nature biotechnology

Brief Communication

https://doi.org/10.1038/s41587-024-02190-7

De novo and somatic structural variant discovery with SVision-pro

Received: 1 August 2023

Accepted: 27 February 2024

Published online: 22 March 2024

Check for updates

Songbo Wang[1,2,3], Jiadong Lin[2,3], Peng Jia[1,2,3], Tun Xu[2,3], Xiujuan Li[2,3], Yuezhuangnan Liu[4], Dan Xu[4], Stephen J. Bush[2,3], Deyu Meng[3,5,6,7] & Kai Ye[1,2,3,4,8,9]

Long-read-based de novo and somatic structural variant (SV) discovery remains challenging, necessitating genomic comparison between samples. We developed SVision-pro, a neural-network-based instance segmentation framework that represents genome-to-genome-level sequencing differences visually and discovers SV comparatively between genomes without any prerequisite for inference models. SVision-pro outperforms state-of-the-art approaches, in particular, the resolving of complex SVs is improved, with low Mendelian error rates, high sensitivity of low-frequency SVs and reduced false-positive rates compared with SV merging approaches.

Long-read sequencing (LRS) technologies have greatly facilitated the detection of SVs[1], including simple SVs (SSV)[2–5] and complex SVs (CSVs)[6], which typically comprise several internal SSV subcomponents. Given that de novo and somatic SVs[7,8] are responsible for Mendelian disorders[9,10] and development of cancers[11,12], comparative SV discovery between genomes (for example, comparing a proband genome against parent genomes to identify de novo SVs) has generally been attempted by either callset-merge or read-inference strategies. Callset-merge strategies[13–15] (for example, Jasmine) extract genome-specific calls from merged callsets and hence inevitably incorporate the miscalls from callers, leading to many false positives. In contrast, read-inference strategies[16] (for example, nanomonsv) directly search differential alignments between genomes and construct SV inference models. However, this is typically limited to SSVs, and CSV modeling cannot be accommodated due to the unexplored CSV types and nested internal components[17]. Although sequencing-to-image and deep-learning-based callers have improved CSV characterization[6,18], two principal issues hinder their application to comparative SV discovery. First, existing sequencing-to-image schemas can represent SVs only of an individual genome, whereas comparative SV discovery requires additional image features that can represent SV differences between genomes. Second, comparative SV discovery demands several recognition tasks to detect and genotype SV between genomes simultaneously, while current single-task deep-learning callers classify one entire image into either a specific SV type[6,19] or genotype[20].

Here we propose SVision-pro, comprising two key modules: a sequence-to-image representation module encoding genomic features from two samples in a single image, from which a neural-network recognition module comparatively recognizes SVs as well as their intergenome differences. SVision-pro integrates SV detection and genotyping between genomes as a one-stop neural-network-based image instance segmentation task, facilitating the discovery of both de novo and somatic SSVs and CSVs.

The sequence-to-image representation module first takes as input aberrant genome loci identified from LRS data. In contrast to traditional LRS-based callers, which search for SV-specific alignment signatures, SVision-pro summarizes each read into a series of symbols (Extended Data Fig. 1a–d and Methods). These one-dimensional (1D)-symbol series are obtained directly from read alignment results without any SV-type-oriented preprocessing, and then clustered together iteratively as candidate aberrant loci (Extended Data Fig. 1e). This process, without matching known SV types, ensures the comprehensive capture of SV loci, especially for unexplored CSVs. The SV-type-classification task is delegated to subsequent representation and recognition modules.

[1]Department of Gynecology and Obstetrics, Center for Mathematical Medical, The First Affiliated Hospital of 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]MOE Key Lab for Intelligent Networks & Networks Security, Faculty of Electronic and Information Engineering, Xi’an Jiaotong University, Xi’an, China. [4]School of Life Science and Technology, Xi’an Jiaotong University, Xi’an, China. [5]School of Mathematics and Statistics, Xi’an Jiaotong University, Xi’an, China. [6]Macau Institute of Systems Engineering, Macau University of Science and Technology, Taipa, Macau. [7]Pazhou Laboratory (Huangpu), Guangzhou, Guangdong, China. [8]Faculty of Science, Leiden University, Leiden, The Netherlands. [9]Genome Institute, The First Affiliated Hospital of Xi’an Jiaotong University, Xi’an, China. e-mail: kaiye@xjtu.edu.cn

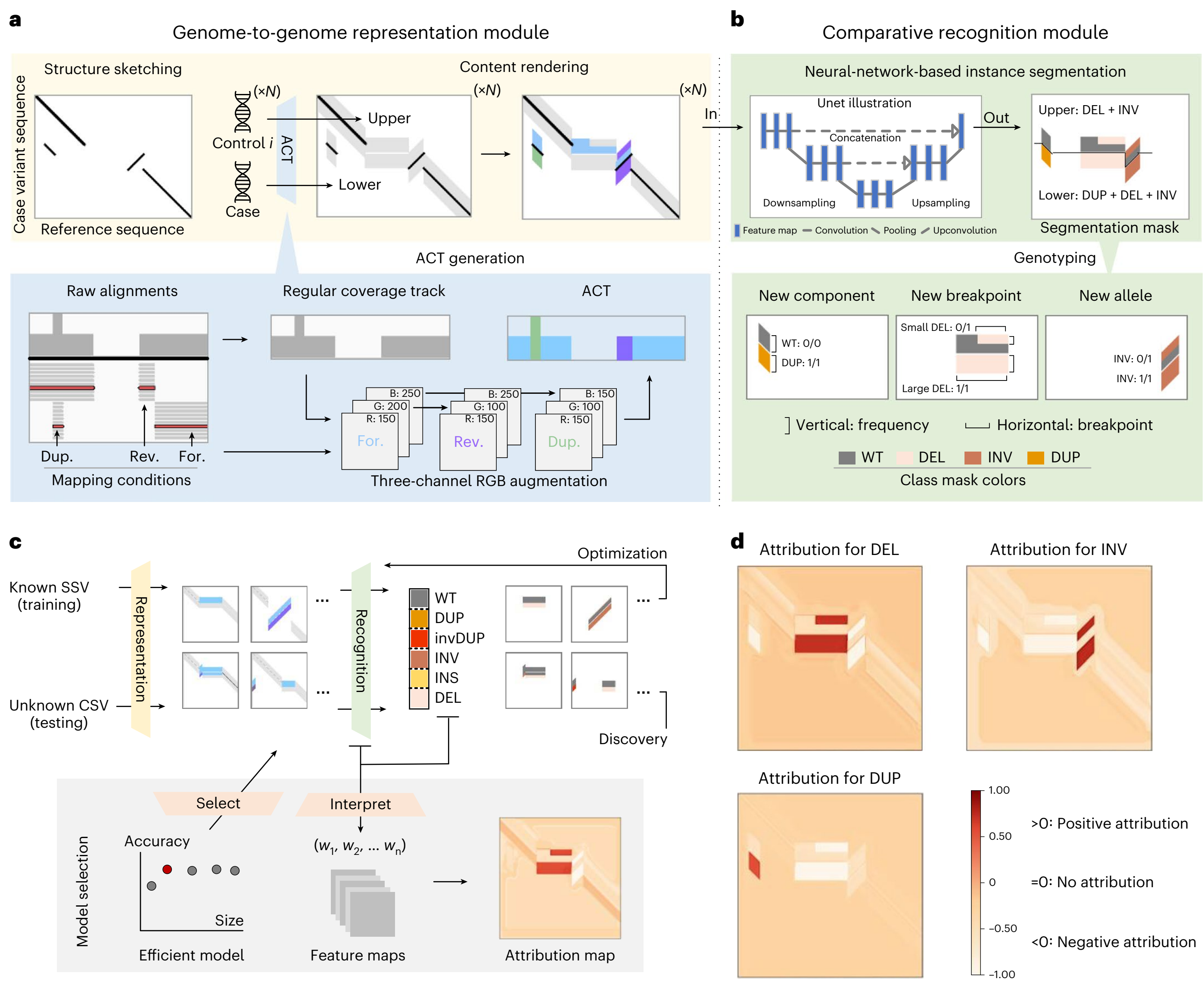


Fig. 1 | SVision-pro overview. **a**, Overview of the sequence-to-image representation module in SVision-pro. SVision-pro sketches the structures of a candidate SV locus and renders ACTs (above) into the sparse image regions. The ACT is generated from mapped alignments by the three-channel RGB augmentation (below). Dup., duplicated-matching; Rev., reversed-matching; For., forward-matching. **b**, Overview of the comparative recognition module in SVision-pro. The neural-network-based instance segmentation framework outputs a segmentation mask, providing intuitive SV types (above). By comparative genotyping analysis of the colored regions in the upper and lower panels (below), we can determine the SV differences between case and control genomes. **c**, Neural-network model training and selection strategy of SVision-pro. SVision-pro was trained with five basic SV subcomponent types along with wild type (identical to reference genome) and was able to recognize CSVs with several internal subcomponents (above). To select an efficient instance segmentation models (red solid circle), we leveraged three factors: validation accuracy, parameter size and interpretability. **d**, Attribution maps of the Lite-Unet model. Pixels relevant for a certain prediction class are highlighted. DEL, deletion; DUP, duplication; INV, inversion; INS, insertion; invDUP, inverted-duplication; WT, wild type; R, red; G, green; B, blue; w_1, w_2 and w_n, parameter weights.

The sequence-to-image representation module then compares two genomes (termed as case and control genome) in two steps (Fig. 1a): structure sketching and content rendering. For an aberrant locus in the case genome (for example, from child or tumor tissue), the structure sketching step directly transforms the 1D read symbol series into a two-dimensional (2D) similarity image (Extended Data Fig. 2a), which uses segments and gaps to measure the structural similarity of the reference sequence and the variant feature sequence from the case genome in an image (Extended Data Fig. 2b). The content rendering step (Methods) fills the sparse image regions with augmented coverage tracks (ACTs), which represent genomic differences between the case and control genome. First, we color the raw coverage track according to the forward-, inverted- and duplicated-matching conditions of alignments in three image channels (Fig. 1a and Extended Data Fig. 3a). Then, we use a fixed-height track above these structures (upper track) to encode the normalized ACT from the control genome (for example, from parent samples or normal tissue) while the track below (lower track) encodes ACT from the case genome (Extended Data Fig. 3b). This representation strategy facilitates genome-to-genome comparison, simultaneously encoding both SV structures (via segments and gaps) and their intergenome differences (via contrasting ACTs in lower and upper tracks), thereby requiring a multitask neural-network framework that can perform the detection and genotyping tasks simultaneously.

We integrated those many tasks into a one-stop neural-network-based image instance segmentation framework instead of utilizing several deep-learning classification modules (Fig. 1b and

Extended Data Fig. 4; Methods). Briefly, this framework takes in an encoded image and generates a pixel-level segmentation mask, classifying image areas in the upper and lower tracks into five basic SV component classes (Fig. 1b and Extended Data Fig. 4a), and one wild-type reference (REF). The other image regions, such as the flanking sequence encoding region, were classified as Background. SV types are predicted directly by joining components together in both the case and control tracks. Moreover, this instance segmentation framework enables a three-task comparison of SV component types, breakpoints and allele frequencies (AFs) between the case and control genomes (Fig. 1b). Specifically, for each SV component in the segmentation mask, the horizontal span of the masked pixels represents its breakpoint span, while the vertical span represents its AF (Extended Data Fig. 4b). Apart from the widely used genotyping tags (1/1, 0/1 and 0/0) derived from AF, SVision-pro generated four distinct categories by contrasting each SV component presented in the case genome with that of the control genome (Extended Data Fig. 4b; Methods). These categories are: (1) 'Germline,' indicating the presence of the SV subcomponent in the control genome with the same allele frequency as that of the case; (2) 'New component,' indicating the absence of the SV subcomponent in the control genome; (3) 'New breakpoint,' indicating the presence of the SV subcomponent in the control genome but with a different breakpoint span to the case and (4) 'New alleles,' indicating the presence of the SV subcomponent in the control genome but with a different AF to the case. In the scenarios for de novo SV discovery, SVision-pro will output the differences between the case genome and each control genome (Extended Data Fig. 4c). SVision-pro offers flexible image properties for different sensitivity requirements. Currently, SVision-pro enables a minimum detection AF of 0.01. Larger image sizes result in lower minimum representable and detectable AFs (Extended Data Fig. 4d; Methods).

To identify an appropriate instance segmentation model (Fig. 1c), five well-known models of different parameter sizes, including Unet[21], Fully-Convolutional-Network[22], Deeplab v.3 (ref. 23), Lite-Unet and mini-Unet were trained and compared on simulated data (Supplementary Note 1). The default model, Lite-Unet, achieved a balance between accuracy and model size (Extended Data Fig. 5a,b) while also exhibiting strong model interpretability (Fig. 1d and Extended Data Fig. 5c,d).

We benchmarked the performance of SVision-pro and other approaches using both simulated and publicly available datasets (Supplementary Table 1), covering high-fidelity (HiFi), Oxford nanopore (ONT) and continuous long reads (CLR). The computational resource usages were assessed on both a personal computer and a cluster node (Supplementary Note 2 and Supplementary Table 14).

SVision-pro outperformed other callers on HG002 groundtruth SSVs and simulated CSVs (Extended Data Fig. 6a,b and Supplementary Table 2; Methods). Moreover, SVision-pro achieved 96–98% accuracy in CSV subcomponent accuracy (Extended Data Fig. 6c and Supplementary Table 3; Methods), improving, on average, 15% compared with SVision—the state-of-the-art CSV caller. Further experimental validations (Supplementary Table 4, Supplementary File 1 and Supplementary Note 3) supported that SVision-pro has high sensitivity and a low false-positive rate for CSV detection.

We next compared SVision-pro with callset-merge strategies on six families, including a ChineseQuartet[24] (Methods). SVision-pro achieved the highest Mendelian consistency (97.3–98.4% on HiFi reads and 94.5%-97.6% on ONT reads) and the lowest discordancy (0.7%) between monozygotic twins (Fig. 2a and Supplementary Tables 5 and 6; Methods). When restricted to high-confidence regions (Methods), SVision-pro continued to outperform other approaches: the Mendelian consistency improved to 98.4–99.3% and 96.8–98.8% for HiFi and ONT, respectively, and the twin discordancy decreased to 0.3% (Supplementary Tables 5 and 6 and Extended Data Fig. 7). On a simulated trio harboring de novo/inherited CSVs (Supplementary Note 4), SVision-pro achieved 96.6% and 93.3% Mendelian genotype accuracy on HiFi and ONT long reads, respectively, while the second-best approach, SVision (followed by Jasmine merging), achieved 53.2% and 33.5% (Fig. 2b and Supplementary Table 7).

The high genotyping accuracy of SVision-pro led to reliable discoveries in Mendelian samples. For instance, a 32,549 bp deletion, encompassing the genes *LCE3B* and *LCL3C* and associated with increased risk of psoriasis[25,26], was incorrectly genotyped by Sniffles2 (ref. 15) yet was correctly genotyped by SVision-pro in the six families (Extended Data Fig. 8 and Supplementary File 2). Another complex locus, which was mis-called by all other approaches, comprised two SV alleles: an SSV (insertion) and an CSV (insertion–deletion) (Extended Data Fig. 9a–c). SVision-pro correctly genotyped these two alleles (Fig. 2c and Extended Data Fig. 9d), consistent with visual verification on HiFi reads and published assemblies (Supplementary File 3).

In the six families, SVision-pro reported 26 de novo SVs, including 13 insertions and 13 deletions (Supplementary Table 8), all of which were validated manually (Supplementary File 4). LRS enabled the discovery of a larger proportion of de novo insertions compared with SRS, and further annotation of the reported de novo SVs revealed that 20 of them featured repeat expansions or contractions (Supplementary Table 8). By contrast, Sniffles2 reported 90 whereas Jasmine/SURVIVOR reported many more redundant calls: 5,831–12,468 de novo SVs in total (Fig. 2d). We overlapped these 90 de novo calls of Sniffles2 with SVision-pro (Fig. 2e and Supplementary Table 9): among the 59 nonoverlapping calls, only one true-positive de novo SV was confirmed by manual inspection. Of the remaining 31 overlapped calls, 19 were identified as germline by both SVision-pro and manual curation (Supplementary File 5), indicating that they are false positives. Additional experimental validations (Supplementary Note 3, Supplementary Files 6 and 7 and Supplementary Table 10) further supported that SVision-pro effectively reduced false-positive calls in Mendelian samples and reported high-quality de novo SVs.

To assess the somatic detection performance, we simulated a subclonal tumor genome, which harbored somatic SSVs and CSVs with AFs ranges from 0.01 to 0.10 (Supplementary Note 4). For SSVs, the F1-scores of SVision-pro were 0.98 (HiFi) and 0.94 (ONT), leading the other two somatic-capable callers, Sniffles2 and nanomonsv[16], by 0.03 to 0.45 (Extended Data Fig. 10a). For CSVs, the F1-scores were 0.95 and 0.91. As expected, as the AF decreased, the detection accuracy exhibited a decreasing trend (Extended Data Fig. 10b). Nevertheless, for somatic SSVs and CSVs with AF = 0.01, SVision-pro still achieved average accuracies of 95.3% and 90.4% on HiFi and ONT reads (Supplementary Table 11). SVision-pro maintained consistent high-performance with various numbers of simulated events and coverages (Supplementary Table 12).

We next assessed SVision-pro using normal-tumor paired cell lines, HCC1395 and HCC1395BL, across three sequencing technologies, including HiFi, ONT and CLR (Methods). SVision-pro detected 87–90% of the published somatic SSV loci[27], while Sniffles2 detected 66–81% and nanomonsv detected 6–29% (Fig. 2f). Through computational validation using Vapor[28] on the detected somatic calls, SVision-pro demonstrated a much lower false-positive rate (4.3–8.7%; Fig. 2g, Supplementary Table 13 and Supplementary Note 5) compared with Sniffles2 (9.8–40.3%). Taken together, these results show that SVision-pro detects somatic SVs with higher sensitivity and lower false-positive rates compared with Sniffles2 and nanomonsv[16].

Moreover, SVision-pro resolved eight CSVs that were previously reported as SSVs (Supplementary File 8; Methods), including a dispersed duplication-deletion-inversion where the deletion component was missed and the dispersed duplication component was classified as a translocation (Extended Data Fig. 10c,d). SVision-pro also identified a nonsomatic complex locus, which was previously reported as a somatic SSV (Fig. 2h). SVision-pro revealed that the paired normal genome comprised one SSV allele and one CSV allele (deletion-inversion), whereas

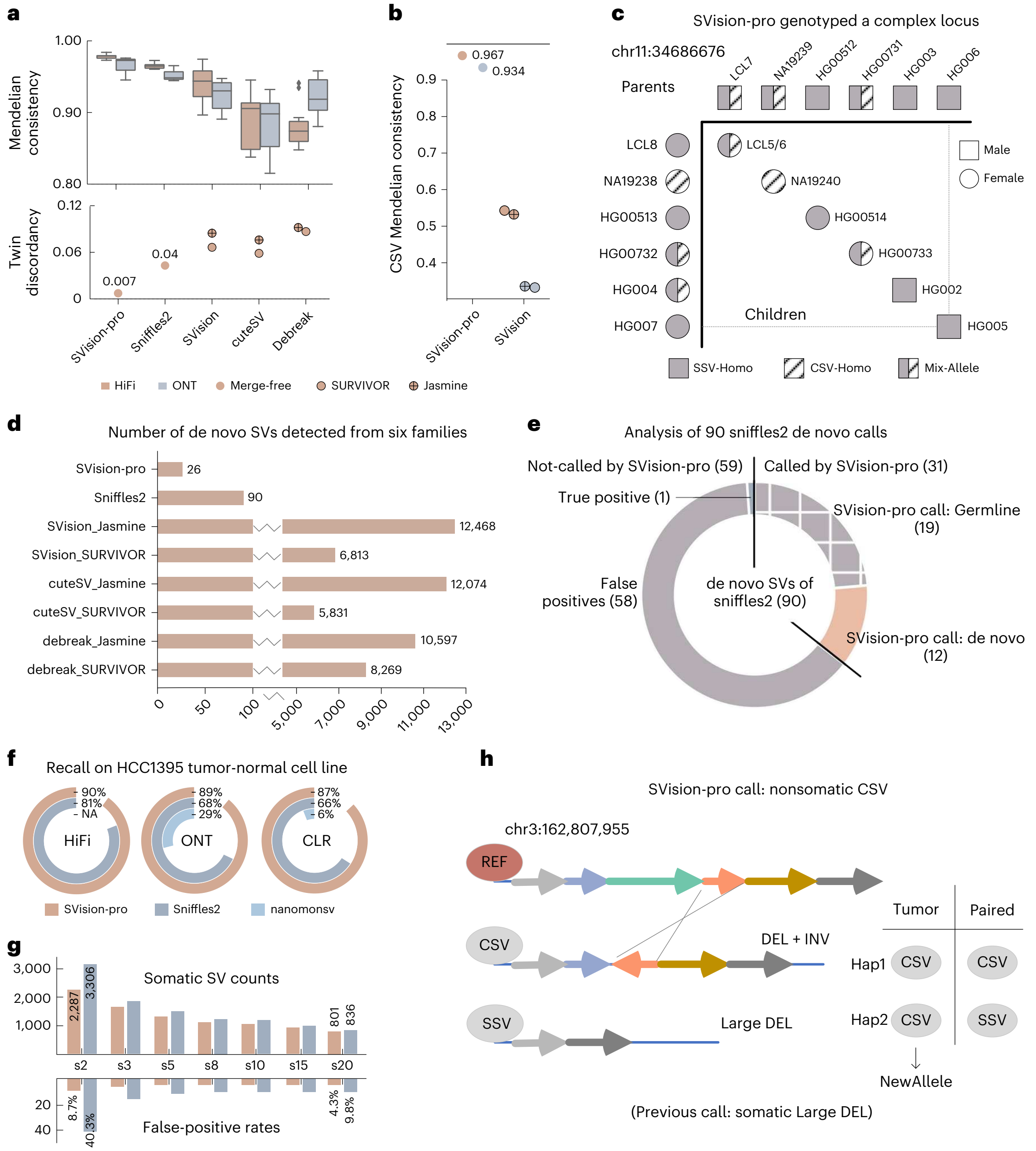


Fig. 2 | Performance comparison. a, Comparison of the Mendelian consistency in six family datasets (above) and the twin discordancy in the ChineseQuartet (below). SVision-pro is compared with Sniffles2 (multisample mode) and SVision, cuteSV and debreak (followed by SURVIVOR and Jasmine merging). Each box contains six and three values for HiFi and ONT, respectively (Supplementary Table 5). The boxplot defines the median (Q2, 50th percentile), first quartile (Q1, 25th percentile) and third quartile (Q3, 75th percentile). The bounds of the boxplot, representing interquartile range (IQR), are between Q1 and Q3. The minimum and maximum values are defined as Q1 − 1.5× IQR and Q3 + 1.5× IQR, respectively. The whiskers are values between minima and Q1 and between Q3 and maxima. Values falling outside the Q1–Q3 range are plotted as outliers of the data. **b**, Comparison of the CSV Mendelian genotype consistency on the simulated trio data. SVision-pro was compared with state-of-the-art CSV caller SVision (followed by SURVIVOR and Jasmine merge). **c**, In the six families, SVision-pro correctly genotyped a complex locus comprising both an SSV and a CSV. Three distinct alleles are found by SVision-pro, including homologous SSV, homologous CSV and mixed heterozygous SSV and CSV. **d**, Comparison of the number of de novo calls in the six family datasets. **e**, Overlapping of 90 de novo calls produced by Sniffles2 with all calls produced by SVision-pro. **f**, Recall values on the previously published somatic SV callset of HCC1395 tumor-normal paired cell lines. **g**, The number of somatic SVs and the false-positive rates produced by Vapor validation decrease as the supporting read number increases. **h**, SVision-pro identified a nonsomatic complex locus that had been reported as a somatic SSV. SVision-pro revealed that the paired normal genome exhibited a heterozygous SSV and CSV, whereas the tumor genome exhibited homozygous CSV.

the tumor genome lost the SSV allele and acquired a homozygous CSV (Extended Data Fig. 10e).

In summary, SVision-pro is an accurate and interpretable approach for comparative SV detection and genotyping, addressing the challenges in de novo and somatic SV discovery from long-read data. SVision-pro visually compares genomic features encoded from sequencing alignments, and so avoids the error-prone merging process intrinsic to a callset-level strategy, hence resulting in high-quality calls.

The instance segmentation framework removes the requirement for prebuilding inference models for SV types, thereby providing high CSV resolution. We conducted experimental validation for the findings of SVision-pro, in which certain events were deemed inconclusive due to PCR failure, characterized by the absence of notable PCR band or the presence of noisy PCR bands. This ambiguity raises the possibility that these events could be false positives, necessitating an orthogonal technique capable of validating SVs identified by LRS. Future work would develop merging- and model-free approaches for population-scale SV characterization to further improve discovery of the human SV spectrum.

Online content

Any methods, additional references, Nature Portfolio 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/s41587-024-02190-7.

References

1. Ebert, P. et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, eabf7117 (2021).
2. Sedlazeck, F. J. et al. Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).
3. Jiang, T. et al. Long-read-based human genomic structural variation detection with cuteSV. *Genome Biol.* **21**, 189 (2020).
4. Chen, Y. et al. Deciphering the exact breakpoints of structural variations using long sequencing reads with DeBreak. *Nat. Commun.* **14**, 283 (2023).
5. 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. *Nat. Methods* **20**, 550–558 (2023).
6. Lin, J. et al. SVision: a deep learning approach to resolve complex structural variants. *Nat. Methods* **19**, 1230–1233 (2022).
7. Koboldt, D. C. Best practices for variant calling in clinical sequencing. *Genome Med.* **12**, 91 (2020).
8. Li, Y. et al. Patterns of somatic structural variation in human cancer genomes. *Nature* **578**, 112–121 (2020).
9. Brandler, W. M. et al. Frequency and complexity of de novo structural mutation in autism. *Am. J. Hum. Genet.* **98**, 667–679 (2016).
10. Sanchis-Juan, A. et al. Complex structural variants in Mendelian disorders: identification and breakpoint resolution using short- and long-read genome sequencing. *Genome Med.* **10**, 95 (2018).
11. Aganezov, S. et al. Comprehensive analysis of structural variants in breast cancer genomes using single-molecule sequencing. *Genome Res.* **30**, 1258–1273 (2020).
12. van Belzen, I., Schonhuth, A., Kemmeren, P. & Hehir-Kwa, J. Y. Structural variant detection in cancer genomes: computational challenges and perspectives for precision oncology. *NPJ Precis. Oncol.* **5**, 15 (2021).
13. 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).
14. Kirsche, M. et al. Jasmine and Iris: population-scale structural variant comparison and analysis. *Nat. Methods* **20**, 408–417 (2023).
15. Smolka, M. et al. Detection of mosaic and population-level structural variants with Sniffles2. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-023-02024-y (2024).
16. Shiraishi, Y. et al. Precise characterization of somatic complex structural variations from tumor/control paired long-read sequencing data with nanomonsv. *Nucleic Acids Res.* **51**, e74 (2023).
17. Ho, S. S., Urban, A. E. & Mills, R. E. Structural variation in the sequencing era. *Nat. Rev. Genet.* **21**, 171–189 (2020).
18. Popic, V. et al. Cue: a deep-learning framework for structural variant discovery and genotyping. *Nat. Methods* **20**, 559–568 (2023).
19. Ma, H., Zhong, C., Chen, D., He, H. & Yang, F. cnnLSV: detecting structural variants by encoding long-read alignment information and convolutional neural network. *BMC Bioinf.* **24**, 119 (2023).
20. Poplin, R. et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983–987 (2018).
21. Ronneberger, O., Fischer, P. & Brox, T. U-Net: convolutional networks for biomedical image segmentation. Preprint at https://doi.org/10.48550/arXiv.1505.04597 (2015).
22. Long, J., Shelhamer, E. & Darrell, T. Fully convolutional networks for semantic segmentation. Preprint at https://doi.org/10.48550/arXiv.1411.4038 (2014).
23. Chen, L.-C., Papandreou, G., Schroff, F. & Adam, H. Rethinking atrous convolution for semantic image segmentation. Preprint at https://doi.org/10.48550/arXiv.1706.05587 (2017).
24. Jia, P. et al. Haplotype-resolved assemblies and variant benchmark of a Chinese Quartet. *Genome Biol.* **24**, 277 (2023).
25. de Cid, R. et al. Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis. *Nat. Genet.* **41**, 211–215 (2009).
26. Pajic, P., Lin, Y. L., Xu, D. & Gokcumen, O. The psoriasis-associated deletion of late cornified envelope genes LCE3B and LCE3C has been maintained under balancing selection since human Denisovan divergence. *BMC Evol. Biol.* **16**, 265 (2016).
27. Talsania, K. et al. Structural variant analysis of a cancer reference cell line sample using multiple sequencing technologies. *Genome Biol.* **23**, 255 (2022).
28. Zhao, X. F., Weber, A. M. & Mills, R. E. A recurrence based approach for validating structural variation using long-read sequencing technology. *Gigascience* **6**, 1–9 (2017).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

SVision-pro methodology

Overall workflow of SVision-pro. SVision-pro initiates by searching the case genome for candidate SV loci, after which a sequence-to-image module encodes genome-to-genome image to visually compare the case and control genomes. Then, the neural-network-based instance segmentation framework recognizes basic SV component types from the encoded image and determines the genomic differences between the case genome and the control genome. Note that, if several control genomes (*N* and *N* > 1) are specified, SVision-pro works in a 1-to-*N* mode and generates representation images for the case genome and each control genome. Consequently, the instance segmentation framework outputs the SV differences between the case genome and each control genome.

Candidate SV locus searching from case genome. SVision-pro identifies candidate SV loci by collecting and clustering abnormal read alignments in a model-free way that avoids searching for specific aberrant patterns of read alignments (Extended Data Fig. 1). Specifically, SVision-pro converts each read into a series of signature symbols, which can be extracted directly from a BAM file: M indicates directly mapping of alignment to the reference genome, V indicates reversed mapping and I indicates an additional sequence in read. Moreover, several properties are allocated to each signature symbol, including its span on the reference sequence, span on the read sequence, subsequence length and read name. Typically, symbols M and V are converted from split read alignments (primary and supplementary alignments) according to their reference span (reference start and end position) and mapping orientation. The symbol I is derived from both intraread alignments, by examining the CIGAR string, and inter-read alignments, by retrieving unmapped sequence between split alignments (Extended Data Fig. 1a). Note that for I, if the unmapped sequence is aligned to a distal location on the reference sequence, SVision-pro marks it as a mapped I by recoding the additional source reference span. Finally, each read is converted into a series of symbols arranged in their read order. For example, if a read does not span any SVs, there will be only one symbol M (Extended Data Fig. 1b). If a read spans a deletion, the read will be converted into symbol series MM, where there is a gap between the reference end position of the first M and the reference start position of the last M (Extended Data Fig. 1c). For complex events, such as a deletion associated with an inversion, the event-supporting read is converted into symbol series MVM (Extended Data Fig. 1d). By adopting this convention, we are able to cluster similar read symbol series iteratively and identify any abnormal ones (Extended Data Fig. 1e). A read with the converted symbol series M is considered a normal read, otherwise, it will be marked as an aberrant one. If the number of reads supporting the same aberrant symbol series surpasses the minimum requirement (default ten reads), the genomic region covered by the aberrant symbol series is considered a candidate SV locus.

Image representation at candidate SV loci. To generate representation images, SVision-pro takes two main steps: structure sketching (Extended Data Fig. 2) and content rendering (Extended Data Fig. 3).

(1) Structure sketching: for a candidate SV locus, the structure sketching step directly converts the 1D read symbol series into a 2D similarity image (Extended Data Fig. 2a), which uses segments and gaps to visually measure the mapping similarity between reference sequence (*x* axis) against variant feature sequence (*y* axis). The reference axis ranges from the start reference position of the first symbol to the end reference position of the last symbol. The read axis ranges from 0 to the length of the read. Typically, segments are derived from symbols M, V and mapped I, whereas gaps are derived from the unmapped symbol I and reference gaps between M and V symbols. Segments and gaps, excluding those converted from M symbols, are marked with aberrant flags for subsequent content rendering step (Extended Data Fig. 2b). This type of similarity image makes it easy for humans and machines to visualize SV structures.

(2) Content rendering: SVision-pro fills the sparse region in the similarity image with ACTs originated from both case and control genomes.

Generating ACTs. Inspired by the regular coverage track commonly used in Integrative Genomics Viewer (IGV)[29], SVision-pro introduces the ACT. In brief, the regular coverage track is a 2D grayscale barplot, where the *x* axis indicates reference positions and *y* axis indicates the coverage values, which are computed by counting the number of mapped alignments at each reference position (Extended Data Fig. 3a). The ACT in SVision-pro utilizes an RGB (red, green and blue) stacked barplot to encode additional genomic information that reflects SV signatures. Before constructing the ACT (Extended Data Fig. 3a), we count the number of alignments along with their mapping conditions. The mapping conditions of alignments include forward mapping, reversed mapping, duplicated mapping and reverse-duplicated mapping. Forward and reversed mapping conditions are retrieved directly from the aligner's outputs and duplicated mapping is determined by checking whether an alignment is encompassed by other alignments from the same read (Extended Data Fig. 3a).

Next, we convert the count table into a three-channel RGB image. We use the RGB color values (135, 206, 255) to plot the coverage value of forward-mapped alignments. For the coverage value of reversed alignments, we subtract 100 from the color value in the second channel (Supplementary Fig. 1a). Likewise, for the coverage value of duplicated alignments, we subtract 100 from the color value in the third channel (Supplementary Fig. 1b). In cases of reverse-duplicated alignments, both the second and third channels undergo a subtraction of 100 (Supplementary Fig. 1c). In brief, we use the second image channel to depict the reverse signatures and the third image channel to depict the duplication signatures. By leveraging this RGB stacked barplot in the ACT, SVision-pro provides a more comprehensive representation of the coverage information, incorporating distinct color variations to depict different types of alignments and their contribution to the SV signature.

Filling ACTs into similarity image. Genome-to-genome comparison requires comparative representation features to contrast the SV differences between the case genome and the control genome. Therefore, we utilize the sparse regions within the similarity image to fill the two ACTs originating from the case and control genomes (Extended Data Fig. 3b). To accomplish this, we first create two fixed-height and empty tracks along these sketched segments and gaps: one track (upper track) above and one track below (lower track). The upper track is used to fill the ACT of the control genome whereas the lower track is used to fill the ACT of the case genome. For a sketched similarity image i, we generate ACTs in both case and control genomes by fetching all read alignments from i.reference_start to i.reference_end. This ensures that the reference span of the sketched similarity image matches that of the ACTs. Next, we fill ACTs into upper/lower tracks that surround aberrant segments and gaps by aligning the reference coordinates. Contrasting ACTs in upper and lower tracks show apparent SV differences between the case and control genomes. Moreover, this kind of similarity image and ACTs maintains readability for both human and machines for further analysis.

Insertion-associated SV representation. Insertions and insertion-related SVs involve additional sequence present in the read sequence that is not in the reference sequence, leading to vertical gaps in the sketched similarity images (Supplementary Fig. 2a). Therefore, for insertions, we create two empty tracks located on the left (used to fill the ACT of the control genome) and right (used to fill the ACT of the case genome) sides of these insertion-induced vertical gaps (Supplementary Fig. 2b). Unlike deletions, inversion and duplications, where we count the alignment mapping conditions against the reference genome, for insertions, we count the alignments at read-level to calculate the number of reads that contain the inserted sequence (Supplementary Fig. 2c). Then, we generate vertical ACTs for both case genome and control genome and fill them into the right and left empty tracks, respectively. For insertion-associated CSVs, such as insertion-associated inversion, alignments are counted at both read-level and reference-level (Supplementary Fig. 2d).

One-to-*N* mode. The genome-to-genome representation module in SVision-pro allows for the comparison of one case genome with one control genome within a single image. However, in certain applications, such as de novo SV discovery, several control genomes are involved. To accommodate such scenarios, SVision-pro employs a One-to-*N* mode to generate images between case genome and each control genome. For example, de novo SV discovery in a trio comprises three genomes: child, father and mother. For a candidate SV locus, SVision-pro generates one image that compares the child genome with the father genome, and another that compares the child genome with the mother genome. This process results in two images that can be utilized by the subsequent instance segmentation framework for further analysis. By employing the One-to-*N* mode, SVision-pro enables direct comparison of the case genome with several control genomes. Moreover, SVision-pro can identify any genome-specific SVs among several genomes by taking one genome as the case genome and all others as control genomes.

Flexible properties of representation image. The image sizes, colors and track heights are flexible and can be customized to meet various application scenarios. Currently, SVision-pro offers three optional image sizes for different sensitivity requirements, including 256, 512 and 1,024, whose track height for rendering contents is 25, 50 and 100 pixels, respectively. Thereby, the minimum representable (1 pixel) and detectable AFs (one per track height) of the three image sizes are 0.04, 0.02 and 0.01, respectively. Note that AF 0.01 is not the lowest detection limit of SVision-pro, and that the track heights and images sizes can be customized to meet lower AF detection requirements.

SV detection and genotyping by instance segmentation. The encoded representation images are directly fed into a neural-network-based instance segmentation framework without any manual or knowledge-oriented preprocessing. Since CSVs typically comprise several internal subcomponents, the instance segmentation framework in SVision-pro is designed to recognize five basic subcomponent types, including insertion (INS), deletion (DEL), inversion (INV), duplication (DUP) and inverted duplication (invDUP). In cases where there is no SV present in the control genome, a recognition type reference (REF) is included to denote that the control genome is identical to the reference genome. Specifically, the instance segmentation framework recognizes these six instance types in the encoded image and generates a segmentation mask. The mask assigns each pixel in the image to either a predicted specific type or the background type, segmenting the image regions and providing quantitative information about the presence and location of various SV subcomponents (Extended Data Fig. 4a). The horizontal span of the masked regions represents the breakpoint span of the subcomponents, while the vertical span represents the allele frequency (Extended Data Fig. 4b). Finally, in respective panels, we obtain the final SV type of the candidate locus by directly jointing together these subcomponents in their read order. By contrasting the lower and upper panels in the segmentation mask image, SVision-pro can determine whether a SV subcomponent is (Extended Data Fig. 4b) Germline, indicating that the SV subcomponent is present in the control genome with same allele frequency; (2) New allele, indicating that the SV subcomponent is present in the control genome at a different allele frequency; (3) New component, indicating that the SV subcomponent is absent from the control genome or (4) New breakpoint, indicating that the SV subcomponent is present in the control genome with a different breakpoint span. If several control genomes are provided, such as the father and mother genome in the scenarios for de novo SV discovery, SVision-pro will output the differences between the case genome and each control genome (Extended Data Fig. 4c).

Performance benchmarking methodology

SSV detection benchmark in HG002 groundtruth. The groundtruth SSVs (HG002_SVs_Tier1_v0.6.vcf.gz, highly confident insertions and deletions) of HG002 (Ashkenazim Trio, son), were applied to benchmark the SSV detection performance of callers. The detailed data generation steps were identical to those described in cuteSV[3] paper. Briefly, both raw HiFi and ONT reads were aligned to human genome GRCh37 using Minimap2 (ref. 30) with parameter '-x pacbio/ont'. Seven state-of-the-art callers, including SVision-pro, SVision[6], Sniffles2 (ref. 15), cuteSV[3], debreak[4], pbsv and SVDSS[5], were applied to the aligned reads with the minimum SV supporting read number set to ten. Truvari[31] was employed to calculate precision, recall and F1-score between the groundtruth and the callset. Please refer to Supplementary Note 6 for the specific versions and parameters of each caller.

CSV detection benchmark in simulated data. The CSV simulation set, which contains 3,000 CSVs crossing ten frequently reported types, was obtained directly from our previous SVision paper[6]. We followed the same procedure described in this paper to generate both HiFi and ONT reads and performed subsequent alignment to GRCh38 by NGMLR[2]. The five highest-performing callers on the HG002 groundtruth dataset (SVision-pro, SVision, Sniffles2, cuteSV and debreak) were employed for the subsequent Truvari region-based comparison. Type-based comparison was performed by examining the CSV subcomponent accuracy. To accomplish this (Supplementary Fig. 3a), we first extracted the matched SV record pairs between the groundtruth and callset from Truvari output files, namely TP-base.vcf and TP-call.vcf, which respectively enumerated the groundtruth record and matched callset record, respectively. Then, for each matched record pair, if any SV component from the groundtruth record was absent from the called record, this record pair was marked as inaccurate (Supplementary Fig. 3b). Note that, only SVision-pro and SVision reported SV component types. For the remaining callers, since they only reported SSVs and limited number of CSV types, we treated their output type directly as a component type.

Mendelian consistency analysis in six families. We collected 19 Mendelian samples from six previously published families, including the Ashkenazim Trio, Chinese Trio, YRI Trio, CHS Trio, PUR Trio and Chinese Quartet (Supplementary Table 1). All six families were sequenced using HiFi reads, with the Ashkenazim Trio, Chinese Trio and Chinese Quartet also sequenced with ONT reads. All reads were aligned to GRCh38 genome using Minimap2. We utilized five callers, including SVision-pro, SVision, Sniffles2, cuteSV and debreak, and two merging approaches, including Jasmine and SURVIVOR. For SVision-pro, we considered the child sample as the case genome and parent samples as control genomes. Sniffles2 was employed in multisample calling mode, following official instructions. For the remaining three callers that required merging approaches, we first applied them independently to generate callsets for each sample, including child(ren), father and mother. Then, we merged these callsets (for example, for ChineseQuartet, there were four callsets) together by Jasmine and SURVIVOR with the default or recommended parameters (Supplementary Note 2). To measure the Mendelian consistency within each family, we extracted the child and parent genotypes from each SV record in the VCF. If the genotypes of child, father and mother adhered to the Mendelian Law, we marked this record as a consistent one. Finally, we computed the Mendelian consistency rate by dividing the number of consistent records by the total number of records.

Twin discordancy analysis in Chinese Quartet. A common assumption is that the genomes of monozygotic twins are almost identical[32]. Therefore, the monozygotic twins (termed as child1 and child2) in the Chinese Quartet were used to calculate the twin discordancy. In brief, if one SV was present in the child1 genome while absent from the child2 genome, we would consider this SV as a discordant one between the twins. As such, for each SV record, we extracted the outputted genotypes of both child1 and child2 and examined whether they were identical. Finally, we computed the twin discordancy by dividing the number of discordant records by the total number of records.

De novo SV analysis in six families. For SVision-pro, de novo SVs were extracted by checking whether the comparison results of child-to-father and child-to-mother were both 'New Component.' For Sniffles2 and the merging approaches, de novo SV records were extracted by checking whether the SUPP_VEC equaled 100, indicating this SV record presented only in the child genome. Moreover, we compared the de novo SVs between SVision-pro and Sniffles2. De novo SV calls from Sniffles2 were overlapped with all SV calls from SVision-pro using the BEDtools[33] intersect option with reciprocal overlap fraction set to 0.5. Since merging approaches resulted in many more redundant de novo SVs, we verified manually only the de novo SVs called by SVision-pro and Sniffles2 using IGV[29] (Supplementary Files 4 and 5).

Somatic SV analysis in tumor-normal paired cell line HCC1395. A previous study[27] utilized several sequence technologies and established a consensus somatic SV callset of 1,788 SVs on cell line HCC1395 and its normal pair HCC1395BL. We download the published HiFi, ONT and PacBio CLR long reads of the two cell lines and aligned them to human genome GRCh38 by Minimap2 with parameter '-x pacbio.' Three callers that could detect somatic SVs were employed on this tumor-normal paired cell line, including SVision-pro, Sniffles2 and nanomonsv. SVision-pro took the tumor cell line as the case genome and normal cell line as the control genome. Sniffles2 was employed in its nongermline mode and nanomonsv was employed according to official instructions. For the three callers, the minimum number of supporting reads was set to 2 and the minimum detectable AF was set to 0.01.

High-confidence region filter. The raw high-confidence regions (HG002_SVs_Tier1_v0.6.bed) were hg19-based. Therefore, following the instruction of SVDSS paper[5], we first used liftOver to convert these regions into hg38-based coordinates. Then we applied BEDtools intersect option with reciprocal overlap fraction set to 0.5 to filter out SV calls that were not located within high-confidence regions.

Reporting summary

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

Data availability

The sources of HiFi, ONT and CLR reads of the six family datasets and HCC1395 normal-tumor paired cell are listed in Supplementary Table 1. The human reference genome GRCh37 was downloaded from http://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/.

Code availability

SVision-pro (v.1.6) is available at GitHub (https://github.com/songbowang125/SVision-pro.git)[34]. The scripts for model training, performance valuation and simulate data generation are available at GitHub (https://github.com/songbowang125/SVision-pro-Utils.git)[35]. Both repositories are available under a GNU General Public License v.3.0, and are free for noncommercial use by academic, government and nonprofit/not-for-profit institutions.

References

29. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
30. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
31. English, A. C., Menon, V. K., Gibbs, R. A., Metcalf, G. A. & Sedlazeck, F. J. Truvari: refined structural variant comparison preserves allelic diversity. *Genome Biol.* **23**, 271 (2022).
32. van Dongen, J., Slagboom, P. E., Draisma, H. H., Martin, N. G. & Boomsma, D. I. The continuing value of twin studies in the omics era. *Nat. Rev. Genet.* **13**, 640–653 (2012).
33. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
34. Wang, S. songbowang125/SVision-pro: SVision-pro. *GitHub* https://github.com/songbowang125/SVision-pro.git (2023).
35. Wang, S. songbowang125/SVision-pro-Utils: SVision-pro. *GitHub* https://github.com/songbowang125/SVision-pro-Utils.git (2023).
36. Krumsiek, J., Arnold, R. & Rattei, T. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics* **23**, 1026–1028 (2007).

Acknowledgements

We thank the L. Shi laboratory for providing the same batch of DNAs used for sequencing. K.Y. is supported by National Science Foundation of China (grant nos. 32125009 and 32070663) and the National Key R&D Program of China (grant no. 2022YFC3400300). D.X. is supported by National Science Foundation of China (grant nos. 32070134 and 32270188). D.M. is supported by National Science Foundation of China (grant nos. 12226004 and 62272375). J.L. is supported by National Science Foundation of China (grant no. 62302386).

Author contributions

K.Y. designed and supervised the research. S.W. developed the SVision-pro algorithm and performed the performance evaluation. D.M. contributed to the assessment and analysis of the deep-learning model. P.J. and T.X. contributed to the sequencing data processing. D.X. designed the experimental validation. X.L. and Y.L. performed the experimental validation. S.W., J.L., S.J.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

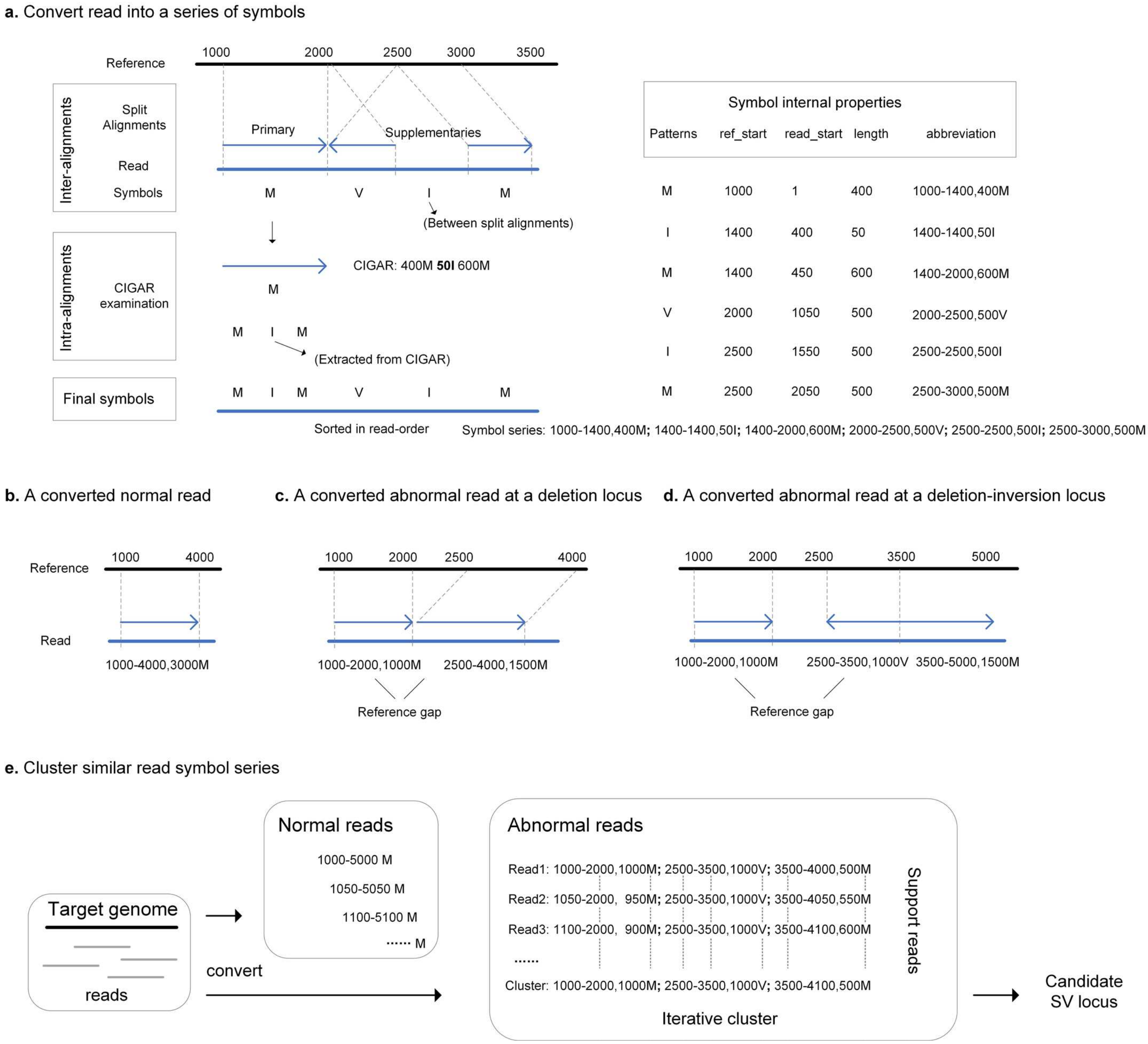
Extended data is available for this paper at https://doi.org/10.1038/s41587-024-02190-7.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-024-02190-7.

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

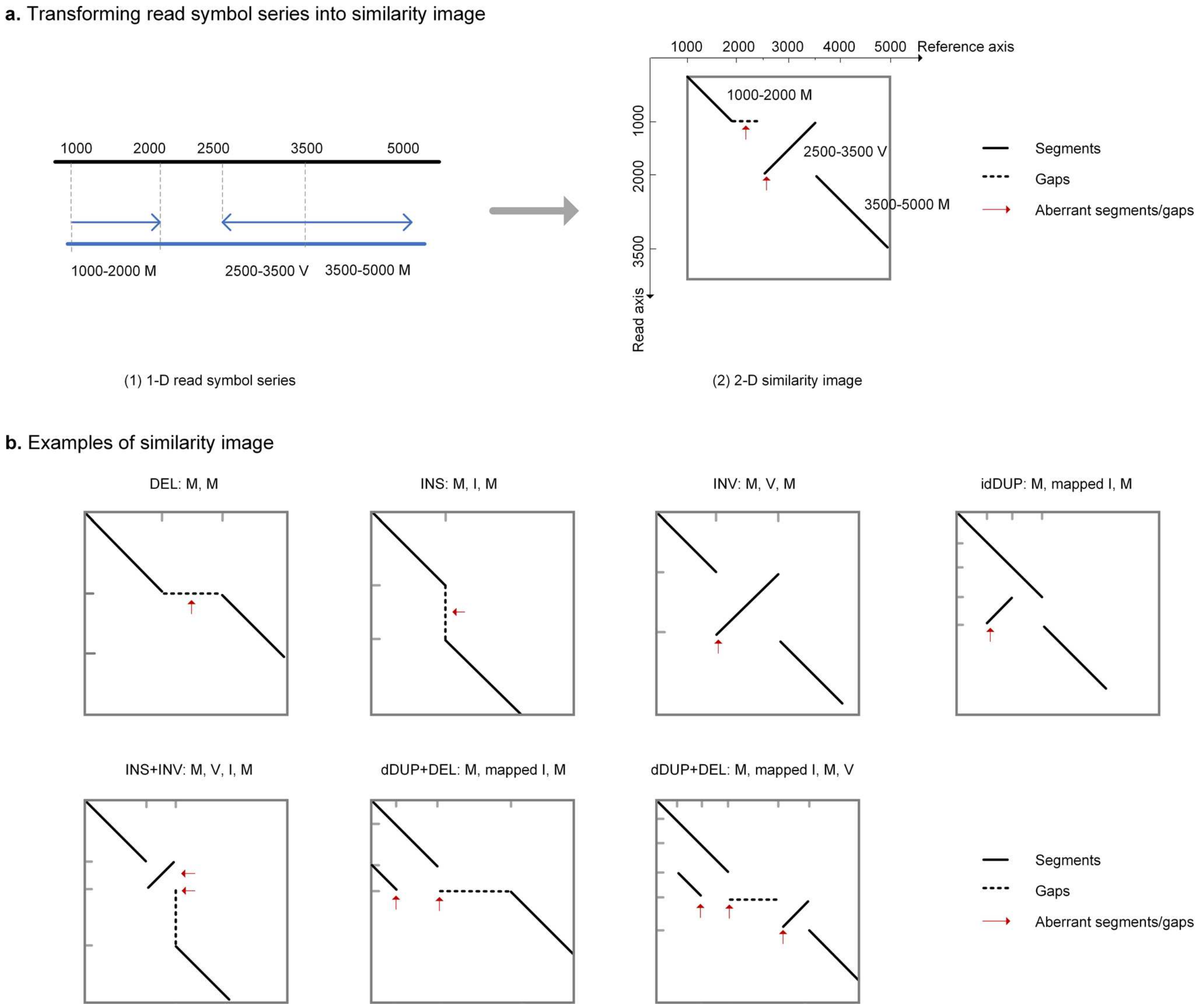
Peer review information *Nature Biotechnology* thanks the anonymous reviewers for their contribution to the peer review of this work.

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

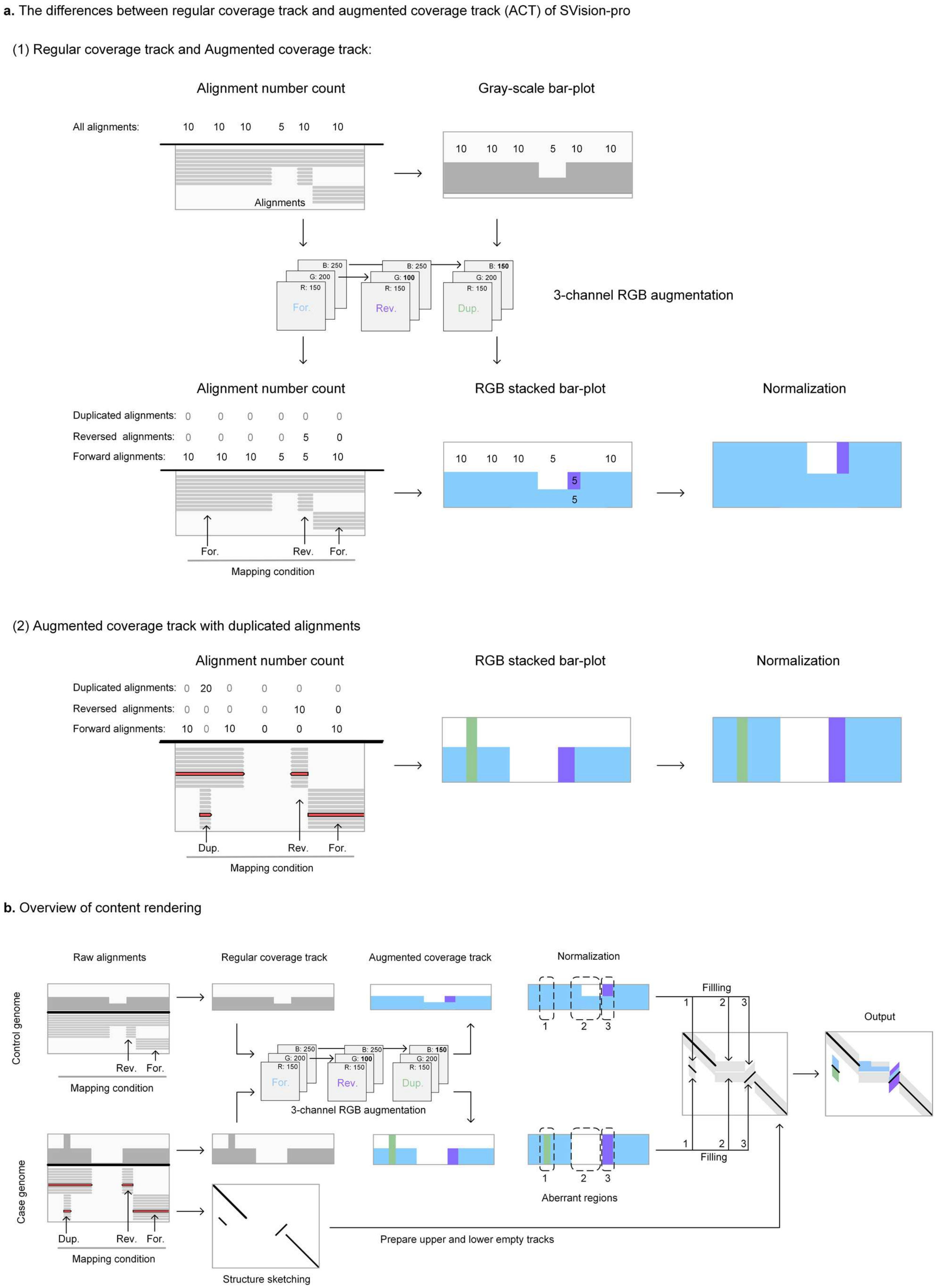

a. Convert read into a series of symbols
Reference
1000 2000 2500 3000 3500
Inter-alignments
Split Alignments
Primary
Supplementaries
Read
Symbols
M V I M
(Between split alignments)
Intra-alignments
CIGAR examination
CIGAR: 400M 50I 600M
M
M I M
(Extracted from CIGAR)
Final symbols
M I M V I M
Sorted in read-order
Symbol internal properties
Patterns ref_start read_start length abbreviation
M 1000 1 400 1000-1400,400M
I 1400 400 50 1400-1400,50I
M 1400 450 600 1400-2000,600M
V 2000 1050 500 2000-2500,500V
I 2500 1550 500 2500-2500,500I
M 2500 2050 500 2500-3000,500M
Symbol series: 1000-1400,400M; 1400-1400,50I; 1400-2000,600M; 2000-2500,500V; 2500-2500,500I; 2500-3000,500M
b. A converted normal read
Reference
1000 4000
Read
1000-4000,3000M
c. A converted abnormal read at a deletion locus
1000 2000 2500 4000
1000-2000,1000M 2500-4000,1500M
Reference gap
d. A converted abnormal read at a deletion-inversion locus
1000 2000 2500 3500 5000
1000-2000,1000M 2500-3500,1000V 3500-5000,1500M
Reference gap
e. Cluster similar read symbol series
Target genome
reads
convert
Normal reads
1000-5000 M
1050-5050 M
1100-5100 M
...... M
Abnormal reads
Read1: 1000-2000,1000M; 2500-3500,1000V; 3500-4000,500M
Read2: 1050-2000, 950M; 2500-3500,1000V; 3500-4050,550M
Read3: 1100-2000, 900M; 2500-3500,1000V; 3500-4100,600M
......
Cluster: 1000-2000,1000M; 2500-3500,1000V; 3500-4100,500M
Support reads
Iterative cluster
Candidate SV locus

Extended Data Fig. 1 | Illustration of the candidate SV locus searching step in SVision-pro. a, SVision-pro converts each read into a series of symbols, including 'M', 'V' and 'I', based on the aligner's output. Staring with inter-alignment examination, primary alignment and supplementary alignments of the read are directly converted into 'M' and 'V' according to their mapping orientation. Unmapped sequence between split alignments are converted into 'I'. For each alignments, SVision-pro further examine their CIGAR string (intra-alignment) to retrieve more 'I's. Consequently, a read is converted into a series of symbols arranged in their occurrence on read sequence. Each symbol contains several inner properties, including start position on reference sequence, start position on read sequence and its length. Each symbol can be abbreviated as 'reference_start-reference_end, length and symbol type' for subsequent clustering step. **b**, An example of converting a normal read into a symbol series. **c**, An example of converting an abnormal read, which spans a deletion, into a symbol series. **d**, An example of converting an abnormal read, which spans a CSV deletion-inversion, into a symbol series. **e**, For a genome locus, normal reads, which contain only one 'M' in their symbol series, are filtered out. The remaining abnormal reads are iteratively clustered together by comparing their symbol series to identify candidate SV loci.

https://doi.org/10.1038/s41587-024-02190-7



Extended Data Fig. 2 | Illustration of the structure sketching step in SVision-pro. a, SVision-pro directly transforms the 1-dimensional symbol series into a 2-dimensional similarly image, which utilizes segments and gaps to sketch the structure of the SV. Segments, derived from symbol 'M' and 'V', are represented in solid lines while gaps, derived from symbol 'I', are represented in dash lines. Gaps along with segments converted from symbol 'V' are mark with an aberrant flag (red arrows) for subsequence process. **b**, Several examples for transforming symbol series that span SSVs or CSVs, into similarity images.



a. The differences between regular coverage track and augmented coverage track (ACT) of SVision-pro

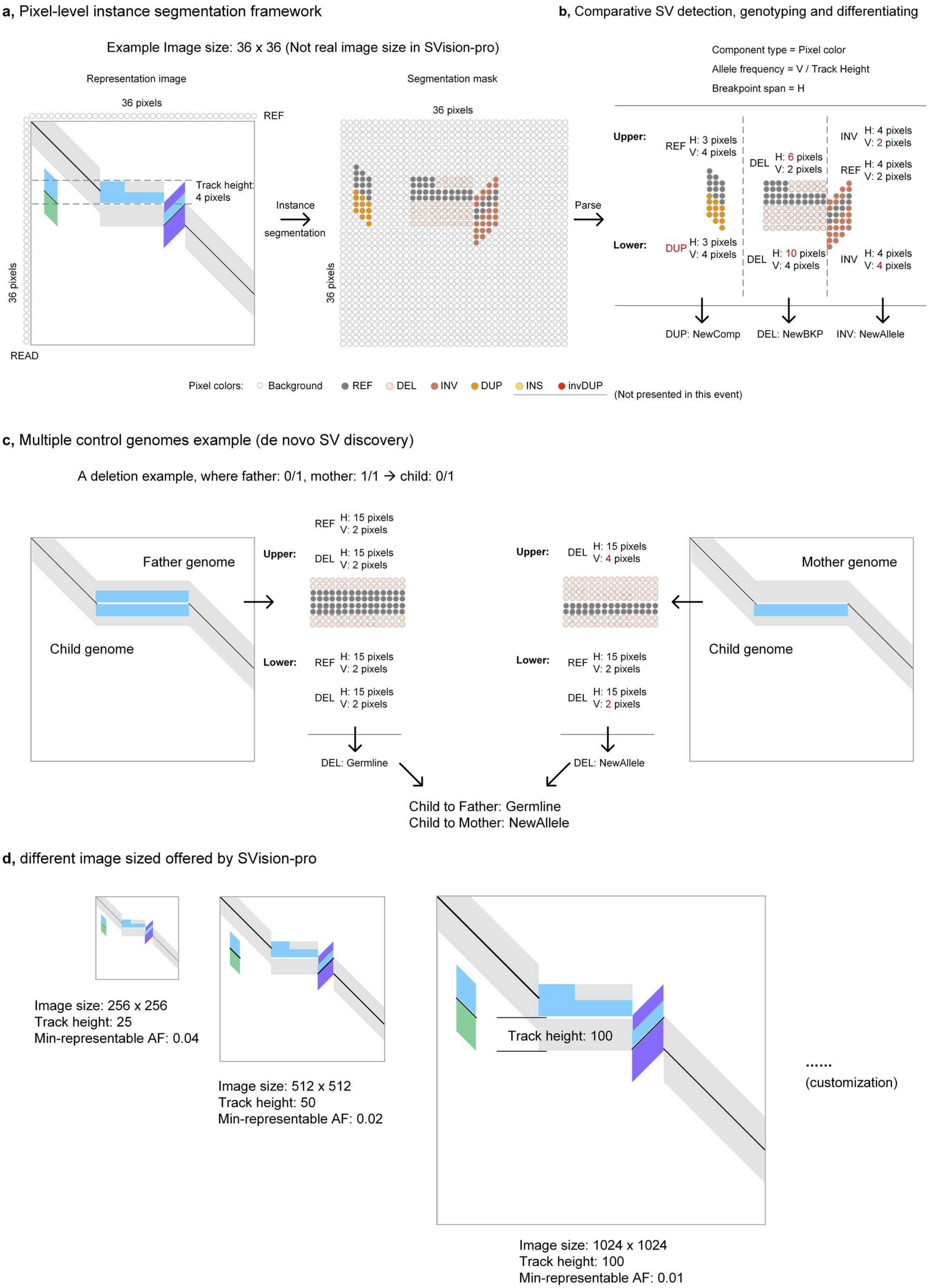
(1) Regular coverage track and Augmented coverage track:

(2) Augmented coverage track with duplicated alignments

b. Overview of content rendering

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

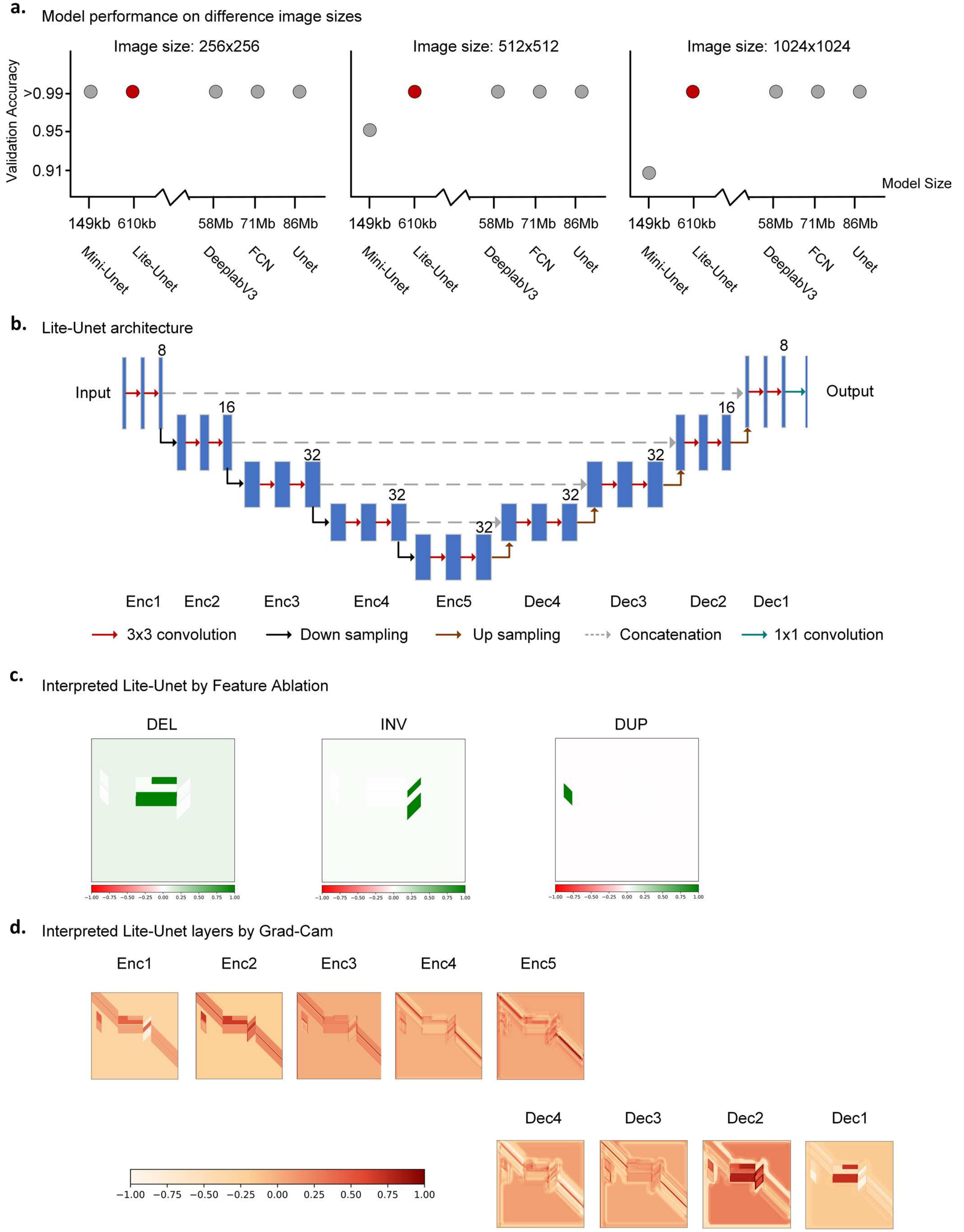
Extended Data Fig. 3 | Illustration of the content rendering step in SVision-pro. a, Comparison of regular coverage track and the augmented coverage track (ACT) in SVision-pro. The ACTs are generated by 3-channel RGB augmentation. SVision-pro counts read alignments according to their mapping conditions and generates a RGB stacked bar-plot, where different mapping conditions are represented in their respective RGB colors. **b**, Overview of the content rendering step. For both control and case genomes, the ACTs are generated, normalized, and further filled into the upper/lower tracks around aberrant segments and gaps in the similarity. Abbreviations: 'Dup.' denotes duplicated mapping; 'Rev.' denotes reversed mapping; 'For.' denotes forward mapping.


a, Pixel-level instance segmentation framework
Example Image size: 36 x 36 (Not real image size in SVision-pro)
Representation image
36 pixels
REF
Track height: 4 pixels
36 pixels
READ
Instance segmentation
Segmentation mask
36 pixels
36 pixels
Parse
Pixel colors: ○ Background ● REF ● DEL ● INV ● DUP ● INS ● invDUP (Not presented in this event)
b, Comparative SV detection, genotyping and differentiating
Component type = Pixel color
Allele frequency = V / Track Height
Breakpoint span = H
Upper:
REF H: 3 pixels V: 4 pixels
DEL H: 6 pixels V: 2 pixels
INV H: 4 pixels V: 2 pixels
REF H: 4 pixels V: 2 pixels
Lower:
DUP H: 3 pixels V: 4 pixels
DEL H: 10 pixels V: 4 pixels
INV H: 4 pixels V: 4 pixels
DUP: NewComp
DEL: NewBKP
INV: NewAllele
c, Multiple control genomes example (de novo SV discovery)
A deletion example, where father: 0/1, mother: 1/1 → child: 0/1
Father genome
Child genome
Upper:
REF H: 15 pixels V: 2 pixels
DEL H: 15 pixels V: 2 pixels
Lower:
REF H: 15 pixels V: 2 pixels
DEL H: 15 pixels V: 2 pixels
DEL: Germline
Upper:
DEL H: 15 pixels V: 4 pixels
Lower:
REF H: 15 pixels V: 2 pixels
DEL H: 15 pixels V: 2 pixels
DEL: NewAllele
Mother genome
Child genome
Child to Father: Germline
Child to Mother: NewAllele
d, different image sized offered by SVision-pro
Image size: 256 x 256
Track height: 25
Min-representable AF: 0.04
Image size: 512 x 512
Track height: 50
Min-representable AF: 0.02
Track height: 100
Image size: 1024 x 1024
Track height: 100
Min-representable AF: 0.01
......
(customization)

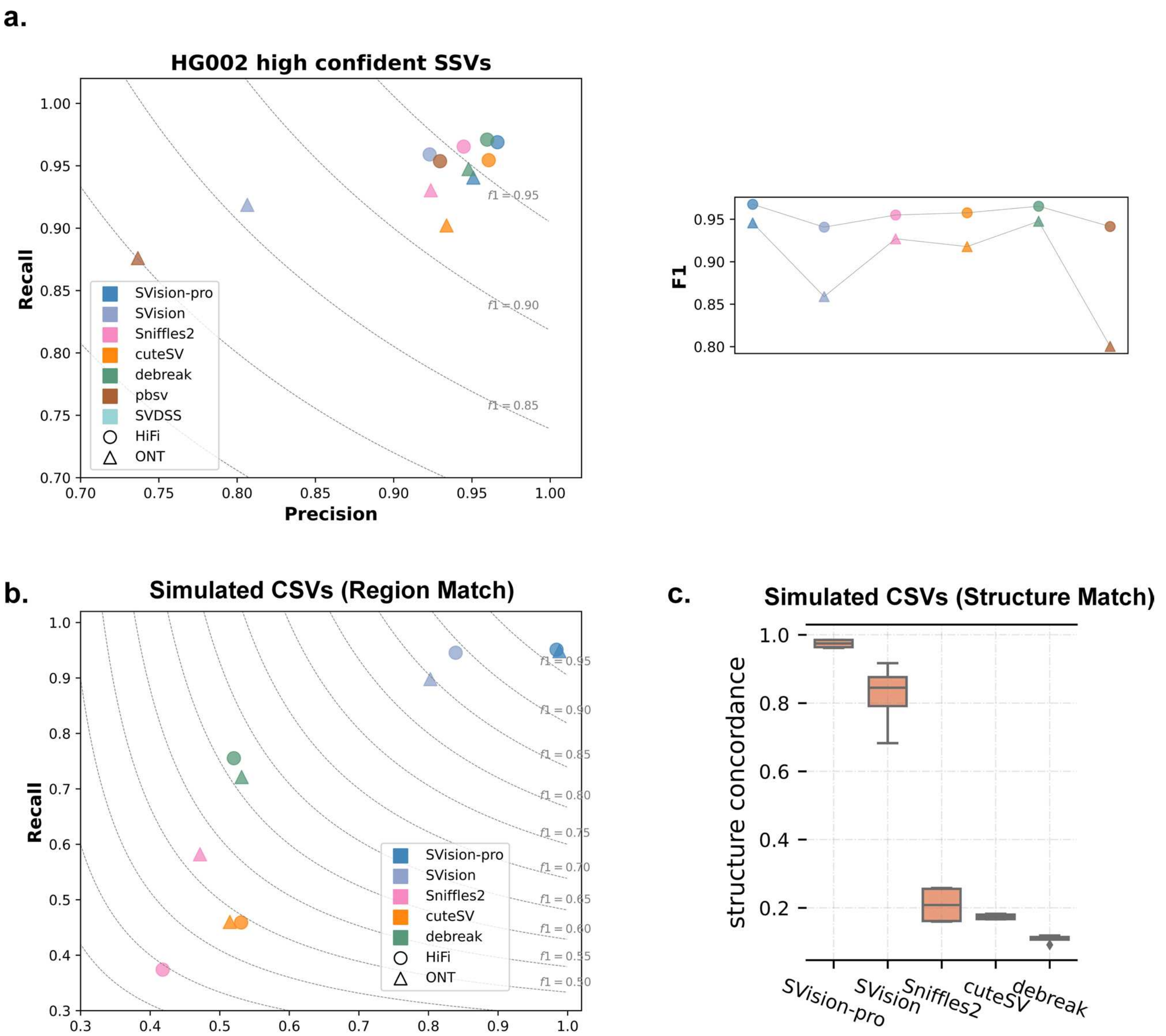
Extended Data Fig. 4 | See next page for caption.

https://doi.org/10.1038/s41587-024-02190-7

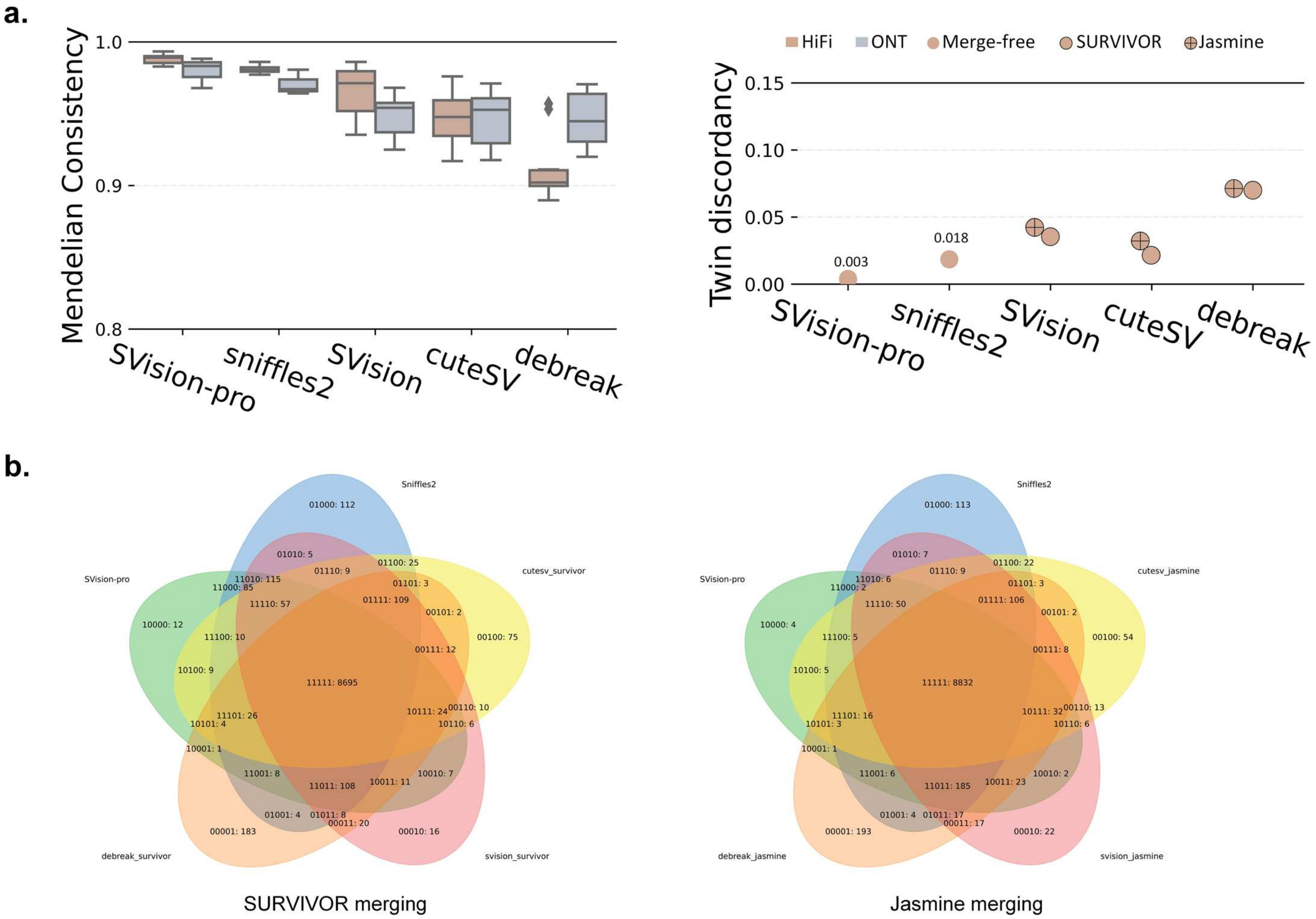
Extended Data Fig. 4 | Illustration of the image instance segmentation framework in SVision-pro. a, At the pixel level, the segmentation process predicts each image pixel as either belonging to the background or a specific variant type in the segmentation mask image **b**, The segmentation mask provides obvious comparison in SV subcomponent type, breakpoint, and allele frequency (AF) by contrasting the lower and upper track. Mask color comparison indicates the differences in SV subcomponent type. Horizontal comparison indicates the differences in SV subcomponent breakpoint span. Vertical comparison indicated the differences in SV subcomponent AF. Consequently, SVision-pro outputs four distinct comparison types to depict the SV difference between the case genome and the control genome, including germline, new components, new breakpoints and new alleles. **c**, In the scenarios where multiple control genomes are provided (such as the parent genomes in de no SV discovery), the instance segmentation framework predicts each image and outputs the SV difference between case genome and each control genome. Abbreviation: 'NewComp' for new component; 'NewBKP' for new breakpoint; 'NewAllele' for new allele frequency. **d**, SVision-pro currently provides three different image sizes. Larger image sizes lead to larger track heights, and thereby lower minimum representable allele frequencies (AFs). Moreover, the properties of the representation image, such as image size, track height and colors, can be customized for user-specific applications.



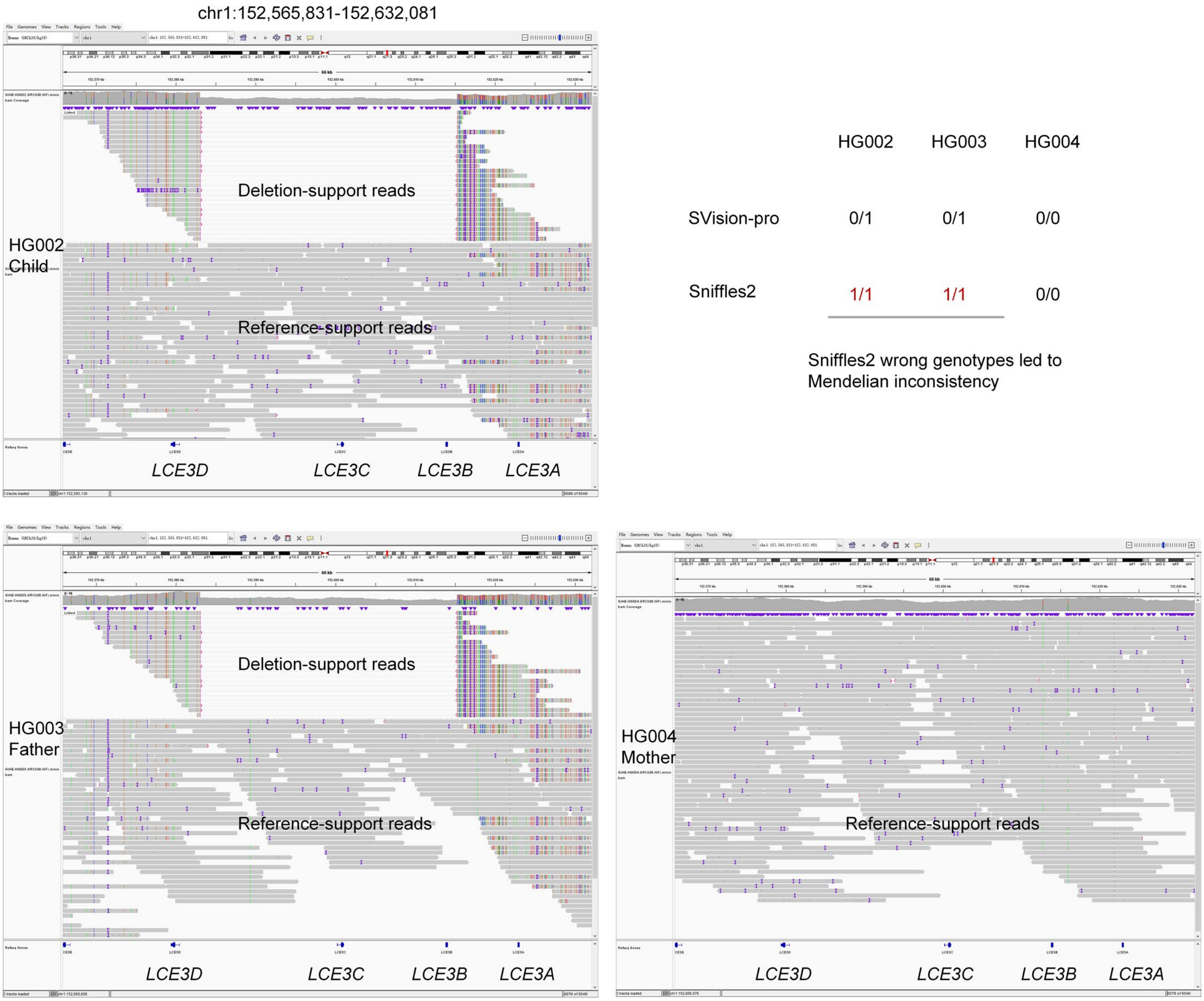
Extended Data Fig. 5 | Comparison and interpretation of the neural-network-based instance segmentation frameworks. a, Comparison of the accuracy (y-axis) on validation dataset among the five models (x-axis). The models are arranged based on their parameter sizes. **b**, the network architecture of the default Lite-Unet model. **c**, A heatmap to illustrate the Feature Ablation interpretation of the Lite-Unet model. Positives values (in green) indicates positive attrition to the specific prediction while negative values are shown in red. **d**, Using Grad-Cam to generate attribution maps of each layer in Lite-Unet.



Extended Data Fig. 6 | Performance evaluation of SSV and CSV calling among callers. a, SSV detection performance on HG002 groundtruth HiFi and ONT dataset. Recall, precision and F1-score were compared among callers. **b**, CSV detection performance on simulated 3,000 CSV HiFi and ONT dataset. Five of the highest-performing callers at SSV detection were chosen for a CSV performance comparison. Since only SVision-pro and SVision were equipped with CSV characterization ability, we utilized the region matching strategy to avoid the comparison of CSV types. **c**, CSV structure concordance evaluation among callers. Each box contains four values (Supplementary Table 3). 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. Values falling outside the Q1 – Q3 range are plotted as outliers of the data.



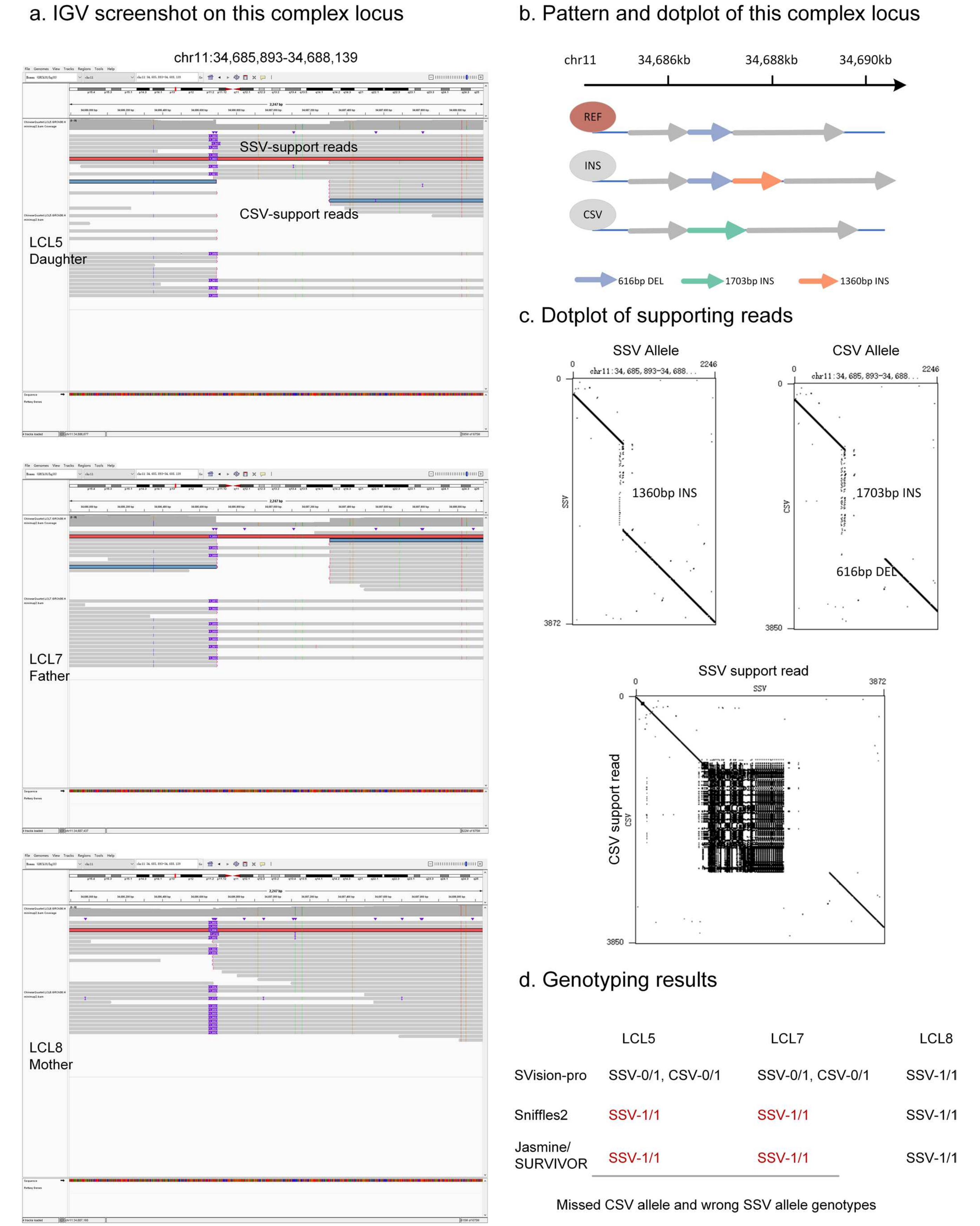
Extended Data Fig. 7 | Performance evaluation of Mendelian sample calling within high-confidence regions. a, In high-confidence regions, comparison of the Mendelian consistency in six family datasets (left) and the twin discordancy in the ChineseQuartet (right). SVision-pro is compared to Sniffles2 (multi-sample mode) and SVision, cuteSV and debreak (followed by SURVIVOR and Jasmine merging). Each box contains six and three values for HiFi and ONT, respectively (Supplementary Table 5). 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. Values falling outside the Q1 – Q3 range are plotted as outliers of the data. **b**, Venn diagrams show the overlapping results of high-confidence calls among approaches. We overlapped these high-confidence calls from each approach in AshkenazimTrio. there were only several unique calls (n = 12 and 4 when overlapping with SURVIVOR and Jasmine, respectively) from SVision-pro (9,348 in total), indicating that the leading consistency in Mendelian samples was attribute to the higher genotyping accuracy of SVision-pro compared to merging approaches.



	HG002	HG003	HG004
SVision-pro	0/1	0/1	0/0
Sniffles2	1/1	1/1	0/0

Sniffles2 wrong genotypes led to Mendelian inconsistency

Extended Data Fig. 8 | IGV screenshot of the 32,549 bp deletion in chromosome 1. The Ashkenazim Trio (HG002, HG003 and HG004) from GIAB was used to illustrate the various genotypes of this deletion. Sniffles calculated incorrect genotypes in this trio, leading to mendelian inconsistency. SVision-pro correctly genotyped this locus in the trio dataset, revealing that both the child genome (HG002) and the father genome (HG003) exhibited a heterozygous deletion, while the mother genome (HG004) contained no SV in this locus.



a. IGV screenshot on this complex locus

b. Pattern and dotplot of this complex locus

c. Dotplot of supporting reads

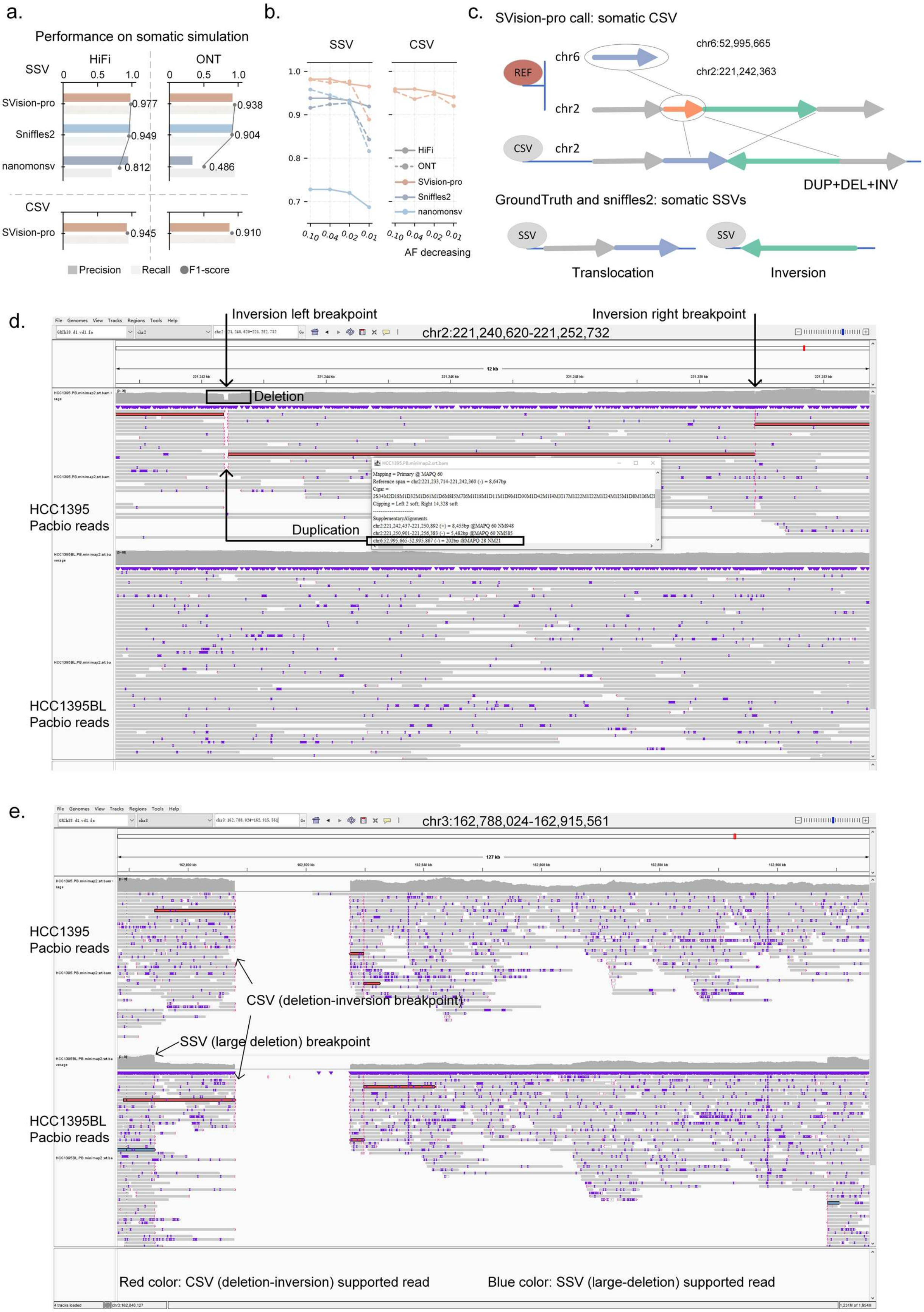
d. Genotyping results

	LCL5	LCL7	LCL8
SVision-pro	SSV-0/1, CSV-0/1	SSV-0/1, CSV-0/1	SSV-1/1
Sniffles2	SSV-1/1	SSV-1/1	SSV-1/1
Jasmine/ SURVIVOR	SSV-1/1	SSV-1/1	SSV-1/1

Missed CSV allele and wrong SSV allele genotypes

Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Illustration of the complex locus in chromosome 11. **a**, IGV screenshot on this complex locus in the ChineseQuartet. This complex locus comprised of two alleles, including one SSV deletion and one CSV deletion-insertion. Read that supported the SSV allele was marked in red while read that supported the CSV allele was marked in blue. **b**, The summarized pattern at this complex locus. **c**, Gepard Dotplots[36] were used to show the differences between the SSV allele and CSV allele. **d**, SVision-pro correctly genotyped the two alleles, outputting the correct genotype of each allele. Sniffles2 and callset-merging strategies missed the CSV allele and incorrectly genotyped the SSV allele as homozygous in the child and father genome.



```
HCC1395.PB.minimap2.srt.bam
Mapping = Primary @ MAPQ 60
Reference span = chr2:221,233,714-221,242,360 (-) = 8,647bp
Cigar =
2S34M2D18M1D32M1D61M1D6M8I5M7I6M1I18M1D11M1D9M1D30M1D42M1I4M3I17M1I22M1I22M1I24M1I5M1D8M1I6M2I
Clipping = Left 2 soft; Right 14,328 soft
----------------------
SupplementaryAlignments
chr2:221,242,437-221,250,892 (+) = 8,455bp @MAPQ 60 NM948
chr2:221,250,901-221,256,383 (-) = 5,482bp @MAPQ 60 NM585
chr6:52,995,665-52,995,867 (-) = 202bp @MAPQ 28 NM21
```

Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Somatic detection evaluation and discovery. a, The Precision, Recall, and F1-score of SVision-pro, Sniffles2, and nanomonsv on the simulated somatic SSVs and CSVs. **b**, The recall values of various low-frequency SSVs and CSVs in the simulation. **c**, A somatic CSV locus in chromosome 2 of HCC1395 cell line. SVision-pro reported this locus as somatic CSV, dispersed duplication-deletion-inversion, while in the previous published somatic SV set, the deletion component was missed and the dispersed duplication component was classified into translocation. **d**, IGV screenshot supported the CSV outputted by SVision-pro. **e**, IGV screenshot supported the homozygous CSV in the tumor genome and heterozygous SSV and CSV in the paired normal genome. The SSV large deletion breakpoint present in the paired normal genome while absent from the tumor genome.

nature portfolio

Corresponding author(s): Kai Ye

Last updated by author(s): Feb 21, 2024

Reporting Summary

Nature Portfolio 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 Portfolio 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	wget (GNU v1.14) was used to download all published datasets from their official ftps.
Data analysis	Minimap2 (v2.20-r1061) and NGMLR (v0.2.7) were used for long read alignment. SVision-pro (v1.6, https://github.com/songbowang125/SVision-pro.git), SVision (v1.3.9), Sniffles2 (v2.0.7), cuteSV (v2.0.2), pbsv (v2.9.0), debreak (v1.0.2), SVDSS (v1.0.5) and nanomonsv (v0.5.0) were used for structural variant calling. Jasmine (v1.1.5) and SURVIVOR (v1.0.7) were used for callset merging. Truvari (v3.5.0) and BEDtools (v2.30.0) were used for performance evaluation. IGV (v2.16.2) and Gepard (v1.4.0) were used to create dotplot for manual inspection. Vapor (version default) was used to perform computational validation SVision-pro (v1.6) is available at GitHub (https://github.com/songbowang125/SVision-pro.git) The scripts for model training, custom performance valuation and simulated data generation are available at GitHub (https://github.com/songbowang125/SVision-pro-Utils.git).

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 Portfolio 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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The human reference genome GRCh37: http://ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/reference/phase2_reference_assembly_sequence/hs37d5.fa.gz
The human reference genome GRCh38: http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/GRCh38_reference_genome/
NA19238 HiFi: http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/working/20191205_YRI_PacBio_NA19238_HIFI/
NA19239 HiFi: http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/working/20191205_YRI_PacBio_NA19239_HIFI/
NA19240 HiFi: http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/working/20191005_YRI_PacBio_NA19240_HiFi/
HG00512 HiFi: http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/working/20191031_CHS_PacBio_HG00512_HiFi/
HG00513 HiFi: http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/working/20191031_CHS_PacBio_HG00513_HiFi/
HG00514 HiFi: http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/working/20200731_CHS_PacBio_HG00514_HiFi_reseq/
HG00731 HiFi: http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/working/20190925_PUR_PacBio_HiFi/
HG00732 HiFi: http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/working/20190925_PUR_PacBio_HiFi/
HG00733 HiFi: http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/working/20190925_PUR_PacBio_HiFi/
HG002 HiFi: https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG002_NA24385_son/PacBio_CCS_15kb_20kb_chemistry2/
HG003 HiFi: https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG003_NA24149_father/PacBio_CCS_15kb_20kb_chemistry2/
HG004 HiFi: https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG004_NA24143_mother/PacBio_CCS_15kb_20kb_chemistry2/
HG005 HiFi: https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/ChineseTrio/HG005_NA24631_son/PacBio_CCS_15kb_20kb_chemistry2/
HG006 HiFi: https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/ChineseTrio/HG006_NA24694-huCA017E_father/PacBio_CCS_15kb_20kb_chemistry2
HG007 HiFi: https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/ChineseTrio/HG007_NA24695-hu38168_mother/PacBio_CCS_15kb_20kb_chemistry2
Chinese Quartet LCL5 HiFi: https://chinese-quartet.org/#/data/download/quartet-genomics
Chinese Quartet LCL6 HiFi: https://chinese-quartet.org/#/data/download/quartet-genomics
Chinese Quartet LCL7 HiFi: https://chinese-quartet.org/#/data/download/quartet-genomics
Chinese Quartet LCL8 HiFi: https://chinese-quartet.org/#/data/download/quartet-genomics
HG002 ONT: https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG002_NA24385_son/UCSC_Ultralong_OxfordNanopore_Promethion/
HG003 ONT: https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG003_NA24149_father/UCSC_Ultralong_OxfordNanopore_Promethion/
HG004 ONT: https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG004_NA24143_mother/UCSC_Ultralong_OxfordNanopore_Promethion/
HG005 ONT: https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/ChineseTrio/HG005_NA24631_son/UCSC_Ultralong_OxfordNanopore_Promethion/
HG006 ONT: https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/ChineseTrio/HG006_NA24694-huCA017E_father/UCSC_Ultralong_OxfordNanopore_Promethion/
HG007 ONT: https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/ChineseTrio/HG007_NA24695-hu38168_mother/UCSC_Ultralong_OxfordNanopore_Promethion/
Chinese Quartet LCL5 ONT: https://chinese-quartet.org/#/data/download/quartet-genomics
Chinese Quartet LCL6 ONT: https://chinese-quartet.org/#/data/download/quartet-genomics
Chinese Quartet LCL7 ONT: https://chinese-quartet.org/#/data/download/quartet-genomics
Chinese Quartet LCL8 ONT: https://chinese-quartet.org/#/data/download/quartet-genomics
HCC1395 CCS: https://downloads.pacbcloud.com/public/revio/2023Q2/HCC1395/HCC1395/
HCC1395 ONT: https://www.ncbi.nlm.nih.gov/sra?term=SRP162370
HCC1395 CLR: https://www.ncbi.nlm.nih.gov/sra?term=SRP162370
HCC1395BL CCS: https://downloads.pacbcloud.com/public/revio/2023Q2/HCC1395/HCC1395-BL/
HCC1395BL ONT: https://www.ncbi.nlm.nih.gov/sra?term=SRP162370
HCC1395BL CLR: https://www.ncbi.nlm.nih.gov/sra?term=SRP162370
HG002 SV callset: https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/analysis/NIST_SVs_Integration_v0.6/HG002_SVs_Tier1_v0.6.vcf.gz
HCC1395 SV callset: https://static-content.springer.com/esm/art%3A10.1186%2Fs13059-022-02816-6/MediaObjects/13059_2022_2816_MOESM4_ESM.xlsx

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	Not relevant to our study
Population characteristics	Not relevant to our study
Recruitment	Not relevant to our study
Ethics oversight	Not relevant to our study

Note that full information on the approval of the study protocol must also be provided in the manuscript.

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

Life sciences study design

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

Sample size	For de novo structural variant analysis, we collected 19 already-published samples from six family datasets, including Ashkenazim Trio, Chinese Trio, YRI Trio, CHS Trio, PUR Trio and Chinese Quartet. For somatic structural variant analysis, we collected one already-published normal-tumor paired cell line, including HCC1395 and HCC1395BL. These sample sizes were chosen based on the accessibility of already-published long-read-sequencing data. These sample sizes were sufficient for performance benchmarking of SVision-pro and other callers due to they possessed ground-truth callsets.
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-pro and its application to de novo and somatic structural variant detection with long-read sequencing data.
Randomization	Randomization was not relevant to our study. SVision-pro 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 involved. Besides, in this study, all softwares are deterministic and do not take advantages from knowing the origin of the input data.

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
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

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