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 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.

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

Mice

Inbred Founder Mice

Diversity Outbred mice

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

Genotyping

Diversity outbred mice

All DO mice were genotyped as described in Svenson et al.˜(2012) using the Mouse Universal Genotyping Array (MUGA) (7854 markers), and the MegaMUGA (77,642

markers) (GeneSeek, Lincoln, NE). All animal procedures were approved by the Animal Care and Use Committee at The Jackson Laboratory (Animal Use Summary # 06006).

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.

Measurement of gene expression

Inbred Founders

Diversity outbred mice

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

Measurement of Chromatin Modifications

Measurement of DNA Methylation

Percent DNA methylation was measured using reduced representation bisulfite sequencing.

Data Processing

Filtering transcripts

We remove transcripts with extremely low read counts, by filtering out transcripts whose mean read count was less than five across all individuals.

We then used the R package sva {[}sva{]} 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.

Chromatin modifications

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

DNA methylation

Annat's stuff to get bed files.

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 this 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 (Methods). Figure XXX gives an overview of these states.

The states were distributed non-randomly around known functional elements in the mouse genome (Figure XXX Genomic Enrichment). 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 3) were primarily found in intergenic regions, three states were enriched around known enhancers, and one was enriched predominantly near the TES. The majority of these states were also associated with variation in gene expression (Figure XXX Expression Effects). The within-strain effects on expression (gray) were stronger in magnitude than the across-strain effects (colored by states), but the direction of the effects matched with only one exception.

The two states that were primarily located in intergenic regions were associated with strong downregulation of genes when associated with genes, suggesting that there is differential epigenetic silencing of genes in hepatocytes across strains. That the majority of chromatin states were associated with variation in expression across strains suggests 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.

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

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).

Each chromatin state had a unique pattern of spatial abundance (Figure XXXA), and an overlay of all states together emphasizes the difference in abundance between the most abundant states , and the remaining states, which were relatively rare.

Each chromatin state had a characteristic distribution pattern relative to gene bodies (Figure XXXA). For example, state 3, 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 chromatin modification. However, its abundance increased steadily to a peak at the TES. In contrast, states 11 and 10 were both concentrated at the TSS. Whereas state 11 was very narrowly concentrated right at the TSS, whereas state 10 was more broadly abundant both upstream and downstream of the TSS. These differential distribution patterns may speak to distinct functional roles in regulation of expression.

The remaining states were relatively low in abundance compared to states 3, 10, and 11 (Figure XXXB), but also showed specific distributions relative to the gene body. For example, state 12, which overall was enriched around annotated TESs, was highly concentrated just downstream of the TES, depleted at the TSS, and concentrated immediately up and downstream of the TSS. In contrast, states 8 and 7 were most abundant inside the gene body, and were relatively depleted at the TSS.

DNA methylation also showed strong positional effects (Figure XXXC and D). Across all genes, the TSS had densely packed CpG sites relative to the region between the TSS and TES (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 a strain (Figure XXX first column) and the effect of each chromatin state on the variation in gene expression across strains (Figure XXX second column).

Imputed chromatin state was correlated with gene expression in DO mice

To further investigate the relationship between genotype, epigenetic features, and gene expression, we imputed chromatin state, DNA methylation, and SNPs into a population of DO mice described previously [Svenson, Tyler] (Methods). Gene expression was measured in whole livers in these mice giving us the opportunity to explore the extent to which local chromatin state corresponded with varation in gene expression across individuals.

In addition to chromatin state, we also imputed DNA methylation state and SNPs as comparators.

SNP imputation is almost certainly ground truth in the DO mice is chromatin state is determined by (local?) genotype? how well does this predict gene expression in animals with mixed up genomes of multiple founders DNA methylation probably not determined by local genetics gives lower bound on expectations for imputations if feature is not related to gene expression

We compared the percent variance explained by local haplotype to the maximum percent variance explained by local chromatin state, local percent DNA methylation, and local SNPs for each transcript (Figure XXXA).

Overall, local haplotype explained the largest amount of variance in gene expression ($R^2 = 0.31$). The variance explained by local chromatin state was very highly correlated with that of haplotype (Pearson $r = 0.95$) and also explained a relatively high proportion of variation in gene expression ($R^2 = 0.28$). Individual SNPs were less correlated with haplotype (Pearson $r = 0.69$), and explained less overall variance in gene expression ($R^2 = 0.12$). DNA methylation, which previous results suggest is not genetically determined, had the lowest correlation with haplotype (Pearson $r = 0.55$) and explained the least variance in gene expression ($R^2 = 0.07$).

An example of how different functional genomic features are associated with gene expression is shown in Figure XXX.

finding supports idea that chromatin state is defined by local genetics imputed chromatin almost as good as measured haplotype in explaining variation in gene expression, but with higher resolution

that this worked in a population of genetically unique mice with completely mixed up genomes speaks to just how much gene expression is determined by local genetics.

Powerful observation that we can impute chromatin state in 500 DO mice from measuring chromatin state in a handful of inbred mice

Gives us higher resolution than haplotype, without the loss of explanatory power we get from SNP analysis

haplotype includes all genetically determined functional elements in a relatively large region

SNPs do tag the haplotype, but are

Further, because chromatin modifications are measured at extremely high density, we can map high-density chromatin effects in the DO mice, which may help prioritize functional SNPs within gene bodies and in regulatory regions.

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

Variation in expression in an outbred population maps to imputed chromatin state

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

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

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

DNA methylation varied across the gene body

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

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

Chromatin state but not DNA methylation correlated with gene expression across strains

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

Annotation of chromatin states

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.

Discussion

imputation gives us a way to do a very limited, gene-based GWAS? not good wording, but we can potentially increase the resolution right around gene bodies

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

Haplotype and chromatin state represent broader regions of genome than SNPs and DNA methylation, which are measured at the base pair level. The measurements that

represent larger regions of the genome are more predictive of local gene expression than the point-wise measurements.

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

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

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

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

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

RRBS discussion - In humans estimates of heritability of DNA methylation are relatively low (0.1 to 0.3). It is estimated that around 10% of methylation sites are highly heritable. heritability estimates are age- and population-specific.

Even if there are inherited patterns of DNA methylation, do they have any effect on gene expression? Keep in mind that we are only looking at local effects.

human studies have shown that trimodal sites (0, 0.5, 1) have relatively high heritability (0.8), and almost half were associated with eQTLs.

no evidence for trans-generational inheritance of DNA methylation in humans.

Positional information is interesting

We observed interesting spatial patterns of chromatin state distribution and

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

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

Acknowledgements

This work was funded by XXX.

Data and Software Availability

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

Supplemental Figure Legends

Fig 1

Supplemental Table Descriptions

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

References