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

Lauren Kuffler, Daniel A. Skelly, Anne Czechanski, Steven C. Munger, Christopher L. Baker, Laura G. Reinholdt, Gregory W. Carter

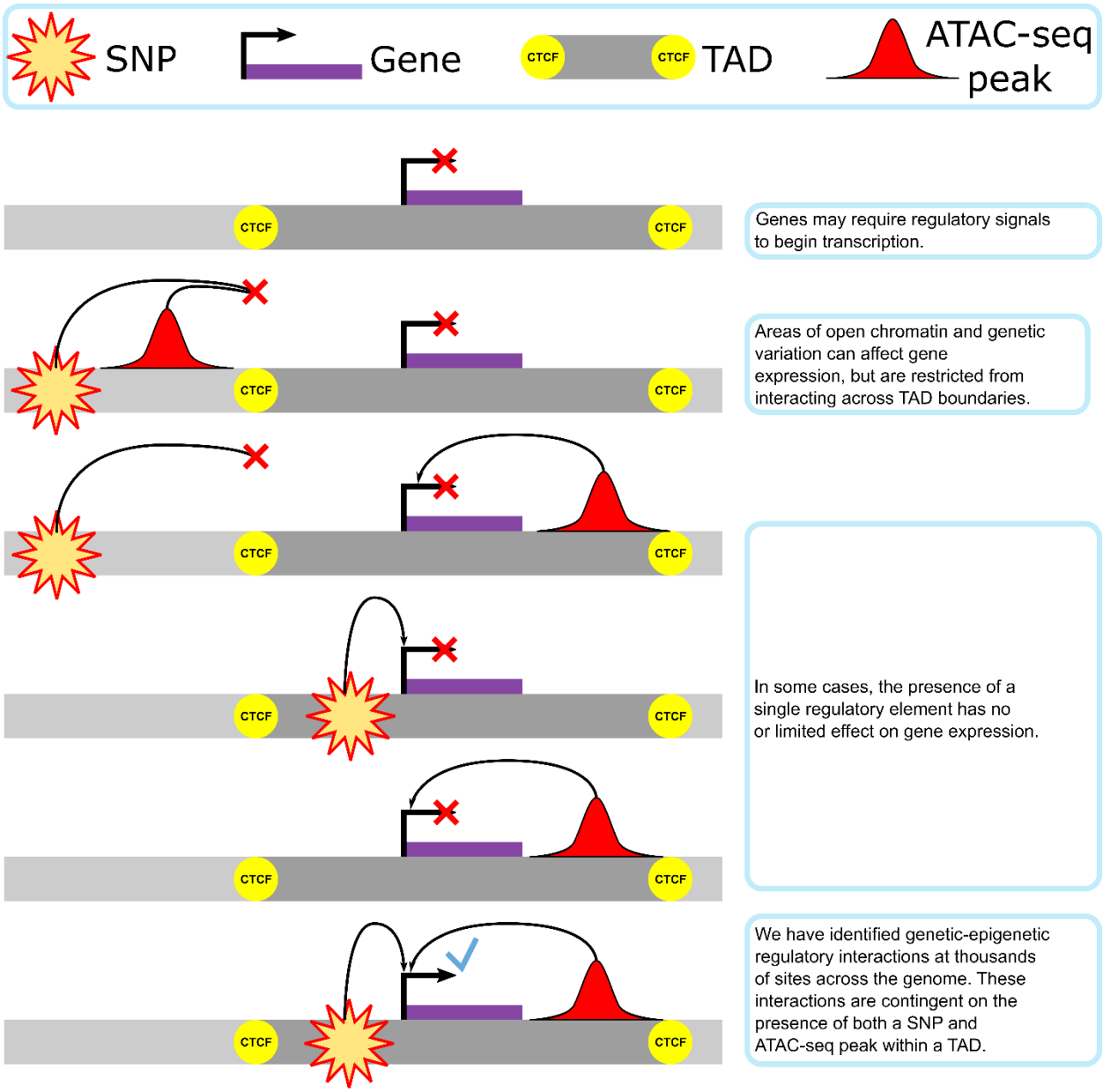
The Jackson Laboratory, Bar Harbor, ME, USA

Corresponding author: Gregory W. Carter, gregory.carter@jax.org

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 openness 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 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 regulatory elements in genetically diverse populations provide a means to infer 3D genome structure. 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 TAD boundaries operating during early development. These findings provide evidence that genetic and epigenetic factors operate within the context of three-dimensional chromatin structure.

Graphical Abstract



Introduction

Local regulatory mechanisms within the genome and their interaction with chromatin structure give rise to subtle variations in gene expression. 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. Studies are generally limited to examination of individual regions or overlapping single-nucleotide polymorphisms (SNPs) and open chromatin peaks with limited investigation into how these regulatory elements combine to affect gene transcription [1,2]. Genetic analyses are a powerful approach that allow the study of these interactions. Conversely, 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.

Physical distance along a linear genome is a common metric for determining whether a putative regulatory element will affect a given gene's transcription. This has led to the concept of a "local area" in which most regulatory interactions take place. Many have used a 100 to 500 kb flanking window around a gene to encapsulate the landscape of local epigenetic effects [3–5]. This interval size is motivated more by convenience than rigorous evidence. Incidentally, the estimated scale of this local area is approximately equivalent to the size of a topologically associated domain (TAD). TADs are chromatin loops with an average length of approximately 1.1 Mb that are joined via the architectural protein CTCF and the cohesin protein complex at known binding motifs [6,7]. TADs are believed to be mostly cell-type invariant and conserved across species [6,8,9]. The current hypothesis is that TADs 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. Reverse genetic approaches like saturation mutagenesis can be used to introduce functionally informative genetic variation. But such screens are difficult in mammals, and even modern CRISPR-Cas9 based approaches are still limited in their scope and application outside of immortalized cell lines [10].

Samples from genetically diverse populations permit comprehensive study of complex genetic interaction and non-coding regulation on a genome-wide scale [11]. The Diversity Outbred mouse population (DO) is a heterogeneous stock composed of mice derived from eight founder strains, including three wild-derived strains [12]. 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 [13,14]. 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 generation 19 DO embryos, 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 [15].

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 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, and identify potential regulatory functions underlying a set of interactions.

Methods

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 [15]. 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).[16] Genotyping was performed by Giga Mouse Universal Genotyping Array (GigaMUGA) [17]. Aneuploidies were removed with the argyle R package [18]. QTL2 haplotype reconstruction, normalization and pseudoprobe processing was carried out as previously described[15]. Samples with X0 genotypes were removed, and the union of all samples with the required data types resulted in 176 samples.

SNPs were subset to reflect linkage disequilibrium and haplotype prediction confidence, selecting only the closest SNP to grid points of GigaMUGA haplotype probabilities, evenly spaced by genetic distance [19,20]. This produced a subset of 68,413 SNPs which were most likely to match the actual haplotype for each sample at a given locus (Supplementary Information).

RNA-seq, ATAC-seq and haplotype data were processed into SQLite databases and fit to a regression model using the stats::step() function in R. We modeled the expression of each gene by linear regression with a SNP-by-ATAC interaction term:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon_i \qquad \text{(Equation 1)}$$

where y_i =RNA abundance ($\log_2$TPM), x_1 = SNP, and x_2 = ATAC intensity (TMM). SNPs were coded as 0 for reference, 1 for heterozygote, and 2 for homozygote. Multi-allelic variants were not present in the subset used and were therefore not accounted for in the method. Models that retained the $\beta_3 x_1 x_2$ term by passing the default Akaike Information Criterion cutoff were deemed to be interacting. These were used as the input for all further analyses.

This regression model was compared to two null models: one in which neither a genetic variant nor open chromatin element affect a gene's expression additively or interactively ($\beta_1 = \beta_2 = \beta_3 = 0$ in Equation 1), and one where both affect gene expression, but they do so independently of each other ($\beta_3 = 0$ in Equation 1). The former null model produced more conservative results

and was consistent with the latter null model. Significance of the full model (Equation 1) was estimated relative to the two null models with an F-ratio test. F statistics were calculated with the R function `pf()`. Given the possibility of overfitting as the number of model terms increases and the high number of models tested (see Results), a Bonferroni adjusted p value cutoff of $p < 1x10^{-7}$ was selected (hereafter referred to as "significant models"). Bonferroni corrections were calculated via `stats::p.adjust()`.

mESC TAD data was derived from previously published work [6]. in a liftover from mm9 to mm10 via UCSC genome browser [21]. Chromosome information for mm10 retrieved from the Integrative Genomics Viewer [22].

DNA motif and binding site discovery was performed using the MEMEsuite tools MEME, STREME, and Tomtom, all set to default parameters (See Supplemental Data) [23–26]. SNP data were imputed from DO founder genomes referenced from release v3 (REL-1303-SNPs_Indels-GRCm238) of the Mouse Genome Project through the Sanger Institute [27,28].

We downloaded the CTCF ChIP-seq in C57BL/6 Bruce4 and 129/Ola E14TG2a.4 mESCs from the ENCODE portal[29] with the following identifiers: ENCSR000CCB and ENCSR362VNF [30]. These were also used to estimate the locations of Smad3 activity. CTCF and SMAD3 binding motifs were retrieved from HOCOMOCO [31]. with the identifiers CTCF_MOUSE.H11MO.0.A and SMAD3_MOUSE.H11MO.0.B. These were overlapped with our ATAC-seq dataset via FIMO [27].

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

Results

Genetic-Epigenetic Interactions are Pervasive and enriched within TAD boundaries

Although there are numerous reports detailing examples of genetic variants and chromatin state altering gene expression, the interactions between these features have yet to be systematically explored. We aimed to determine the extent to which genetic-epigenetic regulatory interactions occur on a systemic level, and whether these interactions exhibit any bias in their location within the genome. To explore these interactions, we utilized paired transcriptome and chromatin accessibility profiling of Diversity Outbred mESCs [15]. These samples incorporate genetic backgrounds from eight well-characterized founder strains from three *M. musculus* subspecies, randomly outbred for 19 generations. Thus, the resultant data constitute a unique resource for addressing questions of genetic-epigenetic regulatory interactions.

To search for interaction effects on a global scale, we developed a method of analysis that could be broadly applied, and could permit discrimination between co-occurrence and interaction effects. To that end, we combined regression model fitting (Equation 1) and model selection via information criteria in order to test whether gene expression was affected by a

genetic variant, an open chromatin region, both independently of each other, or both working in concert.

To account for the resolution limits on SNP effects imposed by linkage disequilibrium, we subsetted our SNP data to 68,413 high confidence SNPs evenly spaced by genetic distance to reflect this resolution limit (Methods). We collated these with 102,173 ATAC-seq peaks and 13,631 genes with RNA-seq coverage within our sample set.

To test the interaction models, randomly selected trios of intrachromosomal genes, SNPs, and ATAC-seq peaks were tested with the interaction model. A total of 102,104 ATAC-seq peaks, 68,413 SNPs, and 13,631 genes were used to model 39,021,625 combinations. This resulted in 27,509 significant models (Bonferroni adjusted $p < 1x10^{-7}$), or 0.007% of all models, involving 19,145 ATAC-seq peaks, 14,476 SNPs, and 1,286 genes (18.75%, 21.16%, and 8.52% of the total input respectively). Complete results are tallied in Tables S1 and S2. The number of significant models was four orders of magnitude greater than expected, thus indicating the presence of true associations (Fig. 1A). The locations of significant associations indicated interactions were much more likely to occur within the linkage disequilibrium block that contains the gene (Fig. 1B), matching prior work that the majority of detectable genetic effects originate in local regulatory regions [32–34]. Approximately half (49%) of all significant models contained a SNP or ATAC peak within 4 Mb of the gene they affected.

As a result of these findings, we began considering potential chromatin features that might influence the local area for genetic-epigenetic interactions. We hypothesized that TADs may play a role in determining the prevalence of interactions, as they are a 3D genome structure known to enable enhancer access to local genes. Therefore, we generated a focused subset of regression models, testing all potential combinations between each gene and every SNP and ATAC peak that fall within one TAD of its transcription start site in either direction.

Models with at least one interacting element within the same TAD as the gene were 4.5 times more likely to reach our threshold, compared to models that did not (3.9% versus 0.9%), involving 3836 SNPs and 3725 ATAC-seq peaks affecting the expression of 896 genes (Table S1-S2). ATAC-seq peaks involved in interactions favored results between CTCF binding sites associated with TAD formation (Fig. S1A). Regression models that did not identify interaction involvement also show this local, TAD-bound organization of ATAC-seq peaks (Fig. S1B). Interacting regulatory elements as defined by our model are thus clustered within the same TAD as the gene they affect. This provides a possible structural explanation for previously observed local regulatory ranges.

Interacting elements in genetically diverse samples escape conventional discovery methods

We next evaluated whether genetic-epigenetic interactions provide information that genotyping or ATAC-seq alone cannot. To this end, we compared interacting models, found to represent

32.29% of significant intra-TAD models (Fig. 1C), with models associated with SNPs or ATAC-seq peaks without interactions. We broke down the overlap between intra-TAD regression model results by SNP and ATAC-seq peaks separately. We found that 27.10% of SNPs are shared between interacting models and additive models and 22.59% are shared with models that contain no ATAC-seq peak contribution. In contrast, few models contain *only* an ATAC-seq peak contribution, and 37.45% of ATAC-seq peaks are exclusive to interacting models (Fig. 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. S1C-D). These results suggest that although genetic variants are the primary driver of variation in expression, 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, this suggests that ATAC-seq data alone is a less effective predictor of gene expression in a genetically diverse population.

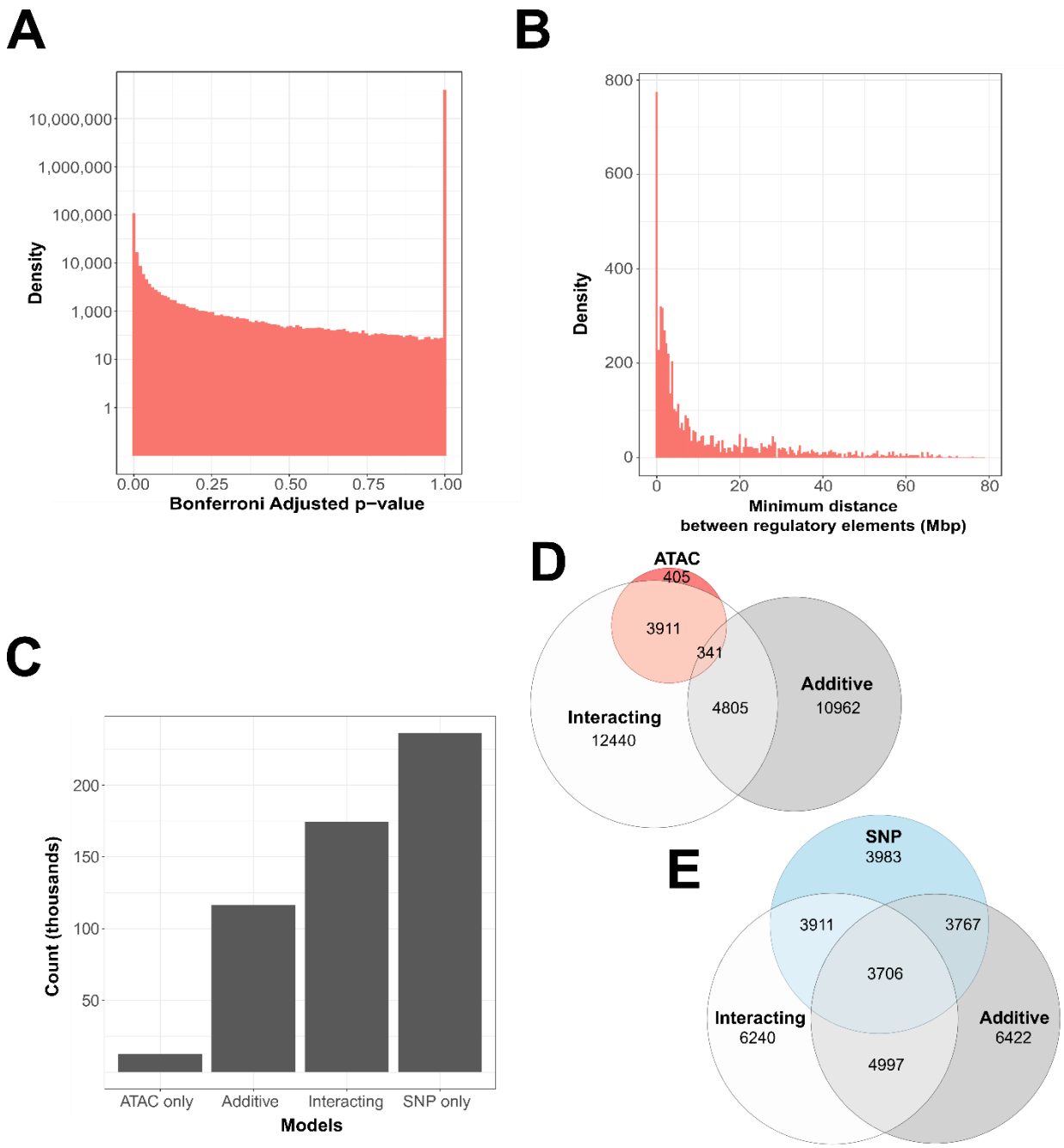


Figure 1. Interacting and additive models are abundant and favor local genes. **A.** Density of Bonferroni Adjusted p-value distribution in randomly sampled intrachromosomal models. **B.** Density of interacting models where adj. $p < 1x10^{-7}$, by minimum distance between regulatory elements. **C.** Distribution of model term retention across intra-TAD models. **D & E.** Intra-TAD ATAC-seq peak and SNP participation in single-term, additive, and interacting models.

TAD boundaries limit genetic-epigenetic interactions

The specific preference for 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. 2A), similar to reported segregation of enhancer elements [35] . 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.

We found that TADs contained more ATAC peaks that interact with local SNPs to affect expression of a resident gene (Fig. 2B). 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. S2). We also identified 29 individual ATAC-seq peaks that were involved in more than 200 interactions each, which are distinctly visible within the aggregate view (Fig. 2B). 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 [36].

Revising constraints on local regulator 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 CTCF Hi-C or ChIA-PET datasets, further limiting many researchers that may want to study genetic-epigenetic 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 a gene TSS was required (Fig. 2C). The density of results using linear DNA sequence does not necessarily experience a linear drop-off over this distance, due to variable TAD lengths (Fig. 3E-F).

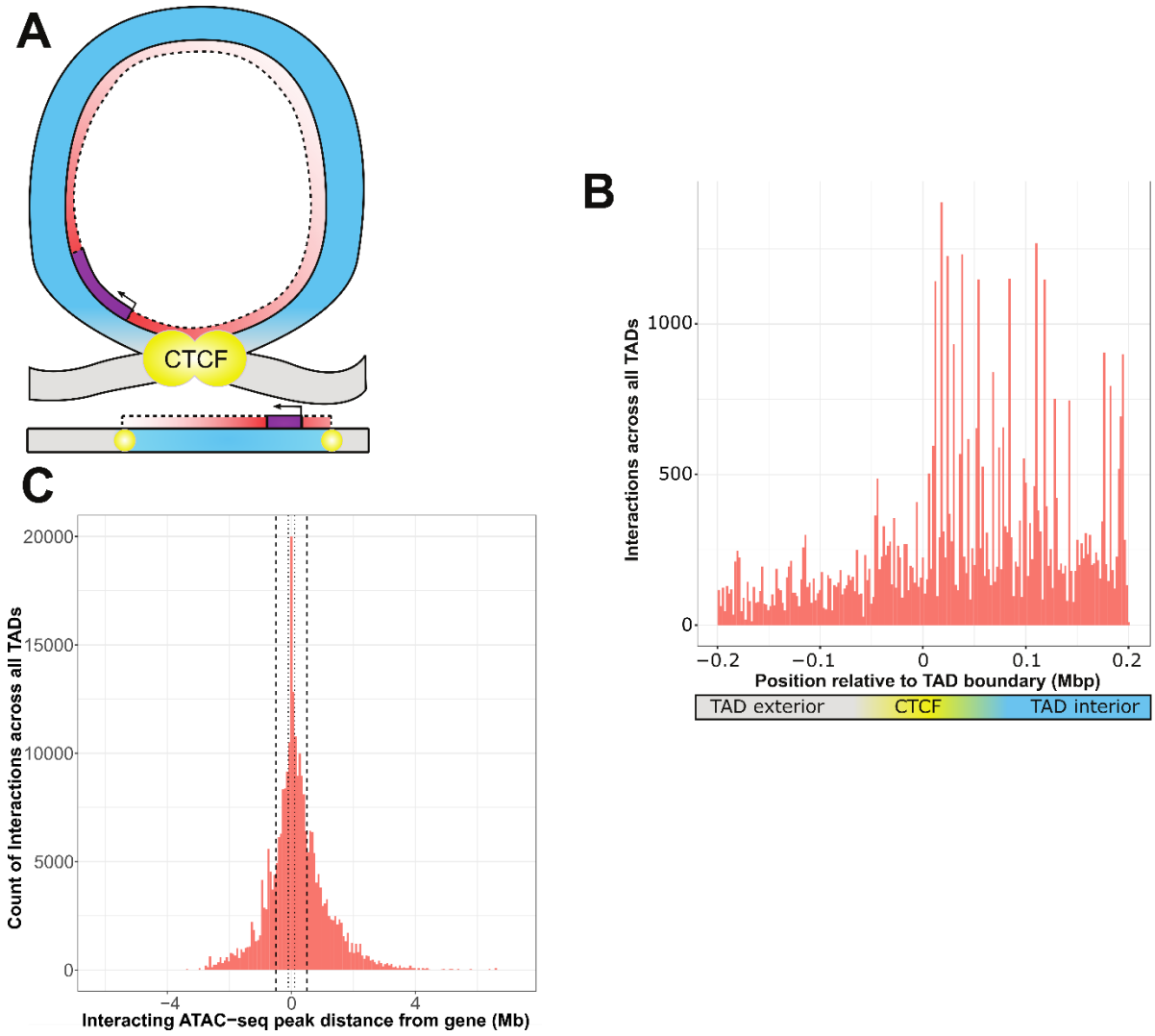


Figure 2. ATAC-seq peaks that interact with SNPs generally reside within the affected gene's TAD. **A.** Schematic of a TAD loop, including gene (purple) and density of interacting model elements (red). Loop interior is in blue, exterior DNA is gray, and CTCF binding sites are in yellow. **B.** Location 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. Local area cutoffs of 100 kb and 500 kb flanking regions are marked.

We next considered the location of these interacting elements with regards to genes and known regulatory elements. ATAC-seq peaks involved in genetic-epigenetic interactions were analyzed by overlap with mm10 NCBI RefSeq genes and selected Enhancer Atlas 2 datasets (Fig. S3A-B). 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. S3C), indicating that this was not simply reflecting the relative density of genes within various TADs across the genome.

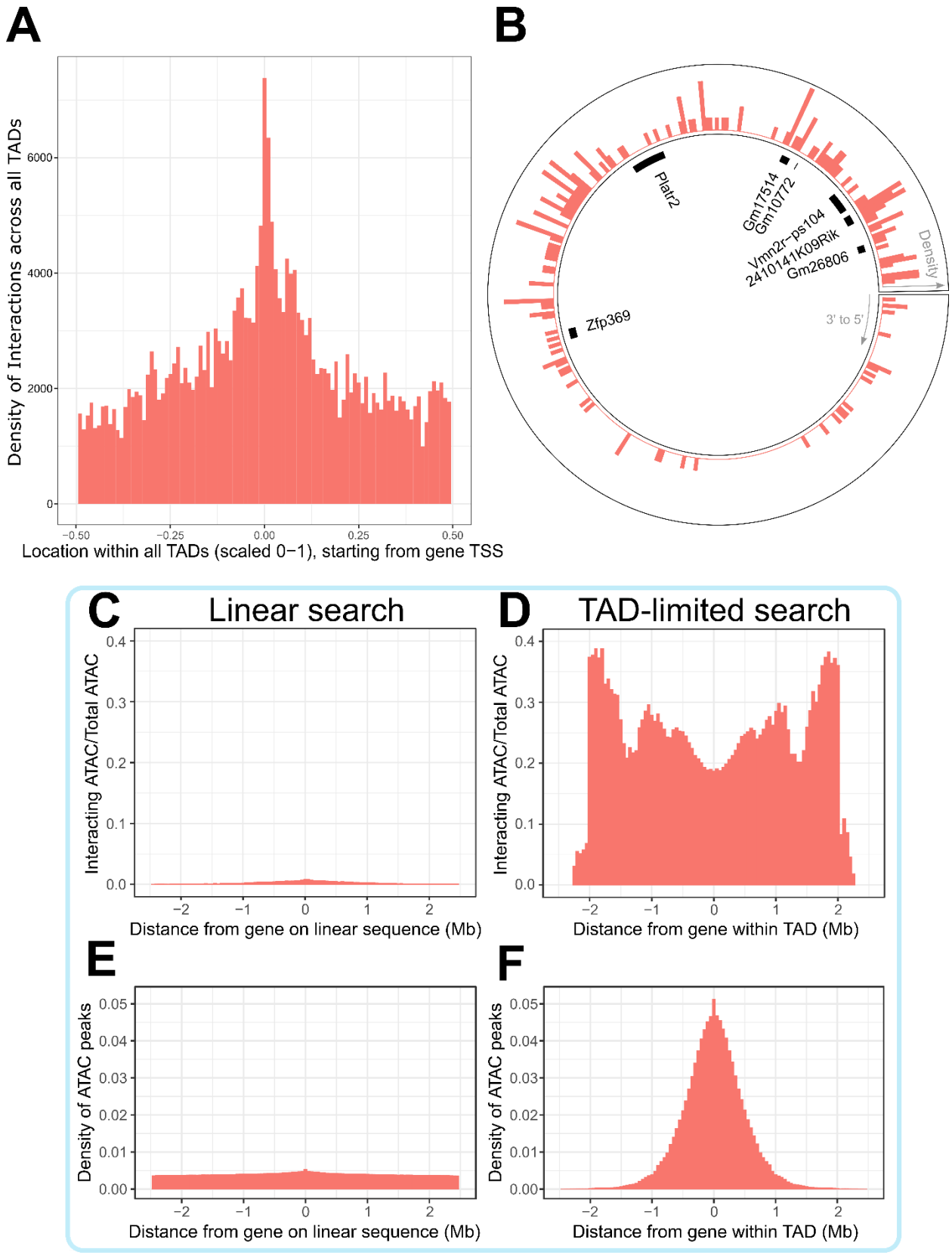

A
Density of Interactions across all TADs
6000
4000
2000
0
−0.50
−0.25
0.00
0.25
0.50
Location within all TADs (scaled 0−1), starting from gene TSS
B
Platr2
Gm17514
Gm10772
Vmn2r−ps104
2410141K09Rik
Gm26806
Zfp369
3' to 5'
Density
C
Linear search
Interacting ATAC/Total ATAC
Distance from gene on linear sequence (Mb)
D
TAD-limited search
Interacting ATAC/Total ATAC
Distance from gene within TAD (Mb)
E
Density of ATAC peaks
Distance from gene on linear sequence (Mb)
F
Density of ATAC peaks
Distance from gene within TAD (Mb)

Figure 3. TADs provide context for interactions and increase interaction search efficacy. **A.** Counts of intra-TAD ATAC-seq peaks involved in all 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.

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 interactions centered at the gene promoter that reaches its minimum halfway around the TAD loop from the gene (Fig. 3A).

We sought to determine whether gene-centered, TAD-internal searches for interacting elements were potentially a more efficient method for searching for 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).

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. 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.

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 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. 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. S3D).

Interestingly, there is a depletion in the interaction rate of ATAC-seq peaks flanking the gene (Fig. 3D), which is not noted in raw interaction counts (Fig 3A). 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.

Our results indicate that the local area for genetic-epigenetic interaction is not only constrained by TADs, it is 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.

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 interaction effects would display patterns that corresponded to positive and negative regulatory roles with varying strengths.

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 [37,38]. However, in the context of 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.

We found that models indicating redundancy or interference were the most common overall, totaling 39.54% of all models (Fig. S4A). Synergistic effects were found in 16.74% of all models. These two observations, suggest that a greater proportion of regulatory interactions attenuate gene expression, rather than strengthening it. While ATAC-seq peaks are often correlated with increased gene expression [39], we were surprised to find that an increase in ATAC-seq signal had a negative effect on gene expression in 40.73% of models (Fig. S4A). 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.

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 suppressive regulatory factors.

To test this hypothesis, we tested for DNA motif enrichment analysis using MEMEsuite (Fig. 4A). We selected the subset of ATAC-seq peaks involved in 10 or more significant 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.

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 ($p < 1.5x10^{-8}$). 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).

We next examined whether these peak locations contain SNPs that might alter CTCF binding potential. Imputing from the founder genomes of the Diversity Outbred 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. 4B). However, in motifs associated with regression models, the density of SNPs was approximately equal across all bases (Fig. 4C). These bases have previously been identified as having high protein-DNA binding energies with the canonical sequence [39,40].

Most genomic sequences matching the CTCF binding motif are not known to be bound, according to ChIP-seq experiments [41]. 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 [42,43], 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 1kb 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.

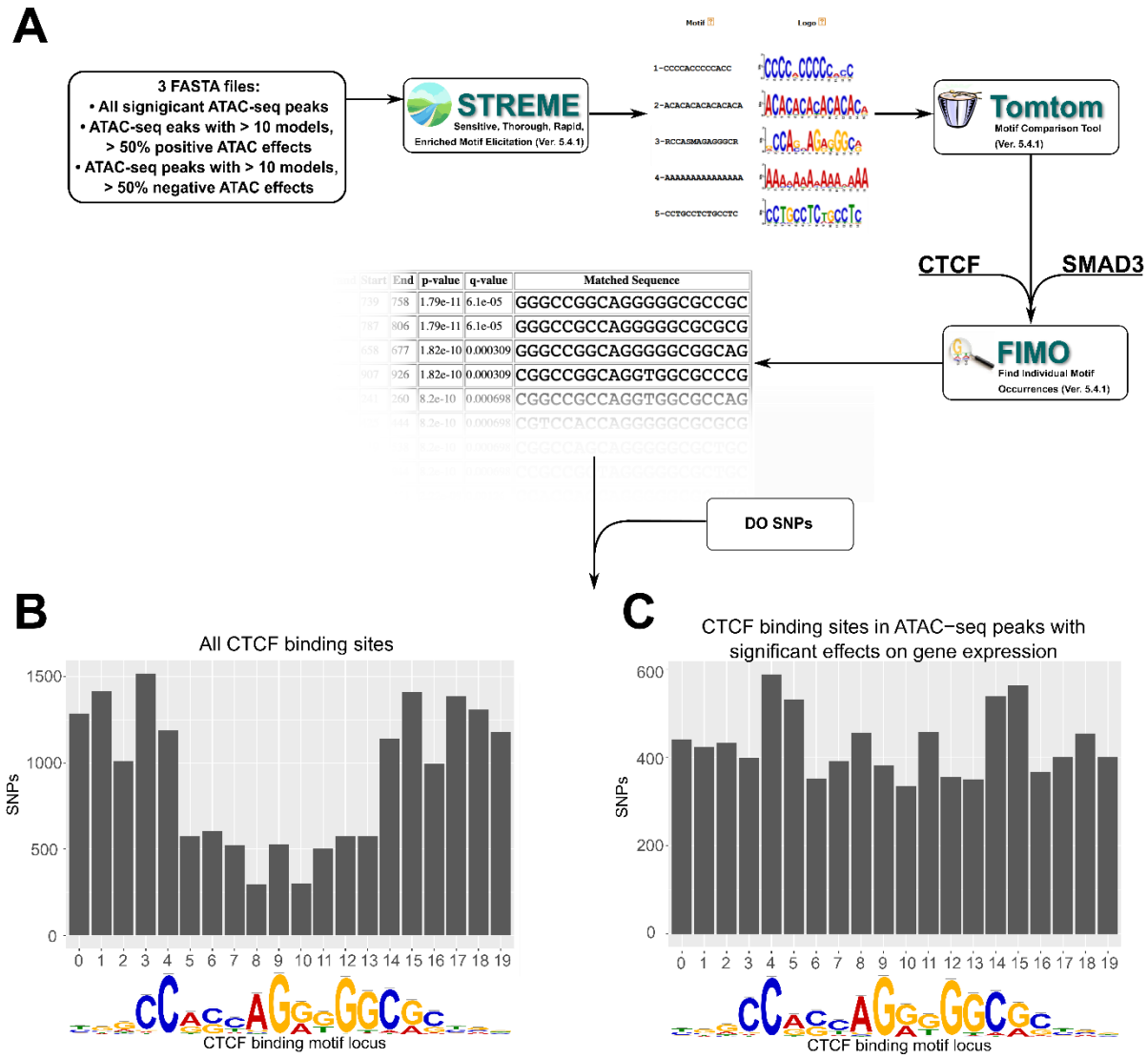


Figure 4. Motif analysis identifies differences in interacting CTCF binding motifs. **A.** A schematic of our motif analysis through MEMEsuite. FASTA files derived from interacting ATAC-seq peaks are used to identify enriched motifs, identify protein binding sequences, and locate the sequences within 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.

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*'s TAD contains a high concentration of confidently called regression models in our analysis (Fig 3B), with 1031 models of SNP-ATAC interaction reaching our significance cutoff for resident genes, of which 179 were models of *Platr2* expression. It also contains 22 haplotype markers, allowing a certain level of model SNP localization. Previous studies have found a group of genes affected by *trans* eQTLs mapped to *Platr2* [15]. These 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.

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 (Fig. S4B). As discussed above, this potentially indicates functional redundancy 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. S4 C-D, Supplemental File 2), which may suggest modulation of CTCF binding strength. These results suggest that *Platr2* may have differential regulation patterns governed by changes in TAD formation.

Discussion

Integrating analysis of multi-omics data in a genetically diverse population allows for greater specificity and modelling of complex regulatory interactions. Populations that segregate natural genetic variation provide dense, randomized 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.

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 [36,44], this marks the first time that the phenomenon has been observed genome-wide, with the genetic mapping resolution provided by outbred, heterozygous samples.

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.

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. 4B-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.

We further analyzed the interplay between direction of effect for SNPs, ATAC-seq peaks, and their interactive components (as in Fig S4A), 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 [6,37]. Other models are more cryptic, but contain variation in beta coefficients associated with these models (Fig. S5), suggesting that functional subtypes that can be investigated within each group.

Our approach can be directly extended to specific gene targets. Mediation analysis in the DO by Skelly et al. revealed several mESC genes that act as mediators of downstream gene networks [15,45]. 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. 3B). ATAC-seq peaks were enriched for Smad3 binding sites, a component of the CTCF binding complex. This provides further regulatory information and potential targets for experimental manipulation.

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*.

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 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, Genetic-epigenetic interactions may underpin some of these failures, particularly those that have initially shown strong SNP effects. 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 [46]. 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.

Small changes in TAD boundary location are previously indicated to be involved in developmental and differentiation processes [35,47,48], and other studies have shown that

chromatin rearrangement that alters TAD boundaries is linked to cancer development [6,37,49]. 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.

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.

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.

References

1. Ackermann, A. M., Wang, Z., Schug, J., Naji, A. & Kaestner, K. H. Integration of ATAC-seq and RNA-seq identifies human alpha cell and beta cell signature genes. *Molecular Metabolism* **5**, 233–244 (2016).
2. Wu, J. *et al.* The landscape of accessible chromatin in mammalian preimplantation embryos. *Nature* **534**, 652–657 (2016).
3. Veyrieras, J.-B. *et al.* High-Resolution Mapping of Expression-QTLs Yields Insight into Human Gene Regulation. *PLoS Genet.* **4**, (2008).
4. Stranger, B. E. *et al.* Population genomics of human gene expression. *Nat. Genet.* **39**, 1217 (2007).
5. Dixon, A. L. *et al.* A genome-wide association study of global gene expression. *Nat. Genet.* **39**, 1202 (2007).
6. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
7. Pombo, A. & Dillon, N. Three-dimensional genome architecture: players and mechanisms. *Nat. Rev. Mol. Cell Biol.* **16**, 245 (2015).
8. Harmston, N. *et al.* Topologically associating domains are ancient features that coincide with Metazoan clusters of extreme noncoding conservation. *Nat. Commun.* **8**, (2017).
9. Yakushiji-Kaminatsui, N. *et al.* Similarities and differences in the regulation of HoxD genes during chick and mouse limb development. *PLoS Biol.* **16**, (2018).
10. Serebrenik, Y. V. & Shalem, O. CRISPR mutagenesis screening of mice. *Nat. Cell Biol.* **20**, 1235 (2018).
11. Svenson, K. L. *et al.* High-Resolution Genetic Mapping Using the Mouse Diversity Outbred Population. *Genetics* **190**, 437 (2012).
12. Churchill, G. A., Gatti, D. M., Munger, S. C. & Svenson, K. L. The Diversity Outbred Mouse

Population. *Mamm. Genome* **23**, 713 (2012).
13. Broman, K. W. Haplotype Probabilities in Advanced Intercross Populations. *G3: Genes|Genomes|Genetics* **2**, 199 (2012).
14. Chesler, E. J. *et al.* Diversity Outbred Mice at 21: Maintaining Allelic Variation in the Face of Selection. *G3: Genes|Genomes|Genetics* **6**, 3893 (2016).
15. Skelly, D. A. *et al.* Mapping the Effects of Genetic Variation on Chromatin State and Gene Expression Reveals Loci That Control Ground State Pluripotency. *Cell Stem Cell* **27**, (2020).
16. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology* vol. 11 R25 (2010).
17. Morgan, A. P. *et al.* The Mouse Universal Genotyping Array: From Substrains to Subspecies. *G3: Genes, Genomes, Genetics* **6**, 263–279 (2016).
18. Morgan, A. P. argyle: An R Package for Analysis of Illumina Genotyping Arrays. *G3* **6**, (2015).
19. Bubier, J. A. *et al.* Genetic variation regulates opioid-induced respiratory depression in mice. *Sci. Rep.* **10**, 1–15 (2020).
20. Chick, J. M. *et al.* Defining the consequences of genetic variation on a proteome–wide scale. *Nature* **534**, 500 (2016).
21. Kuhn, R. M., Haussler, D. & Kent, W. J. The UCSC genome browser and associated tools. *Briefings in Bioinformatics* vol. 14 144–161 (2013).
22. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
23. Bailey, T. L. *et al.* MEME Suite: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202–W208 (2009).
24. Bailey, T. L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2**, (1994).
25. Bailey, T. L. STREME: Accurate and versatile sequence motif discovery. *Cold Spring Harbor Laboratory* 2020.11.23.394619 (2020) doi:10.1101/2020.11.23.394619.
26. Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L. & Noble, W. S. Quantifying similarity between motifs. *Genome Biol.* **8**, 1–9 (2007).
27. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* vol. 27 1017–1018 (2011).
28. Keane, T. M. *et al.* Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* **477**, 289.
29. Sloan, C. A. *et al.* ENCODE data at the ENCODE portal. *Nucleic Acids Res.* **44**, D726–D732 (2015).
30. Consortium, T. E. P. & The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* vol. 489 57–74 (2012).
31. Kulakovskiy, I. V. *et al.* HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res.* **46**, D252–D259 (2017).
32. Ronald, J., Brem, R. B., Whittle, J. & Kruglyak, L. Local regulatory variation in Saccharomyces cerevisiae. *PLoS Genet.* **1**, (2005).
33. Su, J., Teichmann, S. A. & Down, T. A. Assessing Computational Methods of Cis-Regulatory Module Prediction. *PLoS Comput. Biol.* **6**, e1001020 (2010).
34. Nelson, C. E., Hersh, B. M. & Carroll, S. B. The regulatory content of intergenic DNA shapes genome architecture. *Genome Biol.* **5**, 1–15 (2004).
35. Rodríguez-Carballo, E. *et al.* Chromatin topology and the timing of enhancer function at the

HoxD locus. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 31231–31241 (2020).
36. Krijger, P. H. & de Laat, W. Regulation of disease-associated gene expression in the 3D genome. *Nat. Rev. Mol. Cell Biol.* **17**, (2016).
37. Tyler, A. L., Donahue, L. R., Churchill, G. A. & Carter, G. W. Weak Epistasis Generally Stabilizes Phenotypes in a Mouse Intercross. *PLoS Genet.* **12**, e1005805 (2016).
38. Andrew C. Leon, M. H. Sample Sizes Required to Detect Interactions between Two Binary Fixed-Effects in a Mixed-Effects Linear Regression Model. *Comput. Stat. Data Anal.* **53**, 603 (2009).
39. Cao, J. *et al.* Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* **361**, 1380 (2018).
40. Zuo, Z., Roy, B., Chang, Y. K., Granas, D. & Stormo, G. D. Measuring quantitative effects of methylation on transcription factor–DNA binding affinity. *Science Advances* (2017) doi:10.1126/sciadv.aao1799.
41. Maurano, M. T. *et al.* Role of DNA Methylation in Modulating Transcription Factor Occupancy. *Cell Rep.* **12**, (2015).
42. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
43. Shen, Y. *et al.* A map of the cis-regulatory sequences in the mouse genome. *Nature* **488**, (2012).
44. Kumasaka, N., Knights, A. J. & Gaffney, D. J. Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nat. Genet.* **48**, 206–213 (2015).
45. Skelly, D. A. *et al.* Genetic variation influences pluripotent ground state stability in mouse embryonic stem cells through a hierarchy of molecular phenotypes. *Cold Spring Harbor Laboratory* 552059 (2019) doi:10.1101/552059.
46. Fasolino, M. *et al.* Genetic Variation in Type 1 Diabetes Reconfigures the 3D Chromatin Organization of T Cells and Alters Gene Expression. *Immunity* **52**, 257–274.e11 (2020).
47. Takayama, N. *et al.* The Transition from Quiescent to Activated States in Human Hematopoietic Stem Cells Is Governed by Dynamic 3D Genome Reorganization. *Cell Stem Cell* **28**, (2021).
48. Andrey, G. *et al.* A Switch Between Topological Domains Underlies HoxD Genes Collinearity in Mouse Limbs. *Science* **340**, (2013).
49. Menghi, F. *et al.* The Tandem Duplicator Phenotype Is a Prevalent Genome-Wide Cancer Configuration Driven by Distinct Gene Mutations. *Cancer Cell* **34**, 197–210.e5 (2018).