ANALYSIS OF CHROMOSOME SPATIAL ORGANIZATION DATA AND INTEGRATION WITH GENE MAPPING FOR COMPLEX TRAITS

Cheynna A. Crowley

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Public Health in the Department of Biostatistics in the Gillings School of Global Public Health.

Chapel Hill
2020

Approved by:

Yun Li

Michael I. Love

Yuchao Jiang

Di Wu

Paola Giusti-Rodríguez

ABSTRACT

Cheynna A. Crowley: Analysis of Chromosome Spatial Organization Data and Integration with Gene Mapping for Complex Traits
(Under the direction of Yun Li)

Studying the 3D chromosomal organization is crucial to understanding processes of transcription, histone modifications, and DNA repair and replication. Chromatin conformation shapes molecular functions beyond genetic variation at the sequence level and epigenetic footprints along the one-dimensional genome. DNA spatial organization features can influence molecular and organism-level phenotypes, from regulation of the expression of target genes (which can be megabases [Mb] away), to the development of various diseases including autoimmune diseases, neurological diseases, and cancer.

The genome-wide chromosome conformation capture technology Hi-C captures genomic interactions of all loci, genome-wide. Hi-C data allows us to investigate chromatin organization at various levels and resolutions, including the Mb resolution chromosome compartments and topologically associated domains (TADs), 10-40Kb resolution frequently interacting regions (FIREs), and 1-40Kb resolution chromatin loops and long-range chromatin interactions.

FIREs have been demonstrated to provide valuable information for tissue or cell type-specific transcriptional regulation, characteristics unique from other domain features observed in the 3D genome. Until now, there is no stand-alone software package for the detection of FIREs. To fill in this gap, I first present a user-friendly R-package to identify FIREs and the clustering of FIREs (super-FIREs), accessible to the general scientific community.

Next, I further explore the 3D genome and analyze brain tissue Hi-C data from 3 fetal and 3 adult human cortex samples with a total of 10.4 billion raw reads, the most deeply sequenced human brain tissue Hi-C datasets we are aware of to date. My analysis of this Hi-C data (identifying compartments, TAD boundaries, FIREs, and long-range chromatin interactions) generated mechanistic insights at GWAS loci for psychiatric disorders, brain-based traits, and neurological conditions, particularly schizophrenia.

Lastly, as incorporating annotation can provide insights at GWAS loci, I annotate 148,019 variants identified in a recent trans-ethnic analysis for hematological traits in 746,667 participants. I present my findings in an R Shiny app, ABCx: Annotator for Blood Cell Traits, which highlights variants 1D epigenomic signatures, impact on gene expression, and chromatin conformation information to aid in further functional follow up.

ACKNOWLEDGEMENTS

I would like to gratefully acknowledge my advisor, Yun Li, for her support during my time at UNC. I am extremely thankful for all that she has taught me. Additionally, I would like to acknowledge my doctoral committee, Michael Love, Di Wu, Yuchao Jiang, and Paola Giusti-Rodríguez. I would also like to thank Ming Hu and the Li Lab group. Lastly, I would like to thank Melissa Hobgood, as well as my friends and family for their guidance throughout my educational career. The research in Chapters 2, 3, and 4 is supported by the National Institute of Health grant R01HL129132 (awarded to Yun Li).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

3C	Chromosome Conformation Capture
3D	Three-dimensional
AA	African Ancestry
AD	Adrenal Gland
AO	Aorta
BASO	Basophil Count
BL	Bladder
CP	Cortical Plate
DLPFC	Dorsolateral Prefrontal Cortex
E-P	Enhancer-Promoter
EA	European Ancestry
EAS	East Asian
EM	Expectation-Maximization
EMR	Electronic Medical Records
EOS	Eosinophil Count
eQTL	Expression Quantitative Trait Loci
FDR	False Discovery Rate
FIREs	Frequently Interacting Regions
GWAS	Genome-wide Association Studies
GZ	Germinal Zone
H3K27ac	H3K27 acetylation

H3K4me1	H3 lysine 4 monomethylation
HA	Hispanic Ancestry
HCT	Hematocrit
hESC	Human Embryonic Stem Cell
HGB	Hemoglobin Concentration
HGNC	HUGO Gene Nomenclature
Hi-C	High-throughput Chromatin Conformation Capture
Hippo	Brain Hippocampus
Kb	Kilobase
LCL	Lymphoblastoid Cell Lines
LD	Linkage Disequilibrium
LG	Lung
LI	Liver
LRCI	Long-range Chromatin Interactions
LV	Heart Left Ventricle
LYM	Lymphocyte Count
MACS	Model-based Analysis of ChIP-Seq
Mb	Megabase
MCH	Mean Corpuscular Hemoglobin
MCHC	Mean Corpuscular Hemoglobin Concentration
MCV	Mean Corpuscular Volume
MES	Mesoderm

MHC	Major Histocompatibility Complex
MONO	Monocyte Count
MPRA	Massively Parallel Reporter Assay
MPV	Mean Platelet Volume
MSC	Mesenchymal Stem Cell
NB	Negative Binomial
NEU	Neutrophil Count
NPC	Neural Progenitor Cell
OR	Odds Ratio
OV	Ovary
PA	Pancreas
PC	Principal Component
PCA	Principal Component Analysis
pheWAS	Phenome-wide Association Studies
PLT	Platelet Count
PO	Psoas Skeletal Muscle
pQTL	Protein Quantitative Trait Loci
RBC	Red Blood Cell Count
RDW	Red Blood Cell Distribution Width
RV	Heart Right Ventricle
SA	South Asian
SB	Small Intestine (Bowel)

SCC	Stratum-adjusted Correlation Coefficient Statistic
SCZ	Schizophrenia
sQTL	Splicing Quantitative Trait Loci
SX	Spleen
TADs	Topologically Associated Domains
Trans	Trans ethnic
TRO	Trophoblast-like Cell
TSS	Transcription Start Site
VTE	Venous Thromboembolism
WBC	White Blood Cell Count
WGSA	Whole Genome Sequencing Annotator

CHAPTER 1: LITERATURE REVIEW

1.1 Introduction

This section provides background information and reviews existing literature related to the methodologies and applications presented in Chapters 2, 3, and 4. Information reviewed in this section falls under the general umbrella of integration of gene mapping approaches with functional genomics information, with an emphasis on chromosome spatial organization information, for revealing molecular mechanisms underlying complex human diseases and traits. Specifically, we will review relevant literature for genome-wide association studies (GWAS), and the studies of the 3D genome with a deeper dive into the Hi-C technology. The Hi-C technology is the genome-wide, unbiased version of the chromosome conformation capture (3C) derived technologies, that allows interrogation of intra-chromosomal interactions between theoretically all pairs of genomic loci. The main emphasis of this review is the integration of multi-level genomic and epigenomic data to help the interpretation and prioritization of GWAS findings.

1.2 Background on Genome-Wide Association Studies (GWAS)

GWAS are performed to assess genotype-phenotype associations and are widely used to identify genetic determinates of complex traits [58]. They utilize expansive variants (hundreds of thousands to tens of million) across the genome for a relatively large number of individuals (typically thousands to even up to more than a million) to identify genes or genetic variants that are associated with a phenotype or trait of interest. Overall, GWAS have been largely successful; as of October 07, 2020, the NHGRI-EBI catalog of published GWAS contained 4,741 publications and 212,730 unique associations between genetic variants [7] and various diseases and traits

[87]. As one example, a recent GWAS of schizophrenia (SCZ) meta-analyzing 40,675 cases and 64,643 controls from multiple studies, identified 146 loci associated with SCZ, among which 50 were novel loci that have not been previously reported [62].

GWAS results become complicated when they identify genetic variants that may not reside in any genes or encode protein sequences. Non-coding regions make up a vast majority (more than 90%) of GWAS-identified variants and have largely unknown functional consequences. Converting the rich body of GWAS findings into insights connecting health and disease entails the elucidation of molecular mechanisms in general, and the identification of causal functional variants and their target gene(s).

In this post-GWAS era, serious efforts have been invested to generate functional mechanistic hypotheses behind GWAS associations. For instance, the above SCZ study attempted to move from beyond GWAS associations to causal genes by integrating fine-mapping results with brain expression and chromosome conformation data, leading to the identification of candidate causal genes within 33 loci [62].

The overarching goal of GWAS is to detect associations between genotypes and phenotype(s)/trait(s) of interest. GWAS are primarily conducted using the following components: identifying some phenotype(s) to be studied, collecting information from some cohort(s) of samples, with cohort loosely defined to encompass samples from various study designs including a case-control design, rather unselected population samples, and those from electronic medical records (EMR) or biobanks.

The selection of the study cohort(s) can impact the scope as well as the final findings. In general, a more diverse population and larger samples sizes are recommended, because, among other reasons, risk loci can and have been found to show differences in frequency, effect sizes, and correlation structure among neighboring genetic variants (linkage disequilibrium or LD in genetic terminology) across diverse population [83, 58]. For example, if a particular genetic variant is genome-wide significant in one population and not in another, it may be due to a difference in allele frequency, a difference in effect sample sizes from the two populations, or differential

LD (for example, the unmeasured causal variant is highly correlated with the particular genetic variants in one population but not the other). In addition, the non-European populations remain severely underrepresented, creating an opportunity to identify population or ancestry-specific risk variants for diverse populations. More specifically, when examining populations, ensuring multiple or diverse ancestries are represented in all groups and excluding individuals that exhibit extreme difference in genetic background from the rest of the samples will mitigate a residual substructure such that type 1 error can be better controlled.

A larger sample size is almost always preferred, as long as this size does not incur much additional heterogeneity in terms of phenotype subtypes or non-genetic risk factors, as sample size is the most crucial factor in determining power of association studies. For most complex traits, the effect sizes of individual genetic variants have been found to be modest or small, requiring large sample sizes for GWAS. Although, the first GWAS for age-related macular degeneration identified the GWAS locus CFH with merely 96 cases and 50 controls [42]. Noting a typical GWAS and meta-analysis involves thousands to even millions of individuals to have reasonable power to identify the rather moderate effect sizes that most associated genetic variants exert on complex traits. For example, it is known that psychiatric diseases and disorders for mental health involve a rather ”flat” genetic architecture with almost no genetic variants exhibiting large effect sizes identifiable from thousands of individuals. Such flatness was partly reflected by the findings from the 2007 WTCCC flagship study, which performed GWAS for seven complex traits, identified genetic loci for the geneticists’ nightmare, type-II diabetes, but none for bipolar disorder, with a similar sample size of 2,000 cases for each disease and 3,000 shared controls [14].

As previously highlighted, most signals map to non-coding regions complicating the interpretation of GWAS results. To aid interpretation, fine-mapping studies are typically carried out, attempting to identify causal variants from the pool of associated variants and their LD tags. Fine-mapping in this document refers to the determination of the potential causal genetic variant (or variants) and identification of independent signals at known loci, on top of the single-variant GWAS results, from one or multiple cohorts. Fine-mapping can also be carried at gene level rather than at the genetic variant level, as performed in recent transcriptome-wide association studies [55], but will not be further discussed. Fine-mapping methods assume that for the given GWAS evidence of an association in a genomic region for a trait, there is at least one causal variant in or near that region [70]. The overarching goal is to determine which variant(s) are most likely to be functional and to quantify the strength of evidence. There are various forms and variations of fine-mapping approaches. I will briefly comment on the two key aspects of fine-mapping: trans-ethnic fine-mapping and credible sets.

Trans-ethnic fine-mapping has been proven powerful due to the differential LD structure across populations. Complex traits have been reported to be relatively consistent in terms of underlying populations and contain a similar direction of effect of alleles [78, 83, 70, 82]. Thus, assuming the simple, but realistic scenario that the set of causal SNPs is the same for all populations, differential LD patterns across populations would result in narrowing the region where the causal variant(s) reside. As such, incorporating multiple ethnicities in GWAS fine-mapping can help pinpoint the causal variant(s) and neutralize the association of proxy-correlated loci [70, 58].

One standard deliverable out of fine-mapping analysis is the identification of a credible set at each associated locus. Several methods have been proposed to identify credible sets of variants. Most of which adopt a Bayesian framework that uses the posterior probabilities to determine a credible set. Credible sets are useful as they aim to identify sets of minimal size that contain all the causal variants with a pre-defined minimal confidence (typically 95% confidence) [10]. Therefore, variants in the credible sets serve as the minimal (in terms of the number of variants to be evaluated), but comprehensive (in terms of coverage of causal variants) pool for which we can integrate functional annotations to identify causal variants and their molecular function further.

1.2.1 Integrating Functional Annotations

As a large portion of genetic variation associated with complex diseases and traits reside in non-coding regions, augmenting the associated loci with functional regulatory genomic information can provide insights into the true causal variants(s), and the potential consequences of such variant(s). For example, by connecting variants to genes, we can begin to hypothesize the underlying molecular mechanisms. Comprehensive functional annotations can entail intersecting with epigenetic features, gene expression maps, and chromosome contact maps to find evidences of potential regulatory regions through 1D epigenetic features such as open chromatin, histone modifications, transcription factor binding, association with expression of a specific gene(s), and chromatin looping particularly when involving enhancer-promoter interactions. The integrating of chromosomal structure information can also provide insight into how chromosome spatial organization orchestrates transcriptional regulation as genome folding forms loops that bring enhancers and target genes into close proximity [64, 43]. At the organism level, various diseases and disorders (including cancer, immunodeficiencies, limb malformation, autoimmune diseases, neurological diseases, and thrombotic disorders) have been noted as a consequence of erroneous wiring of regulatory circuitry between enhancers and their target genes [43]. Publicly available databases such as ENCODE [13], Epigenome RoadMap [4], GTEx [51], and the 4D Nucleome Project [17] provide various functional annotation information across multiple tissues, cell lines or cell types. Integration with GWAS variants could identify and prioritize genes and variants that are likely to play a causal role, and generate corresponding mechanistic hypothesis, speeding up mechanistic characterization using experimental approaches.

1.3 3D Chromosomal Organization

There are approximately 3 billion base pairs in the human genome (around 1 meter in length) that fit inside the nucleus ($2 * 10^{-}6$ meters in diameter). The way in which the chromosomes fits inside the nucleus (3D architecture), determines the functionality of *cis*-regulatory elements and their genes [48]. It is widely accepted that there are chromosomal territories; however, the internal structure is not fully understood [81, 45].

Uncovering the knowledge of the 3D genome can reveal the functional consequences of disease-associated variants and chromosomal rearrangements [81, 23], leading to insight of the molecular mechanisms underlying disease [90, 45]. It is known that regulatory variants impact enhancer regions that regulate target genes through the formation of chromosomal looping [67]. In addition, gene transcription is initiated at promoter sequences immediately upstream of a gene, and differential gene expression in development and disease is also controlled by additional regulatory elements [43, 69, 88].

Enhancers are sequence modules, approximately a couple hundred base pairs in size, that contain motifs and are around transcription factor binding motifs [43]. It has been observed that the binding and co-recruitment of associated transcription factors activate enhancers. This often includes increased DNA accessibility, histone H3 lysine 4 monomethylation (H3K4me1) and H3K27 acetylation (H3K27ac) marks [37, 69, 43]. More specifically, developmental genes are usually controlled by multiple enhancers, where the enhancers behavior varies and determines the genes expression. Enhancers can be active in the same tissue, or different tissues, and at a variety of distances. Adding to this complexity, the closest gene in proximity to an enhancer is not necessarily the gene it regulates [43]. These complex enhancers are present within chromosomal territories [81, 36], and as such, a thorough understanding of the 3D architecture can provide further disease insight.

1.4 Chromosome Conformation Capture (3C)

Chromosome conformation capture (3C) technologies study the 3D architecture of the genome and reveal chromosomal organization within the nucleus [61, 81]. To better understand the chromosomal organization, 3C-based methods attempt to reveal the organization within the nucleus by determining the physical proximity of sets of points along the chromatin [81, 45]. Overall, 3C technologies crosslink DNA fragments to quantify ligation efficiencies as a measure of their contact frequency in a cell population [48]. Understanding the 3D chromosomal organization within the genome is crucial to understanding processes of transcription, histone modifications, repair, and replication as the conformation coincides with the variation of genomic sequencing. As DNA sequence variation is influence by several 3D chromatin conformation features such as loop strength, contact insulation, contact directionality, and density of *cis*-contacts

[81], it is further hypothesized that 3D conformation elements overlapping with transcriptional, regulatory, and epigenomic features could be then connected to disease risk [43]. While there are multiple 3-C technologies, Hi-C detects the genomic interactions of all loci genome-wide (termed 'All to All') [81, 45, 48]. The resulting 'contacts' discovered through Hi-C displays the interactions of all genomic loci in a tissue or cell line, and is commonly summarized in the form of a genome-wide contact frequency matrix [48].

1.4.1 Hi-C Technology

In a typical Hi-C experiment cells are crosslinked with formaldehydes, and DNA is digested with a restriction enzyme that leaves a 5' overhang. The 5' overhang is filled with a biotinylated residue, and the resulting blunt-end fragments are ligated under conditions favorable to the cross-linking of ligation events. The resulting DNA samples contain ligation products consisting of fragments that were originally in close spatial proximity in the nucleus, marked with biotin at the junction. The resulting Hi-C library is created by shearing the DNA, selecting the biotin-contained fragments with streptavidin beads, and then conducting massive parallel DNA sequencing [81].

1.4.2 Hi-C Mapping

The resulting Hi-C library is typically evaluated for quality using a Phred Score and then mapped to a reference genome using a sequencing aligner. The mapped read is assigned to a single restriction fragment according to its 5' mapped position with limits on the minimum and maximum distance from the restriction site depending on the molecule's size distribution [81, 45]. The resulting fragment-mapped reads are filtered based on whether it is a same-fragment pair resulting from an unligated fragment or circularized fragment, a duplicate read due to PCR with identical pair-end sequence, or identical 5' alignment positions of the pair [81, 45]. The resulting catalog contains information pertaining to whether the loci interaction is due to the chromosome interacting within itself or across other chromosomes (intra-interaction, or inter-

interaction, respectively). Fragments resulting from this alignment can be used as an analysis unit for further discovery.

Binning at a fixed resolution of the interactions is often preferred and beneficial in summarizing the fragment length information due the high sequencing costs and biases. The vast space of all interactions makes achieving sufficient coverage for a maximum resolution challenging. Reducing the interaction space by aggregating restriction fragments into fixed-size bins (i.e. 5Kb, 10Kb, 40Kb, 100Kb, 1Mb) functionally reduces the complexity of the data, and smooths out the potential noise associated with this type of data [81, 45, 48], thus increasing the effective coverage. Intra-chromosome, binned interactions will be the focus of the paper moving forward.

1.4.3 Contact Matrix

Contact matrices summarize the loci-to-loci interactions along the chromosome from the resulting Hi-C and are often visually represented by heatmaps with intensity displaying the interaction frequency. The contact matrix, M, is constructed by dividing the genome into b bin sized consecutive non-overlapping regions for each chromosome. The resolution is determined based on sequencing depth, as a low sequencing depth would create unreliable results at higher resolutions ($<$ 20Kb) [48, 67, 37]. The corresponding symmetric *n*x*n* matrix, where each matrix entry, $M_{i,j}$, corresponds to the number of interactions between locus i and locus j.

The contact frequency between loci is dependent on the linear genomic distance and is further influenced by restriction enzyme fragment size, GC content, and mappability [48]. The amount of 'noise' present in the data is often associated with the sequencing depth, such as low signal, high repeat content due to telomeres or centromeres, or low mappability [81, 45, 48]. Additional balancing procedures of the raw contact matrix to account for these biases are often conducted [34, 35].

Normalization of the contact matrix, M, takes into account the effect of distance and sequence proximity on the contact probability, based on the assumption that by investigating the genome for interactions, each fragment (or bin) is observed approximately the same number of times in the experiment [48]. Normalization is conducted by dividing each of the i,j entry in the contact matrix by the genome-wide average contact probability for the loci at that genomic distance. The resulting normalized matrix, $M*$, when visualized by a heatmap, shows large blocks of enriched and depleted interactions. Investigating the grouping further, it can be deduced that since a pair of loci are nearby in space, they may be related, such as sharing neighboring loci or having correlated interaction profiles. As such, a correlation matrix C where $c_{i,j}$ is the Pearson correlation between the i^{th} row and j^{th} column of $M*$ is calculated. This visually and dramatically enhances the block patterns ('plaid'), inferring that each chromosome can be decomposed into two sets of loci of enriched and depleted sets [48].

1.4.4 Readouts from Hi-C Data

Analyzing Hi-C data has led to the discovery of structural readouts at a cascade of resolutions, including: (1) A/B compartments which are multiple megabases (Mb) to hundreds of kilobases (Kb) in size and largely correspond to active or inactive chromatin [48]; (2) topologically associating domains (TADs) which are on average approximately 1Mb in size and serve as the basic structural and functional unit of the genome to constrain long-range chromatin interactions largely within TADs [20]; (3) frequently interacting regions (FIREs) which at 10-40Kb highlight functional and regulatory areas with cell type and tissue specificity [71]; (4) chromatin loops at Kb resolution, which are nested within TADs and anchored by a pair of convergent CTCF motifs [67]; and (5) statistically significant long-range chromatin interactions (LRCI), again at Kb resolution, are pairs of chromatin segments brought in physical proximity more often than expected by random chromatin looping or collisions [3, 54, 92, 91]. Among these Hi-C readouts, TADs and chromatin loops are largely conserved across cell types [20, 67], while A/B compartments and LRCI's exhibit moderate levels of cell type-specificity [48].

1.4.5 A/B Compartments

The largest-scale, position-specific interaction patterns are known as A/B compartments. A/B compartments display the enrichment and depletion of interactions across the chromosome and tend to vary between cell types and throughout development stages (thus, moderately cell type and tissue type-specific) [81]. Through the creation of a correlation contact matrix, it was observed that the first principal component (PC) corresponded largely to the plaid pattern of enriched and depleted interactions. A compartments are highly enriched for active chromatin, and B compartments are enriched for inactive chromatin [48]. Interactions between loci are primarily constrained to occur between loci belonging to the same compartment. This then leads to the conclusion that the nucleus is segregated into two compartments corresponding to open and closed chromatin, and throughout the genome occupy different spatial compartments in the nucleus [48]. Genomic compartments have also been found to be correlated with chromatin states such as DNase accessibility, gene density, active genes, GC content, and histone marks [67, 48, 69, 18].

1.4.6 Topologically Associated Domains (TADs)

TAD's are often trivially described by observing the triangle structure highlighted along the diagonal of a raw contact matrix heatmap displaying higher levels of interaction each spanning about 1Mb in distance. In reality, TADs are actually quite complicated due to intricate interaction patterns and containing multiple hierarchies of overlapping block-like structures [45].

TAD's are mostly conserved. There is some evidence that TAD's are stable across different cell types and experimental conditions as well as conserved in related species [20, 22]. There is clear evidence that gene clusters are organized into individual TAD's, indicating that genes with functional needs for co-regulation tend to reside or are nested within the same TAD [20]. Coordinated regulatory function has been attributed to the presence of TADs. Chromosomal rearrangements often distort TADs and as such, previously separated genes and enhancers can potentially enter each other's contact space and form new regulatory circuits [43].

There are two main methods to calculate TAD's. The first method calculates the difference between the diagonal's average upstream and downstream interactions. This difference is used to define a directionality index (based on a Chi-square statistic), where the TAD boundaries are significantly different than the TAD regions [20]. The second method uses an insulation score that captures the density of signal in the Hi-C contact matrix around the diagonal, as a function of genomic position [15].

1.4.7 Frequently Interacting Regions (FIREs)

FIREs are defined at a smaller resolution and by a frequent interacting window 200Kb above and below the diagonal of a raw contact frequency matrix [71]. The process of FIRE identification starts with a contact frequency matrix at a specific bin resolution for each chromosome. These contact matrices are then used to identify the total number of *cis*-interactions. The *cis*-interactions are then filtered for technical biases known to Hi-C data, and then within-sample and across-sample normalized. Residuals from the fit of a Poisson regression are mapped to a Gaussian distribution. The -ln(p-value) is considered the continuous FIRE score, and significant bins at an alpha of 0.05, are deemed the dichotomized FIRE.

Uniquely from TADs, FIREs are tissue and cell type-specific, conserved in human and mouse, are enriched for tissue/cell type-specific enhancers, tissue/cell type-specific gene expression, and GWAS SNPs for relevant traits [71, 29, 28]. More specifically, Gorkin et al., found that FIRE scores were commonly positively correlated with marks of *cis*-regulatory activity, including H3K27ac and H3K4me3, thus confirming the previously reported relationship between FIREs and *cis*-regulatory activity [29]. Compared with the genomic location of chromosomal domains, FIREs commonly occur toward the center of TADs, are involved in the many intra-TAD interactions, and are largely contained in A compartments of active chromatin [71, 29].

1.4.7.1 Super-FIREs

To account for the observation that many FIREs occur in clusters, super-FIREs are created by concatenating together adjacent FIRE bins and ranking them by their continuous FIRE score to quantify the FIRE clusters. FIRE regions are then plotted as a function of their continuous FIRE score, once scaled to 0 and 1, clusters to the right of the inflection point are deemed a super-FIRE [71]. This methodology is adapted from the super-enhancer designation using the ROSE algorithm [88].

1.4.8 Chromatin Loops

Chromatin loops are hypothesized to play a significant genomic role in eukaryotes as they frequently link promoters and enhancers, correlate gene activation, and show conservation across species and cell types [67]. In addition, the DNA binding protein, CTCF, is strongly associated with loops as the loop anchors typically display at domain boundaries and bind to the CTCF [67]. The CTCF behavior is not always consistent with an insulator role and often exhibit behavior similar to characteristics of a transcriptional activator. Chromatin loops occur when contacts between chromosomal sites are stabilized which results in regulatory and architectural loops [67].

1.4.9 Long-Range Chromatin Interaction's (LRCI's)

LRCI's , or commonly referred to as 'peaks', detect the non-random interactions between loci from a 2D contact frequency matrix and have shown considerable relevance to functional regulation [3, 54, 92, 91]. Peak calling can be difficult to efficiently and effectively understand the dependency structure within these hidden peak status. There are a variety of peak calling methods, such as Fit-Hi-C [3], HMRFBayes [92], and FastHiC [91], which will be briefly discussed.

Fit-Hi-C aims to identify significant chromatin interactions at various lengths and significance thresholds [3]. Fit-Hi-C fits nonparametric spline curves across genomic distances, unique from other methods that use discrete binning. After a filtering of non-random collisions, there is a refitting of the spline curves and an incorporation of locus-specific defining factors similar to iterative corrections and eigenvector decomposition. Fit-Hi-C is sensitive to sequencing depth, and often variable across samples [3].

HMRFBayes is a peak calling technology that takes into account the dependency structure of loci. It uses a similar methodology as a Hidden Markov Model or Bayesian Hidden Ising model for peak identification of ChIP-Seq data [92], but extends these 1D methods to a 2D space based on the contact frequency matrix. The success of this method is due to accounting for the local spatial dependency among adjacent fragment pairs and then detecting all 2D peaks by borrowing information from the neighboring fragment pairs [92].

FastHiC uses a simulated field approximation to estimate the joint distribution of hidden peak status by a set of independent random variables using an EM algorithm [91]. In a comparison of these three procedures, it has been shown that Fit-Hi-C runs faster than FastHiC, and FastHiC runs almost 5 times faster than HMRFBayes. FastHiC and HMRFBayes outperformed Fit-Hi-C in peak calling accuracy, and FastHiC achieved higher peak calling accuracy then HMRFBayes [92, 91].

1.4.10 Hi-C Comparison Methods

Hi-C technology is experimental, and a high (Kb) resolution requires massive ($>$ 1 billion raw reads) sequencing efforts, which can be expensive and complicated [37]. Identifying similarities are important for quantifying the quality of the sample, as well as the reproducibility.

Previously, the traditional methods for comparing Hi-C results were not consistent. The standard practice for when biological replicates were not available was to visually inspect heat maps, and investigate the ratio of long-range interaction read pairs. When at least two biological replicates were available, then Pearson or Spearman correlations between multiple Hi-C contact matrices were analyzed [96]. Distance measurements such as L1-Norm and L2-Norm, are also sometimes used to classify similarities.

Visual inspection and ratios are difficult to classify as they are not supported by robust statistics, which creates an ad-hoc interpretation. Pearson correlation, Spearman correlation, L1-Norm distance, and L2 Norm-distance can give information about the similarities (and dis-similarities), but are susceptible to outliers as they do not take into account the unique characteristics of Hi-C data such as domain structures, interacting regions, and distance dependence. As such, other reproducibility methods have been developed to account for dependency and noise present in Hi-C data, such as HiCRep, and HiC-Spector.

HiCRep [95] is a reproducibility measurement that takes into account the unique spatial features of Hi-C data by implementing a 2D smoothing parameter, stratifying by distance, and applying correlations resulting in a stratum-adjusted correlation coefficient statistic (SCC) that is interpreted similarly as any correlation [95, 96]. Smoothing reduces the individual spatial resolution and improves continuity of regions with elevated interactions enhancing domain structures. ”Smoothing” the contact map is done by replacing the counts of each contact in the contact map with the mean count of all the contacts in the genomic neighborhood [95, 96]. This is done when samples are not sufficiently sequenced. The local variation introduced by under sampling makes it hard to capture large domain structures. Smoothing the raw contact matrix in order to reduce local noise in the contact map enhances the visibility of the domain structures. To take into account the distance effect in the reproducibility assessment, it stratifies the contacts by the genomic distance between the loci [95, 96]. The method was validated by comparing results of the ranking of similarity of pseudo-replicates, biological-replicates, and non-biological replicates to the results of Pearson and Spearman correlations. In this comparison, the SCC correctly ranked the samples more often and had more clear ranking when compared to the Pearson and Spearman correlation results [95].

HiC-Spector is a reproducibility metric for quantifying the similarity between contact maps (contact matrices, TADs, loops, FIREs) based on spectral decomposition. The distance metric (Q) is the sum of the Euclidian norms of the difference of the normalized eigenvectors derived from the Laplacian matrix of the chromosomal contact map [94]. This method was confirmed as a reliable comparison method as they tested 33 pairs of pseudo-replicates, 11 biological-replicates, and

110 pairs of different cell lines. They observed a clear distinction between the different samples, and that the distribution of Q-scores had a behavior that was expected [94].

CHAPTER 2: FIRECALLER: DETECTING FREQUENTLY INTERACTING REGIONS FROM HI-C DATA

2.1 Introduction

Chromatin folding in the three-dimensional (3D) space is closely related to genome function [16]. In particular, transcription regulation is orchestrated by a collection of *cis*-regulatory elements, including promoters, enhancers, insulators, and silencers. Alteration of chromatin spatial organization in the human genome can lead to gene dysregulation and consequently, complex diseases including developmental disorders and cancers [43, 47].

High-throughput chromatin conformation capture (Hi-C) has been widely used to measure genome-wide chromatin spatial organization since first introduced in 2009 [79, 96, 48]. Analyzing Hi-C data has led to the discovery of structural readouts at a cascade of resolutions, including A/B compartments [48], topologically associating domains (TADs) [20], chromatin loops [67], and statistically significant long-range chromatin interactions [92, 91, 3, 40]. Among these Hi-C readouts, TADs and chromatin loops are largely conserved across cell types [20, 21, 22, 43], while A/B compartments and long-range chromatin interactions exhibit rather moderate levels of cell type-specificity [3, 48].

As an attempt to identify Hi-C readouts that are better indicative of cell type or tissue-specific chromatin spatial organizations, we have in our previous work [71], identified thousands of frequently interacting regions (FIREs) by studying a compendium of Hi-C datasets across 14 human primary tissues and 7 cell types. We defined FIREs as genomic regions with significantly higher local chromatin interactions than expected under the null hypothesis of random collisions [71].

FIREs are distinct from previously discovered Hi-C structural readouts such as A/B compartments, TADs, and chromatin loops. In general, FIREs tend to reside at the center of TADs, associate with intra-TAD Enhancer-Promoter (E-P) interactions, and are contained within broader regions of active chromatin [71]. FIREs are tissue and cell type-specific, and enriched for tissue-specific enhancers and nearby tissue-specifically expressed genes, suggesting their potential relevance to tissue-specific transcription regulatory programs. FIREs are also conserved between human and mouse. In addition, FIREs have been revealed to occur near cell-identity genes and active enhancers [71]. Thus, FIREs have proven valuable in identifying tissue and cell type-specific regulatory regions, functionally conserved regions such as enhancers shared by human and mouse, and in interpreting genetic variants associated with human complex diseases and traits [71, 8, 29].

Since the discovery of FIREs, we have collaborated with multiple groups to further demonstrate their values in various applications, resulting in multiple recent preprints and publications [29, 33, 31, 28]. For example, in an analysis of adult and fetal cortex Hi-C datasets, FIREs and super-FIREs recapitulated key functions of tissue-specificity, such as neurogenesis in fetal cortex and core neuronal functions in adult cortex [28]. In addition, evolutionary analyses revealed that these brain FIRE regions have stronger evidence for ancient and recent positive selection, less population differentiation, and fewer rare genetic variants [28]. For another example, Gorkin et al. [29] investigated how 3D chromatin conformation in lymphoblastoid cell lines (LCL) varies across 20 individuals. They reported that FIREs are significantly enriched in LCL-specific enhancers, super-enhancers, and immune related biological pathways and disease ontologies, further demonstrating the close relationship between FIREs and *cis*-regulatory elements [29]. In particular, even with the sample size of $\leqslant 20$ individuals, hundreds of FIRE-QTLs (that is, genetic variants associated with the strength of FIRE) have been reported, suggesting that FIREs show strong evidence of genetic regulation.

Despite the importance and utilities of FIREs, only in-house pipelines exist for detecting FIREs, limiting the general application of FIRE analysis and the full exploration of cell type-specific chromatin spatial organization features from Hi-C data. In this work, we developed

FIREcaller, a stand-alone, user-friendly R package for detecting FIREs from Hi-C data as an implementation of the method described in our previous work [71].

2.2 Materials and Methods

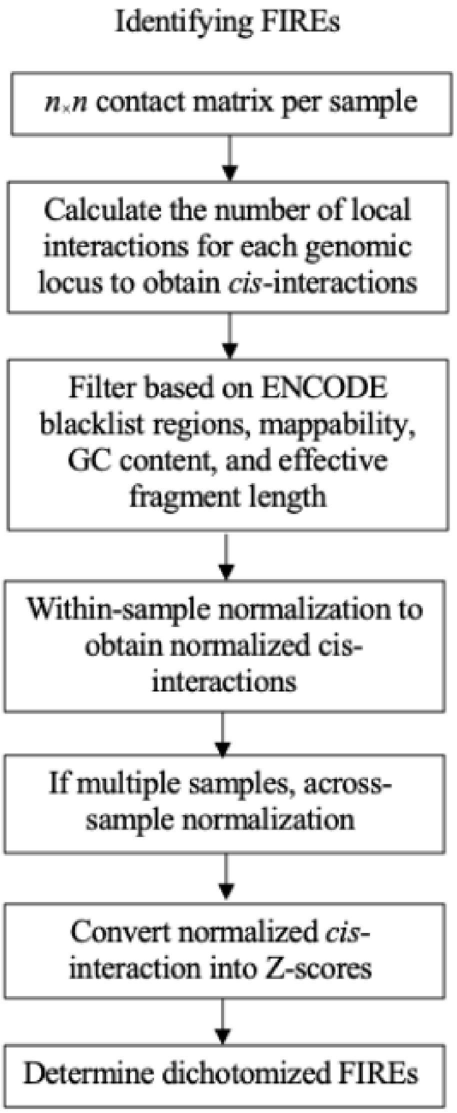

Identifying FIREs
n×n contact matrix per sample
Calculate the number of local interactions for each genomic locus to obtain cis-interactions
Filter based on ENCODE blacklist regions, mappability, GC content, and effective fragment length
Within-sample normalization to obtain normalized cis-interactions
If multiple samples, across-sample normalization
Convert normalized cis-interaction into Z-scores
Determine dichotomized FIREs

Figure 2.1: Workflow of Calling FIREs using the FIREcaller Software

2.2.1 Input Matrix

First, FIREcaller takes an *n*x*n* Hi-C contact matrix as input. The contact matrix *M* is constructed by dividing the genome into bins of size *b* consecutive non-overlapping regions for each chromosome. In our original work [71], *b* was fixed at 40Kb. In this FIREcaller work, we allow *b* to be 10Kb, 20Kb, or the default 40Kb. Each entry in the contact matrix *M*, $m_{i,j}$, corresponds to the number of reads mapped between locus/bin *i* and locus/bin *j*. The corresponding symmetric *n*x*n* matrix reflects the number of mapped intra-chromosomal reads between each pair of loci [48]. We removed all intra-chromosomal contacts within 15Kb to filter out reads due to self-ligation.

Recommendations for the resolution of the input matrix depend on the sequencing depth of the input Hi-C data. Specifically, we recommend using a 10Kb bin resolution for Hi-C data with $\sim$ 2 billion reads, and a 40Kb bin resolution for Hi-C data with $\sim$ 500 million reads [48, 67, 37, 72, 45, 97].

2.2.2 *Cis*-Interaction Calculation

Taking the *n*x*n* contact matrix as input, FIREcaller calculates the total number of local *cis*-interactions for each genomic locus (40Kb bin size by default). Following our previous work [71], we define local to be within 200Kb by default. This threshold is largely driven by empirical evidences showing that contact domains exert influences on transcription regulation within 200Kb. For instance, contact domains reported in human GM12878 from in-situ Hi-C are at a median size of 185Kb [67, 97]. In addition, Jin et al. reported a median distance of E-P interactions at 124Kb [37], Song et al. reported 80% of promoter interacting regions within 160Kb [73], and Jung et al. found promoter centered long-range chromatin interactions with median distance 158Kb [38]. Consistently, an analysis of the dorsolateral prefrontal cortex sample (DLPFC) [46] showed E-P interactions at a median distance of 157Kb, and our study showed adult cortex E-P interactions at a median distance of 190Kb [28]. On the other hand, multiple *cis*-regulatory regions have been shown to control their target genes from longer genomic distances

[47, 28, 97, 24]. To accommodate these longer-range chromatin interactions our FIREcaller package allows a user-specified upper bound of the *cis*-interacting region.

2.2.3 Bin Level Filtering

Bins are then filtered based on multiple criteria that may lead to systematic biases, including effective restriction fragment lengths which measures the density of the restriction enzyme cut sites within each bin, GC content, and sequence uniqueness [34, 93]. FIREcaller removes bins with 0 mappability, 0 GC content, or 0 effective fragment length. We also remove bins for which more than 25$\%$ of their neighborhood (within 200Kb, by default) bins have 0 mappability, 0 GC content or 0 effective fragment length. Finally, any bins overlapped within the MHC (major histocompatibility complex) region or the ENCODE blacklist regions [1] are also filtered out.

2.2.4 Within-sample Normalization

FIREcaller then uses the HiCNormCis method [71] to conduct within-sample normalization. HiCNormCis adopts a Poisson regression approach, adjusting for the three major sources of systematic biases: effective fragment length determined by restriction enzyme cutting frequency, GC content, and mappability [71].

For HiCNormCis, we let U_i, F_i, GC_i, and M_i represent the total *cis*-interactions (15 $\sim$ 200Kb, by default), effective fragment length, GC content, and mappability for bin *i*, respectively. We assume that U_i follows a Poisson distribution, with mean θ_i, where $\log(\theta_i) = \beta_0 + \beta_F F_i + \beta_{GC} GC_i + \beta_M M_i$. After fitting the Poisson regression model, we define the residuals *R*$_i$ as the normalized *cis*-interaction for bin *i*.

FIREcaller fits a Poisson regression model by default. Users can also fit a negative binomial regression model. In practice, both Poisson regression and negative binomial regression model achieved similar effect of bias removal, while Poisson regression is computationally more efficient.

2.2.5 Across-sample Normalization

If the user provides multiple Hi-C datasets, FIREcaller uses the R function *normalize.quantiles* in the ”preprocessCore” package to perform quantile normalization of the normalized *cis*-interactions across samples [6].

2.2.6 Identifying Significant Frequently Interacting Regions

FIREcaller then converts the normalized *cis*-interactions into Z-scores, calculates one-sided p-values based on the standard normal distribution, and classifies bins with p-value < 0.05 as FIREs. The output file contains, for each bin, the normalized *cis*-interactions, the -ln(p-value) (i.e., the continuous FIREscore), and the dichotomized FIRE or non-FIRE classification.

2.2.7 Detecting Super-FIREs

FIREcaller also identifies clustered FIREs, termed as super-FIRE (Figure 2.2). We first concatenate all consecutive FIRE bins, and sum their continuous FIREscores. We then plot the summed continuous FIREscores for the clustered FIREs against their ranks to identify the inflection point where the slope of the tangent line is one. Super-FIREs are defined as all clustered FIREs on the right of the inflection point [71]. This method is adapted from the Ranking of Super Enhancer (ROSE) algorithm [88], which was originally proposed for the identification of super-enhancers.

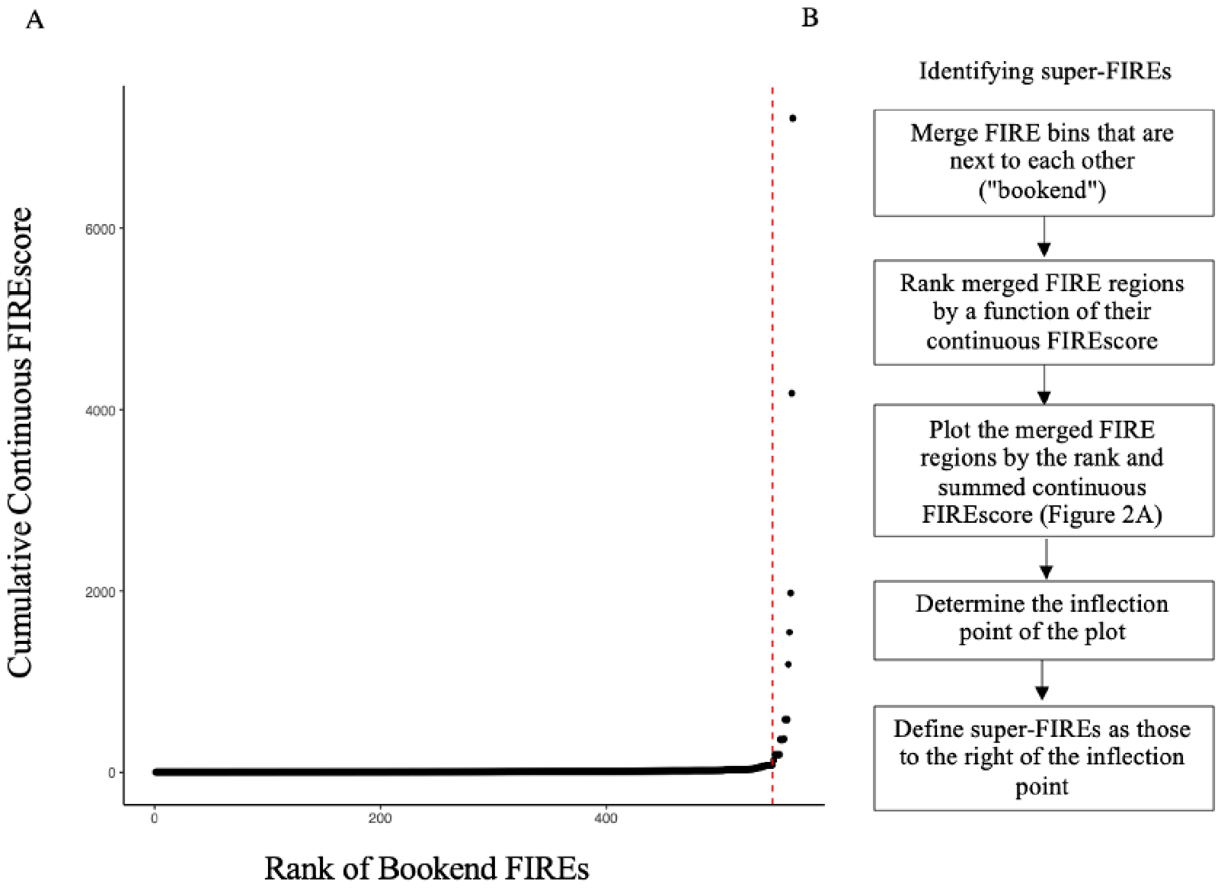


Figure 2.2: Flow Chart for Identifying Super-FIREs. A) Scatterplot of clustered FIREs ranked by their continuous FIREscores, ordered from least interactive (left) to most interactive (right). Red dashed line highlights the inflection point of the curve. B) Flow chart for super-FIRE identification.

2.3 Results

2.3.1 An Illustrative Example

We used the Hi-C data from human hippocampus tissue in our previous study [71] to showcase the utility of FIREcaller. Figure 2.2 shows an illustrative example of a 400Kb super-FIRE (merged from 10 consecutive bins, and marked by the red horizontal bar in the ”FIREs” track), which overlaps with two hippocampus super-enhancers (indicated by the two light blue horizontal bars in the ”Enhancers” track). Notably, this super-FIRE contains a schizophrenia-associated GWAS SNP rs9960767 (black vertical line) [74], and largely overlaps with gene *TCF4* (chr18:

52,889,562-53,332,018; pink horizontal bar depicted at the top), which plays an important role in neurodevelopment [34]. Since rs9960767 resides within a super-FIRE with highly frequent local chromatin interactions, we hypothesize that chromatin spatial organization may play an important role in gene regulation in this region, elucidating potential mechanism by which rs9960767 affects schizophrenia risk.

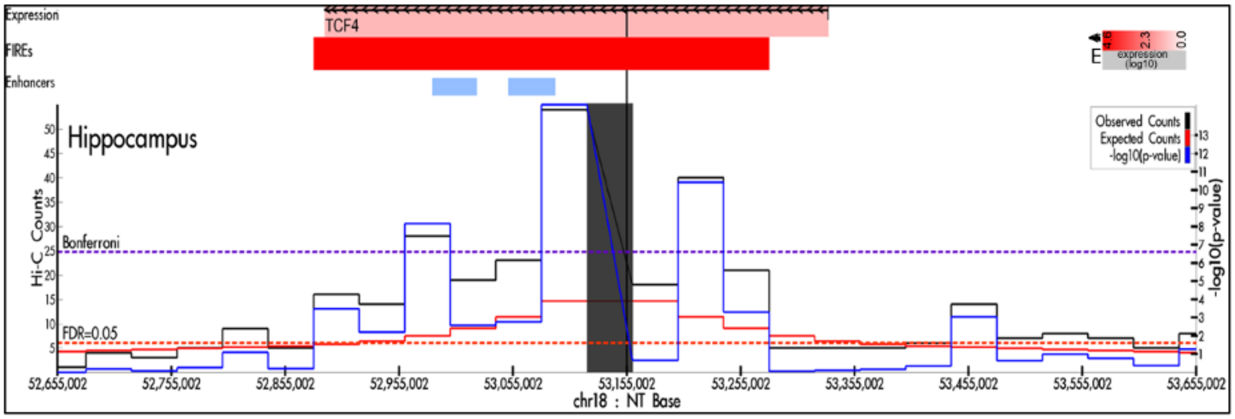


Figure 2.3: An Example of a Super-FIRE in Human Hippocampus tissue. Virtual 4C plot of a 1Mb region (chr18:52,665,002-53,665,002) anchored at the schizophrenia-associated GWAS SNP rs9960767 (black vertical line), visualized by HUGIn [57]. The solid black, red and blue lines represent the observed contact frequency, expected contact frequency, and -log10(p-value) from Fit-Hi-C [3], respectively. The dashed purple and red lines represent significant thresholds corresponding to Bonferroni correction and 5 % false discovery rate (FDR), respectively. The red horizontal bar in the ”FIREs” track depicts the 400Kb super-FIRE region. The two blue horizontal bars in the ”Enhancers” track mark the two hippocampus super-enhancers in the region.

2.3.2 Integrative Analysis of FIREs with Gene Expression in Human Brain Tissue

To investigate the relationship between FIREs and tissue-specifically expressed genes, we applied FIREcaller to Hi-C data from fetal [90] and adult [85] cortical tissues, and identified 3,925 fetal FIREs and 3,926 adult FIREs. Among them, 2,407 FIREs are fetal-specific and 2,408 FIREs are adult-specific (the remaining 1,518 FIREs are shared). We observed massive changes in FIREs between fetal and adult which recapitulate recently reported extensive chromatin rewiring during brain development [85], exemplifying the cell type-specific nature of FIREs. We then overlapped FIREs with gene promoters and found that the dynamics of FIREs across brain developmental stages are closely associated with gene regulation dynamics during brain development (Figure 2.4). Specifically, we examined expression levels of genes whose promoters (defined as ± 500 bp of transcription start site [TSS]) overlap with fetal brain-specific FIREs and are expressed in fetal brain, similarly genes whose promoter overlap with adult brain-specific FIREs and are expressed in adult brain. Gene expression data in both fetal and adult brain cortex are from two of our recent studies [90, 85]. These criteria resulted in 707 and 882 genes in fetal and adult brain, respectively. Among them, 412 are fetal brain-specific, 587 are adult brain specific, and 295 genes are shared (Table 2.1).

FIRE type	N FIREs	N FIREs with gene overlap	N genes with FIRE overlap
Adult-specific	2,408	488	587
Fetal-specific	2,407	338	412
Shared	1,518	258	295

Table 2.1: Tissue-Specific FIREs and Shared FIREs, and Overlapping Genes.

For the 587 genes mapped to adult brain-specific FIREs, the mean gene expression levels, measured by log2(FPKM), are -0.052 and 0.190 in fetal and adult brain cortex, respectively. These 587 genes are significantly up-regulated in adult brain (p-value= $1.3 * 10^{-10}$; Figure 2.4; Table 2.2). Meanwhile, for the 412 genes mapped to fetal brain-specific FIREs, the mean gene expression levels, again measured by log2(FPKM), are 0.551 and 0.209 in fetal and adult brain cortex, respectively. These 412 genes are significantly up-regulated in fetal brain (paired t-test p-value= $7.8 * 10^{-13}$; Figure 2.4; Table 2.2). By contrast, for the 295 genes co-localizing with FIREs shared between fetal and adult cortex, the mean gene expression levels are 0.328 and 0.312 in fetal and adult brain cortex, respectively. These 295 genes show no significant difference in their expression levels in adult and fetal brain (p-value=0.79). Similarly and finally, genes not overlapping any FIREs exhibit no significant expression differences in fetal and adult brains either (p-value=0.96; Figure 2.4; Table 2.2).

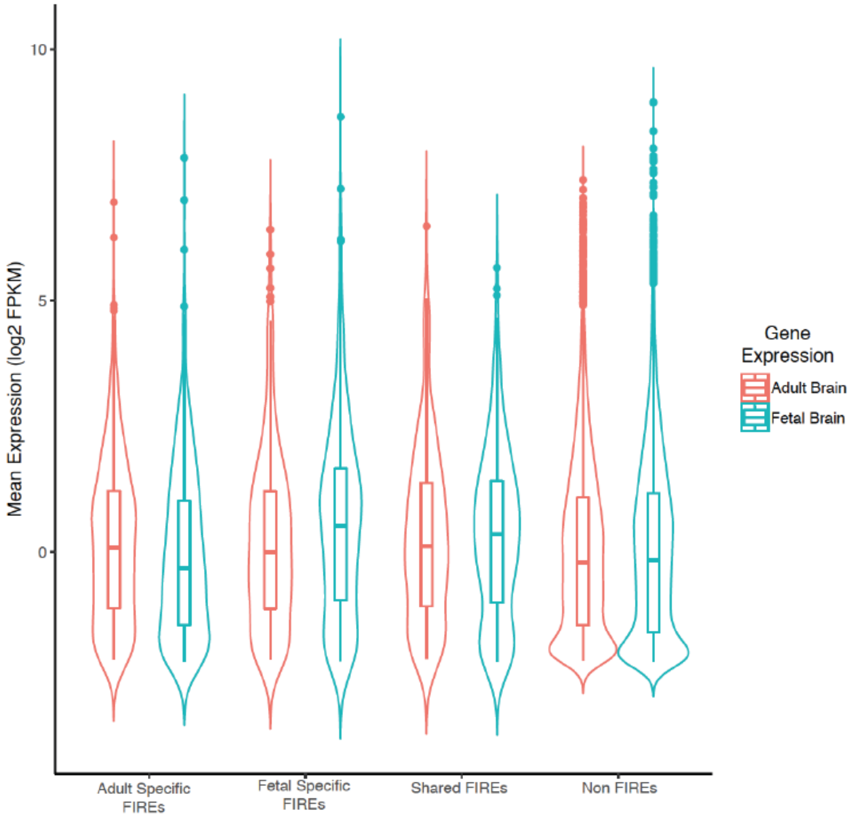


Figure 2.4: Distribution of Expression for Genes Overlapping Fetal or Adult Brain FIREs. The leftmost pair of violin boxplots shows the expression profile of the 587 genes mapped to adult brain-specific FIREs, with expression measured in fetal brain cortex (blue) and adult brain cortex (red), respectively. The second pair of violin boxplots shows the expression profile of the 412 genes mapped to fetal brain-specific FIREs, again in fetal brain cortex (blue) and adult brain cortex (red), respectively. The third pair shows the expression profile of the 295 genes mapped to FIREs shared between fetal and adult brain, yet again in fetal brain cortex (blue) and in adult brain cortex (red). The pair to the farthest right, shows the expression profile of genes not overlapping any FIREs, with a total of 15640 such genes (labeled ”Non FIREs”).

FIRE Type	N Genes	Fetal Brain Gene Expression Mean	Adult Brain Gene Expression Mean	P-Value
Adult-Specific	587	0.052	0.190	$1.26*10^{-10}$
Fetal-Specific	412	0.551	0.209	$7.79*10^{-13}$
Shared	295	0.328	0.312	0.785

Table 2.2: Mean Gene Expression in Adult and Fetal Brain Samples for FIREs and Non-FIREs and Paired T-Test P-Values

2.3.3 Integrative Analysis of FIREs and Enhancer-Promoter Interactions

We used Hi-C data from left ventricle and liver tissues from Schmitt et al study [71], and applied Fit-Hi-C [3] to call significant chromatin interactions at 40Kb bin resolution. We only considered bin pairs within 2Mb distance. Next, we used H3K27ac ChIP-seq peaks [44] in left ventricle and liver tissue to define active enhancers, and used 500 bp upstream / downstream of TSS to define promoters. A pair of 40Kb bins is defined as an E-P interaction if one bin contains a promoter, and the other bin contains an active enhancer. In total, at an FDR of 1%, we identified 41,401 and 30,569 E-P interactions in left ventricle and liver, respectively. Among them, 29,096 are left ventricle-specific, and 18,264 liver-specific.

At the same 40Kb bin resolution, applying our FIREcaller, we identified 3,643 FIREs in left ventricle and 3,642 FIREs in liver, with 1,186 shared between these two tissues. We found that FIREs are enriched for E-P interactions compared to non-FIREs for both liver and left ventricle (liver: odds ratio [OR] = 7.2, Fisher's exact test p-value $< 2.2 * 10^{-16}$; left ventricle: OR = 4.0, p-value $< 2.2 * 10^{-16}$). Comparing between two tissues, we observed that left ventricle-specific E-P interactions are highly enriched in left ventricle-specific FIREs and liver-specific E-P interactions highly enriched in liver-specific FIREs (OR=3.8, p-value $< 2.2 * 10^{-16}$; Table 2.3). Our results demonstrate that the tissue-specificity of FIREs is closely associated with the tissue-specificity of E-P interactions [71].

FIRE type	N Left Ventricle-Specific E-P	N Liver-Specific E-P
Left Ventricle-Specific FIRE	1,093	416
Liver-Specific FIRE	951	1,392

Table 2.3: Tissue-Specific FIREs and Tissue-Specific E-P Interactions in Liver and Left Ventricle Tissues. In the table, we count the numbers of tissue-specific E-P interactions involving tissue-specific FIREs. For example, 1,093 means there are 1,093 left ventricle-specific E-P interactions involving left ventricle-specific FIREs. Similarly for the remaining three counts.

2.3.4 Integrative Analysis of FIREs and ChIP-seq Peaks

Next, we evaluated the relationship between FIREs and histone modifications in cortex [46, 85, 44]. We found that H3K4me3 and H3K27ac ChIP-seq peaks are both enriched at FIRE regions (Figure 2.5).

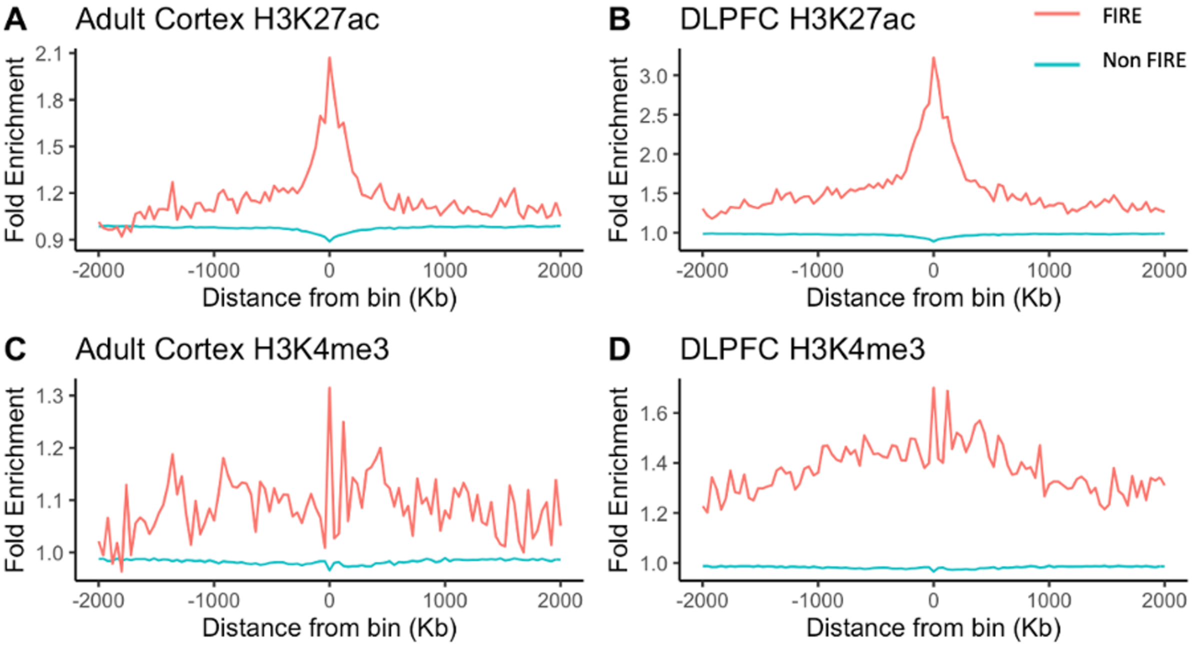


Figure 2.5: H3K4me3 and H3K27ac ChIP-seq Peaks Enrichment at FIREs. X axis is the distance from a bin, with the bins grouped into FIRE bins and non FIRE bins. Y axis is fold enrichment quantified by MACS [98] when applied to the corresponding histone ChIP-seq data.

2.4 Conclusion

In this paper, we present FIREcaller, a user-friendly R package to identify FIREs from Hi-C data. We demonstrate its utilities through applications to multiple Hi-C datasets and integrative analyses with E-P interactions, histone modifications and gene expression. We believe that FIREcaller will become a useful tool in studying cell type or tissue-specific chromatin spatial organization.

CHAPTER 3: ANALYSIS OF CHROMOSOME SPATIAL ORGANIZATION IN ADULT AND FETAL CORTEX SAMPLES

3.1 Introduction

Investigating the genetic architecture can be particularly useful for idiopathic psychiatric disorders, such as schizophrenia, as few genetic studies have resulted in proved or reproducible risk factors [76, 65]. Identifying genetic variants influencing psychiatric disorders has become complex as it has been found to be consistent with a polygenic model [75, 26]. Additional factors of complexity involve the diagnosis of these disorders as they are primarily based on observations of behavior, as well as the notion that different psychiatric disorders overlap, often resulting in re-evaluation. Overall, there is a sizeable clinical importance as psychiatric disorders have a relatively large impact, ranking fifth in global disability.

A majority of variants from GWAS studies of psychiatric disorders map to non-coding regions and are highly correlated [77]. These areas are difficult to directly assign variants to genes as the regions have diverse regulatory functions. Incorporating various genetic functional data can provide insight into the relationship between regulatory regions and genes of interest. Chromatin conformation capture (3C) methods [81] enables the identification of 3D chromatin interactions in vivo and can clarify GWAS findings [75, 62]. Here, we display the pipeline and methodology to transform the resulting fragments from Hi-C of adult and fetal brain cortex into A/B compartments [48], topologically associated domains (TADs) [20], frequently interacting regions (FIREs) [71], and long-range chromatin interactions (LRCI) [3]. In addition, we compare the 'readouts' to other external Hi-C data sources [71, 90], and evaluate potential biological connections. The analysis of the cortex samples chromatin structure was further incorporated into projects spearheaded by Dr. Paola Giusti-Rodríguez and Dr. Patrick Sullivan. More specifically, to understand the complexity of psychiatric GWAS, gene set analysis was conducted based on functional genomic data to provide an expanded view of the biological processes involved in the etiology of schizophrenia and other complex brain traits, which will also be briefly discussed [28].

3.2 Methods

3.2.1 Data Available and Processing

The data consisted of adult (N=3) and fetal (N=3) cortex tissue samples sequenced using ’easy Hi-C’ [52]. The library quality and yield from eHi-C are comparable to conventional Hi-C, but requires much less starting material. Initially, there were 10.4 billion reads generated to allow for a kilobase resolution contact frequency map of the chromatin interactome. For a tissue, this is deeply sequenced. The data was delivered in the form of a bam file, and quality control was performed to include uniquely mapped reads, and exclude inter-chromosomal reads, intra-chromosomal reads less than 15Kb, and PCR-duplicates through various computational procedures. The resulting interactions were summarized in contact matrices of 10Kb, 40Kb, 100Kb, and 1Mb resolution.

3.2.2 Compartments

Compartments were determined by conducting Principal Component Analysis (PCA) incorporated with tissue-specific gene expression at 100Kb and 1Mb resolution. Beginning with the raw *n*x*n* contact matrix, we calculated the genome-wide mean for each fixed 1D genomic distance and evaluated the contact probability for each sample-chromosome pair. As shown in Figure 3.1, the contact probability decreased with genomic 1D distance as expected with some deviation at the tail of the plot. This deviation implies the loci are interacting more often than expected, which could occur due to a possible relationship or grouping.

We transformed the raw nxn matrix into a normalized matrix using the 1D genomic mean. The normalized matrix was then transformed into a correlation matrix by incorporating Pearson correlations for each i^{th} row and j^{th} column.

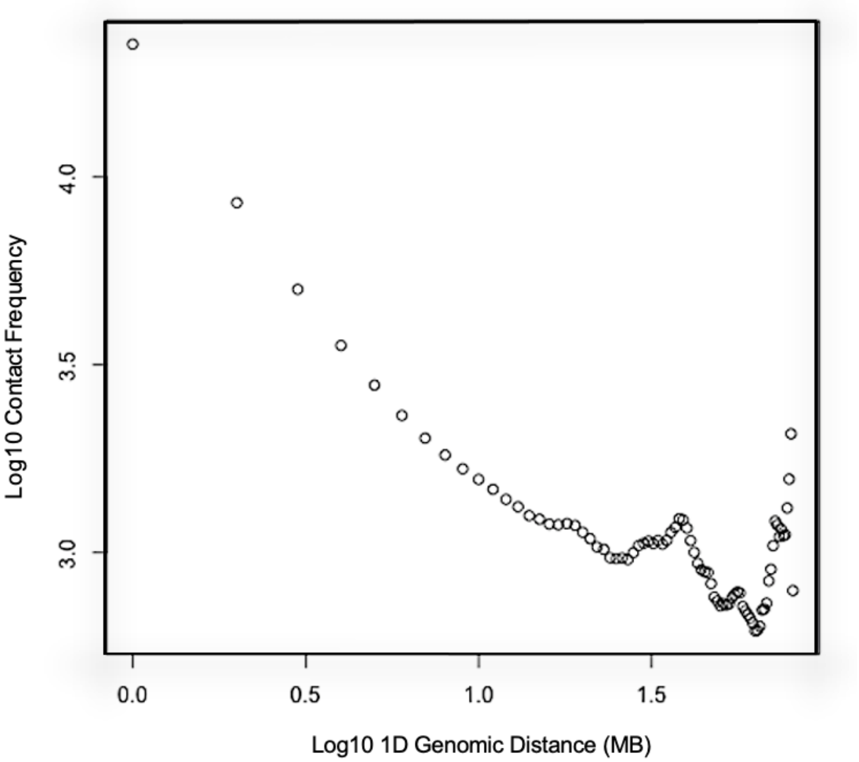


Figure 3.1: Contact Probability Plot at 1Mb for Adult Cortex Chromosome 17

PCA was conducted on the correlation matrix, which resulted in the first three PC's for each 1D genomic location. Beginning with PC1, compartments were loosely defined as 'A' for a positive PC1 value and 'B' for a negative PC1 value. Next, we incorporated the tissue-specific gene expression, and further defined the compartments where if the average gene expression of the 'A' compartments were less than the average gene expression of the 'B' compartments, the compartment designation switched.

Compartments for each chromosome-sample pair were evaluated individually to ensure the correct PC captured the enrichment and depletion of the long arm of the chromosome. Figure 3.2 displays the PC1 change of sign corresponding to the enriched and depleted sections of the correlation matrix heatmap, confirming the selection of the correct PC to determine the compartments.

In the case that PC1 did not accurately reflect the enrichment and depletion visualized by the correlation matrix, PC2 was examined. Post-processing of the compartments and the PC's were conducted by filtering bins that corresponded to the ENCODE blacklist [1] or the MHC region.

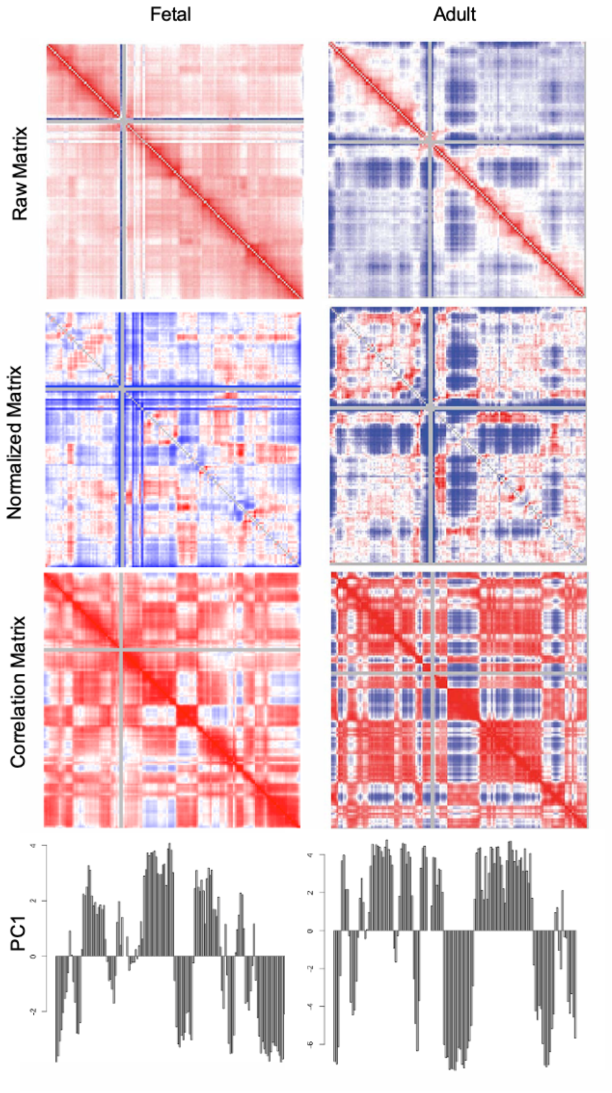


Figure 3.2: Comparison of Correlation Matrix and PC1 at 1Mb for Adult and Fetal Cortex Chromosome 17

3.2.3 TADs

TADs and TAD boundaries were detected by implementing the insulation score method [15]. TAD boundaries were called at a 10Kb and 40Kb resolution and involved a four-step process: (1)

Calculating the Insulation Scores, (2) Normalizing the Insulation Score, (3) Defining the TAD Boundaries, (4) Defining TADs.

Beginning with a raw *n*x*n* contact frequency matrix, HiCNorm [34],which removes biases of Hi-C data by using a Poisson regression, was applied to normalize the chromatin contact map. Next, each bin was treated as an anchor bin independently, and an insulation score was calculated by summing the bins between the anchor bin and all other bins ± 100Mb away in 1D genomic distance. Quantile normalization was performed on all the resulting insulation scores. These normalized insulation scores were then evaluated, and the minima in the ± 1 Mb neighborhood was determined as a TAD boundary at the bin size resolution previously specified (10Kb, 40Kb). The regions between boundaries were deemed TADs.

3.2.4 FIREs and Super-FIREs

Using the statistical software R and the FIREcaller package, FIREs and super-FIREs were identified at resolutions 10Kb and 40Kb. These results implemented the default filtering based on mappability, GC content, and fragment length. A final post-processing of the FIREs was conducted by filtering bins that corresponded to the ENCODE blacklist.

3.2.5 Long-Range Chromatin Interactions

Long-range chromatin interactions were discovered by using a combination of Fit-Hi-C [3] with default parameters along with FastHiC [47]. This combination calling method for long-range interactions mimicked the approach in Won, et al. [90]. At 10Kb resolution, the data was filtered by removing bins that corresponded to the ENCODE blacklist. We ran Fit-Hi-C for all 10Kb regions with restrictions on interactions $\geqslant$ 20Kb and $\leqslant$ 2Mb. Using these results, we then applied FastHiC to the 43,222,677 bin pairs with a Bonferroni correction corresponding to $P < 1.16319^{-10}$. The resulting chromatin interactions represented the 3D peaks presented in the data.

3.2.6 Comparison of Data with Other Publicly Available

With various biases involved and the different pipelines for data analysis, confirming similarities within our samples and with external data sources were deemed essential for reproducibility and confirming the quality of the data. L1-Norm, L2-Norm, Spearman, and Pearson correlations were used as an ad-hoc metric due to the lack of accountability of the dependency nature of Hi-C data. HiC-Spector [94] and HiCRep [95] were used to compare our contact maps as they take additional measures such as spectral decomposition, and smoothing parameters, respectively. MDS plots were used to visualize the results. PCA and Jaccard Similarity Index were evaluated for comparisons with external Hi-C datasets of other fetal cortex [90], and 14 human tissues and 7 cell lines [71].

3.3 Results

3.3.1 General Properties of Chromatin Organization

In this section, we describe the brain Hi-C data we generated and comparison to external datasets to establish their relevance. We generated 10.4 billion reads and, following quality control, identified 1.323 billion high-confidence *cis*-contacts that enabled a 10 Kb resolution map of the chromatin interactome. To our knowledge, these are the deepest Hi-C data from human brain, and contain 2.25X as many *cis*-contacts for adult cortex and 1.56X as many for fetal cortex compared to the next largest datasets [71, 90] (Figure 3.3). Our adult and fetal samples contained the most unique mapped reads (2,131,884,928 and 2,918,232,26, respectively) and the most intra-chromosomal reads (713,524,167 and 609,291,151, respectively) in this collection of Hi-C data.

In Figure 3.3, the point sizes are proportional to the number of informative *cis*-reads represented by the increase in point size along the y-axis. The data from our generation of the human brain Hi-C is represented by the red diamonds. The open red diamond (Fetal) representing the 3 concatenated fetal samples, and the filled red diamond (Adult) representing the 3 concate-

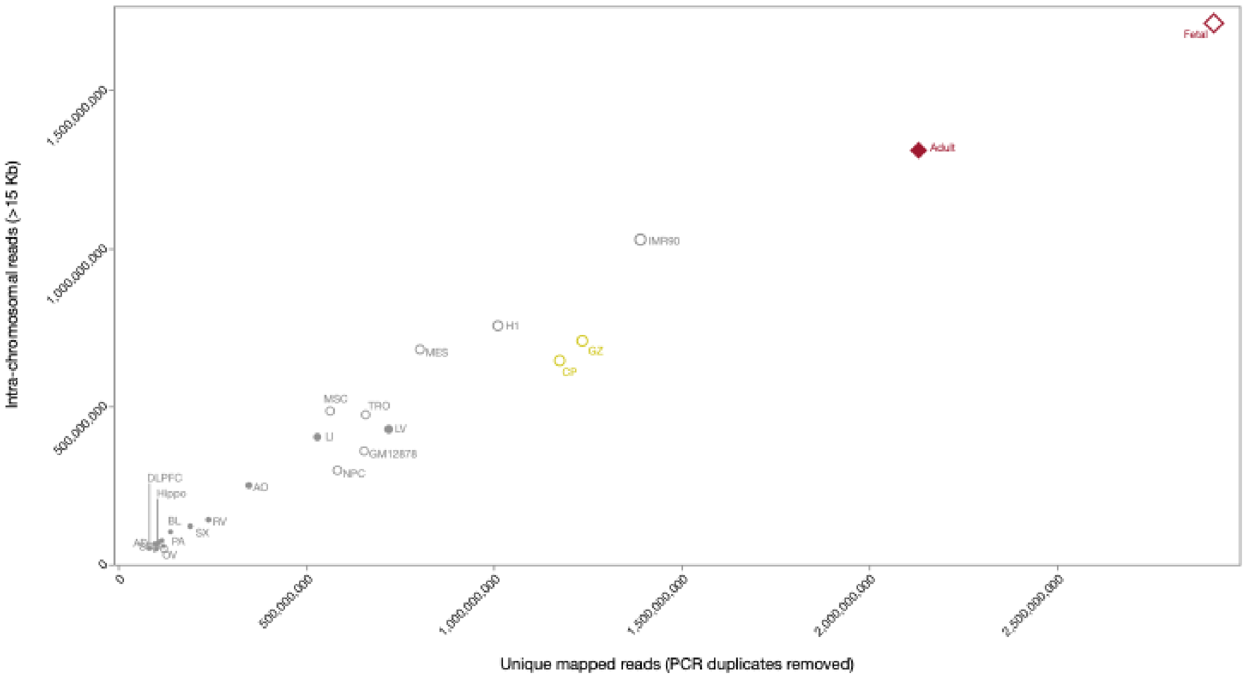


Figure 3.3: Comparison of Hi-C Metrics with External Datasets

nated adult temporal cortex samples. Data from Schmitt et al is represented with the grey circles, with the 7 cell lines visualized with the open grey circles (GM12878=lymphoblast; H1=human embryonic stem cell (hESC); IMR90=lung fibroblast; MES=mesoderm; MSC=mesenchymal stem cell; NPC=neural progenitor cell; TRO=trophoblast-like cell) and the 14 tissues visualized by the filled grey circles (AD=adrenal gland; AO=aorta; BL=bladder; DLPFC=brain dorsolateral prefrontal cortex; Hippo=brain hippocampus; LG=lung; LI=liver; LV=heart left ventricle; OV=ovary; PA=pancreas; PO=psoas skeletal muscle; RV=heart right ventricle; SB=small intestine; SX=spleen). The open yellow circles represent the Won et al data where the fetal germinal zone (GZ) representing the concatenation of the 3 GZ samples, and the cortical plate (CP) representing the concatenation of the 3 CP samples.

3.3.1.1 Contact Maps

We evaluated our Hi-C datasets (N=3 adult temporal cortex, N=3 fetal cortex, and 2 combined samples of all adult and all fetal) by conducting Pearson correlation, Spearman correlation, HiC-Spector, and HiCRep on the raw counts at 40Kb resolution (Figure 3.4). The comparison methods that account for the characteristics of Hi-C data, HiCRep and HiC-Spector, showed clear separation of the two developmental times points, where the blue dots correspond to the adult temporal cortex samples (N=3 individual +1 combined) and red corresponds to the fetal cortex samples (N=3 individual +1 combined).

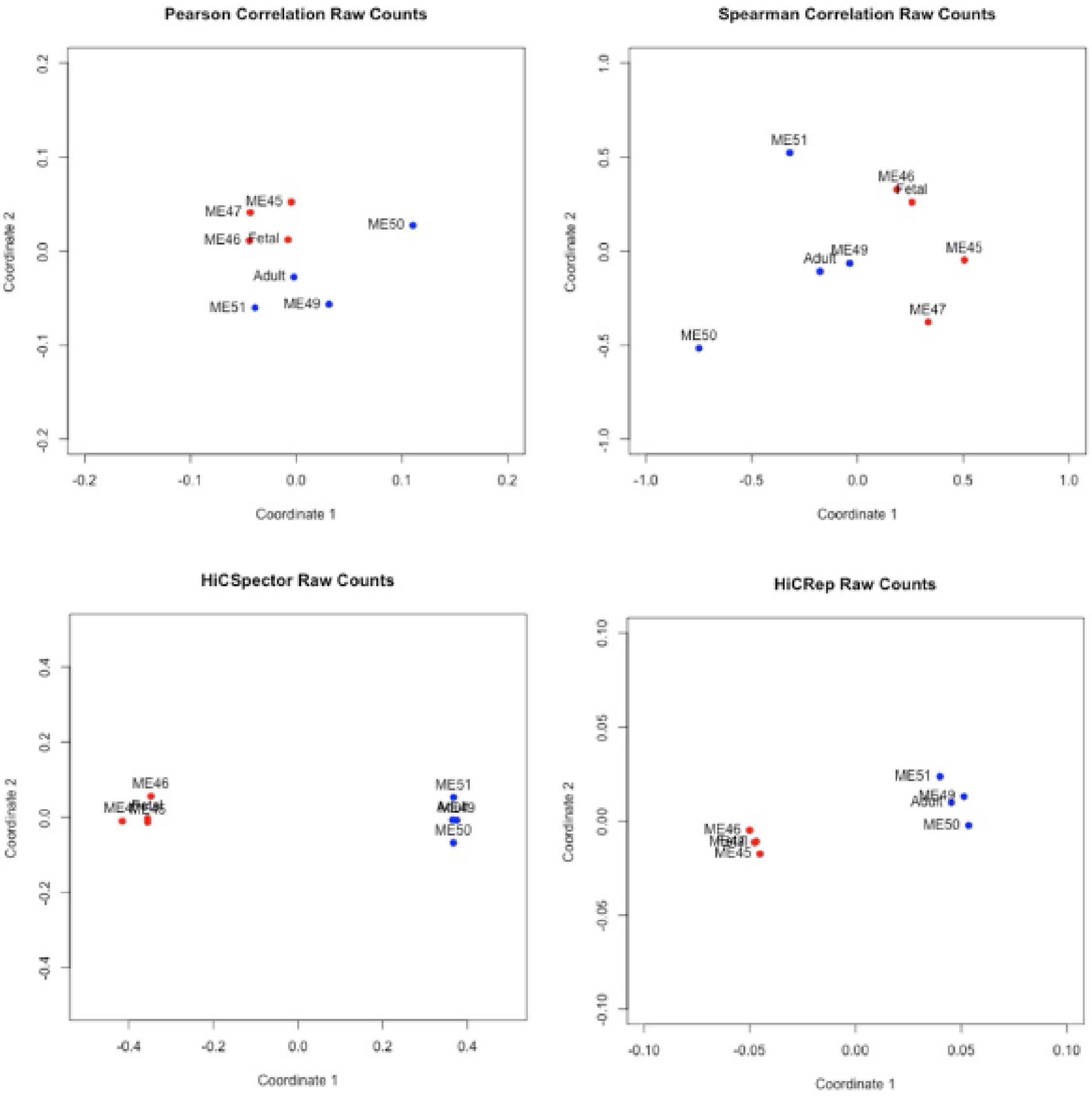


Figure 3.4: Autosomal Chromosomes Contact Frequency Matrix Comparison at 40Kb Resolution for Fetal (blue) and Adult (red) samples

Fetal Compartment	Adult Compartment	N 100Kb bins (%)	N 1Mb bins (%)
A	A	10275(36.7)	1039(36.2)
B	A	2713(9.7)	271(9.4)
A	B	2265(8.1)	196 (6.8)
B	B	12754(45.5)	1365(47.5)

Table 3.1: Compartments of Fetal and Adult Cortex at 100Kb and 1Mb

3.3.1.2 Compartments

Compartments were defined at 100Kb and 1Mb resolution. When comparing the bins defined as A compartments and B compartments between samples, we observe consistency at both resolutions (100Kb:χ^2 p-value $< 2.2 * 10^{-16}$; 1Mb: χ^2 p-value $< 2.2 * 10^{-16}$; Table 3.1)

We compared our results with external datasets at 100Kb resolution. The PC1 was evaluated using PCA (Figure 3.5) and compartment designation was evaluated by the Jaccard Index Similarity (Figure 3.6). The analysis used the PC1 scores integrated with the tissue-specific gene density that defined the A/B compartments. The PC1 (x-axis) had a variance of 74.8 % and PC2 had a variance of 8% (y-axis). The fetal brain samples (CP,GZ, Fetal) all clustered together, and the adult brain samples (Hippo, DLPFC, and Adult) all clustered together. The Jaccard Similarity Index analysis displayed a clustering via the heatmap to describe the degree of overlap of the A/B compartments. The fetal brain samples clustered together (Fetal, GZ, CP; similarity index of 0.83) and had strong overlap. The adult brain samples cluster together (Adult, DLPFC, Hippo; similarity index of 0.83) with a strong overlap. The adult versus fetal brain samples Jaccard Similarity Index was 0.66, which for reference the Jaccard Similarity Index for left heart ventricle versus right heart ventricle was 0.91, as compartments do exhibit moderate cell type-specificity.

3.3.1.3 TADs

There were a total of 1,908 fetal TADs and 2,069 adult TADs identified at a 40Kb resolution. Each sample had 2056 TAD boundaries, with 1340 boundaries that had a 100% overlap. TADs are calculated based on their fixed-bin sized boundary, thus the average TAD length varied. The average adult TAD length was 1.2Mb, and the average fetal TAD length was 1.3 Mb.

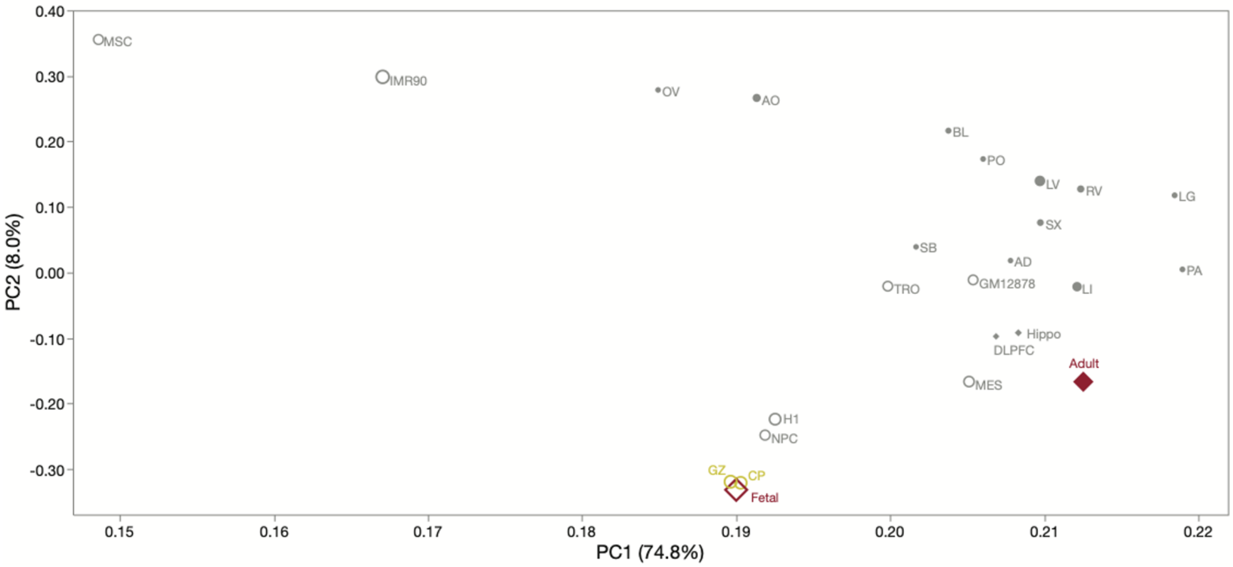


Figure 3.5: PCA Plot Comparing PC1 with External Data at 100Kb

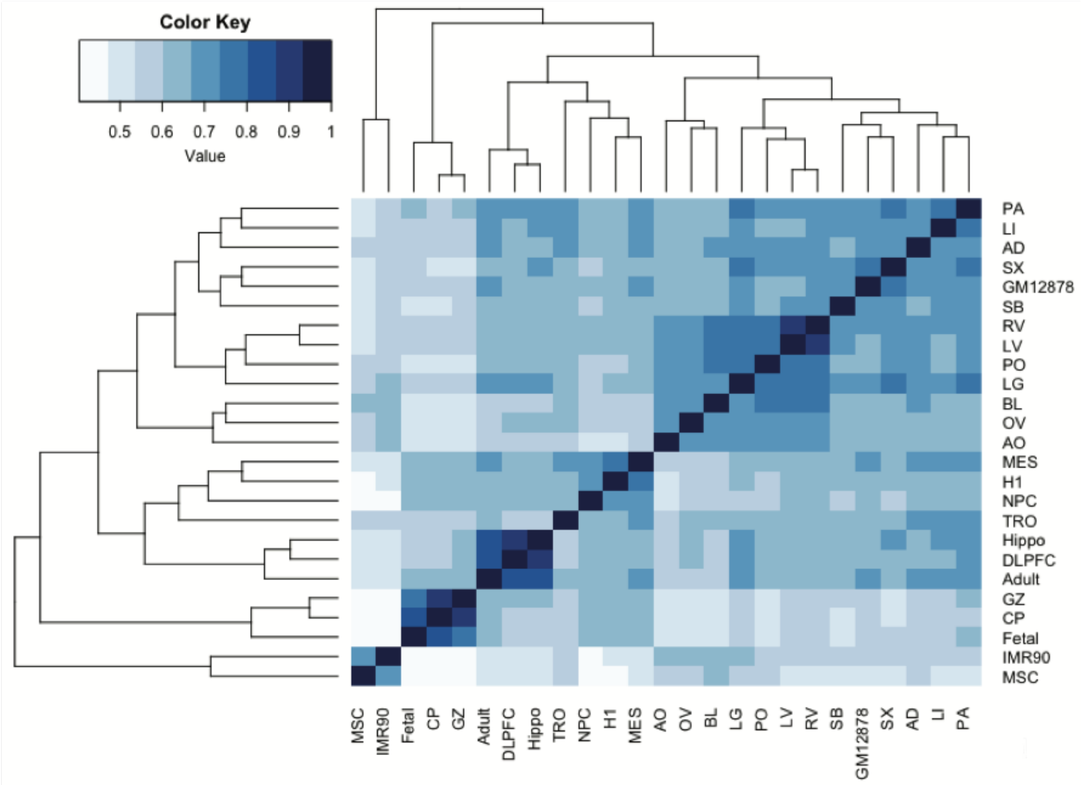


Figure 3.6: Jaccard Similarity Plot Comparing A/B Compartments with External Data at 100Kb

When comparing the PCA of the TAD insulation scores at 40Kb resolution with external datasets (Figure 3.7) there was a 76.6% variance with PC1 (x-axis), and PC2 (y-axis) captured

5.0% variance. The brain samples (Adult, Fetal, CP, GZ) are all clustered on the dominant PC1. Note the other brain samples, DLPFC and Hippos, are not as clustered which could be attributed to a lower sequencing depth. A similar relationship is visible with the clustered heatmap of the Jaccard Similarity Index (Figure 3.8) showing the degree of overlap of the TAD boundaries. Our adult and fetal TAD boundaries clustered very strongly with a Jaccard Index of 0.41 (for comparison, the Jaccard Index for TAD boundaries in left and right heart ventricle was 0.48), which is expected as TADs are cell type invariant.

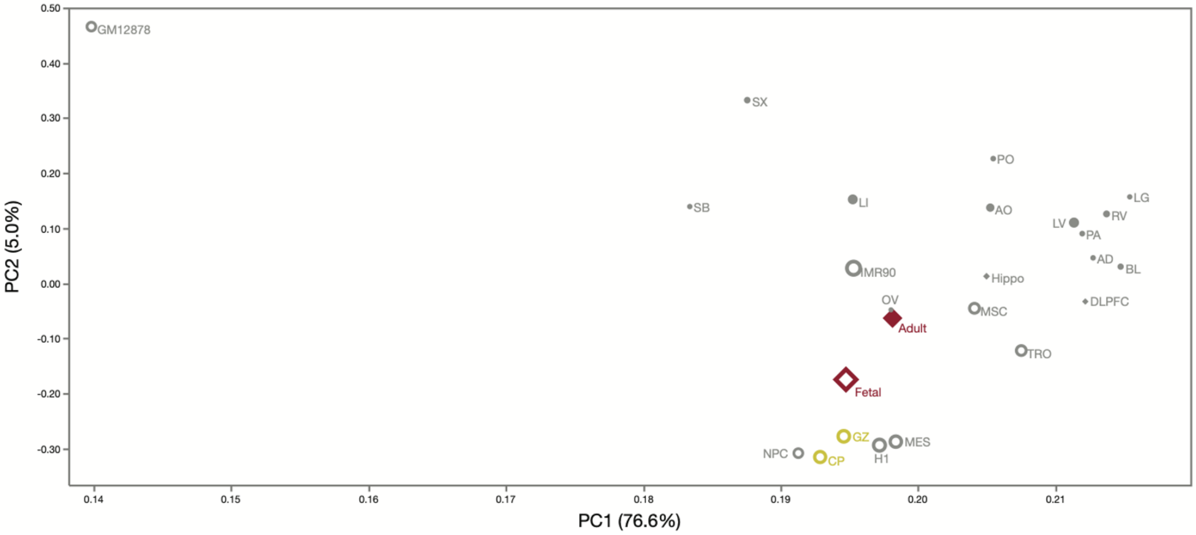


Figure 3.7: PCA Plot Comparing Insulation Scores with External Data at 40Kb

3.3.1.4 FIREs and Super-FIREs

FIREs at 40Kb resolution and associated super-FIREs were calculated. There were 3968 significant fetal FIREs and 3970 adult FIREs, and an overlap of 2365 FIREs. There were 223 significant fetal super-FIREs and 156 adult super-FIREs. Super-FIREs are not constricted at 40Kb, and represent the clustering of FIREs, thus average size of an adult super-FIRE was 352Kb and the average size of a fetal super-FIRE was 269Kb. Adult super-FIREs overlapped with 92 fetal super-FIREs (χ^2 test p-value < 0.0001), and fetal super-FIREs overlapped with 96 adult super-FIREs(χ^2 test p-value < 0.0001).

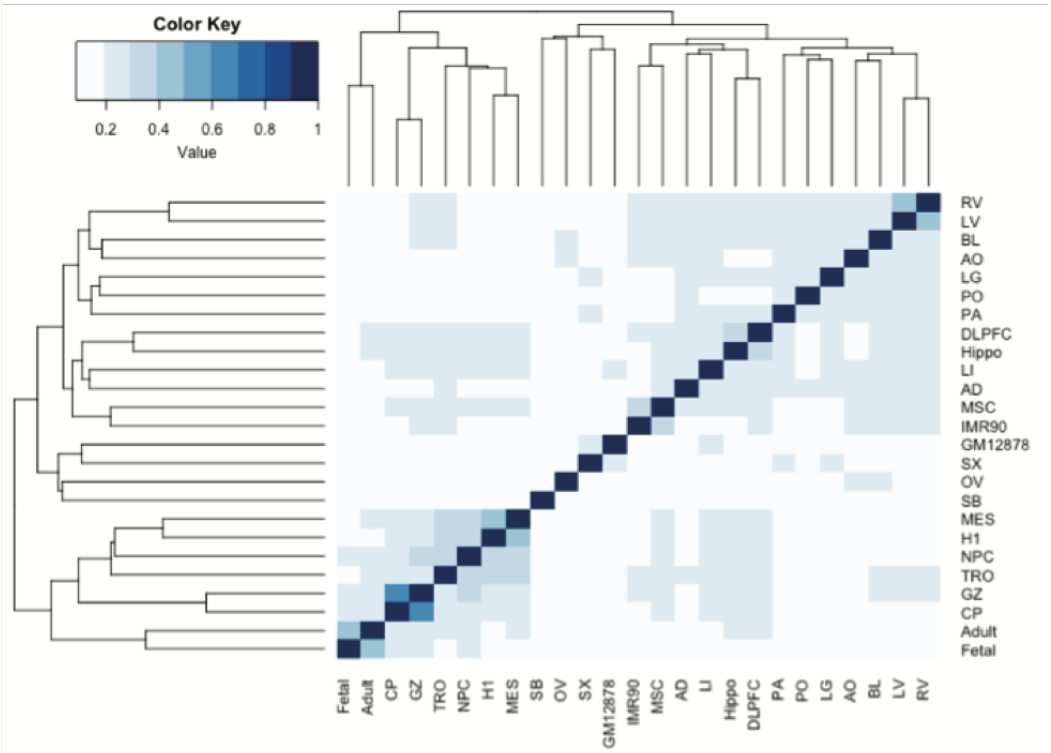


Figure 3.8: Jaccard Similarity Plot Comparing TADs with External Data at 40Kb

PCA of the FIRE scores at 40Kb resolution was conducted to evaluate the comparison with external data (Figure 3.9). PC1 showed 69% variance with (x-axis) and PC2 (y-axis) captured 5.4% variance. The brain samples (CP, GZ, Fetal, Adult) were all clustered together on PC1, with some variance captured for adult by PC2.

3.3.1.5 High Confidence Regulatory Chromatin Interactions

Out of the 43,222,677 possible 10 Kb interactions, there were a total of 969,302 significant interactions identified in fetal cortex and 370,994 significant interactions identified in adult cortex with a Bonferroni correction of $P < 2.31 * 10^{-11}$.

A further analysis was conducted in collaboration to identify anchors that overlapped enhancers or promoters for each developmental period (fetal and adult) by combining eHi-C, RNA-seq, ATAC-seq and ChIP-seq. These enhancers were based on the intersection of the HindIII Hi-C fragment within a 10 Kb anchor, open chromatin in cortex, and either H3K27ac peak or

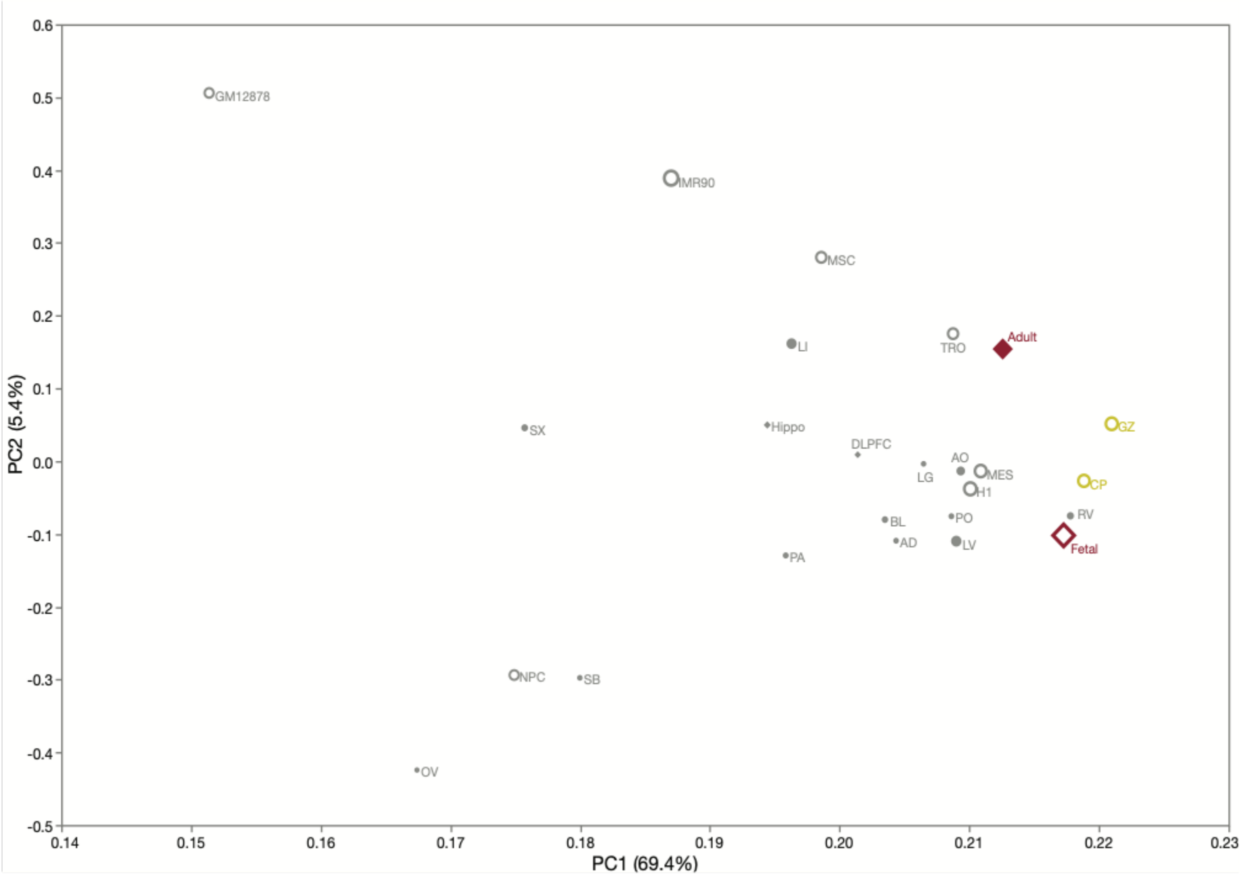


Figure 3.9: PCA of 40Kb FIRE scores with External Datasets

H3K4me3 peak along a brain-expressed TSS. The promoters were based on the overlap of a brain-expressed TSS and open chromatin in cortex. For adult cortex, there were 75,531 high confidence regulatory interactions (HCRCI), and for fetal cortex, there were 75,246 HCRCI.

3.3.2 Evaluation of Biological and Functional Interpretations

To evaluate whether the Hi-C readouts (chromatin interactions, TADs, FIREs, and super-FIREs) captured biologically relevant information, GREAT [59] analysis was conducted. Fetal chromatin interactions showed significant findings for transcription regulation, glial proliferation, oligodendrocyte differentiation, and growth cone, displaying enrichment for transcriptional regulation and core functions of the major cell types (glia and neurons). Adult chromatin interactions showed significant findings for postsynaptic density and excitatory synapse. TADs showed no

GREAT findings of interest, which was not unexpected due to being invariant across cell types. Fetal FIREs had significant findings for CNS axonogenesis, stem cell differentiation, and neural nucleus development, and adult FIREs had significant findings for the regulation of autophagy, phosphoprotein phosphatase, neural nucleus development, and detection of calcium ion. Fetal super-FIREs had significant findings for biological processes such as stem cell differentiation, neuron differentiation, CNS differentiation, neurogenesis, cortex radial cell migration. These results showed that adult and fetal super-FIREs and FIREs recapitulated key functions of the source tissues of differentiation and neurogenesis in the fetal sample and core neuronal functions in the adult sample. Concluding, FIREs, super-FIREs, and chromatin interactions can be connected to biologically relevant information.

We performed multivariate regression to evaluate how various functional genomic features (such as histone modifications, open chromatin, enhancers, insulators) and expression of the corresponding gene associated with chromatin spatial status, including FIREs, involving chromatin interactions, or residing in TAD boundaries. We found that fetal FIREs were enriched for fetal H3K27ac and CTCF marks, while depleted for H3K4me3 marks and TSS of protein-coding genes (model r^2=0.0356). Adult FIREs were enriched for typical enhancers, open chromatin, and H3K27ac in adult cortex, and depleted in H3K4me3 marks (model r^2=0.1317). Fetal chromatin interactions were enriched for H3K27ac and CTCF marks in fetal cortex, and depleted in H3K4me3 and gene expression in fetal cortex (model r^2= 0.0075). Adult chromatin interactions were enriched for open chromatin, typical enhancers, CTCF marks, H3K27ac in adult cortex, and TSS of protein-coding genes while depleted in H3K4me3 adult cortex (model r2=0.0517). Fetal and adult cortex TAD boundaries were enriched for CTCF marks and TSS of protein-coding genes (model r^2= 0.0304 and r^2= 0.0451, respectively). Our results suggest that these chromatin spatial organization features revealed from Hi-C readouts are informative for functional genomics.

3.3.3 Interpreting GWAS loci using 3D Architecture

In this section, we demonstrate the findings of our collaboration with Dr. Paola Giusti-Rodríguez and the Sullivan Lab of using Hi-C read outs as a way to ”connect” significant GWAS loci to specific genes [28]. Identifying a relationship between GWAS loci and specific genes is a crucial deliverable of GWAS for idiopathic disorders of the brain, as there is a strong assumption that GWAS findings connect to the nearest or intersecting gene. In this collaboration, the HCRCI displayed via UCSC browser tracks, Figure 3.10, demonstrates an example of HCRCI and an intergenic loci. This shows how the assumption that the closest gene to the loci may not be an accurate. The regulatory loops, with the exception of one, link the promoters/enhancers to the gene but not gene to the loci. In Figure 3.11, HCRCI are observed connecting the gene *DPYD* to a single locus chr1:98,220,320-98,562,260 (monogenic). In Figure 3.12, HCRCI are observed identify two statistically independent loci (chr14:30,000,405-30,208,630 and chr14:29,466,667-29,506,667). In Figure 3.13, it is observed that this complex, multigenic loci contains dense patterns of HCRCI (chr22:41,027,819-41,753,603 and chr22:39,840,130-40,091,818).

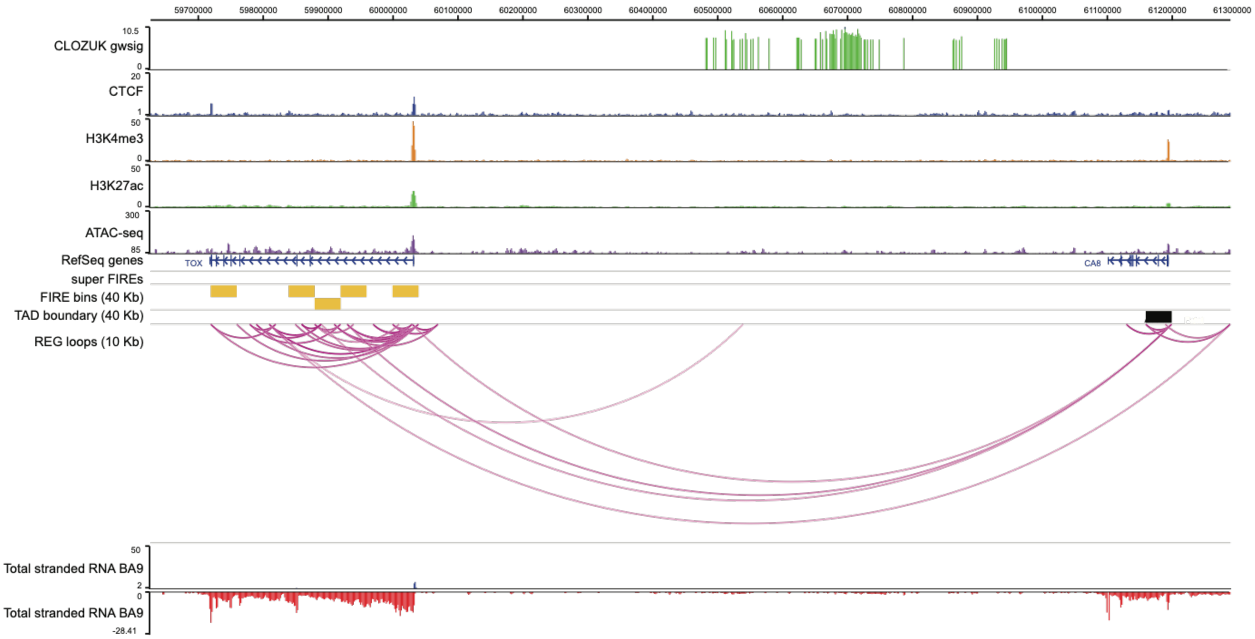


Figure 3.10: UCSC Browser Track: HCRCI Connecting Intergenic Loci

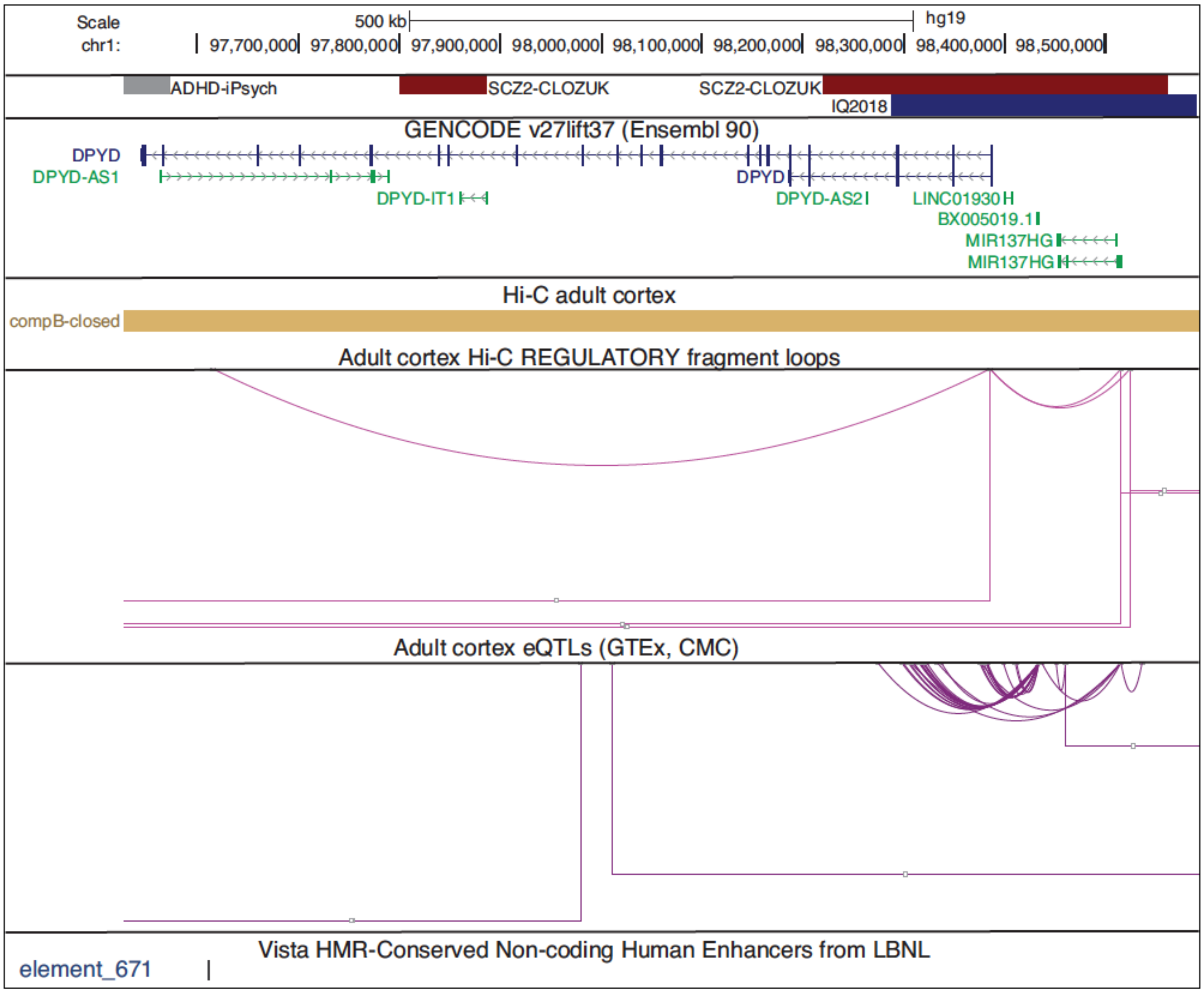


Figure 3.11: UCSC Browser Track: HCRCI Connecting Monogenic Loci

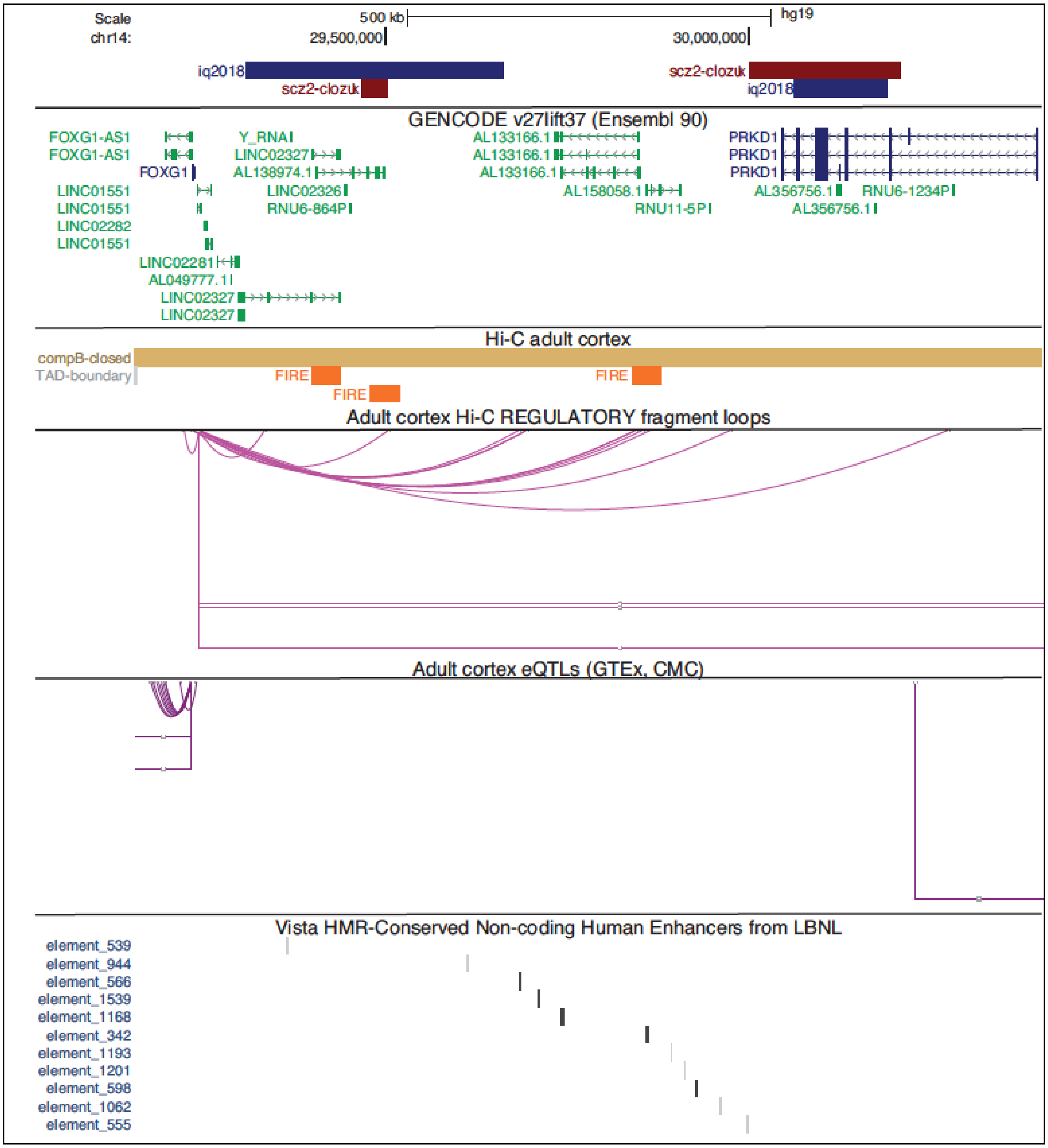


Figure 3.12: UCSC Browser Track: HCRCI Connecting Two Independent Loci

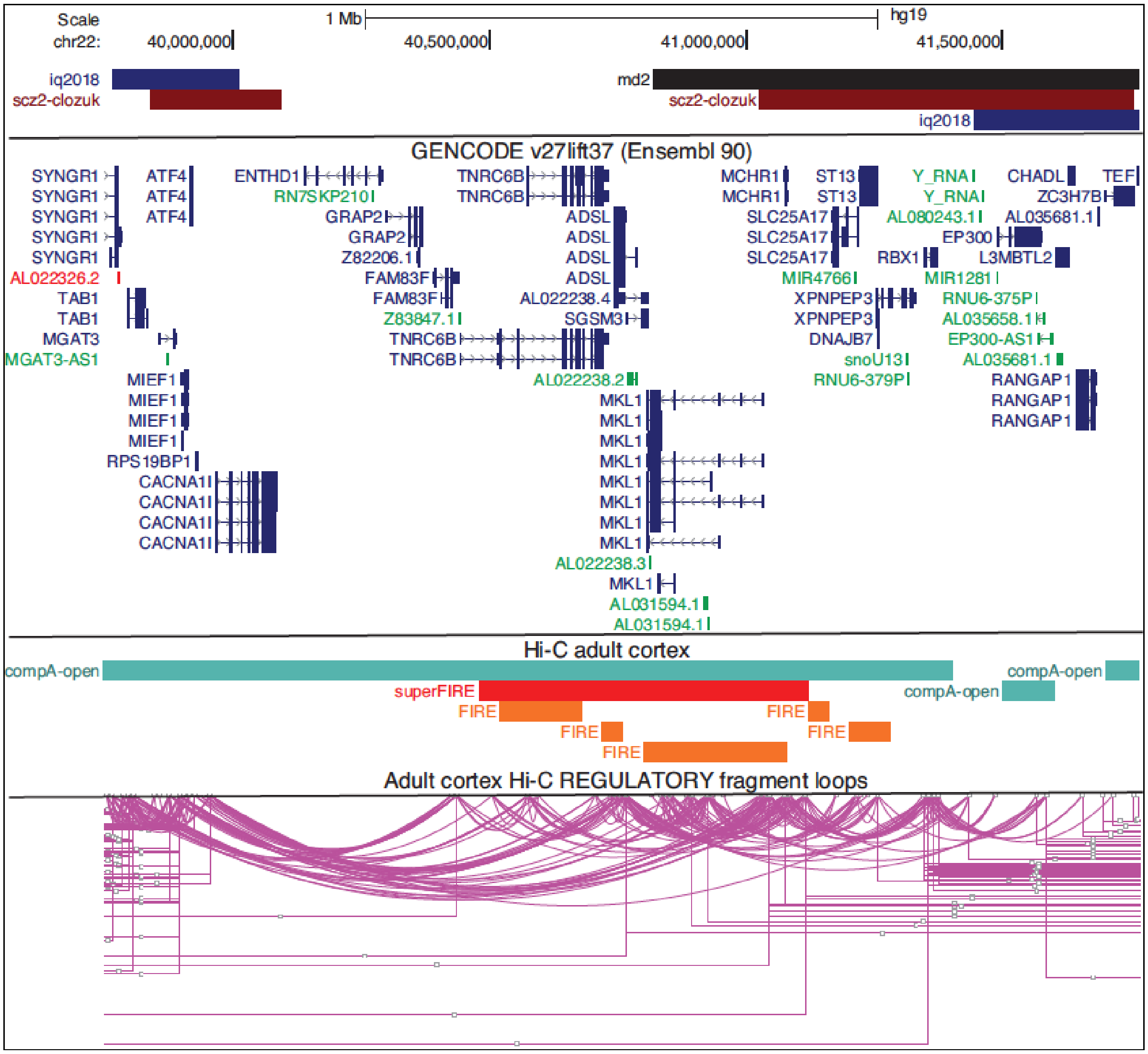


Figure 3.13: UCSC Browser Track: HCRCI Connecting Multigenic Loci

3.4 Conclusion

Our goal was to identify how 3D chromatin interactions can clarify GWAS results. For idiopathic disorders of the brain, this is a crucible deliverable as otherwise, these connections are difficult to obtain. In this paper, we used deep eHi-C datasets from fetal and adult cortex samples. We identified Hi-C readouts and confirmed that they were comparable with other external datasets. We evaluated the biological connection using GREAT, and the connection between functional genomic features such as histone modifications, open chromatin, enhancers, insulators, and expression of these Hi-C readouts. Thus, concluding that Hi-C readouts are informative in functional genomics. As a final result, we investigated the role of HCRCI and their ability to connect significant GWAS loci to specific genes.

CHAPTER 4: GENOMIC FINE MAPPING FOR HEMATOLOGIC TRAITS

4.1 Introduction

Blood cell variants have been associated with a variety of diseases as they make substantial contributions to many important biological processes. For example, abnormalities of function and the formation of blood cells have been associated with cancer, anemia, thrombotic disorders, and various immunodeficiencies [12]. With extensive sequencing and GWAS efforts, there is a need to convert massively-generated variant mapping into human-interpretable knowledge.

Hematological phenotypes (red blood cell, white blood cell, and platelet related traits) are critical physiological intermediaries in oxygen transport, immunity, infection, thrombosis, and hemostasis, as they are associated with many diseases including autoimmunity, asthma, viral infections, and cardiovascular disease. Hematological traits are highly heritable [25] with thousands of loci cumulatively identified from genetic association studies [2, 32, 39, 89]. The distributions of these traits and related complications such as stroke, venous thromboembolism (VTE), and kidney disease differ considerably across ethnicities [68, 11, 5, 63, 30].

We have recently performed a trans-ethnic meta-analysis for a battery of hematological traits in 746,667 participants, including 184,535 non-European ancestry individuals (15,171 African ancestry, 9,368 Hispanic/Latino, 151,807 East-Asian, 8,189 South-Asian) [9]. The trans-ethnic meta-analysis identified 5,552 trait-variant associations at a significance threshold of p $< 5 * 10^{-9}$. For 3,552 loci in which conditional analysis identified a single genome-wide significant variant in a European ancestry-specific analysis, fine-mapping results were generated for each trans-ethnic and ancestry-specific dataset using an approximate Bayesian approach to create 95$\%$ credible sets. In an effort to identify and prioritize variants for functional follow up experiments, we annotated the variants in these 95$\%$ credible sets identified as being significantly associated with seven red blood cell indices (HCT-Hematocrit; HGB-Hemoglobin Concentration; MCH-Mean Corpuscular Hemoglobin; MCHC- Mean Corpuscular Hemoglobin Concentration; MCV-Mean Corpuscular Volume; RBC-Red Blood Cell Count; RDW-Red Blood Cell Distribution Width), six white blood cell indices (BASO- Basophil Count; EOS- Eosinophil Count; LYM- Lymphocyte Count; MONO- Monocyte Count; NEU- Neutrophil Count; WBC-White Blood Cell Count), and two platelet related indices (MPV- Mean Platelet Volume; PLT- Platelet Count) from the above study. More specifically, we leveraged annotations encompassing 1D epigenomic signatures such as open chromatin status, histone modifications and transcription factor binding for non-coding variants and consequences on protein product for coding variants (”1D”); the impact of gene expression using eQTL (expression quantitative trait loci), sQTL (splicing quantitative trait loci), and pQTL (protein quantitative trait loci) information (”2D”); chromatin conformation information from Hi-C and alike technologies all from various blood, blood cell lineages and relevant tissue and cell lines (”3D”), and association information with many other phenotypes from Phenome-wide Association Studies (pheWAS).

In an effort to display and provide these results for further analysis, we created ABCx: Annotator for Blood Cell Traits. ABCx, an R Shiny app, contains four main components and functionalities: (1) an overall summary of the annotation sources incorporated; (2) an overview of the annotation results; (3) a variant level query; and (4) a gene level query. By relaying the annotation results in this format, ABCx provides comprehensive annotation information to prioritize loci for functional follow up experiments, including massively parallel reporter assay (MPRA) [56], genomic and epigenomic editing via CRISPR [86, 27] or CRISPRi [24, 41] experiments.

4.2 Overview of Annotations

4.2.1 Variant Level

Variants identified in the 95% credible sets from 3,552 loci in which conditional analysis identified a single genome-wide significant variant in European-specific ancestry were included. For each related phenotype index (RBC, RDW, HCT, HGB, MCH, MCHC, MCV, MPV, PLT, WBC, EOS, LYM, NEU, BASO, MONO) and ethnic analysis dataset (trans- Trans ethnic; EA- European ancestry; AA- African ancestry; HA-Hispanic American; EAS-East Asian; SA-South Asian) information on the posterior probability from the Bayesian approach, as well as the effect size estimate (β) and p-value from MR-MEGA [60], a meta regression tool for multi ethnic genetic association studies, are presented.

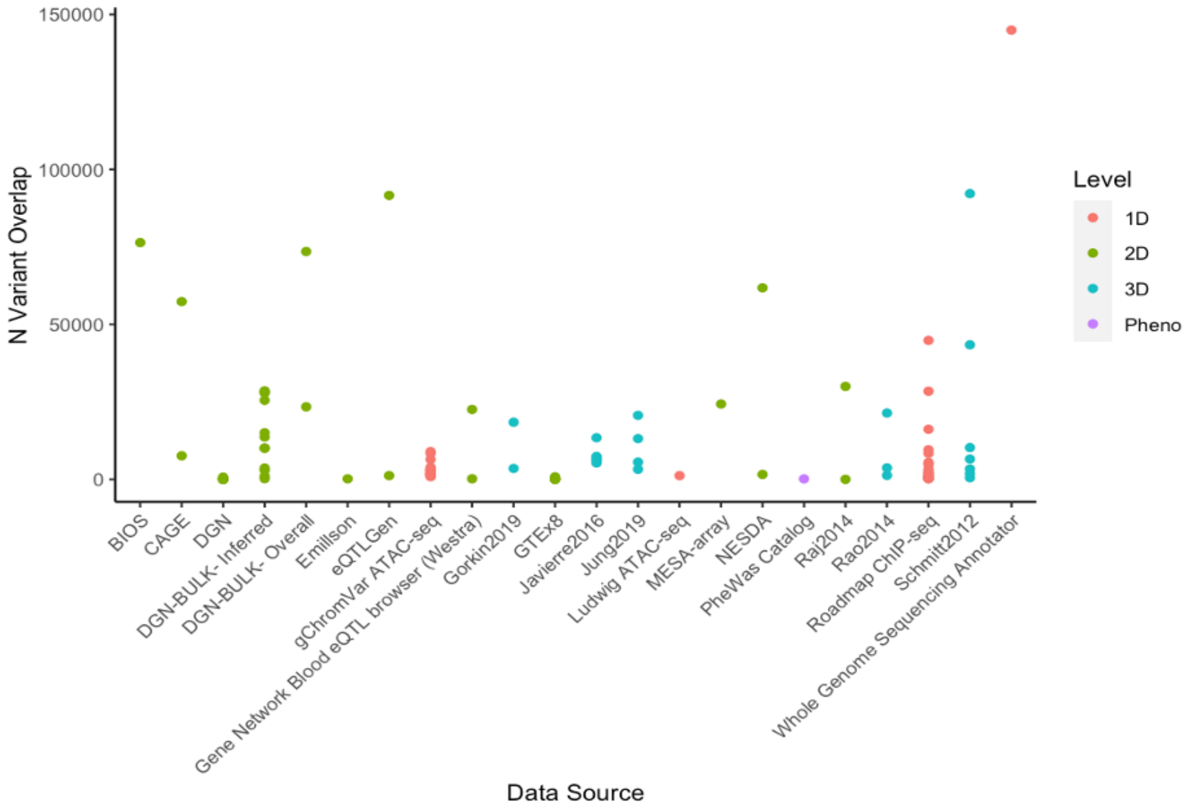


Figure 4.1: Scatter Plot of Annotation Sources and the Number of Individual Variants Present

4.2.2 1D Annotations

For ”1D” epigenomic features, we use output from WGSA, which was explicitly developed to annotate WGS data [49], ATAC-seq peaks, and histone ChIP-seq peaks [44]. WGSA returns many types of annotations including 12 sets of functional prediction scores (e.g., CADD, FATHMM-MKL), nine conservation scores (e.g., bStatistic, GERP++), information from four disease-related databases (ClinVar, COSMIC, GWAS catalog, GRASP2), and epigenomic information from many sources including GeneHancer, FANTOM5, Roadmap, ENCODE and BLUEPRINT. ATAC-seq peaks were gathered from recent studies for blood cell traits [53, 80] and key histone ChIP-seq peaks such as H3K9me3, H3K36me3, H3K4me1, H3K4me3, and H3K27Ac generated for Roadmap [44]. Information is displayed based on the relevance to the blood cell phenotypes. For instance, we included ATAC-seq peak information, GenoSkyline+ scores, key histone ChIP-seq peaks in spleen tissue, erythroid cells, and K562 cell lines for red blood cell related indices; fetal thymus tissue, GM12878 cells, macrophage cells, and T-cells and B-cells for white blood cell related indices; and megakaryocyte cells for platelet related indices. All sources for the 1D level are summarized in Table 4.1. and the number of variants with the related annotation is visualized in Figure 4.1.

Data Source	Type	Information Displayed
Roadmap	ChIP-seq Peaks	Peak Start,Peak End, N Peaks
Ludwig	ATAC-seq Peaks	Peak Start,Peak End, N Peaks
gChromVar	ATAC-seq Peaks	Peak Start,Peak End, N Peaks
WGSA	CADD	Raw, Phred
	fathmm XF	Prediction, Rank score, Raw Score, Non-coding Indicator
	fathmm MKL	Rank score, Raw Score
	Roadmap	GenoSkyline + Scores, CoreMarks, Imputed Mark, ChIP-seq Peaks, DNAse Hotspots
	GeneHancer	

Table 4.1: Summary of 1D Annotation

4.2.3 2D Annotations

For ”2D”, we included the variant impact on gene expression, impact on splicing ratios (eQTL and sQTL information), and impact on protein abundance (pQTL information). Sources of 2D annotations are summarized in Table 4.2 and Figure 4.1. These sources encompass both bulk and cell type specific sources, largely from the public domain (eQTLGen [84], CAGE [50], BIOS [99] for whole blood, and Raj et al for purified CD4+ T cells and monocytes [66]). Information available from these sources ranges, but primary features incorporated include the effect size estimate (β), p-value or FDR, the allele assessed, and the gene or protein involved. Variants across sources were matched based on chromosome, position, and alleles of each variant. Only significant results from the respective sources are relayed.

Data Source	Type	Information Displayed
BIOS	cis-eQTL	Gene Group, Gene Single,P-value, Allele Assessed
Westra	cis/trans-eQTL	Gene Group, Gene Single, FDR, Allele Assessed
NESDA	cis/trans-eQTL	Gene Single, Beta, FDR, Allele Assessed
DGN	cis/trans/distant-eQTL cis/trans/distant-sQTL	Gene Single,-log(p-value) Gene Single,-log(p-value)
Deconvoluted DGN Bulk	cis/trans-eQTL	Gene Single,Beta, FDR, Allele Assessed
GTEx8	cis/trans-eQTL sQTL	Gene Single, Conditional P-value Gene Single, FDR
Emillson	pQTL	Gene Single, Protein Name, Beta, P-value, Allele Assessed
Raj	cis/trans-eQTL	Gene Single, Beta, P-value
eQTLGen	cis/trans-eQTL	Gene Single, Beta, P-value
MESA-array	cis-eQTL	Gene Single, T-Statistic, FDR, Allele Assessed

Table 4.2: Summary of 2D Annotation

4.2.4 3D Annotations

For ”3D”, we include information on the 3D genome conformation to display blood lineage-specific regulatory elements and target genes from various sources summarized in Table 4.3. More specifically, using Hi-C data we incorporated statistically significant long-range chromatin interactions (LRCI) [67, 29, 71] calculated from Fit-Hi-C [3], loops calculated using the HiC-CUPs methodology [67], and superFIREs for related tissues [71]. Two Promoter-Capture Hi-C (PCHi-C) data sources [36, 38] were also incorporated and matched with the 2D results to highlight consistent evidences regarding the effected gene(s) across ”2D” and ”3D” annotations. ABCx displays information on the number of loops, LRCI, PCHi-C ”baits” and ”other ends”, or super-FIREs, as well as significance measures such as p-values, FDR, or CHICAGO scores where applicable. A summary of these sources is displayed in Table 4.3, and the number of variants overlapping with each of these source categories is visualized in Figure 4.1.

Data Source	Type	Information Displayed
Rao	HiCCUPs	Fragment 1, Fragment 2, P-Value, N loops
	Fit-Hi-C	Fragment 1, Fragment 2, P-Value, N LRCI
Gorkin	HiCCUPs	Fragment 1, Fragment 2, P-Value, N loops
	Fit-Hi-C	Fragment 1, Fragment 2, P-Value, N LRCI
Schmitt	Fit-Hi-C	Fragment 1, Fragment 2, P-Value, N LRCI
	Super-FIRE	Super-FIRE region, Cumulative FIREscore
Javierre	PCHi-C	Chicago Score, Gene Name, Bait Position, Other End Position
Jung	PCHi-C	-log(p-value), Gene Name, Bait Position, Other End Position

Table 4.3: Summary of 3D Annotation

4.2.5 PheWAS

In addition, PheWAS related information was incorporated to identify variants that were previously tested across a large number of phenotypes by leveraging EMR and matched genotype data [19]. The number of cases, p-value, odds ratio, Gene Name, PheWAS Code, and an indicator of whether the gene also overlapped with 2D and 3D evidences is given.

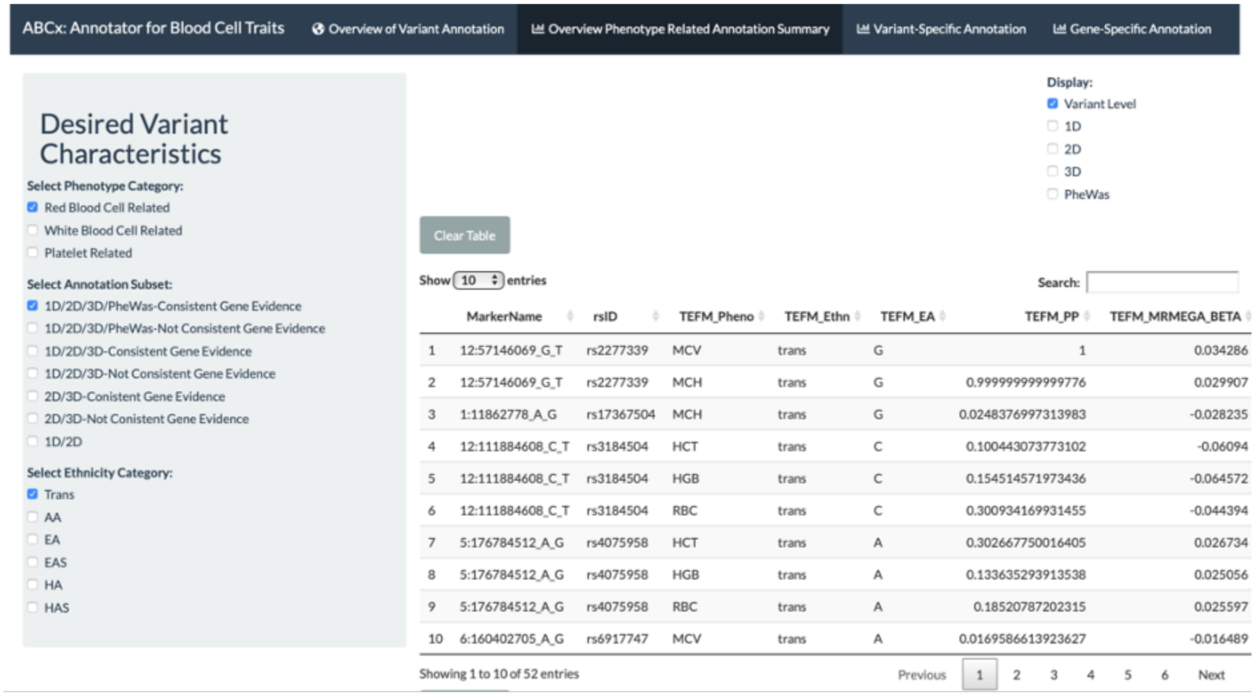
4.3 Overview of ABCx: Annotator for Blood Cell Traits

To visualize and leverage the annotations for further analysis or prioritization of experimental validations, we present ABCx: Annotator for Blood Cell Traits, an R Shiny app. ABCx has the following four main components to effectively and efficiently display and integrate relevant variant information.

First, there is a broad overview of the type of annotation sources incorporated as well as clear definitions of each annotation feature provided and how variant matching was conducted. This information is conveyed via an interactive plot similar to Figure 4.1, with a table that populates specific details.

Next, an overview of variant annotation is presented (Figure 4.2). On the ”Phenotype Annotation Summary” tab, the user can select the phenotype category (Red Blood Cell Related, White Blood Cell Related, and Platelet Related), the ethnicity subset (trans, EA, AA, EAS, HA, SA), the desired annotation features, and display format. Seven annotation categories are available, allowing users to focus on the most likely functional variants for further follow up. Specifically, the seven categories are (1) the most restrictive category, containing variants that have 1D, 2D, 3D and PheWAS evidence where the genes implied by 2D, 3D, and PheWAS are all consistent; (2) containing variants with 1D, 2D, 3D, and PheWAS evidence, but the genes implicated from different resources are not consistent; (3) variants with only 1D, 2D, and 3D but not any pheWAS evidences, and with consistent gene evidence between the 2D and 3D annotations; (4) variants with 1D, 2D, and 3D information (again, no pheWAS evidence) and no consistent gene implied; (5) variants with 2D and 3D evidences and consistent gene evidence from 2D and 3D annotations; (6) variants with 2D and 3D evidences, but no consistent gene evidence; and finally (7) variants with 1D and 2D evidence. As these categories are mutually exclusive, multiple selections are allowed so that users can investigate more than one category of variants. In addition, the user can restrict the amount of information presented by selecting which tables to be displayed. All tables can be exported in either a csv, or tab delimited format.

Next, there is a variant level annotation tab, where the user can input an rsID or chromosome and position of a variant (genome build gb38 and hg19 are both accommodated), and all selected annotation information related to that variant will populate (Figure 4.3). Third, there is a gene level annotation tab, where similar to the variant-specific tab, the user can input a HUGO Gene Nomenclature (HGNC) symbol, and related annotation information will populate.


ABCx: Annotator for Blood Cell Traits
Overview of Variant Annotation
Overview Phenotype Related Annotation Summary
Variant-Specific Annotation
Gene-Specific Annotation
Clear Table
Previous
1
2
3
4
5
6
Next

Desired Variant Characteristics

Select Phenotype Category:
☑ Red Blood Cell Related
□ White Blood Cell Related
□ Platelet Related

Select Annotation Subset:
☑ 1D/2D/3D/PheWas-Consistent Gene Evidence
□ 1D/2D/3D/PheWas-Not Consistent Gene Evidence
□ 1D/2D/3D-Consistent Gene Evidence
□ 1D/2D/3D-Not Consistent Gene Evidence
□ 2D/3D-Conistent Gene Evidence
□ 2D/3D-Not Conistent Gene Evidence
□ 1D/2D

Select Ethnicity Category:
☑ Trans
□ AA
□ EA
□ EAS
□ HA
□ HAS

Display:
☑ Variant Level
□ 1D
□ 2D
□ 3D
□ PheWas

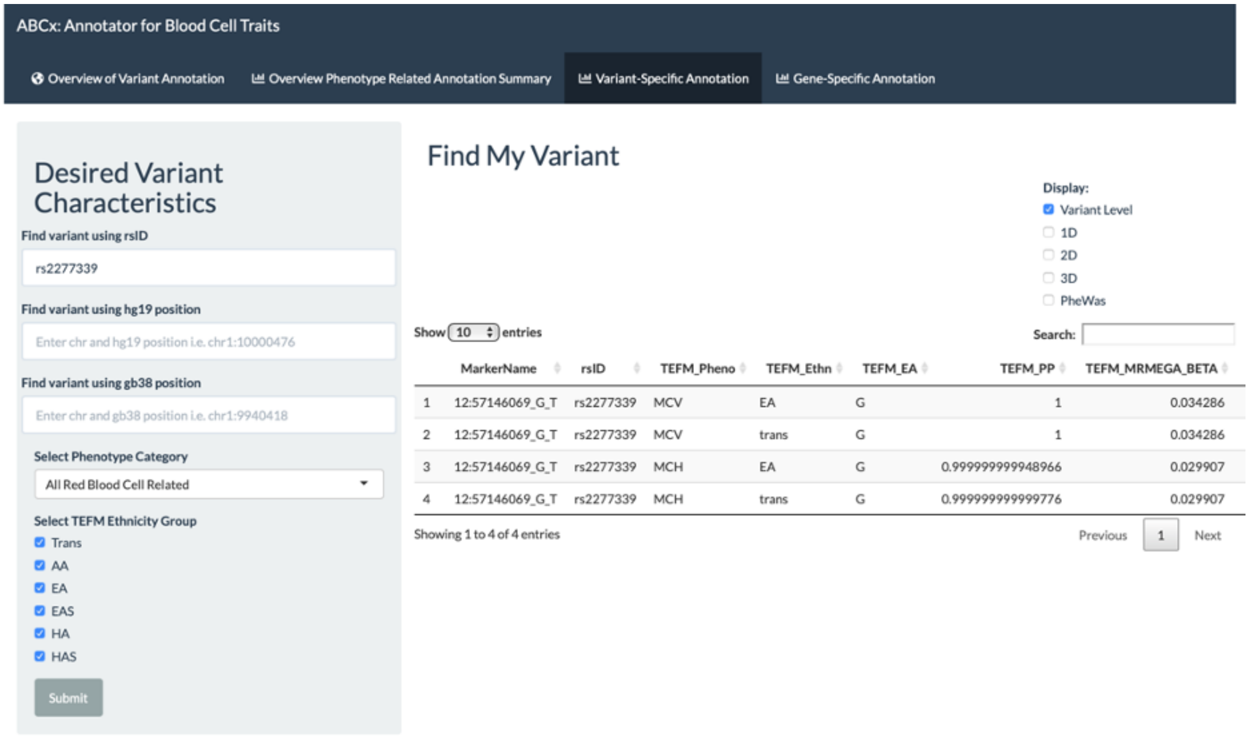
Show 10 entries

Search:

	MarkerName	rsID	TEFM_Pheno	TEFM_Ethn	TEFM_EA	TEFM_PP	TEFM_MRMEGA_BETA
1	12:57146069_G_T	rs2277339	MCV	trans	G	1	0.034286
2	12:57146069_G_T	rs2277339	MCH	trans	G	0.999999999999776	0.029907
3	1:11862778_A_G	rs17367504	MCH	trans	G	0.0248376997313983	-0.028235
4	12:111884608_C_T	rs3184504	HCT	trans	C	0.100443073773102	-0.06094
5	12:111884608_C_T	rs3184504	HGB	trans	C	0.154514571973436	-0.064572
6	12:111884608_C_T	rs3184504	RBC	trans	C	0.300934169931455	-0.044394
7	5:176784512_A_G	rs4075958	HCT	trans	A	0.302667750016405	0.026734
8	5:176784512_A_G	rs4075958	HGB	trans	A	0.133635293913538	0.025056
9	5:176784512_A_G	rs4075958	RBC	trans	A	0.18520787202315	0.025597
10	6:160402705_A_G	rs6917747	MCV	trans	A	0.0169586613923627	-0.016489

Showing 1 to 10 of 52 entries

Figure 4.2: Screenshot of ABCx Interface of Annotation Summary Tab


ABCx: Annotator for Blood Cell Traits
Overview of Variant Annotation
Overview Phenotype Related Annotation Summary
Variant-Specific Annotation
Gene-Specific Annotation
Submit
Previous
1
Next

Desired Variant Characteristics

Find variant using rsID
rs2277339

Find variant using hg19 position
Enter chr and hg19 position i.e. chr1:10000476

Find variant using gb38 position
Enter chr and gb38 position i.e. chr1:9940418

Select Phenotype Category
All Red Blood Cell Related

Select TEFM Ethnicity Group
☑ Trans
☑ AA
☑ EA
☑ EAS
☑ HA
☑ HAS

Find My Variant

Display:
☑ Variant Level
□ 1D
□ 2D
□ 3D
□ PheWas

Show 10 entries

Search:

	MarkerName	rsID	TEFM_Pheno	TEFM_Ethn	TEFM_EA	TEFM_PP	TEFM_MRMEGA_BETA
1	12:57146069_G_T	rs2277339	MCV	EA	G	1	0.034286
2	12:57146069_G_T	rs2277339	MCV	trans	G	1	0.034286
3	12:57146069_G_T	rs2277339	MCH	EA	G	0.999999999948966	0.029907
4	12:57146069_G_T	rs2277339	MCH	trans	G	0.999999999999776	0.029907

Showing 1 to 4 of 4 entries

Figure 4.3: Screenshot of ABCx Interface of Variant-Specific Annotation Tab

4.4 Highlighted Results

ABCx includes a total of 148,019 variants identified in the 95 % credible sets from 3,552 loci in which conditional analysis identified a single genome-wide significant variant in European Ancestry. These 148,019 variants corresponds to various hematologic phenotype indices: red blood cell (N variants=62,613), white blood cell (N variants=61,408) and platelets (N variants=36,810) (Figure 4.4). These variants are also categorized based on their analysis groups, whether ethnic-specific (AA, EA, EAS, HA, SA) or trans-ethnic (trans). Most variants are derived from analysis in European ancestry (N variants=116,122) and from trans-ethnic analysis (N=66,271), with additional variants from other ancestry-specific groups (EAS=10,663; AA=5,790; HA=5,702; SA=3,947; Figure 4.5).

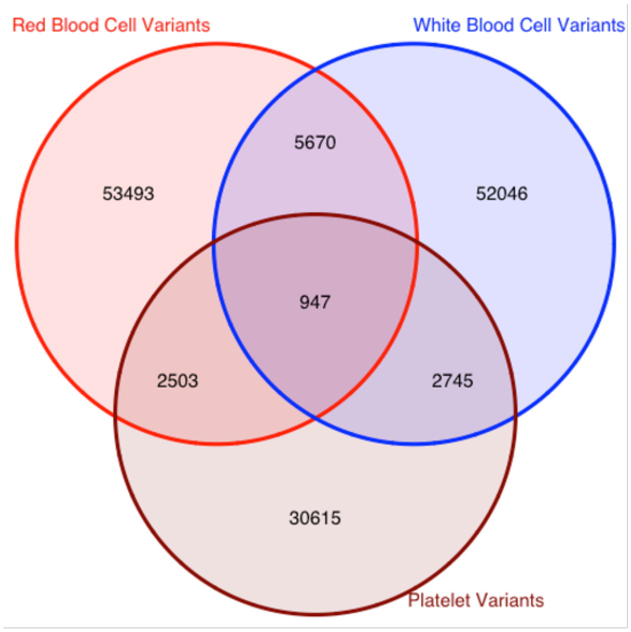


Figure 4.4: Distribution of Variants Across Phenotype Categories

The main characterization of these variants post-annotation manifests in a seven-category scale with the assumption that variants with the most consistent annotation available are the most promising for future functional follow-up (Figure 4.6). For example, if we focus on variants in categories with multiple sources of annotation evidences and consistent target genes implied from various sources, such criteria will result in 210 variants in the ”1D, 2D, 3D, PheWAS- Consis-

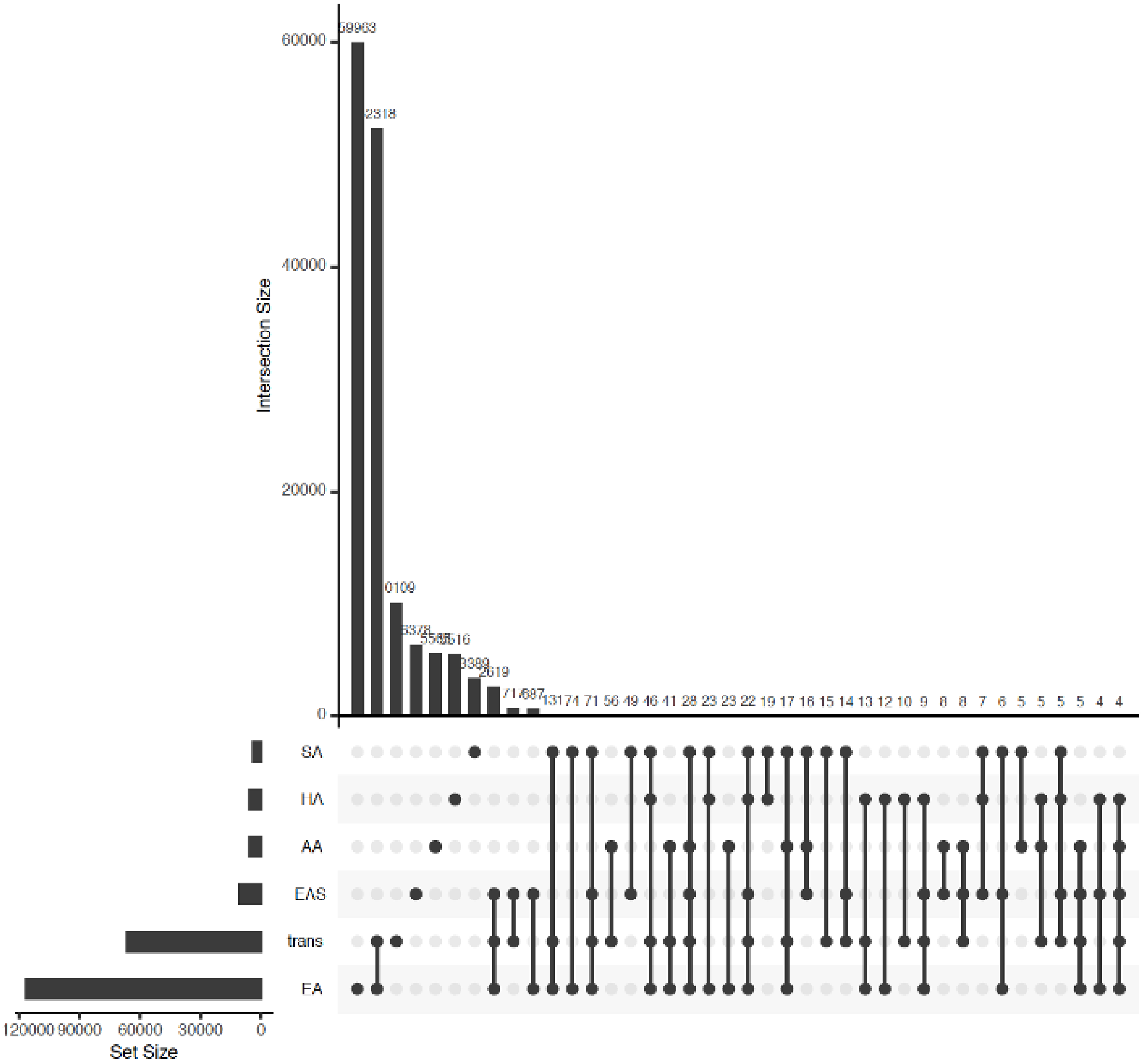


Figure 4.5: Upset Plot of Variants Across Trans-Ethnic Analysis Categories

tent Gene Evidence" category (0.07%), 44,341 in the "1D, 2D, 3D-Consistent Gene Evidence" category (15.1%), and 17,459 in the "2D, 3D-Consistent Gene Evidence" (6%), substantially reducing the set of most promising variants.

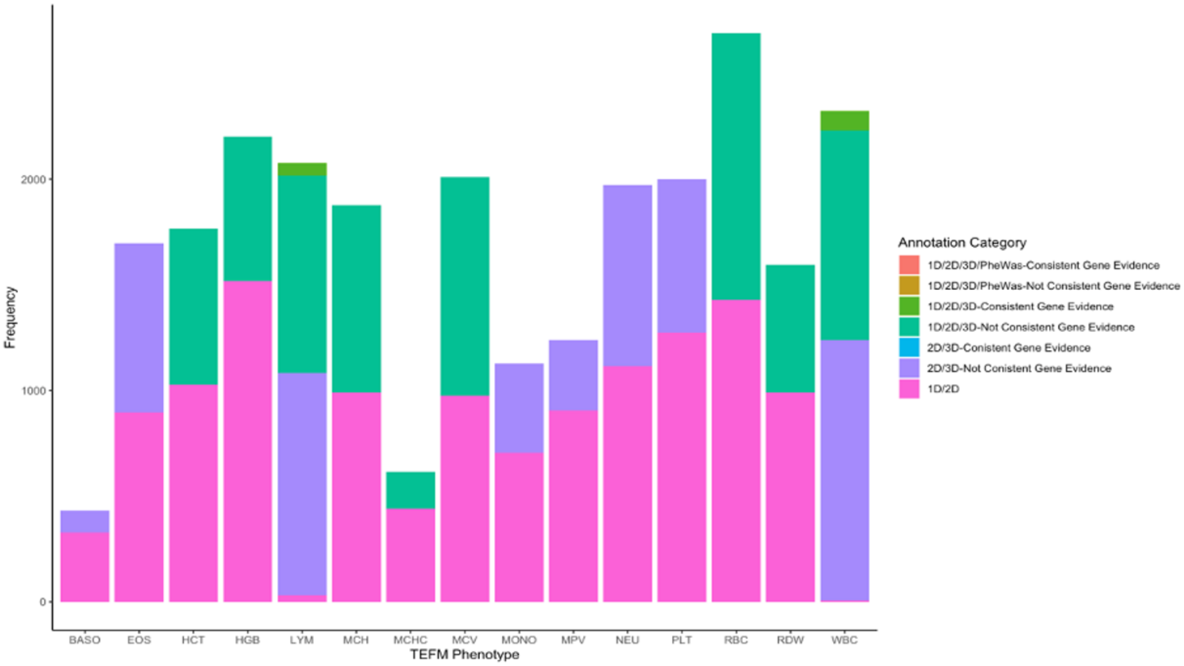


Figure 4.6: Histogram of Variants by Phenotype Index. Colors of the bar correspond to the annotation category.

4.5 Conclusion

Deciphering GWAS results into human interpretable knowledge is complex. By comprehensively annotating variants in the 95% credible sets identified in a recent trans-ethnic meta-analysis, we allow exploring and prioritizing variants with potentially causal effect. By incorporating various levels and sources of functional annotations, our ABCx R Shiny allows us to comb through these variants and their comprehensive sets of annotations to prioritize for further functional follow-up, creating a clearer picture of the functional characteristics of variants underlying hematologic traits.

CHAPTER 5: CONCLUSION

In this dissertation, we explore the importance of understanding the 3D genome and how incorporating multi-level genomic annotations can help clarify GWAS findings, and prioritize variants for further functional follow-up. In Chapter 2, we introduce FIREcaller, a consistent pipeline for the scientific community to discover FIREs and super-FIREs to better understand regulatory and functional processes. There has not been a published software to date, and FIREcaller fills this gap.

In Chapter 3, we further explore the 3D genome by identifying various Hi-C readouts, and comparing the results to other publicly available Hi-C datasets. We justify the use of these fetal and adult brain Hi-C datasets, demonstrate a biological connection of our results, and incorporate the results of high confidence regulatory chromatin interactions to provide insight into the relationship between GWAS loci and specific genes.

In Chapter 4, we provide functional annotation of 148,019 variants identified in 95% credible sets from 3,552 loci determined in a recent trans-ethnic meta-analysis for a battery of hematological traits in 746,667 participants, including 184,535 non-European ancestry individuals. We briefly describe our R Shiny app, ABCx: Annotator for Blood Cell Traits, which relays our annotation results to our collaborators and summarize our findings of incorporating multi-level annotation sources such as epigenomic signatures, information on gene expression, and 3D genome chromatin conformation information.

APPENDIX A: ADDITIONAL RESULTS FOR CHAPTER 2

This chapter contains technical details supplemental to the main text of Chapters 2.

A.1 Choice of Bin Resolutions

We tested different bin resolutions using GM12878 Hi-C data at different sequencing depths. We first created 10Kb, 20Kb and 40Kb Hi-C contact matrices using the full data with $\sim$4.4 billion reads. We then down-sampled the full data using a binomial distribution to create matrices reflecting a sample that would contain $\sim$2.2 billion reads. We repeated this process ten times to generate a total of 20 down-sampled matrices at each of the three resolutions (i.e., 10Kb, 20Kb and 40Kb).

We calculated Pearson correlation of binary FIRE classifications between pairs of replicates, at each resolution. We observed that with $\sim$ 2.2 billion reads per replicate, correlations are all reasonable (mean ρ= 0.980 at 40Kb, mean ρ= 0.973 at 20Kb, mean ρ=0.967 at 10Kb resolution; Figure A.1). We also note that as the resolution increases, the correlation decreases for the obvious reason that it is more challenging and entails more data to detect FIREs at higher resolution. In general, we recommend using 40Kb resolution for Hi-C data with $<$1 billion reads, 20Kb with 1-2 billion reads, and the highest 10Kb resolution only when the Hi-C data has $>$2 billion reads.

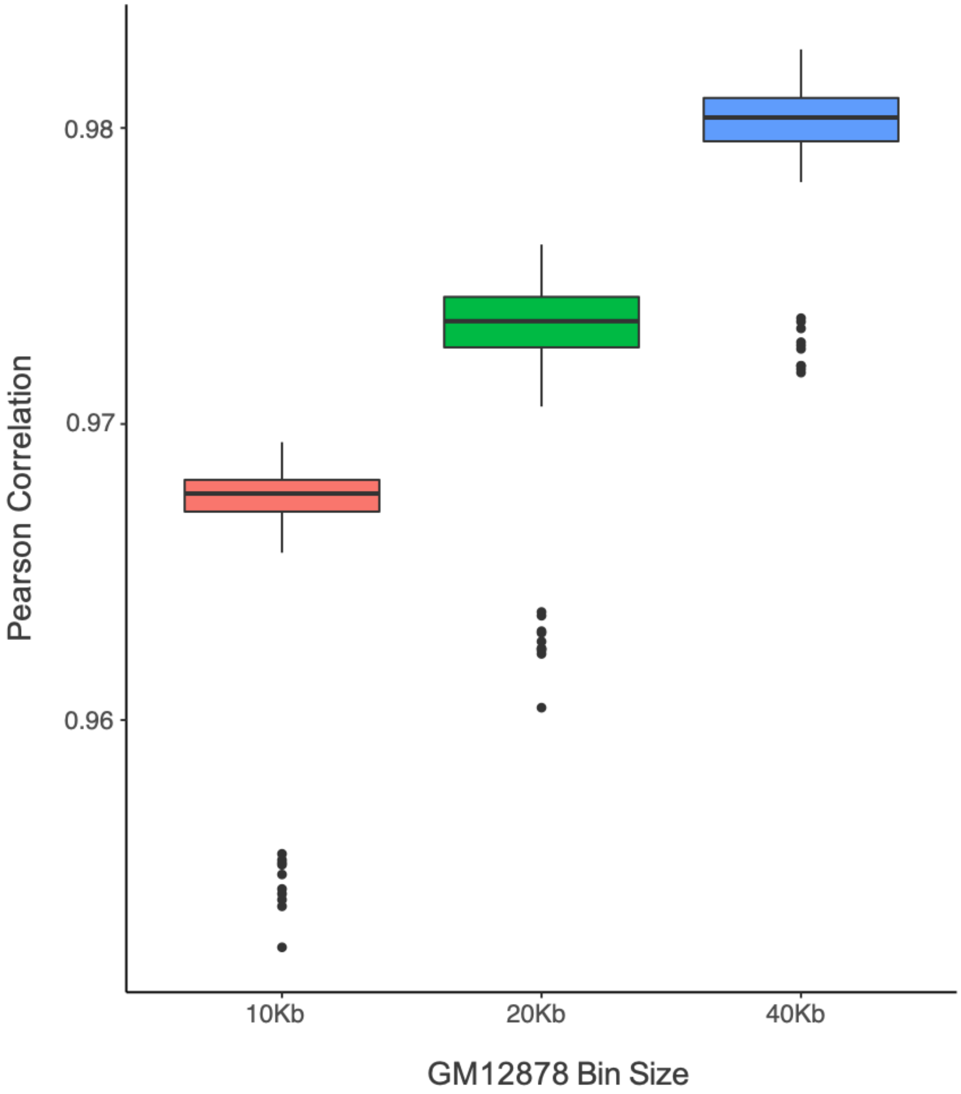


Figure A.1: Boxplots of the Pearson correlations of the dichotomized FIREs for downsampled GM12878 samples at 10Kb, 20Kb, and 40Kb. Each boxplot shows the Pearson correlation across all pairs out of the 20 replicates (i.e., 190 pairs).

A.2 Distance Distribution of Enhancer-Promoter (E-P) Interactions

Previous studies have shown that the majority of E-P interactions are within 200Kb [37, 97, 67, 73, 3]. Here we analyzed interactions reported as high confidence E-P interactions in two brain datasets [28, 46] (Figure A.2). We found that high confidence E-P interactions in both adult cortex [28] and in dorsolateral prefrontal cortex (DLPFC) from PsychENCODE [46] have a median distance less than 200Kb (160Kb and 190Kb, respectively; Figure A.2).

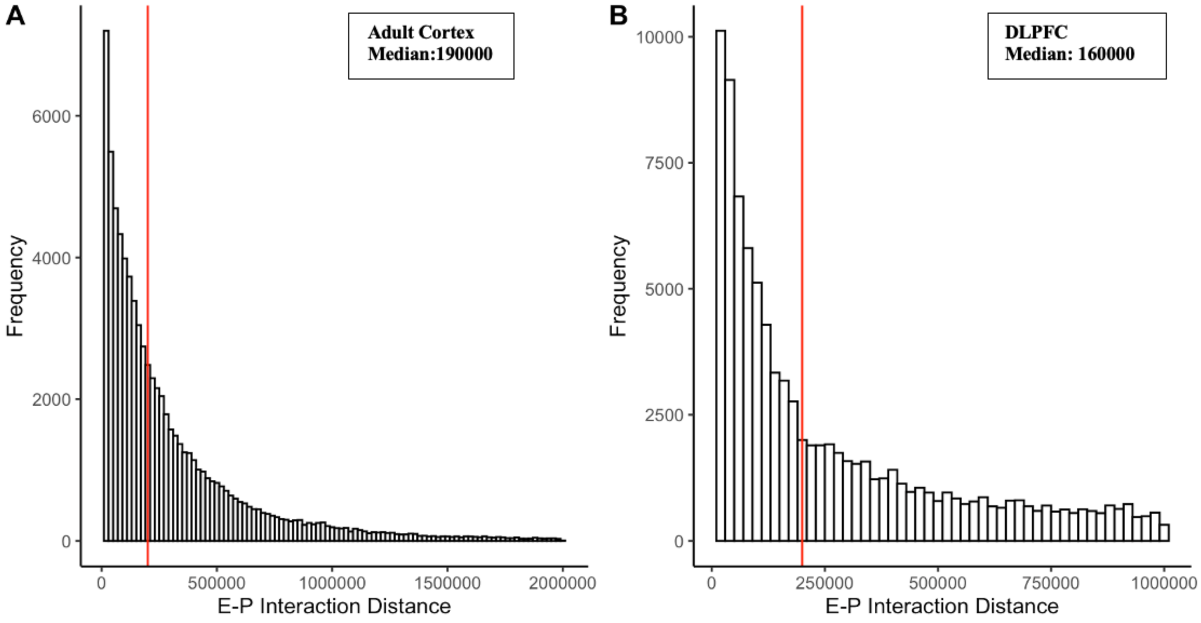


Figure A.2: Histogram of the 1D genomic distance for high confidence E-P interactions. A) Adult Cortex. B) DLPFC. The vertical red line marks the 200Kb distance.

A.3 Poisson versus Negative Binomial Models

To compare the two models: Poisson and negative binomial, we reported the goodness-of-fit statistics (Table A.1), estimation of regression coefficients for major contributors to systematic biases, computational time, and final FIREcaller results (Figure A.3). Specifically, we ran FIREcaller using Hi-C data from hippocampus tissue [71], fitting separately using a Poisson regression model and a negative binomial regression model, both with default values for other parameters.

	Poisson	Negative Binomial
Residual Deviance	536,869	65,740
AIC	944,769	608,301

Table A.1: Model Goodness-of-Fit Statistics for Poisson and Negative Binomial Regression Model. Smaller residual deviance or AIC indicates a better model fit.

In terms of model fit, the negative binomial distribution fits the Hi-C count data better (deviance= 65,740) than Poisson distribution (deviance= 536,869) (Table A.1). As an important intermediate step, FIREcaller aims to remove systematic biases and calculate normalized *cis*-contact frequency. We found that the two models resulted in similar effect size estimates for the three key factors (namely, effective fragment length, GC content and mappability) contributing to systematic biases. When comparing the final binary FIRE regions identified, we found the two models result in highly overlapped FIREs: 3,491 shared out of the 3,642 identified from the Poisson model, and out of the 3,750 from the negative binomial model (odds ratio= 396.5, Fisher's exact P-value $< 2.2 * 10^{-16}$) (Figure A.3).

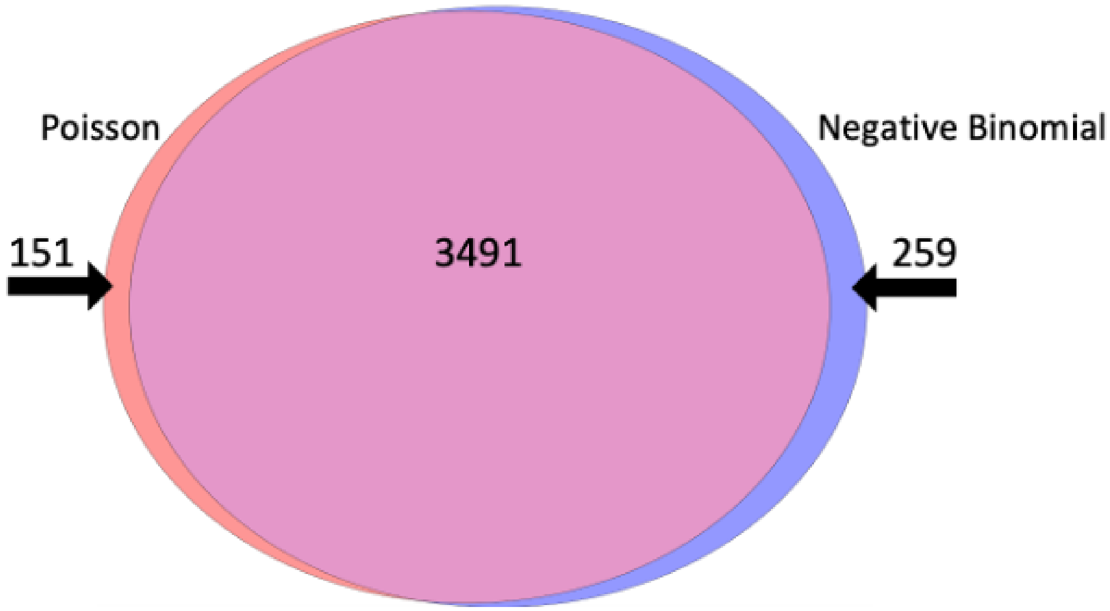
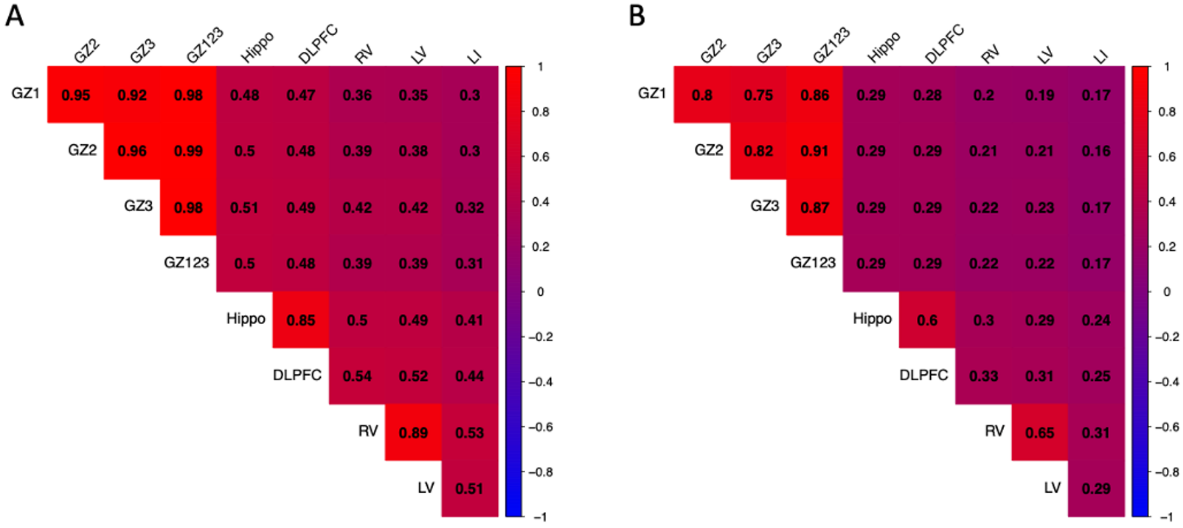


Figure A.3: Venn diagram for FIREs Detected from Negative Binomial or Poisson Model

To compare computational efficiency, we fit the Poisson and the negative binomial regression step of FIREcaller for the Hippocampus data at 40Kb 100 times for each distribution. The average running time was 0.44 seconds (0.028 SD) for Poisson regression and 4.28 seconds (0.17 SD) for negative binomial regression. We note that there is a possible issue of non-convergence when analyzing 10Kb and 20Kb using the negative binomial regression, which we have observed with some of our real data analysis, particularly when sequencing depth is low ($<$500 million raw reads per sample).

A.4 Evaluation of Tissue-Specificity

We applied our FIREcaller to Hi-C data from hippocampus, DLPFC, liver, left ventricle, and right ventricle [71], as well as brain tissues from the germinal zone (and its three replicates) [85]. We then calculated Pearson correlation between any pair of samples, based on either the FIRE continuous scores (Figure A.4) or the dichotomized FIRE calls (Figure A.4). As expected, left and right ventricle are highly correlated with each other (continuous FIRE scores $\rho = 0.89$; dichotomized FIREs $\rho = 0.65$), hippocampus and prefrontal cortex are also highly correlated (continuous FIRE scores $\rho = 0.85$; dichotomized FIREs $\rho = 0.6$) and germinal zone brain tissue combined sample (GZ123) is highly correlated with its replicates (continuous FIRE scores $\rho =$ 0.98, 0.99, 0.98; dichotomized FIREs $\rho = 0.86$, 0.91, 0.87 for GZ1, GZ2, and GZ3 respectively).



A

	GZ2	GZ3	GZ123	Hippo	DLPFC	RV	LV	LI
GZ1	0.95	0.92	0.98	0.48	0.47	0.36	0.35	0.3
GZ2		0.96	0.99	0.5	0.48	0.39	0.38	0.3
GZ3			0.98	0.51	0.49	0.42	0.42	0.32
GZ123				0.5	0.48	0.39	0.39	0.31
Hippo					0.85	0.5	0.49	0.41
DLPFC						0.54	0.52	0.44
RV							0.89	0.53
LV								0.51

B

	GZ2	GZ3	GZ123	Hippo	DLPFC	RV	LV	LI
GZ1	0.8	0.75	0.86	0.29	0.28	0.2	0.19	0.17
GZ2		0.82	0.91	0.29	0.29	0.21	0.21	0.16
GZ3			0.87	0.29	0.29	0.22	0.23	0.17
GZ123				0.29	0.29	0.22	0.22	0.17
Hippo					0.6	0.3	0.29	0.24
DLPFC						0.33	0.31	0.25
RV							0.65	0.31
LV								0.29

Figure A.4: Pearson correlation of continuous FIRE scores and binary FIREs. We calculate the Pearson correlations between each pair of samples (one sample each for hippocampus [Hippo], DLPFC, liver [LI], left ventricle [LV], right ventricle[RV], and four samples from germinal zone [GZ] brains: three replicates [GZ1, GZ2, GZ3] and the pooled sample GZ123), for continuous FIRE scores (panel A on the left), or binary FIREs (panel B on the right).

A.5 Visualization of Hippocampus Tissue Raw Contact Matrix

To further highlight the region of the 400Kb super-FIRE in human hippocampus tissue we visually inspected the raw contact matrix of both the entire chromosome 18 (Figure A.5) and the ∼1Mb region chr18: 52,640,000-53,680,000. As visualized, there is a higher level of *cis*-interactions (red cluster on heatmap) in the region identified as a super-FIRE (red bar).

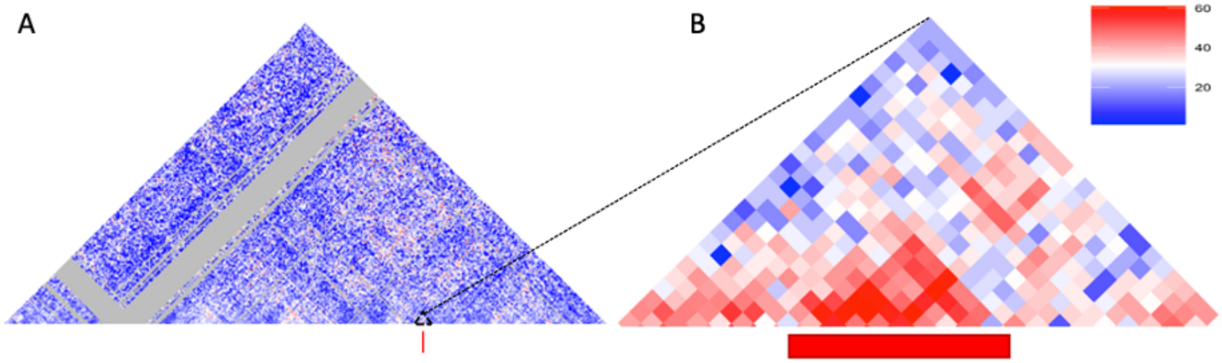


Figure A.5: Heatmap of Hippocampus Tissue Raw Contact Matrix. A) A heatmap of the chromosome 18 contact matrix and B) chromosome 18 contact matrix at region 52,640,000-53,680,000 with the red bar below showing the 400Kb super-FIRE region. Grey regions of the contact matrix indicate counts of 0.

A.6 Comparison of Enhancer and Promoters and FIREs in Liver and Left Ventricle

We used H3K27ac ChIP-seq peaks [44] in left ventricle and liver to define active enhancers, and used 500 bp upstream / downstream of transcription start site (TSS) to define promoters. We then investigated whether a FIRE tends to overlap more with an enhancer or a promoter. In both tissues, we have found that FIREs tend to overlap more with enhancers. Specifically, in liver odds ratio (OR) = 5.22 for enhancers (p-value $< 2.2 * 10^{16}$), OR = 1.98 for promoters (p-value $< 2.2 * 10^{16}$); in left ventricle, OR=4.48 for enhancers (p-value $< 2.2 * 10^{16}$); OR= 1.03 for promoters (with a not significant p-value=0.48) (Figure A.6).

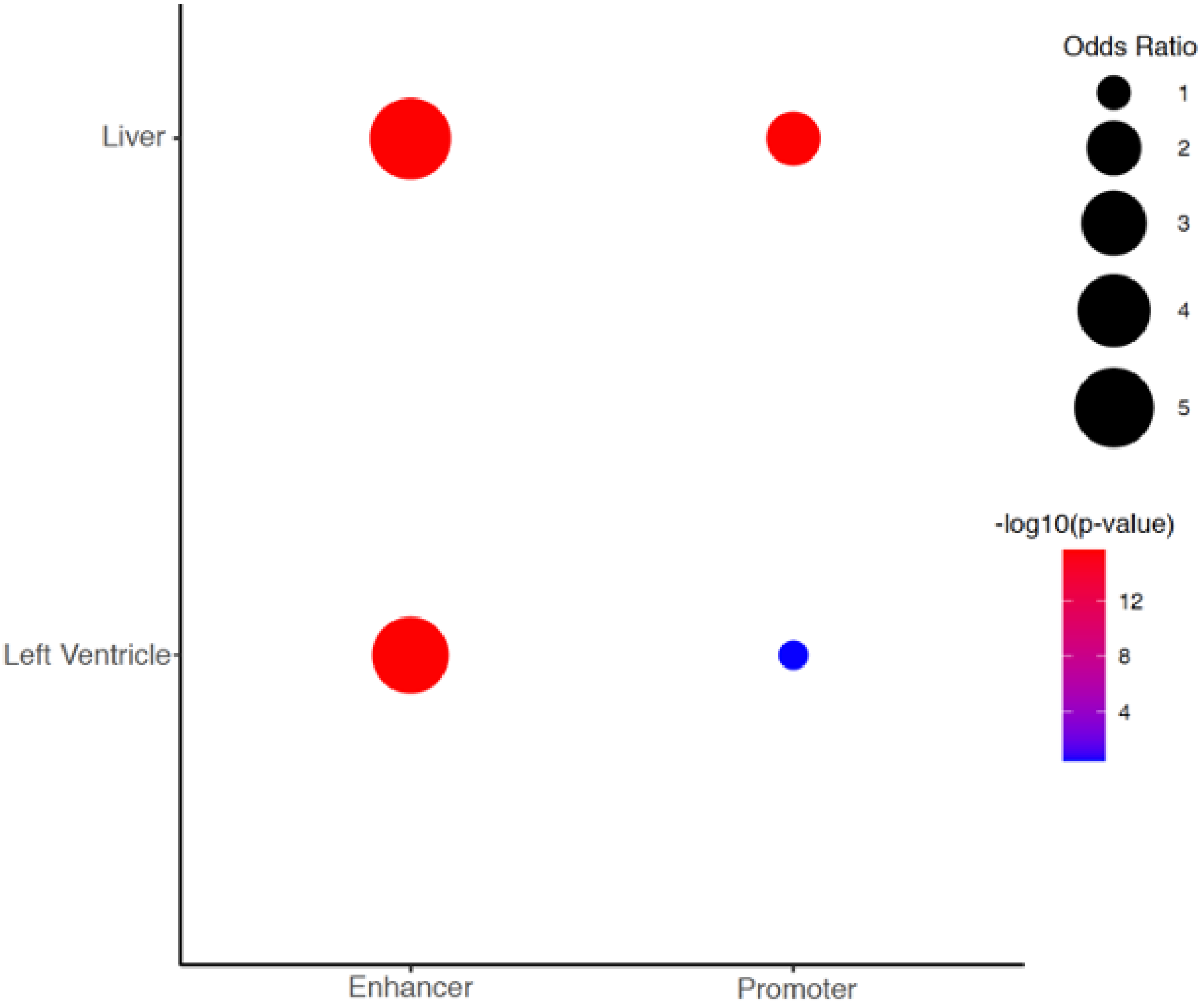


Figure A.6: Dot Plot Showing FIREs Overlapping More with Enhancers than Promoters, in Liver and Left Ventricle tissue. We use the color scale of blue to red to highlight significance (red being most significant and blue least significant), and the size of the bubble to display the odds ratios (larger dot indicating larger effect size).

REFERENCES

Amemiya, H. M., Kundaje, A., and Boyle, A. P. (2019). The encode blacklist: Identification of problematic regions of the genome. *Scientific Reports*, **9**(1), 9354.

Astle, W. J., Elding, H., Jiang, T., Allen, D., Ruklisa, D., Mann, A. L., Mead, D., Bouman, H., Riveros-Mckay, F., Kostadima, M. A., *et al.* (2016). The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell*, **167**(5), 1415–1429.

Ay, F., Bailey, T. L., and Noble, W. S. (2014). Statistical confidence estimation for hi-c data reveals regulatory chromatin contacts. *Genome Research*, **24**(6), 999–1011.

Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M. A., Beaudet, A. L., Ecker, J. R., *et al.* (2010). The nih roadmap epigenomics mapping consortium. *Nature biotechnology*, **28**(10), 1045–1048.

Beutler, E. and Duparc, S. (2007). Glucose-6-phosphate dehydrogenase deficiency and antimalarial drug development. *The American journal of tropical medicine and hygiene*, **77**(4), 779–789.

Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**(2), 185–93.

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., *et al.* (2019). The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research*, **47**(D1), D1005–D1012.

Burgess, D. J. (2017). Deciphering non-coding variation with 3d epigenomics. *Nature Reviews Genetics*, **18**(1), 4–4.

Chen, M. H., Raffield, L. M., Mousas, A., Sakaue, S., Huffman, J. E., Moscati, A., Trivedi, B., Jiang, T., Akbari, P., Vuckovic, D., Bao, E. L., Zhong, X., Manansala, R., Laplante, V., Chen, M., Lo, K. S., Qian, H., Lareau, C. A., Beaudoin, M., Hunt, K. A., Akiyama, M., Bartz, T. M., Ben-Shlomo, Y., Beswick, A., Bork-Jensen, J., Bottinger, E. P., Brody, J. A., van Rooij, F. J. A., Chitrala, K., Cho, K., Choquet, H., Correa, A., Danesh, J., Di Angelantonio, E., Dimou, N., Ding, J., Elliott, P., Esko, T., Evans, M. K., Floyd, J. S., Broer, L., Grarup, N., Guo, M. H., Greinacher, A., Haessler, J., Hansen, T., Howson, J. M. M., Huang, Q. Q., Huang, W., Jorgenson, E., Kacprowski, T., Kahonen, M., Kamatani, Y., Kanai, M., Karthikeyan, S., Koskeridis, F., Lange, L. A., Lehtimaki, T., Lerch, M. M., Linneberg, A., Liu, Y., Lyytikainen, L. P., Manichaikul, A., Martin, H. C., Matsuda, K., Mohlke, K. L., Mononen, N., Murakami, Y., Nadkarni, G. N., Nauck, M., Nikus, K., Ouwehand, W. H., Pankratz, N., Pedersen, O., Preuss, M., Psaty, B. M., Raitakari, O. T., Roberts, D. J., Rich, S. S., Rodriguez, B. A. T., Rosen, J. D., Rotter, J. I., Schubert, P., Spracklen, C. N., Surendran, P., Tang, H., Tardif, J. C., Trembath, R. C., Ghanbari, M., Volker, U., Volzke, H., Watkins, N. A., Zonderman, A. B., Program, V. A. M. V., Wilson, P. W. F., Li, Y., Butterworth, A. S., Gauchat, J. F., Chiang, C. W. K., Li, B., *et al.* (2020). Trans-ethnic and ancestry-specific blood-cell genetics in 746,667 individuals from 5 global populations. *Cell*, **182**(5), 1198–1213 e14.

Chen, W., Larrabee, B. R., Ovsyannikova, I. G., Kennedy, R. B., Haralambieva, I. H., Poland, G. A., and Schaid, D. J. (2015). Fine mapping causal variants with an approximate bayesian method using marginal test statistics. *Genetics*, **200**(3), 719–736.

Chen, Z., Tang, H., Qayyum, R., Schick, U. M., Nalls, M. A., Handsaker, R., Li, J., Lu, Y., Yanek, L. R., Keating, B., *et al.* (2013). Genome-wide association analysis of red blood cell traits in african americans: the cogent network. *Human molecular genetics*, **22**(12), 2529–2538.

Colin, Y., Le Van Kim, C., and El Nemer, W. (2014). Red cell adhesion in human diseases. *Current opinion in hematology*, **21**(3), 186–192.

Consortium, E. P. *et al.* (2004). The encode (encyclopedia of dna elements) project. *Science*, **306**(5696), 636–640.

Consortium, W. T. C. C. *et al.* (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**(7145), 661.

Crane, E., Bian, Q., McCord, R. P., Lajoie, B. R., Wheeler, B. S., Ralston, E. J., Uzawa, S., Dekker, J., and Meyer, B. J. (2015). Condensin-driven remodeling of x-chromosome topology during dosage compensation. *Nature*, **523**(7559), 240–244.

Dekker, J., Marti-Renom, M. A., and Mirny, L. A. (2013). Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat Rev Genet*, **14**(6), 390–403.

Dekker, J., Belmont, A. S., Guttman, M., Leshyk, V. O., Lis, J. T., Lomvardas, S., Mirny, L. A., Oshea, C. C., Park, P. J., Ren, B., *et al.* (2017a). The 4d nucleome project. *Nature*, **549**(7671), 219–226.

Dekker, J., Belmont, A. S., Guttman, M., Leshyk, V. O., Lis, J. T., Lomvardas, S., Mirny, L. A., O'Shea, C. C., Park, P. J., Ren, B., Politz, J. C. R., Shendure, J., Zhong, S., and the 4D Nucleome Network (2017b). Corrigendum: The 4d nucleome project. *Nature*, **552**, 278 EP –.

Denny, J. C., Bastarache, L., Ritchie, M. D., Carroll, R. J., Zink, R., Mosley, J. D., Field, J. R., Pulley, J. M., Ramirez, A. H., Bowton, E., Basford, M. A., Carrell, D. S., Peissig, P. L., Kho, A. N., Pacheco, J. A., Rasmussen, L. V., Crosslin, D. R., Crane, P. K., Pathak, J., Bielinski, S. J., Pendergrass, S. A., Xu, H., Hindorff, L. A., Li, R., Manolio, T. A., Chute, C. G., Chisholm, R. L., Larson, E. B., Jarvik, G. P., Brilliant, M. H., McCarty, C. A., Kullo, I. J., Haines, J. L., Crawford, D. C., Masys, D. R., and Roden, D. M. (2013). Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nature Biotechnology*, **31**(12), 1102–1111.

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**(7398), 376–380.

Dixon, J. R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J. E., Lee, A. Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W., Diao, Y., Liang, J., Zhao, H., Lobanenkov, V. V., Ecker, J. R., Thomson, J. A., and Ren, B. (2015). Chromatin architecture reorganization during stem cell differentiation. *Nature*, **518**(7539), 331–336.

Dixon, J. R., Gorkin, D. U., and Ren, B. (2016). Chromatin domains: The unit of chromosome organization. *Molecular Cell*, **62**(5), 668–680.

Forcato, M., Nicoletti, C., Pal, K., Livi, C. M., Ferrari, F., and Bicciato, S. (2017). Comparison of computational methods for hi-c data analysis. *Nature methods*, **14**(7), 679–685.

Fulco, C. P., Munschauer, M., Anyoha, R., Munson, G., Grossman, S. R., Perez, E. M., Kane, M., Cleary, B., Lander, E. S., and Engreitz, J. M. (2016). Systematic mapping of functional enhancer–promoter connections with crispr interference. *Science*, **354**(6313), 769–773.

Garner, C., Tatu, T., Reittie, J., Littlewood, T., Darley, J., Cervino, S., Farrall, M., Kelly, P., Spector, T., and Thein, S. (2000). Genetic influences on f cells and other hematologic variables: a twin heritability study. *Blood, The Journal of the American Society of Hematology*, **95**(1), 342–346.

Geschwind, D. H. and Konopka, G. (2009). Neuroscience in the era of functional genomics and systems biology. *Nature*, **461**(7266), 908–915.

Gilbert, L. A., Larson, M. H., Morsut, L., Liu, Z., Brar, G. A., Torres, S. E., Stern-Ginossar, N., Brandman, O., Whitehead, E. H., Doudna, J. A., *et al.* (2013). Crispr-mediated modular rna-guided regulation of transcription in eukaryotes. *Cell*, **154**(2), 442–451.

Giusti-Rodrguez, P., Lu, L., Yang, Y., Crowley, C. A., Liu, X., Juric, I., Martin, J. S., Abnousi, A., Allred, S. C., Ancalade, N., Bray, N. J., Breen, G., Bryois, J., Bulik, C. M., Crowley, J. J., Guintivano, J., Jansen, P. R., Jurjus, G. J., Li, Y., Mahajan, G., Marzi, S., Mill, J., ODonovan, M. C., Overholser, J. C., Owen, M. J., Pardias, A. F., Pochareddy, S., Posthuma, D., Rajkowska, G., Santpere, G., Savage, J. E., Sestan, N., Shin, Y., Stockmeier, C. A., Walters, J. T., Yao, S., Crawford, G. E., Jin, F., Hu, M., Li, Y., and Sullivan, P. F. (2019). Using three-dimensional regulatory chromatin interactions from adult and fetal cortex to interpret genetic results for psychiatric disorders and cognitive traits. *bioRxiv*, page 406330.

Gorkin, D. U., Qiu, Y., Hu, M., Fletez-Brant, K., Liu, T., Schmitt, A. D., Noor, A., Chiou, J., Gaulton, K. J., Sebat, J., Li, Y., Hansen, K. D., and Ren, B. (2019). Common dna sequence variation influences 3-dimensional conformation of the human genome. *Genome biology*, **20**(1), 255–255. PMC6883528[pmcid].

Haddy, T. B., Rana, S. R., and Castro, O. (1999). Benign ethnic neutropenia: what is a normal absolute neutrophil count? *Journal of Laboratory and Clinical Medicine*, **133**(1), 15–22.

Halvorsen, M., Huh, R., Oskolkov, N., Wen, J., Netotea, S., Giusti-Rodriguez, P., Karlsson, R., Bryois, J., Nystedt, B., Ameur, A., Kahler, A. K., Ancalade, N., Farrell, M., Crowley, J. J., Li, Y., Magnusson, P. K. E., Gyllensten, U., Hultman, C. M., Sullivan, P. F., and Szatkiewicz, J. P. (2020). Increased burden of ultra-rare structural variants localizing to boundaries of topologically associated domains in schizophrenia. *Nat Commun*, **11**(1), 1842.

Hodonsky, C. J., Jain, D., Schick, U. M., Morrison, J. V., Brown, L., McHugh, C. P., Schurmann, C., Chen, D. D., Liu, Y. M., Auer, P. L., *et al.* (2017). Genome-wide association study of red blood cell traits in hispanics/latinos: The hispanic community health study/study of latinos. *PLoS genetics*, **13**(4), e1006760.

Hu, B., Won, H., Mah, W., Park, R., Kassim, B., Spiess, K., Kozlenkov, A., Crowley, C. A., Pochareddy, S., Li, Y., Dracheva, S., Sestan, N., Akbarian, S., and Geschwind, D. H. (2020). Neuronal and glial 3d chromatin architecture illustrates cellular etiology of brain disorders. *bioRxiv*, page 2020.05.14.096917.

Hu, M., Deng, K., Selvaraj, S., Qin, Z., Ren, B., and Liu, J. S. (2012). Hicnorm: removing biases in hi-c data via poisson regression. *Bioinformatics*, **28**(23), 3131–3133.

Imakaev, M., Fudenberg, G., McCord, R. P., Naumova, N., Goloborodko, A., Lajoie, B. R., Dekker, J., and Mirny, L. A. (2012). Iterative correction of hi-c data reveals hallmarks of chromosome organization. *Nature methods*, **9**(10), 999.

Javierre, B. M., Burren, O. S., Wilder, S. P., Kreuzhuber, R., Hill, S. M., Sewitz, S., Cairns, J., Wingett, S. W., Várnai, C., Thiecke, M. J., *et al.* (2016). Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. *Cell*, **167**(5), 1369–1384.

Jin, F., Li, Y., Dixon, J. R., Selvaraj, S., Ye, Z., Lee, A. Y., Yen, C.-A., Schmitt, A. D., Espinoza, C., and Ren, B. (2013). A high-resolution map of three-dimensional chromatin interactome in human cells. *Nature*, **503**(7475), 290–294.

Jung, I., Schmitt, A., Diao, Y., Lee, A. J., Liu, T., Yang, D., Tan, C., Eom, J., Chan, M., Chee, S., Chiang, Z., Kim, C., Masliah, E., Barr, C. L., Li, B., Kuan, S., Kim, D., and Ren, B. (2019). A compendium of promoter-centered long-range chromatin interactions in the human genome. *Nat Genet*, **51**(10), 1442–1449.

Kanai, M., Akiyama, M., Takahashi, A., Matoba, N., Momozawa, Y., Ikeda, M., Iwata, N., Ikegawa, S., Hirata, M., Matsuda, K., *et al.* (2018). Genetic analysis of quantitative traits in the japanese population links cell types to complex human diseases. *Nature genetics*, **50**(3), 390–400.

Kaul, A., Bhattacharyya, S., and Ay, F. (2020). Identifying statistically significant chromatin contacts from hi-c data with fithic2. *Nat Protoc*, **15**(3), 991–1012.

Klann, T. S., Black, J. B., Chellappan, M., Safi, A., Song, L., Hilton, I. B., Crawford, G. E., Reddy, T. E., and Gersbach, C. A. (2017). Crispr–cas9 epigenome editing enables high-throughput screening for functional regulatory elements in the human genome. *Nature biotechnology*, **35**(6), 561.

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J.-Y., Sackler, R. S., Haynes, C., Henning, A. K., San-Giovanni, J. P., Mane, S. M., Mayne, S. T., *et al.* (2005). Complement factor h polymorphism in age-related macular degeneration. *Science*, **308**(5720), 385–389.

Krijger, P. H. L. and de Laat, W. (2016). Regulation of disease-associated gene expression in the 3d genome. *Nature Reviews Molecular Cell Biology*, **17**(12), 771–782.

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., Wu, Y.-C., Pfenning, A., Wang, X., ClaussnitzerYaping Liu, M., Coarfa, C., Alan Harris, R., Shoresh, N., Epstein, C. B., Gjoneska, E., Leung, D., Xie, W., David Hawkins, R., Lister, R., Hong, C., Gascard, P., Mungall, A. J., Moore, R., Chuah, E., Tam, A., Canfield, T. K., Scott Hansen, R., Kaul, R., Sabo, P. J., Bansal, M. S., Carles, A., Dixon, J. R., Farh, K.-H., Feizi, S., Karlic, R., Kim, A.-R., Kulkarni, A., Li, D., Lowdon, R., Elliott, G., Mercer, T. R., Neph, S. J., Onuchic, V., Polak, P., Rajagopal, N., Ray, P., Sallari, R. C., Siebenthall, K. T., Sinnott-Armstrong, N. A., Stevens, M., Thurman, R. E., Wu, J., Zhang, B., Zhou, X., Abdennur, N., Adli, M., Akerman, M., Barrera, L., Antosiewicz-Bourget, J., Ballinger, T., Barnes, M. J., Bates, D., Bell, R. J. A., Bennett, D. A., Bianco, K., Bock, C., Boyle, P., Brinchmann, J., Caballero-Campo, P., Camahort, R., Carrasco-Alfonso, M. J., Charnecki, T., Chen, H., Chen, Z., Cheng, J. B., Cho, S., Chu, A., Chung, W.-Y., Cowan, C., Athena Deng, Q., Deshpande, V., Diegel, M., Ding, B., Durham, T., Echipare, L., Edsall, L., Flowers, D., Genbacev-Krtolica, O., *et al.* (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, **518**(7539), 317–330.

Lajoie, B. R., Dekker, J., and Kaplan, N. (2015). The hitchhiker's guide to hi-c analysis: Practical guidelines. *Methods (San Diego, Calif.)*, **72**, 65–75.

Li, M., Santpere, G., Imamura Kawasawa, Y., Evgrafov, O. V., Gulden, F. O., Pochareddy, S., Sunkin, S. M., Li, Z., Shin, Y., Zhu, Y., Sousa, A. M. M., Werling, D. M., Kitchen, R. R., Kang, H. J., Pletikos, M., Choi, J., Muchnik, S., Xu, X., Wang, D., Lorente-Galdos, B., Liu, S., Giusti-Rodrguez, P., Won, H., deLeeuw, C. A., Pardias, A. F., Hu, M., Jin, F., Li, Y., Owen, M. J., ODonovan, M. C., Walters, J. T. R., Posthuma, D., Reimers, M. A., Levitt, P., Weinberger, D. R., Hyde, T. M., Kleinman, J. E., Geschwind, D. H., Hawrylycz, M. J., State, M. W., Sanders, S. J., Sullivan, P. F., Gerstein, M. B., Lein, E. S., Knowles, J. A., and Sestan, N. (2018a). Integrative functional genomic analysis of human brain development and neuropsychiatric risks. *Science*, **362**(6420), eaat7615.

Li, Y., Hu, M., and Shen, Y. (2018b). Gene regulation in the 3D genome. *Human Molecular Genetics*, **27**(R2), R228–R233.

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S., and Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**(5950), 289–293.

Liu, X., White, S., Peng, B., Johnson, A. D., Brody, J. A., Li, A. H., Huang, Z., Carroll, A., Wei, P., Gibbs, R., *et al.* (2016). Wgsa: an annotation pipeline for human genome sequencing studies. *Journal of medical genetics*, **53**(2), 111–112.

Lloyd-Jones, L. R., Holloway, A., McRae, A., Yang, J., Small, K., Zhao, J., Zeng, B., Bakshi, A., Metspalu, A., Dermitzakis, M., *et al.* (2017). The genetic architecture of gene expression in peripheral blood. *The American Journal of Human Genetics*, **100**(2), 228–237.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., *et al.* (2013). The genotype-tissue expression (gtex) project. *Nature genetics*, **45**(6), 580.

Lu, L., Liu, X., Peng, J., Li, Y., and Jin, F. (2018). Easy hi-c: A simple efficient protocol for 3d genome mapping in small cell populations. *bioRxiv*, page 245688.

Ludwig, L. S., Lareau, C. A., Bao, E. L., Nandakumar, S. K., Muus, C., Ulirsch, J. C., Chowdhary, K., Buenrostro, J. D., Mohandas, N., An, X., Aryee, M. J., Regev, A., and Sankaran, V. G. (2019). Transcriptional states and chromatin accessibility underlying human erythropoiesis. *Cell Rep*, **27**(11), 3228–3240 e7.

Ma, W., Ay, F., Lee, C., Gulsoy, G., Deng, X., Cook, S., Hesson, J., Cavanaugh, C., Ware, C. B., Krumm, A., Shendure, J., Blau, C. A., Disteche, C. M., Noble, W. S., and Duan, Z. (2015). Fine-scale chromatin interaction maps reveal the cis-regulatory landscape of human lincrna genes. *Nature methods*, **12**(1), 71–78. nmeth.3205[PII].

Mancuso, N., Freund, M. K., Johnson, R., Shi, H., Kichaev, G., Gusev, A., and Pasaniuc, B. (2019). Probabilistic fine-mapping of transcriptome-wide association studies. *Nature genetics*, **51**(4), 675–682.

Maricque, B. B., Dougherty, J. D., and Cohen, B. A. (2017). A genome-integrated massively parallel reporter assay reveals dna sequence determinants of cis-regulatory activity in neural cells. *Nucleic acids research*, **45**(4), e16–e16.

Martin, J. S., Xu, Z., Reiner, A. P., Mohlke, K. L., Sullivan, P., Ren, B., Hu, M., and Li, Y. (2017). Hugin: Hi-c unifying genomic interrogator. *Bioinformatics (Oxford, England)*, **33**(23), 3793–3795. 3861336[PII].

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P., and Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature reviews genetics*, **9**(5), 356–369.

McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., Wenger, A. M., and Bejerano, G. (2010). Great improves functional interpretation of cis-regulatory regions. *Nature biotechnology*, **28**(5), 495.

Mgi, R., Horikoshi, M., Sofer, T., Mahajan, A., Kitajima, H., Franceschini, N., McCarthy, M. I., COGENT-Kidney Consortium, T.-G. C., and Morris, A. P. (2017). Trans-ethnic meta-regression of genome-wide association studies accounting for ancestry increases power for discovery and improves fine-mapping resolution. *Human Molecular Genetics*, **26**(18), 3639–3650.

Naumova, N. and Dekker, J. (2010). Integrating one-dimensional and three-dimensional maps of genomes. *Journal of cell science*, **123**(12), 1979–1988.

Pardiñas, A. F., Holmans, P., Pocklington, A. J., Escott-Price, V., Ripke, S., Carrera, N., Legge, S. E., Bishop, S., Cameron, D., Hamshere, M. L., *et al.* (2018). Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nature genetics*, **50**(3), 381–389.

Patel, K. V., Longo, D. L., Ershler, W. B., Yu, B., Semba, R. D., Ferrucci, L., and Guralnik, J. M. (2009). Haemoglobin concentration and the risk of death in older adults: differences by race/ethnicity in the nhanes iii follow-up. *British journal of haematology*, **145**(4), 514–523.

Phanstiel, D. H., Van Bortle, K., Spacek, D., Hess, G. T., Shamim, M. S., Machol, I., Love, M. I., Aiden, E. L., Bassik, M. C., and Snyder, M. P. (2017). Static and dynamic dna loops form ap-1-bound activation hubs during macrophage development. *Molecular cell*, **67**(6), 1037–1048.

Polderman, T. J., Benyamin, B., De Leeuw, C. A., Sullivan, P. F., Van Bochoven, A., Visscher, P. M., and Posthuma, D. (2015). Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature genetics*, **47**(7), 702.

Raj, T., Rothamel, K., Mostafavi, S., Ye, C., Lee, M. N., Replogle, J. M., Feng, T., Lee, M., Asinovski, N., Frohlich, I., *et al.* (2014). Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. *Science*, **344**(6183), 519–523.

Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S., and Aiden, E. L. (2014). A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**(7), 1665–1680. S0092-8674(14)01497-4[PII].

Reiner, A. P., Lettre, G., Nalls, M. A., Ganesh, S. K., Mathias, R., Austin, M. A., Dean, E., Arepalli, S., Britton, A., Chen, Z., *et al.* (2011). Genome-wide association study of white blood cell count in 16,388 african americans: the continental origins and genetic epidemiology network (cogent). *PLoS genetics*, **7**(6).

Sanyal, A., Lajoie, B., Jain, G., and Dekker, J. (2012). The long-range interaction landscape of gene promoters. *Nature*, **489**(7414), 109–113.

Schaid, D. J., Chen, W., and Larson, N. B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics*, **19**(8), 491–504.

Schmitt, A. D., Hu, M., Jung, I., Xu, Z., Qiu, Y., Tan, C. L., Li, Y., Lin, S., Lin, Y., Barr, C. L., *et al.* (2016a). A compendium of chromatin contact maps reveals spatially active regions in the human genome. *Cell reports*, **17**(8), 2042–2059.

Schmitt, A. D., Hu, M., and Ren, B. (2016b). Genome-wide mapping and analysis of chromosome architecture. *Nature reviews. Molecular cell biology*, **17**(12), 743–755. nrm.2016.104[PII].

Song, M., Yang, X., Ren, X., Maliskova, L., Li, B., Jones, I. R., Wang, C., Jacob, F., Wu, K., Traglia, M., Tam, T. W., Jamieson, K., Lu, S. Y., Ming, G. L., Li, Y., Yao, J., Weiss, L. A., Dixon, J. R., Judge, L. M., Conklin, B. R., Song, H., Gan, L., and Shen, Y. (2019). Mapping cis-regulatory chromatin contacts in neural cells links neuropsychiatric disorder risk variants to target genes. *Nat Genet*, **51**(8), 1252–1262.

Stefansson, H., Ophoff, R. A., Steinberg, S., Andreassen, O. A., Cichon, S., Rujescu, D., Werge, T., Pietiläinen, O. P. H., Mors, O., Mortensen, P. B., Sigurdsson, E., Gustafsson, O., Nyegaard, M., Tuulio-Henriksson, A., Ingason, A., Hansen, T., Suvisaari, J., Lonnqvist, J., Paunio, T., Børglum, A. D., Hartmann, A., Fink-Jensen, A., Nordentoft, M., Hougaard, D., Norgaard-Pedersen, B., Böttcher, Y., Olesen, J., Breuer, R., Möller, H.-J., Giegling, I., Rasmussen, H. B., Timm, S., Mattheisen, M., Bitter, I., Réthelyi, J. M., Magnusdottir, B. B., Sigmundsson, T., Olason, P., Masson, G., Gulcher, J. R., Haraldsson, M., Fossdal, R., Thorgeirsson, T. E., Thorsteinsdottir, U., Ruggeri, M., Tosato, S., Franke, B., Strengman, E., Kiemeney, L. A., (GROUP), G. R., in Psychosis, O., Melle, I., Djurovic, S., Abramova, L., Kaleda, V., Sanjuan, J., de Frutos, R., Bramon, E., Vassos, E., Fraser, G., Ettinger, U., Picchioni, M., Walker, N., Toulopoulou, T., Need, A. C., Ge, D., Yoon, J. L., Shianna, K. V., Freimer, N. B., Cantor, R. M., Murray, R., Kong, A., Golimbet, V., Carracedo, A., Arango, C., Costas, J., Jönsson, E. G., Terenius, L., Agartz, I., Petursson, H., Nöthen, M. M., Rietschel, M., Matthews, P. M., Muglia, P., Peltonen, L., St Clair, D., Goldstein, D. B., Stefansson, K., and Collier, D. A. (2009). Common variants conferring risk of schizophrenia. *Nature*, **460**(7256), 744–747. nature08186[PII].

Sullivan, P. F. and Geschwind, D. H. (2019). Defining the genetic, genomic, cellular, and diagnostic architectures of psychiatric disorders. *Cell*, **177**(1), 162–183.

Sullivan, P. F., Kendler, K. S., and Neale, M. C. (2003). Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. *Archives of general psychiatry*, **60**(12), 1187–1192.

Sullivan, P. F., Daly, M. J., and O'donovan, M. (2012). Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nature Reviews Genetics*, **13**(8), 537–551.

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., and Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, **20**(8), 467–484.

Tjong, H., Li, W., Kalhor, R., Dai, C., Hao, S., Gong, K., Zhou, Y., Li, H., Zhou, X. J., Le Gros, M. A., Larabell, C. A., Chen, L., and Alber, F. (2016). Population-based 3d genome structure analysis reveals driving forces in spatial genome organization. *Proceedings of the National Academy of Sciences of the United States of America*, **113**(12), E1663–E1672. 1512577113[PII].

Ulirsch, J. C., Lareau, C. A., Bao, E. L., Ludwig, L. S., Guo, M. H., Benner, C., Satpathy, A. T., Kartha, V. K., Salem, R. M., Hirschhorn, J. N., Finucane, H. K., Aryee, M. J., Buenrostro, J. D., and Sankaran, V. G. (2019). Interrogation of human hematopoiesis at single-cell and single-variant resolution. *Nat Genet*, **51**(4), 683–693.

van Berkum, N. L., Lieberman-Aiden, E., Williams, L., Imakaev, M., Gnirke, A., Mirny, L. A., Dekker, J., and Lander, E. S. (2010). Hi-c: A method to study the three-dimensional architecture of genomes. *Journal of Visualized Experiments : JoVE*, (39), 1869.

Visscher, P. M., Brown, M. A., McCarthy, M. I., and Yang, J. (2012). Five years of gwas discovery. *The American Journal of Human Genetics*, **90**(1), 7–24.

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., and Yang, J. (2017). 10 years of gwas discovery: biology, function, and translation. *The American Journal of Human Genetics*, **101**(1), 5–22.

Võsa, U., Claringbould, A., Westra, H.-J., Bonder, M. J., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Kasela, S., *et al.* (2018). Unraveling the polygenic architecture of complex traits using blood eqtl meta-analysis. *BioRxiv*, page 447367.

Wang, D., Liu, S., Warrell, J., Won, H., Shi, X., Navarro, F. C. P., Clarke, D., Gu, M., Emani, P., Yang, Y. T., Xu, M., Gandal, M. J., Lou, S., Zhang, J., Park, J. J., Yan, C., Rhie, S. K., Manakongtreecheep, K., Zhou, H., Nathan, A., Peters, M., Mattei, E., Fitzgerald, D., Brunetti, T., Moore, J., Jiang, Y., Girdhar, K., Hoffman, G. E., Kalayci, S., Gümüs, Z. H., Crawford, G. E., Consortium, P., Roussos, P., Akbarian, S., Jaffe, A. E., White, K. P., Weng, Z., Sestan, N., Geschwind, D. H., Knowles, J. A., and Gerstein, M. B. (2018). Comprehensive functional genomic resource and integrative model for the human brain. *Science (New York, N.Y.)*, **362**(6420), eaat8464. 362/6420/eaat8464[PII].

Wang, H., La Russa, M., and Qi, L. S. (2016). Crispr/cas9 in genome editing and beyond. *Annual review of biochemistry*, **85**, 227–264.

Watson, H. J., Yilmaz, Z., Thornton, L. M., Hübel, C., Coleman, J. R., Gaspar, H. A., Bryois, J., Hinney, A., Leppä, V. M., Mattheisen, M., *et al.* (2019). Genome-wide association study identifies eight risk loci and implicates metabo-psychiatric origins for anorexia nervosa. *Nature genetics*, **51**(8), 1207–1214.

Whyte, W. A., Orlando, D. A., Hnisz, D., Abraham, B. J., Lin, C. Y., Kagey, M. H., Rahl, P. B., Lee, T. I., and Young, R. A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, **153**(2), 307–319. S0092-8674(13)00392-9[PII].

Wojcik, G. L., Graff, M., Nishimura, K. K., Tao, R., Haessler, J., Gignoux, C. R., Highland, H. M., Patel, Y. M., Sorokin, E. P., Avery, C. L., *et al.* (2019). Genetic analyses of diverse populations improves discovery for complex traits. *Nature*, **570**(7762), 514–518.

Won, H., de la Torre-Ubieta, L., Stein, J. L., Parikshak, N. N., Huang, J., Opland, C. K., Gandal, M. J., Sutton, G. J., Hormozdiari, F., Lu, D., Lee, C., Eskin, E., Voineagu, I., Ernst, J., and Geschwind, D. H. (2016). Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature*, **538**(7626), 523–527. nature19847[PII].

Xu, Z., Zhang, G., Wu, C., Li, Y., and Hu, M. (2016a). Fasthic: a fast and accurate algorithm to detect long-range chromosomal interactions from hi-c data. *Bioinformatics (Oxford, England)*, **32**(17), 2692–2695. btw240[PII].

Xu, Z., Zhang, G., Jin, F., Chen, M., Furey, T. S., Sullivan, P. F., Qin, Z., Hu, M., and Li, Y. (2016b). A hidden markov random field-based bayesian method for the detection of long-range chromosomal interactions in hi-c data. *Bioinformatics (Oxford, England)*, **32**(5), 650–656. btv650[PII].

Yaffe, E. and Tanay, A. (2011). Probabilistic modeling of hi-c contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat Genet*, **43**(11), 1059–65.

Yan, K.-K., Yardimci, G. G., Yan, C., Noble, W. S., and Gerstein, M. (2017). Hic-spector: a matrix library for spectral and reproducibility analysis of hi-c contact maps. *Bioinformatics*, **33**(14), 2199–2201. btx152[PII].

Yang, T., Zhang, F., Yardimci, G. G., Song, F., Hardison, R. C., Noble, W. S., Yue, F., and Li, Q. (2017). Hicrep: assessing the reproducibility of hi-c data using a stratum-adjusted correlation coefficient. *Genome Res*, **27**(11), 1939–1949. 28855260[pmid].

Yardmc, G. G., Ozadam, H., Sauria, M. E. G., Ursu, O., Yan, K.-K., Yang, T., Chakraborty, A., Kaul, A., Lajoie, B. R., Song, F., Zhan, Y., Ay, F., Gerstein, M., Kundaje, A., Li, Q., Taylor, J., Yue, F., Dekker, J., and Noble, W. S. (2018). Measuring the reproducibility and quality of hi-c data. *bioRxiv*.

Yu, M. and Ren, B. (2017). The three-dimensional organization of mammalian genomes. *Annu Rev Cell Dev Biol*, **33**, 265–289.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., and Liu, X. S. (2008). Model-based analysis of chip-seq (macs). *Genome Biol*, **9**(9), R137.

Zhernakova, D. V., Deelen, P., Vermaat, M., Van Iterson, M., Van Galen, M., Arindrarto, W., Van't Hof, P., Mei, H., Van Dijk, F., Westra, H.-J., *et al.* (2017). Identification of context-dependent expression quantitative trait loci in whole blood. *Nature genetics*, **49**(1), 139–145.