Supplementary Materials

An integrative TAD catalog in lymphoblastoid cell lines discloses the functional impact of deletions and insertions in human genomes

Chong Li[1,19], Marc Jan Bonder[2,3], Sabriya Syed[4], Matthew Jensen[5,6], Human Genome Structural Variation Consortium (HGSVC), HGSVC Functional Analysis Working Group, Mark B. Gerstein[5,6], Michael C. Zody[7], Mark J.P. Chaisson[8], Michael E. Talkowski[9,10,11,12], Tobias Marschall[13,14], Jan O. Korbel[15], Evan E. Eichler[16,17], Charles Lee[4,18], and Xinghua Shi[1,19]

[1]Department of Computer and Information Sciences, College of Science and Technology, Temple University, Philadelphia, PA, USA

[2]Department of Genetics, Groningen, University of Groningen, University Medical Center Groningen, 9713 AV, Netherlands

[3]German Cancer Research Center, Division of Computational Genomics and Systems Genetics, Heidelberg, Germany

[4]The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA

[5]Department of Molecular Biochemistry and Biophysics, Yale University, New Haven, CT, USA

[6]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA

[7]New York Genome Center, New York, NY, USA

[8]Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA, USA

[9]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA

[10]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA

[11]Department of Neurology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA

[12]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA

[13]Institute for Medical Biometry and Bioinformatics, Medical Faculty and University Hospital, Heinrich Heine University, Düsseldorf, Germany

[14]Center for Digital Medicine, Heinrich Heine University, Düsseldorf, Germany

[15]European Molecular Biology Laboratory (EMBL), Genome Biology Unit, Heidelberg, Germany

[16]Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA

[17]Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA

[18]Department of Genetics and Genome Sciences, UConn Health, Farmington, CT, USA

[19]Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, PA, USA

Supplemental information

The overlapped TADs between GM12878 released by ENCODE and our Integrative Catalog

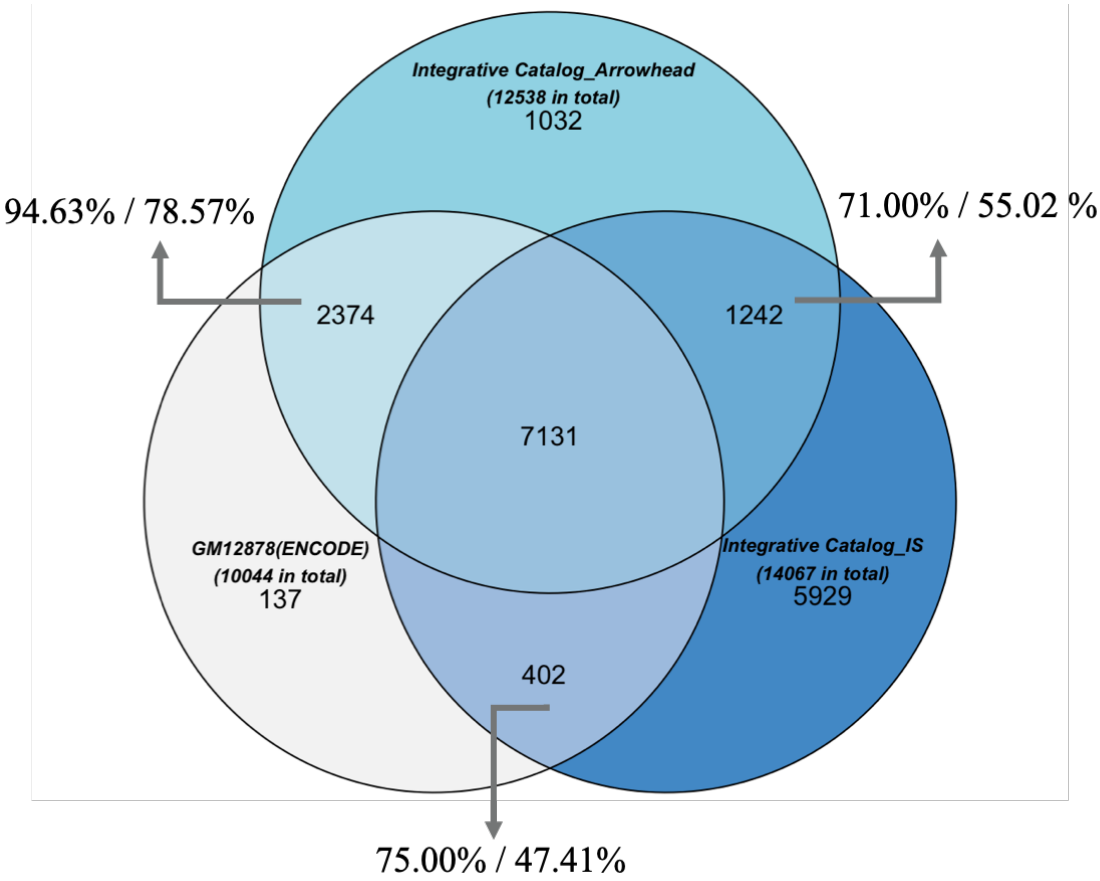
We evaluated the overlapped TADs by different requirements. In the main text, we mainly reported the one bp overlap between GM12878 released by ENCODE (short as GM12878 (ENCODE) in the following) and the Integrative Catalog in our study. To be more strict, we also applied the at least 50% reciprocal overlapped comparison by adding '-f 0.5' and '-r' parameters in the "*BEDTools intersect*" toolset. We found more than 97.76% of the TADs from GM12878 (ENCODE), which can also be detected in our released Integrative Catalog (**Supplementary Fig S2**). We further investigated those less than 2.24% non-overlapped regions (225 TADs) and applied one bp overlap in "*BEDTools intersect*" to calculate the percentage of overlapped regions. We found only one such TAD region (Chr 11-71090000-71430000) from GM12878 (ENCODE) was missing from our Integrative Catalog.

We took one step forward to visualize this TAD and surrounding regions (**Supplementary Fig S3**), and we found missing reads between 70935001 and 71080000 on the Chr 11, which caused the associated TADs to be filtered out by our filter criteria (**Methods**). We also checked the LCLs TAD calls from *Arrowhead*, and we believe it shared the most highly consistent result with the GM12878 (ENCODE), which is also detected by *Arrowhead*. However, we did not find the TAD region in our call set. The nearest TAD we detected is Chr 11:71340000-71440000.

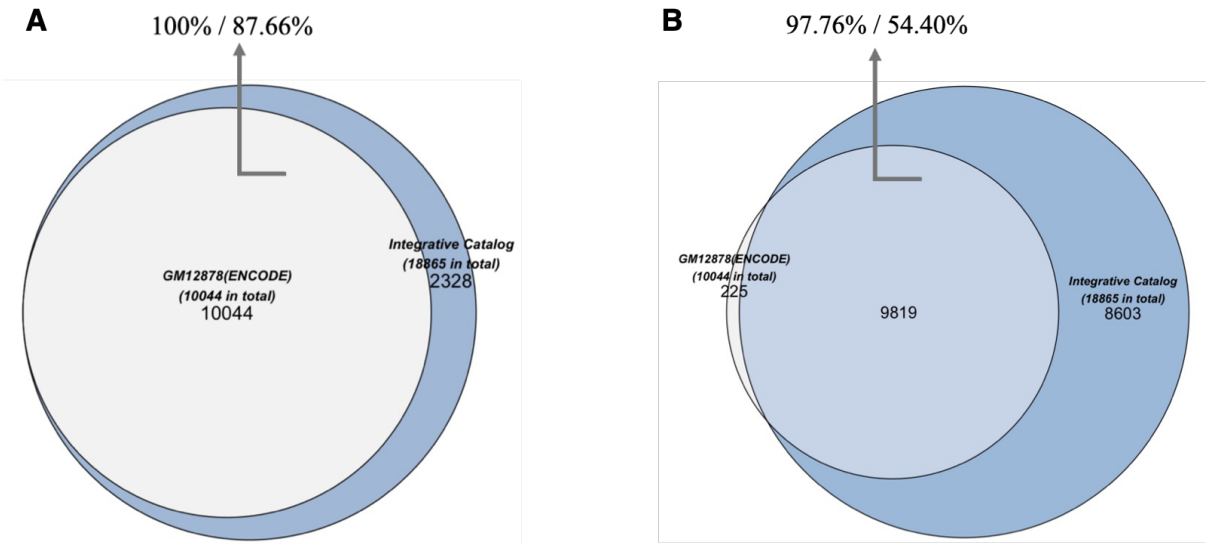
The discordance between the merged and individual TAD boundaries

We recognize that there might be some instances where a novel TAD boundary identified in the merged set may not be truly present in individual samples. We argue that this discrepancy could be due to different resolution selections for calling the merged boundary (5 kb) versus individual boundaries (10 kb). Certain TAD boundaries detectable in the merged dataset might not be observable at the individual sample level due to this difference in resolution. When examining TAD-SVs, we assigned a boundary score of zero to samples lacking an individual boundary at or near the merged boundary loci, indicating the absence of this merged boundary in those individuals and enhancing the accuracy of the statistical test. We found most merged boundaries are detected in the majority of individual samples, with only 85 merged boundaries (0.58%) present in fewer than five samples (11.63%) and 1,319 merged boundaries (8.93%) present in fewer than 21 samples (48.84%) (**Supplementary Fig S7**). We acknowledge that these rare boundaries could be either novel for the samples or artificial for the merging process. We compared these rare boundaries with the TAD boundaries of the benchmark cell line GM12878, derived from the ENCODE-released TAD regions, and found that 76 out of 85 and 975 out of 1,319 rare boundaries are also present in the GM12878 boundary set. Note that the set of GM12878 focuses primarily on TAD regions identified using the *Arrowhead* algorithm, whereas our set targets TAD boundaries using the *IS* algorithm. Therefore, minor differences between the two sets are expected. In the future, we intend to conduct experimental validation to characterize these rare boundaries and expect to directly call TAD boundaries on high-resolution Hi-C data on some of these samples to confirm the presence of these rare boundaries.

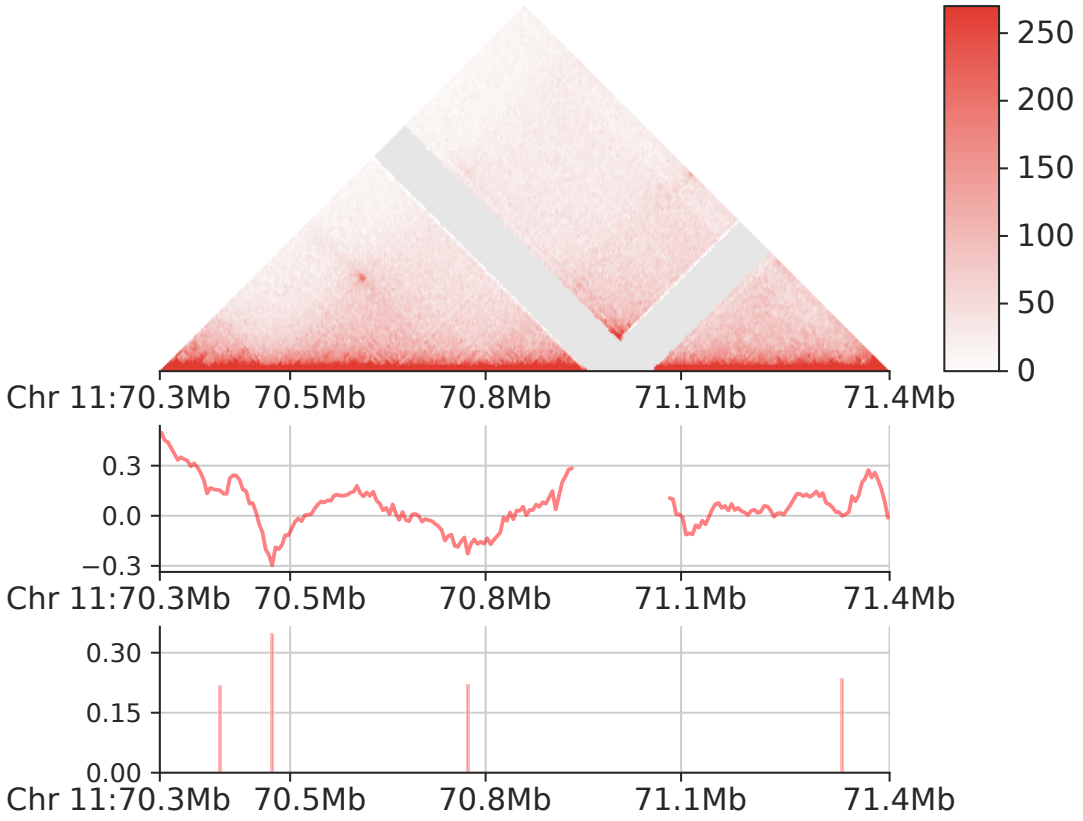
Supplemental Figures



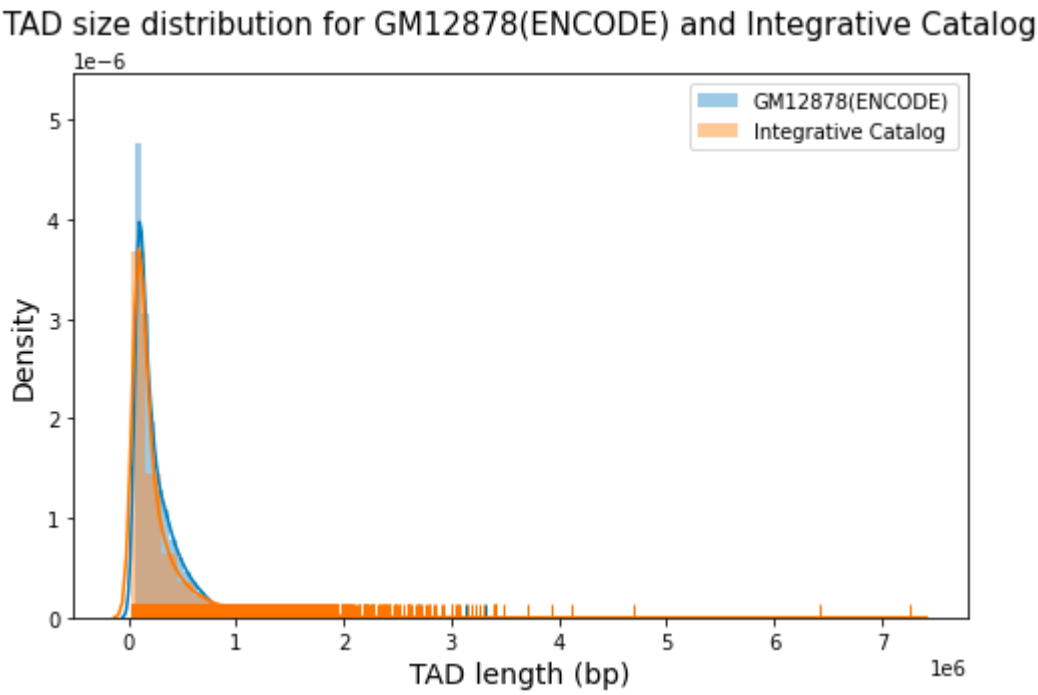
Supplementary Fig S1. The overlaps of TADs between GM12878 released by ENCODE, the TADs called by *Arrowhead,* and *IS* in our study (above 50% reciprocally overlapped). Approximately 71% of the TADs called by *Arrowhead* can also be detected by *IS* in our dataset. We observed that the relatively lower number of overlapped TADs between *Arrowhead* and *IS* is due to the fact that *Arrowhead* tends to identify TADs in larger sizes compared to *IS*, which cannot be recognized when we calculate the reciprocal overlap once the size of the TAD called in *Arrowhead* is more than half of the potential overlapped TAD called by *IS*. Around 94.63% and 75% of the TADs released by ENCODE for GM12878 coincided with the TADs separately detected by *Arrowhead* and *IS* in our study, respectively.



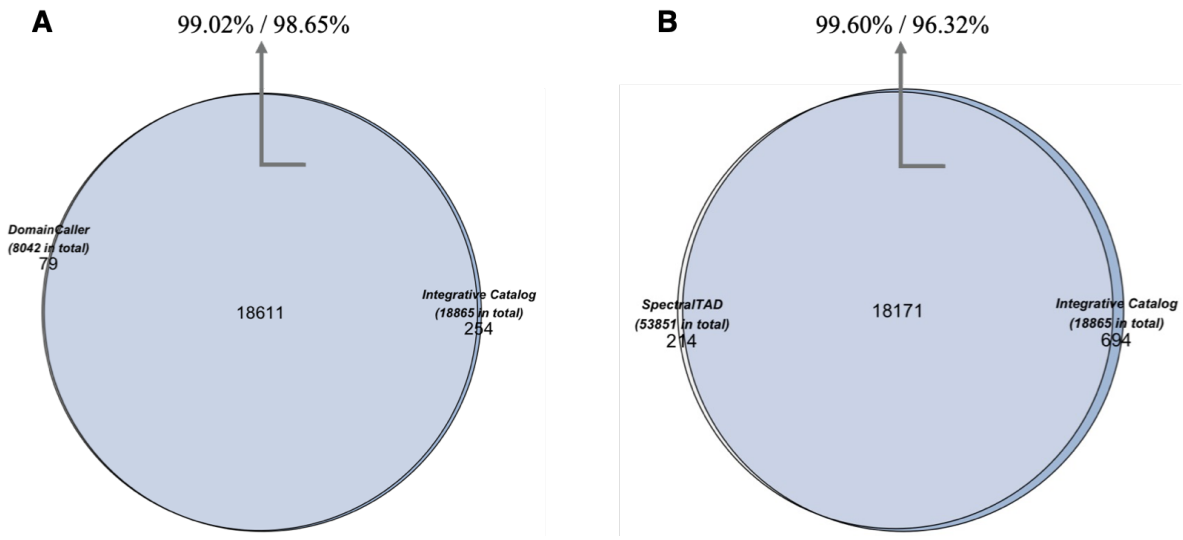
Supplementary Fig S2. The comparison of TADs between GM12878 released by ENCODE and the Integrative TAD Catalog released in our study. (A) The overlap of TADs between GM12878 released by ENCODE and the Integrative TAD Catalog generated using our customized pipeline in this study (one bp overlapped). Note that, for any data and results on GM12878 included in this study, we did not use the original results published by Rao *et al.* but used the results re-processed by ENCODE that mapped to hg38 as the reference genome (**Methods**). The Venn diagram shows that 10,044 TADs of GM12878 released by ENCODE are at least one bp overlapped with 16,537 TADs in our Integrative Catalog (containing sub-TADs). As a result, only the remaining 2,328 TADs that were uniquely released in our Integrative Catalog are shown in the Venn diagram. Since there were no released TAD boundary locations for GM12878, we focused only on comparing the sizes and numbers of the TADs in this study (**Supplementary Table S4**). **(B)** The 50% reciprocally overlapped TADs between GM12878 released by ENCODE and the Integrative TAD Catalog. Over 97.7% of the TADs released by ENCODE for GM12878 overlap more than half of the TADs from our Integrative Catalog call set when 50% reciprocal overlap criteria are applied.



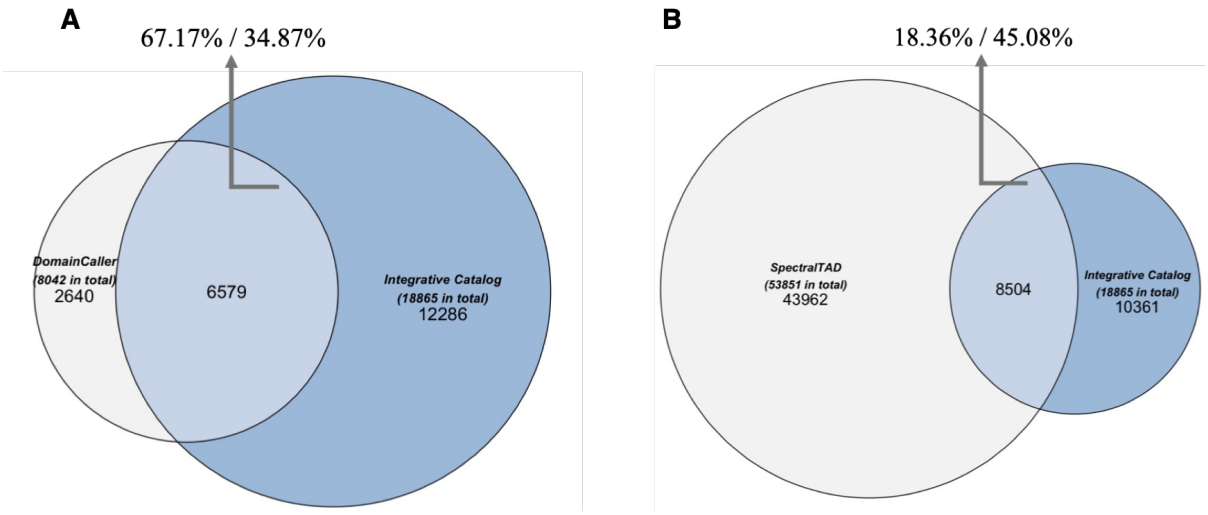
Supplementary Fig S3. The visualization of the TAD (Chr 11-71090000-71430000) was identified in GM12878 (ENCODE) but not in the Integrative Catalog. From top to bottom, the figure shows the Hi-C contact maps, the insulation scores, and the corresponding boundary with the boundary scores over this plotted region. The gray areas represent missing reads from 70935001 to 71080000 on Chr 11.



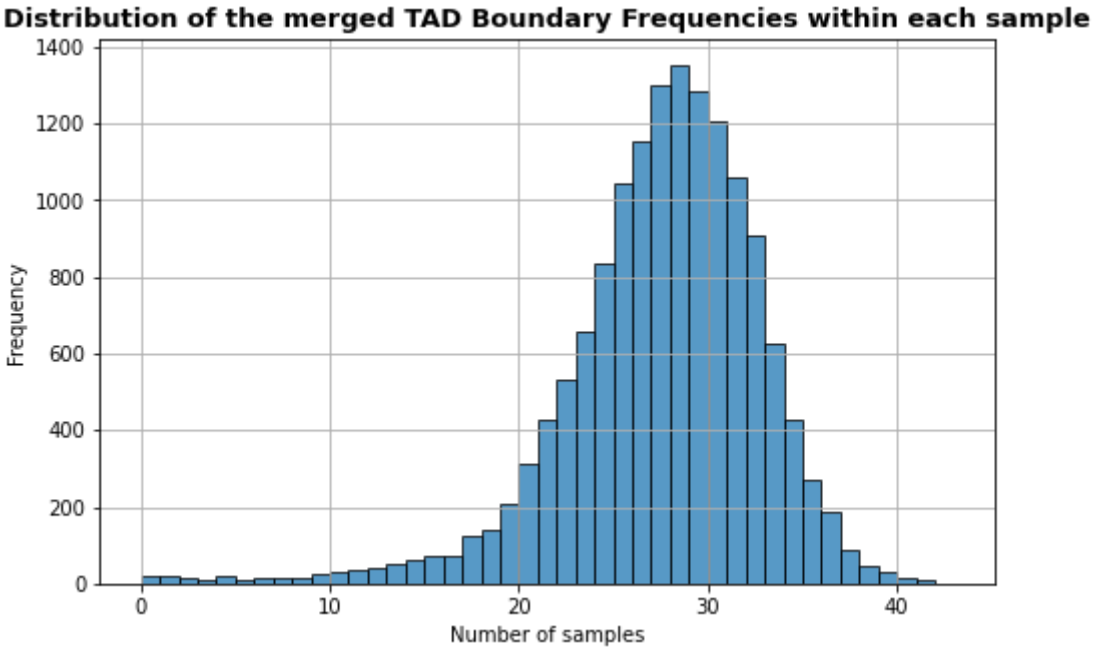
Supplementary Fig S4. The length distribution of the TADs identified from GM12878 (ENCODE) and our Integrative Catalog. Light blue represents the TADs from GM12878 preprocessed by ENCODE, while orange represents the TADs from our Integrative Catalog.



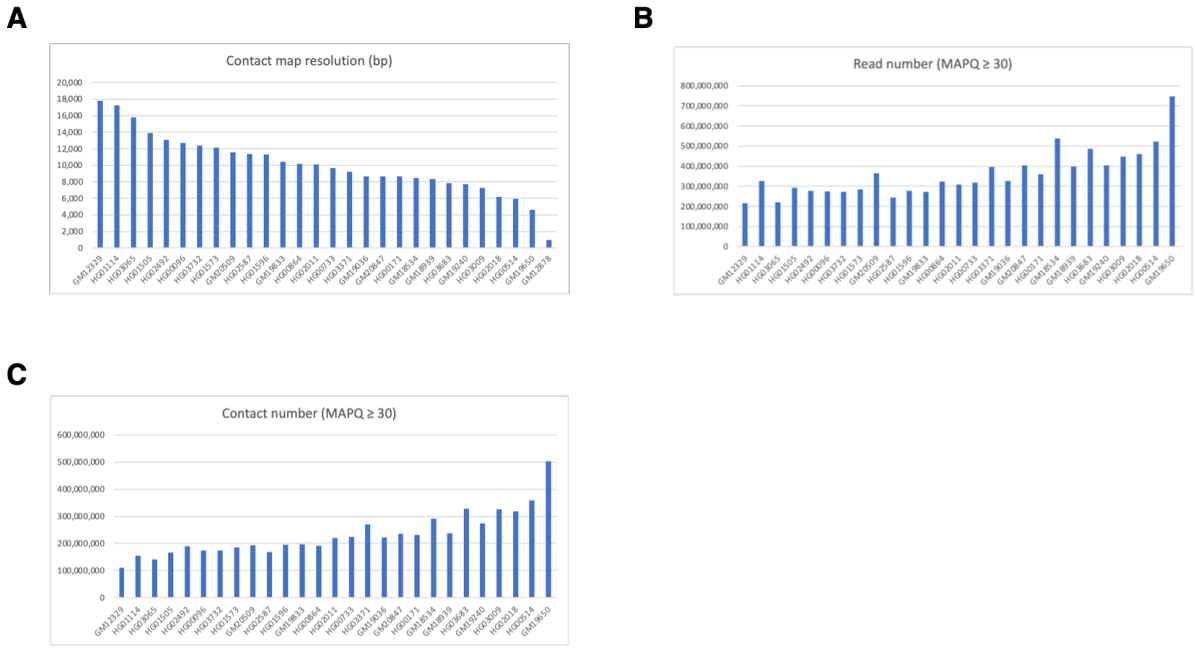
Supplementary Fig S5. The comparison of TADs identified between *DomainCaller*, *SpectralTAD*, and the Integrative TAD Catalog released in our study. (A) The overlap of TADs detected by *DomainCaller* and the Integrative TAD Catalog generated using our customized pipeline in this study (one bp overlapped). A total of 99.02% of the TADs identified by *DomainCaller* are at least one bp overlapped with all of the TADs identified in our Integrative Catalog. **(B)** A similar comparison was conducted between TADs detected by *SpectralTAD* and the Integrative TAD Catalog generated using our customized pipeline in this study (one bp overlapped). 99.60% of TADs identified by *SpectralTAD* are at least one bp overlapped with 96.32% of TADs in our Integrative Catalog. A reciprocal overlap ratio of 50% was also applied to conduct a more strict comparison, as shown in Supplementary Fig S6.



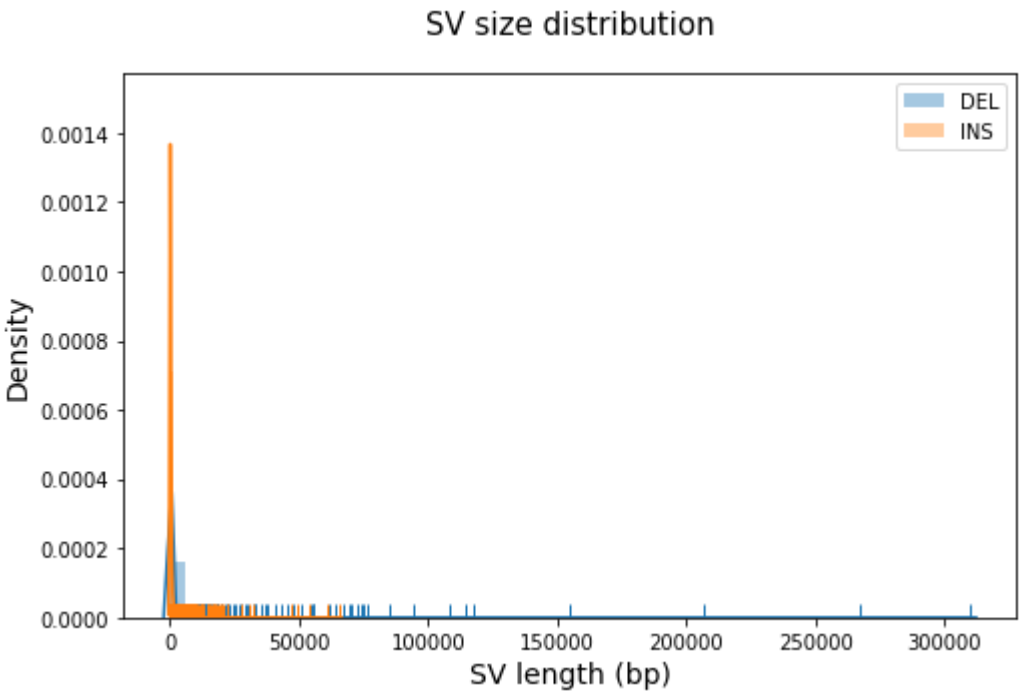
Supplementary Fig S6. A more stringent comparison of TADs identified from *DomainCaller*, *SpectralTAD*, and our Integrative TAD Catalog. (A) The 50% reciprocal overlap of TADs was detected by *DomainCaller*, and the Integrative TAD Catalog was generated in this study. 67.17% of TADs identified by *DomainCaller* are 50% reciprocally overlapped with 34.87% of TADs in our Integrative Catalog. **(B)** A similar comparison was conducted between TADs detected by *SpectralTAD* and the Integrative TAD Catalog. 18.36% of TADs identified by *SpectralTAD* are at least 50% reciprocal overlapping with 45.08% of TADs in our Integrative Catalog.



Supplementary Fig S7. The frequency of occurrence distribution for each merged TAD boundary at the individual level. The x-axis represents the number of samples that detected the TAD boundary at the same location as the merged TAD boundary (GM12878 was excluded as we did not call it directly on its individual level). The y-axis indicates the corresponding count of these merged TAD boundaries. Our analysis revealed that the most merged boundaries are also detected in the majority of individual samples. We observed that only 85 merged boundaries (0.58%) are present in fewer than five samples (11.63%), and 1,319 merged boundaries (8.93%) are present in fewer than 21 samples (48.84%).

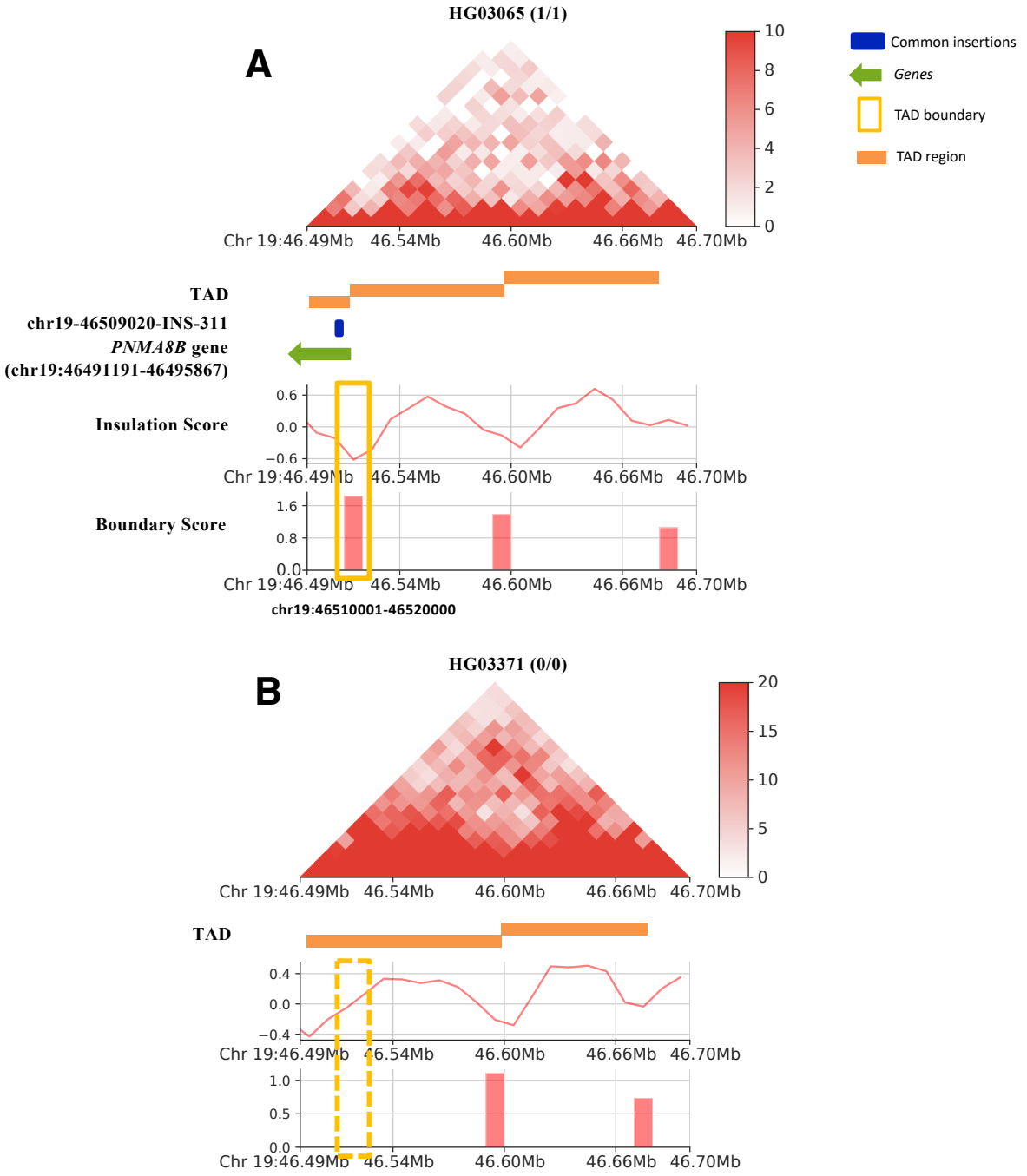


Supplementary Fig S8. **Map resolution, read numbers, and contact numbers of each sample. (A)** The map resolutions of our 27 samples were compared with those for GM12878 in the last column. The x-axis represents the sample ID, from left to right, and the resolution is low to high; the y-axis shows the value of resolution in the base pair (bp). **(B)** The number of read pairs of each sample with the filtered alignment quality based on the mapping quality score (MAPQ) ≥ 30. **(C)** The number of Hi-C contacts of each sample with the filtered alignment quality for MAPQ ≥ 30.

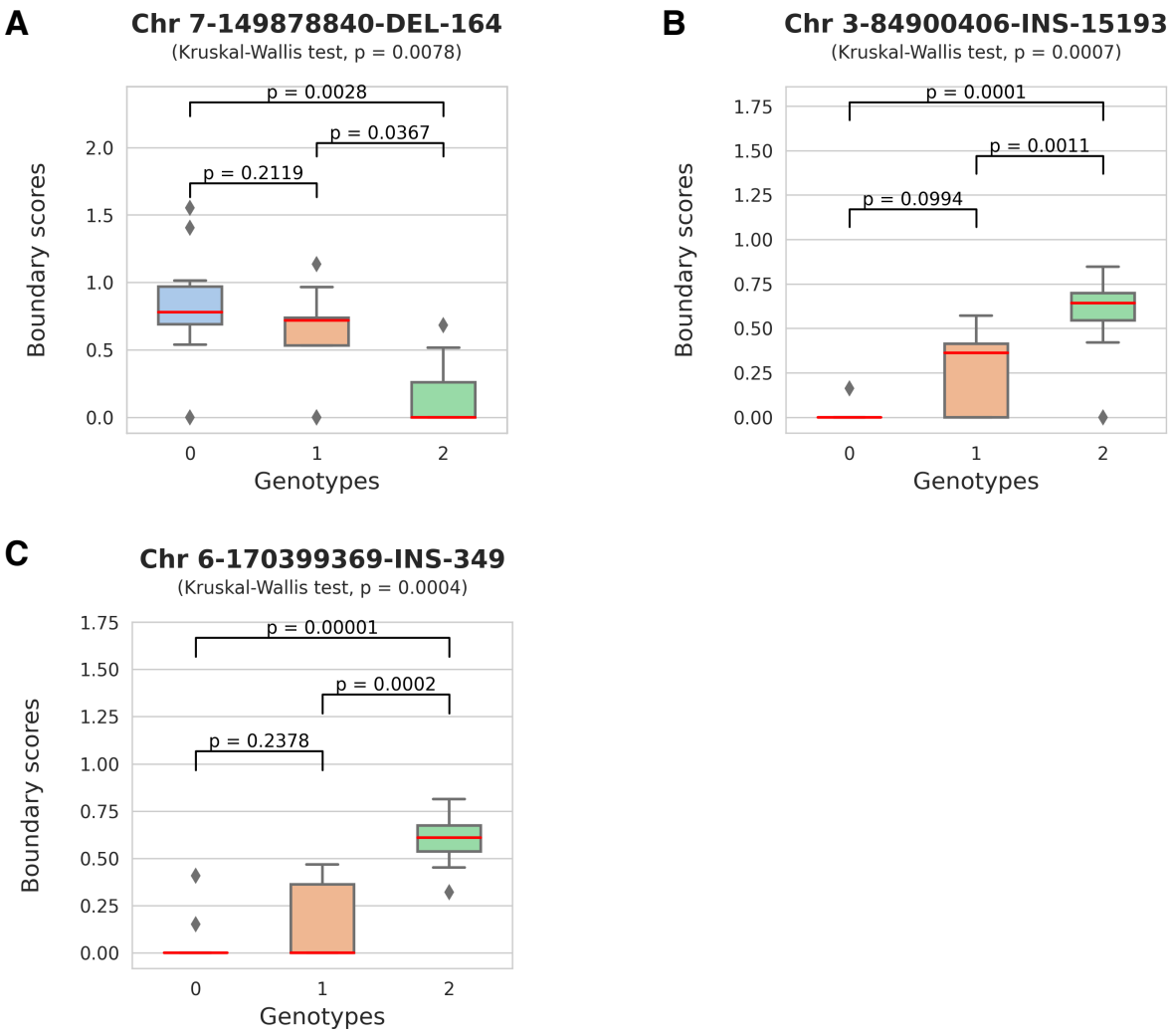


156

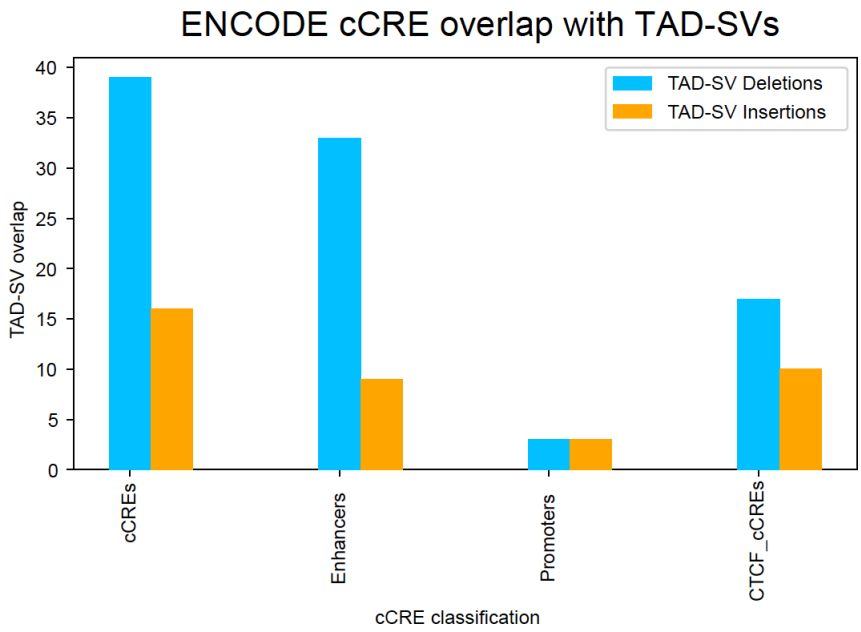
157 **Supplementary Fig S9. The length distribution of the SVs included in our study.** The light blue
158 color represents the deletions from the *PanGenie* SV genotyped call set preprocessed in this study,
159 while the orange color represents the insertions from the same SV sets after filtering.



Supplementary Fig S10. Visualization of an insertion disrupting TAD boundaries with significant changes in boundary strength. The figure includes a comparison of an insertion (Chr 19-46509020-INS-311) between the individual HG03065 (genotype 1/1) and HG03371 (genotype 0/0). The boundary score panel shows that the HG03065 sample, which carries the genomic insertion, exhibits a TAD boundary at the insertion site.



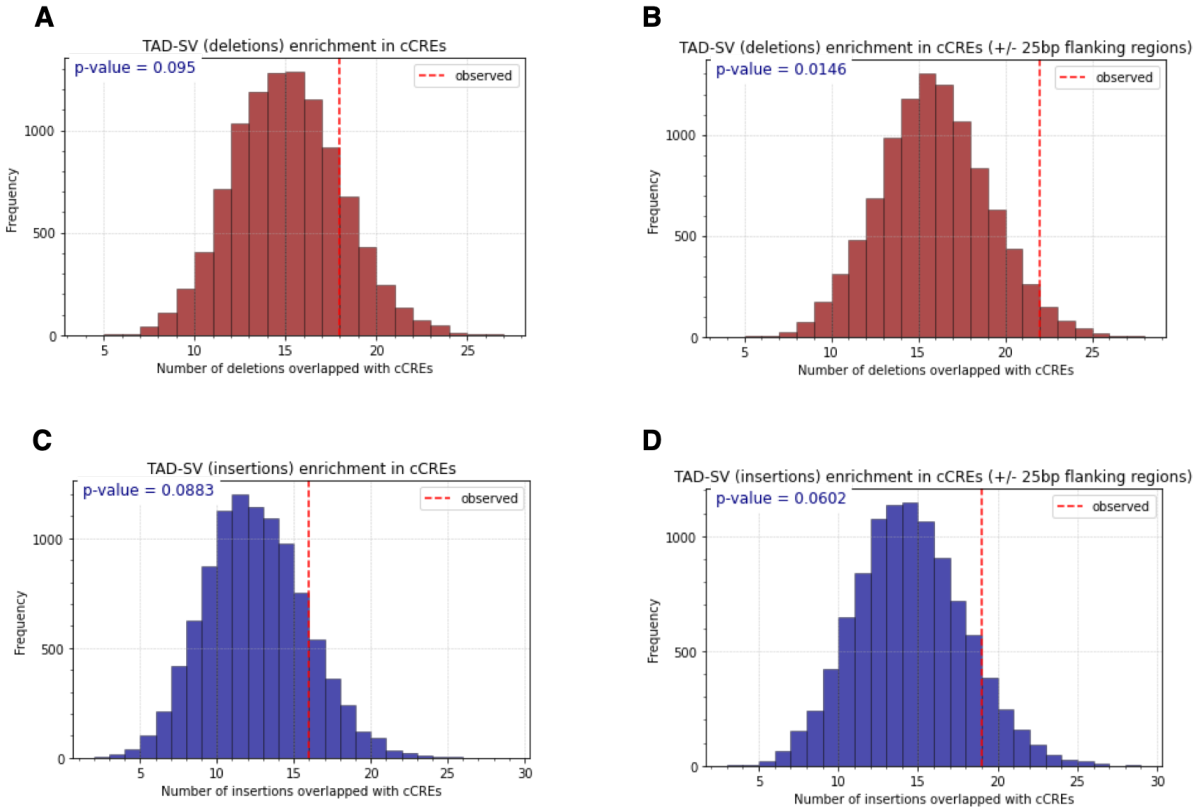
Supplementary Fig S11. Boxplots showing the different impacts of homozygous and heterozygous SVs on the 3D chromatin organization. Subfigures **A** to **C** present the data for one deletion (Chr 7-149878840-DEL-164), two insertions (Chr 3-84900406-INS-15193 and Chr 6-170399369-INS-349). These SVs exhibit significant differences among the genotype groups: 0 (genotype 0/0), 1 (genotypes 0/1 and 1/0), 2 (genotypes 1/1), and also demonstrate significantly different impacts on the boundary scores between the heterozygous genotype group 1 and the homozygous genotype group 2. Notably, homozygous SVs consistently exert a stronger effect on chromatin organization than heterozygous SVs.



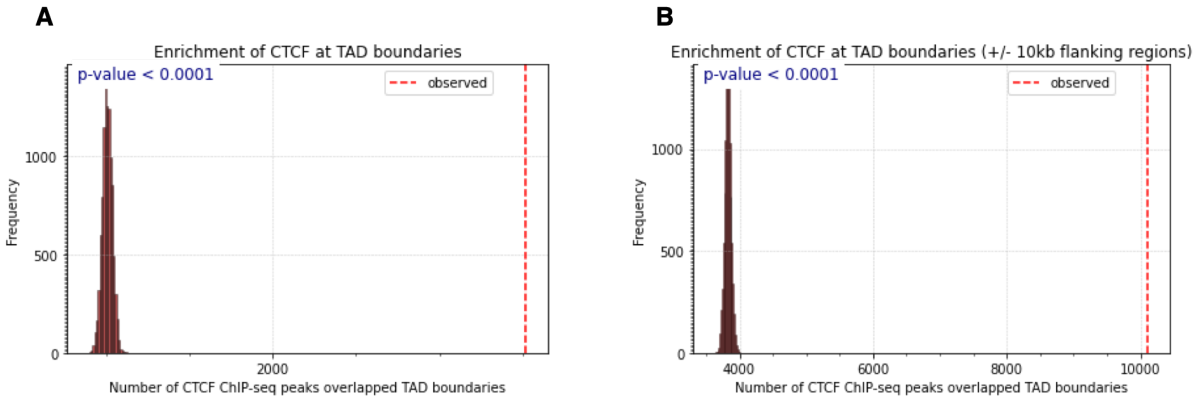
174

175 **Supplementary Fig S12. Overlap between TAD-SVs and ENCODE cCREs**. Bar plots show the
176 number of ENCODE v3 cCREs that overlap with the identified TAD-SVs. cCRE datasets are labeled
177 as follows: all cCREs, cCREs with enhancer-like signatures, cCREs with promoter-like signatures, and
178 cCREs with high CTCF binding activity.

179



Supplementary Fig S13. Enrichment of TAD-SVs in ENCODE cCRE regions. Bar plots show the count distribution for each of the 10,000 sets of randomly permuted TAD-SVs (**A** and **B** for deletions and **C** and **D** for insertions) overlapped with the ENCODE v3 cCRE sets. The red vertical lines indicate the observed counts of the TAD-SVs (18 for deletions and 16 for insertions) identified in this study intersected with the cCRE regions. We observed a modest but notable significance in the enrichment of TAD-SVs within the precise coordinates of cCRE (subfigures **A** and **C**). This significance became more pronounced when we expanded our analysis to include regions extending 25 bp (less than 10% of the median lengths of the cCRE datasets) both the upstream and downstream of each cCRE region (subfigures **B** and **D**). This finding could be attributed to the relatively smaller number of our identified TAD-SVs. The enhanced significance observed in the flanking regions suggests that the enrichment of TAD-SVs may not be confined strictly to the exact cCRE locations but may also extend to their adjacent genomic neighborhood.



193

194 **Supplementary Fig S14. Enrichment of CTCF at TAD boundaries identified in the Integrative**
195 **Catalog.** In each plot, we assess the observed count of overlapped TAD (red vertical line) overlapped
196 with (**A**) exact TAD boundaries and (**B**) TAD boundaries with added 10 kb flanking regions before and
197 after the start and end position by comparing it to a distribution of counts obtained from 10,000
198 randomly permuted sets of deletions. We observed a strong significant enrichment of CTCF at our
199 identified LCL TAD boundaries.

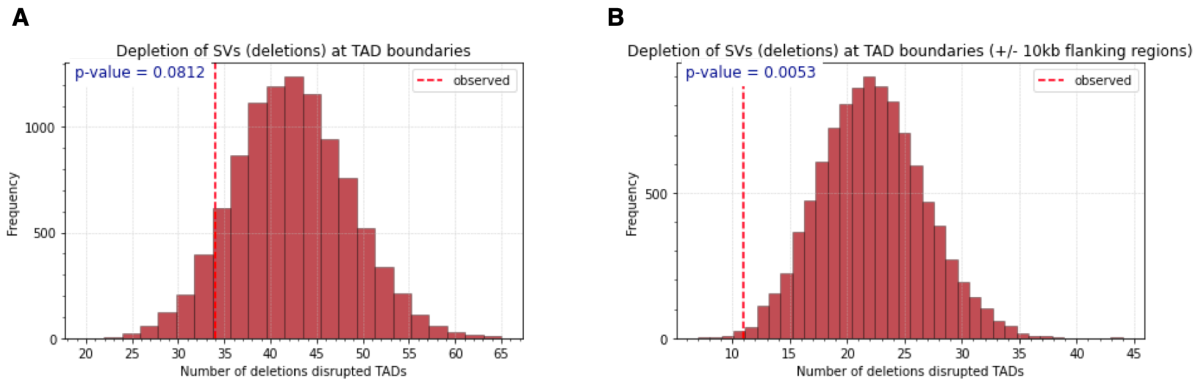
200

201

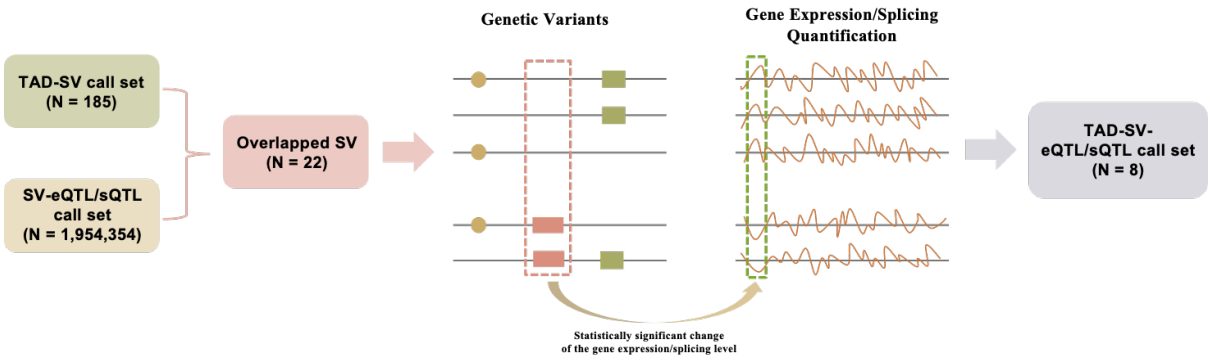
202

203

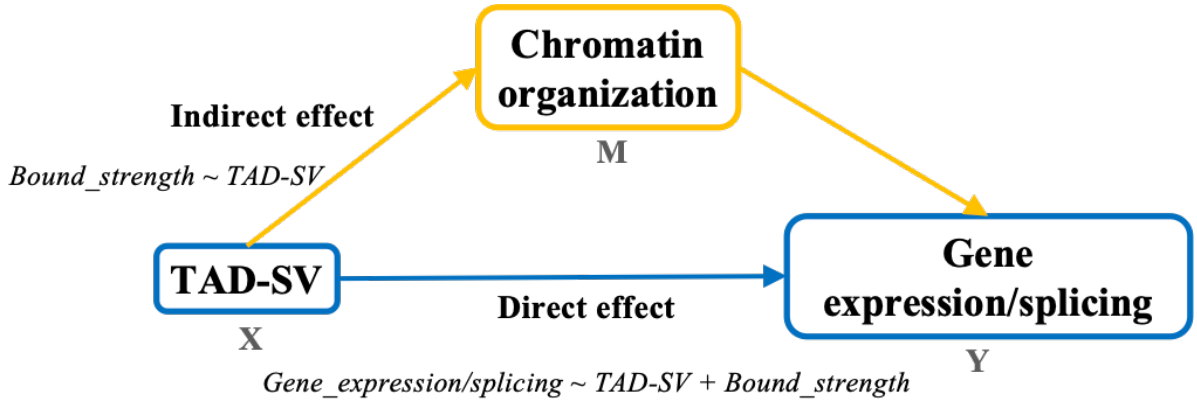
204



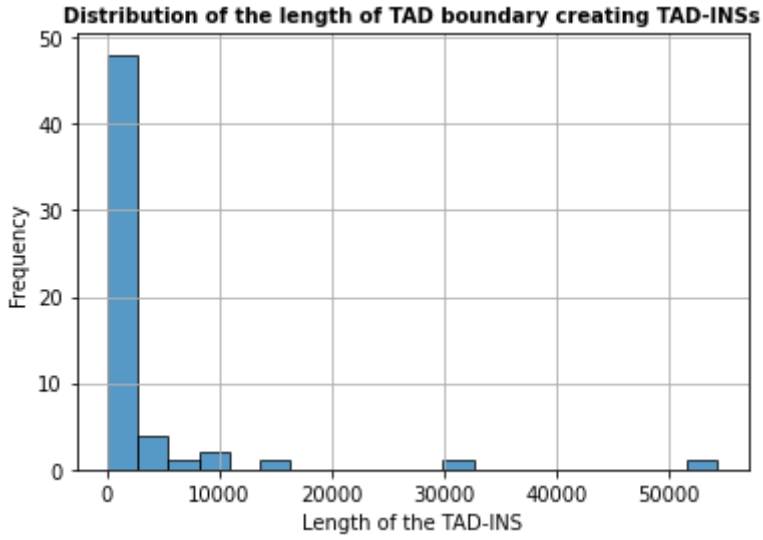
Supplementary Fig S15. Depletion of deletions at the identified TAD boundaries. In each plot, we assess the observed count of deletions (red vertical line) 50% reciprocal overlapped with (**a**) exact TAD boundaries and (**b**) TAD boundaries with added 10 kb flanking regions before and after the start and end position by comparing it to a distribution of counts obtained from 10,000 randomly permuted sets of deletions. We observed a modest yet significant depletion of genomic deletions within the TAD boundaries. Notably, the significance of this depletion was markedly enhanced when the analysis included the 10 kb flanking regions around the TAD boundaries. This finding could be attributed to the relatively smaller sizes of the deletions in our call set compared to the average length of TAD boundaries. The increased significance observed in the flanking regions indicated that the impact of deletions on TAD boundaries may not be limited to the immediate boundary area but could also affect the surrounding genomic landscape.



Supplementary Fig S16. An illustrated figure depicting the pipeline for association analysis between the TAD-SV and the SV-QTL to generate the TAD-SV-QTL set. We overlapped the TAD-SVs identified in this study with SV expression quantitative trait loci (eQTLs) and SV splicing quantitative trait loci (sQTL) characterized in Ebert's study. For each intersecting SV-eQTL and SV-sQTL, we extracted their corresponding quantifications of gene expression and transcript splicing in the 26 HGSVC2 samples used in this study. We employed the Wilcoxon rank-sum test (Mann–Whitney *U* test) to analyze gene expression and transcript splicing for each of those overlapped SV-eQTLs and SV-sQTLs, respectively, to compare the 26 samples with their homozygous reference (0/0) and heterozygous/homozygous deletions (1/1, 0/1, and 1/0). This comparison aims to elucidate TAD-SVs' impact on gene regulation and splicing patterns in these samples, thereby enhancing our understanding of the functional consequences of SVs on the TAD level.



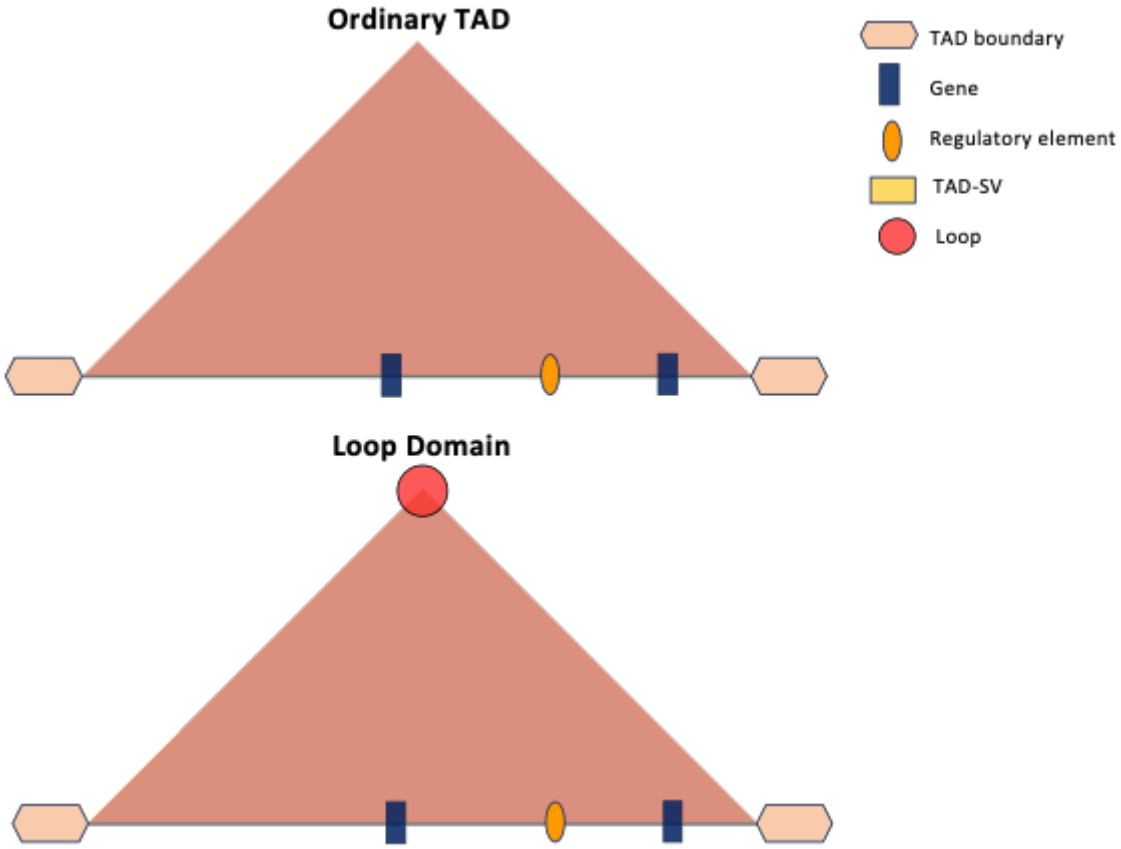
Supplementary Fig S17. An illustrated figure depicting the causal mediation analysis among TAD-SV, Chromatin organization, and gene regulation. We conducted a Causal Mediation Analysis to test whether these TAD-SVs (variable X) directly affect gene expression and splicing levels (variable Y), or if their effects are mediated through alterations in chromatin organization (variable M).



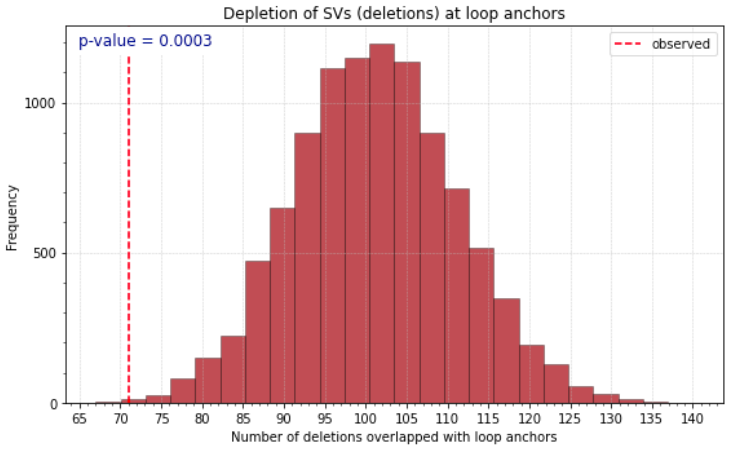
252

253 **Supplementary Fig S18. The size distribution for the TAD-INSs which are likely to create or**
254 **strengthen TAD boundaries.** The x-axis represents the length range of the 58 TAD-INSs, which
255 exhibit a tendency to introduce new TAD boundaries or strengthen the existing weaker boundaries. The
256 y-axis indicates the count number of the respective TAD-INS.

257



Supplementary Fig S19. An illustrated cartoon figure depicting the structure of the ordinary topologically associating domain (TAD) and the loop domain. The top section of the figure represents a conventional TAD characterized by its triangular shape. In contrast, the bottom section highlights loop domains, which are distinct chromatin structures that align with corner dots at their apexes, indicating points of chromatin looping.



Supplementary Fig S20. Depletion of deletions at the identified loop anchors. In the plot, we assess the observed count of deletions (red vertical line) 50% reciprocal overlapped with unique loop anchors by comparing it to a distribution of counts obtained from 10,000 randomly permuted sets of deletions.