JOINT ANALYSIS OF SQTL AND HI-C REVEALS SPATIAL PROXIMITY BETWEEN SQTLS AND TARGET GENES ACROSS MULTIPLE TISSUES

A Thesis

by

Batuhan Eralp

Submitted to the
Graduate School of Sciences and Engineering
In Partial Fulfillment of the Requirements for
the Degree of

Master of Science

in the
Department of Artificial Intelligence

Özyeğin University
May 2023

JOINT ANALYSIS OF SQTL AND HI-C REVEALS SPATIAL PROXIMITY BETWEEN SQTLS AND TARGET GENES ACROSS MULTIPLE TISSUES

Approved by:

Asst. Prof. Emre Sefer, Özyeğin University
Dept. of Computer Science
Özyeğin University

Assoc. Prof. İlknur Eruçar Fındıkçı
Dept. of Natural and Mathematical Sciences
Özyeğin University

Asst. Prof. Öznur Taştan Okan
Computer Science and Engineering - Molecular Biology, Genetics and Bioengineering
Sabanci University

Date Approved: 07 April 2023

To my Family, Friends, and Cafés And Teashops I Constantly and Obsessively Regulared at Nights.

ABSTRACT

Gene expression and regulation with or without alternative splicing are crucial for tissues and cells to properly function. They have been studied from three almost independent perspectives at the genome level: 1- Recognition of splicing quantitative trait loci (sQTLs), 2- Expression quantitative trait loci (eQTLs) recognition, and 3- Recognition of longer-range physical chromatin interactions between genome segments which model 3D dynamics of cells and tissues. Even though the associations between eQTLs and longer range chromatin interactions have been previously studied, similar relationship between sQTLs and chromatin interactions has not been previously analyzed. In this case, it is crucial to analyze whether sQTLs control the alternative splicing of their target genes mRNA via physically-interacting genome segments. Even though chromatin interactions are part of the principal processes governing eQTLs functioning, similar analysis is missing from sQTLs perspective.

We have jointly analyzed high-throughput chromatin conformation capture (Hi-C) and sQTL datasets over 8 different human cancer tissues. We have discovered the existence of positive association between the number of genes having sQTLs and chromatin interaction frequency. Such positive association still exists when we also control for eQTLs. Additionally, sQTLs and their target genes generally exist inside identical topologically associating domain (TAD). Those findings are observed over the whole set of analyzed cancer types and over different functional subsets of sQTL dataset such as survival-related sQTLs. Furthermore, tissue-specific sQTLs are statistically enriched in tissue-specific frequently interacting regions (FIREs) in 6 out of 8 human cancer tissues (Chronic Myeloid Leukemia, Colon Adenocarcinoma, Acute Myeloid Leukemia, Lung Adenocarcinoma, Prostate Cancer, Sarcoma).

Our sQTL and Hi-C datasets have shown the existence of closer spatial distance between sQTLs and their target genes with possible alternative splicing across a number of different cancer types in human. Such closer spatial distance also exists independent of whether we integrate eQTLs into the analysis. We found that sQTLs regulate the alternative splicing through chromatin interactions.

ÖZETÇE

Gen ifadesi ve alternatif splicing ile birlikte gen ifadesi ve düzenlemesi, dokuların ve hücrelerin düzgün bir şekilde işlev görmesi için önemlidir. Genom düzeyinde, bu konular genellikle üç ayrı perspektiften incelenir: 1- Splicing niceliksel özellik lokuslarının (sQTL'lerin) tanınması, 2- İfade niceliksel özellik lokuslarının (eQTL'lerin) tanınması, ve 3- Genom segmentleri arasındaki daha uzun mesafeli fiziksel kromatin etkileşimlerinin (Hi-C verileriyle modellediği) hücre ve dokuların üç boyutlu dinamiklerini tanımlama. Daha önce eQTL'ler ile daha uzun mesafeli kromatin etkileşimleri arasındaki ilişkiler çalışılmış olsa da, sQTL'ler ile kromatin etkileşimleri arasındaki benzer ilişki henüz analiz edilmemiştir. Bu durumda, sQTL'lerin hedef genlerinin mRNA'sının splicingini fiziksel olarak etkileşen genom segmentleri aracılığıyla kontrol edip etmediğini analiz etmek önemlidir. Kromatin etkileşimleri, eQTL'lerin işleyişini yöneten temel süreçlerin bir parçası olmasına rağmen, benzer bir analiz sQTL'lerin perspektifinden eksiktir.

Biz 8 farklı insan kanser dokusunda yüksek çıktılı kromatin konformasyon yakalama (Hi-C) ve sQTL veri setlerini birlikte analiz ettik. SQTL'lere sahip gen sayısı ile kromatin etkileşim sıklığı arasında pozitif bir ilişkinin varlığını keşfettik. Bu pozitif ilişki, eQTL'leri de kontrol ettiğimizde hala mevcuttur. Ayrıca, sQTL'ler ve hedef genler genellikle aynı topolojik ilişkili alanlarda (TAD) bulunmaktadır. Bu bulgular, analiz edilen kanser tiplerinin tüm kümesi ve hayatta kalım ile ilgili sQTL'ler gibi farklı fonksiyonel alt kümeleri üzerinde gözlenmektedir. Ayrıca, doku özgü sQTL'ler, 8 insan kanser dokusunun 6'sında (Kronik Myeloid Lösemi, Kolon Adenokarsinomu, Akut Myeloid Lösemi, Akciğer Adenokarsinomu, Prostat Kanseri, Sarkom) doku özgü sık etkileşen bölgelerde (FIRE'lar) istatistiksel olarak

zenginleştirilmiştir.

sQTL ve Hi-C veri setlerimiz, insanlarda farklı kanser tiplerinde splicing olasılığına sahip sQTL'lerin ve hedef genlerin daha yakın bir mekansal mesafede bulunduğunu göstermektedir. Bu daha yakın mekansal mesafe, eQTL'leri analize dahil edip etmememize bağlı olarak da mevcuttur. sQTL'lerin alternatif splicing'i kromatin etkileşimleri aracılığıyla düzenlediğini bulduk.

ACKNOWLEDGEMENTS

I would like to extend my thanks to Asst. Prof. Emre Sefer for welcoming me to Ozyegin University, expanding my horizon to pursue bioinformatics and for his kind and understanding as well as masterful approach to everything about this thesis.

I also like to thank my colleagues, Alara ERENEL, M.Sc. (Ph.D. Student,Department of Biological and Medical Sciences, Biosciences,Faculty of Health and Life Sciences,Oxford Brookes University, Oxford, United Kingdom) for helping me with the theoretical background wherever I needed the assistance as well as Alper BÜLBÜL Msc. (Graduate School of Health Sciences, Department of Biostatistics and Bioinformatics, Acibadem University, Istanbul, Turkey) for helping me with the theoretical backgrounds of the algorithms and statistics that are used in this thesis. Last but not least, I would like to thank my family for their endless support for me to get the best education that I can get and my friends who were with me during my master's process.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Gene expression, gene regulation, and alternative splicing (AS) of mRNA for protein synthesis are crucial for proper functioning of tissues and cells. 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 [1]. 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 [2]. As an average, every protein-encoding gene encodes for 7 transcript variants [3]. 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 [4].

As shown by the accumulating evidence, alternative splicing events can be modulated by sequence variants throughout the genome. In this case, splicing quantitative trait loci (sQTLs), which are genetic variants affecting the possible alternative splicing events, are crucial towards complete comprehension of gene expression and regulation, as a result have an impact on response to drugs and susceptibility to diseases [5].

sQTLs can be caused by single-nucleotide polymorphisms (SNPs) which change alternative splicing patterns, such as via modifying the splicing factor's binding affinity to pre-mRNA which is specific to the splicing factor sequence. Previous studies have found that no less than 20% of disease-originating SNPs impact the splicing [6]. sQTLs are crucial to completely understand the impact of genetic variants over disease progression. For instance, alternative splicing is frequently modified across cancer tissues to generate both functional and non-functional abnormal proteins which have an impact on cancer development. We often observe abnormal splicing motifs in cancer beginning, progression, and prognosis, and these aberrant motifs have a significant impact on metastasis, dedifferentiation and carcinogenesis [7]. In particular, the gene which encodes spleen tyrosine kinase protein has a different alternatively-spliced transcript isoform over breast cancer tissues which is not expressed in normal cells [8]. In this case, a superior comprehension of AS misregulation will be important in developing specific gene targets to pharmacologically intervene in cancer progression.

Recent development of high-throughput sequencing techniques enables us to analyze gene expression, gene regulation, and alternative splicing at genome level across various standpoints. Some examples of these high-throughput techniques are SNP chips which are DNA microarrays testing genetic variation around more than 100,000 specific locations simultaneously throughout the genome, DNA sequencing, and RNA sequencing techniques [1, 9, 10]. Capturing expression quantitative trait loci (eQTLs) has been investigated by many studies [11, 12], where eQTLs represent variants at the genome that are related with gene expression levels among individuals according to robust statistics. Similarly, as defined previously, splicing quantitative trait loci (sQTLs) are genome loci which control pre-mRNA's alternative splicing that might be identified over RNA-seq datasets [1, 9]. A number of computational approaches have been introduced to extract sQTLs from RNA-seq data [13, 14]. As a result, the inference of sQTLs, in particular over cancer cells, may be crucial step in

complete comprehension of the impact of genetic variants in cancer progression and tumorigenesis.

The significance of the identified genomic variants across gene regulation and functioning motivates many of the sQTL studies. Cis-regulatory modules (CRMs) are non-coding DNA regions regulating the next-door genes transcription. By modifying cis-regulatory modules (CRMs) across the genome, these sQTLs modulate their target genes expression. There are a number of examples of cis-regulatory elements, some of which are mediators, insulators, promoters, and enhancers [15, 16]. Meanwhile, the studies on three-dimensional spatial chromatin shape have shown the significance of interactions between different chromatin segments for regulation of genes [17, 18]. For instance, cis-regulatory elements could regulate the expression of genes that are further away from the regulatory element, via making up longer range interactions between corresponding chromatin segments [19, 20]. Widely-adopted high-throughput chromatin conformation capture technique like Hi-C provides interactions between chromatin segments at a genome-wide scale [21, 22, 23]. Such Hi-C interaction dataset is generally composed of interaction frequencies between different chromatin segments, and it is in general represented as a contact matrix where genome is partitioned into equal size bins. In this case, each matrix entry shows chromatin interaction frequency (CIF), which is the read pair count aligned to a given bin pairs.

Further analysis of Hi-C experiments have revealed A and B compartments which correspond to the partition of spatial genome into open and closed chromatin respectively [22]. Among these compartments, A compartment is related to easily accessible, transcriptionally active euchromatin. Similarly, B compartment is related to condense, transcriptionally inactive heterochromatin. Resulting analysis of the Hi-C interaction matrix at a higher resolution has discovered topologically associating domains (TADs) which are frequently-interacting, sequential, closely-located interaction matrix areas [24, 25, 26, 27]. TADs are pervasive genome organization unit which are

stable attributes of Hi-C matrices. TADs match greatly with cellular differentiation and long-range transcriptional control [28, 29].

sQTL and Hi-C studies are complementary methods focusing on non-identical characteristics of gene expression and regulation. sQTL studies require a linked single-nucleotide polymorphism (SNP) which is the variation at one position in a DNA sequence among individuals. As a result, sQTL results are statistically defined among individuals. On the other hand, for a given sample, chromatin interactions are instead physical interactions and polymorphism does not need to exist in the dataset. Integrating the outcome from both of these methods will be useful to properly understand the connection between them. In this case, one open question will be how important the chromatin interactions become while sQTLs regulating target genes. 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 have been answered. As an example, Duggal et al. [30] 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 that are detected overlapping withs CRMs and eQTLs that are detected inside the same topologically associating domain (TAD). 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 [11] 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, [31] 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 [32], while the others focus on predicting enhancer-promoter interactions [33, 34].

Even though interactions between chromatin segments are considered as one of the fundamental processes defining eQTLs and sQTLs, any direct proof for the existence of such relationship for sQTLs does not exist for tissues. sQTLs are tissue-specific [14, 35]. Furthermore, [36] has 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, and they are defined as bins which often interact with the closeby segments that are less than 200 kb apart. In this case, the degree of overlap between tissue-specific sQTLs and FIREs is also not known.

sQTL and Hi-C datasets have recently and systematically become available for a number of human cell lines and primary tissues, especially cancer tissues. For instance, [35] come up with a database called CancerSplicingQTL which has sQTL datasets for 33 cancer types. Hi-C interaction datasets for many cancer tissues have also been generated by various experiments [21, 37, 38, 39, 40, 41, 42]. Similarly, [43] has systematically mapped eQTLs at a genome-wide scale over 33 cancer tissues. Among these cancer resources, 8 cancer tissues overlap as seen in Table 1, which facilitates a robust assessment of the connection between sQTLs and chromatin interactions over a number of cancer tissues, while also controlling for eQTLs. Table 1 in Results section summarizes these datasets.

In this paper, we have jointly analyzed the relationship between sQTLs and Hi-C datasets across 8 different cancer tissues while also taking into account eQTLs. We have identified that there is a positive association between the frequency of sGenes and CIF, where sGene is a gene which expression is significantly related to an sQTL. Such positive association still exists when we also control for eGenes, that represent

genes which expression is significantly related to an eQTL instead of an sQTL. Additionally, sQTLs colocalize with their target genes more frequently inside the same TAD, compared to randomly created control dataset. This colocalization result is also valid when we consider subset of sQTLs, such as GWAS sQTLs [35] which overlap with the identified GWAS (Genome-wide association study) regions that are associated with linkage disequilibrium regions, and survival sQTLs [35] which are impactful in patient's survival time for a given cancer. All our results are compatible over all the evaluated cancer types and tissues. Due to tissue specificity of both FIREs and sQTLs [36], we have analyzed the degree of association between tissue-specific FIREs and tissue-specific sQTLs, and identified a positive relationship between them across most of the cancer tissues.

To our best knowledge, this is the first study to analyze the relation between sQTLs and chromatin interactions over a number of human cancer types, and to analyze the association between tissue-specific FIREs and tissue-specific sQTLs. Our results are important to enhance our comprehension of the roles of sQTLs and chromatin interactions on gene expression and regulation through alternative splicing.

CHAPTER II

MATERIALS AND METHODS

2.1 Data Description

We utilized sQTL, eQTL, and Hi-C datasets over 8 human cancer tissues; Chronic Myeloid Leukemia (CML), Colon Adenocarcinoma (COAD), Glioblastoma (GBM), Acute Myeloid Leukemia (LAML), Lung Adenocarcinoma (LUAD), Multiple Myeloma (MM), Prostate Cancer (PRAD), Sarcoma (SARC) [35, 43, 44, 21, 37, 38, 39, 40, 41, 42]. More details about these datasets can be found in Tables 1-2. In our study, we only analyze autosomal chromosomes where hg19 is the reference genome. Similarly, Figure 1 shows log-transformed Hi-C interaction matrix at 10 Kb for GBM's chromosome 20.

Table 1: Data summary for all analyzed cancer types.

Cancer Type	Description	sQTL Dataset	eQTL Dataset	Hi-C Dataset
CML	Chronic Myeloid Leukemia	[35]	[43]	GSE63525 [21]
COAD	Colon Adenocarcinoma	[35]	[43]	GSE133928 [37]
GBM	Glioblastoma	[35]	[43]	GSE121601 [38]
LAML	Acute Myeloid Leukemia	[35]	[43]	GSE63525 [21]
LUAD	Lung Adenocarcinoma	[35]	[43]	GSE92819 [39]
MM	Multiple Myeloma	[35]	[44]	GSE87585 [40]
PRAD	Prostate Cancer	[35]	[43]	GSE118629 [41]
SARC	Sarcoma	[35]	[43]	GSE117582 [42]

The number of raw intra-chromosomal unique read pairs on 8 human cancer tissues is over 3 billion, where over 800 million pairs have a distance greater than 15

Table 2: Data characteristics for all analyzed cancer types.

Cancer Type	Number of interactions	Number of TADs	Number of FIREs	Number of genes with FPKM > 1	Number of tested genes	Number of sGenes
CML	30,731,911	8,205	3,444	15,257	23,244	4,903
COAD	22,823,424	8,210	3,471	15,679	23,567	3,680
GBM	145,013,101	8,101	2,892	14,075	20,155	7,177
LAML	106,069,045	8,075	3,359	14,349	22,366	8.918
LUAD	25,773,838	7,834	3,961	16,371	24,470	11,005
MM	34,430,763	7,922	3,616	15,.052	22,953	3,157
PRAD	25,843,694	7,687	3,407	15,221	22,376	6,233
SARC	134,867,731	8,656	3,296	13,733	21,053	3,886

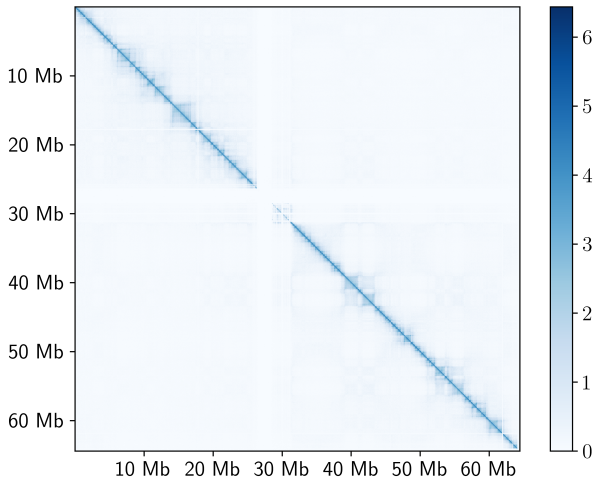


Figure 1: Log-transformed Hi-C interaction matrix of GBM at 10 kb resolution, for chromosome 20

kb between their endpoints. We have analyzed Hi-C datasets by binning at 10 kb resolution. Figures 2 show the distribution of distance between gene and sQTL over all autosomal chromosomes for all the considered cancer tissues. We have mainly used Armatus [45] by setting $g = 0.5$ to infer TAD boundaries over the analyzed cancer tissues. We have also inferred TADs by Arrowhead [21] via using its default parameter and MrTADFinder [46] by setting $res = 1.5$ parameter in order to test the robustness of results. Similarly, we have used FIREcaller [47] to infer FIREs over cancer tissues. 2070 TADs and 3682 FIREs exist on average per cancer tissue sample.

2.2 Regression analysis of the association between interaction frequencies and sQTLs

First of all, we have used regression methods to analyze the association between sQTL results and chromatin interaction frequency (CIF) similar to the previous analysis on eQTLs [31]. Every SNP-splicing gene pair in CancerSplicingQTL [35] database is mapped to a single bin pair. We remove the SNP-gene pairs mapped to the same interaction matrix bin from the analysis. In this case, if a tested gene maps in one interaction matrix bin, such gene should have a corresponding sQTL and SNP in one of the remaining bins. For every bin pair (i, j) where $i < j$, we have the following attributes: 1- $F_{\text{Hi-C}}$: Chromatin interaction frequency (CIF) between bins i and j, 2- G_{sGene}: sGenes (splicing genes) frequency with transcription start site (TSS) located in bin i or bin j, 3- G_i^s: the count of tested genes over sQTL dataset where transcription start site is located in bin i, 4- G_j^s: the count of tested genes over sQTL dataset where transcription start site is located in bin j, 5- G_{eGene}: eGenes (eQTL-gene associations) frequency with transcription start site (TSS) located in bin i or bin j, 6- $D = |i - j|$: genomic distance between bins i and j. In our analysis, we add the constraint $G_i^s + G_j^s > 0$ which focuses on bin pairs containing at least one tested gene.

To better understand the relationship between sQTL results and CIF, we have applied linear and negative binomial regression where CIF is regressed against the sGenes count in each human cancer tissue. As the genomic distance between two loci affects the CIF between them [22], we have incorporated distance D as an additional covariate in the following regression models:

$$\ln(F_{\text{Hi-C}}) \approx G_{\text{sGene}} + D \tag{1}$$

$$F_{\text{Hi-C}} \approx G_{\text{sGene}} + D \tag{2}$$

where Equations 1-2 define the negative binomial and linear regressions of CIF again

sGenes count respectively, and ln represents the corresponding log transformation in negative binomial regression. Moreover, chromatin interactions are observed more frequently inside the same compartment than between a gene-poor and gene-dense compartments [22]. Even though the compartments are observed at multi-megabase scale, gene density might play a critical role in chromatin interactions according to this observation. As a result, we have further controlled for nonuniformity in the genes distributions between bins i and j. We have controlled such nonuniformity by adding gene difference between bins i and j $G_{\text{Diff}} = |G_i^s - G_j^s|$ to our regression models as another feature. In this case, higher G_{Diff} values indicate genes corresponding to the bin pairs i and j are distributed less uniformly. Below, we define the regression equation only for nonnegative binomial regression, but the linear regression will be defined similarly except taking the logarithm:

$$\ln(F_{\text{Hi-C}}) \approx G_{\text{sGene}} + G_{\text{Diff}} + D \tag{3}$$

We have extended the negative binomial regression equation above by taking into account the pairwise interaction and confounding effects between the covariates as seen below in Equation 4. In these scenarios, we did not find the confounding effects between covariates to be statistically significant. As a result, we mainly report the results without considering the covariate interactions:

$$\ln(F_{\text{Hi-C}}) \approx G_{\text{sGene}} + G_{\text{Diff}} + D + G_{\text{sGene}}G_{\text{Diff}} + G_{\text{Diff}}D + G_{\text{sGene}}D \tag{4}$$

In the results section, we present the estimated coefficients obtained by fitting to these equations with bar plots, where "effect" on the y-axis title "Effect on chromatin interaction frequency" means the estimated coefficient. We have also applied stratified analysis to subsets stratified by the distance between sQTL and splicing position of the genes, where stratification partitions the whole population as subsets which are called strata. Then, an independent sample is chosen within each of strata. As discussed in Results section, even after stratification, we have found a significant

positive relationship between the number of sGenes and CIF. We have also applied a similar regression analysis for the number of non-sGenes, which is defined as tested genes without any sQTLs, and compared its coefficient in the regression with the one from sGenes. By running another regression on non-sGenes, we are primarily interested in understanding whether such strong associations are primarily due to sGenes. Let $G_{\text{non-sGene}}$ be the number of non-sGenes, regression on non-sGenes is as follows:

$$\ln(F_{\text{Hi-C}}) \approx G_{\text{non-sGene}} + G_{\text{Diff}} + D \tag{5}$$

Lastly, we have also carried out sensitivity analysis via running other regression models where input variables (independent variables) are either as categorical variables, or they are defined in logarithmic scale in a linear regression. Such sensitivity analysis ensures that our conclusions are not mainly impacted by the model choices. Even with such categorization and logarithmic transformation, we have still identified a significant positive relationship between CIF and number of sGenes as discussed in Results section.

2.3 Enrichment study of sQTL-gene relationships in topologically associating and compartment domains

We have also assessed whether TADs are enriched for sQTL-gene associations for all human cancer tissues. We have carried out such enrichment analysis as follows: We have generated a matching pseudo-pair as a control for each tested SNP-gene pair. While keeping the gene's transcription start site (TSS) position, we flipped SNP's direction such that it becomes on the opposite side of transcription start site without changing the distance from the transcription start site. In this case, if the actual SNP is 90Kb upstream of the transcription start site, its flipped position in the control dataset is 90Kb downstream of the transcription start site. As a result of this procedure, the distribution of SNP-TSS distance and the distribution of gene

locations will be same across the pseudo SNP-gene pairs and actual SNP-gene pairs. In case flipped location is no longer belongs to the chromosome, we remove both actual and matching pseudo pair from our analysis.

We analyzed SNP-gene pairs via 2 attributes: 1- whether TSS of a gene and SNP belong to the same TAD, 2- whether SNP-gene pair is an actual or pseudo one. As a result, we obtain a 2×2 matrix, and observe if the probability of SNP-gene pairs switching from locating within the same TAD to locating in different TADs is significantly greater. We apply McNemar's test to test this hypothesis. Similarly, we have also evaluated the relation between those 2 attributes by performing Fisher's exact test.

Additionally, we bring distance into the equation via applying logistic regression $Y \approx X + D$, where D defines the distance between the gene's TSS and SNP, X defines whether the SNP-gene pair is actual, and Y defines whether the pair locates within the same TAD. We have also stratified the dataset by genomic distance changing from 10 Kb to 250 Kb, and applied stratified analyses.

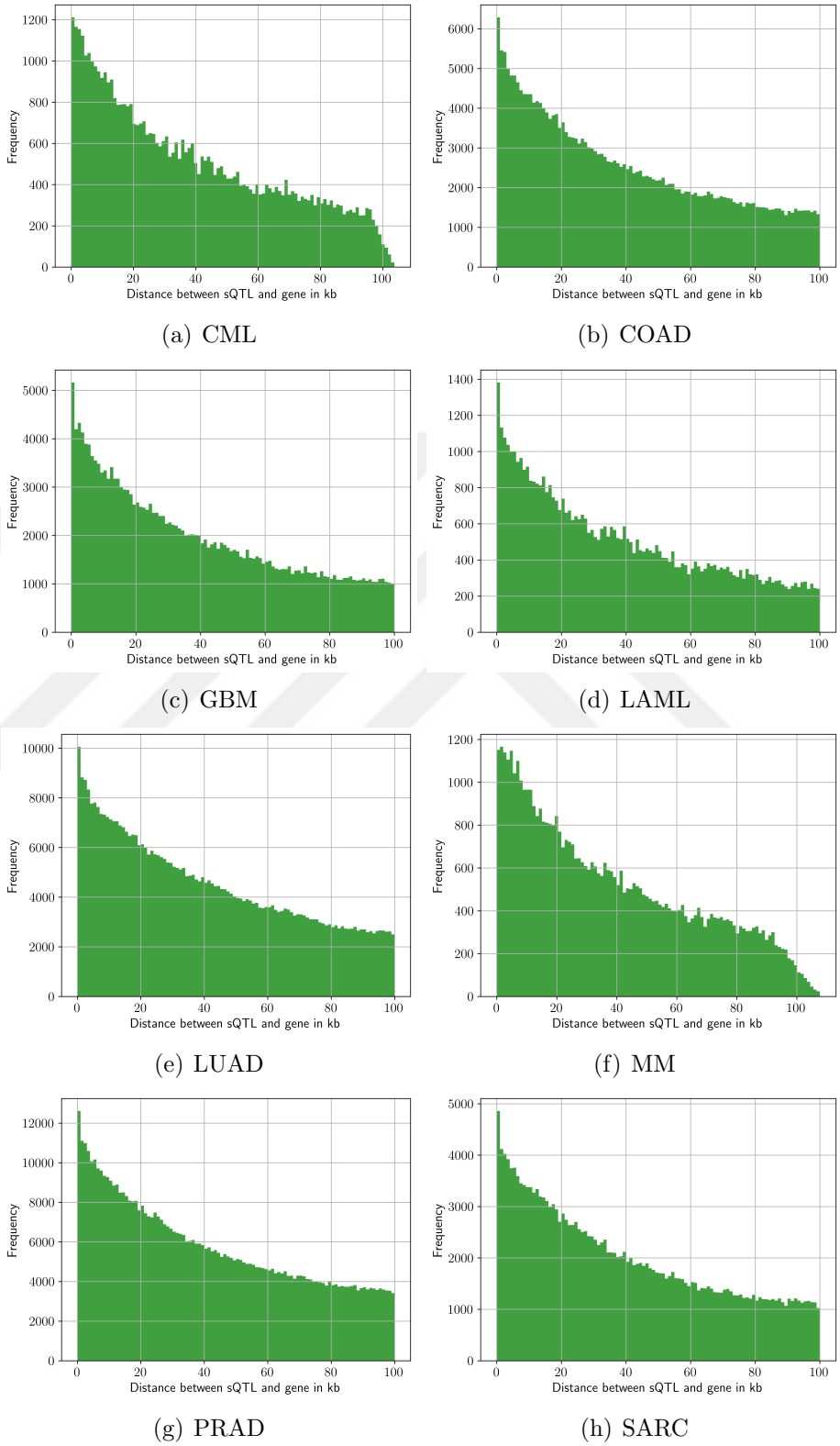
We have also applied a similar procedure to compartment domains, and while analyzing survival-sQTLs and GWAS-sQTLs.

2.4 Analysis of cancer tissue-specific FIREs and cancer tissue-specific sQTLs

Lastly, we have analyzed whether Hi-C dataset is tissue-specific. We have characterized FIREs that are tissue-specific, if these FIREs are detected only in the specific cancer tissue but not in the remaining cancer tissues. Then, we have analyzed whether tissue-specific FIREs are enriched near genes, and analyzed their degree of relationship with tissue-specific sQTLs. We define tissue-specific sQTLs similarly by CancerSplicingQTL database results.

We have analyzed whether tissue-specific sQTLs are enriched in tissue-specific

FIREs for each cancer tissue. As part of this enrichment analysis, we have partitioned all sQTLs according to whether sQTL overlaps with tissue-specific FIRE and whether the sQTL is tissue-specific. We have also calculated the odds ratio where the numerator is the proportion of tissue-specific sQTLs mapped to tissue-specific FIREs and the denominator is the proportion of tissue-specific sQTLs mapped to remaining FIREs of the same tissue. We have also calculated Bonferroni corrected p-value and confidence interval at 95% confidence level after Bonferroni correction (i.e. 99.375% nominal level, where $0.99375 = 1–0.05/8$).



(a) CML (b) COAD

(c) GBM (d) LAML

(e) LUAD (f) MM

(g) PRAD (h) SARC

Figure 2: The distribution of distance between gene and sQTL over all autosomal chromosomes for (a) CML, (b) COAD, (c) GBM, (d) LAML, (e) LUAD, (f) MM, (g) PRAD, (h) SARC cancer tissues.

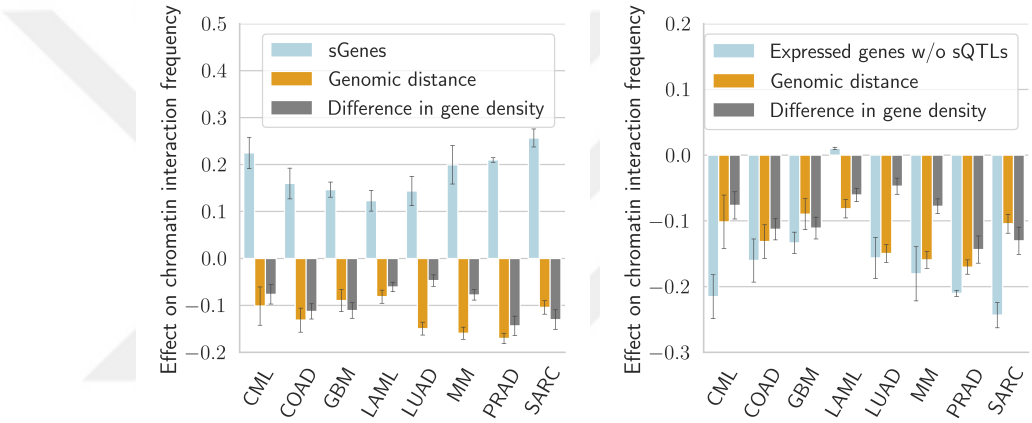
CHAPTER III

RESULTS

3.1 The number of chromatin interactions is significantly positively related with the number of splicing genes

We can expect a pair of loci located as an sQTL-gene association to interact more frequently if 3D chromatin organization impacts the sQTLs regulation of their target genes. We test this hypothesis by fitting negative binomial and linear regressions, and evaluating the association between CIF and number of sGenes (splicing gene-sQTL) associations between two loci over interaction dataset with 10kb resolution by using these regression models. As discussed in Methods section, we have ignored the interactions within same bins, and only used sQTL-gene pairs that are located in different bins. Once we have controlled for genomic distance between any loci pair, the effect of number of sGenes on CIF become significantly positive (p-value < 0.001) over all human cancer tissues as seen in Figure 4(a). For instance, in LAML cancer tissue, according to negative binomial regression, the effect of number of sGenes on CIF is 0.21 (p-value < 0.001). According to this effect, CIF would have been $1.23 = e^{0.21}$ times greater for each additional sGene in single pair of bin. This effect coefficient changes across different cancer tissues, taking values between 0.05 and 0.21. Across these experiments, effect sizes are quite similar to each other for relatively more similar tissues such as CML (Chronic Myeloid Leukemia) and LAML (Acute Myeloid Leukemia). Furthermore, in the same regression models, the coefficient between CIF and genomic distance is significantly negative across all cancer tissues. Such negative coefficient is expected since interaction frequency between two loci in general decreases as distance between these loci increases [22]. If we change

the resolution of binning for both interaction and sQTL datasets, we still observe similar significance results as seen in Figure 4(b) for both 5 kb, 10 kb and 40 kb resolutions, where we only plot the coefficients of sGenes. Similar to the 10 kb resolution results above, across different resolutions, effect sizes are quite similar to each other for relatively more similar tissues such as CML and LAML. We also observe similar statistical significance results when we run linear regression instead of negative binomial regression as shown in Supplementary Materials.



(a) The effect of sGenes over multiple cancer tissues

(b) CIF is related with number of sGenes, not with the total gene count

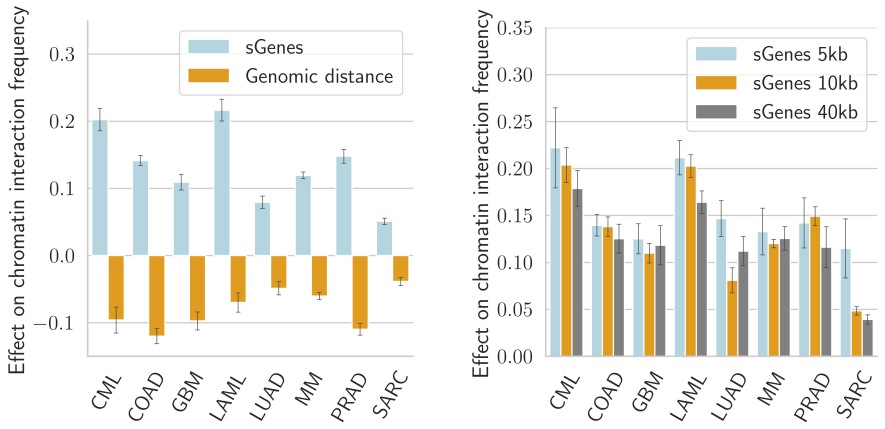
Figure 3: (a) The effect of sGenes frequency on chromatin interaction frequency, estimated by fitting negative binomial regression where we also adjust for gene density difference and genomic distance covariates. (b) Estimated coefficients while regressing CIF against the number of genes without sQTLs and controlling for gene density difference and genomic distance. ± standard errors are represented by the error bars. (***) stands for $p < 0.001$, whereas (**) stands for $p < 0.01$ significance.

[22] has come up with the partition of genome into A and B compartments. These compartments have been related to higher gene density and lower gene density regions respectively. They have also discovered that a pair of genome regions with different compartment types has a lower interaction frequency than regions inside the same compartment types. According to this result, gene density might have a critical role in chromatin interactions. To control for gene density's critical role in our analysis, as seen in Figure 3(a), we have integrated the absolute difference in tested gene counts

among the pair of bins as an extra attribute into the regression framework. We have discovered that the effect of difference in gene density between pair of bins on CIF is significantly negative in all cancer tissues. The correlation between number of sGenes and gene density difference is small as well, Pearson correlation ≤ 0.05 in all cancer tissues. Furthermore, the correlation between number of sGenes and CIF is still significantly positive where the estimated effects are higher than the previous model which does not take into account the gene density difference. Moreover, we have run stratification analysis by stratifying the data via the gene density difference. Stratification divides the population into subsets (that are called strata) within each of which an independent sample is selected. In this analysis, gene count difference differs between 0 and 8, and sample sizes are nearly tiny for the strata with difference > 5. The results were similar to the ones in Figures 4(a)-3(a) for many of the analyzed strata. In a few number of strata, the estimated coefficients for sGene count were negative but without any statistical significance.

We have also analyzed the datasets by incorporating the number of genes without an associated sQTL into the regression model instead of number of sGenes. According to Figure 3(b), the coefficients are mostly negative across all cancer tissues. For LAML tissue, the effect of the number of genes without an associated sQTL on the CIF is positive. However, such positive coefficient is remarkably smaller than the coefficients for the number of sGenes. As a result, CIF is related to number of sGenes, not with the total gene count. We also observe similar statistical significance results when we run linear regression instead of negative binomial regression as shown in Supplementary Material.

Previous research [31] have found that number of eGenes (eQTL-gene associations) is also positively associated with chromatin interaction frequency for a number of human tissues. We test the robustness of our sQTL results by applying similar regressions while controlling for eGenes as well. In more detail, we test our sQTL



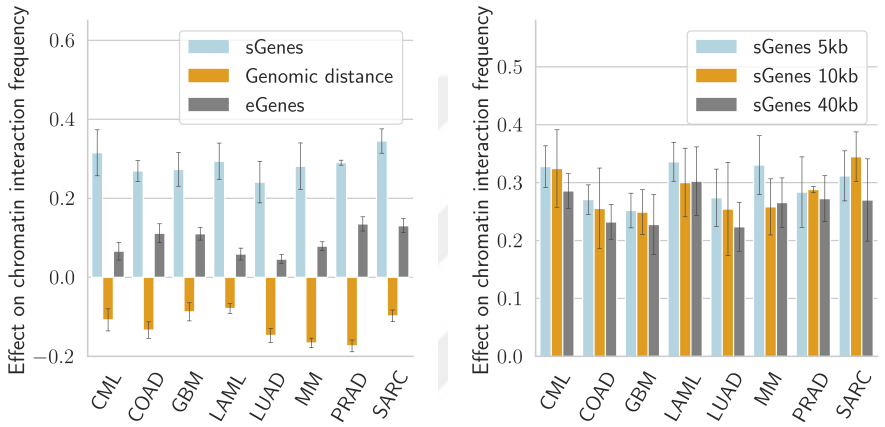
(a) The effect of sGenes over multiple cancer tissues

(b) The effect of sGenes across different resolutions

Figure 4: The effect of sGenes quantity on chromatin interaction frequency, estimated by fitting negative binomial regression. While estimating the effect of sGenes on chromatin interaction frequency, we also control for genomic distance. ± standard errors are represented by the error bars.

hypothesis by fitting negative binomial and linear regressions, and evaluating the association between CIF and number of sGenes associations between two loci over interaction dataset with 10kb resolution while controlling for eGenes and genomic distance. In this case, the effect of number of sGenes on CIF is also significantly positive (p-value < 0.0001) over all human cancer tissues as seen in Figure 5(a). For instance, in CML cancer tissue, according to negative binomial regression, the effect of number of sGenes on CIF is 0.32 (p-value < 0.0001). According to this effect, CIF would have been $1.38 = e^{0.32}$ times greater for each additional sGene in single pair of bin. This effect coefficient changes across different cancer tissues, taking values between 0.24 and 0.32. Across these experiments, similar to the case without eGenes, effect sizes are quite similar to each other for relatively more similar tissues such as CML (Chronic Myeloid Leukemia) and LAML (Acute Myeloid Leukemia). Similar to the case without eGenes above, the coefficient between CIF and genomic distance is significantly negative across all cancer tissues. If we change the resolution of binning for both interaction and sQTL datasets, we still observe similar significance results as

seen in Figure 5(b) for both 5 kb, 10 kb and 40 kb resolutions, where we only plot the coefficients of sGenes. Similar to the 10 kb resolution results above, across different resolutions, effect sizes are quite similar to each other for relatively more similar tissues such as CML and LAML. We also observe similar statistical significance results when we run linear regression instead of negative binomial regression as shown in Supplementary Material.



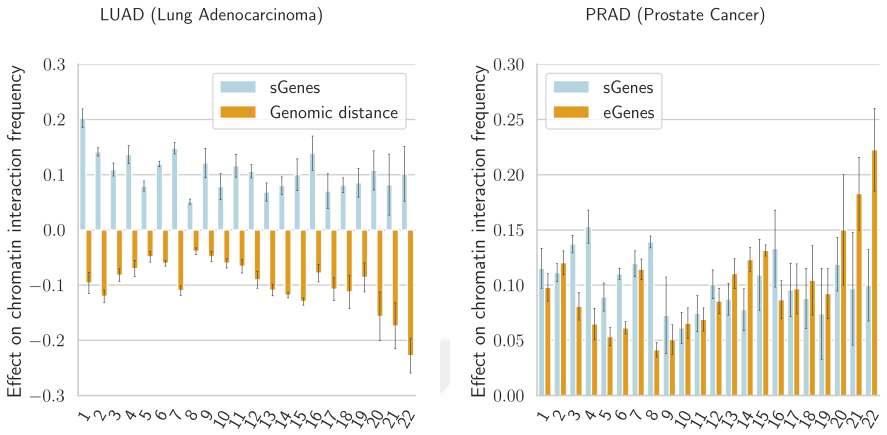
(a) The effect of sGenes over multiple cancer tissues

(b) The effect of sGenes across different resolutions

Figure 5: The effect of sGenes quantity on chromatin interaction frequency, estimated by fitting negative binomial regression. While estimating the effect of sGenes on chromatin interaction frequency, we also control for eGenes and genomic distance. $\pm$ standard errors are represented by the error bars. (***) stands for $p < 0.001$, whereas (**) stands for $p < 0.01$ significance.

We have also analyzed the significance of positive association between chromatin interaction frequency and sGenes across individual autosomal chromosomes for lung and prostate cancer types via negative binomial regression over interaction dataset with 10kb resolution as seen in Figures 6(a)-6(b) respectively. In terms of lung cancer, positive association between CIF and sGenes is statistically significant for all autosomal chromosomes except 17, 21, 22 (p-value < 0.001) as seen in Figure 6(a). Similarly, such positive associations are statistically significant for all prostate cancer chromosomes except 9, 19 (p-value < 0.001) as seen in Figure 6(b). Detailed

chromosomal results for the remaining cancer types can be found in Supplementary Material.



(a) LUAD (b) PRAD

Figure 6: The effect of sGenes quantity on chromatin interaction frequency per chromosome, estimated by fitting negative binomial regression while controlling for eGenes and genomic distance as well. Results are in (a) LUAD, (b) PRAD cancer tissues. $\pm$ standard errors are represented by the error bars. (***) stands for $p < 0.001$, whereas (**) stands for $p < 0.01$ significance.

We also report the adjusted R^2 values for negative binomial regressions seen in Figures 4(a), 3(a) and 5(a) respectively as seen in Table 3. In general, adding the covariates to model nonuniformity in the genes distributions between the bins (G_{Diff}) and eQTL-gene associations frequency with transcription start site (TSS) located in bin i or bin j (G_{eGene}) to the negative binomial regression leads to a better prediction results. We have also tested whether taking the confounding effects into account between covariates improves the performance as seen in Table 4. Here, we have extended the negative binomial regression equation 3 by taking into account the pairwise interaction and confounding effects between the covariates as in Eq 4. In terms of adjusted R^2 values, the scores frequently stay the same or decrease as a result of extending the equation. As a result, we mainly report the results without considering the covariate interactions.

We have also ensured that our conclusions so far are not impacted by the model

Table 3: Adjusted R^2 values for negative binomial regressions seen in Figures 4(a) ($\ln(F_{\text{Hi-C}}) \approx G_{\text{sGene}} + D$), 3(a) ($\ln(F_{\text{Hi-C}}) \approx G_{\text{sGene}} + G_{\text{Diff}} + D$), and 5(a) ($\ln(F_{\text{Hi-C}}) \approx G_{\text{sGene}} + G_{\text{eGene}} + D$) respectively.

Cancer	$\ln(F_{\text{Hi-C}}) \approx G_{\text{sGene}} + D$	$\ln(F_{\text{Hi-C}}) \approx G_{\text{sGene}} + G_{\text{Diff}} + D$	$\ln(F_{\text{Hi-C}}) \approx G_{\text{sGene}} + G_{\text{eGene}} + D$
CML	0.245	0.254	0.251
COAD	0.232	0.249	0.265
GBM	0.243	0.261	0.266
LAML	0.288	0.299	0.302
LUAD	0.275	0.288	0.291
MM	0.255	0.272	0.274
PRAD	0.297	0.311	0.315
SARC	0.267	0.275	0.283

Table 4: Adjusted R^2 values for negative binomial regressions without (Eq. 3) and with confounding effects (Eq. 4).

Cancer	Without confounding effects, Eq 3	With confounding effects, Eq 4
CML	0.254	0.251
COAD	0.249	0.250
GBM	0.261	0.252
LAML	0.299	0.289
LUAD	0.288	0.277
MM	0.272	0.269
PRAD	0.311	0.305
SARC	0.275	0.286

choices, by performing similar analyses over linear regression model, over a log-transformation model where we also log-transform the independent variables. In all cases, we have identified a significant positive relationship between CIF and number of sGenes.

3.2 TADs and compartment domains are enriched in terms of sQTL-gene associations

Genome regions inside the same TAD tend to interact more frequently than regions in different TADs [48, 24]. By taking this result into account, we have analyzed whether TADs are enriched for sQTL-gene associations. Firstly, figure 7 shows the

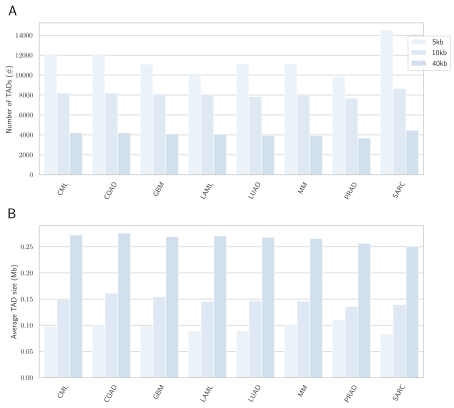
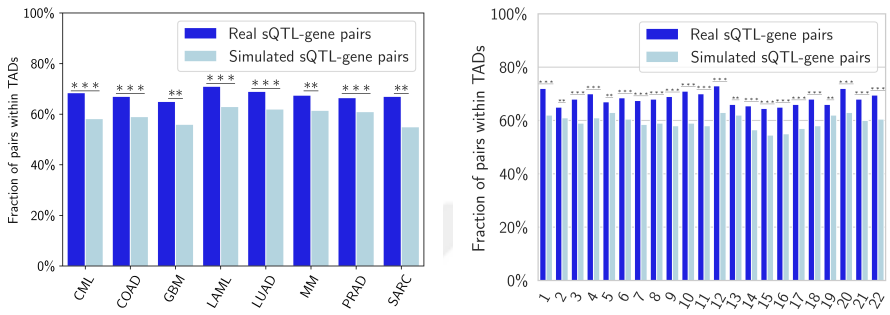


Figure 7: Average number of TADs and the mean TAD size in Mb inferred by Armatus for different cancer types across 5 kb, 10 kb and 40 kb resolutions.

average number of TADs and the mean TAD size in Mb inferred by Armatus for different cancer types across the resolutions. As discussed in Methods in detail, we have generated a pseudo SNP-gene pair matching every sQTL-gene pair in the dataset without modifying the position of TSS of the gene but instead putting SNP to the opposite side of TSS for all tissues. The higher proportion of sQTL-gene pairs remained within the same TAD after putting SNP to the opposite side, but a relatively significant number of sQTL-gene pairs have started to locate across different TADs, such as between 10% and 15% across all tissues. However, almost none of the sQTL-gene pair that are located in different TADs has started to locate in the same TAD according to McNemar's test with p-value $< 1 \times 10^{-16}$ for all cancer tissues. Additionally, the number of sQTL-gene pairs locating in the same TAD in real dataset is statistically significantly higher than the ones in simulated dataset according to Fisher's exact test with p-value $< 1 \times 10^{-16}$ for all cancer types as seen in Figure 8(a). For instance, 58.23% pseudo sQTL-gene pairs in CML tissue were inside TADs, whereas it was 68.45% sQTL-gene pairs in the real dataset.

We have also analyzed the ratio of SNP-gene pairs inside topologically associating domains per chromosome for pseudo and real LUAD cancer datasets as seen in Figure 8(b). Similarly, we found the number of sQTL-gene pairs locating in the same

TAD in real LUAD dataset to be statistically significantly higher than the ones in simulated dataset according to Fisher's exact test for all chromosomes. We observe similar statistically significant enrichment profiles across different chromosomes for other cancer types as well.

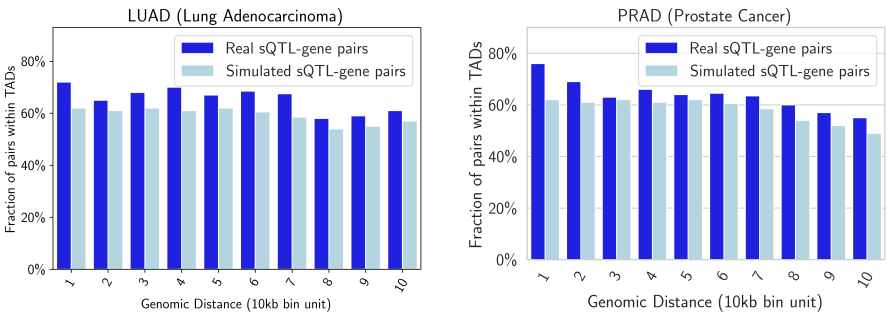


(a) Across cancer types (b) Across LUAD chromosomes

Figure 8: Enrichment of sQTL-gene associations in topologically associating domains. The ratio of SNP-gene pairs inside topologically associating domains for pseudo and real data, (***) $p < 0.001$ for Fisher's exact test, (**) $p < 0.01$ for Fisher's exact test. (a) Over all cancer types, (b) Over all chromosomes of LUAD cancer type.

These associations are also valid if genomic distances are also taken into account. As discussed in Methods section, we have applied a joint analysis with distance as an independent variable and applied a stratified analyses where distance is the stratifying factor. When logistic regression is applied, the odds-ratio for a SNP-gene pair being in the same TAD is statistically significantly greater for real datasets than for simulated datasets across all cancer tissues as seen in Table 5. The stratified analyses results for lung (LUAD) and prostate (PRAD) cancer tissues are shown in Figures 9(a)-9(b) where we stratify via the distance between sQTL and splicing position of the genes. All our TAD enrichment results are robust across different TAD callers.

Similar to TAD enrichment analysis, we have also analyzed whether compartment domains [22] are enriched for sQTL-gene associations. As discussed in Methods in detail, we have generated a pseudo SNP-gene pair matching every sQTL-gene pair in the dataset, similar to the generation process for TADs discussed above. We have



(a) LUAD (b) PRAD

Figure 9: Enrichment of sQTL-gene associations in topologically associating domains. 2 examples, LUAD and PRAD, displaying the results in greater detail after we stratify via the distance between sQTL and splicing position of the genes (x-axis, in 10Kb). All comparisons have $p < 0.001$.

Table 5: Odds ratio and 95% confidence interval (CI) for SNP-gene pairs that map to the identical topologically associating domain among simulated and real datasets

Cancer Type	Odds Ratio	%95 CI	p-value
CML	1.68	[1.66, 1.70]	< 0.001
COAD	1.96	[1.93, 1.99]	< 0.001
GBM	1.83	[1.81, 1.85]	< 0.001
LAML	2.02	[1.97, 2.07]	< 0.001
LUAD	1.77	[1.73, 1.81]	< 0.001
MM	1.86	[1.80, 1.92]	< 0.001
PRAD	1.61	[1.51, 1.71]	< 0.001
SARC	1.73	[1.68, 1.78]	< 0.001

found the number of sQTL-gene pairs locating in the same compartment domain in real dataset to be statistically significantly higher than the ones in simulated dataset according to Fisher's exact test with p-value $< 1 \times 10^{-16}$ for all cancer types as seen in Figure 10. For instance, 59.67% pseudo sQTL-gene pairs in CML tissue were inside TADs, whereas it was 78.25% sQTL-gene pairs in the real dataset.

We have also analyzed the distance distribution between significant Hi-C peaks and significant sQTL-gene associations. We have utilized Fit-Hi-C [49] to estimate the distance between the sQTL and the splicing sites for the whole set of significant sQTL-gene associations and between the significant Hi-C peaks pair of bins. We have observed quite close resulting distributions across different cancer types.

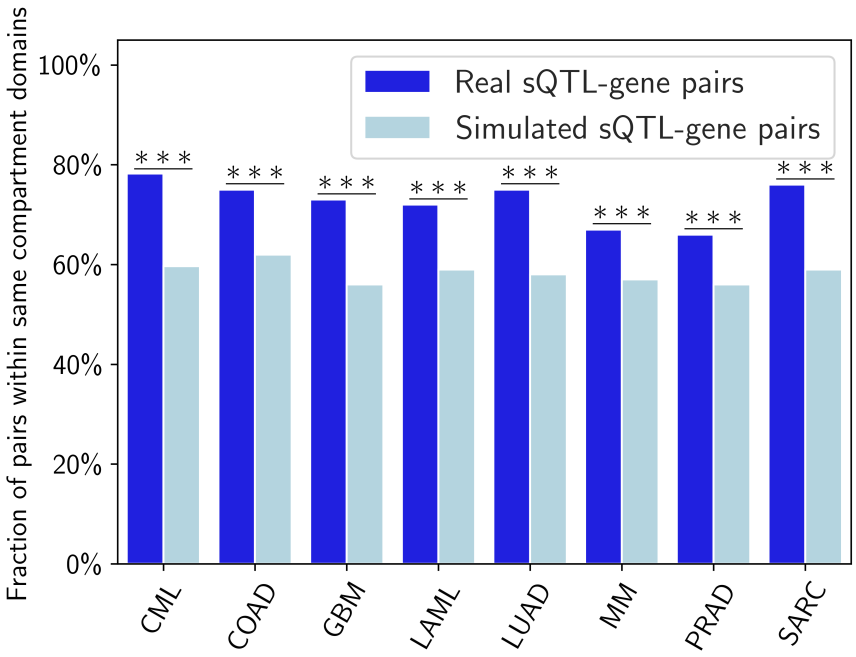


Figure 10: Enrichment of sQTL-gene associations in compartment domains over different cancer types. The ratio of SNP-gene pairs inside compartment domains for simulated and real data, (***) $p < 0.001$ for Fisher's exact test.

3.3 TADs and compartment domains are enriched in terms of GWAS and cancer survival sQTL-gene associations

Similar to the analysis in the previous section, we have also analyzed whether TADs are enriched in terms of GWAS and cancer survival sQTL-gene associations [35]. GWAS sQTLs overlap with the identified GWAS regions that are associated with linkage disequilibrium regions, and they are inferred by integrating the sQTLs with the correct GWAS risk locations. GWAS sQTLs help us in interpreting the genetic variants function. On the other hand, survival sQTLs are sQTLs that are associated with patient's survival time for a given cancer. Identification of survival sQTLs is important in prioritizing sQTLs that may be influencing the cancer prognosis. According to Table 6, for both GWAS sQTLs and cancer survival sQTLs, the number of sQTL-gene pairs locating in the same TAD in real dataset is statistically significantly higher than the ones in simulated dataset according to Fisher's exact test with for

almost all cancer types. Similar to TAD analysis results for the whole set of sQTLs, the identified sQTL-gene associations over TADs are also valid if genomic distances are also taken into account. When logistic regression is applied, the odds-ratio for a SNP-gene pair being in the same TAD is statistically significantly greater for real datasets than for simulated datasets across all cancer tissues.

Table 6: Enrichment of sQTL-gene associations in topologically associating domains over GWAS sQTLs and survival sQTLs. The ratio of SNP-gene pairs inside topologically associating domains for simulated and real data, (***) $p < 0.001$ for Fisher's exact test, (**) $p < 0.01$ for Fisher's exact test.

	GWAS sQTLs			Survival sQTLs		
Cancer Type	Real	Simulated	p-value	Real	Simulated	p-value
CML	70.41%	62.83%	***	72.21%	59.25%	***
COAD	71.11%	61.43%	***	69.75%	60.14%	***
GBM	67.71%	60.88%	***	66.42%	62.34%	**
LAML	69.33%	65.63%	**	70.85%	66.21%	**
LUAD	68.55%	58.95%	***	69.65%	60.07%	***
MM	67.42%	63.39%	**	68.09%	61.12%	***
PRAD	66.41%	59.34%	***	68.19%	61.19%	***
SARC	73.85%	69.48%	**	69.25%	62.66%	***

3.4 Cancer tissue-specific sQTLs are enriched in cancer tissue-specific FIREs

sQTLs are significantly tissue-specific [35]. By taking this result into account, we have also analyzed whether tissue-specific sQTLs are enriched in tissue-specific FIREs. The total number of tissue-specific sQTLs were 443,567 for all 8 cancer tissues we have analyzed. Calculated odds ratio for 7 out of 8 tissues are > 1, which indicates tissue-specific sQTLs are enriched in tissue-specific FIREs. Among these tissues, 6 out of 8 tissues (CML, COAD, LAML, LUAD, PRAD, SARC) are statistically significantly have odds ratio greater than 1 after Bonferroni correction as seen in Table 7, and GBM has nearly satisfied the criteria. Such statistically significant enrichment of tissue-specific sQTLs show that those tissue-specific sQTLs might achieve their splicing effects via tissue-specific chromatin interactions, significant in at least 6 tissues.

However, we have identified a significant negative association in lung cancer tissue. Such negative significant relation can be due to multiple reasons: 1- The cancer sample that is used to create interaction and eQTL dataset might have been collected over different people, or may have not been collected from the same lung cancer tissue locations; 2- A more complex associations could take place between chromatin organization and sQTLs in lung cancer tissue compared to other issues; 3- These samples may not be homogeneous, comprising of various cell type ratios. In order to evaluate the impact of those factors, we need additional experimental datasets. Even though we have not identified a consistent direction for results over cancer tissues, many cancer tissues might exhibit a positive relation between tissue-specific sQTLs and tissue-specific FIREs than a negative relation.

Table 7: sQTLs that are specific to tissues are enriched in tissue-specific FIREs for most cancer types. Odds ratio, 95% confidence interval, and p-value from Fisher's exact test are estimated where confidence interval and p-value are Bonferroni corrected.

Cancer Type	Odds Ratio	%95 CI	p-value
CML	1.58	[1.36, 1.80]	< 0.001
COAD	1.92	[1.72, 2.12]	< 0.001
GBM	1.35	[0.95, 1.75]	0.11
LAML	2.01	[1.83, 2.19]	< 0.001
LUAD	1.33	[1.21, 1.45]	< 0.001
MM	0.85	[0.36, 1.34]	0.75
PRAD	1.42	[1.37, 1.47]	< 0.001
SARC	1.53	[1.44, 1.62]	< 0.001

CHAPTER IV

CONCLUSIONS

Gene regulation is understood to be affected by sQTLs and spatial organization of chromatin. In this research, we have focused on identifying the association between spatial chromatin interactions and sQTL-gene associations across 8 human cancer tissues in a systematic way. We have identified CIF to have positive association with eGenes count. Furthermore, TADs are enriched for sQTL-gene associations. We observed these results in all 8 assessed human cancer tissues. To our best knowledge, this is the first study to come up with well-grounded statistical support for these results across multiple cancer tissues. Our results indicate that sQTLs could achieve regulation of corresponding target genes via interactions between chromatin segments.

Tissue-specific sQTLs are statistically significantly enriched in tissue-specific FIREs in 6 out of 8 human cancer tissues. According to that result, tissue-specific sQTLs could operate via chromatin interactions which are tissue-specific as well, in these 6 tissues. But, the association between tissue-specific sQTLs and chromatin interactions is significantly negative in the lung adenocarcinoma tissue. Such results can be explained by many factors. More complex functioning in the lung cancer tissue can be the principal underlying mechanism. Another factor can be tissue cell types, or sample sources being heterogeneous. Noise may have been presented the possible heterogeneity into our analysis, and we should have obtained more robust results if we have run our analysis on more homogeneous datasets.

The degree of association between tissue-specific chromatin interactions and tissue-specific sQTLs is useful to identify the genes regulated by sQTLs via chromatin interactions over the analyzed cancer tissue. For instance, we have analyzed

3961 tissue-specifc FIREs and 11005 tissue-specific sQTLs in lung cancer tissue. When we consider both of those factors, we have extracted 534 lung adenocarcinoma-specific sQTLs revealed in those lung adenocarcinoma-specific FIREs. We found those sQTLs to be statistically significantly associated with six genes: XPA (DNA Damage Recognition And Repair Factor), EGFR (Epidermal growth factor receptor), KRAS (Kirsten rat sarcoma virus), DDR2 (Discoidin domain-containing receptor 2), ALK (Anaplastic lymphoma kinase), and ADGRB2 (Adhesion G protein-coupled receptor B2). Among these genes, 4 of them, EGFR, KRAS, DDR2, and ALK have been found to have significant roles specifically in lung adenocarcinoma [50].

Transmembrane protein EGFR is an epidermal growth factor receptor regulating the cellular pathways that are involved in cellular proliferation. It is expressed in more than 50% of lung adenocarcinoma types, and is one of the principal focused genes for therapeutics of those tumors [51]. KRAS gene proteins carry growth or differentiate signal from cell exterior to inside cell, especially towards nucleus. Significant increase in the expression of KRAS gene by mutation is a well-studied event in lung adenocarcinoma as well as in different epithelial cancers [52], which is more resistant to various types of cancer treatments. DDR2 mutations are also responsible for adenocarcinoma development, and they have been characterized as common target genes in lung cancer therapy [53]. On the other hand, ALK gene is mainly responsible for gut development during the embyronic development, but it has also roles in lung cancer [54]. Similarly, ADGRB2 exhibits tissue-specific gene expression patterns during the development by encoding a signaling receptor [24], and expression levels increase as the lung develops further [55]. ADGRB2's transcription start site is at 32,192,718 on chromosome 1 which is approximately 47 kb away from tissue-specific FIRE which is between 32,240,000 and 32,320,000 positions of chromosome 1.

We observe similar results for survival-sQTLs and GWAS-sQTLs. We have found those sQTLs to be statistically significantly associated with the same four genes

specific to lung adenocarcinoma development; EGFR, KRAS, DDR2, and ALK. Apart from these four genes, sQTLs are also significantly associated with APC, CASP8, and TP53 genes for survial-sQTLs. Among these genes, APC and CASP8 are associated with the metastatis of lung cancer [56]. APC is a tumor suppressor gene that takes part in cell adhesion, and its mutations are found to exist in lung adenocarcinoma [57]. CASP8 polymorphisms are important in advanced lung cancer prognosis even after therapy [58]. On the other hand, frequent mutations in TP53 gene are well-known markers for a great number of human cancer types, not specific to the lung cancer.

We have performed our analyses over Hi-C datasets with 10kb resolution, which is remarkably higher than the resolution of datasets used in similar joint eQTL and Hi-C analyses. sQTL analysis strength is mainly impacted by sample size. Due to sample size in our datasets, we may have overlooked a number of SNP-gene splicing associations in this study. Additionally, the analysis on topologically associating domain enrichment have not considered linkage disequilibrium (LD) among sQTLs. If there is any effects of LD, they have cancelled out between simulated and real datasets with a very high probability.

As a summary,

- We have discovered sQTLs are located spatially near their target genes with possible alternative splicing across a number of human cancer tissues. Such closer spatial distance also exists independent of whether we integrate eQTLs into the analysis.

- We found that sQTLs regulate the alternative splicing through chromatin interactions.

There results can be used to comprehend the complementary outcomes of sQTLs and chromatin interactions over gene regulations via alternative splicing in detail.

References

[1] A. Takata, N. Matsumoto, and T. Kato, "Genome-wide identification of splicing qtls in the human brain and their enrichment among schizophrenia-associated loci," *Nature Communications*, vol. 8, no. 1, p. 14519, 2017.

[2] S. Djebali, Davis, *et al.*, "Landscape of transcription in human cells," *Nature*, vol. 489, pp. 101–8, 09 2012.

[3] J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, I. Barnes, A. Bignell, V. Boychenko, T. Hunt, M. Kay, G. Mukherjee, J. Rajan, G. Despacio-Reyes, G. Saunders, C. Steward, R. Harte, M. Lin, C. Howald, A. Tanzer, T. Derrien, J. Chrast, N. Walters, S. Balasubramanian, B. Pei, M. Tress, J. M. Rodriguez, I. Ezkurdia, J. van Baren, M. Brent, D. Haussler, M. Kellis, A. Valencia, A. Reymond, M. Gerstein, R. Guigó, and T. J. Hubbard, "Gencode: The reference human genome annotation for the encode project," *Genome Research*, vol. 22, no. 9, pp. 1760–1774, 2012.

[4] E. Wang, R. Sandberg, S. Luo, I. Khrebtukova, k. Mayr, S. Kingsmore, G. Schroth, and C. Burge, "Alternative isoform regulation in human tissue transcriptomes," *Nature*, vol. 456, pp. 470–6, 12 2008.

[5] E. Lalonde, K. C. Ha, Z. Wang, A. Bemmo, C. L. Kleinman, T. Kwan, T. Pastinen, and J. Majewski, "Rna sequencing reveals the role of splicing polymorphisms in regulating human gene expression," *Genome Research*, vol. 21, no. 4, pp. 545–554, 2011.

[6] N. A. Faustino and T. A. Cooper, "Pre-mrna splicing and human disease," *Genes & Development*, vol. 17, no. 4, pp. 419–437, 2003.

[7] A. Sveen, S. Kilpinen, A. Ruusulehto, R. Lothe, and R. Skotheim, "Aberrant rna splicing in cancer; expression changes and driver mutations of splicing factor genes," *Oncogene*, vol. 35, 08 2015.

[8] L. Wang, L. Duke, P. S. Zhang, R. B. Arlinghaus, W. F. Symmans, A. Sahin, R. Mendez, and J. L. Dai, "Alternative splicing disrupts a nuclear localization signal in spleen tyrosine kinase that is required for invasion suppression in breast cancer," *Cancer Research*, vol. 63, no. 15, pp. 4724–4730, 2003.

[9] E. Wu, T. Nance, and S. B. Montgomery, "SplicePlot: a utility for visualizing splicing quantitative trait loci," *Bioinformatics*, vol. 30, pp. 1025–1026, 12 2013.

[10] R. Lowe, N. Shirley, M. Bleackley, S. Dolan, and T. Shafee, "Transcriptomics technologies," *PLOS Computational Biology*, vol. 13, pp. 1–23, 05 2017.

[11] F. Aguet, Ardlie, *et al.*, "Genetic effects on gene expression across human tissues," *Nature*, vol. 550, pp. 204–213, 10 2017.

[12] Y. Gilad, S. A. Rifkin, and J. K. Pritchard, "Revealing the architecture of gene regulation: the promise of eqtl studies," *Trends in Genetics*, vol. 24, no. 8, pp. 408–415, 2008.

[13] C. Trapnell, Williams, *et al.*, "Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation," *Nature Biotechnology*, vol. 28, no. 5, pp. 511–515, 2010.

[14] D. Garrido-Martín, B. Borsari, M. Calvo, F. Reverter, and R. Guigó, "Identification and analysis of splicing quantitative trait loci across multiple tissues in the human genome," *Nature Communications*, vol. 12, no. 1, p. 727, 2021.

[15] F. W. Albert and L. Kruglyak, "The role of regulatory variation in complex traits and disease," *Nature Reviews Genetics*, vol. 16, no. 4, pp. 197–212, 2015.

[16] C. D. Brown, L. M. Mangravite, and B. E. Engelhardt, "Integrative modeling of eqtls and cis-regulatory elements suggests mechanisms underlying cell type specificity of eqtls," *PLOS Genetics*, vol. 9, pp. 1–19, 08 2013.

[17] N. Heidari, D. H. Phanstiel, C. He, F. Grubert, F. Jahanbani, M. Kasowski, M. Q. Zhang, and M. P. Snyder, "Genome-wide map of regulatory interactions in the human genome," *Genome Research*, vol. 24, no. 12, pp. 1905–1917, 2014.

[18] A. Dean, "In the loop: long range chromatin interactions and gene regulation," *Briefings in Functional Genomics*, vol. 10, pp. 3–10, 01 2011.

[19] G. Li, X. Ruan, and R. others, "Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation," *Cell*, vol. 148, no. 1, pp. 84–98, 2012.

[20] Y.-C. Hwang, Q. Zheng, B. D. Gregory, and L.-S. Wang, "High-throughput identification of long-range regulatory elements and their target promoters in the human genome," *Nucleic Acids Research*, vol. 41, pp. 4835–4846, 03 2013.

[21] S. S. P. Rao, Huntley, *et al.*, "A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping," *Cell*, vol. 159, pp. 1665–1680, 2022/03/03 2014.

[22] E. Lieberman-Aiden, N. L. van Berkum, *et al.*, "Comprehensive mapping of long-range interactions reveals folding principles of the human genome," *Science*, vol. 326, no. 5950, pp. 289–293, 2009.

[23] F. Jin, Y. Li, J. R. Dixon, S. Selvaraj, Z. Ye, A. Y. Lee, C.-A. Yen, A. D. Schmitt, C. A. Espinoza, and B. Ren, "A high-resolution map of the three-dimensional chromatin interactome in human cells," *Nature*, vol. 503, no. 7475, pp. 290–294, 2013.

[24] J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren, "Topological domains in mammalian genomes identified by analysis of chromatin interactions," *Nature*, vol. 485, no. 7398, pp. 376–380, 2012.

[25] E. Sefer, G. Duggal, and C. Kingsford, "Deconvolution of ensemble chromatin interaction data reveals the latent mixing structures in cell subpopulations," *Journal of Computational Biology*, vol. 23, no. 6, pp. 425–438, 2016. PMID: 27267775.

[26] E. Sefer, "A comparison of topologically associating domain callers over mammals at high resolution," *BMC Bioinformatics*, vol. 23, no. 1, p. 127, 2022.

[27] E. Sefer and C. Kingsford, "Semi-nonparametric modeling of topological domain formation from epigenetic data," *Algorithms for Molecular Biology*, vol. 14, no. 1, p. 4, 2019.

[28] E. Sefer, "Hi–c interaction graph analysis reveals the impact of histone modifications in chromatin shape," *Applied Network Science*, vol. 6, no. 1, p. 54, 2021.

[29] E. Sefer, "Probc: joint modeling of epigenome and transcriptome effects in 3d genome," *BMC Genomics*, vol. 23, no. 1, p. 287, 2022.

[30] G. Duggal, H. Wang, and C. Kingsford, "Higher-order chromatin domains link eQTLs with the expression of far-away genes," *Nucleic Acids Research*, vol. 42, pp. 87–96, 10 2013.

[31] J. Yu, M. Hu, and C. Li, "Joint analyses of multi-tissue hi-c and eqtl data demonstrate close spatial proximity between eqtls and their target genes," *BMC Genetics*, vol. 20, no. 1, p. 43, 2019.

[32] A. Das, Morley, *et al.*, "Bayesian integration of genetics and epigenetics detects causal regulatory snps underlying expression variability," *Nature Communications*, vol. 6, no. 1, p. 8555, 2015.

[33] B. He, C. Chen, L. Teng, and K. Tan, "Global view of enhancer-promoter interactome in human cells," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 111, 05 2014.

[34] G. Ron, Y. Globerson, D. Moran, and T. Kaplan, "Promoter-enhancer interactions identified from hi-c data using probabilistic models and hierarchical topological domains," *Nature Communications*, vol. 8, no. 1, p. 2237, 2017.

[35] J. Tian, Wang, *et al.*, "CancerSplicingQTL: a database for genome-wide identification of splicing QTLs in human cancer," *Nucleic Acids Research*, vol. 47, pp. D909–D916, 10 2018.

[36] A. Schmitt, M. Hu, I. Jung, Z. Xu, Y. Qiu, C. Tan, Y. Li, S. Lin, Y. Lin, C. Barr, and B. Ren, "A compendium of chromatin contact maps reveals spatially active regions in the human genome," *Cell Reports*, vol. 17, no. 8, pp. 2042–2059, 2016.

[37] S. E. Johnstone, Reyes, *et al.*, "Large-scale topological changes restrain malignant progression in colorectal cancer," *Cell*, vol. 182, pp. 1474–1489.e23, 2021/03/31 2020.

[38] M. J. Johnston, Nikolic, *et al.*, "High-resolution structural genomics reveals new therapeutic vulnerabilities in glioblastoma," *Genome Research*, vol. 29, no. 8, pp. 1211–1222, 2019.

[39] I. Dunham, Kundaje, *et al.*, "The encode project consortium: An integrated encyclopedia of dna elements in the human genome. 2012. nature 489: 57–74," *Nature*, vol. 489, pp. 57–74, 09 2012.

[40] P. Wu, T. Li, R. Li, L. Jia, P. Zhu, Y. Liu, Q. Chen, D. Tang, Y. Yu, and C. Li, "3d genome of multiple myeloma reveals spatial genome disorganization associated with copy number variations," *Nature Communications*, vol. 8, 12 2017.

[41] S. K. Rhie, A. A. Perez, F. D. Lay, S. Schreiner, J. Shi, J. Polin, and P. J. Farnham, "A high-resolution 3d epigenomic map reveals insights into the creation of the prostate cancer transcriptome," *Nature Communications*, vol. 10, no. 1, p. 4154, 2019.

[42] O. Kantidze, Luzhin, *et al.*, "The anti-cancer drugs curaxins target spatial genome organization," *Nature Communications*, vol. 10, p. 1441, 03 2019.

[43] J. Gong, Mei, *et al.*, "PancanQTL: systematic identification of cis-eQTLs and trans-eQTLs in 33 cancer types," *Nucleic Acids Research*, vol. 46, pp. D971–D976, 09 2017.

[44] A. Macauda, Piredda, *et al.*, "Expression quantitative trait loci of genes predicting outcome are associated with survival of multiple myeloma patients," *International Journal of Cancer*, vol. 149, no. 2, pp. 327–336, 2021.

[45] D. Filippova, R. Patro, G. Duggal, and C. Kingsford, "Identification of alternative topological domains in chromatin," *Algorithms for Molecular Biology*, vol. 9, no. 1, p. 14, 2014.

[46] K.-K. Yan, S. Lou, and M. Gerstein, "Mrtadfinder: A network modularity based approach to identify topologically associating domains in multiple resolutions," *PLOS Computational Biology*, vol. 13, pp. 1–22, 07 2017.

[47] C. Crowley, Y. Yang, *et al.*, "Firecaller: Detecting frequently interacting regions from hi-c data," *Computational and Structural Biotechnology Journal*, vol. 19, pp. 355–362, 2021.

[48] J. Dixon, D. Gorkin, and B. Ren, "Chromatin domains: The unit of chromosome organization," *Molecular Cell*, vol. 62, no. 5, pp. 668–680, 2016.

[49] F. Ay, T. L. Bailey, and W. S. Noble, "Statistical confidence estimation for hi-c data reveals regulatory chromatin contacts," *Genome Research*, vol. 24, no. 6, pp. 999–1011, 2014.

[50] A. El-Telbany and P. C. Ma, "Cancer genes in lung cancer: racial disparities: are there any?," *Genes & cancer*, vol. 3, pp. 467–480, 07 2012.

[51] G. da Cunha Santos, F. A. Shepherd, and M. S. Tsao, "Egfr mutations and lung cancer," *Annual Review of Pathology: Mechanisms of Disease*, vol. 6, no. 1, pp. 49–69, 2011. PMID: 20887192.

[52] P. M. K. Westcott and M. D. To, "The genetics and biology of kras in lung cancer," *Chinese Journal of Cancer*, vol. 32, pp. 63–70, 02 2013.

[53] H. Terai, L. Tan, E. M. Beauchamp, J. M. Hatcher, Q. Liu, M. Meyerson, N. S. Gray, and P. S. Hammerman, "Characterization of ddr2 inhibitors for the treatment of ddr2 mutated nonsmall cell lung cancer," *ACS Chemical Biology*, vol. 10, pp. 2687–2696, 12 2015.

[54] T. Le and D. E. Gerber, "Alk alterations and inhibition in lung cancer," *Seminars in Cancer Biology*, vol. 42, pp. 81–88, 02 2017.

[55] A. A. Gad and N. Balenga, "The emerging role of adhesion gpcrs in cancer," *ACS pharmacology & translational science*, vol. 3, pp. 29–42, 01 2020.

[56] Y. Lu, Lemon, *et al.*, "A gene expression signature predicts survival of patients with stage i non-small cell lung cancer," *PLoS medicine*, vol. 3, pp. e467–e467, 12 2006.

[57] H. Ohgaki, J. M. Kros, Y. Okamoto, A. Gaspert, H. Huang, and M. O. Kurrer, "Apc mutations are infrequent but present in human lung cancer," *Cancer Letters*, vol. 207, no. 2, pp. 197–203, 2004.

[58] D. Liu, W. Xu, X. Ding, Y. Yang, Y. Lu, K. Fei, and B. Su, "Caspase 8 polymorphisms contribute to the prognosis of advanced lung adenocarcinoma patients after platinum-based chemotherapy," *Cancer Biology & Therapy*, vol. 18, pp. 948–957, 12 2017.

VITA

Mr. Batuhan Eralp his bachelors in Acıbadem Mehmet Ali Aydınlar University Faculty of Engineering and Natural Sciences Biomedical Engineering department in 2020. After graduating, he co-founded Solidus Medical Products and Software Informatics Co. Ltd. and works on clinical decision support systems. He also joined Ozyegin University and started working with Asst. Prof. Emre Sefer