Integrative QTL analysis of gene expression and chromatin accessibility identifies multi-tissue patterns of genetic regulation

Gregory R. Keele[*,†,‡,***], **Bryan C. Quach**[*,†,‡,†††], **Jennifer W. Israel**[‡], **Yihui Zhou**[‡‡], **Grace A. Chappell**[§§], **Lauren Lewis**[§§], **Alexias Safi**[††], **Jeremy M. Simon**[‡], **Paul Cotney**[‡], **Gregory E. Crawford**[††], **Fred A. Wright**[‡‡], **William Valdar**[‡,**], **Ivan Rusyn**[§§,1] **and Terrence S. Furey**[‡,§,**,1]

[*]Authors contributed equally, [†]Curriculum in Bioinformatics and Computational Biology, [‡]Department of Genetics, [§]Department of Biology, [**]Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, [††]Department of Pediatrics, Center for Genomic and Computational Biology, Duke University, Durham, NC, [‡‡]Departments of Statistics and Biological Sciences, North Carolina State University, Raleigh, NC, [§§]Department of Veterinary Integrative Biosciences, Texas A&M University, College Station, TX, [***]The Jackson Laboratory, Bar Harbor, ME, [†††]Center for Omics Discovery and Epidemiology, Research Triangle Institute (RTI) International

ORCID IDs: 0000-0002-1843-7900 (G.R.K.), 0000-0002-1094-3104 (B.C.Q.), 0000-0003-3906-1663 (J.M.S.), 0000-0002-2419-0430 (W.V.)

ABSTRACT

Many factors such as genetic variation and chromatin accessibility modulate the transcriptional profiles of genes across tissues. We analyzed gene transcript and chromatin accessibility levels from liver, lung, and kidney tissues for 47 strains of the Collaborative Cross (CC) mouse population to characterize expression and chromatin state dynamics across tissues, with particular attention paid to their genetic regulation through the identification of expression quantitative trait loci (eQTL) and chromatin QTL (cQTL). We detected 1,101 genes and 133 chromatin regions that possessed QTL with correlated allele effects across multiple tissues. Tissue-specific QTL were also detected genes and chromatin, including *Pik3c2g* with distinct local-eQTL observed in all three tissues. Leveraging the overlapping measurement of gene expression and chromatin accessibility on the same mice from multiple tissues, we used mediation analysis to identify chromatin and gene expression intermediates of eQTL effects. Based on QTL and mediation analyses over multiple tissues, we propose a causal model for the distal genetic regulation of *Akr1e1*, a gene involved in glycogen metabolism, through a zinc finger protein and chromatin intermediates. This analysis demonstrates the complexity of transcriptional and chromatin dynamics and their regulation over multiple tissues, as well as the value of the CC and related genetic resource populations for identifying specific regulatory mechanisms in biological data.

KEYWORDS multiparental populations (MPP), Collaborative Cross (CC), Diversity Outbred (DO), mediation, genome-wide association (GWA)

Determining the mechanisms by which genetic variants drive molecular, cellular, and physiological phenotypes has proved to be challenging (Schadt 2009). Genome-wide experiments that provide data on variations in molecular and cellular states in genotyped individuals can contribute to elucidating these mechanisms, but most analyses of these data are largely observational, due in part to constraints of specific populations (*i.e.* humans), the ongoing development of experimental technologies, and the challenge of coordinating large numbers of experiments with multiple levels of data (Schaid *et al.* 2018). More recently, complementary genome-wide datasets from the same individuals have been paired with modern statistical mediation analysis to draw inferences on the relationships that connect these data. These analyses can identify putative causal relationships rather than correlational interactions, providing meaningful and actionable targets in terms of downstream applications in areas such as medicine and agriculture.

Manuscript compiled: Wednesday 27th March, 2019

[1]Corresponding authors: 4458 TAMU, Texas A & M University, College Station, TX 77843.
E-mail: irusyn@tamu.edu
120 Mason Farm Rd., Genetic Medicine Bldg., Suite 5113, University of North Carolina, Chapel Hill, NC 27599.
E-mail: tsfurey@email.unc.edu

In human data, co-occurence of QTL across various multi-omic data has been used to assess the potential for related and connected biological processes [examples include gene expression with chromatin accessibility (Degner *et al.* 2012), regulatory elements (Pai *et al.* 2015), and ribosome occupancy and protein (Battle *et al.* 2015)]. More formal integration through statistical mediation analyses have also been used to investigate relationships between levels of human biological data [*e.g.*distal-eQTL with proximal gene expression (Battle *et al.* 2014), and eQTL with regulatory elements (Alasoo *et al.* 2018; Roytman *et al.* 2018; Wu *et al.* 2018)].

Despite these successes, using data from human populations for genotype-phenotype analyses is challenging, in part due to an inability to tightly control experimental conditions and genetic backgrounds to reproducibly and unambiguously detect casual linkages. To aid in this effort, genetically-diverse mouse population resources have been established, including the Collaborative Cross (CC) (Churchill *et al.* 2004; Collaborative Cross Consortium 2012; Srivastava *et al.* 2017) and the Diversity Outbred (DO) mice (Churchill *et al.* 2012). Derived from the same eight founder strains, the CC are recombinant inbred strains while the DO are largely heterogeneous outbred animals, bred with a random mating strategy that seeks to maximize diversity. Chick *et al.* (2016) used a genome-wide mediation approach to link transcriptional and post-translational regulation of protein levels in 192 DO mice, and also used the CC to confirm results by showing that estimates of founder allele effects from each of the related populations corresponded. Recently, Skelly *et al.* (2019) used mediation analysis to connect chromatin accessibility with gene expression in embryonic stem cells derived from DO mice. In comparison to the DO, the CC is more restricted in the number of unique genomes and reduced mapping resolution, but it uniquely provides the ability for replicate observations under varying experimental conditions while maintaining a constant genetic background.

Here we use a sample composed of a single male mouse from 47 strains of the CC to investigate tissue-specific dynamics between gene expression and chromatin accessibility, as determined by Assay for Transposase Accessible Chromatin sequencing (ATAC-seq), in lung, liver, and kidney tissues. We detect QTL underlying gene expression and chromatin accessibility variation across the strains, and assess the support for mediation of the effect of eQTL through chromatin accessibility, using a novel implementation of methods adapted from Chick *et al.* (2016). Additionally, we detect gene mediators of distal-eQTL. These findings demonstrate the experimental power of the CC for integrative analysis of multi-omic data to determine genetically-driven phenotype variation, despite limited sample size, and provides support for continued use of the CC in larger experiments going forward.

Materials and Methods

Animals

Adult male mice (8-12 weeks old) from 47 CC strains (University of North Carolina Systems Genetics Core) were maintained on an NTP 2000 wafer diet (Zeigler Brothers, Inc., Gardners, PA) and water *ad libitum*. The housing room was maintained on a 12-h light-dark cycle. Our experimental design sought to maximize the number of strains relative to within-strain replications based on the power analysis for QTL mapping in mouse populations (Kaeppler 1997); therefore, one mouse was used per strain. Mice were euthanized and lungs, liver and kidney tissues were collected, flash frozen in liquid nitrogen, and stored at -80°C. These studies were approved by the Institutional Animal Care and Use Committees at Texas A&M University and the University of North Carolina. **Figure 1** is a diagram of the experiment and subsequent analyses performed for this study.

mRNA sequencing and processing

Total RNA was isolated from flash-frozen tissue samples using a Qiagen miRNeasy Kit (Valencia, CA) according to the manufacturer's protocol. RNA purity and integrity were evaluated using a Thermo Scientific Nanodrop 2000 (Waltham, MA) and an Agilent 2100 Bioanalyzer (Santa Clara, CA), respectively. A minimum RNA integrity value of 7.0 was required for RNA samples to be used for library preparation and sequencing. Libraries for samples with a sufficient RNA integrity value were prepared using the Illumina TruSeq Total RNA Sample Prep Kit (Illumina, Inc., San Diego, USA) with ribosomal depletion. Single-end (50 bp) sequencing was performed (Illumina HiSeq 2500).

Sequencing reads were filtered (sequence quality score $\geq$ 20 for $\geq$ 90% of bases) and adapter contamination was removed (TagDust). Reads were mapped to strain-specific pseudo-genomes (Build37, http://csbio.unc.edu/CCstatus/index.py?run=Pseudo) and psuedo-transcriptomes (C57BL/6J RefSeq annotations mapped to pseudo-genomes) using RSEM with STAR (v2.5.3a). Uniquely aligned reads were used to quantify expression as transcripts per million (TPM) values.

ATAC-seq processing

Flash frozen tissue samples were pulverized in liquid nitrogen using the BioPulverizer (Biospec) to break open cells and allow even exposure of intact chromatin to Tn5 transposase (Buenrostro *et al.* 2015). Pulverized material was thawed in glycerol containing nuclear isolation buffer to stabilize nuclear structure and then filtered through Miracloth (Calbiochem) to remove large tissue debris. Nuclei were washed and directly used for treatment with Tn5 transposase. Paired-end (50 bp) sequencing was performed (Illumina HiSeq 2500).

Reads were similarly filtered as with RNA-seq. Reads were aligned to the appropriate pseudo-genome using GSNAP (parameter set: -k 15, -m 1, -i 5, –sampling=1, –trim-mismatch-score=0, –genome-unk-mismatch=1, –query-unk-mismatch=1). Uniquely mapped reads were converted to mm9 mouse reference genome coordinates using the associated MOD files (UNC) to allow comparison across strains. Reads overlapping regions in the mm9 blacklist (UCSC Genome Browser) were removed. Exact sites of Tn5 transposase insertion were determined as the start position +5 bp for positive strand reads, and the end position -5 bp for negative strand reads (Buenrostro *et al.* 2013). Peaks were called using F-seq with default parameters. A union set of the top 50,000 peaks (ranked by F-seq score) from each sample was derived. Peaks were divided into overlapping 300 bp windows as previously described (Shibata *et al.* 2012). Per sample read coverage of each window was calculated using coverageBed from BedTools (Quinlan and Hall 2010).

Sequence trait filtering for QTL analysis

Trimmed mean of M-values (TMM) normalization [edgeR, (Robinson *et al.* 2010)] was applied to TPM values from read counts of genes and chromatin windows, respectively. Genes with TMM-normalized TPM values $\leq$ 1 and chromatin windows with normalized counts $\leq$ 5 for $\geq$ 50% of samples were excluded. For each gene and chromatin window, we applied

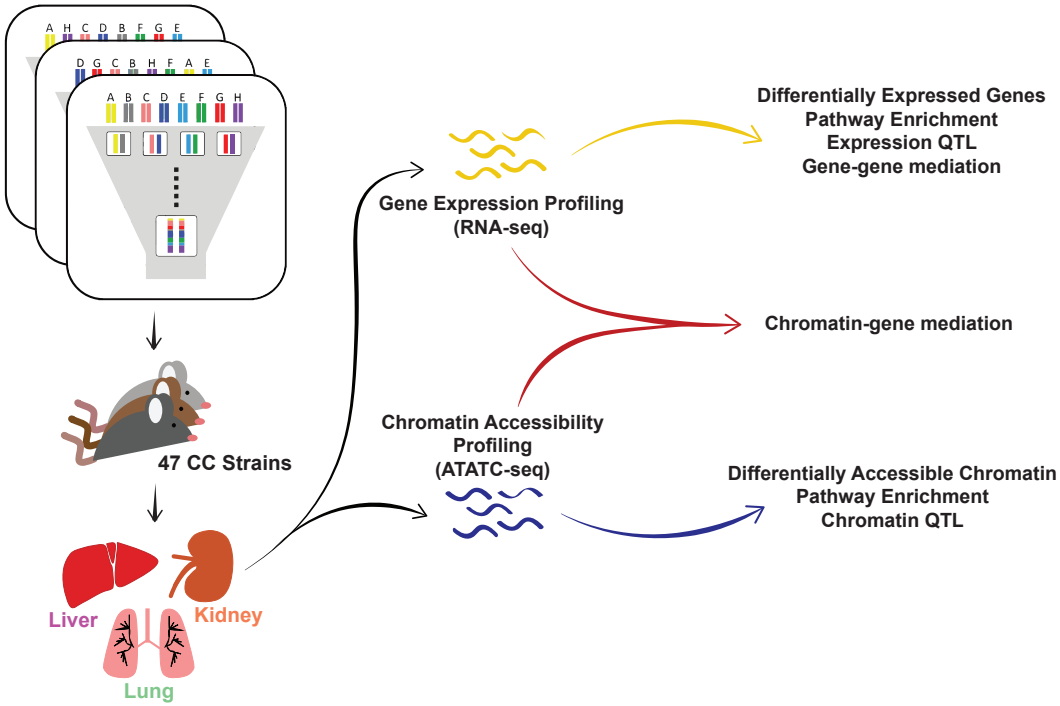

Figure 1 Diagram of the experiment and analyses. RNA-seq and ATAC-seq were collected from liver, lung, and kidney tissues of males from 47 CC strains. Each CC strain was derived from an inbreeding funnel, and thus represents a recombinant inbred mosaic of the initial eight founder haplotypes. Differential analyses with pathway enrichment analyses were performed to identify biological pathways enriched in differentially expressed genes and accessible chromatin regions. QTL and mediation analyses were performed to identify regions that causally regulate gene expression and chromatin accessibility.

K-means clustering with $K = 2$ to identify outcomes containing outlier observations that could cause spurious, outlier-driven QTL calls. Any gene or chromatin window where the smaller K-means cluster had a cardinality of 1 was removed.

Founder mosaic reduction

CC genomes are mosaics of the founder strain haplotypes and were previously reconstructed by the UNC Systems Genetics Core (http://csbio.unc.edu/CCstatus/index.py?run=FounderProbs) with the Hidden Markov Model of Fu *et al.* (2012a) on genotype calls [MegaMUGA array (Morgan *et al.* 2016)] from multiple animals per strain. To reduce the number of statistical tests, adjacent genomic regions were merged through averaging if the founder mosaics for all mice were similar, defined as L2 distance $\leq 10\%$ of the maximum L2 distance ($\sqrt{2}$ for a probability vector).

Differential expression and accessibility analyses

Read counts for each sample were converted to counts per million (CPM) and normalized using TMM normalization (*edgeR*). Windows with > 70 % of samples with CPM ≤ 1 were removed. Differentially expressed genes and accessible chromatin windows were determined using *limma* (Ritchie *et al.* 2015) with the linear model:

$$y_i = \mu + \text{strain}_i + \text{batch}_i + \text{tissue}_i + \varepsilon_i, \tag{1}$$

where y_i is the TMM-normalized CPM value in tissue_i (lung, liver, or kidney) from individual i; μ is the intercept, representing the overall mean; batch_i is effect of sequencing center; strain_i is the strain-specific effect; and ε_i is an error term, distributed as $\text{N}(0, \sigma^2)$.

To account for mean-variance relationships in gene expression and chromatin accessibility data, precision weights were calculated using the *limma* function *voom* and incorporated into the linear modeling procedure. The p-values were adjusted using an false discovery rate (FDR) procedure (Benjamini *et al.* 1995), and differentially expressed genes and accessible chromatin windows were called based on the q-value ≤ 0.01 and $\log_2$ fold-change ≥ 1. Adjacent significantly differential chromatin windows in the same direction were merged with a p-value computed using Simes' method (Sarkar and Chang 1997), and chromatin regions were re-evaluated for significance using the Simes p-values.

Gene set association analysis

GSAASeqSP (Xiong *et al.* 2014) with Reactome Pathway Database annotations (July 24, 2015 release) was used to identify biological pathways enriched with differentially expressed genes or accessible chromatin. A list of assayed genes were provided as input to GSAASeqSP along with a weight for each gene g, calculated as:

$$\text{weight}_g = \text{sign}(\text{fc}_g) * (1 - q_g), \tag{2}$$

where $\text{sign}(\text{fc}_g)$ is the sign of the expression fold change, and q_g is the FDR-adjusted p-value. Pathways with gene sets of cardinality < 15 or > 500 were excluded. For chromatin accessibility, each region was mapped to a gene using GREAT v3.0.0 (*basal plus extension* mode, *5 kb upstream*, *1 kb downstream*, and no distal extension). Each gene was associated with the chromatin region with the most significant FDR-adjusted q-value and gene weights were calculated as above with $\text{sign}(\text{fc}_g)$, the sign of

chromatin accessibility fold-change and q_g the FDR-adjusted q-value of the chromatin region representing gene g.

QTL analysis

A single locus approach to QTL mapping was used for both gene expression and chromatin accessibility. The CC mice have well-characterized founder haplotypes, allowing the association analysis to be haplotype-based, also referred to as interval mapping or linkage analysis (Lander and Botstein 1989). Assessing association between haplotype descent and phenotype has advantages to variant association, such as implicitly modeling multiple local variants simultaneously and more closely reflecting the linkage disequilibrium (LD) decay. Because haplotype state is not directly observed but rather probabilistically inferred (*e.g.*Lander and Green 1987; Mott *et al.* 2000; Liu *et al.* 2010; Fu *et al.* 2012a; Gatti *et al.* 2014; Zheng *et al.* 2015), formal interval mapping requires an computationally inefficient Expectation-Maximization (EM) algorithm (Dempster *et al.* 1977). A computationally efficient regression approximation (Haley and Knott 1992; Martínez and Curnow 1992) is commonly used instead for multiparental populations (MPP) (*e.g.*Valdar *et al.* 2006b, 2009; Svenson *et al.* 2012; Baud *et al.* 2013, 2014), including the CC (*e.g.*Aylor *et al.* 2011; Kelada 2016; Mosedale *et al.* 2017; Donoghue *et al.* 2017). This efficiency is particularly important in the context of a study of genome-wide molecular phenotypes.

General QTL model. For a genome scan, the alternative model is fit at positions across the genome and compared to the null model of no locus effect. The general alternative model for gene expression and chromatin accessibility is the same:

$$y_i = \mu + \mathrm{QTL}_i + \mathrm{batch}_i + \varepsilon_i, \tag{3}$$

where y_i is the trait, either levels of gene expression or chromatin accessibility for individual i. μ is an intercept term, representing the sample mean. QTL_i is the genetic locus effect, here fit according to an additive fixed effects model. The effect of sequencing center is modeled as the covariate batch_i. ε_i is the error term for individual i, such that $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$. y are transformed to better satisfy the regression assumption that the residuals are normally distributed; here we take the normal quantiles from the ranks of the trait, in order to be strongly reduce the influence of potential extreme observations in a sample of 47 individuals. The null model is the same for all tested loci and is equivalent to Eq 3 with $\mathrm{QTL}_i = 0$ for all individuals. The two models are compared statistically at each position, and summarized with an ANOVA p-value. QTL are detected if the improvement in fit of the alternative model compared to the null model surpasses a set threshold of significance, which are described in **Appendix A**.

Mapping procedures. Three mapping procedures were used to detect QTL, each with unique features:

1. Multi-stage conditional regression with genome-wide family-wide error rate (FWER) and FDR control (Method 1). This procedure is the most statistically stringent, can identify both local and distal-QTL, and potentially multiple QTL per trait in an unbiased FDR control.

2. Single-stage regression with chromosome-wide FWER and FDR control (Method 2). This procedure is oriented towards identification of local-QTL by using the much more lenient chromosome-wide FWER adjustment. Putative distal-QTL that are located on the same chromosome as the target but are not proximal to the trait's genomic coordinate are also leniently detected, referred to here as intra-chromosomal distal-QTL.

3. Single-stage regression with genome-wide and chromosome-wide FWER control (Method 3). This procedure only detects local-QTL, leveraging strong biological precedent for local genetic regulation to reduce the significance threshold. FWER adjustment, either at genome or chromosome-wide levels, is performed for the maximum logP in the local window (defined here as 10Mb upstream or downstream of target's coordinate), without FDR control.

See **Table 1** for simple descriptions of the three procedures and **Appendix B** for greater detail on each.

QTL founder allele effects and effect size. The founder allele effects at a QTL, as fit in the mapping model (Eq 3), can help distinguish which genetic variants drive an observed QTL. However, these effects, when pulled directly from the mapping procedures, are comprised of a vector with one fixed effect regression coefficient per founder strain and can be unstable when there are few observations per founder strain contribution at a locus, resulting in artificially extreme effects (Zhang *et al.* 2014). To conservatively constrain the effect estimates, the model in Eq 3 was re-fit at the detected QTL, but with the QTL effect as a random effect vector with corresponding variance component (Wei and Xu 2016), such that $\mathrm{QTL}_i = \mathbf{x}_i\boldsymbol{\beta}$ with $\boldsymbol{\beta} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}\tau^2)$. $\mathbf{x}_i$ is the founder haplotype dosage vector for individual i at the QTL, $\boldsymbol{\beta}$ is the QTL effect vector, and τ^2 is the variance component corresponding to the QTL effect vector. The best linear unbiased predictors of the allele effects [$\widehat{\boldsymbol{\beta}}_{\mathrm{BLUP}}$] (Robinson 1991) were calculated, centered and scaled, and used for further comparison of QTL across tissues.

A point estimate of the proportion of the variance explained by the QTL effect, referred to as the QTL effect size ($\widehat{h^2}_{\mathrm{QTL}}$), is calculated as:

$$\widehat{h^2}_{\mathrm{QTL}} = 1 - \frac{\mathrm{RSS}_{\mathrm{QTL}}}{\mathrm{RSS}_0}, \tag{4}$$

where $\mathrm{RSS} = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$ is the residual sum of squares for the alternative (Eq 3) and null models, and $n = 47$. For comparison, we also calculated a conservative, constrained estimate of QTL effect size, estimated from the QTL model with QTL parameters constrained as random effects:

$$\widehat{h^2}_{\mathrm{QTL\ ranef}} = \frac{\widehat{\tau^2}}{\widehat{\tau^2} + \widehat{\sigma^2}}, \tag{5}$$

QTL with effect sizes that are particularly reduced by the random effects model are more likely to be the result of extreme data points or imbalanced founder contributions at the locus.

Cross-tissue comparison of QTL. To evaluate patterns of the genetic regulation of gene expression and chromatin accessibility across tissues, correlations between the founder allele effects of QTL that map to approximately the same region of the genome for the same traits but in different tissues were calculated. Pairs of local-QTL were required to both be detected within the 20Mb window around the gene TSS or the chromatin window midpoint. For distal-QTL, the QTL positions had to be within 10Mb of each other. All detected QTL were considered, including QTL from Methods 1 and 2 controlled at an FDR of 0.2, allowing for

Table 1 QTL mapping procedures

Procedure	QTL type	FWER control	FDR control	QTL per trait
Method 1	local & distal	genome-wide	yes	≥ 1
Method 2	local & intra-chromosomal distal	chromosome-wide	yes	1
Method 3	local	genome & chromosome-wide	no	1

consistent signal across tissues to provide further evidence for putative QTL with marginal significance within a single tissue.

For a pair of matched QTL j and k from different tissues, the Pearson correlation coefficient was calculated as $r_{jk} = \text{cor}(\widehat{\boldsymbol{\beta}}^{j}_{\text{BLUP}}, \widehat{\boldsymbol{\beta}}^{k}_{\text{BLUP}})$. $\widehat{\boldsymbol{\beta}}_{\text{BLUP}}$ represent 8-element vectors, thus corresponding r are distributed such that $\frac{r\sqrt{6}}{1-r^2} \sim t_6$ according to the null model of independent variables. Testing the alternative models that $r_{jk} > 0$ and $r_{jk} < 0$ were performed for each pair of QTL, producing p-values. An FDR procedure (Benjamini-Hochburg) was then used to retrieve two q-values per QTL pair, $q_{ij}^{\{r>0\}}$ and $q_{ij}^{\{r<0\}}$, which were used to classify pairs of QTL as being significantly correlated or anti-correlated, respectively.

Variant association. Variant association was performed within the genomic regions surrounding overlapping QTL in order to detect similarities and differences in patterns of association between tissues. Variant genotypes from Build38 were obtained using the ISVdb (Oreper *et al.* 2017) for the CC strains, which were converted to Build37 coordinates with the liftOver tool (Lawrence *et al.* 2009). Variants were filtered out if their minor allele frequencies ≤ 0.1 or they were not genotyped in one of the CC founder strains to avoid false signals.

Mediation analysis

Detection of mediation is dependent on a number of assumptions about the underlying variables, their relationships, and the directionality of those relationships, many of which cannot be satisfied in a system as complex as chromatin accessibility and transcriptional regulation in whole living organisms. However, signals in the data that are consistent with the mediation model could identify putative causal factors that regulate gene expression. **Figure 2** represents the mediation models tested herein.

Genome-wide mediation. Mediation analysis have recently been used with genomic data, including in humans (*e.g.*Battle *et al.* 2014). We use a similar approach to Chick *et al.* (2016) and Keller *et al.* (2018) to assess statistically significant mediation of eQTL effects on gene expression.

The mediation relationships are evaluated in a genome-wide scan, allowing for the detection of FWER-corrected significant mediators that satisfy relationships 8, 9, and 10. Similar to the QTL mapping genome scan described in Eq 3 and Eq 7, the mediation scan involves a comparison of an alternative and a null model at loci across the genome. The alternative model is

$$y_i^{G_j} = \mu + \text{eQTL}_i^{G_j} + m_{ik} + \text{batch}_i + \varepsilon_i, \tag{6}$$

which is compared to the null model:

$$y_i^{G_j} = \mu + m_{ik} + \text{batch}_i + \varepsilon_i$$

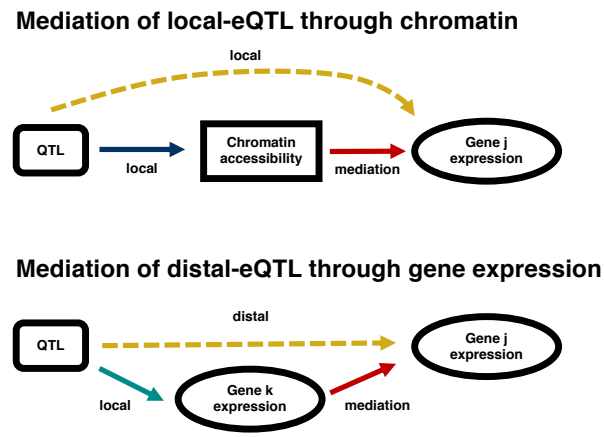

Figure 2 Simple mediation models for the genetic regulation of gene expression. The top model is the mediation of local-eQTL through chromatin accessibility in the region of the gene. This model of mediation is consistent with genetic variation influencing the accessibility of gene j to the transcriptional machinery. The bottom model represents distal-eQTL that are mediated through the transcription of genes proximal to the QTL. Mediation through proximal gene expression is consistent with distal-eQTL that result from as transcription factor activity.

where $y_i^{G_j}$ is the expression levels for individual i of gene j with an eQTL, μ is the intercept term, $\text{eQTL}_i^{G_j}$ is the eQTL effect on gene j, m_{ik} is the effect of the kth mediator, either chromatin accessibility or gene expression depending on the mediation model being evaluated, batch_i is the effect of the sequencing center, and ε_i is random noise. Conditioned loci can be included for certain eQTL, as in Eq 7, but not included here for clarity. Whereas with the QTL genome scan, the locus effect was changed at each position, for the mediation scan, it is fixed at the eQTL locus, and the mediator is changed.

Because the eQTL is always included in the alternative model but not the null model, the logP of the mediation scan should fluctuate around the observed eQTL logP. At mediators that possess some or all of the information present in the eQTL, the logP will drop. Significant logP drops represent candidate mediators. Significance thresholds to control FWER were determined by performing mediation scans on 1,000 permutations of the mediator variable. See **Appendix C** for further discussion and greater detail on the mediation analysis, including permutation approach to significance thresholds and the formal criterion for mediation detection.

Software and Data Availability

All statistical analyses were conducted with the R statistical programming language (R Core Team 2019). The R package

miQTL was used for all the mapping and mediations analyses, and is available on GitHub at https://github.com/gkeele/miqtl. A static version of miQTL is also provided in the **Supplement** (File S2).

The data, results files, and scripts to generate the results are provided in the **Supplement**, described in File S1. All files are available at figshare. File S3 contains the founder haplotype mosaics used for this study in a format that works with miQTL. Files S4 - S9 are the processed gene expression and chromatin accessibility data for liver, lung, and kidney used in the differential and QTL analyses. Files S10 - S12 are the results tables of differentially expressed genes for (liver/lung), (liver/kidney), and (lung/kidney); and Files S13 - S15 are results tables for differentially accessible regions. Files S16 - S18 are the local-eQTL results tables for liver, lung, and kidney, respectively; Files S19 - S21 are distal-eQTL results tables; Files S22 - S24 are local-cQTL results tables; and Files S25 - S27 are distal-cQTL results tables. Files S28 - S30 are results tables of local-eQTL with chromatin mediators. File S31 is the supplemental tables and figures.

Results

Tissue-specific genome-wide gene expression and chromatin accessibility in three tissues

Gene expression, using RNA-seq, and chromatin accessibility, using ATAC-seq, were measured in lung, liver, and kidney tissues of 47 unique strains in the Collaborative Cross (CC) mouse population. As the function of each tissue is quite distinct, we expected that these data would reflect tissue-specific differences. Principal components analysis (PCA) on each of the gene expression and chromatin accessibility profiles clearly show the clustering of all samples by tissue (**Figure S1**).

To further validate the quality of these data, we performed pairwise differential expression (DE) and differentially accessible region (DAR) analysis between the three tissues (**Table S1**). We found between 3,564 - 5,709 DE genes, and 28,048 - 40,797 DARs across the comparisons (FWER < 0.1). For both expression and chromatin accessibility, liver and kidney tissues were the most similar, and lung and liver were the most distinct, also reflected in the PCA plots. Pathway analyses showed many between-tissue differences related to metabolic and immune-related pathways (FWER < 0.1) that reflects the distinct demands of each tissue. Energy metabolism pathways were more active in liver and kidney and immune-related pathways were more pronounced in lung. We compared the concordance between DE genes and DARs genome-wide and observed that most DE gene promoters do not show significant differences in chromatin accessibility (**Figure S2**). In cases where there is significantly variable accessibility in the promoter of a DE gene, though, the vast majority agree in direction (i.e. higher expression and greater accessibility).

Tissue-specific expression QTL

To evaluate the impact of genetic variation on gene expression, three approaches to detect eQTL were used (**Table 1**), summarized here and described more completely in **Appendix B**. Method 1 was the most stringent and involved a multi-stage conditional regression procedure paired with an FDR control that allowed for the detection of potentially multiple eQTL per gene and requiring significance at a genome-wide level. Method 2 required FDR-controlled significance at the less stringent chromosome-wide level and only used results from the first stage of the multi-stage analysis, which we refer to as a single-step analysis. Method 3 was also a single-step analysis and employed an even more lenient significance criteria based on the genome-wide and chromosome-wide adjusted FWER p-value. We classified local-eQTL as being within 10Mb of the transcription start site (TSS) of the gene, and distal-eQTL as being greater than 10Mb from the TSS on the same chromosome (intra-chromosomal) or on a different chromosome (inter-chromosomal). For Method 2, only local-eQTL and intra-chromosomal distal-eQTL were detected due to our use of chromosome-wide significance criteria. For Method 3, the least stringent method, we only detected local-eQTL.

After filtering lowly expressed genes, a total of 8401, 11357, and 10092 genes were considered for liver, lung, and kidney tissues, respectively; a breakdown of the genes analyzed per tissue for QTL analysis is provided in **Figure S3 [left]** using an UpSet plot (Conway *et al.* 2017). Positions of local-eQTL detected for each tissue using any of the three methods (Method 1, genome-wide q-value < 0.1; Method 2, chromosome-wide q-value < 0.1; Method 3, genome-wide and chromosome-wide FWER p-value < 0.05) are shown in **Figure 3 [top]** (light colored dots), as well as summarized in **Table S2**. We also identified eQTL with Method 1 using a more relaxed significance threshold (q-value < 0.2; **Table S3**). Increasing the FDR threshold from 0.1 to 0.2 increases the number of distal-eQTL detected to a greater extent than local-eQTL. Using Method 3 and requiring genome-wide significance, the percentage of tested genes with local-eQTL are 6.3%, 8.4%, and 9.5% for lung, liver, and kidney, respectively (**Table S2**). These percentages increase to 16.6%, 19.8%, and 20.8% when criteria were less stringent with only chromosome-wide significance. The majority of genes with local-eQTL were only observed in one tissue (**Figure S5 [left]**).

For each eQTL detected by any of our methods, we estimated the effect size of the eQTL by two methods, either the coefficient of determination (R^2) for the fixed effect fitting of the QTL term (**Methods**; Eq 4) or as the proportion of variance explained based on the variance component estimate from a random effect fitting of the QTL term (**Methods**; Eq 5). We primarily report results from the fixed effect approach because the estimates were largely consistent with plausible QTL effect sizes given the size of the study, outlined in Keele *et al.* (2018), namely we were not well-powered to detect QTL with effect sizes < 50% at genome-wide significance. As expected, Method 1 detects eQTL with large effects (**Figure S7**, red dots). The less stringent Methods 2 and 3 result in greater power to detect local-eQTL with reduced effects (**Figure S7**, gray and blue dots). Distal-eQTL discovered with the multi-stage mapping procedure (Method 1; q-value < 0.1) were detected for $\leq$ 1.6% of tested genes (**Table S2**; **Figure 3 [top]**, dark colored dots), and, consistent with previous studies (*eg* Chick *et al.* 2016), have weaker effects than local-eQTL (**Figure S8 [top]**). For both Methods 1 and 2, almost three times as many local-eQTL are detected compared to distal-eQTL, likely reflecting these stronger effects. For some distal-eQTL, the effect sizes estimated through random effects models approach zero, potentially resulting from highly influential data points in the fixed effect regression procedure and are thus likely to be false positives (**Figure S10**).

To assess whether distance measures to classify intra-chromosomal local- vs distal-eQTL were reasonable, we plotted the statistical association for all intra-chromosomal eQTL from Method 1 (**Figure S11 [top]**) and the less stringent Method 2 (**Figure S12 [top]**). In addition to finding that the majority are local-eQTL, we see a drastic reduction in the strength of the

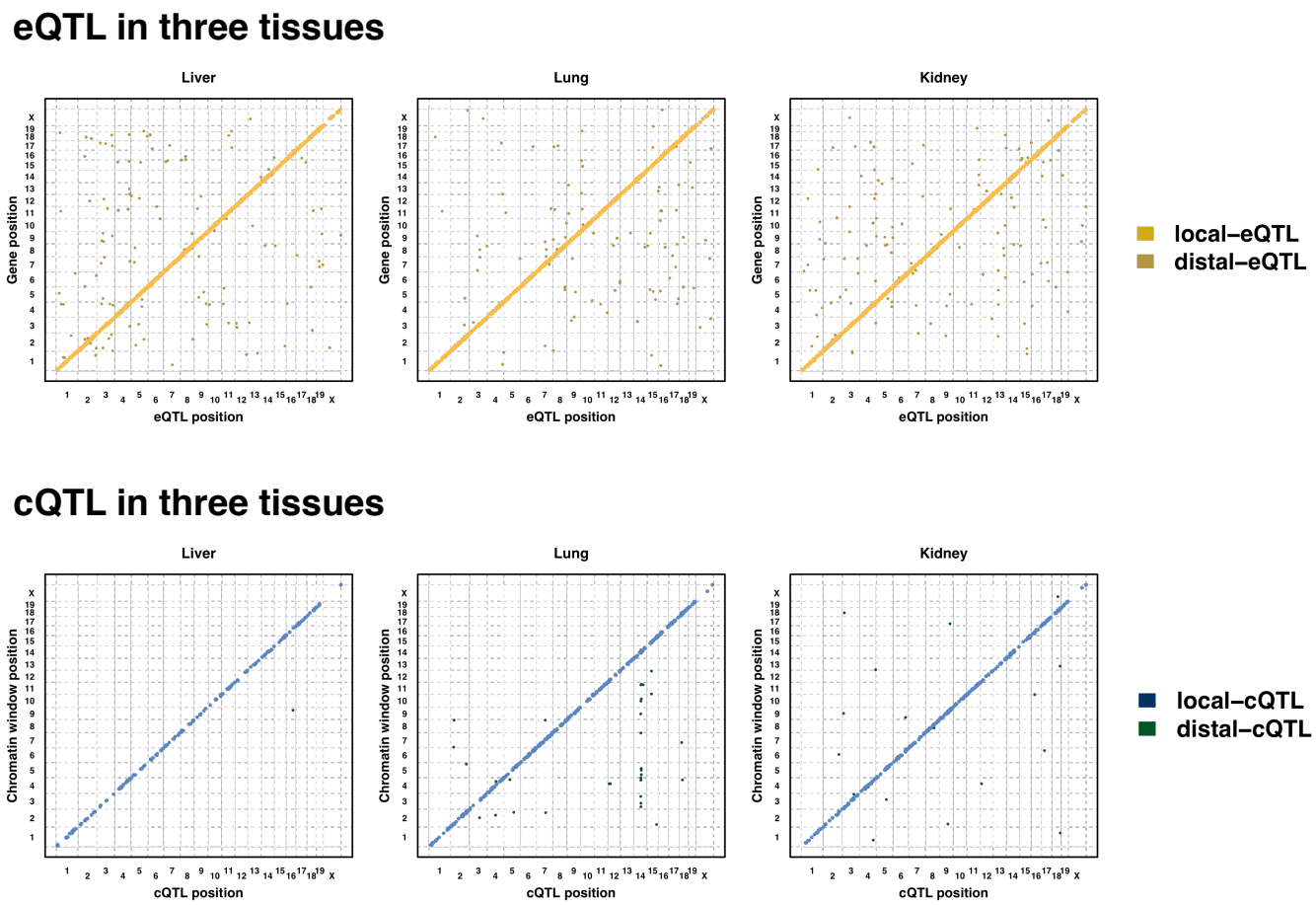

Figure 3 Detected QTL are largely local for both gene expression and chromatin accessibility. Detected QTL from Method 1 (multi-stage FDR) and Method 3 (genome-wide and chromosome-wide) are included, excluding intra-chromosomal distal-QTL detected through Method 2. The y-axis represents the genomic position of the gene or chromatin site, and the x-axis represents the genomic position of the QTL. Local-QTL appear as dots along the diagonal, and distal-QTL off of it.

statistical association outside of the defined local region. This suggests that intra-chromosomal distal-eQTL are more similar to inter-chromosomal distal-eQTL than local-eQTL, and that our 10Mb boundaries are appropriate.

Founder allele effects were estimated for all eQTL, using a random effects model to constrain unstable estimates. Consistent with previous studies (*e.g.*Aylor *et al.* 2011), the CAST and PWK alleles tended to have more extreme effects than the classical inbred strains. This pattern was observed for both local and distal eQTL (**Figure S13**).

Tissue-specific chromatin QTL

To determine genetic effects on chromatin structure, genomic regions were divided into ∼300 base pair windows, and used in chromatin accessibility QTL (cQTL) analysis with the same methods used for eQTL. We tested 11448, 24426, and 17918 chromatin regions in liver, lung, and kidney, respectively. The overlap in chromatin windows across tissues is described in **Figure S3 [right]**. Overall, there were substantially fewer cQTL detected compared to eQTL for all tissues (**Figure 3 [bottom]**;**Table S4**). As with eQTL, cQTL are more likely to be local (66%-94.1% for Method 1; 75%-90% for Method 2). Consistent with eQTL, the effect sizes of local-cQTL are on average higher than distal-cQTL (**Figure S8 [bottom]**), likely contributing to the to the small number of detected distal-cQTL. cQTL with low effect size are primarily distal-cQTL and may represent false positives. Likewise, most intra-chromosomal cQTL detected using Method 1 or 2 are local-cQTL, and the significance of association is much greater for the local-cQTL (**Figure S11 [bottom]** and **Figure S12 [bottom]**). Similar to local-eQTL, though to a lesser extent, the majority of chromatin windows with local-cQTL were only observed in one tissue (**Figure S5 [right]**). Again we found that the CAST and PWK founder alleles have more extreme effects than the other strains for local-cQTL, though the pattern is less pronounced than with local-eQTL, likely due to the reduced number of cQTL (**Figure S14**). The numbers of detected distal-cQTL are low, and no clear trends are obvious.

Paired QTL detected in multiple tissues

For a given trait, QTL from different tissues were paired based on co-localizing to approximately the same genomic region. For local-QTL, both had to be within the local window, defined at 10Mb around the gene TSS, resulting in a maximum distance of 20Mb between QTL. For distal-eQTL, both had to be within 10Mb of each other. For local-eQTL, 761 (liver/lung), 1206 (liver/kidney), and 1025 (lung/kidney) pairs were detected.

With distal-eQTL, 61 (liver/lung), 120 (liver/kidney), and 59 (lung/kidney) were identified. For cQTL, the vast majority of pairs were local, with 55 (liver/lung), 56 (liver/kidney), and 142 (lung/kidney). Only 4 distal-cQTL pairs were observed, all between lung and kidney. The effect sizes of paired QTL vary across tissues, though they were significantly correlated (**Figure S15**).

Correlations between the founder allele effects for pairs of QTL were calculated (FDR $\leq$ 0.1; **Figures S17**), resulting in 345 (liver/lung), 623 (liver/kidney), and 498 (lung/kidney) local-eQTL pairs and 21 (liver/lung), 34 (liver/kidney), and 16 (lung/kidney) distal-eQTL pairs found to be positively correlated. Highly correlated QTL pairs were highly proximal to each other (**Figure S16**), suggesting that the underlying causal variants are the same or linked. For local-cQTL, 47 (liver/lung), 48 (liver/kidney), and 118 (lung/kidney) positively correlated pairs were identified. The correlations between distal-cQTLs were not formally tested because there were only four pairs; however, three of the four pairs had correlations greater than 0.5. No negatively correlated QTL pairs were detected, after accounting for multiple testing.

Cytochrome c oxidase subunit 7C (*Cox7c*) and ubiquitin C (*Ubc*) are examples of genes that possess local-eQTL with highly correlated effects in all three tissues (**Figure 4A**). For *Cox7c* local-eQTL were driven by high expression when the CAST allele was present, intermediate expression with 129, NOD, and NZO alleles, and low expression with A/J, B6, PWK, and WSB alleles. The *Ubc* local-eQTL were driven by high expression with the B6, NOD, NZO, and WSB alleles. Though the founder effects are consistent across tissues for both genes, we note that expression of *Cox7c* was significantly higher in liver compared to both lung ($q = 5.41\text{e}{-18}$) and kidney ($q = 6.25\text{e}{-19}$), and for *Ubc*, expression in liver and lung were considered significantly differential ($q = 0.022$).

QTL pairs that are uncorrelated potentially represent distinct tissue-specific QTL where genetic variants influence gene expression or chromatin accessibility in a tissue-specific pattern. For example, the solute carrier family 44, member 3 (*Slc44a3*) gene has correlated local-eQTL effects in lung and kidney, but unique effects in liver (**Figure 4B**). Notably, the effects in liver are anti-correlated with the effects in lung and kidney, suggesting the liver eQTL could be transgressive (Rieseberg *et al.* 1999) to the eQTL in lung and kidney, whereby the effects of the founder alleles are reversed. For *Slc44a3*, CAST, PWK, and WSB alleles result in higher expression in liver, but lower expression in lung and kidney. The local-eQTL for *Slc44a3* were more similar in location in lung and kidney, whereas the liver eQTL was more distal to the gene TSS. Overall, the expression data, estimated effects, and patterns of association are consistent with lung and kidney sharing a causal local-eQTL, and liver possessing a unique one.

Another example is phosphatidylinositol-4-phosphate 3-kinase catalytic subunit type 2 gamma (*Pik3c2g*), a gene of interest for diabetes-related traits (Braccini *et al.* 2015). *Pik3c2g* has local-eQTL in all three tissues but that are all uncorrelated (**Figure 5**). Expression of *Pik3c2g* varies at statistically significant levels for liver versus lung ($q = 3.12\text{e}{-13}$) and lung versus kidney ($q = 3.55\text{e}{-6}$). The presence of tissue-specific local-eQTL is further supported by the *Mus musculus* lineages in the genomic region local to *Pik3c2g*. The CC founder strains all possess contributions from three subspecies of *M. musculus*: *domesticus* (*dom*), *castaneus*, (*cast*), and *musculus* (*mus*) (Yang *et al.* 2011). Crowley *et al.* (2015) found that allele-specific gene expression in mice descended from the CC founders often followed patterns that matched the subspecies inheritance at the gene regions. In lung, the CAST allele, which represents the *cast* subspecies lineage at the locus, results in higher expression, consistent with a *cast* versus *dom*/*mus* allelic series. In kidney, the B6, 129, NZO, and WSB alleles result in higher expression, whereas A/J, NOD, CAST, and PWK alleles show almost no expression, mostly consistent with a *dom* versus *cast*/*mus* allelic series. The PWK founder has a small *dom* haplotype block at the QTL peak in a broader region that is largely *mus*. The expression data are highly consistent with PWK having the *mus* allele, suggesting that the causal variant is located in a nearby region, supported by the variant association. Notably, the B6 allele expression appears intermediate to the other founders that possess *dom* inheritance, representing a potentially multi-allelic QTL that haplotype association can better identify in comparison to bi-allelic variant association. In liver, the B6 allele resulted in lower expression. The B6 founder does not have a unique subspecies lineage at the locus, but instead possesses a recombination event between *dom* and *mus*, potentially disrupting the local regulation specific to each subspecies lineage and explaining the reduced expression.

Founder allele effects that correlate across tissues can provide stronger confirmation of distal-QTL, even those with marginal significance. For example, the Aldo-keto reductase family 1, member E1 (*Akr1e1*; **Figure 6A**) gene is located on chromosome 13 and had no local-eQTL detected in any of the tissues. Distal-eQTL for *Akr1e1* were detected in all tissues that localize to the same distal region of chromosome 4. The founder allele effects of the distal-eQTL are all highly correlated, with A/J, B6, 129, and CAST being the more highly expressed alleles. The overall magnitude of expression varies significantly across the tissues, with liver and kidney having significantly higher expression than lung ($q = 1.63\text{e}{-19}$ and $q = 4.40\text{e}{-34}$, respectively). Another example is the Period circadian clock 2 (*Per2*; **Figure 6B**) gene, which possesses intra-chromosomal distal-eQTL that are detected in all three tissues that are approximately 100Mb away from its TSS. Despite being detected with the highly lenient Method 2 (FDR $\leq$ 0.2), the founder allele effects are significantly correlated among the tissues, characterized by high expression with the NOD and PWK alleles present. Collectively, these findings provide strong validation for the distal-eQTL, which would commonly not even be detected based on tissue-stratified analyses.

QTL analysis in multiple tissues also allows for the detection of tissue-specific distal-QTL. The Ring finger protein 13 (*Rnf13*; **Figure 6C**) gene was lowly expressed in all tissue but with a strong local-eQTL detected in liver. The multi-stage conditional regression approach, conditioning on the local-eQTL, detected a distal-eQTL on chromosome 12. Comparing the residuals after regressing out the local-eQTL and founder allele effects suggested that mice with B6 contributions at the distal locus had higher *Rnf13* expression. In lung tissue, a distal-eQTL was detected on the X chromosome with a different founder allele pattern, notably high expression when A/J, B6, and PWK alleles were present, and low expression with NOD and NZO alleles. These examples demonstrate that analyzing multiple tissues can provide greater evidence of distal-QTL, as well as potentially identify tissue-specific genetic regulation.

Mediation of eQTL

Measurement of gene expression and chromatin accessibility data in the same mice and tissues enables the use of mediation

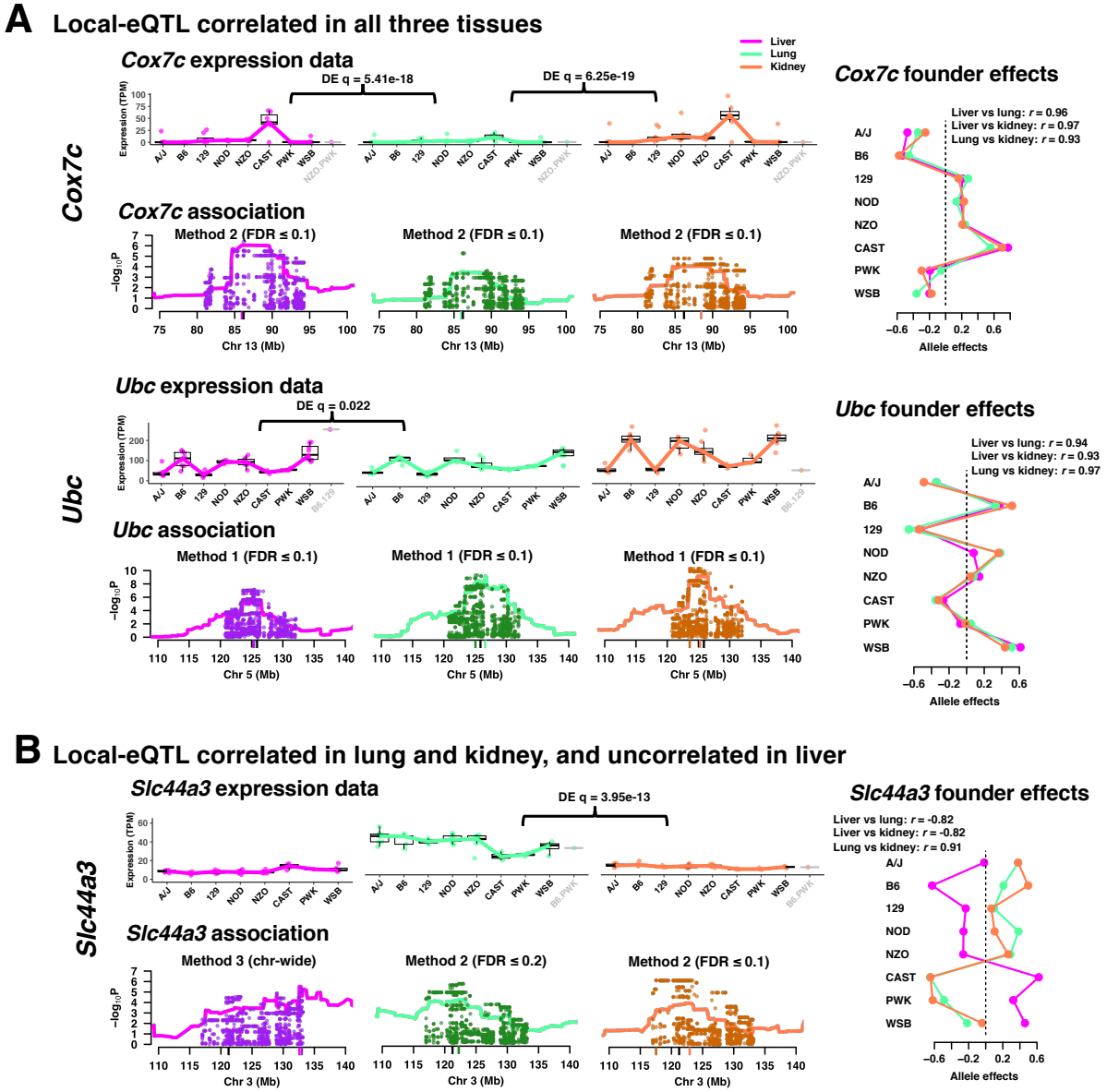

Figure 4 Examples of genes with local-eQTL observed in all three tissues. *Cox7c* and *Ubc* possess local-eQTL with highly correlated founder alllele effects across all three tissues, supportive of shared causal origin (A). *Slc44a3* has a more complicated pattern of local-eQTL founder effects across the tissues, with correlated effects shared between lung and kidney, and transgressive effects in the liver eQTL by comparison, consistent with distinct casual variants comparing liver to lung and kidney. (B). For each gene, the expression data is plotted with boxplots based on most likely founder haplotype pair (diplotype), and differential expression between tissues is highlighted. The haplotype association for each tissue is also included near the gene TSS with variant association overlaid. The most statistically rigorous method to detect the QTL is also included. The black tick represents the gene TSS and the colored ticks represent haplotype and variant peaks. Founder effects estimated as constrained BLUPs are also included with their pairwise correlations. Estimated effects were generally consistent with the expression data.

analysis to elucidate the relationships between genotype, chromatin accessibility, and gene expressions. We assessed evidence for two mediation models of eQTL effects (**Figure 2**). First, we tested for proximal chromatin state as a mediator of the effect of local-eQTL on gene expression using an approach adapted from Chick *et al.* (2016) to detect mediation through chromatin accessibility of local-eQTL detected through Method 3 (genome-wide and chromosome-wide; see **Appendix C** for greater detail).

Across the three tissues, we found from 13-42 local-eQTL showed evidence of mediation through proximal chromatin accessibility at genome-wide significance, and 35-106 at chromosome-wide significance (**Table S6**). The coiled-coil domain containing 137 gene (*Ccdc137*) is a strong example of mediation through local chromatin accessibility. **Figure 7** includes the genome scans that identify local-QTL for *Ccdc137* expression (eQTL) and a proximal chromatin accessibility window (cQTL). A large drop in eQTL significance at the location of the cQTL represents a strong mediation signal. A closer inspection of the

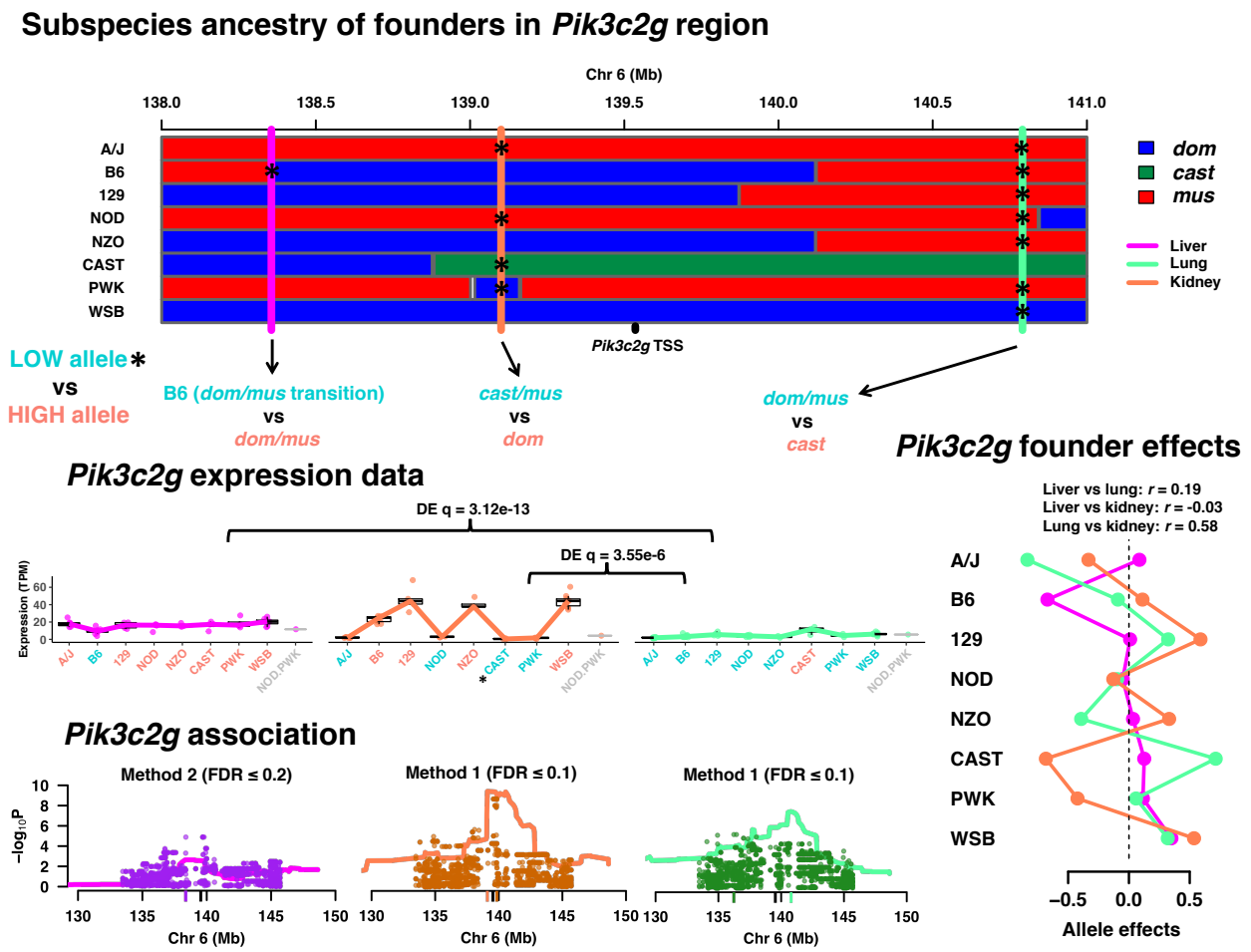

Figure 5 *Pik3c2g* possesses tissue-specific local-eQTL. Local-eQTL for *Pik3c2g* are detected in all three tissues in the 3Mb region surrounding its TSS. The genomes of the CC founders can be simplified in terms of contributions from three subspecies lineages of *M. musculus*: *dom* (blue), *cast* (green), and *mus* (red). The effects of each local-eQTL matched the subspecies contributions near the eQTL coordinates, with low subspecies alleles colored teal and high alleles colored salmon, consistent with local-eQTL for *Pik3c2g* being distinct and tissue-specific. The gene expression data is represented as boxplots, categorized based on most likely diplotype at the eQTL for each CC strain. Haplotype and variants associations are included for each tissue, with the black tick representing the *Pik3c2g* TSS and colored ticks representing haplotype and variant peaks. The most rigorous procedure to detect each QTL is reported. Founder effects, estimated as constrained BLUPs, were consistent with the expression data, and uncorrelated across the tissues.

region shows that the significance of the eQTL and cQTL are similar but notably stronger for chromatin, as expected by the proposed mediator model.

Co-localization of eQTL and cQTL is not sufficient for mediation, shown in **Figure S18**. For both *Hdhd3* in liver (**Figure S18 [left]**) and the acyl-Coenzyme A binding domain containing 4 gene (*Acbd4*) in kidney (**Figure S18 [right]**), there are local-eQTL with a co-localizing cQTL. Comparing the statistical associations at the genomic locations of the eQTL and cQTL as well as the founder haplotype effects for both the eQTL and cQTL, the correspondence in both is better for *Hdhd3* compared to *Acbd4*. As with *Ccdc137*, a strong mediation signal is detected for *Hdhd3*, indicated by the decrease in the eQTL association when conditioning on the chromatin state. Mediation is not detected for *Abcd4*, which is consistent with the genetic regulation of both gene and chromatin site not stemming from the same causal origin.

Next, we tested for evidence of expression levels of genes proximal to distal-eQTL acting as mediators, using an approach similar to Keller *et al.* (2018) for distal-eQTL detected through Method 1. Eight genes were identified with mediated distal-eQTL (**Table S7**). As example, the distal-eQTL of cyclin Y-like 1 (*Ccnyl1*) is mediated by the expression of zinc finger protein 979 (*Zfp979*), a putative transcription factor (**Figure S19**). The founder haplotype effects at the distal-eQTL of *Ccnyl1* are highly correlated with the local-eQTL effects for *Zfp979*, though with a reduction in overall strength, in accordance with the proposed mediation model.

Proposed model for the genetic regulation of Akr1e1 expression

As detailed above, *Akr1e1* possesses a strong distal-eQTL, observed in all three tissues with significantly correlated founder effects. Additionally, we found that this effect is strongly me-

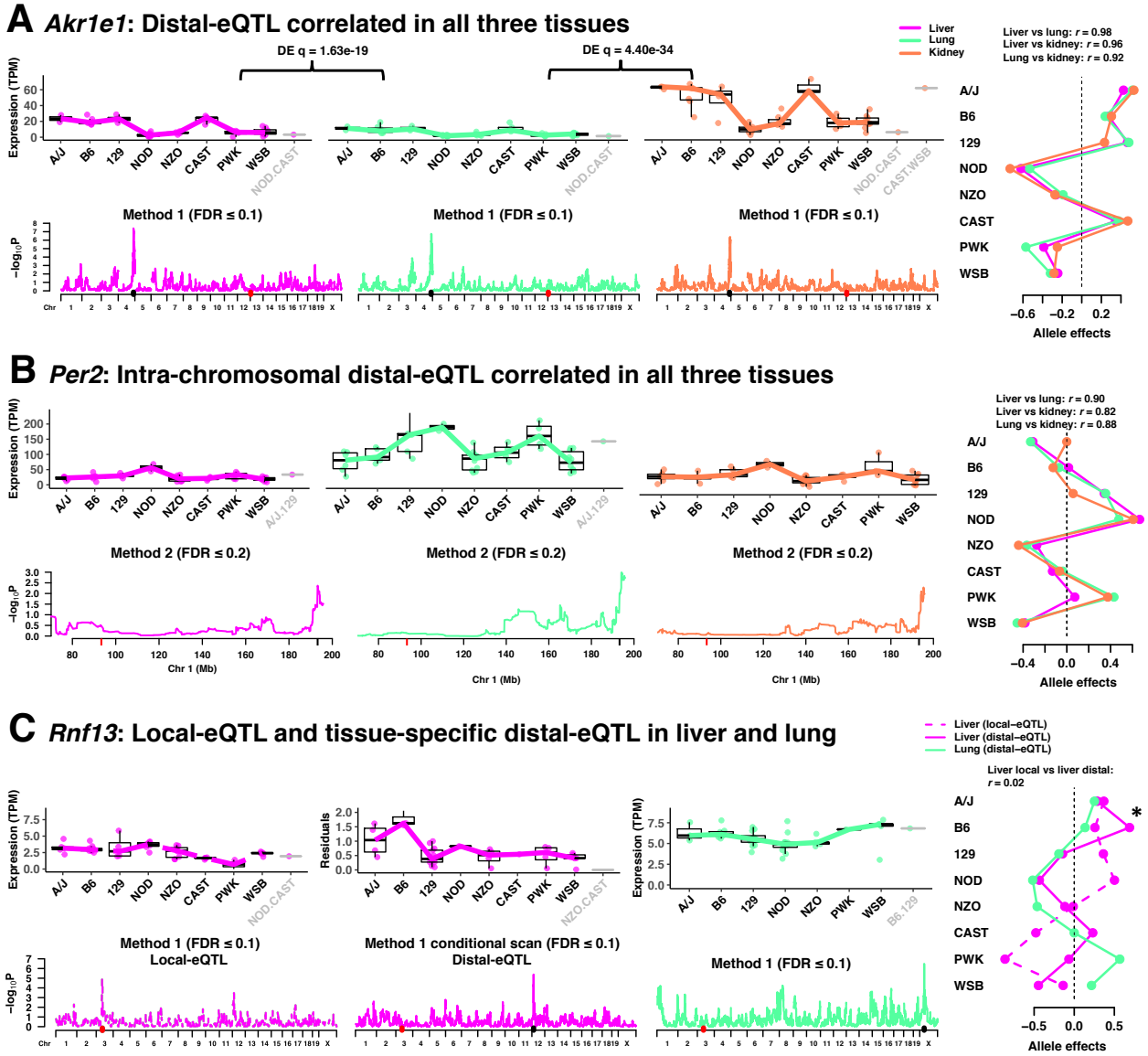

Figure 6 Examples of genes with distal-eQTL effect patterns across tissues. *Akr1e1* has highly significant distal-eQTL detected on chromosome 4 in all three tissues with correlated founder effects (A). *Per2* has intra-chromosomal distal-eQTL leniently detected 100Mb away from the TSS, also with highly correlated founder effects across the tissue, providing further support that the distal-eQTL are real (B). *Rnf13* has a liver-specific distal-eQTL on chromosome 13 detected after conditioning on its local-eQTL on chromosome 2, and a lung-specific distal-eQTL on chromosome X, each with distinct founder effect patterns (C). Expression data are represented as boxplots for most likely diplotype, with differential expression noted when significant. Haplotype associations for each tissue distal-eQTL combination are shown, with the most rigorous statistical procedure for detection reported. Red ticks signify the gene TSS and black ticks represent that eQTL peak. Fit founder effects, estimated as BLUPs, are included, along with the pairwise correlations of the eQTL. The black asterisk marks the high B6 effect that drives the distal-eQTL in liver after conditioning on the local-eQTL for *Rnf13*.

diated through Zinc finger protein 985 (*Zfp985*) expression in lung. The *Zfp985* TSS is located on chromosome 4 at 147.6Mb, while the *Akr1e1* distal-eQTL coordinates are 142.5Mb, 143.2Mb, and 148.6Mb in liver, lung, and kidney, respectively (**Figure S20**). *Akr1e1* and *Zfp985* expression are negatively correlated ($r = -0.69$), which is consistent with *Zfp985* inhibiting *Akr1e1* expression. This same distal-eQTL and mediator relationship for *Akr1e1* was observed in pancreatic islet tissue in DO mice (Keller *et al.* 2018). The presence of distal genetic regulation for *Akr1e1* was previously described in Hamilton-Williams *et al.* (2013). The distal-eQTL we detect are highly proximal to their *Idd9.2* region (Hamilton-Williams *et al.* 2010), originally defined as spanning 145.5 - 148.57Mb of chromosome 4 and subsequently using NOD mice congenic with C57BL/10 (B10) introgressions Hamilton-Williams *et al.* (2013). The B10 allele was found to be protective against the development of diabetes, characteristic of NOD mice. *Akr1e1* is involved in glycogen metabolism and the larger *Idd9* region harbors immune-related genes, implicating *Akr1e1* and other genes located in and regulated by *Idd9* as potential diabetes-related genes. Consistent with these studies,

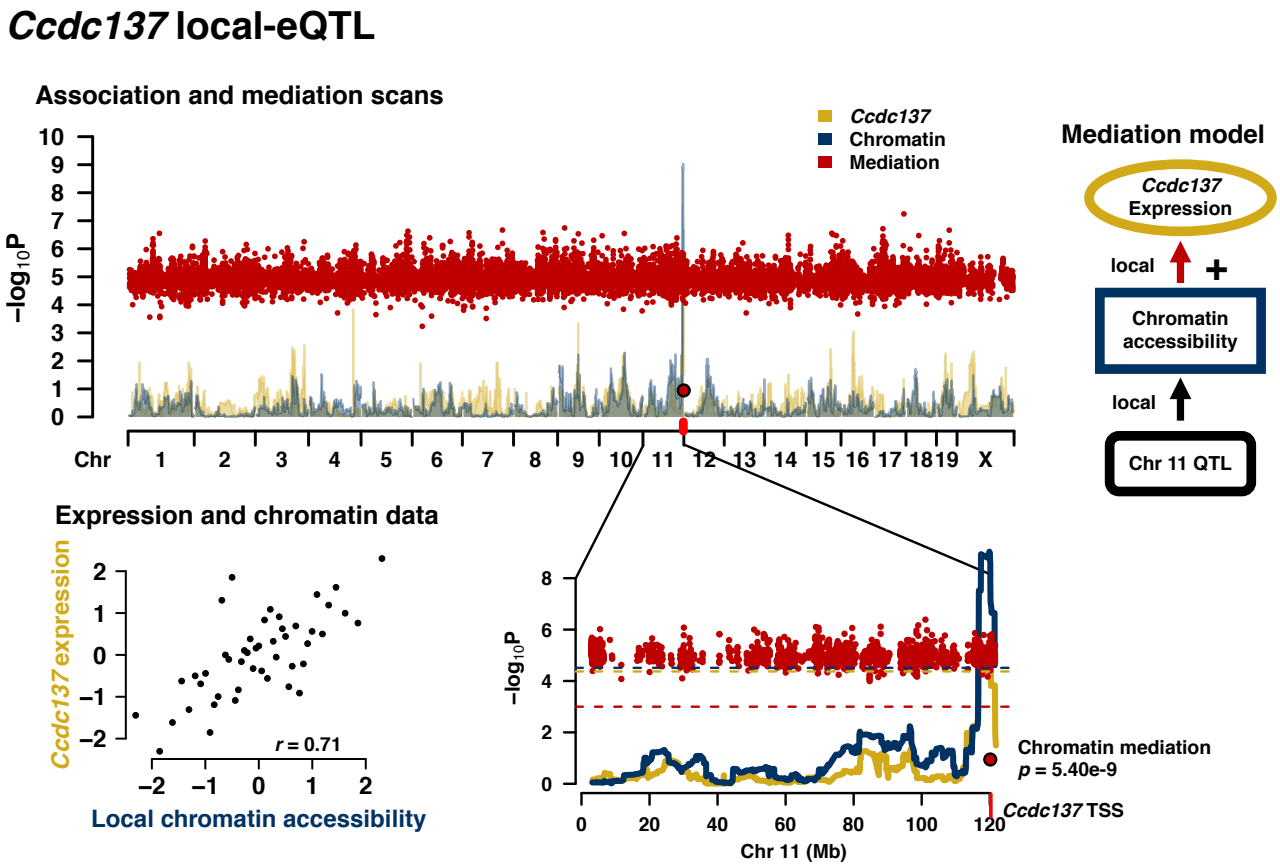

Figure 7 ***Ccdc137*** **local-eQTL is mediated by proximal chromatin accessibility.** *Ccdc137* expression and the chromatin accessibility in the proximal region are highly correlated ($r = 0.71$). Genome scans for *Ccdc137* expression (yellow), nearby chromatin accessibility (blue), and chromatin mediation of the *Ccdc137* local-eQTL (red) in lung tissue. The local-eQTL and local-cQTL for the chromatin region at the TSS of *Ccdc137* (red tick) are over-lapping. The steep drop in the statistical association, represented as logP, for the chromatin site in the mediation scan supports chromatin mediation of *Ccdc137* expression, depicted as a simple graph in the topright. The QTL and mediation signals are detected at genome-wide significance.

CC strains with NOD inheritance at the distal-eQTL have low expression of *Akr1e1*. Genetic variation from B10 is not present in the CC, but the closely related B6 founder has high expression of *Akr1e1* like B10. Similar to NOD, NZO, PWK, and WSB alleles result in low *Akr1e1* expression, while A/J, 129, and CAST alleles join B6 as driving high expression (**Figure 8**). Lowly expressed *Zfp985* is a strong candidate for driving these effects on *Akr1e1*.

Further elucidation of the genetic regulation of *Akr1e1* was observed in kidney tissue, where a strong distal-cQTL is mapped to the *Idd9.2* region for the chromatin site proximal to *Akr1e1*. Based on the close proximities of the chromatin window to *Akr1e1* and the distal-cQTL to the distal-eQTL, we mediated the distal-eQTL effect on the chromatin window, which was highly significant (permP = 2.18e−13). The founder effects for the distal-cQTL are correlated with the distal-eQTL effects correlated ($r = 0.92$). The relative magnitudes of the QTL effect sizes and mediation p-values, shown in **Figure S21**, support a causal model whereby *Zfp985* expression reduces the expression of *Akr1e1* by altering chromatin accessibility in the *Akr1e1* proximal promoter region (**Figure 8**).

Overview of QTL and mediation analyses

All QTL and mediation results for all three tissues are depicted as circos plots (Gu *et al.* 2014) in **Figure 9**. Notably, local-QTL and chromatin mediation are not evenly distributed across the genome, but tend to aggregate in pockets. In particular, there is a high concentration of cQTL and chromatin mediation detected along chromosome 17, observed across all three tissues, which happens to correspond to the immune-related major histocompatibility (MHC) region in mouse. However, we note that most of the chromatin-mediated genes in this region are not histocompatibility genes (**Files S28 - S30**). The patterns of distal-QTL and gene mediation appear to be tissue-specific, though there are consistent regions observed across the tissues, such as the *Idd9* region (Hamilton-Williams *et al.* 2010) of chromosome 4, which contains multiple zinc finger proteins and regulates genes such as *Akr1e1* and *Ccnyl1*. Previously discussed genes are highlighted for each tissue.

Discussion

We performed QTL mapping on gene expression and chromatin accessibility data in 47 CC strains. We have described the use of multiple procedures of varying statistical stringency to allow for detection of lower effect local-QTL, important in a study with a restricted sample size. Furthermore, we described statistical approaches for two models of the mediation of the eQTL effects, one through chromatin accessibility for local-eQTL, and the

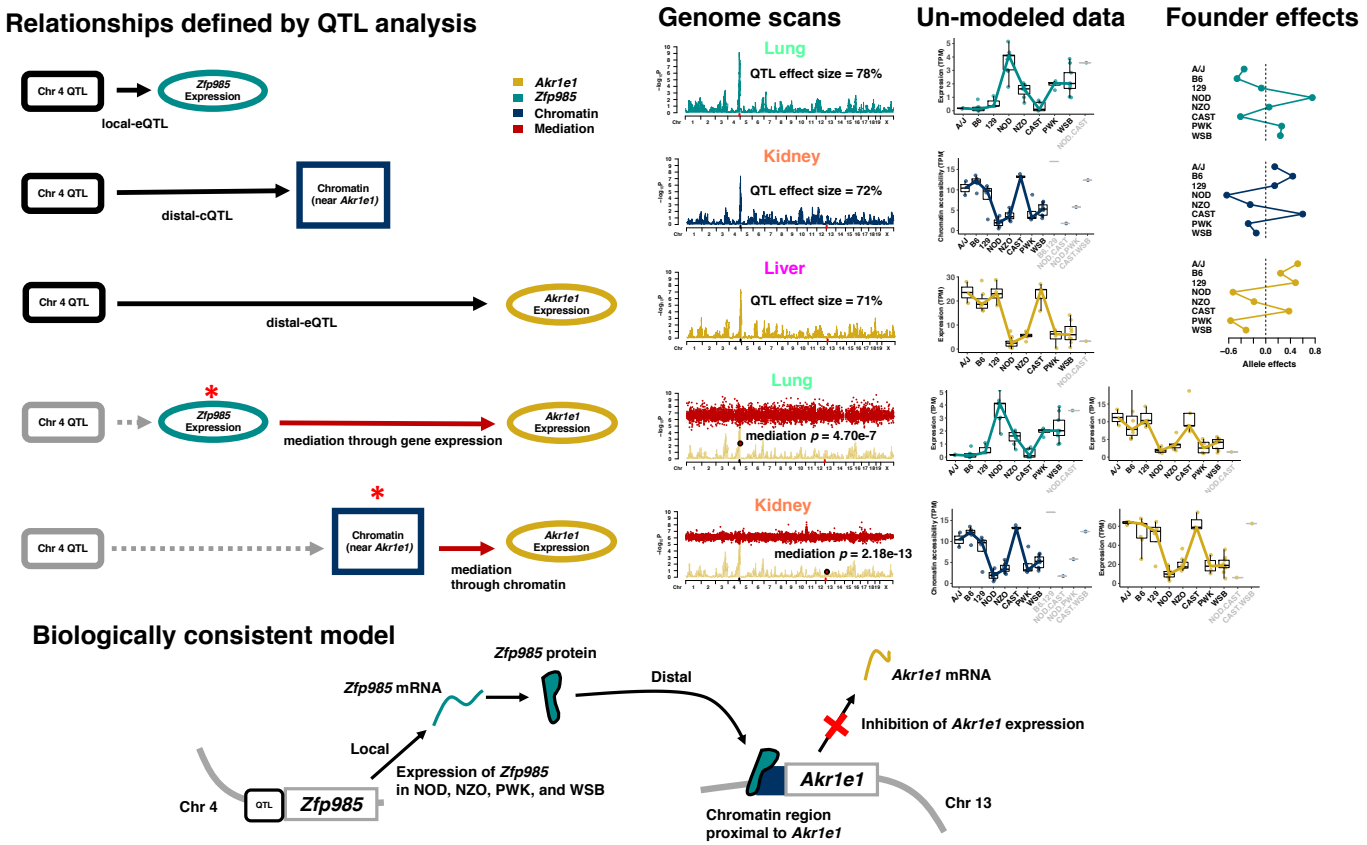

Figure 8 Complete mediation model for *Akr1e1* distal-eQTL. The genetic regulation of *Akr1e1* expression is reconstructed based on relationships observed across the three tissues. Distal-eQTL are detected in all tissues at similar levels of significance. A local-eQTL for *Zfp985* that is proximal to the *Akr1e1* distal-eQTL was observed in lung, and *Zfp985* expression is detected as anti-correlated mediator of the distal-eQTL, consistent with *Zfp985* suppressing *Akr1e1* expression. The chromatin site proximal to the *Akr1e1* TSS has a distal-cQTL detected in kidney. The chromatin accessibility at the site was found to be a significant correlated mediator of *Akr1e1* expression. Combining associations across tissues supports a biological model whereby *Zfp985*, expressed in mice with NOD, NZO, PWK, and WSB alleles, inhibits the proximal chromatin accessibility at the *Akr1e1* TSS. QTL and mediation genome scans are included, along with sequence phenotypes as boxplots categorized according to most likely diplotype, and modeled founder effects fit as BLUPs. The relative magnitudes of the QTL effect sizes (estimated from Eq 4) and mediation significance are consistent with the proposed model, with *Zfp985* local-eQTL > chromatin distal-eQTL > *Akr1e1* distal-eQTL and chromatin mediation > mediation through *Zfp985* expression.

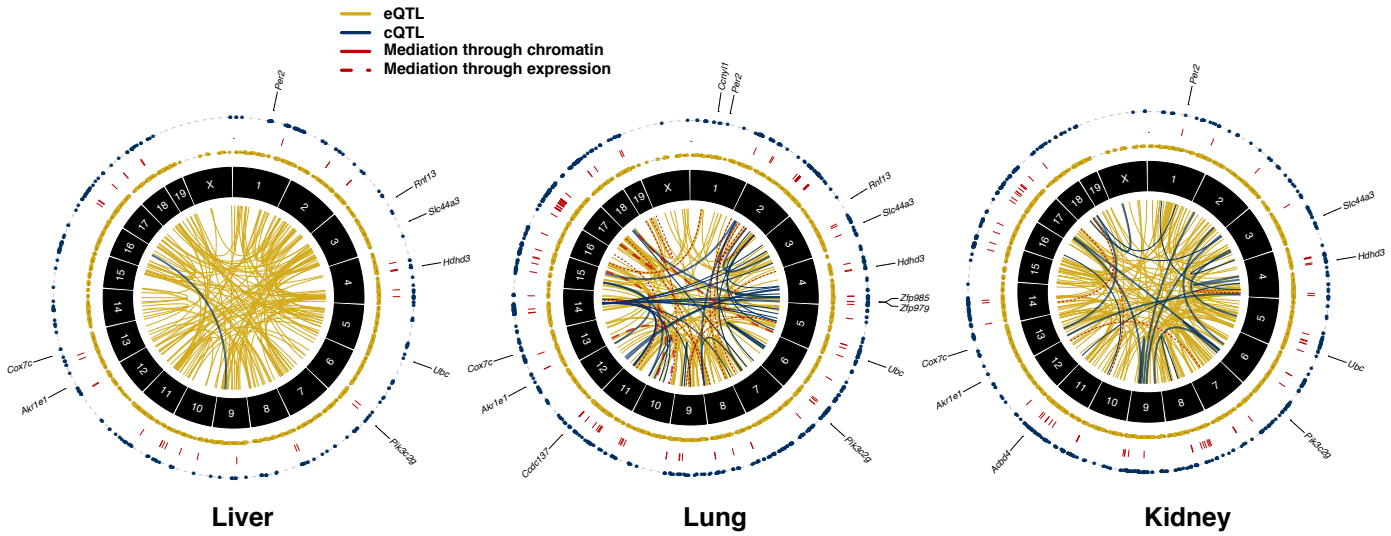

Figure 9 Summaries of QTL and mediation analyses. Circos plots of eQTL (yellow), cQTL (blue), and mediation (red) in lung, liver, and kidney. The two outer rings of dots represent local-eQTL and local-cQTL detected by Method 3 at chromosome-wide significance, with red lines between connecting genes and chromatin sites for which chromatin mediation was detected. The inner circle contains connections representing distal-eQTL, distal-cQTL, and gene-gene mediation from Method 1. Thick lines represent QTL and mediators with permutation-based p-value (permP) $< 1e{-}5$. The detected signals are primarily local, which also tend to be stronger than the observed distal signals. Fewer QTL and mediators are detected in liver tissue. Genes highlighted previously, such as *Akr1e1*, are indicated at their genomic coordinates.

other through activity of proximal genes as measured through their expression for distal-eQTL. We oriented our mediation through chromatin accessibility on local-eQTL, for which this study is better powered, and find significant support that mediation through chromatin is present (4.2-9.6% of local-eQTL across the tissues). Certain regions of the genome could not be assessed for cQTL or chromatin mediation due to ATAC-seq read alignment ambiguity, further restricting our ability to assess mediation. Despite these constraints, our results are consistent with the role of accessible chromatin regions in the regulation of gene expression (Klemm *et al.* 2019). Despite being underpowered to detect gene expression mediators of distal-eQTL we do find a number of strong candidates, including zinc finger protein genes.

Multiple procedures for QTL detection

The use of multiple procedures for QTL detection allowed us to assess varying degrees of statistical support for their associations, and of particular importance, leverage prior belief in the predominance of local associations in order to detect higher numbers of QTL for a sample of 47 mice. The multi-stage procedure (Method 1) allowed for genome-wide detection of multiple QTL per outcome while incorporating both FWER and FDR adjustments. Both the chromosome-wide FDR (Method 2) and local-only (Method 3) analyses strongly increase power to detect local-QTL, but lose the ability to identify distal-QTL located on other chromosomes.

Estimates of QTL effect size

Estimation of the variance explained by the QTL, here referred to as the effect size, can provide a useful and interpretable point summary for haplotype-based QTL mapping, analogous to the effect of the minor allele, commonly used in SNP association. We used two estimates of effect size, one based on a fixed effect fitting of the QTL term (Eq 4) whereas the other included the QTL as a random term (Eq 5). Both estimates of effect size possessed unique and appealing qualities. We primarily report results from the fixed effects estimate because these were largely consistent with the expectations prescribed by Keele *et al.* (2018) for a study of this size, whereas the conservative shrinkage-based estimate seemed overly reduced, likely due to the sample size. The mapping results are largely consistent with the genetic regulation of these molecular traits being largely Mendelian, with large effect sizes (> 60 %), though more complex genetic regulation is suggested by QTL with smaller effect sizes (< 50%) when detecting local-QTL with reduced stringency.

The shrunken effect size estimates can potentially identify putative QTL with unstable fixed effects from influential data points. Notably, a number of the distal-eQTL have very low random effect size estimates (**Figure S10**) compared to their fixed effects-based estimates. Notably, QTL with this pattern of effect size estimates are largely distal. A mapping procedure that fits the QTL effect as a random effect, as in Wei and Xu (2016), would likely not detect these QTL. However, a random effects procedure is more computationally intensive, which is particularly unappealing in the context of genome-wide molecular traits, thus justifying a follow-up approach in which effect sizes are evaluated for the detected QTL.

Joint QTL analysis in multiple tissues

Recent research, such as the GTEx project (Aguet *et al.* 2017), comprised of expression data from 44 human tissues, represents a growing interest in interrogating and characterizing differences in gene expression and its regulation across a wide range of tissues. Previous studies in mice (*e.g.*Huang *et al.* 2009) have used tissue-specific or stratified QTL analyses, similar to as

done herein, and then characterized the overlap across tissues. Our use of formal statistical tests of the correlation coefficient between the founder allele effects of overlapping QTL is, to our knowledge, a novel approach to formally defining overlapping QTL, allowing for the identification of proximal QTL with unique founder effects patterns, likely representing distinct tissue-specific QTL, such as observed with three tissue-specific local-eQTL for the gene *Pik3c2g* (**Figure 5**).

However, examples like the gene *Per2* (**Figure 6B**), which possessed marginally significant intra-chromosomal distal-eQTL with consistent founder allele effects observed across the three tissues, suggest that mapping QTL in all the tissues jointly could better leverage the full extent of the data and detect QTL more powerfully. Formal joint analysis approaches have been proposed, largely implemented for SNP association, including meta-analysis on summary statistics (*e.g.*Fu *et al.* 2012b; Sul *et al.* 2013) and fully joint analysis, including Bayesian hierarchical models (Flutre *et al.* 2013) and mixed models (Acharya *et al.* 2016). Extending such statistical frameworks to haplotype-based analysis in MPP, as used here, poses some challenges, including how to best generalize these approaches to the more complex genetic model, and in the context of the CC where the number of unique genomes is limited. Nevertheless, such approaches would certainly be more powerful and principled for detecting and calling QTL present across multiple tissues, and should be considered for future related research in the CC and similar populations, such as the DO. When multiple levels of molecular traits are measured on each individual, as is the case for this study, there is also the potential to incorporate joint analysis approaches for QTL mapping into the mediation analysis, which could presumably improve power.

Correlated founder effects suggest multi-allelic QTL

Haplotype-based association in MPP allows for the detection of multi-allelic QTL (Aylor *et al.* 2011), such as observed with the kidney local-eQTL of *Pik3c2g* in which mice with B6 contributions in the region have an intermediate level of expression. The correlation coefficient between the founder allele effects for QTL pairs provides an interesting summary, generally not possible in the simpler bi-allelic setting commonly used in variant association. The extent of the correlation between the founder effects for QTL pairs of certain genes, such as *Cox7c* and *Ubc* (**Figure 4A**) suggests that these QTL are at least subtly multi-allelic. Correlated founder effects are consistent with genetic regulation, even local to the gene TSS, being potentially complex with strain-specific modifiers.

Correlated founder effects in the presence of differential expression

We detected numerous genes with QTL pairs with correlated founder effects across tissues despite the overall expression patterns being significantly different based on the differential analysis; examples include *Cox7c*, *Ubc*, *Slc44a3*, and *Akr1e1*. We propose two primary explanations for the unique co-occurence of consistent genetic regulation of genes with expression profiles that vary significantly across tissues. The samples used in the study represent bulk tissue samples, and thus possess cellular heterogeneity that likely varies between liver, lung, and kidney tissues. DE signal may simply represent the varying cellular compositions of the tissues. Single cell sequence experiments would reduce noise from tissue heterogeneity, though the results would no longer reflect averages across the overall tissue. Statistical approaches that aim to estimate the cellular composition of a tissue (Aran *et al.* 2017) have been developed, which could provide some insights into how the tissues vary from each other. Alternatively, DE genes that share QTL across tissues could be real, and suggest that other tissue-specific modifiers of expression are active that alter the magnitude of the expression profile. Single cell sequencing experiments could be used to confirm such patterns in the genetic regulation of expression.

Mediation analysis

Our approach to mediation analysis is related to previous studies in DO mice (Chick *et al.* 2016; Keller *et al.* 2018; Skelly *et al.* 2019), though with modification, including the use of QTL effect sizes to establish consistency with the directionality of the mediator relationship and the formal calculation of a mediation p-value through permutation (permP). To our knowledge, such mediation analyses have not been used in the CC, though related conditional regression approaches were used in the incipient CC lines (Rutledge *et al.* 2014; Kelada *et al.* 2014). Mediation within the field of system genetics is a growing area of interest.

Determining causality, particularly in a biological system as complex as transcriptional regulation in bulk tissue samples, is impossible. These mediation analyses will largely reflect the correlations between the variables after adjusting for additional sources of variation, such as covariates and batch effects. Our additional step of requiring the mediator QTL to have a larger effect size than outcome QTL aims to identify trios that are consistent with the proposed mediation models. It is possible that the measurement of the mediator is a noisier process compared to the measurement of the outcome, which would result in the effect size of the mediator QTL being reduced, and thus potentially a false negative mediation result. Complex relationships with many factors or levels of feedback are also unlikely to be detected.

Despite these limitations, mediation analysis can provide specific candidates underlying detected QTL, particularly in comparison to considering all genes near the QTL peak. The genetic regulation of *Akr1e1* (**Figure 8**) provides a strong example case in which a single gene, *Zfp985*, in the eQTL region is identified as a strong mediator. It is possible that *Zfp985* is not the true mediator, but is instead strongly correlated with it. For comparison, Hamilton-Williams *et al.* (2013) proposed *Rex2* as a candidate, which was filtered out for having low expression from all three tissues in this study, whereas *Zfp985* was lowly expressed but successfully quantified in lung. Disentangling the relationships underlying QTL poses a strong challenge, and mediation analyses can provide strong candidates for downstream experiments for validation. Additionally, mediation analysis provides strong support that the causal mediator inhibits *Akr1e1* expression by reducing chromatin accessibility near its TSS.

Conclusion

In summary, in this study we have performed QTL and mediation analyses of gene expression and chromatin accessibility data in liver, lung, and kidney tissue samples from 47 strains of the CC. We use the formal statistical testing of correlation coefficient between founder allele effects of overlapping QTL to identify QTL that are likely functionally active in multiple tissues as well as likely distinct tissue-specific QTL, as is the case with the gene *Pik3c2g*. We detect extensive evidence of chromatin mediation of local-eQTL as well as gene expression mediators underlying distal-eQTL. One unique example is the elucidation of the ge-

netic regulation of *Akr1e1* expression, a gene that plays a role in glycogen metabolism, involving distal inhibition by a zinc finger protein mediator that reduces chromatin accessibility at the TSS. These findings highlight the ability of integrative QTL approaches, including mediation analysis, to identify interesting biological findings, including the ability to identify functional candidates for further downstream analysis.

Acknowledgments

This work was funded by grants from the National Institute of Environmental Health Sciences (NIEHS) (R01-ES023195 to IR and GEC and P30-ES025128). We are grateful for discussions with Gary A. Churchill for on mediation analysis that influenced this work. Additionally GRK was partially supported through funding from Dr. Churchill and The Jackson Laboratory.

Author contributions: GRK, BCQ, WV, and TSF wrote the manuscript. GRK, BCQ, and WV developed the statistical methods for QTL analysis, and GRK wrote the software to perform the respective analyses. BCQ processed the sequence data and performed the DE and DAR analyses. GRK and BCQ performed all QTL and mediation analyses. The authors declare no conflicts of interest.

Literature Cited

Acharya, C. R., J. M. McCarthy, K. Owzar, and A. S. Allen, 2016 Exploiting expression patterns across multiple tissues to map expression quantitative trait loci. BMC Bioinformatics **17**: 257.

Aguet, F., A. A. Brown, S. E. Castel, J. R. Davis, Y. He, *et al.*, 2017 Genetic effects on gene expression across human tissues. Nature **550**: 204–213.

Alasoo, K., J. Rodrigues, S. Mukhopadhyay, A. J. Knights, A. L. Mann, *et al.*, 2018 Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. Nature genetics p. 102392.

Aran, D., Z. Hu, and A. J. Butte, 2017 xCell: digitally portraying the tissue cellular heterogeneity landscape. Genome biology **18**: 220.

Aylor, D. L., W. Valdar, W. Foulds-mathes, R. J. Buus, R. A. Verdugo, *et al.*, 2011 Genetic analysis of complex traits in the emerging Collaborative Cross. Genome research **21**: 1213–22.

Baron, R. M. and D. a. Kenny, 1986 The Moderator-Mediator Variable Distinction in Social The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations. Journal of Personality and Social Psychology **51**: 1173–1182.

Battle, A., Z. Khan, S. H. Wang, A. Mitrano, M. J. Ford, *et al.*, 2015 Genomic variation. Impact of regulatory variation from RNA to protein. Science (New York, N.Y.) **347**: 664–7.

Battle, A., S. Mostafavi, X. Zhu, J. B. Potash, M. M. Weissman, *et al.*, 2014 Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Research **24**: 14–24.

Baud, A., V. Guryev, O. Hummel, M. Johannesson, Rat Genome Sequencing and Mapping Consortium, *et al.*, 2014 Genomes and phenomes of a population of outbred rats and its progenitors. Scientific data **1**: 140011.

Baud, A., R. Hermsen, V. Guryev, P. Stridh, D. Graham, *et al.*, 2013 Combined sequence-based and genetic mapping analysis of complex traits in outbred rats. Nature genetics **45**: 767–75.

Benjamini, Y., Y. Hochberg, R. S. Society, Y. Benajmini, Y. Hochberg, *et al.*, 1995 Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society, Series B **57**: 289–300.

Braccini, L., E. Ciraolo, C. C. Campa, A. Perino, D. L. Longo, *et al.*, 2015 PI3K-C2γ is a Rab5 effector selectively controlling endosomal Akt2 activation downstream of insulin signalling. Nature Communications **6**: 7400.

Buenrostro, J. D., P. G. Giresi, L. C. Zaba, H. Y. Chang, and W. J. Greenleaf, 2013 Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nature methods **10**: 1213–8.

Buenrostro, J. D., B. Wu, H. Y. Chang, and W. J. Greenleaf, 2015 ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. In *Current Protocols in Molecular Biology*, volume 2015, pp. 21.29.1–21.29.9, John Wiley & Sons, Inc., Hoboken, NJ, USA.

Chick, J. M., S. C. Munger, P. Simecek, E. L. Huttlin, K. Choi, *et al.*, 2016 Defining the consequences of genetic variation on a proteome-wide scale. Nature **534**: 500–5.

Churchill, G., D. Gatti, S. Munger, and K. Svenson, 2012 The Diversity outbred mouse population. Mammalian Genome **23**: 713–8.

Churchill, G. A., D. C. Airey, H. Allayee, J. M. Angel, A. D. Attie, *et al.*, 2004 The Collaborative Cross, a community resource for the genetic analysis of complex traits. Nature Genetics **36**: 1133–1137.

Collaborative Cross Consortium, 2012 The genome architecture of the Collaborative Cross mouse genetic reference population. Genetics **190**: 389–401.

Conway, J. R., A. Lex, and N. Gehlenborg, 2017 UpSetR: an R package for the visualization of intersecting sets and their properties. Bioinformatics **33**: 2938–2940.

Crowley, J. J., V. Zhabotynsky, W. Sun, S. Huang, I. K. Pakatci, *et al.*, 2015 Analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance. Nature genetics **47**: 353–60.

Degner, J. F., A. A. Pai, R. Pique-Regi, J.-B. Veyrieras, D. J. Gaffney, *et al.*, 2012 DNase I sensitivity QTLs are a major determinant of human expression variation. Nature **482**: 390–4.

Dempster, A., N. Laird, and D. B. Rubin, 1977 Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society Series B Methodological **39**: 1–38.

Didelez, V. and N. Sheehan, 2007 Mendelian randomization as an instrumental variable approach to causal inference. Statistical methods in medical research **16**: 309–30.

Doerge, R. and G. Churchill, 1996 Permutation tests for multiple loci affecting a quantitative character. Genetics **142**: 285–94.

Donoghue, L. J., A. Livraghi-Butrico, K. M. McFadden, J. M. Thomas, G. Chen, *et al.*, 2017 Identification of trans Protein QTL for Secreted Airway Mucins in Mice and a Causal Role for Bpifb1. Genetics **207**: 801–812.

Dudbridge, F. and B. P. Koeleman, 2004 Efficient Computation of Significance Levels for Multiple Associations in Large Studies of Correlated Data, Including Genomewide Association Studies. The American Journal of Human Genetics **75**: 424–435.

Flutre, T., X. Wen, J. Pritchard, and M. Stephens, 2013 A statistical framework for joint eQTL analysis in multiple tissues. PLoS genetics **9**: e1003486.

Fu, C.-P., C. E. Welsh, F. P.-M. de Villena, and L. McMillan, 2012a Inferring ancestry in admixed populations using microarray probe intensities. In *Proceedings of the ACM Conference on Bioin-*

formatics, Computational Biology and Biomedicine - BCB '12, pp. 105–112, New York, New York, USA, ACM Press.

Fu, J., M. G. M. Wolfs, P. Deelen, H.-J. Westra, R. S. N. Fehrmann, *et al.*, 2012b Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. PLoS genetics **8**: e1002431.

Gatti, D. M., K. L. Svenson, A. Shabalin, L.-Y. Wu, W. Valdar, *et al.*, 2014 Quantitative Trait Locus Mapping Methods for Diversity Outbred Mice. G3 (Bethesda, Md.) **4**: 1623–1633.

Gu, Z., L. Gu, R. Eils, M. Schlesner, and B. Brors, 2014 circlize implements and enhances circular visualization in R. Bioinformatics **30**: 2811–2812.

Haley, C. S. and S. A. Knott, 1992 A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity **69**: 315–24.

Hamilton-Williams, E. E., D. B. Rainbow, J. Cheung, M. Christensen, P. A. Lyons, *et al.*, 2013 Fine mapping of type 1 diabetes regions Idd9.1 and Idd9.2 reveals genetic complexity. Mammalian genome : official journal of the International Mammalian Genome Society **24**: 358–75.

Hamilton-Williams, E. E., S. B. J. Wong, X. Martinez, D. B. Rainbow, K. M. Hunter, *et al.*, 2010 Idd9.2 and Idd9.3 Protective Alleles Function in CD4+ T-Cells and Nonlymphoid Cells to Prevent Expansion of Pathogenic Islet-Specific CD8+ T-Cells. Diabetes **59**: 1478–1486.

Huang, G.-J., S. Shifman, W. Valdar, M. Johannesson, B. Yalcin, *et al.*, 2009 High resolution mapping of expression QTLs in heterogeneous stock mice in multiple tissues. Genome Research **19**: 1133–1140.

Jansen, R., J.-J. Hottenga, M. G. Nivard, A. Abdellaoui, B. Laport, *et al.*, 2017 Conditional eQTL analysis reveals allelic heterogeneity of gene expression. Human molecular genetics **26**: 1444–1451.

Kaeppler, S. M., 1997 Quantitative trait locus mapping using sets of near-isogenic lines: Relative power comparisons and technical considerations. Theoretical and Applied Genetics **95**: 384–392.

Keele, G. R., W. L. Crouse, S. N. P. Kelada, and W. Valdar, 2018 Determinants of qtl mapping power in the realized collaborative cross. bioRxiv .

Kelada, S. N. P., 2016 Plethysmography Phenotype QTL in Mice Before and After Allergen Sensitization and Challenge. G3 (Bethesda, Md.) **6**: 2857–2865.

Kelada, S. N. P., D. E. Carpenter, D. L. Aylor, P. Chines, H. Rutledge, *et al.*, 2014 Integrative genetic analysis of allergic inflammation in the murine lung. American journal of respiratory cell and molecular biology **51**: 436–45.

Keller, M. P., D. M. Gatti, K. L. Schueler, M. E. Rabaglia, D. S. Stapleton, *et al.*, 2018 Genetic Drivers of Pancreatic Islet Function. Genetics **209**: 335–356.

Klemm, S. L., Z. Shipony, and W. J. Greenleaf, 2019 Chromatin accessibility and the regulatory epigenome. Nature reviews. Genetics p. 1.

Lander, E. S. and D. Botstein, 1989 Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics **121**: 185–99.

Lander, E. S. and P. Green, 1987 Construction of multilocus genetic linkage maps in humans. Proceedings of the National Academy of Sciences **84**: 2363–2367.

Lawrence, M., R. Gentleman, and V. Carey, 2009 rtracklayer: an r package for interfacing with genome browsers. Bioinformatics **25**: 1841–1842.

Liu, E. Y., Q. Zhang, L. McMillan, F. P.-M. de Villena, and W. Wang, 2010 Efficient genome ancestry inference in complex pedigrees with inbreeding. Bioinformatics **26**: i199–i207.

MacKinnon, D. P., A. J. Fairchild, and M. S. Fritz, 2007 Mediation analysis. Annual review of psychology **58**: 593–614.

Martínez, O. and R. N. Curnow, 1992 Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. Theor. Appl. Genet. **85**: 480–488.

Morgan, A. P., C.-P. Fu, C.-Y. Kao, C. E. Welsh, J. P. Didion, *et al.*, 2016 The Mouse Universal Genotyping Array: From Substrains to Subspecies. G3: Genes, Genomes, Genetics **6**: 263–279.

Mosedale, M., Y. Kim, W. J. Brock, S. E. Roth, T. Wiltshire, *et al.*, 2017 Candidate Risk Factors and Mechanisms for Tolvaptan-Induced Liver Injury Are Identified Using a Collaborative Cross Approach. Toxicological Sciences **156**: kfw269.

Mott, R., C. J. Talbot, M. G. Turri, A. C. Collins, and J. Flint, 2000 A method for fine mapping quantitative trait loci in outbred animal stocks. PNAS **97**: 12649–54.

Oreper, D., Y. Cai, L. M. Tarantino, F. P.-M. de Villena, and W. Valdar, 2017 Inbred Strain Variant Database (ISVdb): A Repository for Probabilistically Informed Sequence Differences Among the Collaborative Cross Strains and Their Founders. G3 (Bethesda, Md.) **7**: 1623–1630.

Pai, A. A., J. K. Pritchard, and Y. Gilad, 2015 The genetic and mechanistic basis for variation in gene regulation. PLoS genetics **11**: e1004857.

Quinlan, A. R. and I. M. Hall, 2010 BEDTools : a flexible suite of utilities for comparing genomic features **26**: 841–842.

R Core Team, 2019 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rieseberg, L. H., M. A. Archer, and R. K. Wayne, 1999 Transgressive segregation, adaptation and speciation. Heredity **83**: 363–372.

Ritchie, M. E., B. Phipson, D. Wu, Y. Hu, C. W. Law, *et al.*, 2015 limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research **43**: e47.

Robinson, G., 1991 That BLUP is a Good Thing: The Estimation of Random Effects. Statistical Science **6**: 15–51.

Robinson, M. D., D. J. McCarthy, and G. K. Smyth, 2010 edger: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics **26**: 139–140.

Roytman, M., G. Kichaev, A. Gusev, and B. Pasaniuc, 2018 Methods for fine-mapping with chromatin and expression data. PLoS genetics **14**: e1007240.

Rutledge, H., D. L. Aylor, D. E. Carpenter, B. C. Peck, P. Chines, *et al.*, 2014 Genetic regulation of Zfp30, CXCL1, and neutrophilic inflammation in murine lung. Genetics **198**: 735–45.

Sarkar, S. K. and C.-K. Chang, 1997 The Simes Method for Multiple Hypothesis Testing With Positively Dependent Test Statistics. Journal of the American Statistical Association **92**: 1601.

Schadt, E. E., 2009 Molecular networks as sensors and drivers of common human diseases. Nature **461**: 218–223.

Schaid, D. J., W. Chen, and N. B. Larson, 2018 From genome-wide associations to candidate causal variants by statistical fine-mapping. Nature Reviews Genetics **19**: 491–504.

Shibata, Y., N. C. Sheffield, O. Fedrigo, C. C. Babbitt, M. Wortham, *et al.*, 2012 Extensive evolutionary changes in regulatory element activity during human origins are associated with altered gene expression and positive selection. PLoS

genetics **8**: e1002789.

Skelly, D. A., A. Czechanski, C. Byers, S. Aydin, C. Spruce, *et al.*, 2019 Genetic variation influences pluripotent ground state stability in mouse embryonic stem cells through a hierarchy of molecular phenotypes. bioRxiv .

Srivastava, A., A. P. Morgan, M. L. Najarian, V. K. Sarsani, J. S. Sigmon, *et al.*, 2017 Genomes of the mouse collaborative cross. Genetics **206**: 537–556.

Storey, J. D. and R. Tibshirani, 2003 Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences of the United States of America **100**: 9440–5.

Sul, J. H., B. Han, C. Ye, T. Choi, and E. Eskin, 2013 Effectively Identifying eQTLs from Multiple Tissues by Combining Mixed Model and Meta-analytic Approaches. PLoS Genetics **9**: e1003491.

Svenson, K. L., D. M. Gatti, W. Valdar, C. E. Welsh, R. Cheng, *et al.*, 2012 High-resolution genetic mapping using the Mouse Diversity outbred population. Genetics **190**: 437–47.

Valdar, W., J. Flint, and R. Mott, 2006a Simulating the collaborative cross: power of quantitative trait loci detection and mapping resolution in large sets of recombinant inbred strains of mice. Genetics **172**: 1783–97.

Valdar, W., C. C. Holmes, R. Mott, and J. Flint, 2009 Mapping in structured populations by resample model averaging. Genetics **182**: 1263–77.

Valdar, W., L. C. Solberg, D. Gauguier, S. Burnett, P. Klenerman, *et al.*, 2006b Genome-wide genetic association of complex traits in heterogeneous stock mice. Nature Genetics **38**: 879–887.

Wei, J. and S. Xu, 2016 A random-model approach to QTL mapping in multiparent advanced generation intercross (MAGIC) populations. Genetics **202**: 471–486.

Wu, Y., J. Zeng, F. Zhang, Z. Zhu, T. Qi, *et al.*, 2018 Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits. Nature Communications **9**: 918.

Xiong, Q., S. Mukherjee, and T. S. Furey, 2014 GSAASeqSP: a toolset for gene set association analysis of RNA-Seq data. Scientific reports **4**: 6347.

Yang, H., J. R. Wang, J. P. Didion, R. J. Buus, T. A. Bell, *et al.*, 2011 Subspecific origin and haplotype diversity in the laboratory mouse. Nature genetics **43**: 648–55.

Zhang, Z., W. Wang, and W. Valdar, 2014 Bayesian modeling of haplotype effects in multiparent populations. Genetics **198**: 139–56.

Zheng, C., M. P. Boer, and F. A. van Eeuwijk, 2015 Reconstruction of Genome Ancestry Blocks in Multiparental Populations. Genetics **200**: 1073–1087.

Appendix A: Significance thresholds

Thresholds of significance in the context of genome-wide scans for a large number of traits need to account for the heavy multiple testing burden. In order to efficiently detect QTL, we use three procedures to define statistical significance over varying stringency, allowing greater leniency for putative proximal genetic regulators. These procedures address the family-wide error rate (FWER) and false discovery rate (FDR) differently.

At the level of a single trait, FWER is the probability of a false positive across all tests that are performed. The FWER is controlled through adjustment of an observed -$\log_{10}$ p-value (logP) through 1,000 permutations of the trait, producing an FWER-adjusted p-value (permP). This adjustment can be performed for genome-wide and more lenient chromosome-wide significance.

To account for multiple testing over many traits, an FDR adjustment (Benjamini *et al.* 1995) can be used to produce q-values, which are interpreted as the probability that a detected QTL is a false discovery.

Family-wide error rate

The family-wide error rate (FWER) is the probability of detecting a single false positive across all genome-wide tests. Controlling FWER instead of the single test type I error rate (α) is necessary because FWER increases with the number independent tests, reaching 1 for $\alpha = 0.05$ in the context of genome-wide tests. Controlling FWER is more conservative than controlling the false discovery rate (FDR) for multiple testing correction, and appropriate for genome scans of a single trait because there are not expected to be an excess number of QTL per outcome.

Based on simulations of the breeding design (Valdar *et al.* 2006a), the CC panel was expected to be balanced in terms of founder haplotype contributions. This balance was realized to large extent (Srivastava *et al.* 2017), though the PWK and CAST founders have reduced contributions in comparison to the other founders. Keele *et al.* (2018) performed simulations from the realized CC strain genomes to examine QTL mapping power, and found low levels of population structure in the CC genomes, violating the exchangeability assumption made by the permutation procedure, which characterizes a null distribution for the test statistic(Doerge and Churchill 1996). Nevertheless, we use a permutation approach, given that we expect the presence of strong effect QTL for the molecular traits that are detectable in the presence of subtle population structure.

Specifically, 1000 permutations of the sample identities were taken, and then genome scans performed on each permutation of each trait. For each trait, the minimum p-value from each permutation genome scan is recorded, transformed to the -$\log_{10}$ scale (logP), and then used fit a null extreme value distribution (EVD) (Dudbridge and Koeleman 2004; Valdar *et al.* 2006a). Genome-wide permutation-based p-values (permP) are obtained by calculating the probability of a more extreme logP than the observed from the cumulative distribution function of the EVD.

False discovery rate

Analyzing a large numbers of traits, as is the case with gene expression and chromatin accessibility, also contributes significantly to a multiple testing burden. Not only is an association for a trait being tested at positions across the genome, but the whole process is repeated per trait. The expected number of significant tests is also different globally across all the traits, with many traits expected to have QTL in comparison to there being few QTL for a single trait.

An FDR procedure can address this multiple testing dynamic, accounting for the additional testing due to genome-wide traits while being more lenient than strict FWER control across traits. As in Chick *et al.* (2016), an FDR procedure (Benjamini *et al.* 1995; Storey and Tibshirani 2003) is applied to the per-trait FWER-adjusted permP, producing q-values. QTL are detected if q-value $\leq \alpha_{\text{FDR}}$, where α_{FDR} is the specified false discovery rate. QTL were called using both $\alpha_{\text{FDR}} = 0.1$ and $\alpha_{\text{FDR}} = 0.2$.

Appendix B: Mapping procedures

Method 1: Multi-stage conditional regression with genome-wide FWER and FDR

The combined per-trait FWER control and across-trait FDR control restrict the mapping procedure's ability to detect multiple QTL per outcome in a statistically valid manner because the EVD used for FWER adjustment is fit from the maximum statistical score per trait. To include additional strong statistical scores is potentially biasing, as would be the case when there are multiple associations above the FWER threshold, increasing the enrichment in small p-values. This inflations in small p-values is precisely the pattern of association the FDR procedure detects. To avoid this problem, a multi-stage conditional fitting approach is used (Jansen *et al.* 2017). The procedure is described in the following steps:

1. For a given trait, conduct genome scan according to the model described in Eq 3.

2. Perform genome scans for the permuted trait to characterize an EVD. Calculate FWER-adjusted permP from the observed maximum logP of the genome scan of the trait. The permP is stored to be used as input to an FDR method.

3. Specify a genome-wide α_{step} for determining whether subsequent conditional scans should be conducted for the trait. We set $\alpha_{\text{step}} = 0.1$. If the permP $> \alpha_{\text{step}}$, no further conditional scans are conducted.

4. If permP $\leq \alpha_{\text{step}}$, steps 1-3 are repeated for additional conditional scans of the outcome. For $j > 1$, j^{th} conditional scan use the same form of alternative and null model as described in Eq 3, except with the inclusion of locus effects from the peak associations from previous stages. Generally, the alternative model for conditional stage scan J follows as:
$$y_i = \mu + \text{QTL}_i + \sum_{j=1}^{J-1} \text{locus}_i^j + \text{batch}_i + \varepsilon_i, \qquad (7)$$
with locus_i^j representing the locus effect of the peak association from the jth stage scan of the outcome for individual i, and is also included in the null model of conditional scans. Now steps 2-4 are repeated.

A multi-stage conditional scan approach could have issues with over-fitting 47 data points per trait, given that each QTL and locus effect represents the estimation of seven fixed effects. However, the permutations were found to appropriately compensate for this potential in the recalculation of the EVD based on permutations of the conditional scans in step 2. **Figure S4** provides a clear example in which the gene *Gpn3* has a local-eQTL detected after a strong distal-eQTL is conditioned for within the model in lung tissue. This multi-stage conditional procedure is referred to as Method 1.

Method 2: Single-stage regression with chromosome-wide FWER and FDR control

Given that the data represent 47 individual mice, there may be poor power to detect genome-wide significant QTL. Because of the strong biological prior evidence for local-QTL, significant genetic associations that meet chromosome-wide significance for the local chromosome of the trait were also detected and recorded. This was accomplished by fitting a chromosome-wide EVD, and thus producing chromosome-wide FWER-adjusted permP for each trait. Chromosome-wide results were restricted to the first stage of the multi-stage procedure, thus referred to as a single-stage analysis, and allowing only a single chromosome-wide permP per trait. The same FDR then used as in Method 1, resulting in chromosome-wide q-values. This more lenient single-stage procedure is referred to as Method 2.

Method 3: Single-stage regression with genome-wide and chromosome-wide FWER control for local-QTL

A final local-QTL exclusive single step analysis is used. Instead of calculating genome-wide and chromosome-wide permP from the overall maximum logP, FWER-controlled permP are adjusted from the maximum logP within the proximal window of the trait from the first stage scans from Method 1. An FDR procedure is not used, in part because the permP are likely to be poorly distributed with respect to the expectations of FDR. A local-QTL is detected when permP < 0.05. This procedure, now referred to as Method 3, is less conservative for local-QTL because there is no FDR control, instead leveraging the strong biological precedent for proximal genetic regulation. It also does not leniently detect intra-chromosomal distal-QTL, as with Method 2.

Appendix C: Mediation analysis

Baron and Kenny (1986) give a description of the relationships that need be tested to declare mediation. For this study, simple models that consist of three variables are used: the independent variable X, the mediator M, and the dependent variable Y (MacKinnon *et al.* 2007). Superscripts are used to indicate whether a variable is the QTL, chromatin accessibility at site k (C_k), or expression of gene j (G_j) in the model. When an eQTL is detected, the following relationship is supported:

$$X^{QTL} \quad \rightarrow \quad Y^{G_j} \tag{8}$$

When cQTL or proximal eQTL to distal-eQTL are detected, a similar relationship is detected:

$$X^{QTL} \quad \rightarrow \quad M \tag{9}$$

The previous traits (Y) can also be viewed as potential mediators (M). Here we consider two models of mediation: 1) the effect of an eQTL is mediated through local chromatin state (M^{C_k}) and 2) the effect of the distal-eQTL is mediated through a proximal gene (M^{G_k}), depicted in **Figure 2**. Co-localizing QTL are consistent with the co-occurrence of $X \rightarrow Y^{G_j}$ and $X \rightarrow Y^{C_k}$ at a locus. A true mediation relationship has an additional requirement, specifically that:

$$X^{QTL} \quad \perp\!\!\!\perp \quad Y^{G_j} \quad | \quad M \tag{10}$$

where "A $\perp\!\!\!\perp$ B" indicates that A and B are independent of each other, and "A | B" is A conditioned on variable B. X being fully independent of Y given M is consistent with X acting on Y completely through M, referred to as full mediation. In practice, the true relationship can be complex and obscured by noise, instead resulting in partial mediation, whereby X affects Y both directly and through M. Genome-wide data present a significant challenge to systematically testing these relationships.

Mediation models

Mediation of local-eQTL through chromatin state. Mediation through chromatin accessibility was evaluated for all local-eQTL detected through the lenient Method 3, leveraging prior knowledge of the predominance of local genetic regulation and the effects of proximal chromatin state. Candidate mediators are detected at genome and chromosome-wide significance through 1,000 permutations to define EVD for FWER adjustment.

Mediation of distal-eQTL through proximal gene expression. Mediation of the effect of genome-wide significant distal-eQTL through proximal gene expression was evaluated. Genome-wide significance is required for distal-eQTL because there is not additional strong biological information to support increased leniency as is the case with proximal signal. As with mediation through chromatin state, 1,000 permutations are used to define mediation significance thresholds at the genome-wide level.

Mediation of cQTL through other chromatin sites and gene expression were also evaluated, but did not result in the detection of any candidate mediators.

Mediation significance thresholds

Empirical significance thresholds for the mediation scans can be defined through permutations, similarly to as is done with the QTL scans. Instead of permuting the trait outcome, the pairing of trait to QTL is maintained, and the mediator is permuted. EVD are then fit for potential mediators from the minimum logP from permutation scans, instead of the maximum logP in QTL scans. FWER-adjusted mediation permP are extracted from the EVD, which can be calculated in terms of genome-wide and chromosome-wide significance. An FDR adjustment is not used with mediation because it is only evaluated for detected QTL, which could potentially violate the uniform null distribution expectations of FDR.

Formal mediation criterion

Detection of mediation has some caveats and pitfalls compared to the statistical analyses used for QTL mapping, in part due to the fact that the relationships 8 and 9 could be complex, whereas the directionality of the relationships for QTL is fixed because the genetic state at the QTL cannot change due because of the trait. This does not hold for the mediation model, where Y could in fact causally influence M ($Y \rightarrow M$). Didelez and Sheehan (2007) note that false mediation signals can be detected when the association between the M and Y are particularly strong, whereby M acts as a surrogate for X. Based on concerns over fitting relatively complex models to data comprising 47 observations, we use stringent requirements for declaring evidence for mediation:

1. Detection of a significant eQTL (relationship 8).
   (a) Mediation through chromatin: chromosome-wide permP < 0.05 from Method 3
   (b) Mediation through expression: genome-wide q-value < 0.1 from Method 1
2. Detection of significant mediator within a 20Mb window centered around the eQTL from previous step.

3. Detection of a significant QTL for the mediator from previous step.

   (a) Mediation through chromatin: chromosome-wide permP < 0.05 from Method 3 and within 10Mb upstream or downstream of gene TSS

   (b) Mediation through expression: genome-wide q-value < 0.1 from Method 1 and within 10Mb upstream or downstream of the eQTL position

4. It is not possible to distinguish $M \rightarrow Y$ from $M \leftarrow Y$ given these data. If M is an intermediary of X on Y, the expectation is that the effect of $X \rightarrow M$ is greater than $X \rightarrow Y$. We check that the effect size of the QTL for step 1 ($X \rightarrow Y$) is less than the effect size of the QTL from step 3 ($X \rightarrow M$), based on Eq 4. Failure of this step does not disprove the proposed mediation model, but it could suggest that the relationship between M and Y is more complicated than the simple models tested here, or that greater noise is incurred in the measurement of M than Y.