Transcripts with high distal heritability mediate genetic effects on complex traits

Abstract

The transcriptome is increasingly viewed as a bridge between genetic risk factors for complex disease and their associated pathophysiology. Powerful insights into disease mechanism can be made by linking genetic variants affecting gene expression (expression quantiatitive trait loci - eQTLs) to phenotypes.

Introduction

In the quest to understand genetic contributions to complex traits, gene expression is an important bridge between genotype and phenotype. By identifying transcripts that mediate the effect of genetic loci on traits, we get one step closer to a mechanistic understanding of the influence of genetic variants on traits. There is evidence from genome-wide association studies (GWAS) that regulation of gene expression accounts for the bulk of the genetic effect on complex traits, as most trait-associated variants lie in gene regulatory regions [1, 2, 3, 4, 5, 6, 7]. It is widely assumed that these variants influence local transcription, and methods such as transcription-wide association studies (TWAS) [8, 9, 10, 11] summary data-based Mendelian randomization (SMR) [10], and others [cite] have capitalized on this idea to identify genes associated with multiple disease traits [cite many]

Despite the great promise of these methods, however, they have not been as widely successful as it seemed they could have been, and the vast majority of complex trait heritability remains unexplained. Although trait-associated variants tend to lie in non-coding, regulatory regions, they often do not have detectable effects on gene expression [12] and tend not to co-localize with expression quantitative trait loci (eQTLs) [13, 14].

One possible explanation for these observations is that we are not measuring gene expression in the appropriate cell types and thus are unable to detect true eQTLs influencing traits [12]. An alternative explanation that has been discussed in recent years is that heritability of these variants is mediated not through local regulation of gene expression, but through distal regulation [14, 15, 16, 17]. Yao *et al.* [15] observed that genes with

low local heritability explained more expression-mediated disease heritability than genes with high local heritability. This observation is consistent with principles of robustness in complex systems. If a transcript were both important to a trait and subject to strong local regulation, a population would be susceptible to extremes in phenotype that might frequently cross the threshold to disease. Indeed, strong disruption of highly trait-relevant genes is the cause of Mendelian disease.

Rather, observations suggest that genes near GWAS hits and have obvious functional relevance to a trait tend to have highly complex regulatory landscapes under strong selection pressures [14]. In contrast, genes with strong local regulation tend to be depleted of functional annotations and are under looser selection constraints [14]. These observations and others led Liu et al. [18] to suggest that most heritability of complex traits is driven by weak trans-eQTLs. They proposed a framework of understanding heritability of complex traits in which massive polygenicity is distributed across common variants in both functional "core genes'', as well as more peripheral genes that may not seem obviously related to the trait.

Assessing the role of wide-spread distal gene regulation on complex traits requires a large, dedicated data set that includes high-dimensional clinically relevant phenotyping, dense genotyping in a highly recombined population, and transcriptome-wide measurements of gene expression in multiple tissues. Measuring gene expression in multiple tissues is key to adequately assess the extent to which local gene regulation varies across multiple tissues and whether such variablilty might account for previous failed attempts to identify trait-relevant local eQTL. Thus to investigate further the role of local and distal gene regulation on complex traits, we have generated such a data set in a large population of diversity outbred (DO) mice [19] in a population model of diet-induced obesity and metabolic disease.

The DO mice were derived from eight inbred founder mouse strains, five classical lab strains, and three strains more recently derived from wild mice [19]. They represent three subspecies of mouse *Mus musculus domesticus*, *Mus musculus musculus*, and *Mus musculus casteneus*, and capture 90% of the known variation in laboratory mice [cite]. They are maintained with a breeding scheme that ensures equal contributions from each founder across the genome [19]. We placed a population of 500 male and female mice from generations 18, 19, and 21 on a high-fat, high-sugar diet to induce diet-associated obesity and metabolic disease [20]. Over the experimental period of 18 weeks multiple metabolic traits were measured longitudinally, including body weight, plasma levels of insulin and glucose, and plasma lipids. At the end of the experiment, we used RNASeq to measure gene expression in 384 mice in four tissues relevant to metabolic disease: adipose tissue, pancreatic islets, liver, and skeletal muscle. The mice were also genotyped using the Mouse Universal Genotyping Array (GigaMUGA).

To assess the role of gene regulation in mediating variation in metabolic traits in this population, we propose a high-dimensional mediation (HDM) approach. This is in contrast to the univariate versions of mediation analysis [21] that have been used previously to identify trait relevant eQTLs [20, 22, 23, 17], and are the basis of TWAS, SMR, and other Mendelian randomization approches. In these approaches each transcript is tested independently for mediation of a local variant on a trait. Because genetic effects on traits are widespread and small, we expect that any given transcript may mediate a small amount of the genetic effect on the trait. Transcripts that mediate more are more important to the trait variation seen at the population level and may be candidates for intervention in the case of disease traits. In these univariate approaches local and distal regulation must be treated separately with huge numbers of statistical tests, which is compuatationally expensive, requires strict corrections for multiple testing, and assumes independence of genetic variants and transcripts. Such methods are therefore limited to detecting only the largest statistical effects and are biased toward local gene regulation.

We developed high-dimension mediation to test the omnigenic model with a more holistic approach. This model 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. This hypothesis lends itself well to systems approaches that can account for arbitrarily complex gene regulation, as well as the interconnectedness and redundancy of the transcriptome without explicitly modeling them. The HDM approach we propose here tests the hypothesis of the omnigenic model using regularized and generalized canonical correlation analysis (RGCCA) [24], an extension of canonical correlation analysis (CCA) [cite] that allows for more than two data sets with arbitrary relationships among them. Here we analyze the correlations among three data sets, genotype, transcriptome, and phenotype explicitly modeling mediation in which the transcriptome (T) mediates the effect of the genome (G) on the phenome (P).

$$G \rightarrow T \rightarrow P$$

Because the genome, transcriptome, and phenome had different dimensions, we kernelized each data set prior to running HDM. This step ensured that each set will contribute equally to the solution. The result is a set of three vectors representing a composite transcriptome (T_C) that perfectly mediates the effect of the composite genome (G_C) on the composite phenome (P_C). That is, the partial correlation between G_C and P_C is 0 when T_C is accounted for. Because of the central dogma of molecular biology, information flow is directed out of the genome, and not back into it. Thus, the otherwise undirected relationships between genome, transcriptome, and phenome can be inferred as a causal mediation by the transcriptome of the

effects of the genome on the phenome.

P_C represents a highly heritable combination of traits that is mediated by a heritable combination of transcripts represented by T_C. The loadings of each transcript and each trait on T_C and P_C respectively provide interpretable transcript-level and trait-level contributions to these composite variables. We further combined HDM with heritability analysis in the traits and transcripts to investigate whether the mediating transcripts were regulated primarily locally or distally.

Results

Genetic variation contributes to wide phenotypic variation

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 [20]. Each individual was assessed longitudinally for multiple metabolic measures including fasting glucose levels, glucose tolerance, insulin levels, body weight, and blood lipid levels (Methods).

Although the environment was consistent across animals, 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 B6 body weight to several times the body weight of a B6 adult in both sexes (Fig. 1A). Fasting blood glucose (FBG) also varied considerably (Fig. 1B) although few of the animals had FBG levels that would indicate pre-diabetes (animals,), or diabetes (7 animals, 1.4) according to previously developed cutoffs (pre-diabetes: FBG $\geq$ 250 mg/dL, diabetes: FBG $\geq$ 300, mg/dL) [25]. 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 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 $R^2 = 0.51$, $p = 1.5 \times 10^{-75}$) and fasting blood glucose (FBG) (Fig. 1E, $R^2 = 0.25$, $p = 2 \times 10^{-32}$) suggesting a link between behavioral factors and metabolic disease. However, the heritability of this trait and others (Fig. 1F) indicates that background genetics contribute substantially to correlates of metabolic disease in this population.

The landscape of trait correlations (Fig. 1G) shows that most of the metabolic trait pairs were relatively weakly correlated indicating complex relationships among the measured traits. This low level of redundancy suggests a broad sampling of multiple heritable aspects of metabolic disease including overall body weight, glucose homeostasis, pancreatic composition and liver function.

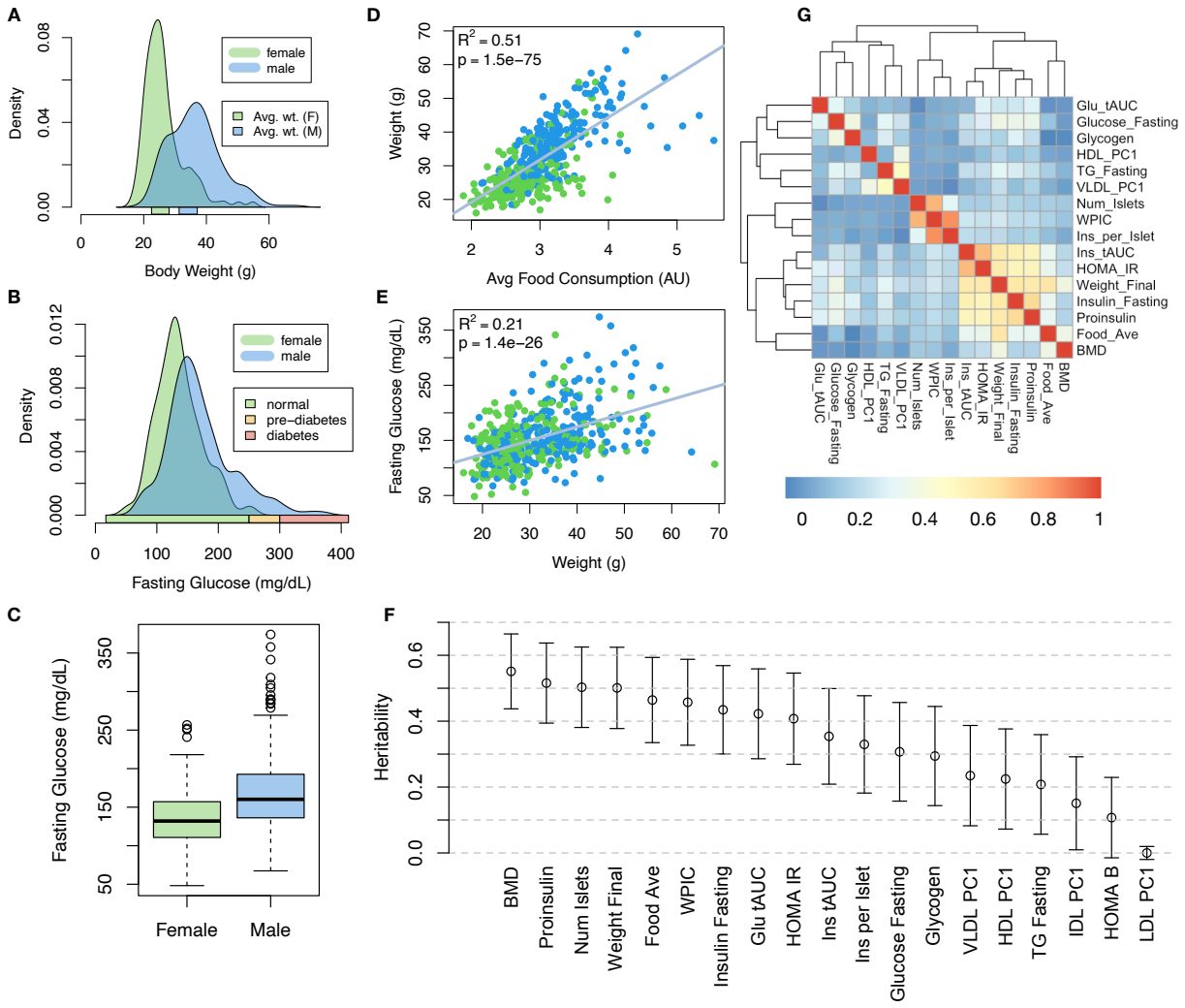

Figure 1: Clinical overview. **A.** Distributions of final body weight in the diversity outbred mice. Sex is indicated by color. The average B6 male and female adult weights at 24 weeks of age are indicated by blue and green bars on the x-axis. **B.** The distribution of final fasting glucose across 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. **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.

Distal Heritability Correlates with Phenotype Relevance

To elaborate the mechanistic details of genetic effects on metabolic phenotypes in the DO population, we also measured gene expression in four tissues known to be involved in metabolic disease: adipose, pancreatic islet, liver, and skeletal muscle. To confirm the heritability of transcript levels, we performed expression QTL analysis using R/qtl2 [cite] (Methods) and identified both local and distal eQTL for transcripts in each tissue (Supp. Fig 9). Significant local eQTLs far outnumbered distal eQTLs (Supp. Fig. 9F) and tended to be

shared across tissues (Supp. Fig. 9G) whereas the few significant distal eQTL we identified tended to be tissue-specific (Supp. Fig. 9H)

To better compare the relative contribution of local and distal genetics to transcript levels, we performed a heritability analysis for each transcript (Methods). Overall, local and distal factors contributed approximately equally to transcript abundance. In all tissues, both local and distal factors explained between 13 and 19% of the variance in the median transcript (Fig 2A).

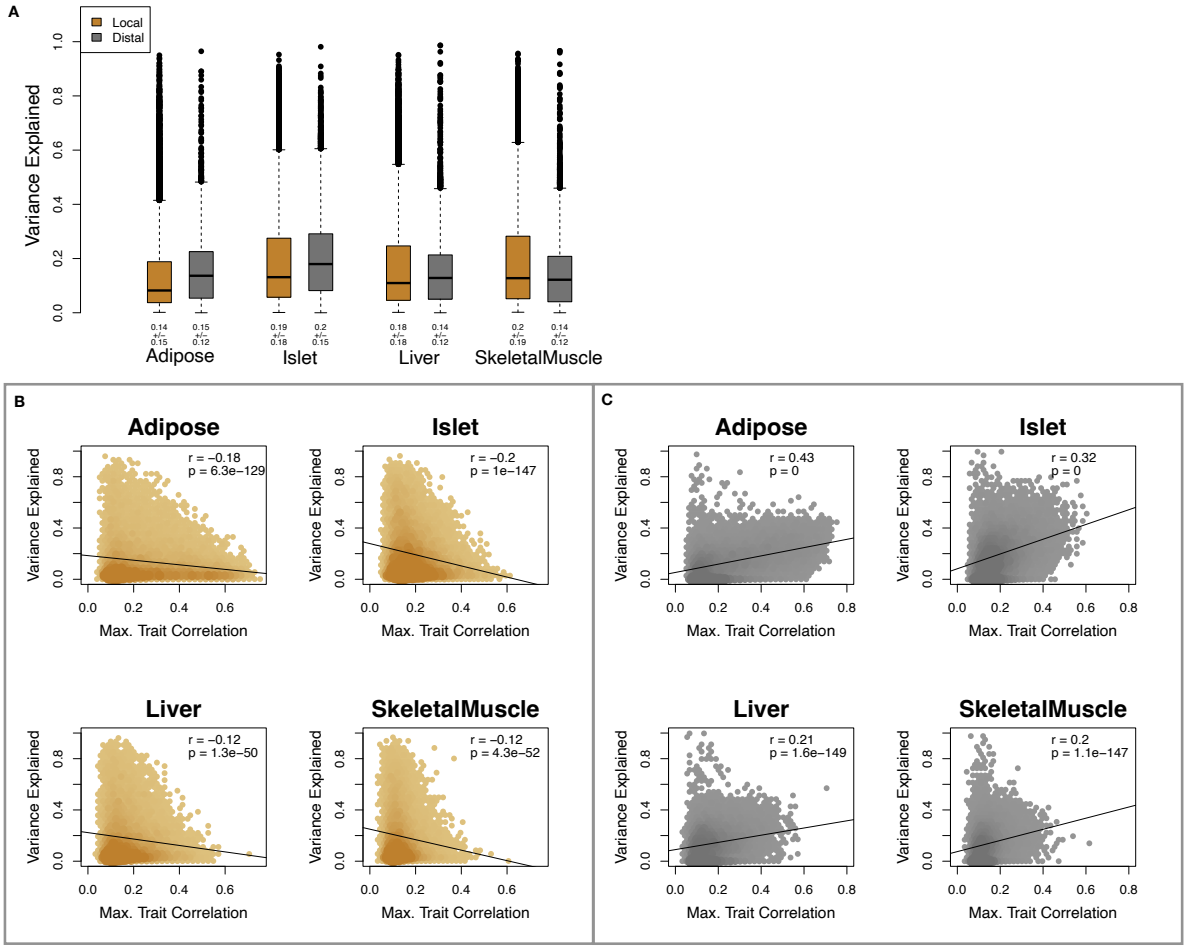

Figure 2: Transcript heritability and trait relevance. **A.** Distributions of distal and local heritability of transcripts across the four tissues. Overall local and distal factors contribute equally to transcript heritability. The relationship between (**B.**) local and (**C.**) distal heritability and trait relevance across all four tissues. Here trait relevance is defined as the maximum correlation between the transcript and all traits. Local heritability is negatively correlated with trait relevance, and distal heritability is positively correlated with trait relevance. Pearson (r) and p values for each correlation are shown in the upper-right of each panel.

Local heritability of transcripts was negatively correlated with their trait relevance, defined as the maximum correlation of a transcript across all traits (Fig. 2B). This suggests that the more local genotype influenced

transcript abundance, the less effect variation in transcript abundance was related to the measured traits. Conversely, distal heritability of transcripts was positively correlated 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. That trait-correlated transcripts have low local heritability is consistent with previous observations that low-heritability transcripts explain more expression-mediated disease heritability than high-heritability transcripts [15]. However, the positive relationship between trait correlation and distal heritability suggests that there are alternative mechanisms through which genetic regulation of transcripts may influence traits.

High-Dimensional Mediation identifies composite transcript that perfectly mediates composite trait

To identify mechanisms through which genetic regulation of transcripts influences heritable traits, we propose high-dimensional mediation (HDM) (Fig. 3). In this process we kernelize each of the genome, transcriptome, and phenome, and perform regularized and sparse generalized canonical correlation analysis (RGCCA) [cite] in which we explicitly model the mediation by the transcriptome of the effect of the genome on the phenome (Methods, Fig. 3). RGCCA is an extended form of canonical correlation analysis (CCA) [cite] in which multiple data sets can be analyzed simultaneously with explicit relationships.

The result of this process is three vectors representing the composite genome (G_C), composite transcriptome (T_C) and the composite phenome (P_C) where the composite transcriptome perfectly mediates the effect of the composite genome on the composite phenome. Each vector is of length n where n is the number of individual mice. Fig. 3A shows the partial correlations between all pairs of composite vectors. The partial correlation r between G_C and T_S was 0.46, and the partial correlation between T_S and P_S was 0.78. However, when the transcriptome was taken into account, the partial correlation between G_S and P_S was effectively 0 (-0.01).

Standard CCA is prone to over-fitting because in any two large matrices it can be trivial to identify highly correlated composite vectors. To assess whether RGCCA was similarly prone to over-fitting in a high-dimensional space, we performed permutation testing. We permuted the individual labels on the transcriptome kernel matrix 1000 times and recalculated the path coefficient, which is the partial correlation of G_C and T_C multiplied by the partial correlation of T_C and P_C. This represents the path from G_C to P_C that is mediated through T_C. The null distribution of the path coefficient is shown in Fig. 3B, and the observed path coefficient from the original data is indicated by the red line. The observed path coefficient was well outside the null distribution generated by permutations. 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. The red dot shows that

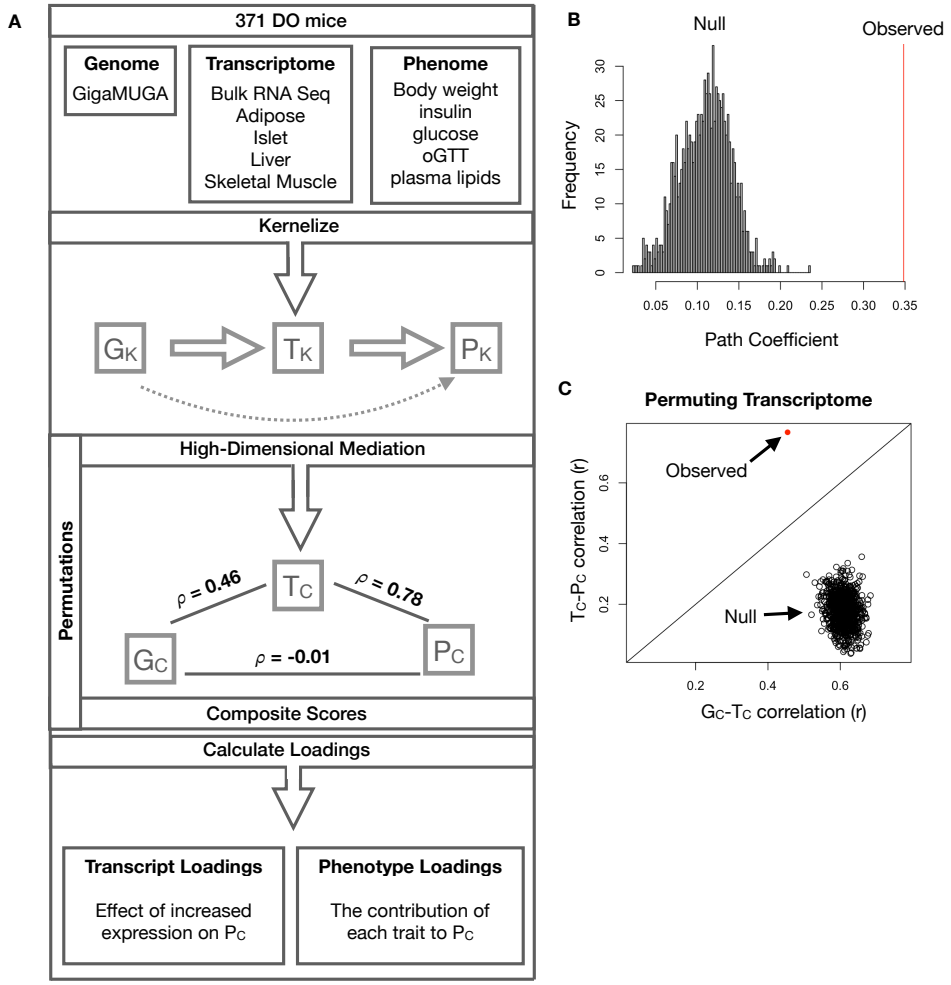

Figure 3: High-dimensional mediation. **A.** Workflow indicating major steps of high-dimensional mediation. The genotype, transcriptome, and phenotype matrices were kernelized to yield single matrices representing the relationships between all individuals for each data modality (G_K = genome kernel, 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 compared to the observed path coefficient (red line). **C.** The null distribution of the G_C-T_C correlation vs. the T_C-P_C correlation compared with the observed value (red dot).

in the real data both the G_C-T_C correlation and the T_C-P_C correlation could be maximized simultaneously suggesting that that path from genotype to phenotype through transcriptome is highly non-trivial and identifiable in this case. These results suggest that these composite vectors represent genetically determined variation in phenotype that is mediated through genetically determined variation in transcription.

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

The loadings of each measured trait onto P_C indicate how much each contributed to P_C. Final body weight contributed the most to P_C (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 index and greater metabolic disease, including higher body weight and higher insulin resistance. We refer to P_C as the metabolic index going forward. Traits contributing the least to the metabolic index were measures of cholesterol and pancreas composition. Thus, when we interpret the transcriptomic signature identified by HDM, we are explaining primarily transcriptional mediation of body weight and insulin resistance, as opposed to cholesterol measurements.

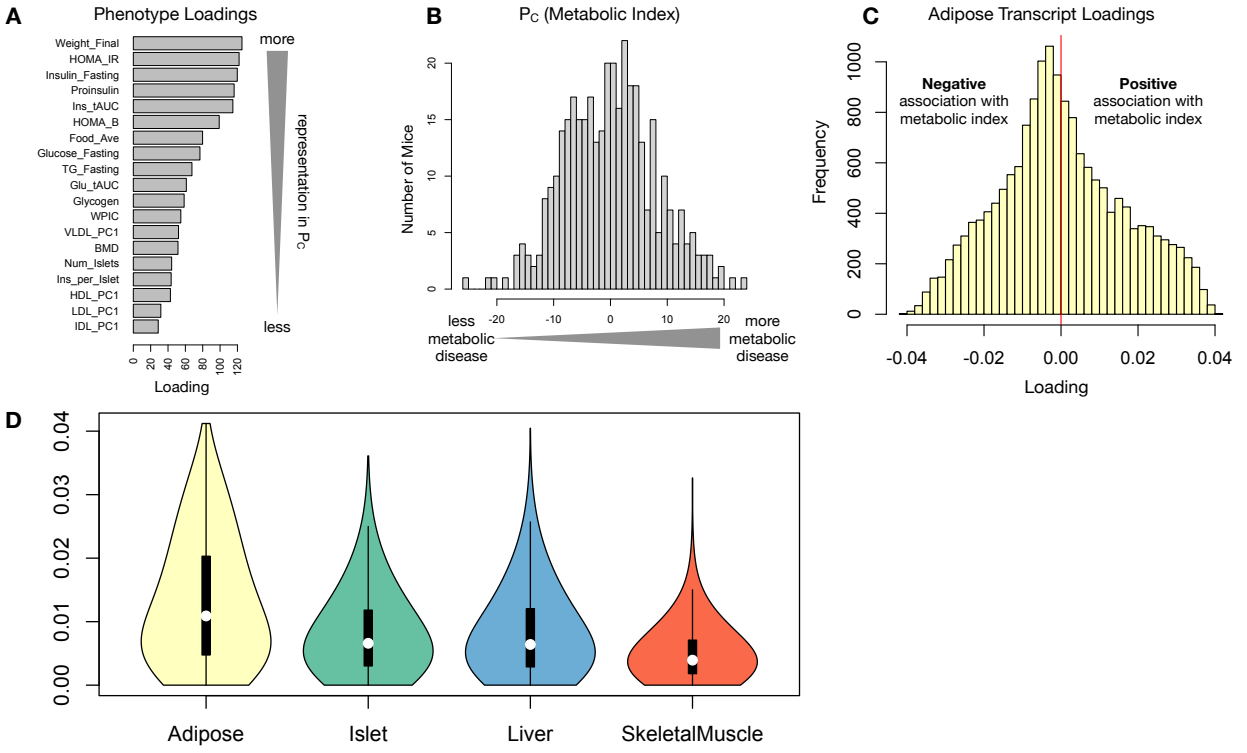

Figure 4: Interpretation of loadings. **A.** Loadings across traits. Body weight and insulin resistance contributed the most to the composite trait. **B.** Phenotype scores across individuals. Individuals with large positive phenotype scores had higher body weight and insulin resistance than average. 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.** Distribution of absolute value of transcript loadings across tissues. Transcripts in adipose tissue had the largest loadings indicating that transcripts in adipose tissue were the best mediators of the genetic effects on body weight and insulin resistance.

High-loading transcripts have low local heritability, high distal heritability, and are linked mechanistically to obesity

We interpreted large loadings onto transcripts as indicating strong mediation of the effect of genetics on metabolic index. Large positive loadings indicate that inheriting higher expression was associated with a higher metabolic index (i.e. higher risk of obesity and metabolic disease on the high-fat diet) (Fig. 4C). Conversely, large negative loadings indicate that inheriting lower expression of these transcripts was associated with a lower metabolic index (i.e. lower risk of obesity and metabolic disease on the high-fat diet) (Fig. 4C). We used GSEA 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 and KEGG pathway enrichments pointed to an axis of inflammation and metabolism (Supp. Fig. 10 and 11). Processes and pathways associated with inflammation, particularly macrophage infiltration were positively associated with metabolic index, indicating that increased expression in inflammatory pathways was associated with a higher metabolic index. It is well established that adipose tissue in obese individuals is highly inflamed [cite] and infiltrated by macrophages [cite], and the results here suggest that this may be a heritable component of metabolic disease.

The strongest negative enrichments in adipose tissue were related to mitochondial activity in general, and thermogenesis in particular (Supp. Fig. 10 and 11). It has been shown mouse strains with greater thermogenic potential are also less susceptible to obesity on a high-fat diet.

Transcripts associated with the citric acid (TCA) cycle as well as the catabolism of branched-chain amino acids (BCAA), valine, leuceine, and isoleucine also had strong negative enrichment in the adipose tissue (Supp. Fig. XXX). Expression of genes in both pathways (for which there is some overlap) has been previously associated with insulin sensitivity [20, 26, 27], suggesting that impairment in these pathways may be associated with insulin resistance. Selective PPARγ modulation by insulin-sensitizing thiazolidinedione drugs has further been shown to influence both inflammation and BCAA metabolism in obese rats suggesting a relationship between these pathways and insulin resistance [28]. BCAA levels are also related to insulin resistance in human subjects and are elevated in insulin-resistant obese individuals relative to weight-matched non-insulin reisistant individuals [29]. In the DO mice studied here, inheriting increased expression of genes involved in BCAA catabolism was associated with reduced body weight and insulin resistance.

Transcripts in the adipose tissue had the largest loadings, both positive and negative, of all tissues, suggesting that much of the effect of genetics on body weight and insulin reisistance is mediated through gene expression in adipose tissue (Fig. 5A). The 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. Across all tissues, trahscripts with the largest loadings tended to have relatively high distal heritability compared with local heritability (Fig. 5A). Transcripts with the highest local heritability tended to have very weak loadings and were 3.6 times less likely to be associated with diabetes and obesity in the literature than transcripts with high loadings (Fig. fig:loading_heritabilityB, Methods). TWAS-nominated transcripts also had relatively weak loadings and high local heritabilty (Fig. 4C). They were half as likely as transcripts with the highest loadings to be associated with diabetes and obesity in the literature (Fig. 4C).

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 in the DO population (Fig. 6A). In this figure, the importance of immune activation specifically in the adipose tissue is apparent. There are also other tissue-specific processes. Positive loadings on lipid metabolism in liver suggest that inheriting high liver expression of genes in this cluster is positively associated with 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 [30]. Agonists of Pparg, such as Thiazolidinediones, which are FDA-approved to treat type II diabetes, reduce inflammation and adipose hyptertrophy [30]. Consistent with this role, the loading for *Pparg* in adipose tissue is slightly negative, suggesting that upregulation is associated with leaner mice (Fig. 6B). In contrast, *Pparg* has a large positive loading in liver, where it plays 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 [31, 32]. Overexpression of *Pparg* in the livers of mice with a *Ppara* knockout, causes upregulation of genes involved in adipogenesis [33]. In the livers of both mice and humans [34, 35] High *Pparg* expression is associated with hepatocytes that accumulate large lipid droplets and have gene expression profiles similar to adipocytes.

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 fairly high, but the loading is low, suggesting that variability of expression in the islet does not drive phenotypic variation. 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 Supplemental Files XXX.

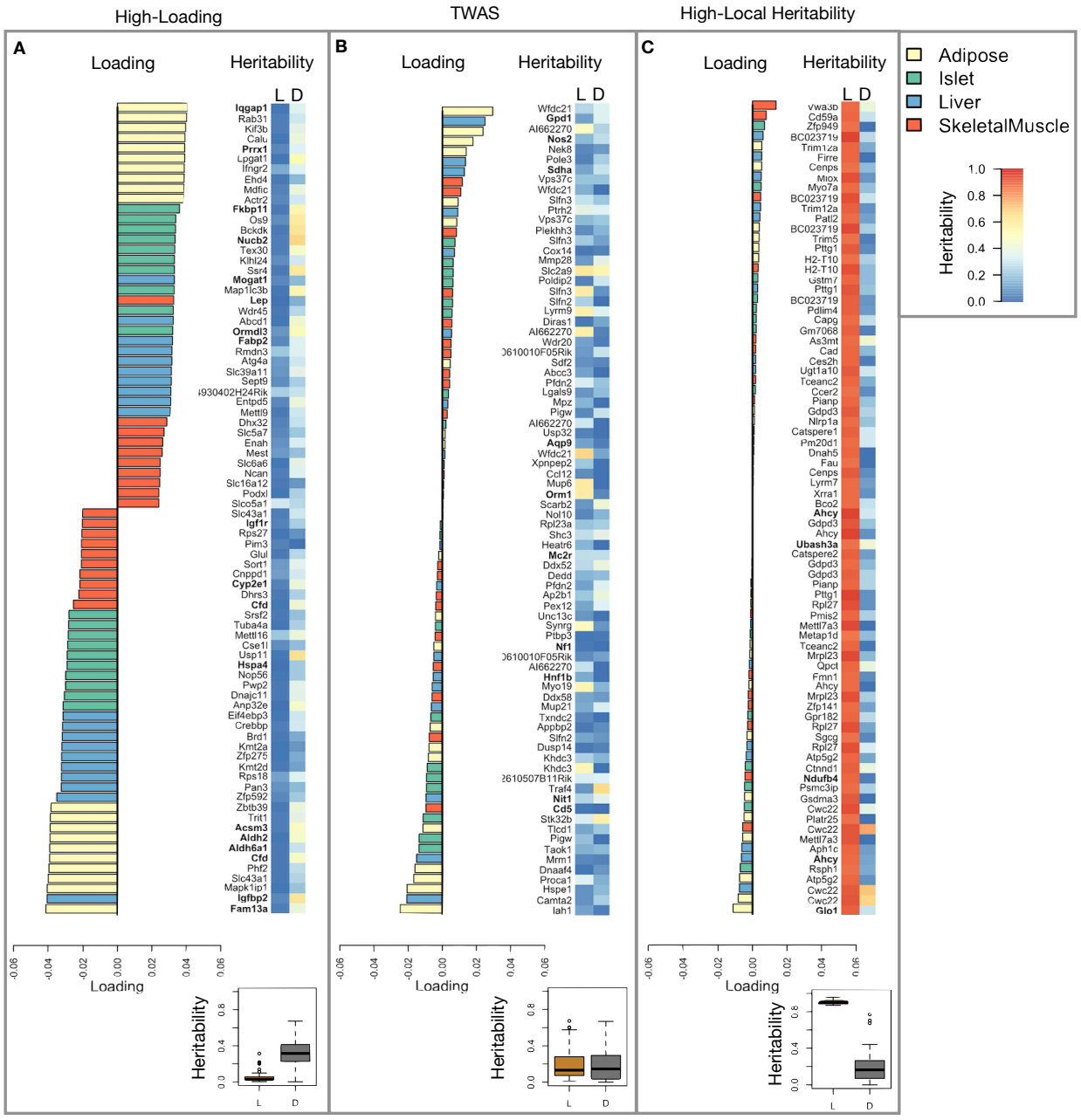

Figure 5: Transcripts with high loadings have high distal heritability and literature support. Each panel has a bar plot showing the loadings of transcripts selected by different criteria. Bar color indicates the tissue of origin. The heat map shows the local (L - left) and distal (D - right) heritability 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. **B.** Loadings of TWAS candidates with the 10 largest positive correlations with traits and the largest negative correlations with traits across all four tissues. **C.** The transcripts with the largest local heritability (top 20) across all four tissues.

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

The loading of each transcript indicates how inherited expression levels influence metabolic phenotypes. If local regulation is the predominant factor influencing gene expression, we should be able to predict an

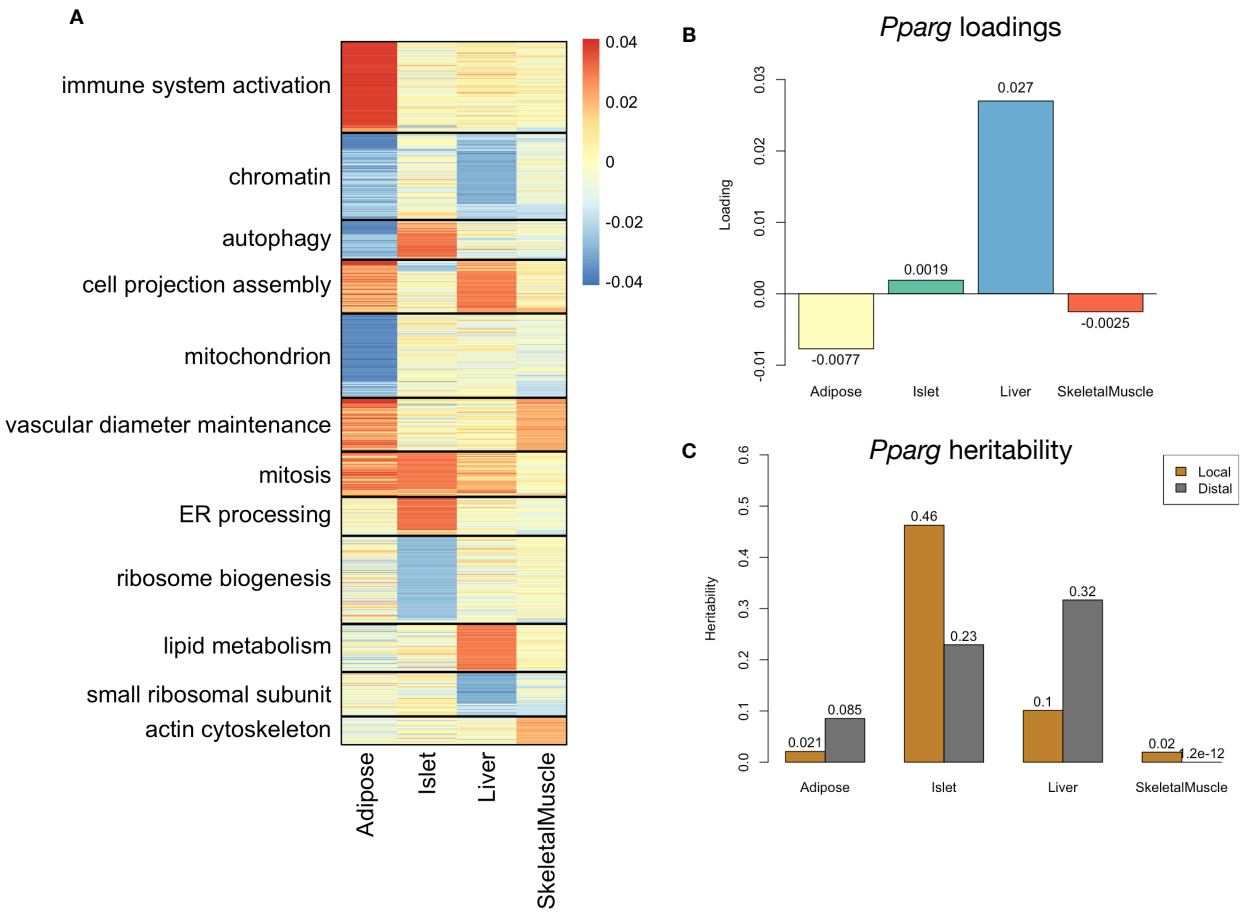

Figure 6: Tissue-specific transcriptional programs were associated with obesity and insulin resistance. **A** Heat map showing the loadings of all transcripts with loadings greater than 2.5 standard deviations from the mean in any tissue. The heat map was clustered using k medoid clustering. Functional 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.

individual's phenotype based on their genotypes across all local eQTLs. We tested this hypothesis in an independent population of F1 mice generated through multiple pairings of Collaborative Cross (CC) [cite] strains (Fig. 7A) (Methods).

We first tested whether the transcript loadings derived from HDM in the DO were relevant to the relationship between the transcriptome and the phenome in the CC-RIX. To do this, we multiplied the transcript loadings derived from HDM in the DO mice by transcript measurements in the CC-RIX standardized across individuals. This created a transcript vector weighted by importance to metabolic disease as determined in the DO. The mean of this vector was the predicted metabolic index for the animal based on its transcription in either adipose tissue, liver, or skeletal muscle. Across all three tissues, weighted transcription values were significantly correlated with metabolic index in the CC-RIX population measured as body weight (Fig. 7B left

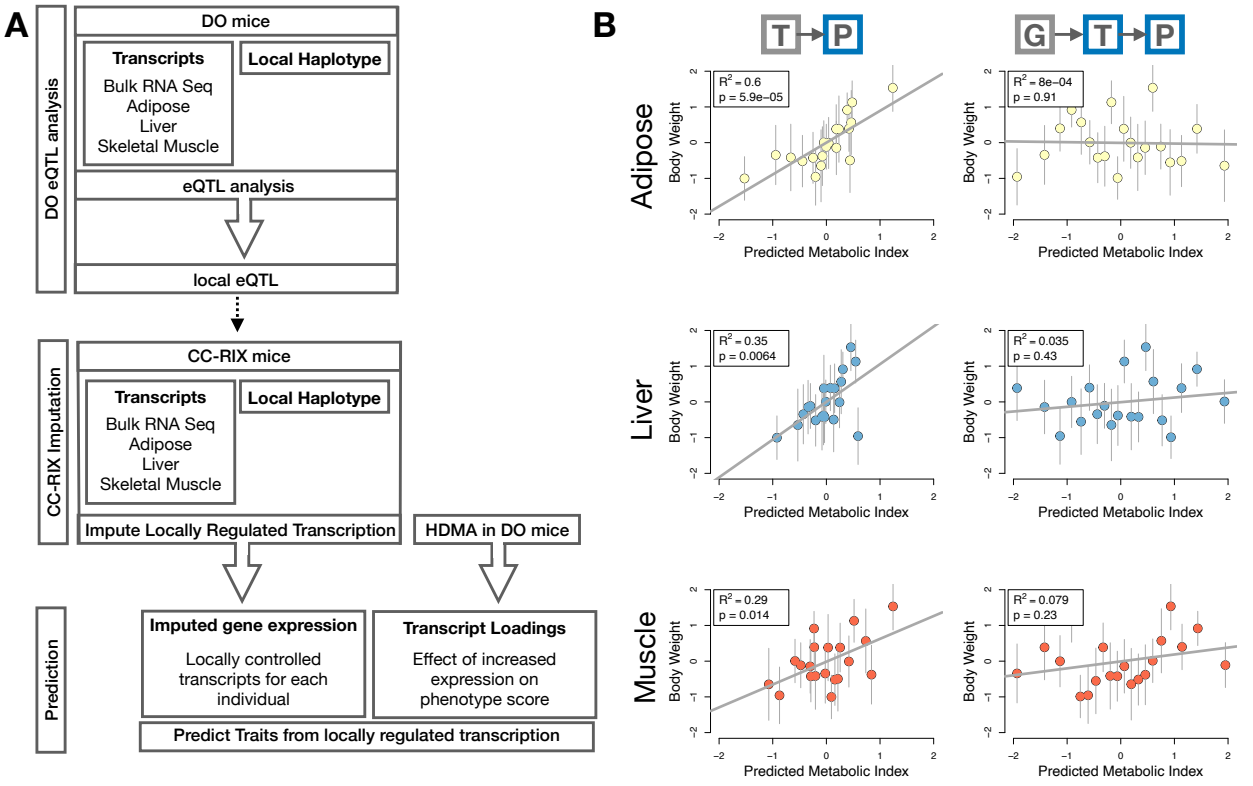

Figure 7: Transcription, but not local genotype, predicts phenotype in the CC-RIX. **A.** Workflow showing procedure for translating HDM results to an independent population of mice. **B.** Relationships between the predicted metabolic index and measured body weight. The left column shows the predictions 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.

column). Adipose tissue transcription yielded the most accurate prediction (stats). This result confirms the validity and translatability of the transcript loadings determined in the DO population and their relationship to metabolic disease. It also supports the observation that transcription in adipose tissue is the strongest mediator of genetic effects on metabolic index.

We then tested whether this mediation signal was encoded by local genotype. To do this, we imputed gene expression in the CC-RIX using local genotype. We were able to estimate variation in gene transcription robustly. The correlation between measured gene expression and imputed gene expression across all tissues was close to $R = 0.5$, and the variance explained by local genotype was comparable in the DO and CC-RIX (Supp. Fig. 12). However, when weighted with the loadings derived from HDM in the DO population, these imputed transcripts across all tissues failed to predict metabolic index in the CC-RIX (Fig. 7B right column).

Taken together, these results support the hypothesis that distal, rather than local genetic factors are primarily

driving complex-trait related variation in gene expression.

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

Interpretation of global distal genetic influences on gene expression and phenotype is potentially more challenging than interpretation and translation of local genetic influences. Effects can not be located to individual gene variants or transcripts, but because we have a measure of importance across all transcripts in multiple tissues, we can look at global patterns. We noted earlier that functional enrichments of transcripts with large positive loadings in the adipose tissue, suggested that the obese mice in the population had a genetic predisposition toward elevated macrophage infiltration into the adipose tissue. This suggests heritabl variability in cell-type composition of the adipose tissue. We investigated this further bioinformatically by comparing the loadings of cell-type-specific transcripts (Methods). For adipose tissue we used a list of cell-type specific genes identified in human adipose tissue

In adipose tissue, the mean loading of macrophage-specific genes was substantially greater than 0 (Fig. 8A), indicating that obese mice were genetically predisposed to have high levels of macrophage infiltration in adipose tissue in response to the high-fat, high-sugar diet.

In islet, the mean loadings for alpha-cell specific transcripts were significantly positive, while the mean loadings for delta- and endothelial-cell specific genes were significantly negative (Fig. 8B). These results suggest that obese mice had inherited higher proportions of alpha cells, and lower proportions of endothelial and delta cells in their pancreatic islets.

The loadings for pancreatic beta cell-type specific loadings was not significantly different from zero. This does not reflect on the function of the beta cells in the obese mice, but rather suggests that mice prone to obesity were not obese because they inherited fewer beta cells than non-obese mice.

Biological interpretation of alpha, endothelial, delta cells??

Distally heritable transcriptomic signatures translate to human disease

Ultimately, the distally heritable transcriptomic signatures that we identified in DO mice will be useful if they inform pathogenicity 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.

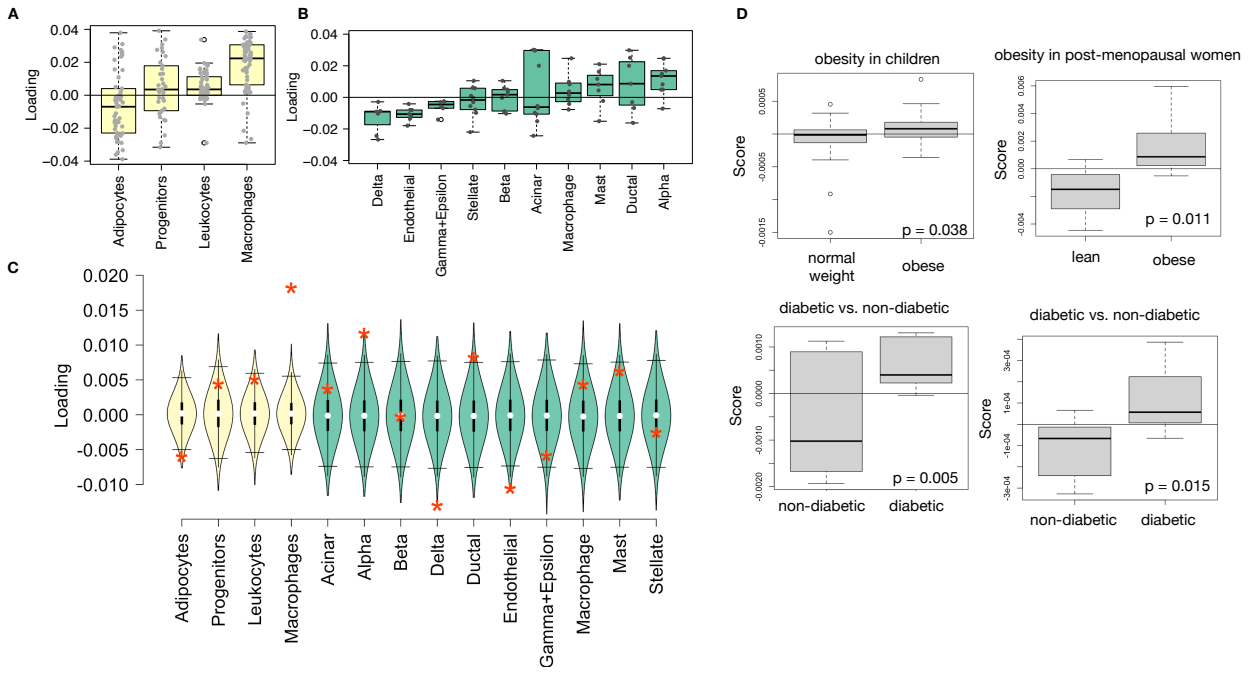

Figure 8: HDM results translate to humans. **A.** Distribution of loadings for cell-type-specific transcripts in adipose tissue. **B.** Distribution of loadings for cell-type-specific transcripts in pancreatic islets (green). **C.** Null distributions for the mean loading of randomly selected transcripts 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 metabolic disease than the lean/non-diabetic patients based on the HDM results from DO mice.

We calculated a predicted obesity score 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 obesity scores were higher on average for individuals in the obese and diabetic groups compared with the lean and non-diabetic groups, indicating that the distally heritable signature of obesity identified in DO mice is relevant to obesity and diabetes in human subjects.

Targeting gene signatures

Although high-loading transcripts are likely good candidates for understanding specific biology related to obesity, we emphasize that the transcriptome overall is highly interconnected and redundant, and that focusing on individual transcripts for treatment may be less effective than using a broader transcriptomic signature. The ConnectivityMap (CMAP) database [cite] developed by the Broad Institute allows us to query thousands of compounds that reverse or enhance transcriptomic signatures as a whole in multiple different cell types. By identifying drugs that reverse pathogenic transcriptomic signatures as a whole rather than

targeting individual genes, we can potentially increase efficacy of tested compounds.

We thus queried the CMAP database through the CLUE online query tool developed by The Broad Institute [cite] (Methods).

Alternatively, we can target the gene signature as a whole using CMAP. Identifying drugs to target gene signatures is possible through CMAP. We put our loadings from islet into CMAP. The top hit was PPAR receptor agonist. Rosiglitazone, a widely used diabetes drug, is a PPAR receptor agonist. Another class of drugs on the list was sufonylureas, which are another major class of drugs for type 2 diabetes.

- **Supplemental Table** results from CMAP

Discussion

- distal heritability correlates with phenotype relevance
- others who use local eQTL to associate genotype with traits often say "we nominated this gene" even though other nearby genes have higher eQTL LOD scores (27019110, 31465442) Our method supports the idea that the transcripts with the strongest local regulation are less likely to be functionally related to the trait

Data Availability

Here we tell people where to find the data

Acknowledgements

Here we thank people

Supplemental Figures

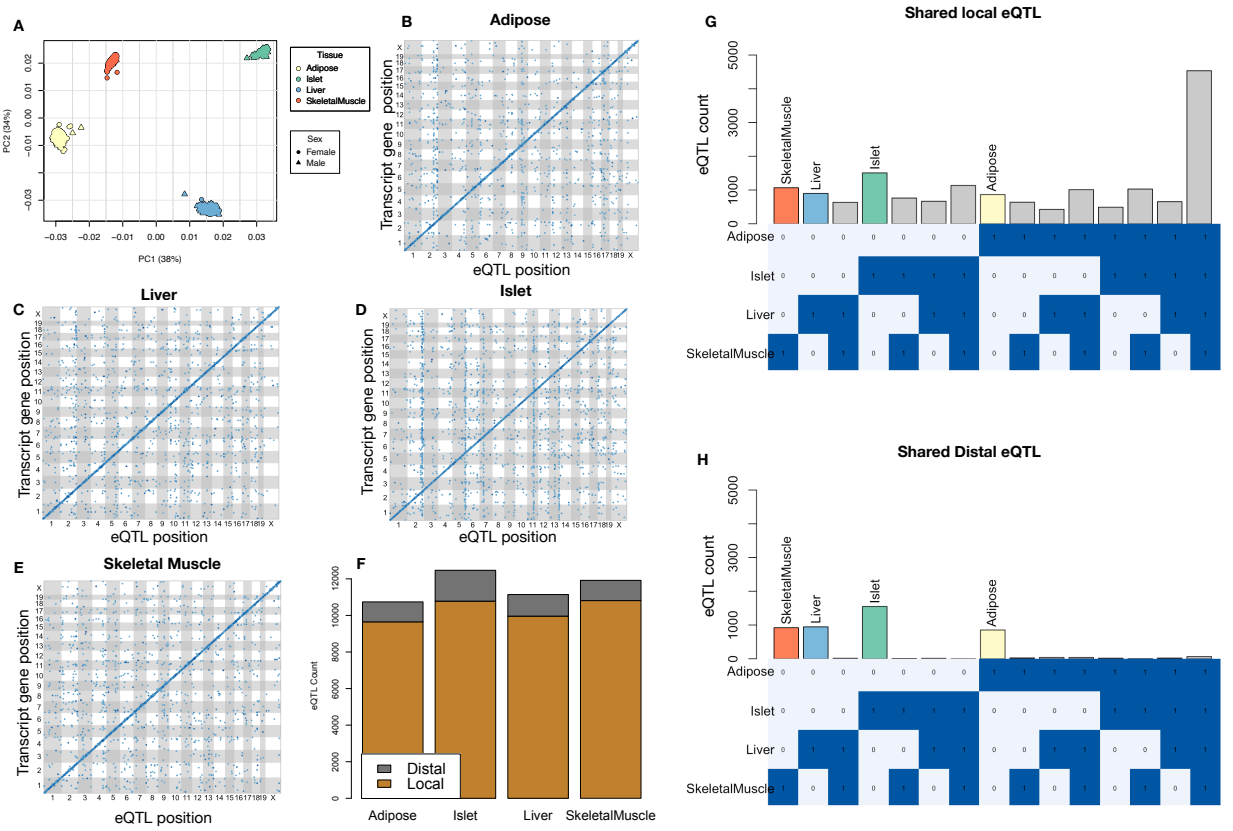

Figure 9: Overview of eQTL analysis in DO mice. **A.** RNA seq samples from the four different tissues clustered by tissue. **B.-E.** eQTL maps are shown for each tissue. The x-axis shows the position of the mapped eQTL, and the y-axis shows the physical position of the gene encoding each mapped transcript. Each dot represents an eQTL with a minimum LOD score of 8. 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 eQTL are much more numerous than distal eQTL. **G.** Counts of transcripts with local eQTL shared across multiple tissues. The majority of local eQTL 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.

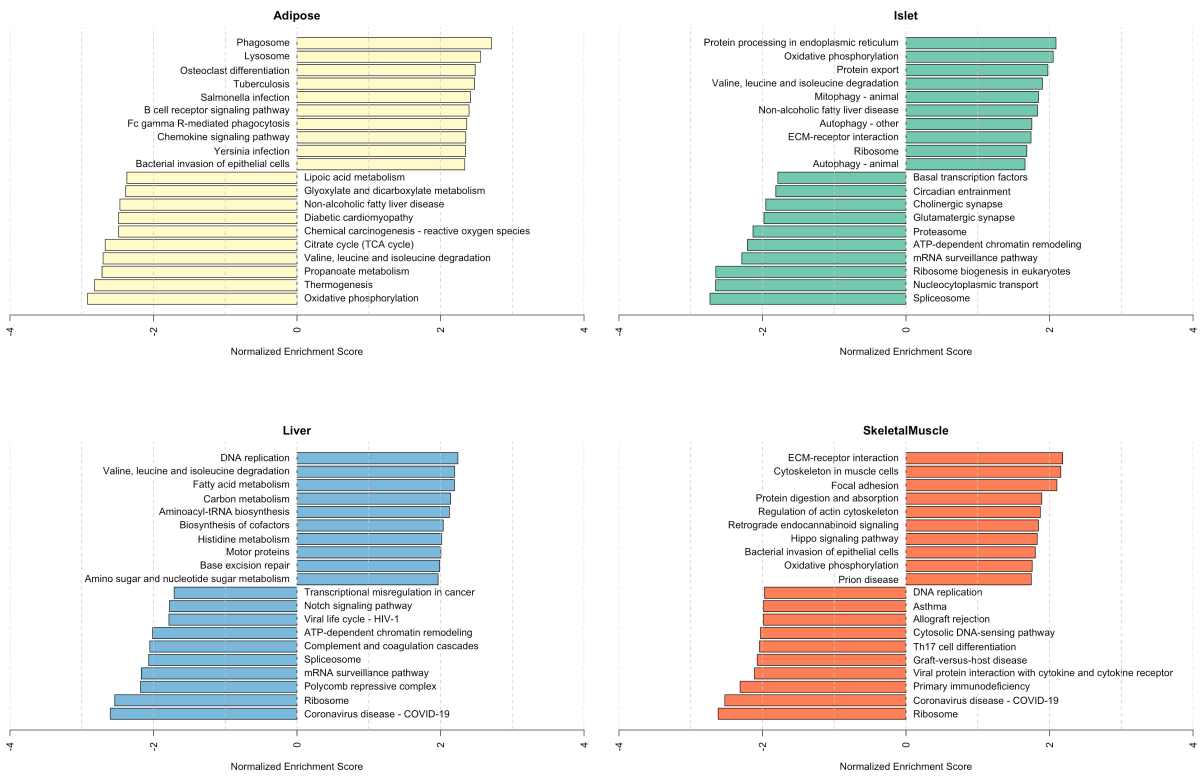

Figure 10: Bar plots showing normalized enrichment scores (NES) for KEGG pathways as determined by fast gene score enrichment analysis (fgsea). Only the top 10 positive and top 10 negative scores are shown. Colors indicate tissue. The name beside each bar shows the name of each enriched KEGG pathway.

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

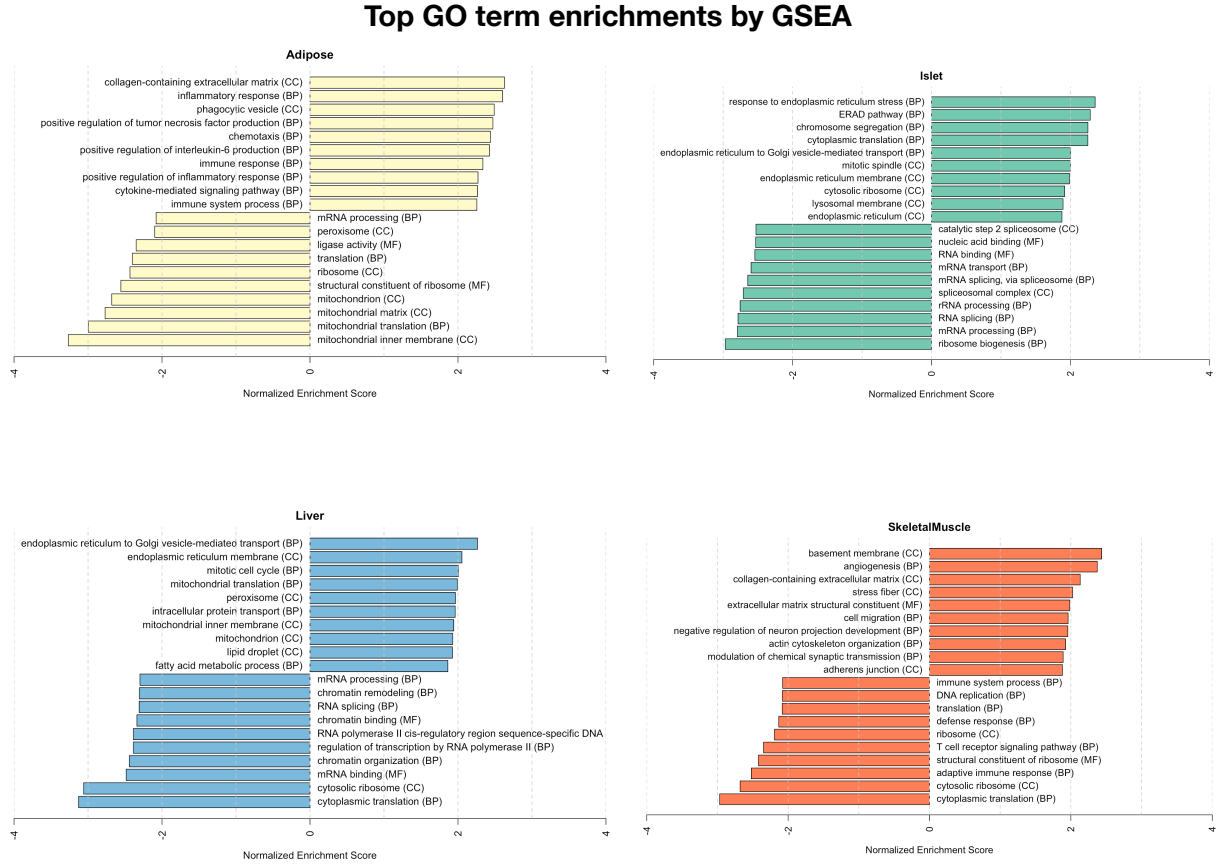

Figure 11: Bar plots showing normalized enrichment scores (NES) for GO terms as determined by fast gene score enrichment analysis (fgsea). Only the top 10 positive and top 10 negative scores are shown. Colors indicate tissue. The name beside each bar shows the name of each enriched GO term. The letters in parentheses indicate whether the term is from the biological process ontology (BP), the molecular function ontology (MF), or the cellular compartment ontology (CC).

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.

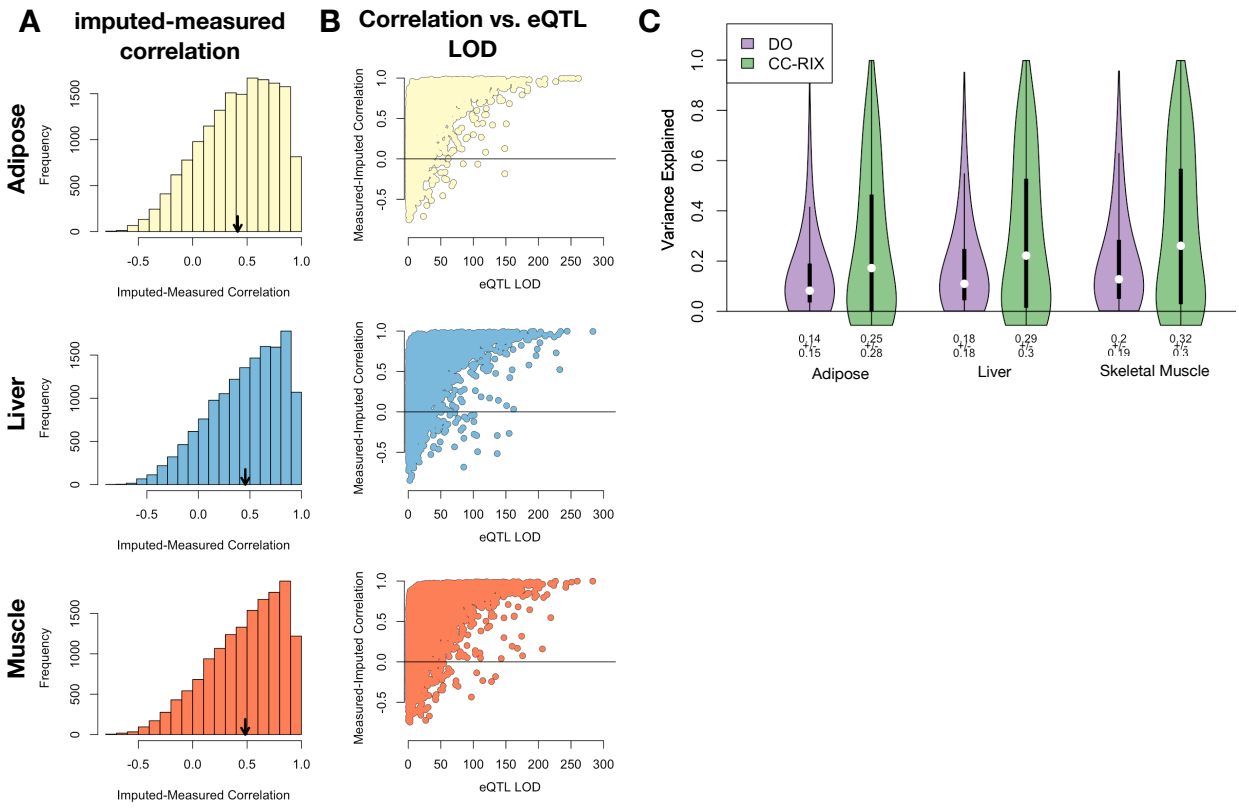

Figure 12: Validation of transcript imputation in the CC-RIX. **A.** Distributions of correlations between imputed and measured transcripts in the CC-RIX. The mean of each distribution is shown by the red line. All distributions were skewed toward positive correlations and had positive means near a Pearson 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 eQTL. **C.** Variance explained by local genotype in the DO and CC-RIX.

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] D. P. MacKinnon, A. J. Fairchild, and M. S. Fritz. Mediation analysis. *Annu Rev Psychol*, 58:593–614, 2007.

[22] 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.

[23] 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.

[24] Fabien Girka, Etienne Camenen, Caroline Peltier, Arnaud Gloaguen, Vincent Guillemot, Laurent Le Brusquet, and Arthur Tenenhaus. *RGCCA: Regularized and Sparse Generalized Canonical Correlation Analysis for Multiblock Data*, 2023. R package version 3.0.3.

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

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

[27] 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.

[28] G. Hsiao, J. Chapman, J. M. Ofrecio, J. Wilkes, J. L. Resnik, D. Thapar, S. Subramaniam, and D. D. Sears. modulation of insulin sensitivity and metabolic pathways in obese rats. *Am J Physiol Endocrinol Metab*, 300(1):E164–174, Jan 2011.

[29] D. E. Lackey, C. J. Lynch, K. C. Olson, R. Mostaedi, M. Ali, W. H. Smith, F. Karpe, S. Humphreys, D. H. Bedinger, T. N. Dunn, A. P. Thomas, P. J. Oort, D. A. Kieffer, R. Amin, A. Bettaieb, F. G. Haj, P. Permana, T. G. Anthony, and S. H. Adams. Regulation of adipose branched-chain amino acid catabolism enzyme expression and cross-adipose amino acid flux in human obesity. *Am J Physiol Endocrinol Metab*, 304(11):E1175–1187, Jun 2013.

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

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.