FUNCTIONAL ANALYSIS OF STRUCTURAL VARIATION IN THE 2D AND 3D HUMAN GENOME

by

Conor Mitchell Liam Nodzak

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Bioinformatics and Computational Biology

Charlotte

2019

Approved by:

Dr. Xinghua Mindy Shi

Dr. Rebekah Rogers

Dr. Jun-tao Guo

Dr. Adam Reitzel

ABSTRACT

CONOR MITCHELL LIAM NODZAK. Functional analysis of structural variation in the 2D and 3D human genome. (Under the direction of DR. XINGHUA MINDY SHI)

The human genome consists of over 3 billion nucleotides that have an average distance of 3.4 Angstroms between each base, which equates to over two meters of DNA contained within the 125 μm^3 volume diploid cell nuclei. The dense compaction of chromatin by the supercoiling of DNA forms distinct architectural modules called topologically associated domains (TADs), which keep protein-coding genes, noncoding RNAs and epigenetic regulatory elements in close nuclear space. It has recently been shown that these conserved chromatin structures may contribute to tissue-specific gene expression through the encapsulation of genes and cis-regulatory elements, and mutations that affect TADs can lead to developmental disorders and some forms of cancer. At the population-level, genomic structural variation contributes more to cumulative genetic difference than any other class of mutation, yet much remains to be studied as to how structural variation affects TADs. Here, we study the functional effects of structural variants (SVs) through the analysis of chromatin topology and gene activity for three trio families sampled from genetically diverse populations from the Human Genome Structural Variation Consortium. We then leverage clinically-relevant recurrent genomic rearrangements in acute lymphoblastic leukemia and propose a machine learning approach to identify the rare Philadelphia-like subtype based on the gene activities within lymphoblastoid chromatin domains. This analysis has found that TADs may improve our understanding of how SVs contribute to diverse gene expression patterns in health and disease.

DEDICATION

I would like to dedicate this work to my colleagues, friends and family for their valuable support and feedback. Above all, to my loving wife, *Natasha*, who gifted me this opportunity through support that remains straightforward, and without complexities.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Xinghua Mindy Shi, for her continual encouragement, support and guidance throughout the duration of my studies. I have tremendous gratitude towards my committee members, Dr. Rebekah Rogers, Dr. Jun-tao Guo and Dr. Adam Reitzel, for their valuable constructive feedback in the development of this work. Additionally, I would like to acknowledge the several colleagues in the HGSV consortium and at the Laboratorio de Biologia Molecular, Centro Infantil Boldrini, Campinas, Sao Paulo, Brazil for providing the data, feedback and opportunity to conduct the enriching collaborations necessary to produce this work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ALL	Acute lymphoblastic leukemia
ASE	Allele-specific expression
BAM	Binary sequence alignment map
BED	Browser extendable data
BP	Base-pair
DI	Directionality Index
DNA	Deoxyribonucleic acid
DNA-Seq	Deoxyribonucleic acid sequencing
FDR	False discovery rate
GTF	Gene transfer format
Hi-C	High-throughput chromatin conformation capture
INDEL	Small Insertion/Deletion
LNCRNA	Long, noncoding ribonucleic acid
RNA	Ribonucleic acid
RNA-Seq	Ribonucleic acid sequencing
SAM	Sequence alignment map
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
SV	Structural Variant

TAD	Topologically associated domain
VCF	Variant call file

CHAPTER 1: INTRODUCTION

The human genome can be varied in a number of different ways. Typically, a scientific investigation of human genetic variation would catalogue single nucleotide polymorphisms (SNPs), small insertions and deletions (INDELs), or larger structural variants (SVs). Each class of mutation is defined by the nucleotide length of the change, where a SNP contains a single base change and may result in a transitions or transversions depending upon the nitrogenous bases involved. Transitions and transversions may further be annotated as silent, frameshift, missense, or nonsense alterations to the genic readout, each with different affects on downstream protein product synthesis. The delineation of variant classes greater than one nucleotide base is arbitrarily defined by the amount of nucleotide content modified, however, in common practice a standardized system is used based on size where an INDEL refers to any insertion or deletion less than 50 bases long, and a structural variant describes any genomic event greater in size. Although subjectively defined, these length-based classifications have come to be widely accepted by researchers interested in human genomic variation, moreover it has been shown that each class contributes variable amounts of cumulative nucleotide-content changed between genomes when viewed across global populations. Indeed, efforts by the 1000 Genomes, the Welcome Trust Case-Control consortium and other groups have determined SNPs contribute about 0.3% cumulative difference, INDELs contribute 0.3%, while SVs contribute the largest amount of cumulative locus-length difference with 0.8% of the genome covered [1, 2, 3]. This high level of variability caused by SVs suggests that they represent a natural mutational phenomenon, which leads to evolutionary divergence across populations. Furthermore, the amount of naturally occurring cumulative change attributable to

SVs implicates such cellular generational acquisition of somatic events would lead to modifications to the genome over time and be associated with aberrant gene expression phenotypes.

Mechanistically, there are common biological processes which lead to the generation of intrachromosomal structural variants. These mutational modes include non-allelic homologous recombination, nonhomologous end-joining, fork stalling and template switching, and L1-mediated retrotransposition [4]. Additionally, microhomology-mediated break-induced replication at breakpoint contributes rearrangements of chromosomal segments [5]. Each modality represents an endogenous cellular process or the activity of naturally occurring mobile elements. These may manifest as SVs by creating unbalanced copy number variations (insertions, duplications, deletions) or as balanced copy number neutral variants (translocations, inversions).

In order to study structural variation in the human genome, we currently rely upon arrays and sequencing based methods to allow for their detection and subsequent genotyping. Each approach carries with it a unique set of strengths and weaknesses towards this end. For instance, SNP-Microarrays utilize the hybridization of sequencing reads to probes followed by computational signal detection in order to infer copy number variation. These SNP-array platforms may suffer in terms of signal-to-noise ratios, yet offer a cost efficient method to detect copy-number differences for large sample cohorts in population genetics and biomedical research. With the advent of modern high-throughput sequencing technologies, limitations remain for the detection and genotyping of structural variants imposed by sequencing depth and read length [6]. Sufficient depth is required to resolve deletions and duplications as compared to the reference genome, while the short sequencing read length makes the detection of SVs in repetitive regions of the genome especially cumbersome. The key advantages for computational biologists using a sequencing-based approach comes in the form of genomic localization and increased resolution by means of four main strategies including read-depth, paired-reads, split-reads and de-novo assembly. A read-depth approach compares aligned reads to the region in order to detect anomalies in the pile-ups of sequencing reads. A region with fewer than expected aligned reads may be interpreted as a deletion, while a tandem duplication would be recognized by roughly twice as many reads than either flanking region of the SV. Paired-read analysis detects deletions and insertions when read pairs are aligned to the reference and compares relative distances to the expected library insert size. Similarly, tandem duplication appear as read pairs in an unexpected order, while inversions show unexpected orientation and translocations appear as pairs mapped to different chromosomes. Another method maps breakpoints by splitting reads when aligned to the reference and detecting gaps in the reads or reference after alignment. Finally, de-novo assembly methods do not use a reference genome and instead build contigs from novel insertion sequence reads. The usefulness of each of these methods for each type of SV is inconsistent, which has led to the development of algorithms, software, and experimental designs that employ one or more strategies to allow for more robust SV detection.

The importance and significant interest in detecting SVs comes form the inherent large contribution that they make to genomic variation at the population level, as well as their long-recognized role in the development of human diseases, such as cancer. Recently, collaborators in the Human Genome Structural Variation Consortium (HGSVC) have used a multi-platform approach that combines several different long-read and short-read sequencing technologies to better resolve structural variation for trios of families sampled by the 1000 Genomes Project. The goal of the HGSVC is to produce an well-characterized and highly validated set of naturally occurring SVs with methodological development to integrate these technologies and multiple computational techniques for better detection and characterization. This work has developed a multitude of algorithmic approaches to resolve consensus breakpoints, as well as produced a wealth of multi-omics data in addition to the genotyped structural variants. In order to better understand the phenotypic consequences of these structural variant, we present here a functional analysis and contributions made towards the goal of the HGSVC through the development of bioinformatic techniques to detect phenotypic changes in gene function and regulation through the integration of expression and epigenetic datasets. From the analysis of these data, we hope to glean a deeper appreciation by which SVs affect chromatin architecture, regulatory elements as well as the overall impact on the activity of protein-coding and long, noncoding RNA genes.

In tumor oncology, structural variation plays a well-known role in the development and progression of human cancers. Copy-number variants (CNVs), which are structural variants such as deletions and duplications that affect the number of gene copies per-genome can lead to dosage or positional effects, and undergo either positive or purifying selective pressures throughout evolutionary processes at the cellular and species levels [7]. Through missteps in processes of replication and recombination, these CNVs accumulate 10^3 times faster than single nucleotide variants (SNVs) and rapidly change the structure and function of a cells genome, and promote oncogenesis [8]. The relatively inexpensive array-based platforms have allowed for the clinical detection and interpretation of CNVs involved in cancer predisposition syndromes clinically-actionable targets [9]. In studies of pediatric B-cell leukemia, copy-number alterations were found to be common and frequently recurrent at loci of genes involved in key checkpoints of B-cell development, with cancer progression consequential to the aberrant expression [10].

Similarly, we can leverage common recurrent sets of translocations to better understand how SVs create subtype-specific expression patterns in pediatric patients diagnosed with precursor acute lymphoblastic leukemia (ALL). We begin with an analysis of global expression patterns with a particular emphasis on the rare Philadelphia-like (Ph-like) subtype of this disease, which is recognized as possessing a similar transcriptional profile to those patients with a reciprocal translocation between chromosome 9 and 22, Philadephia-chromosome positive (Ph(+)), yet the Ph-like samples lack this particular SV. Given that this translocation produces a fusion protein kinase with unregulated capacity to phosphorylate proteins in the JAK/STAT signalling pathway and thus activate growth cascades, we hypothesize that the observed similarities between these two subtypes may be driven by SV-mediated disruptions to the chromatin architecture, which causes irregular activity of genes through epigenetic modulation and disruption of gene regulatory elements. To evaluate this proposition, CNVs will be detected for samples with ALL and subsequently evaluated for their effects on topologically associated domains (TADs) and the coding and non-coding genes localized in proximal nuclear space with cis-regulatory elements.

CHAPTER 2: BACKGROUND

Functional investigations into the human genome generally fall into two categories, either genetics, which considers how the sequence of nucleotides contributes to an organism's development, phenotype, and intra-species variation, or epigenetics, which has an overall goal to study aspects of genomic regulation that do not entail a change to the DNA sequence content. While molecular genetics has provided an immense knowledgebase from which we better understand evolution and cellular function, epigenetics has filled some of the gaps as to how tissue-specific expression patterns are possible given an organisms cells share the same genome. When we consider trans-generational inheritance of allele silencing, such as imprinted loci, or induction of gene activity by environmental stimuli, we can begin to appreciate the complex milieu of regulatory process that must be at work in the cellular epigenetics cycle, and how endogenous and exogenous effects exhibit phenotypic changes. The field of epigenetics has a long scientific history linking genome structure and functional state, stemming from the discovery of proteins that provide structural integrity to DNA may also dictate the accessibility of transcriptional machinery in order to regulate gene activity.

This relationship has been further characterized by the discovery of untranslated RNA molecules that take part in various biochemical mechanisms as protein scaffolds or exhibiting direct and indirect regulation of mRNA transcription and translation. Moreover, studies have recently linked the three-dimensional chromatin architecture to yet another layer of genomic regulation, which has showed how genes and regulatory elements tend to be proximally localized along chromosome domains and active chromatin co-localizes in nuclear compartments. The connection between genome activity, structure, and noncoding RNAs provides a highly complex swathe of modes by which eukaryotic genomes may be self-regulated to create a vast array of cell-types, tissue-specific expression and even be altered to promote within-species diversity and disease phenotypes. Here we attempt to contextualize the current state of epigenetic gene regulation through descriptions of various DNA-binding proteins, the noncoding genome, and detail the importance of structural variants, methods for their detection and functional interpretation, as well as known cases in which structural variants alter epigenetic factors to lead to developmental effects and cancer.

2.1 Epigenetics, Epigenomics and the non-Coding Genome

2.1.1 Introduction

In the early 1940s, prior to Oswald Avery's discovery that DNA was the main molecular mode of heredity, revolutionary geneticist Conrad Hal Waddington published a theoretical description for the contemporary state of knowledge in the fields of embryology and developmental biology that espoused the delineation of an organism's genetic makeup into categories of organizers and genes [11, 12]. This fundamental representation of the genome led Waddington to develop the construct of the epigenetic landscape, where cellular differentiation may be illustrated as a stream of water flowing downward over *chreodes*, a neologism for a multitudinous set of divergences in paths, towards a final resting position that represented the multiple state-space transitions taken as an cell experiences environmental fluctuations and the embryo progresses towards a developed organism with fully differentiated tissues [13]. Waddington elaborated on this pioneering visualization for development with descriptions of cellular specification as *individuation*, and variable *competence* of inductive molecular signals triggered by the environment that guide cellular specification down common *canalizations* to minimize perturbations to the system [14]. Furthermore, the *epigenotype* was presented as a model for canalization to explain observations wherein simple mutation was not a sufficient causal explanation for observed evolutionary inheritance patterns, and further suggests that organismal development could be described by a complex interaction between the environmental stimulation of effectors and genetic makeup [15, 16].

At the time, the burgeoning field of epigenetics was therefore a loosely defined conceptual exercise surrounding resultant observable phenotypic variation with respect to environmental interaction forces. Many interpretations of the principle focus have been proposed, while the research community has been acquiescent towards an emphasis on the mechanisms of genetic regulation that underlie tissue and developmentally specific patterns of gene expression capable of being produced by cells that share a single genome within an individual [17].

A modern formalization for epigenetics was published by *Berger et al.*, which promulgated a set of chromatin centric mechanisms for the distinct molecular regulation required for the transient production, maintenance and transmission of phenotypic variation not consequent to DNA sequence mutation. As such, this work introduced relevant terminology for biological pathways that act to produce these epigenetic effects as *epigenators*, *initiators* and *maintainers* of chromatin state. The jargon of epigenetic analysis loosely reflects the philosophy of Waddington's theoretical epigenetic landscape, where an *epigenator* denotes any temporary environmental signal or extracellular milieu that leads to activation of the *initiator*. For example, the *initiator* may be a noncoding RNA or other DNA binding factor that is capable of generating a shift in the chromatin state towards an epigenetic phenotype of either repression or activation at specific genomic locations. The *maintainer* may then be a histone acetyltransferase or methyltransferase enzyme that propagates the epigenetic phenotype through time, with some maintenance through generations collectively referred to as epigenetic memory [18]. These standardized definitions for epigenetics represents a rapprochement of the founding ideological principles with advanced technological understanding to succinctly capture the regulatory modules of genes that produce phenotypic variability.

With the advent of high-throughput sequencing technology, a plethora of epigenetic datasets have been produced that probe the modifications to histones and expression of noncoding RNA and allow for a genome-wide perspective. As epigenomic regulatory connections continue to be resolved, a growing set of initiators and maintainers have been identified by the consequence to the expression of genes during development and disease. Large epigenetics consortiums such as the Encyclopedia of DNA Elements (ENCODE) and the NIH's Roadmap Epigenomics Project have produced expansive reference datasets for the epigenetics research community [19, 20]. Additionally, analyses by ENCODE research groups have provided dynamic representations of chromatin state modulation over several distinct cell lines [21]. Taken together, these resources may be utilized for the identification of regulatory modules and subsequent comparative and functional analyses across a spectrum of tissues in healthy and diseased samples.

2.1.2 Mechanistic Regulation of Chromatin State

Epigeneticists attempt to resolve how a single genome can create a vast array of cellular phenotypes, each with different patterns of gene expression. Biological processes of transgenerational inheritance must be at work other than random evolutionary DNA sequence mutation and sites of polymorphic nucleotides. Scientific inquiry at multiple levels of genome organization and structural composition led to the discovery that chemical modification of specific amino acid residues on histone octamer subunits can alter the activity and accessibility of chromosomal loci, referred to as chromatin state, and lead to differences in expression of genes not caused by DNA sequence mutation [17, 22]. Post-translational modifications may manifest in the form of variable amounts of methyl, acetyl or phosphoryl functional groups added to a collection of histone amino acid residues, which have been shown to be characteristic of specific regulatory elements and chromatin state. A definitive list of several histone subunit chemical modifications has been published by the ENCODE consortium that provides markers with known associations with heterochromatin, euchromatin, bivalent chromatin and regulatory capacity of promoters and enhancers [19].

Therefore, in order to maintain expression phenotypes across an organismal body plan, somatic cellular division must be concomitant with a mechanism for the tissue-specific epigenetic patterning and regulation of genes. In humans, Groth and colleagues were able to demonstrate chromatin state stability is mediated by the dynamic interplay of histone recycling and synthesis at the head of the replication fork progression that combines parental transfer of bound, modified histones to the daughter strand mediated by the *Asf1* protein with associated deposition of newly translated histones. [23]. The mechanism of chaperone guided histone recycling provides a means for chromatin state propagation. For example, existing methylation of the ninth lysine of histone subunit 3 (H3K9) deposited by the methyltransferase *Suv39 H1* is associated with heterochromatin, which further allows for the binding of *HP1*, which can in turn recruit the *Suv39 H1* protein to maintain transgenerational cellular inheritance of the closed-form chromatin state at specific regions of the genome [24].

Some of the most important protein complexes in epigenetic regulation and chromatin state modification are the Polycomb Group proteins, *PcG*, and the Trithorax Group, *trxG*, which respectively mediate gene silencing and activation [25]. Polycomb Group proteins are responsible for maintenance of the silent chromatin state marker, trimethylation of histone 3 at lysine 27 (H3K27me3), throughout development by methyltransferase activity [26]. *PcG* proteins are vital regulators of developmental cell-fate specification along body axes through HOX genes transcriptional repression [27]. The *PcG* complex is capable of histone modifications at specific chromatin sites through interactions with various other proteins, and in some known cases the precise localization is mediated by long noncoding RNA recruitment [28, 22]. This capability was shown by Zhang et al. who found that a Polycomb gene encodes an RNA binding protein, which suggests there may be additional regulatory long noncoding RNAs that need to be investigated [29]. Additional evidence was also shown whereby Polycomb proteins preferentially bind a chromatin domain based on the presence of noncoding RNA molecules [30]. Similarly, the active state of chromatin may be delineated by the regional presence of histone 3 lysine 4 trimethylation (H3K4me3), which can be induced and maintained by action of several proteins that constitute *trxG* complexes. [31].

2.1.3 Gene Regulation in Nuclear Space

Classical genetics has long demonstrated the important role that location and distance may exert on observable phenotypic variation. Muller's 1930 experimental findings showed that variegated eye coloration in *Drosophila* could be attributed to the position along a chromosome of induced mutations by X-ray exposure, the event dubbed position effect variegation (PEV) [32]. Two decades later, Mclintock's groundbreaking discovery that transposable elements control variable kernel coloration in maize through insertion at random locations about the genome was published, further suggesting a critical regulatory role in the relative positions of genomic elements. [33].

The importance of genomic position has continued to underlie contemporary research discoveries in genetics. With the adoption of modern high-throughput sequencing techniques (DNAseq and RNAseq), chromatin immunoprecipitation assays (ChIPseq) and chromatin conformation capture technologies (3C, 4C, Hi-C, etc.) provide a more complete lens to study the complexity of human genetics. It is now possible to obtain the sequence, determine the state of chromatin markers across whole genomes, identify exact loci of mutations ranging from single-nucleotides to large structural variants, and finally contextualize this information within the 3-dimensional compartmentalization of the nucleus. A recent pan-cancer study of all samples from The Cancer Genome Atlas (*TCGA: http://cancergenome.nih.gov/*) integrated knowledge of cell-line specific chromatin topology with the sample-level copy number variants in a computational framework to identify structural variants that significantly altered gene expression [34, 35]. In this study, Weischenfeldt et al. were able to show recurrent duplications and deletions at the boundaries of topologically associated domains capable of driving aberrant gene expression through enhancer hijacking events. Enhancer hijacking refers to the phenomenon whereby genomic structural variation disrupts the three-dimensional topology of the chromatin leading to altered contact frequencies between loci. Cognate epigenetic regulators of specific genes, i.e. enhancers, are no longer able to associate with the promoter region and a reduction in expression may be observed. On the other hand, a gene may be over-expressed if the altered chromatin topology introduces the possibility for increased contact frequency between a non-cognate enhancer and the promoter.

2.1.4 Epigenetic Regulation of Hematopoiesis and Cancer

Hematopoiesis encompasses the combined developmental processes that lead to the production of various cell-type constituents of blood, including those of the adaptive and innate immune systems [36]. During embryonic development, a phase of definitive hematopoiesis occurs in which a population of multipotent hematopoietic stem cells (HSC) are ultimately established in the bone marrow that allow for the production of all blood-cell lineages well into adulthood [37]. The overall process of HSC production is regulated by complex time-dependent actions of several sets of adhesion molecules, tyrosine kinase receptors, growth factors, and transcription factors [37]. The developmental cascade activated during definitive hematopoiesis leads to the production of both myeloid and lymphoid progenitor cells, which in turn may become erythrocytes, mast cells, platelets, neutrophils, eosinophils, macrophages, basophils, and T-cells and B-cells [38]. Throughout the cell-fate specification of these mature types, each cell-type has a unique set of regulators with context-dependent expression patterns, leaving multiple causative avenues for hematological malignancies when disrupted. [36]. For an example of this context dependency, the activation and up-regulation of the *PU.1* gene leads to the generation of macrophages, yet low levels of this gene will lead to B-cell development [39]. In addition, it is well-established that the transcription factors *E2A*, and *Pax5* are critical to the production of B-cells and loss of potency, as knock-down experiments confirmed that the progenitor may develop into a variety of cell-types in the absence of the *Pax5* transcription factor [40]. These transcription factors allow for a precise regulation of progenitors leading to programming of the multiple blood cell types, however there is mounting evidence showing epigenetic regulation provides another level of regulation including the action of micro-RNAs (miRNAs) to impede or drive certain cell types by targeting various transcription factors [41]. The work by Xiao et al. and others confirmed stage-specific expression patterns of miRNAs is involved in lymphocyte development and likely function where loss of the *miR-150* miRNA leads to an arrested B-cell stage during development and expansion of immature cells, analogous to observations during leukemias [42, 43, 44].

Moreover, there is increasing evidence of the emerging role of long noncoding RNAs (lncRNAs) in the epigenetic regulation of processes of hematopoiesis and malignancies. These lncRNAs have been implicated in both epigenetic activation and chromatin silencing through a variety of mechanisms that drive aberrant expression of proto-oncogenes and tumor suppressors [45]. For example, studies of chronic lymphocytic leukemia found that the lncRNA *DLEU2* exhibits tumor suppressor activity in B-cells through regulation the cell-cycle and induction of apoptosis, and is often silenced in B-cell of patients with the disease [46]. Additionally, myeloproliferative disorders appear to be induced by the suppression of the lncRNA *Xist*, leading to modification of the chromatin state and activity of developmental genes and progression of blood cancers associated with X-chromosome dosage [47].

The lineage programming carried out during hematopoiesis is highly dependent upon key transitions of histone marks, deemed the chromatin state dynamics. Chromatin immunoprecipitation experiments showed that over 17,000 enhancer elements are established to promote lineage-specific transcription patterns characterized by a complex network of transcription factors that control the overall chromatin state and developmental progression [48]. Disruptions of epigenetic networks is a known causal driver of several types of disease, especially cancers [49, 50]. Indeed, malignancies of myeloid cells and leukocytes appear to show epigenetic changes to the DNA, both cases exhibit global hypomethylation and increased expression of methyltransferases, indicating a strong connection with blood malignancies and chromatin state [51]. Furthermore, the activity of recurrent genomic translocation of leukemias and fusion gene products have been shown to specifically target and modify the chromatin at loci of lineage-specific regulators of hematopoiesis through heterochromatinization as well as decondensation by off-target acetyltranferase activity [52].

2.2 Long Noncoding RNAs

2.2.1 Introduction

Recently, the influence of the noncoding genome on gene expression, development, phenotypic variation and disease susceptibility has been widely accepted and continues to be increasingly understood as a paramount genomic regulatory constituent [53, 54, 55, 56]. One such area of noncoding gene regulation that has spurred much interest is in the function of long noncoding RNAs, an transcribed constituent of the noncoding portion of the human genome, which in and of itself comprises 98% of a person's DNA [57]. As of the most recently published release, there are 15,779 long noncoding RNA (lncRNA) gene annotations recognized by the GENCODE Project on the basis of being a noncoding transcript greater than 200 ribonucleotides in length [58]. Furthermore, the biological significance of lncRNA functions are continually discovered in almost every organ system in humans, and their disruption has been implicated in the cellular development and several complex disorders including neurodegeneration, cardiovascular disease and cancers [59, 60, 61, 62].

2.2.2 Definition and classes of lncRNAs

The classes of lncRNA genes recognized by GENCODE comprise several distinct annotation types, which are primarily based on the characterizations of their genomic locus, orientation, relative relationship to protein coding counterparts, and presence of computationally predicted open reading frame. The process of lncRNA discovery and annotation is not simple, where a general workflow includes transcript assembly, assessment of protein-coding potential, followed by subcellular localization studies and further characterization of secondary structure [63]. The 5,501 antisense RNAs represent a subclass of lncRNAs with a locus on the opposing strand of an annotated protein coding gene. With 7,490 lncRNAs, the long interleaving noncoding RNAs (lincRNAs) represent the largest type of lncRNA. The designation of lincRNA is made for transcripts found in a conserved, intergenic regions. The lncRNA *HELLPAR* is over 200Kb in length and represents a class of large unspliced transcripts, a macro lncRNA, unto itself. GENCODE annotations also include 3 lncRNAs as simply non coding, where they are known to be in fact non-coding transcripts. The 3-prime overlapping ncRNAs are a set of 32 experimentally supported noncoding transcripts located within the 3-prime UTR of genes. There are 47 non-coding genes that are located within protein coding gene promoters and transcribed in the opposing polarity, referred to as bidirectional promoter lncRNA. 556 lncRNAs are annotated as processed transcripts, wherein they are genes that produce transcripts longer than 200 nucleotides and do not contain an open reading frame. Along the same strand as protein coding genes the sense intronic annotation represent over 899 lncRNA genes that are located exclusively within introns, while 183 sense overlapping genes contain a protein coding gene in an intron of the lncRNA. Additionally, GENCODE recognizes 1087 TEC lncRNAs, were the experimental validation for coding potential is yet to be confirmed for the transcript, however there is the presence of a poly-A tail.

2.2.3 lncRNAs as epigenetic regulators of gene expression

The quintessential examples of epigenetic gene regulation provide our current explanation for genomic imprinting and human sex chromosome dosage compensation, which are both lncRNA mediated processes. In biologically human females a process of random X inactivation must occur early during development to achieve dosage compensation. This process of nonspecific selection and silencing of an X chromosome is governed by the actions of two long noncoding transcripts in *cis*, which run antisense to one another, *Xist* and *Tsix*. The expression of the human *Xist* gene persists exclusively on the inactive X chromosome, is 17-kilobases of ribonucleotides in length and remains relegated to the sub-nuclear location from where it is derived [64]. This lncRNA was found to be conserved among the mammalian clade and was further demonstrated to be required to achieve X chromosome inactivation [65]. Subsequent investigation provided evidence that the additional monoallelic expression of the lncRNA *Tsix* by the future active X chromosome also contributes to the initiation of the process [66]. The opposing actions of the antisense lncRNAs guide chromatin modifications whereby specific sequence domains of the *Xist* mediate association and spreading across the inactive X chromosome [67]. More facets of *Xist* biology are continually being discovered, and recent findings have shown that the lncRNA binds 81 different proteins throughout distinct phases of X-chromosome inactivation, and the formation of a ribonucleoprotein complex with *HnrnpK* is required to recruit Polycomb proteins for chromatin silencing [68].

For a subset of genes, a process of selective parent of origin expression, or imprinting, results from the actions of lncRNAs that contribute to silencing of one of the alleles. The lncRNA that functions to produce the effect is generally adjacent to a cluster of known imprinted genes within the linear spatial domain [69]. In humans and mice, the domain *Kcnq1* Imprinting Control Region (*Kcnq1* ICR) comprises four protein coding genes and one antisense lncRNA, *Kcnq1ot1*. The imprinted effect of genes within the *Kcnq1* ICR is controlled by the paternal expression of *Kcnq1ot1*, which drives the H3K9me3 mark of heterochromatinization through an interaction with the neighboring chromatin and occlusion to the perinucleolar region where repressive PcG complexes are more abundantly concentrated [70, 71, 72]. Subsequent deletion of specific *Kcnq1ot1* domains led to loss of the lncRNA function and biallelic expression was observed [71].

From these examples, we can see a clear relationship exists whereby normal allele-specific expression events may be the direct results of epigenetic gene regulation by lncRNA expression and association with chromatin. Disruptions of such events were shown by experimentally induced knockouts of lncRNA *Kcnq1ot1* domains and requirement of *Xist* to achieve chromosomal silencing, which demonstrate the varied functional significance of lncRNAs in gene regulation. Functional lncRNAs therefore play a critical role in normal human physiology and development, although the specific action that they perform is unknown for the vast majority of annotations available. Nuclear segregation of separate alleles is shown to be a process that underlies monoallelic expression and in some cases is the direct result of lncRNA expression and function. These findings suggest that nuclear partitioning is a process at least in part regulated by lncRNAs, and underscores the importance of spatial relationships within the nucleus and gene expression as well as the need to unravel genomic spatial complexity in human physiology and pathology [73]. More work is therefore necessary to identify these relationships between lncRNAs and protein coding genes within nuclear territories that drive cellular phenotypes. Further work is immediately needed to explore how structural variation may disrupt the regulatory roles of lncRNAs in gene expression for normal and disease cases. The significance of such studies is evident in the possibility to detect novel therapeutic biochemical pathways with greater specificity as to the underlying epigenetic mechanisms.

Long noncoding RNAs are also capable of direct recruitment of histone modifying proteins to facilitate regulatory targeting of homeobox (HOX) genes. The highly conserved HOX genes are responsible for development of an organisms anterior-posterior axes that lead to cell-fate specification and are subject to tight coordination of expression. Rinn et al. were the first to conclusively demonstrate such a regulatory capacity for the lncRNA *HOTAIR* located in the HOXC gene cluster, which acts in *trans* to silence the expression of the HOXD gene cluster [28]. The *HOTAIR* lncRNA is required for the recruitment of Polycomb repressive complex 2 (PCR2) to the specific targeting of the HOXD loci for histone 3 lysine-27 trimethylation (H3K27me3) by influencing the *SUZ12* subunit [28, 74]. Later, research demonstrated that increased activity of *HOTAIR* in breast carcinoma was a predictor of invasiveness and metastasis, and the recruitment of *PRC2* by *HOTAIR* led to an epigentically reprogrammed embryonic chromatin state in the cancer cell [75]. Furthermore, immunoprecipitation experiments demonstrated distinct domains of the *HOTAIR* transcript were found to serve as a scaffold that links the PRC2 and LSD1 repressive chromatin complexes to modulate histone methylation of HOX genes [76]. It was also found that tissue-specific noncoding isoforms of *SRA* act as a scaffold forming ribonucleoprotein complexes of steroid nuclear receptors and function as transcriptional coactivators [77]. In doing so, the *SRA* lncRNA component of these steroid receptors stimulates a variety of physiological responses in the cell, including gene expression, proliferation and apoptosis [78]. These works were critical to frame our current understanding of lncRNA functions as regulators of gene expression, which may act in *cis* or in *trans* to guide, tether, and link chromatin modifying machinery to protein coding targets, and greatly increase the complexity of learning the processes epigenetic regulation, which requires further work be done for the majority of lncRNAs that remain poorly understood and likely have context-specific functions.

As evidenced by the GENCODE Project annotations, long noncoding RNAs are often found antisense to protein coding genes [58]. Additionally, researchers that worked with the consortium for the Functional Annotation of the Mammalian Genome (FANTOM) found that over 70% of the mammalian genome may contain sense-antisense transcriptional units that are regulated in a tissue-specific reciprocal or congruent manner [79]. In acute lymphoblastic leukemia, the tumor suppressor gene Cyclin Dependent Kinase Inhibitor 2B, *CDKN2B* was shown to be repressed when there is expression of the antisense lncRNA *ANRIL*, which contributes to heterochromatin formation [80]. Antisense transcription is thought to contribute to methylation of sense-promoters through lncRNA sequence homology and secondary structure mediated recruitment of chromatin modifying machinery [69]. Moreover, single nucleotide variants (SNVs) within the *ANRIL* locus are also known associated high-risk alleles for atherosclerosis and homozygosity of which leads to reduced expression of the antisense lncRNA and disrupts the development of arterial smooth muscle cells [81]. Put another way, disrupted lncRNA activity may increase the possibility for irregular vascularization, for which *Tie-1*, a key tyrosine kinase involved in angiogenesis and proper endothelial cell development, was found to be under direct antagonistic regulation by mRNA-lncRNA hybridization with its antisense-lncRNA counterpart [82]

The mechanism by which these SNVs within the *ANRIL* locus led to a decreased abundance of *CDKN2B* was later confirmed by expression analysis followed by a variant of chromatin conformation capture (3C). This study examined lymphoblastoid cell lines and demonstrated an inhibition of *ANRIL* when the cognate transcription factor *STAT1* was bound to the region, and an inability for *STAT1* to bind for cells harboring high risk alleles because of altered chromatin contacts, which led to reduced DNA accessibility [83]. In some cases of the blood disorder α-Thalassemia, a structural deletion was shown to cause abnormal positioning of a noncoding variant of the *LUC7L* gene, and the antisense transcription of which leads to epigenetic silencing of the sense globin gene *HBA2*, and therefore reduced bioavailability of alpha globin

[84].

Furthermore, concordant expression of lncRNAs with pathway-specific mRNA has been implicated in a wide array of human biological mechanisms relevant to complex diseases. As a model of type 2 diabetes mellitus, mice fed a diet rich in fats resulted in elevated expression of the lncRNA *MEG3* through acetylation of the locus, which led to *FOXO1* upregulation causing hyperglycemia and insulin resistance in primary hepatocytes [85]. Additionally, a pleiotropic role of *MEG3* was discovered by neuroscientists that studied expression subsequent to cerebral ischemia. Following a stroke event, activity of the lncRNA was greatly increased in neurons and specifically bind to the *p53* protein DNA binding domain, stimulating cellular apoptosis [86]. A focus on lncRNAs enriched in human cancers found that upregulated *MEG3* selectively targets and binds *TGF-β* genes through computationally predicted RNA-DNA triplexes via GA-rich repeats in the lncRNA [87]. The formation of RNA-DNA triple helices has been recognized for quite some time, and the computational prediction of RNA-DNA triplexes may be performed using the Triplexator tool, which performs sequence-based hydrogen bonding analysis of the RNA molecule for enriched features that contribute to chromatin association for DNA targets and can be used to predict lncRNA regulation [88, 89, 90]. Notably, a Nobel Prize was awarded to Blackburn, Greider and Szostak in 2009 for the compendium of their research towards our understanding of chromosomal telomeres and the function of the ribonucleoprotein telomerase [91]. Their collective works in part determined that a catalytic lncRNA *TERRA* was a required element for the telomerase enzyme to synthesize appropriate DNA sequences at the chromosome terminals by containing the necessary nucleotide template [92]. *TERRA* was later demonstrated to be localized at telomeric chromatin through G-quadruplex formation, a DNA-RNA hybridized structure composed of looped, hydrogen bonded guanine tetrads, to provide a protective effect and recruitment of enzymes for telomere maintenance [93, 94, 95].

Long noncoding RNAs may play an important, yet incompletely understood, role in a multitude of pathologies of unregulated cellular growth such as cancer [45]. Found in the promoters of genes, so called promoter-lncRNA expression was found to be tightly co-regulated with the associated protein coding genes and may be altered in tumor cells. In one case, the promoter-lncRNA *PANDA* was shown to be induced by *p53* as part of cellular response pathway to exogenous DNA damage to actively binds and sequester *NF-YA*, a transcription factor responsible for activation of apoptotic genes [96]. Additionally, research has demonstrated promoter-lncRNA regulation of a key enzyme that catalyzes folate metabolism as well as precursor DNA synthesis [97]. Furthermore, glucocorticoid-mediated cellular growth arrest is modulated in part by the action of the *Gas5* lncRNA by serving as a glucocorticoid response element decoy sequence, effectively titrating the glucocorticoid receptor away from its DNA target sequences. Thus, lncRNAs may play a systematic role to actively inhibit cellular death and drive key metabolic reactions, whereby the coordinated expression of lncRNAs with cell-cycle control genes suggests yet undiscovered instances of their disruption among cancers [98]. Indeed, an integrated study of RNA expression, chromatin state and chromatin contacts in T-cell acute lymphoblastic leukemia found the oncogenic proliferative signaling of *NOTCH1* specifically targets the lncRNA *LUNAR1* with downstream effect of stimulation of expression at the insulin-like growth factor 1 receptor locus, thereby providing a clear example of a lncRNA-mediated process required for tumor growth and sustainability [99]. One key aspect that this study lacks is an assessment of the variability of the lncRNAs among samples due to genomic structural variation, which may modulate the process among samples of a specific tumor type.

These findings elucidate one important facet of lncRNA biology as a regulator of genomic stability, which when abrogated could lead to the development of cancer, and lend credence to take a closer look at other lncRNAs that act in physiological pathways. The binding capacity of lncRNAs for important enzymes and implicated roles in a wide variety of cellular functions demonstrate their biological significance, which necessitates further investigation when impacted by large structural variation and their disruption during complex genetic disorders. The literature on these topics remains minimal compared to the quantity of lncRNAs discovered and the availability of newer sequencing technologies to capture high-order relationships is ever expanding making it now required that the field establish computational methods for their integration to capture higher-order relationships of lncRNAs in genomics [91].

The binding capacity of lncRNA for other RNA molecules is also a method of action by which lncRNA molecules may contribute to the regulation of gene expression. Researchers have shown that in gall bladder carcinoma cell lines, the lncRNA *TUG1* sequesters the micro RNA (miRNA) *miR-300* by acting as a miRNA sponge, effectively reducing the negative gene regulation of the miRNA and thereby indirectly increasing the availability of the miRNA target mRNA, which promotes epithelial-mesenchymal transition [100]. *TUG1* has also been implicated in the pathology of ischemic stoke and neuronal cell death by acting as a miRNA sponge of *miR-9*, a regulator of pro-apoptosis gene mRNA [101].

2.2.4 Databases for Curated lncRNA Annotations

There are currently several databases that curate information on lncRNAs that may be used for experimental and bioinformatics analyses. Table 2.1 describes the available information on lncRNAs to use in these types of experiments. lncRNAtor provides a comprehensive overview of over 34,000 human and model organism lncRNAs and includes experimentally validated interactions with proteins and expression profiles from TCGA [102]. The LNCipedia database of 111,685 human lncRNAs with computationally predictions of RNA-RNA interactions [103]. Similarly, the 113,513 lncRNAs collected in the lncRNome databank have computational predictions of their potential to interact with proteins [104]. Finally, the lncRNAdb offers detailed, man-

Table 2.1: Long noncoding RNA Databases

Information	LNCipedia	lncRNome	lncRNAdb	lncRNAtor
lncRNA Count	111,685	113,513	287	34,605
Genomic Loci	Y	Y	Y	Y
Sequences	Y	Y	Y	N
Secondary Structure	Y	Y	N	N
Expression Profile	N	N	Y	Y
RNA-protein Interactions	N	Y (predicted)	Y (literature)	Y (experimental)
RNA-RNA Interactions	Y (predicted)	N	Y (literature)	N

ually curated information on 287 human lncRNAs with profiles from the human expression atlas, and interaction data from published literature [105].

2.2.5 Computational Tools for lncRNA Identification and Functional Prediction

The activity of known long noncoding RNAs can be identified by their expression levels from lncRNA-specific probes on Affymetrix arrays [106]. In order to identify known or novel lncRNAs from an RNAseq experiment, transcriptome assembly must be performed using a computational tool such as Cufflinks [107]. If the goal is to determine whether a new transcript is a lncRNA, one must then assess the coding potential of the RNA using either ribosomal profiling experiments or the comparative genomics method implemented by the PhyloCSF software [108]. Once low coding potential is determined, the functional assessment begins with computational prediction of the secondary structure, which give lncRNAs functional domains and binding sequences to interact with RNA or proteins [109]. Several methods have been developed for secondary structure prediction, which can be a computationally expensive task due to the inherent large size of lncRNAs. One such tool, lncRNAScan-SVM implements a support vector machine, a type of machine learning based method, in order to predict the secondary structure, while CROSSalign is capable of nucleotide-level resolution of lncRNA structure and implements a dynamic time warping method to align variable lengths of sequences [110, 111]. This information can glean valuable insight into subsequent prediction with other RNAs, DNA sequences, and proteins in order to characterize the function of lncRNAs, which can partially be performed using a tool such as lncTar [112].

2.3 Architectural Features of Human Chromatin

2.3.1 Introduction

Within the nucleus of a eukaryotic cell is a highly condensed form of DNA, which for diploid humans amounts to nearly two meters of DNA confined in a 125 μm^3 volume [113]. In order to achieve this level of compaction, the DNA is wrapped around histone proteins with regular intervals to form nucleosomes which resemble a beads on a string. This building block of chromatin can be regionally condensed by the antagonistic actions of methylation or acetylation of the histone proteins leading to the formation of heterochromatin and euchromatin. Furthermore, megabases of nucleosomes may form a helical structure and coil round itself in a process of supercoiling, leading to ultimate chromosome structure [114]. Here, we will provide an overview of the genomic architecture of human chromatin and present recent findings toward our current understanding by which this topology contributes the regulation of gene expression.

2.3.2 Overview of chromatin structure at multiple scales

The content of the human genome is a multilevel, highly organized DNA-protein structure that enables enormous compaction of billions of DNA molecules into the microscopic physical dimension of the nucleus within the cell. The double stranded DNA helix first associates with a histone-3:histone-4 tetramer (H3-H4), followed by two histone 2A and histone 2B (H2A-H2B) dimers to constitute the DNA nucleosome octamer core particle, which is the primary building block of chromatin [115]. This process is propagated across the entire chromosome, which further self-associate to create chromatin nucleofilaments that are comprised of closely linked nucleosomes. At higher levels of genome assembly, the nucleofilaments form supercoiled chromatin which in turns forms radial helices leading to the high density of whole chromosomes within the nucleus.

Recent advances in biophysical technologies have made it possible to characterize the organization of DNA with greater clarity and detail, and have led to a more robust understanding of genome spatial relationships and regulation. Chromatin conformation capture technology was the first kind of such methods and let to accurate quantification of chromatin contacts frequencies for specific regions of DNA [116]. The spatial organization of the genome determined from this method was found using a formaldehyde cross-linkage of short, digested DNA fragments followed by quantitative PCR to identify chromatin that occupies proximal space in the nucleus [116]. A wide variety of 3C derivatives have since been proposed with wide and narrow focal ranges, and Leiberman-Aiden et al. leveraged high-throughput sequencing in their method, denoted Hi-C, which enables one to assess the chromatin dynamics and long-range DNA interactions at the genome-wide level [117]. This study calculated intrachromosomal contact frequencies and determined that chromatin had a higher propensity to come in physical contact with certain regions within a compartment and less frequently with loci outside the a compartment boundary. Moreover, they found that the genome could be stratified by contact frequencies and inferred activity levels into so-called A and B compartments, for open and closed chromatin state, respectively [117]. Furthermore, Hi-C analyses confirmed a long-held concept from the nineteenth century of chromosomal partitioning, and was found to be congruent with florescence microscopy experiments that showed the nuclear DNA is organized into distinct chromosome territories during interphase of the cell cycle [118].

With a method to probe the organization of eukaryotic genomes, subsequent investigations have led to the discovery that stability of the organization is inversely related to the reproducibility of chromosomal interactions. Genomic organization may be described by the differential dynamics model of chromosomal connections. The model describes multiple scales of stability and proposes that high-level chromosomal territories and compartmentalization remains stable during a single cell cycle, yet are unlikely to be maintained into future cycles, whereas lower-level chromatin loops and topologically associated domains are highly dynamic through a cell cycle but highly reproducible in future cycles [119]. Topologically associated domains (TADs) are highly conserved organizational features of high contact frequency, linear stretches of DNA generally less than one megabase in length, are comprised of smaller chromatin loops and are therefore thought to be the basic structure of genome organization and regulatory modules of the genes within them [120]. This level of genome organization is stable through evolutionary time, where they have been shown through comparative Hi-C to be highly conserved through syntenic regions, bolstering their significance in gene expression and development [121]. Indeed, genome organization is tightly linked to regulation of gene expression and cellular differentiation, where dynamic reorganization of chromatin architecture takes place between embryonic stem cells and all levels of development across tissues [122].

In terms of epigenetic gene regulation, the topologically associated domain may contain within it several chromatin loops and therefore TADs may then be a consequence of their formation. The loop structure is a layer of chromatin organization where one or two genes and epigenetic elements such as enhancers may be confined within adjoined, linearly distant CTCF protein bound DNA motifs that are maintained by a cohesin complex in a process called loop extrusion [123, 124]. Furthermore, the orientation of CTCF binding motifs must have a convergent orientation in order for loop formation to occur normally, suggesting a natural mechanism for disruption of chromatin looping through mutation or structural variation, albeit not all genes are found within insulated neighborhoods [125]. These architectural proteins maintain accessibility of DNA and form self-contained epigenetic regulatory modules of genes called insulated neighborhoods, which have been shown can be eliminated with engineered loss of cohesin, and they are known to drive aberrant gene expression in disease when the structure is disrupted [126, 127, 128].

2.3.3 The Three-Dimensional Genome and Human Disease

With the regulatory role of enhancer-promoter interactions mediated by chromatin architecture being widely established, there is the need to learn more about how disruptions in chromatin structure can contribute to disease etiology. There are now several diseases that have been characterized using conformation capture technologies that have demonstrated disruptions in topology leading to epigenetic rearrangements and altered gene expression. Deletion of TAD boundaries could lead to the merging of tow adjacent TADs and irregular enhancer-promoter contacts, while rearrangements may lead to neo-TAD formation or dissociation of the gene with an enhancer [129]. In some cases of familial acute myelogenous leukemia, a translocation on chromosome 3 creates exactly such a situation where the enhancer located in the same TAD as the normally active *GATA2* gene is repositioned in a neo-TAD with the *EVI1* gene, increasing the myeloid cell stem-ness and driving oncogenesis [130]. T-Cell acute lymphoblastic leukemia was found to be enriched for microdeletions at insulated neighborhood boundaries that contain common oncogenes, while mutagenic alteration at these sites was sufficient to promote tumorigenesis in nonmalignant cells [126].

Multiple myeloma is a type of cancer of circulating plasma cells which exhibits aneuploidy, or partial to whole chromosome duplication with high copy number variation. In a recent integrative analysis of copy number variants (CNV), GM12878 ChIP-seq data, gene expression, and Hi-C data, it was found that hematopoietic lineage-specific cytokine signaling pathways were upregulated and driving immune system evasion due to altered TAD boundaries by CNVs [131]. This work has provided a clearer interpretation in the way aneuploidy may be further characterized by architectural changes in the chromatin and directly influence the activity of oncogenic gene expression. In another study of myeloma, a recurrent translocation found in 20% of patients studied led to TAD disruption and the juxtaposition of important lineage-specific enhancers of immunoglobulin genes, dubbed super-enhancers, near a known proto-oncogene *MYC* [132].

The significance of chromatin architectural rearrangements extends to developmental and demyelinating disorders, which may also result from TAD-shuffling, neo-TAD formation, as well as TAD fusions [133]. These types of structural changes represent a rearrangement due to translocations, a new domain formation, and a loss of a boundary between adjacent, non-interacting TADs, respectively. The shortened digits characteristic of Cooks syndrome result from a duplication intersecting a TAD boundary, which leads to a new TAD containing the *KCNJ2* and *KCNJ16* genes and the *Sox9* enhancer-promoter interaction within the isolated domain [134]. As a result of a TAD boundary deletion, a fused TAD causes multiple enhancers normally active in the forebrain to be hijacked by the *LMNB1*, causing its overexpression and concomitant downregulation of myelin sheath proteins sensitive to *LMNB1*, and the progressive phenotype in adult-onset demyelinating leukodystrophy [135].

With the amount of available chromatin conformation currently data and continually being produced, we are now at the verge of discovering a more complete understanding of how genomic structural variation contribute to human diseases. By contextualizing genomic rearrangements and variants in terms of their impact on the three-dimensional architecture of chromosomes, a clearer understanding may be drawn wherein these variants may alter the spatial regulatory modules of genes and promote the aberrant expression patterns that are characteristic of specific diseases. It may also show how subtype-specific transcriptional patterns develop and contribute to varied responses to treatment protocols.

2.4 Structural Variation in the Human Genome

2.4.1 Introduction

Since the publication of the human genome draft sequence, it readily became apparent there existed a great deal of variation, beginning with a focus on single nucleotide polymorphisms [136]. With Chromosomal structural variation (SV) is thought to be a naturally occurring phenomenon that contributes to the course of human evolution and diversification, and in some cases may lead to disease[4]. Structural variation is any change to the chromatin with arbitrary cut-off being larger than 50 basepairs, and make a contribution of roughly 18 megabases difference between two human genomes, more than any other type of variant [1]. Furthermore, the functional impact of SVs on the expression of genes has been shown to be much greater than that of single nucleotide variants (SNVs) [1, 137]. Due to their apparent importance in genomic biology, it is no surprise that a large amount of research has been devoted to their detection, how they originate, and their impact in natural variation and disease.

2.4.2 A History of Structural Variant Detection and Analysis

Multiple efforts have been make to tackle the difficult challenge of SV detection and functional analysis, although a gold standard still remains to be developed as a reference [6]. Kidd et al. developed a resource from 17 human genomes and resolved the breakpoints of 1,054 SVs in order to determine the mechanism behind their formation, which was primarily limited by the inability of short-read sequencing technology to capture all of the SV junctions [138]. Researchers working with the 1000 Genomes Project made an effort to catalogue unbalanced genomic structural variants, or copy number variants (CNVs), from population based sequencing and found over 28,000 deletions, insertion and duplications across 185 human genomes, with an average of about 1000 per genome [2, 3]. The mutational mechanisms that generate SVs were then inferred by the breakpoints, and are described by four different mechanisms. These mechanisms are nonhomologous end-joining (NHEJ), nonallelic homologous recombination (NAHR), L1-mediated retrotransposition, and fork stalling and template switching (FoSTeS) [7]. NHEJ occurs due to a failure in a DNA repair mechanism, NAHR is the result of a crossing over event between genomic regions that share sequence homology, while FoSTeS occurs due to an error in replication and L1 retrotransposition denotes the activity of a biologically active retrotransposon [139]. Recent work by the Human Genome Structural Variation Consortium, a subgroup of the 1000 Genomes Project, has created a haplotype-resolved map of structural variation in Han Chinese, Yoruban, and Puerto Rican trios [140]. This work has taken advantage of multiple sequencing platforms and technologies, as well as a numerous SV-detection algorithms in order to resolve these variants in great detail, and has created a rich dataset to explore the functional consequence of SVs in healthy individuals. This orthogonal approach to structural variant detection has provided the most comprehensive assessment to date, making use of both long-read and short-read sequencing technology to resolve SVs with nucleotide resolution.

From these studies and more, we now fully appreciate the complexity of genomic structural variation, which may come in the form of intrachromosomal variants and interchromosomal translocations. Intrachromosomal variants occur along one chromosome in coding and noncoding regions, and encompass deletions, duplications, insertions, as well as inversions [4]. A translocation may be balanced or unbalanced, wherein either both chromosomes or only one may exchange portions of their genetic material and join through NHEJ, respectively. These translocations may result in the abnormal concatenation of genetic sequences leading to fusion genes and fusion gene protein products. Additionally, it has been noted that the repositioned DNA induces transcriptional changes in the genes along exchanged chromosomes, likely due to nuclear repositioning and changes in the global architecture of the chromatin [141].

The functional impact of SVs in the human genome is varied across a severity spectrum, and SVs have been implicated in a wide variety of disorders in every tissue. Research has shown that CNVs may play a critical role in several neurocognitive disorders, autism spectrum disorder and cognitive delay [139]. On the other side of the spectrum, homozygous deletions of whole genes are known to naturally occur from population based studies, indicating their relative dispensability based upon their preponderance in normal variation in healthy samples [1]. One of the most destructive types of structural variation, chromothripsis, occurs in cancer genomes and is a catastrophic shattering of whole chromosomes [142]. As aforementioned, the 1000 Genomes Project and others have noted that structural variants contribute more to genetic difference than any other type of variant in humans, with the cumulative nucleotide content affected representing roughly 0.1% of the genome [3, 2, 139].

2.5 Acute Lymphoblastic Leukemia

2.5.1 Introduction

B-cell acute lymphoblastic leukemia is one of the most common diagnoses for blood disorders among afflicted pediatric patients. In the precursor form, a cancer manifests itself as a circulating population of cancerous innate-immune cells resulting from clonal expansion of a B-lineage lymphoblasts. The clinical morphological characteristics used at diagnosis include the small size of a relatively homogeneous pool of lymphoblasts that exhibit elevated nucleocytoplasmic ratios and a visibly condensed chromatin state [143].

Pre-B ALL represents a major success story in the treatment of cancer, with an approximate median five-year survival rate among children diagnosed of over 80% [144, 145]. The overall incidence of ALL, according to the National Cancer Institute's Surveillance, Epidemiology, and End Results Program (SEER), is about 1.7 per 100,000 individuals for the U.S. population. The median age of onset is 15 with a bimodal distribution around two and sixty years old, indicative that the 84,000 people afflicted each year are more frequently children [146]. The exact epidemiological root of the leukemia remains poorly understood, as predisposition syndromes account for less than 5% of all reported diagnoses and no conclusively demonstrated environmental cause or carcinogenic factor, while the greatest percentage of childhood ALL cases are affected committed precursor B-cell hematopoietic lineages [147]. A reflection on the current state of acute lymphoblastic leukemia treatment, which for many may include a regiment of antileukemic agents like glucocorticoids, asparaginase, tyrosine kinase inhibitors, chemotherapy or allogeneic stem cell transplantation, may lead one to the conclusion that prognostic factors may direct clinical treatment [44]. However, the post-genomic era has produced several insights of commonly occurring translocations and genomic rearrangements now recognized by the World Health Organization that patients may be classified by and serve as the basis of treatment protocols [148]. Treatment for the 20% of patients who do not respond well to current medical care may therefore require more translocation-group specific research to identify possible regulatory pathways amenable to targeted therapies.

2.5.2 Characterization of cytogenetic subtypes

There are several subtypes of acute lymphoblastic leukemia that may be defined based on recurrent translocations and large rearrangements with shared loci. While these aberrations are often drivers of oncogenesis, it is important to note that they have been found to exist in samples derived from healthy umbilical blood at much higher frequency than the reported oncogenic incidence, suggesting they may not be sufficient and a second genetic lesion is likely necessary for leukemogenesis to occur [149]. A genome-wide analysis of pediatric B-lymphocyte ALL patients found that there is often an additional gain or loss at genes known to regulate B-lineage development that act as cooperating mutations with recurrent translocations to promote leukemia [10]. Further work demonstrated somatic alterations were enriched in genes that contribute to B-cell development, *Janus* kinases, the *tp53* tumor suppressor pathway, as well as other cellular growth signaling cascades [150]. Additionally, a study of DNA methylation found consistent epigenetic silencing via hypermethylation at promoters of these genes, however this study did not make use of immunoprecipitation data to investigate the effects on chromatin state due to somatic mutations [151, 152]. From these studies we can glean that B-cell specific areas of the genome contribute to the pathogenic state of leukemia, however what lacks is a look into the noncoding genome that may be disrupted, and how the chromatin topology is affected by structural variation with respect to shared translocation groups. Furthermore, studies of HOX genes among leukemias showed they contribute to oncogenesis through disturbed hematopoiesis, yet are not sufficient to discriminate between subgroups based on combined expression alone [153, 154]. These expression based studies may be limited by this approach, while integration with epigenetic data types could improve the discriminatory capacity based on capturing subtype-specific regulatory relationships between the coding and noncoding genome.

Hematological cancers represent a class of disorders affecting cells of the blood and lymphatic system [155]. The specific diagnosis and classification is first based upon the cell lineage from which the hematopoietic stem cell has committed, and may further be described by the maturation state of the cancerous cell [156]. Acute varieties of leukemia denote that premature myelogenous or lymphoblast cells rapidly proliferate and are released into circulation [157]. In contrast, patients suffering from chronic leukemia often involve late-stage, mature blood cells, which reflects the generally poorer prognosis of the acute variety [158]. Acute lymphoblastic leukemia may then be categorized using a variety of commonly occurring genomic rearrangements and ploidy levels [159]. The focus of this study will be B-cell Philadelphia chromosome-like acute lymphoblastic leukemia, a rare form of hematological cancer specifically involving the immature antigen-presenting cells of the adaptive immune system [160]. The name itself comes from the likeness in genetic expression to the well-characterized translocation event between the long arms of chromosome 9 and 22, first described by Peter Nowell at the University of Pennsylvania in 1960 [161, 162].

2.5.3 Recurrent Chromosomal Rearrangements and Chromatin Topology in Hematological Cancers

Recent integrative analyses of the three dimensional chromatin contacts in nuclear space has led to the discovery of mechanisms of oncogenesis by which genomic rearrangements can reposition promoter/enhancer regulatory elements near oncogenes leading to aberrant expression. With high throughput sequencing, chromatin immunoprecipitation assays and advanced chromatin conformation capture technologies, one may obtain the sequence, determine the chromatin state, identify exact loci of mutations ranging from single nucleotides to large structural variants, and visualize this information with respect to the 3-dimensional compaction of nuclear DNA to identify networks of regulatory elements that act upon genes within proximal nuclear space and identify disturbances in cancer. Cell-line chromatin topology has been shown to be useful as a reference to study sample-level copy number variation (CNV) affects on topologically associated domains and the disrupted regulatory relationships within these regions normally in frequent physical contact with one another [34]. These authors produced a computational framework to perform a expression quantitative trait loci analysis of amplifications and deletions for all cancer samples in The Cancer Genome Atlas (*TCGA: http://cancergenome.nih.gov/*) and found examples of enhancer hijacking events common to a variety of cancers [35]. Enhancer hijacking refers to the phenomenon whereby genomic structural variation disrupts the chromatin topology and the contact frequencies between loci. When a hijacking event occurs, two possibilities may occur and lead to changes in gene expression. If enhancers are no longer able to physically associate with promoters, there may be a reduced expression of the normally regulated gene. Alternatively, a gene may be over-expressed if the altered chromatin topology introduces the possibility for increased contact frequency between a non-cognate enhancer and promoter.

Groschel et al. concluded that patients with acute myeloid leukemia (AML) that harbor a recurrent inversion along the long arm of chromosome 3 exhibit increased stemness with *EVI1* over-expression due to a repositioned strong enhancer for *GATA2*, with its concomitant haploinsufficiency [130]. This finding was important as it provides direct link between a large, recurrent structural variant that not only rearranged the epigenetic landscape when compared to ENCODE reference data, but was the causal mechanism for unmitigated growth a subgroup of patients with AML. The importance of this discovery underlies the notion of chromatin gene regulation as inherently a structurally mediated process. T-cell acute lymphoblastic leukemia provides another unique case where cancer results from structural variants in key positions that define chromatin topology. Control of tissue-specific gene expression has been suggested as a process where distal regulatory elements conform as stretches of DNA in close three-dimensional space, with boundaries defined by bound CTCF proteins and the cohesin complex, forming a structure denoted as an insulated neighborhood or loop [126]. Recurrent microdeletions in T-cell ALL intersect these insulated neighborhood boundaries, as defined by the Jurkat cell line as a reference, and where found to be a direct activation mechanism for known T-cell oncogenes [163]. Currently there is not an analogous study for acute leukemia of the B-cell lineage, and there has yet to be any work taken as to subtype-specific differences of gene expression with respect to the impact on chromatin topology.

2.5.4 Philadelphia-like Acute Lymphoblastic Leukemia

In acute lymphoblastic leukemia, a rare subtype was discovered that had poor prognostic outcomes relative to other cytogentic subtypes. Found to occur in roughly 9-15% of cases, the Ph-like form of ALL was named for it's global expression similarity to Philadelphia chromosome positive (*BCR-ABL1* fusion) cases, yet lacks the hallmark kinase fusion gene product and associated phosphorylation activity [164].

Patients lacking the *BCR-ABL1* fusion gene are known to exhibit alterations to *JAK1* and *JAK2* signaling kinases, and be associated with poor outcome due to pathway misregulation [165]. In fact, disruption of *JAK* signaling is a found in a numerous forms of hematological cancers due to their pivotal roles in cytokine receptor signaling [166]. Recently, some evidence exists that support the idea that *JAK-STAT* signaling pathway is activated via various genomic mutations observed in Ph-like patients and researchers have called for greater understanding of this activity and it's use as a actionable target for treatment [167]. Furthermore, the diagnosis of Ph-like leukemia has been a challenging undertaking due to the lack of a characteristic recurrent mutation. Thus, more research is needed to better understand this pathway activation in the absence of a singular aberration, and it is likely to be a combination of both mutational change and the disruption of key regulatory elements.

CHAPTER 3: FUNCTIONAL ANALYSIS OF HAPLOTYPE-RESOLVED STRUCTURAL VARIANTS AMONG HEALTHY HUMAN TRIOS

3.1 Introduction

In recent years, large-scale population sequencing efforts such as the 1000 Genomes Project have undoubtedly bestowed a surfeit of information from which to study the functional effects of various mutations, an undertaking with an initial focus on single nucleotide polymorphisms (SNPs) and has since expanded to provide a worldwide reference set of human genomic variation [168, 169, 170]. As a subsidiary of the 1000 Genomes Project, the Human Genome Structural Variation Consortium (HGSVC) took the helm to produce an analogous integrated set of structural variation on the same 2,504 samples from 26 global populations [1]. The continued work of the HGSVC research group has focused on creating a high resolution set of structural variation through the use of a multitude of different technologies including long and short reads, as well as jumping libraries from Pacbio, Illumina, 10X and others. Contemporary sequencers have unique sets of strengths and weaknesses that when employed together allow for a combined analysis of structural variants in a manner that is amenable to multiple strategies for detection such as read depth, split read and de novo assembly. With this strategy in mind, the HGSVC working group aimed to produce a set of refined map of structural variation with advanced methods based on the integration of multiple sequencing methods and computational SV detection algorithms.

3.2 Methods

3.2.1 Significance of SVs that Engulf Protein Coding Genes

Initially, a preliminary assessment of the functional consequence of SVs on coding genes was performed using a permutation enrichment analysis with the results shown in Figure 3.2. Genomic regions were randomly permuted for the length of genes found to be engulfed by duplications and inversions. Put another way, these genes were hypothesized to be completely overlapped and thus affected spatially by SVs by either complete duplication of their sequence of repositioned in an opposing orientation. The intersections and permutations were performed using BEDtools and analysis of expression of these genes was conducted on RPKM normalized RNAseq counts.

3.2.2 Allele Specific Expression of Protein Coding Genes

In order to do determine whether structural variants impacted allele-specific expression in healthy individuals, we developed an SV ASE analysis pipeline with the following three steps for PB SVs and IL SVs, respectively, and shown in Figure 3.1. First, we established a set of candidate SVs gene pairs by taking the intersection of heterozygous SVs (het SVs) with previously reported SNP ASE genes. Second, phased RNAseq reads were filtered following criteria whereby reads will have an NM less than or equal to six, a base quality greater than or equal to ten, a mapping quality score above twenty, and total read counts above eight. Read counts of the genes were calculated for each sample's two haplotypes using BEDtools multicov [171]. Third, the significance of SV genes pairs was then obtained by applying a binomial test to the read counts of the two haplotypes with multitest correction using FDR 5%.

3.2.3 Identification of Topologically Associated Domains

The task of annotating domains from Hi-C data was performed by calculating the directionality index (DI) values for windows across contact matrix bins, the TAD calling algorithm proposed by Dixon et al. [120]. Reference sets of cell-line domains

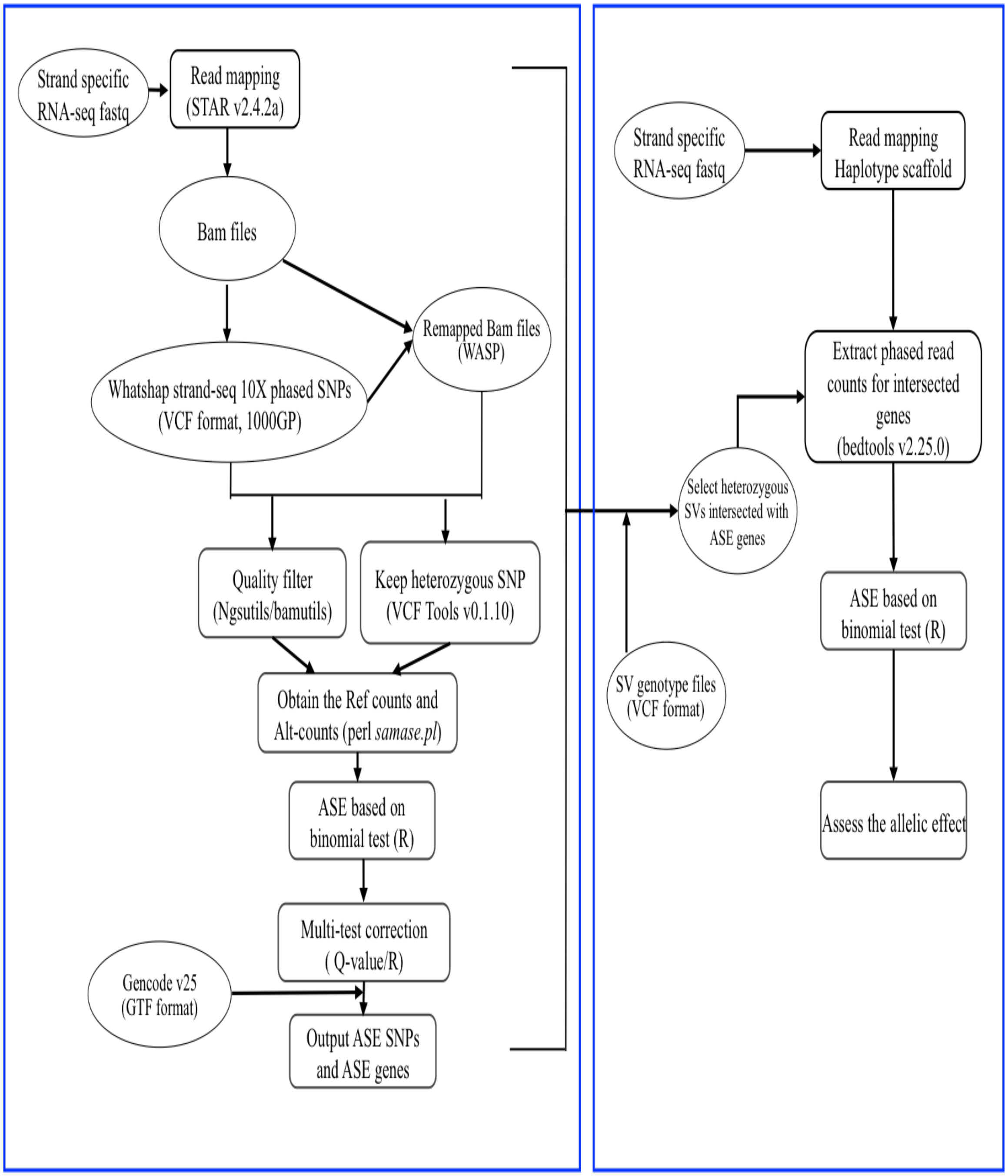
Strand specific RNA-seq fastq
Read mapping (STAR v2.4.2a)
Bam files
Remapped Bam files (WASP)
Whatshap strand-seq 10X phased SNPs (VCF format, 1000GP)
Quality filter (Ngsutils/bamutils)
Keep heterozygous SNP (VCF Tools v0.1.10)
Obtain the Ref counts and Alt-counts (perl samase.pl)
ASE based on binomial test (R)
Multi-test correction (Q-value/R)
Gencode v25 (GTF format)
Output ASE SNPs and ASE genes
Strand specific RNA-seq fastq
Read mapping Haplotype scaffold
Extract phased read counts for intersected genes (bedtools v2.25.0)
Select heterozygous SVs intersected with ASE genes
SV genotype files (VCF format)
ASE based on binomial test (R)
Assess the allelic effect

Figure 3.1: Workflow to detect allele specific expression from SVs.

were obtained from Feng Yue's 3D-Genome Browser to serve as a comparative and contrasting sets of domains, from which we can determine the changes to chromosomal architecture associated with the genomic structural variants. For the unphased Hi-C datasets, the bin size is simply the resolution of the quantile-normalized contact matrices, which was 40kb for all nine Hi-C experiments of each sample. A window size of $\pm 2MB$ was used which was identical to that used in the reference Hi-C calculations and corresponds to the binning resolution from which to ascertain directional bias. The directionality index provides a quantitative measure of the contact bias for a given bin with bins upstream and downstream within a given window, and has been widely utilized in order to detect topologically associated chromatin domains. The exact formula for the directionality index of any given bin is given in the supplemental of the Dixon et al. paper as equation 1, where $DI = \frac{(R_d - R_u)}{|(R_d - R_u)|} * [\frac{(R_u - \mu)^2}{\mu} + \frac{(R_d - \mu)^2}{\mu}]$. To calculate the degree of upstream or downstream bias for a given bin we need to find three values for our DI formula. We need the number of reads within the window mapped upstream, R_u, and downstream, R_d, and the average or null expectation number of reads within the $\pm 2MB$, given as μ. The DI quantifications can then be used as input observations to a hidden Markov model to calculate the posterior probabilities of a given chromosomal region being a TAD. The TAD calling workflow is the same software utilized by 3D-Genome Browser Hi-C reference dataset domain calls and may be downloaded from the Ren lab website at UCSD [172].

3.2.4 Identification of TADs from Haplotype-Resolved Chromatin Contact Maps

The same dilution Hi-C dataset as mentioned previously was statistically phased to each sample haplotype in order to detect allele-specificity in chromatin contacts and boundaries. The aforementioned procedure of calculating the directionality index was performed in order so as to detect upstream vs. downstream contact biases. Due to the process of phasing of the reads, roughly 10% of the Hi-C sequencing coverage remains, and thus we chose to follow a common practice in which we decreased the resolution to 100kb with the same window size of $\pm 2MB$ in order to detect more long-range interactions from shallower sequencing depth [173, 174].

3.2.5 Quantification of Chromatin Topology Dissimilarity

The determination of chromatin topology across multiple samples naturally leads to the question of how to quantify the degree of similarity across samples. For this study, we began with a simple determination of overlapping TAD boundaries. The degree of sample-TAD overlap was calculated first using a reciprocal overlap cutoffs of 50% and 75% and 100% using BEDtools [171]. In addition to simple pairwise tests of shared chromatin interaction regions, the pairwise Jaccard distances between samples were calculated for each set of TADs on a per-chromosome basis. The Jaccard distance is a measurement which represents the degree of dissimilarity between two sets after taking the intersection over the union and can be given by the following formula, $d_J(D_1, D_2) = 1 - J(D_1, D_2) = \frac{|D_1 \cup D_2| - |D_1 \cap D_2|}{|D_1 \cup D_2|}$, for two sets of TAD coordinates given as D_1 and D_2 [175, 176]. Recently, a bipartite matching method has been proposed to provide a refined quantification for the distance or dissimilarities between particular sets of chromosomal segments, called the BP score [177]. The BP score will complement traditional distance metrics and allow for the discernment of noise from true signal in the calculation of the DI, and accounts for the size of the changes to chromatin. In addition, the bipartite matching method will allow for a more precise look at specific loci with varying TAD structure and SVs. The BP score method treats two sets of TADs (i.e. reference vs. sample, sample vs. sample) as a bipartite graph where TADs from one set nodes of the first component and TADs in the second set are nodes of the second component. The algorithm allows for each node to have multiple edges and makes connections between nodes from the first component with a node from the second component, if and only if, they have a non-empty intersection. The outcome of the bipartite graph describes how similar chromosomal TAD partitioning is between Hi-C experiments and corresponds to a BP score of TAD set dissimilarity for each chromosome.

3.2.6 Mapping Structural Variation on Local Chromatin Topology

The integrated Illumina callset of structural variants was first filtered to identify only unique SVs that had a heterozygous genotype. These were then mapped to regions within TAD structures using BEDtools intersect where the SV was completely within the bounds of the TAD [171].

3.2.7 Identification of Enhancers and Target Genes Within Chromatin Domains

The identification of epigenetic regulators of genes is a difficult task that often relies upon evidence from enhancer-RNA coexpression or expression quantitative trait loci (eQTL) studies powered by large sample sizes. In order to assess how gene regulation may be modulated by structural variants within domain boundaries, enhancer and target gene relationships were downloaded from the GeneHancer database via the UCSC Genome Table Browser [178, 179]. The GeneHancer annotations provide the unique ability to map enhancers with associated target genes within topological boundaries using the genome arithmetic provided by the BEDtools suite intersect and windows [171]. These intersections were made using TADs known to harbor structural variants, thus allowing for the combined detection of epigenetic regulatory impact of SVs on enhancers as well as possible allelic impact to chromatin topology with phased Hi-C data. Gene target names of enhancers that were intersected by heterozygous SVs were assessed for an allelic imbalance in expression activity associated with regulatory disruption.

3.2.8 Enhancer Mediated Allele-Specific Expression Analysis

These target gene loci were obtained using the R bioconductor biomaRt package and downloaded from the ensembl database using the HGNC symbol as a query term [180, 181]. Phased read counts were obtained for these loci in order to determine allele specific expression events associated with disrupted enhancer functionality using

BEDtoos Multicov. These haplotype specific read counts across whole gene regions were then tested for significant difference using a binomial test with an FDR correction level of 5%.

3.2.9 Annotation of lncRNAs and Expression Analysis

Paired end, strand specific RNA-seq data for each of nine samples from HGSV consortium was sequenced on an Illumina Hiseq platform and mapped to the GRCh38 reference genome with the GSNAP aligner [182]. This data was downloaded from the HGCV ftp website to the high-performance computing cluster at UNCC Department of Bioinformatics and Genomics. PCR duplicates were removed and reads were filtered on base quality and read quality scores using the Picard software from the Broad Institute [183]. These filtered, stand-specific RNA-Seq reads were then used for transcriptome assembly by Cufflinks against the GENCODE version 28 reference GTF to quantify expression of long noncoding RNA transcripts for each sample [107, 58]. Unguided transcriptome assembly was also performed using Cufflinks on each sample, and the resulting fasta files were then filtered for length to include only transcripts at least 200 nucleotides long to be tested for coding potential. Transcripts that passed filtering and with low coding potential were blasted against RNACentral database of known lncRNAs to identify the presence of novel and known transcripts [184].

3.3 Results

Samples are given along the x-axis of each plot while vertical bars depict the average -log2(q-values) calculated from the group t-tests between the RPKM normalized expression values of genes engulfed by structural variants and that of genes engulfed by permuted chromosomal regions. The top left panel shows results for the integrated Illumina deletions (IL-DELs) for all 9 individuals, while the top right panel gives the results for PacBio deletions (PB-DELs) in trio daughters. The bottom left illustrates the results from the analysis of the integrated Illumina duplications (IL-DUPs) for the 9 samples, and the bottom right panel shows results from the analysis of Illumina inversions (IL-INVs) engulfed genes for the trio daughters. The position of the horizontal line in each panel corresponds to the corrected significance threshold using the Benjamini-Hochberg method to control the false discovery rate at 5%.

Our results of the allele specific expression analysis showed that the majority of heterozygous SVs tested significantly affected the target gene expression in allele specific manner. Specifically, in the pacbio SV set, a total of 144 SVs (70 insertions and 73 deletions) showed ASE effect on 60 genes, out of the 199 heterozygous SVs intersected with 78 SNP-ASE genes for NA19240; a total of 196 SVs (88 insertions and 108 deletions) showed ASE effect on 77 genes, out of the 220 heterozygous SVs intersected with 85 SNP-ASE genes for HG00514; and a total of 219 SVs (141 insertions and 78 deletions) showed ASE effect on 89 genes, out of the 274 heterozygous SVs intersected with 106 SNP ASE genes for HG00733. In the illumina SV set, 58 SVs (7 insertions, 48 deletions and 3 inversions) demonstrated ASE effect on 59 genes, out of the 83 heterozygous SVs intersected with 62 SNP ASE genes for HG00514; 60 SVs (10 insertions, 45 deletions and 5 inversions) demonstrated ASE effect on 60 genes, out of the 108 heterozygous SVs intersected with 78 SNP ASE genes for HG00733; and 57 SVs (6 insertions, 48 deletions and 3 inversions) demonstrated ASE effect on 44 genes, out of the 79 heterozygous SVs intersected with 55 SNP ASE genes for NA19240. Our SV ASE results prompted us to address whether or not the observed allelic imbalance at SV ASE genes was attributable to a local haplotype along the gene region and an example is shown in Figure 3.3. For this, we calculated the LD (R2 values) between the SVs and SNPs with ASE effect on the same gene. We illustrate the haploblock analysis to assess the allelic effect resulting from a heterozygous deletion belonging to HG00514 within a transcription factor binding site on exon 5 of the ZNF717 gene. We further ruled against a haploblock effect driving the allelic imbalance between the haplotype from low R2 values for the sample's variants and those from the 1000GP

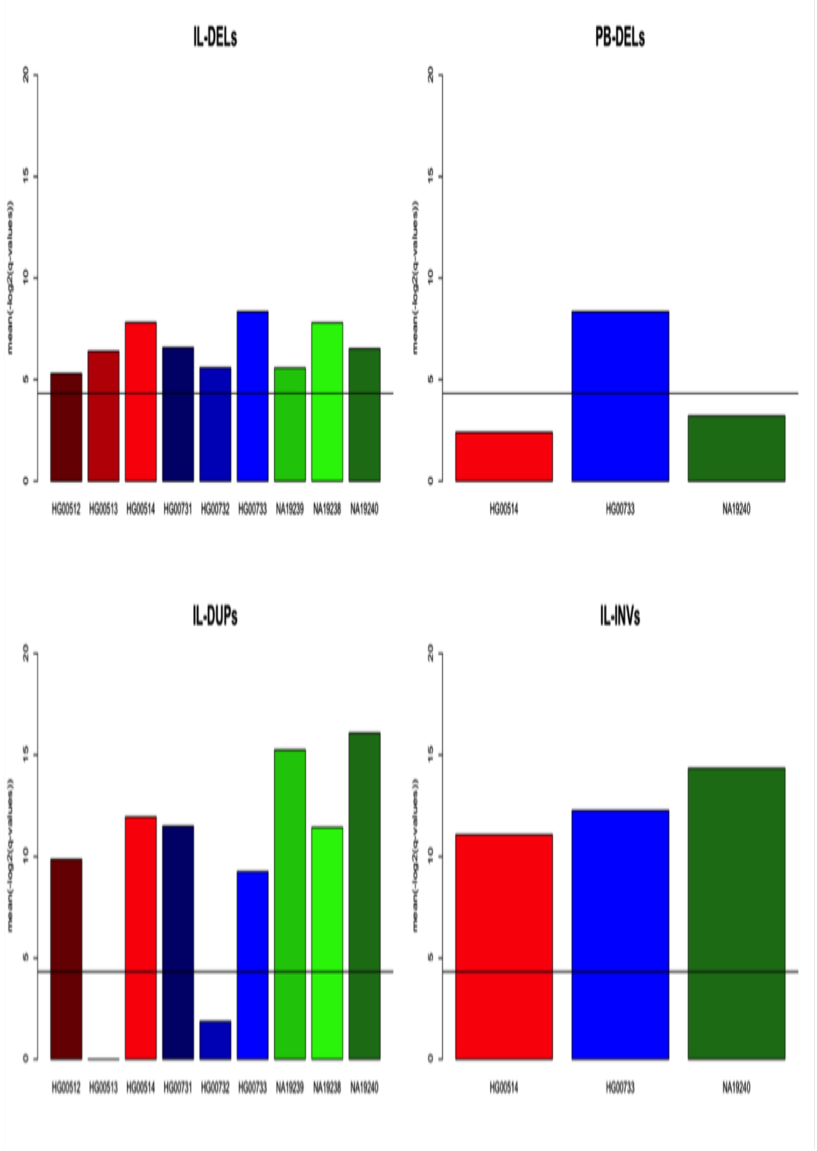

Figure 3.2: Permutation analysis of SVs that engulf protein coding genes.

phase3 CHS population within a 100kb window of the gene, and showed that there were few variants with high R2 within the exon as well.

Our investigation into chromatin structure for HGSVC trio samples began with an overall assessment of filtered, unphased Hi-C reads. Chromatin contact maps were constructed by our colleagues out of the Ren and Sebat labs at UCSD, which were normalized to removed intrinsic biases using a Poisson regression method implemeneted by HiCNorm [185]. It has been well documented that the chromatin structure appears to be stable across chromosomes within cell and tissue-types, yet the degree to which TADs may be identified is strongly dependent upon experimental conditions, such as the limitation of sequencing read depth upon the resolution of contact detection, as well as the detection algorithm [186]. In order to asses the ability of our method to detect lymphoblastoid TADs in our trio populations, we began by comparing the tad counts per-chromosome for each sample with that of the GM12878 reference cell line at high (5kb) and low (100kb) resolutions from Rao et al. and Leiberman et al., respectively [187, 117]. In order to avoid imposing any algorithmic bias in our comparative analysis, all samples and reference TADs were identified using the directionality index (DI) method. Given that our sample contact maps were constructed at 40kb resolution, the expectation would be that the numerical counts of TADS would fall somewhere between the high and low resolutions. Indeed, as depicted in Figure 3.4, we found this pattern to be true across all autosomes and the X chromosome. Notably, the number of TADs identified per-chromosome appears to be consistent for all samples, and within-family counts appear to be highly similar. From these preliminary comparisons we felt confident that the TAD detection using the directionality index calculation was consistent and appropriate for our sample data and would make for informative downstream assessments of SVs on genomic structure and function.

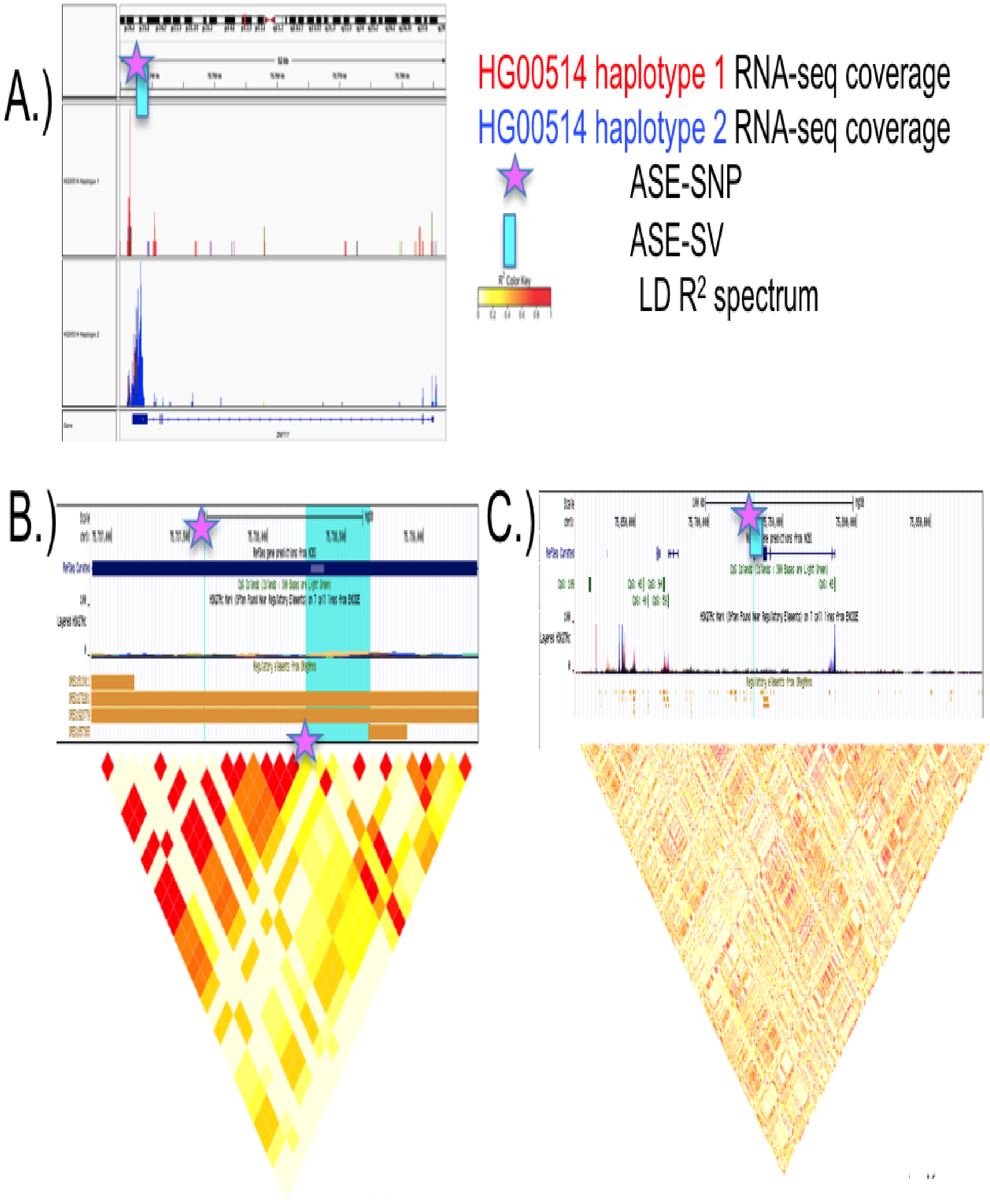

Figure 3.3: Haploblock analysis of loci surrounding ASE-SV.

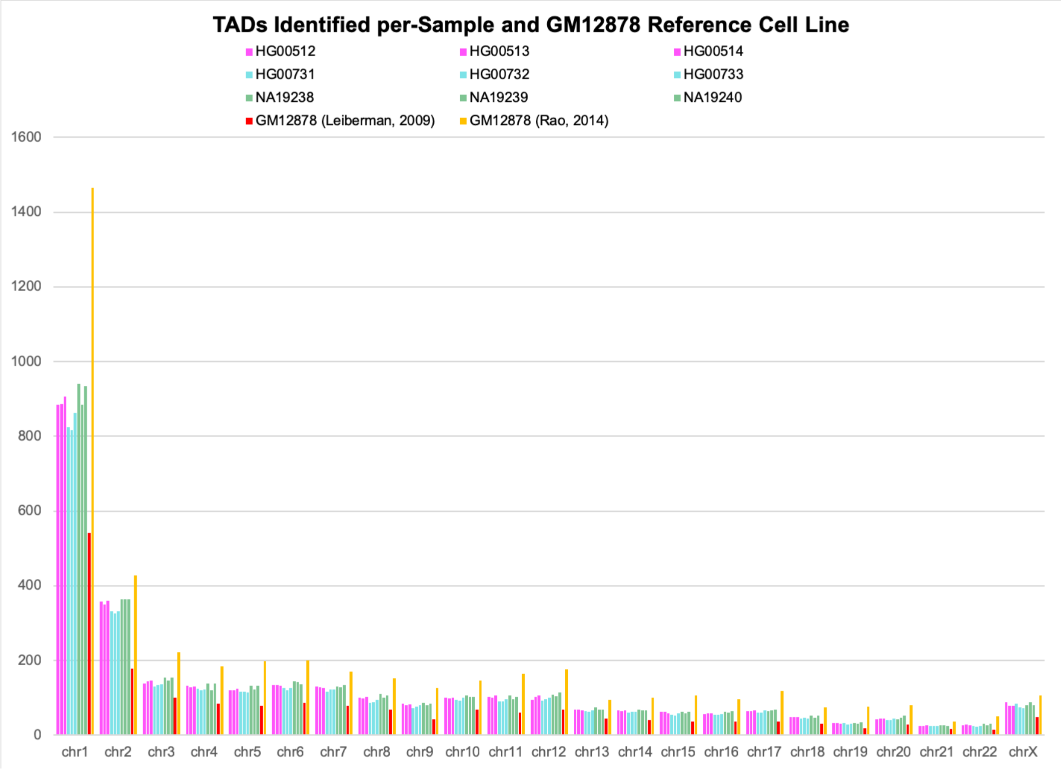

Figure 3.4: Quantitative comparison of HGSV TADs identified per chromosome with GM12878 reference cell line at multiple resolutions.

Next, we continued our comparative analysis of chromatin structures beyond simple count metrics in order to assess how well our data characterized genomic chromatin structure given the lower resolution that our reference dataset. We quantified the dissimilarity via the by the bipartite graph based approach to give BP scores. This method assesses chromosome-level TAD partitioning distances between two samples and provides scores between zero and one, where lower values indicate more similarity in TAD sets, and thus overall chromatin structure. Our results found that across all chromosomes and all samples, when comparing our sample TAD calls to those made for GM12878 at 5kb resolution, the mean BP score was 0.354 with a variance of 0.000159, indicating a high degree of similarity in the regions of the chromosome identified with similar contact frequency profiles.

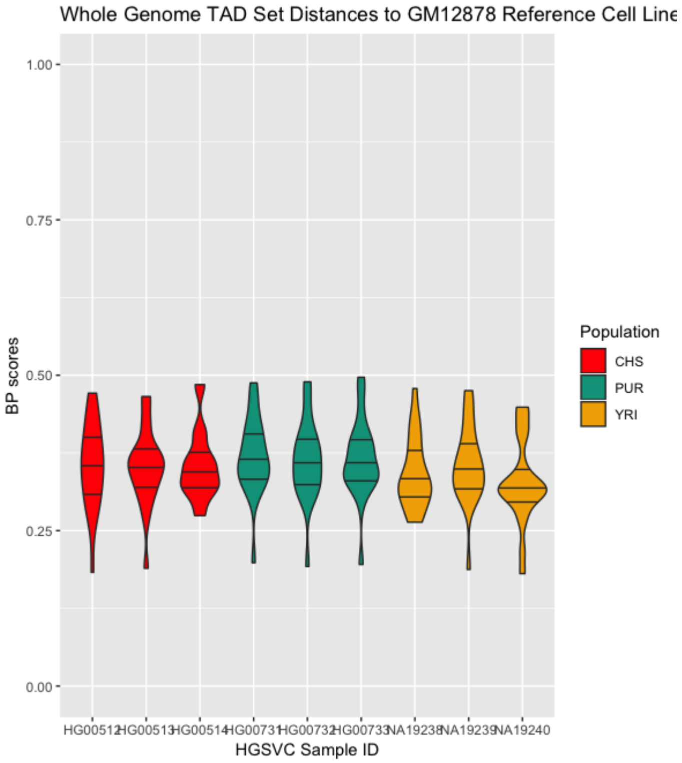

Figure 3.5: BP scores of whole genome TAD set distances compared with GM12878 TADs.

Additionally, pairwise-comparisons of TAD sets show a tight distribution of BP scores that range from 0.09 to 0.305, with over 50% of scores clustered around 0.2, indicating a high degree of TAD set similarity. This finding is inline with current understanding of TAD conservation and developmental changes and should be expected for samples of the same cell type. As shown in Figure 3.6, the violin plots show the distributions of pairwise BP scores per chromosome for each contrast between daughter sof each trio. We can clearly show that all of the trios have similar genomic TAD calls, while each contrast has a unique distribution of scores indicated by the density of the violin plots. The pairwise assessment of the Yoruban trio TADs with the Han Chinese shows a tight similarity profile, with a range of 0.15 to 0.25. The Puerto Rican sample is representative of a more admixed population than either of the Han Chinese or Yoruban, and displays somewhat wider tails in the plot which would indicate regions of greater distance in the overall chromatin topology. The greatest amount of distance between the populations was for the TADs called along the gene dense chromosome 11, between the Han Chinese and Puerto Rican samples, which may indicate a greater amount of topological diversity between these samples, yet the mechanism behind this observation remains only speculative.

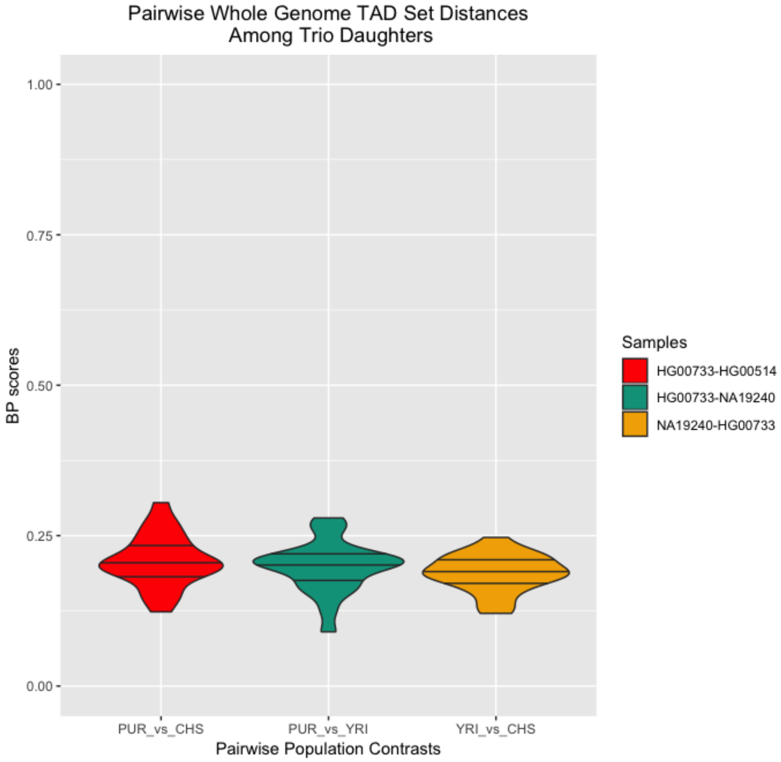

Figure 3.6: Pairwise assessment of trio TAD set distances.

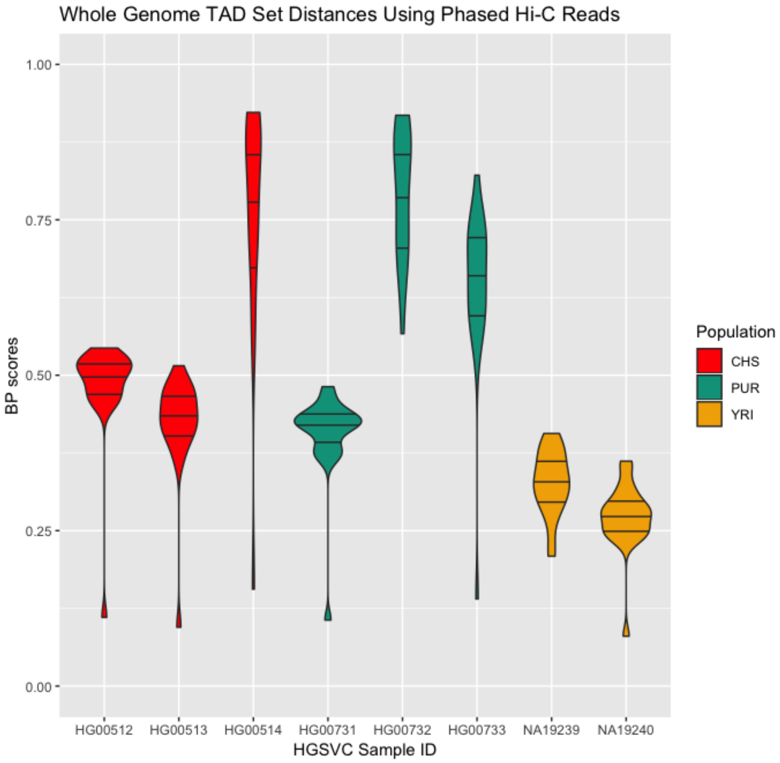

Figure 3.7: BP scores of whole genome TAD set distances among haplotypes using phased Hi-C.

The same analysis segmentation set analysis was applied to statistically phased Hi-C reads. The phasing process greatly reduced the number of Hi-C sequencing reads, and thus left for a difficult process of detecting TAD boundaries at anything other than a poor resolution. The is to say, roughly 10% of reads remain in the resulting bam files as compared to the unphased data. The phased Hi-C bam files are limited in their ability to detect a majority of the chromatin structure due to the decrease in resolution. For instance, due to the reduction in Hi-C reads contacts must be inferred with a decreased resolution of 100kb to boost the signal to detect long range contacts. A comparison of the TAD sets between each haplotype can be seen in Figure 3.7 with the similarity for each sample's chromosome partitioning is given by a BP score. One noteworthy exception is sample NA19238, for which BP scores could not be reliably calculated, likely resulting form the reduced ability to detect non-empty overlaps between the haplotypes given the low resolution of the initial experiment and reduction in Hi-C reads due to statistical phasing. The remaining samples show a similar distribution to that of the unphased data compared to the high-resolution GM12878 reference cell line, which would indicate an analogous ability to utilize phased TADs for downstream analysis. Additionally, the increase the dissimilarity between the haplotype-specific TAD sets may be indicative of possible allele-specific chromatin domain structures resulting from SVs. Overall the mean BP score across chromosomes for all samples was 0.512, with a variance of 0.038. Given that the bipartite graph approach is able to tease our some the subtle noise between the sample haplotypes, these statistics indicate the likely presence of distinct allelic chromatin structures, with the known caveat of poor resolution limiting the precise boundary delineation. Therefore the allelic TAD structures may provide a glimpse into more pronounced differences in the chromatin, while a fine-tuned approach would benefit greatly from increased sequencing depth to boost genomic coverage after phasing.

3.3.1 Functional Analysis of SVs intersecting TADs

We assess the distribution of SVs across chromatin domains we aimed to find how many of the SVs were located within the boundaries of domains. By mapping the heterozygous SVs in this manner we identified that a majority of TADs harbor structural variants of various types, and and furthermore that most SVs tested were located within these boundaries for these samples, which is suggestive of the overall resilience of chromatin architecture and reflects the known detrimental regulatory effects of neo-TAD formation or TAD fusion events. Moreover, there were more deletions found within TAD boundaries than insertions or duplications, which may simply be a reflection that the SV detection algorithms more easily map these variants. The counts of IL-SVs per sample can be found in table 3.1, and shows a similar pattern of intersections for all samples and among each trio. For the Han Chinese family, the daughter possessed 1,704 TADs which were intersected by IL-SVs, while the mother and father had 1,684 and 1,657, respectively. The Puerto Rican trio had the fewest

Table 3.1: Structural variants within TAD boundaries.

Sample	TADs	IL-SVs	SVs int TADs	IL-SV INS	IL-SV DEL	IL-SV DUP
HG00512	1657	7655	6687	921	5177	589
HG00513	1684	7884	6807	935	5250	622
HG00514	1704	7993	6969	945	5401	623
HG00731	1556	8116	7031	946	5480	605
HG00732	1554	7996	6927	951	5395	581
HG00733	1610	7906	6803	920	5270	613
NA19238	1795	9352	8204	1103	6405	696
NA19239	1707	9189	7982	1106	6213	663
NA19240	1797	9507	8270	1150	6389	731

TADs intersected, with 1,556 in the father , 1,554 in the mother and 1,610 for the daughter's genome. Conversely, the Yoruban trio had the most TADs intersected by SVs, where the father had 1,795, the mother had 1,707 and the daughter had 1,797. This observation may have to do with the shear number of SVs detected for each trio was variable, while the Yoruban trio had the most overall.

We next took a look at the overall distribution of genomic SVs with respect to epigenetic regulatory elements, specifically enhancers with curated gene targets within TADs. This method identified 8,529 enhancers-gene possibly impacts by the intersections of SVs within TADs for HG00512, 8,778 for HG00513 and 8,810 in HG00514. For HG00731 there were 8,525, and 8,508 enhancers-gene pairs for HG00732, with 8,307 relationships for HG00733. The numbers of enhancers-gene pairs intersected for the Yoruban trio were 9,573 for NA19238, 9601 for NA19239, and 9,804 for NA19240. The overall breakdown of heterozygous SVs by class and the number of unique enhancers can be seen in table 3.2. Not surprisingly, there were several enhancers impacted within the trios that had identical gene targets, which has been shown by the venn diagrams in Figure 3.8. The gene targets were then assessed for allele specific expression given the allelic imbalance of SVs intersecting enhancer regulators of their expression. In Figure 3.9, we see that there are a a great deal of significant loci that

Table 3.2: Structural variants that intersect enhancers within TADs from Illumina callset.

Sample	TADs	IL-SVs	Enhancers	IL-SV INS	IL-SV DEL	IL-SV DUP
HG00512	1536	2662	7154	321	2068	273
HG00513	1555	2736	7411	332	2121	283
HG00514	1583	2841	7461	351	2220	270
HG00731	1448	2850	7229	337	2244	269
HG00732	1453	2742	7105	333	2141	268
HG00733	1501	2696	6992	297	2122	277
NA19238	1645	3300	8133	385	2631	284
NA19239	1571	3217	8111	403	2527	287
NA19240	1654	3314	8306	405	2600	309

exhibit allele-specific expression patterns when associated enhancer regulatory elements located within TADs are impacted by an SV. One interesting feature of theses observations is that of the Puerto Rican trio enhancer ASE, in which all but one of the daughter loci are shared with the mother. Moreover, only 4.6% of target genes shared by HG00731 and HG00733 exhibit similar ASE patterns, whereas that figure is 21.6% between HG00732 and HG00733.

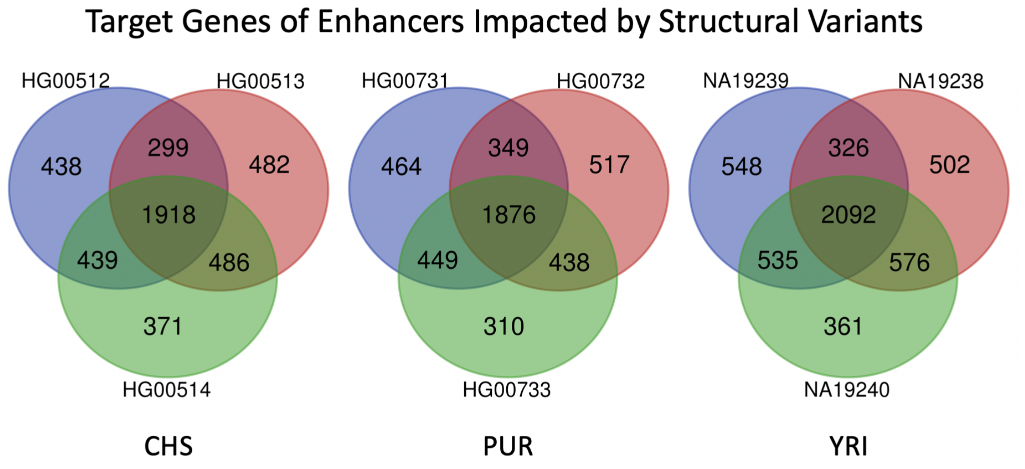

Figure 3.8: Venn diagrams of gene targets with enhancers intersected by SVs.

Table 3.3: Structural variants that intersect enhancers within TADs from Illumina callset and ASE.

Sample	ASE Genes	IL-SVs	IL-SV ASE INS	IL-SV ASE DEL	IL-SV ASE DUP
HG00512	394	482	71	337	73
HG00513	407	489	57	361	70
HG00514	513	614	78	457	78
HG00731	418	505	72	379	53
HG00732	501	622	92	457	72
HG00733	501	622	92	457	72
NA19238	645	796	125	576	94
NA19239	673	824	128	609	86
NA19240	562	681	99	507	740

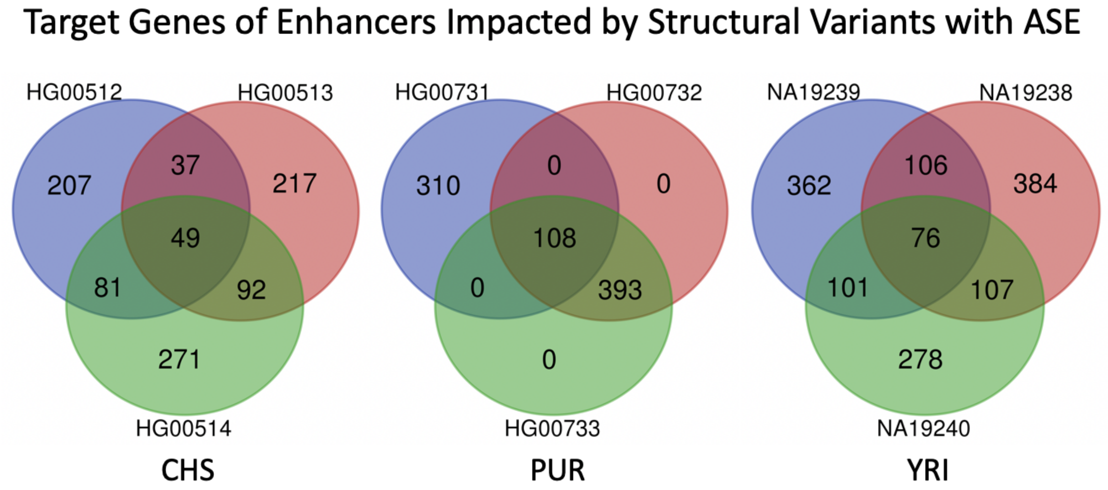

Figure 3.9: Venn diagrams of gene targets with enhancers intersected by SVs that exhibit ASE.

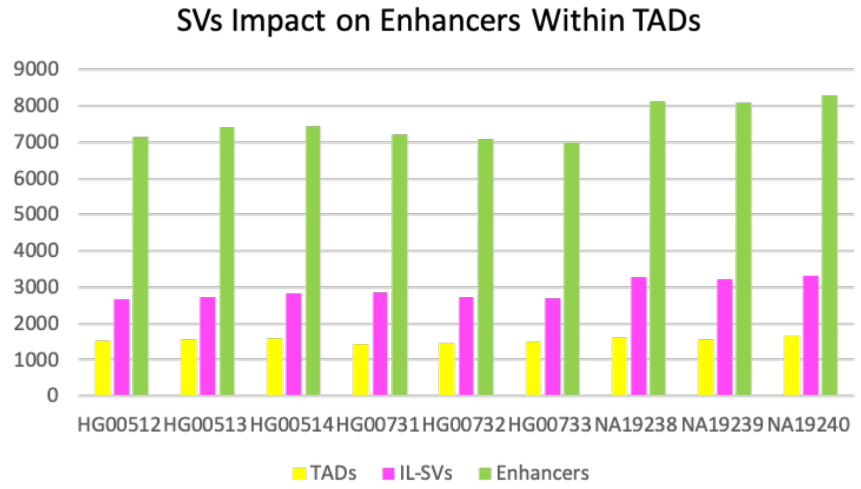

Figure 3.10: Per sample counts of SVs intersect with TAD encapsulated enhancers.

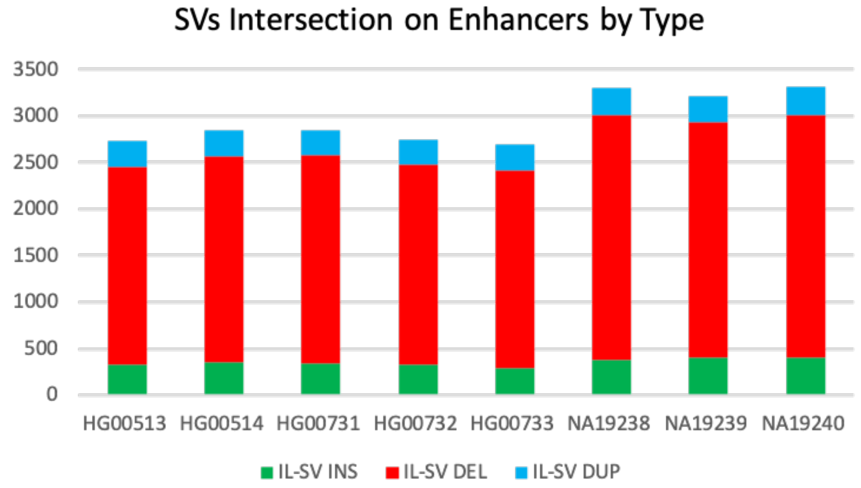

Figure 3.11: Stacked bar chart of SV types that intersect with TAD encapsulated enhancers.

In Figure 3.10, we see the per-sample count distribution of SVs with that intersect enhancers that are located within sample-specific TADs. This bar chart provides a visual representation of the information in Table 3.1 where TADs are in yellow, total SVs are in pink and enhancers are in green. We can clearly see a pattern where large SVs intersect with many enhancers that are located within even larger TAD regions. Additionally, we depict a stacked bar chart to show the the types of SVs that intersect these enhancer elements with target-genes located within sample specific TADs in Figure 3.11. For each sample we follow commonly used color scheme for variants, where insertions are colored in green, deletions in red, and duplications in blue. This chart shows roughly equal numbers of insertions and duplications for each sample, while deletions are much more numerous in each case. This result appears to be non-significant, as the current sequencing technologies and SV calling methods tend to bias the detection of deletions. As the technologies improve the detection of SVs in the future, there may be more variety of SV types found to overlap with enhancers in this way.

One specific case of NA19240 TADs using BEDtools to find 30 sample-specific TADs impacted by insertions and deletions and subsequently determined to contain 27 GM12878 epigenetic enhancers downloaded from UCSC. This preliminary analysis led to the discovery of a 280kb heterozygous deletion along chromosome 2. This region was found by Rao et al. to contain a single TAD in the GM12878 reference cell line, indicating that the deletion within this region disrupted the contact profile of the chromatin at this locus and led to the formation of two distinct TADS [187]. Furthermore, this deletion spans a loci known to contain a transcription factor binding site and enhancer element that lies adjacent to two genes, *IGKJ1* and *AC244205.1*, and analysis of NA19240 RNA-seq data showed significant difference in the expression of both of these genes. The deleted enhancer, *GH02J088855*, catalogued by GeneHancer scores high for direct interactions with the genes exhibiting allele-specific expression. This can be visualized by Figure 3.12, where the top panel shows GM12878 Hi-C contact matrix from Rao et al. and the single TAD indicated by a gold bar, the middle panel shows a UCSC genome browser session at this locus with the deleted region highlighted in red, and the neighboring NA19240-specific TADs highlighted blue and yellow. The bottom panel of Figure 3.12 shows the IGV plot of genomic RNA-seq covereage at this region, with genes indicated by horizontal blue bars.

In another case, a heterozygous 340bp deletion overlapped the enhancer element GH21J029235, which is intronic to, and interacts with the lncRNA *LINC00189*, to modulate the activity of its expression. We diagram this result in Figure 3.13 using the HiCPlotter tool to overlay multiple tracks in the plot. The top-most track shows a heatmap of the normalized Hi-C contact profile as log2 values of the interaction matrix, for the Han Chinese HG00514 along chromosome 21 between 28.5Mb and 30Mb. The color scale of the heatmap ranges from dark red to indicate few contacts to orange and finally bright yellow as the contact frequencies increase. The tack below this shows the view of the genomic region of interest along the diagonal to make for easier visualization of the regional chromatin contacts. Below this is another track of gencode v28 genes as dark blue rectangles, which are mapped used GRCh38 genome coordinates. Underneath this layer are the locations of three enhancers that are annotated by the geneHancer database to have known interactions with the genes depicted above. Through each of these layers of the image is a pale blue highlight which indicates a heterozygous deletion that can be seen to intersect both the lncRNA as well as the enhancer GH21J029235. Finally, we show in the bottom-most track the location of HG00514 TADs as khaki triangles superimposed on GM12878 reference lymphoblastoid cell line TADs in blue and IMR90 TADs in red as an example of healthy lung tissue. We note that the expression of the genes shown in this figure both possess an allelic imbalance, yet in opposing directions. To investigate what impact this structural variant may impose on these genes and explain the observation of both allelic imbalance and opposing effects, we consulted tissue-specific expression patterns for both the lncRNA *LINC00189* as well as the protein coding *LTN1* through the GTEx expression database. We obtained the log expression levels across epstein barr transformed lymphoblastoid cell lines as an anlaogous reference for both GM12878 and our trio samples. We sampled tissues from several body sites and found that a distinct pattern emerged that can be seen in Figure 3.14, which illustrates this point.

We have stratified the samples by gender and sorted by the median expression values, and found that there is a -0.75 Pearson correlation between the median expression values of these genes across the sampled tissues, indicating a strong indirect association in their activity levels. Based on what is known about lncRNA gene regulation this may suggest that there is coordinated gene regulation by the lncRNA at this genetic loci that appears to have been disrupted by the heterozygous deletion at the site of the intronic enhancer element. Once such example of this has already been shown in the literature, in which the intronic enhancer of the cystic fibrosis associated *CFTR* locus is responsible for coordinated activation of the *CFTR* promoter region through a chromatin looping mechanism that produces distinct expression patterns with the adjacent *ASZ1* and *CTTNBP2* genes by means of tissue-specific chromatin looping [188]. While it still remains to be experimentally validated through the use of targeted chromatin looping such as 4C or 5C, the results we observe here with regard to the coordinated heterozygous deletion of this intronic enhancer with the allelic asymmetry in the expression of the *LINC00189* and *LTN1* genes and identify a candidate region with similar chromatin looping-mediated gene regulation. Moreover, the these two genes are antisense, and transcription of the lncRNA could modulate the protein coding gene activity, which warrants further analysis of lncRNA-DNA binding and chromatin state modeling.

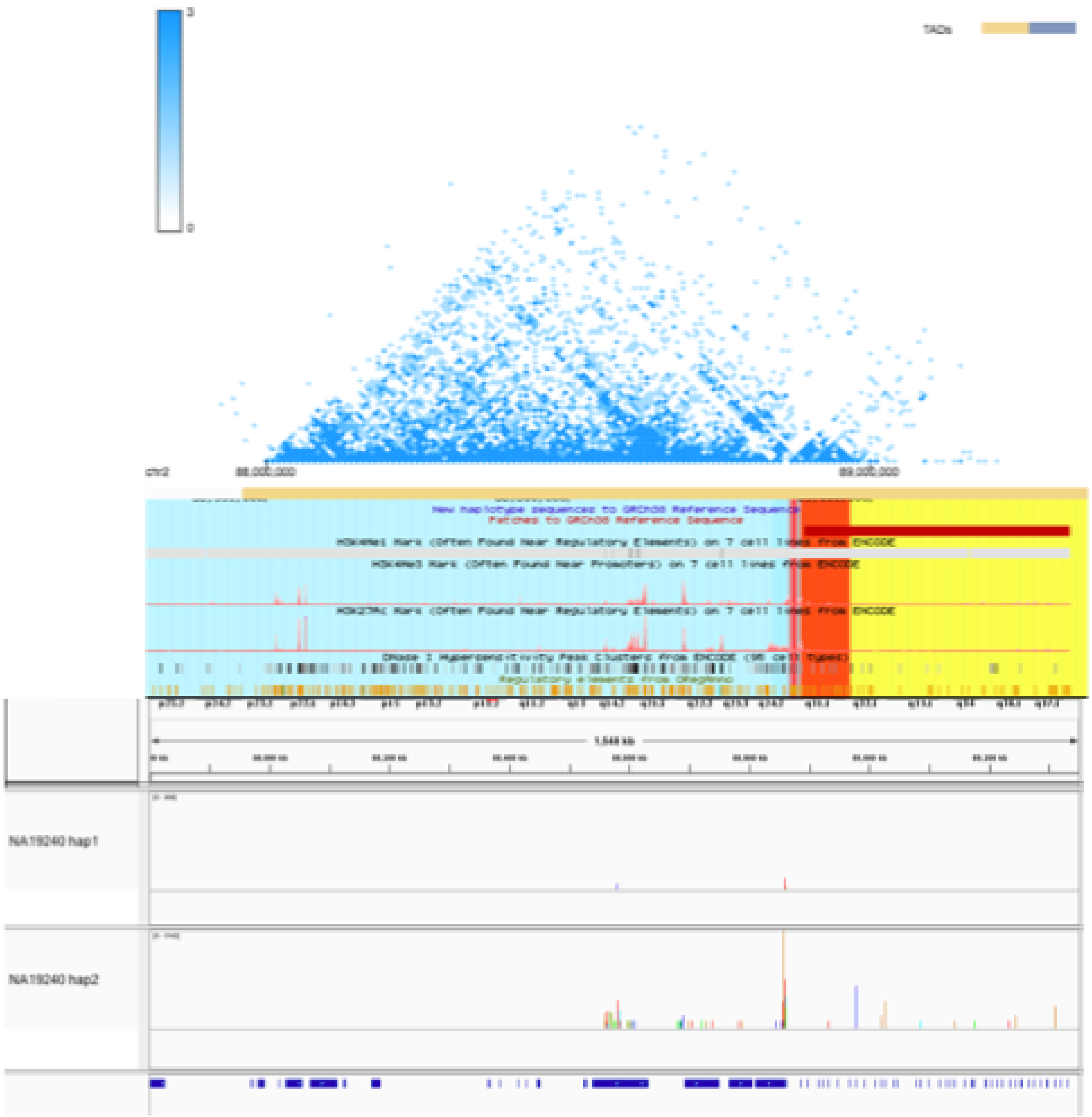

Figure 3.12: Large deletion coincides with altered chromatin topology, altered expression in NA19240.

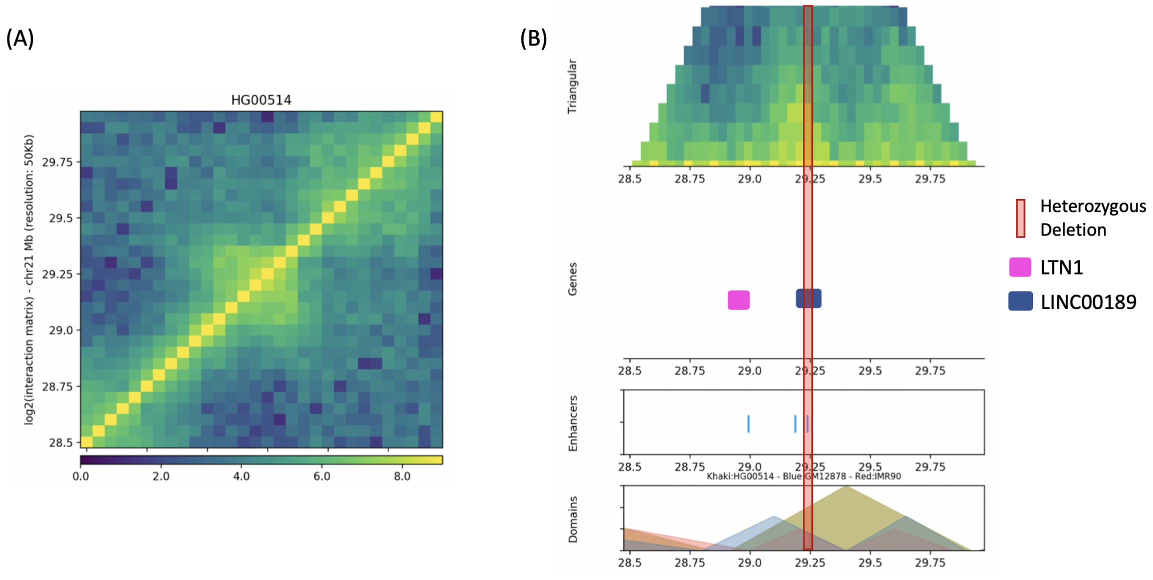

Figure 3.13: Hi-C interaction matrix overlaid with SV impact lncRNA and intronic enhancer element GH21J029235.

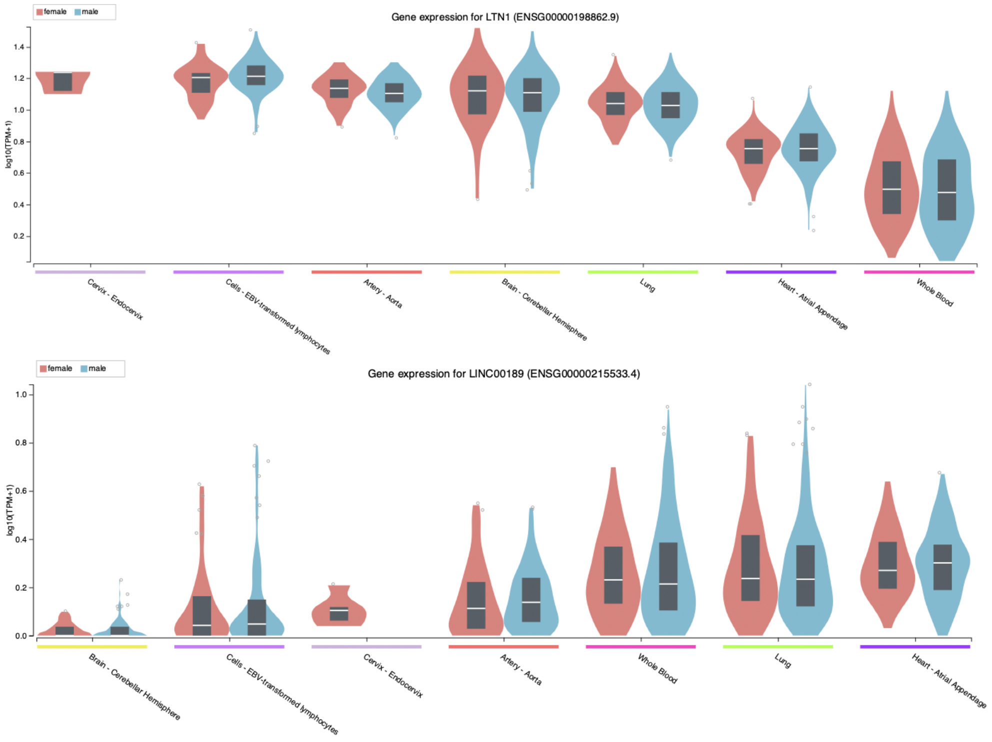

Figure 3.14: Tissue specific expression patterns of LINC00189 and LTN1.

Table 3.4: Allele specific expression in HGSVC Pacbio callset.

Sample	PB-SV Genes	PB-SV ASE Genes	PB-SV ASE INS	PB-SV ASE INS
HG00514	85	77	88	108
HG00733	106	89	141	78
NA19240	78	60	70	73

Table 3.5: Allele specific expression in HGSVC Illumina callset.

Sample	IL-SV Genes	IL-SV ASE Genes	IL-SV ASE INS	IL-SV ASE DEL	IL-SV ASE INV
HG00514	62	59	7	48	3
HG00733	73	60	10	45	5
NA19240	55	44	6	48	3

CHAPTER 4: THE DIFFERENTIAL IMPACT OF STRUCTURAL VARIATION ON GENE REGULATION IN LEUKEMIA SUBTYPES

4.1 Introduction

Pediatric acute lymphoblastic leukemia (ALL) is a hematological cancer that has been experienced a dramatic success in terms interventional strategies with positive clinical outcomes in a majority of afflicted patients. Since the discovery of the Philadelphia-chromosome, large structural variants have come to be recognized as a significant contributing factor in the development of circulating tumors. Patients positive for the Philadelphia-chromosome possess a genome harboring a reciprocal translocation between the long arms of chromosomes 9 and 22 that leads to an expressed fusion gene product of two kinase signaling molecules, ABL1 and BCR. With no endogenous regulatory mechanism, the unmitigated phosphorylation activity by this novel fusion protein stimulates B-cell growth and evasion of apoptosis and provides an archetypal link between structural variant, oncogenesis and tumor development. Similarly, cytogenetic and molecular phenotyping experiments have now found several translocations and genomic rearrangements with recurrent breakpoints adjacent to genes regulating B-cell development in pediatric ALL patients [189].

While these aberrations are often oncogenic, it is important to note that they have been found to exist in samples derived from healthy umbilical blood at much higher frequency than the reported oncogenic incidence, suggesting they may not be sufficient and a second genetic lesion is likely necessary for leukemogenesis to occur [149]. A genome-wide analysis of pediatric B-lymphocyte ALL patients found that there is often an additional copy number differences at loci responsible for B-lineage fate-specification that act as cooperating mutations with recurrent translocations to promote leukemia [10]. Further work demonstrated somatic alterations were enriched in genes that contribute to B-cell development, *Janus* kinases, the *tp53* tumor suppressor pathway, as well as other cellular growth signaling cascades [150]. Additionally, a study of DNA methylation found consistent epigenetic silencing via hypermethylation at promoters of these genes, however this study did not make use of immunoprecipitation data to investigate the effects on chromatin state due to somatic mutations [151, 152]. From these studies we can glean that B-cell specific areas of the genome contribute to the pathogenic state of leukemia, however what lacks is a look into the noncoding genome that may be disrupted, and how the chromatin topology is affected by structural variation with respect to shared translocation groups. Furthermore, studies of HOX genes among leukemias showed they contribute to oncogenesis through disturbed hematopoiesis, yet are not sufficient to discriminate between subgroups based on combined expression alone [153, 154]. These expression based studies may be limited by this approach, while integration with epigenetic data types could improve the discriminatory capacity based on capturing subtype-specific regulatory relationships between the coding and noncoding genome.

In the clinical setting, the cytogenetic reporting of a patient's subtype must be complete in order to guide treatment protocols and pharmaceutical regimentation. This diagnostic workup is essential to the long-term outcome and quality of life of the patient, however it can be seen as a rate limiting step in the overall medical workflow. The rapid distinction between subtypes with known links to genomic rearrangements may be better facilitated by the implementation of bioinformatics methods that are capable of leveraging the multitude of available multi-omics datasets. For rare subtypes, such as the Ph-like group, little is known about the root cause of the observed expression anomalies. Therefore, an analysis based upon integrated copy-number profiles and expression measurements may elucidate the key oncogenic factors contributing to the clinical phenotype behind the Ph-like group.

Table 4.1: Accessions and hyperlinks to data sources for microarrays analyzed.

Database	Accession	Download URL
NCI caArray	EXP-578	ftp://caftpd.nci.nih.gov/pub/caARRAY/experiments/caArray_EXP-578/
NCBI-GEO	GSE11877	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE11877
EMBL-EBI	EGAS00001000654	https://www.ebi.ac.uk/ega/studies/EGAS00001000654

4.2 Methods

4.2.1 Microarray Expression Analysis of Stratified Leukemia Samples

Acute lymphoblastic leukemia expression analysis was performed for a large microarray dataset with special attention given to the Ph-like ALL subtype do ascertain specific, coordinated gene activities in patients with the rare diagnosis. This study began by the identification of differentially expressed genes in the Ph-like ALL subtype with respect to alternative acute lymphoblastic leukemia groups. Differential gene expression analysis was performed and heatmaps generated for the top 50 differentially expressed genes, which can be found in chapter 2 of the appendix. The expression matrices were then used to perform functional annotations as well as an enrichment study of the genes involved in metabolic signalling pathways with significant differences across each of the ALL subtypes. Below, we provide a detailed overview and visual workflow for the expression analysis performed.

4.2.2 Data Acquisition

The HGU133_Plus2 Affymetrix microarray data generated by Roberts et al. was uploaded to multiple online databases under different accessions [190]. Microarray probe intensity measurements were formatted as raw .CEL files and collected from three separate online accessions at EBI, NCBI and NCI caARRAY and downloaded to the high-performance computing cluster at University of North Carolina at Charlotte. The accession numbers and web addresses to the data may be found in Table 4.1.

4.2.3 Identification of Differentially Expressed Genes in the Ph-like Subtype

Differential gene expression analysis was conducted to test for significant differences in the gene-level expression across the two Ph-like groups with respect to the expression levels of the contrast groups. The contrast groups are given by the nine non Ph-like ALL group classifications found in table 2. For any given contrast, a file containing the paths to the .CEL files and a phenotype description reflecting the group name was read by the Bioconductor package simpleaffy [191]. This package allowed the raw .CEL files to be read in as a batch and normalized together using the robust multi-array average algorithm (RMA) to create an expression set (eset) object in the R environment [192]. Limma was used to fit a linear model for each gene in a set of microarrays after filtering out low-variance probe-sets [193]. A contrast matrix was then constructed to perform the contrast fit between the Ph-like group and each of the alternates, followed by empirical Bayes moderation of the statistics for differential expression and on the standard errors as a means to evaluate the log odds of differential expression. Given that the HGU133_Plus2 affymetrix array platform has a large amount of redundancy, in that probes often map with a many-to-one relationship to a gene, the probes were collapsed to the gene level based on the probe with the largest absolute levels of expression for a given gene across arrays. This was done so that the measured differences in expression would be less influenced by inefficient hybridization of probes and a lack of continuity across arrays. These probes were then annotated with their official gene symbol. Genes were selected for further analysis based on a FDR significance threshold of 0.01, and a B-value of 1.5. This set of criteria may be interpreted as having limited the set of genes identified as being differentially expressed to those, where after for multiple hypothesis testing, have a one-percent chance of being a false-positive and a 1.5 odds of being differentially expressed value relative to the contrast group [194].

This dataset was comprised of 1,319 microarray expression files only for those on

Table 4.2: Distributions of ALL subtypes as defined by recurrent genomic rearrangement events.

ALL Group	Samples (N)
Ph-Like	223
BCR-ABL1	87
CRLF2	21
E2A-PBX1	75
ERG	110
ETV6-RUNX1	102
Hyperdiploid	131
Hypodiploid	12
MLL	74
Other	484

the HGU133Plus2 Affymetrix platform. This platform contains 54,675 hybridization probes, which map to 20,606 unique genes [106]. Prior to analysis, the microarrays were divided according to the group classification given by column R of appendix one, table 1 in the supplementary materials provided with the original article. Described in appendix two, figure S18 on page 44, these groups were defined by the source publication using a combination of molecular profiling, cytogenetics, immunophenotyping, and low-density expression arrays. The counts of microarrays for each subtype may be found below in Table 4.2.

4.2.4 Detection of Differential Activity in Metabolic Signaling Pathways across ALL Subtypes

Pathway enrichment analysis is a form of bioinformatics and statistical association analysis of gene expression values across a set of samples. The end goal is to identify coordinated gene expression differences between groups based on known sets of genes that belong to a common biological pathway. To perform pathway analysis, we tested the entire expression matrix for enriched pathways in the Ph-like subtype and the Ph-like with *CRLF2* rearrangent against each of the other groups using the GSEA software. This software ranks genes belonging to particular pathways and associates them with either phenotype via a Kolmogorov-Smirnoff test [195, 196].

In order to probe the impact of recurrent genomic rearrangements on the phenotypes and outcomes associated with subtypes of ALL, we expanded on the differential gene expression by attempting to resolve the enrichment of genes associated with B-cell developmental pathways and oncogenic characteristics. Gene set enrichment analysis (GSEA) is a commonly used tool for the bioinformatic enrichment tests of particular genes. The protocol for this analysis begins with matrix $[N \text{ x } (M + J_i)]_k$ for M Ph-like samples, plus J samples of subtype i, by N quantile-normalized gene expression measurements taken for each contrast, k, of Ph-like and another ALL subtype.

Complete analysis reports may be found in the supplementary materials under the supplementary materials. Within that directory the reports are further divided by the Ph-like group being interrogated for the gene sets tested, and split again by the enrichment contrast group. For GSEA, a negative enrichment score corresponds to whichever phenotype is second in the expression set categories. For all results, the left-right organization of the original expression sets in which the right most samples belonged to the Ph-like groups was maintained throughout and thus, the enrichment scores for the Ph-like groups are all negative. These results reflect that the KS-test ranking of genes assigns negative values when a gene is correlated with the second phenotype by GSEA convention, rather than the intuitive interpretation of negative correlations, and each negative value contributes to the cumulative enrichment score when the gene also belongs to the gene set [197].

4.2.5 Coexpression analysis of lncRNAs with mRNAs

The lncRNA contribution between subtypes can be ranked according to results of LPC or T-scores depending upon the predictive advantage of the lassoed principal components approach computed using the R package LPC. Each contrast between Ph-like and other subtypes may be treated as several two-class cases in the model.

Selected distinguishing lncRNAs can then be identified via conventional differential expression and build networks around their co-expression with differentially expressed protein coding genes to identify differences in lncRNA patterns between ALL Ph-like subtype.

4.2.6 Integration with Topologically Associated Domains

We have previously studied the relationships between lncRNAs and protein coding genes located within chromatin domains as a means to study differential regulatory impacts resulting from large-scale chromatin rearrangements in the form of classically recurrent translocation. We then proposed the idea that samples of precursor B-cell acute lymphoblastic leukemia (ALL) may be classifiable with a narrows focus on those genes within close three-dimensional proximity. Raw expression data available from the published research by researchers at St. Jude working in conjunction with the TARGET initiative was obtained for 1,319 samples in the form of HG-U133plus2 Affymetrix array CEL files. According to the results of this study, subtype annotations could be made for samples with respect to the presence of 9 recurrent genomic rearrangements. These aberrations included the BCR-ABL1 gene fusion that results from a reciprocal translocation along the long arms of chromosomes 9 and 22 known as the Philadelphia chromosome (Ph(+)), fusions of E2A-PBX1 and ETV6-RUNX1, and rearrangements at the CRLF2, MLL and ERG loci. Samples that cluster together with the Ph(+) samples yet lack the archetypal translocation of the class are considered to belong to the rare ALL-subtype, Philadelphia-like (Ph-like). The remaining classes may be grouped as hyperdiploid/hypodiploid or Other. Microarray probes intensity measurements were normalized with the robust multi-array average (RMA) method and low variance probes were filtered. The remaining probe sets were collapsed to the gene level using a criterion of the maximal median expression value measured across all samples. The expression matrix that resulted from preprocessing contained measurements for 11,976 genes across 1,319 samples. Annotations from gencode.v19, GRCh37 reference genome, contained 13,870 long non-coding RNA genes. These annotations were used to identify the 282 lncRNAs that remained in the expression matrix. Rao et al. (Cell, 2014) published a map of the intrachromosomal interactions observed from high coverage in-situ Hi-C experiments mapped to the GRCh37 reference at 1kb resolution for the GM12878 lymphoblastoid cell line. In total, 9,275 intrachromosomal topologically associated domains (TADs) were called by this group with the Arrowhead algorithm and may be obtained under the GEO accession GSE63525. Previous studies have shown both the conservation of TAD structure across mouse and human three-dimensional genomes, as well as the presence of cell-type and tissue-specific variation in the local chromosomal architecture of the nucleus. For these reasons, several studies have used cell line TAD structure as a reference to compare case samples against [34, 163, 130]. These finding lend credence to the use of GM12878 TAD calls as a three dimensional reference map for the acute lymphoblastic leukemia samples. From the 282 lncRNAs present in the filtered expression matrix, 163 were found within 204 of these reference TADs. GM12878 domains were used again and identified 607 encompassed protein coding genes, while 350 were found in the expression matrix. The intersection of these genes resulted in a final set of 141 lncRNA and 350 protein coding genes within 193 defined TAD boundaries for use as predictors in our models. We found that the lncRNA:TAD ratios were 181 single-lncRNA TADs, 11 two-lncRNA TADs, and one TAD with three TADs.

Several different classification models were tested for accuracy using identical test set of (n = 989) and training set (n = 330). A random forest model was implemented as well as regularized multinomial regression models with varying alpha parameter settings ranging from 0 (ridge) to 1.0(Lasso) with a step size of 0.1. Parameter optimization to train each model was performed with 10-fold cross validation procedure with constant fold-ids on the training data in order to make comparisons of the models more reliable. The hyperparameter optimization performed for all models was

10-fold cross validated grid search, with the exception of the elastic net implementations, which implemented coordinate descent for tuning. The cross-validated models were then compared for the ability to predict the genomic rearrangement of the test set using the model-specific optimum lambda-min values learned during training. In addition we tested these against a naive Bayes classifier, simple logistic regression, a decision tree, random forest classifier, and K-nearest neighbors. In addition we tested the models on randomly permuted sets of 3,207 to evaluate the usefulness of our lncRNA-mRNA pair expression values as predictors for class status.

4.3 Discussion

The overarching aim of the differential gene expression analysis was to determine if there were any distinct differences between the Ph-like subtypes of precursor B-cell acute lymphoblastic leukemia. Through the use the R statistical programming language, commonly adapted microarray normalization workflow was implemented, moderated test-statistics were calculated and significance levels corrected for multiple-testing errors using Benjamini-Hochberg method off false discovery rate correction. This method allowed for the selection of differentially expressed genes in the Ph-like group based on a one-percent false positive rate and 1.5 odds of differential expression. These statistically significant genes were then clustered based on common hits to KEGG pathways, which served as the means to narrow the search for gene set enrichment analysis with GSEA. By testing the gene sets identified by gene pathway clustering, the expression sets for the differentially expressed genes could be analyzed across the groups for pathway enrichment. The results of the analysis by GSEA show that, for the enriched pathways, the number of genes belonging to the pathway gene set and also found in the expression set is greater than ten genes for all contrasts. This cutoff helps eliminate the inflation of enrichment scores from low-counts. The pathway enrichment analysis provides a description of the biochemical level differences in the Ph-like groups that may be inferred from the expression data. This analysis showed that there was no significant difference between the two Ph-like groups and those patients classified as possessing BCR-ABL1 and CRLF2 rearrangements and those shown to be hypodiploid. While the first two groups also showed the lowest number of gene counts in table 3, the relatively low number of patients found to be hypodiploid suggests these results may be improved in future works with a greater sample size. Here, it was shown that several pathways were unregulated in the Ph-like ALL samples, especially when compared to the ERG, ETV6-RUNX1, and E2A-PBX1 and Other groups. These results would suggest that that there are distinct patterns in cellular proliferation and disease progression for the Ph-like variety of precursor B-cell acute lymphoblastic leukemia. For Ph-like (with CRLF2) samples, this pathway was significantly enriched compared to the E2A-PBX1, ETV6-RUNX1, Other and the ERG groups. It is worth noting that according GSEA user manual, an FDR value cutoff up to 25% may yield significantly interesting results due to the inherent differences across expression sets, and by this metric the contrast between Ph-like and hyperdiploid is borderline enriched for this cytokine interaction pathway. This pathway is shown to be involved in regulating cellular growth and is also implicated in stimulating the mobility of the lymphocytes [198].

For both types of Ph-like samples, the *JAK-STAT* signaling pathway was enriched against samples belonging to the Other, ERG, and E2A-PBX1 groups. The constitutive activation of this pathway has been shown to be a driver for other types of leukemia[199]. This pathway also is the main target for the BCR-ABL1 fusion gene product expressed by the Philadelphia chromosome [200]. In healthy cells, the JAK-STAT pathway is a primary target for cytokine signaling, which suggests a possible area for further research to identify whether the interplay of these two up-regulated pathways is involved in both the similar expression profile with Philadelphia- chromosome positive leukemia and the progression of the disease [201, 166].

The analysis of Ph-like differential gene expression showed clear evidence of significant enrichment of the *JAK-STAT* signaling pathway. This finding is consistent with prior literature citing the activation of this cytokine activation pathway through the similar outcome of distinct mutational patterns in non BCR-ABL1 positive samples [202, 166]. The implication of this result is indicative in that even with respect to each of the other subtypes studied in this analysis, treatment of the Ph-like subtype may respond better with inhibitors and modulators of *JAK-STAT* signaling. In addition, there were several pairs of antisense lncRNAs and sense, coding counterparts identified by differential gene expression that would indicate a possible relationship between the structural changes to the chromatin resulting from somatic rearrangements and copy-number changes, and the expression of mRNA and regulatory RNAs. One avenue for future investigation would be to further characterize these antisense RNAs in terms of their coexpression, predicted and molecular interactions, and impact on chromatin state in order to validate function. The model with the highest prediction accuracy was found to be the grouped elastic net regularized multinomial model with alpha of 0.1, which had a mean square error of 0.0603. This model was found to have 330 non-zero coefficients, which made use of 97 lncRNAs in the classification of leukemia genomic rearrangement subtypes. Only 13 of these lncRNAs were identified by differential expression analysis, which suggests the possibility of other contributions from the relationships of lncRNA and coding genes located within the same topologically associated domain. Furthermore, the large-scale variations in the genomic landscape of patients with acute lymphoblastic leukemia may results in architectural differences of their chromosomes which can be exploited to classify samples based on a reduced transcriptomic analysis alone. Work is still to be done to further increase the accuracy of our model, with a priori knowledge of the specific domains afflicted for each recurrent aberration class, for example, as well as improved ability to specifically discriminate between Ph-like samples and Ph(+). However, chromatin topology guided transcriptomic analysis may prove to be a more useful way to delineate between cancer samples for which there known sets of recurrent genomic rearrangements.

4.4 Conclusion

The Philadelphia-like subtype of acute lymphoblastic leukemia is a rare hematological cancer with poor prognosis relative to the much more common form of BCR-ABL1 fusion gene expression. It remains unclear as to which specific mechanisms drive the similar global expression patterns between the two subtypes, however with a more defined chromatin topology there may be hope to tease out event of enhancer hijacking that have been reported in various other forms of cancer. Here, we showed that analysis of both lncRNAs and mRNA expression that colocalize within TADs enables a high degree of accuracy in expression-mediated classification of pediatric B-cell acute lymphoblastic leukemia. With more data, such as greater sample sizes, epigenetic chromatin marks and sample-specific chromatin topology, there may be further improvement in expression-based diagnosis of patients. These results show promise for the development of a quick, cost effective gene panel for rapid diagnosis of ALL patients based on the expression of genes with consideration of chromatin topology. Ph-like leukemia has been traditionally difficult to diagnose stemming from the occurrence of a wide variety of somatic mutations leading to the archetypal expression pattern.

In this study we looked at overall differential expression patterns and found a significant amount of variation across all subtypes of the disease. These findings may suggest that there a multitude of possible genomic aberrations that contribute to ALL progression and poor outcomes beyond just the cytogenetic phenotyping. With more deep high-throughput sequencing of ALL samples, we may be able to fine map the structure of the chromatin as it has been undoubtedly been altered by large, recurrent translocations. What is clear, is that the *JAK-STAT* metabolic signalling pathway shows significant enrichment in the Ph-like samples relative to all other subtypes. There is likely a link to be found between the somatic copy number variation in the patients and the overexpression of this signaling cascade. Moreover, as more evidence continually identifies lncRNAs as key regulators of protein coding targets, chromatin state modeling will become increasingly more important in order to identify the overall functional significance of SVs in cancer as modification to the genes as well as their regulatory sequences.

Given that the classification scheme of hematological cancers is based primarily on chromosomal rearrangement events and is cell-type specific, it may not be surprising to find differences in gene expression among the different subtypes of precursor B-cell acute lymphoblastic leukemia [203, 204]. What may be more interesting is finding particular pathways that define a rearrangement-based subtype by means of consistent perturbation relative to other forms of the same cell-specific disease [205]. By further defining a given disease by the biochemical pathways that are consistently up or down regulated, an avenue for precise therapeutic intervention may be found. It was the aim of this project to define the Ph-like precursor B-cell ALL in this way, and doing so identified verifiable enrichment in pathways involved in cellular growth and cytoskeletal motility. The particular pathways involved may indicate an interplay between increased cytokine activity and kinase activation resulting in rapid proliferation of immature lymphocytes [206]. These results may indicate contributing factors to the overall poorer prognosis of this rare subtype relative to others and may be an avenue for further research focus. [35]

CHAPTER 5: CONCLUSIONS

Through the development of a framework to perform chromatin topology guided transcriptomic analysis, this work aimed to contextualize the effects of structural variation in the human genome in terms of the impact to topologically associated domains and genetic elements within these architectural regulatory modules. This type of multi-omics analysis will become more common as the cost of sequencing continues to become less expensive and the chromatin conformation technology becomes more accurate. Chromatin conformation capture methods have already shown the reliability and reproducibility of their results and ability to resolve the complex spatial relationships of chromatin and the long-range physical interactions of specific loci [207, 208, 132]. By pairing this data with RNA-seq measurements, one can assess the activity levels of genes and their locations within topologically associated domains [209]. Research has already demonstrated that genomic structural variants can cause neo-TAD formation, enhancer adoption by non-cognate target genes, TAD fusion or TAD shuffling in several human pathologies. Analogous work is needed to analyze these types of effects that lead to the natural variation in gene activity in healthy individuals. Furthermore, the emerging importance of lncRNA regulatory capacity of epigenetic state and recruitment of trancscriptional enzymes to adjacent, antisense promoters necessitates a comprehensive analysis of their interactions with the genes in a common physical space.

Acute lymphoblastic leukemia provides a set of recurrent chromosomal translocations and genomic rearrangements, which would suggest a high likelihood for the disruption of normal TAD boundaries. The direct effects of these rearrangements are mensurable by the transcriptional differences and enriched pathways across different subtypes. A detailed analysis of the Ph-like subtype compared to Ph(+) with the BCR-ABL1 fusion gene may lead to insights into the observed similarity of the respective global expression patterns, even without the highly active fusion protein kinase. This fusion protein has no naturally occurring down-regulation mechanism and leads to aberrant cell signalling in the tumor cell promoting cancer progression, while the Ph-like samples expression pattern is less clear.

This study aims to provide a comprehensive evaluation of the functional effects structural variation in the human genome. Through the use of a highly-validated set of insertions, deletions, duplications, and inversions from the HGSV consortium and RNA-seq data, Hi-C contacts and GM12878 markers of epigenetic regulation, SVs will be interrogated and statistically evaluated for allele-specific effects, observed expression variation and consequence to chromatin topology. Additionally, RNA-seq based transcriptome assembly will provide a set of lncRNAs for the nine samples in this dataset that can be used to predicted cis-interactions with the promoters, miRNA sequestration, as well as direct mRNA binding for genes located within the same TAD.

Additionally, the acute lymphoblastic leukemia dataset provides SNP array files for multiple genomic rearrangement subtypes of the disease, which can be used to infer copy number variation (CNV) for each sample. CNVs may be in variable locations of the genome and create novel TAD formation and abnormal chromatin contacts in tumor cells and drive aberrant expression and promote cancer phenotypes. These differences may be captured by specific patterns of gene expression for common sets of genes that are found within shared TADs defined by the GM12878 lymphoblast cell line as a reference. Prior analyses have used analogous cell line TADs as a reference for comparison to cancer cell lines, yet no one has yet studied the effects of SVs in this manner for acute lymphoblastic leukemia or across any of the recurrent translocation subtypes of hematological cancers.

As a matter of investigating the impact of structural variants on genomic function, it readily seems apparent that narrowing the focus to particular regulatory domains provided the key to identifying enhancer-mediated allele-specific expression. This indirect association between variants and their epigenetic regulatory sequences is a difficult task. chromatin capture technologies allow us to look beyond the genome at the two-dimensional level. Rather, topologically associated domains provide a lens with which to view chromatin contacts and the encapsulation of enhancers with target genes. With the low resolution Hi-C datasets currently available, it is immediately worth mentioning promising work that is currently being done into applying deep-learning methods to boost the resolution of contact matrices [210]. This would circumvent the need for costly deep-sequencing libraries, and allow for the more rapid analysis of currently available data. Even with that in mind, it seems quite clear that the regulatory capacity of TADs appears to be robust across evolution, yet the overall distribution of epigenetic regulatory elements encapsulated within TAD boundaries is modified by the germline and somatic structural variants that in turn affect expression phenotypes.

REFERENCES

[1] P. H. Sudmant, T. Rausch, E. J. Gardner, R. E. Handsaker, A. Abyzov, J. Huddleston, Y. Zhang, K. Ye, G. Jun, M. Hsi-Yang Fritz, M. K. Konkel, A. Malhotra, A. M. Stütz, X. Shi, F. Paolo Casale, J. Chen, F. Hormozdiari, G. Dayama, K. Chen, M. Malig, M. J. P. Chaisson, K. Walter, S. Meiers, S. Kashin, E. Garrison, A. Auton, H. Y. K. Lam, X. Jasmine Mu, C. Alkan, D. Antaki, T. Bae, E. Cerveira, P. Chines, Z. Chong, L. Clarke, E. Dal, L. Ding, S. Emery, X. Fan, M. Gujral, F. Kahveci, J. M. Kidd, Y. Kong, E.-W. Lameijer, S. McCarthy, P. Flicek, R. A. Gibbs, G. Marth, C. E. Mason, A. Menelaou, D. M. Muzny, B. J. Nelson, A. Noor, N. F. Parrish, M. Pendleton, A. Quitadamo, B. Raeder, E. E. Schadt, M. Romanovitch, A. Schlattl, R. Sebra, A. A. Shabalin, A. Untergasser, J. A. Walker, M. Wang, F. Yu, C. Zhang, J. Zhang, X. Zheng-Bradley, W. Zhou, T. Zichner, J. Sebat, M. A. Batzer, S. A. McCarroll, R. E. Mills, M. B. Gerstein, A. Bashir, O. Stegle, S. E. Devine, C. Lee, E. E. Eichler, J. O. Korbel, and J. O. Korbel, "An integrated map of structural variation in 2,504 human genomes," *Nature*, vol. 526, pp. 75–81, oct 2015.

[2] R. E. Mills, K. Walter, C. Stewart, R. E. Handsaker, K. Chen, C. Alkan, A. Abyzov, S. C. Yoon, K. Ye, R. K. Cheetham, A. Chinwalla, D. F. Conrad, Y. Fu, F. Grubert, I. Hajirasouliha, F. Hormozdiari, L. M. Iakoucheva, Z. Iqbal, S. Kang, J. M. Kidd, M. K. Konkel, J. Korn, E. Khurana, D. Kural, H. Y. K. Lam, J. Leng, R. Li, Y. Li, C.-Y. Lin, R. Luo, X. J. Mu, J. Nemesh, H. E. Peckham, T. Rausch, A. Scally, X. Shi, M. P. Stromberg, A. M. Stütz, A. E. Urban, J. A. Walker, J. Wu, Y. Zhang, Z. D. Zhang, M. A. Batzer, L. Ding, G. T. Marth, G. McVean, J. Sebat, M. Snyder, J. Wang, K. Ye, E. E. Eichler, M. B. Gerstein, M. E. Hurles, C. Lee, S. A. McCarroll, J. O. Korbel, and . G. 1000 Genomes Project, "Mapping copy number variation by population-scale genome sequencing.," *Nature*, vol. 470, pp. 59–65, feb 2011.

[3] D. F. Conrad, D. Pinto, R. Redon, L. Feuk, O. Gokcumen, Y. Zhang, J. Aerts, T. D. Andrews, C. Barnes, P. Campbell, T. Fitzgerald, M. Hu, C. H. Ihm, K. Kristiansson, D. G. MacArthur, J. R. MacDonald, I. Onyiah, A. W. C. Pang, S. Robson, K. Stirrups, A. Valsesia, K. Walter, J. Wei, C. Tyler-Smith, N. P. Carter, C. Lee, S. W. Scherer, and M. E. Hurles, "Origins and functional impact of copy number variation in the human genome," *Nature*, vol. 464, pp. 704–712, apr 2010.

[4] B. Weckselblatt and M. K. Rudd, "Human Structural Variation: Mechanisms of Chromosome Rearrangements.," *Trends in genetics : TIG*, vol. 31, pp. 587–599, oct 2015.

[5] F. Zhang, M. Khajavi, A. M. Connolly, C. F. Towne, S. D. Batish, and J. R. Lupski, "The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans," *Nature Genetics*, vol. 41, pp. 849–853, jul 2009.

[6] C. Alkan, B. P. Coe, and E. E. Eichler, "Genome structural variation discovery and genotyping.," *Nature reviews. Genetics*, vol. 12, pp. 363–76, may 2011.

[7] F. Zhang, W. Gu, M. E. Hurles, and J. R. Lupski, "Copy number variation in human health, disease, and evolution.," *Annual review of genomics and human genetics*, vol. 10, pp. 451–81, 2009.

[8] J. R. Lupski, "Genomic rearrangements and sporadic disease," *Nature Genetics*, vol. 39, pp. S43–S47, jul 2007.

[9] A. Shlien and D. Malkin, "Copy number variations and cancer.," *Genome medicine*, vol. 1, p. 62, jun 2009.

[10] C. G. Mullighan, S. Goorha, I. Radtke, C. B. Miller, E. Coustan-Smith, J. D. Dalton, K. Girtman, S. Mathew, J. Ma, S. B. Pounds, X. Su, C.-H. Pui, M. V. Relling, W. E. Evans, S. A. Shurtleff, and J. R. Downing, "Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia," *Nature*, vol. 446, pp. 758–764, apr 2007.

[11] O. T. Avery, C. M. Macleod, and M. McCarty, "Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types: Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated From Pneumococcus Type III," *The Journal of experimental medicine*, vol. 79, pp. 137–58, feb 1944.

[12] E. Tronick and R. G. Hunter, "Waddington, Dynamic Systems, and Epigenetics," *Frontiers in Behavioral Neuroscience*, vol. 10, p. 107, jun 2016.

[13] C. H. Waddington, *Organizers and Genes.* Cambridge, UK: The Cambridge University Press, 1940.

[14] J. M. W. Slack, "Conrad Hal Waddington: the last Renaissance biologist?," *Nature Reviews Genetics*, vol. 3, pp. 889–895, nov 2002.

[15] C. H. Waddington, "The Epigenotype. 1942.," *International Journal of Epidemiology*, vol. 41, pp. 10–13, feb 2012.

[16] C. H. Waddington, "Canalization of Development and the Inheritance of Acquired Characters," *Nature*, vol. 150, pp. 563–565, nov 1942.

[17] G. Felsenfeld, "A Brief History of Epigenetics," *Cold Spring Harbor Perspectives in Biology*, vol. 6, jan 2014.

[18] S. L. Berger, T. Kouzarides, R. Shiekhattar, and A. Shilatifard, "An operational definition of epigenetics.," *Genes & development*, vol. 23, pp. 781–3, apr 2009.

[19] T. E. P. Consortium, "An integrated encyclopedia of DNA elements in the human genome," *Nature*, vol. 489, pp. 57–74, sep 2012.

[20] Roadmap Epigenomics Consortium, A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M. J. Ziller, V. Amin, J. W. Whitaker, M. D. Schultz, L. D. Ward, A. Sarkar, G. Quon, R. S. Sandstrom, M. L. Eaton, Y.-C. C. Wu, A. R. Pfenning, X. Wang, M. Claussnitzer, Y. Liu, C. Coarfa, R. A. Harris, N. Shoresh, C. B. Epstein, E. Gjoneska, D. Leung, W. Xie, R. D. Hawkins, R. Lister, C. Hong, P. Gascard, A. J. Mungall, R. Moore, E. Chuah, A. Tam, T. K. Canfield, R. S. Hansen, R. Kaul, P. J. Sabo, M. S. Bansal, A. Carles, J. R. Dixon, K.-H. H. Farh, S. Feizi, R. Karlic, A.-R. R. Kim, A. Kulkarni, D. Li, R. Lowdon, G. Elliott, T. R. Mercer, S. J. Neph, V. Onuchic, P. Polak, N. Rajagopal, P. Ray, R. C. Sallari, K. T. Siebenthall, N. A. Sinnott-Armstrong, M. Stevens, R. E. Thurman, J. Wu, B. Zhang, X. Zhou, A. E. Beaudet, L. A. Boyer, P. L. De Jager, P. J. Farnham, S. J. Fisher, D. Haussler, S. J. M. Jones, W. Li, M. A. Marra, M. T. McManus, S. Sunyaev, J. A. Thomson, T. D. Tlsty, L.-H. H. Tsai, W. Wang, R. A. Waterland, M. Q. Zhang, L. H. Chadwick, B. E. Bernstein, J. F. Costello, J. R. Ecker, M. Hirst, A. Meissner, A. Milosavljevic, B. Ren, J. A. Stamatoyannopoulos, T. Wang, M. Kellis, C. Roadmap Epigenomics, A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M. J. Ziller, V. Amin, J. W. Whitaker, M. D. Schultz, L. D. Ward, A. Sarkar, G. Quon, R. S. Sandstrom, M. L. Eaton, Y.-C. C. Wu, A. R. Pfenning, X. Wang, M. Claussnitzer, Y. Liu, C. Coarfa, R. A. Harris, N. Shoresh, C. B. Epstein, E. Gjoneska, D. Leung, W. Xie, R. D. Hawkins, R. Lister, C. Hong, P. Gascard, A. J. Mungall, R. Moore, E. Chuah, A. Tam, T. K. Canfield, R. S. Hansen, R. Kaul, P. J. Sabo, M. S. Bansal, A. Carles, J. R. Dixon, K.-H. H. Farh, S. Feizi, R. Karlic, A.-R. R. Kim, A. Kulkarni, D. Li, R. Lowdon, G. Elliott, T. R. Mercer, S. J. Neph, V. Onuchic, P. Polak, N. Rajagopal, P. Ray, R. C. Sallari, K. T. Siebenthall, N. A. Sinnott-Armstrong, M. Stevens, R. E. Thurman, J. Wu, B. Zhang, X. Zhou, A. E. Beaudet, L. A. Boyer, P. L. De Jager, P. J. Farnham, S. J. Fisher, D. Haussler, S. J. M. Jones, W. Li, M. A. Marra, M. T. McManus, S. Sunyaev, J. A. Thomson, T. D. Tlsty, L.-H. H. Tsai, W. Wang, R. A. Waterland, M. Q. Zhang, L. H. Chadwick, B. E. Bernstein, J. F. Costello, J. R. Ecker, M. Hirst, A. Meissner, A. Milosavljevic, B. Ren, J. A. Stamatoyannopoulos, T. Wang, and M. Kellis, "Integrative analysis of 111 reference human epigenomes," *Nature*, vol. 518, no. 7539, pp. 317–329, 2015.

[21] J. Ernst, P. Kheradpour, T. S. Mikkelsen, N. Shoresh, L. D. Ward, C. B. Epstein, X. Zhang, L. Wang, R. Issner, M. Coyne, M. Ku, T. Durham, M. Kellis, and B. E. Bernstein, "Mapping and analysis of chromatin state dynamics in nine human cell types," *Nature*, vol. 473, pp. 43–49, may 2011.

[22] C. D. Allis and T. Jenuwein, "The molecular hallmarks of epigenetic control," *Nature Reviews Genetics*, vol. 17, pp. 487–500, aug 2016.

[23] A. Groth, A. Corpet, A. J. L. Cook, D. Roche, J. Bartek, J. Lukas, and G. Almouzni, "Regulation of Replication Fork Progression Through Histone Supply

and Demand," *Science*, vol. 318, pp. 1928 LP – 1931, dec 2007.

[24] A. J. Bannister, P. Zegerman, J. F. Partridge, E. A. Miska, J. O. Thomas, R. C. Allshire, and T. Kouzarides, "Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain," *Nature*, vol. 410, pp. 120–124, mar 2001.

[25] B. Papp and J. Müller, "Histone trimethylation and the maintenance of transcriptional ON and OFF states by trxG and PcG proteins.," *Genes & development*, vol. 20, pp. 2041–54, aug 2006.

[26] J. Müller, C. M. Hart, N. J. Francis, M. L. Vargas, A. Sengupta, B. Wild, E. L. Miller, M. B. O'Connor, R. E. Kingston, and J. A. Simon, "Histone Methyltransferase Activity of a Drosophila Polycomb Group Repressor Complex," *Cell*, vol. 111, pp. 197–208, oct 2002.

[27] L. Morey and K. Helin, "Polycomb group protein-mediated repression of transcription," *Trends Biochem. Sci*, vol. 35, 2010.

[28] J. L. Rinn, M. Kertesz, J. K. Wang, S. L. Squazzo, X. Xu, S. A. Brugmann, L. H. Goodnough, J. A. Helms, P. J. Farnham, E. Segal, and H. Y. Chang, "Functional Demarcation of Active and Silent Chromatin Domains in Human HOX Loci by Noncoding RNAs," *Cell*, vol. 129, pp. 1311–1323, 2007.

[29] H. Zhang, A. Christoforou, L. Aravind, S. W. Emmons, S. van den Heuvel, and D. A. Haber, "The C. elegans Polycomb gene SOP-2 encodes an RNA binding protein.," *Molecular cell*, vol. 14, pp. 841–7, jun 2004.

[30] E. Bernstein, E. M. Duncan, O. Masui, J. Gil, E. Heard, and C. D. Allis, "Mouse polycomb proteins bind differentially to methylated histone H3 and RNA and are enriched in facultative heterochromatin.," *Molecular and cellular biology*, vol. 26, pp. 2560–9, apr 2006.

[31] B. Schuettengruber, D. Chourrout, M. Vervoort, B. Leblanc, and G. Cavalli, "Genome Regulation by Polycomb and Trithorax Proteins," *Cell*, vol. 128, pp. 735–745, feb 2007.

[32] H. Muller, "Types of visible variations induced by X-rays in Drosophila," *J.Genet.*, vol. 22, no. 3, pp. 299—-334, 1930.

[33] B. McClintock, "The origin and behavior of mutable loci in maize.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 36, pp. 344–55, jun 1950.

[34] J. Weischenfeldt, T. Dubash, A. P. Drainas, B. R. Mardin, Y. Chen, A. M. Stütz, S. M. Waszak, G. Bosco, A. R. Halvorsen, B. Raeder, T. Efthymiopoulos, S. Erkek, C. Siegl, H. Brenner, O. T. Brustugun, S. M. Dieter, P. A. Northcott, I. Petersen, S. M. Pfister, M. Schneider, S. K. Solberg, E. Thunissen, W. Weichert, T. Zichner, R. Thomas, M. Peifer, A. Helland, C. R. Ball, M. Jechlinger,

R. Sotillo, H. Glimm, and J. O. Korbel, “Pan-cancer analysis of somatic copy-number alterations implicates IRS4 and IGF2 in enhancer hijacking,” *Nature Genetics*, vol. 49, p. 65, nov 2016.

[35] “TCGA Research Network.”

[36] M. Jagannathan-Bogdan and L. I. Zon, “Hematopoiesis.,” *Development (Cambridge, England)*, vol. 140, pp. 2463–7, jun 2013.

[37] A. Cumano and I. Godin, “Ontogeny of the Hematopoietic System,” *Annual Review of Immunology*, vol. 25, pp. 745–785, apr 2007.

[38] S. H. Orkin and L. I. Zon, “Hematopoiesis: An Evolving Paradigm for Stem Cell Biology,” *Cell*, vol. 132, pp. 631–644, feb 2008.

[39] R. P. DeKoter and H. Singh, “Regulation of B Lymphocyte and Macrophage Development by Graded Expression of PU.1,” *Science*, vol. 288, pp. 1439–1441, may 2000.

[40] S. L. Nutt and B. L. Kee, “The transcriptional regulation of B cell lineage commitment.,” *Immunity*, vol. 26, pp. 715–25, jun 2007.

[41] C. Xiao, D. P. Calado, G. Galler, T.-H. Thai, H. C. Patterson, J. Wang, N. Rajewsky, T. P. Bender, and K. Rajewsky, “MiR-150 controls B cell differentiation by targeting the transcription factor c-Myb.,” *Cell*, vol. 131, pp. 146–59, oct 2007.

[42] S. P. Fahl, R. B. Crittenden, D. Allman, and T. P. Bender, “c-Myb is required for pro-B cell differentiation.,” *Journal of immunology (Baltimore, Md. : 1950)*, vol. 183, pp. 5582–92, nov 2009.

[43] M. D. Thomas, C. S. Kremer, K. S. Ravichandran, K. Rajewsky, and T. P. Bender, “c-Myb is critical for B cell development and maintenance of follicular B cells.,” *Immunity*, vol. 23, pp. 275–86, sep 2005.

[44] T. Terwilliger and M. Abdul-Hay, “Acute lymphoblastic leukemia: a comprehensive review and 2017 update,” *Blood Cancer Journal*, vol. 7, p. e577, 2017.

[45] M. Huarte, “The emerging role of lncRNAs in cancer,” *Nature Medicine*, vol. 21, pp. 1253–1261, nov 2015.

[46] U. Klein, M. Lia, M. Crespo, R. Siegel, Q. Shen, T. Mo, A. Ambesi-Impiombato, A. Califano, A. Migliazza, G. Bhagat, and R. Dalla-Favera, “The DLEU2/miR-15a/16-1 cluster controls B cell proliferation and its deletion leads to chronic lymphocytic leukemia.,” *Cancer cell*, vol. 17, pp. 28–40, jan 2010.

[47] E. Yildirim, J. E. Kirby, D. E. Brown, F. E. Mercier, R. I. Sadreyev, D. T. Scadden, and J. T. Lee, “Xist RNA is a potent suppressor of hematologic cancer in mice.,” *Cell*, vol. 152, pp. 727–42, feb 2013.

[48] D. Lara-Astiaso, A. Weiner, E. Lorenzo-Vivas, I. Zaretsky, D. A. Jaitin, E. David, H. Keren-Shaul, A. Mildner, D. Winter, S. Jung, N. Friedman, and I. Amit, "Chromatin state dynamics during blood formation.," *Science (New York, N.Y.)*, vol. 345, pp. 943–9, aug 2014.

[49] G. Egger, G. Liang, A. Aparicio, and P. A. Jones, "Epigenetics in human disease and prospects for epigenetic therapy," *Nature*, vol. 429, pp. 457–463, may 2004.

[50] A. H. Shih, O. Abdel-Wahab, J. P. Patel, and R. L. Levine, "The role of mutations in epigenetic regulators in myeloid malignancies," *Nature Reviews Cancer*, vol. 12, pp. 599–612, sep 2012.

[51] R. Claus and M. Lübbert, "Epigenetic targets in hematopoietic malignancies," *Oncogene*, vol. 22, pp. 6489–6496, sep 2003.

[52] L. D. Croce, "Chromatin modifying activity of leukaemia associated fusion proteins," *Human Molecular Genetics*, vol. 14, no. 1, 2005.

[53] D. Smedley, M. Schubach, J. O. B. Jacobsen, S. Köhler, T. Zemojtel, M. Spielmann, M. Jäger, H. Hochheiser, N. L. Washington, J. A. McMurry, M. A. Haendel, C. J. Mungall, S. E. Lewis, T. Groza, G. Valentini, and P. N. Robinson, "A Whole-Genome Analysis Framework for Effective Identification of Pathogenic Regulatory Variants in Mendelian Disease.," *American journal of human genetics*, vol. 99, pp. 595–606, sep 2016.

[54] P. Makrythanasis and S. Antonarakis, "Pathogenic variants in non-protein-coding sequences," *Clinical Genetics*, vol. 84, pp. 422–428, nov 2013.

[55] A. Rivera-Reyes, K. E. Hayer, and C. H. Bassing, "Genomic Alterations of Non-Coding Regions Underlie Human Cancer: Lessons from T-ALL.," *Trends in molecular medicine*, vol. 22, pp. 1035–1046, dec 2016.

[56] M. Kellis, B. Wold, M. P. Snyder, B. E. Bernstein, A. Kundaje, G. K. Marinov, L. D. Ward, E. Birney, G. E. Crawford, J. Dekker, I. Dunham, L. L. Elnitski, P. J. Farnham, E. A. Feingold, M. Gerstein, M. C. Giddings, D. M. Gilbert, T. R. Gingeras, E. D. Green, R. Guigo, T. Hubbard, J. Kent, J. D. Lieb, R. M. Myers, M. J. Pazin, B. Ren, J. A. Stamatoyannopoulos, Z. Weng, K. P. White, and R. C. Hardison, "Defining functional DNA elements in the human genome.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 111, pp. 6131–8, apr 2014.

[57] J. S. Mattick, "Non-coding RNAs: the architects of eukaryotic complexity.," *EMBO reports*, vol. 2, pp. 986–91, nov 2001.

[58] 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, pp. 1760–74, sep 2012.

[59] R. A. A. Flynn and H. Y. Y. Chang, "Long Noncoding RNAs in Cell-Fate Programming and Reprogramming," *Cell Stem Cell*, vol. 14, no. 6, pp. 752–761, 2014.

[60] Y. Ma, W. Ma, L. Huang, D. Feng, and B. Cai, "Long non-coding RNAs, a new important regulator of cardiovascular physiology and pathology," *International Journal of Cardiology*, vol. 188, pp. 105–110, jun 2015.

[61] A. Rotini, E. Martínez-Sarrà, E. Pozzo, and M. Sampaolesi, "Interactions between microRNAs and long non-coding RNAs in cardiac development and repair," *Pharmacological Research*, vol. 127, pp. 58–66, jan 2018.

[62] P. Wu, X. Zuo, H. Deng, X. Liu, L. Liu, and A. Ji, "Roles of long noncoding RNAs in brain development, functional diversification and neurodegenerative diseases," *Brain Research Bulletin*, vol. 97, pp. 69–80, aug 2013.

[63] J. S. Mattick and J. L. Rinn, "Discovery and annotation of long noncoding RNAs," 2015.

[64] C. J. Brown, B. D. Hendrich, J. L. Rupert, R. G. Lafrenière, Y. Xing, J. Lawrence, and H. F. Willard, "The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus.," *Cell*, vol. 71, pp. 527–42, oct 1992.

[65] G. D. Penny, G. F. Kay, S. A. Sheardown, S. Rastan, and N. Brockdorff, "Requirement for Xist in X chromosome inactivation," *Nature*, vol. 379, pp. 131–137, jan 1996.

[66] J. Lee, L. Davidow, and D. Warshawsky, "Tsix, a gene antisense to Xist at the X-inactivation centre," *Nat. Genet*, vol. 21, 1999.

[67] A. Wutz, T. P. Rasmussen, and R. Jaenisch, "Chromosomal silencing and localization are mediated by different domains of Xist RNA," *Nature Genetics*, vol. 30, pp. 167–174, feb 2002.

[68] C. Chu, Q. C. C. Zhang, S. T. da Rocha, R. A. A. Flynn, M. Bharadwaj, J. M. M. Calabrese, T. Magnuson, E. Heard, H. Y. Y. Chang, S. T. Da Rocha, R. A. A. Flynn, M. Bharadwaj, J. M. M. Calabrese, T. Magnuson, E. Heard, and H. Y. Y. Chang, "Systematic discovery of Xist RNA binding proteins," *Cell*, vol. 161, pp. 404–416, apr 2015.

[69] J. Whitehead, G. K. Pandey, and C. Kanduri, "Regulation of the mammalian epigenome by long noncoding RNAs," *Biochimica et Biophysica Acta (BBA) - General Subjects*, vol. 1790, no. 9, pp. 936–947, 2009.

[70] N. Thakur, V. K. Tiwari, H. Thomassin, R. R. Pandey, M. Kanduri, A. Göndör, T. Grange, R. Ohlsson, and C. Kanduri, "An antisense RNA regulates the bidirectional silencing property of the Kcnq1 imprinting control region.," *Molecular and cellular biology*, vol. 24, pp. 7855–62, sep 2004.

[71] F. Mohammad, R. R. Pandey, T. Nagano, L. Chakalova, T. Mondal, P. Fraser, and C. Kanduri, "Kcnq1ot1/Lit1 noncoding RNA mediates transcriptional silencing by targeting to the perinucleolar region.," *Molecular and cellular biology*, vol. 28, pp. 3713–28, jun 2008.

[72] D. Umlauf, Y. Goto, R. Cao, F. Cerqueira, A. Wagschal, Y. Zhang, and R. Feil, "Imprinting along the Kcnq1 domain on mouse chromosome 7 involves repressive histone methylation and recruitment of Polycomb group complexes," *Nature Genetics*, vol. 36, pp. 1296–1300, dec 2004.

[73] P. K. Yang and M. I. Kuroda, "Noncoding RNAs and intranuclear positioning in monoallelic gene expression.," *Cell*, vol. 128, pp. 777–86, feb 2007.

[74] J. L. Rinn, "lncRNAs: Linking RNA to Chromatin," *Cold Spring Harbor Perspectives in Biology*, vol. 6, aug 2014.

[75] R. Gupta, N. Shah, K. Wang, J. Kim, and H. Horlings, "Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis," *Nature*, vol. 464, 2010.

[76] M.-C. Tsai, O. Manor, Y. Wan, N. Mosammaparast, J. K. Wang, F. Lan, Y. Shi, E. Segal, and H. Y. Chang, "Long noncoding RNA as modular scaffold of histone modification complexes.," *Science (New York, N.Y.)*, vol. 329, pp. 689–93, aug 2010.

[77] R. Lanz, N. McKenna, S. Onate, U. Albrecht, and J. Wong, "A steroid receptor coactivator, SRA, functions as an RNA and is present in an SRC-1 complex," *Cell*, vol. 97, 1999.

[78] R. B. Lanz, S. S. Chua, N. Barron, B. M. Söder, F. DeMayo, and B. W. O'Malley, "Steroid receptor RNA activator stimulates proliferation as well as apoptosis in vivo.," *Molecular and cellular biology*, vol. 23, pp. 7163–76, oct 2003.

[79] S. Katayama, Y. Tomaru, T. Kasukawa, K. Waki, M. Nakanishi, M. Nakamura, H. Nishida, C. C. Yap, M. Suzuki, J. Kawai, H. Suzuki, P. Carninci, Y. Hayashizaki, C. Wells, M. Frith, T. Ravasi, K. C. Pang, J. Hallinan, J. Mattick, D. A. Hume, L. Lipovich, S. Batalov, P. G. Engström, Y. Mizuno, M. A. Faghihi, A. Sandelin, A. M. Chalk, S. Mottagui-Tabar,

Z. Liang, B. Lenhard, C. Wahlestedt, RIKEN Genome Exploration Research Group, Genome Science Group (Genome Network Project Core Group), and FANTOM Consortium, "Antisense transcription in the mammalian transcriptome.," *Science (New York, N.Y.)*, vol. 309, pp. 1564–6, sep 2005.

[80] W. Yu, D. Gius, P. Onyango, K. Muldoon-Jacobs, J. Karp, A. P. Feinberg, and H. Cui, "Epigenetic silencing of tumour suppressor gene p15 by its antisense RNA," *Nature*, vol. 451, pp. 202–206, jan 2008.

[81] C. E. Burd, W. R. Jeck, Y. Liu, H. K. Sanoff, Z. Wang, and N. E. Sharpless, "Expression of Linear and Novel Circular Forms of an INK4/ARF-Associated Non-Coding RNA Correlates with Atherosclerosis Risk," *PLoS Genetics*, vol. 6, dec 2010.

[82] K. Li, Y. Blum, A. Verma, Z. Liu, K. Pramanik, N. R. Leigh, C. Z. Chun, G. V. Samant, B. Zhao, M. K. Garnaas, M. A. Horswill, S. A. Stanhope, P. E. North, R. Q. Miao, G. A. Wilkinson, M. Affolter, and R. Ramchandran, "A noncoding antisense RNA in tie-1 locus regulates tie-1 function in vivo.," *Blood*, vol. 115, pp. 133–9, jan 2010.

[83] O. Harismendy, D. Notani, X. Song, N. G. Rahim, B. Tanasa, N. Heintzman, B. Ren, X.-D. Fu, E. J. Topol, M. G. Rosenfeld, and K. A. Frazer, "9p21 DNA variants associated with coronary artery disease impair interferon-γ signalling response," *Nature*, vol. 470, pp. 264–268, feb 2011.

[84] C. Tufarelli, J. A. S. Stanley, D. Garrick, J. A. Sharpe, H. Ayyub, W. G. Wood, and D. R. Higgs, "Transcription of antisense RNA leading to gene silencing and methylation as a novel cause of human genetic disease," *Nature Genetics*, vol. 34, pp. 157–165, jun 2003.

[85] X. Zhu, Y.-B. Wu, J. Zhou, and D.-M. Kang, "Upregulation of lncRNA MEG3 promotes hepatic insulin resistance via increasing FoxO1 expression," *Biochemical and Biophysical Research Communications*, vol. 469, no. 2, pp. 319–325, 2016.

[86] H. Yan, J. Yuan, L. Gao, J. Rao, and J. Hu, "Long noncoding RNA MEG3 activation of p53 mediates ischemic neuronal death in stroke," *Neuroscience*, vol. 337, pp. 191–199, 2016.

[87] T. Mondal, S. Subhash, R. Vaid, S. Enroth, S. Uday, B. Reinius, S. Mitra, A. Mohammed, A. R. James, E. Hoberg, A. Moustakas, U. Gyllensten, S. J. M. Jones, C. M. Gustafsson, A. H. Sims, F. Westerlund, E. Gorab, and C. Kanduri, "MEG3 long noncoding RNA regulates the TGF-β pathway genes through formation of RNA-DNA triplex structures," *Nature Communications*, vol. 6, p. 7743, jul 2015.

[88] G. Felsenfeld, D. R. Davies, and A. Rich, "Formation of a Three-Stranded Polynucleotide Molecule," *Journal of the American Chemical Society*, vol. 79, pp. 2023–2024, apr 1957.

[89] R. Besch, C. Giovannangeli, C. Kammerbauer, and K. Degitz, "Specific inhibition of ICAM-1 expression mediated by gene targeting with Triplex-forming oligonucleotides.," *The Journal of biological chemistry*, vol. 277, pp. 32473–9, sep 2002.

[90] F. A. Buske, D. C. Bauer, J. S. Mattick, and T. L. Bailey, "Triplexator: detecting nucleic acid triple helices in genomic and transcriptomic data.," *Genome research*, vol. 22, pp. 1372–81, jul 2012.

[91] L. Lipovich, R. Johnson, and C. Lin, "MacroRNA underdogs in a microRNA world: evolutionary, regulatory, and biomedical significance of mammalian long non-protein-coding RNA," *Biochimica et Biophysica Acta*, vol. 1799, 2010.

[92] C. W. Greider and E. H. Blackburn, "A telomeric sequence in the RNA of Tetrahymena telomerase required for telomere repeat synthesis," *Nature*, vol. 337, pp. 331–337, jan 1989.

[93] E. Cusanelli and P. Chartrand, "Telomeric repeat-containing RNA TERRA: a noncoding RNA connecting telomere biology to genome integrity," *Frontiers in Genetics*, vol. 6, p. 143, apr 2015.

[94] Y. Xu and M. Komiyama, "Structure, function and targeting of human telomere RNA," *Methods*, vol. 57, pp. 100–105, may 2012.

[95] D. Oliva-Rico and L. A. Herrera, "Regulated expression of the lncRNA TERRA and its impact on telomere biology," *Mechanisms of Ageing and Development*, vol. 167, pp. 16–23, oct 2017.

[96] T. Hung, Y. Wang, M. F. Lin, A. K. Koegel, Y. Kotake, G. D. Grant, H. M. Horlings, N. Shah, C. Umbricht, P. Wang, Y. Wang, B. Kong, A. Langerød, A.-L. Børresen-Dale, S. K. Kim, M. van de Vijver, S. Sukumar, M. L. Whitfield, M. Kellis, Y. Xiong, D. J. Wong, and H. Y. Chang, "Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters," *Nature Genetics*, vol. 43, pp. 621–629, jul 2011.

[97] I. Martianov, A. Ramadass, A. Serra Barros, N. Chow, and A. Akoulitchev, "Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript," *Nature*, vol. 445, pp. 666–670, feb 2007.

[98] T. Kino, D. E. Hurt, T. Ichijo, N. Nader, and G. P. Chrousos, "Noncoding RNA Gas5 Is a Growth Arrest- and Starvation-Associated Repressor of the Glucocorticoid Receptor," *Science Signaling*, vol. 3, pp. ra8–ra8, feb 2010.

[99] T. Trimarchi, E. Bilal, P. Ntziachristos, G. Fabbri, R. Dalla-Favera, A. Tsirigos, and I. Aifantis, "Genome-wide Mapping and Characterization of Notch-Regulated Long Noncoding RNAs in Acute Leukemia," *Cell*, vol. 158, pp. 593–606, jul 2014.

[100] F. Ma, S.-h. Wang, Q. Cai, L.-y. Jin, D. Zhou, J. Ding, and Z.-w. Quan, "Long non-coding RNA TUG1 promotes cell proliferation and metastasis by negatively regulating miR-300 in gallbladder carcinoma," *Biomedicine & Pharmacotherapy*, vol. 88, pp. 863–869, 2017.

[101] S. Chen, M. Wang, H. Yang, L. Mao, Q. He, H. Jin, Z. ming Ye, X. ying Luo, Y. peng Xia, and B. Hu, "LncRNA TUG1 sponges microRNA-9 to promote neurons apoptosis by up-regulated Bcl2l11 under ischemia," *Biochemical and Biophysical Research Communications*, vol. 485, no. 1, pp. 167–173, 2017.

[102] C. Park, N. Yu, I. Choi, W. Kim, and S. Lee, "lncRNAtor: a comprehensive resource for functional investigation of long non-coding RNAs," *Bioinformatics*, vol. 30, pp. 2480–2485, sep 2014.

[103] P.-J. Volders, K. Helsens, X. Wang, B. Menten, L. Martens, K. Gevaert, J. Vandesompele, and P. Mestdagh, "LNCipedia: a database for annotated human lncRNA transcript sequences and structures," *Nucleic Acids Research*, vol. 41, pp. D246–D251, jan 2013.

[104] D. Bhartiya, K. Pal, S. Ghosh, S. Kapoor, S. Jalali, B. Panwar, S. Jain, S. Sati, S. Sengupta, C. Sachidanandan, G. P. S. Raghava, S. Sivasubbu, and V. Scaria, "lncRNome: a comprehensive knowledgebase of human long noncoding RNAs.," *Database : the journal of biological databases and curation*, vol. 2013, p. bat034, 2013.

[105] X. C. Quek, D. W. Thomson, J. L. V. Maag, N. Bartonicek, B. Signal, M. B. Clark, B. S. Gloss, and M. E. Dinger, "lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs.," *Nucleic acids research*, vol. 43, pp. D168–73, jan 2015.

[106] Affymetrix, "Affymetrix Data Sheets," 2007.

[107] C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn, and L. Pachter, "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.," *Nature protocols*, vol. 7, pp. 562–78, mar 2012.

[108] M. F. Lin, I. Jungreis, and M. Kellis, "PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions," *Bioinformatics*, vol. 27, pp. i275–i282, jul 2011.

[109] J. Iwakiri, M. Hamada, and K. Asai, "Bioinformatics tools for lncRNA research," *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, vol. 1859, no. 1, pp. 23–30, 2016.

[110] L. Sun, H. Liu, L. Zhang, and J. Meng, “lncRScan-SVM: A Tool for Predicting Long Non-Coding RNAs Using Support Vector Machine,” *PLOS ONE*, vol. 10, p. e0139654, oct 2015.

[111] R. D. Ponti, A. Armaos, S. Marti, and G. G. Tartaglia, “A method for RNA structure prediction shows evidence for structure in lncRNAs,” *bioRxiv*, p. 284869, apr 2018.

[112] J. Li, W. Ma, P. Zeng, J. Wang, B. Geng, J. Yang, and Q. Cui, “LncTar: a tool for predicting the RNA targets of long noncoding RNAs,” *Briefings in Bioinformatics*, vol. 16, pp. 806–812, sep 2015.

[113] R. Milo, P. Jorgensen, U. Moran, G. Weber, and M. Springer, “BioNumbers–the database of key numbers in molecular and cell biology.,” *Nucleic acids research*, vol. 38, pp. D750–3, jan 2010.

[114] D. E. Olins and A. L. Olins, “Chromatin history: our view from the bridge,” *Nature Reviews Molecular Cell Biology*, vol. 4, pp. 809–814, oct 2003.

[115] P. Ridgway and G. Almouzni, “Chromatin assembly and organization.,” *Journal of cell science*, vol. 114, no. Pt 15, pp. 2711–2712, 2001.

[116] J. Dekker, K. Rippe, M. Dekker, and N. Kleckner, “Capturing Chromosome Conformation,” *Science*, vol. 295, no. 5558, pp. 1306–1311, 2002.

[117] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker, “Comprehensive mapping of long-range interactions reveals folding principles of the human genome.,” *Science (New York, N.Y.)*, vol. 326, pp. 289–93, oct 2009.

[118] T. Cremer and M. Cremer, “Chromosome territories.,” *Cold Spring Harbor perspectives in biology*, vol. 2, no. 3, pp. 1–22, 2010.

[119] J. H. Gibcus and J. Dekker, “The Hierarchy of the 3D Genome,” *Molecular Cell*, vol. 49, pp. 773–782, mar 2013.

[120] 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, pp. 376–380, apr 2012.

[121] M. Vietri Rudan, C. Barrington, S. Henderson, C. Ernst, D. T. Odom, A. Tanay, and S. Hadjur, “Comparative Hi-C Reveals that CTCF Underlies Evolution of Chromosomal Domain Architecture,” *Cell Reports*, vol. 10, no. 8, pp. 1297–1309, 2015.

[122] J. R. Dixon, I. Jung, S. Selvaraj, Y. Shen, J. E. Antosiewicz-Bourget, A. Y. Lee, Z. Ye, A. Kim, N. Rajagopal, W. Xie, Y. Diao, J. Liang, H. Zhao, V. V. Lobanenkov, J. R. Ecker, J. A. Thomson, and B. Ren, "Chromatin architecture reorganization during stem cell differentiation," *Nature*, vol. 518, no. 7539, 2015.

[123] A. Goloborodko, J. F. Marko, and L. A. Mirny, "Chromosome Compaction by Active Loop Extrusion," *Biophysical Journal*, vol. 110, no. 10, pp. 2162–2168, 2016.

[124] G. Fudenberg, M. Imakaev, C. Lu, A. Goloborodko, N. Abdennur, and L. A. Mirny, "Formation of Chromosomal Domains by Loop Extrusion," *Cell Reports*, vol. 15, no. 9, pp. 2038–2049, 2016.

[125] E. de Wit, E. S. Vos, S. J. Holwerda, C. Valdes-Quezada, M. J. Verstegen, H. Teunissen, E. Splinter, P. J. Wijchers, P. H. Krijger, and W. de Laat, "CTCF Binding Polarity Determines Chromatin Looping," *Molecular Cell*, vol. 60, pp. 676–684, nov 2015.

[126] D. Hnisz, D. S. Day, and R. A. Young, "Insulated Neighborhoods: Structural and Functional Units of Mammalian Gene Control," *Cell*, vol. 167, pp. 1188–1200, nov 2016.

[127] S. S. Rao, S. C. Huang, B. Glenn St Hilaire, J. M. Engreitz, E. M. Perez, K. R. Kieffer-Kwon, A. L. Sanborn, S. E. Johnstone, G. D. Bascom, I. D. Bochkov, X. Huang, M. S. Shamim, J. Shin, D. Turner, Z. Ye, A. D. Omer, J. T. Robinson, T. Schlick, B. E. Bernstein, R. Casellas, E. S. Lander, and E. L. Aiden, "Cohesin Loss Eliminates All Loop Domains," *Cell*, vol. 171, no. 2, pp. 305–320.e24, 2017.

[128] J. Dowen, Z. Fan, D. Hnisz, G. Ren, B. Abraham, L. Zhang, A. Weintraub, J. Schuijers, T. Lee, K. Zhao, and R. Young, "Control of Cell Identity Genes Occurs in Insulated Neighborhoods in Mammalian Chromosomes," *Cell*, vol. 159, no. 2, pp. 374–387, 2014.

[129] A.-L. Valton and J. Dekker, "TAD disruption as oncogenic driver," *Current Opinion in Genetics & Development*, vol. 36, pp. 34–40, feb 2016.

[130] S. Gröschel, M. A. Sanders, R. Hoogenboezem, E. de Wit, B. A. Bouwman, C. Erpelinck, V. H. van der Velden, M. Havermans, R. Avellino, K. van Lom, E. J. Rombouts, M. van Duin, K. Döhner, H. B. Beverloo, J. E. Bradner, H. Döhner, B. Löwenberg, P. J. Valk, E. M. Bindels, W. de Laat, and R. Delwel, "A Single Oncogenic Enhancer Rearrangement Causes Concomitant EVI1 and GATA2 Deregulation in Leukemia," *Cell*, vol. 157, pp. 369–381, apr 2014.

[131] 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, no. 1, 2017.

[132] B. A. Walker, C. P. Wardell, A. Brioli, E. Boyle, M. F. Kaiser, D. B. Begum, N. B. Dahir, D. C. Johnson, F. M. Ross, F. E. Davies, and G. J. Morgan, "Translocations at 8q24 juxtapose MYC with genes that harbor superenhancers resulting in overexpression and poor prognosis in myeloma patients," *Blood Cancer Journal*, vol. 4, pp. e191–e191, mar 2014.

[133] M. Spielmann, D. G. Lupiáñez, and S. Mundlos, "Structural variation in the 3D genome," *Nature Reviews Genetics*, vol. 7, pp. 1–15, apr 2018.

[134] M. Franke, D. M. Ibrahim, G. Andrey, W. Schwarzer, V. Heinrich, R. Schöpflin, K. Kraft, R. Kempfer, I. Jerković, W.-L. Chan, M. Spielmann, B. Timmermann, L. Wittler, I. Kurth, P. Cambiaso, O. Zuffardi, G. Houge, L. Lambie, F. Brancati, A. Pombo, M. Vingron, F. Spitz, and S. Mundlos, "Formation of new chromatin domains determines pathogenicity of genomic duplications," *Nature*, vol. 538, pp. 265–269, oct 2016.

[135] E. Giorgio, D. Robyr, M. Spielmann, E. Ferrero, E. Di Gregorio, D. Imperiale, G. Vaula, G. Stamoulis, F. Santoni, C. Atzori, L. Gasparini, D. Ferrera, C. Canale, M. Guipponi, L. A. Pennacchio, S. E. Antonarakis, A. Brussino, and A. Brusco, "A large genomic deletion leads to enhancer adoption by the lamin B1 gene: a second path to autosomal dominant adult-onset demyelinating leukodystrophy (ADLD).," *Human molecular genetics*, vol. 24, pp. 3143–54, jun 2015.

[136] E. Lander, L. Linton, B. Birren, C. Nusbaum, and M. Zody, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, 2001.

[137] M. J. P. Chaisson, R. K. Wilson, and E. E. Eichler, "Genetic variation and the de novo assembly of human genomes.," *Nature reviews. Genetics*, vol. 16, pp. 627–40, nov 2015.

[138] J. M. Kidd, T. Graves, T. L. Newman, R. Fulton, H. S. Hayden, M. Malig, J. Kallicki, R. Kaul, R. K. Wilson, and E. E. Eichler, "A Human Genome Structural Variation Sequencing Resource Reveals Insights into Mutational Mechanisms," *Cell*, vol. 143, pp. 837–847, nov 2010.

[139] D. Malhotra and J. Sebat, "CNVs: Harbingers of a Rare Variant Revolution in Psychiatric Genetics," *Cell*, vol. 148, pp. 1223–1241, mar 2012.

[140] M. J. P. Chaisson, A. D. Sanders, X. Zhao, A. Malhotra, D. Porubsky, T. Rausch, E. J. Gardner, O. Rodriguez, L. Guo, R. L. Collins, X. Fan, J. Wen, R. E. Handsaker, S. Fairley, Z. N. Kronenberg, X. Kong, F. Hormozdiari, D. Lee, A. M. Wenger, A. Hastie, D. Antaki, P. Audano, H. Brand, S. Cantsilieris, H. Cao, E. Cerveira, C. Chen, X. Chen, C.-S. Chin, Z. Chong, N. T. Chuang, D. M. Church, L. Clarke, A. Farrell, J. Flores, T. Galeev, G. David, M. Gujral, V. Guryev, W. Haynes-Heaton, J. Korlach, S. Kumar, J. Y. Kwon, J. E. Lee, J. Lee, W.-P. Lee, S. P. Lee, P. Marks, K. Valud-Martinez, S. Meiers, K. M.

Munson, F. Navarro, B. J. Nelson, C. Nodzak, A. Noor, S. Kyriazopoulou-Panagiotopoulou, A. Pang, Y. Qiu, G. Rosanio, M. Ryan, A. Stutz, D. C. J. Spierings, A. Ward, A. E. Welsch, M. Xiao, W. Xu, C. Zhang, Q. Zhu, X. Zheng-Bradley, G. Jun, L. Ding, C. â. L. Koh, B. Ren, P. Flicek, K. Chen, M. B. Gerstein, P.-Y. Kwok, P. M. Lansdorp, G. Marth, J. Sebat, X. Shi, A. Bashir, K. Ye, S. E. Devine, M. Talkowski, R. E. Mills, T. Marschall, J. Korbel, E. E. Eichler, and C. Lee, "Multi-platform discovery of haplotype-resolved structural variation in human genomes," *bioRxiv*, jan 2017.

[141] L. Harewood, F. Schütz, S. Boyle, P. Perry, M. Delorenzi, W. A. Bickmore, and A. Reymond, "The effect of translocation-induced nuclear reorganization on gene expression.," *Genome research*, vol. 20, pp. 554–64, may 2010.

[142] J. O. Korbel and P. J. Campbell, "Criteria for inference of chromothripsis in cancer genomes," *Cell*, vol. 152, no. 6, pp. 1226–1236, 2013.

[143] S. Chiaretti, G. Zini, and R. Bassan, "Diagnosis and subclassification of acute lymphoblastic leukemia.," *Mediterranean journal of hematology and infectious diseases*, vol. 6, no. 1, p. e2014073, 2014.

[144] L. B. Silverman, R. D. Gelber, V. K. Dalton, B. L. Asselin, R. D. Barr, L. A. Clavell, C. A. Hurwitz, A. Moghrabi, Y. Samson, M. A. Schorin, S. Arkin, L. Declerck, H. J. Cohen, and S. E. Sallan, "Improved outcome for children with acute lymphoblastic leukemia: results of Dana-Farber Consortium Protocol 91-01.," *Blood*, vol. 97, pp. 1211–8, mar 2001.

[145] C.-H. Pui and W. E. Evans, "Treatment of Acute Lymphoblastic Leukemia," *New England Journal of Medicine*, vol. 354, pp. 166–178, jan 2006.

[146] A. Noone, N. Howlader, M. Krapcho, D. Miller, A. Brest, M. Yu, J. Ruhl, Z. Tatalovich, A. Mariotto, D. Lewis, H. Chen, E. Feuer, and K. Cronin, "SEER Cancer Statistics Review, 1975-2015, https://seer.cancer.gov/csr/1975_2015/, based on November 2017 SEER data submission, posted to the SEER web site, April 2018.," tech. rep., National Cancer Institute, Bethesda, MD, 2018.

[147] C.-H. Pui, L. L. Robison, and A. T. Look, "Acute lymphoblastic leukaemia," *The Lancet*, vol. 371, pp. 1030–1043, mar 2008.

[148] D. A. Arber, A. Orazi, R. Hasserjian, J. Thiele, M. J. Borowitz, M. M. Le Beau, C. D. Bloomfield, M. Cazzola, and J. W. Vardiman, "The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia," *Blood*, vol. 127, pp. 2391 LP – 2405, may 2016.

[149] T. Raff and M. Brüggemann, "Leukemia-Initiating Cells in Acute Lymphoblastic Leukemia," in *Cancer Stem Cells* (V. K. Rajasekhar, ed.), ch. 12, pp. 161–170, John Wiley & Sons, Inc., 2014.

[150] J. Zhang, C. G. Mullighan, R. C. Harvey, G. Wu, X. Chen, M. Edmonson, K. H. Buetow, W. L. Carroll, I.-M. Chen, M. Devidas, D. S. Gerhard, M. L. Loh, G. H. Reaman, M. V. Relling, B. M. Camitta, W. P. Bowman, M. A. Smith, C. L. Willman, J. R. Downing, and S. P. Hunger, "Key pathways are frequently mutated in high-risk childhood acute lymphoblastic leukemia: a report from the Children's Oncology Group.," *Blood*, vol. 118, pp. 3080–7, sep 2011.

[151] J. Roman-Gomez, J. A. Castillejo, A. Jimenez, M. Barrios, A. Heiniger, and A. Torres, "The Role of DNA Hypermethylation in the Pathogenesis and Prognosis of Acute Lymphoblastic Leukemia," *Leukemia & Lymphoma*, vol. 44, pp. 1855–1864, jan 2003.

[152] J. Roman-Gomez, A. Jimenez-Velasco, M. Barrios, F. Prosper, A. Heiniger, A. Torres, and X. Agirre, "Poor prognosis in acute lymphoblastic leukemia may relate to promoter hypermethylation of cancer-related genes," *Leukemia & Lymphoma*, vol. 48, pp. 1269–1282, jan 2007.

[153] J. Starkova, B. Zamostna, E. Mejstrikova, R. Krejci, H. A. Drabkin, and J. Trka, "HOX gene expression in phenotypic and genotypic subgroups and low HOXA gene expression as an adverse prognostic factor in pediatric ALL," *Pediatric Blood & Cancer*, vol. 55, pp. 1072–1082, dec 2010.

[154] J. Soulier, E. Clappier, J.-M. Cayuela, A. Regnault, M. García-Peydró, H. Dombret, A. Baruchel, M.-L. Toribio, and F. Sigaux, "HOXA genes are included in genetic and biologic networks defining human acute T-cell leukemia (T-ALL).," *Blood*, vol. 106, pp. 274–86, jul 2005.

[155] W. Schulz, "Leukemias and Lymphomas," in *Molecular Biology of Human Cancers*, pp. 219–242, Dordrecht: Springer Netherlands, 2005.

[156] X. G. Thomas, "Rare Acute Leukemias," in *Rare Hematological Malignancies*, pp. 149–191, Boston, MA: Springer US, 2008.

[157] J. N. Nichol, S. Assouline, and W. H. Miller, "The Etiology of Acute Leukemia," in *Neoplastic Diseases of the Blood*, pp. 177–198, New York, NY: Springer New York, 2013.

[158] B. J. Bain, "Pathology of the Chronic Myeloid Leukemias," in *Neoplastic Diseases of the Blood*, pp. 19–28, New York, NY: Springer New York, 2013.

[159] S. H. Swerdlow, E. Campo, S. A. Pileri, N. L. Harris, H. Stein, R. Siebert, R. Advani, M. Ghielmini, G. A. Salles, A. D. Zelenetz, and E. S. Jaffe, "The 2016 revision of the World Health Organization classification of lymphoid neoplasms.," *Blood*, vol. 127, pp. 2375–90, may 2016.

[160] R. C. Ribeiro, M. Abromowitch, S. C. Raimondi, S. B. Murphy, F. Behm, and D. L. Williams, "Clinical and Biologic Hallmarks of the Philadelphia Chromosome in Childhood Acute Lymphoblastic Leukemia," *Blood*, vol. 70, no. 4, 1987.

[161] K. Kobayashi, N. Miyagawa, K. Mitsui, M. Matsuoka, Y. Kojima, H. Takahashi, K. Ootsubo, J. Nagai, H. Ueno, T. Ishibashi, S. Sultana, Y. Okada, S. Akimoto, H. Okita, K. Matsumoto, H. Goto, N. Kiyokawa, and A. Ohara, "TKI dasatinib monotherapy for a patient with Ph-like ALL bearing ATF7IP/PDGFRB translocation," *Pediatric Blood & Cancer*, vol. 62, pp. 1058–1060, jun 2015.

[162] G. A. Koretzky, "The legacy of the Philadelphia chromosome.," *The Journal of clinical investigation*, vol. 117, pp. 2030–2, aug 2007.

[163] D. Hnisz, A. S. Weintraub, D. S. Day, A.-L. Valton, R. O. Bak, C. H. Li, J. Goldmann, B. R. Lajoie, Z. P. Fan, A. A. Sigova, J. Reddy, D. Borges-Rivera, T. I. Lee, R. Jaenisch, M. H. Porteus, J. Dekker, and R. A. Young, "Activation of proto-oncogenes by disruption of chromosome neighborhoods," *Science*, mar 2016.

[164] M. L. Den Boer, M. van Slegtenhorst, R. X. De Menezes, M. H. Cheok, J. G. Buijs-Gladdines, S. T. Peters, L. J. Van Zutven, H. B. Beverloo, P. J. Van der Spek, G. Escherich, M. A. Horstmann, G. E. Janka-Schaub, W. A. Kamps, W. E. Evans, and R. Pieters, "A subtype of childhood acute lymphoblastic leukaemia with poor treatment outcome: a genome-wide classification study," *The Lancet Oncology*, vol. 10, pp. 125–134, feb 2009.

[165] H. Inaba, M. Greaves, and C. G. Mullighan, "Acute lymphoblastic leukaemia," *The Lancet*, vol. 381, no. 9881, pp. 1943–1955, 2013.

[166] W. Vainchenker and S. N. Constantinescu, "JAK/STAT signaling in hematological malignancies," *Oncogene*, vol. 32, pp. 2601–2613, may 2013.

[167] Y. Ofran and S. Izraeli, "BCR-ABL (Ph)-like acute leukemia Pathogenesis, diagnosis and therapeutic options," *Blood Reviews*, vol. 31, pp. 11–16, mar 2017.

[168] R. M. Durbin and et al., "A map of human genome variation from population-scale sequencing," *Nature*, vol. 467, pp. 1061–1073, oct 2010.

[169] T. . G. P. Consortium, "An integrated map of genetic variation from 1,092 human genomes," *Nature*, vol. 491, pp. 56–65, nov 2012.

[170] R. A. Gibbs, E. Boerwinkle, and T. . G. P. Consortium, "A global reference for human genetic variation," *Nature*, vol. 526, pp. 68–74, oct 2015.

[171] A. R. Quinlan and I. M. Hall, "BEDTools: A flexible suite of utilities for comparing genomic features," *Bioinformatics*, vol. 26, pp. 841–842, mar 2010.

[172] J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren, "Domaincall Software : http://bioinformatics-renlab.ucsd.edu/collaborations/sid/domaincall_software.zip," 2012.

[173] C. Brenner, “Homer Software and Data Download,” tech. rep., Salk Institute, UCSD, 2010.

[174] S. Heinz, C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh, and C. K. Glass, “Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities,” *Molecular Cell*, vol. 38, pp. 576–589, may 2010.

[175] P. Jaccard, “Etude comparative de la distribution florale dans une portion des Alpe’s et du Jura,” *Bulletin de la Societe Vaudoise des Sciences Naturelles*, vol. 37, no. 142, pp. 547–579, 1901.

[176] M. Levandowsky and D. Winter, “Distance between Sets,” *Nature*, vol. 234, pp. 34–35, nov 1971.

[177] R. Zaborowski and B. Wilczyński, “BPscore: An Effective Metric for Meaningful Comparisons of Structural Chromosome Segmentations,” *Journal of Computational Biology*, p. cmb.2018.0162, feb 2019.

[178] S. Fishilevich, R. Nudel, N. Rappaport, R. Hadar, I. Plaschkes, T. Iny Stein, N. Rosen, A. Kohn, M. Twik, M. Safran, D. Lancet, and D. Cohen, “GeneHancer: genome-wide integration of enhancers and target genes in GeneCards.,” *Database : the journal of biological databases and curation*, vol. 2017, 2017.

[179] D. Karolchik, A. S. Hinrichs, T. S. Furey, K. M. Roskin, C. W. Sugnet, D. Haussler, and W. J. Kent, “The UCSC Table Browser data retrieval tool,” *Nucleic Acids Research*, vol. 32, pp. 493D–496, jan 2004.

[180] S. Durinck, Y. Moreau, A. Kasprzyk, S. Davis, B. De Moor, A. Brazma, and W. Huber, “BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis,” *Bioinformatics*, vol. 21, pp. 3439–3440, aug 2005.

[181] S. Durinck, P. T. Spellman, E. Birney, and W. Huber, “Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt.,” *Nature protocols*, vol. 4, no. 8, pp. 1184–91, 2009.

[182] T. D. Wu, J. Reeder, M. Lawrence, G. Becker, and M. J. Brauer, “GMAP and GSNAP for Genomic Sequence Alignment: Enhancements to Speed, Accuracy, and Functionality,” in *Statistical Genomics*, pp. 283–334, Humana Press, New York, NY, 2016.

[183] “Picard: http://broadinstitute.github.io/picard. Broad Institute.”

[184] A. I. Petrov, S. J. E. Kay, I. Kalvari, K. L. Howe, K. A. Gray, E. A. Bruford, P. J. Kersey, G. Cochrane, R. D. Finn, A. Bateman, A. Kozomara, S. Griffiths-Jones, A. Frankish, C. W. Zwieb, B. Y. Lau, K. P. Williams, P. P. Chan,

T. M. Lowe, J. J. Cannone, R. Gutell, M. A. Machnicka, J. M. Bujnicki, M. Yoshihama, N. Kenmochi, B. Chai, J. R. Cole, M. Szymanski, W. M. Karlowski, V. Wood, E. Huala, T. Z. Berardini, Y. Zhao, R. Chen, W. Zhu, M. D. Paraskevopoulou, I. S. Vlachos, A. G. Hatzigeorgiou, L. Ma, Z. Zhang, J. Puetz, P. F. Stadler, D. McDonald, S. Basu, P. Fey, S. R. Engel, J. M. Cherry, P.-J. Volders, P. Mestdagh, J. Wower, M. B. Clark, X. C. Quek, and M. E. Dinger, "RNAcentral: a comprehensive database of non-coding RNA sequences," *Nucleic Acids Research*, vol. 45, pp. D128–D134, jan 2017.

[185] M. Hu, K. Deng, S. Selvaraj, Z. Qin, B. Ren, and J. S. Liu, "HiCNorm: removing biases in Hi-C data via Poisson regression," *Bioinformatics*, vol. 28, pp. 3131–3133, dec 2012.

[186] M. Forcato, C. Nicoletti, K. Pal, C. M. Livi, F. Ferrari, and S. Bicciato, "Comparison of computational methods for Hi-C data analysis," *Nature Methods*, vol. 14, p. 679, jun 2017.

[187] S. S. Rao, M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, and E. L. Aiden, "A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping," *Cell*, vol. 159, pp. 1665–1680, dec 2014.

[188] C. J. Ott, N. P. Blackledge, J. L. Kerschner, S.-H. Leir, G. E. Crawford, C. U. Cotton, and A. Harris, "Intronic enhancers coordinate epithelial-specific looping of the active CFTR locus.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, pp. 19934–9, nov 2009.

[189] C. G. Mullighan, X. Su, J. Zhang, I. Radtke, L. A. Phillips, C. B. Miller, J. Ma, W. Liu, C. Cheng, B. A. Schulman, R. C. Harvey, I.-M. Chen, R. J. Clifford, W. L. Carroll, G. Reaman, W. P. Bowman, M. Devidas, D. S. Gerhard, W. Yang, M. V. Relling, S. A. Shurtleff, D. Campana, M. J. Borowitz, C.-H. Pui, M. Smith, S. P. Hunger, C. L. Willman, J. R. Downing, and t. C. O. Group, "Deletion of IKZF1 and Prognosis in Acute Lymphoblastic Leukemia," *New England Journal of Medicine*, vol. 360, pp. 470–480, jan 2009.

[190] K. G. Roberts, Y. Li, D. Payne-Turner, R. C. Harvey, Y.-L. Yang, D. Pei, K. McCastlain, L. Ding, C. Lu, G. Song, J. Ma, J. Becksfort, M. Rusch, S.-C. Chen, J. Easton, J. Cheng, K. Boggs, N. Santiago-Morales, I. Iacobucci, R. S. Fulton, J. Wen, M. Valentine, C. Cheng, S. W. Paugh, M. Devidas, I.-M. Chen, S. Reshmi, A. Smith, E. Hedlund, P. Gupta, P. Nagahawatte, G. Wu, X. Chen, D. Yergeau, B. Vadodaria, H. Mulder, N. J. Winick, E. C. Larsen, W. L. Carroll, N. A. Heerema, A. J. Carroll, G. Grayson, S. K. Tasian, A. S. Moore, F. Keller, M. Frei-Jones, J. A. Whitlock, E. A. Raetz, D. L. White, T. P. Hughes, J. M. Guidry Auvil, M. A. Smith, G. Marcucci, C. D. Bloomfield, K. Mrózek, J. Kohlschmidt, W. Stock, S. M. Kornblau, M. Konopleva, E. Paietta, C.-H. Pui, S. Jeha, M. V. Relling, W. E. Evans, D. S. Gerhard, J. M. Gastier-Foster, E. Mardis, R. K. Wilson, M. L. Loh, J. R. Downing, S. P.

Hunger, C. L. Willman, J. Zhang, and C. G. Mullighan, “Targetable Kinase-Activating Lesions in Ph-like Acute Lymphoblastic Leukemia,” *New England Journal of Medicine*, vol. 371, pp. 1005–1015, sep 2014.

[191] C. Miller, “No Titlesimpleaffy: Very simple high level analysis of Affymetrix data. http://www.bioconductor.org,” 2018.

[192] R. Irizarry, “Exploration, Normalization, and Summaries of High Density oligonucleotide Array Probe Level Datatle,” *Biostatistics*, vol. 4, no. 2, pp. 249–264, 2003.

[193] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth, “limma powers differential expression analyses for RNA-sequencing and microarray studies,” *Nucleic Acids Research*, vol. 43, pp. e47–e47, apr 2015.

[194] R. Efron and R. Tibshirani, “Empirical Bayes Methods and False Discovery Rates for Miroarrays.,” *Genetic Epidemiolgy*, vol. 23, no. 1, pp. 70–86, 2002.

[195] V. K. Mootha, C. M. Lindgren, K.-F. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, M. Ridderstråle, E. Laurila, N. Houstis, M. J. Daly, N. Patterson, J. P. Mesirov, T. R. Golub, P. Tamayo, B. Spiegelman, E. S. Lander, J. N. Hirschhorn, D. Altshuler, and L. C. Groop, “PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes,” *Nature Genetics*, vol. 34, pp. 267–273, jul 2003.

[196] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, “Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, pp. 15545–50, oct 2005.

[197] Broad Institute, “GSEA User Guide,” 2016.

[198] Broad Institute, “Gene Set: KEGG CYTOKINE-CYTOKINE RECEPTOR INTERACTION,” 2016.

[199] P. P. Zenatti, D. Ribeiro, W. Li, L. Zuurbier, M. C. Silva, M. Paganin, J. Tritapoe, J. A. Hixon, A. B. Silveira, B. A. Cardoso, L. M. Sarmento, N. Correia, M. L. Toribio, J. Kobarg, M. Horstmann, R. Pieters, S. R. Brandalise, A. A. Ferrando, J. P. Meijerink, S. K. Durum, J. A. Yunes, and J. T. Barata, “Oncogenic IL7R gain-of-function mutations in childhood T-cell acute lymphoblastic leukemia,” *Nature Genetics*, vol. 43, pp. 932–939, oct 2011.

[200] M. B. Sonbol, B. Firwana, A. Zarzour, M. Morad, V. Rana, and R. V. Tiu, “Comprehensive review of JAK inhibitors in myeloproliferative neoplasms.,” *Therapeutic advances in hematology*, vol. 4, pp. 15–35, feb 2013.

[201] Broad Institute, “Gene Set: KEGG JAK-STAT SIGNALINGPATHWAY,” 2016.

[202] G. Tamamyan, C. Shi, K. Roberts, H. Ma, C. Mullighan, and M. Konopleva, "Targeting of JAK-STAT signaling pathway and BCL-2 family proteins in Ph-like acute lymphoblastic leukemia," *Clinical Lymphoma Myeloma and Leukemia*, vol. 15, p. S4, sep 2015.

[203] C.-H. Pui, "Genomic and pharmacogenetic studies of childhood acute lymphoblastic leukemia," *Frontiers of Medicine*, vol. 9, pp. 1–9, mar 2015.

[204] C. Wu and W. Li, "Genomics and pharmacogenomics of pediatric acute lymphoblastic leukemia," *Critical Reviews in Oncology/Hematology*, vol. 126, pp. 100–111, jun 2018.

[205] J. E. Cortes and H. M. Kantarjian, "Acute lymphoblastic leukemia. A comprehensive review with emphasis on biology and therapy.," *Cancer*, vol. 76, pp. 2393–417, dec 1995.

[206] H. Yu, D. Pardoll, and R. Jove, "STATs in cancer inflammation and immunity: a leading role for STAT3," *Nature Reviews Cancer*, vol. 9, pp. 798–809, nov 2009.

[207] D. C. Wang, W. Wang, L. Zhang, and X. Wang, "A tour of 3D genome with a focus on CTCF," *Seminars in Cell & Developmental Biology*, jul 2018.

[208] J. Qian, Q. Wang, M. Dose, N. Pruett, K. R. Kieffer-Kwon, W. Resch, G. Liang, Z. Tang, E. Mathé, C. Benner, W. Dubois, S. Nelson, L. Vian, T. Y. Oliveira, M. Jankovic, O. Hakim, A. Gazumyan, R. Pavri, P. Awasthi, B. Song, G. Liu, L. Chen, S. Zhu, L. Feigenbaum, L. Staudt, C. Murre, Y. Ruan, D. F. Robbiani, Q. Pan-Hammarström, M. C. Nussenzweig, and R. Casellas, "B cell super-enhancers and regulatory clusters recruit AID tumorigenic activity," *Cell*, vol. 159, no. 7, pp. 1524–1537, 2014.

[209] P. B. Talbert, M. P. Meers, and S. Henikoff, "Old cogs, new tricks: the evolution of gene expression in a chromatin context," *Nature Reviews Genetics*, p. 1, mar 2019.

[210] Y. Zhang, L. An, J. Xu, B. Zhang, W. J. Zheng, M. Hu, J. Tang, and F. Yue, "Enhancing Hi-C data resolution with deep convolutional neural network HiC-Plus," *Nature Communications*, vol. 9, no. 1, pp. 1–9, 2018.

APPENDIX A: TRANSCRIPTOMIC ANALYSIS OF HGSVC TRIOS

The description of the data processing and command line arguments for SV-ASE are detailed below. The full results of the structural variant allele-specific expression analysis for each trio daughter are available as a supplemental data file formatted as an excel workbook titled *HGSVC_ SV-ASE.results.xlsx*, with the columns described at the end of this section. Command line arguments and detailed description of SV-ASE analysis methods are available in the supplementary file *README.HGSV.SVASE.txt*. HGSVC_SV-ASE.results.xlsx provides the significant SV-intersected ASE-SNP genes that showed an allele specific effect for IL-SVs and PB-SVs. Below, table A.1 provides an overview of the supplemental file providing the results of the allele specific analysis.

Table A.1: Descriptions of HGSV ASE results supplemental file.

COLUMN	VALUE
1	gene chromosome
2	gene start position
3	gene stop position
4	Ensemble ID of gene
5	gene symbol
6	sv chromosome
7	sv start position
8	sv end postion
9	reference allele
10	alternate allele (given as sv type)
11	quality score
12	filter value
13	SV genotype
14	haplotype 1 RNAseq read counts
15	haplotype 2 RNAseq read counts
16	Read count ratio between haplotypes
17	Read count total for both haplotypes
18	P-value result from binomial tests
19	FDR corrected p-value

APPENDIX B: TRANSCRIPTOMIC ANALYSIS OF ACUTE LYMPHOBLASTIC LEUKEMIA

B.1 Heatmaps of differentially expressed genes in Ph-like leukemia

Below are images that depict the top 50 differentially expressed genes in the Ph-like ALL samples relative to each of the alternative subtypes as heatmaps. The color bar at the top of each heatmap gives the Ph-like samples as gold bars, while the contrast group is given by a pink bar. Low gene activity is given by a blue color, whereas increasing levels span the color spectrum from blue to yellow to red. The names of the genes for which the heatmap rows are drawn are given on the right hand side of each image.

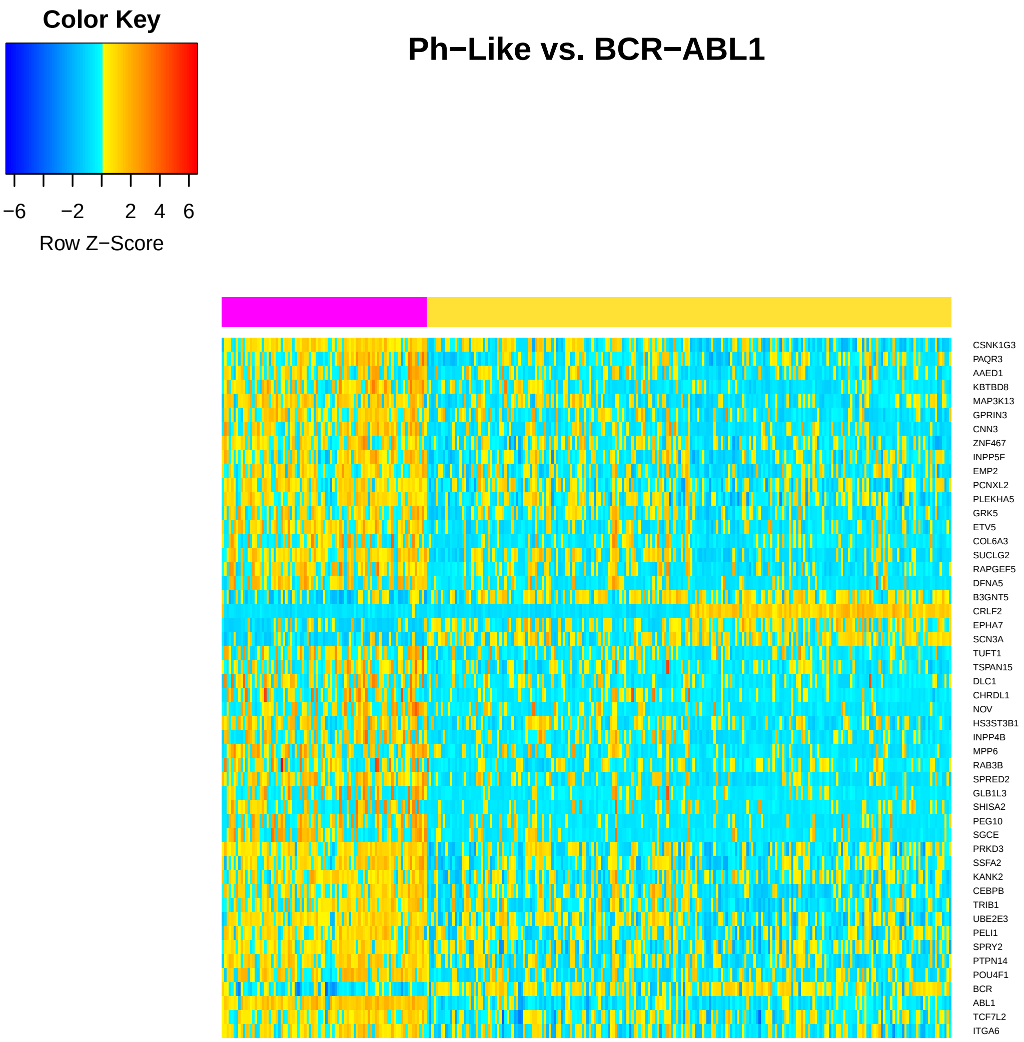

Figure B.1: Expression heatmap BCR-ABL1 vs. Ph-like ALL

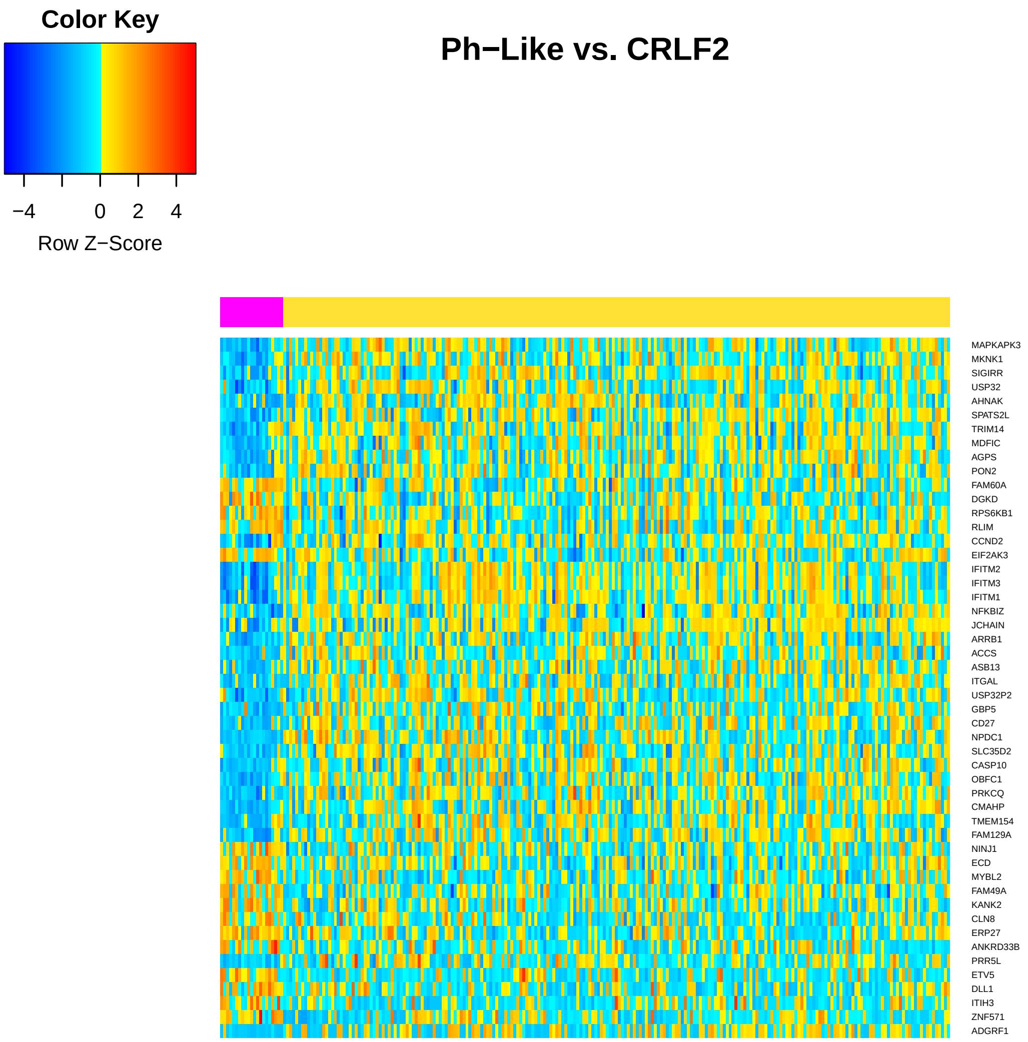

Figure B.2: Expression heatmap CRLF2 vs. Ph-like ALL

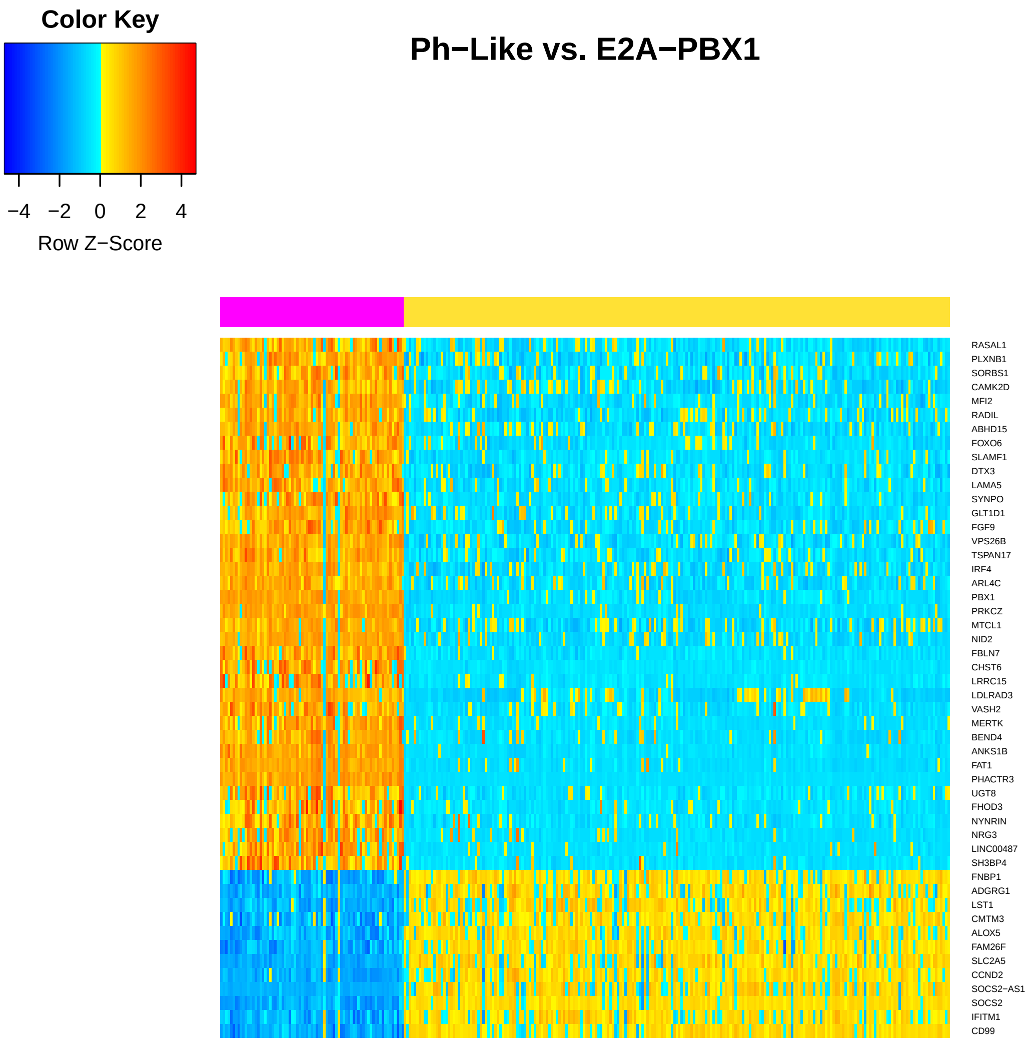

Figure B.3: Expression heatmap E2A-PBX1 vs. Ph-like ALL

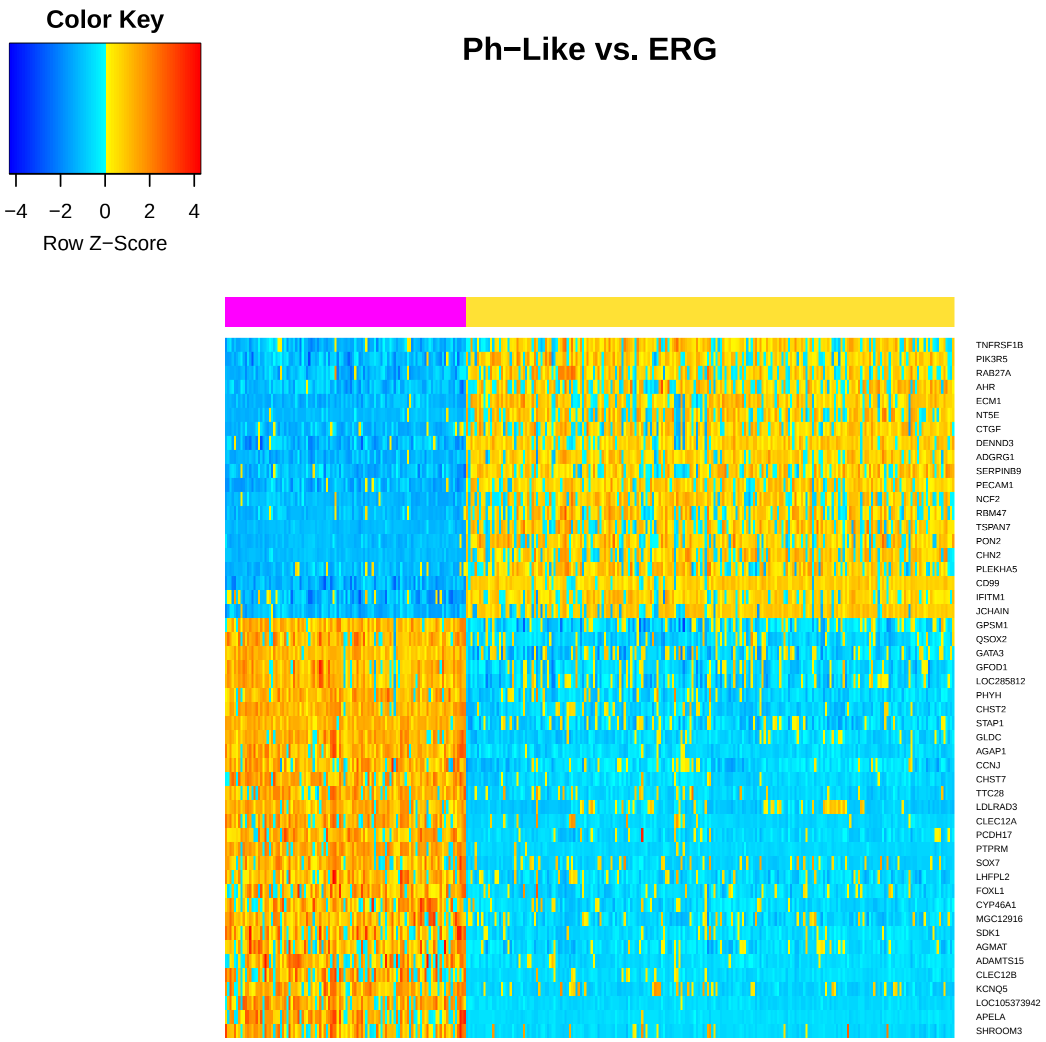

Figure B.4: Expression heatmap ERG vs. Ph-like ALL

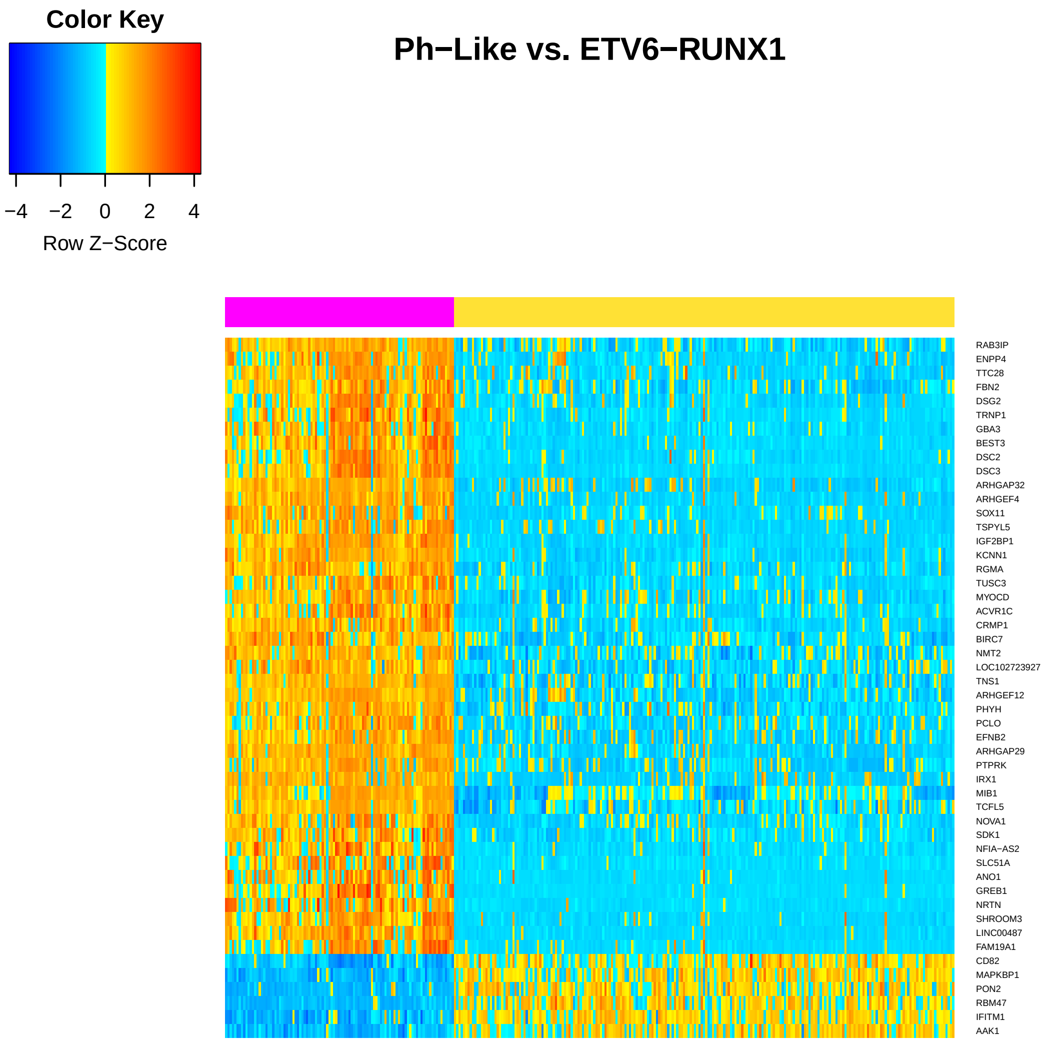

Figure B.5: Expression heatmap ETV6-RUNX1 vs. Ph-like ALL

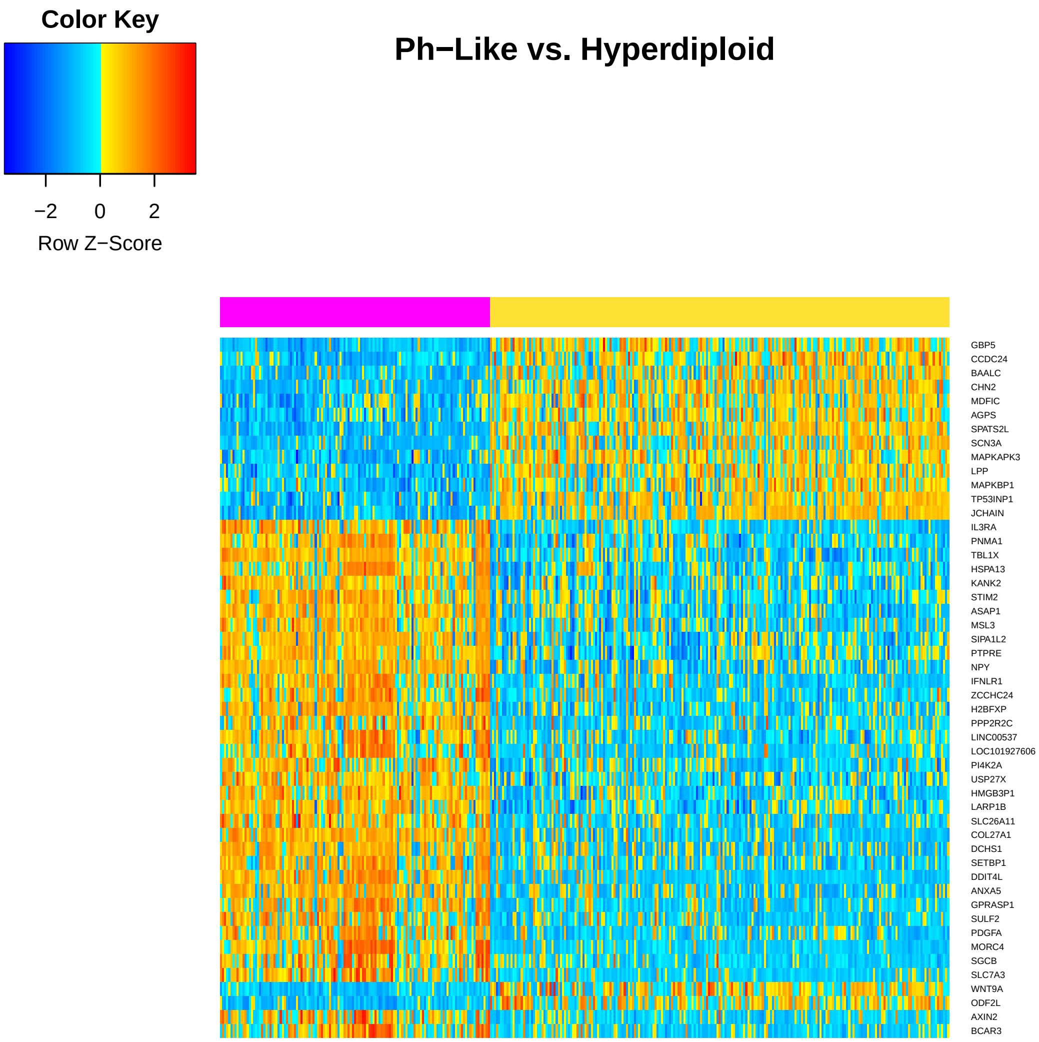

Figure B.6: Expression heatmap Hyperdiploid vs. Ph-like ALL

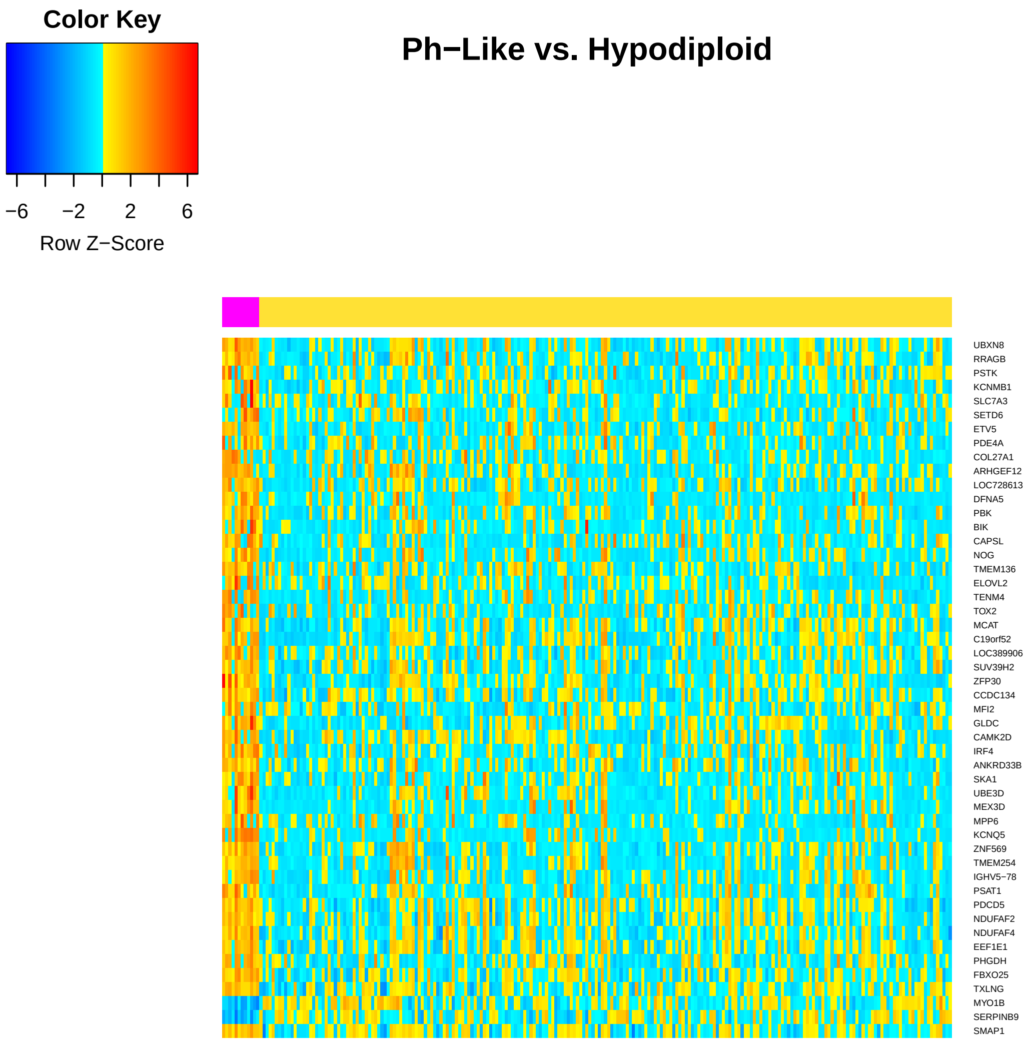

Figure B.7: Expression heatmap Hypodiploid vs. Ph-like ALL

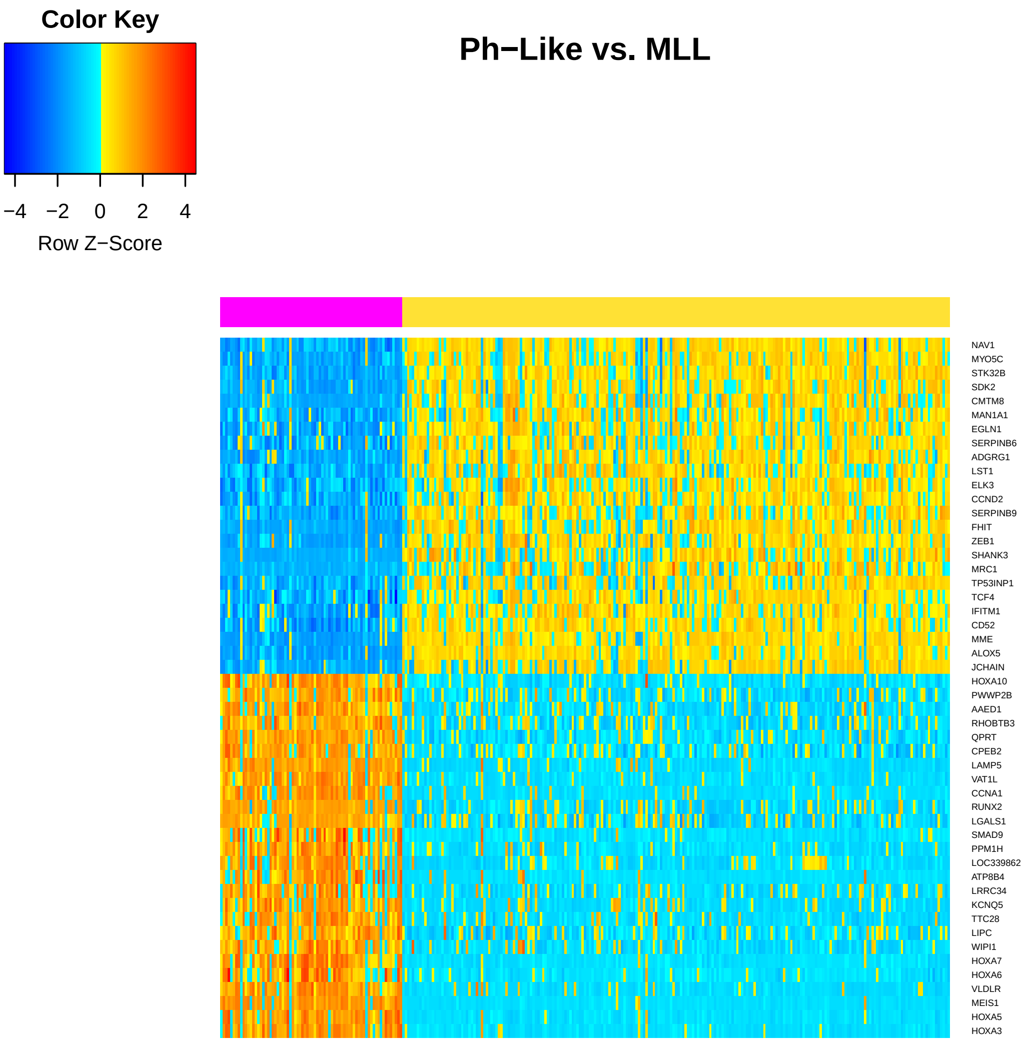

Figure B.8: Expression heatmap MLL vs. Ph-like ALL

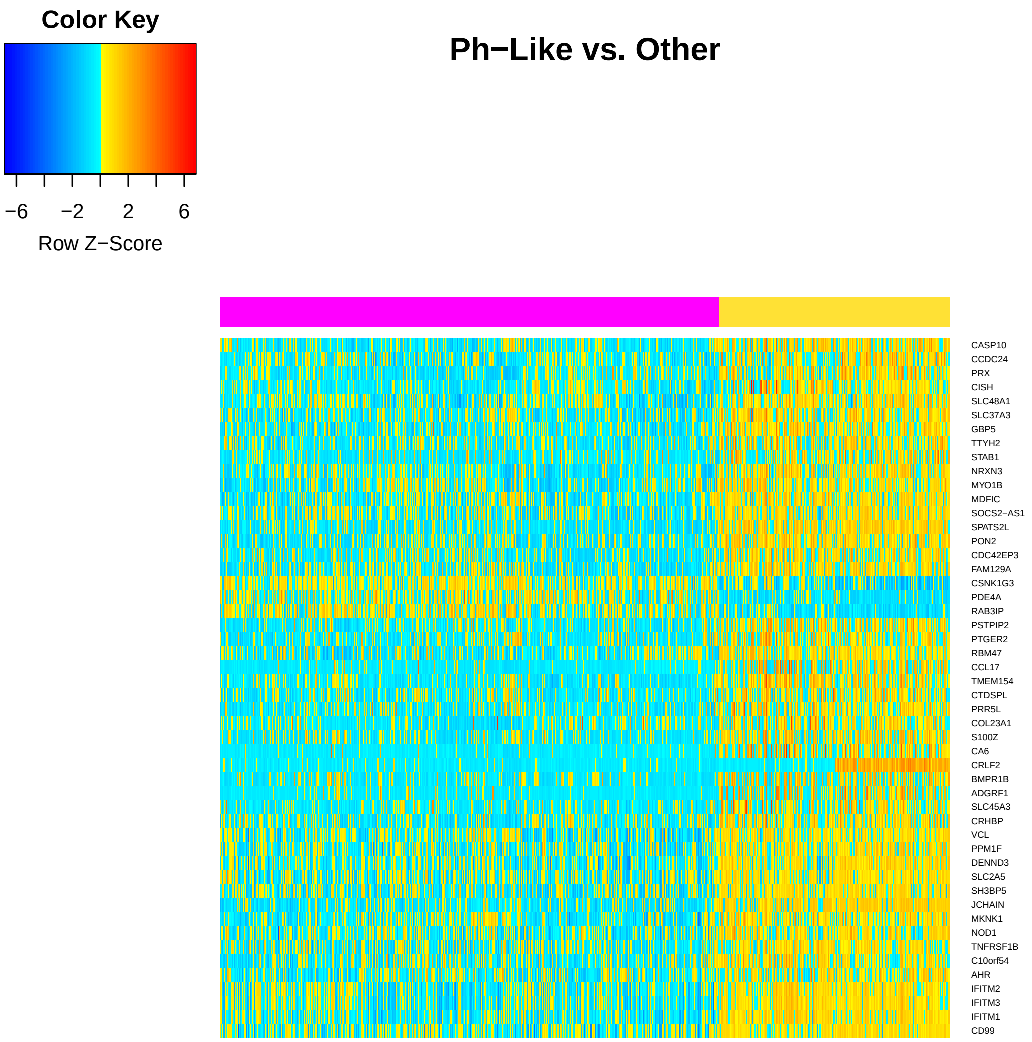

Figure B.9: Expression heatmap Unspecified vs. Ph-like ALL

B.2 Plots of Multinomial Deviance vs. log(Lambda) in elastic-Net regression models

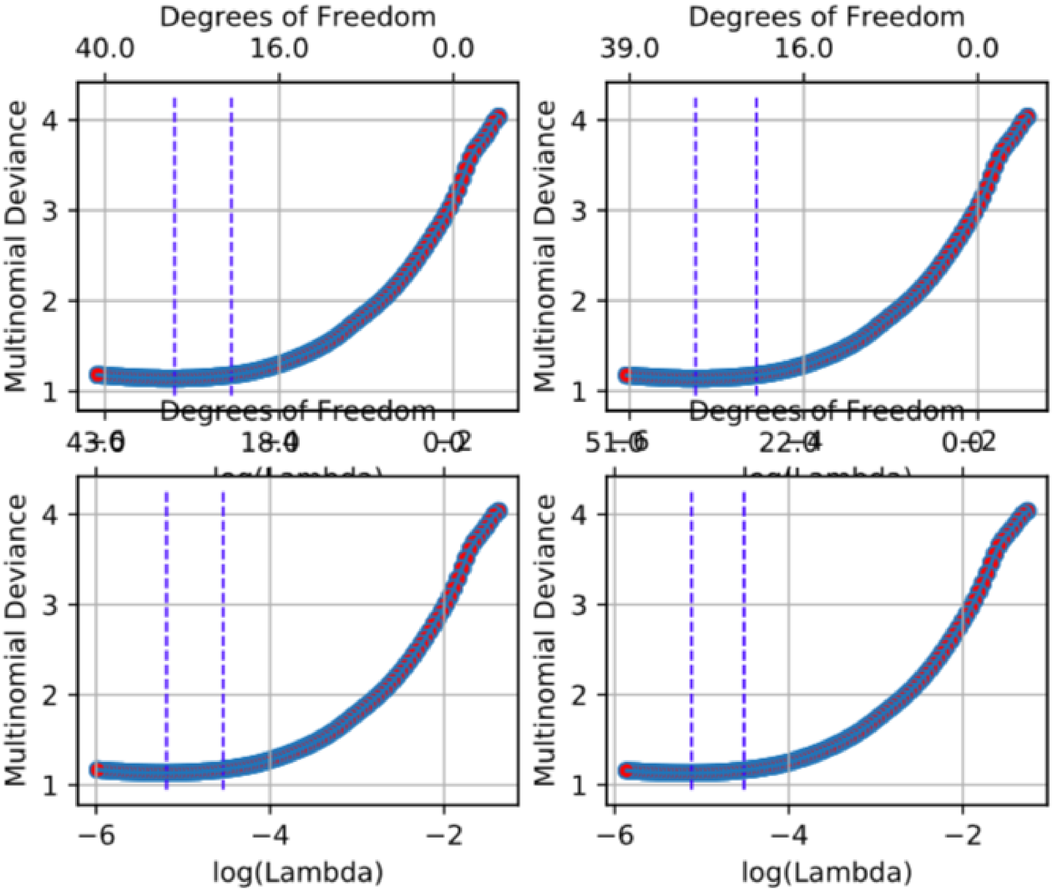

Figure B.10: multinomial deviance vs. log(λ) 1

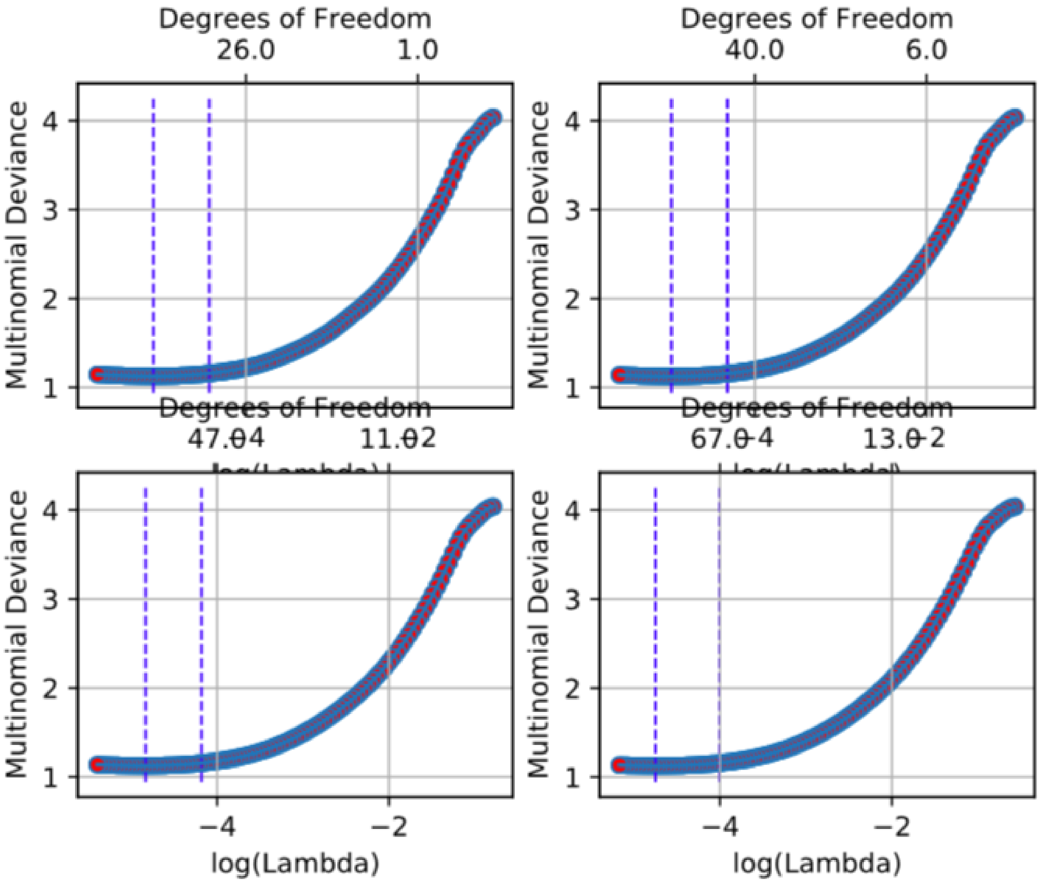

Figure B.11: multinomial deviance vs. log(λ) 2

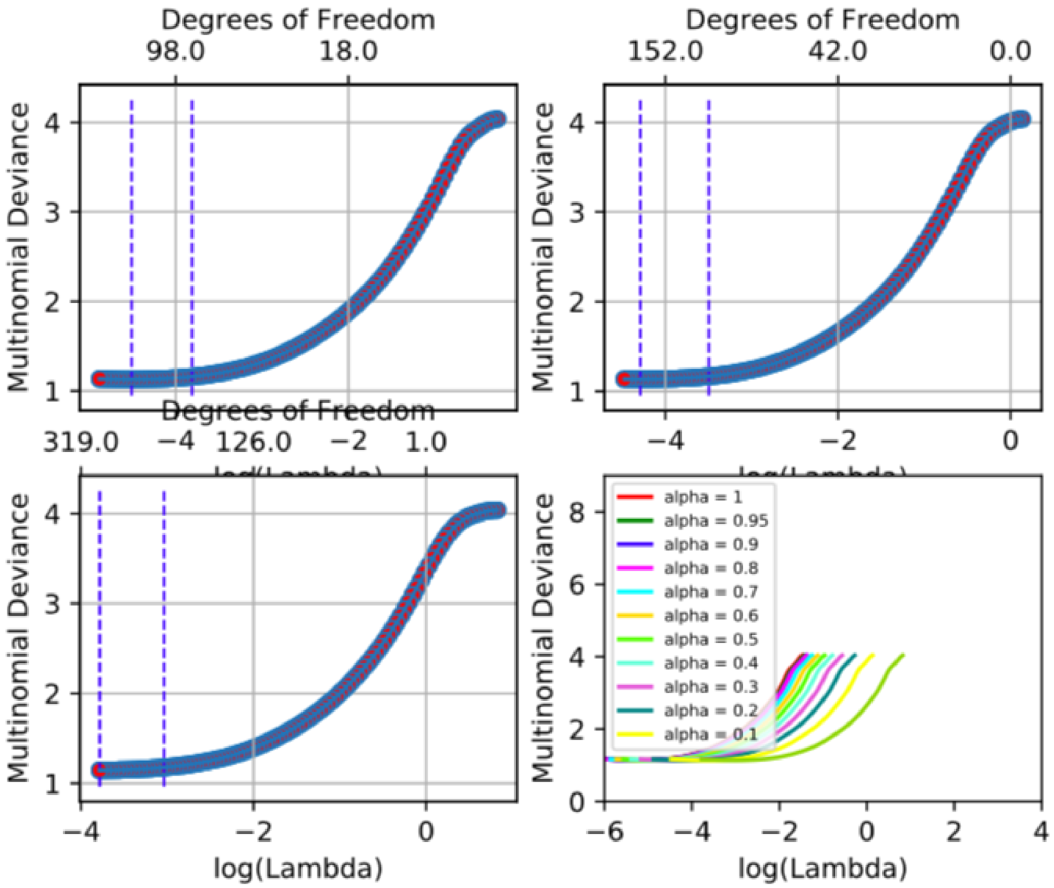

Figure B.12: multinomial deviance vs. $\log(\lambda)$ 3