scientific reports

Check for updates

OPEN

# Analysis of chromatin structure reveals the connection between sQTLs and the splicing of distant genes

Batuhan Eralp & Emre Sefer✉

**Gene expression and regulation with or without alternative splicing are key factors for cells to properly function. Distant splicing quantitative trait loci (distant sQTLs) are genomic mutations that impact the alternative splicing patterns of far-away genes. Nevertheless, the procedures causing a distant sQTL to regulate the alternative splicing of genes are not well defined. Higher resolution chromosome conformation capture experiments like Micro-C or Hi-C together with an expanding number of sQTL datasets on humans help us in understanding the spatial processes governing distant sQTL relationships at a genome-wide scale. In this study, we focus on analyzing whether spatial closeness helps in regulating sQTL-gene interactions over high-order chromatin topological domain structure, which is inferred from chromosome conformation experiments. We discover larger-scale chromatin shape to be in line with sQTL associations. In detail, sQTLs are generally spatially near their splicing genes in 3D, they frequently appear near topologically associating domain (TAD) and frequently interacting region (FIRE) boundaries, and are favorably related to genes over TADs and FIREs. Additionally, we discover that inside-domain sQTLs accompanied by functional regulatory elements, including enhancers and promoters, are spatially closer than all inside-domain sQTLs. This result suggests that spatial closeness between sQTLs and their distant splicing genes obtained from chromatin's TAD structure has major importance in regulating alternative splicing and thus in gene regulation. Our results are robust across different experiments such as Hi-C and Micro-C, different TAD inference methods, different Hi-C binning resolutions, different alternative splicing events, and once we control for eQTLs, which are shown to be spatially close to their genes.**

**Keywords** sQTL, Gene regulation, Hi-C, Chromatin Organization, Topological Domains, eQTL

Gene expression, gene regulation, and alternative splicing (AS) of mRNA for protein synthesis are crucial for proper functioning of tissues and cells[1–4]. Gene regulation and gene expression differ between multiple tissues over a given species, which are mainly accountable for morphological and cellular differences. Similarly, alternative splicing is also accountable for gene expression differences and resultingly for many cellular differences, where AS is defined as the alternative splicing of mRNA throughout gene expression by which each gene can encode different proteins[5]. Alternative splicing, as a molecular process, creates a number of different transcript isoforms over a unique gene, which as a result increases the diversity of the proteome and the diversity of transcripts in terms of structure. The transcript frequency over human transcriptome greatly surpasses the frequency of genes which encode for proteins[6]. As an average, every protein-encoding gene encodes for 7 transcript variants[7]. Across human, alternative splicing occurs over greater than 90% of the overall genes in a species-, condition-, or cell type-specific fashion. Such splicing is considered to significantly expand the possible proteins count over the available set of genes throughout the genome[8].

Splicing quantitative trait loci (sQTL) experiments map genetic mutations to changes in alternative splicing[5,9], and as a result, changes in gene expression and regulation[10,11]. A number of regulators regulating the gene expression have been discovered[12] via these experiments. Similarly, these sQTL experiments have also discovered genes related with point mutations called single nucleotide polymorphisms (SNPs) in diseases, and efficient targets for cancer therapies[13,14]. Recognizing regulators and developing therapeutic targets for

Ozyegin University, Artificial Intelligence and Data Engineering Department, Istanbul 34794, Turkey. ✉email: emre.sefer@ozyegin.edu.tr

disease-related genes shows our capability in determining the processes via which alternative splicing and gene expression is modulated by a mutation[15–17].

In this study, we investigate whether high-order three-dimensional chromatin structure is important in controlling sQTL associations at a genome-wide scale via bringing sQTLs near their genomically far away genes in 3D. To our best knowledge, this question has still not been answered, but similar question on the relationship between chromatin interactions and eQTLs instead of sQTLs has been answered. Results from some of the previous studies state that spatial closeness is important in regulating genes for a specific subset of eQTLs[18–20]. In another work,[21] has found that eQTLs are statistically near their target genes in 3D over human IMR90 fibroblasts and ESCs Hi-C datasets. Such statistical closeness is more important for eQTLs (expression quantitative trait loci) that are detected overlapping withs CRMs (Cis Regulatory Module) and eQTLs that are detected inside the same topologically associating domain (TAD)[22–27]. According to their study, genomic segments with eQTLs in general have a greater chromatin interaction frequency (CIF). Similar to the findings by Duggal et al., another research[13] has found that CRM enriched eQTLs are spatially close to target gene promoters over Hi-C datasets. Hi-C datasets in both research are solely for cell lines, whereas eQTL conclusions were derived over both cell line and tissue datasets. Lastly,[28] has analyzed the proximity between Hi-C and eQTL datasets, and their results have shown the spatial closeness between eQTLs and their corresponding genes among multiple human primary tissues. The remaining studies accepted the association between eQTL and Hi-C data results apriori, without providing any evidence, to show their results correctness. For instance, some of these studies focus on detecting regulatory SNPs[29], while the others focus on predicting enhancer-promoter interactions[30,31].

Nevertheless, the importance and the degree of such importance is not known for sQTLs. Chromatin's topological structure might be related with sQTLs according to a number of genome-wide studies. For instance, SNPs identified by genome-wide association studies are depleted over DNA fragments of chromatin interaction networks which are mediated by RNA polymerase[32]. This approach is specific to RNA polymerase, and it does not consider sQTLs and their target genes but only takes into account SNPs over genome-wide association studies. Additionally, a number of studies have examined eQTLs instead of sQTLs in respect of the predictive power of a number of chromatin features for the target genes associated with eQTLs[33]. On the other hand, sQTLs have also been analyzed in connection to multiple different chromatin marks such as promoters, transcription factor binding sites (TFBSs), and DNase I hypersensitive sites[9]. Among these marks, some of them such as DNase I hypersensitive sites show that chromatin's topological structure can be connected to sQTL associations. However, the problem of whether high-order chromatin structure are globally connected to sQTL-gene associations is still open. We can compare sQTL relationships with 3D chromatin structure at a genome-wide scale by utilizing high-coverage Hi-C and Micro-C datasets[34,35], which are chromosome conformation capture experiments that infer interactions between genome segments at a high resolution as high as 1 kb, and at a genome-wide scale[36].

We analyze whether spatial proximity in 3D has a significant role for sQTLs to regulate the splicing of their target genes. To analyze such relationship, we place sQTLs with regards to spatially-compact contiguous topologically associating domains (TADs) as seen in Fig. 1 which have been preserved over species and lasting across cell types[37–39]. High correlation exists between those topological domains and many chromatin marks related to gene regulation and alternative splicing. As a result, these domains can be related to sQTLs locations over the whole genome. More explicitly, we focus on testing the subsequent hypotheses independent of the cell type, TAD inference method, and alternative splicing event type: 1- sQTL fragments frequently interact with the remaining fragments; 2- sQTLs exist near topological domain boundaries throughout the genome; 3- sQTLs are spatially near the genes of which they regulate the alternative splicing in 3D, specifically inside topological domains; 4- sQTLs are frequently associated with genes across topological domains; 5- Inside-domain sQTLs that also have functional regulatory elements are near their target genes of which they regulate alternative splicing. We find that high order chromatin structure is paired with sQTL associations, and provide evidence for all of those hypotheses. Additionally, we still provide the evidence for all of those hypotheses when we control for eQTLs.

In this study, we have collected total number of 1,048,124 sQTL-gene pairs from Garrido-Martin et al.[9] for 11 human cell types: Adrenal Gland (AD), Dorsolateral Prefrontal Cortex (CO), Hippocampus (HC), Lung (LG, IMR90), Ovary (OV), Pancreas (PA), Small Bowel (Intestine) (SB), Spleen (SX), Liver (LI), Aorta (AO), Heart Left Ventricle (LV). Among those sQTL-gene pairs, we have chosen 739,112 intrachromosomal pairs, corresponding to almost 70% of all sQTLs, for which the distance between a single-nucleotide polymorphism and the boundary of the associated gene is greater than 50 kb. We then map those sQTL-gene associations to corresponding Hi-C interaction networks of single chromosomes at 40 and 5 kb resolutions using datasets from multiple Hi-C and Micro-C experiments[34,35] respectively, as discussed in Methods section in detail. In the text, we will mostly refer Micro-C and Hi-C datasets singly as Hi-C dataset. Since resolution is inherently limited for Hi-C datasets, we have combined sQTL-gene associations which could not be separated in regards to chromatin interactions, via putting the associations corresponding to the same Hi-C bins into the same group. As a result of such grouping, we obtain totally a set of 156,777 eQTL–gene equivalence classes, which is explained in Methods section in detail. We come up with a framework for optimal matching which relates sQTLs and topological domains with randomly-generated equivalents, and apply statistical tests over this framework. As part of this matching, we control for a number of confounding factors such as TAD lengths, sQTLs genomic attributes, distribution of the existing eQTLs, and interaction patterns that are commonly observed in Hi-C datasets as discussed in Methods section in detail. This robust matching method eliminates the fact that our analysis results are artifacts of less fascinating directions of chromatin interaction and sQTL datasets. Even when we follow a stringent statistical approach by controlling many factors, we observe statistically significant associations between chromatin structure, sQTLs, and their target genes. These associations suggest a strong connection between the landscape of sQTL-gene associations and spatial structure of chromatin at a genome-wide scale.

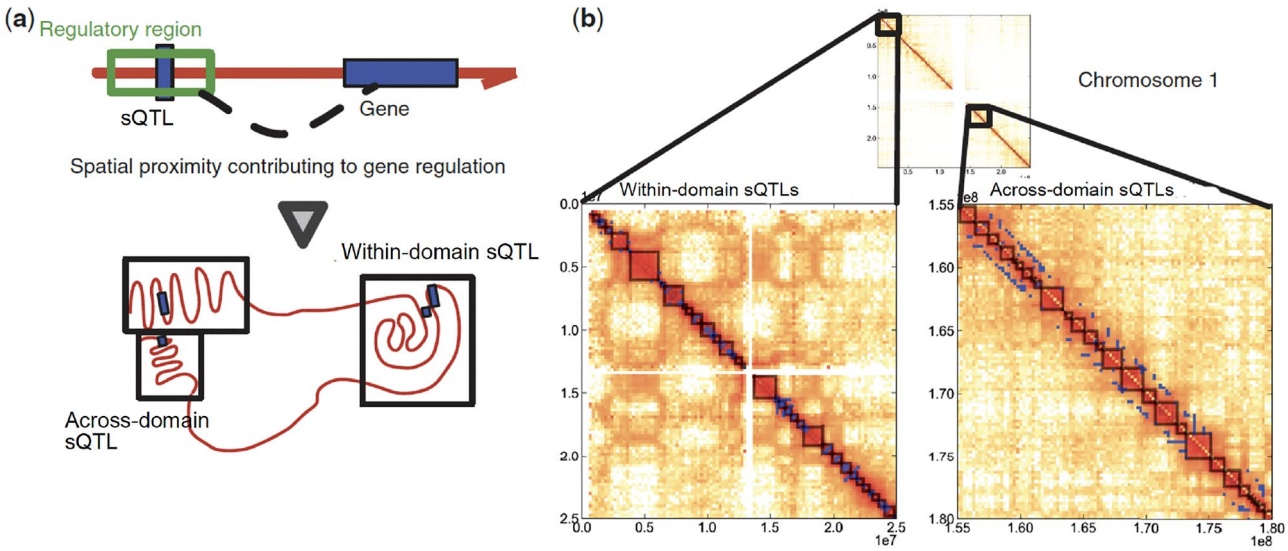



**Fig. 1**. Spatial closeness of sQTLs and corresponding target genes in chromatin's high-order Topologically Associating Domain (TAD) structure. (**a**) Illustration on how regulatory regions mutations could impact the spatially adjacent target genes expression. Close-range high-frequency interactions can promote to inside-TAD gene regulation and long-range cross-TAD interactions can temporarily lead 2 TADs to be in spatial closeness to promote drastic longer-range regulation. (**b**) Multiple evidences from experiments suggest gene expression's both inter-TAD and intra-TAD regulation through single nucleotide polymorphism mutations. 2 parts of a heatmap of Hi-C chromatin interactions for lung fibroblast's chromosome 1 obtained via Schmitt et al.[34] exhibit the prevalence of both intra-TAD interactions (bottom left) and inter-TAD interactions (bottom right). Every axis in interaction matrices corresponds to a locus on chromosome 1. High-frequency interactions between a loci pair are shown in red. Interactions between sQTLs and corresponding genes are displayed in blue, whereas TADs are surrounded by black squares along the matrix's diagonal.

To our best knowledge, this is the first study to systematically analyze the spatial closeness between sQTLs and their distant splicing genes obtained from chromatin's TAD structure. Such closeness has major importance in regulating alternative splicing and thus in gene regulation. Our results are important to enhance our comprehension of the roles of sQTLs and chromatin interactions on gene expression and regulation through alternative splicing.

## Results

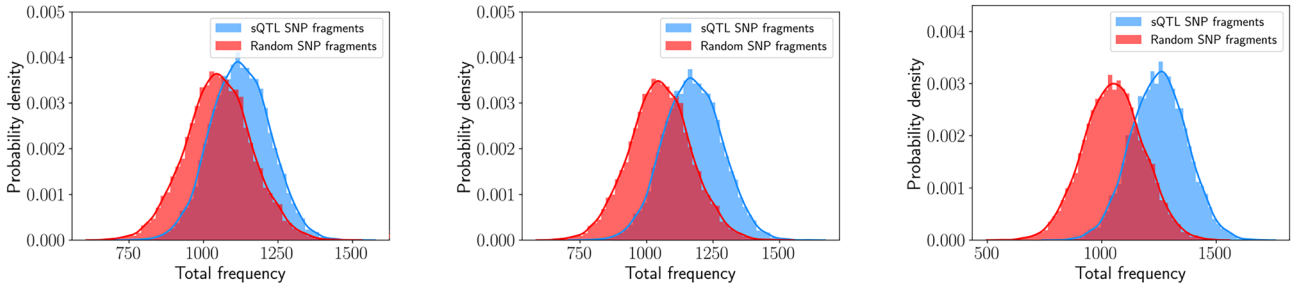
Even though we run our experiments over all of the considered 11 human cell types, we mainly report the results from IMR90 lung fibroblasts which uses finer resolution Micro-C interaction dataset and liver using Hi-C interaction dataset unless otherwise mentioned (See Methods section for details on datasets). In general, results are similar for almost all human cell types. 'Across-domain' sQTL-gene associations are defined as the overlapping SNP-gene intervals that are not fully enclosed inside a TAD, which are described previously in the Methods section. 'Non-crossing' sQTLs are sQTLs that are associated with genes across a TAD. In addition, we can divide non-crossing sQTLs into 2 subgroups: 1- Non-crossing sQTLs that associate inside-TAD, and 2- Non-crossing sQTLs that exist in gapped intervals between TADs. We have found that 6456 non-crossing sQTL associations out of 6554 associations ($98.5\%$) locate inside TADs for IMR90 cells. Similarly, 6136 non-crossing sQTL associations out of 6314 associations ($97\%$) locate inside TADs for liver cells. As a result, our conclusions for non-crossing associations in general apply to within-TAD as well.

### sQTLs frequently interact

We have discovered that sQTL SNP bins (fragments) $v$ have on average greater total interaction frequencies $t(v)$ than the background collection of whole set of Hi-C bins, as seen in Figs. 2ba-2cfor IMR90 Micro-C, IMR90 Hi-C, and liver Hi-C interaction datasets respectively with $p < 10^{-3}$ for all three cell types according to Wilcoxon rank sum test. $p < 10^{-3}$ is also observed when we use a one-sided test. When we repeat these distribution analyses over replicate datasets, we obtain $p = 0.0005$, $p = 0.0005$, and $p = 0.0003$ by using Wilcoxon rank sum test for IMR90 Micro-C, IMR90 Hi-C, and liver Hi-C interaction datasets, respectively. Additionally, the distribution of spatial proximity for both sIMR90 Hi-C and Micro-C datasets appear to be quite similar, which strengthens our conclusions.

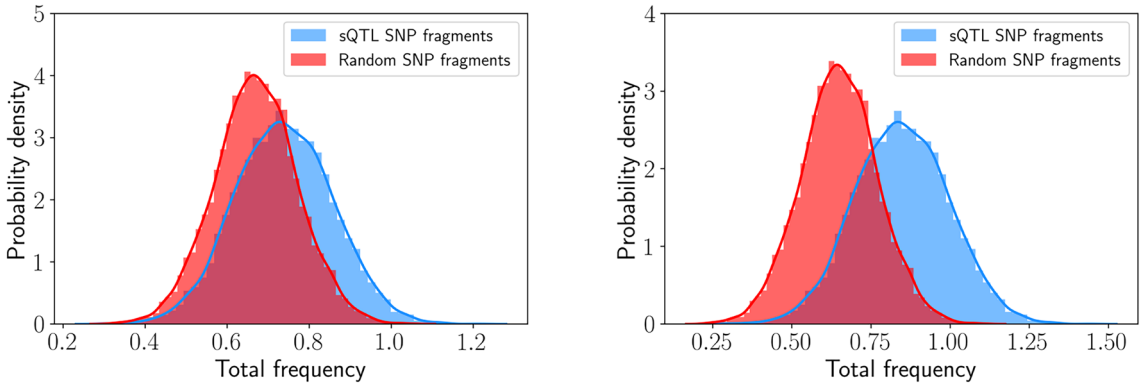
We also assess the impact of the Hi-C normalization technique by analyzing the distribution and spatial proximity of sQTL bins and associated genes nearby TAD boundaries via two other normalization techniques as seen in Figs. 3a-3b

over IMR90 cells: Knight-Ruiz (KR) matrix balancing[40] and Iterative Correction and Eigenvector decomposition (ICE) normalization[41]. We obtain $p = 0.064$ and $p = 0.021$ by the Wilcoxon rank sum test




**(a)** IMR90 cells, Micro-C **(b)** IMR90 cells, Hi-C **(c)** Liver cells, Hi-C

**Fig. 2**. The distribution and spatial proximity of sQTL bins and associated genes nearby TAD boundaries. Distributions of the total frequency of the whole set of Micro-C/Hi-C bins are compared against the same distribution for sQTL bins in **a**) IMR90 cells with Micro-C, **b**) IMR90 cells with Hi-C, **c**) Liver cells with Hi-C datasets. We have obtained the distribution shapes by kernel density estimations for each distribution.

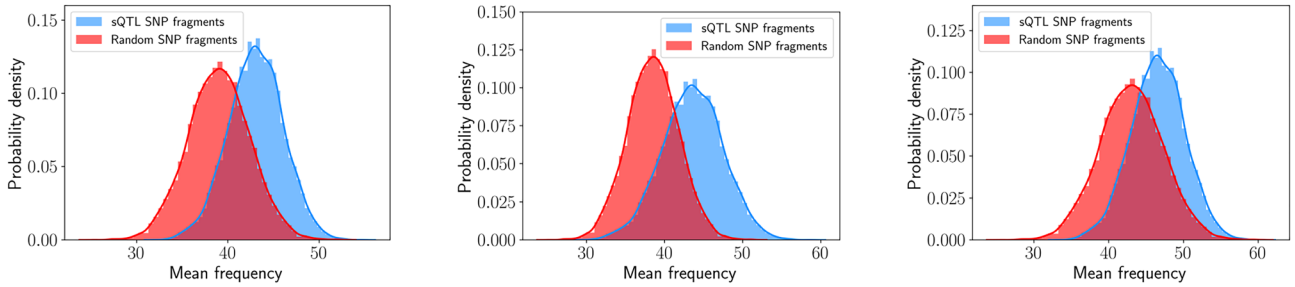



**(a)** IMR90 cells, KR Normalization **(b)** IMR90 cells, ICE Normalization

**Fig. 3**. The distribution and spatial proximity of sQTL bins and associated genes nearby TAD boundaries. Distributions of the total frequency of the whole set of Hi-C bins are compared against the same distribution for sQTL bins in human IMR90 cells: **a**) After KR normalization, **b**) After ICE normalization. We have obtained the distribution shapes by kernel density estimations for each distribution.

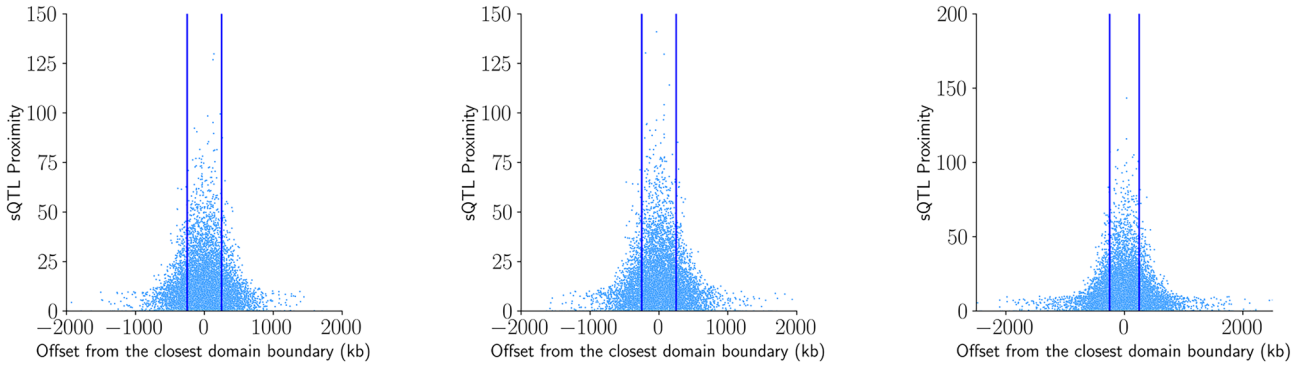
for KR and ICE normalization approaches, respectively, so statistical significance is relatively lower for KR normalization. Both normalization approaches are designed to equalize total interaction frequencies across bins, which might result in attenuating the true Hi-C signal for sQTL analysis. Such attenuation after normalization can be due to the following factors: 1- Normalization might result in loss of global contact strength: Matrix balancing methods enforce equal visibility across bins, which may erase real biological differences in contact propensity (e.g., very open chromatin vs. heterochromatin which are related with sQTL distribution), 2- Normalization might result in overcorrection: If biases correlate with biology (e.g., GC-rich regions are both highly expressed and more mappable), normalization may attenuate genuine biological signal along with technical noise, 3- Normalization may amplify noise since balancing can inflate noise in low-coverage bins by forcing them to match global expectations where sQTLs are not distributed uniformly among all bins.

Similarly, sQTL SNP bins enclosing TAD boundaries have greater total interaction frequencies than the background collection of the whole set of Hi-C bins with $p < 10^{-3}$ according to Wilcoxon rank sum test. Similar to the previous case, statistical significance via $p < 10^{-3}$ is also observed when we use a one-sided test. Moreover, we have discovered that bins enclosing both TAD boundaries and sQTLs showed greater mean total interaction frequencies than the collection of whole set of sQTL bins with $p < 0.01$ according to the Wilcoxon rank sum test, as shown in Figs. 4a-4cfor IMR90 Micro-C, IMR90 Hi-C, and liver Hi-C interaction datasets respectively. When those 6 results are considered jointly, TAD boundaries and bins enclosing sQTLs interact more frequently with the remaining chromosome intervals. As a result, bins with sQTLs might be near the remaining bins in the 3D structure of the genome. We observe similar distribution results for the remaining human cell types as well. In the following section, we particularly analyze whether sQTLs are spatially near their splicing genes.




**(a)** IMR90 cells, Micro-C **(b)** IMR90 cells, Hi-C **(c)** Liver cells, Hi-C

**Fig. 4**. The distribution and spatial proximity of sQTL bins and associated genes enclosing TAD boundaries. Distributions of the mean frequency of the whole set of Micro-C/Hi-C bins are compared against the same distribution for sQTL bins in **a**) IMR90 cells with Micro-C, **b**) IMR90 cells with Hi-C, **c**) Liver cells with Hi-C datasets. We have obtained the distribution shapes by kernel density estimations for each distribution.



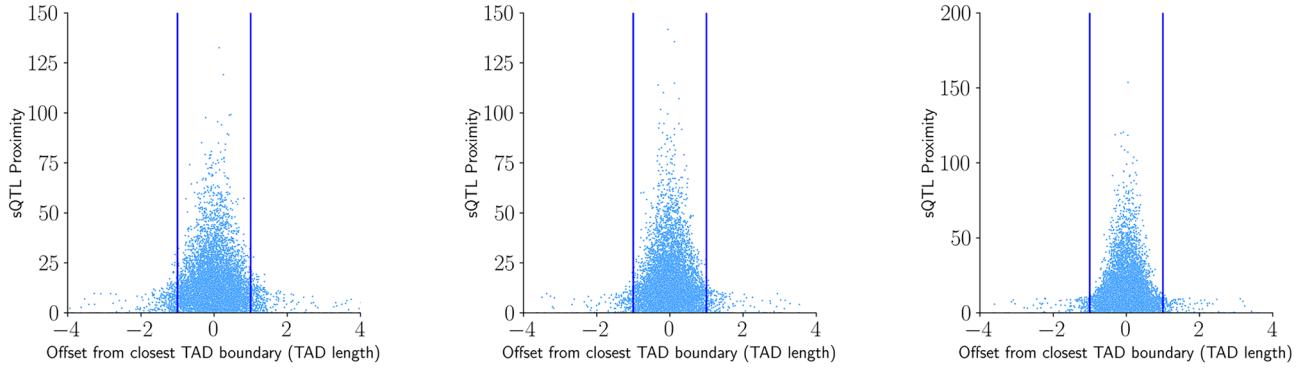

**(a)** IMR90 cells, Micro-C **(b)** IMR90 cells, Hi-C **(c)** Liver cells, Hi-C

**Fig. 5**. The associations between sQTL spatial closeness and the genomic distance to the nearest TAD boundary, plotted as two-dimensional histogram for **a**) IMR90 cells with Micro-C, **b**) IMR90 cells with Hi-C, and **c**) Liver cells with Hi-C datasets. Vertical blue lines show -250 kb and 250 kb distances from TAD boundary.

## sQTLs are genomically near TAD boundaries

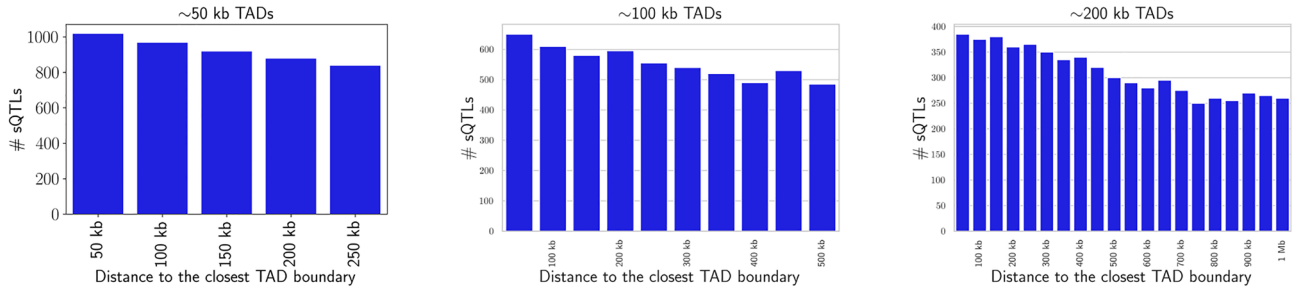
We have investigated whether sQTL bins are near TAD boundaries throughout the genome. In all our experiments, "near" indicates a closer distance compared to a control group. Although sQTLs may locate as distant as 3.5 Mb apart from the beginning of a TAD, most sQTL bins, 6208 out of 10805 (approximately $57.45\%$), locate inside only 250 kb of a TAD boundary with $p < 0.001$ for randomly-shuffled TADs as seen in vertical blue lines in Fig. 5cafor human IMR90 cells by using Micro-C dataset. We observe similar significance results when we repeat the same process over IMR90 by using Hi-C dataset instead of Micro-C as seen in Fig. 5cb. Similarly, for liver cells, most sQTL bins, 4912 out of 10243 (approximately $48\%$), locate inside only 250 kb of a TAD boundary with $p < 0.001$ for randomly-shuffled TADs as seen in vertical lines in Fig. 5c. According to these Fig. 5ca-5c, the spatial closeness $p(\tilde{q})$ of an sQTL-gene association is greater nearby the TAD boundaries. We also generate similar plots, as presented in Fig. 6a-6c, after normalizing the TAD lengths and plotting the relative position of sQTLs as a ratio of the closest TAD length. Such an analysis also provides a clearer and more generalizable illustration of eQTL proximity to TAD boundaries, as well as being in line with the previous figures without TAD length normalization.

Additionally, we have also discovered that the preferential existence of sQTLs near TAD boundaries has not been restricted to small- or large-length TADs. According to Fig. 7a-7cfor IMR90 cells, for 3 domain lengths , the distribution of the distances to the nearest TAD boundary are skewed towards 0 with $p < 10^{-4}$ according to Pearson's chi-squared test for the whole set of lengths that has been adjusted with Bonferroni correction. TAD lengths for IMR90 cells are 50 kb, 100 kb, 200 kb with the allowed standard deviation for TAD lengths is $\pm 10$ kb. We observe similar skew towards 0 for TADs from the remaining cell types Hi-C interaction datasets as well. For instance, in terms of TAD lengths 500 kb, 1 Mb, 2 Mb with $\pm 100$ kb standard deviation for liver cells, the distribution of the distances to the nearest TAD boundary are again skewed towards 0 with $p < 10^{-4}$ according to Pearson's chi-squared test as seen in Fig. 8a-8c. Plotted TAD lengths for IMR90 cells are smaller than liver cells



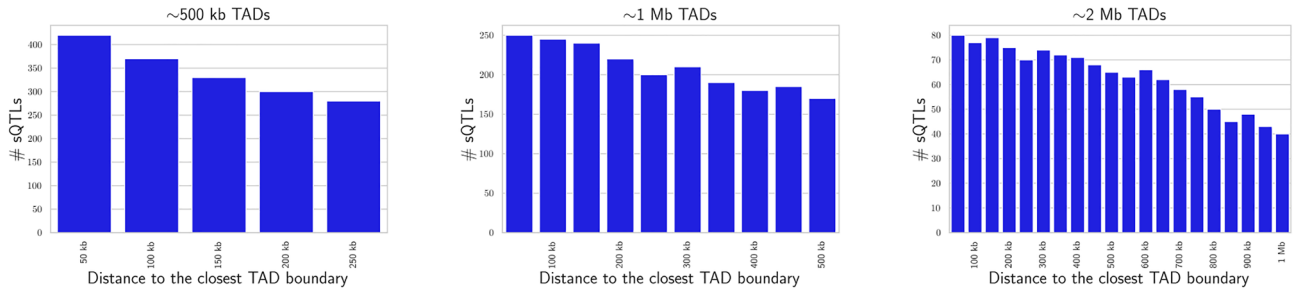

**(a)** IMR90 cells, Micro-C **(b)** IMR90 cells, Hi-C **(c)** Liver cells, Hi-C

**Fig. 6**. The associations between sQTL spatial closeness and the genomic distance to the nearest TAD boundary, plotted as a two-dimensional histogram for **a**) IMR90 cells with Micro-C, **b**) IMR90 cells with Hi-C, and **c**) Liver cells with Hi-C datasets. Vertical blue lines show the -1 closest TAD length and the 1 closest TAD length distances from the closest TAD boundary.




**(a)** 50 kb TAD **(b)** 100 kb TAD **(c)** 200 kb TAD

**Fig. 7**. Histograms of distance in genome between a SNP bin and its nearest TAD boundary for IMR90 SNP bins for that appear inside $\frac{s}{2}$ of a TAD boundary where $s$ is TAD size. The plotted TAD lengths (s) are **a**) 50 kb, **b**) 100 kb, **c**) 200 kb respectively.




**(a)** 500 kb TAD **(b)** 1 Mb TAD **(c)** 2 Mb TAD

**Fig. 8**. Histograms of distance in genome between a SNP bin and its nearest TAD boundary for liver SNP bins for that appear inside $\frac{s}{2}$ of a TAD boundary where $s$ is TAD size. The plotted TAD lengths (s) are **a**) 500 kb, **b**) 1 Mb, **c**) 2 Mb respectively.

since we use finer resolution Micro-C dataset at 5 kb resolution for IMR90 cells, whereas the experiments are run over Hi-C dataset with 50 kb resolution for the remaining cell types including the plotted liver cells.

## sQTLs are spatially near their splicing genes

We analyzed whether sQTL bins from sQTL-gene associations are spatially close to their target gene bins by utilizing the spatial closeness definition as mentioned previously. Since sQTLs could impact genes of distinct length over multiple distinct distances, and bins as part of sQTL-gene associations could differ remarkably in their total frequencies, we have paired every sQTL–gene association with its best-matching randomly-defined counterpart. In this case, a better matching corresponds to a tiny Euclidean distance between the attribute vector for sQTL and the other attribute vector associated with the random counterpart as discussed in Methods section. Each attribute vector is composed of the attributes of the equivalent sQTL: maximum total gene bin frequency, total SNP bin frequency, gene length, SNP-gene distance, and eQTL count as discussed in Methods section. In this way, we compare each sQTL with its best-maching random counterpart. We have robustly controlled for confounding factors. Once disregarded, these factors can yield less relevant, and possibly inaccurate results for sQTL-gene spatial closeness.

We define sQTL-gene associations with spatial closeness $\geq 20$ as 'high proximity' pairs. We have observed such high proximity pairs more frequently in the collection of true sQTLs than in the collection of randomly-defined optimally-matched counterparts as in Fig. 9 for IMR90 cells. We observe that pairs with spatial proximity $< 20$ occur more often in the collection of random sQTLs. Generally, the collection of matched association pairs varied statistically significantly from one another regarding spatial closeness with $p < 10^{-10}$ by Wilcoxon signed rank test for matched pairs. We observe $p < 10^{-3}$ when we use one-sided test instead of two-sided test. According to Fig. 10a-10e, the attribute vector's independent dimensions are matched quite accurately across the whole set of sQTLs where Pearson correlation coefficient for maximum total gene frequency, total SNP fragment frequency, gene length, SNP-distance, and eQTL count are $\rho = 0.96, 0.965, 0.975, 0.99, 0.99$ respectively for IMR90 cells. The Pearson correlations are remarkably soaring, even if they are not 1. This is because, the Euclidean distance between attribute vectors is used in calculating the best match, and such Euclidean distance-based matching does not need matches to be jointly perfect in each dimension of the vector. Practically, we discover that sQTLs match quite accurately with their random counterparts in multiple dimensions concurrently. Similarly, Table 1 shows the Pearson correlation coefficient for these attributes over all tested human tissues. According to this table, we still observe such remarkably soaring Pearson correlation coefficients for all remaining tissues. Additionally, Hi-C and Micro-C datasets over lung tissue have almost the same Pearson correlation coefficients.

Closer non-crossing sQTL-gene associations take place more frequently in the collection of true sQTLs. The opposite of this conclusion is correct for across-domain pairs as seen in Fig. 11a -11b for IMR90 cells, where closer across-domain pairs occur less often than the matched pairs. As a result, the general spatial closeness of sQTLs is mainly owing to non-crossing sQTL associations. When those results are considered all together, interactions between sQTLs and their far-away, inside-TAD genes, are in general enriched for higher frequency Hi-C interactions even when we control for a number of confounding variables.

## sQTLs frequently connect to splicing genes across TADs

We have analyzed the frequency of interaction between sQTL-gene pairs which do not cross TAD boundaries versus those cross TAD boundaries. Among the whole set of sQTL-gene pairs in human IMR90 cells, we have identified 9128 interactions between sQTL-gene components across TADs, whereas 6143 interactions which do not cross TADs. Even though sQTL-gene pairs that are across TADs have interaction over prolonged distances and have shorter distance in three-dimensional space than the pairs which belong to the same TAD, sQTL-gene pairs with comparably longer distances yet interact across TADs, where 1056 such sQTL-gene pairs are between 10 and 30 apart as seen in Fig. 12a. We have tested the statistical significance of the frequency of sQTLs across

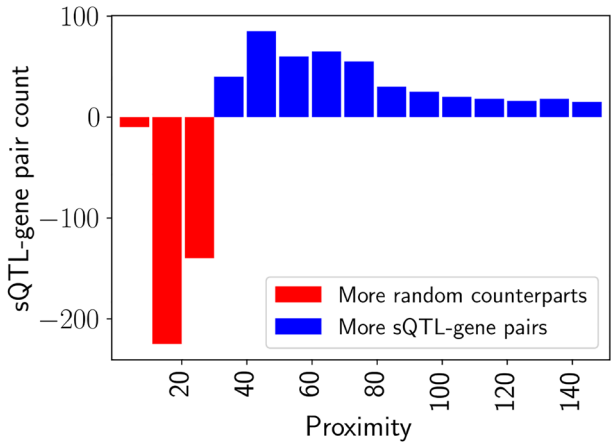



**Fig. 9**. Spatial closeness of sQTL–gene pairs once compared with their optimally bipartite matching randomly-defined counterparts in IMR90 cells. Spatial closeness bins in x-axis are plotted and colored according to whether we observe more equivalent sQTLs (blue) or more randomly chosen sQTL–gene pairs (red).

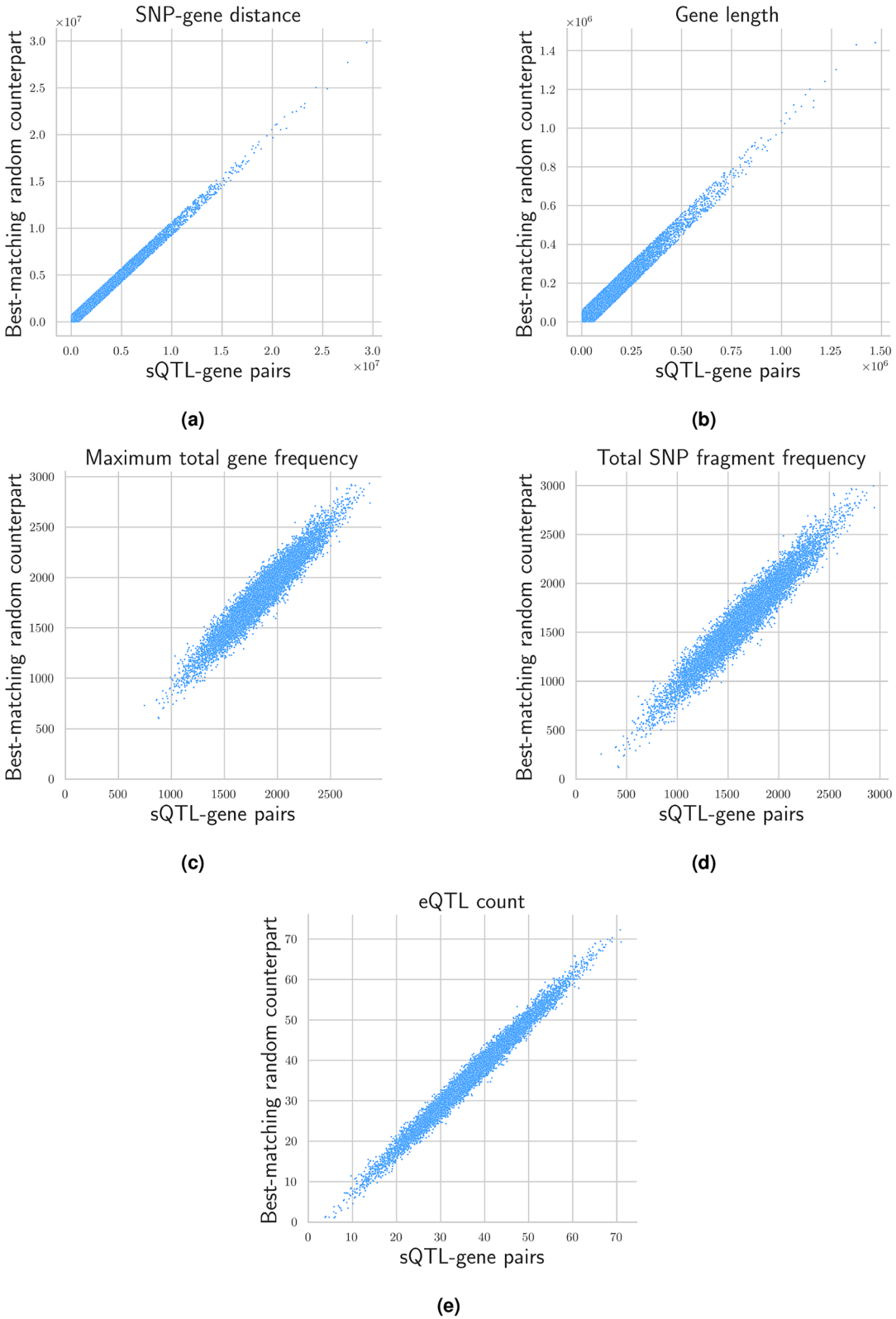



**Fig. 10**. Spatial closeness of sQTL–gene pairs once compared with their optimally bipartite matching randomly-defined counterparts in IMR90 cells. 2D correlation plots where true sQTL control attributes are on the x-axis and their random best matching counterparts are on the y-axis. We control for the following attributes: **a**) SNP–gene distance throughout the genome, (**b**) Gene length, **c**) Maximum total gene Micro-C interaction frequency, **d**) Total SNP Micro-C interaction frequency, and **e**) eQTL count.

| Cell type | SNP-gene distance | Gene length | Max total gene frequency | Total SNP frequency | eQTL count |
|---|---|---|---|---|---|
| Adrenal Gland (AD) | 0.987 | 0.985 | 0.961 | 0.975 | 0.991 |
| Dorsolateral Prefrontal Cortex (CO) | 0.983 | 0.976 | 0.962 | 0.973 | 0.992 |
| Hippocampus (HC) | 0.994 | 0.978 | 0.962 | 0.971 | 0.993 |
| Lung (LG, IMR90), Hi-C | 0.991 | 0.975 | 0.961 | 0.965 | 0.992 |
| Lung (LG, IMR90), Micro-C | 0.989 | 0.976 | 0.962 | 0.963 | 0.990 |
| Ovary (OV) | 0.992 | 0.968 | 0.967 | 0.971 | 0.985 |
| Pancreas (PA) | 0.981 | 0.971 | 0.968 | 0.968 | 0.983 |
| Small Bowel (Intestine) (SB) | 0.985 | 0.972 | 0.958 | 0.969 | 0.986 |
| Spleen (SX) | 0.992 | 0.973 | 0.971 | 0.963 | 0.979 |
| Liver (LI) | 0.993 | 0.965 | 0.973 | 0.961 | 0.988 |
| Aorta (AO) | 0.991 | 0.968 | 0.961 | 0.959 | 0.987 |
| Heart Left Ventricle (LV) | 0.978 | 0.969 | 0.964 | 0.958 | 0.986 |

**Table 1**. Pearson correlation coefficient $\rho$ between sQTLs and the attribute vector's independent dimensions (maximum total gene frequency, total SNP fragment frequency, gene length, SNP-distance, and eQTL count) across all tested human tissues.

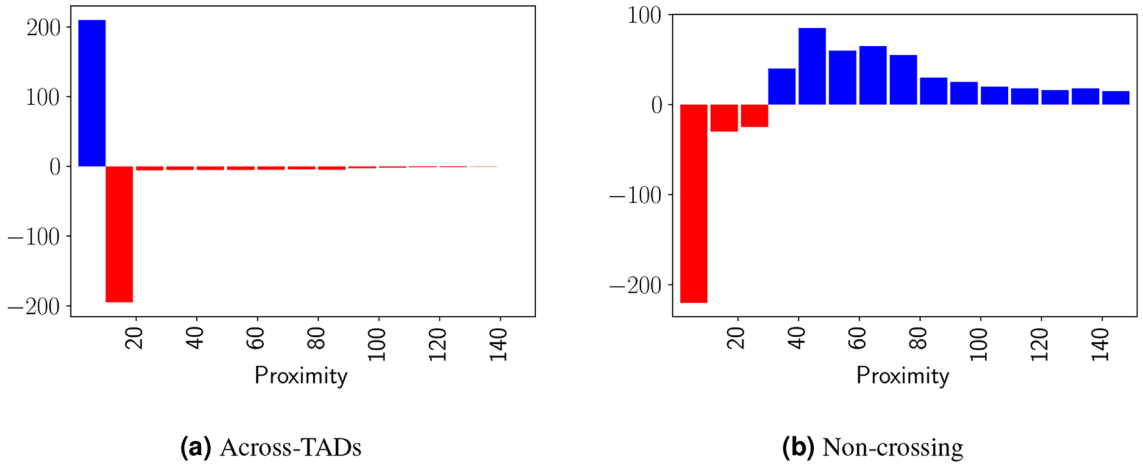



**Fig. 11**. Spatial closeness of sQTL–gene pairs once compared with their optimally bipartite matching randomly-defined counterparts in IMR90 cells for **a**) Across-TADs, **b**) Non-crossing. Spatial closeness bins in x-axis are plotted and colored according to whether we observe more equivalent sQTLs (blue) or more randomly chosen sQTL–gene pairs (red).

TAD boundaries by randomly generating ordered set of TADs while controlling for a number of attributes derived from the true set of TADs. These attributes are the frequency of SNP parts in genome crossing TADs, frequency of genes crossing TADs, the ordered list of TADs and non-TADs, and TAD and non-TAD lengths as discussed in Methods section. Similarly, Fig. 12banalyzes the frequency of interaction between sQTL-gene pairs which do not cross FIRE boundaries versus those cross FIRE boundaries. Procedure to test the statistical significance for FIREs is similar to the one for TADs.

By comparing the frequency of sQTL-gene pairs across true TADs with the frequency of matching sQTL-gene pairs crossing controlled randomly-generated domains, we have found that sQTL relationships crossing TAD boundaries appear significantly more frequently than by chance for 21 chromosomes out of 22 according to $p < 0.05$ after Benjamini–Hochberg correction as seen in Fig. 13a

for IMR90 Micro-C interaction dataset at 5 kb resolution where TADs are inferred via Armatus. Similarly, sQTL relationships crossing TAD boundaries appear significantly more frequently than by chance for 20 chromosomes out of 22 according to $p < 0.05$ after Benjamini–Hochberg correction as seen in Fig. 13b

for the same IMR90 Micro-C interaction dataset where TADs are inferred via MrTADFinder instead. Fig. 14a-14bshow the similar distribution characteristics between sQTL-gene pairs and TADs inferred via Armatus and MrTADFinder respectively for liver Hi-C interaction dataset at 40 kb resolution. Lastly, we have also found that SNP fragments that interact with multiple genes in general occur either completely across TADs or completely inside a TAD. For instance, 423 out of 3845 multiple-gene fragments, only $11\%$, include both across TAD and within TAD associations.

Table 2 shows the mean across-domain sQTL counts for the set of originally inferred TADs via Armatus over all tested human cell types according to different major alternative splicing event types, including exon skipping,

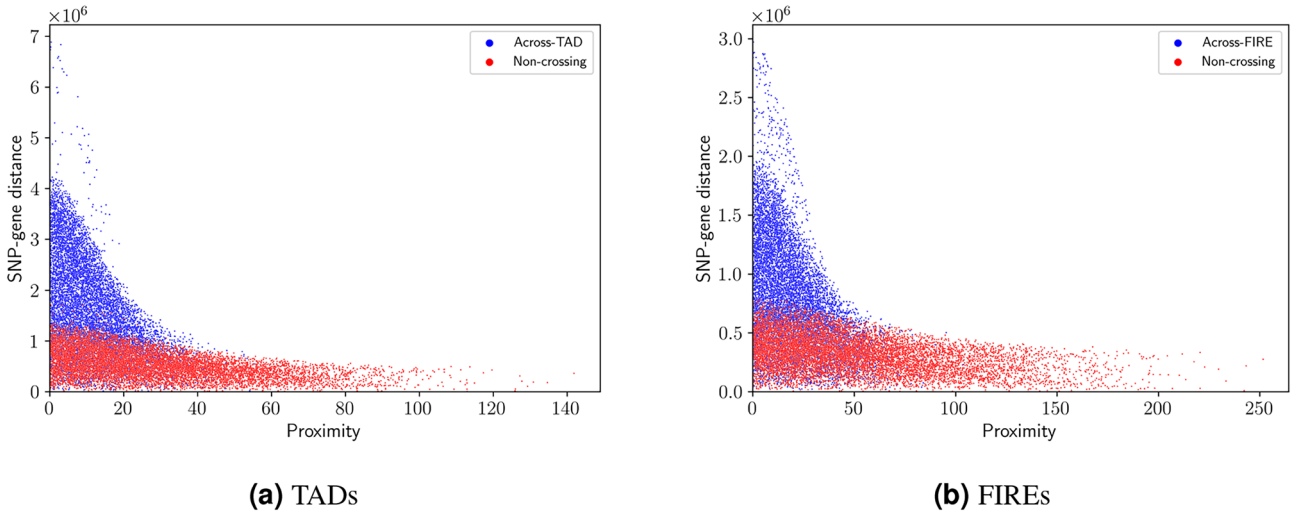



**Fig. 12**. The distribution of frequency of interaction between sQTL-gene pairs which **a**) do not cross TAD boundaries versus those cross TAD boundaries, **b**) do not cross FIRE boundaries versus those cross FIRE boundaries.

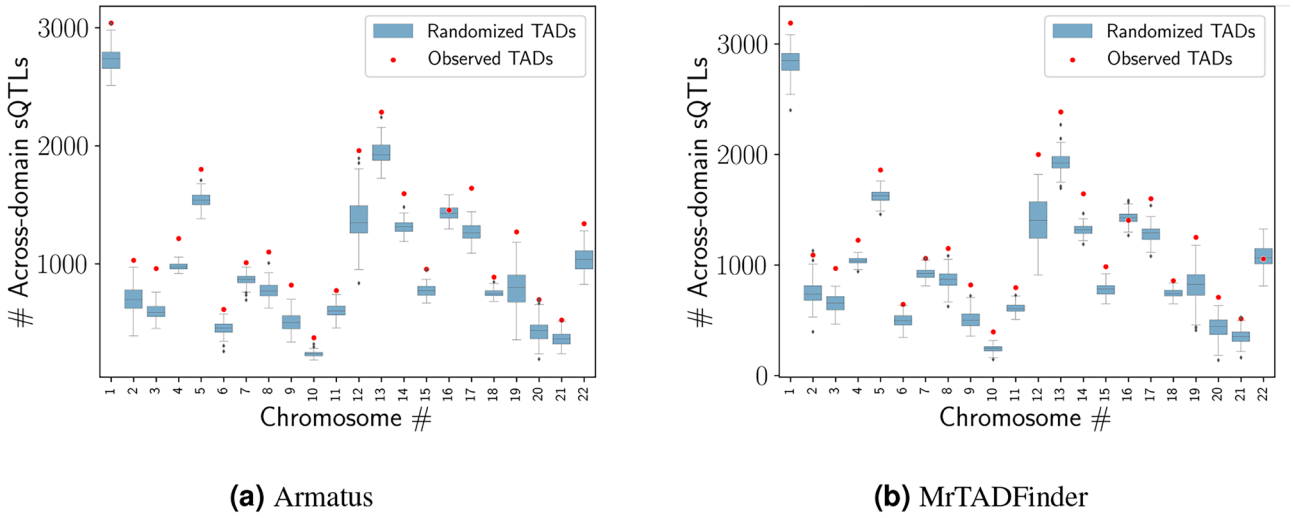



**Fig. 13**. The distribution of across-domain sQTL counts for the set of originally inferred TADs (red) versus randomly shuffled TADs (blue) for individual chromosomes ranging from 1 to 22 over IMR90 Micro-C interaction dataset at 5 kb resolution across different TAD callers: **a**) Armatus, **b**) MrTADFinder.

mutually exclusive exons, alternative donor/acceptor, and intron retention. According to the table, distributions of alternative splicing event types are in general similar across all Hi-C datasets, whereas it is different in the lung Micro-C dataset.

## Inside-TAD sQTLs with functional elements are near the corresponding splicing genes

We have discovered that distant inside-TAD sQTL interactions, including functional regulatory elements, are spatially closer than the collection of all sQTLs interacting inside TAD. In this case, we use the functional regulatory elements in ENCODE[42] that are CTCF, RAD21, JUND, TCF12[43]. Among them, CTCF is a zinc-finger architectural protein that, together with the cohesin complex (including RAD21), facilitates chromatin looping and the establishment of TADs, effectively bringing distant regulatory elements, including variants, into spatial proximity with gene bodies. In mouse embryonic stem cells, targeted degradation of CTCF led to increased inclusion of upstream exons, suggesting that CTCF binding can influence alternative splicing-possibly by inducing transcriptional pausing that impacts splice-site selection[44]. RAD21's loading dynamics strongly affect loop extrusion: overexpression leads to altered 3D structure, impacting transcription globally[45]. CTCF/cohesin-mediated looping provides plausible mechanistic support for how sQTLs at distal regulatory sites might influence splicing and gene expression-including through transcriptional pausing. Though direct evidence

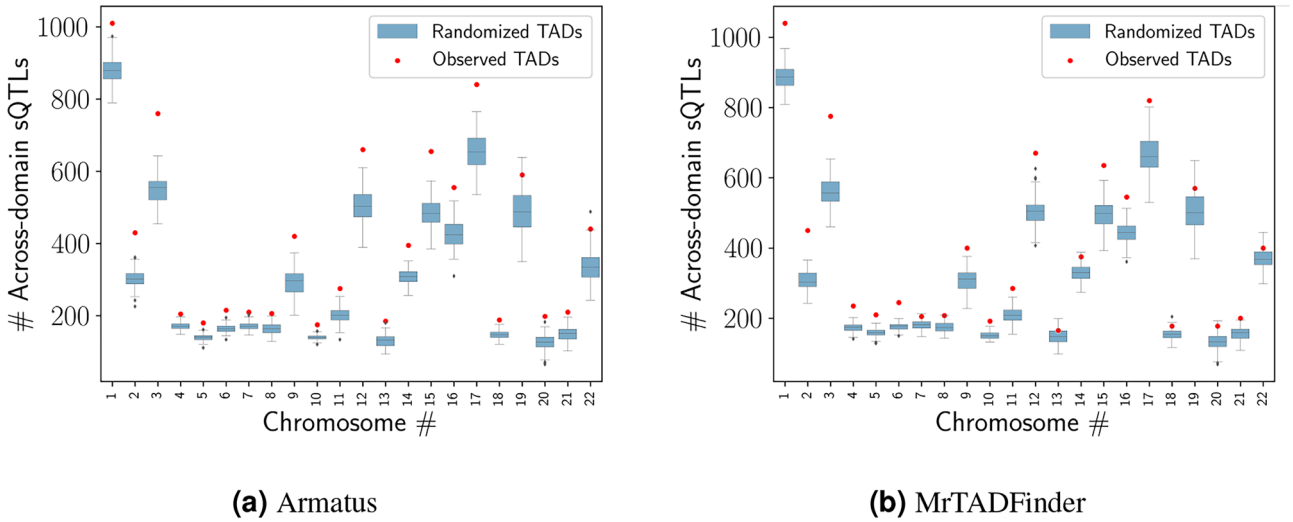



**Fig. 14**. The distribution of across-domain sQTL counts for the set of originally inferred TADs (red) versus randomly shuffled TADs (blue) for individual chromosomes ranging from 1 to 22 over the liver Hi-C interaction dataset at 40 kb resolution across different TAD callers: **a**) Armatus, **b**) MrTADFinder.

| Cell type | Exon Skipping | Mutually exclusive exons | Alternative donor/acceptor | Intron retention |
|---|---|---|---|---|
| Adrenal Gland (AD) | 675.45 | 56.34 | 61.23 | 31.45 |
| Dorsolateral Prefrontal Cortex (CO) | 665.12 | 59.12 | 63.45 | 32.91 |
| Hippocampus (HC) | 685.13 | 55.11 | 65.88 | 32.55 |
| Lung (LG, IMR90), Hi-C | 678.13 | 58.11 | 63.33 | 32.01 |
| Lung (LG, IMR90), Micro-C | 2124.24 | 123.56 | 165.32 | 67.23 |
| Ovary (OV) | 661.18 | 55.35 | 60.90 | 31.56 |
| Pancreas (PA) | 664.23 | 55.35 | 57.25 | 32.35 |
| Small Bowel (Intestine) (SB) | 691.11 | 57.75 | 65.09 | 34.16 |
| Spleen (SX) | 680.06 | 57.72 | 62.63 | 32.99 |
| Liver (LI) | 711.41 | 66.21 | 71.34 | 39.56 |
| Aorta (AO) | 688.27 | 58.15 | 64.44 | 35.68 |
| Heart Left Ventricle (LV) | 668.12 | 63.45 | 68.99 | 32.55 |

**Table 2**. The average across-domain sQTL counts for the set of originally inferred TADs via Armatus over all tested human cell types according to different alternative splicing event types, including exon skipping, mutually exclusive exons, alternative donor/acceptor, and intron retention.

tying CTCF or RAD21 to specific sQTL-mediated splicing events is still limited, the chromatin contexts they establish are functionally consistent with regulatory mediation. On the other hand, JUND is part of the AP-1 transcription factor family. It binds regulatory DNA and affects chromatin accessibility and transcriptional dynamics. TCF12 is a bHLH transcription factor (E-protein family), widely expressed in immune tissues, brain, muscle, and known to form heterodimers with other E-proteins to regulate lineage-specific gene expression. While JUND and TCF12 may co-locate with splicing-associated genomic regions, there's no clear evidence that they specifically mediate sQTL effects on alternative splicing.

There are 523 sQTLs related with those functional regulatory elements in IMR90 cells. Among them, 206 relate with genes that are more than 10 kb apart from the SNP, making up 165 equivalence classes. Among these classes, 123 are non-crossing, exist frequently inside TADs, and are associated with a remarkably closer distance than the remaining collection of non-crossing sQTLs as seen in Fig. 15 for IMR90 cells with $p < 0.001$ according to Wilcoxon rank sum test as well as one-sided test. We observe similar results for the remaining cell types other than IMR90 cells as well. As shown in Fig. 16, we also observe a similar relationship across non-crossing eQTLs and non-crossing eQTLs with functional elements when we repeat the same analysis for eQTLs, which has also been discussed in[28,46].

To reconcile the observation that many functional regulatory element (FRE)–sQTL pairs reside within the same TAD with the broader cross-TAD enrichment, we stratified sQTL–gene pairs by whether the sQTL overlapped annotated FREs (enhancers, promoters, and sites of CTCF, RAD21, JUND, or TCF12 binding) and compared their 3D proximity patterns. FRE-linked sQTLs showed a strong intra-TAD bias, with 72% of

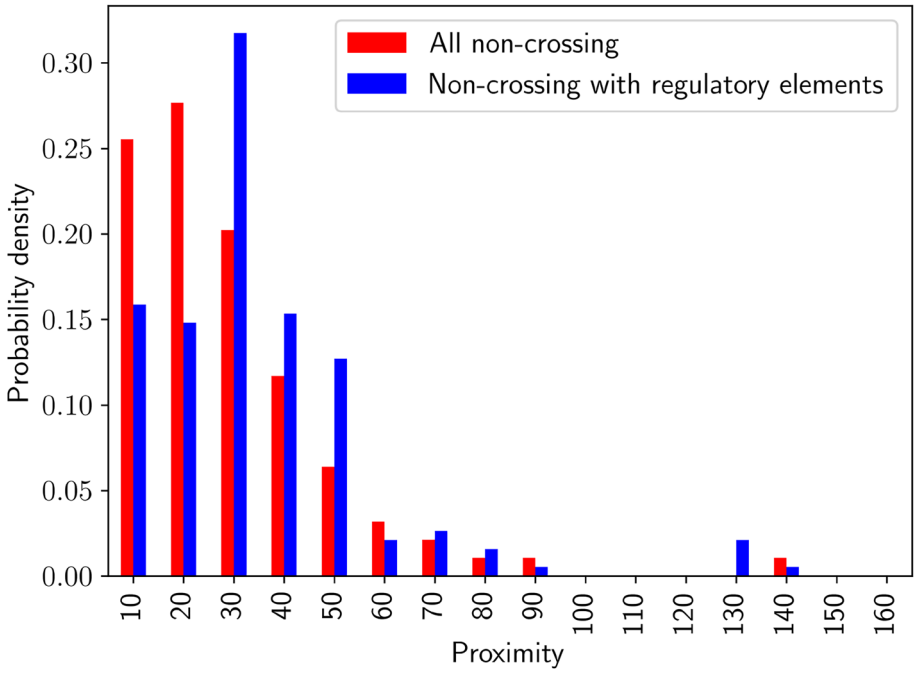



**Fig. 15**. Distribution of the set of all non-crossing sQTLs in red color with non-crossing sQTLs with functional regulatory elements in blue color in IMR90 cells.

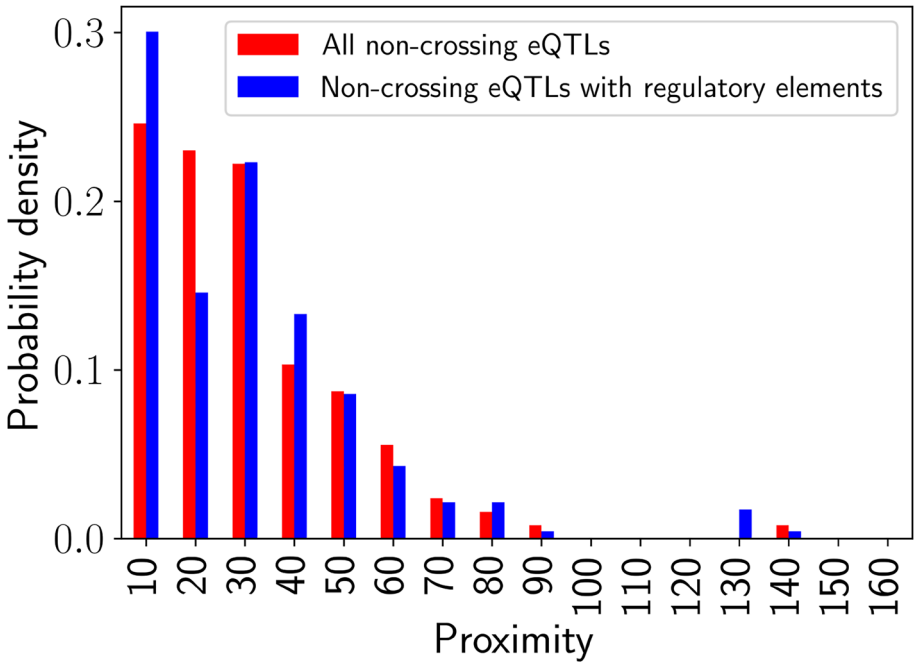



**Fig. 16**. Distribution of the set of all non-crossing eQTLs in red color with non-crossing eQTLs with functional regulatory elements in blue color in IMR90 cells.

pairs remaining within the same TAD compared to $48\%$ for non-FRE sQTLs (Fisher's exact test, $p \leq 10^{-6}$). In contrast, non-FRE sQTLs exhibited higher cross-TAD representation, explaining much of the overall cross-TAD enrichment previously reported. When we repeated this analysis for eQTL–gene pairs in the same cell types, the intra-TAD fraction was markedly higher ($81\%$) and the cross-TAD fraction lower than for sQTLs, indicating that splicing QTLs display more long-range contacts than eQTLs. These stratified analyses demonstrate that the apparently conflicting patterns reflect different classes of regulatory interactions: FRE-associated sQTLs act largely within TADs, while a distinct subset of non-FRE sQTLs contributes to cross-TAD connectivity.

To further investigate the regulatory context of sQTL loci, we performed an in silico analysis using publicly available ENCODE ChIP-seq datasets for CTCF, RAD21, JUND, and TCF12. We overlapped sQTL bins with ChIP-seq peaks in the corresponding cell types and compared the frequencies against matched random controls generated by our bipartite matching framework. sQTL bins were significantly enriched for CTCF (fold enrichment $2.3\times$, $p < 10^{-6}$) and RAD21 ($1.9\times$, $p < 10^{-4}$) relative to matched controls. Moderate but significant enrichment was also observed for JUND ($1.4\times, p = 0.002$) and TCF12 ($1.3\times, p = 0.006$). In addition, motif enrichment analysis using HOMER revealed overrepresentation of CTCF and RAD21-associated cohesin motifs within $\pm 500$ bp of sQTL sites ($q < 0.01$ after Benjamini–Hochberg correction). Stratifying spatial proximity by factor overlap showed that sQTL–gene pairs intersecting CTCF or RAD21 sites are significantly closer in 3D than other inside-TAD sQTLs (Wilcoxon signed-rank test $p < 10^{-3}$). These findings suggest that the structural context created by loop extrusion and boundary insulation contributes to the spatial coupling between sQTLs and their splicing targets.

## Comparison with regression-based approach

To analyze the robustness of our results, we also compared our proposed approach with the modified version of another regression-based approach for eQTLs[28]. In comparison,[28] normally does not take into account sQTLs, where their study is intended to indirectly analyze the relationship between only eQTLs and genome 3D structure. To make the comparison analysis fair, we modified their regression-based approach and added sQTLs as another covariate. Table 3 shows how statistical significance changes across different cell types. Even though we find significant results to be similar across several cell types, their approach does find statistically significant relationships in some cell types and experiments, such as Adrenal Gland (AD) Hi-C and Lung Micro-C (LG, IMR90). In those cell types, our approach finds statistically significant results exhibiting the promising nature of our statistical approach. Our proposed method identifies significant associations in scenarios where regression-based approaches fail. Given the stricter equivalent counterpart controls used in this study, the key contributing factor of our significance seems to be the resampling-based optimal bipartite matching strategy employed by the authors, which carefully controls for confounding factors without depending on restrictive analytical null models. This design likely enhances the method's ability to detect subtle spatial relationships between sQTLs and chromatin features, which might be overlooked by regression-based approaches that rely on more global, parametric assumptions.

To ensure that the higher number of significant sQTL–gene pairs detected by our matching framework does not result from inflated type I error, we applied Benjamini–Hochberg correction to all p-values. The majority of additional associations remained significant at FDR $\leq 5\%$, comparable to or better than the regression baseline. We further examined the subset of sQTL–gene pairs detected only by our approach. These additional pairs were significantly enriched for functional regulatory marks (enhancers, DNase I hypersensitive sites; fold enrichment $2.1\times, p \leq 10^{-5}$) and for binding sites of CTCF and RAD21 compared to matched random controls, supporting their biological relevance. While our framework identifies more significant associations than a regression-based baseline, we took steps to ensure this increase reflects a genuine signal. Additional pairs pass stringent FDR control, replicate across TAD-calling strategies and chromatin datasets, and show strong enrichment for functional regulatory features. Together, these checks suggest that our higher sensitivity does not arise from false positives but rather from improved detection of subtle but biologically meaningful spatial relationships.

## Discussion

sQTL-gene relationships are mapped to chromatin's three-dimensional genome structure by utilizing chromosome conformation capture interaction graph. Via such mappings, our research brings forth strong evidence that spatial closeness is an important element in the way a mutation impacts its correlated gene's

| Cell type | sQTL coef. and significance by modified[28] | Significance by our approach |
|---|---|---|
| Adrenal Gland (AD) | 0.162 ($p = 0.12$) | $p = 0.00074$ *** |
| Dorsolateral Prefrontal Cortex (CO) | 0.173 ($p = 0.0071$ **) | $p = 0.024$ ** |
| Hippocampus (HC) | 0.158 ($p = 0.0062$ **) | $p = 0.00066$ *** |
| Lung (LG, IMR90), Hi-C | 0.167 ($p = 0.017$ *) | $p = 0.0072$ ** |
| Lung (LG, IMR90), Micro-C | 0.154 ($p = 0.16$) | $p = 0.032$ * |
| Ovary (OV) | 0.183 ($p = 0.0075$ **) | $p = 0.0091$ ** |
| Pancreas (PA) | 0.167 ($p = 0.022$ *) | $p = 0.0089$ ** |
| Small Bowel (Intestine) (SB) | 0.175 ($p = 0.0065$ **) | $p = 0.00077$ *** |
| Spleen (SX) | 0.177 ($p = 0.043$ *) | $p = 0.034$ * |
| Liver (LI) | 0.183 ($p = 0.034$ *) | $p = 0.041$ * |
| Aorta (AO) | 0.170 ($p = 0.0082$ **) | $p = 0.0054$ ** |
| Heart Left Ventricle (LV) | 0.125 ($p = 0.0088$ **) | $p = 0.00021$ *** |

**Table 3**. Comparison of our statistical approach with the modified version in[28] where we provide both sQTL coefficient and its statistical significance. In terms of statistical significance, * represents $p < 0.05$, ** represents $p < 0.01$, and *** represents $p < 0.001$.

expression. This impact is mainly achieved by regulatory genomic segments inside TADs. We have defined several confounding factors which would have resulted in biased inferences while associating sQTLs with the chromatin's three-dimensional spatial structure. We have been in control of such factors by utilizing a resampling-based optimal bipartite matching framework. Our optimal matching-based approach is robust due to introducing a direct way to control for confounding factors. Additionally, it also does not depend upon a null model's analytical expression. Under this controlled situation, justification of such an analytical definition would have been very tough. Such an analytical definition can also straightforwardly lead to statistical significance when the introduced resampling procedure would not lead[47].

Non-crossing sQTL relationships can practically explain nearly the whole set of enrichment in the observed three-dimensional spatial closeness when we compare them with their best-matching shuffled controls as seen in Fig. 9. According to this result, those sQTL components are highly probably near of each other in three-dimensional space. sQTL components interacting across TADs can achieve such behavior by additional temporary interactions among tightly-coupled TADs. In alignment with such observation, the fact that major proportion of sQTLs interacting entirely not across or across TADs can be intimately related to TADs tightly-packed characteristics. sQTLs which interact on further tightly-packed chromatin regions, have increased chance of being not accessible to chromatin regions that are further away throughout the genome than sQTLs over a region with increased accessibility near TAD boundaries. Such finding is in line with TAD boundaries being more accessible, which is brought about via DNase I sensitivity[37]. Hi-C interactions are enriched at TAD boundaries by definition, which in part clarifies our finding of sQTL fragments having in general inflated frequency since sQTLs occur nearby TADs. However, such high occurrence of sQTLs near TAD boundaries cannot be clearly explained. Lastly, sQTLs across domains are prevalent which shows that the difference in gene alternative splicing and the resulting expression patterns can be remarkably explained via mutations over open chromatin neighborhoods nearby TAD boundaries.

We have observed the robustness of our findings with respect to dissimilar sQTL subsets, TAD arrangements, and chromatin interaction datasets. When we use Hi-C interaction dataset for remaining cell types including liver instead of IMR90 Micro-C interaction dataset, we have again found almost indistinguishable results. Nonetheless, the distributions of spatial closeness have been not so much segregated, but each distribution has again been significantly distinct from the other distribution. The collection of tested sQTLs have been specific to the individual cell type, but we have observed that analysis results over the aggregated collection of sQTLs among different cell types are quite similar. Even though we have utilized a number of different experiments with fewer samples that have reproduced the same conclusions, they have not always generated the similar domain-length-specific TAD boundaries charts as in Fig. 7a-7c. Those experiments have resulted only in a smaller set of chromosomes to be enriched for across-domain sQTLs as seen in Fig. 13a-13b. We have reached out to a similar conclusion when we repeat the same experiments over alternative collection of TADs from different approaches[24,37,48,49]. Additionally, the estimated test statistic for number of cross-domain sQTLs have been significant to a greater extent than the statistic estimated over initial collection of TADs. When considered jointly, those results show that the observed features are not due to specific TAD arrangements, interaction dataset, or collection of sQTLs.

The spatial closeness of sQTLs across and inside TADs can support the correlations between sQTLs and other chromatin attributes which are discussed and analyzed only for eQTLs instead[33,50].[33] supports this conclusion by utilizing Hi-C datasets to infer functional regulatory elements and these elements promoters. Via our experiments, we have figured out that sQTLs nearby functional regulatory parts of genome have been enriched in terms of spatial closeness to the corresponding targets. More clearly, such sQTLs are nearer to their inside-TAD target splicing genes than the collection of the whole set of sQTLs occurring inside TAD. sQTLs could have also been nearby functional regulatory parts of genome which are yet to be inferred. As a result, such closeness test may demonstrate the specificity of spatial closeness enrichment results to the functional regulatory parts and transcription factors for identification of those parts, in regards to our dataset. We also observe higher correlation between SNPs of sQTL associations and single point mutations over DNaseI sensitivity sQTLs, where DNaseI sensitivity sQTLs correspond to mutations that are related with chromatin accessibility variations. Such high correlation has previously been observed for SNPs of eQTL associations instead[51]. As a result, these DNaseI sensitivity QTLs can also be related to TADs, similar to how we have identified sQTLs as associated to TADs.

One of our main observation of correlation between TADs and sQTL locations may underpin a different up-to-date hypothesis: Disruption of the established TAD structure nearby an sQTL may end up with a remarkable difference in the alternative splicing pattern of its associated gene. For instance, if an sQTL's splicing gene locates inside a TAD and such sQTL is also linked to an enhancer (for instance, an sQTL from the collection of inside-TAD sQTLs associating with functional regulatory elements that are examined earlier), we can test such a hypothesis over an experimental method which can split up the corresponding sQTL parts such that the splicing gene and mutation will not be near each other in 3D space. The experimental method in[52] allows one to test this hypothesis by separating and merging TADs.

Chromatin organization's three-dimensional geometry has layed out many intuitions regarding nuclear organization, cancer, and gene expression regulation[36,53–62]. In this study, we find strong evidence in favor of coupling three-dimensional chromatin shape with sQTLs location and distribution throughout the genome. We can also apply the methods developed in this research to different genomic datasets where there are pairwise interacting items such as promoter-enhancer pairs. Our approach may also be modified to control for extra confounding factors without much effort. The higher number of high-resolution interaction datasets specific to tissues have been generated, the proposed approach can be more beneficial in understanding the differential influence of chromatin structure onto pairs of associations over different cell types in more detail.

Lastly, our enrichment analysis supports a model in which CTCF and RAD21, as core components of the loop extrusion machinery, help establish TAD boundaries and bring distal regulatory variants into proximity

with splicing genes. This agrees with prior evidence that perturbing CTCF alters exon inclusion patterns and chromatin contacts. The observed JUND and TCF12 enrichment indicates that AP-1 and bHLH transcription factors may mark regulatory hubs of open chromatin where sQTLs preferentially occur, potentially influencing splice-site selection through local transcriptional dynamics. Although these observations are correlative, they embed our spatial proximity findings within known chromatin biology and suggest testable hypotheses for future perturbation studies. Incorporating these in silico validations strengthens the biological interpretation of our increased sensitivity relative to regression models. While direct experimental disruption of these factors is outside the scope of the current work, our results highlight plausible molecular intermediates linking long-range chromatin architecture to splicing regulation.

## Methods

### Chromatin interactions, topologically associating domains, and FIREs

We have built chromatin interaction graphs for multiple different 11 cell types in human[34,35]; Adrenal Gland (AD), Dorsolateral Prefrontal Cortex (CO), Hippocampus (HC), Lung (LG, IMR90), Ovary (OV), Pancreas (PA), Small Bowel (Intestine) (SB), Spleen (SX), Liver (LI), Aorta (AO), Heart Left Ventricle (LV) as summarized in Table 4 which include sQTL and eQTL datasets for these cell types as well. Among these interaction datasets, all of them except the lung tissue are Hi-C datasets whereas the lung tissue is from both Hi-C[34] and finer resolution Micro-C[35] experiments. We have constructed the graphs for every single chromosome from 1 to 22 independently, since we can analyze intrachromosomal chromatin interactions at a significantly higher resolution than interchromosomal ones[63].

We represent raw Hi-C or Micro-C dataset for every chromosome as a weighted graph $G = (V, E, I)$. In this graph, $V$ define set of restriction sites in the experiment, $E$ define collection of observed interactions among $V$, and $I(e)$ for every edge $e = \{a, b\} \in E$ define the interaction frequency between restriction sites $a$ and $b$. In this graph, $I(e)$ correspond to the Hi-C/Micro-C read pair count mapping to the restriction sites $a$ and $b$. Hi-C data inherently has a resolution limit, so we have built a modified graph $G' = (V', E', I')$ where all set of chromatin interactions are binned at 40 kb resolution. Similarly, we have built the modified graph for Micro-C dataset at 5 kb resolution. In this graph, nodes in $V'$ divide up a chromosome into 40 and 5 kb bins for Hi-C and Micro-C datasets respectively which correspond to contiguous intervals or fragments on the chromosome. The interaction frequency $I'(e') = I'(u, v)$ where $e' = \{u, v\} \in E'$ is defined as the sum of chromatin interaction counts for all interactions where read pairs are mapped inside the bins $u$ and $v$. $t(v) = \sum_{u \in \mathrm{Neighbors(v)}} I'(u, v)$ represents the bin $v$'s total frequency where $\mathrm{Neighbors(v)}$ is the set of neighbors of bin $v$ in the binned interaction graph $G'$. We have normalized our high coverage datasets by using the error-correction method by[65].

We have inferred TADs over each chromosome at a genome-wide scale by a number of TAD identification methods[24,37,48,49,66]. Among them,[37] follows hidden markov model-based method while inferring topological domain boundaries $D$. Each topological domain in $D$ represents a contiguous interval $[a_i, b_i], 1 \leq a_i < b_i \leq n$ where none of different TADs intersect with each other and $n = |V'|$ is the size of bins in $G'$. As an another method, we have also run Armatus[24] by setting $g = 0.5$ to infer TAD boundaries over the analyzed human cell types. Armatus[24] defines a quality score for each possible domain, and infers TADs by optimizing the overall quality score via dynamic programming. In addition, Armatus infers TADs persisting across different scales. Lastly, we have also inferred TADs by Arrowhead[48] via using its default parameter and MrTADFinder[49] by setting $res = 1.5$ parameter in order to test the robustness of results. All these methods generate different TADs, but all of these TADs are also significantly enriched for same chromatin features at a genome-wide scale. Among these methods, we mainly report the results by using Armatus unless mentioned otherwise.

Furthermore,[67] has recently extracted frequently interacting special regions called Frequently Interacting Regions (FIREs) that are also hotspots of local interactions between chromatin segments over Hi-C dataset. The existence and distribution of those FIREs are strongly tissue-specific, they are closely associated with cell-type-specific gene regulation, and they are defined as bins which often interact with the closeby segments that are less than 200 kb apart. In our case, we have inferred FIREs over each chromosome at a genome-wide scale for all 11 human cell types by running FIREcaller[68] with the default parameters.

| Cell type | Hi-C/Micro-C | sQTL | eQTL |
|---|---|---|---|
| Adrenal Gland (AD) | 34 | 9 | 64 |
| Dorsolateral Prefrontal Cortex (CO) | 34 | 9 | 64 |
| Hippocampus (HC) | 34 | 9 | 64 |
| Lung (LG, IMR90) | 34,35 | 9 | 64 |
| Ovary (OV) | 34 | 9 | 64 |
| Pancreas (PA) | 34 | 9 | 64 |
| Small Bowel (Intestine) (SB) | 34 | 9 | 64 |
| Spleen (SX) | 34 | 9 | 64 |
| Liver (LI) | 34 | 9 | 64 |
| Aorta (AO) | 34 | 9 | 64 |
| Heart Left Ventricle (LV) | 34 | 9 | 64 |

**Table 4**. Data summary for all analyzed human tissues.

## eQTLs

In this study, we have focused on analyzing the relationship between sQTLs and high order 3D chromatin structure. But, we also control for eQTLs in our analysis. We have obtained eQTLs for all these 11 cell types from GTEx study[64], which includes eQTLs shared across multiple human cell types as well as the ones specific to each considered human cell type. As part of preprocessing, we have overlapped the IDs in Ensembl database[69] with gene names in this GTEx dataset.

## sQTL equality classes

We have obtained sQTLs from[9], which includes sQTLs shared across multiple human tissues as well as the ones specific to each considered human tissue (https://doi.org/10.5281/zenodo.4058759). As part of preprocessing, we have overlapped the IDs in Ensembl database[69] with gene names in this dataset. Then, we have generated a set $Q$ of sQTLs by selecting genes which has a distinct range in the Ensembl database. We have also analyzed the relationship between 3D chromatin structure and sQTLs for different splicing event types such as exon skipping, alternative 5' and 3' splice sites, intron retention, mutually exclusive exons, alternative first and last exons, alternative 5' and 3' UTRs, tandem 5' and 3' UTRs, complex splicing events (complex combinations of events affecting internal exons) and complex 5'/3' events (complex combinations of events affecting 5'/3' termini)[9]. Some of these events are not explicitly involving splicing, but alternative transcription initiation and termination sites. We have detected the splicing event type associated with each sQTL via sQTLseekeR[70].

An sQTL-gene pair $q \in Q$ represents a SNP coupled with its associated gene. Each individual gene may be associated with multiple SNPs where all such SNPs are located on the same bin that is called as a SNP or sQTL bin. In the binned chromatin interaction graph, all of those sQTL associations would be indistinguishable, and as a result the interaction frequencies for all of them will be the same throughout the interaction network. Consequently, we have come up with equivalence classes $\tilde{q} = \{q' \in Q : q' \sim q\}$ where $q' \sim q$ when SNPs in $q$ and $q'$ map to same node in the interaction graph $G'$ and their mapped target genes are same. Then, we have defined a new collection of gene-sQTL pairs $\tilde{Q} = \{\tilde{q}, \forall\, q \in Q\}$ via equivalent gene-sQTL pairs subsets. Every gene-sQTL equality class $\tilde{q}$ includes a gene name, SNP bin, and a collection of bins where the gene occurs.

Let $t(v)$ represent the sum of edge frequencies incident to $v$ on $G'$. In our analysis, considering a gene-sQTL pair $\tilde{q}$, we define the following 6 attributes:

1. **Spatial proximity** ($p(\tilde{q})$): This attribute defines the closeness between SNP and gene centred on chromatin interaction graph. It defines the largest frequency $f'(e)$ of the whole set of interactions $e \in E'$ between the single nucleotide polymorphism bin and bins intersecting with the gene.
2. **Total SNP frequency** ($t_s(\tilde{q})$): This attribute defines the magnitude of interaction of the SNP associated with the sQTL with the remaining bins in the binned interaction graph. It is $t(s)$ where $s$ is the SNP bin in $G'$ associated with $\tilde{q}$.
3. **Maximum total gene frequency** ($t_g(\tilde{q})$): This attribute defines the magnitude of interaction of the gene associated with the sQTL with the remaining bins in the binned interaction graph. $t_g(\tilde{q})$ is the largest $t(v)$ of any bin $v \in G'$ intersecting with the gene over that sQTL.
4. **Gene length** ($l(\tilde{q})$): the genomic distance between the middle points of the initial and rearmost gene bins.
5. **eQTL count** ($e(\tilde{q})$): the number of eQTLs across fragments associated with gene-sQTL pair $\tilde{q}$. This attribute is mainly to control for eQTL effects.
6. **Gene-SNP distance** ($d(\tilde{q})$): the distance across genome between the middle point of SNP bin and the middle location of the nearest gene bin to the SNP.

## Optimal bipartite matching

We tested for statistical significance by comparing the sequence of TADs $D$ and the collection of sQTL-gene pairs $Q$ to their randomly-generated equivalents. This optimal matching framework and the subsequent analysis are similar to the one discussed in[21]. As part of our matching, we control for confounding factors associated with the genomic features of sQTL–gene pairs, features associated with TADs, and interaction patterns common across Hi-C experiments. Then, we have matched the sequence of TADs $D$ with a collection of randomly defined domain sequences $\Omega_D$ and have matched the sQTLs in $Q$ with randomly defined sQTLs $\Omega_{\tilde{Q}}$ as the confounding factors between the matched items are kept as similar as possible.

As part of our matching process, each observed item $d \in D$ or $i \in \tilde{Q}$ is associated with a confounding factors numerical attribute vector $x_i$. Similarly, each item $j$ in the sample spaces $\Omega_D$ or $\Omega_{\tilde{Q}}$ is also associated with a numerical attribute vector of the same confounding factors $x'_j$. Attributes for $x$ are defined differently for domains and sQTLs. We match each observed item $i$ to $k$ random items $j$ by using a maximum weighted bipartite matching solver[71], such that the optimization algorithm in the solver minimizes the summation of Euclidean distances $\sum_{i,j} \|x_i - x'_j\|$ for all matched pairs $(i, j)$ as seen in Fig. 17a.

In our optimal matching framework, for each real sQTL–gene pair, we define an attribute vector capturing biologically relevant features. These features were selected because they represent major known drivers of 3D contact probability and regulatory confounding: genomic distance and gene size strongly influence Hi-C contact maps; local interaction frequency proxies for inherent chromatin accessibility; and eQTL density helps adjust for expression-linked confounding. Matching on these controls makes the random "control loci" as comparable as possible to the real sQTLs in basic genomic and interaction features before testing for spatial proximity. While our matching framework is robust and more flexible than regression-based null models, it is still constrained by the features available. The approach does not depend upon a null model's analytical expression, but defining a perfect null would be very tough because other unmeasured epigenomic factors could still influence results. In addition, the matching was not extended to all possible functional marks or epigenomic states, since the focus

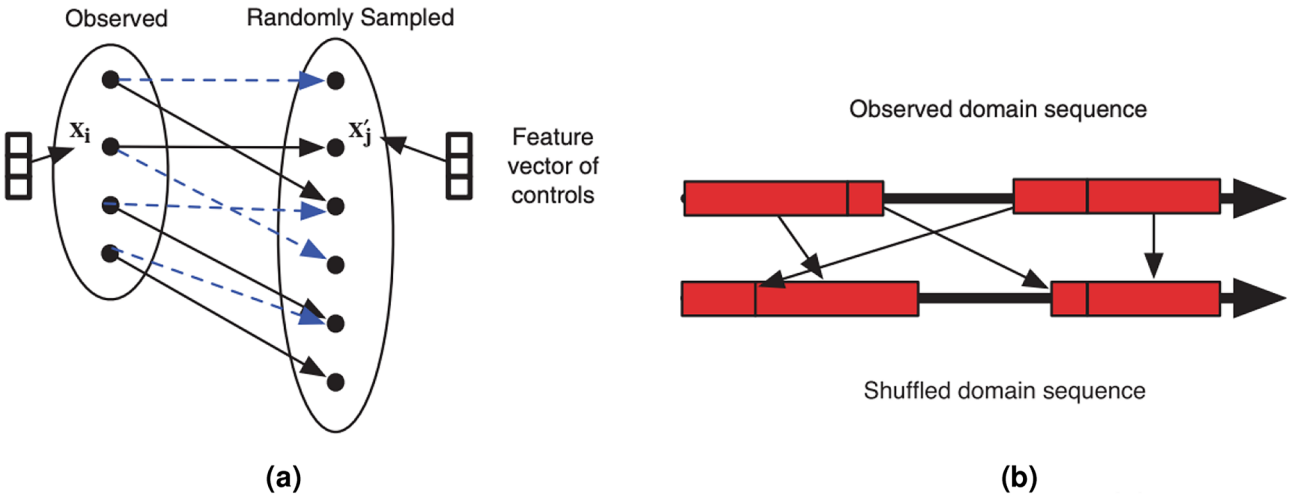



**Fig. 17**. Optimal bipartite matching structure and TAD shuffling process. **a**) We pair the observed items (either sQTLs or an arrangement of TADs) with their 2 ($k = 2$) best-matching equivalents in the control set: dotted arrows in blue color represent the second-best match, wheras the arrows in black color represent the best match. Each true and controled item (node in the graph) has an attribute vector consisting of controlled factors, and smaller Euclidean distance between true and controlled item's vectors mean a better match. **b**) A demonstration of a TAD shuffling process where TADs are shown in red, whereas gaps between TADs (non-TADs) throughout the chromosome are shown in black. We perform the shuffling in a way that TAD and gap lengths are preserved, as well as the true sequence of TAD and non-TADs are also preserved.

was on key structural and regulatory covariates (distance, length, local interaction, and eQTLs). Other features, such as histone modifications or DNA methylation, were not explicitly included and might refine the control design if added.

### Sample space of spatial proximity

In order to increase the likelihood of closer matches between the randomly-generated and real sQTLs over a certain chromosome, the sample space $\Omega_{\tilde{Q}}$ is selected as substantially greater than sQTLs collection $\tilde{Q}$. Let $n_p$ be a ratio of the randomly-generated collection of sQTLs size to the size of true collection of sQTLs. As part of our procedure, we have generated random sQTL equality classes and added them onto $\Omega_{\tilde{Q}}$ by the method outlined in Algorithm 1 for each chromosome. In this algorithm, $c$ represents the length of a given chromosome. Probabilistic determination of whether a control SNP takes place in the gene's downstream or upstream region depends on a SNP's probability of occurring at gene's upstream or downstream across the original dataset.

**while** $|\Omega_{\tilde{Q}}| \leq |\tilde{Q}|n_p$ **do**
    Build a control gene by selecting its length $l$ uniformly among the collection of gene lengths detected over the given chromosome
    Uniformly choose a distance $d$ between SNP and gene among distances detected over the given chromosome
    Probabilistically determine whether the SNP takes place in the gene's downstream or upstream
    **if** the SNP is located in the downstream **then**
        Choose a SNP position $s$ uniformly in $[l+d, c]$
    **else**
        Assign $s$ a uniform position in $[1, c-l-d]$ range
    **end if**
    Construct a random sQTL $q'$ from the identified variables and decide on its equality class $\tilde{q}'$.
    Append $\tilde{q}'$ to $\Omega_{\tilde{Q}}$
**end while**

**Algorithm 1**. Sample space construction for spatial proximity.

Practically, we reach the target size $|\tilde{Q}|$ in less than 20 million iterations by setting $n_p = 100$ in all tests. We match to items in $\Omega_{\tilde{Q}} - \tilde{Q}$, by ensuring that an sQTL is not matched to itself via our bipartite matching solver.

### Statistical analysis of sQTLs

While determining whether a collection of sQTLs is statistically significantly spatially closer than its random expectation, we construct each attribute vector for an equivalent sQTL-gene pair $\tilde{q}$ in a way to contain 5 attributes that are defined previously: $x_{\tilde{q}} = [d(\tilde{q}), I(\tilde{q}), e(\tilde{q}), t_s(\tilde{q}), t_g(\tilde{q})]$. We test whether sQTLs in $\tilde{Q}$ are significantly spatially closer after an optimal matching using $k = 1$ with their equivalents in $\Omega_{\tilde{Q}} - \tilde{Q}$ by non-parametric Wilcoxon signed-rank test where the test is calculated on $p(\tilde{q})$ for matched pairs. Wilcoxon signed-rank test is a typical test for matched case-control studies[72]. In our analysis, we have also applied one-sided test in addition to this two-sided test, where one-sided test's null hypothesis is that the simulated data is spatially closer than the real data.

### Sample space of TADs

We define a sample space for TADs in order to decide on in case a test metric over true collection of TADs and sQTLs vary significantly from random topological domain counterparts. Using the sample space for sQTLs defined in previous sections to explicitly determine in case a test metric on true collection of TADs and sQTLs vary significantly from the same metric over random sQTLs, would have been an alternative method. Nevertheless, when the sample space for TADs is used, we will be mainly investigating whether the real observed TAD structure is significant rather than the observed collection of sQTLs. For a given chromosome, we have randomly shuffled domain and non-domain intervals from $D$ on the chromosome to generate an item of the sample space for TADs. As part of this shuffling process, we preserve the lengths and observed patterns of non-domains and domains as seen in Fig. 17b. For a given chromosome, let $l_i$ be length of interval $i$, the approach defined in Algorithm 2 chooses from 2 shuffled collections of non-domain and domain lengths while preserving the aforementioned features:

1: Partition the chromosome into 'domain' and 'non-domain' tagged intervals corresponding to TADs, and gaps between TADs respectively.
2: Create ('domain', $l_i$) or ('non-domain', $l_i$) pair for every tagged interval $i$. Define $O$ as a collection of those pairs which constitutes the true sequence of non-domain and domain ranges over the chromosome.
3: Generate 2 collections over the set $O$: a collection using solely non-domains $O_{\text{non}}$, and remaining one only with domains: $O_{\text{domain}}$.
4: Shuffle these two collections $O_{\text{non}}$, $O_{\text{domain}}$ randomly.
5: Create modified TAD locations via iterating over 0. If we counter a domain in $O$, we delete the rearmost 'domain' range out of $O_{\text{domain}}$ and add it to $L$ which is a collection of interval lengths. If not, we delete the rearmost 'non-domain' interval out of $O_{\text{non}}$ and add it to $L$.
6: We assign every domain in $L$ to a beginning location based on the summation of interval lengths taking place prior to the domain in $L$.

**Algorithm 2**. Sample space construction for TADs.

Let $\Omega_D$ be the collection of $n_d$ shuffled domain sequences, and we use $n_d = 10000$ in all the experiments.

### Statistical analysis of TADs

While determining whether TADs are crossed by sQTL-gene pairs more frequently than its random expectation, we have estimated the probability of detecting TAD crossings of count $n$ or a higher number for a collection of TADs. More explicitly, we characterize the 'SNP–gene interval' as the length between the middle point of SNP bin for an sQTL-gene association and the middle point of the nearest bin of its regulating gene. SNP-gene interval is defined to 'cross' the TAD if SNP-gene interval of an sQTL-gene association intersects with a TAD, but it is not fully enclosed inside a TAD. $d(\tilde{q})$ is the distance of this SNP-gene interval for an sQTL $\tilde{q}$. Such probability is estimated as we control for 2 random factors: 1- $N_G$: the gene count in sQTL–gene associations which crosses a random collection of TADs; 2- $N_S$: the SNP fragment count in sQTL–gene associations which crosses a random collection of TADs.

Let $N$ be a random factor defining the total number of sQTL–gene associations which crosses a random collection of TADs. Then, the $p$-value or probability that $n$ or more sQTL–gene associations that are crossing TADs is conditioned on the fact that the number of SNPs and genes crossing TADs is inside a given range of observed number of SNP bin and genes crossing TADs:

$$Pr(N \geq n | \gamma^{min} \leq N_G \leq \gamma^{max}, \sigma^{min} \leq N_s \leq \sigma^{max})$$

where $\sigma$ and $\gamma$ let the features of a random collection of TADs to be similar, without being absolutely equal to, the real collection of TADs $D$.

As part of calculating this $p$-value, we have initially determined the lengths for $\gamma$ and $\sigma$ by generating a set $\Omega_D^k$, and then by using $\Omega_D^k$ in determining the sample space of TADs $\Omega_D^*$. Particularly, we have compared a collection of TADs $D$ with its $k = 1000$ topmost matches in the sampled space $\Omega_D^k \subset \Omega_D$ to determine these parameters $\gamma$ and $\sigma$ values. The $x_D = [n_g(D), n_s(D)]$ is the attribute vector utilized to match sampled TADs

to real TADs, containing the gene count in sQTL–gene pairs crossing TADs and SNP-bin count in sQTL–gene pairs crossing TADs.

The parameter range for $\gamma$ is between ($\gamma^{min}$ and $\gamma^{max}$), based on lengths in $\Omega_D^k$ as in:

$$\gamma^{min} = \min_{D' \in \Omega_D^k} n_g(D'), \quad \gamma^{max} = \max_{D' \in \Omega_D^k} n_g(D'),$$

We perform an almost identical approach for $\sigma$ utilizing $n_s(D')$ rather than $n_g(D')$. Then, we define the conditional sample space $\Omega_D^*$ to calculate *p*-value as a subgroup of the sample space $\Omega_D$:

$$\Omega_D^* = \{D' \in \Omega_D : \gamma^{min} \leq n_g(D') \leq \gamma^{max}, \\ \sigma^{min} \leq n_s(D') \leq \sigma^{max}\}$$

This procedure determines the topmost-*k* matches in the control set and restricts the parameters to the intervals seen in this set of *k* size. This procedure ensures the existence of minimum *k* samples over $\Omega_D^*$, where it lets the difference between parameter ranges and the real TAD collection to be as close as possible. This procedure is generally more conservative in the calculated *p*-values than the *p*-values calculated without any restriction over the parameters. In this case, obtaining statistical significance from confounding factors is of very low probability. Every time, we include the real TAD collection in $\Omega_D^*$ calculation in order to ensure a positive probability.

## Data availability

All the data employed in this study is publicly available. Hi-C datasets analyzed during the current study are available in GSE87112. Raw and processed Micro-C files are available in the 4DN data portal (https://data.4dnucleome.org/) under the accession numbers 4DNESWST3UBH, 4DNES21D8SP8, 4DNES2R6PUEK, and 4DNESRJ8KV4Q. sQTL catalog and related datasets analyzed during the current study are available at https://doi.org/10.5281/zenodo.4058759. eQTL datasets analyzed are available GTEx database which is obtained from dbGaP (https://www.ncbi.nlm.nih. gov/gap), accession phs000424. In case needed, you can contact the corresponding author Emre Sefer (emre.sefer@ozyegin.edu.tr).

Received: 22 August 2024; Accepted: 7 November 2025
Published online: 17 December 2025

## References

1. Spitz, F. & Furlong, E. E. Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.* **13**, 613–626 (2012).
2. Soylu, N. N. & Sefer, E. Deepptm: protein post-translational modification prediction from protein sequences by combining deep protein language model with vision transformers. *Curr. Bioinform.* **19**, 810–824 (2024).
3. Eralp, B. & Sefer, E. Reference-free inferring of transcriptomic events in cancer cells on single-cell data. *BMC cancer* **24**, 607 (2024).
4. Sefer, E. Hi-c interaction graph analysis reveals the impact of histone modifications in chromatin shape. *Appl. Netw. Sci.* **6**, 54 (2021).
5. Takata, A., Matsumoto, N. & Kato, T. Genome-wide identification of splicing qtls in the human brain and their enrichment among schizophrenia-associated loci. *Nat. Commun.* **8**, 14519 (2017).
6. Djebali, S. et al. Landscape of transcription in human cells. *Nature* **489**, 101–8 (2012).
7. Harrow, J. et al. Gencode: The reference human genome annotation for the encode project. *Genome Res.* **22**, 1760–1774 (2012).
8. Wang, E. et al. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–6 (2008).
9. Garrido-Martín, D., Borsari, B., Calvo, M., Reverter, F. & Guigó, R. Identification and analysis of splicing quantitative trait loci across multiple tissues in the human genome. *Nat. Commun.* **12**, 727 (2021).
10. Gilad, Y., Rifkin, S. A. & Pritchard, J. K. Revealing the architecture of gene regulation: the promise of eqtl studies. *Trends Genet.* **24**, 408–415 (2008).
11. Zhang, Y. et al. Regional variation of splicing qtls in human brain. *Am. J. Hum. Genet.* **107**, 196–210. https://doi.org/10.1016/j.ajhg.2020.06.002 (2020).
12. Cheung, V. G. & Spielman, R. S. Genetics of human gene expression: mapping dna variants that influence gene expression. *Nat. Rev. Genet.* **10**, 595–604 (2009).
13. Cookson, W., Liang, L., Abecasis, G., Moffatt, M. & Lathrop, M. Mapping complex disease traits with global gene expression. *Nat. Rev. Genet.* **10**, 184–194 (2009).
14. Tian, J. et al. CancerSplicingQTL: a database for genome-wide identification of splicing QTLs in human cancer. *Nucleic Acids Res.* **47**, D909–D916, https://doi.org/10.1093/nar/gky954 (2018). https://academic.oup.com/nar/article-pdf/47/D1/D909/27436247/gky954.pdf.
15. Sefer, E. Drgat: Predicting drug responses via diffusion-based graph attention network. *J. Comput. Biol.* **32**, 330–350 (2025).
16. Sefer, E. & Kingsford, C. Metric labeling and semimetric embedding for protein annotation prediction. *J. Comput. Biol.* **28**, 514–525, https://doi.org/10.1089/cmb.2020.0425 (2021). **PMID: 33370163**
17. Sefer, E. & Kingsford, C. Metric labeling and semi-metric embedding for protein annotation prediction. In *Research in Computational Molecular Biology, 392–407 (Springer* (eds Bafna, V. & Sahinalp, S. C.) (Berlin Heidelberg, Berlin, Heidelberg, 2011).
18. Davison, L. J. et al. Long-range DNA looping and gene expression analyses identify DEXI as an autoimmune disease candidate gene. *Hum. Mol. Genet.* **21**, 322–333 (2011).
19. Meyer, K. B. et al. A functional variant at a prostate cancer predisposition locus at 8q24 is associated with pvt1 expression. *PLOS Genetics* **7**, 1–11 (2011).
20. Sotelo, J. et al. Long-range enhancers on 8q24 regulate c-myc. *Proc. Natl. Acad. Sci.* **107**, 3001–3005 (2010).
21. Duggal, G., Wang, H. & Kingsford, C. Higher-order chromatin domains link eQTLs with the expression of far-away genes. *Nucleic Acids Res.* **42**, 87–96. https://doi.org/10.1093/nar/gkt857 (2013).
22. Sefer, E. A comparison of topologically associating domain callers over mammals at high resolution. *BMC Bioinformatics* **23**, 127. https://doi.org/10.1186/s12859-022-04674-2 (2022).
23. Sefer, E. & Kingsford, C. Semi-nonparametric modeling of topological domain formation from epigenetic data. *Algorithms Mol. Biol.* **14**, 4. https://doi.org/10.1186/s13015-019-0142-y (2019).

24. Filippova, D., Patro, R., Duggal, G. & Kingsford, C. Identification of alternative topological domains in chromatin. *Algorithms Mol. Biol.* **9**, 14 (2014).
25. Sefer, E. Probc: joint modeling of epigenome and transcriptome effects in 3d genome. *BMC Genomics* **23**, 287. https://doi.org/10.1186/s12864-022-08498-5 (2022).
26. Duggal, G. et al. Resolving spatial inconsistencies in chromosome conformation measurements. *Algorithms Mol. Biol.* **8**, 1–10 (2013).
27. Sefer, E., Duggal, G. & Kingsford, C. Deconvolution of ensemble chromatin interaction data reveals the latent mixing structures in cell subpopulations. *J. Comput. Biol.* **23**, 425–438 (2016).
28. Yu, J., Hu, M. & Li, C. Joint analyses of multi-tissue hi-c and eqtl data demonstrate close spatial proximity between eqtls and their target genes. *BMC Genetics* **20**, 43. https://doi.org/10.1186/s12863-019-0744-x (2019).
29. Das, A. et al. Bayesian integration of genetics and epigenetics detects causal regulatory snps underlying expression variability. *Nat. Commun.* **6**, 8555 (2015).
30. He, B., Chen, C., Teng, L. & Tan, K. Global view of enhancer-promoter interactome in human cells. *Proc. Natl. Acad. Sci. U. S. A.* **111**(21), E2191–E2199 (2014).
31. Ron, G., Globerson, Y., Moran, D. & Kaplan, T. Promoter-enhancer interactions identified from hi-c data using probabilistic models and hierarchical topological domains. *Nat. Commun.* **8**, 2237 (2017).
32. Sandhu, K. et al. Large-scale functional organization of long-range chromatin interaction networks. *Cell Rep.* **2**, 1207–1219 (2012).
33. Wang, D., Rendon, A. & Wernisch, L. Transcription factor and chromatin features predict genes associated with eQTLs. *Nucleic Acids Res.* **41**, 1450–1463 (2012).
34. Schmitt, A. et al. A compendium of chromatin contact maps reveals spatially active regions in the human genome. *Cell Rep.* **17**, 2042–2059 (2016).
35. Krietenstein, N. et al. Ultrastructural details of mammalian chromosome architecture. *Mol. Cell* **78**, 554-565.e7 (2020).
36. Gibcus, J. & Dekker, J. The hierarchy of the 3d genome. *Mol. Cell* **49**, 773–782 (2013).
37. Dixon, J. R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
38. Zhang, R., Zhou, T. & Ma, J. Multiscale and integrative single-cell hi-c analysis with higashi. *Nat. Biotechnol.* **40**, 254–261. https://doi.org/10.1038/s41587-021-01034-y (2022).
39. Xiong, K. & Ma, J. Revealing hi-c subcompartments by imputing inter-chromosomal chromatin interactions. *Nat. Commun.* **10**, 5069. https://doi.org/10.1038/s41467-019-12954-4 (2019).
40. Knight, P. A. & Ruiz, D. A fast algorithm for matrix balancing. *IMA J. Numer. Anal.* **33**, 1029–1047 (2013).
41. Imakaev, M. et al. Iterative correction of hi-c data reveals hallmarks of chromosome organization. *Nat. Methods* **9**, 999–1003 (2012).
42. Bernstein, B. et al. An integrated encyclopedia of dna elements in the human genome. *Nature* **489**, 57–74. https://doi.org/10.1038/nature11247 (2012).
43. Karimzadeh, M. & Hoffman, M. M. Virtual chip-seq: predicting transcription factor binding by learning from the transcriptome. *Genome Biol.* **23**, 126. https://doi.org/10.1186/s13059-022-02690-2 (2022).
44. Yang, B. & Aounallah, N. Ctcf degradation causes increased usage of upstream exons in mouse embryonic stem cells. arXiv preprint arXiv:1810.03205 (2018).
45. Sun, Y. et al. Rad21 is the core subunit of the cohesin complex involved in directing genome organization. *Genome Biol.* **24**, 155 (2023).
46. Lu, L. et al. Robust hi-c maps of enhancer-promoter interactions reveal the function of non-coding genome in neural development and diseases. *Mol. Cell* **79**, 521–534 (2020).
47. Witten, D. M. & Noble, W. S. On the assessment of statistical significance of three-dimensional colocalization of sets of genomic elements. *Nucleic Acids Res.* **40**, 3849–3855 (2012).
48. Rao, S. S. P. et al. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
49. Yan, K.-K., Lou, S. & Gerstein, M. Mrtadfinder: A network modularity based approach to identify topologically associating domains in multiple resolutions. *PLoS Comput. Biol.* **13**, 1–22. https://doi.org/10.1371/journal.pcbi.1005647 (2017).
50. Hwang, Y.-C., Zheng, Q., Gregory, B. D. & Wang, L.-S. High-throughput identification of long-range regulatory elements and their target promoters in the human genome. *Nucleic Acids Res.* **41**, 4835–4846 (2013).
51. Degner, J. F. et al. Dnase i sensitivity qtls are a major determinant of human expression variation. *Nature* **482**, 390–394 (2012).
52. Nora, E. P., Dekker, J. & Heard, E. Segmental folding of chromosomes: A basis for structural and regulatory chromosomal neighborhoods?. *BioEssays* **35**, 818–828 (2013).
53. Bickmore, W. & van Steensel, B. Genome architecture: Domain organization of interphase chromosomes. *Cell* **152**, 1270–1284 (2013).
54. Fudenberg, G., Getz, G., Meyerson, M. & Mirny, L. A. High order chromatin architecture shapes the landscape of chromosomal alterations in cancer. *Nat. Biotechnol.* **29**, 1109–1113 (2011).
55. Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F. & Chen, L. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat. Biotechnol.* **30**, 90–98 (2012).
56. Zhang, Y. et al. Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. *Cell* **148**, 908–921 (2012).
57. Yang, J. et al. Analysis of chromatin organization and gene expression in t cells identifies functional genes for rheumatoid arthritis. *Nat. Commun.* **11**, 4402. https://doi.org/10.1038/s41467-020-18180-7 (2020).
58. Choy, M.-K. et al. Promoter interactome of human embryonic stem cell-derived cardiomyocytes connects gwas regions to cardiac gene networks. *Nat. Commun.* **9**, 2526. https://doi.org/10.1038/s41467-018-04931-0 (2018).
59. Sefer, E. Biocode: A data-driven procedure to learn the growth of biological networks. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **19**, 3103–3113 (2022).
60. Patro, R. et al. The missing models: a data-driven approach for learning how networks grow. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, 42–50, https://doi.org/10.1145/2339530.2339541 (Association for Computing Machinery, New York, NY, USA, 2012).
61. Patro, R. et al. Parsimonious reconstruction of network evolution. In *Algorithms in Bioinformatics* (eds Przytycka, T. M. & Sagot, M.-F.) 237–249 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2011).
62. Patro, R. et al. Parsimonious reconstruction of network evolution. *Algorithms Mol. Biol.* **7**, 25 (2012).
63. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
64. Aguet, F. et al. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
65. Yaffe, E. & Tanay, A. Probabilistic modeling of hi-c contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat. Genet.* **43**, 1059–1065 (2011).
66. Cetin, S. & Sefer, E. The significance of chromosome conformation capture in 3d genome architecture comprehension. *Comput. Biol. Chem.* **119**, 108534 (2025).
67. Schmitt, A. et al. A compendium of chromatin contact maps reveals spatially active regions in the human genome. *Cell Rep.* **17**, 2042–2059 (2016).

68. Crowley, C. et al. Firecaller: Detecting frequently interacting regions from hi-c data. *Comput. Struct. Biotechnol. J.* **19**, 355–362 (2021).
69. Flicek, P. et al. Ensembl 2012. *Nucleic Acids Res.* **40**, D84–D90 (2011).
70. Monlong, J., Calvo, M., Ferreira, P. G. & Guigó, R. Identification of genetic variants associated with alternative splicing using sqtlseeker. *Nat. Commun.* **5**, 4698. https://doi.org/10.1038/ncomms5698 (2014).
71. Huang, B. & Jebara, T. Fast $b$-matching via sufficient selection belief propagation. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics* (eds Gordon, G., Dunson, D. & Dudík, M.) vol. 15 of *Proceedings of Machine Learning Research*, 361–369 (PMLR, Fort Lauderdale, FL, USA, 2011).
72. Pearce, N. Analysis of matched case-control studies. *BMJ* **352**, i969 (2016).

## Acknowledgements
The author is grateful for the anonymous reviewers' valuable comments and suggestions.

## Author contributions
B.E. has contributed to the study design, analyzed the data, interpreted the results, prepared the figures, and drafted the manuscript. E.S. has contributed to the study design, interpreted the results, and finalized the manuscript. All authors reviewed the manuscript.

## Funding
This research was funded by TUBITAK (Scientific and Technological Research Council of Turkey) 3501 Project with grant number 122E706.

## Declarations

## Competing interests
The authors declare no competing interests.

## Additional information
**Correspondence** and requests for materials should be addressed to E.S.

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

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

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025