Variation in epigenetic state correlates with gene expression across nine inbred strains of mice

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

Abstract goes here.

Author summary

The author summary goes here if we submit to a journal that has one.

Abstract

It is well established that epigenetic features, such as histone modifications and DNA methylation, are associated gene expression across cell types. However, it is not well known how variation in genotype affects epigenetic state, or to what extent such variation contributes to variation in gene expression across genetically distinct individuals. Here we investigated the relationship between heritable epigenetic variation and gene expression in hepatocytes across nine inbred mouse strains. Eight of the inbred strains were founders of the diversity outbred (DO) mice, and the ninth was DBA/2J, which, along with C57BL/6J, is one of the founders of the BxD recombinant inbred panel of mice. We surveyed four histone modifications, H3K4me1, H3K4me3, H3K27me3 and H3K27ac, as well as DNA methylation. We used ChromHMM to identify 14 chromatin states representing distinct combinations of the four measured histone modifications. We found that variation in chromatin state mirrored genetic variation across the inbred strains. Furthermore, epigenetic variation was correlated with gene expression across strains. The correspondence between epigenetic state and gene expression was replicated in an independent population of DO mice in which we imputed local epigenetic state. In contrast, we found that DNA methylation did not vary across inbred strains and was not correlated with variation in expression in DO mice. This work suggests that chromatin state is highly influenced by local genotype and may be a primary mode through which expression quantitative trait loci (eQTLs) are mediated. We further demonstrate that strain variation in chromatin state, paired with gene expression is useful for annotation of functional regions of the mouse genome. Finally, we provide, to our knowledge, the first data resource that documents variation in chromatin state across genetically distinct individuals.

Introduction

Epigenetic modifications to DNA and its associated histone proteins influence the accessibility of DNA to transcription machinery, and are associated with up- and down-regulation of gene expression [1–3]. Across cell types, unique combinatorial patterns of histone modifications mark chromatin states that establish cell type-specific patterns of gene expression [4]. Similarly, the methylation of CpG sites around gene promoters and enhancers influences transcription in a cell type-specific manner [6,7].

These patterns of histone modifications and DNA methylation are established during development. The result is a canonical epigenetic landscape for coordination of major patterns of gene expression for each cell type [sources about development]. As an organism ages and responds to its environment, patterns of both histone modifications [citation] and of DNA methylation change [citation]. Such changes have been linked to scenescence [Horvath clock] and cancer [citations].

Epigenetic modifications coordinate the usage of a single genome to be used for many different types of cells with diverse morphology and physiology. This amazing feature of epigenetic modifications has been intensely studied, and the variation in epigenetic landscapes across cell types has been extensively documented [citations]. Less well understood, however, is the role that genetic variation plays in determining epigenetic landscapes.

Across genetically diverse populations of humans or mice, individual cell types, such as hepatocytes, or cardiomyocytes, have globally similar gene expression profiles that define their role within the greater organism. However, it is also true that across individuals, gene expression varies widely within the global constraints of cell type. This variation can increase or decrease an organism's risk of developing disease. Variation in gene expression has been extensively mapped to variation in genetic loci, or expression quantitative trait loci (eQTL). Large, coordinated efforts, such as the Genotype-Tissue Expression (GTEx) Project [32913073, 32913075] have identified and catalogued many such loci in humans, and countless independent studies have identified eQTL in mice and other model organisms.

Although the link between genetic variation and gene expression has been well studied, there is relatively little known about inter-individual variation in epigenetic modifications, and how these variations are related to variations in genotype and gene expression. The generation of a more complete picture of inter-individual variation in epigenetic modifications has the potential to increase our understanding of the mechanisms of gene regulation, provide insights into the mechanisms establishing cell type-specific epigenetic landscapes, and to improve the functional annotation of the genome as it relates to the regulation of gene expression. The vast majority of SNPs associated with human disease traits are located in non-coding regions, suggesting that they influence gene regulation, rather than protein function [citation]. However, annotation of these regions is difficult without additional genomic features, such as histone modifictions and DNA methylation. Overlaying a map of variation in epigenetic features has the potential to provide a picture of how genetic variation changes functional elements, like enhancers and insulators, in the genome [citation].

Advances in chromatin immunoprecipitation (ChIP) and sequencing technologies now enable genome-wide surveys of histone modifications with relatively few cells [8], thus opening the door to the possibility of cataloging epigenetic variation across cell types and individuals. Here, we performed a survey of epigenetic variation in hepatocytes across nine inbred mouse strains. We included the eight founders of the Diversity Outbred/Collaborative Cross (DO/CC) [9] mice, as well as DBA/2J, which, along with C57BL/6J, is one of the founders of the widely used BxD recombinant inbred panel of mice [10]. We assayed four histone modifications (H3K4me1, H3K4me3, H3K27me3, and H3K27ac), as well as DNA methylation. We used ChromHMM [11] to identify 14 chromatin states, classified by unique combinations of the four histone marks, and investigated the association between variation in these states and variation in gene expression across the nine strains. We separately investigated the relationship between DNA methylation and gene expression across strains.

We further investigated the relationship between epigenetic state and gene expression by imputing the 14 chromatin states and DNA methylation into a population of DO mice. We then mapped gene expression to the imputed epigenetic states to assess the extent to which eQTLs are driven by variation in epigenetic modification. We thus linked genetically controlled variation in epigentic modifications to variation in gene expression in mice, and we provide the first resource documenting epigenetic variation across a wide panel of genetically diverse mice.

Materials and Methods

Ethics Statement

Ethics Statement All animal procedures followed Association for Assessment and Accreditation of Laboratory Animal Care guidelines and were approved by Institutional Animal Care and Use Committee (The Jackson Laboratory, Protocol XXX).

Inbred Mice

Three female mice from each of nine inbred strains were used. Eight of these strains (129S1/SvImJ, A/J, C57BL/6J, CAST/EiJ, NOD/ShiLtJ, NZO/HlLtJ, PWK/PhJ, and WSB/EiJ) are the eight strains that served as founders of the Collaborative Cross/Diversity Outbred mice [12]. The ninth strain, DBA/2J, will facilitate the interpretation of existing and forthcoming genetic mapping data obtained from the BxD recombinant inbred strain panel. Samples were harvested from the mice at 12 weeks of age.

Liver perfusion

To purify hepatocytes from the liver cell population, the mouse livers were perfused with 87 CDU/mL Liberase collagenase with 0.02% CaCl2 in Leffert's buffer to digest the liver into a single-cell suspension, and then isolated using centrifugation.

We aliquoted $5x10^6$ cells for each RNA-Seq and bisulfite sequencing, and the rest were cross-linked for ChIP assays. Both aliquots were spun down at 200 rpm for 5 min, and resuspended in $1200\mu L$ RTL+BME (for RNA-Seq) or frozen as a cell pellet in liquid nitrogen (for bisulfite sequencing). In the sample for ChIP-Seq, protein complexes were cross-linked to DNA using 37% formaldehyde in methanol. All cell samples were stored at -80°C until used (See Supplemental Methods for more detail).

Hepatocyte histone binding and gene expression assays

Hepatocyte samples from 30 treatment and control mice were used in the following assays:

1. RNA-seq to quantify mRNA and long non-coding RNA expression, with approximately 30 million reads per sample.
2. Reduced-representation bisulfate sequencing to identify methylation states of approximately two million CpG sites in the genome. The average read depth was 20-30x.
3. Chromatin immunoprecipitation and sequencing to assess binding of the following histone marks:

a. H3K4me3 to map active promoters

b. H3K4me1 to identify active and poised enhancers
c. H3K27me3 to identify closed chromatin
d. H3K27ac, to identify actively used enhancers
e. A negative control (input chromatin)

Samples are sequenced with $\sim$ 40 million reads per sample.

The samples for RNA-Seq in RTL+BME buffer were sent to The Jackson Laboratory Gene Expression Service for RNA extraction and library synthesis.

Histone chromatin immunoprecipitation assays

After exdtraction, hepatocyte cells were lysed to release the nuclei, spun down, and resuspended in 130ul MNase buffer with 1mM PMSF and 1x protease inhibitor cocktail (Roche) to prevent histone protein degradation. The samples were then digested with 15U of micrococcal nuclease (MNase), which digests the exposed DNA, but leaves the nucleosome-bound DNA intact. We confirmed digestion of nucleosomes into 150bp fragments with with agarose gel. The digestion reaction was stopped with EDTA and samples were used immediately in ChIP assay. The ChIP assay was performed with Dynabead Protein G beads and histone antibodies. After binding to antibodies, samples were washed to remove unboud chromatin and then eluted with high-salt buffer and proteinase K to digest protein away from DNA-protein complexes. The DNA was purified using the Qiagen PCR purification kit. Quantification was performed using the Qubit quantification system (See Supplemental Methods).

Diversity Outbred mice

We used previously published data from a population of 478 diversity outbred (DO) mice [9] to compare to the data collected from the inbred mice. The DO population included males and females from DO generations four through 11. 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.

Genotyping

All DO mice were genotyped as described in [9] 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).

Founder haplotypes were inferred from SNPs using a Hidden Markov Model as described in [13]. The MUGA and MegaMUGA arrays were merged to create a final set of evenly spaced 64,000 interpolated markers.

Tissue collection and gene expression

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). Transcript sequences were aligned to strain-specific genomes, and we used an expectation maximization algorithm (EMASE) to estimate read counts (`https://github.com/churchill-lab/emase`).

Data Processing

Sequence processing

The raw sequencing data from both RNA-Seq and ChIP-Seq was put through the quality control program FastQC (version), and duplicate sequences were removed before downstream analysis. Reads from each sample were mapped to strain-specific pseudogenomes, which integrate known SNPs and indels from each strain, and the B6 samples were aligned directly to the reference mouse genome. The pseudogenomes were created using EMASE \url(https://github.com/churchill-lab/emase). We used Bowtie [14] to align and map reads from the RNA-Seq and ChIP-Seq experiments.

Transcript quantification

We quantified gene expression using edgeR [15], and transcripts with less than 1 CPM in two or more replicates were filtered out. Transcripts were further filtered to include only protein-coding transcripts.

We used the R package sva [16] to perform a variance stabilizing transformation (vst) on the RNA-Seq read counts from both inbred and outbred mice. In the inbred mice we used a blind transformation, while in the outbred mice, we included DO wave and sex in the model. For eQTL mapping, we performed rank Z normalization on the RNA-Seq read counts across transcripts from the outbred mice.

ChIP-Seq quantification

We used MACS 1.4.2 [17] to identify peaks in the ChIP-Seq sequencing data, with a significance threshold of $p \leq 10^{-5}$. In order to compare peaks across strains, we converted the MACS output peak coordinates to common B6 coordinates using g2g tools (https://churchill-lab.github.io/g2gtools/).

Quantifying DNA methylation

RRBS data were processed using a bismark-based pipeline modified from [18]. The pipeline uses Trim Galore! 0.6.3 (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) for QC, followed by the trimRRBSdiversityAdaptCustomers.py script from NuGen for trimming the diversity adapters. This script is available at: https://github.com/nugentechnologies/NuMetRRBS

All samples had comparable quality levels and no outstanding flags. Total number of reads was 45-90 million, with an average read length of about 50 bp. Quality scores were mostly above 30 (including error bars), with the average above 38. Duplication level was reduced to < 2 for about 95% of the sequences.

High quality reads were aligned to a custom strain pseudogenomes, using bowtie2 as implemented in Bismark 0.22 [19]. The pseudogenomes were created by incorporating strain-specific SNPs and indels into the reference genome using g2gtools (https://github.com/churchill-lab/g2gtools), allowing a more precise characterization of methylation patterns. Bismark methylation extractor tool was then used for creating a bed file of estimated methylation proportions for each animal, which was then translated to the reference mouse genome (GRCm38) coordinates using g2gtools. Unlike other liftover tools, g2gtools does not throw away alignments that land on indel regions. B6 samples were aligned directly to the reference mouse genome.

Analysis of histone modifications

Identification of chromatin states

We used ChromHMM [20] to identify chromatin states, which are unique combinations of the four chromatin modifications, for example, high levels of both H3K4me3 and H3K4me1, and low levels of the other two modifications. We conducted all subsequent analyses at the level of the chromatin state.

To ensure we were analyzing the most biologically meaningful chromatin states, we calculated chromatin states for all numbers of states between four and 16, which is the maximum number of states possible with four binary chromatin modifications (2^n). We aligned states across the models by assigning each to one of the sixteen possible binary states using an emissions probability of 0.3 as the threshold for presence/absence of the histone mark. We then investigated the stability of three features across all states: the emissions probabilities (Supp Fig1), the abundance of each state across transcribed genes (Supp Fig2), and the effect of each state on transcription (Supp Fig3). Methods for each of these analyses are described separately below. All measures were remarkably consistent across all models, but the 14-state model was characterized by a wide range of relatively abundant states with relatively strong effects on expression. We used this model for all subsequent analyses. For more details on how the different models were compared, see Supplemental Methods.

Genome distribution of chromatin states

We investigated genomic distributions of chromatin states in two ways. First, we used the ChromHMM function OverlapEnrichment to calculate enrichment of each state around known functional elements in the mouse genome. We analyzed the following features:

- **Transcription start sites (TSS) -** Annotations of TSS in the mouse genome were provided by RefSeq [21] 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 [22] using the UCSC genome browser [23] on May 4, 2021.
- **Promoters -** We downloaded promoter coordinates provided by the eukaryotic promoter database [24,24], 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 [25] 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.

In addition to these enrichments around individual elements, we also calculated chromatin state abundance relative to the main anatomical features of a gene. For each transcribed gene, we normalized the base pair positions to the length of the gene such that the transcription start site (TSS) was fixed at 0, and the transcription end site (TES) was fixed at 1. We also included 1000 bp upstream of the TSS and 1000 bp downstream of the TES, which were converted to values below 0 and above 1 respectively.

To map chromatin states to the normalized positions, we binned the normalized positions into 41 bins defined by the sequence from -2 to 2 incremented by 0.1. If a bin encompassed multiple positions in the gene, we assigned the mean value of the feature of interest to the bin. 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.

Chromatin state and gene expression

We calculated the effect of each chromatin state on gene expression. We did this both across genes and across strains. The across-gene analysis identified states that are associated with high expression and low expression within the hepatocytes and independent of strain. The across-strain analysis investigated whether variation in chromatin state across strains contributed to variation in gene expression across strains.

For each transcribed gene, we calculated the proportion of the gene body that was assigned to each chromatin state. We then fit a linear model separately for each state to calculate the effect of state proportion with gene expression:

$$y_e = \beta x_s + \epsilon$$

where y_e is the rank Z normalized gene expression of the full transcriptome in a single inbred strain, and x_s is the rank Z normalized proportion of each gene that was assigned to state s. We fit this model for each strain and each state to yield one β coefficient with 95% confidence interval. The effects were not different across strains, so we averaged the effects and confidence intervals across strains to yield one summary effect for each state.

To calculate the effect of each chromatin state across strains, we first standardized transcript abundance across strains for each transcript. We also standardized the proportion of each chromatin state for each gene across strains. We then fit the same linear model, where y_e was a rank Z normalized vector concatenating all standardized expression levels across all strains, and x_s was a rank Z normalized vector concatenating all standardized state proportions across all strains. We fit the model for each state independently yielding a β coefficient and 95% confidence interval for each state.

In addition to calculating the effect of state proportion across the full gene body, we also performed the same calculations in a position-based manner. This second analysis yielded an effect of each state at multiple points along the gene body and a more nuanced view of the effect of each state.

Analysis of DNA methylation

Creation of DNA methylome

We combined the DNA methylation data into a single methylome cataloging the methylated sites across all strains. For each site, we averaged the percent methylation across the three replicates in each strain. The final methylome contained 5,311,670 unique sites across the genome. Because methylated CpG sites can be fully methylated, unmethylated, or hemi-methylated, we rounded the average percent methylation at each site to the nearest 0, 50, or 100%.

Distribution of CpG sites

We used the enrichment function in ChromHMM described above to identify enrichment of CpG sites around functional elements in the mouse genome. We further performed a gene-based analysis of abundance similar to that in the chromatin states. As a function of relative position on the gene body, we calculated the density of CpG sites as the average distance to the next downstream CpG site, as well as the percent methylation at each site.

Effects of DNA methylation on gene expression

As with chromatin state, we assessed the effect of DNA methylation on gene expression both across genes and across strains. We used the same linear model described above, except that y_s became the rank Z normalized percent methylation either across genes or across strains. Because the effect of DNA methylation on gene expression is well-known to be dependent on position, we only calculated a position-dependent effect on expression.

Imputation of genomic features in Diversity Outbred mice

To assess the extent to which chromatin state and DNA methylation were responsible for local expression QTLs, we imputed local chromatin state and DNA methylation into a population of diversity outbred (DO) mice described above and in [9]. We compared the effect of the imputed epigenetic features to imputed SNPs.

All imputations 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 (Supp. Fig. 11).

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 14, for SNPs feature state consisted of the genotypes A,C,G, and T.

We then 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, the genomic feature in the second dimension, and genomic position in the third dimension. This array is analagous to the genoprobs object in R/qtl2 [26]. The genomic position dimension included all positions from 1 kb upstream of the TSS to 1 kb downstream of the TES. SNP data for the DO founders in mm10 coordinates were downloaded from the Sanger SNP database [27,28], on July 6, 2021.

To calculate the effect of each imputed genomic feature on gene expression in the DO population, we fit a linear model. From this linear model, we calculated the variance explained (R^2) by each genomic feature, thereby relating gene expression in the DO to each position of the imputed feature in and around the gene body.

Results

Gene expression varies widely and reproducibly across inbred strains of mice. This is seen as a clustering of individuals from the same strain in a principal component plot of the hepatocyte transcriptome across strains (Figure 1A). Patterns of DNA methylation (Figure 1B) and individual histone modifications (Figure 1C-F) cluster in a similar pattern. This suggests that these epigenetic features may relate to gene expression in a manner that is consistent with genetic background.

Chromatin state overview

To investigate this association, we used ChromHMM to identify 14 chromatin states composed of unique combinations of four histone modifications in the hepatocytes of nine inbred strains of mice. Panel A in Figure 2 shows the representation of each histone modification across the states.

The states were distributed non-randomly around known functional elements in the mouse genome (Figure 2B). The majority of the states were enriched around the TSS, and other TSS-related functional elements, such as promoters and CpG islands. Two states (states 13 and 14) were primarily found in intergenic regions. Three states (states 6, 2, and 4) were enriched around known enhancers, and one (state 9) was enriched predominantly near the TES. The majority of these states were also associated with variation in gene expression. The colored bars in Figure 2C) show the effect of each state on gene expression across the inbred strains. For reference, the paired tan bars show the effect of each chromatin state on gene expression in hepatocytes. These effects tend to be of the same sign and greater magnitude than the across-strain effects.

The states in Figure 2 are shown in order of their effect on expression, which helps illustrate several patterns in the data. The state with the largest negative effect on gene expression, state 14, is the absence of all measured modifications. The next few states all contain the repressive mark H3K27me3, and are all associated with reduced gene expression. The states with the largest positive effects on expression all have some combination of the activating marks, H3K4me3, H3K4me1, and H3K27ac. The repressive mark is less commonly seen in these activating states.

By merging the information from Figure 2A-C), we were able to suggest annotations for many of the 14 chromatin states (Figure 2D). States with the strongest effects on expression had the clearest annotations, while states with weaker effects remained unannotated.

Spatial distribution of epigenetic modifications around gene bodies

In addition to looking for enrichment of chromatin states near annotated functional elements, we characterized the fine-grained spatial distribution of each state around gene bodies (Figure 3A-B). We similarly characterized the distribution of CpG sites and their percent methylation at this gene-level scale (Figure 3C-D).

The spatial patterns of the individual chromatin states are shown in (Figure 3A), and an overlay of all states together (Figure 3B) emphasizes the difference in abundance between the most abundant states (states 1, 3, and 14), and the remaining states, which were relatively rare.

Each chromatin state had a characteristic distribution pattern relative to gene bodies. For example, state 14, which was characterized by the absence of all measured histone modifications, was strongly depleted near the TSS, indicating that this region is commonly subject to histone modification. However, its abundance increased steadily through the gene to a peak at the TES. In contrast, states 3 and 1 were both concentrated at the TSS. State 3 was very narrowly concentrated right at the TSS, whereas state 1 was more broadly abundant both upstream and downstream of the TSS. Both were associated overall with increased expression in the inbred mice (indicated by red shading), suggesting promoter or enhancer functions. The third state in this group of high-expressing states, state 2, was depleted nere the TSS, but enriched within the gene body, suggesting that this state may mark active intragenic enhancers.

States with weaker effects on expression (indicated by grayer shades) were of lower abundance. However, they still had distinct distribution patterns around the gene body suggesting the possibility of distinct functional roles in the regulation of gene expression.

There were similarly dramatic spatial patterns in DNA methylation (Figure 3C-D). Across all genes, the TSS had densely packed CpG sites relative to the gene body (Figure 3C). As expected, the median CpG site near the TSS was consistently hypomethylated relative to the median CpG site in intergenic regions (Figure 3D). CpG sites within the gene body were slightly hypermethylated compared to intergenic CpGs.

Spatially resolved effects on gene expression

The distinct spatial distributions of the chromatin states and methylated CpG sites around the gene body raised the question as to whether the effects of these states on gene expression could also be spatially resolved. To investigate this possibility we tested the association between both chromatin state and DNA methylation and gene expression with spatially resolved models (Methods). We tested the effect of each chromatin state on expression across genes within hepatocytes (Figure 4A) and the effect of each chromatin state on the variation in gene expression across strains (Figure 4B).

All chromatin states demonstrated spatially dependent effects on gene expression within hepatocytes. For many of the states, the effects on expression were concentrated at or near the TSS, while in the other states effects were seen across the whole gene. The direction of the effects matched the overall effects of each state seen previously (Figure 2). Remarkably, the spatial effects were recapitulated for almost every state when we measured across strains. That is, variation in chromatin state across strains contributed to variation in gene expression in the same manner that cell-type expression was being established. One notable exception was state 6, whose presence upregulated genes within hepatocytes, but did not contribute to expression variation across strains.

We also examined the effect of percent DNA methylation across genes within hepatocytes, and across strains (Figure 5). As expected, methylation at the TSS was associated with lower expression in hepatocytes. However, percent DNA methylation did not contribute at all to expression variation across strains, implying that although percent DNA methylation is used in gene regulation within a cell type, it is not heritable and does not contribute to variation in gene expression across genetically diverse individuals.

Imputed chromatin state explained expression variation in diversity outbred mice

Thus far, we have used inbred strains of mice to identify correlations between local chromatin state and gene expression. However, we cannot establish causality in this population. For that we need a mapping population in which we can associate genetic or epigenetic variation at a single locus with changes in gene expression. A mapping population also allows us to establish the extent to which variation in epigenetic factors contributes to observed expression quantitative trait loci (eQTL).

To compare the contribution of genetic and epigenetic features to eQTLs in a gentically diverse population, we imputed chromatin state, DNA methylation, and SNPs into a population of DO mice (Methods). Chromatin state is largely determined by local genotype, especially early in life [REF], and can thus be reliably imputed from local genotype. Further, we have shown here that local chromatin state correlates with variation in gene expression across inbred strains. DNA methylation, on the other hand, is known not to be highly heritable [REF], and thus cannot be reliably imputed from local genotype. We have also shown here that DNA methylation is not correlated with variation in gene expression across inbred strains. The imputation of DNA methylation thus serves as an estimate of a lower bound the ability of a feature imputed from local haplotype to explain gene expression in a new population.

For each transcript in the DO population, we imputed the local chromatin state across the gene body based on the gene's local founder haplotype and the chromatin state at the corresponding position in the inbred mice. We did the same for DNA methylation and SNPs.

After imputing each genomic feature into the DO population, we mapped gene expression to the imputed features and calculated the variance explained. Examples of each genomic feature and the mapping results for the gene *Pkd2* are shown in Figure 6. There are two particularly interesting regions in this gene. One is at the TSS and the immediately surrounding area, and the other is just downstream of the TSS.

These two regions are colored red, indicating that they are marked by chromatin states with a positive effect on gene expression. The order of the rows in this panel helps illustrate that the strains with the most red in chromatin state space contributed the highest-expressing alleles to the DO (Figure 6E). The two haplotypes with the strongest negative effect on gene expression in the DO have mostly blue chromatin states in these two regions. These two strains also had the lowest expression among the inbred mice (Figure 6F). The concordance between chromatin state and gene expression in the DO is seen as the blue pluses in Figure 6A that are aligned with the two red regions, which we suggest are putative enhancer regions.

The spatial patterns in the SNPs only partially mirror those in chromatin state (Figure Figure 6C). SNPs underlying the putative enhancer regions could potentially influence gene expression by altering chromatin state. But SNPs downstream of this region underly invariant chromatin.

Percent DNA methylation does not vary across the strains in either of these putative enhancer regions, and does not contribute to variation in expression across genetically distinct individuals (Figure 6D).

The overall distributions of variance explained by each feature across all transcripts is shown in Figure 7. These distributions show the haplotype effect for the marker nearest each transcript compared with the maximum effect across the gene body for each of the other imputed features. Overall, local haplotype explained the largest amount of variance of gene expression in the DO ($R^2 = 0.17$). The variance explained by local chromatin state was very highly correlated with that of haplotype (Pearson $r = 0.96$) and explained almost as much variance in gene expression in the DO as local haplotype ($R^2 = 0.15$).

The mean variance explained by SNPs was lower ($R^2 = 0.13$) than that explained by haplotype and was not as highly correlated with local haplotype as chromatin state was (Pearson $r = 0.93$). DNA methylation, the lower bound for variance explained by a feature imputed from local haplotype, explained the lowest amount of expression variance in the DO population ($R^2 = 0.09$), and had a much lower correlation to haplotype than either chromatin state or SNPs (Pearson $r = 0.74$).

Discussion

In this sudy we showed that variation in histone modifications in inbred mice mirrors genetic variation, and we further showed that this variation was highly related to variation in gene expression across strains. These observations suggest that cell type-specific patterns of histone modifications are determined by local genotype, and may be a major mechanism through which expression QTL (eQTL) are generated. This hypothesis was supported by the high concordance between chromatin state, which was imputed from local genotype, and gene expression in an independent outbred population of mice.

The high resolution of the chromatin states combined with spatial patterns of abundance and effect on gene expression offers opportunities for the annotation of functional elements in and around genes. For example, the chromatin state patterns in the gene *Pkd2*, suggest two enhancers – one at the TSS, and the other just downstream of the TSS inside the gene body. The positive effects of these putative enhancer regions in the inbred mice were replicated in outbred mice suggesting that these effects are robust and contribute to variation in gene expression seen in diverse populations.

The putative enhancers are not apparent in the SNP patterns or in the patterns or DNA methylation, which suggests that chromatin modification is the primary mechanism through which gene expression is regulated by these regions. Further, the richness of the information in this chromatin state layer provides data with which to further annotate the effects of SNPs underlying these regions. There are SNPs throughout the gene, as seen in Figure 6, and many of them are associated with variation in gene expression. However, while the SNPs within the putative enhancer regions may change expression by altering histone modifications placed in those regions, SNPs futher downstream may work through another mechanism, such as through directly dirsupting transcription, or by altering the transcript such that it is processed differently post transcriptionally. The intermediate resolution of the chromatin state between that of SNPs and haplotype thus provides a highly informative layer of information between genotype and gene expression.

In contrast to chromatin state, percent DNA methylation was not associated with variation in gene expression across inbred strains or in the outbred population. This was largely due to a lack of variation in methylation across strains. An example of this observation is shown in panel D of Figure 6. Despite strain variation in both genotype and chromatin state at the TSS of *Pkd2*, DNA methylation is invariant – the CpG island at the TSS is unmethylated in all strains. Thus, although chromatin state appears to be highly influenced by local genotype, percent DNA methylation is not.

Similar observations have been made in human studies [33931130]. Multiple twin studies have estimated the average heritability of individual CpG sites to be roughly 0.19 [27051996, 24183450, 22532803], with only about 10% of CpG sites having a heritability greater than 0.5 [24183450, 22532803, 24887635]. Trimodal CpG sites, i.e. those with methylation percent varying among 0, 50, and 100%, have been shown in human brain tissue to be more heritable than unimodal, or bimodal sites ($h^2 = 0.8 \pm 0.18$), and roughly half were associated with local eQTL [20485568]. Here, we did not see an association between trimodal CpG sites and gene expression across strains (Supplemental Figure XXX).

The diversity in the effects observed in the 14 chromatin states highlights the importance of analyzing combinatorial states as opposed to individual histone modifications. To illustrate this point, consider the three states with the largest positive effects on transcription. Each of these three states had a distinct combination of the three histone marks associated with transcriptional activation: H3K4me1, H3K4me3, and H3K27ac. State 3 was characterized by high levels of H3K4me3 and H3K27ac, and low levels of H3K4me1. State 2 was characterized by high levels of H3K4me1 and H3K27ac, and low levels of H3K4me3. And state 1 was characterized by high levels of all three activating marks (Figure XXX). Although all three states were associated with increased gene expression, each had a completely distinct spatial distribution. State 3 was distributed in a very narrow band centered on the TSS, while state 1 was distributed across a much broader region centered upstream of the TSS. State 2 had a completely different distribution – it was depleted at the TSS, and most abundant within the gene body and near the TES. This variation in spatial distribution was mirrored in the spatial effects on transcription. State 3, which we annotated as an active promoter, was positively associated with transcription when it was present at the TSS. In contrast, states 2 and 1, which we annotated as enhancers, were associated with increased transcription when present anywhere in the gene body (Figure XXX). We would not be able to detect such patterns if analyzing the histone modifications in isolation. These results highlight the complexity of the histone code and the importance at analyzing combinatorial states.

While we were able to annotate several states, particularly those with the strongest effects on gene expression, other states were more difficult to annotate. This raises the intruiguing possibility of identifying new modes of expression regulation through histone modification. One of these unannotated states, state 9, had a weak, but consistent negative effect on gene transcription centered within the gene body, downstream of the TSS. This state was characterized by high levels of H3K4me3 and low levels of the other three modifications.

The modification H3K4me3 is most frequently associated with increased transcriptional activity [citation], so the association with state 9 with reduced transcription is a deviation from the dominant paradigm. The physical distribution of this state is also interesting. It was depleted at the TSS, and enriched just upstream and just downstream of the TSS (Figuree XXX). It was also enriched just downstream of the TES, although it did not appear to influence transcription at this location (Figure XXX). The group of genes marked by state 9 were enriched for functions such as stress response, DNA damage repair, and ncRNA processing suggesting that this state may be used to regulate subsets of genes involved in responses to environmental stimuli.

There were other states that we were able to annotate, but were not necessarily expecting to see in this study. We detected two bivalent states, which are states that combine an activating histone modification and a repressing histone modificaction and are usually associated with undifferentiated cells [citation]. Here we identified two bivalent states in adult mouse hepatocytes, and annotated them as a poised enhancer (state 12) and a bivalent promoter (state 11). Both states were associated with downregulation across inbred strains when present near the TSS; however this effect was not replicated in the outbred mice. The lack of replication was perhaps because the effect was too weak to detect given the number of animals in the population.

Both bivalent promoters and poised enhancers are dynamic states that change over the course of differentiation and in response to external stimuli [citation]. Bivalent promoters have been studied primarily in the context of development. They are abundant in undifferentiated cells, and are typically resolved either to active promoters or to silenced promoters as the cells differentiate into their final state [23788621, 22513113]. These promoters have also been shown to be important in the response to changes in the environment. Their abundance increases in breast cancer cells in response to hypoxia [27800026]. Poised enhancers are also observed during differentiation and in differentiated cells [32432110]. In concordance with these previous observations, the genes marked by states 12 and 11 were enriched for vascular development and morphogenesis. That we identified these states in differentiated hepatocytes may indicate that a subset of developmental genes retain the ability to be activated under certain circumstances, such as during liver regeneration in response to damage. It is also possible these states were induced in the inbred strains in respose to stress, rather than genetically coded. This could explain why the negative effect on gene expression was not replicated in the outbred mice. However, given that we detected this state in all nine inbred strains in relatively equal proportions, this latter hypothesis seems less likely.

Broadly, local variation in chromatin state was highly correlated with variation in gene expression across individuals, an observation that was replicated in an independent population of genetically diverse, outbred mice. The percent variance explained by chromatin state closely matched that of haplotype, and exceeded that of individual SNPs. These results suggest two things. First, a large portion of the effect of local haplotype on gene expression in mice is likely mediated through variation in chromatin state. Second, the intermediate resolution of chromatin state between that of individual SNPs and broad haplotypes carries important imfornation that cannot be resolved at the other levels. Individual SNPs, although, sometimes causally linked to trait variation, are highly redundant and cannot be readily used to annotate functional elements in the genome. Haplotypes aggregate genomic information over broad regions and are a powerful tool to link genomic variation to trait variation. However, they are usually too broad to be used to annotate regions less than a few megabases in length. By combining the mapping power of haplotypes, the high resolution of SNPs, and the intermediate resolution of chromatin states, we can begin to build mechanistic hypotheses that link genetic variation to variation in physiology. Understanding the role that genetic variation plays in modifying the chromatin state landscape will be critical in making these links. Through this survey we are providing one of the first rigorous resources that explores the connection between genetic variation and epigenetic variation.

work this paragraph in... That states 14 and 13 were associated with reduced gene expression both within hepatocytes and across strains suggests that there may be differential epigenetic silencing of genes in hepatocytes across strains. Further, the majority of chromatin states were associated with variation in expression across strains, suggesting that epigenetic regulation of gene expression through histone modification may contribute substantially to variation in gene expression across genetically distinct individuals. That most states have the same effects across genes within a cell type and across strains suggests that the mechanisms that are used to regulate cell type specificity also contribure to variation in genetically distinct individuals.

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?).

Figure Legends

Fig 1. The first two principle components of each genomic feature across nine inbred strains of mouse. In all panels each point represents an individual mouse, and strain is indicated by color as shown in the legend at the bottom of the figure. Each panel is labeled with the data used to generate the PC plot. (A) Hepatocyte transcriptome - all transcripts sequenced in isolated hepatocytes. (B) DNA methylation - the percent methylation at all CpG sites shared across all individuals. (C-F) Histone modifications - the peak heights of the indicated histone modification for sites shared across all individuals.

Fig 2. Overview of chromatin state composition, genomic distribution, and effect on expression. The left most panel shows the emission probabilites for each histone modification in each chromatin state. Blue indicates the absence of the histone modification, and red indicates the presence of the modification. The panel labeled genomic enrichment shows the distribution of each state around functional elements in the genome. Red indicates that the state is more likely to be found near the annotated functional element than expected by chance. Blue indicates that the state is less likely to be found near the annotated functional element than expected by chance. Abbreviations are as follows: TFBS = transcription factor binding sites, cCRE = candidate cis-regulatory element [@pmid32728249], TSS = transcription start site, TES = transcription end site. The panel labeled Expression Effects shows the effect of variation in the state on gene expression. Bars are colored based on the size and direction the state's effect on expression. Darker bars show the effects on expression of chromatin state variation across strains. Tan bars show the effects on expression of chromatin state variation across genes. The final column of the figure shows plausible annotations for each state based on combining the data in the previous three panels. The numbers in parentheses indicate the percent of the genome that was assigned to each state.

Fig 3. Relative abundance of chromatin states and methylated DNA. A. Each panel shows the abundance of a single chromatin state relative to gene TSS and TES. The y-axis in each panel is the proportion of genes containing the state. Each panel has an independent y-axis to better show the shape of each curve. The x-axis is the relative gene position. The TSS and TES are marked as vertical gray dashed lines. B. The same data shown in panel A, but with all states overlayed onto a single set of axes to show the relative abundance of the states. C. The density of CpG sites relative to the gene body. The y-axis shows the number of CpG sites per base pair. The density is highest near the TSS. CpG sites are less dense within the gene body and in the intergenic space. D. Percent methylation relative to the gene body. The y-axis shows the median percent methylation at CpG sites, and the x-axis shows relative gene position. CpG sites near the TSS are unmethylated relative to intragenic and intergenic CpG sites.

Supplemental Figure Legends

References

1. Lawrence M, Daujat S, Schneider R. Lateral thinking: How histone modifications regulate gene expression. Trends in Genetics. Elsevier; 2016;32: 42–56.

2. Jones PA. Functions of dna methylation: Islands, start sites, gene bodies and beyond. Nature Reviews Genetics. Nature Publishing Group; 2012;13: 484–492.

3. Moore LD, Le T, Fan G. DNA methylation and its basic function. Neuropsychopharmacology. Nature Publishing Group; 2013;38: 23–38.

4. Ernst J, Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat Biotechnol. 2010;28: 817–825.

5. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 2011;473: 43–49.

6. Wiench M, John S, Baek S, Johnson TA, Sung MH, Escobar T, et al. DNA methylation status predicts cell type-specific enhancer activity. EMBO J. 2011;30: 3028–3039.

7. Ji H, Ehrlich LI, Seita J, Murakami P, Doi A, Lindau P, et al. Comprehensive

Fig 4. Effects of chromatin states on gene expression. Each column shows the effect of each chromatin state on gene expression in a different experimental context. The first column shows the effect across genes in the inbred mice showing how chromatin states are used within a single tissue to increase the expression of some genes and decrease the expression of other genes. The second column shows the effect of chromatin state on gene expression across strains, showing how variation in chromatin state across strains leads to variation in expression of individual genes across strains. The third column shows the effect of imputed chromatin state on gene expression in a population of diversity outbred mice. These plots show the effect on local gene expression of variation in chromatin state across genetically diverse individuals. Each column of panels is plotted on a single scale for the y-axis so the magnitude of the effects in a single column can be compared directly to each other. Across a single row, the scale of the y-axis varies to highlight the similarity of the shape of each curve in each different setting. The final column shows the annotation of each state for comparison with its effects on gene expression. All y-axes show the β coefficient from the linear model shown in equation [REF]. All x-axes show the relative position along the gene body running from just upstream of the TSS to just downstream of the TES. Vertical gray dashed lines mark the TSS and TES in all panels.

Fig 5. Effect of DNA methylation on gene expression (A) across gene expression in hepatocytes and (B) across inbred strains. Dark gray line shows estimate of the effect of percent DNA methylation on gene expression. The x-axis is normalized position along the gene body running from the transcription start site (TSS) to the transcription end site (TES), marked with vertical gray dashed lines. The horizontal solid black line indicates an effect of 0. The shaded gray area shows 95% confidence interval arond the model fit.

methylome map of lineage commitment from haematopoietic progenitors. Nature. 2010;467: 338–342.

8. Collas P. The current state of chromatin immunoprecipitation. Mol Biotechnol. 2010;45: 87–100.

9. Svenson KL, Gatti DM, Valdar W, Welsh CE, Cheng R, Chesler EJ, et al. High-resolution genetic mapping using the Mouse Diversity outbred population. Genetics. 2012;190: 437–447.

10. Ashbrook DG, Arends D, Prins P, Mulligan MK, Roy S, Williams EG, et al. The expanded BXD family of mice: A cohort for experimental systems genetics and precision medicine. bioRxiv. 2019;139: 387–64.

11. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nature Publishing Group. 2012;9: 215–216.

12. Chesler EJ, Miller DR, Branstetter LR, Galloway LD, Jackson BL, Philip VM, et al. The Collaborative Cross at Oak Ridge National Laboratory: developing a powerful resource for systems genetics. Mammalian Genome. 2008;19: 382–389.

13. Gatti DM, Svenson KL, Shabalin A, Wu L-Y, Valdar W, Simecek P, et al. Quantitative trait locus mapping methods for diversity outbred mice. G3 (Bethesda, Md). 2014;4: 1623–1633.

14. Langmead B. Aligning short sequencing reads with Bowtie. Curr Protoc Bioinformatics. 2010;Chapter 11: Unit 11.7.

15. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26: 139–140.

16. Leek JT, Johnson WE, Parker HS, Fertig EJ, Jaffe AE, Zhang Y, et al. Sva: Surrogate variable analysis. 2020.

Fig 6. Example of epigenetic states and imuptation results for a single gene, *Pkd2*. (A) The variance in DO gene expression explained at each position along the gene body by each of the imputed genomic features: SNPs - red X's, Chromatin State - blue plus signs, and Percent Methylation - green circles. The horizontal dashed line shows the variance explained by the haplotype. For reference, the arrow below this panel runs from the TSS of *Pkd2* to the TES and shows the direction of transcription. (B) The chromatin states assigned to each 200 bp window in this gene for each inbred mouse strain. States are colored by their effect on gene expression in the inbred mice. Red indicates a positive effect on gene expression, and blue indicates a negative effect. Each row shows the chromatin states for a single inbred strain, which is indicated by the label on the left. (C) SNPs along the gene body for each inbred strain. The reference genotype is shown in gray. SNPs are colored by genotype as shown in the legend. (D) Percent DNA methylation for each inbred strain along the *Pkd2* gene body. Percentages are binned into 0% (blue) 50% (yellow) and 100% (red). (E) Haplotype effects for expression of *Pkd2* in the DO. Haplotype effects are colored by from which each allele was derived. (F) *Pkd2* expression levels across inbred mouse strains. For ease of comparison, all panels B through F are shown in the same order as the haplotype effects.

Fig 7. Chromatin state explains variation in gene expression in an outbred population. A. Distributions of gene expression variance explained by different genomic features: local haplotype, local imputed chromatin state, local SNP genotype, and local imputed DNA methylation status. B. Direct comparisons of variance explained by local haplotype, and the three other genomic features: imputed chromatin state, SNP genotype, and imputed DNA methylation status. Blue lines show $y = x$. Each point is a single transcript.

17. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9: R137.

18. Thompson MJ, Chwiałkowska K, Rubbi L, Lusis AJ, Davis RC, Srivastava A, et al. A multi-tissue full lifespan epigenetic clock for mice. Aging (Albany NY). 2018;10: 2832–2854.

19. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics. 2011;27: 1571–1572.

20. Ernst J, Kellis M. Chromatin-state discovery and genome annotation with ChromHMM. Nat Protoc. 2017;12: 2478–2492.

21. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44: D733–745.

22. Lesurf R, Cotto KC, Wang G, Griffith M, Kasaian K, Jones SJ, et al. ORegAnno 3.0: a community-driven resource for curated regulatory annotation. Nucleic Acids Res. 2016;44: D126–132.

23. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome Res. 2002;12: 996–1006.

24. Dreos R, Ambrosini G, Groux R, Cavin Périer R, Bucher P. The eukaryotic promoter database in its 30th year: focus on non-vertebrate organisms. Nucleic Acids Res. 2017;45: D51–D55.

25. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489: 57–74.

26. Broman KW, Gatti DM, Simecek P, Furlotte NA, Prins P, Sen Ś, et al. R/qtl2: Software for Mapping Quantitative Trait Loci with High-Dimensional Data and Multiparent Populations. Genetics. 2019;211: 495–502.

Fig 8. Comparison of emissions probabilities across all ChromHMM models. Each row contains data for a single ChromHMM model fit to the number of states indicated on either side of the row. Each set of four columns shows data for each of the four histone modifications. Each set is separated from the next by a column of gray for ease of visualization. The bottom row, the reference row, shows the ideal state that all model states are being compared to. Blue indicates absence of the histone mark and red indicates presence. For each ChromHMM model, each state was assigned to one of the reference states using an emissions probability of 0.3 as a threshold for presence of the histone modification. If a state was not present in the given model, the corresponding area is shown in gray. Emissions probabilities near 0 are shown in blue, and probabilities near 1 are shown in red. Orange and yellow indicate intermediate probabilities. Aligning the states across all models shows a remarkable stability in the emissions across models, seen as vertical bars of consistent color.

Fig 9. Comparison of state abundance across all ChromHMM models. The left-most column shows the annotation for each state. Unannotated states are marked with a dash. The binary heatmap indicates which histone modifications were present in each state: 1 indicates presence, and 0 indicates absence. The histone modifications are labeled at the bottom of each column. The continuous heatmap shows the abundance of each state (in rows) in each ChromHMM model (in columns). The abundance is the proportion of transcribed genes with the state present. Less abundant states are shaded blue, and more abundant states are shaded yellow, orange, and red. The number of states in the model is indicated at the bottom of each column. The black box highlights the model used in this study – the 14-state model. State abundance was remarkably stable across the different models.

27. Jockenhövel F, Grandt D, Weber F, Fritschka E, Philipp T. [Plasma exchange as therapy of recurrent hemolytic-uremic syndrome (HUS) in adults]. Med Klin (Munich). 1991;86: 419–422.

28. Yalcin B, Wong K, Agam A, Goodson M, Keane TM, Gan X, et al. Sequence-based characterization of structural variation in the mouse genome. Nature. 2011;477: 326–329.

Fig 10. Comparison of state effect across all ChromHMM models. This figure is identical to Figure 9, except that the cells in the continuous heatmap show the effect of each state on gene expression across all ChromHMM models. The effect was the β coefficient derived from a linear model. Similar to state abundance, the effects were remarkably stable across models.

Fig 11. Schematic for imputation of histone modifications into the DO mice. For a single transcript imputation was made by multiplying a three-dimensional array, containing chromatin state by strain by position, by a two-dimensional array, contatining haplotype probabilities by DO individual, to create a three-dimensional array, containing individual by position by chromatin state probability.