Imputation of 3D genome structure by genetic-epigenetic interaction modeling in mice

A thesis submitted by

Lauren Kuffler

In partial fulfillment of the requirements for the degree of

PhD

in

Genetics

Tufts University

Graduate School of Biomedical Sciences

August 2023

Advisor: Gregory W. Carter, PhD

Abstract

Gene expression is known to be affected by interactions between local genetic variation and DNA accessibility, with the latter organized into three-dimensional chromatin structures. Analyses of these interactions has previously been limited, obscuring their regulatory context, and the extent to which they occur throughout the genome. Here we undertake a genome-scale analysis of these interactions in a genetically diverse population to systematically identify global genetic-epigenetic interaction, and reveal constraints imposed by chromatin structure. We establish the extent and structure of genotype-by-epigenotype interaction using embryonic stem cells derived from Diversity Outbred mice. This mouse population segregates millions of variants from eight inbred founders, enabling precision genetic mapping with extensive genotypic and phenotypic diversity. With 176 samples profiled for genotype, gene expression, and open chromatin, we used regression modeling to infer genetic-epigenetic interactions on a genome-wide scale. Our results demonstrate that statistical interactions between genetic variants and chromatin accessibility are common throughout the genome. We found that these interactions occur within the local area of the affected gene, and that this locality corresponds to topologically associated domains (TADs). The likelihood of interaction was most strongly defined by the three-dimensional (3D) domain structure rather than linear DNA sequence. We show that stable 3D genome structure is an effective tool to guide searches for regulatory elements and, conversely, that interacting regulatory elements in genetically diverse populations provide a means to infer 3D genome structure. We confirmed this finding with CTCF ChIP-seq that revealed strain-specific binding in the inbred founder mice. In stem cells, open chromatin participating in the most significant regression models demonstrated an enrichment for developmental genes and the TAD-forming CTCF binding complex, providing an opportunity for statistical inference of shifting TAD boundaries operating during early development.

These findings provide evidence that genetic and epigenetic factors operate within the context of three-dimensional chromatin structure (Reproduced with permission from Kuffler et al., 2023).

Acknowledgments

We thank G. Churchill for helpful comments. This study was funded by the National Institute of General Medical Sciences of the National Institutes of Health under award number R01GM115518. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Table of Contents

List of Tables

List of Figures

List of Copyrighted Materials Used

Kuffler Lauren, Fortin Haley J., Skelly Daniel A., Czechanski Anne, Munger Steven C., Reinholdt Laura G., Baker Christopher L., Carter Gregory W. (2023) Imputation of 3D genome structure by genetic-epigenetic interaction modeling in mice eLife 12:RP88222 https://doi.org/10.7554/eLife.88222.1.

List of Abbreviations

1i medium: mESC growth medium containing only G3SK inhibitor
ATAC-seq: Assay of Transposase-Accessible Chromatin sequencing
CC: Collaborative Cross mouse program
CC-RIX: CC-recombinant intercross mice
CTCF: CCCTC-binding factor
Ctcfl: CTCF-like
DO: Diversity Outbred mouse program
(e)QTL: (expression) quantitative trait locus
GigaMUGA: third generation Mouse Universal Genotyping Array
Kb(p): kilobase(pair)
Main effects: single effect regression model terms
Mb(p): megabase(pair)
mESC: mouse embryonic stem cell
RNA-seq: RNA sequencing
SNP: Single Nucleotide Polymorphism
TAD: Topologically Associated Domain
TMM: trimmed mean of *M* values
TPM: log2-transformed transcripts per million
TSS: transcription start site

Chapter 1: Introduction

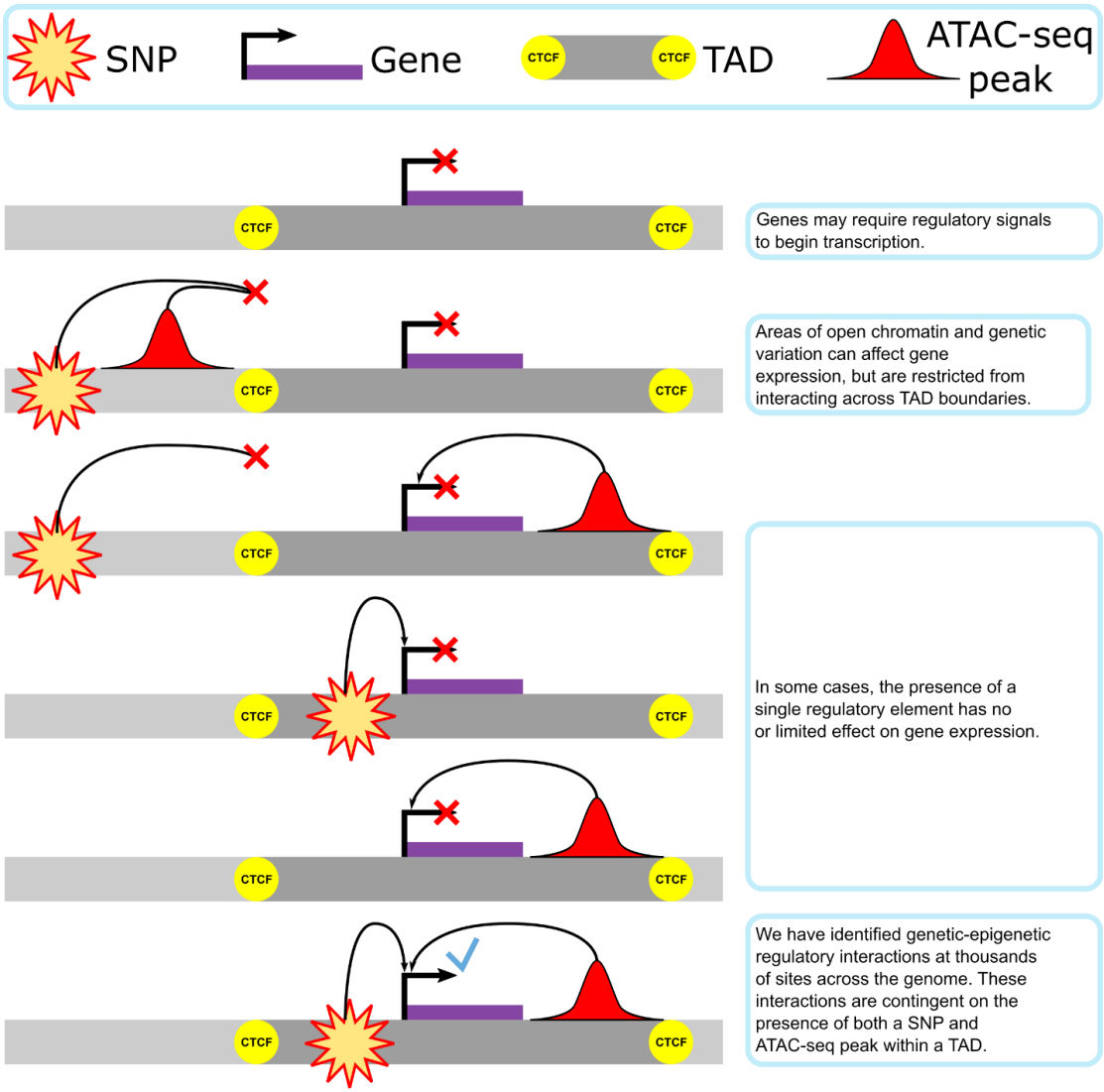


Figure 1.1. A graphical abstract describing genetic-epigenetic interactions, and how TAD boundaries constrain their occurrence. Reprinted with permission from *Kuffler L, Fortin HJ, Skelly DA, Czechanski A, Munger SC., Reinholdt LG, Baker CL, Carter GW. (2023) Imputation of 3D genome structure by genetic-epigenetic interaction modeling in mice eLife 12:RP88222 https://doi.org/10.7554/eLife.88222.1.*

1.1 Regulation of Gene Expression

Genotypic variation may affect gene expression by altering either coding or non-coding DNA. Alteration of coding regions can affect the gene's expression through directly altering the chance of successfully completing transcription. As biological sciences have

progressed, the field has shifted from a tight focus on genes and their DNA sequence. Where previously non-coding regions were once referred to as "junk DNA", we now understand that there are many levels of genetic regulation that depend on the genetic sequence, protein binding, and chromatin structure of non-coding DNA. Without the information contained in these regions and epigenetic factors, cells cannot maintain their identity, nor develop and differentiate into new tissues.

Non-coding sequence variation affects regulatory factors that alter chromatin accessibility, recruitment of transcription factors, and formation of local 3D genome structure that promotes transcription. These are forms of epigenetic modification, a broad class of regulatory factors that can more generally include any DNA-protein binding, DNA conformation, or DNA modification. Some of these regulatory factors are consistent enough in their identity that they can be termed an 'epigenotype'(Waddington, 2012).

The physical conformation of chromatin provides a vital layer of gene regulation. It is common to assay the location and modifications on histones, which provide information on both chromatin accessibility, as well as priming or suppression in protein binding and transcription(Waddington, 2012).

Nucleosomes are an organizing factor on the level of 146 bp, and act as physical barriers to other proteins. These are the most fundamental and therefore lowest order of chromatin organization. Higher orders of chromatin organization produce varied effects of their own, but were historically not so commonly studied, outside of a few broad categories.

Since the 1920s, researchers have noticed the difference between highly compacted heterochromatin and more open euchromatin(Babu & Verma, 1987; Berger, 2019). With the advent of Hi-C, the similar categories of chromatin compartments A and B were identified through principle component analysis of chromatin accessibility, above or below expected signal strength (Lieberman-Aiden et al., 2009). Cell type-specific patterns of chromatin compartments were noted(H. Gong et al., 2021). These compartments are on a multi-megabase scale, and were eventually identified as defined by a number of chromatin states roughly correlating to euchromatin and heterochromatin.

1.2 Regulatory Interactions

The presence or absence of epigenetic factors are often predicated on underlying genotype, and local epigenotype may determine the ability of a genetic variation to affect gene expression. These interactions are by definition non-additive, as these factors are either dependent on one another, or not fully independent. This is similar to epistatic mechanisms in gene-gene networks: regulatory factors may work together, rely on one another, interfere with each other's function, or even mixtures of these events. These interactions have previously been demonstrated in many contexts(Bekris et al., 2012; Gacita et al., 2021; Perez-Martinez et al., 2011). Yet the interacting effects that genetic and epigenetic factors produce on gene transcription are rarely studied at a genome-wide scale, leaving us without global information on a key step between the genetic code and the phenotype.

This limited scope has likewise limited the conclusions that can be drawn from the available literature. These studies are unable to determine whether these interactions are common and widespread, or limited to specific genes that are already tractable to study by current methods.

1.3 Multi-omics Analysis

To identify genetic-epigenetic non-additive interactions and characterize their behavior on a genome-wide scale, multiple data types are required. Genetic analyses, epigenetic analyses, and quantification of gene expression in genetically diverse populations are required, but such multiomics datasets are extremely rare. At the start of this project, there were no publicly available datasets containing such multi-omics data in genetically diverse samples, and no known studies prioritizing their creation.

This limits our ability to comprehend phenotype on a fundamental level. Heritable phenotypic variation in genetically diverse populations is a result of both genetic and epigenetic factors operating in tandem. Understanding the scope and landscape of these interactions on a genome-wide scale is a vital step towards deciphering the genetic regulation of gene expression and, in turn, the mechanisms of non-coding variation on phenotypic outcomes. This project was undertaken to fill this gap.

Epigenetic analyses must determine the likelihood of a gene interacting with a putative regulatory element. To do so, most researchers default to the linear genome distance between the two features. This is supported by decades of results showing the importance of direct physical interactions between nearby regulatory elements and genes(Bartkuhn & Renkawitz, 2008; Carter et al., 2002; Fraser & Grosveld, 1998;

Greenwald et al., 2019), as well as alterations to chromatin compaction and recruitment of transcriptional machinery to nearby genes(Workman & Kingston, 1998; You et al., 2021).

1.4 Local Regulatory Area

These short-range interactions have led to the concept of a "local area". This is generally defined as the region in which direct chromatin contact with regulatory elements is expected, or indirect regulatory effects will consistently reach a given gene. Beyond this local area, interactions are far less frequent, and considered to either be uniformly indirect, random, or another class of distal interaction.

Local areas have been defined in a number of different ways, based off of previous experimental findings and the limits of computational power. Many have used 100 to 500 kb flanking windows around a gene to define a local area(Gate et al., 2018; Liu et al., 2017; Luo et al., 2016; West et al., 2016). This produces consistent results(Kim et al., 2014), but the interval size is not linked to any particular biological feature. It is a construct of convenience.

Arbitrary limits on analysis scope may be appropriate for research that does not primarily focus on local regulatory function, but it limits the completeness and rigor of the result. Statistical distributions in biological science are often abstractions of discrete functions, rather than a stochastic distribution. One classic cell biology example is the mechanism of action potential across the neuromuscular junction. These phenomena were observed to have a remarkably precise and replicable quantization of their behavior, which turned out to be the result of the packaging of acetylcholine into vesicles of consistent size for

transmembrane release(Del Castillo & Katz, 1954). Another example from genetics would be the location of recombination hotspots. These regions were identifiable through study of linkage disequilibrium, but their mechanism was unknown(Jeffreys et al., 2001). These hotspots were found to be the result of PRDM9, a single, fast-evolving gene(Baudat et al., 2010; Myers et al., 2010; Parvanov et al., 2010).

1.5 Topologically Associating Domains

Early epigenetic analysis identified that physical proximity between genes and their regulatory elements was often key to their expression(Brown et al., 2006). As a result, chromatin looping was identified as a potential insulator and facilitator of gene-enhancer interactions(Chambeyron & Bickmore, 2004).

One important chromatin loop structure is the topologically associated domain (TAD). TADs were first identified via chromosome conformation capture(Dekker et al., 2002), and named in the context of the mouse X-inactivation(Nora et al., 2012). Their definition has shifted over time, and differs between model organisms. In older literature, a TAD may refer to a region of self-associating chromatin, essentially small, self-contained regions of chromatin compartment A that are cell type invariant, building from their initial identification around X-inactivation. This loose definition led to different domains of study converging on multiple distinct epigenetic features being identified as "TADs", depending on the model organism in use(Hou et al., 2012; Le et al., 2013; Sexton et al., 2012).

In mammals, TADs are now usually defined as a loop of DNA held together by cohesin, which is stabilized and anchored by the CCCTC-binding factor (CTCF). Cohesin complexes attach to DNA, quickly extruding chromatin in a loop(Maji et al., 2020). As

this loop lengthens, it brings genes and regulatory regions transiently into contact. If the cohesin encounters CTCF in the correct orientation, extrusion will halt along one direction on the strand(Kagey et al., 2010). When CTCF in the opposing orientation is encountered on the other end of the extrusion, CTCF will form a homodimer and anchor the newly formed TAD loop(Yusufzai et al., 2004).

CTCF can remain bound to chromatin from anywhere between 72 to more than 144 hours(Khoury et al., 2020), and has eleven zinc finger domains, creating a DNA binding site with a potential for variation. This long-lasting binding creates stable domains of open chromatin that allows active gene transcription.

CTCF binding creates a boundary between intra-TAD chromatin and extra-TAD chromatin, suppressing the ability of regulatory elements to cross this boundary(Hou et al., 2008). This behavior has been modeled as a block copolymer, as a potential explanation for why TAD compartmental segregation occurs(Nuebler et al., 2018). TADs contain more heterochromatin than non-TAD regions, allowing more regulation of and initiation of gene transcription. The reason why TAD-internal chromatin is more likely to be open is still unclear(Shinkai et al., 2016). However, it has been noted that regions containing super-enhancers tend to experience higher degrees of insulation than other regions(Y. Gong et al., 2018).

Within TADs, sub-TAD loops can form, sometimes referred to as chromatin domains(Shinkai et al., 2016). These can be created by a number of mechanisms, including CTCF binding(Tang et al., 2015). These loops are often less stable than TADs, acting as transient regulatory mechanisms to affect local gene expression.

TADs have been described as mostly cell-type invariant and conserved across species, which can be used to argue for a relatively inert, structural role in DNA organization. However, these studies identified substantial minorities of differential TAD binding sites, and also were subject to limited resolution of TAD binding site occupancy. CTCF binding sites are plentiful across relevant genomes, with many remaining unoccupied in any given cell type. It is not uncommon to see more than a third of binding sites unoccupied(Holzmann et al., 2019).

Other studies have found vital roles for shifts in CTCF binding patterns during embryonic development and determining cell fate(Arzate-Mejía et al., 2018). In chickens, mice, and humans, the HoxD gene cluster is subject to regulation during embryonic development by the shifting of a shared boundary between two TADs that contain the gene cluster(Le Caignec et al., 2020; Rodríguez-Carballo et al., 2017; Yakushiji-Kaminatsui et al., 2018). This boundary shift exposes the genes to different regulatory elements at different time points and in different developing tissues, resulting in orderly forelimb development. When this TAD boundary is abolished, congenital disorders of forelimb digit formation arise. Further studies have found that CTCF has profound effects during the creation of induced pluripotent stem cells, first silencing somatic genes, then encouraging enhancer-promoter interactions and increasing chromatin accessibility. This context-sensitive, genome-wide behavior is indicative of the multifaceted use of this protein in the restructuring of both DNA and cell type(Song et al., 2022).

Alteration of TAD structure is also found to produce pathogenic dysregulation in some contexts(Y. Gong et al., 2018). Various studies have shown that chromosomal

rearrangements that abolish or shift TAD boundaries are more likely to be carcinogenic, indicating that a disregulated cell type is created by their alteration(Akdemir et al., 2020; Taberlay et al., 2016). The creation of new TADs is also associated with chromosomal rearrangements in cancer cells(J. R. Dixon et al., 2018).

Because of the properties described above, TADs are increasingly hypothesized to provide physical boundaries for local regulatory function, but functional analysis of regulatory interactions within TADs has been limited by the lack of informative genetic variation in the studies that led to TAD discovery and the characterization of TAD behavior. Individual examples have been demonstrated in the publications described above, but these are likewise limited in their scope, and cannot speak to whether these patterns hold true genome-wide.

Many authors have argued against the regulatory importance of TADs outside of specific applications, particularly in adult organisms. Some publications have shown that when CTCF is depleted, TAD structures still form in mammalian genomes(Kubo et al., 2017). Some studies contest this, implying that these studies had not fully depleted CTCF that was constituently bound to the genome(Wutz et al., 2017). Other authors have contended that TAD boundary disruption may be important in the development of some cancers(Aitken et al., 2018), but other diseases seem to have no link to TAD-based regulation. However, the general use of isogenic models and lack of TAD boundary data makes it difficult to judge whether TADs are truly uninvolved in these diseases.

1.6 Model Systems

In many model organisms, reverse genetic approaches like saturation mutagenesis can be used to introduce functionally informative genetic regulation. These approaches are challenging in mammalian models at scale(Gupta & Varadarajan, 2018; Mason et al., 2018), requiring more time and with fewer viable results.

Even modern CRISPR-Cas9 based approaches are still limited in their scope and application outside of immortalized cell lines and stem cells (Serebrenik & Shalem, 2018). By comparison, samples from genetically diverse populations permit comprehensive study of complex genetic interaction and non-coding regulation on a genome-wide scale(Svenson et al., 2012).

Mice are excellent models for these studies, with well-characterized isogenic strains available from a variety of subspecies. Many of these strains have deeply sequenced genomes with comprehensive variant catalogs. Increasingly complex interbred and outbred lines have been created in mice, beginning with two-parent crosses like B6 and DBA(Davis et al., 2005), and expanding into more ambitious projects. The Collaborative Cross recombinant inbred panel (CC mice) is the result of eight founder strains from three subspecies: A/J, C57BL/6J, 129Sv/ImJ, NOD/LtJ, NZO/H1J, CAST/EiJ, PWK/PhJ, and WSB/EiJ(The Complex Trait Consortium, 2004). These grandparent strains were intercrossed for three generations and then offspring were inbred to create homozygous inbred lines.

The CC mice are a powerful tool for reproducibly studying the effects of genetic variation, but are subject to some difficulties. Infertility resulted in a loss of more lines

than expected (Shorter et al., 2017). This also means that the CC is not ideal for studying models of reproductive or developmental dysfunction. There is also unequal contribution from the CAST and PWK lines in the remaining CC strains, selection has occurred on chromosome X and Y, and in mitochondria(Srivastava et al., 2017).

The lack of heterozygosity in the CC population also cut off study of haplotype effects, parental chromosome imprinting, and other benefits of crossing model lines. The creation of CC-recombinant intercross mice (CC-RIX) was intended to address this problem, creating F1 crosses of CC lines(Y. Gong & Zou, 2012; Zou et al., 2005). Because this process is often undertaken by institutions that have their own CC colonies, CC-RIX are not necessarily fully comparable to each other, due to differing experimental design and lab environments. They allow a higher degree of precision in genetic mapping than the CC, but are still limited in their ability to model the diversity of a wild population.

Studying wild populations is complicated and expensive, lacking the ability to standardize environment. Only some populations are amenable to long-term monitoring and dense sampling that allows the creation of genetic pedigrees(Slate et al., 2010). This means that mapping heritable traits can be complicated and costly, and resulting QTL studies can be of variable quality(Slate, 2004). To simulate wild populations more accurately, cheaply, and controllably in mice, the Diversity Outbred mouse project was created.

1.6.1 Diversity Outbred Mouse Project

The Diversity Outbred mouse project arose as an offshoot of the Collaborative Cross, beginning as a random cross between CC lines, and continuing as a random outcross. This creates unique individuals in each litter(Svenson et al., 2012). They contain 45 million segregating SNPs, creating a more genetically diverse population than humans.

Genotyping is undertaken using the third generation of the Mouse Universal Genotyping Array (GigaMUGA)(Morgan et al., 2016). GigaMUGA contains 141,090 SNP probes across the genome, intended to be used for the identification of laboratory mice down to the substrain level. Because the DO population ultimately traces its roots back to eight well-characterized inbred strains, GigaMUGA genotyping data can be used to create probability matrices of local haplotype for each DO mouse. With a pseudoprobe grid of 69,005 predictions, GigaMUGA can resolve haplotype in highly recombinant areas down to under 100 bp in some regions(Skelly et al., 2020).

DO mice are not replicable like CC and CC-RIX, but they offer a greater diversity of phenotypes than either(Svenson et al., 2012), with similar reliability to inbred mouse lines(Tuttle et al., 2018). This includes the study of more developmental phenotypes(Katz et al., 2020). DO mice have also allowed for a method of mapping increasingly precise quantitative trait associations due to increasingly small linkage disequilibrium blocks and balanced allele frequencies (Solberg Woods, 2014).

The ongoing process of random outbreeding produces a mouse population segregating millions of precisely-mapped SNPs, with well-balanced allele frequencies, and increasingly small blocks of linkage disequilibrium that now approach the size of TADs

almost genome-wide(Broman, 2012; Chesler et al., 2016). Genetic mapping depends on the ability to distinguish between traits that segregate by genotype. Genotypes segregate by recombination. Thus, the more recombination events that separate a case from its controls, and the more controls are available, the more precisely one can map a trait.

Samples from the DO have been used to create extensive resources and data sets, including mouse embryonic stem cell (mESC) lines. A recent study used mESCs derived from 19th generation DO embryos by Predictive Biology, grown in the absence of ERK inhibition to study the gene regulatory networks that stabilize pluripotent states *in vitro*. This revealed phenotypic variability in features of ground state pluripotency(Skelly et al., 2020). 2i media allow for the stable preservation of mESC culture, but necessarily limit the colonies' abilities to differentiate and display developmental phenotypes. 1i mediums allow researchers to explore early differentiation phenotypes and expression patterns(Acurzio et al., 2021; Atlasi et al., 2013), while maintaining healthy culture conditions. As models of early development, they are well-suited to profiling the effects of genetic-epigenetic interactions, and the role of 3D chromatin conformation as a regulatory mechanism (Adapted with permission from Kuffler et al., 2023, expanded to provide context).

1.7 Project Goals

Here we determine the extent to which genetic-epigenetic interactions between single nucleotide polymorphisms (SNPs) and regions of open chromatin are present, and uncover the biological basis of their distribution in three-dimensional (3D) genomic organization. Our results show that genetic-epigenetic interactions were found across

the genome and involve regulatory elements that would not be identified with single-omics data. These interacting elements cluster together within TADs bounded by previously identified active CTCF binding sites. We infer chromatin structure from interaction data, analyze interaction contributions from the main and interaction regression elements, identify potential regulatory functions underlying a set of interactions, and correlate these interaction behaviors to CTCF binding differences in inbred founder lines (Reproduced with permission from Kuffler et al., 2023).

Chapter 2: Methods and Materials

2.1 Dataset Selection and DO mESC Generation

DO mESC production was performed by Predictive Biology. Cultures were grown with G3SK inhibitor (1i medium). Bulk RNA and ATAC sequencing were performed and normalized as previously described(Skelly et al., 2020).

Expression data was formatted and analyzed as log2-transformed transcripts per million (TPM). ATAC-seq data was formatted as trimmed mean of *M* values (TMM)(M. D. Robinson & Oshlack, 2010). Genotyping was performed by Giga Mouse Universal Genotyping Array (GigaMUGA)(Morgan et al., 2016). Aneuploidies were removed with the argyle R package(Morgan, 2016). R/QTL2 haplotype reconstruction, normalization and pseudoprobe processing was carried out as previously described(Skelly et al., 2020). Samples with XO genotypes were removed, and the union of all samples with the required data types resulted in 176 samples (Reproduced with permission from Kuffler et al., 2023).

These data were selected for this project due to their unique suitability for the study of non-additive genetic-epigenetic interactions. In fact, the generation of these data made the project conceivably possible in the first place—no sufficiently large publicly available datasets combined ATAC-seq and RNA-seq in any genetically diverse model system at the time of the project's conception. We were aware of this at the time, and considered the potential drawbacks.

We would not be able to confirm our findings by recapitulating them in another model system, unless another group independently produced a suitable equivalent while we

worked on the project. However, we felt that we would be able to perform targeted confirmation experiments within mESCs to support our findings. In fact, we were later able to perform a systemic confirmation experiment in DO founder mESCs, thanks to collaboration with our co-authors (see below, and in Kuffler et al. 2023).

Another potential complication was the culture protocol for the mESCs. The standard 2i medium for mESC culture combines inhibition of MEK and GSK3(Silva et al., 2008). This allows cells harvested from the inner cell mass of a mouse zygote to be maintained in a ground state of pluripotency in most strains, demethylated and with extensive open chromatin. The 1i medium is non-standard, meaning that the results would not necessarily be comparable to publicly available mESC data. However, we considered this to be a strength of the dataset, as it would expose expression phenotypes that we would otherwise not see in mESC culture. Given that our intent at the beginning of this project was to assay the locations, commonness, and properties of genetic-epigenetic interactions, this was useful for us. We would have more variation in gene expression, and presumably more variation in chromatin openness. This was confirmed by Skelly et al in their identification of significant expression QTLs and chromatin accessibility QTLs(Skelly et al., 2020). This essentially enhanced our chances of identifying regulatory interactions. Concomitantly active genes are actively worked upon by multiple regulatory factors, but these interactions cannot be identified without precisely targeted assays to identify the mechanisms at play. We wanted to cast a wider net, and the 1i medium was very useful for this.

It was also worth considering the genomic stability of the Diversity Outbred mice. The DO population has begun experiencing genetic selection as it has progressed. While the

generation of this data predates much of these issues, R2d2 had begun demonstrating meiotic drive on chromosome 2, which resulted in a favoring of the WSB allele at that locus(Chesler et al., 2016). To avoid factoring this into our initial exploratory data, we made sure to examine chromosome 2 separately and in comparison with other chromosomes.

Stem cell culture is known to produce a number of effects on genome stability, including aneuploidies that can confer growth advantages, leading to aneuploid cells outcompeting the original line they were derived from(Enver et al., 2005). Detection of these events is not a simple process, and relevant data on these DO mESCs was not publicly available at time of processing. However, we received in-progress aneuploidy predictions from Alex Stanton (results currently unpublished), and avoided using examples from chromosomes that were prone to duplication when assessing the location and prevalence of interactions.

With these limitations in mind, we decided that the potential of this dataset far outweighed its limitations. Genetically diverse samples with genetic profiling and multiple epigenetic data types are very rare, and the well-characterized nature of the DO founder strains allowed us a high degree of mapping precision.

2.2 Subsetting Genetic Markers

Haplotype predictions were represented using the closest SNP to grid points of GigaMUGA haplotype probabilities, regularly spaced by genetic distance(Chick et al., 2016; Kingra et al., 2020). The gigaMUGA array provides 143,259 SNP probes via the Illumina Infinium II platform(Morgan et al., 2016). The haplotype probability calculation

on this basis created a subset of 68,413 SNPs which were most likely to match the actual haplotype for each sample at a given locus (See data repository, linked in section 2.6 below) (Adapted with permission from Kuffler et al., 2023, edited for clarity).

It is worth noting that these genotypes cannot truly resolve to SNP-level accuracy. While the 19 generations of DO outcrossing that preceded the generation of these cultures provided a high level of recombination, blocks of linkage disequilibrium created large areas of uncertainty in many LD blocks. This is not always the case, with some blocks being on the order of a few kilobases in length, and containing low numbers of SNPs. Even in these cases, the SNPs we chose to represent the LD blocks were merely representative of their local genotype, at the locus we can most confidently predict to match the pseudoprobe's haplotype. The vast majority of genetic variants between the DO founder strains are SNPs, and all representative genetic variants fall within this category. Therefore in the interest of brevity, this manuscript often refers to "SNP effects" when the genetic component of our interaction modelling is discussed.

However, there are other variants within the genomes of the DO that need to be acknowledged. The aforementioned R2d2 locus and mESC aneuploidies aside, there are also polymorphic regions that show structural variation in copy number, 2006 of which are captured in the gigaMUGA panel. While these are accounted for in the gigaMUGA panel, they were not considered in these experiments, and can lead to background-specific increases or decreases in signal strength during genetic and/or epigenetic profiling(Morgan et al., 2016). These have been recently cataloged by Ferraj et al., and may be worth integrating into future analyses(Ferraj et al., 2023).

2.3 Regression Modelling

RNA-seq, ATAC-seq and haplotype data were repackaged into SQLite databases and fit to a regression model using the R stats::step() function. Essentially, this function identifies a line of best fit for a dataset, with the user providing the terms that should be considered as potential independent variables, and how they interact. The function then assigns constants ("beta coefficients" or "effect coefficients") to these variables, removing terms whose beta coefficients approach zero or are otherwise not needed to describe the data.

We modeled the gene expression by linear regression, including a SNP-by-ATAC non-additive interaction term:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon_i$$

(Equation 1)

where y_i =RNA abundance (log_2TPM), x_1 = SNP, and x_2 = ATAC intensity (TMM) for each gene. SNPs were coded as 0 for reference genotype, 1 for heterozygote, and 2 for homozygote. No multi-allelic variants were present in the subset. Models passed the default Akaike Information Criterion cutoff to retain the $\beta_3 x_1 x_2$ term were deemed to be interacting. This means that the R stats::step() function determines whether dropping the term would lose too much ability to describe the underlying data, then the term would be kept. If it did not meet this threshold, it would be discarded to prevent overfitting(Stoica & Selen, 2004). The R stats::step() function aims to select beta coefficients that allow the Akaike Information Criterion to equal Mallow's C_p, which is a function that assesses the fit of a regression model(Hocking, 1976). The resultant database was the input for all

further analyses (Adapted with permission from Kuffler et al., 2023, wording altered for readability).

This is a relatively simple regression model, but it contains more terms than some researchers may be familiar with. Simple linear regression often uses only one variable term, taking the form $y_i = \beta_0 + \beta_1 x_1 + \varepsilon_i$. The goal of these analyses is to find the values of β_0 and β_1 that best fit the observed data, by attempting to minimize the sum of squared residuals. This means that the differences between the predicted and actual values of y should be minimized. This method is common in genetics, and used for fitting linear trends. When multiple terms are modelled, it is often done so in an additive manner: more terms are added, assuming they are not only independent from the dependent variable (in this case RNA abundance), but also independent from each other. This would look something like $y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon_i$. Any number of terms can be added to these models. This is appropriate if features are expected to only affect the gene, and not each other. All terms function independently, assuming that the dependent variable's value will have a direct relation to the magnitude of each term.

However, we wish to examine if the relationship of these independent variables change depending on each other's values. Thus, we added the interaction term $\beta_3 x_1 x_2$. This allows us to model SNP and ATAC-seq effects that produce greater or lesser results than expected, when both are present in a single sample. This in itself is not a novel technique, nor is it computationally complex. In fact, it was the simplest available method of modeling the potential interaction between SNPs and ATAC-seq peaks. This simplicity in itself was an asset in this case: we were modeling interactions in a statistical manner,

rather than performing targeted experiments to assay physical interaction between these regions. We were doing this for the first time on a genome-wide scale. This meant that we had to be careful not to overstate the power of our models, and we could not risk overfitting that could undercut the biological effects we are attempting to demonstrate. Therefore, retaining a linear model with a single extra term was decided to be the correct choice.

This regression model was tested with two null models: one where neither potential regulators affected a gene's expression ($\beta_1 = \beta_2 = \beta_3 = 0$ in Equation 1), and one where both independently affect gene expression ($\beta_3 = 0$ in Equation 1). The former null model was non-standard, more stringent, and intended to determine if there were any true effects present in the dataset. This produced results consistent with the latter model, and was ultimately discarded. Full model significance (Equation 1) was estimated relative to the null model with an F-ratio test. F statistics were calculated with the R function `pf()`. Given the possibility of overfitting and the number of models tested (see Results), a Bonferroni adjusted p value cutoff of $p < 1x10^{-7}$ was chosen (hereafter "significant models"). Bonferroni adjustments were calculated via `stats::p.adjust()` (Adapted with permission from Kuffler et al., 2023, edited for clarity).

2.3.1 Significance Cutoff Selection

One of the complications of performing a novel analysis is the lack of literature to guide the selection of a p value cutoff. Too forgiving, and false positives will drown out any true signal, too stringent, and your data becomes sparse and the cutoff may appear as if it is chosen simply to impress the audience, rather than responding to the features of your dataset. We chose our adjusted p value cutoff based on the observed density of our

adjusted p values (Fig. 3.1), which we believed confidently indicated an enrichment of p values under $1x10^{-7}$. We erred on the side of restraint in this initial analysis with this figure. We would recommend that as more analyses of this nature are produced, this cutoff could be relaxed.

2.4 TAD Data Selection

mESC TAD data was downloaded from previously published work(J. R. Dixon et al., 2012) with a liftover from mm9 to mm10 via the UCSC genome browser(Kuhn et al., 2013). Chromosome information for mm10 was retrieved from the Integrative Genomics Viewer(J. T. Robinson et al., 2011) (Adapted with permission from Kuffler et al., 2023 with edits to readability).

This TAD data set was initially selected as part of a screen of potential biological bases for the local regulatory area (see Results). There are inherent caveats in using publicly available data when comparing to the DO population. One will never be able to create a perfect match to the DO samples themselves, as they are unique individuals. The DO does provide a useful basis to start from, however: the DO founders, all of which are well-documented and commercially available inbred strains(Svenson et al., 2012).

However, public data does not have the breadth necessary to capture all the DO founders, especially in mESCs. Up until recently, mESCs were limited to generation in only B6 and 129 strains, which are only two of the eight founders of the DO, and both from the same subspecies background(Evans & Kaufman, 1981; Ledermann & Bürki, 1991; Martin, 1981; Schnabel et al., 2012). In the case of TAD boundary predictions, we were only able to find data for B6. This inherently limited the accuracy of our analysis, as

we could not identify which TAD boundaries might vary based on genetic background. Given the previously published caQTL data from these cells, we knew that there were significant and genetically-driven differences in chromatin accessibility with mediating effects on gene expression, something that could be explained by differing TAD boundaries(Skelly et al., 2020). Without chromatin conformation data, we could not determine if this was the case.

We also could not find exact matches for our culture conditions. The 1i medium used benefitted us in our search for genetic-epigenetic interactions, but it necessarily means allowing the cells to begin priming for differentiation and demonstrating their cell fate tendencies. Many TADs are conserved across cell types(McArthur & Capra, 2021), but we have no way to determine which are conserved, which are not, and limited ability to identify which DO mESC cultures may be heading towards which cell fate. Regardless, granular data on this level was not available, but it limited the accuracy we could expect from overlaying data from B6 mESCs grown in 2i.

On top of this, the TAD boundary imputation methods used in this publicly available dataset had a limited resolution set to a 40kb grid. This was limited by the underlying Hi-C data, but of course presented a problem for our calculations. CTCF binding sites that underpin TADs are ubiquitous across the genome(Ziebarth et al., 2012), and many can fall in a region of 40kb. Lacking any other appropriate data source, we took this data and performed all calculations based on this 40kb grid. At the time, this was the best resolution we could hope to achieve in this analysis.

2.5 DNA Motif Analysis

DNA motif and binding site discovery were performed using MEME Suite tools (MEME, STREME, and Tomtom) set to default parameters (See 2.6 for data repository)(Bailey, 2020; Bailey et al., 2009; Bailey & Elkan, 1994). SNP data were imputed from DO founder genomes as reported in release v3 (REL-1303-SNPs_Indels-GRCm238) of the Mouse Genome Project through the Sanger Institute(Grant et al., 2011; Keane et al., 2011) (Adapted with permission from Kuffler et al., 2023 with edits to word choice)

MEME Suite was found to be a generally reliable motif and binding site toolset, with a few caveats. Its web interface does not have an API. The servers are relatively reliable, but there are no alternate servers, and no ways to access older versions of MEME Suite without running them locally. This was not possible for us, as it could not be successfully integrated with the local high-performance computing cluster. This limits our ability to ensure the motif searches are truly reproducible, but it is theoretically possible for others to do so.

2.5.1 Publicly Available CTCF ChIP-seq Analysis

To investigate CTCF binding in mESCs as a proxy for TAD boundary activity, we downloaded the CTCF ChIP-seq in C57BL/6 Bruce4 and 129/Ola E14TG2a.4 mESCs from the ENCODE portal(Sloan et al., 2016) with the following identifiers: ENCSR000CCB and ENCSR362VNF. These were also used to estimate the locations of SMAD3, an optional component of the CTCF complex. CTCF and SMAD3 binding motifs were retrieved from HOCOMOCO(Kulakovskiy et al., 2018). with the identifiers CTCF_MOUSE.H11MO.0.A and SMAD3_MOUSE.H11MO.0.B. These were overlapped

with our ATAC-seq dataset via FIMO(Grant et al., 2011) (Adapted with permission from Kuffler et al., 2023 with edits to readability).

As before, these datasets are not ideal models for our DO mESC samples. B6 Bruce4 is an mESC line that exhibits significant heterozygosity from B6, and is prone to aneuploidy(Gacita et al., 2021). E14TG2a.4 is a widely used cell line, which seemed to have no particular complaints associated with it upon investigation(Hooper et al., 1987; Stryke, 2003). However, they were grown in feeder free 2i medium, which is again not comparable to our 1i samples.

2.6 Generating Novel CTCF ChIP-seq From DO Founders

We undertook an analysis of CTCF binding site activity by use of chromatin immunoprecipitation sequencing (ChIP-seq). Potential options for CTCF ChIP-seq alternatives were considered. CUT&RUN ChIP was considered, a variation on standard ChIP-seq(Skene & Henikoff, 2017). The opinion of our collaborators was that while CUT&RUN required fewer cells, it was less accurate compared to standard ChIP-seq. Classic ChIP-seq requires 5-10 million cells per replicate, which was considered reasonable. The cells were grown in collaboration with Laura Reinholdt's lab, and ChIP-seq protocols were performed by Haley Fortin in Christopher Baker's lab.

CTCF ChIP-seq was undertaken under previously established protocols(Oomen et al., 2019), with antibodies tested by Haley Fortin. Three cell lines were run in four isogenic DO founder mESC strains: C57BL/6J, CAST/EiJ, PWK/Ph, and WSB/EiJ. Cultures were grown in 1i medium. One sequencing run failed, resulting in the loss of one PWK and one WSB sample. This was disappointing, but we proceeded ahead. We had already

decided to employ an analysis method that did not require three replicates: we would count ChIP-seq peaks as evidenced within a strain only if they appeared in two or more samples. The loss of two samples lowered our detection chances, but did not affect our detection reliability (Adapted with permission from Kuffler et al., 2023, expanded to provide details).

FastQC, Bowtie alignment and g2gtools liftover from imputed strain genomes to BL/6J were performed using a modification of a now publicly available ATAC-seq pipeline, accessed in development 15 December 2020(Sanderson et al., n.d.). Mitochondrial read filtering steps were removed. The g2gtools liftover tool is a production of the Chuchill lab, and is quite unique in its ability to perform liftovers in either direction between a reference genome and a non-reference genome. We used imputed genomes for the non-reference strains, to permit direct comparison to B6. Results were filtered to those ChIP-seq binding locations that appeared in two or more isogenic strains. The mitochondrial read filtering steps were removed due to their lack of relevance for ChIP-seq. CTCF does not form TADs on mitochondrial DNA, therefore it would not appear in the data and would not need to be removed from it (Adapted with permission from Kuffler et al., 2023, expanded to provide context).

All study data and scripts are available at Synapse (https://doi.org/10.7303/syn26534443).

2.7 Attribution

Analyses in Chapter 2.2-2.5.1 were performed solely by the author. 2.1 was the work of Skelly et al. as previously described. DO founder mESC generation in 2.6 was

performed by the lab of Laura Reinholdt. CTCF ChIP-seq generation and initial QC analysis was performed by the lab of Christopher Baker, with antibody testing and CTCF ChIP-seq preparation performed by Haley Atwell. Code for processing CTCF ChIP-seq data was adapted from Sanderson et al., with advice from Michael Lloyd.

Chapter 3: Results

3.1 Genetic-Epigenetic Interactions are Pervasive and Enriched Within TAD Boundaries

3.1.1 Introduction

Despite numerous reports showing that genetic variants and chromatin state alter gene expression, non-additive interactions between these regulators have not been systematically explored. We wanted to determine how commonly genetic-epigenetic regulatory interactions are observed on a systemic level, and whether these interactions show any bias towards particular locations within the genome. To examine these interactions, we used paired transcriptome and chromatin accessibility profiling of Diversity Outbred mESCs(Skelly et al., 2020). These DO mESCs incorporate genetic backgrounds from eight well-understood isogenic founder strains from three *M. musculus* subspecies, taken from generation 19 of the DO project. Thus, these data constitute unique resources that can address questions of genetic-epigenetic regulatory interaction (Adapted with permission from Kuffler et al., 2023 with edits to word choice)

3.1.2 Regression Modelling

To impute non-additive interaction effects on a global scale, we used a method of analysis that could be applied to co-occurrence and interaction effects, and could permit discrimination between them. To that end, we used regression model fitting (Equation 1) alongside model selection using information criteria, to test if gene expression was affected by a local haplotype, an open chromatin region, both as independent regulators, or both in interaction (Adapted with permission from Kuffler et al., 2023 with edits to word choice).

Linkage disequilibrium places resolution limits when discerning the activity of genetic variants. Thus, we created a subset of our genetic data limited to 68,413 high confidence SNPs, spaced evenly by genetic distance to reflect the resolution limit (Methods). We collated these haplotypes with 102,173 ATAC-seq peaks and 13,631 genes that had RNA-seq coverage within our sample set (Adapted with permission from Kuffler et al., 2023 with edits to word choice).

We first did this with a null model of RNA ~ 1, i.e. no relation between expression and SNP or ATAC. This is the default for most regression analyses, but it was potentially not appropriate for our work, as we have an interaction term. We redid it with a null model of RNA ~ SNP + ATAC, i.e. no interaction detected. These were found to have minimal effect on results. We ran both for every analysis up to Figure 3.7. For consistency with the intent of our paper, we used the second null model.

3.1.3 Random Model Generation And Analysis

To search for non-additive interaction models, we randomly selected intrachromosomal genes, SNPs, and ATAC-seq peaks to test with the interaction model. In total, 102,104 ATAC-seq peaks, 68,413 SNPs, and 13,631 genes were modeled in 39,021,625 combinations. The result was 27,509 significant models (Bonferroni adjusted $p < 1x10^{-7}$), or 0.07% of all models, associated with 19,145 ATAC-seq peaks, 14,476 SNPs, and 1,286 genes (18.75%, 21.16%, and 8.52% of the total input respectively). Full results are tallied in Table 3.1. The number of significant results was four orders of magnitude over expectations, thus indicating true associations were present (Fig. 3.1A) (Adapted with permission from Kuffler et al., 2023, correcting errata and with edits to word choice).

Randomly generated sample database	Only SNP effect	Only ATAC effect	Additive effect	Interactive effect	No effect	All Models
SNPs (p < 1x10^-7)	9261	N/A	3627	3836	N/A	68413
ATAC-seq peaks (p < 1x10^-7)	N/A	670	3486	3735	N/A	102104
Genes (p < 1x10^-7)	927	221	823	896	0	13631
Models	8059819	6952343	228616	560484	21166363	39021625
Models where p < 1x10^-7	14683	2465	4921	5110	0	27509
% Models where p < 1x10^-7	0.18%	0.04%	2.15%	0.91%	0.00%	0.07%

Table 3.1. Counts and percentages within a database of randomly generated regression models. Adapted with permission from *Kuffler L, Fortin HJ, Skelly DA, Czechanski A, Munger SC., Reinholdt LG, Baker CL, Carter GW. (2023) Imputation of 3D genome structure by genetic-epigenetic interaction modeling in mice eLife 12:RP88222 https://doi.org/10.7554/eLife.88222.1. Changes include correction of errata.*

To test the structure of our significant results further, we looked at different ways to quantify the distance between interacting elements. We took the sum of the distance between the gene and both the ATAC-seq peak and the representative SNP, versus the distance between the gene and whichever interacting element was closest to it. We quantified this both by bp distance, and by the number of TADs between the interacting elements. We found that looking at the sum of the distances produced a spread across the chromosome, while the minimum distance to an interacting element was tightly focused around the gene itself. The locations of significant associations indicated interactions were much more likely to occur within the linkage disequilibrium block that contains the gene (Fig. 3.1B), matching prior work that the majority of detectable genetic effects originate in local regulatory regions(Nelson et al., 2004; Oomen et al., 2019; Ronald et al., 2005; Su et al., 2010). This was echoed in the TAD-based distance

measurements. Approximately half (49%) of all significant models contained a SNP or ATAC peak within 4 Mb of the gene they affected.

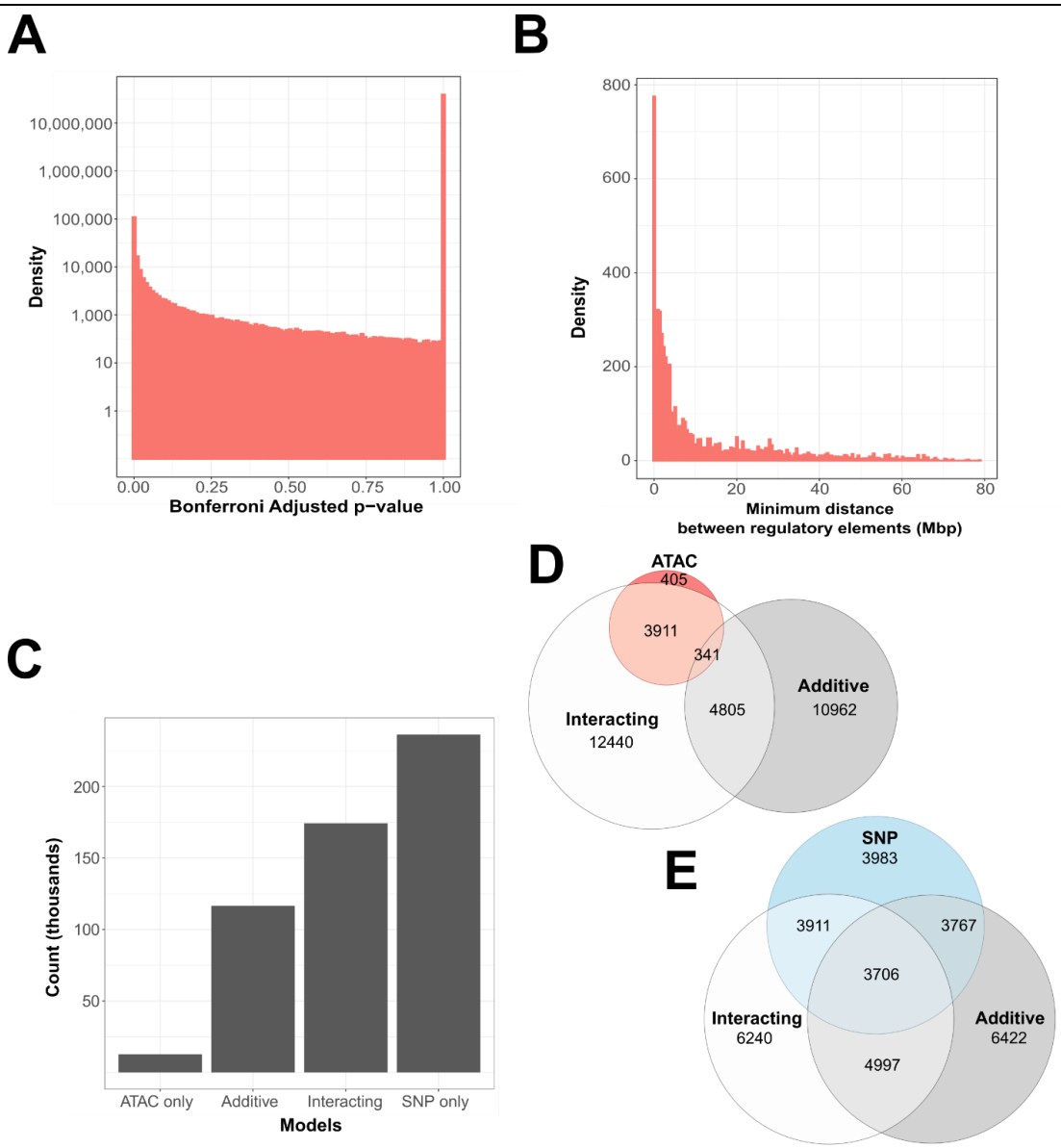


Figure 3.1. Interacting and additive models are abundant and favor local genes. A. Density of Bonferroni Adjusted p-value distribution in randomly sampled intrachromosomal models. **B.** Density of non-additive interacting models where adj. $p < 1x10^{-7}$, by minimum distance between regulatory elements. **C.** Distribution of model term retention across intra-TAD models. **D & E.** Intra-TAD ATAC-seq peak and SNP participation in single-term, additive, and non-additive interacting models. Reprinted with permission from *Kuffler L, Fortin HJ, Skelly DA, Czechanski A, Munger SC., Reinholdt LG, Baker CL, Carter GW. (2023) Imputation of 3D genome structure by genetic-epigenetic interaction modeling in mice eLife 12:RP88222 https://doi.org/10.7554/eLife.88222.1.*

We hypothesize that the reason why the sum of distances is so much greater is due to the fact that these randomly chosen trios of elements may contain co-regulated features. For example, an ATAC-seq peak may be present due to the binding of a particular protein, which also binds to other sides across the genome. Distant genotype can appear significant due to a distal QTL, or it could be due to strain segregation of SNPs that is identical to a more local feature.

3.2 TADs Constitute the Boundaries of Local Regulatory Areas

Due to these results, we investigated what chromatin features might underlie the local area we saw for genetic-epigenetic interactions. We hypothesized that TADs might be involved in the presence of non-additive interactions, as they are 3D genome structures previously seen to enable gene-enhancer interaction. Thus, we created a more focused subset of regression models, which tested every potential combination that each gene could have with every SNP and ATAC peak that fell within one TAD of its transcription start site (Adapted with permission from Kuffler et al., 2023 with edits for clarity).

Models possessing at least one interacting element in the same TAD as the gene were found to be 4.5 times more likely to meet our significance cutoff, when compared to models that had neither interacting element within the TAD (3.9% versus 0.9%). This subset was drawn from our randomly generated data, featuring 3,836 SNPs and 3,725 ATAC-seq peaks affecting the expression of 896 genes (Table 3.1). ATAC-seq peaks that were involved in non-additive interactions favored genes that fell between CTCF binding sites previously identified as TAD forming in B6 mESCs (Fig. 3.2A). Regression models without interaction terms also show this local, TAD-bounded ATAC-seq peak

involvement (Fig. 3.2B). We can therefore state that interacting regulatory elements inferred by our regression model are localized to the same TAD as the gene they affect. This produces a potential structural reason for local regulatory areas previously observed (Adapted with permission from Kuffler et al., 2023 with edits to word choice).

3.2.1 Interacting Elements in Genetically Diverse Samples Escape Conventional Discovery Methods

Our next analysis evaluated whether genetic-epigenetic, non-additive interactions could provide information that standard analysis methods cannot. Thus, we subset our imputed interacting models, constituting 32.29% of significant intra-TAD models (Fig. 3.1C), and compared them with models associated with models that featured SNPs and/or ATAC-seq peaks that were not interacting. We examined the overlap between intra-TAD regression model results by SNP and ATAC-seq peaks by themselves (Adapted with permission from Kuffler et al., 2023 with edits to word choice).

These were compared to randomly selected models (Fig. 3.3). In this set, models with only SNP effects predominate. We hypothesize that this is because genetic effects are often the most easily identified, especially over long distances. Randomly selected trios will primarily feature models with extremely distant results, making this essentially akin to a search for eQTLs.

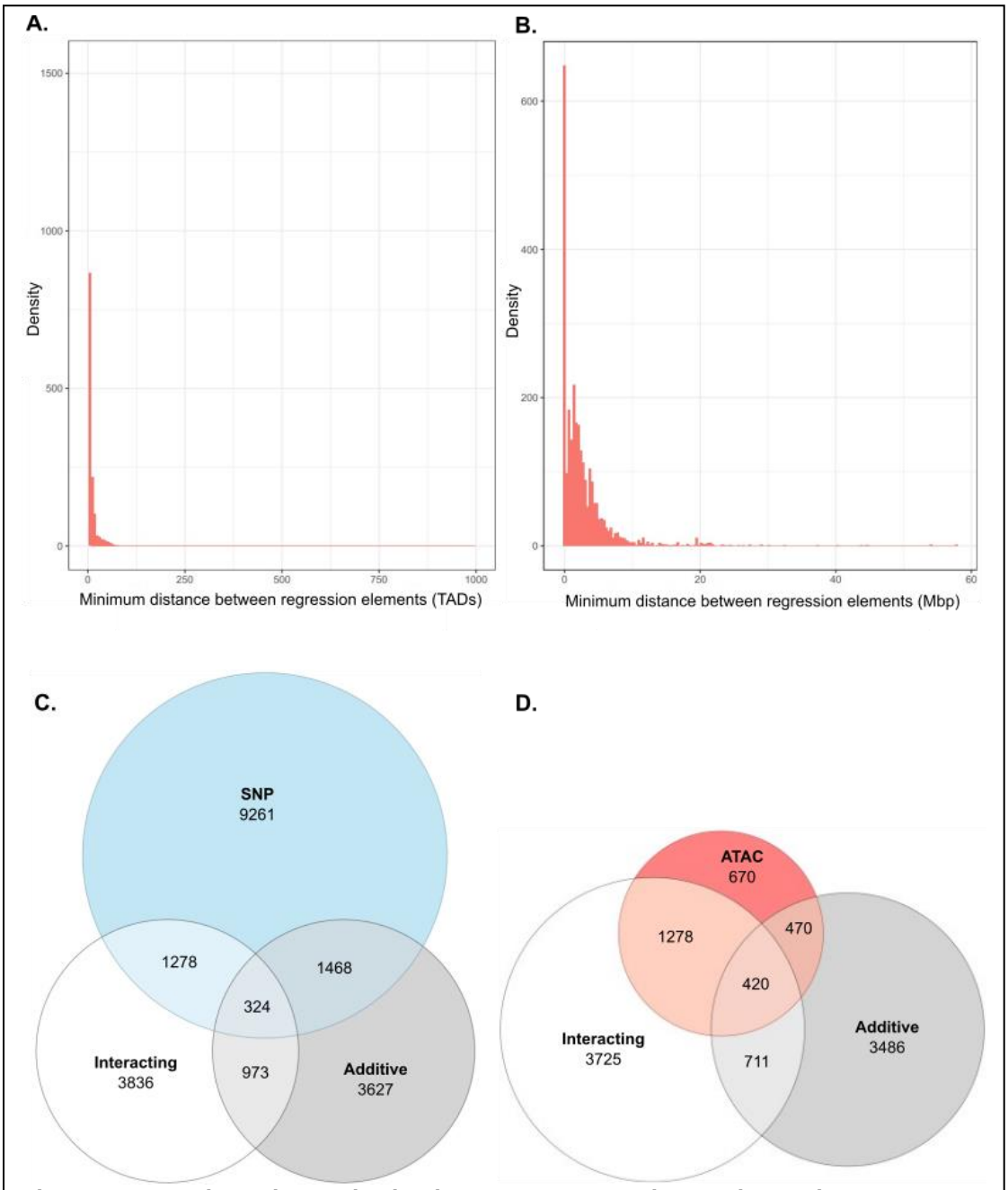


Figure 3.2. TAD-bound organization is common across interacting and non-interacting models and include many exclusively interactive regulators A. Density of interacting models where adj. $p < 1x10^{-7}$, by minimum distance between regulatory elements, measured in relation to number of TADs between elements. **B.** Density of additive models where adj. $p < 1x10^{-7}$, by minimum bp distance between regulatory elements. **C-D.** ATAC-seq peak and SNP participation in single-term, additive, and interacting models.

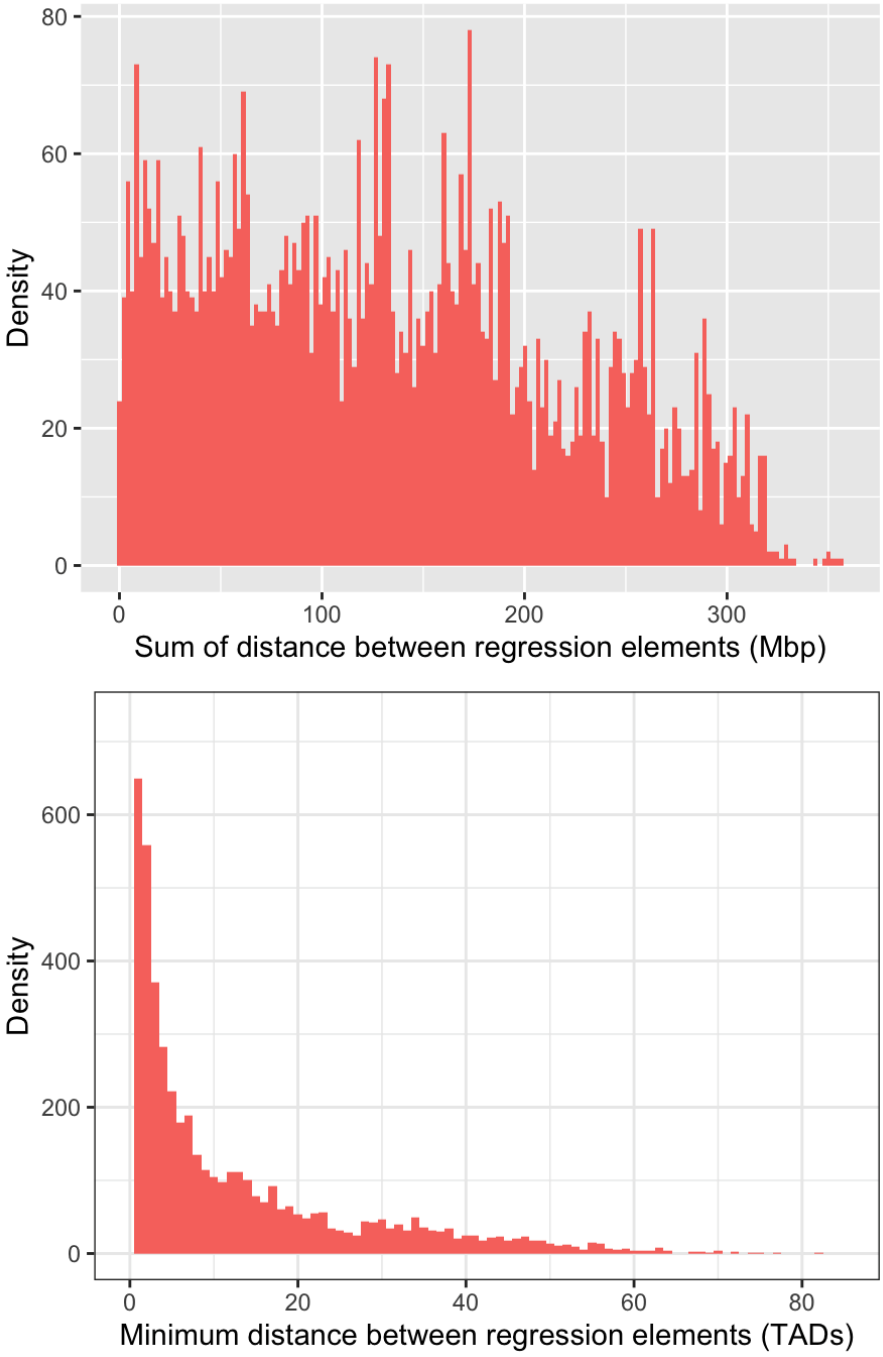


Figure 3.3. Interactions favor local regulators. A. Sum of distance between every significant gene-SNP-ATAC trio (in Mbp), showing long-range associations. **B.** Minimum distance between the same, measured in TADs, showing preferential short-range interaction.

It is worth noting that while this is a higher raw number of model counts, the method of calling models we used will drop terms if it can find a satisfactory fit with fewer terms, at which point it applies a p value to the result. The distribution of p values per model type

is thus also instructive, particularly examining what percentage of models reach our p value cutoff for significance. When viewed through this lens, interactive effects show the highest percentage of confidently called models, followed by additive, SNP-only, and ATAC-only (Table 3.2). Many models ended up classed as having no detectable effect, but p values were not assigned to these.

TAD-focused sample database	Only SNP effect	Only ATAC effect	Additive effect	Interactive effect	No effect	All Models
SNPs (p < 1x10^-7)	7208	N/A	12268	11933	N/A	66214
ATAC-seq peaks (p < 1x10^-7)	N/A	480	22370	25916	N/A	100805
Genes (p < 1x10^-7)	748	171	1120	1094	N/A	13631
Models	47724300	20117970	16991360	14262787	52294007	151390424
Models where p < 1x10^-7	752027	32720	303285	460804	0	1548836
% Models where p < 1x10^-7	1.60%	0.16%	1.78%	3.23%	0.00%	1.02%

Table 3.2. Counts and percentages within a database of all possible regression models where all SNPs and ATAC peaks are within +/- one TAD of the gene they interact with. Adapted with permission from *Kuffler L, Fortin HJ, Skelly DA, Czechanski A, Munger SC., Reinholdt LG, Baker CL, Carter GW. (2023) Imputation of 3D genome structure by genetic-epigenetic interaction modeling in mice eLife 12:RP88222 https://doi.org/10.7554/eLife.88222.1. Changes include correction of errata.*

We found 27.10% of SNPs were found in both non-additive interacting models and additive models, and 22.59% were shared with models containing no contribution from an ATAC-seq peak. In comparison, few models predicted *only* an ATAC-seq peak contribution, and 37.45% of ATAC-seq peaks are exclusive to interacting models (Fig. 3.1D). This is also in contrast to results from our randomly sampled gene-ATAC-SNP trios, which show a preference for SNP-only models (Fig. 3.2C-D). These results

suggest that although genetic variants are the primary driver of variation in expression, non-additive interactions with chromatin states further reveal the origins of expression variation. Our findings also suggest that genetic variation and open chromatin data cannot be used independently to capture these regulatory features. Conversely, these findings suggest that ATAC-seq data alone is a less effective predictor of gene expression in a genetically diverse population (Reproduced with permission from Kuffler et al., 2023).

3.3 TAD Boundaries Limit Genetic-Epigenetic Interactions

The specific preference for non-additive interactions within active CTCF binding sites warranted further study. We sought to determine if active CTCF binding sites provided an appreciable boundary to interactions (Fig. 3.4A), similar to reported segregation of enhancer elements(Rodríguez-Carballo et al., 2017). We carried out regression analyses across the genome for all possible models involving each gene-SNP-ATAC combination within the gene's TAD and nearest flanking TADs or intra-TAD regions. Although resolution of causal SNP location limited by linkage disequilibrium, ATAC-seq peaks that interact with SNPs could be confidently localized in relation to TAD boundaries (Reproduced with permission from Kuffler et al., 2023).

We found that TADs contained more ATAC peaks that interact with local SNPs to affect expression of a resident gene (Fig. 3.4B). With a window of 200 kb on either side of an active CTCF boundary, we found 52,321 ATAC-seq peaks sharing a TAD with their interacting gene, versus 24,145 peaks in different TADs or inter-TAD regions This constitutes a significant enrichment of intra-TAD gene-peak interactions (an odds ratio of 1.27, null expectation $p < 2.2x10^{-16}$) (Fig. 3.5). We also identified 29 individual ATAC-

seq peaks that were involved in more than 200 non-additive interactions each, which are distinctly visible within the aggregate view (Fig. 3.4B). These were exclusively affecting genes within the same TAD as the interacting ATAC-seq peak, which suggests that our results are not due to higher density of open chromatin within TAD structures. This is consistent with previous findings that the effects of enhancers and other regulatory elements are constrained by TADs(Krijger & De Laat, 2016). (Reproduced with permission from Kuffler et al., 2023).

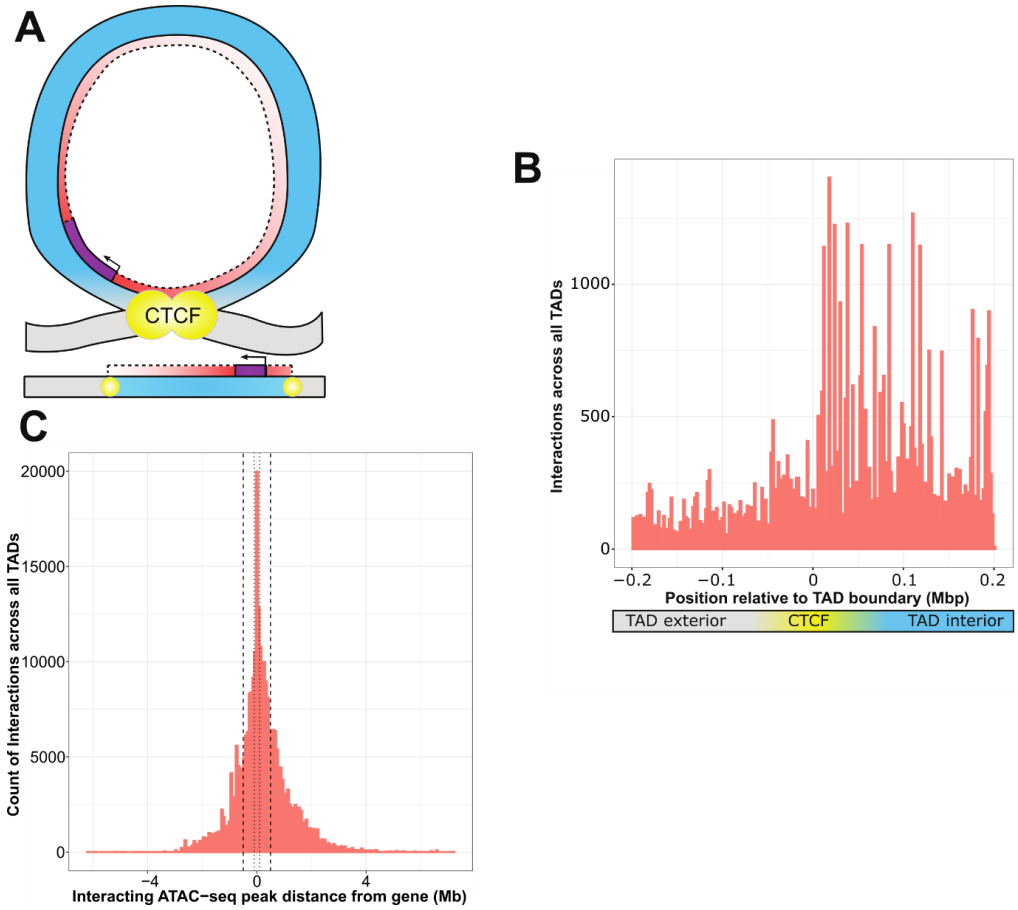


Figure 3.4. ATAC-seq peaks that interact with SNPs generally reside within the affected gene's TAD. A. Schematic of a TAD loop, including gene (purple) and density of interacting model elements (red). Loop interior is in blue, exterior DNA is gray, and CTCF binding sites are in yellow. **B.** Location of interacting ATAC-seq peaks relative to TAD boundary location, merged across all genes. TAD interior denotes the TAD in which the dependent gene was found. **C.** Interacting ATAC-seq peaks by distance from associated gene TSS. Local area cutoffs of 100 kb and 500 kb flanking regions are marked. Reprinted with permission from *Kuffler L, Fortin HJ, Skelly DA, Czechanski A, Munger SC., Reinholdt LG, Baker CL, Carter GW. (2023) Imputation of 3D genome structure by genetic-epigenetic interaction modeling in mice eLife 12:RP88222 https://doi.org/10.7554/eLife.88222.1.*

TADs contain greater amounts of open chromatin than intra-TAD DNA (Fig. 3.5). Thus, we needed to ensure that these results were not simply a recapitulation of the underlying density of open chromatin, and constituted a true enrichment.

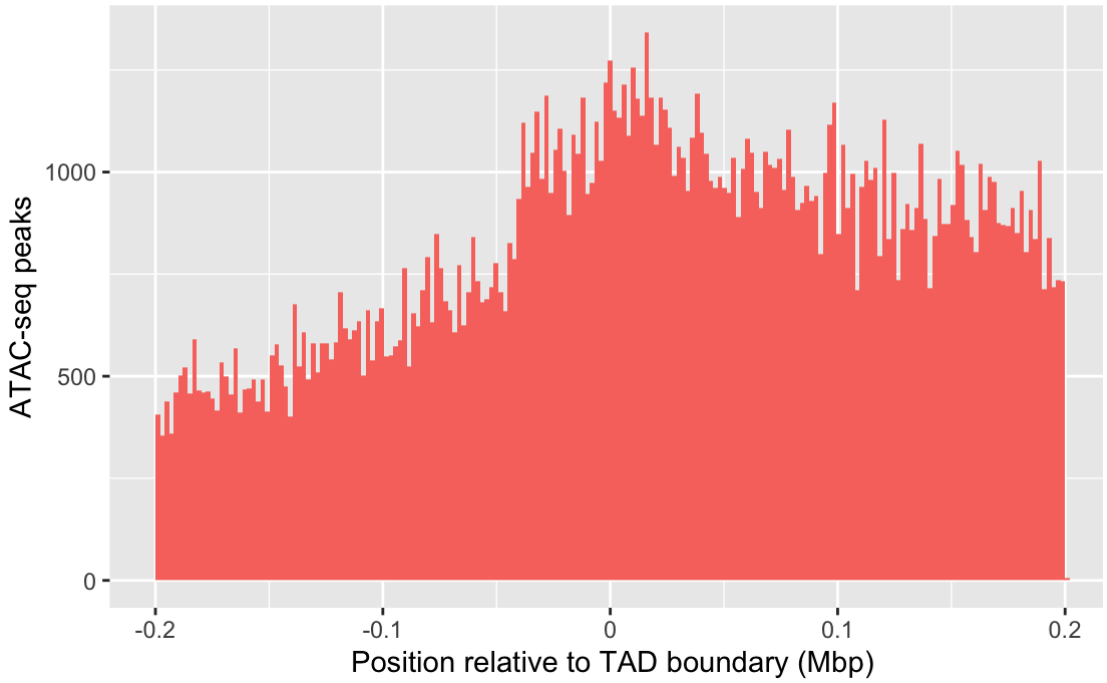


Figure 3.5. Location of all ATAC-seq peaks relative to TAD boundary location, merged across all genes. Reprinted with permission from *Kuffler L, Fortin HJ, Skelly DA, Czechanski A, Munger SC., Reinholdt LG, Baker CL, Carter GW. (2023) Imputation of 3D genome structure by genetic-epigenetic interaction modeling in mice eLife 12:RP88222 https://doi.org/10.7554/eLife.88222.1.*

Furthermore, our models count each interaction with an ATAC-seq peak, genotype, and gene, meaning ATAC-seq peaks may be counted multiple times. Therefore, we also needed to not simply compare the locations of interacting elements, but also compare to the number of possible interactions each ATAC-seq peak could theoretically be involved in. We expected that peaks within the same TAD as the gene they affected would be proportionally more likely to have a significant interaction with the gene, compared to an ATAC-seq peak outside of that TAD.

Therefore, we compared the ratio of intra-TAD vs. extra-TAD models, first for all possible interactions, and then just for those that reached our significance threshold. We found a ratio of 1.47 for all models, and 1.55 for significant models. This is a significant difference according to a chi-squared test = 5.56, p = 0.018.

3.4 Revising Constraints on Local Regulatory Area Using Genetic Data

Definitions of local regulatory area are used to establish the scope of analyses. Many cell types and animal models do not have publicly available Hi-C or CTCF ChIA-PET datasets, further limiting many researchers that may want to study genetic-epigenetic non-additive interactions. Our data supports that TADs may act as a biologically-defined boundary. Therefore, we wanted to determine how far the TAD-defined local area was likely to extend from any given gene. We found that to capture 95% of inter-TAD genetic-epigenetic interactions at adjusted $p < 1x10^{-7}$, a window of 2,071,826 bp upstream and 2,579,188 bp downstream of the gene transcription start site (TSS) was required (Fig. 3.4). The density of results using linear DNA sequence does not necessarily experience a linear drop-off over this distance, due to variable TAD lengths (See Section 3.4.2 and Fig. 3.7E-F). (Reproduced with permission from Kuffler et al., 2023).

3.4.1 Comparisons To Previous Local Area Estimates

We wanted to compare to some of the common definitions of local area that have been previously used in the literature. A window of +/- 100kb, usually centered on the transcription start site was previously thought to be acceptable. But we found that this only captured 14.1% of our significant interacting intra-TAD ATAC-seq peaks. the +/-

500kb window predominated when this project began, but that only captured 47.2% of significant interacting intra-TAD ATAC-seq peaks.

This means that previous studies of local regulatory areas were likely truncated, missing the majority of a gene's local regulatory context. While TADs can range in size, their average length in mice has been calculated to be 1.09 Mb, meaning that TADs can extend further than the +/- 500kb window could cover. In addition, genes are usually not centered within TADs, and thus the gene-centric window makes the incorrect assumption that DNA on either side of a gene is always going to be equally likely to contain relevant regulatory elements. This will be discussed further in the following section.

3.4.2 Interacting ATAC-seq Peaks Overlap With Enhancers, Gene Bodies, Other Features

We next considered the location of these interacting elements with regards to genes and known regulatory elements. ATAC-seq peaks involved in non-additive genetic-epigenetic interactions were analyzed by overlap with mm10 NCBI RefSeq genes and selected Enhancer Atlas 2 datasets (Table 5.1-5.2). 65.22% of peaks on each chromosome fell within gene bodies. Normalized for length, exons were especially enriched, with 34.39% of peaks falling within exons, half of which fell within the first (or only) exon. 6.19% of peaks fall within enhancers curated for ESC R, R1, KH2 embryonic stem cell lines. There was no inverse relationship between gene body vs. enhancer peak percentage (Fig. 3.6A), indicating that this was not simply reflecting the relative density of genes within various TADs across the genome (Reproduced with permission from Kuffler et al., 2023).

It should be noted that these figures are determined by overlap with the features in question, meaning that some ATAC-seq peaks fell over multiple features. This resulted in double-counts of many ATAC-seq peaks across multiple features.

Therefore, it is worth examining not only the percentage of peaks that fall within a certain feature, but also what their density is within those features. Introns can be very long, but open chromatin within them may not be as strongly tied to active gene transcription as open chromatin within an exon.

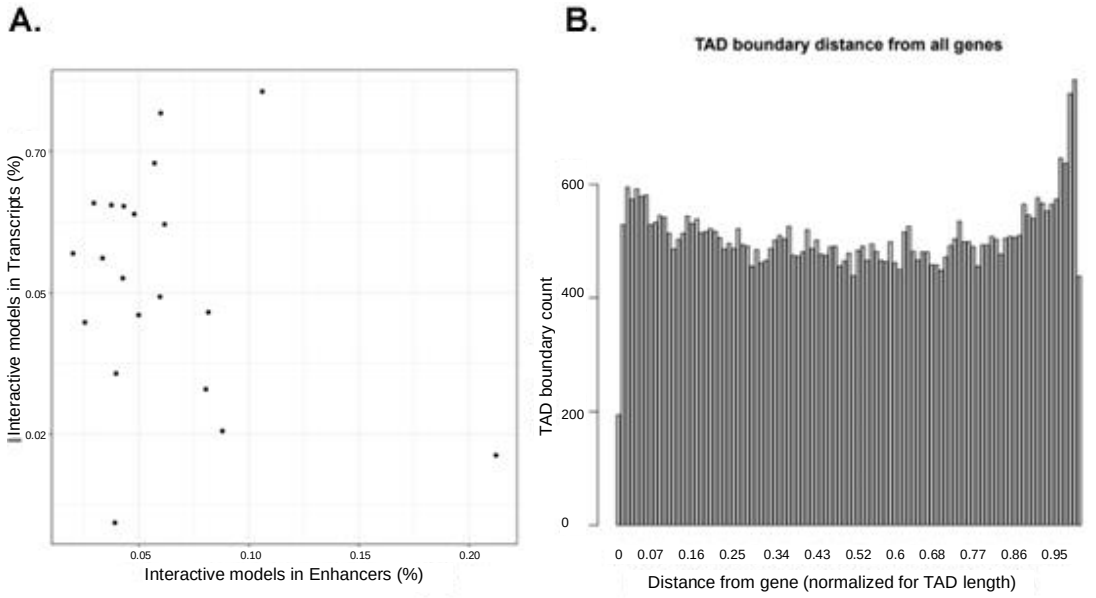


Figure 3.6. Comparisons of ATAC-seq peak and TAD boundary locations relative to genomic features A. Per-chromosome comparison of percentages of interacting ATAC-seq peaks in transcripts vs. in enhancers. **B.** TAD boundary locations relative to distance from each gene contained within them, normalized for TAD length. Reprinted with permission from *Kuffler L, Fortin HJ, Skelly DA, Czechanski A, Munger SC., Reinholdt LG, Baker CL, Carter GW. (2023) Imputation of 3D genome structure by genetic-epigenetic interaction modeling in mice eLife 12:RP88222 https://doi.org/10.7554/eLife.88222.1.*

The density of interacting ATAC-seq peaks within introns is 0.45, while the density within exons is 0.22 per exon. Within the first or only exon, the density is 1.11 ATAC-seq peaks

per exon. Single exon genes have 1.21 ATAC-seq peaks on average. Enhancers have 1.87.

We hypothesize that the density of interacting ATAC-seq peaks within genes themselves likely reflects the most classic forms of non-additive interaction, as it is generally understood: chromatin must be open over a gene for transcriptional machinery to access it. Therefore, open chromatin at a gene has a positive effect on gene expression. A genetic variant can affect the likelihood of gene transcription if it falls within a gene body or close proximal features like promoter regions. These variants can increase the chance of gene expression by making it easier or harder for RNA polymerase to bind or successfully transcribe a gene. These effects will only become evident if the gene is primed for transcription in the first place, thus the genetic variant is dependent on the open chromatin to achieve its effects.

There are also other possible ways that genetic variation and open chromatin within a gene could affect gene transcription. Intron-mediated enhancement is one such method(Callis et al., 1987; Palmiter et al., 1991). While open chromatin in these areas may indicate a primed state rather than active transcription, our bulk ATAC-seq and RNA-seq data means that we are more likely to capture the connection between priming and transcription compared to single cell data, in which these temporally distinct events are separated from each other.

It is also possible that a gene with multiple transcription start or end sites could be made more likely to complete transcription if a genetic variant alters the probability of transcription for a specific portion of a gene. If an omitted region is associated with

transcriptional stalling, then reducing its transcription would increase the number of transcripts successfully produced.

However, while these variants may be very interesting, the interactions themselves were more likely to be well-characterized already. In later sections, we will transition from describing the location of these regulatory features relative to the genes they affect, and focus more on the identity of DNA binding sites that overlap with these features, focusing on some that do not directly alter transcription.

3.5 Density of Interactions Between Genomic Elements Is Defined By 3D Context

If TADs act as a constraint to local interaction, then their 3D looping structure should be reflected in the regulatory patterns of genes found within them. At the most basic level of TAD structural organization, the CTCF binding site brings distant areas of chromatin into close physical association. We analyzed genome-wide distribution of SNP-interacting ATAC peaks relative to genes within their respective TADs. This revealed that the gene-centric increase in regulatory interaction density responded to the 3D context of the TAD loop, crossing CTCF boundaries irrespective of linear DNA sequence. This creates a distribution of non-additive interactions centered at the gene promoter that reaches its minimum halfway around the TAD loop from the gene (Fig. 3.7A) (Reproduced with permission from Kuffler et al., 2023).

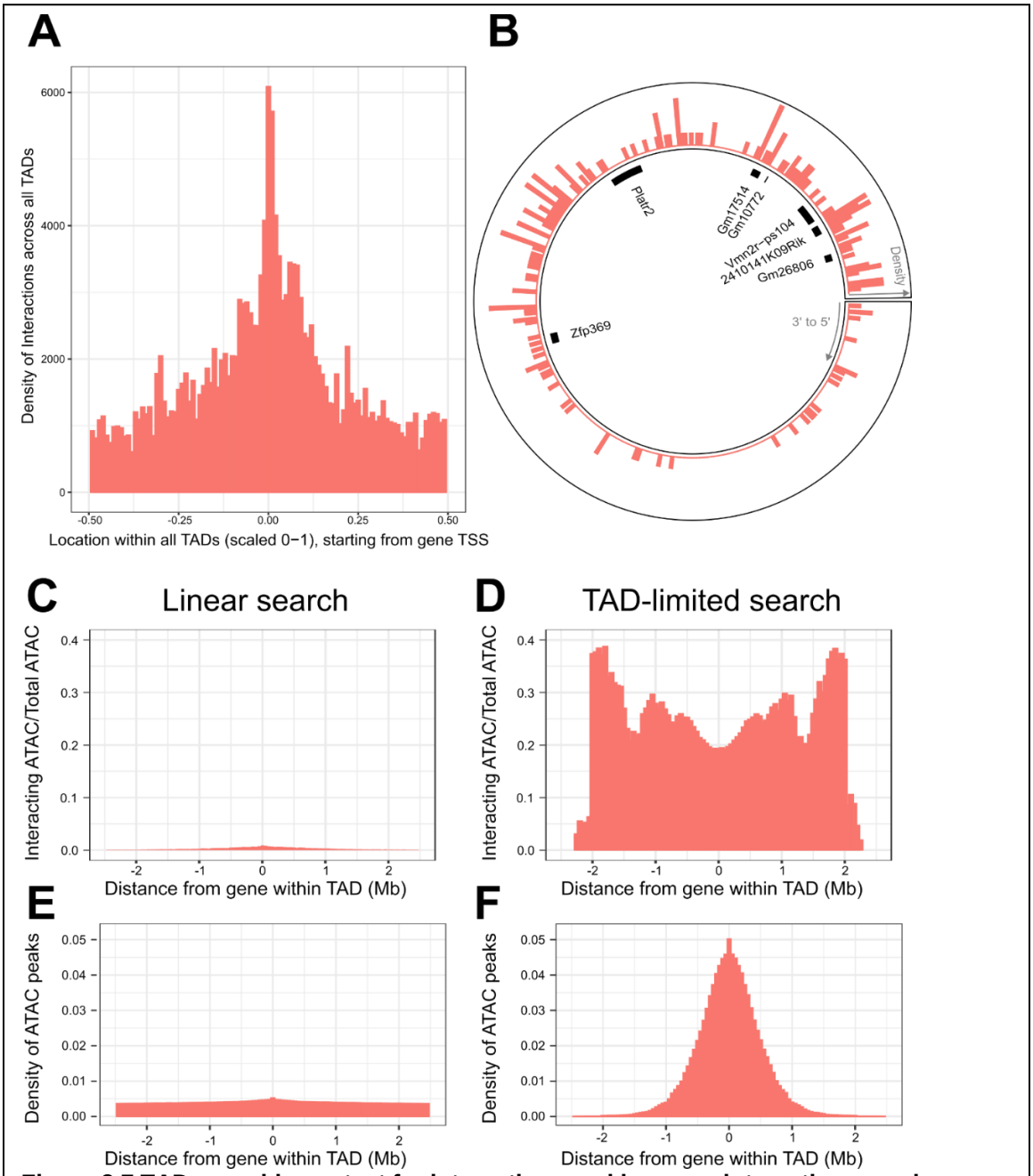


Figure 3.7 TADs provide context for interactions and increase interaction search efficacy. A. Counts of intra-TAD ATAC-seq peaks involved in all non-additive interactive models, centered on the TSS of the gene affected by the SNP-ATAC interaction. Coordinates transformed to a standard scale. **B.** Example TAD, displaying interacting ATAC peak density and gene locations. Peak relevance generally decays relative to intra-TAD distance rather than linear chromosomal distance. **C-F.** A comparison between linear sequence-based and TAD-limited search methods for interacting ATAC-seq peaks. C and D compare percentage of significantly interacting ATAC-seq peaks at each gene-relative locus. E and F compare density of ATAC-seq peaks at each locus. TAD-based search shows a higher density of interactions and places limits on search distance due to testing only TAD-internal ATAC-seq peaks. Reprinted with permission from *Kuffler L, Fortin HJ, Skelly DA, Czechanski A, Munger SC., Reinholdt LG, Baker CL, Carter GW. (2023) Imputation of 3D genome structure by genetic-epigenetic interaction modeling in mice eLife 12:RP88222*

3.5.1 Scaling TADs For Genome-wide Analysis

To plot and quantify this distribution, we could not use linear DNA sequence or unscaled bp distance. This is due to a number of complicating factors regarding TADs and the location of genes within them. TADs do not have a fixed length. Therefore, interacting ATAC-seq peaks will necessarily tend to be closer than 1.09 Mb away from a gene, because the TADs themselves average that distance in length.

Without the scaling factor, we see this distorting effect in action (Fig. 3.8). Interactions cluster within a confined plateau around +/- 250kb, with the majority well within +/- 500kb from the gene, eventually dwindling to nothing as more and more TADs fall out the further one travels from the gene.

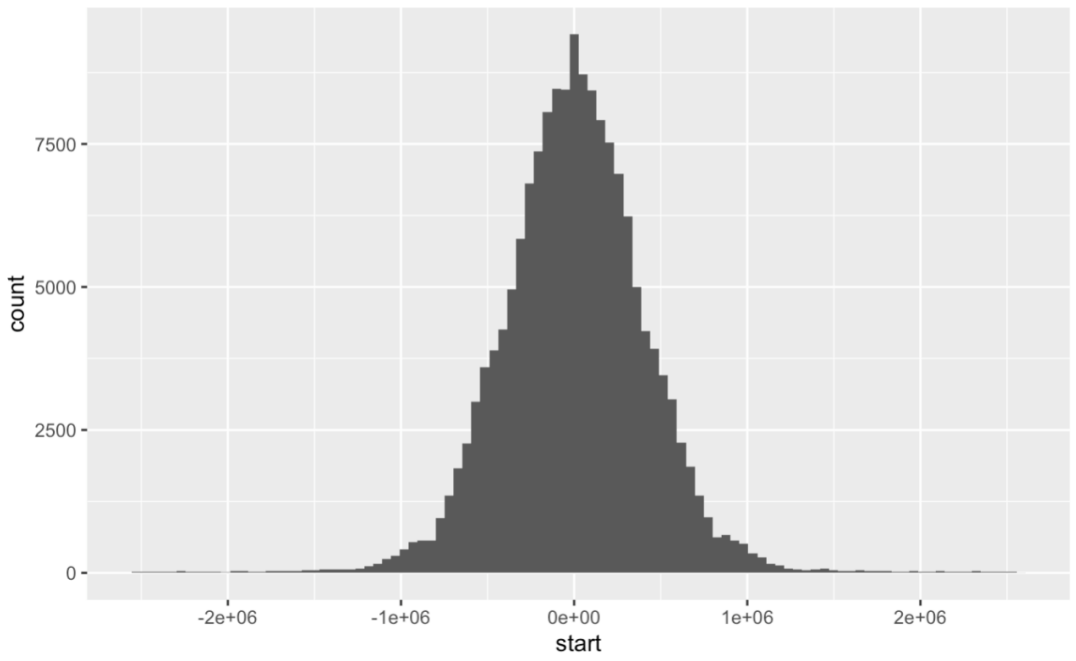


Figure 3.8. Unscaled gene-centric plot of all interactions per ATAC-seq peak. Measured in Mbp. The y-axis measures the number of interactions per locus.

Genes are also not necessarily centrally located within a TAD (see section 3.4.2 for

more details). This distorts the distance between a gene and an interacting feature within 3D space, when viewed purely by linear distance. If a gene lies next to the start of a TAD and an ATAC-seq peak lies next to the end of the TAD, the two features will be brought into close proximity in 3D space, and are thus more likely to physically interact and engage in regulatory behavior.

Therefore, we had to incorporate the structure of the TAD into our calculations, as well as introduce a scaling factor for each TAD. This resulted in a plot which centers the gene TSS, placing an equal amount of intra-TAD DNA upstream and downstream of the TSS. The coordinates were then set to a standard scale.

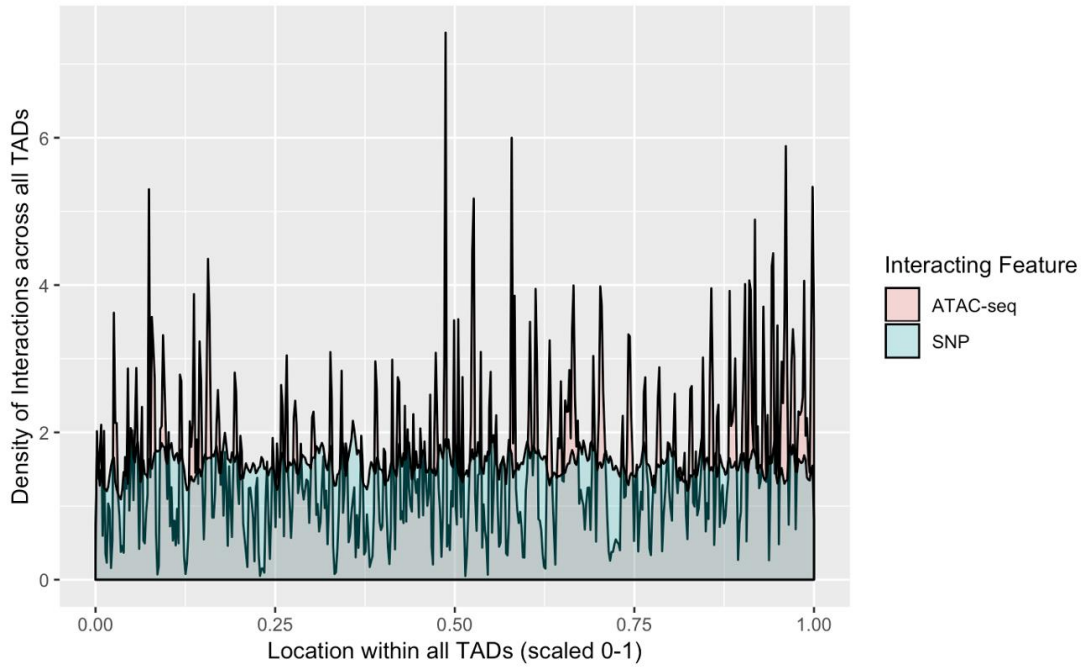


Figure 3.9 Intra-TAD interacting feature locations by linear DNA sequence. Location of interacting elements within TADs, scaled 0-1 to remove falloff over distance. Does not show comparable levels of organization versus gene-centric plots (Fig. 3.7A). SNP locations have high degrees of inaccuracy due to linkage disequilibrium. For more on interacting ATAC-seq location, see Fig. 3.10

These centered, scaled graphs allow us to confirm that intra-TAD interactions center around genes. This is as expected, but could not be taken for granted. To that end, we also created a version of this graph centered on the TAD binding sites, with the standard distance scale still intact. This revealed no similar level of organization to interacting ATAC-seq location.

3.5.2 Comparing Search Efficiency With TAD Data Versus Without

We had thus far seen evidence that TADs contained the majority of genetic-epigenetic interactions, in a gene-centric manner that spanned TAD binding sites. We were interested to determine whether this could be used to improve the speed and efficiency of searches for relevant regulatory elements.

Thus, we sought to determine whether gene-centered, TAD-internal searches for interacting elements were potentially a more efficient method for searching for non-additive interactions, when compared to local area searches along the linear DNA sequence. We compared the two methods, running a pair of genome-wide, gene-centric analyses out to the limit of our TAD-internal search (+/- ~2.5 Mb). We calculated the percentage of ATAC-seq peaks within this region that participated in significant interactions. (Fig 3C-D). The overall density of ATAC-seq peaks within these regions was also compared (Fig. E-F) (Reproduced with permission from Kuffler et al., 2023).

We had previously identified that the majority of interacting regulatory elements lay outside of previously reported, arbitrary local area constraints. This was due to the length of TADs, and gene placement within them. We therefore wished to go further to

describe more practical reasons why using TADs in analyses of gene regulation is useful for researchers.

Despite using a publicly available dataset that only matched the genotype of one of the eight DO founder strains, the TAD boundary data we used showed a strong dropoff in interaction density at TAD boundaries. Capturing the majority of these interactions without TAD data requires searching a large window of linear DNA coordinates. We hypothesized that it would likely be more efficient to perform searches within TADs, rather than use the linear DNA sequence.

Therefore, we decided to compare the standard search method for epigenetic regulatory factors--searching outwards from a gene along the linear DNA sequence--versus our TAD only approach. Comparing these approaches posed some challenges, due to the differing qualities of the search spaces.

Because we had already calculated the significance of all potential interactions within the local TAD area, this was not technically a search, meaning we already knew the extent to which we would find TAD-inclusive results. Therefore, we set the limits of the linear DNA sequence plot equal to the limits of our most distant intra-TAD interactions. This would draw from extra-TAD interactions, thanks to our previous analyses.

To calculate search efficiency, we also needed to examine overall ATAC-seq peak density. This allowed us to consider the quantity of significant results as a proportion of features examined. We expected that ATAC-seq peak density would remain relatively

constant in the linear search, while the intra-TAD search would see a steady decline in ATAC-seq peak density the further one goes, as TAD boundaries are reached.

We hypothesized that we would see a higher proportion of interacting peaks in the intra-TAD search, likely recapitulating the structure seen in Fig. 3.7A. We also expected to see a similar distribution in the search along the linear DNA sequence, but with a sharper falloff.

We found that TAD-based searches produce consistently higher percentages of interacting ATAC-seq peaks across the analysis, and higher ATAC-seq peak density out to +/- ~1Mb (Fig. 3.7). TAD-based searches experience a progressive falloff in ATAC-seq peak density when compared to linear sequence-based searches. This is largely due to dropout of smaller TADs from the analysis. This results in more variable interaction percentages past ~1Mb from the gene. However, the decreasing number of TADs out past +/- 500kb does not fully explain the fluctuations that take place past this point. They have symmetry that does not follow a simple curve, and we are uncertain why this occurs (Reproduced with permission from Kuffler et al., 2023).

We also found that there was a depression in the proportion of interacting ATAC-seq peaks immediately around the gene. We suspect this may be due to the fact that some ATAC-seq peaks may be associated with constitutive transcriptional activity, or transcriptional priming. These are of course regulatory processes, but they may not be in non-additive interactions with any local sequence variants that alter gene expression. This is an excellent reminder that our methods are not seeking to capture all regulatory

activity, only regulation that is otherwise difficult to access without this multiomics approach.

3.5.3 Analyzing Gene Distribution Within TADs

Linear search results show a consistently lower percentage of interacting ATAC-seq peaks, and a lower density of ATAC-seq peaks per locus. This was unexpected, as the immediate vicinity of the gene was expected to contain a similar density and non-additive interaction potential of ATAC-seq peaks when compared to the TAD-based search method. We therefore hypothesized that TAD boundaries in close proximity to some genes might create this discrepancy (Reproduced with permission from Kuffler et al., 2023).

As expected, we calculated that in absolute terms, genes tended to be very close to TAD boundaries, with half of all genes lying under 670kb away from the closest TAD boundary. However, this calculation requires a normalization of TAD length. The shorter a TAD, the less distance there can possibly be between a gene's TSS and the boundary. After normalizing for TAD length, we found an interesting distribution.

We found that while the majority of genes were randomly distributed within their TADs, 12,437 or 24.31% of genes were found to have a TAD boundary close upstream to their transcription start site (Fig. 3.6B) (Reproduced with permission from Kuffler et al., 2023).

This can also be visualized as an excess or deficit relative to the mean distribution of gene TSS (Fig. 3.9). This highlights some odd features of the distribution, particularly a strong bias against transcription start sites immediately downstream of the gene, which

is present but not fully mirrored at the other end of the TAD. The inverse is true for the preference for gene TSS lying within the final or initial 10% of the TAD's length.

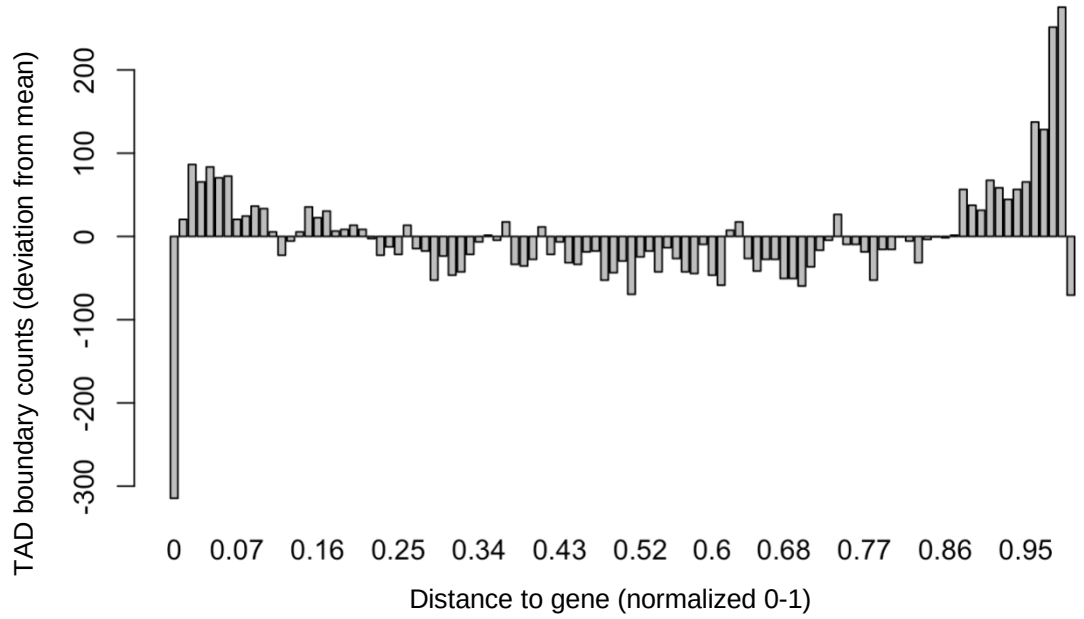


Figure 3.10. TAD boundary locations relative to distance from each gene contained within them, normalized for TAD length and scaled for deviation from the mean.

It is worth noting that our TAD boundary locations have a 40kb window of uncertainty due to the resolution limitations of the public dataset we used. It therefore makes intuitive sense that genes would be less likely to fall within regions where they may be bisected by a TAD boundary. It also makes sense that genes would favor a location that implies distant, potentially regulatory DNA sequences are brought into close proximity by the TAD loop structure. However, without more precise Hi-C, ChIA-PET, other chromatin conformation and/or protein binding data, we could not examine this in greater detail.

3.5.4 ATAC-seq Peaks Are Less Likely To Be Interacting Near Genes

Interestingly, when examining TAD-internal searches for ATAC-seq peaks there is a depletion in the interaction rate of ATAC-seq peaks flanking the gene (Fig. 3.7D), which is not noted in raw interaction counts (Fig 3A). It is worth noting that the area of

depression is a window of approximately +/- 500kb, similar to both the standard arbitrary window used to define local areas, and the average length of TADs in mice. This would indicate that for the region where we have > 50% of TADs remaining, the percentage of interacting ATAC-seq peaks is steadily increasing (Adapted with permission from Kuffler et al., 2023, edited for clarity).

This may be due to the necessary presence of open chromatin at the promoter and the gene body during priming and transcription regardless of interaction. There are also genes which are concomitantly active within our samples, and do not show variation in expression, thus experiencing activity without detectable regulatory interaction.

Alternatively, these areas may not respond to local regulatory signals and are solely responsive to regulators acting in trans, which may be unaffected by local genome sequence. Furthermore, given the fact that some regions have large, TAD-spanning blocks of linkage disequilibrium, there may be regions where ATAC-seq peaks exist, but we were unable to determine with confidence what genotype was present at that location, due to its distance from the closest gigaMUGA pseudoprobe. This means that there are regions where we were unable to resolve any haplotypes, thus we cannot have any interacting models.

Our results indicate that the local area for non-additive genetic-epigenetic interaction is not only constrained by TADs, but also shaped according to the overall 3D genome structure of the TAD loop. Analytical methods which reflect this are more efficient at discovering interacting regulatory elements (Reproduced with permission from Kuffler et al., 2023).

3.6 Genetic-Epigenetic Interactions Influence Gene Expression More Than Epigenetic Factors Alone

The coefficients estimated by our regression model quantify the effects of the genetic and epigenetic features on the expression of each gene (Equation 1). We postulated that our ATAC-seq, SNP, and non-additive interaction effects would display patterns that corresponded to positive and negative regulatory roles with varying strengths (Reproduced with permission from Kuffler et al., 2023).

In previous analyses of SNP-SNP interactions, the importance of relative magnitude of interaction effects has been a subject of debate, as single effect terms ("main effects") are normally of greater magnitude than interaction effects(Leon & Heo, 2009; Tyler et al., 2016). However, in the context of non-additive genetic-epigenetic interactions we observed interaction effects of greater absolute value than ATAC-seq effects 70.52% of the time, greater than SNP coefficients 13.42% of the time, and greater than both main effects 14.00% of the time. This suggests that ATAC-seq peaks are generally a modifier of underlying genetics, so the majority of ATAC-seq peak effects are smaller than interaction effects. This was supported by the relatively low contribution of ATAC-seq peaks to single-effect regression models (Fig 1D). Their quantitative variation produces relatively subtle effects. This stands in comparison to the binary presence or absence of SNPs, which also directly capture regulatory sequence.

The directionality of effects on gene expression specifies the positive or negative influences from SNPs, ATAC peaks, and their combinations. Positive and negative effects represent correlation or anti-correlation of SNP presence or ATAC-seq score with

gene expression. When all effects have the same sign (positive or negative), the total effect is synergistic and the pairwise combination of SNP and ATAC peak alter expression beyond the sum of the two. When the interaction effect has the opposite sign of the main effects, this indicates redundancy or interference, based on the idea that redundant factors create an “or” logic that yields a combined result that is less than the sum-based expectation. Alternatively, mixed main effects with interactions can also signify a suppression of one outcome in favor of the other, in which the sign of the interaction term serves to move the additive expectation nearer to one of the two marginal outcomes (Reproduced with permission from Kuffler et al., 2023).

3.6.1 Interacting Model Behavior Indicates Many Interaction Subtypes

We wanted to assess the potential functions underlying interactions as described above. Therefore, we split all significant models by their ATAC-seq, SNP, and interaction signs.

It is worth noting that the positive or negative sign of the interaction term does not necessarily correlate with increased or decreased expression of the target gene, when compared to the ATAC-seq or SNP effects by themselves. Instead, the interaction term's sign indicates a greater or lesser effect of the two main effects in combination, versus what would be expected in a purely additive model. For example, if the expected level of gene expression without regulatory activity is 1, its level in the presence of an ATAC-seq peak is 3, and its expected level in the presence of a non-reference SNP is 3, these would be positive main effects, as they both add a score of 2 to gene expression. One would expect an additive effect of these two regulatory elements, equaling a result of 5 in total. If the actual result is a score of 6, that would mean a positive interaction effect, as the two factors working in combination exceed expectations. If the actual result is a

score of 4, then the interaction effect is negative, because the contributions of the two factors are not performing to expectations. It is worth noting that we cannot determine which of the main effects is primarily responsible for the interaction with the other, as the regression model is agnostic as to the order or directionality of its terms.

We can also arrange models by the magnitude of their effects. For example, if a gene's expression level is 3 in the presence of an ATAC-seq peak, 4 in the presence of the SNP, and 6 in the presence of both, that would mean that the SNP effect is largest, then the ATAC-seq peak, then the interaction term is smallest, because the presence of both factors only adds 1 above the expected additive effect of the two if they were independent of each other. This may imply different functions underpinning the models depending on which effects are strong than others.

Breaking down these models is a complicated task, and one that we cannot fully complete with our dataset as it stands. However, we can describe the properties of the models we have, and the frequency with which we encountered them. This analysis can be found in the appendix (Chapter 5.3).

This analysis details 64 potential configurations of magnitude and sign that these models allow. Out of these possible configurations, 63 contained results. These were not evenly distributed across configurations, with some patterns of effect sign heavily favoring certain relative magnitudes of effect terms, and vice versa. These gave hints to the potential mechanisms that might underlie these models, which are summarized in the analysis for those who may wish to explore our data more thoroughly.

it is worth noting that in most models, Interaction terms are of greater magnitude than ATAC-seq terms, and even outweigh SNP terms a surprising amount of the time. This is against expectations--usually main terms predominate, while interaction terms have a modulating effect. In this case, ATAC-seq peaks seem to have taken up that role.

This may be due to the nature of the interactions themselves. Perhaps this indicates a causal relationship between the SNP and ATAC-seq peak, or colocalization. However, neither are accessible from our data.

We found that models indicating redundancy or interference were the most common overall, totaling 39.54% of all models (Table 3.3). Synergistic effects were found in 16.74% of all models. These two observations, suggest that a greater proportion of regulatory non-additive interactions attenuate gene expression, rather than strengthening it. While ATAC-seq peaks are often correlated with increased gene expression(Cao et al., 2018), we were surprised to find that an increase in ATAC-seq signal had a negative effect on gene expression in 40.73% of models (Table 3.3). Due to the high proportion of ATAC-seq peaks found within gene bodies and the association between open chromatin and gene transcription, this result warranted further investigation (Reproduced with permission from Kuffler et al., 2023).

Effect	Effect sign (all interacting models, adj. p < 1x10^-7)							
ATAC-seq	+	+	+	+	-	-	-	-
SNP	+	+	-	-	+	+	-	-
Interaction	+	-	+	-	+	-	+	-
% of total	16.44%	27.76%	10.75%	4.33%	12.39%	16.26%	11.78%	0.30%

Table 3.3. **Model percentages calculated by distribution of effect signs for all significant interacting models.** Reprinted with permission from *Kuffler L, Fortin HJ, Skelly DA, Czechanski A, Munger SC., Reinholdt LG, Baker CL, Carter GW. (2023) Imputation of 3D genome structure by genetic-epigenetic interaction modeling in mice eLife 12:RP88222 https://doi.org/10.7554/eLife.88222.1.*

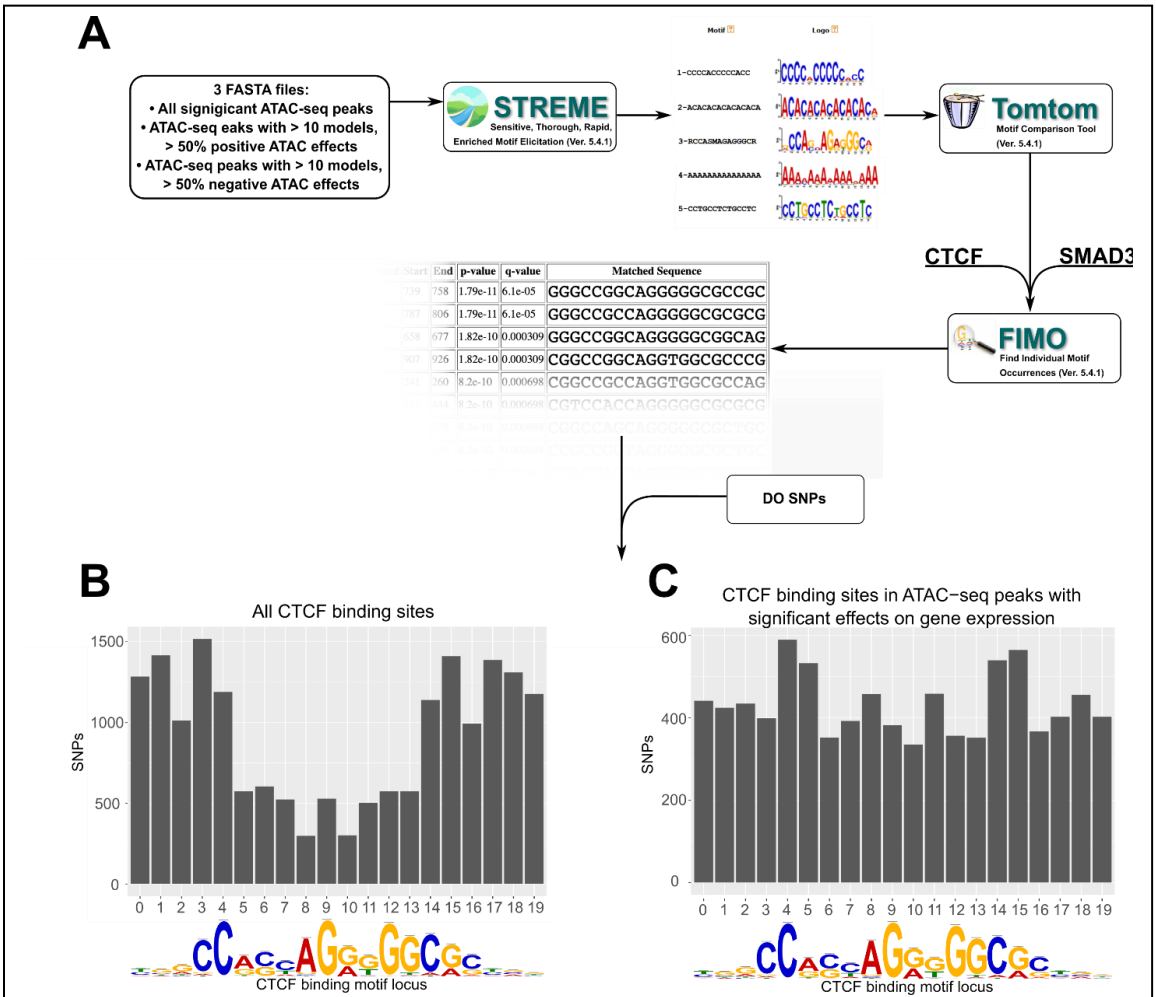
3.7 Motif Enrichment Analysis Reveals CTCF Complex Participation in Genetic-Epigenetic Interactions

We looked next for potential binding sites or functional motifs underlying our results to provide clues as to the mechanistic underpinnings of model effects. We were especially interested in the subset of ATAC-seq peaks with negative effects on gene transcription, as these results were counter to our expectations. We hypothesized that these areas of open chromatin might expose binding sites of repressive regulatory factors (Reproduced with permission from Kuffler et al., 2023).

To test this hypothesis, we tested for DNA motif enrichment analysis using MEME Suite (Fig. 3.10). We selected the subset of ATAC-seq peaks involved in 10 or more significant non-additive interactions, at least 50% of which have negative ATAC-seq effects (negative effectors). This subset was compared versus all ATAC peaks and against shuffled control sequences via STREME analysis, which finds enriched ungapped motifs in provided sequences (Reproduced with permission from Kuffler et al., 2023).

Results showed the negative effector subset was enriched for 49 motifs (Supplemental File 1), including the CTCF binding site ($p < 2.0x10^{-14}$) and SMAD3 binding site ($p < 1.5x10^{-8}$), an optional component of the CTCF complex. These motifs were present, but less significantly enriched in positive effectors, or in all significant ATAC-seq peaks. This suggested that a portion of negative ATAC-seq effects can be functionally explained by altered behavior of CTCF binding sites carried by specific ancestries in the DO. To complement this analysis, we also quantified the CTCF motif occupancy in negative effectors versus other ATAC-seq peaks. FIMO motif scanning showed that 53.3% of top

negative ATAC-seq sequences had at least one CTCF binding motif in them, compared to 35.3% in ATAC-seq peaks with positive effectors on gene transcription (See deposited data) (Reproduced with permission from Kuffler et al., 2023).



Start	End	p-value	q-value	Matched Sequence
739	758	1.79e-11	6.1e-05	GGGCCGGCAGGGGGCGCCGC
787	806	1.79e-11	6.1e-05	GGGCCGCCAGGGGGCGCGCG
658	677	1.82e-10	0.000309	GGGCCGGCAGGGGGCGGCAG
907	926	1.82e-10	0.000309	CGGCCGGCAGGTGGCGCCCG
241	260	8.2e-10	0.000698	CGGCCGCCAGGTGGCGCCAG

Figure 3.11 Motif analysis identifies differences in interacting CTCF binding motifs. A. A schematic of our motif analysis through MEME Suite. FASTA files derived from interacting ATAC-seq peaks are used to identify enriched motifs, identify protein binding sequences, and locate the sequences within the ATAC-seq peaks. **B-C.** Binding sites found within significant motifs are less protected from genetic variation. SNP counts are shown at each locus in the CTCF binding sequence, comparing motifs within interacting ATAC-seq peaks versus all CTCF binding sites. Reprinted with permission from *Kuffler L, Fortin HJ, Skelly DA, Czechanski A, Munger SC., Reinholdt LG, Baker CL, Carter GW. (2023) Imputation of 3D genome structure by genetic-epigenetic interaction modeling in mice eLife 12:RP88222 https://doi.org/10.7554/eLife.88222.1.*

We next examined whether these peak locations contain SNPs that might alter CTCF binding potential. Imputing from the founder genomes of the DO population, we analyzed the locations of SNPs in CTCF binding motifs associated with regression models versus

all CTCF motifs. Out of CTCF motifs within this negative subset, 96.1% of these were found to contain a DO founder SNPs, versus 18.0% for positive effect-associated CTCF motifs ($p < 2.2x10^{-16}$). Similar results were found for SMAD3. We also found that across all motifs, the density of SNPs favored the start and end of the binding sequence (Fig. 3.10B). However, in motifs associated with regression models, the density of SNPs was approximately equal across all bases (Fig. 3.10C). These bases have previously been identified as having high protein-DNA binding energies with the canonical sequence(Cao et al., 2018; Zuo et al., 2017) (Reproduced with permission from Kuffler et al., 2023).

Most genomic sequences matching the CTCF binding motif are not known to be bound, according to ChIP-seq experiments(Maurano et al., 2015). We wanted to determine the overlap between CTCF binding sites found in our data and known active binding sites in mESCs. Comparing to available ENCODE CTCF ChIP-seq in C57BL/6 Bruce4 and 129/Ola E14TG2a.4 mESCs(Shen et al., 2012; The ENCODE Project Consortium, 2012), we found 2.04% and 1.87% of these overlapped with negative ATAC effect CTCF binding sites that contained SNPs, versus 17.45% and 16.20% overlapping with positive effect CTCF binding sites that contained SNPs. Expanding the scope to attempt to find flanking Smad3 regions netted consistently low results, with a 1 kb flanking window returning between 1.20% and 0.41% for Bruce4 and E14TG2a.4 respectively. These findings show that the majority of CTCF binding sites found within our significant models are not captured in previous analyses of ESCs two different *M. musculus musculus* strains (Reproduced with permission from Kuffler et al., 2023).

3.8 Putative Developmental Regulator Platr2 is Regulated By Multiple Redundant Elements

To provide an example of our analytical method and probe a gene previously proposed to be important in stem cell development, we examined *Platr2* and its TAD. *Platr2* is a long non-coding RNA gene transcribed in mESCs. Genetic variation in and around this gene has been observed to have distal regulatory effects on a number of embryonic ectoderm-associated genes in stem cell culture(Skelly et al., 2020). While our interest in this project was focused on systems-level effects of non-additive regulator interactions and chromatin organization, we felt it was important to produce a specific example of these phenomena for those who are more comfortable with individual models, and may be interested in applying these analytical techniques to individual genes (Adapted with permission from Kuffler et al., 2023, edited for clarity).

Platr2's TAD contains a high concentration of confidently called regression models in our analysis (Fig 3B), with 1,031 models of non-additive SNP-ATAC interaction reaching our significance cutoff for resident genes, of which 179 were models of *Platr2* expression. The TAD is 2,480,372 bp long, and it also contains 22 haplotype markers, allowing a certain level of model SNP localization. These haplotype markers range from 2.5 kb to 411 kb, with a median length of 6.3 kb. 14 of these regions have genotype effects on interacting models that affect *Platr2*. The gene itself lies in a relatively conserved region approximately 195 kb and 108 kb away from its flanking predicted LD boundaries. Previous studies have found a group of genes regulated in *trans* by expression quantitative trait loci (eQTLs) mapped to *Platr2*(Skelly et al., 2020). These target genes are associated with embryonic ectoderm, indicating *Platr2* may act as a regulator of

stem cell state. These factors made it a target of interest for further exploration (Adapted with permission from Kuffler et al., 2023, edited to add further analysis).

Effect	Effect sign (Platr2 interacting models, adj. p < 1x10^-7)							
ATAC-seq	+	+	+	+	-	-	-	-
SNP	+	+	-	-	+	+	-	-
Interaction	+	-	+	-	+	-	+	-
% of total	4.31%	38.55%	1.13%	12.47%	12.47%	9.07%	21.77%	0.23%

Table 3.4. Model percentages calculated by distribution of effect signs for the gene Platr2.

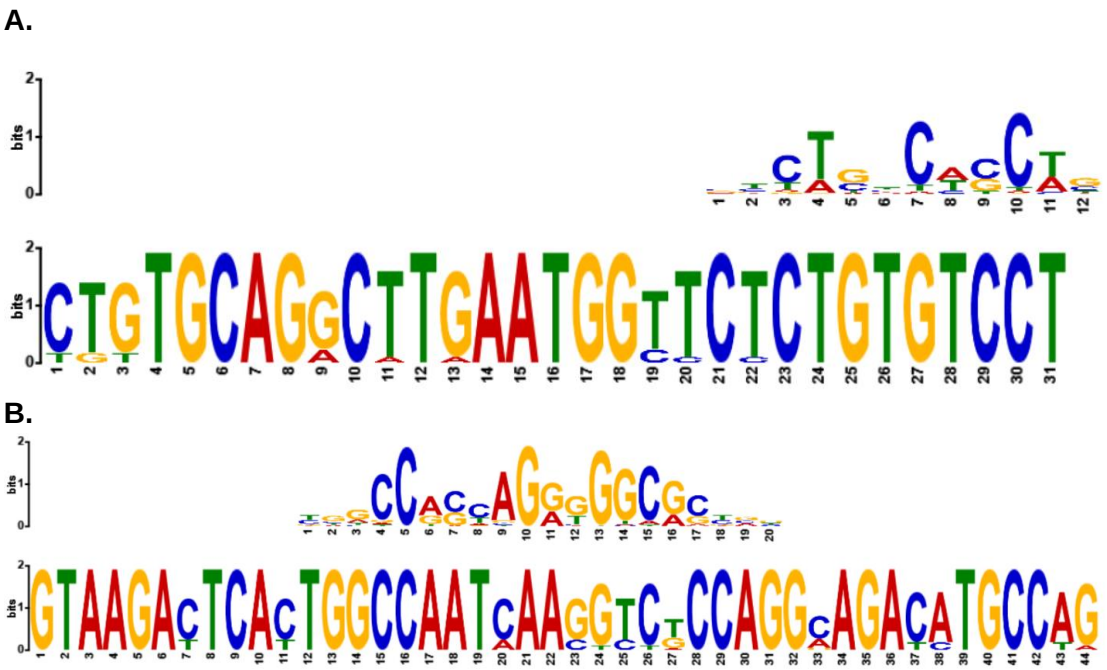


Figure 3.12. **Smad3 and Ctcf binding sites within motifs identified in Platr2's TAD. A-B.** Smad3 and Ctcf binding sites within motifs identified in Platr2's TAD. Reprinted with permission from *Kuffler L, Fortin HJ, Skelly DA, Czechanski A, Munger SC., Reinholdt LG, Baker CL, Carter GW. (2023) Imputation of 3D genome structure by genetic-epigenetic interaction modeling in mice eLife 12:RP88222 https://doi.org/10.7554/eLife.88222.1.*

When the direction of effects for *Platr2* were analyzed, the distribution showed a shift toward models where ATAC and SNP effects agreed with each other, but not with the interaction term (Table 3.4). As discussed above, this potentially indicates functional

redundancy or interference between haplotype and chromatin openness at these sites. Motif enrichment analysis of interacting ATAC-seq peaks identified a sequence at 16 sites which contain Smad3 binding motifs, and another sequence at 15 sites that contain CTCF binding motifs (Fig. 3.11A-B), which may suggest modulation of CTCF binding strength. These results suggest that *Platr2* may have differential regulation patterns governed by changes in TAD formation (Reproduced with permission from Kuffler et al., 2023).

Publicly available data could not suffice to analyze and confirm this. Data available through ENCODE is limited to B6 and 129 ESCs. These are both from the same subspecies. The cell culture preparations were also different from each other, and different from our own DO mESCs. The 1i medium used for the DO mESCs is not standard, and allows more potential for samples to begin priming for differentiation according to their lineage bias in cell culture.

To test our hypothesis, we performed CTCF ChIP-seq on mESCs derived from four of the eight DO founder strains, including representatives from the three subspecies contributing to the DO population. C57BL/6J was also included as the standard reference (Adapted with permission from Kuffler et al., 2023, edited for clarity).

Difficulties were encountered in sequencing. One batch of three samples failed to enrich for unknown reasons, leading to only two samples in B6, PWK, and WSB. These were rerun. Upon running quality control checks, it was found that the rerun PWK sample appeared to actually be a spurious B6 sample of unknown cell line. This was discarded. Initial QC of the data appeared to show that the WSB and B6 reruns had successfully

enriched, and analysis proceeded. However, PCA analysis of ChIP-seq binding intensity showed that the rerun WSB was largely driving PC1. Upon closer examination, this sample had failed to enrich. It was discarded. The loss of these samples did not change our analysis method going forward: we had already decided to primarily analyze ChIP-seq peaks that appeared in at least two samples. Losing a sample in a given strain necessarily lowers our detection capability, but does not impact the significance of consensus peaks used in our analysis (see Methods, Chapter 2.6).

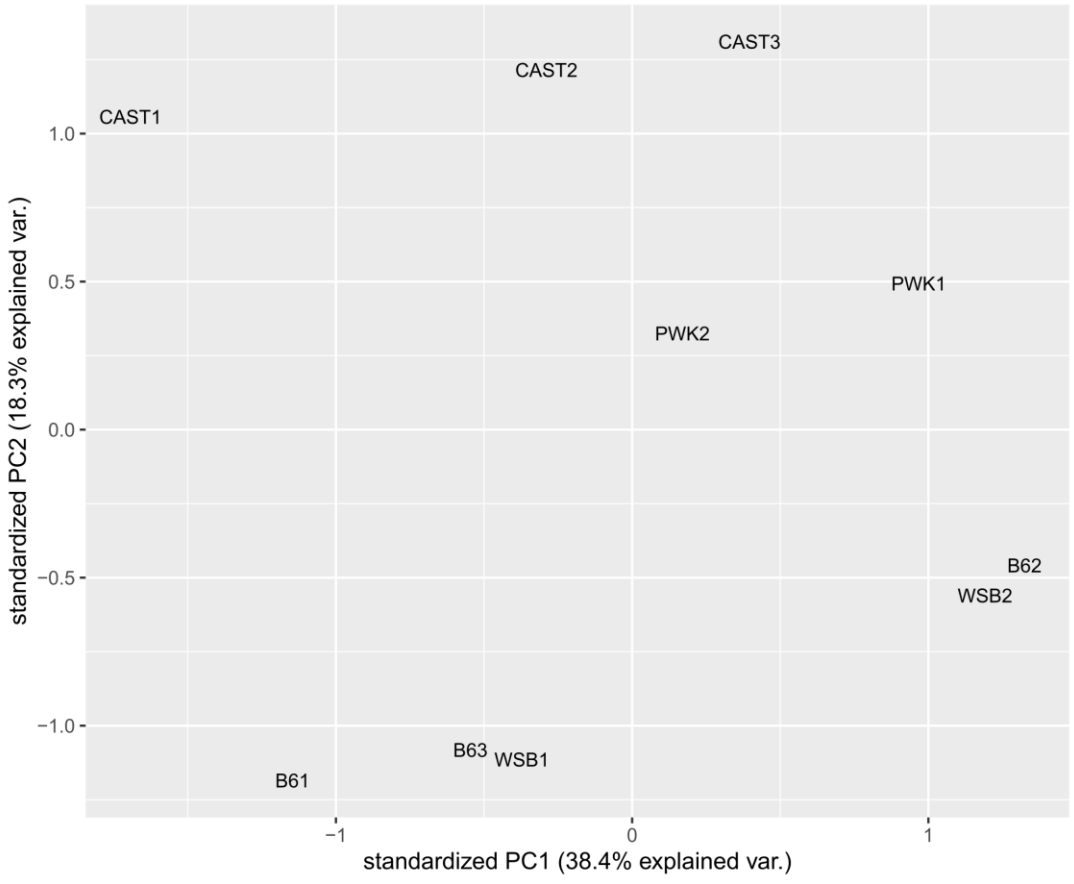


Figure 3.13. Principle component analysis of CTCF ChIP-seq binding intensity.

Subsequent ChIP-seq binding intensities clustered by strain in PCA, and were separated by subspecies as expected (Fig. 3.12). Notably, PC2 (18.3% of explained variation) separates by subspecies, but PC1 (38.4% of explained variation) does not. It separates sample CAST1 from CAST2 and CAST3, and clusters B61, B63, and WSB1 together,

separately from WSB2 and B62. This might seem to be a batch effect, but the sample IDs do not actually cluster by sequencing run or any other batch-related factor we have been able to determine (Adapted with permission from Kuffler et al., 2023, edited for clarity and to expand context).

We undertook a number of exploratory analyses at this time. We wanted to learn the properties of our data, and understand how to best proceed with predictions of the CTCF ChIP-seq results from our DO data.

Across all samples, 65,541 peaks made it to our significance threshold (found in at least two samples in at least one strain), or 68.21% of results. We found 23,887 of these overlapped with CTCF binding sites that had been previously identified, or about 1/3rd overlap. 16,951 overlap with CTCF ChIP-seq previously retrieved from ENCODE, and among those, 7,880 that overlap with significant, interacting ATAC-seq peaks in our DO mESC data. About 80% of these are found in all four strains.

We wanted to examine this subset first, because it overlapped with the most publicly available data, and had potential links to our DO mESC data. While most of these ChIP-seq peaks were found in all four strains, the variability of CTCF's binding intensity in these locations could also be used to indicate the consistency of CTCF occupancy in a given locus. Binding intensity variance ranged from 1.4 to 28.5 in peaks found in all four strains, with a median of 10.9 and a mean of 11.1. This indicates a potentially interesting level of differential binding strength between strains. This then could be used to subset the data further, to ChIP-seq peaks with highly differential binding intensity. As a

temporary cutoff for exploration, a variance level of 10 or greater was used. This left 478 peaks, 84.9% found in all four strains.

At this point, we reassessed whether this level of stringency was needed in subsetting our data. Subsetting with all these conditions could allow us to point to precedent for our data, but it also would bias toward what is already available in the literature, namely data from B6 mESCs grown in 2i medium, and peaks that were strong enough to pick up by techniques that may be increasingly out of date.

Still, it was not immediately clear what metric should be used to show that CTCF binding in the DO founders is predictable by DO samples. One initial idea was to correlate CTCF binding with SNP and ATAC-seq presence or absence, in CTCF ChIP-seq peaks that overlap with DO ATAC-seq peaks. The theory in mapping a correlation to both SNP and ATAC-seq peak occupancy was to tie ChIP-seq results to both of the main factors of the interacting models, and show that the all three were linked, possibly even characterizing expression of individual genes as examples of putative CTCF presence or absence in our DO samples, and attempt to find causative SNPs. Each of these had to be subset to DO samples where the haplotype was exclusively drawn from the four founder strains that we had generated ChIP-seq data for.

We found a large proportion of CTCF binding in high correlation or anti-correlation with ATAC-seq and RNA-seq presence (Fig. 3.13). A number of these binding sites themselves candidate SNPs in them, which were manually examined to determine whether their haplotype split along lines that matched the CTCF ChIP-seq in the region.

This produced several results that showed promise, but it was not a systems-level or comprehensive analysis. The ATAC-SNP to ChIP-seq correlation was also difficult to clearly explain and were increasingly divorced from the underlying biology that we were trying to represent. We therefore stepped back and reassessed again. We wanted to identify what the relationship was between interacting regions in DO samples and DO founder ChIP-seq, specifically to the strain-specific patterning of CTCF binding. Therefore, we narrowed our focus to whether we could correlate DO genotype to CTCF binding in regions that overlapped with interacting ATAC-seq peaks.

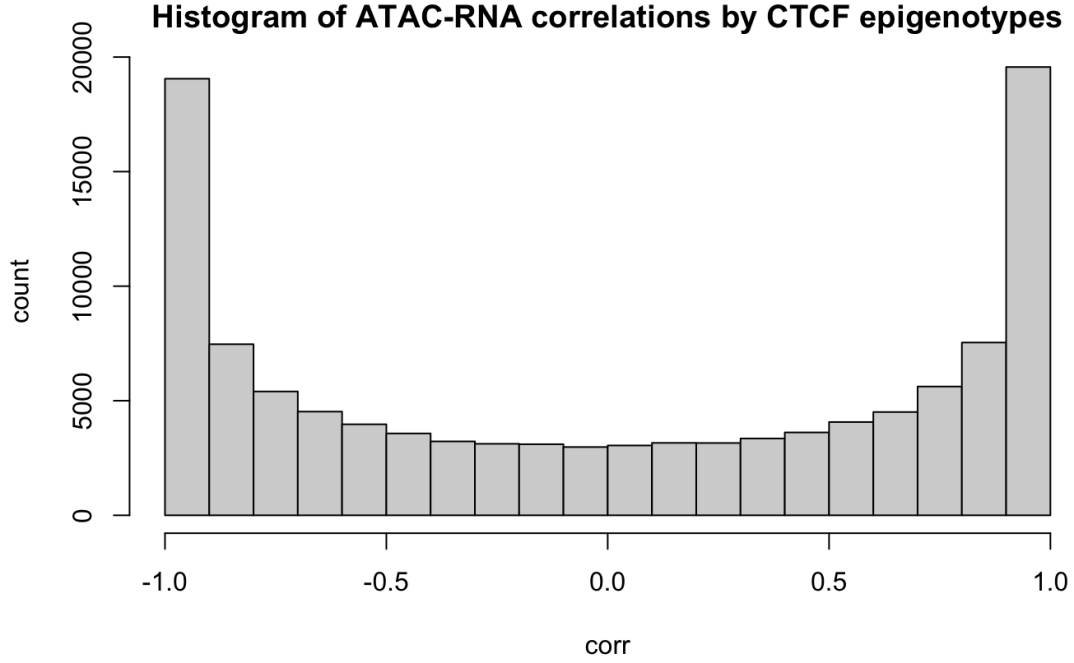


Figure 3.14. Histogram of correlation between ATAC-seq-RNA-seq presence with CTCF binding.

To approach this relationship between the two datasets, we had to first make sure that we could identify and align ChIP-seq results to CTCF binding motifs. The motif is 13-20 bp long, depending on how many highly variable bases are retained on the main binding sequence's flanks. The CTCF ChIP-seq peaks are significantly longer than this, and could theoretically include multiple CTCF binding sites. Therefore, we had to identify the CTCF binding sites that were bound in our data, and the location of SNPs within them.

To begin examining this relationship, we wanted to recreate the SNP density histogram in Fig. 3.10C, looking at SNPs in CTCF binding sites within our CTCF ChIP-seq results. Aligning the plot to the center of the ChIP-seq peak produced unexpectedly messy results, with a broad plateau dense SNPs before falling off at +/- 100 bp (Fig. 3.14).

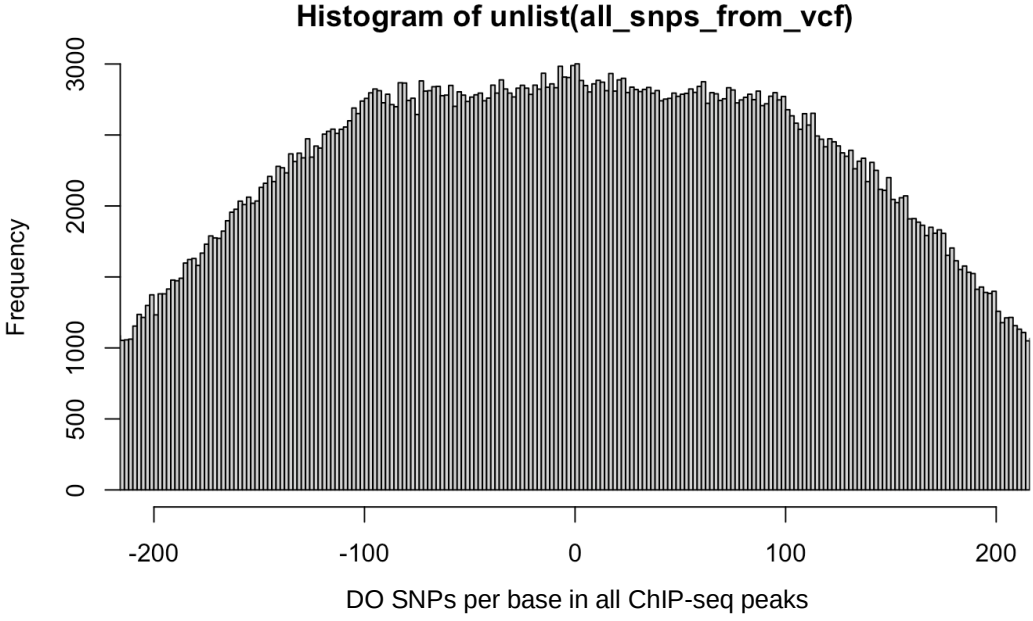


Figure 3.15. Histogram of CTCF binding site SNPs within ChIP-seq peaks, centered on the ChIP-seq peak center point.

This indicated that the CTCF binding sites were not actually centered within the ChIP-seq peaks. Therefore, we plotted the difference between the center of the ChIP-seq peak and the center of overlapping CTCF binding sites identified by FIMO. We found a relatively tight distribution of CTCF binding sites within the ChIP-seq peaks. The median distance from the center of the ChIP-seq peak to an overlapping CTCF binding site was zero, with half of all results lying between -30.5 to 32 bp. The other half included some highly eccentric outliers, ranging from -588 to 918 bp away from the peak center. This was puzzling, given the accuracy of the median result.

We theorized that this wide range of locations might be due to multiple CTCF binding sites identified by FIMO within the ChIP-seq peaks, some of which might not be bound. Another option was potential non-canonical binding sites, which FIMO would not capture. We therefore wanted to determine how many CTCF binding sites were appearing under these ChIP-seq peaks. This would also give us some clues as to whether these regions use adjustments in TAD binding as a regulatory mechanism—Following findings in developmental biology(Rodríguez-Carballo et al., 2017), we would expect to see multiple local CTCF binding sites in areas where TAD binding sites are adjusted to selectively expose local genes to different enhancers.

We found that in all ChIP-seq peaks, FIMO identified 57,150 CTCF binding sites in 96,083 ChIP-seq peaks, indicating 0.5950 canonical CTCF binding sites per CTCF ChIP-seq peak. When subset to the intersection with interacting ATAC-seq peaks, 9,831 binding sites were found in 7,803 peaks, or 1.260 binding sites per ChIP-seq peak.

This potentially indicated a number of explanations. The low proportion of binding sites per peak could indicate off-target antibody binding, despite comprehensive testing of available antibodies had been performed prior to sample processing. Processing of the raw data might also have skewed results toward more or fewer binding sites per peak. The latter seemed unlikely, because the initial “narrowPeak” files were not employed for this analysis, instead erring in favor of data derived from files that include the region of uncertainty around the narrowPeak results. These were further merged to combine overlapping peaks. This would bias toward picking up *more* local CTCF binding sites within a single merged peak. That could explain the greater number of CTCF binding

sites within the subset that interacts with ATAC-seq peaks, but that would only make sense if the whole dataset also contained greater than one binding site per peak.

This indicated that there was a true enrichment in the number of CTCF binding sites per ChIP-seq peak, if they overlapped with interacting ATAC-seq peaks. The low binding site per peak levels across the dataset remains unexplained. We theorize it could possibly come from non-canonical binding of CTCF. There is a similar protein, CTCF-like (*Ctcfl*), which theoretically could be binding to these cryptic regions identified by the ChIP-seq. However, Ctcfl is primarily active in the cytoplasm of spermatocytes, rather than in the nucleus of somatic cells. Thus, it does not seem to be the mechanism in play(Loukinov et al., 2002). *Ctcf* contains 11 zinc finger domains, which provides it with a high degree of flexibility in binding. The protein also forms complexes with multiple optional components and interacts with other DNA binding proteins(Wei et al., 2022; Weth & Renkawitz, 2011). While not an established role, some of these optional interactions could affect the binding affinity of CTCF.

We then investigated the density of CTCF binding sites per ChIP-seq peak in overlap with interacting ATAC-seq peaks. In this subset, 66% of ATAC-seq peaks with CTCF binding sites had one binding site, 24% had two binding sites, 7% had 3, 2% had 4 or more, and one peak had 14 binding sites within it. That peak contained five separate interacting ATAC-seq peaks.

This opened up an avenue of further exploration and interest. While genetic variants are the most classical way to differentiate between strains, the use of epigenotype has become more tenable as an alternative in this case. ATAC-seq peaks are not a binary

present/absent feature, but these ATAC-seq peaks already have beta coefficients that indicate their effects on gene expression. If a cluster of CTCF binding sites does have regulatory function to expose local genes to different enhancers, we would expect to see that ATAC-seq peaks would be associated with increase or decrease in a gene's expression, based on whether the DNA is open when the gene is exposed to enhancers by shifting TAD boundaries. This could not be recapitulated with our SNP beta effects except in very specific loci, due to linkage disequilibrium. This leaves ATAC-seq peak beta effects as a potential means to assay a CTCF binding site cluster for differential effects on gene expression.

We pulled out two CTCF cluster regions where ATAC-seq peaks overlap with CTCF binding sites, which allowed us to look at their effects on gene expression as a potential function of which CTCF binding sites they overlap with (Fig. 3.15). In one region with two ATAC-seq peaks overlapping with CTCF binding sites, the closer ATAC-seq peak had a mostly positive effect on the expression of local genes, while the further ATAC-seq peak had a highly negative effect on gene expression. In the other region plotted, three ATAC-seq peaks overlap with CTCF binding sites that are active in the DO founder samples. Two of the ATAC-seq peaks have minimal difference between their effects, while the remaining one has a significantly lower effect on gene expression. These differential effects of ATAC-seq peaks overlapping with CTCF binding sites are potential indicators of the direct involvement of TAD boundaries in regulation of local gene expression.

These results are very tantalizing, but limited in scope. To confirm their effects, one would need to at least run RNA-seq on the DO founder samples as well, to confirm this particular hypothesis, possibly even employ locally targeted 3D chromatin assays to

ensure that CTCF isn't just binding, it truly is forming chromatin loops. Combined with the lack of systemic reach to this particular combination of data, this is an interesting and promising piece of data, but does not provide the comprehensive evidence our systems-level approach requires.

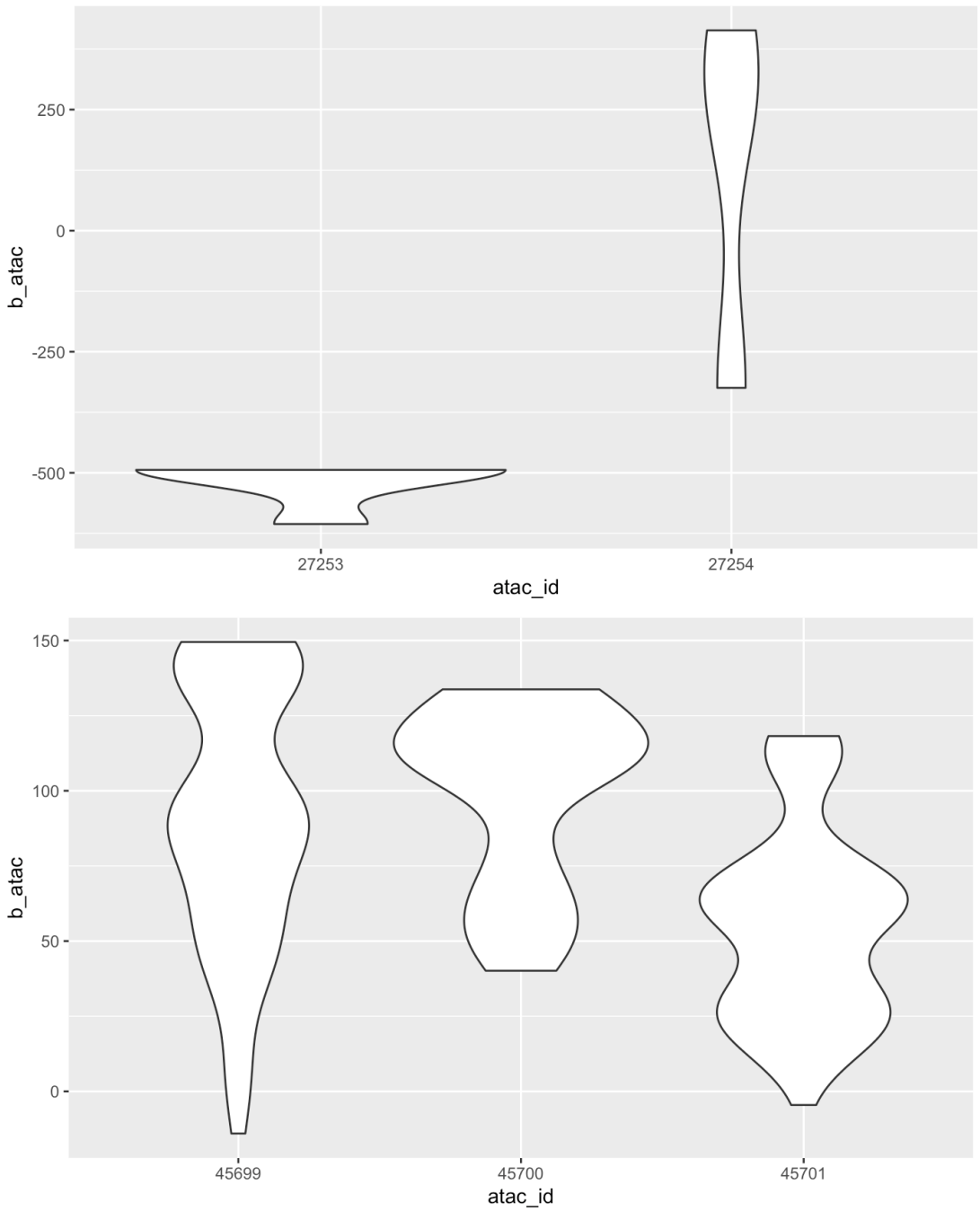


Figure 3.16 Violin plots of two CTCF binding regions A-B. ATAC-seq peak effects in ATAC-seq peaks that overlap CTCF binding sites in our ChIP-seq data in two different TADs.

One potential way to look at how our DO data informs and is informed by the CTCF ChIP-seq is to look at how often SNPs show up within the CTCF binding sites that overlap with ChIP-seq peaks, and whether that is correlated with the presence and sign of non-additive interactions. We could therefore determine whether SNPs are associated overall with increased or decreased presence of CTCF binding. We analyzed the data together to determine the overlap between these subsets. We determined that if a SNP is found in a given strain, a ChIP-seq peak will only be found in that strain 29.6% of the time. When subsetting to only those that overlap with negative effector ATAC-seq peaks, that drops to 25%. However, that sub-group is only composed of about a hundred loci, limiting confidence in whether there is an actual significant difference.

Despite this, considering this method of analysis brought us to a new idea that gave us direction toward a potential solution to how we could truly examine the relationship between observed effects in the DO, and CTCF binding in the DO founders. Previous analysis methods had been too complicated or locally-targeted. The most recent analysis method was both too unfocused to provide meaningful distinctions between groups, and too stringent in other ways to provide much data. Therefore, we looked at another way to describe the relationship between CTCF ChIP-seq and the DO mESCs.

If our DO data truly does imply genetically-determined differential CTCF binding, then we would expect to see strain-specific differences in CTCF peak occupancy. Consensus peaks were identified from replicates of each strain. Fifty-two percent of CTCF ChIP-seq peaks were shared across all four strains, with the remaining 48% found in three or fewer strains (Fig. 3.16A). The limited number of samples could potentially bias our results toward peaks being found in fewer strains. This could potentially be explored by

using peaks found in only one sample, but still produce a bias toward CAST and B6-unique peaks. Temporary use of peaks appearing in one sample for PWK and WSB only would not be a rigorous test, and bias the results to an unknown degree.

One potential alternative is to look at the range of binding intensities for consensus ChIP-seq peaks, found across all strains. ChIP-seq is a bulk analysis, so binding intensity can show the probability of finding a protein bound at a locus within the sample. If CTCF binding is disrupted by strain-specific genetic factors, then its binding intensity will vary. These are also less likely to be affected by the differing numbers of samples, as they all appear in two or more samples each.

Thus, we analyzed differences in CTCF binding intensity. A range of binding intensities were found at individual loci, with 13% having a variance in fold enrichment greater than 10. (Fig. 3.16B). This gave us confidence that there was truly differential binding of CTCF in our samples.

3.10 Non-Additive Interactions are Predictive of CTCF Binding Patterns

With our ChIP-seq data indicating the presence of strain-specific CTCF binding, we hypothesized this differential binding could be predicted from our regression models. We anticipated that inbred founder strain CTCF ChIP-seq would have more SNP effects at binding sites with open chromatin in non-additive interactions with local genotype, as

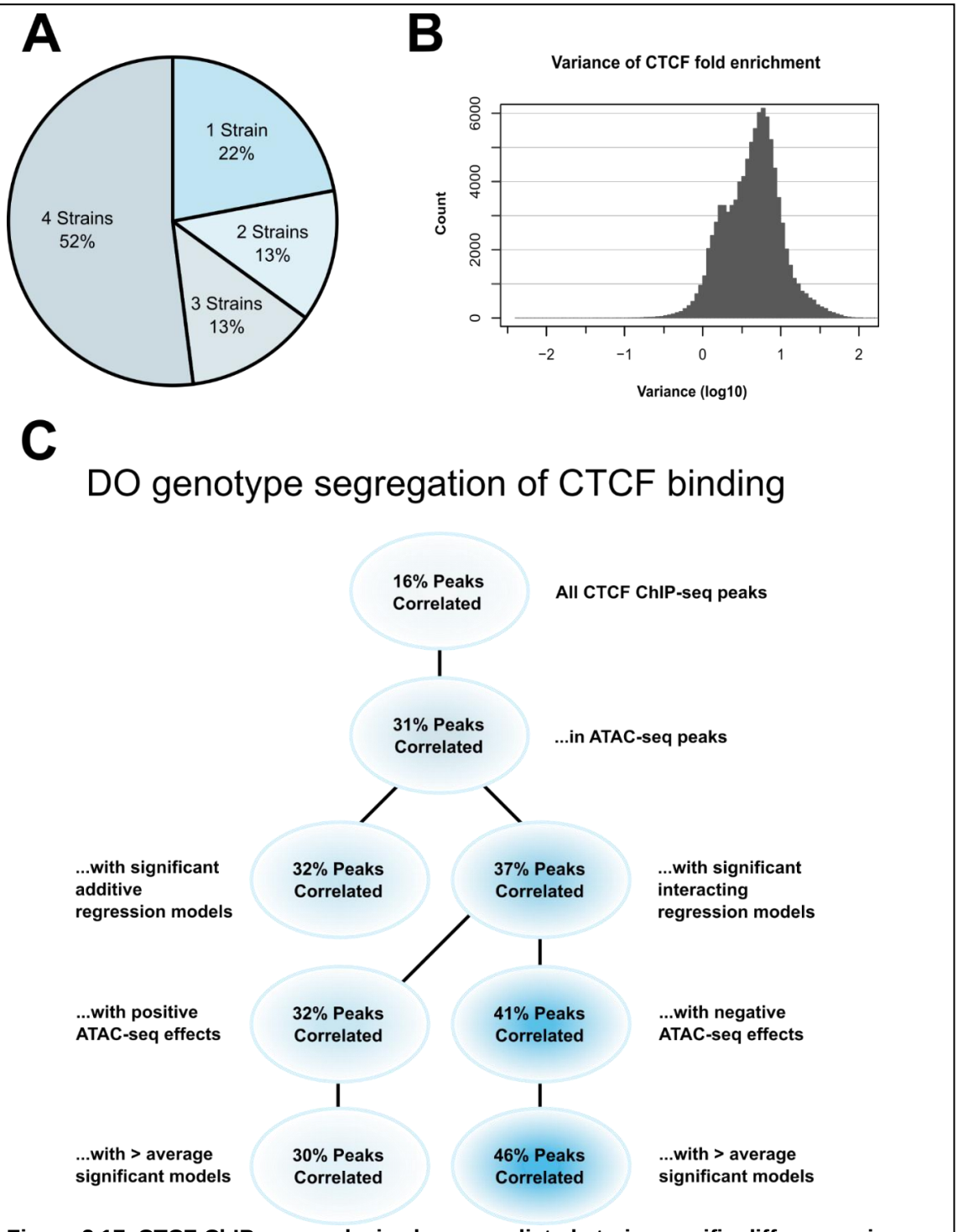


Figure 3.17. CTCF ChIP-seq analysis shows predicted strain-specific differences in binding intensity. A. Percentage of ChIP-seq peaks in surveyed strains. **B.** Variance (log10) in binding intensity fold enrichment for all ChIP-seq peaks. **C.** Percentage of significance in association between DO genotype at CTCF peaks and CTCF binding intensity on inbred ChIP-seq samples, in various subsets. Reprinted with permission from *Kuffler L, Fortin HJ, Skelly DA, Czechanski A, Munger SC., Reinholdt LG, Baker CL, Carter GW. (2023) Imputation of 3D genome structure by genetic-epigenetic interaction modeling in mice eLife 12:RP88222 https://doi.org/10.7554/eLife.88222.1.*

opposed to binding sites without interactions. Both interacting and non-interacting areas contain genetic variation and ATAC-seq peaks, but our models indicated interacting areas had more effects on local gene expression and genetic variability in CTCF binding sites. More specifically, we expected negative effector ATAC-seq peaks to have the greatest predictive power, as we believed their negative effects on local gene expression arise from strain-specific differences in CTCF binding and loop formation, whereas positive effects on gene expression may come from non-localized SNPs in other binding sites (Reproduced with permission from Kuffler et al., 2023).

To test these hypotheses, we retrieved CTCF ChIP-seq locations and identified biallelic SNPs found in the four inbred DO founder strains we used for our ChIP-seq experiments. After some consideration of potentially using hierarchical clustering methods to preserve the relatedness of CTCF ChIP-seq peaks across samples, the correlation of SNP genotype to CTCF binding site occupancy in the DO founders was assayed using an unpaired sample Student's t-test (Fig. 3.16C). This method was intended to determine if ChIP-seq peak intensity at a given CTCF binding site was split by biallelic SNP genotype. This would allow us to cast a wide net to examine these phenomena, removing levels of pre-processing that potentially limited our predictive power, most notably how we'd been folding all CTCF ChIP-seq samples from a given strain into one consensus result that might not accurately reflect its constituents. This allowed us to be sensitive to how individual samples behaved at a given ChIP-seq peak. If a SNP has a tuning effect on CTCF occupancy, then one might see peaks of less consistent strength on a sample-by-sample basis, beyond what might be expected by technical batch effects (Adapted with permission from Kuffler et al., 2023, expanded to provide context).

The proportion of results below $p < 0.05$ in all CTCF ChIP-seq peaks was 16%, establishing a baseline of predictive power for SNPs genome-wide. Subsetting to those CTCF binding sites found in ATAC-seq peaks increased the proportion of significant correlation to 31%. This was expected, as open chromatin in the vicinity of CTCF binding sites is associated with binding site occupancy(Li et al., 2019; Oomen et al., 2019) (Reproduced with permission from Kuffler et al., 2023).

To test whether our theory that non-additive interacting regression models held greater predictive power of CTCF binding intensity, we subset CTCF binding sites in ATAC-seq peaks with significant effects on local gene expression. Binding sites associated with non-additive interacting models had 37% correlation. This outperformed additive models, which had 32% correlation (Reproduced with permission from Kuffler et al., 2023).

ATAC-seq peaks can be associated with significant effects on multiple genes, potentially in combination with multiple SNPs, resulting in some ATAC-seq peaks associating with multiple significant regression models. Interestingly, we found that among ATAC-seq peaks over CTCF binding sites, those associated with additive models were a subset of ATAC-seq peaks with non-additive interacting models. This means that any ATAC-seq peaks that had effects on local gene expression and were localized to CTCF binding sites *always* had interaction effects with the local genotype, and these models are more predictive of CTCF binding intensity (Reproduced with permission from Kuffler et al., 2023).

These results matched our expectations. ATAC-seq peaks co-localized with a candidate polymorphism affecting CTCF binding would be more likely to affect gene expression in a non-additive fashion, as the polymorphism would only affect CTCF binding based on chromatin state at the binding site or any nearby priming factors. This is in contrast to additive models, where genetic effects and ATAC-seq effects are predicted to be independent of each other and are thus less likely to be co-localized (Reproduced with permission from Kuffler et al., 2023).

Subsetting those ATAC-seq peaks that were negative effectors in interacting regression models produced a 41% correlation. This outperformed ATAC-seq peaks with a positive effect, which had 32% correlation. This was in line with our previously stated predictions and suggested that non-additive interactions can be used to evaluate and predict local 3D chromatin structure (Reproduced with permission from Kuffler et al., 2023).

The fact that there was a lower proportion of SNPs correlating with ChIP-seq peaks in regions with positive effects on gene expression was also in line with predictions. Positive effects are what we would expect if a TAD is successfully forming, opening chromatin, and allowing enhancers to activate. SNPs that do not significantly disrupt a TAD binding site are going to be more numerous than SNPs that actively strengthen a TAD binding site's ability to bind CTCF, therefore we cannot expect to see as strong a correlation in these regions.

Lastly, we aimed to determine if areas with the greatest number of DO regression models supported our hypothesis, or if a larger amount of chromatin openness and segregating haplotypes would result in a leveling effect on DO genotype's predictive

power on CTCF binding. We subset results to regions associated with a greater than average number of regression models. We found a wider correlation gap in this subset, with 46% correlation in negative effectors versus 30% in positive effectors. This suggests that our regression model provided detailed data in dense regions of genetic-epigenetic interaction, thus reinforcing our previous findings (Adapted with permission from Kuffler et al., 2023 with edits to word choice).

3.11 Attribution

Analyses in chapter 3.1.2-3.8 were performed solely by the author. 3.1.1 was the work of Skelly et al. as previously described. DO founder mESC generation in 3.9 was performed by the lab of Laura Reinholdt. CTCF ChIP-seq generation and initial QC analysis was performed by the lab of Christopher Baker, with antibody testing and CTCF ChIP-seq preparation performed by Haley Atwell. Code for processing CTCF ChIP-seq data was adapted from Sanderson et al., with advice from Michael Lloyd.

Chapter 4: Discussion

Integrating analysis of multi-omics data in a genetically diverse population allows for greater specificity and modeling of complex regulatory interactions. Populations that segregate natural genetic variation provide dense sequence perturbations from which gene expression variance can be modeled**.** This contrasts with experimental designs that rely on a limited number of engineered perturbations in isogenic cells or animals. By collecting genetic, epigenetic, and transcriptomic data from the same cellular panel, models can be inferred without confounding factors such as different experimental protocols and environments. Our approach to integrating genetic and functional genomics data from 176 Diversity Outbred mouse embryonic stem cell lines allowed us to systematically probe how genetic and epigenetic variation interact to jointly influence local gene expression (Reproduced with permission from Kuffler et al., 2023).

4.1 Overview of Findings

We integrated ATAC-seq peaks and SNPs via an interactive regression model of their effects on gene expression to determine how often these factors are independent of each other, and how often they interact. From this, we were able to infer the wide-scale presence of local interactions between these regulatory mechanisms. This method is well-suited for outbred populations, which are not compatible with methods used for analyzing inbred cohorts, such as differential expression or variance between samples. While similar interactions have been observed in isolated contexts and in co-localized SNPs and ATAC-seq peaks(Krijger & De Laat, 2016; Kumasaka et al., 2016), this marks the first time that the phenomenon has been observed genome-wide, with the genetic mapping resolution provided by outbred, heterozygous samples (Reproduced with permission from Kuffler et al., 2023).

From our analysis of genetic-epigenetic interactions we have put forth several key findings regarding the interactivity, structure, and function in the regulation of gene expression. We discovered that patterns of genetic-epigenetic interaction reflect the structure of topologically associating domains. Our model inferred the presence of frequent interactions between genotypic variation and open chromatin, with a strong preference for coordinates internal to known TADs in mESCs. We demonstrated this inference in several ways, including the clustering of highly interacting areas of open chromatin within TADs, and clustering of interactions based on inter-TAD distance rather than linear DNA sequence. Interacting ATAC-seq peaks were found to cluster within gene bodies and annotated enhancers. These findings confirm that TADs generally define the local area for interactions. While open chromatin density is higher within TADs, we found that open chromatin density is less important for interactivity than TAD boundaries are. Furthermore, linear proximity cannot be used to identify association between genes and regulatory features. SNPs or areas of open chromatin that are segregated by a TAD boundary may be incapable of affecting expression of a nearby gene. Conversely, a distant SNP or ATAC peak may be placed near a gene within a TAD loop (Reproduced with permission from Kuffler et al., 2023).

Interaction effects were further classified based on the magnitude and direction of their effects on gene expression. These models contained an unexpectedly high proportion of increased ATAC-seq peak effects associated with reduced gene expression. Upon further investigation, these were found to be enriched for CTCF binding sites, which were further enriched for genetic variation in Diversity Outbred mice, particularly in the core sequence of CTCF binding sites (Fig. 3.10B-C). This indicates that these CTCF

binding sites may be differentially bound in different samples, resulting in downregulation of transcription, either by cutting off access to nearby enhancers or by abolishing TAD structure (Reproduced with permission from Kuffler et al., 2023).

We further analyzed the interplay between direction of effect for SNPs, ATAC-seq peaks, and their interactive components (see Appendix), allowing us to make hypotheses about the functional significance of various interactions. Models where all effects are positive or negative suggest the open chromatin region and genetic variant enhance each other's effectiveness in increasing or decreasing gene expression. These synergistic effects are indicative of two regulatory factors working together to produce a greater change in gene expression, beyond what either could produce independently. Functional redundancy or interference can be inferred from models where the SNP and ATAC-seq peak have a positive effect on gene expression and the interaction effect is negative, or vice versa. Redundancy is rather common in our genetic-epigenetic interactions, while synergistic effects are relatively rare. This appears to align with previous analyses of purely genetic interactions in other mouse crosses (A. L. Dixon et al., 2007; Tyler et al., 2016). Other models are more cryptic, but contain variation in beta coefficients associated with these models (Fig. 4.1), suggesting that functional subtypes that can be investigated within each group (See Appendix, section 5.3) Our approach can be directly extended to specific gene targets (Adapted with permission from Kuffler et al., 2023, expanded to provide further context).

Mediation analysis in the DO by Skelly et al. revealed several mESC genes that act as mediators of downstream gene networks(Skelly et al., 2020). Our analysis identified several genetic-epigenetic interactions in the area of one mediator gene, *Platr2*, and

provided a targeted list of SNPs and ATAC-seq peaks that may influence the gene's expression (Fig. 3.7B). ATAC-seq peaks were enriched for Smad3 binding sites, a

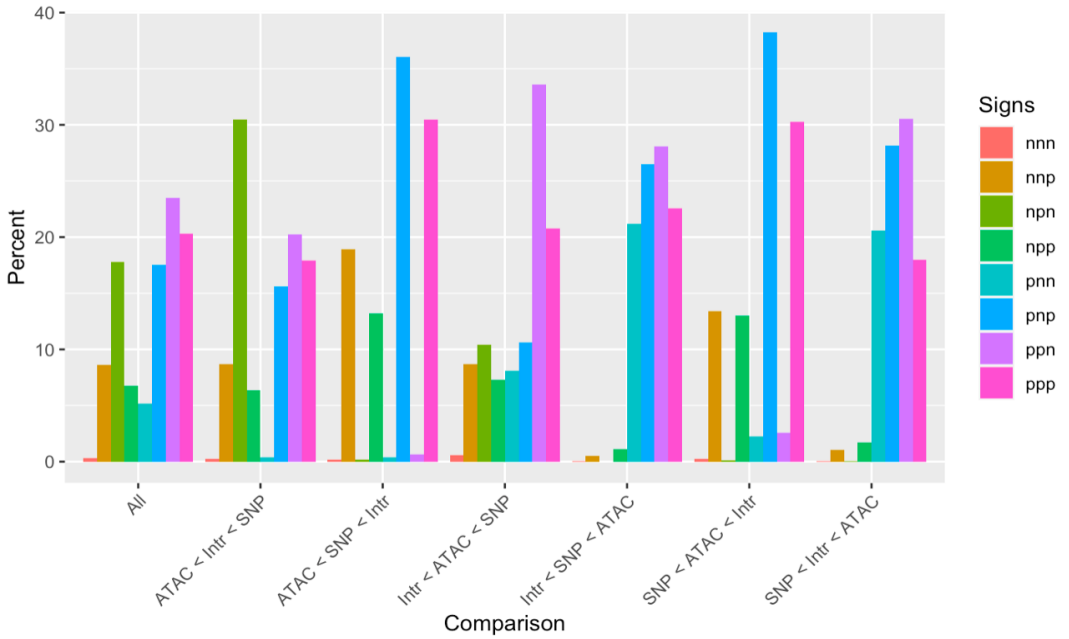


Figure 4.1. Relative effect magnitudes of all significant intra-TAD models. Split by effect signs. Model signs are listed in the order ATAC, SNP, Interaction, and positive (p) or negative (n). Reprinted with permission from *Kuffler L, Fortin HJ, Skelly DA, Czechanski A, Munger SC., Reinholdt LG, Baker CL, Carter GW. (2023) Imputation of 3D genome structure by genetic-epigenetic interaction modeling in mice eLife 12:RP88222 https://doi.org/10.7554/eLife.88222.1.*

component of the CTCF binding complex. This provides further regulatory information and potential targets for experimental manipulation (Adapted with permission from Kuffler et al., 2023, edited for clarity).

Through DNA motif analysis, we identified distal regulatory activity via our interaction modeling. This permits functional analysis of protein factors on gene activity that is otherwise undetected by this method, such as regulatory proteins with no local QTL. This means that while interaction models may achieve the best resolution in areas of low linkage disequilibrium, they can still be used to identify and infer regulatory action of key conserved proteins in *trans* (Reproduced with permission from Kuffler et al., 2023).

4.1.1 Overview of CTCF ChIP-seq Experiments

Genetically diverse populations in mice still contain significant linkage disequilibrium blocks which affect genetic resolution. This limited our ability to identify the causal SNPs for the purposes of this paper, but as demonstrated with our analysis of Platr2's local regulatory area, the variable genetic distance across the DO genome sometimes permits us sufficient resolution to analyze whether interacting elements are colocalized or not.

Our CTCF ChIP-seq experiments found strain-specific differences in CTCF binding, including clusters of CTCF binding sites. ATAC-seq peaks within these clusters showed differential effects on gene expression. We also found that previously identified CTCF binding sites in negative effector ATAC-seq peaks are more likely to contain SNPs that can predictably CTCF binding potential (Adapted with permission from Kuffler et al., 2023, edited for clarity).

4.2 Implications of Results and Future Directions

Our results have several implications for genetic analysis of gene expression in genetically diverse populations. Experiments performed in cell cultures or isogenic models often fail to produce replicable results in other tissues or human trials(Seyhan, 2019). Genetic-epigenetic interactions may underpin some of these failures, particularly those that have initially shown strong SNP effects(Mak et al., 2014). Thus, it is important to consider generating datasets with a combination of genetic mapping, gene expression, and appropriate epigenetic data, either on a local or genome-wide scale. At present, publicly available data that matches these criteria is not available. As research expands towards greater coverage in wild populations and humans, accessibility is likely to increase. Furthermore, ATAC-seq could be substituted for other experimental datasets, such as H3K4me3 ChIP-seq.

In future, our study would benefit from the generation of more precise mapping of active CTCF binding sites, across more genetic backgrounds. Public databases only include CTCF mESC data from two inbred strains, and the enrichment of CTCF binding sites in our interacting ATAC-seq peaks suggest that other strains and subspecies may have different active CTCF binding sites. Mouse strains have previously been shown to have different strengths in CTCF binding in specific regulatory loci, resulting in altered gene expression(Van Ruiten & Rowland, 2021). More comprehensive CTCF profiling in the DO founder strains would permit us to determine whether TAD boundary alteration underlies changes in chromatin accessibility in any of our interactions, whether there are interactions that were obscured in some samples by the blocking effect of a TAD boundary, and whether there are some TADs that are more tolerant of boundary shifts than others (Reproduced with permission from Kuffler et al., 2023).

Small changes in TAD boundary location are previously indicated to be involved in developmental and differentiation processes(Andrey et al., 2013; Li et al., 2019; Su et al., 2010), and other studies have shown that chromatin rearrangement that alters TAD boundaries is linked to cancer development(Aitken et al., 2018; Akdemir et al., 2020; J. R. Dixon et al., 2018). However, many publicly available datasets of CTCF binding activity have limited resolution and/or coverage. With CTCF binding sites scattered across the genome, there could be further subtle shifts in local regulatory areas that have previously gone unnoticed, particularly in developmental contexts or across different tissues.

Male and female samples were not separated during this analysis. Therefore, we have not assayed whether there are any sex effects among these non-additive interactions, or interactions that are only detectible when the samples are split by sex. Given that TADs were first identified in X-chromosome inactivation(Nora et al., 2012), it is almost inevitable that there would be some sex effects we did not see with pooled samples.

TAD locations and functions are known to be conserved between species(Krefting et al., 2018). Movement of TAD boundaries with profound regulatory effects have also been conserved in humans, mice, and chickens(Le Caignec et al., 2020; Rodríguez-Carballo et al., 2017; Yakushiji-Kaminatsui et al., 2018). These conserved mechanisms have been established in TADs that are relatively ubiquitous, or have easily assayable effects. We do not know if similar movement of TAD boundaries is used in active regulation on a broader scale. And we know even less about how the movements of TAD boundaries may affect genetic-epigenetic interactions in these areas. Regulatory regions are more highly variable and faster-evolving than genes themselves(Merkin et al., 2012; Shibata et al., 2012; Villar et al., 2015). We would expect enhancers to be at least as variable, but what might be conserved between species, if anything? What paralogs may develop? Are there patterns to paralog development? We do not have answers to these questions at this time.

It would be possible to produce an equivalent statistical inference of non-additive regulatory interactions with the use of other epigenetic factors. While ATAC-seq peaks have a well-defined role that can be assayed relatively easily, one could substitute this with histone modifications, methylation, or DNA binding of relevant proteins. The ability to localize an epigenetic signal to a relatively specific locus allowed us the chance to

characterize the patterning of these interactions in their 3D chromatin context(Kulakovskiy et al., 2018; Maurano et al., 2015; Shinkai et al., 2016).

While we hypothesize the potential function and role of different interaction subsets based on their effect sizes and signs, we did not focus on confirming these during the project thus far. We found that there were differences between the enrichment of CTCF binding sites under positive effector ATAC-seq peaks versus negative effector peaks (Fig. 3.10). We also found that negative effector peaks were more likely to contain SNPs that had predictive power for CTCF binding in DO founder mESCs, indicating likely colocalization between interacting factors in these regions. Thus, it seems likely that we would find patterns of different mechanisms underlying different subsets of interacting models.

Analyzing these models would require some level of predictive SNP effect analysis, as we are limited by linkage disequilibrium in our ability to determine whether any given SNP is the driver of a genetic effect in an interaction. It would be easiest to start in those regions with extremely small linkage disequilibrium blocks, where the number of potential SNPs is likely to be more manageable. Confirmation by reciprocal biallelic SNP introduction between two strains of DO founder mESCs with locally targeted ATAC-seq and RNA-seq would be a good way to confirm such an analysis.

As we noted in Fig. 3.1D-E, a portion of the SNPs and ATAC-seq peaks that are involved in significant regression models are only observed in interacting models. While those that fall exclusively or inclusively in single-term or additive regression models may have been previously identified for their regulatory role, it is likely that the more

complicated effects of the exclusively interacting regulatory factors have not been characterized. Exploratory analysis might yield interesting results for these factors. Some of these may be colocalized and thus interacting by default, but others may represent attenuating or tuning factors on local regulation that are previously undescribed.

4.2.1 TAD Profiling

Despite all we were able to draw from publicly available Hi-C derived TAD boundaries, the resolution on the boundaries was relatively poor. With 40kb windows of TAD boundary prediction and CTCF binding sites appearing prolifically across the genome, it would be very hard to determine which binding sites were in use from this data. Previous publications and our own ChIP-seq experiments have shown that there can be differential CTCF binding based on genetic variants or different time points, and these can be important for orderly embryonic development. Hi-C results have become more precise since the generation of the data set we accessed, and there are now other options available. ChIP-PET protocols for CTCF have been established(Handoko et al., 2011), though they are still a difficult process that requires specialist training. Another option could be to combine Hi-C with CTCF ChIP-seq, which would allow identification of active CTCF binding sites that overlap with Hi-C signals. This could provide both TAD and sub-TAD resolution, something we were not able to differentiate in our ChIP-seq experiment.

And indeed, these differences in CTCF function can be very important in defining and shaping local regulatory areas. Our ChIP-seq analysis did not distinguish the difference between TAD versus sub-TAD loops, and between more transiently binding CTCF

versus more long-lasting. While some of this could be inferred from ChIP-seq signal strength, another option for more precise experimental determination of CTCF function would be to perform an RNAi knockdown or similar assay(Khoury et al., 2020), combined with ChIP-seq. Doing so would allow one to assay how long it takes for bound CTCF to detach from the DNA, giving a measure of how transient its regulatory role is in normal cell function. While such experiments have been done with CTCF, there has been no known effort to compare whether the patterns of CTCF binding longevity are universal, or whether they are affected by genotype or cell type. Given the sometimes extreme fold change in CTCF ChIP-seq enrichment we observed across our DO founder samples, we would expect to see some differential length of occupancy in some CTCF binding sites that are *technically* ubiquitous across these strains.

4.2.2 Effects of Local Sequence and Interaction With ATAC-seq

One aspect of potential analysis that we spent little time on is how the TAD-dependent, gene-centric distribution of interacting elements overlaid with local regulatory features in the DNA sequence. We identified the density of interacting ATAC-seq peaks overlapping with enhancer and gene features, but we went no deeper with this analysis (Table 5.1 and 5.2). While ATAC-seq peak locations have been surveyed before, we have demonstrated that interacting ATAC-seq peaks are more likely to contribute to meaningful regulation of gene expression, and can be analyzed in a more nuanced fashion. While individual genes are likely to vary widely in the mechanisms underpinning the effects of their interacting ATAC-seq peaks, there may also be more patterns of particular systems that are preferentially involved in these interactions.

We have potential explanations for why the percentage of ATAC-seq peaks involved in interactions is depressed around the gene itself (Fig. 3.7E), but we do not have a complete sense of why this is. This could potentially be accomplished without further laboratory experiments, particularly because this analysis only considered interacting ATAC-seq peaks. Comparing the distribution of ATAC-seq peaks that fall in additive or single-term models, we could potentially see different patterns that would elucidate more on the matter. Combined with further analysis of whether those ATAC-seq peaks fall in different distributions around enhancers, promoters, exons and introns, it could be possible to identify the likely cause of the distribution seen in Fig. 3.7E. We would expect to see a different distribution, based on the fact that ATAC-seq peaks that overlap with a gene may be more closely related to ongoing gene expression itself, and thus not require an interaction term to describe their behavior. However, it may be more difficult to explain why there is symmetry in the fluctuations in Fig. 3.7E past +/- 500kb. Again, this could be explained by looking at other regression models these ATAC-seq peaks could feature in, but the decreasing number of TADs long enough to contribute to the results limits the ability to draw conclusions from the data.

We performed an analysis of motifs within interacting ATAC-seq peaks, which revealed important overlaps with CTCF binding sites (Fig. 3.10). However, we did not deeply investigate the other binding sites identified via these analyses. Given the fact that these DO mESCs were grown in a 1i medium and stem cells of different strain backgrounds have different tendencies toward differentiation in culture, there should be differential developmental pathways activated in these samples, by different DNA binding proteins. During the proposal for this project, we performed a basic transcriptomic analysis that identified some differential expression of developmental markers, but this does not tell

us where they are acting, how they may be affecting gene expression in trans, and how they are interacting with local genetic variation to do so. Many core regulators of development are protected from recombination and sequence variation. Instead of assaying their local regulatory environment through our interaction analysis, we may be able to better characterize what they do once they're at work. Single-cell sequencing approaches could help identify cell populations and facilitate these analyses.

4.2.3 CTCF Binding Clusters and Founder RNA-seq

The clusters of active CTCF binding sites we identified in our ChIP-seq analysis also deserve further examination. We have seen from previous literature that shifts in individual TAD boundaries can be vital for tissue differentiation during development. With the ChIP-seq drawn from DO mESCs that are known to react differently to cell culture conditions(Skelly et al., 2020), it is reasonable to assume that there are distinct cell colony morphology phenotypes that differentiated these samples. Furthermore, our analysis limited the number of closely-spaced CTCF binding sites would be detected as differential, as we merged overlapping ChIP-seq peaks. Should we have more time to disambiguate and analyze these regions, we could examine the nuances of CTCF binding site clusters, functional enrichment of associated local genes, whether the patterns of their activation are cell type-specific, and whether we can detect any signal of this in our DO mESC data. Another option for confirmation would be to perform RNA-seq on the founder mESCs. Samples were frozen and are being held in case this analysis be needed, thus it would be relatively simple and inexpensive to get further confirmation.

Analyzing RNA-seq data in the DO founders would also allow us to compare to our DO mESC RNA-seq, allowing for a means to further confirm our interaction models. While

we would not have access to founder ATAC-seq data, we would know the genotype and expression levels for the founder strains. This could be used to predict a dividing line: genes whose expression could be primarily explained by genetic effects alone, versus those genes that could primarily be explained by interactions. One would expect genotype to be more predictive in the former case over the latter.

This could potentially be handled in a very granular fashion, observing the effects of local founder contributions to a gene's expression in the DO, reconstructing what the gene expression levels *should* be in a founder strain. This could then be compared to the actual result. At the same time, unless ATAC-seq data is also generated, one could not as strongly predict the results in regions where models including ATAC-seq effects were common. ATAC-seq peaks can be correlated to genotype(Waddington, 2012), but they are also linked to cell fate and other dynamic processes.

4.2.4 TAD Cell-Type Specificity and Disease

On this note of cell fate, it would be informative to learn which genetic-epigenetic interactions remain consistent across cell types and model systems. Those that do would likely be tied to core regulatory processes such as cell cycle, and could serve to better identify the regulatory mechanisms that underpin why a cell enters the cycle and why it exits. However, those interactions that are cell type-specific would be extremely useful to understand. The regulatory landscape of any cell type contains vital aspects of its identity, and much of it is still opaque to researchers. However, evidence suggests that TAD boundaries and their manipulation are important in defining a number of cell types(Barrington et al., 2019; Rodríguez-Carballo et al., 2017).

Disorders and diseases that produce aberrant cell types such as foam cells associated with plaque formation(Maguire et al., 2019), symptoms of diabetes(Ding et al., 2006), and cancers are all associated with regulatory dysfunction(Aitken et al., 2018; J. R. Dixon et al., 2018; Taberlay et al., 2016). It is unclear whether they would follow interaction models seen in healthy cells and simply display different input variables, or whether they would establish entirely new interaction models.

4.2.5 Other Interacting Elements And 3D Structures

We also believe it would be informative to examine interactions between different classes of epigenetic regulators, particularly if one wished to examine genes that are protected from recombination and genetic variation. In mESCs, our ability to resolve regression models around may core developmental regulators was limited, due to this protected status. The activity of these genes are not being adjusted by genetic variation, but they do still have to change their expression level to pass through differentiation Directly and comprehensively assaying interactions between epigenetic regulatory factors could be key to understanding their behavior, and the broader context in which their regulatory mechanisms operate.

Other 3D chromatin structures may also be of interest. We exclusively examined local chromatin structure in the form of TADs, but much of the transcription of vital proteins takes place within the structure of the nucleolus, or at the nuclear pore complex. This provides a very different transcriptional environment for these genes. It is unclear how this affects TAD dynamics and the refresh rate of CTCF binding, and we do not know how it may shape local genetic-epigenetic interactions. Are certain subtypes of interactions favored in these regions? Are they more or less prolific than in other

regions? Are there more or fewer models of any kind, when factors not assayed in our data may be influencing gene expression? These are all questions that could be explored if broader chromatin structure were taken into account in future experiments.

4.2.6 TADs as Active Regulators

The role of TADs in responses to environmental challenge in terminally differentiated cells has not been explored. TAD boundaries are refreshed relatively slowly, on the order of hours or days rather than the immediate, acute responses that are tracked in many environmental challenge studies(Khoury et al., 2020)**.** On the more long-term and permanent level of response, many chronic diseases have environmental contributions and triggers, including diabetes, obesity, asthma, and chronic effects from environmental toxicants. These would be particularly suitable for studying changes in 3D chromatin structure, and also how that may effect local regulatory interactions.

While we know TAD boundary alterations can result in differentiation into new cell types, it is not clear what mechanism might block them from *returning* to their previous configurations, should environmental pressures upon the cell be transient. Perhaps TADs could also be capable of a certain level of dynamic response, similar to the more transient sub-TAD loops.

This may be a more long-term goal, as a suitable cell model and challenge would have to be identified, as well as the correct assay to interrogate TAD formation. Hi-C the most well-established, possibly combined with CTCF ChIP-seq or ChIA-PET. On the other hand, monitoring cells over time could call for live imaging of chromatin contacts(Brandão et al., 2021). While the resolution would not be as good, continuous

monitoring of chromatin conformation could provide excellent support and context for any particular region identified as a strong candidate.

Developmental diseases are also a field where 3D chromatin conformation has not been given much consideration. This is despite the proven disorders of development that have been demonstrated when TAD boundary control of regulation is disrupted(Le Caignec et al., 2020; Rodríguez-Carballo et al., 2017). These TAD boundaries are dynamic and important regulatory factors that can affect gross morphology of a developing organism, but they have received only limited study thus far. The literature seems to indicate that most studies tend to acknowledge the potential of regulatory disruption of 3D chromatin conformation only when CTCF binding sites themselves are disrupted by mutation. This is the proximal and most pressing cause of these disorders, but it also means that 3D chromatin conformation is often neglected in this context.

4.2.7 Computational and Analytical Methods

It is worth noting that our analysis of non-additive genetic-epigenetic interactions used a fundamentally simple method. The novelty in our findings was not due to our methods, as regression modeling is an extremely simple technique. This should indicate to the reader that the field of study for this topic is wide open for further advancement, but should potentially be approached with a certain level of conservative methodology. A new subject does not always call for new techniques. It may call for simple ones that are easier to interpret, rather than introducing complications into the analysis of an uncharacterized system. This is, in effect, an exploratory project. Once a baseline is established, more advanced techniques become appropriate.

At this point, more advanced techniques could be applied to identifying patterns within the regression models themselves, rather than necessarily looking for a more complex model of interaction. However, we do see directions in which such analyses could go. One early aim of this project was to continue on from our establishment of non-additive regulatory interactions as a genome-wide local phenomenon, and then attempt to identify its distal effects. This ultimately was not pursued, though it was attempted. The complexity of how one would computational define distal regulation through an interaction proved difficult to tackle.

Out-of-the-box network analysis methods at the time of the project were poorly documented and poorly maintained. Hierarchical Hotnet was initially selected as an analysis method, but its dependencies were no longer fully functional. Hotnet also expected potential pathways to be built into the graph from the start. Running exploratory analysis with all potential gene-gene interactions represented was found to be computationally prohibitive. Therefore, we would have to use pre-existing networks. In genetics, this mostly takes the form of protein-protein or gene-gene interaction networks, which have been identified through experimental techniques. But these publicly available graphs are not based on the cell types and experimental conditions used in our analysis. This would limit our ability to do more than confirm already detected associations between genes, rather than predict novel, distal interaction effects.

The shape of our data also caused complications. While one can think of an interaction as a single event, it is composed of multiple factors, which may or may not be present in a given sample. Each gene can be affected by multiple interactions, which can often co-

exist within a sample. How then should one represent the influence of these interactions on distal genes?

This is the problem we encountered when we attempted to determine a suitable method for identifying distal effects of interactions. Traditional networking models usually work on an edge and node structure, which expects that an edge connects only two nodes(Trudeau & Trudeau, 1993). But with each sample having a freely varying set of genotype, ATAC-seq peak magnitude, and RNA-seq score, how then can a traditional network approach represent this?

We explored the potential use of hypergraphs, wherein multiple nodes can be joined by a single edge(Haussler & Welzl, 1987). Theoretically this could be done in a directional manner, testing each gene's mediation of all other's expression, but this still does not represent the interaction between SNP and ATAC-seq peak as we have previously defined it. A method like this might be able to determine if co-occurrence was linked to downstream gene expression, but not how it does so. Additive models might actually be the only ones detectable via this method. Interaction terms could be added as a sort of filtering node that modifies the SNP and ATAC-seq peak effects, but this continues to layer on more complicated structure and data types into a single system that may not be fit for purpose. However, we were not able to budget enough time to approach this method more practically.

Network analysis could still be leveraged for distal interaction effects. With the heavy focus on computational neural network development in the past five years, techniques carried over from machine learning could be leveraged to pursue this. In the meantime,

one potential way to search for interaction effects could simply be to employ the regression modeling technique we already used: Subset to those genotypes and chromatin effects that were found to be interacting to affect local gene expression, and then set the dependent variable to a *distal* gene's expression. Any significant results could potentially be the result of indirect regulation via local genes. However, steps would need to be taken to identify whether specific ATAC-seq peaks in particular were truly involved in these interactions, or whether their apparent effects were due to co-activation of a binding site local to the distal gene.

In any case, we unfortunately did not have the time to pursue this line of inquiry. We believe it is an important step to take. Local gene regulation does not directly give rise to phenotype, it is mediated through cascading effects through gene networks, protein-protein interactions, and a host of other cellular processes. Our analysis has been able to better define the complex space of local regulatory interaction, but more of that complexity needs to be carried forward to understand biological processes on even a cell-wide scale, let alone tissue- or organism-wide phenotype.

Returning to the subject of 3D chromatin organization, it should also be noted that one technique that was extremely helpful during this project had to do with how the data was visualized. The standard method of representing 3D chromatin contacts privileges the linear genome sequence—in either heat map or contact loop graph forms, the DNA is arranged in a straight line, which preserves linear bp distance between genome features. While this is useful to visualize the sometimes complex and overlapping or nesting structures seen in 3D chromatin structure, it also obscures the basic unit of chromatin organization that we were most interested in for this project. TADs are

rendered quite abstract by these graphs. In the process, one can obscure the continuity of local regulatory areas across TAD boundaries, or the physical proximity between a gene and an enhancer.

A deliberate decision was made to avoid these graphs in the course of this project, which allowed us to examine these 3D structures with some sense of the physical proximity and continuity of these structures. When we rendered TADs in a linear format, we did so around a different organizing principle: we centered the gene in the TAD, regardless of whether this created a discontinuity in the chromosome sequence. This allowed us to not only view the 3D context of the gene and its local regulatory area, but to also focus on this most functional unit of the genome. In future, we would like to incorporate sub-TAD looping structures into our visual understanding of these regions. Just as visualization of 3D protein structures can help researchers understand their method of function, so too might more precise mapping of 3D genome structure give us a greater sense of regulatory action, even possibly the dynamism with which these processes are carried out.

While no computational tools are yet available, it would be useful to expand such modeling of TAD and sub-TAD structure to visualization of 3D genome structure in motion: what does a particular TAD look like when gene transcription is being promoted and initiated? When it's being repressed? Does it perhaps have multiple ways to reach functionality, and how physically stable are they? If there are multiple active sites (i.e. genes), are they all activated together, or are there particular configurations used to activate individual ones? All these are questions we often ask about proteins or non-coding RNA, but we rarely do the same for chromatin, despite knowing its physical

dynamism as a complex, organized structure, with set binding sites and predictable points of self-contact.

4.3 Summation

Overall, this study demonstrates that genetic-epigenetic interaction analysis can reveal 3D genome structure through the positioning of interacting genome features. We see in this how genetics and genome structure can inform each other. These findings imply that including TAD boundaries and TAD loops in analyzing genomic features affecting gene expression, such as chromatin states and genetic variants, can maximize and contextualize results (Reproduced with permission from Kuffler et al., 2023).

4.4 Attribution

Figures presented in Chapter 4 were created solely by the author. Analyses were conducted by the author. Consultation on hypergraph methods was done with Matt Mahoney.

Chapter 5: Appendix

5.1 Breakdown of Interacting ATAC-seq Peak Locations Relative to Gene Features

Chr	Count	In Transcript	%	In Intron	%	Before Coding	%	In Exon	%
chr_1	1898	1231	64.86	1161	61.17	163	8.59	553	29.14
chr_10	1210	842	69.59	803	66.36	157	12.98	460	38.02
chr_11	1770	1263	71.36	1170	66.10	285	16.10	819	46.27
chr_12	721	491	68.10	465	64.49	87	12.07	244	33.84
chr_13	1101	626	56.86	594	53.95	88	7.99	281	25.52
chr_14	1104	706	63.95	664	60.14	87	7.88	304	27.54
chr_15	795	536	67.42	499	62.77	85	10.69	288	36.23
chr_16	584	346	59.25	329	56.34	57	9.76	180	30.82
chr_17	1852	1191	64.31	1139	61.50	196	10.58	592	31.97
chr_18	627	425	67.78	412	65.71	59	9.41	207	33.01
chr_19	660	476	72.12	459	69.55	103	15.61	263	39.85
chr_2	2604	1706	65.51	1628	62.52	250	9.60	778	29.88
chr_3	1425	915	64.21	850	59.65	173	12.14	494	34.67
chr_4	1567	965	61.58	900	57.43	130	8.30	462	29.48
chr_5	1886	1249	66.22	1184	62.78	202	10.71	618	32.77
chr_6	1112	691	62.14	649	58.36	107	9.62	339	30.49
chr_7	2039	1390	68.17	1303	63.90	274	13.44	864	42.37
chr_8	1524	916	60.10	868	56.96	148	9.71	464	30.45
chr_9	836	569	68.06	530	63.40	116	13.88	309	36.96
chr_X	601	399	66.39	372	61.90	97	16.14	282	46.92
Total	25916								

Table 5.1. Interacting ATAC-seq peak locations found in gene features
Reprinted with permission from *Kuffler L, Fortin HJ, Skelly DA, Czechanski A, Munger SC., Reinholdt LG, Baker CL, Carter GW. (2023) Imputation of 3D genome structure by genetic-epigenetic interaction modeling in mice eLife 12:RP88222 https://doi.org/10.7554/eLife.88222.1.*

Chr	After Coding	%	Enhancer	%	Exon1	%	Single Exon	%	outside	%
chr_1	350	18.44	113	5.95	233	12.28	14	0.74	637	33.56
chr_10	240	19.83	69	5.70	207	17.11	11	0.91	354	29.26
chr_11	389	21.98	106	5.99	369	20.85	28	1.58	479	27.06
chr_12	119	16.50	27	3.74	120	16.65	8	1.11	229	31.76
chr_13	166	15.08	43	3.91	121	10.99	11	1.00	459	41.69
chr_14	179	16.21	28	2.54	131	11.87	10	0.91	390	35.33
chr_15	164	20.63	49	6.16	126	15.85	9	1.13	245	30.82
chr_16	102	17.47	124	21.23	73	12.50	9	1.54	200	34.25
chr_17	301	16.25	151	8.15	258	13.93	13	0.70	619	33.42
chr_18	74	11.80	30	4.78	106	16.91	30	4.78	195	31.10
chr_19	165	25.00	70	10.61	116	17.58	11	1.67	175	26.52
chr_2	461	17.70	111	4.26	342	13.13	24	0.92	870	33.41
chr_3	287	20.14	71	4.98	242	16.98	26	1.82	492	34.53
chr_4	226	14.42	126	8.04	186	11.87	22	1.40	546	34.84
chr_5	338	17.92	63	3.34	262	13.89	12	0.64	623	33.03
chr_6	209	18.79	44	3.96	154	13.85	11	0.99	402	36.15
chr_7	454	22.27	60	2.94	382	18.73	25	1.25	638	31.29

Table 5.2. Interacting ATAC-seq peak locations found in gene features and elsewhere. Reprinted with permission from *Kuffler L, Fortin HJ, Skelly DA, Czechanski A, Munger SC., Reinholdt LG, Baker CL, Carter GW. (2023) Imputation of 3D genome structure by genetic-epigenetic interaction modeling in mice eLife 12:RP88222 https://doi.org/10.7554/eLife.88222.1.*

5.2 Breakdown of Regression Models

This section is devoted to a comprehensive breakdown of interacting models, as subdivided by their effect term sizes and magnitudes. As each of these subtypes indicates potential functional distinctions between models, providing predictions for what they may mean and where to expect the genetic contribution to come from can help in mapping and identifying functional features for those interested in examining individual interactions in-depth.

Abbreviations are used for ATAC-seq (A), SNP (S) and Interaction (I) terms. Each section is organized first by effect sign, and then split by the magnitude of each.

5.3 SNP + ATAC +, Interaction +

Number of models: 67735

% of all models: 16.44%

% relative magnitudes of each term (individual terms listed largest to smallest):

49% SIA, 28% SAI, 5% ASI, 4% AIS, 10% ISA, 3% IAS

Notes:

While there are ~4% more results for this model than the average, it is worth noting that this is the classic sort of interaction that is often used as an example: two primary effects interacting to strengthen each other. However, this model type does not predominate.

What is especially notable about this model is its distribution of relative effect magnitudes: SNPs predominate as the most common strongest effect, but interaction effects take a notable second place, and are consistently stronger than ATAC-seq effects. This is contrary to what would normally be expected: main effects, as the least complex terms in a regression model, are usually expected to predominate. These are effects we are directly registering from our data, with no further processing required. But in this case, ATAC-seq peaks are less likely to have strong effects by themselves, compared to interactions with SNPs.

5.3.1 SNP-dominant Models

If SNP effects are strongest, then the genetic variant is in a region with strong regulatory capacity, strengthening gene expression. 52493 models (77%) fall in this subset.

5.3.1.1 Interaction Term Second

If the interaction term is the next strongest, this may be due to colocalization of SNPs and ATAC-seq peaks, which would necessarily make any ATAC-seq peak behavior an

interaction with the SNP. It may also be that the two factors are not colocalized, residing in features that are complementary in their action, which are capable of acting on their own but with only limited effectiveness. 33223 models (49%) follow this pattern.

5.3.1.2 ATAC-seq Term Second

If the ATAC-seq effect is the next strongest term, then the two are likely not colocalized, and are capable of effectively acting independently of each other, but their mechanisms are supportive of each other. 19270 models (28%) follow this pattern.

5.3.2 ATAC-dominant Models

ATAC-seq is strongest: If ATAC effects are strongest, then it may be that it lies over a genetically stable, powerful enhancer or possibly the gene body. 5888 models (9%) fall in this subset.

5.3.2.1 Interaction Term Second

If the interaction term is the next strongest, colocalization with a SNP that improves regulatory performance of the area is possible, or the factors are non-overlapping but complementary. 3514 models (5%) follow this pattern.

5.3.2.2 SNP Term Second

If the SNP term is the next strongest, then the SNP and ATAC-seq peak are likely not colocalized and are capable of acting independently, but can support each other. 2374 models (4%) follow this pattern.

5.3.3 Interaction-dominant Models

Interaction term is strongest: This may result from colocalization that renders one main effect so dependent on the other that they are far less capable to increase gene expression without both appearing together. Another potential mechanism would be effects in two stages of the same regulatory pathway, such as transcription initiation. 9354 models (13%) fall in this subset.

5.3.3.1 SNP Term Second

If the SNP term is the next strongest, it may be causative of the ATAC-seq peak, upstream in a regulatory pathway. 7012 models (10%) follow this pattern.

5.3.3.2 ATAC Term Second

If the ATAC term is next strongest, it may be upstream in the regulatory pathway affected by the SNP, or the SNP may be in a less profoundly influential location. 2342 models (3%) follow this pattern.

5.4 SNP -, ATAC -, Interaction -

Number of models: 1230

% of all models: 0.30%

% relative magnitudes of each term (individual terms listed largest to smallest):

43% SIA, 50% SAI, 1% ASI, <1% AIS, 4% ISA, 2% IAS

One might naïvely suspect that models where main effects reinforce each other would predominate, but that turns out to not be the case. While models where all effects are positive make up about 1/6th of all significant models, results with all negative effects are the rarest model type by a wide margin.

The latter are almost entirely SNP-driven, with 93% of all models having SNP effects as the largest contributor. We can't say more about the magnitudes, especially with a relatively small subset of the data, but it indicates that this is a SNP-driven behavior. We will see this several times, but not with such an even split between the magnitudes of the other two terms.

5.4.1 SNP-dominant Models

If SNP effects are strongest, then the genetic variant disrupts a region with strong regulatory capacity. 1144 models (93%) fall in this subset.

5.4.1.1 Interaction Term Second

If the interaction term is the next strongest, the two main effects work together strongly to decrease gene expression. This may indicate that they are colocalized to a repressive region of some kind, which is made more effective by the presence of the SNP. This may also mean that they are in two different regions that work together to shut down gene expression. 531 models (43%) follow this pattern.

5.4.1.2 ATAC-seq Term Second

If the ATAC-seq effect is the next strongest term, then the two are likely not colocalized, and are capable of effectively acting independently of each other, but their mechanisms are supportive of each other. 613 models (50%) follow this pattern.

5.4.2 ATAC-dominant Models

If ATAC effects are strongest, then it may be that it lies over a genetically stable, powerful repressor or possibly an area where transcription tends to stall. Only 11 models (1%) fall in this subset.

5.4.2.1 Interaction Term Second

If the interaction term is the next strongest, it is possible that colocalization with a SNP that improves regulatory performance of the area of open chromatin, or the factors are non-overlapping but complementary. 9 models (1%) follow this pattern.

5.4.2.2 SNP Term Second

If the SNP term is the next strongest. 2 models (0%) follow this pattern.

5.4.3 Interaction-dominant Models

This may result from This may result from colocalization that renders one main effect so dependent on the other that they are far less capable to decrease gene expression without both appearing together. Another potential mechanism would be effects that

have complementary negative effects on gene expression, making it extremely effective in lowering gene expression. 75 models (6%) fall in this subset.

5.4.3.1 SNP Term Second

If the SNP term is the next strongest, it may be causative of the ATAC-seq peak, upstream in a regulatory pathway, such as transcription initiation or closing of other local chromatin. 53 models (4%) follow this pattern.

5.4.3.2 ATAC Term Second

If the ATAC term is next strongest, it may be upstream in the regulatory pathway affected by the SNP, or the SNP may be in a less profoundly influential location. 22 models (2%) follow this pattern.

5.5 SNP +, ATAC +, Interaction -

Number of models: 114383

% of all models: 27.76%

% relative magnitudes of each term (individual terms listed largest to smallest):

63% SIA, 29% SAI, 4% ASI, 4% AIS, <1% ISA, <1% IAS

This is the most common type of interaction by a wide margin. It indicates some form of interference or redundancy, where the main effects do not produce as much of a result as expected when they're added together. While in mathematical terms these variables are all 'independent', in biological terms they are not: because the SNP and ATAC-seq peak interact with each other in some way, they are necessarily dependent on each other to produce an interaction. But where and how this interaction takes place is not specified by the model. For example, it could be that the ATAC-seq peak and SNP are found in two different enhancers, which activate the same or competing regulatory mechanisms.

As seen before, the vast majority are SNP-driven, with the majority of models being SNP > Interaction > ATAC. SNPs dominate so strongly that there are essentially no models where interactions are the most powerful effect, which no other subset recapitulates. However, interaction terms still beat out ATAC-seq peaks as the second most powerful component.

5.5.1 SNP-dominant Models

If SNP effects are strongest, then the genetic variant is in a region with regulatory capacity, strengthening gene expression, but not as much as expected in combination with an ATAC-seq peak, indicating interference. 104,452 models (92%) fall in this subset.

5.5.1.1 Interaction Term Second

If the interaction term is the next strongest, it means that the two main effects are strongly interfering with each other. It may be that the two factors are not colocalized, residing in features that act upon the same target mechanism. 71,698 models (63%) follow this pattern.

5.5.1.2 ATAC-seq Term Second

If the ATAC-seq effect is the next strongest term, then the interference effect is weak, either due to partial physical interference or the regulatory factors kept relatively non-competitive by usually not activating at the same time. 32,754 models (29%) follow this pattern.

5.5.2 ATAC-dominant Models

If ATAC effects are strongest, then it may be that it lies over a genetically stable, powerful enhancer, and the SNP lies in an interfering regulatory region. 9,287 models (8%) fall in this subset.

5.5.2.1 Interaction Term Second

If the interaction term is the next strongest, the two regulatory factors are non-overlapping but interfering. Colocalization with a SNP that improves regulatory performance of the area is possible, but would require the two to function via some sort of opposing methods. 4,855 models (4%) follow this pattern.

5.5.2.2 SNP Term Second

If the SNP term is the next strongest, then the SNP and ATAC-seq peak are likely not colocalized and are most effective when acting independently, but do not cause strong interference with each other. 4,432 models (4%) follow this pattern.

5.5.3 Interaction-dominant Models

This is very rare, indicating that very few cases of interference are so strong as to actually overpower the positive effects of their individual parts. These would be interesting to examine, to identify their validity and what they act upon. 644 models (1%) follow this pattern.

5.5.3.1 SNP Term Second

If the SNP term is the next strongest, the SNP's presence by itself provides a relatively large increase to gene expression, but the mechanism of this regulation may profoundly disrupt another regulatory function that involves open chromatin (or vice versa), causing transcription to be suppressed when both are present. 220 models (>1%) follow this pattern.

5.5.3.2 ATAC Term Second

If the ATAC term is next strongest, the function may be much the same as above, with the SNP lying in a less profoundly influential location. 424 models (>1%) follow this pattern.

5.6 SNP -, ATAC -, Interaction +

Number of models: 48550

% of all models: 11.78%

% relative magnitudes of each term (individual terms listed largest to smallest):

62% SIA, 18% SAI, <1% ASI, <1% AIS, 14% ISA, 5% IAS

Interestingly, the percentage of the results that involve interference is not symmetrical--results with negative main effects and a positive interaction effect are far less common than their inverse. Perhaps this is due to the relative scarcity of repressors that act by holding open an area of DNA, but as we will see later, that may not be the case. The presence of these models may also be explained by the general principle that transcription can always theoretically be increased, but it cannot go lower than zero--two extremely strong repressive regulators may be completely redundant.

As before, SNPs dominate, but with a surprisingly strong contribution from Interaction terms.

5.6.1 SNP-dominant Models

If SNP effects are strongest, then the SNP is in a relatively strong negative regulatory area, possibly but not necessarily co-localized with the ATAC-seq peak. However, the interaction term shows that there is a floor for how low they can collectively push, either by interference or redundancy. 38,840 models (80%) follow this pattern.

5.6.1.1 Interaction Term Second

If the interaction term is the next strongest, the effects described above are especially pronounced. 29,919 models (62%) follow this pattern.

5.6.1.2 ATAC-seq Term Second

If the ATAC-seq effect is the next strongest term, the SNP and ATAC-seq peak are effective negative regulators, but are still encountering the conditions described above. 8,921 models (18%) follow this pattern.

5.6.2 ATAC-dominant Models

If ATAC effects are strongest, the reverse of the situation described above predominates, with two negative regulatory factors encountering interference or redundancy issues with a powerful ATAC-seq peak predominating. This appears to be exceedingly rare, with only 212 models (0%) following this pattern.

5.6.2.1 Interaction Term Second

If the interaction term is the next strongest, the effects described above are especially pronounced. 51 models (0%) follow this pattern.

5.6.2.2 SNP Term Second

If the SNP term is the next strongest, the interaction effect is weak, but still indicates some level of interference or redundancy between the two main effects. 161 models (0%) follow this pattern.

5.6.3 Interaction-dominant Models

If the interaction term is strongest, this can be interpreted a couple of ways, depending on the actual magnitude of the interaction effect relative to the combined total of the SNP and ATAC effects. Redundancy is no longer a consideration, these models push into actual interference that completely overrides the negative effects of one or both negative main effects. If it overpowers both, then you are left with an actual *increase* in gene expression relative to baseline, perhaps indicating that a positive regulatory pathway is triggered by the co-activation of the SNP and ATAC-seq peak's regions. 9,498 models (20%) of models follow this pattern.

5.6.3.1 SNP Term Second

If the SNP term is the next strongest, then the SNP is the primary negative regulator, which will pull down gene expression if the ATAC-seq peak's region isn't active. 6,966 models (14%) follow this pattern.

5.6.3.2 ATAC Term Second

If the ATAC term is next strongest, the reverse of the above occurs. 2,532 models (5%) follow this pattern.

5.7 SNP-, ATAC +, Interaction +

Number of models: 44300

% of all models: 10.75%

% relative magnitudes of each term (individual terms listed largest to smallest):

44% SIA, 18% SAI, 9% ASI, 8% AIS, 14% ISA, 7% IAS

In these models, the positive effects of open chromatin override a negative SNP effect. This could be from the effects being localized to two different regulatory features, with the ATAC-seq peak's region functionally overriding the SNP's region. Another option is that the ATAC-seq peak and SNP could be colocalized to the same enhancer/promoter/etc., but the effect of the genetic variant is not sufficiently disruptive to abolish the positive effect of the result.

SNPs do not have as strong a hold over these models, resulting in a more mixed set of these. Interactions are again the second-strongest term, in contravention of conventional expectations of main effect dominance. Depending on the relative strengths of the terms, the interaction could ultimately end up pulling expression into the positive, or remain depressed compared to its baseline expression without any factors.

5.7.1 SNP-dominant Models

If SNP effects are strongest, then it may lie in a relatively strong regulatory region, which is modified or interfered with by the ATAC-seq region. 27,389 models (62%) follow this pattern.

5.7.1.1 Interaction Term Second

If the interaction term is the next strongest, the disruptive effects of the ATAC-seq peak are more strongly felt. 19,386 models (44%) follow this pattern.

5.7.1.2 ATAC-seq Term Second

If the ATAC-seq effect is the next strongest term, the disruptive effects are not especially strong, when compared to the strength of the ATAC-seq peak itself. 8,003 models (18%) follow this pattern.

5.7.2 ATAC-dominant Models

If ATAC effects are strongest, then it's overriding the SNP's negative pull. While this doesn't happen a majority of the time, it is still surprisingly common, given how rarely the ATAC-seq effect overrides the other effects: 7,554 models (17%) follow this pattern.

5.7.2.1 Interaction Term Second

If the interaction term is the next strongest, the strong ATAC-seq signal is also accompanied by a strong interaction, possibly indicating that the two main effects are colocalized, and opening the local chromatin overrides the negative effect of the SNP, creating a more positive result than expected. 3,985 models (9%) follow this pattern.

5.7.2.2 SNP Term Second

If the SNP term is the next strongest, the above model still may apply, but with less intense results. 3,569 models (8%) follow this pattern.

5.7.3 Interaction-dominant Models

If the interaction term is strongest, the main effect regions may not be relatively strong regulators in their own right, and their overall positive effect is the result of interference that favors the ATAC-seq regulatory pathway. 9,357 models (21%) follow this pattern.

5.7.3.1 SNP Term Second

If the SNP term is the next strongest, the ATAC-seq peak is a weak regulator on its own, but may be the initiator of a regulatory mechanism that overrides the SNP effect. 6,389 models (14%) follow this pattern.

5.7.3.2 ATAC Term Second

If the ATAC term is next strongest, the above mechanism still holds, but the SNP effect appears to be more weakly modifying local behavior. 2,968 models (7%) follow this pattern.

5.8 SNP+, ATAC-, Interaction-

Number of models: 67021

% of all models: (16.26%)

% relative magnitudes of each term (individual terms listed largest to smallest):

83% SIA, 17% SAI, <1% ASI, 0% AIS, <1% ISA, <1% IAS

The same sort of pattern as the subsection above seems to follow for the inverse arrangement, where a negative ATAC-seq peak seems to override a positive SNP. Interestingly, these are more common than the other way around--a negative SNP overriding a positive ATAC-seq peak. Unlike the previous category, these models are extremely SNP-driven. This is only somewhat surprising, because ATAC-seq peaks are generally associated with positive effects, rather than negative ones. However, given the high proportion of negative SNP effects overall in this analysis, it is notable that this category in particular is almost devoid of strongly regulatory ATAC-seq peaks.

5.8.1 SNP-dominant Models

If SNP effects are strongest, the SNP is rendered less effective as a positive regulator by the presence of the ATAC-seq peak. This may be because the ATAC-seq peak lies in a region that renders the SNP's region less functional, but does not entirely preclude it from use. 66,987 models (100%) fall within this category.

5.8.1.1 Interaction Term Second

If the interaction term is the next strongest, the ATAC-seq peak's local regulatory behavior is not particularly strong on its own compared to the SNP, but its function has a strong modulating effect on the overall increase in gene expression. 55,344 models (83%) follow this pattern.

5.8.1.2 ATAC-seq Term Second

If the ATAC-seq effect is the next strongest term, then the modulatory effect is weaker, producing a n effect that is not as significant as the ATAC-seq peak's negative effect on gene expression by itself. 11,643 models (17%) follow this pattern.

5.8.2 ATAC-dominant Models

If ATAC effects are strongest, the ATAC-seq peak is highly negative by itself, and creates an interaction effect that takes the SNP's co-occurrence and turns it to an overall negative effect. Only 3 models (<1%) follow this pattern.

5.8.2.1 Interaction Term Second

If the interaction term is the next strongest, then the negative effect of the ATAC-seq's co-occurrence with the SNP would override the positive effect it might have had. However, no models follow this pattern.

5.8.2.2 SNP Term Second

If the SNP term is the next strongest, the overall effect of interaction is weakly negative, possibly due to minor interference by the ATAC-seq peak, which drives the overall negativity of the model. Only 3 models (<1%) follow this pattern.

5.8.3 Interaction-dominant Models

the interference by the ATAC-seq peak is profound, overriding the positive effect of the SNP. 31 models (<1%) follow this pattern.

5.8.3.1 SNP Term Second

If the SNP term is the next strongest, the ATAC-seq peak by itself is weakly negative, but it interferes significantly with the region where the SNP effect takes place. 21 models (<1%) follow this pattern.

5.8.3.2 ATAC Term Second

If the ATAC term is next strongest, a strongly negative ATAC-seq peak has a compounding negative effect on gene expression when the SNP is present, indicating potential disruption of a larger regulatory mechanism. 10 models (<1%) follow this pattern.

5.9 SNP-, ATAC+, Interaction-

Number of models: 17853

% of all models: (4.33%)

% relative magnitudes of each term (individual terms listed largest to smallest):

11% SIA, 49% SAI, 19% ASI, 17% AIS, 1% ISA, 2% IAS

This is particularly odd when you consider one of the most simple configurations of interacting elements: a SNP under an ATAC-seq peak. If an ATAC-seq peak would normally have a positive effect on gene expression by exposing an enhancer, and a SNP within that enhancer has a negative effect, one might expect to see an interaction

like this. Perhaps this is due to the fact that one would need a SNP that has an especially negative effect on protein binding to create a negative interaction effect, or that disruption of enhancer binding would disrupt the ATAC-seq peak as well.

After the their generally weak showings in many other model sets, it is worth noting the percentage of ATAC-dominant models in this rare group.

5.9.1 SNP-dominant Models

If SNP effects are strongest, the SNP is a strong negative regulator by itself, likely directly disrupting some important step in facilitating gene expression. The ATAC-seq peak may be co-localized with it, explaining its inability to function as expected. It also has a strong negative effect by itself, meaning that it does not necessarily need open chromatin for its negative effect to be felt, or it also disrupts the chances of chromatin being opened in the first place. 10,656 models (60%) follow this pattern.

5.9.1.1 Interaction Term Second

If the interaction term is the next strongest, the effects of the interaction are particularly strong, with the ATAC-seq peak's intended functionality comprehensively undermined. 1,889 models (11%) follow this pattern.

5.9.1.2 ATAC-seq Term Second

If the ATAC-seq effect is the next strongest term, the open chromatin still signals some level of function, but it is not enough to override the negative effect of the SNP. 8,767 models (49%) follow this pattern.

5.9.2 ATAC-dominant Models

If ATAC effects are strongest, the ATAC-seq peak is strongly positive in spite of either interference with its function or active disruption of the sequence it contains. Overall

results may be positive, but if they are, they are less so than expected. 6,522 models (36%) follow this pattern.

5.9.2.1 Interaction Term Second

If the interaction term is the next strongest, the SNP does not have a strongly negative effect by itself, indicating that it needs the ATAC-seq peak for the implications of its disruption to be felt. 3425 models (19%) follow this pattern.

5.9.2.2 SNP Term Second

If the SNP term is the next strongest, the SNP may be able to stand on its own as a negative regulator, but it has some detectable disruptive effect on the ATAC-seq region. 3,097 models (17%) follow this pattern.

5.9.3 Interaction-dominant Models

The SNP is likely co-localized, and strongly undermines the ATAC-seq peak's functionality. Another possibility is that the two main effects are within the same regulatory pathway, with the ATAC-seq peak and SNP modifying overall function of the pathway. 675 models (3%) follow this pattern.

5.9.3.2 SNP Term Second

If the SNP term is the next strongest, the SNP is a strong negative regulator by itself, indicating that it may be able to disrupt ATAC-seq peak formation, and undermine its effect when it is present. Another possibility is that the SNP is present elsewhere, disrupting the regulatory process in another pathway even when the ATAC-seq peak is absent. 263 models (1%) follow this pattern.

5.9.3.3 ATAC Term Second

If the ATAC term is next strongest, the SNP is not a strong regulator, but it causes disproportionate disruption to the ATAC-seq peak's function. This likely means they are

co-localized, as the SNP does not function much without the ATAC-seq peak. 412 models (2%) follow this pattern.

5.10 SNP+, ATAC-, Interaction+

Number of models: 51040

% of all models: (12.39%)

% relative magnitudes of each term (individual terms listed largest to smallest):

62% SIA, 18% SAI, <1% ASI, <1% AIS, 14% ISA, 5% IAS

The inverse of the previous model type is far more common, a negative ATAC-seq effect seemingly overpowered by a positive SNP. This further adds to the cryptic nature of these models, both in terms of the non-reciprocal nature of these two model types, but also simply due to to the relative rarity of both effects: SNPs that disrupt regulatory mechanisms in a way that increases gene expression is less likely to happen than a SNP that decreases gene expression. The ATAC-seq peaks seem even less likely. While not the most powerful effectors in this subset, the fact that they have this effect is not as expected, based on general findings about ATAC-seq peaks.

5.10.1 SNP-dominant Models

If SNP effects are strongest, its positive effect stands strong by itself, but is buoyed further by the presence of the normally repressive ATAC-seq peak. Perhaps the open chromatin that would normally have a repressive effect also exposes the SNP, or a related part of its regulatory pathway. 40,740 models (80%) follow this pattern.

5.10.1.1 Interaction Term Second

If the interaction term is the next strongest, the interaction described above is influential in bolstering the SNP's action, perhaps through one of the ways described above. 30,437 models (62%) follow this pattern.

5.10.1.2 ATAC-seq Term Second

If the ATAC-seq effect is the next strongest term, the interaction is of lesser importance, meaning that the SNP is more likely to be largely independent from the ATAC-seq peak, and possibly engaging in more distant regulatory interference. 10,303 models (18%) follow this pattern.

5.10.2 ATAC-dominant Models

If ATAC effects are strongest, the negative effect of the open chromatin is uncommonly strong, which is also unexpected. Accordingly, 503 models (1%) follow this pattern.

5.10.2.1 Interaction Term Second

If the interaction term is the next strongest, the ATAC-seq peak and SNP may be co-localized, with the SNP's positive effect only really felt when the chromatin is open, at which point it has a throttling effect on the local regulatory feature(s). 189 models (<1%) follow this pattern.

5.10.2.2 SNP Term Second

If the SNP term is the next strongest, the SNP itself seems to be largely independent of the ATAC-seq peak, with the interaction modifying the negative effects of the ATAC-seq peak. 314 models (<1%) follow this pattern.

5.10.3 Interaction-dominant Models

If the interaction effects are strongest, the SNP is only partially functional without the usually repressive ATAC-seq peak, which may indicate co-localization that decreases the chance of the chromatin opening in the first place, and undermines repressive function in the area when it does open.

5.10.3.2 SNP Term Second

If the SNP term is the next strongest, the effect described above is particularly acute. 7,158 models (14%) follow this pattern.

5.10.3.3 ATAC Term Second

If the ATAC term is next strongest, localization may be one explanation, due to the low effects of the SNP without the ATAC-seq peak. However, the SNP may not disrupt the local repressive function as strongly. Alternatively, if the ATAC-seq peak is not co-localized with the SNP, the SNP could be downstream a regulatory pathway from the ATAC-seq peak, and thus less effective without it. 2,639 models (5%) follow this pattern.

5.11 Attribution

Analyses in chapter 5 were performed solely by the author.

Chapter 6: References

Acurzio, B., Verma, A., Polito, A., Giaccari, C., Cecere, F., Fioriniello, S., Della Ragione, F., Fico, A., Cerrato, F., Angelini, C., Feil, R., & Riccio, A. (2021). Zfp57 inactivation illustrates the role of ICR methylation in imprinted gene expression during neural differentiation of mouse ESCs. *Scientific Reports*, *11*(1), 13802. https://doi.org/10.1038/s41598-021-93297-3

Aitken, S. J., Ibarra-Soria, X., Kentepozidou, E., Flicek, P., Feig, C., Marioni, J. C., & Odom, D. T. (2018). CTCF maintains regulatory homeostasis of cancer pathways. *Genome Biology*, *19*(1), 106. https://doi.org/10.1186/s13059-018-1484-3

Akdemir, K. C., Le, V. T., Chandran, S., Li, Y., Verhaak, R. G., Beroukhim, R., Campbell, P. J., Chin, L., Dixon, J. R., Futreal, P. A., PCAWG Structural Variation Working Group, Akdemir, K. C., Alvarez, E. G., Baez-Ortega, A., Boutros, P. C., Bowtell, D. D. L., Brors, B., Burns, K. H., Campbell, P. J., … Von Mering, C. (2020). Disruption of chromatin folding domains by somatic genomic rearrangements in human cancer. *Nature Genetics*, *52*(3), 294–305. https://doi.org/10.1038/s41588-019-0564-y

Andrey, G., Montavon, T., Mascrez, B., Gonzalez, F., Noordermeer, D., Leleu, M., Trono, D., Spitz, F., & Duboule, D. (2013). A Switch Between Topological Domains Underlies *HoxD* Genes Collinearity in Mouse Limbs. *Science*, *340*(6137), 1234167. https://doi.org/10.1126/science.1234167

Arzate-Mejía, R. G., Recillas-Targa, F., & Corces, V. G. (2018). Developing in 3D: The role of CTCF in cell differentiation. *Development*, *145*(6), dev137729. https://doi.org/10.1242/dev.137729

Atlasi, Y., Noori, R., Gaspar, C., Franken, P., Sacchetti, A., Rafati, H., Mahmoudi, T., Decraene, C., Calin, G. A., Merrill, B. J., & Fodde, R. (2013). Wnt Signaling Regulates the Lineage Differentiation Potential of Mouse Embryonic Stem Cells through Tcf3 Down-Regulation. *PLoS Genetics*, *9*(5), e1003424. https://doi.org/10.1371/journal.pgen.1003424

Babu, A., & Verma, R. S. (1987). Chromosome Structure: Euchromatin and Heterochromatin. In *International Review of Cytology* (Vol. 108, pp. 1–60). Elsevier. https://doi.org/10.1016/S0074-7696(08)61435-7

Bailey, T. L. (2020). *STREME: Accurate and versatile sequence motif discovery* [Preprint]. Bioinformatics. https://doi.org/10.1101/2020.11.23.394619

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W., & Noble, W. S. (2009). MEME SUITE: Tools for motif discovery and searching. *Nucleic Acids Research*, *37*(Web Server), W202–W208. https://doi.org/10.1093/nar/gkp335

Bailey, T. L., & Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings. International Conference on Intelligent Systems for Molecular Biology*, *2*, 28–36.

Barrington, C., Georgopoulou, D., Pezic, D., Varsally, W., Herrero, J., & Hadjur, S. (2019). Enhancer accessibility and CTCF occupancy underlie asymmetric TAD

architecture and cell type specific genome topology. *Nature Communications*, *10*(1), 2908. https://doi.org/10.1038/s41467-019-10725-9

Bartkuhn, M., & Renkawitz, R. (2008). Long range chromatin interactions involved in gene regulation. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, *1783*(11), 2161–2166. https://doi.org/10.1016/j.bbamcr.2008.07.011

Baudat, F., Buard, J., Grey, C., Fledel-Alon, A., Ober, C., Przeworski, M., Coop, G., & De Massy, B. (2010). PRDM9 Is a Major Determinant of Meiotic Recombination Hotspots in Humans and Mice. *Science*, *327*(5967), 836–840. https://doi.org/10.1126/science.1183439

Bekris, L. M., Lutz, F., & Yu, C.-E. (2012). Functional analysis of APOE locus genetic variation implicates regional enhancers in the regulation of both TOMM40 and APOE. *Journal of Human Genetics*, *57*(1), 18–25. https://doi.org/10.1038/jhg.2011.123

Berger, F. (2019). Emil Heitz, a true epigenetics pioneer. *Nature Reviews Molecular Cell Biology*, *20*(10), 572–572. https://doi.org/10.1038/s41580-019-0161-z

Brandão, H. B., Gabriele, M., & Hansen, A. S. (2021). Tracking and interpreting long-range chromatin interactions with super-resolution live-cell imaging. *Current Opinion in Cell Biology*, *70*, 18–26. https://doi.org/10.1016/j.ceb.2020.11.002

Broman, K. W. (2012). Haplotype Probabilities in Advanced Intercross Populations. *G3 Genes|Genomes|Genetics*, *2*(2), 199–202. https://doi.org/10.1534/g3.111.001818

Brown, J. M., Leach, J., Reittie, J. E., Atzberger, A., Lee-Prudhoe, J., Wood, W. G., Higgs, D. R., Iborra, F. J., & Buckle, V. J. (2006). Coregulated human globin genes are frequently in spatial proximity when active. *Journal of Cell Biology*, *172*(2), 177–187. https://doi.org/10.1083/jcb.200507073

Callis, J., Fromm, M., & Walbot, V. (1987). Introns increase gene expression in cultured maize cells. *Genes & Development*, *1*(10), 1183–1200. https://doi.org/10.1101/gad.1.10.1183

Cao, J., Cusanovich, D. A., Ramani, V., Aghamirzaie, D., Pliner, H. A., Hill, A. J., Daza, R. M., McFaline-Figueroa, J. L., Packer, J. S., Christiansen, L., Steemers, F. J., Adey, A. C., Trapnell, C., & Shendure, J. (2018). Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science*, *361*(6409), 1380–1385. https://doi.org/10.1126/science.aau0730

Carter, D., Chakalova, L., Osborne, C. S., Dai, Y., & Fraser, P. (2002). Long-range chromatin regulatory interactions in vivo. *Nature Genetics*, *32*(4), 623–626. https://doi.org/10.1038/ng1051

Chambeyron, S., & Bickmore, W. A. (2004). Does looping and clustering in the nucleus regulate gene expression? *Current Opinion in Cell Biology*, *16*(3), 256–262. https://doi.org/10.1016/j.ceb.2004.03.004

Chesler, E. J., Gatti, D. M., Morgan, A. P., Strobel, M., Trepanier, L., Oberbeck, D., McWeeney, S., Hitzemann, R., Ferris, M., McMullan, R., Clayshultle, A., Bell, T. A., De Villena, F. P.-M., & Churchill, G. A. (2016). Diversity Outbred Mice at 21: Maintaining Allelic Variation in the Face of Selection. *G3 Genes|Genomes|Genetics*, *6*(12), 3893–3902. https://doi.org/10.1534/g3.116.035527

Chick, J. M., Munger, S. C., Simecek, P., Huttlin, E. L., Choi, K., Gatti, D. M., Raghupathy, N., Svenson, K. L., Churchill, G. A., & Gygi, S. P. (2016). Defining the consequences of genetic variation on a proteome-wide scale. *Nature*, *534*(7608), 500–505. https://doi.org/10.1038/nature18270

Davis, R. C., Schadt, E. E., Cervino, A. C. L., Péterfy, M., & Lusis, A. J. (2005). Ultrafine Mapping of SNPs From Mouse Strains C57BL/6J, DBA/2J, and C57BLKS/J for Loci Contributing to Diabetes and Atherosclerosis Susceptibility. *Diabetes*, *54*(4), 1191–1199. https://doi.org/10.2337/diabetes.54.4.1191

Dekker, J., Rippe, K., Dekker, M., & Kleckner, N. (2002). Capturing Chromosome Conformation. *Science*, *295*(5558), 1306–1311. https://doi.org/10.1126/science.1067799

Del Castillo, J., & Katz, B. (1954). Quantal components of the end-plate potential. *The Journal of Physiology*, *124*(3), 560–573. https://doi.org/10.1113/jphysiol.1954.sp005129

Ding, K.-H., Wang, Z.-Z., Hamrick, M. W., Deng, Z.-B., Zhou, L., Kang, B., Yan, S.-L., She, J.-X., Stern, D. M., Isales, C. M., & Mi, Q.-S. (2006). Disordered osteoclast formation in RAGE-deficient mouse establishes an essential role for RAGE in diabetes related bone loss. *Biochemical and Biophysical Research Communications*, *340*(4), 1091–1097. https://doi.org/10.1016/j.bbrc.2005.12.107

Dixon, A. L., Liang, L., Moffatt, M. F., Chen, W., Heath, S., Wong, K. C. C., Taylor, J., Burnett, E., Gut, I., Farrall, M., Lathrop, G. M., Abecasis, G. R., & Cookson, W. O. C. (2007). A genome-wide association study of global gene expression. *Nature Genetics*, *39*(10), 1202–1207. https://doi.org/10.1038/ng2109

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., & Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, *485*(7398), 376–380. https://doi.org/10.1038/nature11082

Dixon, J. R., Xu, J., Dileep, V., Zhan, Y., Song, F., Le, V. T., Yardımcı, G. G., Chakraborty, A., Bann, D. V., Wang, Y., Clark, R., Zhang, L., Yang, H., Liu, T., Iyyanki, S., An, L., Pool, C., Sasaki, T., Rivera-Mulia, J. C., … Yue, F. (2018). Integrative detection and analysis of structural variation in cancer genomes. *Nature Genetics*, *50*(10), 1388–1398. https://doi.org/10.1038/s41588-018-0195-8

Enver, T., Soneji, S., Joshi, C., Brown, J., Iborra, F., Orntoft, T., Thykjaer, T., Maltby, E., Smith, K., Dawud, R. A., Jones, M., Matin, M., Gokhale, P., Draper, J., & Andrews, P. W. (2005). Cellular differentiation hierarchies in normal and culture-adapted human embryonic stem cells. *Human Molecular Genetics*, *14*(21), 3129–3140. https://doi.org/10.1093/hmg/ddi345

Evans, M. J., & Kaufman, M. H. (1981). Establishment in culture of pluripotential cells from mouse embryos. *Nature*, *292*(5819), 154–156. https://doi.org/10.1038/292154a0

Ferraj, A., Audano, P. A., Balachandran, P., Czechanski, A., Flores, J. I., Radecki, A. A., Mosur, V., Gordon, D. S., Walawalkar, I. A., Eichler, E. E., Reinholdt, L. G., & Beck, C. R. (2023). Resolution of structural variation in diverse mouse genomes

reveals chromatin remodeling due to transposable elements. *Cell Genomics*, *3*(5), 100291. https://doi.org/10.1016/j.xgen.2023.100291

Fraser, P., & Grosveld, F. (1998). Locus control regions, chromatin activation and transcription. *Current Opinion in Cell Biology*, *10*(3), 361–365. https://doi.org/10.1016/S0955-0674(98)80012-4

Gacita, A. M., Fullenkamp, D. E., Ohiri, J., Pottinger, T., Puckelwartz, M. J., Nobrega, M. A., & McNally, E. M. (2021). Genetic Variation in Enhancers Modifies Cardiomyopathy Gene Expression and Progression. *Circulation*, *143*(13), 1302–1316. https://doi.org/10.1161/CIRCULATIONAHA.120.050432

Gate, R. E., Cheng, C. S., Aiden, A. P., Siba, A., Tabaka, M., Lituiev, D., Machol, I., Gordon, M. G., Subramaniam, M., Shamim, M., Hougen, K. L., Wortman, I., Huang, S.-C., Durand, N. C., Feng, T., De Jager, P. L., Chang, H. Y., Aiden, E. L., Benoist, C., … Regev, A. (2018). Genetic determinants of co-accessible chromatin regions in activated T cells across humans. *Nature Genetics*, *50*(8), 1140–1150. https://doi.org/10.1038/s41588-018-0156-2

Gong, H., Yang, Y., Zhang, S., Li, M., & Zhang, X. (2021). Application of Hi-C and other omics data analysis in human cancer and cell differentiation research. *Computational and Structural Biotechnology Journal*, *19*, 2070–2083. https://doi.org/10.1016/j.csbj.2021.04.016

Gong, Y., Lazaris, C., Sakellaropoulos, T., Lozano, A., Kambadur, P., Ntziachristos, P., Aifantis, I., & Tsirigos, A. (2018). Stratification of TAD boundaries reveals preferential insulation of super-enhancers by strong boundaries. *Nature Communications*, *9*(1), 542. https://doi.org/10.1038/s41467-018-03017-1

Gong, Y., & Zou, F. (2012). Varying Coefficient Models for Mapping Quantitative Trait Loci Using Recombinant Inbred Intercrosses. *Genetics*, *190*(2), 475–486. https://doi.org/10.1534/genetics.111.132522

Grant, C. E., Bailey, T. L., & Noble, W. S. (2011). FIMO: Scanning for occurrences of a given motif. *Bioinformatics*, *27*(7), 1017–1018. https://doi.org/10.1093/bioinformatics/btr064

Greenwald, W. W., Li, H., Benaglio, P., Jakubosky, D., Matsui, H., Schmitt, A., Selvaraj, S., D'Antonio, M., D'Antonio-Chronowska, A., Smith, E. N., & Frazer, K. A. (2019). Subtle changes in chromatin loop contact propensity are associated with differential gene regulation and expression. *Nature Communications*, *10*(1), 1054. https://doi.org/10.1038/s41467-019-08940-5

Gupta, K., & Varadarajan, R. (2018). Insights into protein structure, stability and function from saturation mutagenesis. *Current Opinion in Structural Biology*, *50*, 117–125. https://doi.org/10.1016/j.sbi.2018.02.006

Handoko, L., Xu, H., Li, G., Ngan, C. Y., Chew, E., Schnapp, M., Lee, C. W. H., Ye, C., Ping, J. L. H., Mulawadi, F., Wong, E., Sheng, J., Zhang, Y., Poh, T., Chan, C. S., Kunarso, G., Shahab, A., Bourque, G., Cacheux-Rataboul, V., … Wei, C.-L. (2011). CTCF-mediated functional chromatin interactome in pluripotent cells. *Nature Genetics*, *43*(7), 630–638. https://doi.org/10.1038/ng.857

Haussler, D., & Welzl, E. (1987). Ɛ-nets and simplex range queries. *Discrete & Computational Geometry*, *2*(2), 127–151. https://doi.org/10.1007/BF02187876

Hocking, R. R. (1976). A Biometrics Invited Paper. The Analysis and Selection of Variables in Linear Regression. *Biometrics*, *32*(1), 1. https://doi.org/10.2307/2529336

Holzmann, J., Politi, A. Z., Nagasaka, K., Hantsche-Grininger, M., Walther, N., Koch, B., Fuchs, J., Dürnberger, G., Tang, W., Ladurner, R., Stocsits, R. R., Busslinger, G. A., Novák, B., Mechtler, K., Davidson, I. F., Ellenberg, J., & Peters, J.-M. (2019). Absolute quantification of cohesin, CTCF and their regulators in human cells. *ELife*, *8*, e46269. https://doi.org/10.7554/eLife.46269

Hooper, M., Hardy, K., Handyside, A., Hunter, S., & Monk, M. (1987). HPRT-deficient (Lesch–Nyhan) mouse embryos derived from germline colonization by cultured cells. *Nature*, *326*(6110), 292–295. https://doi.org/10.1038/326292a0

Hou, C., Li, L., Qin, Z. S., & Corces, V. G. (2012). Gene Density, Transcription, and Insulators Contribute to the Partition of the Drosophila Genome into Physical Domains. *Molecular Cell*, *48*(3), 471–484. https://doi.org/10.1016/j.molcel.2012.08.031

Hou, C., Zhao, H., Tanimoto, K., & Dean, A. (2008). CTCF-dependent enhancer-blocking by alternative chromatin loop formation. *Proceedings of the National Academy of Sciences*, *105*(51), 20398–20403. https://doi.org/10.1073/pnas.0808506106

Jeffreys, A. J., Kauppi, L., & Neumann, R. (2001). Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genetics*, *29*(2), 217–222. https://doi.org/10.1038/ng1001-217

Kagey, M. H., Newman, J. J., Bilodeau, S., Zhan, Y., Orlando, D. A., Van Berkum, N. L., Ebmeier, C. C., Goossens, J., Rahl, P. B., Levine, S. S., Taatjes, D. J., Dekker, J., & Young, R. A. (2010). Mediator and cohesin connect gene expression and chromatin architecture. *Nature*, *467*(7314), 430–435. https://doi.org/10.1038/nature09380

Katz, D. C., Aponte, J. D., Liu, W., Green, R. M., Mayeux, J. M., Pollard, K. M., Pomp, D., Munger, S. C., Murray, S. A., Roseman, C. C., Percival, C. J., Cheverud, J., Marcucio, R. S., & Hallgrímsson, B. (2020). Facial shape and allometry quantitative trait locus intervals in the Diversity Outbred mouse are enriched for known skeletal and facial development genes. *PLOS ONE*, *15*(6), e0233377. https://doi.org/10.1371/journal.pone.0233377

Keane, T. M., Goodstadt, L., Danecek, P., White, M. A., Wong, K., Yalcin, B., Heger, A., Agam, A., Slater, G., Goodson, M., Furlotte, N. A., Eskin, E., Nellåker, C., Whitley, H., Cleak, J., Janowitz, D., Hernandez-Pliego, P., Edwards, A., Belgard, T. G., … Adams, D. J. (2011). Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature*, *477*(7364), 289–294. https://doi.org/10.1038/nature10413

Khoury, A., Achinger-Kawecka, J., Bert, S. A., Smith, G. C., French, H. J., Luu, P.-L., Peters, T. J., Du, Q., Parry, A. J., Valdes-Mora, F., Taberlay, P. C., Stirzaker, C., Statham, A. L., & Clark, S. J. (2020). Constitutively bound CTCF sites maintain 3D chromatin architecture and long-range epigenetically regulated domains. *Nature Communications*, *11*(1), 54. https://doi.org/10.1038/s41467-019-13753-7

Kim, Y., Xia, K., Tao, R., Giusti-Rodriguez, P., Vladimirov, V., Van Den Oord, E., & Sullivan, P. F. (2014). A meta-analysis of gene expression quantitative trait loci in brain. *Translational Psychiatry*, *4*(10), e459–e459. https://doi.org/10.1038/tp.2014.96

Kingra, S. K., Parmar, V., Chang, C.-C., Hudec, B., Hou, T.-H., & Suri, M. (2020). SLIM: Simultaneous Logic-in-Memory Computing Exploiting Bilayer Analog OxRAM Devices. *Scientific Reports*, *10*(1), 2567. https://doi.org/10.1038/s41598-020-59121-0

Krefting, J., Andrade-Navarro, M. A., & Ibn-Salem, J. (2018). Evolutionary stability of topologically associating domains is associated with conserved gene regulation. *BMC Biology*, *16*(1), 87. https://doi.org/10.1186/s12915-018-0556-x

Krijger, P. H. L., & De Laat, W. (2016). Regulation of disease-associated gene expression in the 3D genome. *Nature Reviews Molecular Cell Biology*, *17*(12), 771–782. https://doi.org/10.1038/nrm.2016.138

Kubo, N., Ishii, H., Gorkin, D., Meitinger, F., Xiong, X., Fang, R., Liu, T., Ye, Z., Li, B., Dixon, J. R., Desai, A., Zhao, H., & Ren, B. (2017). *Preservation of Chromatin Organization after Acute Loss of CTCF in Mouse Embryonic Stem Cells* [Preprint]. Genomics. https://doi.org/10.1101/118737

Kuffler, L., Fortin, H. J., Skelly, D. A., Czechanski, A., Munger, S. C., Reinholdt, L. G., Baker, C. L., & Carter, G. W. (2023). *Imputation of 3D genome structure by genetic-epigenetic interaction modeling in mice* [Preprint]. elife. https://doi.org/10.7554/eLife.88222.1

Kuhn, R. M., Haussler, D., & Kent, W. J. (2013). The UCSC genome browser and associated tools. *Briefings in Bioinformatics*, *14*(2), 144–161. https://doi.org/10.1093/bib/bbs038

Kulakovskiy, I. V., Vorontsov, I. E., Yevshin, I. S., Sharipov, R. N., Fedorova, A. D., Rumynskiy, E. I., Medvedeva, Y. A., Magana-Mora, A., Bajic, V. B., Papatsenko, D. A., Kolpakov, F. A., & Makeev, V. J. (2018). HOCOMOCO: Towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Research*, *46*(D1), D252–D259. https://doi.org/10.1093/nar/gkx1106

Kumasaka, N., Knights, A. J., & Gaffney, D. J. (2016). Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nature Genetics*, *48*(2), 206–213. https://doi.org/10.1038/ng.3467

Le Caignec, C., Pichon, O., Briand, A., De Courtivron, B., Bonnard, C., Lindenbaum, P., Redon, R., Schluth-Bolard, C., Diguet, F., Rollat-Farnier, P.-A., Sanchez-Castro, M., Vuillaume, M.-L., Sanlaville, D., Duboule, D., Mégarbané, A., & Toutain, A. (2020). Fryns type mesomelic dysplasia of the upper limbs caused by inverted duplications of the HOXD gene cluster. *European Journal of Human Genetics*, *28*(3), 324–332. https://doi.org/10.1038/s41431-019-0522-2

Le, T. B. K., Imakaev, M. V., Mirny, L. A., & Laub, M. T. (2013). High-Resolution Mapping of the Spatial Organization of a Bacterial Chromosome. *Science*, *342*(6159), 731–734. https://doi.org/10.1126/science.1242059

Ledermann, B., & Bürki, K. (1991). Establishment of a germ-line competent C57BL/6 embryonic stem cell line. *Experimental Cell Research*, *197*(2), 254–258. https://doi.org/10.1016/0014-4827(91)90430-3

Leon, A. C., & Heo, M. (2009). Sample sizes required to detect interactions between two binary fixed-effects in a mixed-effects linear regression model. *Computational Statistics & Data Analysis*, *53*(3), 603–608. https://doi.org/10.1016/j.csda.2008.06.010

Li, Z., Schulz, M. H., Look, T., Begemann, M., Zenke, M., & Costa, I. G. (2019). Identification of transcription factor binding sites using ATAC-seq. *Genome Biology*, *20*(1), 45. https://doi.org/10.1186/s13059-019-1642-2

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., & Dekker, J. (2009). Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science*, *326*(5950), 289–293. https://doi.org/10.1126/science.1181369

Liu, H., Luo, X., Niu, L., Xiao, Y., Chen, L., Liu, J., Wang, X., Jin, M., Li, W., Zhang, Q., & Yan, J. (2017). Distant eQTLs and Non-coding Sequences Play Critical Roles in Regulating Gene Expression and Quantitative Trait Variation in Maize. *Molecular Plant*, *10*(3), 414–426. https://doi.org/10.1016/j.molp.2016.06.016

Loukinov, D. I., Pugacheva, E., Vatolin, S., Pack, S. D., Moon, H., Chernukhin, I., Mannan, P., Larsson, E., Kanduri, C., Vostrov, A. A., Cui, H., Niemitz, E. L., Rasko, J. E. J., Docquier, F. M., Kistler, M., Breen, J. J., Zhuang, Z., Quitschke, W. W., Renkawitz, R., … Lobanenkov, V. V. (2002). BORIS, a novel male germ-line-specific protein associated with epigenetic reprogramming events, shares the same 11-zinc-finger domain with CTCF, the insulator protein involved in reading imprinting marks in the soma. *Proceedings of the National Academy of Sciences*, *99*(10), 6806–6811. https://doi.org/10.1073/pnas.092123699

Luo, S., Lu, J. Y., Liu, L., Yin, Y., Chen, C., Han, X., Wu, B., Xu, R., Liu, W., Yan, P., Shao, W., Lu, Z., Li, H., Na, J., Tang, F., Wang, J., Zhang, Y. E., & Shen, X. (2016). Divergent lncRNAs Regulate Gene Expression and Lineage Differentiation in Pluripotent Cells. *Cell Stem Cell*, *18*(5), 637–652. https://doi.org/10.1016/j.stem.2016.01.024

Maguire, E. M., Pearce, S. W. A., & Xiao, Q. (2019). Foam cell formation: A new target for fighting atherosclerosis and cardiovascular disease. *Vascular Pharmacology*, *112*, 54–71. https://doi.org/10.1016/j.vph.2018.08.002

Maji, A., Padinhateeri, R., & Mitra, M. K. (2020). The Accidental Ally: Nucleosome Barriers Can Accelerate Cohesin-Mediated Loop Formation in Chromatin. *Biophysical Journal*, *119*(11), 2316–2325. https://doi.org/10.1016/j.bpj.2020.10.014

Mak, I. W., Evaniew, N., & Ghert, M. (2014). Lost in translation: Animal models and clinical trials in cancer treatment. *American Journal of Translational Research*, *6*(2), 114–118.

Martin, G. R. (1981). Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. *Proceedings of the National Academy of Sciences*, *78*(12), 7634–7638. https://doi.org/10.1073/pnas.78.12.7634

Mason, D. M., Weber, C. R., Parola, C., Meng, S. M., Greiff, V., Kelton, W. J., & Reddy, S. T. (2018). High-throughput antibody engineering in mammalian cells by CRISPR/Cas9-mediated homology-directed mutagenesis. *Nucleic Acids Research*, *46*(14), 7436–7449. https://doi.org/10.1093/nar/gky550

Maurano, M. T., Wang, H., John, S., Shafer, A., Canfield, T., Lee, K., & Stamatoyannopoulos, J. A. (2015). Role of DNA Methylation in Modulating Transcription Factor Occupancy. *Cell Reports*, *12*(7), 1184–1195. https://doi.org/10.1016/j.celrep.2015.07.024

McArthur, E., & Capra, J. A. (2021). Topologically associating domain boundaries that are stable across diverse cell types are evolutionarily constrained and enriched for heritability. *The American Journal of Human Genetics*, *108*(2), 269–283. https://doi.org/10.1016/j.ajhg.2021.01.001

Merkin, J., Russell, C., Chen, P., & Burge, C. B. (2012). Evolutionary Dynamics of Gene and Isoform Regulation in Mammalian Tissues. *Science*, *338*(6114), 1593–1599. https://doi.org/10.1126/science.1228186

Morgan, A. P. (2016). argyle: An R Package for Analysis of Illumina Genotyping Arrays. *G3 Genes|Genomes|Genetics*, *6*(2), 281–286. https://doi.org/10.1534/g3.115.023739

Morgan, A. P., Fu, C.-P., Kao, C.-Y., Welsh, C. E., Didion, J. P., Yadgary, L., Hyacinth, L., Ferris, M. T., Bell, T. A., Miller, D. R., Giusti-Rodriguez, P., Nonneman, R. J., Cook, K. D., Whitmire, J. K., Gralinski, L. E., Keller, M., Attie, A. D., Churchill, G. A., Petkov, P., … Pardo-Manuel De Villena, F. (2016). The Mouse Universal Genotyping Array: From Substrains to Subspecies. *G3 Genes|Genomes|Genetics*, *6*(2), 263–279. https://doi.org/10.1534/g3.115.022087

Myers, S., Bowden, R., Tumian, A., Bontrop, R. E., Freeman, C., MacFie, T. S., McVean, G., & Donnelly, P. (2010). Drive Against Hotspot Motifs in Primates Implicates the *PRDM9* Gene in Meiotic Recombination. *Science*, *327*(5967), 876–879. https://doi.org/10.1126/science.1182363

Nelson, C. E., Hersh, B. M., & Carroll, S. B. (2004). The regulatory content of intergenic DNA shapes genome architecture. *Genome Biology*, *5*(4), R25. https://doi.org/10.1186/gb-2004-5-4-r25

Nora, E. P., Lajoie, B. R., Schulz, E. G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., Van Berkum, N. L., Meisig, J., Sedat, J., Gribnau, J., Barillot, E., Blüthgen, N., Dekker, J., & Heard, E. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, *485*(7398), 381–385. https://doi.org/10.1038/nature11049

Nuebler, J., Fudenberg, G., Imakaev, M., Abdennur, N., & Mirny, L. A. (2018). Chromatin organization by an interplay of loop extrusion and compartmental segregation. *Proceedings of the National Academy of Sciences*, *115*(29). https://doi.org/10.1073/pnas.1717730115

Oomen, M. E., Hansen, A. S., Liu, Y., Darzacq, X., & Dekker, J. (2019). CTCF sites display cell cycle–dependent dynamics in factor binding and nucleosome positioning. *Genome Research*, *29*(2), 236–249. https://doi.org/10.1101/gr.241547.118

Palmiter, R. D., Sandgren, E. P., Avarbock, M. R., Allen, D. D., & Brinster, R. L. (1991). Heterologous introns can enhance expression of transgenes in mice. *Proceedings of the National Academy of Sciences*, *88*(2), 478–482. https://doi.org/10.1073/pnas.88.2.478

Parvanov, E. D., Petkov, P. M., & Paigen, K. (2010). *Prdm9* Controls Activation of Mammalian Recombination Hotspots. *Science*, *327*(5967), 835–835. https://doi.org/10.1126/science.1181495

Perez-Martinez, P., Delgado-Lista, J., Garcia-Rios, A., Mc Monagle, J., Gulseth, H. L., Ordovas, J. M., Shaw, D. I., Karlström, B., Kiec-Wilk, B., Blaak, E. E., Helal, O., Malczewska-Malec, M., Defoort, C., Risérus, U., Saris, W. H. M., Lovegrove, J. A., Drevon, C. A., Roche, H. M., & Lopez-Miranda, J. (2011). Glucokinase Regulatory Protein Genetic Variant Interacts with Omega-3 PUFA to Influence Insulin Resistance and Inflammation in Metabolic Syndrome. *PLoS ONE*, *6*(6), e20555. https://doi.org/10.1371/journal.pone.0020555

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology*, *29*(1), 24–26. https://doi.org/10.1038/nbt.1754

Robinson, M. D., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, *11*(3), R25. https://doi.org/10.1186/gb-2010-11-3-r25

Rodríguez-Carballo, E., Lopez-Delisle, L., Zhan, Y., Fabre, P. J., Beccari, L., El-Idrissi, I., Huynh, T. H. N., Ozadam, H., Dekker, J., & Duboule, D. (2017). The *HoxD* cluster is a dynamic and resilient TAD boundary controlling the segregation of antagonistic regulatory landscapes. *Genes & Development*, *31*(22), 2264–2281. https://doi.org/10.1101/gad.307769.117

Ronald, J., Brem, R. B., Whittle, J., & Kruglyak, L. (2005). Local Regulatory Variation in Saccharomyces cerevisiae. *PLoS Genetics*, *1*(2), e25. https://doi.org/10.1371/journal.pgen.0010025

Sanderson, B., Fields, P. D., & Lloyd, M. (n.d.). *The Jackson Laboratory Computational Sciences Nextflow based analysis pipelines* (Pre-Release) [Nextflow, Python, R, Perl, Shell, Groovy]. The Jackson Laboratory. Retrieved December 15, 2020, from https://github.com/TheJacksonLaboratory/cs-nf-pipelines

Schnabel, L. V., Abratte, C. M., Schimenti, J. C., Southard, T. L., & Fortier, L. A. (2012). Genetic background affects induced pluripotent stem cell generation. *Stem Cell Research & Therapy*, *3*(4), 30. https://doi.org/10.1186/scrt121

Serebrenik, Y. V., & Shalem, O. (2018). CRISPR mutagenesis screening of mice. *Nature Cell Biology*, *20*(11), 1235–1237. https://doi.org/10.1038/s41556-018-0224-y

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., & Cavalli, G. (2012). Three-Dimensional Folding and

Functional Organization Principles of the Drosophila Genome. *Cell*, *148*(3), 458–472. https://doi.org/10.1016/j.cell.2012.01.010

Seyhan, A. A. (2019). Lost in translation: The valley of death across preclinical and clinical divide – identification of problems and overcoming obstacles. *Translational Medicine Communications*, *4*(1), 18. https://doi.org/10.1186/s41231-019-0050-7

Shen, Y., Yue, F., McCleary, D. F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenkov, V. V., & Ren, B. (2012). A map of the cis-regulatory sequences in the mouse genome. *Nature*, *488*(7409), 116–120. https://doi.org/10.1038/nature11243

Shibata, Y., Sheffield, N. C., Fedrigo, O., Babbitt, C. C., Wortham, M., Tewari, A. K., London, D., Song, L., Lee, B.-K., Iyer, V. R., Parker, S. C. J., Margulies, E. H., Wray, G. A., Furey, T. S., & Crawford, G. E. (2012). Extensive Evolutionary Changes in Regulatory Element Activity during Human Origins Are Associated with Altered Gene Expression and Positive Selection. *PLoS Genetics*, *8*(6), e1002789. https://doi.org/10.1371/journal.pgen.1002789

Shinkai, S., Nozaki, T., Maeshima, K., & Togashi, Y. (2016). Dynamic Nucleosome Movement Provides Structural Information of Topological Chromatin Domains in Living Human Cells. *PLOS Computational Biology*, *12*(10), e1005136. https://doi.org/10.1371/journal.pcbi.1005136

Shorter, J. R., Odet, F., Aylor, D. L., Pan, W., Kao, C.-Y., Fu, C.-P., Morgan, A. P., Greenstein, S., Bell, T. A., Stevans, A. M., Feathers, R. W., Patel, S., Cates, S. E., Shaw, G. D., Miller, D. R., Chesler, E. J., McMillian, L., O'Brien, D. A., & Villena, F. P.-M. D. (2017). Male Infertility Is Responsible for Nearly Half of the Extinction Observed in the Mouse Collaborative Cross. *Genetics*, *206*(2), 557–572. https://doi.org/10.1534/genetics.116.199596

Silva, J., Barrandon, O., Nichols, J., Kawaguchi, J., Theunissen, T. W., & Smith, A. (2008). Promotion of Reprogramming to Ground State Pluripotency by Signal Inhibition. *PLoS Biology*, *6*(10), e253. https://doi.org/10.1371/journal.pbio.0060253

Skelly, D. A., Czechanski, A., Byers, C., Aydin, S., Spruce, C., Olivier, C., Choi, K., Gatti, D. M., Raghupathy, N., Keele, G. R., Stanton, A., Vincent, M., Dion, S., Greenstein, I., Pankratz, M., Porter, D. K., Martin, W., O'Connor, C., Qin, W., … Reinholdt, L. G. (2020). Mapping the Effects of Genetic Variation on Chromatin State and Gene Expression Reveals Loci That Control Ground State Pluripotency. *Cell Stem Cell*, *27*(3), 459-469.e8. https://doi.org/10.1016/j.stem.2020.07.005

Skene, P. J., & Henikoff, S. (2017). An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *ELife*, *6*, e21856. https://doi.org/10.7554/eLife.21856

Slate, J. (2004). INVITED REVIEW: Quantitative trait locus mapping in natural populations: progress, caveats and future directions: QTL MAPPING IN NATURAL POPULATIONS. *Molecular Ecology*, *14*(2), 363–379. https://doi.org/10.1111/j.1365-294X.2004.02378.x

Slate, J., Santure, A. W., Feulner, P. G. D., Brown, E. A., Ball, A. D., Johnston, S. E., & Gratten, J. (2010). Genome mapping in intensively studied wild vertebrate populations. *Trends in Genetics*, *26*(6), 275–284. https://doi.org/10.1016/j.tig.2010.03.005

Sloan, C. A., Chan, E. T., Davidson, J. M., Malladi, V. S., Strattan, J. S., Hitz, B. C., Gabdank, I., Narayanan, A. K., Ho, M., Lee, B. T., Rowe, L. D., Dreszer, T. R., Roe, G., Podduturi, N. R., Tanaka, F., Hong, E. L., & Cherry, J. M. (2016). ENCODE data at the ENCODE portal. *Nucleic Acids Research*, *44*(D1), D726–D732. https://doi.org/10.1093/nar/gkv1160

Solberg Woods, L. C. (2014). QTL mapping in outbred populations: Successes and challenges. *Physiological Genomics*, *46*(3), 81–90. https://doi.org/10.1152/physiolgenomics.00127.2013

Song, Y., Liang, Z., Zhang, J., Hu, G., Wang, J., Li, Y., Guo, R., Dong, X., Babarinde, I. A., Ping, W., Sheng, Y.-L., Li, H., Chen, Z., Gao, M., Chen, Y., Shan, G., Zhang, M. Q., Hutchins, A. P., Fu, X.-D., & Yao, H. (2022). CTCF functions as an insulator for somatic genes and a chromatin remodeler for pluripotency genes during reprogramming. *Cell Reports*, *39*(1), 110626. https://doi.org/10.1016/j.celrep.2022.110626

Srivastava, A., Morgan, A. P., Najarian, M. L., Sarsani, V. K., Sigmon, J. S., Shorter, J. R., Kashfeen, A., McMullan, R. C., Williams, L. H., Giusti-Rodríguez, P., Ferris, M. T., Sullivan, P., Hock, P., Miller, D. R., Bell, T. A., McMillan, L., Churchill, G. A., & De Villena, F. P.-M. (2017). Genomes of the Mouse Collaborative Cross. *Genetics*, *206*(2), 537–556. https://doi.org/10.1534/genetics.116.198838

Stoica, P., & Selen, Y. (2004). Model-order selection. *IEEE Signal Processing Magazine*, *21*(4), 36–47. https://doi.org/10.1109/MSP.2004.1311138

Stryke, D. (2003). BayGenomics: A resource of insertional mutations in mouse embryonic stem cells. *Nucleic Acids Research*, *31*(1), 278–281. https://doi.org/10.1093/nar/gkg064

Su, J., Teichmann, S. A., & Down, T. A. (2010). Assessing Computational Methods of Cis-Regulatory Module Prediction. *PLoS Computational Biology*, *6*(12), e1001020. https://doi.org/10.1371/journal.pcbi.1001020

Svenson, K. L., Gatti, D. M., Valdar, W., Welsh, C. E., Cheng, R., Chesler, E. J., Palmer, A. A., McMillan, L., & Churchill, G. A. (2012). High-Resolution Genetic Mapping Using the Mouse Diversity Outbred Population. *Genetics*, *190*(2), 437–447. https://doi.org/10.1534/genetics.111.132597

Taberlay, P. C., Achinger-Kawecka, J., Lun, A. T. L., Buske, F. A., Sabir, K., Gould, C. M., Zotenko, E., Bert, S. A., Giles, K. A., Bauer, D. C., Smyth, G. K., Stirzaker, C., O'Donoghue, S. I., & Clark, S. J. (2016). Three-dimensional disorganization of the cancer genome occurs coincident with long-range genetic and epigenetic alterations. *Genome Research*, *26*(6), 719–731. https://doi.org/10.1101/gr.201517.115

Tang, Z., Luo, O. J., Li, X., Zheng, M., Zhu, J. J., Szalaj, P., Trzaskoma, P., Magalska, A., Wlodarczyk, J., Ruszczycki, B., Michalski, P., Piecuch, E., Wang, P., Wang, D., Tian, S. Z., Penrad-Mobayed, M., Sachs, L. M., Ruan, X., Wei, C.-L., …

Ruan, Y. (2015). CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. *Cell*, *163*(7), 1611–1627. https://doi.org/10.1016/j.cell.2015.11.024

The Complex Trait Consortium. (2004). The Collaborative Cross, a community resource for the genetic analysis of complex traits. *Nature Genetics*, *36*(11), 1133–1137. https://doi.org/10.1038/ng1104-1133

The ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74. https://doi.org/10.1038/nature11247

Trudeau, R. J., & Trudeau, R. J. (1993). *Introduction to graph theory*. Dover Pub.

Tuttle, A. H., Philip, V. M., Chesler, E. J., & Mogil, J. S. (2018). Comparing phenotypic variation between inbred and outbred mice. *Nature Methods*, *15*(12), 994–996. https://doi.org/10.1038/s41592-018-0224-7

Tyler, A. L., Donahue, L. R., Churchill, G. A., & Carter, G. W. (2016). Weak Epistasis Generally Stabilizes Phenotypes in a Mouse Intercross. *PLOS Genetics*, *12*(2), e1005805. https://doi.org/10.1371/journal.pgen.1005805

Van Ruiten, M. S., & Rowland, B. D. (2021). On the choreography of genome folding: A grand pas de deux of cohesin and CTCF. *Current Opinion in Cell Biology*, *70*, 84–90. https://doi.org/10.1016/j.ceb.2020.12.001

Villar, D., Berthelot, C., Aldridge, S., Rayner, T. F., Lukk, M., Pignatelli, M., Park, T. J., Deaville, R., Erichsen, J. T., Jasinska, A. J., Turner, J. M. A., Bertelsen, M. F., Murchison, E. P., Flicek, P., & Odom, D. T. (2015). Enhancer Evolution across 20 Mammalian Species. *Cell*, *160*(3), 554–566. https://doi.org/10.1016/j.cell.2015.01.006

Waddington, C. H. (2012). The Epigenotype. *International Journal of Epidemiology*, *41*(1), 10–13. https://doi.org/10.1093/ije/dyr184

Wei, C., Jia, L., Huang, X., Tan, J., Wang, M., Niu, J., Hou, Y., Sun, J., Zeng, P., Wang, J., Qing, L., Ma, L., Liu, X., Tang, X., Li, F., Jiang, S., Liu, J., Li, T., Fan, L., … Ding, J. (2022). CTCF organizes inter-A compartment interactions through RYBP-dependent phase separation. *Cell Research*, *32*(8), 744–760. https://doi.org/10.1038/s41422-022-00676-0

West, D. B., Engelhard, E. K., Adkisson, M., Nava, A. J., Kirov, J. V., Cipollone, A., Willis, B., Rapp, J., De Jong, P. J., & Lloyd, K. C. (2016). Transcriptome Analysis of Targeted Mouse Mutations Reveals the Topography of Local Changes in Gene Expression. *PLOS Genetics*, *12*(2), e1005691. https://doi.org/10.1371/journal.pgen.1005691

Weth, O., & Renkawitz, R. (2011). CTCF function is modulated by neighboring DNA binding factors. *Biochemistry and Cell Biology*, *89*(5), 459–468. https://doi.org/10.1139/o11-033

Workman, J. L., & Kingston, R. E. (1998). ALTERATION OF NUCLEOSOME STRUCTURE AS A MECHANISM OF TRANSCRIPTIONAL REGULATION. *Annual Review of Biochemistry*, *67*(1), 545–579. https://doi.org/10.1146/annurev.biochem.67.1.545

Wutz, G., Várnai, C., Nagasaka, K., Cisneros, D. A., Stocsits, R. R., Tang, W., Schoenfelder, S., Jessberger, G., Muhar, M., Hossain, M. J., Walther, N., Koch, B., Kueblbeck, M., Ellenberg, J., Zuber, J., Fraser, P., & Peters, J. (2017). Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins. *The EMBO Journal*, *36*(24), 3573–3599. https://doi.org/10.15252/embj.201798004

Yakushiji-Kaminatsui, N., Lopez-Delisle, L., Bolt, C. C., Andrey, G., Beccari, L., & Duboule, D. (2018). Similarities and differences in the regulation of HoxD genes during chick and mouse limb development. *PLOS Biology*, *16*(11), e3000004. https://doi.org/10.1371/journal.pbio.3000004

You, Q., Cheng, A. Y., Gu, X., Harada, B. T., Yu, M., Wu, T., Ren, B., Ouyang, Z., & He, C. (2021). Direct DNA crosslinking with CAP-C uncovers transcription-dependent chromatin organization at high resolution. *Nature Biotechnology*, *39*(2), 225–235. https://doi.org/10.1038/s41587-020-0643-8

Yusufzai, T. M., Tagami, H., Nakatani, Y., & Felsenfeld, G. (2004). CTCF Tethers an Insulator to Subnuclear Sites, Suggesting Shared Insulator Mechanisms across Species. *Molecular Cell*, *13*(2), 291–298. https://doi.org/10.1016/S1097-2765(04)00029-2

Ziebarth, J. D., Bhattacharya, A., & Cui, Y. (2012). CTCFBSDB 2.0: A database for CTCF-binding sites and genome organization. *Nucleic Acids Research*, *41*(D1), D188–D194. https://doi.org/10.1093/nar/gks1165

Zou, F., Gelfond, J. A. L., Airey, D. C., Lu, L., Manly, K. F., Williams, R. W., & Threadgill, D. W. (2005). Quantitative Trait Locus Analysis Using Recombinant Inbred Intercrosses. *Genetics*, *170*(3), 1299–1311. https://doi.org/10.1534/genetics.104.035709

Zuo, Z., Roy, B., Chang, Y. K., Granas, D., & Stormo, G. D. (2017). Measuring quantitative effects of methylation on transcription factor–DNA binding affinity. *Science Advances*, *3*(11), eaao1799. https://doi.org/10.1126/sciadv.aao1799