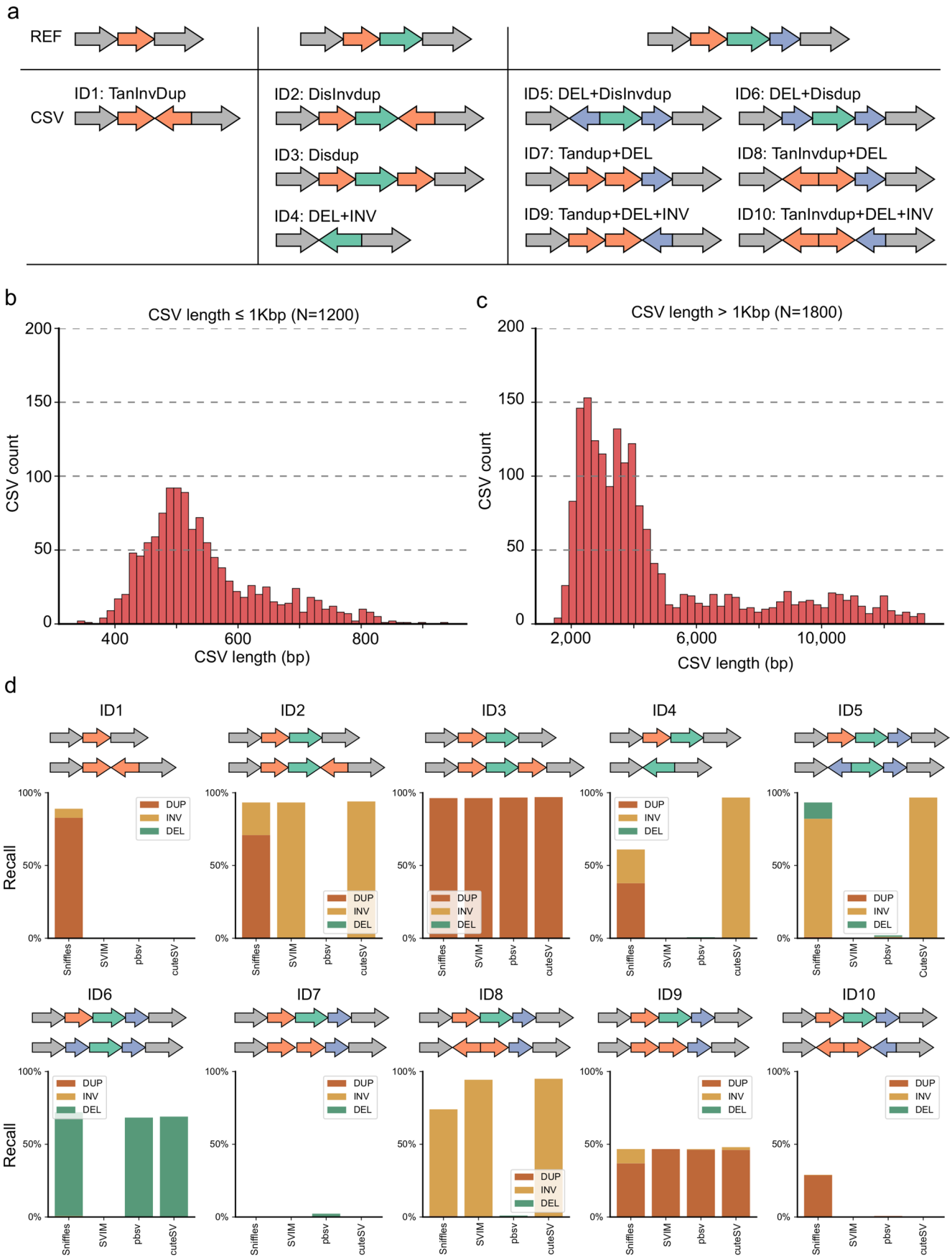
Reprints and permissions information is available at www.nature.com/reprints.



Extended Data Fig. 1 | Diagram of example simple and complex structural variants and their aberrant alignment patterns. a, The diagram and alignment pattern of a simple deletion. **b**, The diagram and alignment pattern of a deletion associated with inversion, where the inverted segment occurred at the 3′ flank region of the deletion.

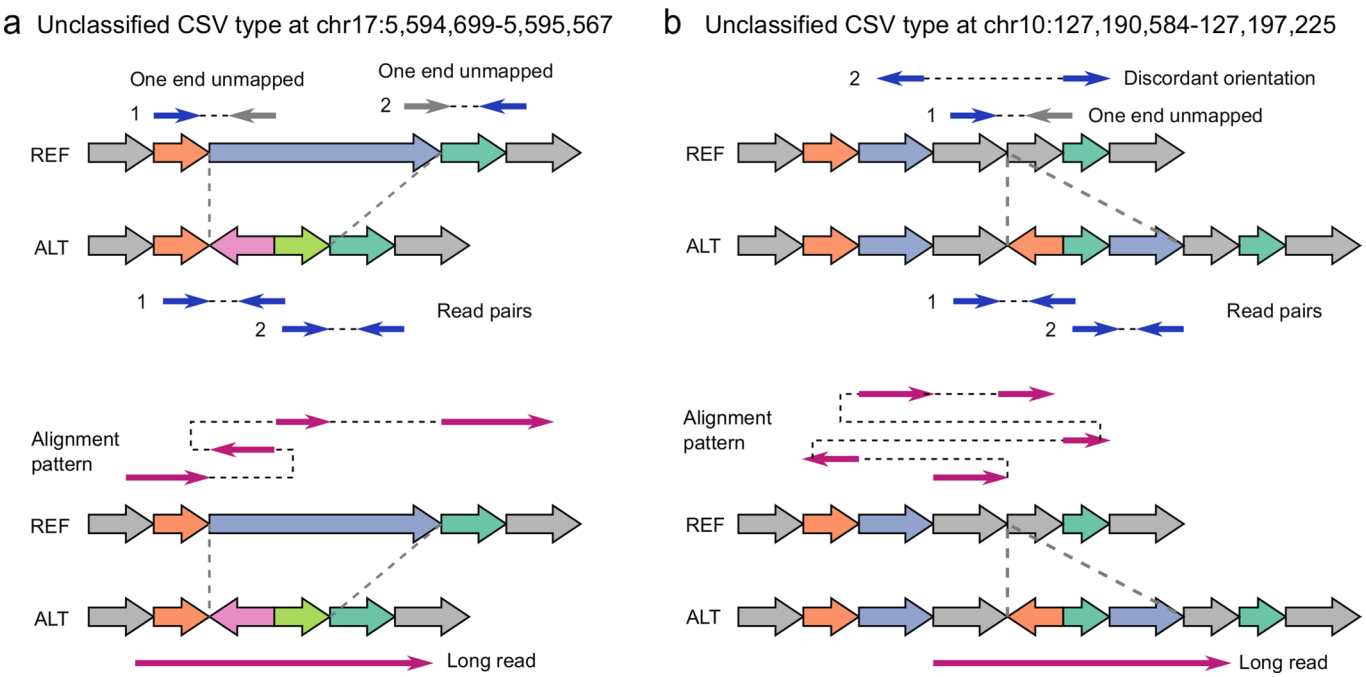


Extended Data Fig. 2 | Performance evaluation of callers with HG002 truthset at different coverages and platforms. a, F-score of callers on different platforms evaluated with Truvari. The boxplot for HiFi data was the F-score measured for each caller at 5X, 10X and 28X coverage, respectively. Each box contains three values, that is, SVision (0.83, 0.89 and 0.90), SVIM (0.83, 0.89 and 0.89), pbsv (0.65, 0.79 and 0.82), CuteSV (0.83, 0.89 and 0.89) and Sniffles (0.72, 0.79 and 0.85). The boxplot for ONT data was the F-score measured for each caller at 5X, 10X and 47X coverage, respectively. Each box also contains three values (n = 3), that is, SVision (0.76, 0.84 and 0.92), SVIM (0.74, 0.82 and 0.89), pbsv (0.67, 0.78 and 0.84), CuteSV (0.77, 0.85 and 0.91) and Sniffles (0.74, 0.82 and 0.90). The boxplot defines the median (Q2, 50th percentile), first quartile (Q1, 25th percentile) and third quartile (Q3, 75th percentile). The bounds of box, that is interquartile range (IQR), of the boxplot is between Q1 and Q3. The minima and maxima values are defined as Q1-1.5*IQR and Q3 + 1.5*IQR, respectively. The whiskers are values between minima and Q1 as well as between Q3 and maxima. **b**, The precision (x-axis), recall (y-axis) and F-score (F, dotted line) measurements of detecting SVs from HiFi data at different coverages. **c**, The precision and recall measurements of detecting SVs from ONT data at different coverages. It should be noted that this evaluation ignored SV genotype, but only evaluated on event level.

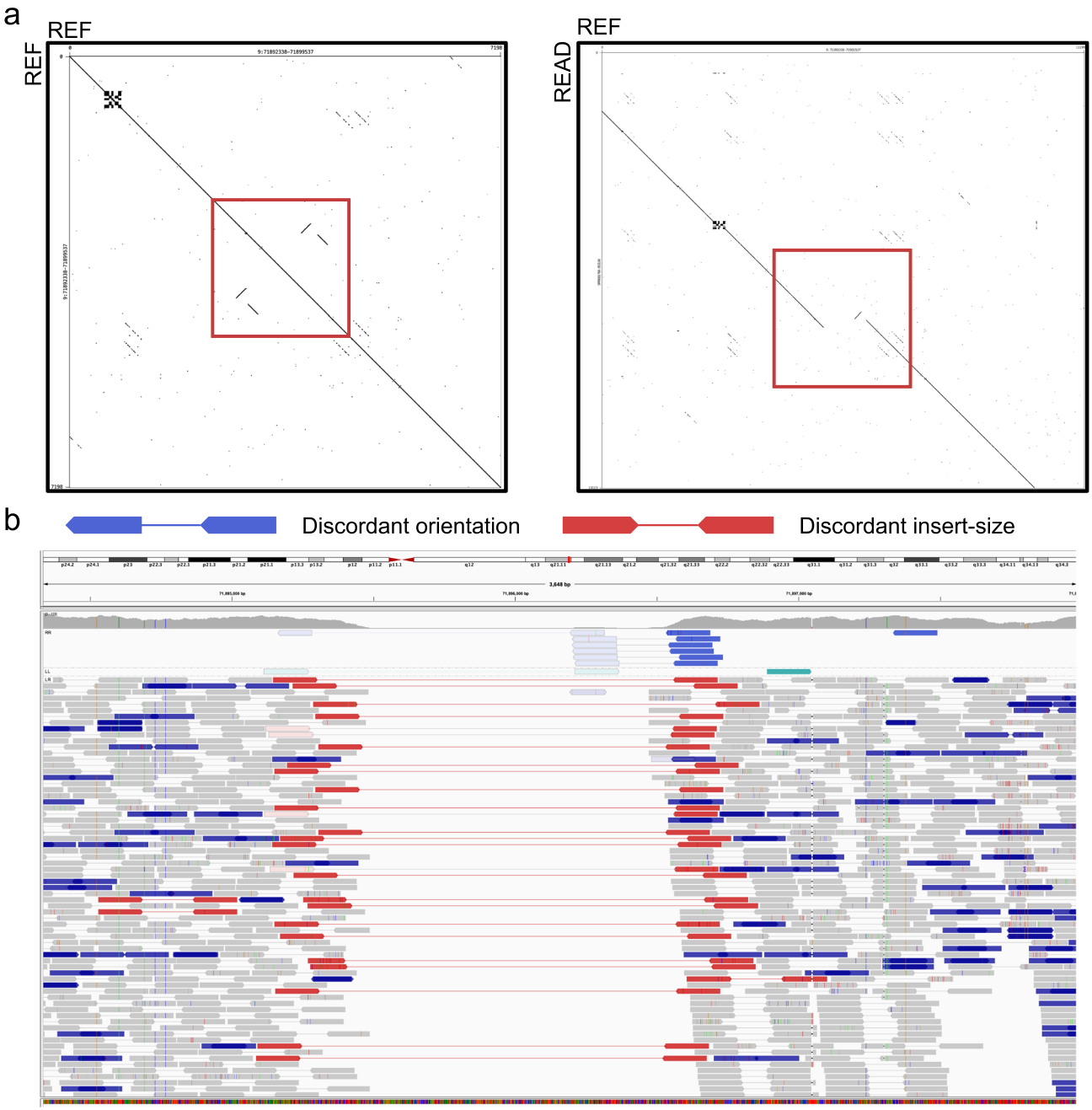


Extended Data Fig. 3 | See next page for caption.

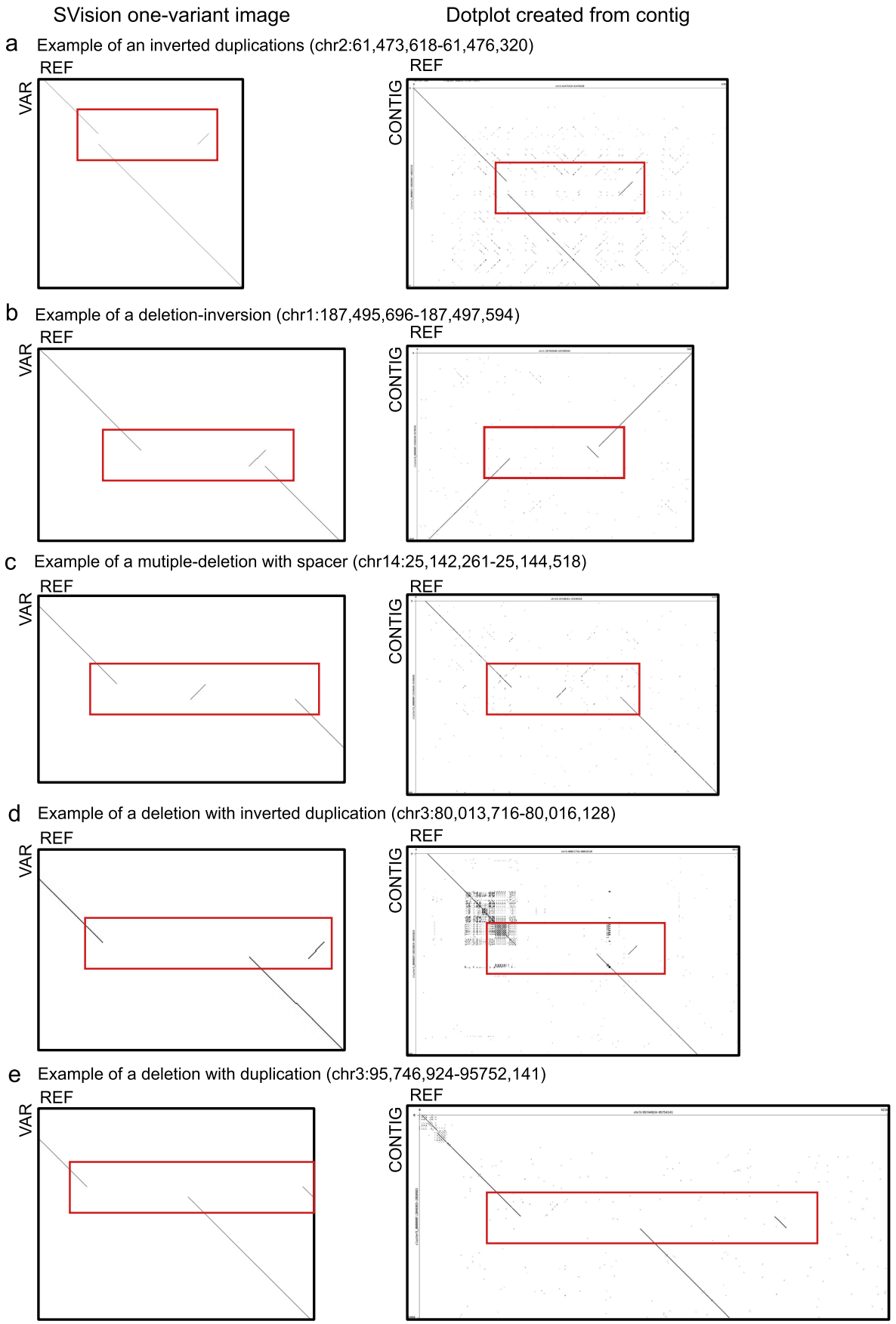
Extended Data Fig. 3 | Simulated complex structural variant types and performance of detecting complex structural variant subcomponents. a, The diagrams of simulated complex structural variants (CSV). Each type has a unique ID and a type definition. **b**, The size distribution of simulated CSVs smaller than 1Kbp (1,200 events). **c**, The size distribution of simulated CSVs larger than 1Kbp (1,800 events). **d**, The region-match recall rates of model-based callers for detecting subcomponents (that is, DUP-duplication, DEL-deletion, INV-inversion) of CSVs.



Extended Data Fig. 4 | The diagrams and alignment patterns of two unclassified complex structural variants. a, SVision correctly detected a deleted sequence replaced with dispersed duplication and inverted duplication. **b**, SVision characterized a complex insertion, consisting of two dispersed duplications and one inverted duplication. Both types of (**a**) and (**b**) are labeled as unclassified (NA) in the 1KGP call set. The top panel of (**a**) and (**b**) are the discordant alignments derived from short-read sequencing (that is, one end unmapped and discordant alignment). The bottom panels of (**a**) and (**b**) describe the abnormal alignments from long-read alignment.

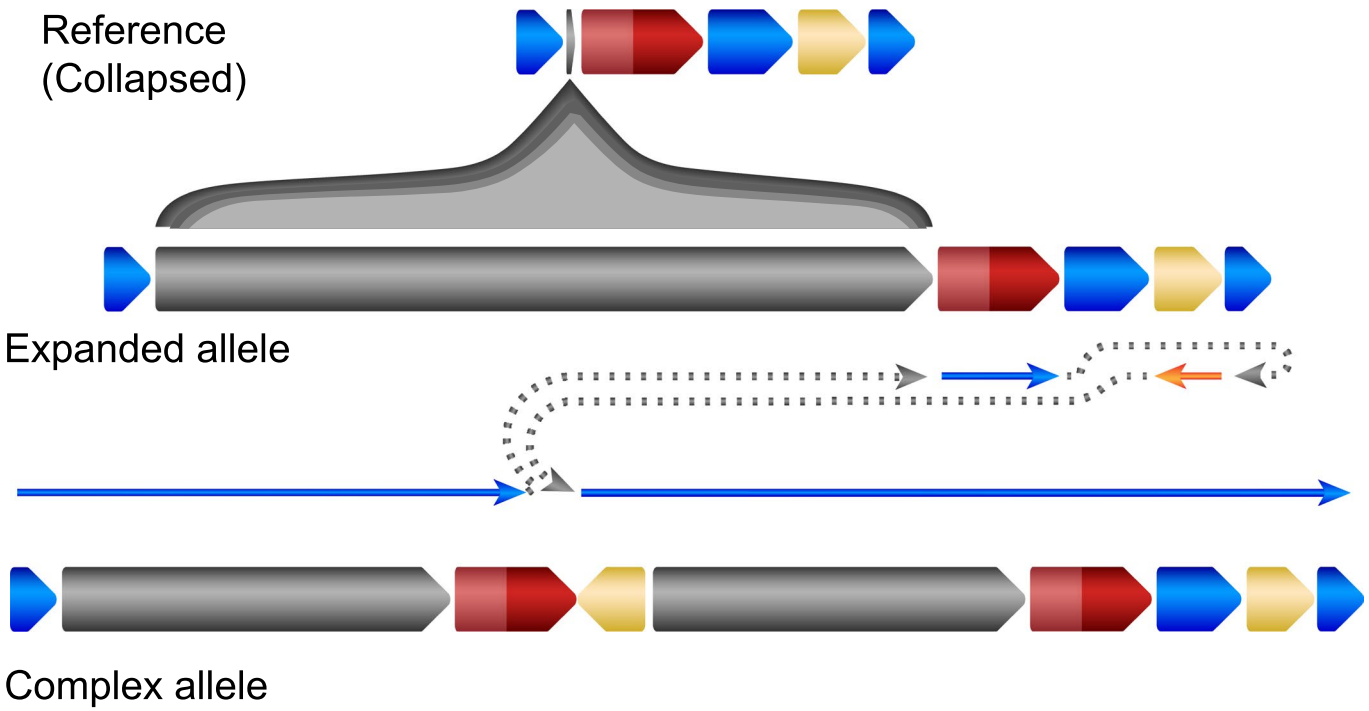


Extended Data Fig. 5 | One example of simple deletion misinterpreted as complex event by short-read data due to local repeats. a, Two Dotplots are created with Gepard to illustrate the local repeats at the variant locus on the reference genome (left) and the breakpoints comparing HiFi read (READ, y-axis) and the reference genome (REF, x-axis). **b**, The IGV view at this locus with reads grouped by pair orientation and colored by insert-size and pair orientation.



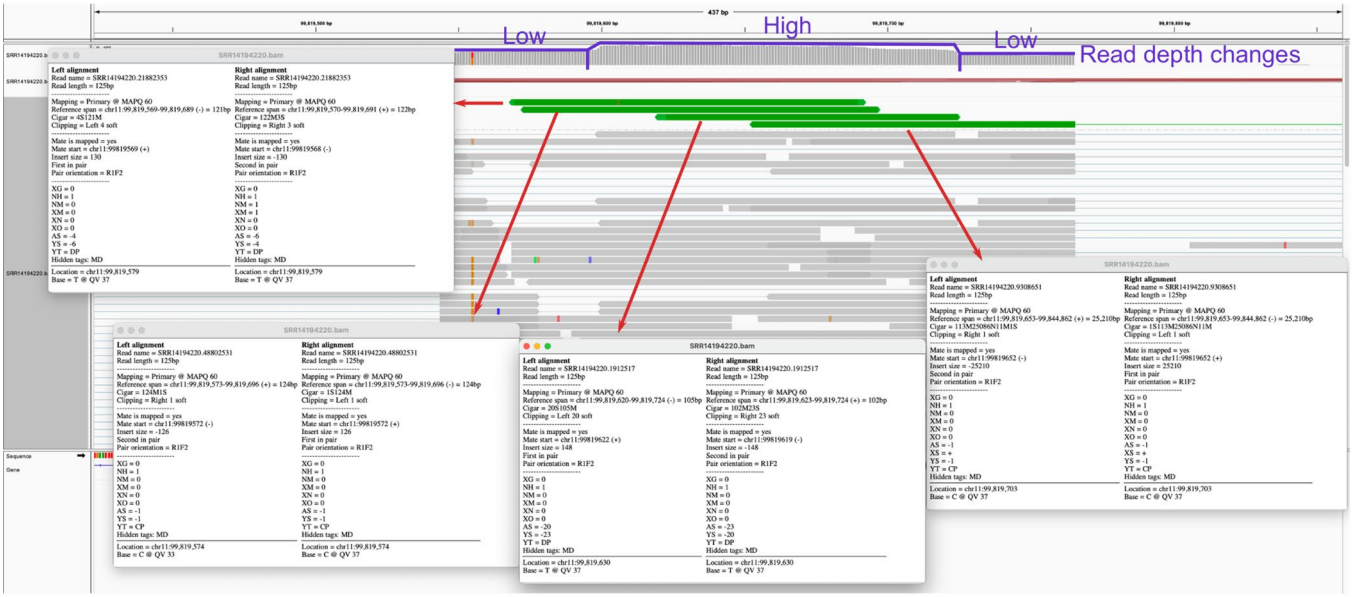
Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Examples of reported complex structural variant types identified by SVision. a, One of the 12 inverted duplication events detected by SVision and classified as CSV graph structure '12', **b**, One of the eight deletion associated with inversion events detected by SVison and classified as CSV graph structure '15'. **c**, One of the five multiple-deletion with spacer events detected by SVision and classified as CSV graph structure '27'. **d**, One of ten deletion with inverted duplication events detected by SVision and classified as CSV graph structure '23'. **e**, One of the five deletion with duplication events detected by SVision and classified as CSV graph structure '28'. From figure (**a**) to (**e**), the Dotplots on the left column are SVision one-variant images created with variant feature sequence (VAR, y-axis) and reference sequence (REF, x-axis) at the variant loci, while the Dotplots on right column are created with variant spanning HiFi assemblies (CONTIG, y-axis) and the reference sequence (REF, x-axis) at the variant loci.

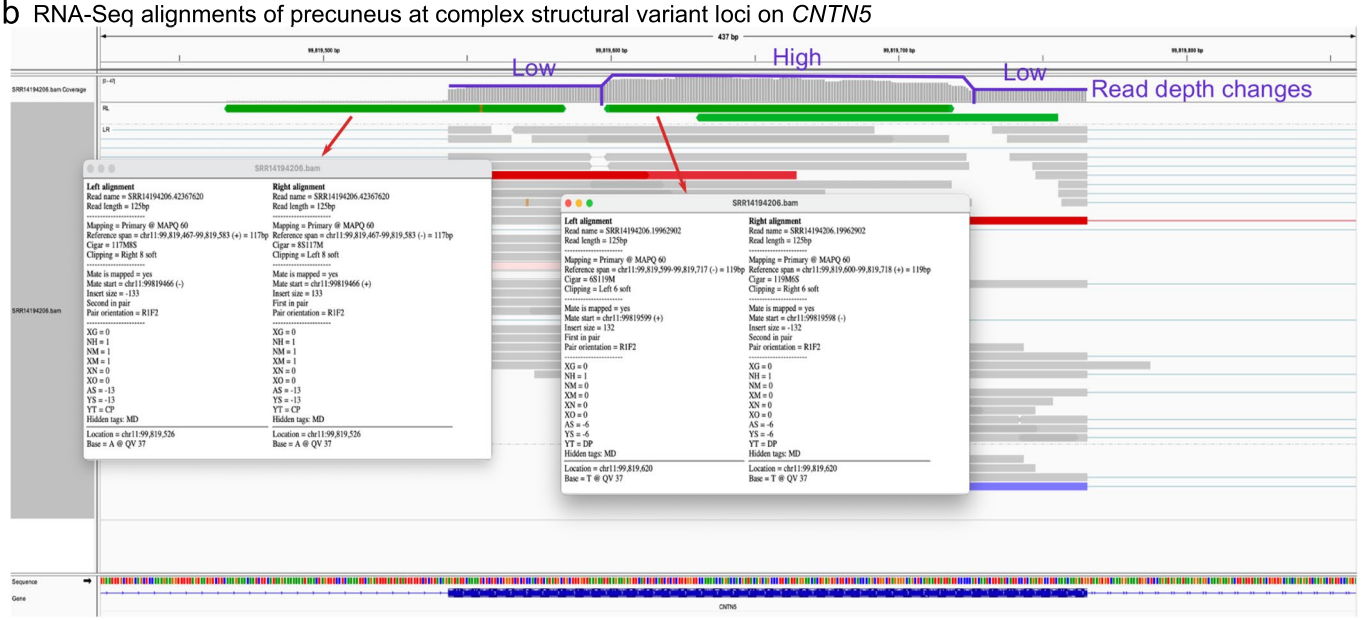


Extended Data Fig. 7 | The HiFi assembly reconstruction of the expanded allele and complex structural variant allele affecting CNTN5. The grey region indicates the repeat expansion. The dark red region indicates exon 4 of CNTN5, while the light red region is the 5′ flanking region of the exon.

a RNA-Seq alignments of primary visual cortex at complex structural variant loci on *CNTN5*

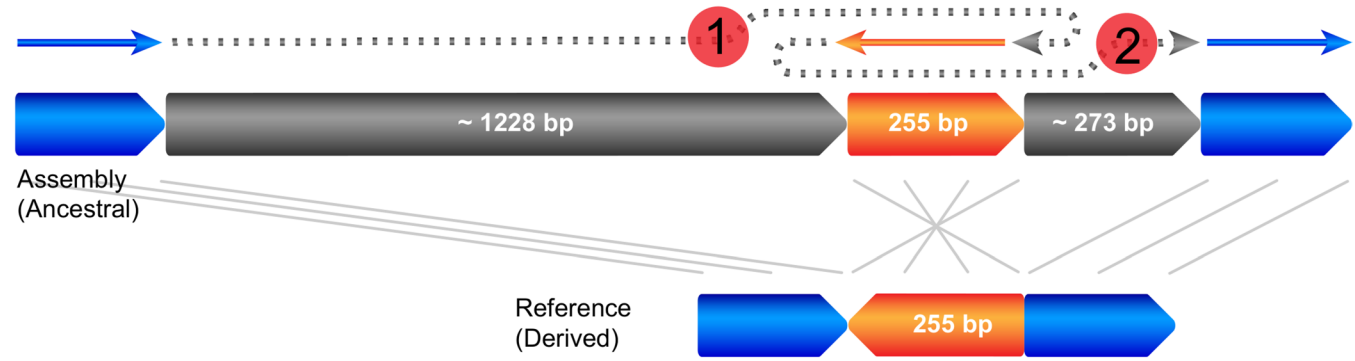


b RNA-Seq alignments of precuneus at complex structural variant loci on *CNTN5*



Extended Data Fig. 8 | The IGV screenshot of duplicated CNTN5 exon signature observed in RNA-Seq data. The RNA-Seq data of the primary visual cortex from an Alzheimer disease female. **b**, The RNA-Seq data of a control male precuneus. In (**a**) and (**b**), the green bars pointed by red arrows are duplication like read-pair signatures, that is, there are 4 supporting discordant read-pairs in (**a**), and 2 in (**b**). Moreover, read depth change (fitted by purple line) on exon is observed in both (**a**) and (**b**). The RNA-Seq data for (**a**) and (**b**) are obtained from Sequence Read Archive (SRA) with accession number SRR14194220 and SRR14194206, respectively.

a



1
CAACTAATCAGCAATGCTGCTTCAGCAAACTGGCCAATTGCAGCCTTTGAAGGTGCCTAAGTGACATCATCAATACGAAGAAAGCATCATTCCTGATCACAAGCAGTT Tig dn +
CTCAATCAATAGCATTTCACCCCTCACCCCTCTCTCACCCTCCTGCCCACTGTAGCCTAAGTGACATCATCAATACGAAGAAAGCATCATTCCTGATCACAAGCAGTT Reference +
CTCAATCAATAGCATTTCACCCCTCACCCCTCTCTCACCCTCCTGCCCACTGTAGCCTCCAATGTCTATTATTCTATTCTGCATGTCCGTGTGTACTCATTGCTTTCC Tig up -

2
ATATTTGAAGAGATAGATTGGTTTTCTCACATTATACATACATCACCTATTAACTAGAGTCACAGTCAACCAAGTTCAGCACTACTGGGCTAAAGATATAATGGGAGA Tig up +
ATATTTGAAGAGATAGATTGGTTTTCTCACATTATACATACATCACCTATTAACTAGCAGGAAGTGGGAACTTATAAAAGCCTGGCGTCACACCCTGACTTGGGTCAG Reference +
AAAGAAAGCTGATGTGTCCACGGTTCTGTTCTTTTCAGGGGTAGATGTAGGTGGGAGCAGGAAGTGGGAACTTATAAAAGCCTGGCGTCACACCCTGACTTGGGTCAG Tig dn -

b

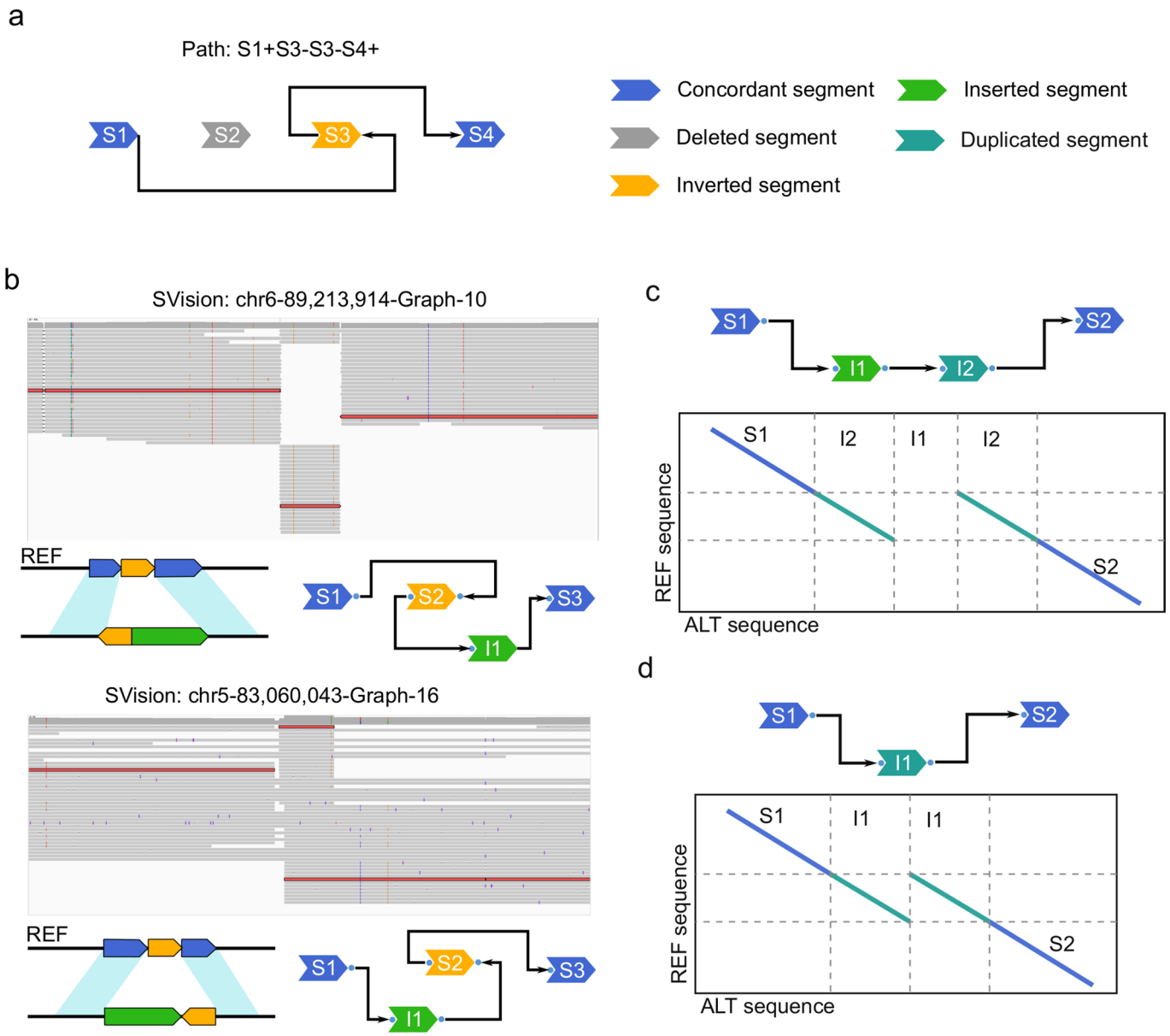
Descriptions | Graphic Summary | Alignments

Sequences producing significant alignments Download | New Manage columns | Show 100

select all *0 sequences selected* GenBank | Graphics | Distance tree of results | New MSA Viewer

Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Pan paniscus isolate Mhudiblu (Carbone #601152) 000048F_17960554_qpds_1_17899071, whole genome shotgun ...	3251	3251	100%	0.0	99.33%	17906892	SSBP03006069.1
Pan troglodytes isolate Yerkes chimp pedigree #C0471 (Clint) 000040F_1_17963911_quiver_pilon, whole genome s...	3251	3251	100%	0.0	99.33%	17982150	NBAG03000246.1
Pan paniscus isolate Ulindi cntg77132, whole genome shotgun sequence	3251	3251	100%	0.0	99.33%	233332	AJFE02076954.1
Pan troglodytes isolate Yerkes chimp pedigree #C0471 (Clint) ctg_18392, whole genome shotgun sequence	3251	3251	100%	0.0	99.33%	10862	AADA01018393.1
Pan troglodytes isolate Yerkes chimp pedigree #C0471 (Clint) Contig278_iteration_1.32, whole genome shotgun seq...	3251	3251	100%	0.0	99.33%	950268	AACZ04029827.1
Gorilla gorilla gorilla breed western lowland gorilla isolate Kamilah (stud number 0661) 000043F_15087491_qpdss17...	3217	3217	100%	0.0	99.00%	13343626	SRLZ01004541.1
Gorilla gorilla gorilla WGS project CABD00000000 data, contig gorGor4_chr9_2634, whole genome shotgun sequence	3212	3212	100%	0.0	98.94%	72364	CABD030064330.1

Extended Data Fig. 9 | The ancestral state of one genome segment revealed by a complex structural variant. a, The structure and breakpoint junction sequence of the variant derived from HiFi assembly. **b**, Blastn results of the inserted sequence mapping to primate genomes, and the top hits include pan troglodytes and gorilla.



Extended Data Fig. 10 | Examples of graph and symmetric graphs as well as two special complex events identified by SVision. a, An example of a complex structural variant (CSV) graph where its graph path is interpreted as S1 + S3-S3-S4+. **b**, Examples of isomorphic graphs representing two different CSV events. **c**, SVision detected CSV classified as local target site duplication. **d**, SVision detected CSV classified as tandem duplication. Though events of structure depicted by (**c**) and (**d**) were computed as complex events, they were considered as simple events from the biological perspective.

nature research

Corresponding author(s): Kai Ye

Last updated by author(s): Jul 4, 2022

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	fastq-dump (v2.9.1) is used to download the RNA-Seq data.
Data analysis	pbmm2 (v1.4.0) and ngmlr (v0.2.7) are used for long-reads alignment. Sniffles (v1.0.12), cuteSV (v1.0.10), pbsv(v2.2.2), SVision (v1.3.6) and SVIM (v1.4.0) are used for SV detection. Truvari (v2.0.1) is used for assessing SV detection performance in HG002 genome. VISOR (v1.1.2) is used for complex events simulation. Gepard (v1.4.0) is used to create Dotplot for manual inspections.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

HG002 ONT reads: ftp://ftp.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG002_NA24385_son/UCSC_Ultralong_OxfordNanopore_Promethion/
HG002 HiFi reads: https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG002_NA24385_son/PacBio_CCS_15kb/
Human hg19 reference: ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/reference/phase2_reference_assembly_sequence/hs37d5.fa.gz
NA12878 HiFi reads: http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/release/v1.0/assemblies/20200628_HHU_assembly-results_CCS_v12/haploid_reads
HG00733 HiFi reads: http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/working/20190925_PUR_PacBio_HiFi/

HG00733 ONT reads: http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/hgsv_sv_discovery/working/20181210_ONT_rebasecalled/
HG00733 phased assembly: http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/working/20200417_Marschall-Eichler_NBT_hap-assm/
Human GRCh38 reference: http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/GRCh38_reference_genome/
HG00733 Phased Assembly Variant calls: http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/working/20210806_PAV_VCF/
Phased Assembly Variant merged calls for 35 samples: http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/release/v2.0/integrated_callset/
RNA-seq data: Sequence-Read-Archive, Porject ID PRJNA720779.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	HG002 genome was used to benchmark simple SV detection. NA12878 genome was used assess complex SV detection from real sample. SVision was applied to HG00733 for novel complex SV discovery. RNA-Seq data of 19 samples (generated by Guennewig et al., PMID: 33649380) and haplotype-aware assembly of 35 samples (generated by Ebert et al., PMID: 33632895) were used to analyze the role of complex SV detected in HG00733.
Data exclusions	No data were excluded in this study.
Replication	Replication was not relevant to our study. This study used deterministic algorithms without statistical analysis, and this study aims to demonstrate SVision and its application with various long-read sequencing data.
Randomization	Randomization was not relevant to our study. SVision is a deterministic method, and all analysis in this study was done with preexisting data sources.
Blinding	Blinding was not relevant to our study. We used publicly available data, no data acquisition or statistical analysis was invovled.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging