Transcripts with high distal heritability mediate genetic effects on complex metabolic traits

Anna L. Tyler[1], J. Matthew Mahoney[1], Mark P. Keller[2], Candice N. Baker[1], Margaret Gaca[1], Anuj Srivastava[1], Isabela Gerdes Gyuricza[1], Madeleine J. Braun[1], Nadia A. Rosenthal[1], Alan D. Attie[2], Gary A. Churchill[1] and Gregory W. Carter[1]

[1]The Jackson Laboratory, Bar Harbor, Maine, USA. [2]University of Wisconsin-Madison, Biochemistry Department, Madison, Wisconsin, USA.

Abstract

Although many genes are subject to local regulation, recent evidence suggests that complex distal regulation may be more important in mediating phenotypic variability. To assess the role of distal gene regulation in complex traits, we combine multi-tissue transcriptomes with physiological outcomes to model diet-induced obesity and metabolic disease in a population of Diversity Outbred mice. Using a novel high-dimensional mediation analysis, we identify a composite transcriptome signature that summarizes genetic effects on gene expression and explains 30% of the variation across all metabolic traits. The signature is heritable, interpretable in biological terms, and predicts obesity status from gene expression in an independently derived mouse cohort and multiple human studies. Transcripts contributing most strongly to this composite mediator frequently have complex, distal regulation distributed throughout the genome. These results suggest that trait-relevant variation in transcription is largely distally regulated, but is nonetheless identifiable, interpretable, and translatable across species.

Introduction

Evidence from genome-wide association studies (GWAS) suggests that most heritable variation in complex traits is mediated through regulation of gene expression. The majority of trait-associated variants lie in gene regulatory regions[1–7], suggesting a relatively simple causal model in which a variant alters the homeostatic expression level of a nearby (local) gene which, in turn, alters a trait. Statistical methods such

as transcriptome-wide association studies (TWAS)[8–11] and summary data-based Mendelian randomization (SMR)[10] have used this idea to identify genes associated with multiple disease traits[12–15]. However, despite the great promise of these methods, explaining trait effects with local gene regulation has been more difficult than initially assumed[16;17]. Although trait-associated variants typically lie in non-coding, regulatory regions, these variants often have no detectable effects on gene expression[16] and tend not to co-localize with expression quantitative trait loci (eQTLs)[17;18]. These observations suggest that the relationship among genetic variants, gene expression, and organism-level traits is more complex than the simple, local model.

In recent years the conversation around the genetic architecture of common disease traits has been addressing this complexity, and there is increased interest in more distant (distal) genetic effects as potential drivers of trait variation[18–20;15;21]. In general, distal effects are defined as being greater than 4 or 5Mb away from the transcription start site of a given gene. We use the terms local and distal rather than *cis* and *trans* because *cis* and *trans* have specific biochemical meanings[22], whereas local and distal are defined only by genomic position. The importance of distal genetic effects is proposed in the omnigenic model, which posits that trait-driving genes are cumulatively influenced by many distal variants. In this view, the heritable transcriptomic signatures driving clinical traits are an emergent state arising from the myriad molecular interactions defining and constraining gene expression. Consistent with this view, it has been suggested that part of the difficulty in explaining trait variation through local eQTLs may arise in part because gene expression is not measured in the appropriate cell types[16], or cell states[23], and thus local eQTLs influencing traits cannot be detected in bulk tissue samples. This context dependence emphasizes the essential role of complex regulatory and tissue networks in mediating variant effects. The mechanistic dissection of complex traits in this model is more challenging because it requires addressing network-mediated effects that are weaker and greater in number. However, the comparative importance of distal effects over local effects is currently only conjectured and extremely challenging to address in human populations.

To assess the role of wide-spread distal gene regulation in the genetic architecture of complex traits, we used genetically diverse mice as a model system. In mice we can obtain simultaneous measurements of the genome, transcriptome, and phenome in all individuals. We used diet-induced obesity and metabolic disease as an archetypal example of a complex trait. In humans, these phenotypes are genetically complex with hundreds of variants mapped through GWAS[24;25] that are known to act through multiple tissues[26;27]. Likewise in mice, metabolic traits are also genetically complex[28] and synteny analysis implicates a high degree of concordance in the genetic architecture between species[28;12]. Furthermore, in contrast to humans, in mice we have access to multiple disease-relevant tissues in the same individuals with sufficient numbers for adequate statistical power.

We generated two complementary data sets: a discovery data set in a large population of Diversity Outbred (DO) mice[29], and an independent validation data set derived by crossing inbred strains from the Collaborative Cross (CC) recombinant inbred lines[30] to form CC recombinant inbred intercross (CC-RIX) mice. Both populations were maintained on a high-fat, high-sugar diet to model diet-induced obesity and metabolic disease[12].

The DO population and CC recombinant inbred lines were derived from the same eight inbred founder strains: five classical lab strains and three strains more recently derived from wild mice[29], representing three subspecies and capturing 90% of the known variation in laboratory mice[31]. The DO mice are maintained with a breeding scheme that ensures equal contributions from each founder across the genome thus rendering almost the whole genome visible to genetic inquiry and maximizing power to detect eQTLs[29]. The CC mice were initially intercrossed to recombine the genomes from all eight founders, and then inbred for at least 20 generations to create recombinant inbred lines[30;32;31]. Because these two populations have common ancestral haplotypes but highly distinct kinship structure, we could directly and unambiguously compare the local genetic effects on gene expression at the whole-transcriptome level while varying the population structure driving distal regulation.

In the DO population, we paired clinically relevant metabolic traits, including body weight and plasma levels of insulin, glucose and lipids[12], with transcriptome-wide gene expression in four tissues related to metabolic disease: adipose tissue, pancreatic islets, liver, and skeletal muscle. We measured similar metabolic traits in a CC-RIX population and gene expression from three of the four tissues used in the DO: adipose tissue, liver, and skeletal muscle. Measuring gene expression in multiple tissues is critical to adequately assess the extent to which local gene regulation varies across the tissues and whether such variability might account for previous failed attempts to identify trait-relevant local eQTLs. The CC-RIX carry the same founder alleles as the DO. Thus, local gene regulation is expected to match between the populations. However, because the alleles are recombined throughout the genome, distal effects are expected to vary from those in the DO, allowing us to directly assess the role of distal gene regulation in driving trait-associated transcript variation. To mechanistically dissect distal effects on metabolic disease, we developed a novel dimension reduction framework called high-dimensional mediation analysis (HDMA) to identify the heritable transcriptomic signatures driving trait variation, which we compared between mouse populations and to human data sets with measured adipose gene expression. Together, these data enable a comprehensive view into the genetic architecture of metabolic disease.

Results

Genetic variation contributed to wide phenotypic variation

Although the environment was consistent across the DO mice, the genetic diversity present in this population resulted in widely varying distributions across physiological measurements (Fig. 1). For example, body weights of adult individuals varied from less than the average adult C57BL/6J (B6) body weight to several times the body weight of a B6 adult in both sexes (Males: 18.5 - 69.1g, Females: 16.0 - 54.8g) (Fig. 1A). Fasting blood glucose (FBG) also varied considerably (Fig. 1B), although few of the animals had FBG levels that would indicate pre-diabetes (19 animals, 3.8%), or diabetes (7 animals, 1.4%) according to previously developed cutoffs (pre-diabetes: FBG $\geq$ 250 mg/dL, diabetes: FBG $\geq$ 300, mg/dL)[33]. Males had higher FBG than females on average (Fig. 1C) as has been observed before suggesting either that males were more susceptible to metabolic disease on the high-fat, high-sugar (HFHS) diet, or that males and females may require different thresholds for pre-diabetes and diabetes.

Body weight was strongly positively correlated with food consumption (Fig. 1D $R^2 = 0.51$, $p < 2.2 \times 10^{-16}$) and FBG (Fig. 1E, $R^2 = 0.21$, $p < 2.2 \times 10^{-16}$) suggesting a link between behavioral factors and metabolic disease. However, the heritability of this trait and others (Fig. 1F) indicates that genetics contribute substantially to correlates of metabolic disease in this population.

The trait correlations (Fig. 1G) showed that most of the metabolic trait pairs were only modestly correlated, which, in conjunction with the trait decomposition (Supplementary Figure 1), suggests complex relationships among the measured traits and a broad sampling of multiple heritable aspects of metabolic disease including overall body weight, glucose homeostasis, and pancreatic function.

Distal Heritability Correlated with Phenotype Relevance

It is widely assumed that variation in traits is mediated through local regulation of gene expression. To test this assumption, we measured transcriptome-wide gene expression in four tissues–adipose, liver, pancreatic islet, and skeletal muscle–in the DO cohort. (Basic results from a standard eQTL analysis[34] (Methods) are available in Supplementary Figure 2). We estimated the local genetic contribution to each transcript as the variance explained by the haplotype probabilities at the genetic marker closest to the gene transcription start site. We estimated the distal heritability as the heritability of the residuals after local haplotype had been accounted for (Methods). Importantly, this estimate was not based on distal eQTL, but rather the unlocalized contribution of the genome after removing the local genetic effect.

Overall, local and distal genetic factors contributed approximately equally to transcript abundance. In all

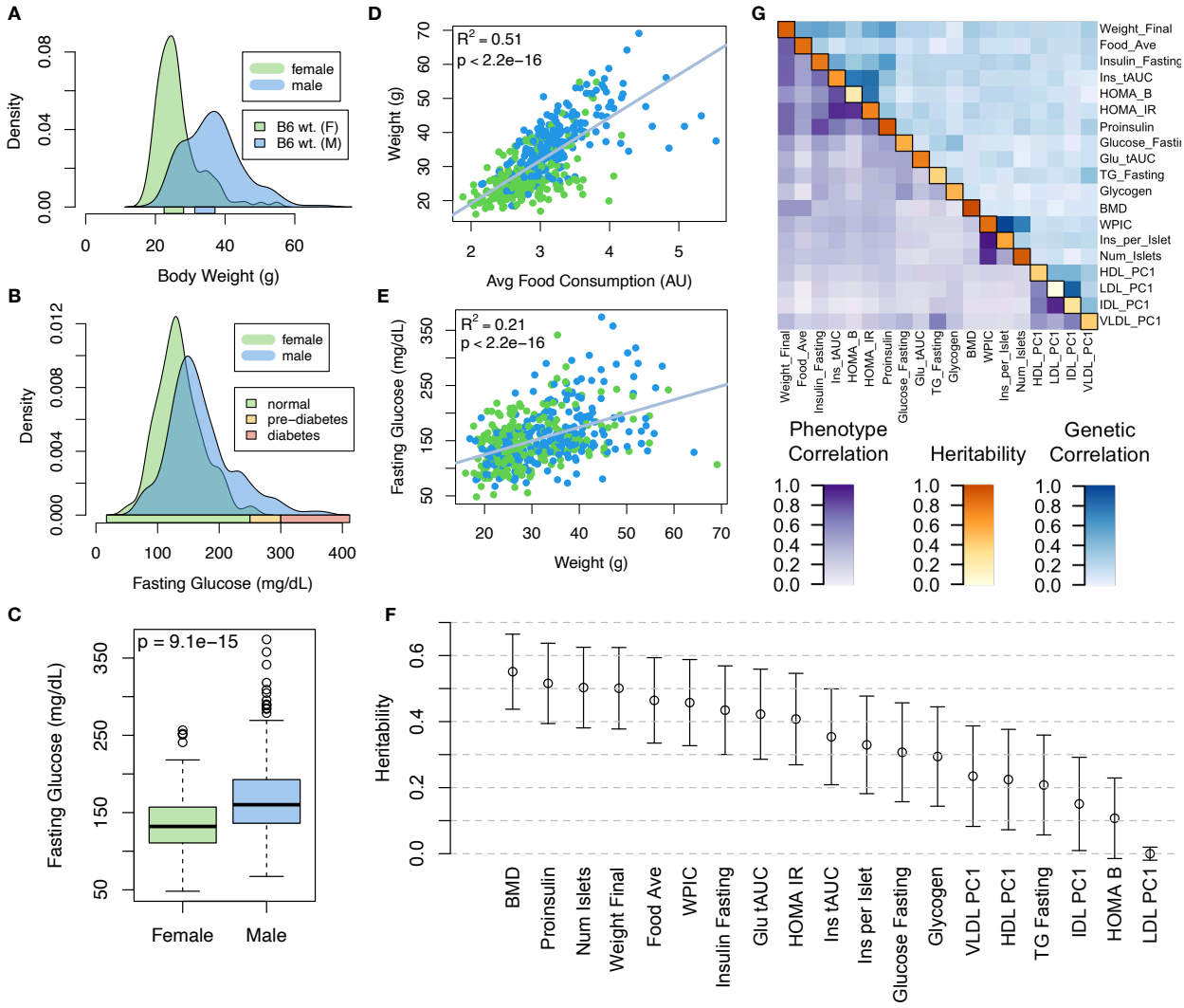


Figure 1: Clinical overview. **A.** Distributions of final body weight in the diversity outbred mice. Sex is indicated by color. The average B6 male and female adult weights at 24 weeks of age are indicated by blue and green bars on the x-axis. **B.** The distribution of final fasting glucose across the population split by sex. Normal, pre-diabetic, and diabetic fasting glucose levels for mice are shown by colored bars along the x-axis. **C.** Males had higher fasting blood glucose on average than females ($p = 9.1 \times 10^{-15}$). **D.** The relationship between food consumption and body weight for both sexes. **E.** Relationship between body weight and fasting glucose for both sexes. **F.** Heritability estimates for each physiological trait. Bars show standard error of the estimate. **G.** Correlation structure between pairs of physiological traits. The lower triangle shows Pearson correlation coefficients between pairs of traits (r). The upper triangle shows the Pearson correlation coefficient (r) between LOD traces of pairs of traits, and diagonal shows the estimated heritability of each trait. BMD - bone mineral density, WPIC - whole pancreas insulin content, Glu tAUC - glucose total area under the curve, HOMA IR - homeostatic measurement of insulin resistance, HOMA B - homeostatic measure of beta cell health, VLDL - very low-density lipoprotein, LDL - low-density lipoprotein, IDL - intermediate density lipoprotein, HDL - high-density lipoprotein, TG - triglyceride.

tissues, both local and distal factors explained between 8 and 17% of the variance in the median transcript (Fig, 2A). This 50% contribution of local genetic variation to transcript abundance contrasts with findings

in humans in which local variants have been found to explain only 20-30% of total heritability, while distal effects explain the remaining 70-80%[35;36]. This discrepancy may arise due to the high degree of linkage disequilibrium in the DO mice compared to human populations and to the high degree of confidence with which we can estimate ancestral haplotypes in this population. At each position in the mice we can estimate ancestral haplotype with a high degree of accuracy. Haplotype at any given genetic marker captures genomic information from a relatively large genomic region surrounding each marker. In contrast, there is a much higher degree of recombination in human populations and ancestral haplotypes are more numerous and more difficult to estimate than in the mice. Thus in the mice, each marker may capture more local regulatory variation than SNPs or estimated haplotypes capture in humans. It has been found that transcripts with muliple local eQTL have higher local heritability than transcripts with single local eQTL[37]. Because of the high diversity in the DO and the high rates of linkage disequilibrium, it is possible that there are more local variants regulating transcription creating a proportionally larger effect of local regulation.

To assess the importance of genetic regulation of transcript levels to clinical traits, we compared the local and distal heritabilities of transcripts to their trait relevance. We defined trait relevance for a transcript as its maximum absolute Spearman correlation coefficient (ρ) across all traits (Methods). The local heritability of transcripts was negatively associated with their trait relevance (Fig. 2B), suggesting that the more local genotype influenced transcript abundance, the less effect this variation had on the measured traits. Conversely, the distal heritability of transcripts was positively associated with trait relevance (Fig. 2C). That is, transcripts that were more highly correlated with the measured traits tended to be distally, rather than locally, heritable. Importantly, this pattern was consistent across all tissues. This finding is also consistent with previous observations that transcripts with low local heritability explain more expression-mediated disease heritability than transcripts with high local heritability[19]. However, the positive relationship between trait correlation and distal heritability demonstrated further that there are diffuse genetic effects throughout the genome converging on trait-related transcripts.

High-Dimensional Mediation Analysis identified a high-heritability composite trait that was mediated by a composite transcript

The above univariate analyses establish the importance of distal heritability for trait-relevant transcripts. However, the number of transcripts dramatically exceeds the number of phenotypes. Thus, we expect the heritable, trait-relevant transcripts to be highly correlated and organized according to coherent, biological processes representing the mediating endophenotypes driving clinical trait variation. To identify these endophenotypes in a theoretically principled way, we developed a novel dimension-reduction technique,

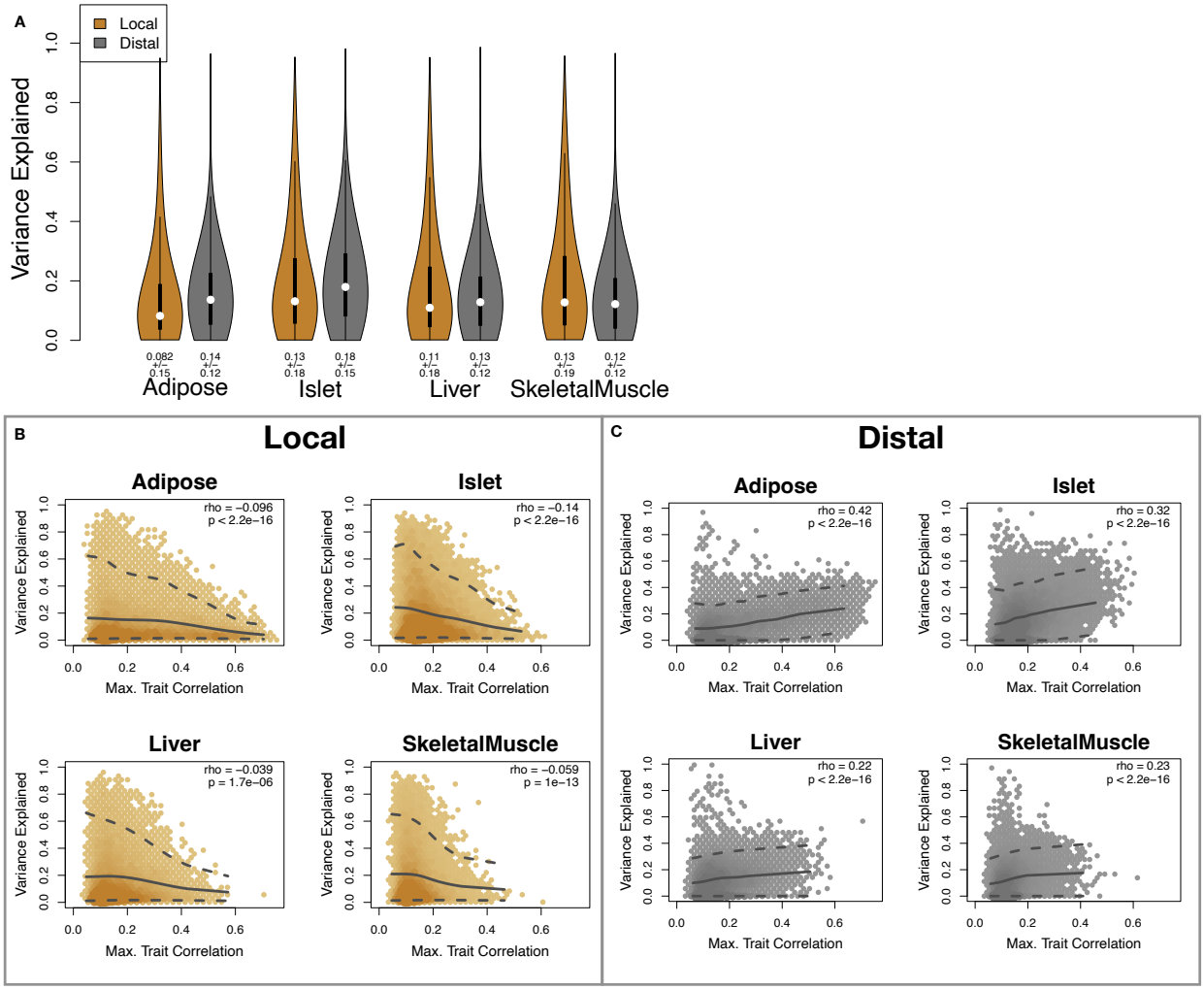


Figure 2: Transcript heritability and trait relevance. **A.** Distributions of local and distal heritability of transcripts across the four tissues. Overall local and distal factors contributed equally to transcript heritability. The relationship between (**B.**) local and (**C.**) distal heritability and trait relevance across all four tissues. Here trait relevance is defined as the maximum correlation between the transcript and all traits. The upper and lower dashed line in each panel show the 95th and 5th percentile correlation. The solid line shows the mean trait correlation in transcripts with increasing variance explained either locally (B) or distally (C). Transcripts that are highly correlated with traits tended to have low local heritability and high distal heritability.

high-dimension mediation analysis (HDMA), that uses the theory of causal graphical models to identify a transcriptomic signature that is simultaneously 1) highly heritable, 2) strongly correlated to the measured phenotypes, and 3) conforms to the causal mediation hypothesis (Fig. 3). In HDMA, we first use a linear mapping called kernelization to dimension-reduce the genome, transcriptome, and phenome to kernel matrices G_K, T_K and P_K respectively, which each have the dimensions n by n where n is the number of individuals (Methods). These kernel matrices describe the relationships among the individual mice in genome space, transcriptome space, and phenome space and ensure that these three omic spaces have the same dimensions, and thus the same weight in the analysis. If not dimension-reduced, the transcriptome would outweigh the

phenome in the model. We then projected these $n \times n$-dimensional kernel matrices onto one-dimensional scores–a composite genome score (G_C), a composite transcriptome score (T_C), and a composite phenome score (P_C)–and used the univariate theory of mediation to constrain these projections to satisfy the hypotheses of perfect mediation, namely that upon controlling for the transcriptomic score, the genome score is uncorrelated to the phenome score. A complete mathematical derivation and implementation details for HDMA are available in the Methods.

Using HDMA we identifed the major axis of variation in the transcriptome that was consistent with mediating the effects of the genome on metabolic traits (Fig 3). Fig. 3A shows the partial correlations (ρ) between the pairs of these composite vectors. The partial correlation between G_C and T_C was 0.42, and the partial correlation between T_C and P_C was 0.78. However, when the transcriptome was taken into account, the partial correlation between G_C and P_C was effectively zero (0.039). P_C captured 30% of the overall trait variance, and its estimated heritability was 0.71 ± 0.084, which was higher than any of the measured traits (Fig. 1F). Thus, HDMA identified a maximally heritable metabolic composite trait and a highly heritable component of the transcriptome that are correlated as expected in the perfect mediation model.

As discussed in the Methods, HDMA is related to a generalized form of canonical correlation analysis (CCA). Standard CCA is prone to over-fitting because in any two large matrices it can be trivial to identify highly correlated composite vectors[38]. To assess whether our implementation of HDMA was similarly prone to over-fitting in a high-dimensional space, we performed permutation testing. We permuted the individual labels on the transcriptome matrix 10,000 times and recalculated the path coefficient, which is the correlation of G_C and T_C multiplied by the correlation of T_C and P_C. This represents the strength of the path from G_C to P_C that is putatively mediated through T_C. The permutations preserved the correlation between the genome and phenome, but broke the correlations between the genome and the transcriptome, as well as between the transcriptome and the phenome. We could thus test whether, given a random transcriptome, HDMA would overfit and identify apparently mediating transcriptomic signatures in random data. The null distribution of the path coefficient is shown in Fig. 3B, and the observed path coefficient from the original data is indicated by a red arrow. The observed path coefficient was well outside the null distribution generated by permutations ($p < 10^{-16}$). Fig. 3C illustrates this observation in more detail. Although we identified high correlations between G_C and T_C, and modest correlations between T_C and P_C in the null data (Fig 3C), these two values could not be maximized simultaneously in the null data. In contrast, the red dot shows that in the real data both the G_C-T_C correlation and the T_C-P_C correlation could be maximized simultaneously suggesting that the path from genotype to phenotype through the transcriptome is highly non-trivial and identifiable in this case.

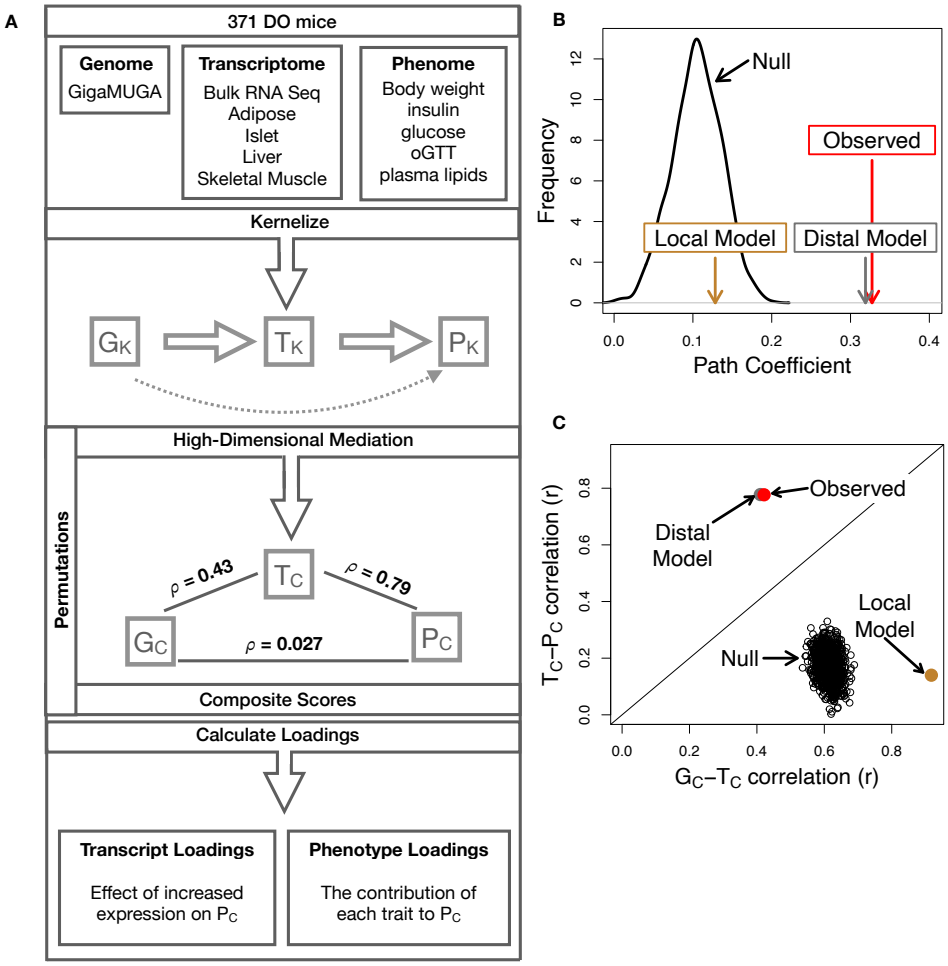


Figure 3: High-dimensional mediation. **A.** Workflow indicating major steps of high-dimensional mediation. The genotype, transcriptome, and phenotype matrices were kernelized to yield single matrices representing the relationships between all individuals for each data modality (G_K = genome kernel, T_K = transcriptome kernel; P_K = phenome kernel). High-dimensional mediation was applied to these matrices to maximize the direct path $G \rightarrow T \rightarrow P$, the mediating pathway (arrows), while simultaneously minimizing the direct $G \rightarrow P$ pathway (dotted line). The composite vectors that resulted from high-dimensional mediation were G_c, T_C, and P_C. The partial correlations ρ between these vectors indicated perfect mediation. Transcript and trait loadings were calculated as described in the methods. **B.** The null distribution of the path coefficient derived from 10,000 permutations. The observed model (red) is compared to models derived from exclusively distal (gray) or local genetic effects (brown). The similarity of the observed and distal models indicates the full model is dominated by distal genetic effects. **C.** The null distribution of the G_C-T_C correlation vs. the T_C-P_C correlation. Comparisons are shown to the observed values (red), and those derived from the distal-only model (gray) and the local-only model (brown).

To test whether the presence of local eQTLs affected the result, we generated two additional transcriptomic kernel matrices. To generate a "distal kernel" we regressed out the effect of local haplotype from each transcript and calculated the kernel from only distally regulated transcription. We generated a "local kernel" using only locally determined gene expression and a "distal kernel" using only distally determined gene expression, i.e. the effects of local haplotype were regressed out. The path coefficient identified using the

local kernel was not significantly different from the null (Fig. 3B), suggesting that locally determined gene expression does not mediate the effects of the genome on the phenome. In contrast, the path coefficient identified using the distal kernel, was highly significant and indistinguishable from that identified using the full transcriptome.

Further, the G_C-T_C and T_C-P_C correlations derived from the distal kernal were indistinguishable from those derived from the original transcriptomic kernel. In contrast, the G_C-T_C correlation derived with the local kernel was extremely high, reflecting the fact that the local transcriptomic kernel was derived directly from local haplotypes. The T_C-P_C correlation, however, was very low (0.14), suggesting that the these locally derived transcripts were not highly related to phenotype. In other words, mice that shared many local eQTL were not highly similar in trait space. Taken together, these results suggest that composite vectors derived from the measured transcriptomic kernel represent genetically determined variation in phenotype that is mediated through genetically determined variation in transcription, and that this genetically deremined variation in transcription is largely driven by distal factors.

Body weight and insulin resistance were highly represented in the expression-mediated composite trait

Each composite score is a weighted combination of the measured variables. The magnitude and sign of the weights, called loadings, correspond to the relative importance and directionality of each variable in the composite score. The loadings of each measured trait onto P_C indicate how much each contributed to the composite phenotype. Body weight contributed the most (Fig. 4), followed by homeostatic insulin resistance (HOMA_IR) and fasting plasma insulin levels (Insulin_Fasting). We can thus interpret P_C as an index of metabolic disease (Fig. 4B). Individuals with high values of P_C have a higher metabolic disease index (MDI) and greater metabolic disease, including higher body weight and higher insulin resistance. We refer to P_C as MDI going forward. Traits contributing the least to MDI were measures of cholesterol and pancreas composition. Thus, when we interpret the transcriptomic signature identified by HDMA, we are explaining primarily the putative transcriptional mediation of body weight and insulin resistance, as opposed to cholesterol measurements.

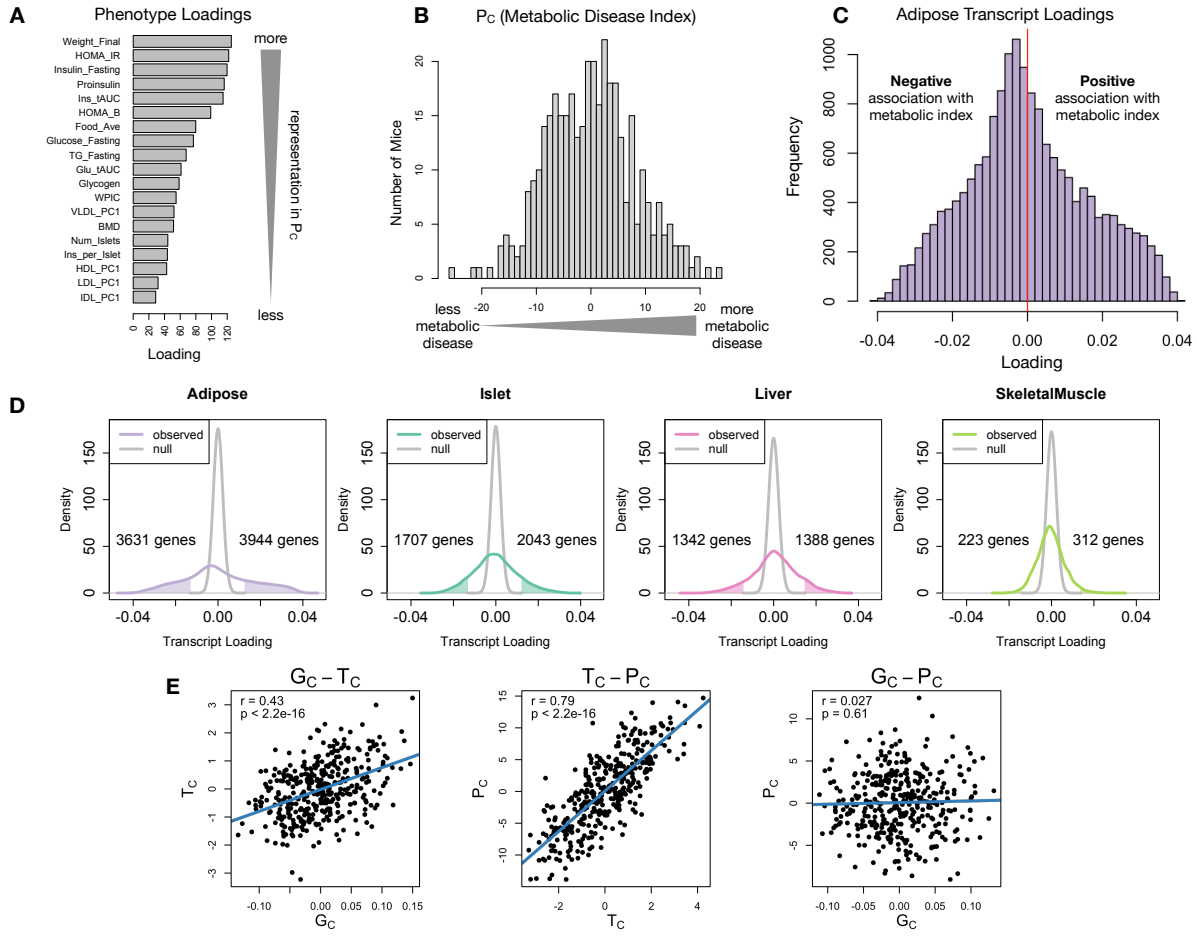


Figure 4: Interpretation of loadings. **A.** Loadings across traits. Body weight and insulin resistance contributed the most to the composite trait. **B.** Phenotype scores across individuals. Individuals with large positive phenotype scores had higher body weight and insulin resistance than average. Individuals with large negative phenotype scores had lower body weight and insulin resistance than average. **C.** Distribution of transcript loadings in adipose tissue. For transcripts with large positive loadings, higher expression was associated with higher phenotype scores. For transcripts with large negative loadings, higher expression was associated with lower phenotype scores. **D.** Distributions of loadings across tissues compared to null distributions. Shaded areas represent loadings that were more extreme than the null distribution. Numbers indicate how many transcripts had loadings above and below the extremes of the null. Transcripts in adipose tissue had the most extreme loadings indicating that transcripts in adipose tissue were the best mediators of the genetic effects on body weight and insulin resistance. **E.** Scatter plots showing correlations between composite vectors for the genome (G_C), the transcriptome (T_C), and the phenome (P_C). The G_C - T_C correlation is high, the T_C - P_C correlation is high, and there is no significant correlation between G_C and P_C. This correlation structure is consistent with perfect mediation.

High-loading transcripts had low local heritability, high distal heritability, and were linked mechanistically to obesity

We interpreted large loadings onto transcripts as indicating strong mediation of the effect of genetics on MDI. Large positive loadings indicate that higher expression was associated with a higher MDI (i.e. higher

risk of obesity and metabolic disease on the HFHS diet) (Fig. 4C-D). Conversely, large negative loadings indicate that high expression of these transcripts was associated with a lower MDI (i.e. lower risk of obesity and metabolic disease on the HFHS diet) (Fig. 4C-D). Fig. 4D compares the observed transcript loading distributions to null distributions and indicates how many transcripts in each tissue had large positive and negative loadings. A direct comparison of the tissues can be seen in Supplementary Figure 3. We used gene set enrichment analysis (GSEA)[39;40] to look for biological processes and pathways that were enriched at the top and bottom of this list (Methods).

In adipose tissue, both GO processes and KEGG pathway enrichments pointed to an axis of inflammation and metabolism (Figs. 4 and 5). GO terms and KEGG pathways associated with inflammation were positively associated with MDI, indicating that increased expression in inflammatory pathways was associated with a higher burden of disease. It is well established that adipose tissue in obese individuals is inflamed and infiltrated by macrophages[41–45], and the results here suggest that this may be a dominant heritable component of metabolic disease.

The strongest negative enrichments in adipose tissue were related to mitochondial activity in general, and thermogenesis in particular (Figs. 4 and 5). Genes in the KEGG oxidative phosphorylation pathway were almost universally negatively loaded in adipose tissue, suggesting that increased expression of these genes was associated with reduced MDI (Supplementary Figure 6). Consistent with this observation, it has been shown previously that mouse strains with greater thermogenic potential are also less susceptible to obesity on an obesigenic diet[46].

Transcripts associated with the citric acid cycle as well as the catabolism of the branched-chain amino acids (valine, leuceine, and isoleucine) were strongly enriched with negative loadings in adipose tissue (Supplementary Figures 4, 7 and 8). Expression of genes in both pathways (for which there is some overlap) has been previously associated with insulin sensitivity[12;47;48], suggesting that heritable variation in regulation of these pathways may influence risk of insulin resistance.

Looking at the 10 most positively and negatively loaded transcripts from each tissue, it is apparent that transcripts in the adipose tissue had the largest loadings, both positive and negative (Fig. 5A bar plot). This suggests that much of the effect of genetics on body weight and insulin reisistance is mediated through gene expression in adipose tissue. This finding does not speak to the relative importance of tissues not included in this study, such as brain, in which transcriptional variation may mediate a large portion of the genetic effect on obesity. The strongest loadings in liver and pancreas were comparable, and those in skeletal muscle were the weakest (Fig. 5A), suggesting that less of the genetic effects were mediated through transcription

in skeletal muscle. As expected, heritability analysis showed that transcripts with the largest loadings had higher distal heritability than local heritability (Fig. 5A heat map and box plot). We also performed TWAS in this population by imputing transcript levels for each gene based on local genotype only and correlating the imputed transcript levels with each trait. In contrast to HDMA, the TWAS procedure tended to nominate transcripts with lower loadings (Fig. 5B), higher local heritability and lower distal heritability. Finally, we focused on transcripts with the highest local heritability in each tissue (Fig. 5C). This procedure selected transcripts with low loadings on average, consistent with our findings above (Fig. 2B).

We performed a literature search for the genes in each of these groups along with the terms "diabetes", "obesity", and the name of the expressing tissue to determine whether any of these genes had previous associations with metabolic disease in the literature (Methods). Multiple genes in each group had been previously associated with obesity and diabetes (Fig. 5 bolded gene names). Genes with high loadings were most highly enriched for previous literature support. They were 2.2 times more likely than TWAS hits and 4 times more likely than genes with high local heritability to be previously associated with obesity or diabetes.

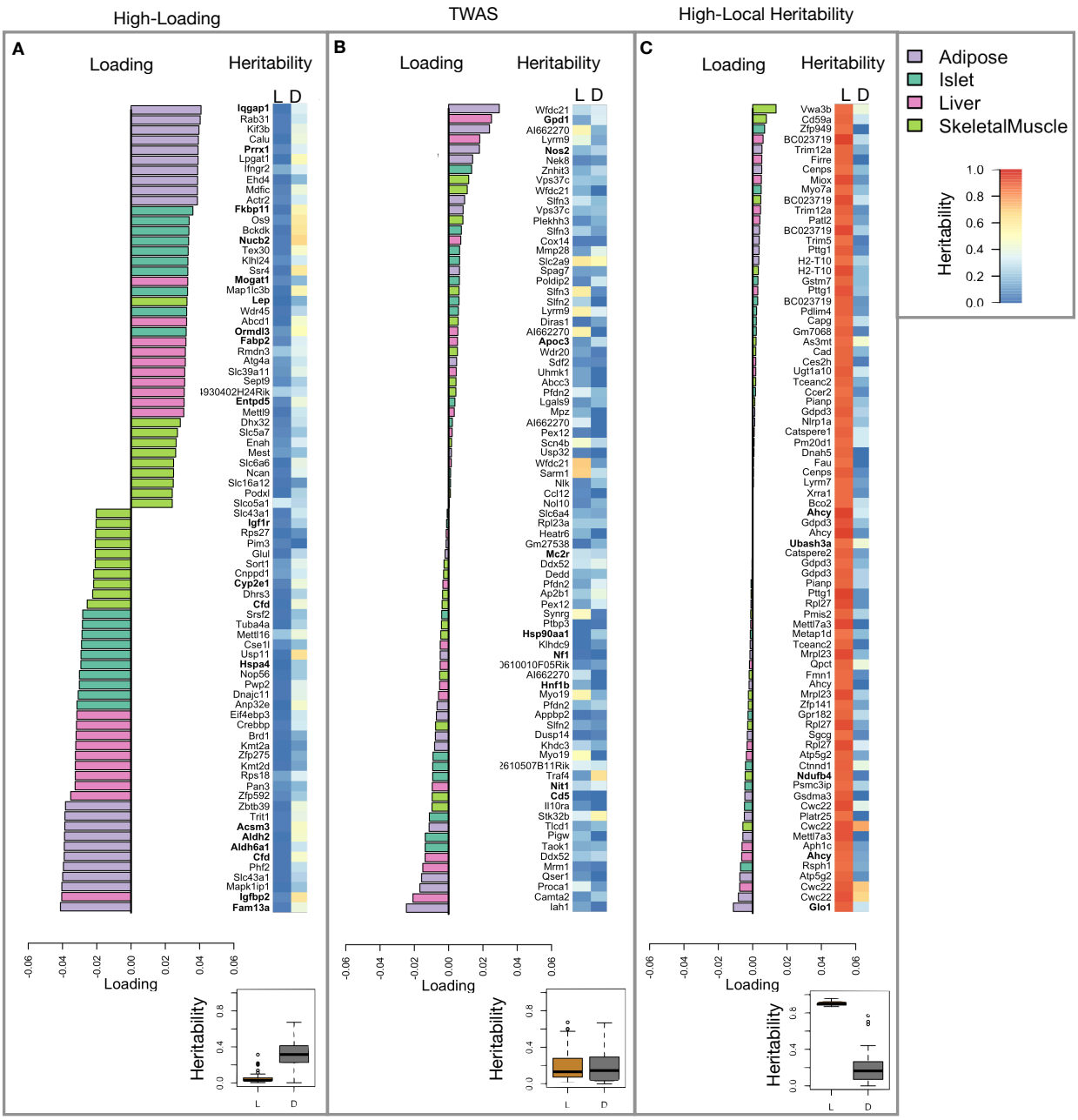


Figure 5: Transcripts with high loadings have high distal heritability and literature support (bolded gene names). Each panel has a bar plot showing the loadings of transcripts selected by different criteria. Bar color indicates the tissue of origin. The heat map shows the local (L - left) and distal (D - right) heritability of each transcript. **A.** Loadings for the 10 transcripts with the largest positive loadings and the 10 transcripts with the largest negative loadings for each tissue. Distal heritability was significantly higher than local heritability (t-test $p < 2.2^{-16}$). **B.** Loadings of TWAS candidates with the 10 largest positive correlations with traits and the largest negative correlations with traits across all four tissues. Local and distal heritability were not significantly different for this group (t-test $p = 0.77$). **C.** The transcripts with the largest local heritability (top 20) across all four tissues. Local heritability was significantly higher than distal heritability of these genes (t-test $p < 2.2^{-16}$)

Tissue-specific transriptional programs were associated with metabolic traits

Clustering of transcripts with top loadings in each tissue showed tissue-specific functional modules associated with obesity and insulin resistance (Fig. 6A) (Methods). The clustering highlights the importance of immune activation particularly in adipose tissue. The "mitosis" cluster had large positive loadings in three of the four tissues potentially suggesting system-wide proliferation of immune cells. Otherwise, all clusters were strongly loaded in only one or two tissues. For example, the lipid metabolism cluster was loaded most heavily in liver. The positive loadings suggest that high expression of these genes, particularly in the liver, was associated with increased metabolic disease. This cluster included the gene *Pparg*, whose primary role is in the adipose tissue where it is considered a master regulator of adipogenesis [49]. Agonists of *Pparg*, such as thiazolidinediones, are FDA-approved to treat type II diabetes, and reduce inflammation and adipose hyptertrophy [49]. Consistent with this role, the loading for *Pparg* in adipose tissue was negative, suggesting that higher expression was associated with leaner mice (Fig. 6B). In contrast, *Pparg* had a large positive loading in liver, where it is known to play a role in the development of hepatic steatosis, or fatty liver. Mice that lack *Pparg* specifically in the liver, are protected from developing steatosis and show reduced expression of lipogenic genes [50;51]. Overexpression of *Pparg* in the livers of mice with a *Ppara* knockout, causes upregulation of genes involved in adipogenesis [52]. In the livers of both mice and humans high *Pparg* expression is associated with hepatocytes that accumulate large lipid droplets and have gene expression profiles similar to that of adipocytes [53;54]. The local and distal heritability of *Pparg* is low in adipose tissue suggesting its expression in this tissue is highly constrained in the population (Fig. 6B). However, the distal heritability of *Pparg* in liver is relatively high suggesting it is complexly regulated and has sufficient variation in this population to drive variation in phenotype. Both local and distal heribatility of *Pparg* in the islet are relatively high, but the loading is low, suggesting that variability of expression in the islet does not drive variation in MDI. These results highlight the importance of tissue context when investigating the role of heritable transcript variability in driving phenotype. Gene lists for all clusters are available in Supplementary File 1.

Gene expression, but not local eQTLs, predicted body weight in an independent population

To test whether the transcript loadings identified in the DO could be translated to another population, we tested whether they could predict metabolic phenotypes in an independent population of CC-RIX mice, which were F1 mice derived from multiple pairings of Collaborative Cross (CC) [55;32;56;57] strains (Fig. 7) (Methods). We tested two questions. First, we asked whether the loadings identified in the DO mice were relevant to the relationship between the transcriptome and the phenome in the CC-RIX. We predicted body weight (a surrogate for MDI) in each CC-RIX individual using measured gene expression in each tissue and the

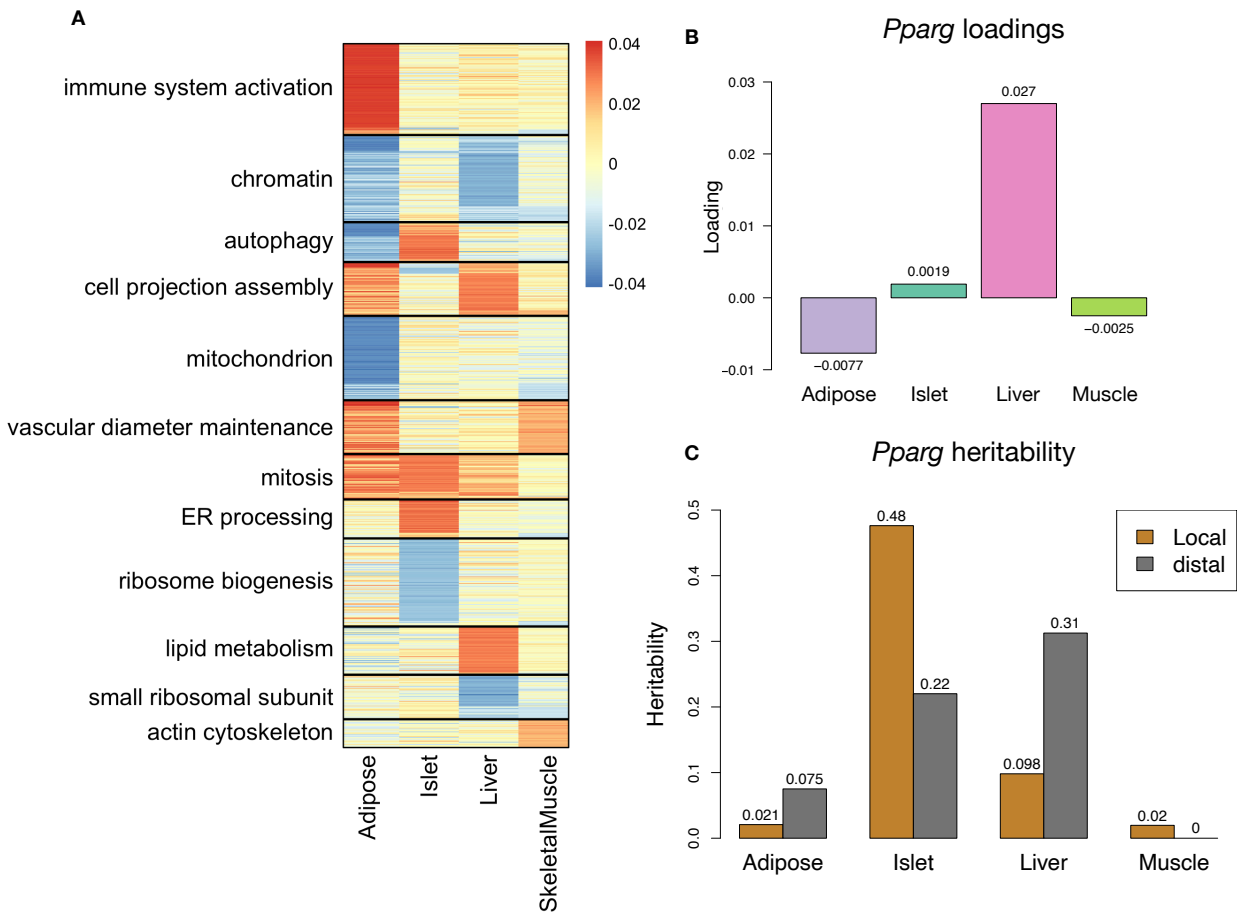


Figure 6: Tissue-specific transcriptional programs were associated with obesity and insulin resistance. **A** Heat map showing the loadings of all transcripts with loadings greater than 2.5 standard deviations from the mean in any tissue. The heat map was clustered using k medoid clustering. Functional enrichments of each cluster are indicated along the left margin. **B** Loadings for *Pparg* in different tissues. **C** Local and distal of *Pparg* expression in different tissues.

transcript loadings identified in the DO (Methods). The predicted body weight and acutal body weight were highly correlated (Fig. 7B left column). The best prediction was achieved for adipose tissue, which supports the observation in the DO that adipose expression was the strongest mediator of the genetic effect on MDI. This result also confirms the validity and translatability of the transcript loadings and their relationship to metabolic disease.

The second question related to the source of the relevant variation in gene expression. If local regulation was the predominant factor influencing trait-relevant gene expression, we should be able to predict phenotype in the CC-RIX using transcripts imputed from local genotype (Fig. 7A). The DO and the CC-RIX were derived from the same eight founder strains and so carry the same alleles throughout the genome. We imputed gene expression in the CC-RIX using local genotype and were able to estimate variation in gene transcription

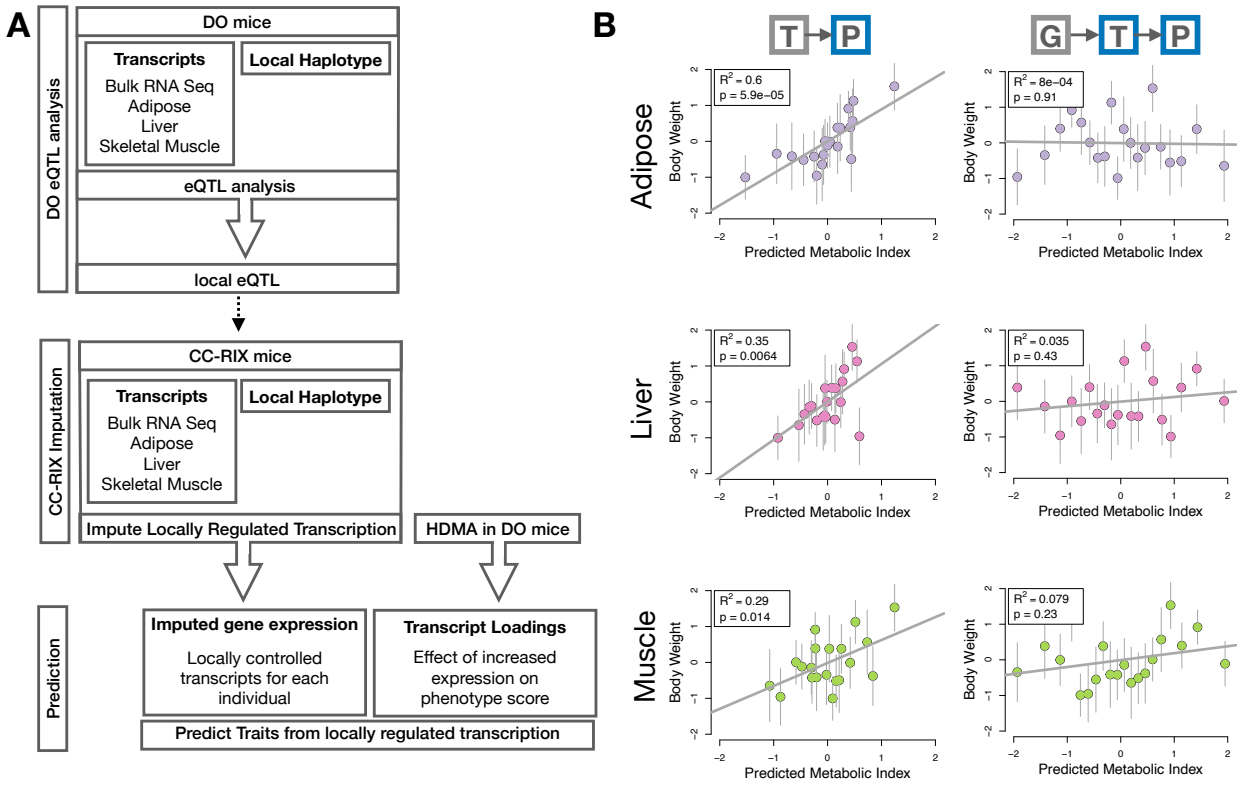


Figure 7: Transcription, but not local genotype, predicts phenotype in the CC-RIX. **A.** Workflow showing procedure for translating HDMA results to an independent population of mice. **B.** Relationships between the predicted metabolic disease index (MDI) and measured body weight in the CC-RIX. The left column shows the predictions using measured transcripts. The right column shows the prediction using transcript levels imputed from local genotype. Gray boxes indicate measured quantities, and blue boxes indicate calculated quantities. The dots in each panel represent individual CC-RIX strains. The gray lines show the standard deviation on body weight for the strain.

robustly (Supplementary Figure 9). However, these imputed values failed to predict body weight in the CC-RIX when weighted with the loadings from HDMA. (Fig. 7B right column). This result suggests that local regulation of gene expression is not the primary factor driving heritability of complex traits. It is also consistent with our findings in the DO population that distal heritability was a major driver of trait-relevant gene expression and that high-loading transcripts had comparatively high distal and low local heritability.

Distally heritable transcriptomic signatures reflected variation in composition of adipose tissue and islets

The interpretation of global genetic influences on gene expression and phenotype is potentially more challenging than the interpretation and translation of local genetic influences, as genetic effects cannot be localized to individual gene variants or transcripts. However, there are global patterns across the loadings that can inform mechanism. For example, heritable variation in cell type composition can be inferred from transcript loadings.

We observed above that immune activation in the adipose tissue was a highly enriched process correlating with obesity in the DO population. In humans, it has been extensively observed that macrophage infiltration in adipose tissue is a marker of obesity and metabolic disease[58]. To determine whether the immune activation reflected a heritable change in cell composition in adipose tissue in DO mice, we compared loadings of cell-type specific genes in adipose tissue (Methods). The mean loading of macrophage-specific genes was significantly greater than 0 (Holm-adjusted two-sided empirical $p < 2 \times 10^{-16}$) (Fig. 8A), indicating that obese mice were genetically predisposed to have high levels of macrophage infiltration in adipose tissue in response to the HFHS diet. Loadings for marker genes for other cell types were not statistically different from zero (Adipocytes: $p = 0.08$, Progenitors: $p = 0.58$, Leukocytes: $p = 0.28$; all Holm-adjusted two-sided empirical p), indicating that changes in the abundance of those cell types was not a mediator of MDI.

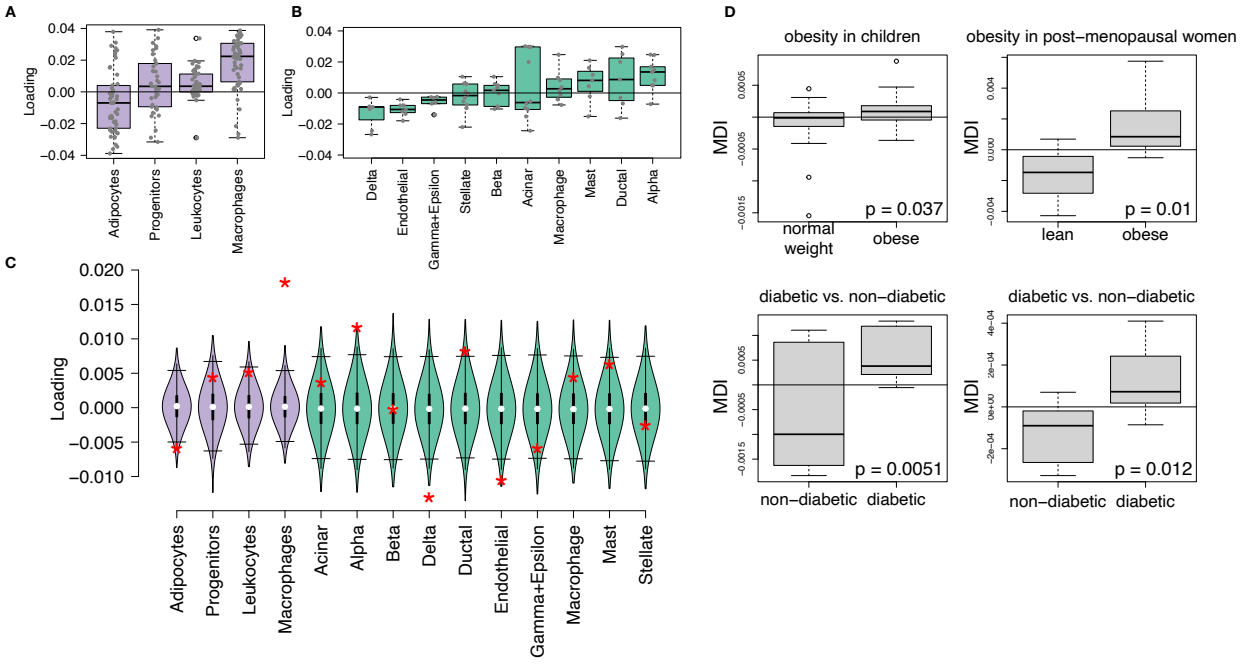


Figure 8: HDMA results translate to humans. **A.** Distribution of loadings for cell-type-specific transcripts in adipose tissue. **B.** Distribution of loadings for cell-type-specific transcripts in pancreatic islets. **C.** Null distributions for the mean loading of randomly selected transcripts in each cell type compared with the observed mean loading of each group of transcripts (red asterisk). **D.** Predictions of metabolic phenotypes in four adipose transcription data sets downloaded from GEO. In each study the obese/diabetic patients were predicted to have greater MDI than the lean/non-diabetic patients based on the HDMA results from DO mice.

We also compared loadings of cell-type specific transcripts in islet (Methods). The mean loadings for alpha-cell specific transcripts were significantly greater than 0 ($p = 0.002$), while the mean loadings for delta- (Holm-adjusted two-sided empirial $p < 2 \times 10^{-16}$) and endothelial-cell (Holm-adjusted two-sided empirical $p = 0.01$) specific genes were significantly less than 0 (Fig. 8B). These results suggest that mice with higher MDI

inherited an altered cell composition that predisposed them to metabolic disease, or that these compositional changes were induced by the HFHS diet in a heritable way. In either case, these results support the hypothesis that alterations in islet composition drive variation in MDI. Notably, the mean loading for pancreatic beta cell marker transcripts was not significantly different from zero (Holm-adjusted two-sided empirical $p = 0.95$). We stress that this is not necessarily reflective of the function of the beta cells in the obese mice, but rather suggests that any variation in the number of beta cells in these mice was unrelated to obesity and insulin resistance, the major contributors to MDI. This is further consistent with the islet composition traits having small loadings in the phenome score (Fig. 4).

Heritable transcriptomic signatures translated to human disease

Ultimately, the heritable transcriptomic signatures that we identified in DO mice will be useful if they inform mechanism and treatment of human disease. To investigate the potential for translation of the gene signatures identified in DO mice, we compared them to transcriptional profiles in obese and non-obese human subjects (Methods). We limited our analysis to adipose tissue because the adipose tissue signature had the strongest relationship to obesity and insulin resistance in the DO.

We calculated a predicted MDI for each individual in the human studies based on their adipose tissue gene expression (Methods) and compared the predicted scores for obese and non-obese groups as well as diabetic and non-diabetic groups. In all cases, the predicted MDIs were higher on average for individuals in the obese and diabetic groups compared with the lean and non-diabetic groups (Fig. 8D). This indicates that the distally heritable signature of MDI identified in DO mice is relevant to obesity and diabetes in human subjects.

Existing therapies are predicted to target mediator gene signatures

Another application of the transcript loading landscape is in ranking potential drug candidates for the treatment of metabolic disease. Although high-loading transcripts may be good candidates for understanding specific biology related to obesity, the transcriptome overall is highly interconnected and redundant. The ConnectivityMap (CMAP) database[59;60] developed by the Broad Institute allows querying thousands of compounds that reverse or enhance the extreme ends of transcriptomic signatures in multiple different cell types. By identifying drugs that reverse pathogenic transcriptomic signatures, we can potentially identify compounds that have favorable effects on gene expression. To test this hypothesis, we queried the CMAP database through the CLUE online query tool (https://clue.io/query/, version 1.1.1.43) (Methods). We identified top anti-correlated hits across all cell types (Supplementary Figures 10 and 11). To get more

tissue-specific results, we also looked at top results in cell types that most closely resembled our tissues. We looked at results in adipocytes (ASC) as well as pancreatic tumor cells (YAPC) regardless of p value (Supplementary Figures 12 and 13).

The CMAP database identified both known diabetes drugs (e.g. sulfonylureas), as well as drugs that target pathways known to be involved in diabetes pathogenesis (e.g. mTOR inhibitors). These findings help support the mediation model we fit here. Although the composite variables we identified here are consistent with mediation, they do not prove causality. However, the results from CMAP suggest that reversing the transcriptomic signatures we found also reverses metabolic disease phenotypes, which supports a causal role of the transcript levels in driving pathogenesis of metabolic disease. These results thus support the mediation model we identified here and its translation to therapies in human disease.

Discussion

Here we investigated the relative contributions of local and distal gene regulation in four tissues to heritable variation in traits related to metabolic disease in genetically diverse mice. We found that distal heritability was positively correlated with trait relatedness, whereas high local heritability was negatively correlated with trait relatedness. We used a novel high-dimensional mediation analysis (HDMA) to identify tissue-specific composite transcripts that are predicted to mediate the effect of genetic background on metabolic traits. The adipose-derived composite transcript robustly predicted body weight in an independent cohort of diverse mice with disparate population structure. It also predicted MDI in four human cohorts. However, gene expression imputed from local genotype failed to predict body weight in the second mouse population. Taken together, these results highlight the complexity of gene expression regulation in relation to trait heritability and suggest that heritable trait variation is mediated primarily through distal gene regulation.

Our result that distal regulation accounted for most trait-related gene expression differences is consistent with a complex model of genetic trait determination. It has frequently been assumed that gene regulation in *cis* is the primary driver of genetically associated trait variation, but attempts to use local gene regulation to explain phenotypic variation have had limited success[16;17]. In recent years, evidence has mounted that distal gene regulation may be an important mediator of trait heritability[19;18;61;62]. It has been observed that transcripts with high local heritability explain less expression-mediated disease heritability than those with low local heritability[19]. Consistent with this observation, genes located near GWAS hits tend to be complexly regulated[18]. They also tend to be enriched with functional annotations, in contrast to genes with simple local regulation, which tend to be depleted of functional annotations suggesting they are less likely to be directly involved in disease traits[18]. These observations are consistent with principles of robustness

in complex systems in which simple regulation of important elements leads to fragility of the system[63–65]. Our results are consistent, instead, with a more complex picture where genes whose expression can drive trait variation are buffered from local genetic variation but are extensively influenced indirectly by genetic variation in the regulatory networks converging on those genes.

Our results are also consistent with the recently proposed omnigenic model, which posits that complex traits are massively polygenic and that their heritability is spread out across the genome[66]. In the omnigenic model, genes are classified either as "core genes," which directly impinge on the trait, or "peripheral genes," which are not directly trait-related, but influence core genes through the complex gene regulatory network. HDMA explicitly models a central proposal of the omnigenic model which posits that once the expression of the core genes (i.e. trait-mediating genes) is accounted for, there should be no residual correlation between the genome and the phenome. Here, we were able to fit this model and identified a composite transcript that, when taken into account, left no residual correlation between the composite genome and composite phenome scores (Fig. 3A, Supplementary Figure 4E).

Unlike in the omnigenic model, we did not observe a clear demarcation between the core and peripheral genes in loading magnitude, but we do not necessarily expect a clear separation given the complexity of gene regulation and the genotype-phenotype map[67].

An extension of the omnigenic model proposed that most heritability of complex traits is driven by weak distal eQTLs that are potentially below the detection threshold in studies with feasible sample sizes[61]. This is consistent with what we observed here. For example, *Nucb2*, had a high loading in islets and was also strongly distally regulated (66% distal heritability) (Fig. 5). This gene is expressed in pancreatic β cells and is involved in inslin and glucagon release[68–70]. Although its transcription was highly heritable in islets, that regulation was distributed across the genome, with no clear distal eQTL (Supplementary Figure 14). Thus, although distal regulation of some genes may be strong, this regulation is likely to be highly complex and not easily localized.

Individual high-loading transcripts also demonstrated biologically interpretable, tissue-specific patterns. We highlighted *Pparg*, which is known to be protective in adipose tissue[49] where it was negatively loaded, and harmful in the liver[50–54], where it was positively loaded. Such granular patterns may be useful in generating hypotheses for further testing, and prioritizing genes as therapeutic targets. The tissue-specific nature of the loadings also may provide clues to tissue-specific effects, or side effects, of targeting particular genes system-wide.

In addition to identifying individual transcripts of interest, the composite transcripts can be used as weighted

vectors in multiple types of analysis, such as drug prioritization using gene set enrichment analysis (GSEA) and the CMAP database. In particular, the CMAP analysis identified drugs which have been demonstrated to reverse insulin resistance and other aspects of metabolic disease. This finding supports the hypothesis that HDMA identified transcripts that truly mediate genetic effects on traits. On its own, HDMA identifies transcriptional patterns that are consistent with a mediation model, but alone does not prove mediation. However, the finding that these drugs act both on the transcriptional patterns and on the desired traits support the mediation model and the hypothesis that these transcripts have a causal role in pathogenesis of metabolic disease.

Together, our results have shown that both tissue specificity and distal gene regulation are critically important to understanding the genetic architecture of complex traits. We identified important genes and gene signatures that were heritable, plausibly causal of disease, and translatable to other mouse populations and to humans. Finally, we have shown that by directly acknowledging the complexity of both gene regulation and the genotype-to-phenotype map, we can gain a new perspective on disease pathogenesis and develop actionable hypotheses about pathogenic mechanisms and potential treatments.

Diversity Outbred Mice

Mice were maintained and treated in accordance with the guidelines approved by the Department of Biochemistry animal vivarium at the University of Wisconsin. Animal husbandry and in vivo phenotyping methods were previously published [28;12].

A population of 500 diversity outbred mice (split evenly between male and female) from generates 18, 19, and 21, was placed on a high-fat (44.6% kcal fat), high-sugar (34% carbohydrate), adequate protein (17.3 % protein) diet from Envigo Teklad (catalog number TD.08811) starting at four weeks of age as described previously [12]. Individuals were assessed longitudinally for multiple metabolic measures including fasting glucose levels, glucose tolerance, insulin levels, body weight, and blood lipid levels.

When mice were harvested at 22 weeks of age, their pancreatic islets were isolated by hand. Insulin per islet was measured, and whole pancreas insulin content was calculated from the insulin per islet measure and the total numer of islets per pancreas [12]. RNA was isolated from the whole islets and sent to The Jackson Laboratory for high-throughput sequencing [12].

Trait measurements

Trait measurements were described previously in [12]. Briefly, body weight was measured every two weeks, and 4-hour fasting plasma samples were collected to measure insulin, glucose, and triglycerides (TG). At around 18 weeks of age, an oral glucose tolerance test (oGTT) was conducted on 4-hour fasted mice to assess changes in plasma insulin and glucose. Glucose (2 g/kg) was given via oral gavage. Blood samples were taken from a retro-orbital bleed before glucose administration, and at 5, 15, 30, 60, and 120 minutes afterward. The area under the curve (AUC) was calculated for glucose and insulin. Glucose was measured using the glucose oxidase method, and insulin was measured by radioimmunoassay.

HOMA-IR and HOMA-B, which are homeostatic model assessments of insulin resistance (IR) and pancreatic islet function (B), were calculated using fasting plasma glucose and insulin values at the start of the oGTT. HOMA-IR = (glucose $\times$ insulin) / 405 and HOMA-B = (360 $\times$ insulin) / (glucose - 63). Plasma glucose and insulin units are mg/dL and mU/L, respectively.

Genotyping

Genotypes at 143,259 markers was performed using the Mouse Universal Genotyping Array (GigaMUGA) [71] at Neogen (Lincoln, NE) as described previously [12;72]. Genotypes were converted to founder strain-haplotype reconstructions using the R/DOQTL software [73] and interpolated onto a grid with 0.02-cM spacing to yield 69,005 pseudomarkers. Individual chromosome (Chr) haplotypes were reconstructed from RNA-seq data using a hidden Markov model [74] (GBRS, https://github.com/churchill-lab/gbrs). Using both methods to call haplotypes provided redundancy for quality control. Three mice had inconsistent calls between the two methods and were excluded from the analysis [12].

Processed DO Data

The DO data used in this study were generated in a previous study [28;12]. We downloaded genotypes, phenotypes, and pancreatic islet gene expression data from Dryad (doi:10.5061/dryad.pj105).

Collaborative cross recombinant inbred mice (CC-RIX)

Mice were cared for and treated following the guidelines approved by the Association for Assessment and Accreditation of Laboratory Animal Care at The Jackson Laboratory. All animals were obtained from The Jackson Laboratory. The mice were kept in a pathogen-free room at a temperature ranging from 20 to 22°C with a 12-hour light/dark cycle. Starting at 6 weeks of age, they were fed either a custom-designed high-fat,

high-sugar (HFHS) diet (Research Diets D19070208) or a control diet (Research Diets D19072203) *ad libitum*. Body weight was measured weekly until the mice were about 16 weeks old, after which measurements were taken every other week. Food intake measurements were collected at 14 weeks, 23 weeks (for 6-month cohorts), 26 weeks (for 12-month cohorts), 38 weeks, and 51 weeks by weighing the grain contents in the cage over a three-day period. Fasted serum was collected at 14 weeks, 28 weeks (for 6-month cohorts), 26 weeks (for 12-month cohorts), 38 weeks, and 56 weeks of age via retro-orbital or submental vein. Sex, diet, and age were used as covariates in all analyses.

Clinical chemistries

CC-RIX animals were fasted for four hours before serum collection via the retro-orbital or submental vein. Whole blood was left at room temperature for 30-60 minutes before being centrifuged for 5 minutes at 12,500 RPM. The serum was then tested for glucose (Beckman Coulter; OSR6121), cholesterol (Beckman Coulter; OSR6116), triglycerides (Beckman Coulter; OSR60118), insulin (MSD; K152BZC-1), or c-peptide (MSD; K1526JK-1).

Intraperitoneal glucose tolerance testing

After a fasting period of 4-6 hours, baseline glucose measurements were taken from CC-RIX mice using an AlphaTrak2 glucometer and test strips (Zoetis) by making a small nick in the tail tip. A bolus intraperitoneal injection of 20% glucose (1g/kg) was then administered, and additional tail tip nicks were performed at 15, 30, 60, and 120 minutes post-injection to measure glucose levels.

Dual Energy X-ray Absorptiometry (DEXA)

To assess bone mineral density in the CC-RIX population at either 27 weeks of age (6-month cohorts) or 55 weeks of age (12-month cohorts), the mice were weighed and anesthetized through continuous inhalation of isoflurane. The Faxitron UltraFocus DXA system was used to emit two energy levels, 40 kV and 80 kV, for capturing images of bone and soft tissue.

Bulk tissue collection

At either 28 weeks of age (for the 6-month cohort) or 56 weeks of age (for the 12-month cohort), CC-RIX animals were humanely euthanized by cervical dislocation. Tissues, including adipose, gastrocnemius, and the left liver lobe, were harvested and flash-frozen in liquid nitrogen for RNA sequencing.

Whole Pancreas Insulin Content

The animals were humanely euthanized at 16 weeks of age and the entire pancreas was removed, ensuring no excess fat or mesentery tissue was included. The pancreas tissue was placed in a pre-weighed 20 mL glass scintillation vial containing acid ethanol (75% HPLC grade ethanol (ThermoFisher; A995-4), 1.\5% concentrated hydrochloric acid (ThermoFisher; A144-212) in distilled water). The weight of the pancreas was measured for normalization. Using curved scissors, the pancreas was chopped for four minutes, and the samples were stored at $-20°C$ until all animals were harvested. For insulin measurements, the contents of the scintillation vials were rinsed with 4 mL PBS (Roche; 1666789) with 1% BSA (Sigma; A7888), neutralized with 65 µL 10N NaOH (Fisher; SS255-1), and vortexed for 30 seconds. The samples were then centrifuged at 4°C for 5 minutes at 2,000 RPM. The samples were diluted 5000X in PBS with 1% BSA, and insulin was measured (MSD; K152BZC-1).

RNA isolation and QC

RNA from both DO and CC-RIX adipose, gastrocnemius, and left liver lobe tissues was isolated using the MagMAX mirVana Total RNA Isolation Kit (ThermoFisher; A27828) and the KingFisher Flex purification system (ThermoFisher; 5400610). The frozen tissues were pulverized with a Bessman Tissue Pulverizer (Spectrum Chemical) and homogenized in TRIzol™ Reagent (ThermoFisher; 15596026) using a gentleMACS dissociator (Miltenyi Biotec Inc). After adding chloroform to the TRIzol homogenate, the RNA-containing aqueous layer was extracted for RNA isolation, following the manufacturer's protocol, starting with the RNA bead binding step using the RNeasy Mini kit (Qiagen; 74104). RNA concentrations and quality were assessed using the Nanodrop 8000 spectrophotometer (Thermo Scientific) and the RNA 6000 Pico or RNA ScreenTape assay (Agilent Technologies).

Library construction

Before library construction, 2 µL of diluted (1:1000) ERCC Spike-in Control Mix 1 (ThermoFisher; 4456740) was added to 100 ng of each RNA sample. Libraries were then constructed using the KAPA mRNA HyperPrep Kit (Roche Sequencing Store; KK8580) following the manufacturer's protocol. The process involves isolating polyA-containing mRNA using oligo-dT magnetic beads, fragmenting the RNA, synthesizing the first and second strands of cDNA, ligating Illumina-specific adapters with unique barcode sequences for each library, and performing PCR amplification. The quality and concentration of the libraries were evaluated using the D5000 ScreenTape (Agilent Technologies) and the Qubit dsDNA HS Assay (ThermoFisher; Q32851), respectively, according to the manufacturers' instructions.

Sequencing

Libraries were sequenced on an Illumina NovaSeq 6000 using the S4 Reagent Kit (Illumina; 20028312). All tissues underwent 100 bp paired-end sequencing, aiming for a target read depth of 30 million read pairs.

Trait selection in DO

We filtered the measured traits in this study to a set of relatively non-redundant measures that were well-represented in the population (having at least 80% of individuals measured). A complete description of trait filtering can be found at Figshare DOI: 10.6084/m9.figshare.27066979 in the file Documents > 1.DO > 1b.Trait_Selection.Rmd.

We took two approaches for traits with multiple redundant measurements, for example longitudinal body weights. In the case of longitudinal measurements, we used the final measurement, as this was the closest physiological measurement to the measurement of gene expression, which was done at the end of the experiment. The labels for these traits have the word "Final" appended to their name. For traits with multiple highly related measurements, such as cholesterol, we used the first principal component of the group of measurements. For example, we used the first principal component of all LDL measurements as the measurement of LDL. For each set of traits, we ensured the first principal component had the correct sign by correlating it with the average of the traits. For correlation coefficients (R) less than 0, we multiplied the principal component by -1. The labels for these traits have the term "PC1" appended to their name.

Processing of RNA sequencing data

We used the Expectation-Maximization algorithm for Allele Specific Expression (EMASE) [75;76] to quantify multi-parent allele-specific and total expression from RNA-seq data for each tissue. EMASE was performed by the Genotype by RNA-seq (GBRS) software package (https://gbrs.readthedocs.io/en/latest/). In the process, R1 and R2 FASTQ files were combined and aligned to a hybridized (8-way) transcriptome generated for the 8 DO founder strains as single-ended reads. GBRS was also used to reconstruct the mouse genotype probabilities along $\sim$ 69K markers, which was used for confirming genotypes in the quality control (QC) process. For the QC process, we used a Euclidean distances method (developed by Greg Keele - Churchill Lab) to compare the GBRS genotype probabilities between the tissues and the genotype probabilities array for all mice. The counts matrix for each tissue was processed to filter out transcripts with less than one read for at least half of the samples. RNA-seq batch effects were removed by regressing out batch as a random effect and considering sex and generation as fixed effects using lme4 R package. RNA-Seq counts were normalized relative to total read counts using the variance stabilizing transform (VST) as implemented

in DESeq2 and using rank normal score.

eQTL analysis

We used R/qtl2[34] to perform eQTL analysis. We used the rank normal score data and used sex and DO generation as additive covariates. We also used kinship as a random effect. We used permutations to find a LOD threshold of 8 for significant QTLs which corresponded to a genome-wide p value of 0.01[77].

To assess whether eQTL were shared across tissues, we considered significant eQTLs within 4Mb of each other to be overlapping. We considered local and distal eQTLs separately. Local eQTL were defined as an eQTL within 4Mb of the transcription start site of the encoding gene.

Local and distal heritability of transcripts

To estimate local and distal heritability of each transcript, we scaled each normalized transcript to have a variance of 1. We then modeled this transcript with the local genotype using the fit1() function in R/qtl. We used the resulting model to predict the transcript values. The variance of the predicted transcript is its local heritability. We then estimated the heritability of the residual of the model fit. The variance of the residual multiplied by its heritability is the distal heritability of the transcript.

We compared local and distal estimates of heritability to measures of trait relevance for each transcript. To calculate trait relevance of a given transcript, we adjusted normalized transcript values for sex, DO wave, and DO generation. We similarly adjusted traits by sex, DO wave, and DO generation. We then calculated all Spearman correlation coefficients (ρ) between adjusted traits and adjusted transcripts. The trait relevance of a given tanscript was the maximum absolute correlation coefficient across all traits.

High-dimensional mediation analysis

In this section we derive the objective function for high-dimensional mediation analysis (HDMA) and present an iterative algorithm to optimize this objective function. Our starting point is the univariate case, where we describe perfect mediation as a constraint on the covariance matrix among variables. We then leverage this constraint to define projections of multivariate data that are maximally consistent with perfect mediation (HDMA). Next, we demonstrate how to *kernelize* HDMA to limit dimensionality of the model and enable non-linear HDMA models.

Perfect mediation as a constraint on covariance matrices

Suppose we have three random variables x, m, and y. Assume they each have unit variance and that they satisfy the following structural equation model (SEM) such that m perfectly mediates the effect of x on y:

$$m = \alpha x + \epsilon_m \tag{1}$$

$$y = \beta m + \epsilon_y \tag{2}$$

From these structural equations, we have the *model-implied covariance matrix*, Σ, given by

$$\Sigma = \begin{bmatrix} 1 & \alpha & \alpha\beta \\ \alpha & 1 & \beta \\ \alpha\beta & \beta & 1 \end{bmatrix} \tag{3}$$

Note that the assumption of perfect mediation forces the covariance between x and y to be $\alpha\beta$. In any finite data set, however, the observed covariance matrix, $S = [S_{ij}]$, will not typically satisfy this constraint.

The general negative log-likelihood fitting function for an SEM is given by

$$L = \operatorname{tr}\left(S\Sigma^{-1}\right) + \log|\Sigma|\,, \tag{4}$$

where $|\cdot|$ denotes the determinant of a matrix and $\operatorname{tr}(\cdot)$ denotes the trace[78]. For the perfect-mediation model, these values are

$$|\Sigma| = (1-\alpha^2)(1-\beta^2) \tag{5}$$

$$\Sigma^{-1} = \begin{bmatrix} 1/(1-\alpha^2) & -\alpha/(1-\alpha^2) & 0 \\ -\alpha/(1-\alpha^2) & (1-\alpha^2\beta^2)/\left((1-\alpha^2)(1-\beta^2)\right) & -\beta/(1-\beta^2) \\ 0 & -\beta/(1-\beta^2) & 1/(1-\beta^2) \end{bmatrix} \tag{6}$$

Plugging these into the likelihood function, we get

$$L = \log\left((1-\alpha^2)(1-\beta^2)\right) - \frac{2\alpha^2\beta^2}{(1-\alpha^2)(1-\beta^2)} + 1 - \frac{2\alpha}{1-\alpha^2}S_{12} - \frac{2\beta}{1-\beta^2}S_{23} \tag{7}$$

To simplify notation, we define

$$F(\alpha,\beta) = \log\left((1-\alpha^2)(1-\beta^2)\right) - \frac{2\alpha^2\beta^2}{(1-\alpha^2)(1-\beta^2)} + 1, \tag{8}$$

so the likelihood function is now

$$L = F(\alpha,\beta) - \frac{2\alpha}{1-\alpha^2}S_{12} - \frac{2\beta}{1-\beta^2}S_{23} \tag{9}$$

Note that this likelihood is maximized by fitting regression coefficients α and β between x and m and m and y, respectively, but the negative log-likelihood formulation is useful for the multivariate extension below.

Projecting multivariate data to identify latent mediators

Suppose now that we have three data matrices, X, M, and Y (individuals by variables) that are mean centered by column. The central assumption of HDMA is that these multivariate data encode *latent variables* that are causally linked according to the perfect-mediation model, in a sense made precise as follows.

We use the log-likelihood function (Eqn. 7) of the perfect mediation model as an objective function to identify latent variables, l_X, l_M, and l_Y, that are are correlated as closely as possible to the constraints of the perfect mediation model, Eqn. (3). We estimate these latent variables as linear combinations of the measured variables

$$l_X = Xa \tag{10}$$

$$l_M = Mb \tag{11}$$

$$l_Y = Yc \tag{12}$$

The coefficient vectors a, b, and c, are called *loadings*, analogous to the terminology in PCA and CCA. Because the data matrices are mean centered, we have

$$\text{mean}(l_X) = \text{mean}(l_M) = \text{mean}(l_Y) = 0, \tag{13}$$

and we assume the loadings are scaled so that each latent variable has unit variance

$$\text{var}(l_X) = \text{var}(l_M) = \text{var}(l_Y) = 1. \tag{14}$$

Plugging these formulae into the objective function (Eqn. 9), we have

$$S_{12} = \text{corr}\,(l_X, l_M) \tag{15}$$

$$S_{23} = \text{corr}\,(l_M, l_Y) \tag{16}$$

$$L(\alpha, \beta, a, b, c) = F(\alpha, \beta) - \frac{2\alpha}{1-\alpha^2}\text{corr}\,(l_X, l_M) - \frac{2\beta}{1-\beta^2}\text{corr}\,(l_M, l_Y) \tag{17}$$

$$= F(\alpha, \beta) - \frac{2\alpha}{1-\alpha^2}\text{corr}\,(Xa, Mb) - \frac{2\beta}{1-\beta^2}\text{corr}\,(Mb, Yc) \tag{18}$$

This yields an objective function of two sets of parameters: the *structural parameters* α and β that define the causal model among latent variables, and the loading vectors a, b, and c, that define the latent variables in terms of the measured variables. The goal of HDMA is to optimize L as a function of all parameters simultaneously. The form of the objective function, Eqn. 17, is effectively a weighted sum of correlation coefficients, connecting it to so-called *sum-of-correlation*, or SUMCOR, optimization problems[79], which we discuss further below.

An algorithm for HDMA

The global optimization of 17 is challenging because it is not a convex problem. However, the decomposition of the variables into structural and loading variables suggests an iterative algorithm, similar to the expectation-maximization algorithm, that converges at least to a stationary point. The overall idea is to use a block-coordinate-ascent strategy that iterates between optimizing a, b, and c, then optimizing α and β.

For fixed a, b, and c, the optimal α and β are simply given by regression coefficients between l_X and l_M and l_M and l_Y, respectively. Given these regression coefficients, α and β, we then optimize a, b, and c. For fixed α and β, the term $F(\alpha, \beta)$ is irrelevant, so minimizing the negative log-likelihood function reduces to maximizing the reduced function

$$L_{red}(a,b,c) = \frac{2\alpha}{1-\alpha^2}\text{corr}\left(Xa, Mb\right) + \frac{2\beta}{1-\beta^2}\text{corr}\left(Mb, Yc\right), \tag{19}$$

which is a weighted sum of correlation coefficients. This is exactly a (weighted) SUMCOR optimization problem[79]. These optimization problems are still not convex, but Tenenhaus *et al.* have recently proved convergence for iterative algorithms that optimize weighted SUMCOR problems[79–81]. These algorithms only guarantee convergence to a stationary point not necessarily a maximum, as is common in other non-convex problems, but this can be overcome with multiple random restarts, if needed. Thus, we have a sub-routine $\texttt{wSUMCOR}(X, M, Y, w_1, w_2)$ that solves the weighted SUMCOR problem

$$L_{wSUMCOR}(a,b,c,w_1,w_2) = w_1\text{corr}\left(Xa, Mb\right) + w_2\text{corr}\left(Mb, Yc\right). \tag{20}$$

Iterating between optimizing the structural parameters and loading parameters, we reduce the negative log-likelihood at each step and converge to a fixed point.

We summarize our optimization procedure in Algorithm 1.

Algorithm 1 High-dimensional mediation analysis

Input: X, M, Y ▷ Data matrices
Output: α, β, a, b, c, l_X, l_M, l_Y ▷ Structural parameters, loadings, scores
  $\alpha \leftarrow 0.5$, $\beta \leftarrow 0.5$ ▷ Initialize structural parameters
  while *converge* $\neq TRUE$ **do**
    $d \leftarrow \frac{2\alpha}{1-\alpha^2} + \frac{2\alpha}{1-\alpha^2}$ ▷ Normalization constant for weights
    $w_1 \leftarrow \frac{1}{d}\frac{2\alpha}{1-\alpha^2}$, $w_2 \leftarrow \frac{1}{d}\frac{2\beta}{1-\beta^2}$ ▷ Set weights (sum to one)
    $(a,b,c) \leftarrow \texttt{wSUMCOR}(X, M, Y, w_1, w_2)$ ▷ Compute loadings
    $l_X \leftarrow Xa$, $l_M \leftarrow Mb$, $l_Y \leftarrow Yc$ ▷ Compute scores
    $\alpha \leftarrow \text{corr}(l_X, l_M)$, $\beta \leftarrow \text{corr}(l_M, l_Y)$ ▷ Update structural parameters
  end while

Kernel HDMA

For large data matrices X, M, and Y, especially with high correlation among variables, as is common for high-throughput biological assays (*e.g.*, ~1M alleles for genotypes, ~20k transcripts), we can further reduce the dimensionality of the HDMA model by requiring that loading vectors lie in the span of the the measured individuals, namely

$$a = X^T \tilde{a} \tag{21}$$

$$b = M^T \tilde{b} \tag{22}$$

$$c = Y^T \tilde{c}. \tag{23}$$

This replaces the full feature data, say X, with the covariances among individuals (aka, Gram matrices), $C_X = XX^T$, and reduces the dimensionality from the number of measured variables down to the number of individuals

$$l_x = XX^T \tilde{a} = C_X \tilde{a} \tag{24}$$

$$l_M = MM^T \tilde{b} = C_M \tilde{b} \tag{25}$$

$$l_M = YY^T \tilde{c} = C_Y \tilde{c}. \tag{26}$$

This reduction is called *kernelization*[81] and is widely applied to other linear models, including CCA, linear regression, and classification.

It is interesting to note that kernelization is often used to convert a linear model to a non-linear model by replacing the covariance matrices, *e.g.* C_X, with more complex *kernel matrices* K_X that encode similarity measures among individuals that are non-linear functions of the measured variables. non-linear model by replacing the covariance matrices, *e.g.* C_X, with more complex *kernel matrices* K_X that encode similarity measures among individuals that are non-linear functions of the measured variables. Promoting a linear model to a non-linear model in this way is called the *kernel trick* and is widely used in the machine learning field. The above considerations show that HDMA is kernelizable in the same way as other linear models, although the exploration of non-linear models is outside the scope of this study.

We generated kernel matrices for the genome, phenome, and transcriptome as described above. To test the effect of the presence of local eQTLs on mediation, we further generated two additional transcriptomic kernels. 1) A distal-only kernel was derived first by regressing out the effect of local haplotype on all transcripts as described above and generating the kernel matrix using the residual expression (distal-affected only). 2) A local-only kernel was derived by imputing transcription levels for each transcript as described above and then calculating the kernel with only these locally derived expression values. We replaced the original

transcriptomic kernel with each of these additional kernels in turn and performed HDMA. We calculated the correlation between all pairs of latent variables and the path coefficient for each instance.

Implementation details

We have implemented HDMA (Algorithm 1) in the R programming language. Tenenhaus *et al.* have implemented their optimizers in the Regularized Generalized Canonical Correlation Analysis (RGCCA) R package[82], which we use as the subroutine `wSUMCOR`. As Tenenhaus *et al.* discuss optimizing the empirical correlation coefficient *per se* is numerically unstable due to the inversion of the covariance matrices of the measured variables (*e.g.*, the transcript-transcript covariance matrix). To overcome this, the RGCCA package uses a regularized form of the covariance matrix developed by Schaeffer and Strimmer[83], which can be estimated rapidly using an analytic formula.

As a convergence criterion, we stop the iterations when both α and β change by less than 10^{-6} from their previous value in one iteration.

All code required to run HDMA is available at Figshare: https://figshare.com/ DOI: 10.6084/m9.figshare.27066979

Enrichment of biological terms

We performed gene set enrichment analysis (GSEA)[40] using the transcript loadings in each tissue as gene weights. GSEA determines enrichment of pathways based on where the contained genes appear in a ranked list of genes. If the genes in the pathway are more concentrated near the top (or the bottom) of the list than expected by chance, the pathway can be interpreted as being enriched with positively (negatively) loaded transcripts. We used the R package fgsea[39] to calculate normalized enrichment scores for all GO terms and all KEGG pathways.

We downloaded all KEGG[84] pathways for *Mus musculus* using the R package clusterProfiler[84]. We then used fgsea to calculate enrichment scores in each tissue using the transcript loadings in each tissue as our ranked list of genes. We reported the normalized enrichment score (NES) for the 10 pathways with the largest positive NES and the 10 pathways with the largest negative NES.

We used the R package pathview[85] to visualize the loadings from each tissue in interesting pathways. We scaled the loadings in each tissue by the maximum absolute value of loadings across all tissues to compare them across tissues.

We downloaded GO term annotations from Mouse Genome Informatics at the Jackson Laboratory[86] https://www.informatics.jax.org/downloads/reports/index.html We removed gene-annotation pairs labeled with

NOT, indicating that these genes were known not to be involved in these GO terms. We also limited our search to GO terms with between 80 and 3000 genes. We used the R package annotate[87] to identify the ontology of each term and the R package pRoloc[88] to convert between GO terms and names. As with the KEGG pathways, we used fgsea to calculate a normalized enrichment score for each GO term and collected loadings for the transcripts in each term to compare across tissues.

TWAS in DO mice

We performed a transcriptome-wide analysis (TWAS)[9;11] in the DO mice to compare to the results of high-dimensional mediation. To perform TWAS, we fit a linear model to explain variation in each transcript across the population using the genotype at the nearest marker to the gene transcription start site (TSS). We used kinship as a random effect and sex, diet, and DO generation as fixed effects. The predicted transcript from each of these models was the imputed transcript based only on the local genotype.

We correlated each imputed transcript with each of the metabolic phenotypes after adjusting phenotypes for sex, diet, and DO generation. To calculate significance of these correlations, we performed permutation testing by shuffling labels of individual mice and recalculating correlation values. Significant correlations were those more extreme than any of the permuted values, corresponding to an empirical p value of 0. These are transcripts whose locally encoded expression level was significantly correlated with one of the metabolic traits. This suggests an association between the genetically encoded transcript level and the trait but does not identify a direction of causation.

Literature support for genes

To determine whether each gene among those with large loadings or large heritability had a supported connection to obesity or diabetes in the literature, we used the R package easyPubMed[89]. We searched for the terms ("diabetes" OR "obesity") along with the tissue name (adipose, islet, liver, or muscle), and the gene name. We restricted the gene name to appear in the title or abstract as some short names appeared coincidentally in contact information. We checked each gene with apparent literature support by hand to verify that support, and we removed spurious associations. For example, FAU is used as an acronym for fatty acid uptake and CAD is used as an acronym for coronary artery disease. Both terms co-occur with the terms diabetes and obesity in a manner independent of the genes *Fau* and *Cad*. Other genes that co-occurred with diabetes and obesity, but not as a functional connection were similarly removed. For example, the gene *Rpl27* is used as a reference gene for quantification of the expression of other genes, and co-occurrence with diabetes and obesity is a coincidence. We counted the abstracts associated with diabetes or obesity and each gene

name and determined that a gene had literature support when it had at least two abstracts linking it to the terms diabetes or obesity in the respective tissue.

Tissue-specific clusters

To compare the top loading genes across tissues, we selected genes with a loading at least 2.5 standard deviations from the mean across all tissues. We made a matrix consisting of the union of these sets populated with the tissue-specific loading for each gene. We used the pam() function in the R package cluster[90] to cluster the loading profiles around k medoids. We tested $k = 2$ through 20 and used silhouette andlysis to compare the separation of the clusters. The best separation was achieved with $k = 12$ clusters. For each cluster we used the R package gprofiler2[91] to identify enriched GO terms and KEGG pathways for the genes in each cluster.

CC-RIX genotypes

We used the most recent common ancestor (MRCA) genotypes for the Collaborative Cross (CC) mice available on the University of North Carolina Computational Systems Biology website: http://www.csbio.unc.edu/CC status/CCGenomes/

To generate CC-RIX genotypes, we averaged the haplotype probabilities for the two parental strains at each locus.

Imputation of gene expression in CC-RIX

To impute gene expression in the CC-RIX, we performed the following steps for each transcript in each tissue (adipose, liver, and skeletal muscle):

1. Calculate diploid CC-RIX genotype for all CC-RIX individuals at the marker nearest the transcription start site of the transcript.
2. Multiply the genotype probabilities by the eQTL coefficients identified in the DO population.

To check the accuracy of the imputation, we correlated each imputed transcript with the measured transcript. The average Pearson correlation (r) was close to 0.5 for all three tissues (Supp. Fig. S7A), and as expected, the correlation between the imputed transcript and the measured transcript was highly positively dependent on the local eQTL LOD score of the transcript (Supp. Fig. S7B).

Prediction of CC-RIX traits

We used both measured expression and imputed expression combined with the results from HDM in the to predict phenotype in the CC-RIX. The traits measured in the DO and the CC-RIX were not identical, so we limited our prediction to body weight, which was measured in both populations, and was the largest contributor to the phenotype score in the DO.

For each CC-RIX individual, we multiplied the transcript abundances across the transcriptome by the loadings derived from the HDM in the DO population (Fig. 7A). This resulted in a vector with n elements, where n is the number of transcripts in the trancriptome. Each element was a weighted value that combined the relative abundance of the transcript with how that abundance affected the phenotype. We averaged the values in this vector to calculate an overall predicted phenotype score for the individual CC-RIX animal.

After calculating this predicted phenotype value across all CC-RIX animals, we correlated the predicted values from each tissue with measured body weight (Fig. 7B).

Cell type specificity

We investigated whether the loadings derived from HDM reflected tissue composition changes in the DO mice prone to obesity on the high-fat diet. To do this, we acquired lists of cell-type specific transcripts from the literature. In adipose tissue, we looked at cell-type specific transcripts for macrophages, leukocytes, adipocyte progenitors, and adipocytes as defined in [29087381]. In pancreatic islets, we looked at cell-type specific transcripts for alpha cells, beta cells, delta cells, ductal cells, mast cells, macrophages, acinar cells, stellate cells, gamma and epsilon cells, and endothelial cells as defined by [36778506]. Both studies defined cell-type specific transcripts based on human cell types. We collected the loadings for each set of cell-type specific transcripts in the respective tissue and asked whether the mean loading for the cell type differed significantly from 0 (Fig. 8). A significant positive loading for the cell type would suggest a genetic predisposition to have a higher proportion of that cell type in the tissue. To determine whether each mean loading differed significantly from 0, we performed permutation tests. We randomly sampled n genes outside of the cell-type specific, where n was the number of genes in the set. We compared the distribution of loading means over 10,000 random draws to that seen in the observed data. We used a significance threshold of 0.01.

Comparison of transcriptomic signatures to human transcriptomic signatures

To compare the transcriptomic signatures identified in the DO mice to those seen in human patients, we downloaded human gene expression data from the Gene Expression Omnibus (GEO)[92;93]. We focused on

adipose tissue because this had the strongest relationship to obesity and insulin resistance in the DO. We downloaded the following human gene expression data sets:

- Accession number GSE152517 - Performed bulk RNA sequencing on visceral adipose tissue resected from seven diabetic and seven non-diabetic obese individuals.
- Accession number GSE44000 - Used Agilent-014850 4X44K human whole genome platform arrays (GPL6480) to measure gene expression in purified adipocytes derived from the subcutaneous adipose tissue of seven obese (BMI>30) and seven lean (BMI<25) post-menopausal women.
- Accession number GSE205668 - Subcutaneous adipose tissue was resected during elective surgery from 35 normal weight, and 26 obese children. Gene expression was measured by RNA sequencing with an Illumina HiSeq 2500.
- Accession number GSE29231 - Visceral adipose biopsies were taken from three female patients with type 2 diabetes, and three non-diabetic female patients. Expression was measured with Illumina HumanHT-12 v3 Expression BeadChip arrays.

We downloaded each data set from GEO using the R package GEOquery [94]. In each case, we verified that gene expression was log transformed and performed the transformation ourselves if it had not already been done. When covariates such as age and sex were available in the metadata files, we regressed out these variables. We mean centered and standardized gene expression across transcripts.

We matched the human gene expression to the mouse gene expression by pairing orthologs as defined in The Jackson Laboratory's mouse genome informatics data base (MGI) [95]. We multiplied each transcript in the human data by the adipose tissue loading of its ortholog in the DO mice. This resulted in a vector of weighted transcript values for each patient based on their own transcriptional profile and the obesity-related transcriptional signature from the DO analysis. The mean of this vector for an individual was the prediction of their obesity status. Higher values indicate a prediction of higher obesity or risk of metabolic disease based on adipose gene expression. We then compared the values across groups, either obese and non-obese, or diabetic and non-diabetic depending on the groups in each study.

Connectivity Map Queries

We queried the transcript loading signatures from adipose tissue and pancreatic islets with the CMAP database. These tissues are the most related to metabolic disease and diabetes respectively.

The gene expression profiles in the Connectivity Map database are derived from human cell lines and human

primary cultures and are indexed by Entrez gene IDs. To query the CMAP database, we identified the Entrez gene IDs for the human orthologs of the mouse genes expressed in each tissue. Each CMAP query takes the 150 most up-regulated and the 150 most down-regulated genes in a signature, however, not all human genes are included in their database. To ensure we had as many genes as possible in the query, we selected the top and bottom 200 genes with the most extreme positive and negative loadings respectively. We pasted these into the CLUE query application available at https://clue.io/query.

We filtered the results in two ways: First, we looked at the most significantly anti-correlated ($-log_{10}(\text{FDR}q) > 15$) hits across all cell types. Second, we looked at the most anti-correlated within the most related cell type to the query and considered hits regardless of $-log_{10}(\text{FDR}p)$. For adipose tissue we looked in normal adipocytes, abbreviated ASC in the CMAP database, and for pancreatic islets we looked in pancreatic cancer cells, abbreviated YAPC in the CMAP database.

Data Availability

DO mice: Genotypes, phenotypes, and pancreatic islet gene expression data were previously published [12]. Gene expression for the other tissues can be found at the Gene Expression Omnibus https://www.ncbi.nlm.nih.gov/geo/ with the following accession numbers: DO adipose tissue - GSE266549; DO liver tissue - GSE266569; DO skeletal muscle - GSE266567. Expression data with calculated eQTLs are available at Figshare https://figshare.com/articles/dataset/Data_and_code_for_High-Dimensional_Mediation_Analysis_HDMA_in_diversity_outbred_mice/27066979 DOI: 10.6084/m9.figshare.27066979

10.6084/m9.figshare.27066979.v1

CC-RIX mice: Gene expression can be found at the Gene Expression Omnibus https://www.ncbi.nlm.nih.gov/geo/ with the following accession numbers: CC-RIX adipose tissue - GSE237737; CC-RIX liver tissue - GSE237743; CC-RIX skeletal muscle - GSE237747. Count matrices and phenotype data can be found at Figshare https://figshare.com/articles/dataset/Data_and_code_for_High-Dimensional_Mediation_Analysis_HDMA_in_diversity_outbred_mice/27066979 DOI: 10.6084/m9.figshare.27066979

Code Availability

Code: All code used to run the analyses reported here are available at Figshare: https://figshare.com/articles/dataset/Data_and_code_for_High-Dimensional_Mediation_Analysis_HDMA_in_diversity_outbred_mice/27066979 DOI: 10.6084/m9.figshare.27066979

Acknowledgements

This project was supported by The Jackson Laboratory Cube Initiative, as well as grants from the National Institutes of Health (grant numbers.: R01DK101573, R01DK102948, and RC2DK125961) (to A. D. A.) and by the University of Wisconsin-Madison, Department of Biochemistry and Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation (to M. P. K.).

We thank the following scientific services at The Jackson Laboratory: Genome Technologies for the RNA sequencing, necropsy services for the tissue harvests, and the Center for Biometric Analysis for metabolic phenotyping.

Supplementary Figures

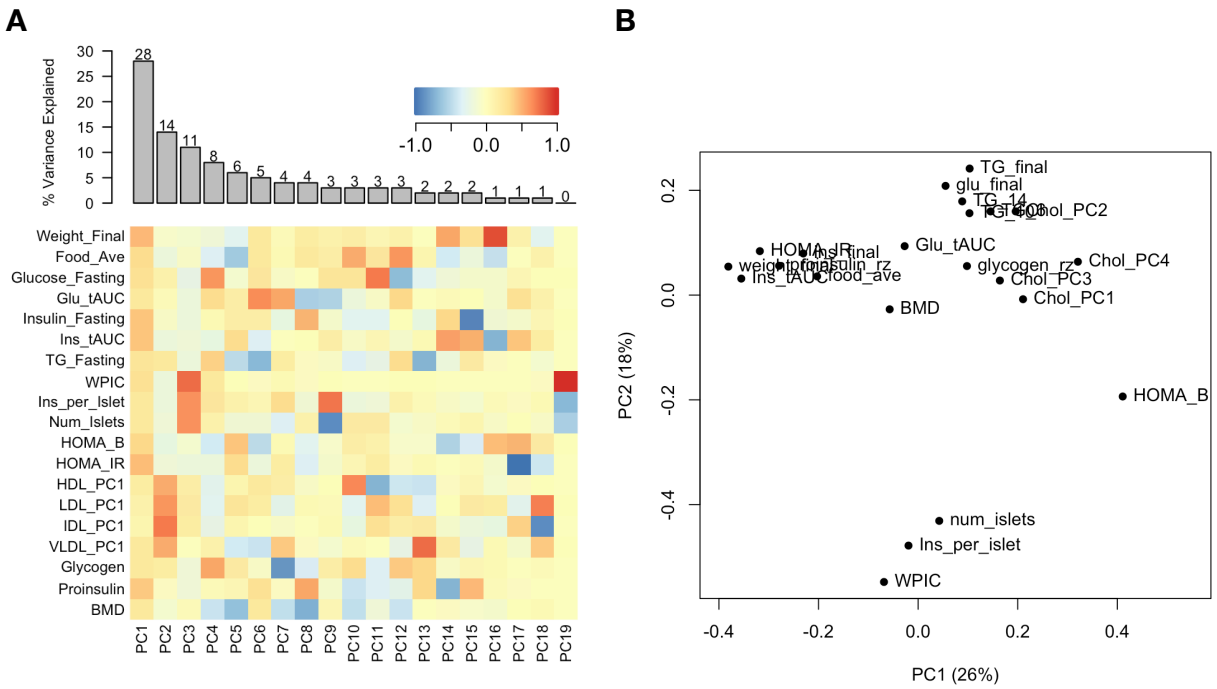


Figure 1: Trait matrix decomposition. **A** The heat map shows the loadings of each trait onto each principal component of the trait matrix. The bars at the top show the percent variance explained for each principal component. **B** Traits plotted by the first and second principal components of the trait matrix. This view shows clustering of traits into insulin- and weight-related traits, lipid-related traits, and ex-vivo pancreatic measurements.

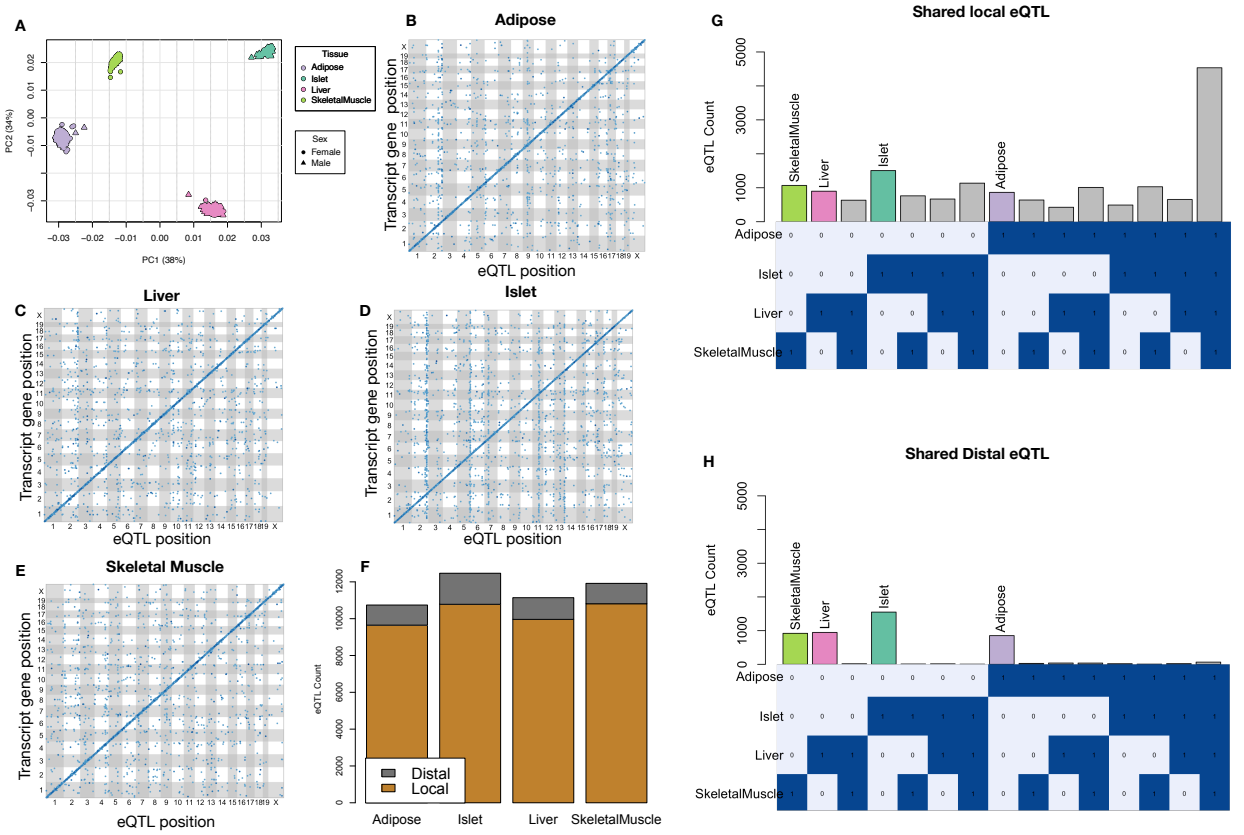


Figure 2: Overview of eQTL analysis in DO mice. **A.** RNA seq samples from the four different tissues clustered by tissue. **B.-E.** eQTL maps are shown for each tissue. The x-axis shows the position of the mapped eQTL, and the y-axis shows the physical position of the gene encoding each mapped transcript. Each dot represents an eQTL with a minimum LOD score of 8, which represents a genome-wide permutation-based threshold of $p < 0.01$. The dots on the diagonal are locally regulated eQTL for which the mapped eQTL is at the within 4Mb of the encoding gene. Dots off the diagonal are distally regulated eQTL for which the mapped eQTL is distant from the gene encoding the transcript. **F.** Comparison of the total number of local and distal eQTL with a minimum LOD score of 8 in each tissue. All tissues have comparable numbers of eQTL. Local eQTLs are much more numerous than distal eQTL. **G.** Counts of transcripts with local eQTL shared across multiple tissues. The majority of local eQTLs were shared across all four tissues. **H.** Counts of transcripts with distal eQTL shared across multiple tissues. The majority of distal eQTL were tissue-specific and not shared across multiple tissues. For both G and H, eQTL for a given transcript were considered shared in two tissues if they were within 4Mb of each other. Colored bars indicate the counts for individual tissues for easy of visualization.

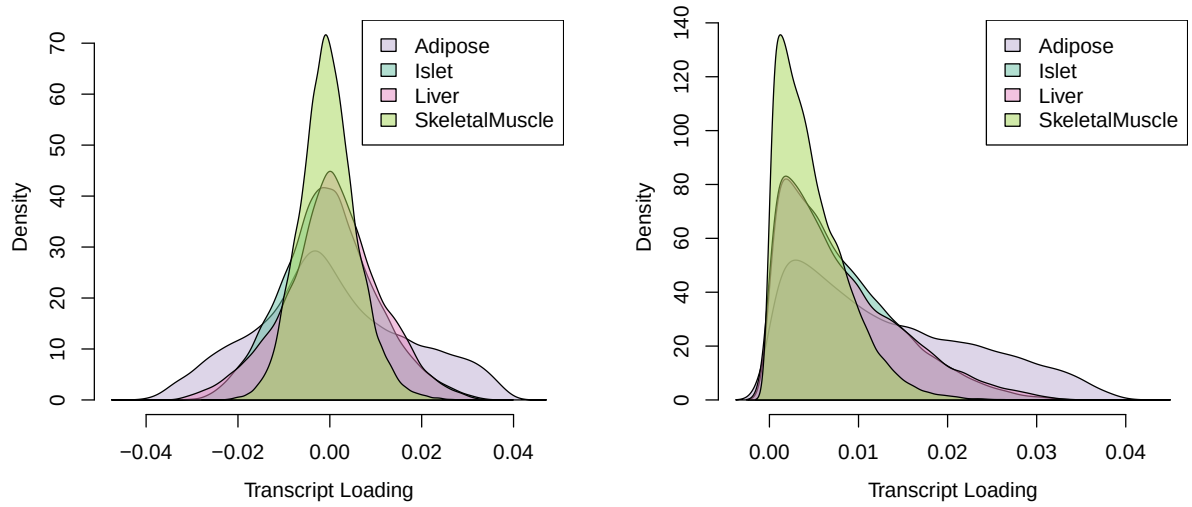


Figure 3: Direct comparisons of transcript loadings across tissues. **A.** Distributions of transcript loadings are shown as density curves and are differentially colored to indicate tissue. Transcripts in adipose tissue had both the largest positive and negative loadings. **B.** Direct comparison of absolute values of transcript loadings across tissues. Transcripts in adipose tissue had the largest loadings overall, while those in skeletal muscle had the smallest.

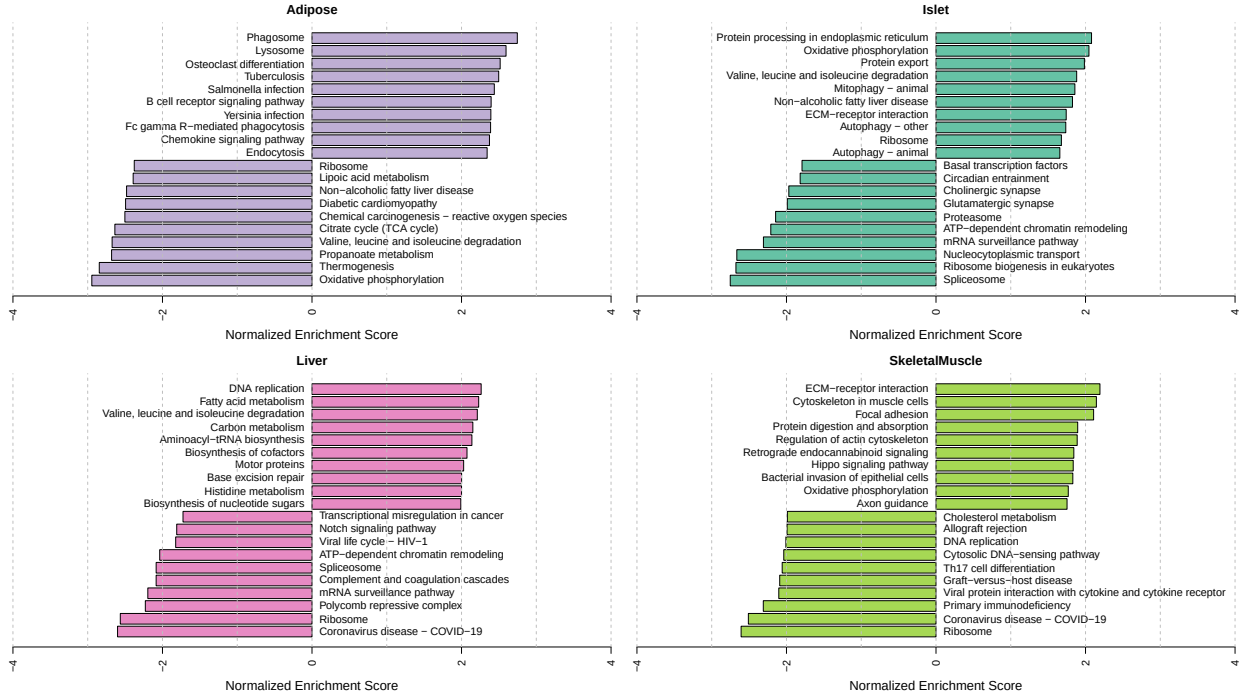


Figure 4: Bar plots showing normalized enrichment scores (NES) for KEGG pathways as determined by fast gene score enrichment analysis (fgsea). Only the top 10 positive and top 10 negative scores are shown. Colors indicate tissue. The name beside each bar shows the name of each enriched KEGG pathway.

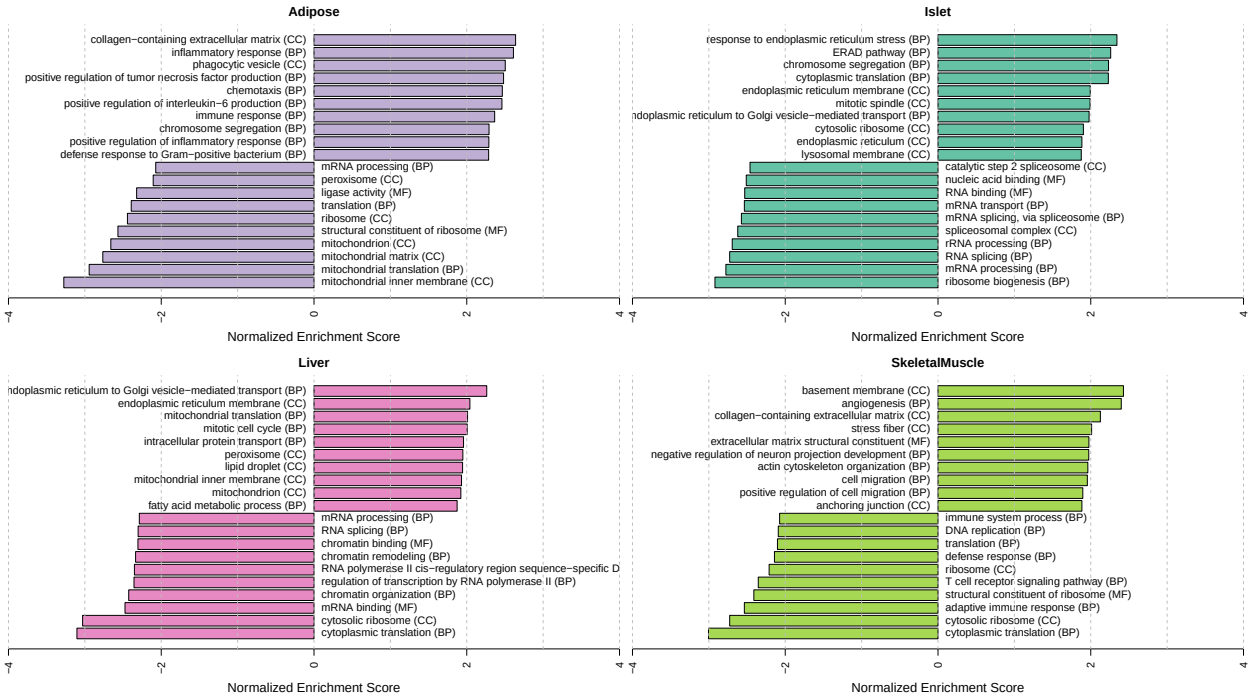


Figure 5: Bar plots showing normalized enrichment scores (NES) for GO terms as determined by fast gene score enrichment analysis (fgsea). Only the top 10 positive and top 10 negative scores are shown. Colors indicate tissue. The name beside each bar shows the name of each enriched GO term. The letters in parentheses indicate whether the term is from the biological process ontology (BP), the molecular function ontology (MF), or the cellular compartment ontology (CC).

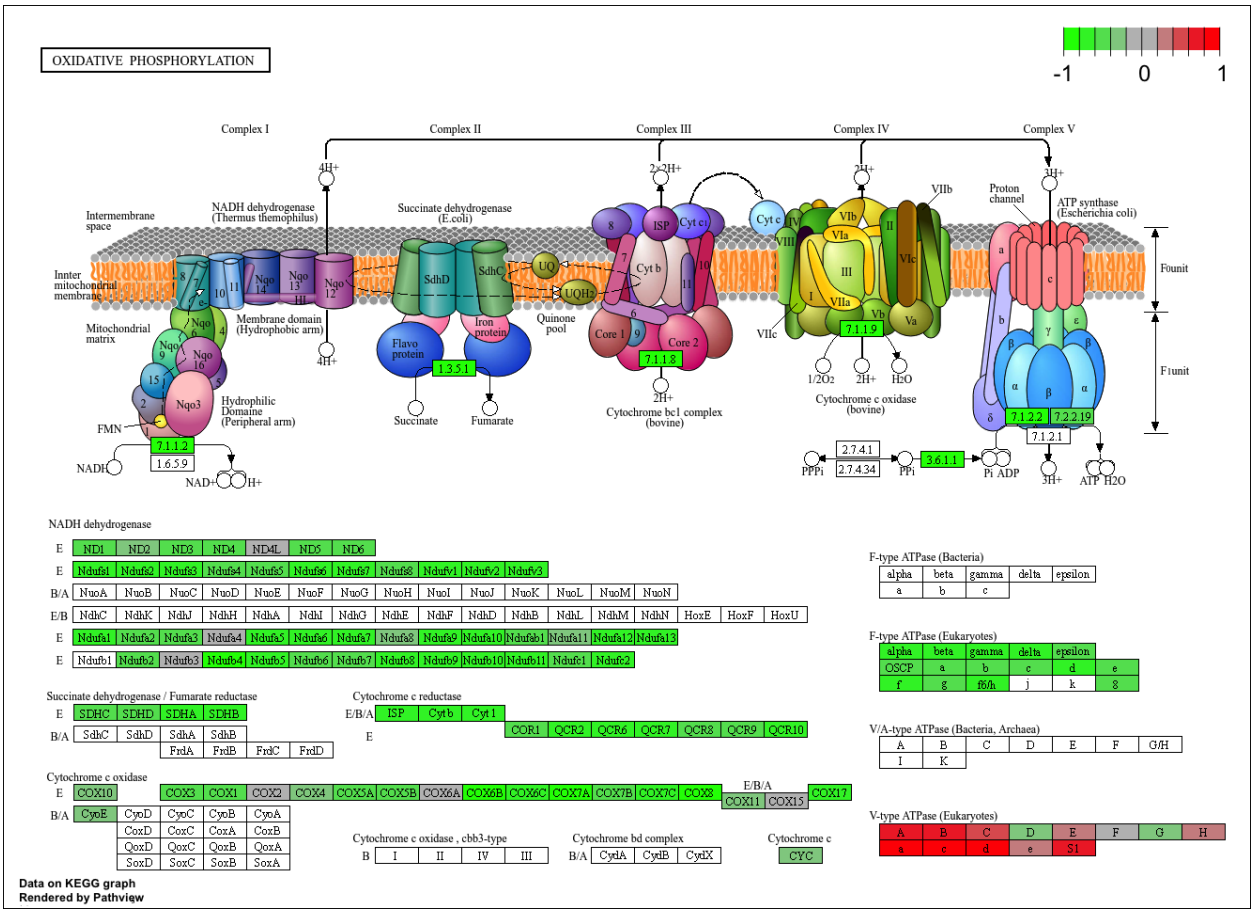


Figure 6: The KEGG pathway for oxidative phosphorylation in mice. Each element is colored based on its HDMA loading from adipose tissue scaled to run from -1 to 1. Genes highlighted in green had negative loadings, and those highlighted in red had positive loadings. Almost the entire pathway was strongly negatively loaded indicating that increased expression of genes involved in oxidative phosphorylation was associated with reduced MDI.

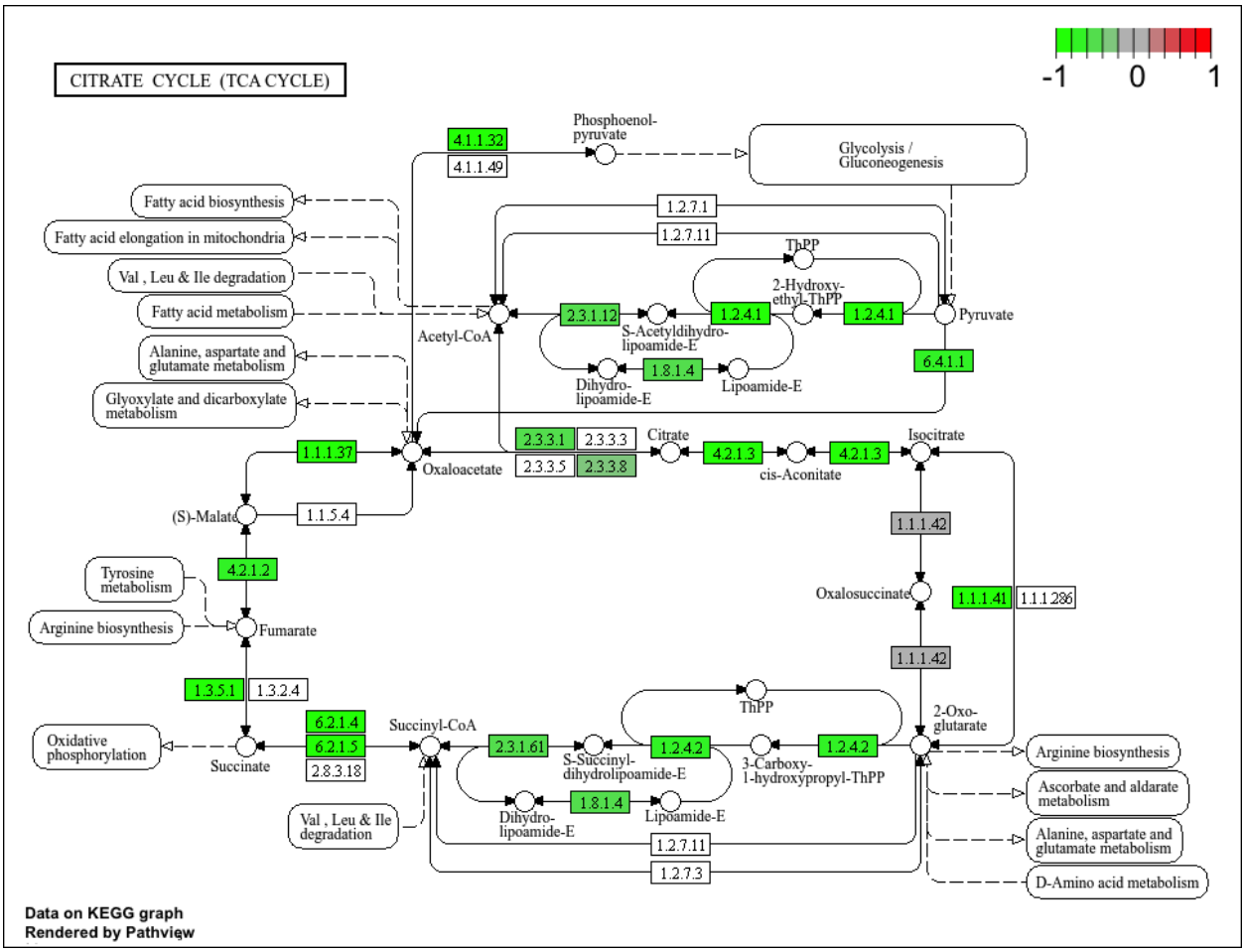


Figure 7: The KEGG pathway for the TCA (citric acid) cycle in mice. Each element is colored based on its HDMA loading from adipose tissue scaled to run from -1 to 1. Genes highlighted in green had negative loadings, and those highlighted in red had positive loadings. Many genes in the cycle were strongly negatively loaded indicating that increased expression of genes involved in the TCA cycle was associated with reduced MDI.

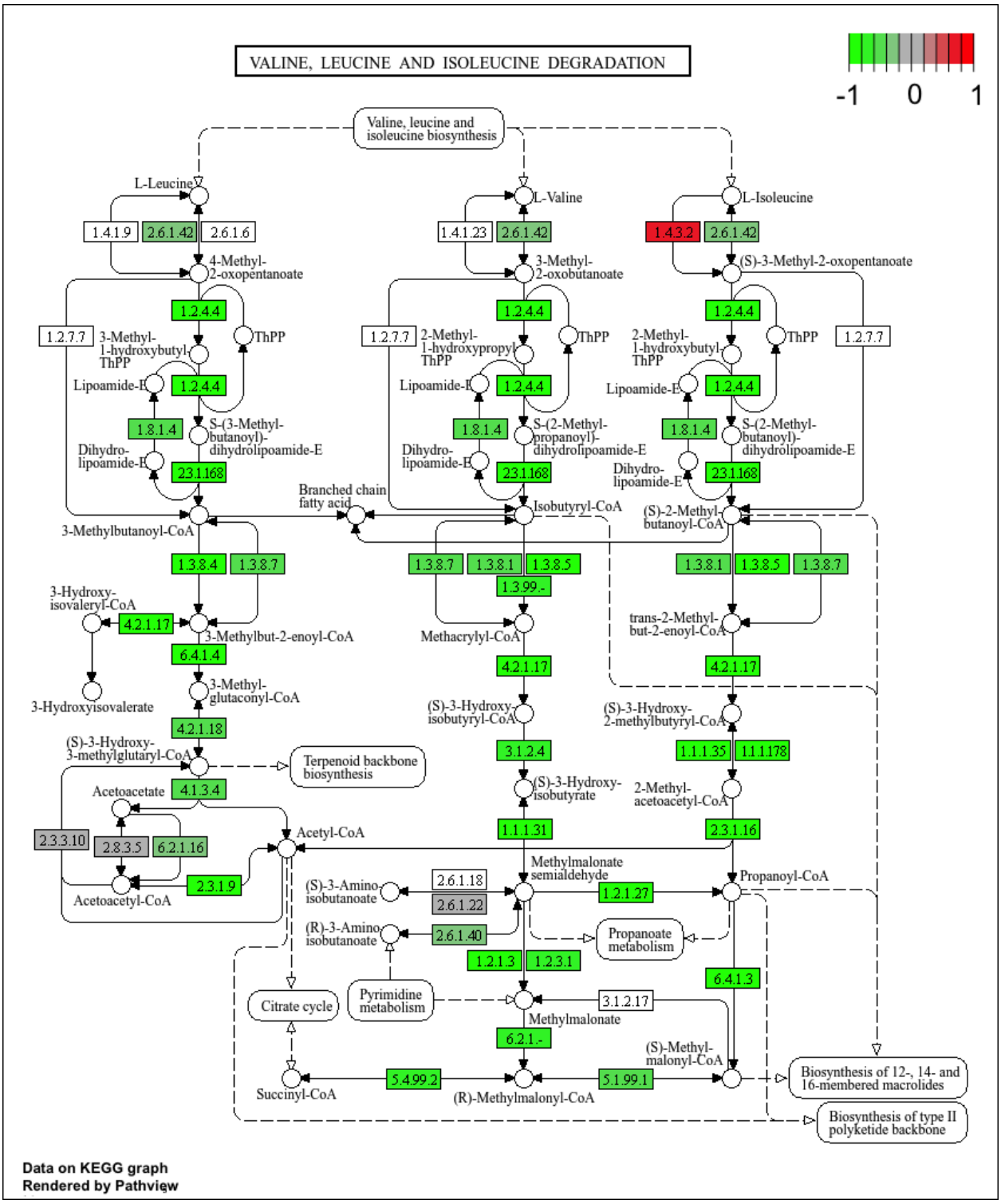


Figure 8: The KEGG pathway for branched-chain amino acid degradation in mice. Each element is colored based on its HDMA loading from adipose tissue scaled to run from -1 to 1. Genes highlighted in green had negative loadings, and those highlighted in red had positive loadings. Almost the entire pathway was strongly negatively loaded indicating that increased expression of genes involved in branched-chain amino acid degradation was associated with reduced MDI.

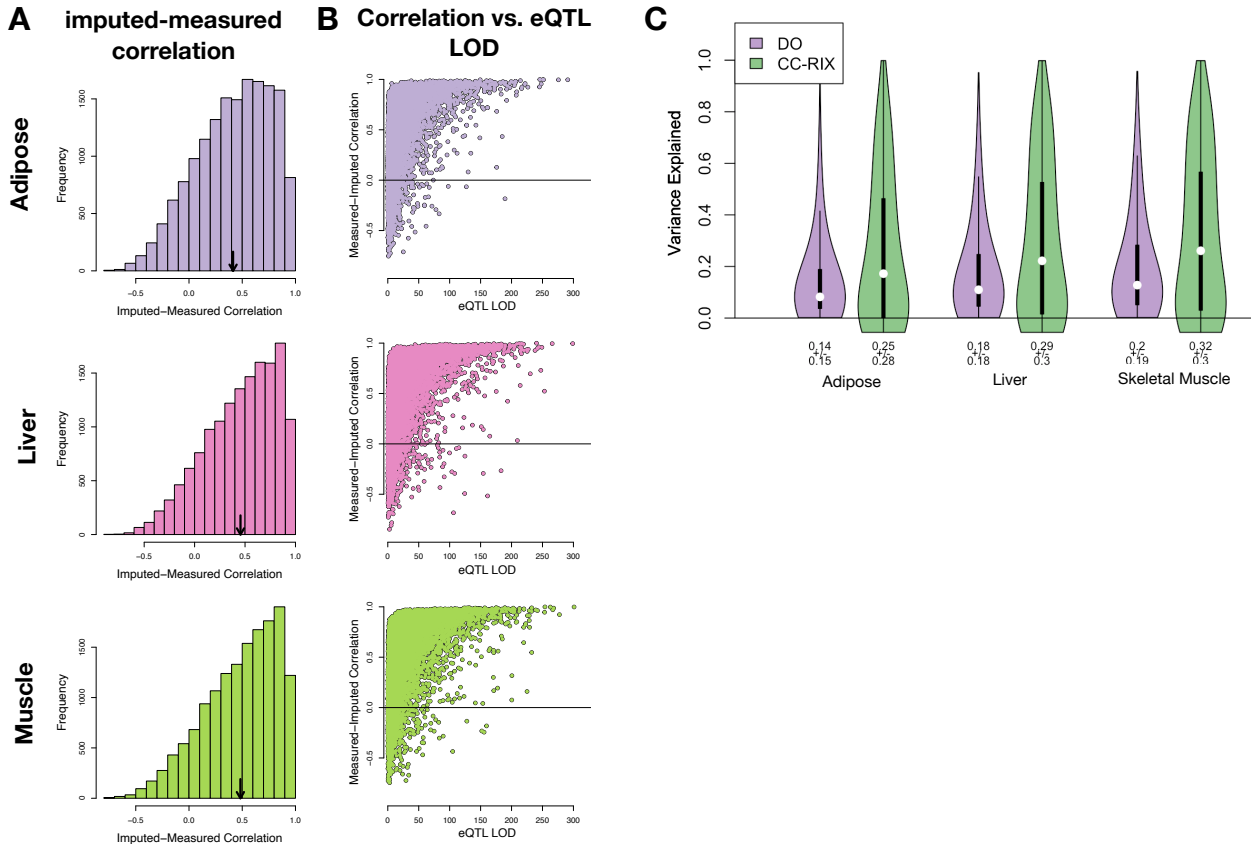
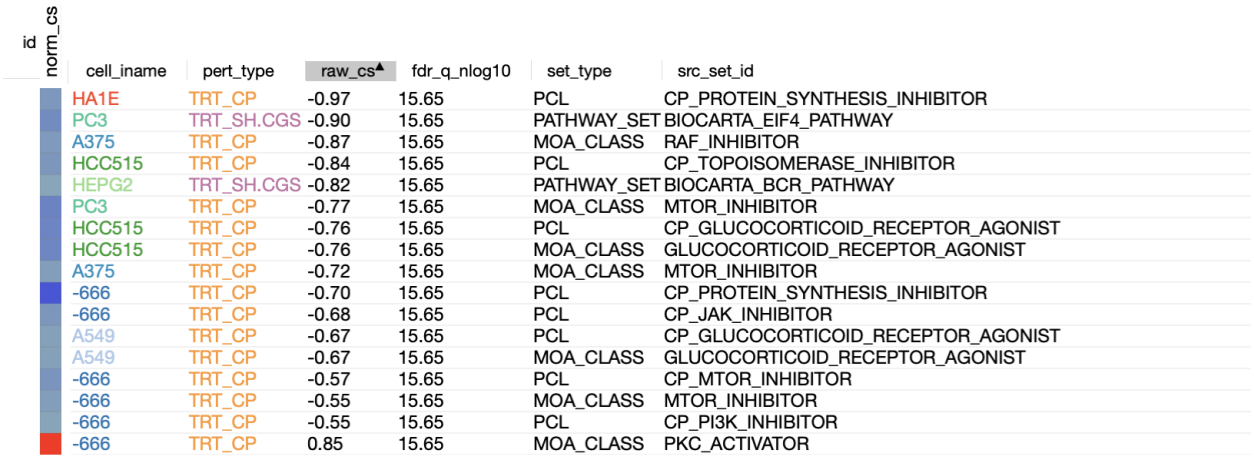


Figure 9: Validation of transcript imputation in the CC-RIX. **A.** Distributions of correlations between imputed and measured transcripts in the CC-RIX. The mean of each distribution is shown by the red line. All distributions were skewed toward positive correlations and had positive means near a Pearson correlation (r) of 0.5. **B.** The relationship between the correlation between measured and imputed expression in the CC-RIX (x-axis) and eQTL LOD score. As expected, imputations are more accurate for transcripts with strong local eQTLs. **C.** Distributions of variance explained by local genotype across all transcripts in the DO and CC-RIX.



id	norm_cs	cell_iname	pert_type	raw_cs	fdr_q_nlog10	set_type	src_set_id
		HA1E	TRT_CP	-0.97	15.65	PCL	CP_PROTEIN_SYNTHESIS_INHIBITOR
		PC3	TRT_SH.CGS	-0.90	15.65	PATHWAY_SET	BIOCARTA_EIF4_PATHWAY
		A375	TRT_CP	-0.87	15.65	MOA_CLASS	RAF_INHIBITOR
		HCC515	TRT_CP	-0.84	15.65	PCL	CP_TOPOISOMERASE_INHIBITOR
		HEPG2	TRT_SH.CGS	-0.82	15.65	PATHWAY_SET	BIOCARTA_BCR_PATHWAY
		PC3	TRT_CP	-0.77	15.65	MOA_CLASS	MTOR_INHIBITOR
		HCC515	TRT_CP	-0.76	15.65	PCL	CP_GLUCOCORTICOID_RECEPTOR_AGONIST
		HCC515	TRT_CP	-0.76	15.65	MOA_CLASS	GLUCOCORTICOID_RECEPTOR_AGONIST
		A375	TRT_CP	-0.72	15.65	MOA_CLASS	MTOR_INHIBITOR
		-666	TRT_CP	-0.70	15.65	PCL	CP_PROTEIN_SYNTHESIS_INHIBITOR
		-666	TRT_CP	-0.68	15.65	PCL	CP_JAK_INHIBITOR
		A549	TRT_CP	-0.67	15.65	PCL	CP_GLUCOCORTICOID_RECEPTOR_AGONIST
		A549	TRT_CP	-0.67	15.65	MOA_CLASS	GLUCOCORTICOID_RECEPTOR_AGONIST
		-666	TRT_CP	-0.57	15.65	PCL	CP_MTOR_INHIBITOR
		-666	TRT_CP	-0.55	15.65	MOA_CLASS	MTOR_INHIBITOR
		-666	TRT_CP	-0.55	15.65	PCL	CP_PI3K_INHIBITOR
		-666	TRT_CP	0.85	15.65	MOA_CLASS	PKC_ACTIVATOR

Figure 10: CMAP results using the *adipose* tissue composite transcript as an input. Table includes results from *all cell types* sorted with a $-log_{10}(q) > 15$. The results are sorted by the correlation of the query to the input with the most negative results at the top.

id	norm_cs	cell_iname	pert_type	raw_cs	fdr_q_nlog10	set_type	src_set_id
		VCAP	TRT_SH.CGS	-0.99	15.65	PATHWAY_SET	REACTOME_DOWNSTREAM_TCR_SIGNALING
		VCAP	TRT_SH.CGS	-0.99	15.65	PATHWAY_SET	REACTOME_NOD1_2_SIGNALING_PATHWAY
		A549	TRT_SH.CGS	-0.92	15.65	PATHWAY_SET	BIOCARTA_TNFR1_PATHWAY
		VCAP	TRT_SH.CGS	-0.92	15.65	PATHWAY_SET	HALLMARK_WNT_BETA_CATENIN_SIGNALING
		HT29	TRT_CP	-0.92	15.65	PCL	CP_TUBULIN_INHIBITOR
		-666	TRT_OE	-0.88	15.65	PCL	OE_CELL_CYCLE_INHIBITION
		VCAP	TRT_SH.CGS	-0.87	15.65	PATHWAY_SET	REACTOME_P75_NTR_RECEPTOR_MEDIATED_SIGNALLING
		HT29	TRT_CP	-0.86	15.65	MOA_CLASS	TUBULIN_INHIBITOR
		MCF7	TRT_CP	-0.85	15.65	PCL	CP_TUBULIN_INHIBITOR
		-666	TRT_CP	-0.81	15.65	PCL	CP_PROTEASOME_INHIBITOR
		-666	TRT_SH.CGS	-0.80	15.65	PATHWAY_SET	REACTOME_DOWNREGULATION_OF_ERBB2_ERBB3_SIGNALING
		HCC515	TRT_CP	-0.80	15.65	PCL	CP_GLUCOCORTICOID_RECEPTOR_AGONIST
		HCC515	TRT_CP	-0.80	15.65	MOA_CLASS	GLUCOCORTICOID_RECEPTOR_AGONIST
		A549	TRT_OE	-0.78	15.65	PATHWAY_SET	REACTOME_RAF_MAP_KINASE_CASCADE
		A549	TRT_OE	-0.78	15.65	PATHWAY_SET	PID_RAS_PATHWAY
		-666	TRT_SH.CGS	-0.78	15.65	PCL	KD_RIBOSOMAL_40S_SUBUNIT
		A549	TRT_OE	-0.76	15.65	PATHWAY_SET	REACTOME_SIGNALLING_TO_P38_VIA_RIT_AND_RIN
		A549	TRT_OE	-0.76	15.65	PATHWAY_SET	REACTOME_PROLONGED_ERK_ACTIVATION_EVENTS
		A549	TRT_OE	-0.73	15.65	PATHWAY_SET	PID_TCR_RAS_PATHWAY
		HA1E	TRT_OE	-0.73	15.65	PATHWAY_SET	REACTOME_SHC_RELATED_EVENTS
		HA1E	TRT_OE	-0.71	15.65	PATHWAY_SET	PID_EPHB_FWD_PATHWAY
		-666	TRT_CP	-0.70	15.65	MOA_CLASS	GLYCOGEN_SYNTHASE_KINASE_INHIBITOR
		HA1E	TRT_OE	-0.70	15.65	PATHWAY_SET	PID_GMCSF_PATHWAY
		A549	TRT_OE	-0.69	15.65	PATHWAY_SET	REACTOME_SIGNALLING_TO_ERKS
		-666	TRT_LIG	-0.69	15.65	PATHWAY_SET	PID_ERBB_NETWORK_PATHWAY
		-666	TRT_CP	-0.67	15.65	MOA_CLASS	PROTEASOME_INHIBITOR
		-666	TRT_CP	-0.66	15.65	PCL	CP_GLYCOGEN_SYNTHASE_KINASE_INHIBITOR
		-666	TRT_CP	0.73	15.65	MOA_CLASS	MTOR_INHIBITOR

Figure 11: CMAP results using the *pancreatic islet* tissue composite transcript as an input. Table includes results from *all cell types* sorted with a $-log_{10}(q) > 15$. The results are sorted by the correlation of the query to the input with the most negative results at the top.

id	norm_cs	cell_iname	pert_type	raw_cs ▲	fdr_q_nlog10	set_type	src_set_id
		ASC	TRT_CP	-0.94	0.79	PCL	CP_PARP_INHIBITOR
		ASC	TRT_CP	-0.94	0.79	MOA_CLASS	PROTEIN_TYROSINE_KINASE_INHIBITOR
		ASC	TRT_CP	-0.84	0.45	MOA_CLASS	BTK_INHIBITOR
		ASC	TRT_CP	-0.81	0.39	MOA_CLASS	LEUCINE_RICH_REPEAT_KINASE_INHIBITOR
		ASC	TRT_CP	-0.81	0.79	PCL	CP_HSP_INHIBITOR
		ASC	TRT_CP	-0.80	0.93	PCL	CP_EGFR_INHIBITOR
		ASC	TRT_CP	-0.79	0.32	MOA_CLASS	T-TYPE_CALCIUM_CHANNEL_BLOCKER
		ASC	TRT_CP	-0.79	1.09	PCL	CP_MTOR_INHIBITOR
		ASC	TRT_CP	-0.76	0.97	PCL	CP_PI3K_INHIBITOR
		ASC	TRT_CP	-0.75	0.20	MOA_CLASS	HISTONE_DEMETHYLASE_INHIBITOR
		ASC	TRT_CP	-0.74	0.42	PCL	CP_IKK_INHIBITOR
		ASC	TRT_CP	-0.74	0.83	PCL	CP_AURORA_KINASE_INHIBITOR
		ASC	TRT_CP	-0.74	0.17	PCL	CP_LEUCINE_RICH_REPEAT_KINASE_INHIBITOR
		ASC	TRT_CP	-0.72	0.36	PCL	CP_BROMODOMAIN_INHIBITOR
		ASC	TRT_CP	-0.71	1.09	MOA_CLASS	TYROSINE_KINASE_INHIBITOR
		ASC	TRT_CP	-0.70	0.82	PCL	CP_PROTEIN_SYNTHESIS_INHIBITOR
		ASC	TRT_CP	-0.67	0.69	PCL	CP_SRC_INHIBITOR
		ASC	TRT_CP	-0.67	0.81	MOA_CLASS	AURORA_KINASE_INHIBITOR
		ASC	TRT_CP	-0.65	0.89	MOA_CLASS	FLT3_INHIBITOR
		ASC	TRT_CP	-0.62	0.40	MOA_CLASS	FGFR_INHIBITOR
		ASC	TRT_CP	-0.59	0.66	MOA_CLASS	MEK_INHIBITOR
		ASC	TRT_CP	-0.59	0.13	MOA_CLASS	SYK_INHIBITOR
		ASC	TRT_CP	-0.58	0.01	PCL	CP_PKC_INHIBITOR
		ASC	TRT_CP	-0.58	0.65	PCL	CP_HDAC_INHIBITOR
		ASC	TRT_CP	-0.58	0.65	PCL	CP_ATPASE_INHIBITOR
		ASC	TRT_CP	-0.53	0.09	PCL	CP_FLT3_INHIBITOR
		ASC	TRT_CP	-0.53	0.42	PCL	CP_P38_MAPK_INHIBITOR
		ASC	TRT_CP	-0.53	0.22	MOA_CLASS	IKK_INHIBITOR
		ASC	TRT_CP	-0.52	0.58	PCL	CP_VEGFR_INHIBITOR
		ASC	TRT_CP	-0.51	-0.00	PCL	CP_T_TYPE_CALCIUM_CHANNEL_BLOCKER

Figure 12: CMAP results using the *adipose* tissue composite transcript as an input. Table includes the top 30 results derived *only from normal adipocytes* (ASC) regardless of significance. The results are sorted by the correlation of the query to the input with the most negative results at the top.

id	norm_cs	cell_iname	pert_type	raw_cs	fdr_q_nlog10	set_type	src_set_id
		YAPC	TRT_CP	-1.00	0.67	MOA_CLASS	ABL_KINASE_INHIBITOR
		YAPC	TRT_CP	-0.99	0.66	PCL	CP_CDK_INHIBITOR
		YAPC	TRT_CP	-0.97	1.41	PCL	CP_TOPOISOMERASE_INHIBITOR
		YAPC	TRT_CP	-0.95	0.70	MOA_CLASS	THYMIDYLATE_SYNTHASE_INHIBITOR
		YAPC	TRT_CP	-0.95	0.62	MOA_CLASS	ADRENERGIC_INHIBITOR
		YAPC	TRT_CP	-0.94	0.50	MOA_CLASS	BENZODIAZEPINE_RECEPTOR_ANTAGONIST
		YAPC	TRT_CP	-0.89	0.63	PCL	CP_RIBONUCLEOTIDE_REDUCTASE_INHIBITOR
		YAPC	TRT_CP	-0.88	0.52	MOA_CLASS	VASOPRESSIN_RECEPTOR_ANTAGONIST
		YAPC	TRT_CP	-0.85	0.63	MOA_CLASS	ANGIOTENSIN_RECEPTOR_ANTAGONIST
		YAPC	TRT_CP	-0.85	0.33	PCL	CP_CANNABINOID_RECEPTOR_AGONIST
		YAPC	TRT_CP	-0.84	0.30	PCL	CP_RETINOID_RECEPTOR_AGONIST
		YAPC	TRT_CP	-0.83	1.19	MOA_CLASS	NFKB_PATHWAY_INHIBITOR
		YAPC	TRT_CP	-0.83	0.54	MOA_CLASS	DNA_ALKYLATING_DRUG
		YAPC	TRT_CP	-0.80	0.50	MOA_CLASS	CHOLESTEROL_INHIBITOR
		YAPC	TRT_CP	-0.79	0.15	MOA_CLASS	SULFONYLUREA
		YAPC	TRT_CP	-0.78	0.52	MOA_CLASS	HIV_INTEGRASE_INHIBITOR
		YAPC	TRT_CP	-0.78	0.13	MOA_CLASS	LEUKOTRIENE_INHIBITOR
		YAPC	TRT_CP	-0.78	0.45	PCL	CP_PPAR_RECEPTOR_AGONIST
		YAPC	TRT_CP	-0.78	0.54	MOA_CLASS	INSULIN_SENSITIZER
		YAPC	TRT_CP	-0.77	0.51	MOA_CLASS	ESTROGEN_RECEPTOR_ANTAGONIST
		YAPC	TRT_CP	-0.77	0.76	MOA_CLASS	DNA_SYNTHESIS_INHIBITOR
		YAPC	TRT_XPR	-0.77	0.67	PATHWAY_SET	BIOCARTA_PARKIN_PATHWAY
		YAPC	TRT_CP	-0.77	0.51	PCL	CP_VEGFR_INHIBITOR
		YAPC	TRT_CP	-0.75	0.39	MOA_CLASS	RNA_SYNTHESIS_INHIBITOR
		YAPC	TRT_CP	-0.72	0.60	MOA_CLASS	BCR-ABL_KINASE_INHIBITOR
		YAPC	TRT_XPR	-0.71	0.66	PATHWAY_SET	BIOCARTA_EIF_PATHWAY
		YAPC	TRT_XPR	-0.69	0.54	PATHWAY_SET	PID_CIRCADIAN_PATHWAY
		YAPC	TRT_CP	-0.68	0.77	MOA_CLASS	TOPOISOMERASE_INHIBITOR
		YAPC	TRT_XPR	-0.64	0.49	PATHWAY_SET	BIOCARTA_CBL_PATHWAY
		YAPC	TRT_CP	-0.64	0.53	MOA_CLASS	TUBULIN_INHIBITOR

Figure 13: CMAP results using the *pancreatic islet* composite transcript as an input. Table includes the top 30 results derived *only from YAPC cells*, which are derived from pancreatic carcinoma cells. Results are shown regardless of significance and are sorted by the correlation of the query to the input with the most negative results at the top.

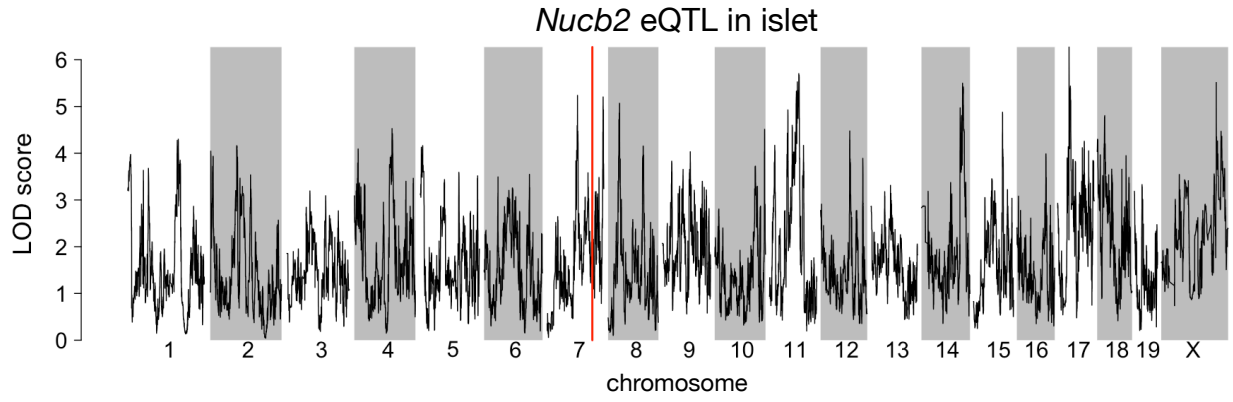


Figure 14: Regulation of *Nucb2* expression in islet. *Nucb2* is encoded on mouse chromosome 7 at 116.5 Mb (red line). In islets the heritability of *Nucb2* expression levels is 69% heritable. This LOD score trace shows that there is no local eQTL at the position of the gene, nor any strong distal eQTLs anywhere else in the genome.

References

[1] M. T. Maurano, R. Humbert, E. Rynes, R. E. Thurman, E. Haugen, H. Wang, A. P. Reynolds, R. Sandstrom, H. Qu, J. Brody, A. Shafer, F. Neri, K. Lee, T. Kutyavin, S. Stehling-Sun, A. K. Johnson, T. K. Canfield, E. Giste, M. Diegel, D. Bates, R. S. Hansen, S. Neph, P. J. Sabo, S. Heimfeld, A. Raubitschek, S. Ziegler, C. Cotsapas, N. Sotoodehnia, I. Glass, S. R. Sunyaev, R. Kaul, and J. A. Stamatoyannopoulos. Systematic localization of common disease-associated variation in regulatory DNA. *Science*, 337(6099):1190–1195, Sep 2012.

[2] K. K. Farh, A. Marson, J. Zhu, M. Kleinewietfeld, W. J. Housley, S. Beik, N. Shoresh, H. Whitton, R. J. Ryan, A. A. Shishkin, M. Hatan, M. J. Carrasco-Alfonso, D. Mayer, C. J. Luckey, N. A. Patsopoulos, P. L. De Jager, V. K. Kuchroo, C. B. Epstein, M. J. Daly, D. A. Hafler, and B. E. Bernstein. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, 518(7539):337–343, Feb 2015.

[3] E. Pennisi. The Biology of Genomes. Disease risk links to gene regulation. *Science*, 332(6033):1031, May 2011.

[4] L. A. Hindorff, P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins, and T. A. Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci*, 106(23):9362–9367, Jun 2009.

[5] J. K. Pickrell. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am J Hum Genet*, 94(4):559–573, Apr 2014.

[6] D. Welter, J. MacArthur, J. Morales, T. Burdett, P. Hall, H. Junkins, A. Klemm, P. Flicek, T. Manolio, L. Hindorff, and H. Parkinson. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res*, 42(Database issue):D1001–1006, Jan 2014.

[7] Y. I. Li, B. van de Geijn, A. Raj, D. A. Knowles, A. A. Petti, D. Golan, Y. Gilad, and J. K. Pritchard. RNA splicing is a primary link between genetic variation and disease. *Science*, 352(6285):600–604, Apr 2016.

[8] D. Zhou, Y. Jiang, X. Zhong, N. J. Cox, C. Liu, and E. R. Gamazon. A unified framework for joint-tissue transcriptome-wide association and Mendelian randomization analysis. *Nat Genet*, 52(11):1239–1246, Nov 2020.

[9] E. R. Gamazon, H. E. Wheeler, K. P. Shah, S. V. Mozaffari, K. Aquino-Michaels, R. J. Carroll, A. E.

Eyler, J. C. Denny, D. L. Nicolae, N. J. Cox, and H. K. Im. A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet*, 47(9):1091–1098, Sep 2015.

[10] Z. Zhu, F. Zhang, H. Hu, A. Bakshi, M. R. Robinson, J. E. Powell, G. W. Montgomery, M. E. Goddard, N. R. Wray, P. M. Visscher, and J. Yang. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet*, 48(5):481–487, May 2016.

[11] A. Gusev, A. Ko, H. Shi, G. Bhatia, W. Chung, B. W. Penninx, R. Jansen, E. J. de Geus, D. I. Boomsma, F. A. Wright, P. F. Sullivan, E. Nikkola, M. Alvarez, M. Civelek, A. J. Lusis, T. ki, E. Raitoharju, M. nen, I. ä, O. T. Raitakari, J. Kuusisto, M. Laakso, A. L. Price, P. Pajukanta, and B. Pasaniuc. Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet*, 48(3):245–252, Mar 2016.

[12] M. P. Keller, D. M. Gatti, K. L. Schueler, M. E. Rabaglia, D. S. Stapleton, P. Simecek, M. Vincent, S. Allen, A. T. Broman, R. Bacher, C. Kendziorski, K. W. Broman, B. S. Yandell, G. A. Churchill, and A. D. Attie. Genetic Drivers of Pancreatic Islet Function. *Genetics*, 209(1):335–356, May 2018.

[13] W. L. Crouse, G. R. Keele, M. S. Gastonguay, G. A. Churchill, and W. Valdar. A Bayesian model selection approach to mediation analysis. *PLoS Genet*, 18(5):e1010184, May 2022.

[14] J. M. Chick, S. C. Munger, P. Simecek, E. L. Huttlin, K. Choi, D. M. Gatti, N. Raghupathy, K. L. Svenson, G. A. Churchill, and S. P. Gygi. Defining the consequences of genetic variation on a proteome-wide scale. *Nature*, 534(7608):500–505, Jun 2016.

[15] H. E. Wheeler, S. Ploch, A. N. Barbeira, R. Bonazzola, A. Andaleon, A. Fotuhi Siahpirani, A. Saha, A. Battle, S. Roy, and H. K. Im. Imputed gene associations identify replicable trans-acting genes enriched in transcription pathways and complex traits. *Genet Epidemiol*, 43(6):596–608, Sep 2019.

[16] B. D. Umans, A. Battle, and Y. Gilad. Where Are the Disease-Associated eQTLs? *Trends Genet*, 37(2):109–124, Feb 2021.

[17] N. J. Connally, S. Nazeen, D. Lee, H. Shi, J. Stamatoyannopoulos, S. Chun, C. Cotsapas, C. A. Cassa, and S. R. Sunyaev. The missing link between genetic association and regulatory function. *Elife*, 11, Dec 2022.

[18] H. Mostafavi, J. P. Spence, S. Naqvi, and J. K. Pritchard. Systematic differences in discovery of genetic effects on gene expression and complex traits. *Nat Genet*, 55(11):1866–1875, Nov 2023.

[19] D. W. Yao, L. J. O'Connor, A. L. Price, and A. Gusev. Quantifying genetic effects on disease mediated by assayed gene expression levels. *Nat Genet*, 52(6):626–633, Jun 2020.

[20] X. Liu, J. A. Mefford, A. Dahl, Y. He, M. Subramaniam, A. Battle, A. L. Price, and N. Zaitlen. GBAT: a gene-based association test for robust detection of trans-gene regulation. *Genome Biol*, 21(1):211, Aug 2020.

[21] H. J. Westra, M. J. Peters, T. Esko, H. Yaghootkar, C. Schurmann, J. Kettunen, M. W. Christiansen, B. P. Fairfax, K. Schramm, J. E. Powell, A. Zhernakova, D. V. Zhernakova, J. H. Veldink, L. H. Van den Berg, J. Karjalainen, S. Withoff, A. G. Uitterlinden, A. Hofman, F. Rivadeneira, P. A. C. ' Hoen, E. Reinmaa, K. Fischer, M. Nelis, L. Milani, D. Melzer, L. Ferrucci, A. B. Singleton, D. G. Hernandez, M. A. Nalls, G. Homuth, M. Nauck, D. Radke, U. lker, M. Perola, V. Salomaa, J. Brody, A. Suchy-Dicey, S. A. Gharib, D. A. Enquobahrie, T. Lumley, G. W. Montgomery, S. Makino, H. Prokisch, C. Herder, M. Roden, H. Grallert, T. Meitinger, K. Strauch, Y. Li, R. C. Jansen, P. M. Visscher, J. C. Knight, B. M. Psaty, S. Ripatti, A. Teumer, T. M. Frayling, A. Metspalu, J. B. J. van Meurs, and L. Franke. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat Genet*, 45(10):1238–1243, Oct 2013.

[22] Y. Gilad, S. A. Rifkin, and J. K. Pritchard. Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends Genet*, 24(8):408–415, Aug 2008.

[23] Aparna Nathan, Samira Asgari, Kazuyoshi Ishigaki, Cristian Valencia, Tiffany Amariuta, Yang Luo, Jessica I Beynor, Yuriy Baglaenko, Sara Suliman, Alkes L Price, et al. Single-cell eqtl models reveal dynamic t cell state dependence of disease loci. *Nature*, 606(7912):120–128, 2022.

[24] E. Sollis, A. Mosaku, A. Abid, A. Buniello, M. Cerezo, L. Gil, T. Groza, O. ş, P. Hall, J. Hayhurst, A. Ibrahim, Y. Ji, S. John, E. Lewis, J. A. L. MacArthur, A. McMahon, D. Osumi-Sutherland, K. Panoutsopoulou, Z. Pendlington, S. Ramachandran, R. Stefancsik, J. Stewart, P. Whetzel, R. Wilson, L. Hindorff, F. Cunningham, S. A. Lambert, M. Inouye, H. Parkinson, and L. W. Harris. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res*, 51(D1):D977–D985, Jan 2023.

[25] R. J. F. Loos and G. S. H. Yeo. The genetics of obesity: from discovery to biology. *Nat Rev Genet*, 23(2):120–133, Feb 2022.

[26] R. K. Singh, P. Kumar, and K. Mahalingam. Molecular genetics of human obesity: A comprehensive review. *C R Biol*, 340(2):87–108, Feb 2017.

[27] P. Arner. Obesity–a genetic disease of adipose tissue? *Br J Nutr*, 83 Suppl 1:9–16, Mar 2000.

[28] Mark P Keller, Mary E Rabaglia, Kathryn L Schueler, Donnie S Stapleton, Daniel M Gatti, Matthew

Vincent, Kelly A Mitok, Ziyue Wang, Takanao Ishimura, Shane P Simonett, et al. Gene loci associated with insulin secretion in islets from nondiabetic mice. *The Journal of Clinical Investigation*, 129(10):4419–4432, 2019.

[29] G. A. Churchill, D. M. Gatti, S. C. Munger, and K. L. Svenson. The Diversity Outbred mouse population. *Mamm Genome*, 23(9-10):713–718, Oct 2012.

[30] Elissa J Chesler, Darla R Miller, Lisa R Branstetter, Leslie D Galloway, Barbara L Jackson, Vivek M Philip, Brynn H Voy, Cymbeline T Culiat, David W Threadgill, Robert W Williams, et al. The collaborative cross at oak ridge national laboratory: developing a powerful resource for systems genetics. *Mammalian Genome*, 19:382–389, 2008.

[31] Michael C Saul, Vivek M Philip, Laura G Reinholdt, and Elissa J Chesler. High-diversity mouse populations for complex traits. *Trends in Genetics*, 35(7):501–514, 2019.

[32] D. W. Threadgill, D. R. Miller, G. A. Churchill, and F. P. de Villena. The collaborative cross: a recombinant inbred mouse population for the systems genetic era. *ILAR J*, 52(1):24–31, 2011.

[33] S. M. Clee and A. D. Attie. The genetic landscape of type 2 diabetes in mice. *Endocr Rev*, 28(1):48–83, Feb 2007.

[34] K. W. Broman, D. M. Gatti, P. Simecek, N. A. Furlotte, P. Prins, Ś. Sen, B. S. Yandell, and G. A. Churchill. R/qtl2: Software for Mapping Quantitative Trait Loci with High-Dimensional Data and Multiparent Populations. *Genetics*, 211(2):495–502, Feb 2019.

[35] Klaasjan G Ouwens, Rick Jansen, Michel G Nivard, Jenny van Dongen, Maia J Frieser, Jouke-Jan Hottenga, Wibowo Arindrarto, Annique Claringbould, Maarten van Iterson, Hailiang Mei, et al. A characterization of cis-and trans-heritability of rna-seq-based gene expression. *European Journal of Human Genetics*, 28(2):253–263, 2020.

[36] Alkes L Price, Agnar Helgason, Gudmar Thorleifsson, Steven A McCarroll, Augustine Kong, and Kari Stefansson. Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. *PLoS genetics*, 7(2):e1001317, 2011.

[37] Julien Bryois, Alfonso Buil, David M Evans, John P Kemp, Stephen B Montgomery, Donald F Conrad, Karen M Ho, Susan Ring, Matthew Hurles, Panos Deloukas, et al. Cis and trans effects of human genomic variants on gene expression. *PLoS genetics*, 10(7):e1004461, 2014.

[38] M. Helmer, S. Warrington, A. R. Mohammadi-Nejad, J. L. Ji, A. Howell, B. Rosand, A. Anticevic,

S. N. Sotiropoulos, and J. D. Murray. On the stability of canonical correlation analysis and partial least squares with application to brain-behavior associations. *Commun Biol*, 7(1):217, Feb 2024.

[39] Gennady Korotkevich, Vladimir Sukhov, and Alexey Sergushichev. Fast gene set enrichment analysis. *bioRxiv*, 2019.

[40] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102(43):15545–15550, Oct 2005.

[41] S. Subramanian and A. Chait. The effect of dietary cholesterol on macrophage accumulation in adipose tissue: implications for systemic inflammation and atherosclerosis. *Curr Opin Lipidol*, 20(1):39–44, Feb 2009.

[42] I. Akoumianakis, N. Akawi, and C. Antoniades. Exploring the Crosstalk between Adipose Tissue and the Cardiovascular System. *Korean Circ J*, 47(5):670–685, Sep 2017.

[43] I. S. Stafeev, A. V. Vorotnikov, E. I. Ratner, M. Y. Menshikov, and Y. V. Parfyonova. Latent Inflammation and Insulin Resistance in Adipose Tissue. *Int J Endocrinol*, 2017:5076732, 2017.

[44] I. P. Fischer, M. Irmler, C. W. Meyer, S. J. Sachs, F. Neff, M. de Angelis, J. Beckers, M. H. p, S. M. Hofmann, and S. Ussar. A history of obesity leaves an inflammatory fingerprint in liver and adipose tissue. *Int J Obes (Lond)*, 42(3):507–517, Mar 2018.

[45] S. Chung, H. Cuffe, S. M. Marshall, A. L. McDaniel, J. H. Ha, K. Kavanagh, C. Hong, P. Tontonoz, R. E. Temel, and J. S. Parks. Dietary cholesterol promotes adipocyte hypertrophy and adipose tissue inflammation in visceral, but not in subcutaneous, fat in monkeys. *Arterioscler Thromb Vasc Biol*, 34(9):1880–1887, Sep 2014.

[46] V. Kus, T. Prazak, P. Brauner, M. Hensler, O. Kuda, P. Flachs, P. Janovska, D. Medrikova, M. Rossmeisl, Z. Jilkova, B. Stefl, E. Pastalkova, Z. Drahota, J. Houstek, and J. Kopecky. Induction of muscle thermogenesis by high-fat diet in mice: association with obesity-resistance. *Am J Physiol Endocrinol Metab*, 295(2):E356–367, Aug 2008.

[47] C. B. Newgard. Interplay between lipids and branched-chain amino acids in development of insulin resistance. *Cell Metab*, 15(5):606–614, May 2012.

[48] D. D. Sears, G. Hsiao, A. Hsiao, J. G. Yu, C. H. Courtney, J. M. Ofrecio, J. Chapman, and S. Subramaniam.

Mechanisms of human insulin resistance and thiazolidinedione-mediated insulin sensitization. *Proc Natl Acad Sci U S A*, 106(44):18745–18750, Nov 2009.

[49] R. Stienstra, C. Duval, M. ller, and S. Kersten. PPARs, Obesity, and Inflammation. *PPAR Res*, 2007:95974, 2007.

[50] O. Gavrilova, M. Haluzik, K. Matsusue, J. J. Cutson, L. Johnson, K. R. Dietz, C. J. Nicol, C. Vinson, F. J. Gonzalez, and M. L. Reitman. Liver peroxisome proliferator-activated receptor gamma contributes to hepatic steatosis, triglyceride clearance, and regulation of body fat mass. *J Biol Chem*, 278(36):34268–34276, Sep 2003.

[51] K. Matsusue, M. Haluzik, G. Lambert, S. H. Yim, O. Gavrilova, J. M. Ward, B. Brewer, M. L. Reitman, and F. J. Gonzalez. Liver-specific disruption of PPARgamma in leptin-deficient mice improves fatty liver but aggravates diabetic phenotypes. *J Clin Invest*, 111(5):737–747, Mar 2003.

[52] D. Patsouris, J. K. Reddy, M. ller, and S. Kersten. Peroxisome proliferator-activated receptor alpha mediates the effects of high-fat diet on hepatic gene expression. *Endocrinology*, 147(3):1508–1516, Mar 2006.

[53] S. E. Schadinger, N. L. Bucher, B. M. Schreiber, and S. R. Farmer. PPARgamma2 regulates lipogenesis and lipid accumulation in steatotic hepatocytes. *Am J Physiol Endocrinol Metab*, 288(6):E1195–1205, Jun 2005.

[54] W. Motomura, M. Inoue, T. Ohtake, N. Takahashi, M. Nagamine, S. Tanno, Y. Kohgo, and T. Okumura. Up-regulation of ADRP in fatty liver in human and liver steatosis in mice fed with high fat diet. *Biochem Biophys Res Commun*, 340(4):1111–1118, Feb 2006.

[55] A. Srivastava, A. P. Morgan, M. L. Najarian, V. K. Sarsani, J. S. Sigmon, J. R. Shorter, A. Kashfeen, R. C. McMullan, L. H. Williams, P. guez, M. T. Ferris, P. Sullivan, P. Hock, D. R. Miller, T. A. Bell, L. McMillan, G. A. Churchill, and F. P. de Villena. Genomes of the Mouse Collaborative Cross. *Genetics*, 206(2):537–556, Jun 2017.

[56] A. Roberts, F. Pardo-Manuel de Villena, W. Wang, L. McMillan, and D. W. Threadgill. The polymorphism architecture of mouse genetic resources elucidated using genome-wide resequencing data: implications for QTL discovery and systems genetics. *Mamm Genome*, 18(6-7):473–481, Jul 2007.

[57] G. A. Churchill, D. C. Airey, H. Allayee, J. M. Angel, A. D. Attie, J. Beatty, W. D. Beavis, J. K. Belknap, B. Bennett, W. Berrettini, A. Bleich, M. Bogue, K. W. Broman, K. J. Buck, E. Buckler, M. Burmeister, E. J. Chesler, J. M. Cheverud, S. Clapcote, M. N. Cook, R. D. Cox, J. C. Crabbe,

W. E. Crusio, A. Darvasi, C. F. Deschepper, R. W. Doerge, C. R. Farber, J. Forejt, D. Gaile, S. J. Garlow, H. Geiger, H. Gershenfeld, T. Gordon, J. Gu, W. Gu, G. de Haan, N. L. Hayes, C. Heller, H. Himmelbauer, R. Hitzemann, K. Hunter, H. C. Hsu, F. A. Iraqi, B. Ivandic, H. J. Jacob, R. C. Jansen, K. J. Jepsen, D. K. Johnson, T. E. Johnson, G. Kempermann, C. Kendziorski, M. Kotb, R. F. Kooy, B. Llamas, F. Lammert, J. M. Lassalle, P. R. Lowenstein, L. Lu, A. Lusis, K. F. Manly, R. Marcucio, D. Matthews, J. F. Medrano, D. R. Miller, G. Mittleman, B. A. Mock, J. S. Mogil, X. Montagutelli, G. Morahan, D. G. Morris, R. Mott, J. H. Nadeau, H. Nagase, R. S. Nowakowski, B. F. O'Hara, A. V. Osadchuk, G. P. Page, B. Paigen, K. Paigen, A. A. Palmer, H. J. Pan, L. Peltonen-Palotie, J. Peirce, D. Pomp, M. Pravenec, D. R. Prows, Z. Qi, R. H. Reeves, J. Roder, G. D. Rosen, E. E. Schadt, L. C. Schalkwyk, Z. Seltzer, K. Shimomura, S. Shou, M. J. ä, L. D. Siracusa, H. W. Snoeck, J. L. Spearow, K. Svenson, L. M. Tarantino, D. Threadgill, L. A. Toth, W. Valdar, F. P. de Villena, C. Warden, S. Whatley, R. W. Williams, T. Wiltshire, N. Yi, D. Zhang, M. Zhang, and F. Zou. The Collaborative Cross, a community resource for the genetic analysis of complex traits. *Nat Genet*, 36(11):1133–1137, Nov 2004.

[58] J. Y. Huh, Y. J. Park, M. Ham, and J. B. Kim. Crosstalk between adipocytes and immune cells in adipose tissue inflammation and metabolic dysregulation in obesity. *Mol Cells*, 37(5):365–371, May 2014.

[59] J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat, M. J. Wrobel, J. Lerner, J. P. Brunet, A. Subramanian, K. N. Ross, M. Reich, H. Hieronymus, G. Wei, S. A. Armstrong, S. J. Haggarty, P. A. Clemons, R. Wei, S. A. Carr, E. S. Lander, and T. R. Golub. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313(5795):1929–1935, Sep 2006.

[60] Aravind Subramanian, Rajiv Narayan, Steven M Corsello, David D Peck, Ted E Natoli, Xiaodong Lu, Joshua Gould, John F Davis, Andrew A Tubelli, Jacob K Asiedu, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6):1437–1452, 2017.

[61] X. Liu, Y. I. Li, and J. K. Pritchard. Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. *Cell*, 177(4):1022–1034, May 2019.

[62] U. Vosa, A. Claringbould, H. J. Westra, M. J. Bonder, P. Deelen, B. Zeng, H. Kirsten, A. Saha, R. Kreuzhuber, S. Yazar, H. Brugge, R. Oelen, D. H. de Vries, M. G. P. van der Wijst, S. Kasela, N. Pervjakova, I. Alves, M. J. é, M. Agbessi, M. W. Christiansen, R. Jansen, I. ä, L. Tong, A. Teumer, K. Schramm, G. Hemani, J. Verlouw, H. Yaghootkar, R. nmez Flitman, A. Brown, V. Kukushkina, A. Kalnapenkis, S. eger, E. Porcu, J. Kronberg, J. Kettunen, B. Lee, F. Zhang, T. Qi, J. A. Hernandez,

W. Arindrarto, F. Beutner, J. Dmitrieva, M. Elansary, B. P. Fairfax, M. Georges, B. T. Heijmans, A. W. Hewitt, M. nen, Y. Kim, J. C. Knight, P. Kovacs, K. Krohn, S. Li, M. Loeffler, U. M. Marigorta, H. Mei, Y. Momozawa, M. ller Nurasyid, M. Nauck, M. G. Nivard, B. W. J. H. Penninx, J. K. Pritchard, O. T. Raitakari, O. Rotzschke, E. P. Slagboom, C. D. A. Stehouwer, M. Stumvoll, P. Sullivan, P. A. C. 't Hoen, J. Thiery, A. njes, J. van Dongen, M. van Iterson, J. H. Veldink, U. lker, R. Warmerdam, C. Wijmenga, M. Swertz, A. Andiappan, G. W. Montgomery, S. Ripatti, M. Perola, Z. Kutalik, E. Dermitzakis, S. Bergmann, T. Frayling, J. van Meurs, H. Prokisch, H. Ahsan, B. L. Pierce, T. ki, D. I. Boomsma, B. M. Psaty, S. A. Gharib, P. Awadalla, L. Milani, W. H. Ouwehand, K. Downes, O. Stegle, A. Battle, P. M. Visscher, J. Yang, M. Scholz, J. Powell, G. Gibson, T. Esko, L. Franke, P. A. C. 't Hoen, J. van Meurs, J. van Dongen, M. van Iterson, M. A. Swertz, and M. Jan Bonder. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat Genet*, 53(9):1300–1310, Sep 2021.

[63] B. Hallgrimsson, R. M. Green, D. C. Katz, J. L. Fish, F. P. Bernier, C. C. Roseman, N. M. Young, J. M. Cheverud, and R. S. Marcucio. The developmental-genetics of canalization. *Semin Cell Dev Biol*, 88:67–79, Apr 2019.

[64] M. L. Siegal and A. Bergman. Waddington's canalization revisited: developmental stability and evolution. *Proc Natl Acad Sci U S A*, 99(16):10528–10532, Aug 2002.

[65] A. B. Paaby and G. Gibson. Cryptic Genetic Variation in Evolutionary Developmental Genetics. *Biology (Basel)*, 5(2), Jun 2016.

[66] E. A. Boyle, Y. I. Li, and J. K. Pritchard. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell*, 169(7):1177–1186, Jun 2017.

[67] Naomi R Wray, Cisca Wijmenga, Patrick F Sullivan, Jian Yang, and Peter M Visscher. Common disease is more complex than implied by the core gene omnigenic model. *Cell*, 173(7):1573–1580, 2018.

[68] H. Shimizu and A. Osaki. Nesfatin/Nucleobindin-2 (NUCB2) and Glucose Homeostasis. *Curr Hypertens Rev*, pages Nesfatin/Nucleobindin–2 (NUCB2) and Glucose Homeostasis., Jul 2014.

[69] M. Nakata and T. Yada. Role of NUCB2/nesfatin-1 in glucose control: diverse functions in islets, adipocytes and brain. *Curr Pharm Des*, 19(39):6960–6965, 2013.

[70] M. Riva, M. D. Nitert, U. Voss, R. Sathanoori, A. Lindqvist, C. Ling, and N. Wierup. Nesfatin-1 stimulates glucagon and insulin secretion and beta cell NUCB2 is reduced in human type 2 diabetic subjects. *Cell Tissue Res*, 346(3):393–405, Dec 2011.

[71] A. P. Morgan, C. P. Fu, C. Y. Kao, C. E. Welsh, J. P. Didion, L. Yadgary, L. Hyacinth, M. T. Ferris, T. A. Bell, D. R. Miller, P. Giusti-Rodriguez, R. J. Nonneman, K. D. Cook, J. K. Whitmire, L. E. Gralinski, M. Keller, A. D. Attie, G. A. Churchill, P. Petkov, P. F. Sullivan, J. R. Brennan, L. McMillan, and F. Pardo-Manuel de Villena. The Mouse Universal Genotyping Array: From Substrains to Subspecies. *G3 (Bethesda)*, 6(2):263–279, Dec 2015.

[72] K. L. Svenson, D. M. Gatti, W. Valdar, C. E. Welsh, R. Cheng, E. J. Chesler, A. A. Palmer, L. McMillan, and G. A. Churchill. High-resolution genetic mapping using the Mouse Diversity outbred population. *Genetics*, 190(2):437–447, Feb 2012.

[73] D. M. Gatti, K. L. Svenson, A. Shabalin, L. Y. Wu, W. Valdar, P. Simecek, N. Goodwin, R. Cheng, D. Pomp, A. Palmer, E. J. Chesler, K. W. Broman, and G. A. Churchill. Quantitative trait locus mapping methods for diversity outbred mice. *G3 (Bethesda)*, 4(9):1623–1633, Sep 2014.

[74] Kwangbom Choi, Hao He, Daniel M Gatti, Vivek M Philip, Narayanan Raghupathy, Steven C Munger, Elissa J Chesler, and Gary A Churchill. Genotype-free individual genome reconstruction of multiparental population models by rna sequencing data. *bioRxiv*, pages 2020–10, 2020.

[75] Narayanan Raghupathy, Kwangbom Choi, Matthew J Vincent, Glen L Beane, Keith S Sheppard, Steven C Munger, Ron Korstanje, Fernando Pardo-Manual de Villena, and Gary A Churchill. Hierarchical analysis of rna-seq reads improves the accuracy of allele-specific expression. *Bioinformatics*, 34(13):2177–2184, 2018.

[76] S. C. Munger, N. Raghupathy, K. Choi, A. K. Simons, D. M. Gatti, D. A. Hinerfeld, K. L. Svenson, M. P. Keller, A. D. Attie, M. A. Hibbs, J. H. Graber, E. J. Chesler, and G. A. Churchill. RNA-Seq alignment to individualized genomes improves transcript abundance estimates in multiparent populations. *Genetics*, 198(1):59–73, Sep 2014.

[77] Gary A Churchill and Rachel W Doerge. Empirical threshold values for quantitative trait mapping. *Genetics*, 138(3):963–971, 1994.

[78] Kenneth A Bollen. *Structural equations with latent variables*. John Wiley & Sons, 2014.

[79] Arthur Tenenhaus and Michel Tenenhaus. Regularized generalized canonical correlation analysis. *Psychometrika*, 76:257–284, 2011.

[80] Michel Tenenhaus, Arthur Tenenhaus, and Patrick JF Groenen. Regularized generalized canonical correlation analysis: a framework for sequential multiblock component methods. *Psychometrika*, 82(3):737–777, 2017.

[81] Arthur Tenenhaus, Cathy Philippe, and Vincent Frouin. Kernel generalized canonical correlation analysis. *Computational Statistics & Data Analysis*, 90:114–131, 2015.

[82] Fabien Girka, Etienne Camenen, Caroline Peltier, Arnaud Gloaguen, Vincent Guillemot, Laurent Le Brusquet, and Arthur Tenenhaus. Multiblock data analysis with the rgcca package. *Journal of Statistical Software*, pages 1–36, 2023.

[83] Juliane Schäfer and Korbinian Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, 4(1), 2005.

[84] M. Kanehisa, M. Furumichi, Y. Sato, M. Kawashima, and M. Ishiguro-Watanabe. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res*, 51(D1):D587–D592, Jan 2023.

[85] W. Luo and C. Brouwer. Pathview: an R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics*, 29(14):1830–1831, Jul 2013.

[86] J. A. Blake, R. Baldarelli, J. A. Kadin, J. E. Richardson, C. L. Smith, C. J. Bult, A. V. Anagnostopoulos, J. S. Beal, S. M. Bello, O. Blodgett, N. E. Butler, J. Campbell, K. R. Christie, L. E. Corbani, M. E. Dolan, H. J. Drabkin, M. Flores, S. L. Giannatto, A. Guerra, P. Hale, D. P. Hill, J. Judd, M. Law, M. McAndrews, D. Miers, C. Mitchell, H. Motenko, L. Ni, H. Onda, J. Ormsby, M. Perry, J. M. Recla, D. Shaw, D. Sitnikov, M. Tomczuk, L. Wilming, and Y. ' Zhu. Mouse Genome Database (MGD): Knowledgebase for mouse-human comparative biology. *Nucleic Acids Res*, 49(D1):D981–D987, Jan 2021.

[87] Jeff Gentry. *annotate: Annotation for microarrays*, 2024. R package version 1.82.0.

[88] L. Gatto, L. M. Breckels, S. Wieczorek, T. Burger, and K. S. Lilley. Mass-spectrometry-based spatial proteomics data analysis using pRoloc and pRolocdata. *Bioinformatics*, 30(9):1322–1324, May 2014.

[89] Damiano Fantini. *easyPubMed: Search and Retrieve Scientific Publication Records from PubMed*, 2019. R package version 2.13.

[90] Martin Maechler, Peter Rousseeuw, Anja Struyf, Mia Hubert, and Kurt Hornik. *cluster: Cluster Analysis Basics and Extensions*, 2023. R package version 2.1.6 — For new features, see the 'NEWS' and the 'Changelog' file in the package source).

[91] Liis Kolberg, Uku Raudvere, Ivan Kuzmin, Jaak Vilo, and Hedi Peterson. gprofiler2– an r package for gene list functional enrichment analysis and namespace conversion toolset g:profiler. *F1000Research*, 9 (ELIXIR)(709), 2020. R package version 0.2.3.

[92] E. Clough, T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, H. Lee, N. Zhang, N. Serova, L. Wagner, V. Zalunin, A. Kochergin, and A. Soboleva. NCBI GEO: archive for gene expression and epigenomics data sets: 23-year update. *Nucleic Acids Res*, 52(D1):D138–D144, Jan 2024.

[93] R. Edgar, M. Domrachev, and A. E. Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*, 30(1):207–210, Jan 2002.

[94] Sean Davis and Paul Meltzer. Geoquery: a bridge between the gene expression omnibus (geo) and bioconductor. *Bioinformatics*, 14:1846–1847, 2007.

[95] R. M. Baldarelli, C. L. Smith, M. Ringwald, J. E. Richardson, C. J. Bult, A. Anagnostopoulos, D. A. Begley, S. M. Bello, K. Christie, J. H. Finger, P. Hale, T. F. Hayamizu, D. P. Hill, M. N. Knowlton, D. M. Krupke, M. McAndrews, M. Law, I. J. McCright, L. Ni, H. Onda, D. Sitnikov, C. M. Smith, M. Tomczuk, L. Wilming, J. Xu, Y. Zhu, O. Blodgett, J. W. Campbell, L. E. Corbani, P. Frost, S. C. Giannatto, D. B. Miers, H. Motenko, S. B. Neuhauser, D. R. Shaw, N. E. Butler, and J. E. Ormsby. Mouse Genome Informatics: an integrated knowledgebase system for the laboratory mouse. *Genetics*, 227(1), May 2024.