Correcting for relatedness in standard mouse mapping populations; and something about epistasis

Catrina Spruce [1], Anna L. Tyler [1], Many more people, Gregory W. Carter [1]*

1 600 Main St. Bar Harbor, ME, 04609

* Corresponding author: Gregory.Carter@jax.org

Abstract

The abstract goes here

Author summary

The author summary goes here

Introduction

There is evidence that, especially early in life, chromatin modifications are genetically determined [cite].

Materials and Methods

Mice

Inbred Founder Mice

Diversity Outbred mice

The genomic features we collected from inbred founders: chromatin state, percent DNA methylation, and SNPs were imputed into a population of DO mice based on local haplotypes. These mice were described previously in (Svenson 2012). The study population included males and females from DO generations four through eleven. Mice were randomly assigned to either a chow diet (6% fat by weight, LabDiet 5K52, LabDiet, Scott Distributing, Hudson, NH), or a high-fat, high-sucrose (HF/HS) diet (45% fat, 40% carbohydrates, and 15% protein) (Envigo Teklad TD.08811, Envigo, Madison, WI). Mice were maintained on this diet for 26 weeks (CITE).

Genotyping

Diversity outbred mice

All DO mice were genotyped as described in Svenson et al. (2012) using the Mouse Universal Genotyping Array (MUGA) (7854 markers), and the MegaMUGA (77,642 markers) (GeneSeek, Lincoln, NE). All animal procedures were approved by the Animal Care and Use Committee at The Jackson Laboratory (Animal Use Summary # 06006).

Measurement of gene expression

Inbred Founders

Diversity outbred mice

At sacrifice, whole livers were collected and gene expression was measured using RNA-Seq as described in (Chick, Munger et al. 2016, and Tyler et al. 2017)

Measurement of Chromatin Modifications

Measurement of DNA Methylation

Percent DNA methylation was measured using reduced representation bisulfite sequencing.

Data Processing

Chromatin modifications

Annat's stuff to get fastq files to bam files bam to bed binarize bed files

DNA methylation

Vivek's stuff to get bed files.

Analysis

Identification and characterization of chromatin states

We used ChromHMM [29120462] to identify chromatin states corresponding to the presence and absence of the four chromatin modifications. We calculated states for all numbers of states between four and 16, which is the maximum number of states possible with four binary chromatin modifications.

Selecting the number of chromatin states

To identify the ChromHMM model that corresponded best with gene expression, we compared the correlation of each state with gene expression across all ChromHMM models (Supp Fig. XXX). Across all models, the correlations between gene expression and chromatin state could be binned roughly into five bins: low, moderately low, no correlation, moderately high correlation, and high correlation. The nine-state model had states in each of these categories with the lowest redundancy. Furthermore, state seven in the nine-state model had the maximum correlation with gene expression. Therefore, we chose the nine-state model for downstream analysis.

Chromatin state composition

Emissions probabilities derived from ChromHMM indicate the contribution of each histone modification to each chromatin state (Figure XXX).

Positional enrichments of chromatin states

We used the ChromHMM function OverlapEnrichment to identify correlations between chromatin state position and annotated functional elements of the genome. We used functional annotations for the following features:

- **Transcription start sites (TSS)** - Annotations of TSS in the mouse genome were provided by RefSeq [26553804] and included with the release of ChromHMM, which we downloaded on December 9, 2019 [29120462].
- **Transcription end sites (TES)** - Annotations of TES in the mouse genome were provided by RefSeq and included with the release of ChromHMM.
- **Transcription factor binding sites (TFBS)** - We downloaded TFBS coordinates from OregAnno [26578589] using the UCSC genome browser [12045153] on May 4, 2021.
- **Promoters** - We downloaded promoter coordinates provided by the eukaryotic promoter database [27899657,25378343], through the UCSC genome browser on April 26, 2021.
- **Enhancers** - We downloaded annotated enhancers provided by ChromHMM through the UCSC genome browser on April 26, 2021.
- **Candidates of cis regulatory elements in the mouse genome (cCREs)** - We downloaded cCRE annotations provided by ENCODE [22955616] through the UCSC genome browser on April 26, 2021.
- **CpG Islands** - Annotations of CpG islands in the mouse genome were included with the release of ChromHMM.

Downloading SNP data

We downloaded SNP data for the eight inbred DO/CC founders from the Sanger SNP database [1921910, 21921916] on July 6, 2021. We downloaded SNPs ranging from 1kb upstream of the TSS to 1kb downstream of the TES for each gene in our expression data set.

Aligning positions relative to gene bodies

For multiple analyses in this paper, we quantified genomic feature abundance or correlation to gene expression, based on the feature's relative position to the gene body. To do this, we normalized all gene coordinates to run from 0 at the transcription start site to 1 at the transcription end site. Upstream regulatory regions were assigned negative coordinates and downstream regulatory regions were assigned coordinates greater than 1. Base pair positions of genomic features were first centered on the TSS by subtracting the base pair position of the TSS. Centered positions were then divided by the length of the gene in base pairs, defined as the distancr from the gene TSS to the gene TES. These relative positions were grouped into 41 positions defined by the sequence from -2 to 2 incremented by 0.1. If multiple positions were grouped together, the mean value across positions was used.

To avoid potential contamination from regulatory regions of nearby genes, we only included genes that were at least 2kb from their nearest neighbor, for a final set of 14048 genes.

Assessing physical distributions of epigenetic modifications

We used the above normalized base pair positions to calculate the relative abundance of each chromatin state across all gene bodies. For each relative position, we calculated the proportion of genes that contained each chromatin state.

We also calculated the median percent DNA methylation across all gene bodies at each normalized position. To calculate density of CpG sites, we first used absolute base pair positions to calculate the distance from each CpG site to the next. We then normalized the positions relative to gene TSS and TES as described above. The final

result was a distance to the next CpG site in base pairs for each normalized position across the gene bodies.

Correlating genomic features with gene expression in inbred mice

We correlated both chromatin state and percent DNA methylation with gene expression in nine strains of inbred mice which included the DO/CC founders and DBA/2J. We looked for relationships between epigenetic features and gene expression both within-strain, and across strains.

Within-strain correlations between epigenetic features and gene expression suggest some role for the epigenetic feature in regulation of expression. Across-strain correlations of epigenetic features and gene expression further suggest a role for local genotype in determining the regulatory epigenetic features.

To assess the within-strain relationship between chromatin state and gene expression, we scaled transcript levels of all genes in each strain. We used the normalized base pair positions calculated above to average expression of all genes assigned to each chromatin state at each position along the gene body.

Similarly, we calculated the correlation between gene expression and percent methylation at each normalized position across the gene body.

To identify across-strain correlations between epigenetic modifications and transcript abundance, we scaled transcript abundance across strains separately for each transcript. We then correlated this scaled abundance with either chromatin state or percent DNA methylation at each normalized position across the gene body.

Imputing genomic features in Diversity Outbred mice

To further investigate the effect of genetic and epigenetic features on local gene expression, we imputed chromatin state, SNPs, and DNA methylation into a population of diversity outbred (DO) mice (CITE).

Each imputation followed the same basic procedure: For each transcript, we identified the haplotype probabilities in the DO mice at the genetic marker nearest the gene transcription start site. This matrix held DO individuals in rows and DO founder haplotypes in columns.

For each transcript, we also generated a three-dimensional array representing the genomic features derived from the DO founders. This array held DO founders in rows, feature state in columns, and genomic position in the third dimension. The feature state for chromatin consisted of states one through nine, for SNPs feature state consisted of the genotypes A,C,G, and T, and for DNA methylation, feature state consisted of percent DNA methylation rounded to the nearest 0%, 50%, or 100%. Rounding reduced the direct transfer of haplotype effects to functionally irrelevant variation in DNA methylation across strains at individual positions.

We then matrix multiplied the haplotype probabilities by each genomic feature array to obtain the imputed genomic feature for each DO mouse. This final array held DO individuals in rows, genomic feature in the second dimension, and genomic position in the third dimension. This array is analagous to the genoprobs object in R/qtl2 (CITE). The genomic position dimension included all positions between the transcription start site and the transcription end site ($\pm 1kb$). SNP data for the DO founders in mm10 coordinates were downloaded from the Sanger SNP database (CITE), on July 6, 2021.

To calculate the effect of each genomic feature on gene expression, we fit the following linear model:

$$y = \beta_0$$

where blah blah blah

From this linear model, we calculated the variance explained (R^2) by each genomic feature. We thus related gene expression in the DO to each position of imputed chromatin state, SNPs, or DNA methylation in and around the gene body.

Results

Description of Chromatin States

We identified nine chromatin states corresponding to nine distinct combinations of histone modifications (Figure XXXA). These states were differentially distributed near functionally annotated genomic elements (Figure XXXB). For example, State 1, which corresponded to the absence of all four histone modifications, found mainly in intergenic regions. States 5 and 9 were enriched near enhancers and TES respectively. Finally, states 3 and 7 were highly enriched near the TSS and other functional elements that also occur near the TSS, such as cis-regulatory regions (cCREs), transcription factor binding sites (TFBS), and promoters.

A subset of the chromatin states was also correlated with gene expression across strains (Figure XXXC) in a manner that concorded with both their histone modification profiles and their enrichments near functional genomic elements. For example, the histone modification H3K27me3 has been previously shown to be associated with reduced transcription [CITE]. We found this modification in both states 2 and 3 (Figure XXXA), which were both correlated with reduced transcription (Figure XXXC). State 3, furthermore, was enriched near the TSS, promoters, TFBS, and other regulatory elements supporting a possible role in transcriptional regulation.

State 5 was characterized by the presence of two histone modifications previously associated with increased transcription: H3K27ac [CITE], and H3K4me1 [CITE]. The enrichment of this state in enhancers coincides with previous work showing the presence of these two modifications in active enhancers [21106759, 21160473, 29273804], and supports the role of this state in upregulation of gene transcription.

States 6 and 7 were also associated with increased transcription and the presence of transcriptionally activating histone modifications. State 6 was modestly enriched in enhancers, while state 7 was enriched not only in enhancers, but also stongly near the TSS and other associated functional elements (Figure XXXB).

Spatial distribution of chromatin states and DNA methylation

In addition to looking for enrichment of chromatin states near annotated functional elements, we also characterized the relative spatial distribution around gene bodies of both chromatin states and DNA methylation. To do this, we scaled base pair positions of our measured genomic features to run between 0 at the TSS and 1 at the TES of their containing gene (Methods) (Figure XXX).

Each chromatin state had a characteristic distribution pattern relative to gene bodies (Figure XXXA and B). For example, state 1 was strongly depleted near the TSS, indicating that this region is commonly subject to chromatin modification. However, its abundance increased steadily to a peak at the TES. In contrast, state 7 was present in over 60% of TSS, but decreased to almost 0% near the TES.

The remaining states were relatively low in abundance compared to states 1 and 7, but also showed specific distributions relative to the gene body. State 8, was depleted at the TSS, but enriched immediately downstream of the TSS. State 9 had slight enrichments immediately upstream of the TSS and immediately downstream of the TES.

These patterns agree with the enrichments around functional genomic elements shown in Figure XXXB, and add interesting resolution around both the TSS and TES.

DNA methylation also showed strong positional enrichments (Figure XXXC and D). Across all genes, the TSS had densely packed CpG sites cytosines relative to the region between the TSS and TES (Figure XXXC). As expected, the median CpG site near TSS were consistently hypomethylated relative to the median CpG site in intergenic regions. CpG sites within the gene body were slightly hypermethylated compared to intergenic CpGs (Figure XXXD).

Within-strain correlation of epigenetic features and gene expression

To investigate the relationship between epigenetic modifications and gene expression, we calculated the relationship both chromatin state and DNA methylation and transcript abundance (Methods) (Figure XXX). Across all genes within individual inbred strains of mice, the percent methylation at CpG sites was negatively correlated with transcript abundance primarily at the TSS.

The correlation of each chromatin state matched that seen in Figure XXXC, and the spatial resolution of the relationships added an interesting dimension to this observation. For example, state 3, which was overall correlated with lower transcript abundance in Figure XXXC, was shown to have this negative effect primarily when it was positioned at the TSS. In contrast, state 2, was also negatively correlated with transcription overall, but this effect was constant throughout the gene body. State 2 is characterized by H3K27me3, which has previously been associated with reduced transcription [CITE], but was not enriched near any annotated functional genomic elements.

Across-strain correlation of epigenetic features and gene expression

To investigate whether the previously identified relationships between epigenetic features and gene expression were related to local genotype, we looked for correlations between gene expression and epigenetic features for each transcript across all inbred strains. That is, for any given transcript, did variation in chromatin state or DNA methylation across strains correlate with gene expression?

Interestingly, the relationship between gene expression and chromatin state across strains was nearly identical to the relationship within-strain, although positional effects were perhaps even more pronounced. States 3 and 9 had negative correlations with gene expression across strains that were localized to the TSS. States 5 and 7 had strong positive correlations with gene expression throughout the gene body, and states 1 and 2 had strong negative correlations with gene expression throughout the gene body.

In stark contrast, DNA methylation was completely uncorrelated with variation in gene expression across strains (Figure XXXB). This lack of correlation is likely due to the low variability of percent methylation across strains at any given position. Figure XXXC shows the standard deviation in percent DNA methylation at normalized positions across the gene body. It is strikingly low everywhere, with the standard deviation being around 6%, which is likely below any biologically functional threshold. The variation dips even lower, to around 4% at the TSS, indicating that for the most part that DNA methylation does not vary across strains and is not contributing to strain difference in gene expression.

Imputed chromatin state correlated with local gene expression in Diversity Outbred mice

To further investigate the relationship between genotype, epigenetic features, and gene expression, we imputed chromatin state, DNA methylation, and SNPs into a population of DO mice described previously [Svenson, Tyler].

We investigated the extent to which chromatin state imputed into DO mice explained variation in expression across individuals. Although local genetic variation explains a large amount of variation in gene expression [cite], chromatin state may offer further insight into regulation of gene expression at the local level. [more compelling stuff here]

We imputed genome-wide chromatin states in a population of DO mice based on their genotype (Methods) and compared the percent variance explained by local genotype to the maximum percent variance explained by local chromatin state for each transcript (Figure XXX.) The two measurements were very tightly correlated (Pearson R = 0.95) indicating that chromatin state determined by genetics is an excellent approximation of the genetic effect on gene expression. The imputation further allowed us to observe the effects of chromatin state across [500] genetically diverse mice by measuring chromatin modifications in a handful of inbred mice. Further, because chromatin modifications are measured at extremely high density, we can map high-density chromatin effects in the DO mice, which may help prioritize functional SNPs within gene bodies and in regulatory regions.

For example, Figure XXX shows chromatin states across the gene Irf5 in the inbred founders along with the LOD score and chromatin state effects at each position along the gene body as calculated in the DO population. The LOD scores and allele effects highlight variation at the TSS, and at several internal positions in the gene as potentially regulating gene expression.

DNA methylation varied across the gene body

In addition to chromatin state, we examined the distribution of DNA methylation across the gene body, as well as the relationship between DNA methylation and gene expression in both inbred mice and DO mice.

As expected, methylated cytosines were densely packed near the gene TSS (Figure XXX). They were relatively sparse within the gene body, and had intermediate spacing outside of gene bodies.

Outside of gene bodies, percent methylation was measured at an average of 50%, whereas there was very low DNA methylation at the gene TSS (Figure XXX). Percent DNA methylation within gene bodies was higher than the surrounding intergenic spaces, reaching a maximum of around 80% near the gene TES.

Within each strain, percent methylation at the gene TSS was slightly negatively correlated with gene expression (Pearson r for all strains was about -0.2). However, there was very little variation in DNA methylation across strains, particularly at the TSS, and consequently, there was no relationship between percent methylation and gene expression across strains.

Discussion

The enrichment of these states in regulatory regions indicates the possibility that these states are used for regulating expression levels whereas states 7 and 3 at the transcription start site may be primarily related to switching gene transcription on and off.

Haplotype and chromatin state represent broader regions of genome than SNPs and DNA methylation, which are measured at the base pair level. The measurements that represent larger regions of the genome are more predictive of local gene expression than the point-wise measurements.

While local haplotype is the best predictor of gene expression, it has poor resolution. SNPs and DNA methylation have very high resolution, but are relatively poor predictors of gene expression. Chromatin state sits in the middle ground. It is almost as good a predictor of gene expression as haplotype, but has resolution down to 200 base pairs, thus offering the potential for dissecting mechanisms of local gene expression at a higher resolution than is possible with haplotype alone.

There is clearly a lot going on at the TSS, but there these results show correlations between gene expression

Perhaps by overlaying all modalities, particularly with measurements of open chromatin, we can come up with examples of this kind of inference? Are there any anecdotes that illustrate this?

Local chromatin state was highly correlated with local gene expression in the DO/CC founders. This was true across genes within each strain, as well as for individual genes across strains, suggesting that variation in chromatin modifications may be a major mechanism of local gene expression regulation.

(Alternatively, chromatin state aligns well with the true local mechanism of gene regulation, but is not itself a mechanism.)

Positional information is interesting

We observed interesting spatial patterns of chromatin state distribution and correlation with gene expression. States 3 and 7 were particularly abundant around transcription start sites (TSS), while all other states were depleted at the TSS. State 8 peaked in abundance immediately downstream of the TSS, and state 9 peaked immediately upstream of the TSS.

State 5 had relatively low abundance. However, it was concentrated within gene bodies where it had a relatively strong positive correlation with gene expression. This indicates that (?)

Acknowledgements

This work was funded by XXX.

Data and Software Availability

All data used in this study and the code used to analyze it are avalable as part of a reproducible workflow located at... (Figshare?, Synapse?).

Supplemental Figure Legends

Fig 1

Fig 2. Correlations between traits and the first PC of the kinship matrix.

Supplemental Table Descriptions

References