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

This abstract is in the yaml header. The easier-to-edit one is below.

Author summary

The author summary goes here

Abstract

It is well known that epigenetic modifications, such as histone modifications, and DNA methylation are a major mode of regulating gene transcription.

It is not well known how variation in epigenetic modifications across genetically distinct individuals contributes to heritable variation in gene expression.

When we map an eQTL, how much of the effect of the eQTL is mediated through epigenetic modifications?

We investigated this question in genetically diverse mice.

local imputed histone modifications matched eQTL extremely well, suggesting that a large portion of variation in gene expression mapped to local genotype is mediated through histone modifications.

In contrast percent DNA methylation is not determined by local genetics, and does not contribute to eQTLs.

Introduction

It is well established that epigenetic modifications, such as histone modifications, and DNA methylation influence gene expression [26704082, 22641018, 22781841]. Across cell types, unique combinatorial patterns of histone modifications mark chromatin states that establish cell type-specific patterns of gene expression [20657582, 21441907]. Similarly, the methylation of CpG sites around gene promoters and enhancers influences transcription in a cell type-specific manner [21701563, 20720541].

Cell type-specific 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. 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 in 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 countlessindependent 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 improve our understanding of the mechanics of gene regulation, improve our understanding of how cell type-specific epigenetic landscapes are established, and to improve the functional annotation of the genome as it relates to the regulation of gene expression.

Advances in chromatin immunoprecipitation (ChIP) and sequencing technologies now enable genome-wide surveys of histone modifications with relatively few cells [20077036], thus opening the door to the

Such variation can be mapped to genetic variation

Estimates of the heritability

It is not well known how variation in epigenetic modifications across genetically distinct individuals contributes to heritable variation in gene expression.

Estimates of the heritability of epigenetic features range widely

Early in life, the landscape of epigenetic patterns is

When we map an eQTL, how much of the effect of the eQTL is mediated through epigenetic modifications?

We investigated this question in genetically diverse mice.

We conducted a survey of four histone modifications known to be correlated with gene transcription across nine inbred strains of mice. We also surveyed DNA methylation in these strains.

We looked at how both histone modifications and DNA methylation were associated with transcription variation across strains. We further imputed epigenetic states in a population of diversity outbred mice to more directly investigate the extent to which eQTLs are driven by variation in epigenetic modifications

histone modifications, at least early in life, are determined by local genotype.

DNA methylation is not determined genetically

GWAS hits tend to be in non-coding regions of the genome estimated that most common disease variants work by altering gene expression rather than protein function

These disease-associated SNPs likely fall into functional regions of the genome

eQTLs - what are we measuring when we measure eQTL?

The identity of a hepatocyte is deterimined through patterns of gene expression. Patterns of gene expression are determined in part through patterns of genotype, DNA methylation, and chromatin modifications.

Within a given cell type, how do variations in local genetics and epigenetics influence gene expression?

Across mouse strains, gene expression in hepatocytes is largely similar. For the most part, genes that are highly expressed in one strain are highly expressed in another. However, there are subtle variations in gene expression that are based on strain background.

This variation in gene expression across strains is related to genetic and epigenetic factors. Here we explore how local genotype, chromatin modifications, and DNA methylation influence strain differences in gene expression.

patterns of chromatin state in hepatocytes varied across strains patterns of DNA methylation in hepatocytes varied across strains patterns of gene expression in hepatocytes varied across strains

major axes of variation were similar in all cases, i.e. PWK and CAST were most divergent, while other strains clustered together

Each of these epigenetic-expression patterns represent functioning hepatocytes these are "good enough" solutions to make hepatocytes [22859671]

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

Materials and Methods

Inbred Mice

information about housing, animal use, etc.

Hepatocyte acquisition

Samples were taken from 12-week female mice of nine inbred mouse strains: 129S1/SvImJ, A/J, C57BL/6J, CAST/EiJ, DBA/2J NOD/ShiLtJ, NZO/HlLtJ, PWK/PhJ, and WSB/EiJ. Eight of these strains are the eight strains that served as founders of the Collaborative Cross/Diversity Outbred mice [REF]. The ninth strain, DBA/2J, will facilitate the interpretation of existing and forthcoming genetic mapping data obtained from the BxD recombinant inbred strain panel [REF]. Mice were aged and processed in groups to maintain a steady sample preparation workflow. Mice were housed, born, and aged in the same mouse room, with uniformity in timing, diet, and all other possible conditions. Female mice were used for all experiments due to potentially confounding effects from variation in testosterone among males that can affect liver gene expression, as well general experience that female expression is less variable than male in multiple tissues. This will also facilitate the analysis of maternal effects on offspring in later studies. Three mice were used from each strain.

Liver perfusion

To purify hepatocytes from the liver cell population, the mouse livers were perfused with collagenase to digest the liver into a single-cell suspension, and then isolated using centrifugation. Mice were harvested at 9:00 AM and sacrificed by cervical dislocation. Mice were placed over a stack of paper towels in preparation to catch excess liquid, and the appendages were pinned out to hold the body in place. to keep the fur from contaminating the liver sample later, the fur was wiped down with 70% ethanol. The mouse skin was then cut open and peeled back to the appendages to allow clear access to the abdominal cavity. The fascia was cut open and back to the ribs, being careful to not nick the liver. Moving the intestines and stomach to the right side, the vena cava and hepatic portal vein should be clearly visible below the liver.

For the perfusion, a 23G x $\frac{3}{4}$'' BD Vacutainer Safety-Lok needle (REF 367297) was attached to 1.6mm ID BioRad Tygon tubing (R-3603) connected to a Pharmacia peristaltic pump that allows a flow of up to 8 ml/min. The liver will be processed with three solutions: 5mM EGTA in Leffert's buffer, Leffert's buffer wash, and 87 CDU/mL Liberase collagenase with 0.02% CaCl2 in Leffert's buffer. The three solutions were at 37°C before perfusion.

The needle was placed into the vena cava for the perfusion superior to the kidneys and inferior to the liver. With the peristaltic pump running slowly, the vena cava was pierced at shallow 15° angle and the needle was inserted to a shallow depth (around 2-3mm of the needle tip in the vena cava). Once the needle is inserted into the vena cava, the volume on the peristaltic pump is increased to 5-7mL/min. The liver will immediately blanch, and the hepatic portal vein is immediately severed to allow flushing of the liver.

The 1x EGTA buffer was used to flush the blood out of the liver and start the digestion of the desmosomes connecting the liver cells. To help with the perfusion, pressure was applied to the hepatic portal vein for 5 second intervals causing more solution to be forced through the liver, which can be seen visually by the liver swelling. After 35ml of the 1x EGTA solution is passed through the liver, the solution was switched to the 1x Leffert's buffer. The pump was turned off during the switch to prevent air from being sucked into the tubing while the tubing is transferred to the new solution. To wash, 7-10ml of the Leffert's buffer was passed through the liver to flush out the EGTA, which otherwise chelates the calcium ions necessary for collagenase activity in the next buffer. The pump was turned off again to switch to the Liberase solution. To digest the liver, 25-50mL of Liberase solution (~ 4.3 wunsch units) was passed through the liver. Throughout the perfusion process, periodic pressure was applied to the hepatic portal vein to help pump the buffers more completely through the liver. As the liver was digested with the Liberase, it will swell and look soggy and limp. Over-digestion leads to increased contamination with non-hepatocyte cell types, and further reduces cell viability.

After perfusion, which takes around 15-20min to complete, the liver was carefully cut out of the abdominal cavity and placed in a petri dish with 35 mL ice-cold Leffert's buffer with 0.02% $CaCl_2$. The digested liver was passed through Nitex 80 μm nylon mesh (cat #03-80/37) into a 50mL conical, using additional ice-cold Leffert's buffer with 0.02% CaCl2 if necessary, and a rubber policeman. After the liver cells from both animals were collected, they were put through two wash and spin cycles to purify the hepatocytes and remove other types of cells. To isolate the hepatocytes, the much larger size of the hepatocyte cells was exploited in very slow 4 min, 50 x g spins that leave smaller other cell types in suspension. After each spin, the solution was decanted as waste, and the enriched cell pellet of hepatocytes was resuspended in 30ml ice-cold Leffert's buffer with 0.02% $CaCl_2$. After the second spin, the solution should be almost clear, indicating that other cell types have been removed. The hepatocytes are resuspended in room temperature PBS, counted, and volume adjusted to $1x10^6$ cells/mL.

We aliquoted $5x10^6$ cells for each RNA-Seq and bisulfite sequencing, and the rest were cross-linked for ChIP assays. Two $5x10^6$ aliquots (5mLs) of liver cells were removed into two 15mL conicals. These 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). Meanwhile, 37% formaldehyde in methanol (VENDOR) were added to the remaining cells to a final concentration of 1%. The cells were rotated at room temperature for 5 min to cross-link protein complexes to the DNA bound to them. After cross-linking, 10x glycine (VENDOR) is added to a final concentration of 125 mM and rotated for 5 min to quench the formaldehyde and stop cross-linking. The cells were spun down at 2000 rpm for 5 min, decanted, and resuspended in PBS to $5x10^6$ cells/mL. The cells were divided into $5x10^6$ aliquots in 2mL tubes. The tubes were spun down again at 5000 x g for 5 min, decanted, and the cell pellets frozen in liquid nitrogen. All cell samples were stored at -80°C until used.

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 is 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 Lab Gene Expression Service for RNA extraction and library synthesis.

Histone chromatin immunoprecipitation assays

The H3K4me1 and H3K4me3 histone chromatin immunoprecipitation assays were performed on cross-linked hepatocytes using similar protocols. For all histone ChIP assays, the crosslinked chromatin was prepared the same way. First, the aliquot of $5x10^6$ hepatocyte cells was lysed to release the nuclei by rotating the sample in hypotonic buffer for 20 min at 4°C. The cells were pelleted by spinning for 10min, 10K x G, at 4°C. The cells were resuspended in 130ul MNase buffer with 1mM PMSF (VENDOR) and 1x protease inhibitor cocktail (Roche VENDOR) to prevent histone protein degradation, then digested with 15U of MNase. The micrococcal nuclease digests the exposed DNA, but leaves the nucleosome-bound DNA intact. After 10min of incubation at 37°C, the chromatin was digested into primarily mononucleosomes. This was confirmed by DNA-purification of the MNase-digested chromatin run out on an agarose gel, which yielded mostly 150bp fragments, and few 300bp fragments. The MNase digestion was stopped by adding EDTA to 10mM, and incubating on ice for 5 min. The digested chromatin was purified by spinning out insoluble parts at top speed for 10 min at 4°C. The chromatin was transferred to a new tube and spun again to further remove impurities and reduce background in the ChIP assays. The final chromatin was transferred to a fresh tube, and used immediately in the ChIP.

To prepare for the ChIP, $20\mu L/1x10^6$ cells Dynabead Protein G beads were aliquoted into an Eppendorf tube. A magnetic tube holder was used to attract the beads to the wall of the tube, and then the solution was carefully pipetted off, leaving only the beads behind. The beads were washed twice with buffer to prepare them for binding to the antibody. For this binding step and the chromatin binding step, the buffer used was either RIPA buffer for the H3K4me3 and K3K27me3 ChIPs, or ChIP buffer (VENDOR) for the H3K4me1 ChIP. The ChIP buffer was gentler and less stringent than RIPA buffer, which was better for the weaker binding of the H3K4me1 antibody that was used. The buffers were supplemented with 50 mg/mL BSA (VENDOR) and 0.5 mg/mL Herring Sperm DNA, both of which are blocking agents that reduce background and non-specific binding. The ChIP assays also varied in the amount of input chromatin and corresponding size of the reaction that was necessary to yield sufficient DNA for sequencing. H3K4me3 ChIP needed only $1.5x10^6$ cells, and H3K4me1 and K3K27me3 ChIP used $4x10^6$ cells. To perform the ChIP, 20μL of Dynabeads per $1x10^6$ cells is incubated with $5\mu L$ of histone antibody for $>$ 20min in $50\mu L/1x10^6$ cells RIPA (or ChIP) buffer supplemented with 50 mg/mL BSA, 0.5 mg/mL Herring Sperm DNA, 1xPIC, and 1mM PMSF. The antibodies used were (XXX). Once the antibody was bound to the Dynabeads, the beads were washed twice with $100\mu L/1x10^6$ cells RIPA buffer with BSA and Herring Sperm DNA.

Next, the MNase-digested chromatin were added, which was at a concentration of $1x10^6$ cells/$25\mu L$. The ChIP reaction was incubated overnight with rotation at 4°C, to allow the histone protein to bind to the antibody, which was bound to the magnetic beads. In order to calculate enrichment for each ChIP sample, a known amount (10 or $20\mu L$) of MNase-digested input chromatin was saved.

The next morning, the ChIPs underwent a series of washes to remove unbound chromatin. The H3K4me3 and H3K27me3 ChIPs were washed 3x with $100\mu L/1x10^6$ cells RIPA buffer, and the H3K4me1 ChIP was washed with a low salt wash (0.1% SDS, 1% Triton X-100, 2mM EDTA, 20mM Tris-HCl pH 8, 150 mM NaCl), a high salt wash (0.1% SDS, 2% Triton X-100, 2mM EDTA, 20mM, Tris-HCl, pH 8, 500mM NaCl), and a LiCl wash (0.25 MLiCl, 1% IGEPAL-CA630, 1% deoxycholic acid (sodium salt), 1 mM EDTA, 10 mM Tris-HCl pH 8). After three washes, the ChIPs were washed twice with TE buffer and transferred to a new tube during the last TE wash to reduce background. At this point, the histone of interest and the histone-bound DNA fragment had been purified from the MNase-digested, cross-linked chromatin, and was bound by histone-specific antibody to the magnetic Dynabeads. In the next step, a high-salt elution buffer is used to degrade the antibody binding interactions to the beads and the histone, and concurrently, proteinase K is added to digest the protein away from the DNA-protein complexes. The ChIP was incubated with the elution buffer and proteinase K at 68°C for $>$ 6 hours to liberate the DNA. At the same time, the saved input chromatin was also digested in the same buffer. Afterwards, the beads were removed using the magnet, and the DNA was purified using the Qiagen PCR purification kit. Quantification was performed using the Qubit quantification system, which is accurate to $0.02ng/\mu L$ and only requires a small amount of sample to measure concentration. The ChIP sample was enriched for only DNA that was bound to the histone of interest. The goal for each ChIP was to yield 10 ng of ChIP DNA for sequencing. Not all samples met this criterion, and the H3K4me1 ChIPs often had a total yield of $\sim 2ng$ of DNA.

To test the efficiency of the ChIPs, quantitative PCR using QuantiFAST was performed. Two sets of primers were used, one set in a known region of histone binding (positive control), and one set in a region without histone binding (negative control). The qPCR was performed both on the ChIP DNA and the input DNA. Then the relative enrichment of positive vs negative assays was compared between the ChIP and input DNA.

The ChIP DNA was submitted to The Jackson Lab GES service for library preparation and sequencing. Libraries were made using the Kapa Hyper Prep kit with adapters at $0.6\mu M$. The libraries were amplified by 10 cycles of PCR. These libraries were not size selected, although most fragments were $\sim$ 150 bp due to MNase-digestion. The samples were sequenced with 40 or more million reads per sample, which is almost 2x more reads than the ENCODE project, which sequenced using 20 million reads.

Diversity Outbred mice

We used previously published data from a population of diversity outbred (DO) mice [Svenson et al. 2012] 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 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).

Founder haplotypes were inferred from SNPs using a Hidden Markov Model as described in Gatti$_{et\ al.}$2014. 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

Sequencing

The raw sequencing data from both RNA-Seq and ChIP-Seq was put through the quality control program FastQC. FastQC identifies problems or biases in either the sequencer run or the starting library material. The FastQC readout includes total number of reads, sequence quality, duplication level, and overrepresented sequences. All of our samples had comparable quality levels and no outstanding flags. However, the ChIP-Seq data was flagged for having a high level of duplicate reads. This can be explained by the use of MNase to shear the DNA into 150 bp fragments. If the binding positions of nucleosomes are fixed, then the MNase enzyme will cleave the DNA in the same place in multiple cells, resulting in duplicate pieces of DNA. Despite evidence that the duplication rate has a biological explanation, duplicates were removed before downstream analysis, as is typical in sequencing workflows, to avoid potential biases caused by starting libraries that have less diversity.

For the sequence analysis, reads from each sample were mapped to strain-specific pseudogenomes that integrate known SNPs from each strain. While the B6 samples were aligned directly to the reference mouse genome, the other samples were from genetically different strains. Strain-specific sequence variation in transcripts can affect alignment quality and result in biased estimates of abundance. To counteract potential strain biases, sequencing data from each strain were aligned to a custom strain pseudogenome, allowing a more precise characterization of gene expression and histone binding. The pseudogenomes were created using the EMASE computational program [REF] designed to construct customized genomes based on known SNP and indel attributes. The resulting custom genomes are called pseudogenomes, because they are based on inserting small known variations into the reference genome, but do not attempt whole genome sequencing for each strain and complete rebuild the entire genomic sequence from the scaffold up. The strain-specific pseudogenomes were then used in the Bowtie mapping algorithm to align and map reads from the RNA-Seq and ChIP-Seq experiments.

Quantifying gene expression

Once the sequencing data was mapped to the custom genomes, edgeR is used to quantify transcripts. The edgeR program uses a Trimmed Mean of M-values (TMM), which adjusts each sample for library size and RNA composition using the assumption that most genes are not differentially expressed. The output is sample read count for each of the ENSMUSG transcript ID's. Next, transcripts with less than 1 CPM in two or more replicates were filtered to remove lowly expressed genes. Also, the data were trimmed to include only protein-coding transcripts.

ChIP-Seq quantification:

After the ChIP-Seq sequencing data were mapped to the custom pseudogenomes, peaks were called in each sample using MACS 1.4.2 [REF], with a significance threshold of $p \leq 10^{-5}$. In order to compare peaks across strains, the MACS output peak coordinates were converted to common B6 coordinates using g2g tools [REF].

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

Quantifying DNA methylation

Annat's stuff to get bed files.

Analysis

Filtering transcripts

For all gene expression data, we remove transcripts with extremely low read counts, by filtering out those whose mean read count across all individuals was less than five.

We used the R package sva [REF] 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.

Analysis of histone modifications

Identification of chromatin states

We used ChromHMM [29120462] to identify *chromatin states*, which are unique combinations of the four chromatin modifications, for example, the presence of both H3K4me3 and H3K4me1, but the absence 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 then investigated a number of features of each state in each model: presence/absence of histone modifications, distribution patterns across the genome, and the effect of each state on gene expression. We compared chromatin states from the different models based on these analyses and selected the 14-state model. Each of these analyses, and the model comparison, are described below.

Emission probabilities

Emission probabilites are a primary output of ChromHMM (Figure XXXA). They define the probability that each histone mark is present in each detected state. Low probabilities suggest absence, or low levels of the mark, and high probabilities suggest presence. To compare states to each other and to annotate states, we declared a histone mark to be present in a state if its emission probability was 0.3 or higher.

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 [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.

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 generated a chromatin state matrix with genomic position in rows, and mouse strains in columns. Each cell contained the chromatin state assignment for a 200 base pair (bp) window, defined by ChromHMM, for each strain. We normalized these bp positions for each gene, such that they ran from 0 at the transcription start site (TSS) to 1 at the transcription end site (TES). 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 normalize the coordinates, we first centered all coordinates on the TSS of the gene by subtracting off the base pair position of the TSS. Centered positions were then divided by the length of the gene in base pairs from the TSS to the TES. We then binned the relative 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 first analysis identifies states that are associated with high expression and low expression within the hepatocytes, and the second analysis investigates whether variation in chromatin state across strains contributes 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.

Selecting the most biologically meaningful model

We performed the above analyses on all states from the four-state model to the 16-state model to find the most meaningful clustering of histone modifications. Across all models, the states were remarkably stable (Supplemental Figure XXX). As we increased the number of states detected by the model, new states appeared, but previously detected states were not disrupted. This stability was apparent in all state measures: emissions probability patterns, overall abundance, effect on expresssion, and localization along the genome. The one exception to this stability was that highly abundant state (present in 65% of transcribed genes) detected first in the four-state model was split into two distinct states in the 10-state model. These states were also highly abundant (appearing in 40% and 41% of transcribed genes), and had distinct genomic distributions and emissions probabilities (Supplemental Figure XXX). These two states remained stable with increasing numbers of clusters through to the 16-state model. States arising after the 10-state model were of lower abundance, appearing in 2% or less of transcribed genes.

All of the higher abundance states were established in the 10-state model. However, as we moved toward higher numbers of clusters, the resolution on the lower-abundance states improved in terms of the emission probability profiles, and strength of the correlation with gene expression. For example, the 14-state model better resolved a state that had appeared in the 10-state model but was not strongly correlated with gene expression. In the 14-state model, the emission patterns were closer to binary, and the strength of the correlation with expression was increased. Beyond 14 clusters, the new states identified were extremely rare (1% of transcripts or less), and were not strongly correlated with gene expression. We thus selected the 14-state model and the model with the most biologically meaningful clusters.

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 within strains (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. However, unlike with the chromatin states, we only calculated the effects of DNA methylation on gene expression in a position-dependent manner.

Imputation of genomic features in Diversity Outbred mice

To assess the extent to which chromatin state and DNA methylation are 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 Svenson et al. 2012. 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.

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 (Fig XXX?).

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 (CITE). 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 [1921910, 21921916], 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 XXX). The effect of genotype on this variation can be measured in a mapping population as expression quantitative trait loci (eQTL), which associate genetic variation with variation in gene expression. In this study we investigated the extent to which variation in epigenetic modifications, such as histone modifications and DNA methylation, are associated with genetically controlled variation in gene expression.

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. Ppanel A in Figure XXX 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 XXXB). 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 2 and 1) were primarily found in intergenic regions, three states (states 9, 13, and 11) were enriched around known enhancers, and one (state 6) was enriched predominantly near the TES. The majority of these states were also associated with variation in gene expression (Figure XXXC). The colored bars in this panel show the effect of each state on gene expression variation across the inbred strains. For reference, the paired tan bars show the effect of each chromatin state on gene expression in hepatocytes. These effects are the same sign as the across-strain effects, for the most part, and tend to be stronger.

By merging histone modification patterns with enrichments near functional elements and the effects of each state on gene expression, we were able to suggest annotations for many of the 14 chromatin states (Figure XXXD). A more detailed description of these annotations is in Table XXX.

The states in Figure XXX are shown in order of their effect on expression, which helps illustrate several patterns in the data. The state with the most negative effect on gene expression, state 1, is the absence of all measured modifications. The next few states all contain the repressive mark H3K27me2, and are all associated with reduced gene expression. The states with the most positive effects on expression all have some combination of the activating marks, H3K4me3, H3K4me1, and H3K27ac, and the repressive mark is less commonly seen in these states with positive effects. Maybe unsurprisingly, we were also better able to annotate the states with strong negative effects and strong positive effects, but the states in the middle with weak effects were more difficult to annotate.

That states 1 and 2 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.

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 XXXA-B). We also characterized the distribution of CpG sites and their percent methylation at this gene-level scale (Figure XXXC-D).

The spatial patterns of the individual chromatin states are shown in (Figure XXXA), and an overlay of all states together emphasizes the difference in abundance between the most abundant states (states 14, 12, and 1), and the remaining states, which were relatively rare (Figure XXXB).

Each chromatin state had a characteristic distribution pattern relative to gene bodies. For example, state 1, 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 to a peak at the TES. In contrast, states 12 and 14 were both concentrated at the TSS. State 12 was very narrowly concentrated right at the TSS, whereas state 14 was more broadly abundant both upstream and downstream of the TSS. Both were associated with increased expression in the inbred mice, suggesting promoter or enhancer functions. The state third state in this group of high-expressing states, state 13, was depleted nere the TSS, but enriched within the gene body, suggesting that this state may mark active intragenic enhancers.

The states with weaker effects on expression, the states with more gray shades, were also of lower abundance, but still had distinct patterns of abundance around the gene body suggesting the possibility of distinct functional roles in the regulation of gene expression.

DNA methylation also showed strong positional effects (Figure XXXC and D). Across all genes, the TSS had densely packed CpG sites relative to the gene body (Figure XXXC). As expected, the median CpG site near the TSS was 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).

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 XXXA) and the effect of each chromatin state on the variation in gene expression across strains (Figure XXXB).

All chromatin states demonstrated spatially dependent effects on gene expression within hepatocytes. For about half 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 body or predominantly in the intragenic region. The direction of the effects matched the overall effects of each state seen previously. Remarkably, the spatial effects were recapitulated for almost every state when we looked 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 9, 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 as it contributred to expression variation across the inbred strains (Figure XXX). As expected, hypomethylation 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 DNA methylation is used in gene regulation within a cell type, it it not heritable and does not contribute to variation in gene expression across genetically diverse individuals.

Imputed chromatin state explained varation in expression 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 will also allow 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 described previously [Svenson, Tyler] (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 (Figure XXX).

After imputing each genomic feature into the DO population, we mapped gene expression to the imputed features and calculated the variance explained. An example of the data used for the imputation and the results of the mapping are shown for the gene *Pkd2* in Figure XXX. Penels B, C, and D show the chromatin state, SNP genotype, and DNA methylation status at multiple positions along the gene body. The boundaries of the gene body along with the direction of transcription are shown by the arrow under panel A. Panel A shows the variance explained by each genomic feature at each position with a distinct symbol. The variance explained by the local haplotype is shown as the dashed line at the top of thie panel. 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. Both regions have high variation in chromatin state that are highly correlated with gene expression in the DO (blue plus signs in top panel). Just upstream of the TSS there are SNPs (red x's in the top panel) that also explain large amounts of variation in expression, although not as much as chromatin state. A similar pattern is seen at the internal region just downstream of the TSS. Percent DNA methylation does not vary across the strains in either of these critical regions, and does not contribute to variation in expression across genetically distinct individuals.

The overall distributions of variance explained by each feature across all transcripts is shown in Figure XXX. 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 least 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 study we surveyed strain variation in two types of epigenetic modifications across nine inbred strains of mouse: histone modifications, and DNA methylation. We identified 14 chromatin states representing different combinations of four histone modifications and showed that the presence of these states was associated with variation in gene expression across the inbred strains. This relationship between chromatin state and gene expression varied along the gene body and was concentrated both at the TSS and within the gene body. In addition to the correlation betweeen chromatin state and gene expression across inbred strains, we found that chromatin state, imputed from local founder haplotype, explained a large amount of the variation in gene expression in an independent outbred mouse population, almost completely accounting for the variance explained by local eQTLs.

In contrast, although percent DNA methylation was associated with downregulation of genes within hepatocytes, it was not associated with gene expression variation across inbred strains or in the outbred population. Panel D in Figure XXX gives a visual example of why this is the case. Despite strain variation in both genotype and chromatin state at the TSS of *Pkd2*, DNA methylation does not vary at all. The CpG island at the TSS is unmethylated in all strains, and thus all strains express *Pkd2* in hepatocytes. Rather, it is the variation in histone modifications that contributes to the variation in expression levels across the strains.

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 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 spatial resolution of chromatin state gives a finer grained picture of local gene regulation than haplotype without the loss of explanatory power seen with SNPs. In the example gene *Pkd2*, chromatin state shows two clear regulatory regions, one surrounding the TSS, and the other just downstream inside the gene body. These regulatory regions are not apparent in the SNP patterns or in the patterns or DNA methylation. The chromatin state map not only helps identify putative regulatory regions, but may also help annotate the SNPs underyling these regions. SNPs underlying the two regulatory regions may alter gene expression by affecting histone modificactions in regulatory regions, whereas the SNPs that are further downstream of this region may alter gene expression through other mechanisms, such as through altering transcription itself, or the processing of the transcript. Thus the intermediate resolution of the chromatin state between that of SNPs and haplotype provides a highly informative layer of information between genotype and gene expression.

There were a few individual states that were of particular interest.

more detailed discussion of a few individual states: combine spatial distribution, spatial effects, literature, and enrichments, we can find out interesting things about the states and generate hypotheses about the biology of hepatocytes.

individual states to discuss: states with combo activating and repressing:

bivalent promoter: typically resolved during development, but may stay for regulation of some processes? tissue regeneration? abundance is right at TSS, consistent with promoter annotation. Pretty rare state, but marks it genes that are enriched for developmental processes, so probably real. (?) enrichments:ube morphogenesis, tube development, vasculature development, regulation of locomotion, blood vessel development, cell adhesion, biological adhesion, regulation of cell migration, blood vessel morphogenesis, anatomical structure formation, regulation of cell differentiation

poised enhancer: frequency similar to bivalent promoter. Abundance peaks upstream of the TSS, so consistent with enhancer annotation. enrichments: vasculature development, circulatory system development, cell adhesion, biological adhesion, blood vessel development, regulation of cell migration, regulation of cell motility, cell-cell adhesion, protein localization to cell periphery, blood vessel morphogenesis, locomotion, cell migration, smooth muscle cell differentiation, regulation of immune system process

both of these states have the same negative regulation of expression at the TSS in hepatocytes and across inbred strains, but the effect does not translate to DO mice. What do we make of this?

activating states: active promoter and enhancer have different distributions, the enhancer is broader around the TSS, the promoter state is right at the TSS. both have positive effects on transcription that match their abundance profile. Effects go all the way through to DO mice

active promoter enrichments: response to cytokine, brush border, lipid metabolic process, cellular lipid metabolic process, actin cytoskeleton, cluster of actin-based cell projections, regulation of intracellular signal transduction, lipid binding, fatty acid metabolic process, drug transport, brush border membrane, MAPK cascade, positive regulation of MAPK cascade, monocarboxylic acid metabolic process, regulation of cell migration, Cortisol synthesis and secretion, regulation of MAPK cascade, chemical homeostasis

active enhancer enrichments: chromatin binding, tube morphogenesis, tube development, regulation of establishment of protein localization, regulation of protein localization, beta-catenin binding, nucleotide transport, mitochondrial membrane, neuron projection, purine nucleotide transport, purine ribonucleotide transport, regulation of cellular component movement, regulation of cell migration, regulation of anatomical structure morphogenesis

intragenic enhancer: abundance and positive effect peak inside gene body effect carries all the way through to DO mice enrichments: cellular macromolecule localization, cellular protein localization, intracellular transport, establishment of protein localization, protein transport, RNA processing, cellular response to stress.

repressor state: marks developmental genes

Genes that are marked with these different states are regulated differently across the inbred strains of mice.

intragenic enhancer that has effect in hepatocytes, but not across strains: genes in hepatocytes with this state in the gene body are more highly regulated than genes without this state. The effect is quite strong in hepatocytes, but does not extend to variation across strains, even in the inbred mice. This state marks genes that are important in hepatocytes, but are not differentially regulated across strains. enrichments: small molecule metabolic process, oxoacid metabolic process, organic acid metabolic process, carboxylic acid metabolic process, lipid metabolic process, organic substance catabolic process, monocarboxylic acid metabolic process

By combining the spatial distribution and effects of the chromatin states with known information about the effects of histone modifications, we can

variation in local genotype -> variation in epigenotype -> variation in expression local eigenetics explains more variation in expression than individual SNPs - implies that individual SNPs may not be the best unit of analysis (too polemical?)

for the majority of transcripts we still only explain about 10% of the variance with haplotype and local epigenetics other variation explained by more complex genetic architecture? stochasticity? other genomic features?

Annotation of chromatin states

Make this section into a table

We identified 14 chromatin states corresponding to 14 distinct combinations of histone modifications (Figure XXXA). To annotate these states to functional elements, we combined previously known annotations with functional enrichments and relationship to gene expression. The characterizations are summarized in Figure XXX. Figure XXXA further shows the relative abundance of each state in and around the gene body. This high-resolution image of abundance helped further refine the annotations of each state. Figure XXXB shows that overall states 1 and 7 were the most abundant states with state 7 being highly enriched at the TSS, and state 1 being strongly depleted at the TSS, but enriched within the gene body and in intragenic spaces. We describe the reasoning behind the annotation of each state below:

State 1 - heterochromatin was characterized by the absence of all measured marks, enrichment in intergenic regions, and strong downregulation of gene expression. This state was strongly depleted at the TSS of expressed genes (Figure XXXA), but the most abundant state in the gene body and outside the gene body. This state may multiple different states that could be resolved with the measurement of more histone modifications. For example, intergenically, state 1 may mark heterochromatin, which is characterized by H3K9 trimethylation [12867029], which was not measured here. However, state 1 was also highly abundant in the gene bodies of expressed genes, but was associated with reduced expression. This could suggest differential distribution of heterochromatin across strains, or could represent an additional transcriptionally repressive state.

State 2 - repressed chromatin was characterized by the presence of H3K27me3, which has been shown previously to correlate with transcriptional silencing [REF]. This state was not enriched in any particular functional element, but was associated with strong downregulation of transcription.

State 3 - poised enhancer was primarily characterized by the presence of H3K27me3, a mark associated with polycomb silencing [REF], and H3K4me1 a mark associated with enhancers [REF]. The co-occurence of these opposing marks has previously been associated with a functional element known as a poised enhancer [21160473].

This element has been studied mostly in the context of development. Bivalent promoters are abundant in undifferentiated cells, and are resolved either to active promoters or silenced promoters as the cells differentiate into their final state [REF]. These promoters have also been shown to be important in the response of cancer cells to environmental disturbances such as hypoxia [REF]. The presence of bivalent promoters in adult mouse hepatocytes is interesting. They may mark genes poised for expression during liver regeneration, or for responding to a particular environmental stimulus. There were XXX genes that were marked with this bivalent promoter state at the TSS across all strains. This group of genes was enriched for developmental processes as well as alcohol metabolism (Fig? Table?).

State 4 - intragenic enhancer was characterized by the presence of H3K4me1, which is known to mark cell type-specific enhancers, both active and poised [REF]. The presence of H3K4me1 alone, in the absence of H3K27ac, as it occurs in state 4, has been shown to mark inactive, or poised enhancers [21106759]. The addition of H3K27ac can then activate the enhancer to increase transcription. When present within the gene body, this state acts as an intragenic enhancer, which acts as an alternative promoter, and can be transcribed bidirectionally to produce short RNAs known as eRNA [20393465]. This state was modestly enriched in known enhancers and was associated with slightly increased gene expression. The presence of H3K4me1 in the absence of H3K4me3 has been shown to mark intragenic enhancers and to be associated with increased transcription, as these regions can be transcribed independently of the full gene [Kowalczyk et al. 2012]. We annotated this state as a weak enhancer.

State 5 - active enhancer was characterized by the co-occurence of H3K4me1, which marks cell type-specific enhancers, and H3K27ac, which specifically marks active enhancers [21106759, 21160473]. This state was strongly enriched in known enhancers, and its presence had a strong postive effect on transcription. We thus annotated this state as a strong enhancer.

state 0100 (just H3K4me3 absence of others) - interesting abundance pattern. Just up and downstream of the TSS, and just downstream of the TES. Associated with slight downregulation of expression. Could be antisense transcription [22768981]? This fits the negative association, but the position of the negative association is just downstream of the TSS. can't really find any other sources on this. The probably with most studies is that they are looking only at a single mark, and not noting the absence of the other three marks.

state 0110 -

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

Supplemental Table Descriptions

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

References