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

Measurement of Chromatin Modifications

Measurement of DNA Methylation

Percent DNA methylation was measured using reduced representation bisulfite sequencing.

Data Processing

Identifying chromatin states

We used ChromHMM 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.

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.

Correlating genomic features with gene expression in DO founders

We correlated both chromatin state and percent DNA methylation with gene expression in the DO founders.

Chromatin State

To correlate chromatin state with gene expression, we calculated the proportion of the gene body that was assigned to each chromatin state across the nine inbred founders. We then correlated the proportion of each state with the mean gene expression across the founders. We calculated these correlations across all states and all state models.

DNA Methylation

To correlate percent DNA methylation with gene expression,

Assessing abundance of chromatin states across gene bodies

We calculated the relative abundance of each chromatin state across all gene bodies.

Assessing correlation of chromatin state with expression across gene body

We used the normalized gene coordinates calculated above to calculate position-based correlations between chromatin state and gene expression. To do this we used a sliding window across the gene body from normalized coordinates -1 to +2 and correlated state proportion within each window with gene expression.

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). These mice, described previously in (CITE), 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).

All 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).

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%.

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 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.

We used R/qtl2 [cite] to calculate the effect of each genomic feature on gene expression in the DO. We calculated LOD scores to relate gene expression in the DO to each position with imputed chromatin state, SNPs, or DNA methylation in and around the gene body.

Results

Chromatin State and Gene Expression

The nine-state model had the highest correlation with gene expression

To identify the ChromHMM model that corresponded best with gene expression, we compared the correlation of each state with gene expression (Methods) 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.

Genomic position enrichments aided interpretation of chromatin states

The correlation between chromatin state and gene expression, as well as positional enrichments around functionally annotated positions in the genome aided annotation of individual chromatin states. For example, chromatin state 5, which was characterized by the co-occurence of H3K27ac and H3K4me1, had enriched representation around known enhancers. Its presence was also positively correlated with gene expression. This combination of chromatin modifications has previously been associated with active enhancers [21106759, 21160473, 29273804].

State 7, which was characterized by the presence of three activating marks, H3k27ac, H3k4me1, and H3K4me3. It was enriched around transcription start sites (TSS) and promoters and was also positively correlated with gene expression.

State 3 was characterized by the presence of H3K4me1 and H3k27me3. H3K27me3 is associated with downregulation of transcription and poised enhancers. When paired with H3K4me1

also enriched around promoters and TSS, but was most strongly enriched at transcription factor binding sites,

Table?? Figure?? Supp Figure??

Promoter and enhancer states correlated with increased expression

Chromatin states 5 and 7, which were enriched around enhancers and promoters respectively, were correlated with increased expression. This was true both across the liver transcriptome, as well as across strains. That is, that within a single strain, genes with higher proportions of states 5 or 7 across the gene body were more highly expressed than genes with low proportions of these states. And for individual genes, strains with higher proportions of states 5 and 7 had higher expression than strains with lower proportions of these states (Figure?).

State 6 was also correlated with increased gene expression to a lesser degree. This state was also enriched at enhancers, though less so than state 5.

H3K27me3 and the absence of measured marks correlated with decreased expression

High abundance of states 1, 2, and 3 were correlated with decreased gene expression. State 1 was the absence of all measured chromatin modifications, and states 2 and 3 were the only states with the suppressing modification H3K27me3. State 1 was enriched in intergenic regions and state 3 was enriched around transcription factor binding sites. Positional enrichment of state 2 was not clear.

Differential spatial distribution of states around gene bodies

Chromatin states were each distributed in a specific pattern across gene bodies (Methods) (Figure XXX). 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 gene-body specific distribution patterns. 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. 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.

Correlation of chromatin state and gene expression was differentially distributed across the gene body

We examined whether there was a spatial component to the correlation between chromatin state and gene expression (Methods).

Figure XXX shows the Pearson correlation between expression and chromatin state across all windows for each chromatin state. The most prominant position-specific correlations between state and gene expression were for state 3 and state 9, which were

both negatively correlated with gene expression exclusively at the TSS. There were no TES-specific correlations for any state.

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

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

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.)

The most abundant states were 7 and 1, followed by states 8, 9, and 3. States 4, 5, and 6 were the lowest abundant states. Interestingly, despite its low abundance, state 5, which was concentrated within gene bodies had a relatively strong positive correlation with gene expression when present in this region, indicating that (?)

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 was of lower abundance than the other states were concentrated within the gene body.

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.

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?).

CAPE is available at CRAN...

Supplemental Figure Legends

Fig 1. Correlations between traits and the first principal component (PC) of the kinship matrix. Traits with high correlation to the kinship matrix may be highly polygenic and thus be susceptible to test statistic inflation due to many true positives. To reduce this risk, we selected traits with low correlation with the first kinship matrix PC. This figure shows the distribution of correlations between traits and the first kinship PC across populations.

Fig 2. Reducing n reduces inflation. This figure is identical to Fig. **??** except that we have added a column for the F2 that has been subsampled to the same n as the Backcross. This subsampling reduces power to detect effects, and thus reduces inflation to roughly the same level as that seen in the backcross.

Supplemental Table Descriptions

Fig 3. Correlations between traits and the first PC of the kinship matrix.

References