

1 Transcripts with high distal heritability mediate genetic effects on  
2 complex traits

3

4 **Abstract**

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

8 **Introduction**

9 In the quest to understand genetic contributions to complex traits, gene expression is an important bridge  
10 between genotype and phenotype. The majority of variants identified in GWAS are in regulatory regions of  
11 the genome [cite], suggesting that they influence clinical phenotypes through regulation of gene expression.  
12 Consistent with this idea, powerful insights into disease mechanism can be made by linking genetic variants  
13 affecting gene expression (expression quantitative trait loci - eQTLs) to phenotypes. In particular, mediation  
14 analysis has been used to identify transcripts that mediate the effect of genetic variants on phenotypes. In  
15 mice... [1, 2, 3] (bmediatR) [4] In humans... [5, 6]

16 Thus far, the primary focus of expression mediated traits has been on local genetic variation; that is genetic  
17 variation that influences the transcription of local genes, thereby causing variation in traits. However, there  
18 is evidence that the bulk of disease heritability is mediated by the distal component of gene expression,  
19 rather than the local component [7]. Yao et al. [cite] observed that genes with low local heritability explain  
20 more expression-mediated disease heritability than genes with high local heritability. We have observed a  
21 similar pattern in mice, which we describe here. Thus, identifying heritable components of complex traits  
22 that are mediated through distally regulated variation in gene expression may provide important insights  
23 into mechanisms regulating complex traits.

24 Identification of distal factors influencing gene expression and traits is challenging, as the multiple testing  
25 corrections are much more severe for distal effects [8]. However, systems approaches that consider the entire

26 transcriptome simultaneously and avoid univariate testing provide promising avenues for identification of  
27 broad transcriptomic patterns influencing complex traits that provide both biological insight and targets  
28 for therapeutics. Here we propose high-dimensionaional mediation (HDM) as one such systems approach for  
29 identification of the heritable portion of the transcriptome that mediates the effect of the genome on phenotype.  
30 HDM uses a regularized and generalized canonical correlation analysis (RGCCA) [cite], which is an extension  
31 of canonical correlation analysis (CCA) that allows for more than two data sets with an arbitrary relationship  
32 among them. Thus, we can identify linear combinations of the genome, transcriptome, and phenotype, that  
33 describe the mediation of the genetic effects on the phenotype through the transcriptome. Because of the  
34 central dogma of molecular biology, information flow is directed out of the genome, and not back into it.  
35 Thus, the otherwise undirected relationships between genome, transcriptome, and phenotype can be inferred  
36 as a causal mediation by the transcriptome of the effects of the genome on the phenotype.

37 Here we apply HDM

## 38 Results

### 39 Genetic variation contributes to wide phenotypic variation

40 A population of 500 diversity outbred mice (XXX male and XXX female), was placed on a high-fat (XXX/%),  
41 high-sugar (XXX/%) diet starting at XXX weeks of age as described previously [cite]. Each individual was  
42 assessed longitudinally for multiple metabolic measures including fasting glucose levels, glucose tolerance,  
43 insulin levels, body weight, and blood lipid levels (Methods).

44 Although the environment was consistent across animals, the genetic diversity present in this population  
45 resulted in widely varying distributions across physiological measurements (Fig. 1). For example, body  
46 weights of adult individuals varied from less than the average adult B6 body weight to several times the  
47 body weight of a B6 adult in both sexes (Fig. 1A). Fasting blood glucose also varied considerably (Fig. 1B)  
48 although few of the animals had FBG levels that would indicate pre-diabetes ( animals, ), or diabetes (7  
49 animals, 1.4) according to previously developed cutoffs (pre-diabetes:  $\text{FBG} \geq 250 \text{ mg/dL}$ , diabetes:  $\text{FBG}$   
50  $\geq 300, \text{ mg/dL}$ ) [9]. Males had higher FBG than females on average (Fig. 1C) as has been observed before  
51 suggesting either that males were more susceptible to metabolic disease on the high-fat diet, or that males  
52 and females may require different thresholds for pre-diabetes and diabetes.

53 Body weight was strongly positively correlated with food consumption (Fig. 1D  $R^2 = 0.51, p = 1.5 \times 10^{-75}$ )  
54 and fasting blood glucose (FBG) (Fig. 1E,  $R^2 = 0.25, p = 2 \times 10^{-32}$ ) suggesting a link between behavioral  
55 factors and metabolic disease. However, the heritability of this trait and others (Fig. 1F) indicates that

56 background genetics contribute substantially to correlates of metabolic disease in this population.

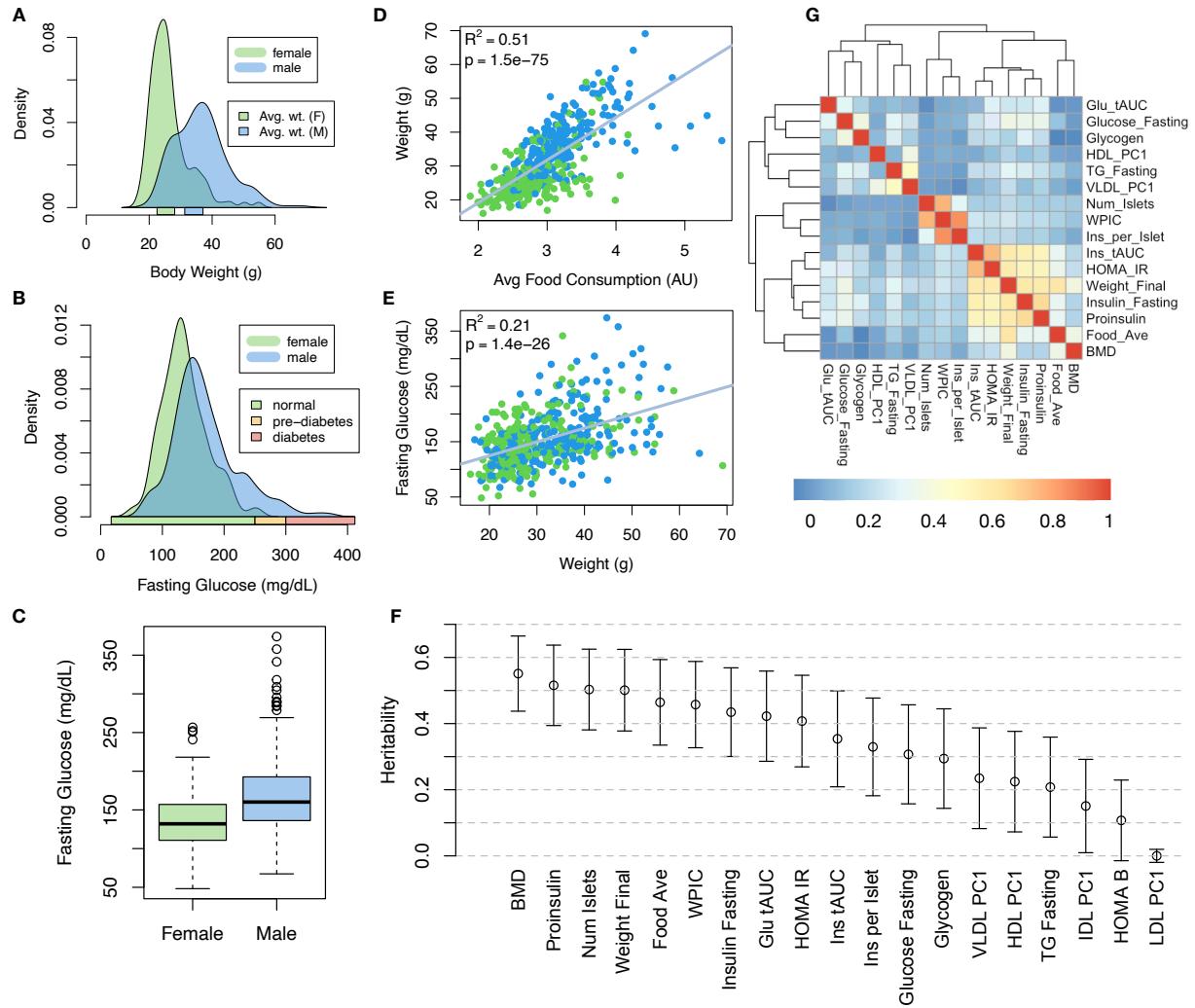


Figure 1: Clinical overview. **A.** Distributions of final body weight in the diversity outbred mice. Sex is indicated by color. The average B6 male and female adult weights at 24 weeks of age are indicated by blue and green bars on the x-axis. **B.** The distribution of final fasting glucose across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

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

61 **Distal Heritability Correlates with Phenotype Relevance**

62 To elaborate the mechanistic details of genetic effects on metabolic phenotypes in the DO population, we  
63 also measured gene expression in four tissues known to be involved in metabolic disease: adipose, pancreatic  
64 islet, liver, and skeletal muscle. To confirm the heritability of transcript levels, we performed expression QTL  
65 analysis using R/qtl2 [cite] (Methods) and identified both local and distal eQTL for transcripts in each tissue  
66 (Supp. Fig 8). Significant local eQTLs far outnumbered distal eQTLs (Supp. Fig. 8F) and tended to be  
67 shared across tissues (Supp. Fig. 8G) whereas the few significant distal eQTL we identified tended to be  
68 tissue-specific (Supp. Fig. 8H)

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

73 Local heritability of transcripts was negatively correlated with their trait relevance, defined as the maximum  
74 correlation of a transcript across all traits (Fig. 2B). This suggests that the more local genotype influenced  
75 transcript abundance, the less effect variation in transcript abundance was related to the measured traits.  
76 Conversely, distal heritability of transcripts was positively correlated with trait relevance (Fig. 2C). That is,  
77 transcripts that were more highly correlated with the measured traits tended to be distally, rather than locally,  
78 heritable. That trait-correlated transcripts have low local heritability is consistent with previous observations  
79 that low-heritability transcripts explain more expression-mediated disease heritability than high-heritability  
80 transcripts [7]. However, the positive relationship between trait correlation and distal heritability suggests  
81 that there are alternative mechanisms through which genetic regulation of transcripts may influence traits.

82 **High-Dimensional Mediation identifies composite transcript that perfectly mediates composite  
83 trait**

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

90 The result of this process is three vectors representing the composite genome ( $G_C$ ), composite transcriptome

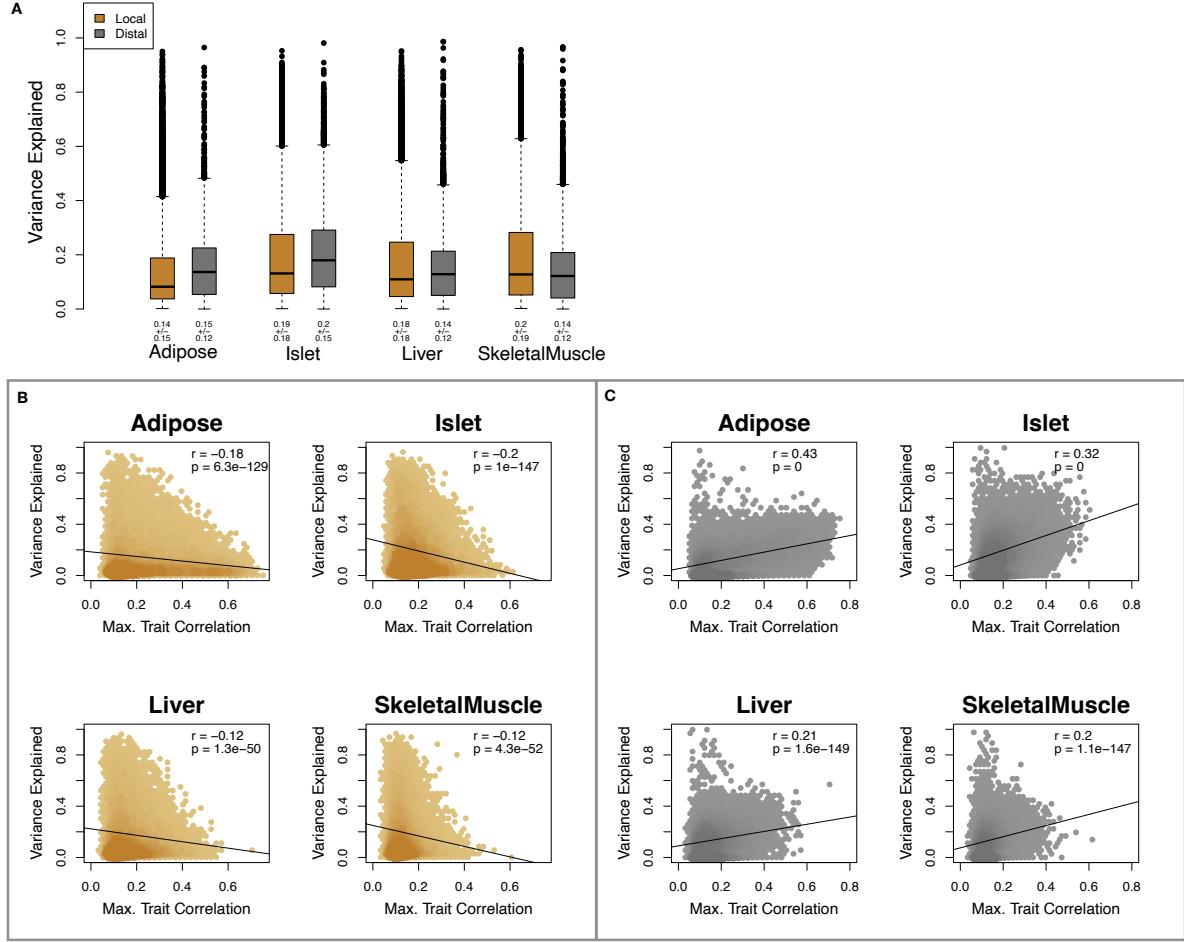


Figure 2: Transcript heritability and trait relevance. **A.** Distributions of distal and local heritability of transcripts across the four tissues. Overall local and distal factors contribute equally to transcript heritability. The relationship between **(B.)** local and **(C.)** distal heritability 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.

91 ( $T_C$ ) and the composite phenotype ( $P_C$ ) where the composite transcriptome perfectly mediates the effect of the  
 92 composite genome on the composite phenotype. Each vector is of length  $n$  where  $n$  is the number of individual  
 93 mice. Fig. 3A shows the partial correlations between all pairs of composite vectors. The partial correlation  $r$   
 94 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  
 95 transcriptome was taken into account, the partial correlation between  $G_S$  and  $P_S$  was effectively 0 (-0.01).  
 96 Standard CCA is prone to over-fitting because in any two large matrices it can be trivial to identify  
 97 highly correlated composite vectors. To assess whether RGCCA was similarly prone to over-fitting in a  
 98 high-dimensional space, we performed permutation testing. We permuted the individual labels on the

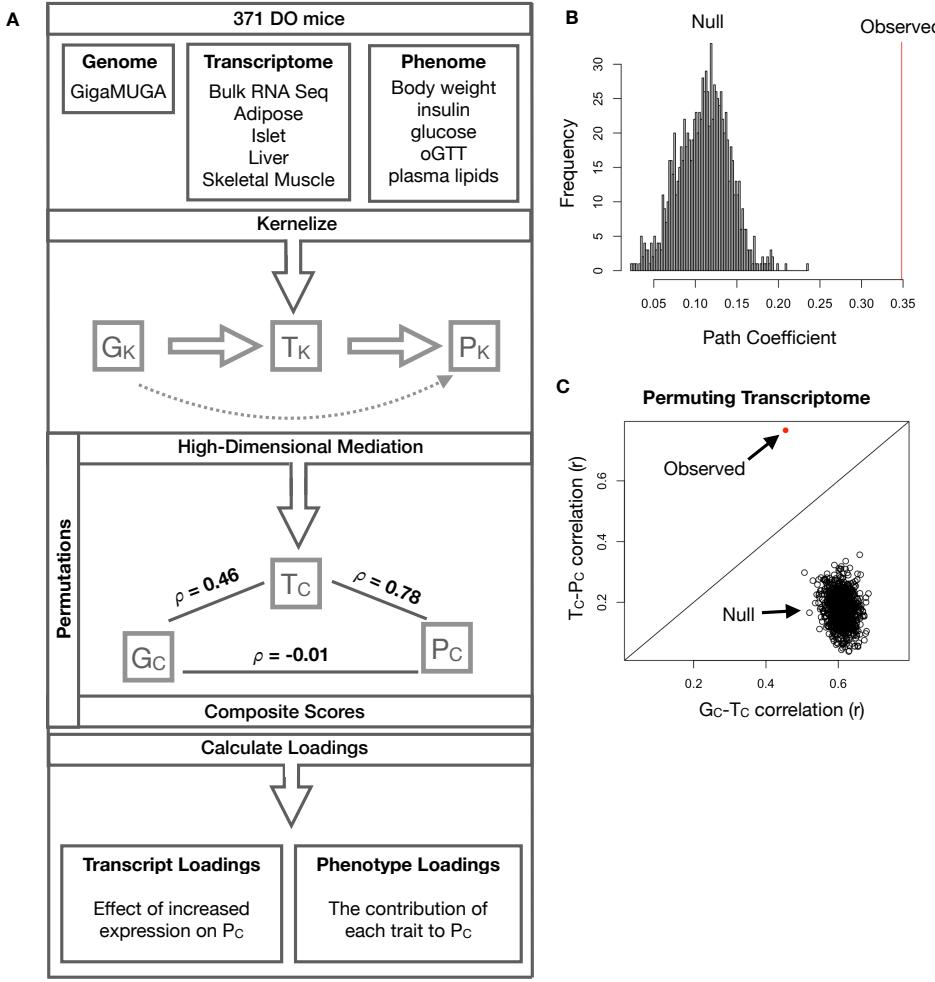


Figure 3: High-dimensional mediation. **A.** Workflow indicating major steps of high-dimensional mediation. The genotype, transcriptome, and phenotype matrices were kernelized to yield single matrices representing the relationships between all individuals for each data modality ( $G_K$  = genome kernel,  $T_K$  = 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  $\rho$  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).

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

105  $P_C$  in the null data (Fig 3C), these two values could not be maximized simultaneously. The red dot shows that  
106 in the real data both the  $G_C-T_C$  correlation and the  $T_C-P_C$  correlation could be maximized simultaneously  
107 suggesting that that path from genotype to phenotype through transcriptome is highly non-trivial and  
108 identifiable in this case. These results suggest that these composite vectors represent genetically determined  
109 variation in phenotype that is mediated through genetically determined variation in transcription.

110 **Body weight and insulin resistance were highly represented in the expression-mediated com-**  
111 **posite trait**

112 The loadings of each measured trait onto  $P_C$  indicate how much each contributed to  $P_C$ . Final body weight  
113 contributed the most to  $P_C$  (Fig. 4), followed by homeostatic insulin resistance (HOMA\_IR) and fasting  
114 plasma insulin levels (Insulin\_Fasting). We can thus interpret  $P_C$  as an index of metabolic disease (Fig. 4B).  
115 Individuals with high values of  $P_C$  have a higher metabolic index and greater metabolic disease, including  
116 higher body weight and higher insulin resistance. We refer to  $P_C$  as the metabolic index going forward. Traits  
117 contributing the least to the metabolic index were measures of cholesterol and pancreas composition. Thus,  
118 when we interpret the transcriptomic signature identified by HDM, we are explaining primarily transcriptional  
119 mediation of body weight and insulin resistance, as opposed to cholesterol measurements.

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

122 Transcripts that most strongly correlated with  $T_C$  were the best mediators of the effect of genetics on  
123 metabolic index. Large positive loadings indicate that inheriting higher expression was associated with a  
124 higher metabolic index (i.e. higher risk of obesity and metabolic disease on the high-fat diet) (Fig. 4C).  
125 Conversely, large negative loadings indicate that inheriting lower expression of these transcripts was associated  
126 with a lower metabolic index (i.e. lower risk of obesity and metabolic disease on the high-fat diet) (Fig. 4C).  
127 We used GSEA to look for biological processes and pathways that were enriched at the top and bottom of  
128 this list (methods).

129 In adipose tissue, both GO processes and KEGG pathway enrichments pointed to an axis of inflammation  
130 and metabolism that was strongly related to metabolic index (Supp. Fig. 9 and 10). Processes and pathways  
131 associated with inflammation, particularly macrophage infiltration were positively associated with metabolic  
132 index, indicating that increased expression in inflammatory pathways was associated with a higher metabolic  
133 index.

134 The strongest negative enrichments in adipose tissue were related to mitochondrial activity in general, and

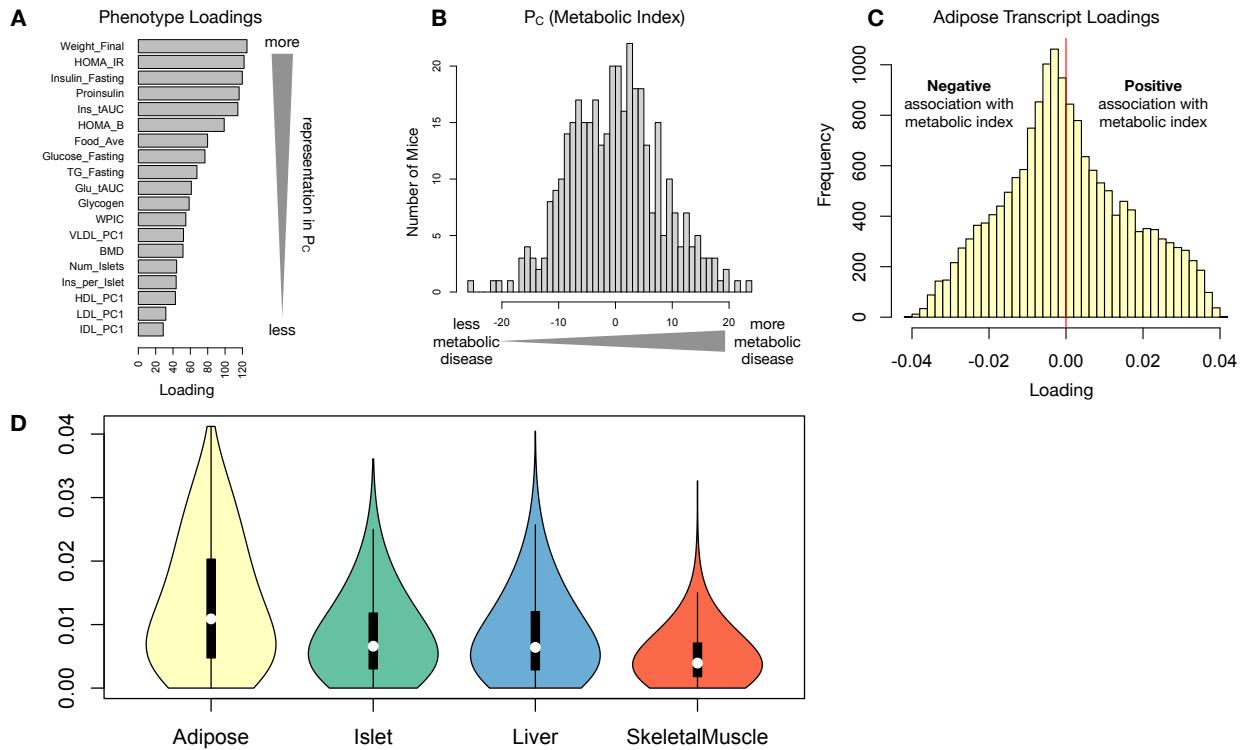


Figure 4: Interpretation of loadings. **A.** Loadings across traits. Body weight and insulin resistance contributed the most to the composite trait. **B.** Phenotype scores across individuals. Individuals with large positive phenotype scores had higher body weight and insulin resistance than average. 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.

135 thermogenesis in particular.

136 and represent known biology of obesity and diabetes. In adipose tissue, for example, the transcripts most  
 137 strongly correlated with  $T_C$  were enriched for immune system signaling and cell motility. It is well established  
 138 that adipose tissue in obese individuals is highly inflamed [cite] and infiltrated by macrophages [cite]. The  
 139 transcripts most strongly negatively correlated with  $T_C$  were enriched for metabolism of the branched-chain  
 140 amino acids (BCAA), valine, leucine, and isoleucine. BCAA are used in adipose tissue in lipogenesis, and  
 141 inhibiting BCAA catabolism inhibits adipogenesis [10]. BCAA levels are also related to insulin resistance and  
 142 are elevated in insulin-resistant obese individuals relative to weight-matched non-insulin resistant individuals  
 143 [11]. In the DO mice studied here, inheriting reduced expression of genes involved in BCAA catabolism was  
 144 associated with reduced body weight and insulin resistance.

145 Transcripts in the adipose tissue had the largest loadings, both positive and negative, of all tissues, suggesting  
146 that much of the effect of genetics on body weight and insulin resistance is mediated through gene expression  
147 in adipose tissue (Fig. 5A). The loadings in liver and pancreas were comparable, and those in skeletal muscle  
148 were the weakest (Fig. 5A), suggesting that less of the genetic effects were mediated through transcription in  
149 skeletal muscle. Across all tissues, transcripts with the largest loadings tended to have relatively high distal  
150 heritability compared with local heritability (Fig. 5A). Transcripts with the highest local heritability tended  
151 to have very weak loadings and were 3.6 times less likely to be associated with diabetes and obesity in the  
152 literature than transcripts with high loadings (Fig. fig:loading\_heritabilityB, Methods). TWAS-nominated  
153 transcripts also had relatively weak loadings and high local heritability (Fig. 4C). They were half as likely as  
154 transcripts with the highest loadings to be associated with diabetes and obesity in the literature (Fig. 4C).

155 **Tissue-specific transcriptional programs are associated with metabolic traits**

156 Clustering of transcripts with top loadings in each tissue shows tissue-specific functional modules associated  
157 with obesity and insulin resistance in the DO population (Fig. 6). Many of these modules, such as leptin  
158 signaling in adipose tissue [cite] and skeletal muscle [cite], as well as apelin signaling [cite] have well established  
159 functional roles in diabetes and obesity.

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

161 The loading of each transcript indicates how inherited expression levels influence metabolic phenotypes.  
162 If local regulation is the predominant factor influencing gene expression, we should be able to predict an  
163 individual's phenotype based on their genotypes across all local eQTLs. We tested this hypothesis in an  
164 independent population of F1 mice generated through multiple pairings of Collaborative Cross (CC) [cite]  
165 strains (Fig. A) (Methods).

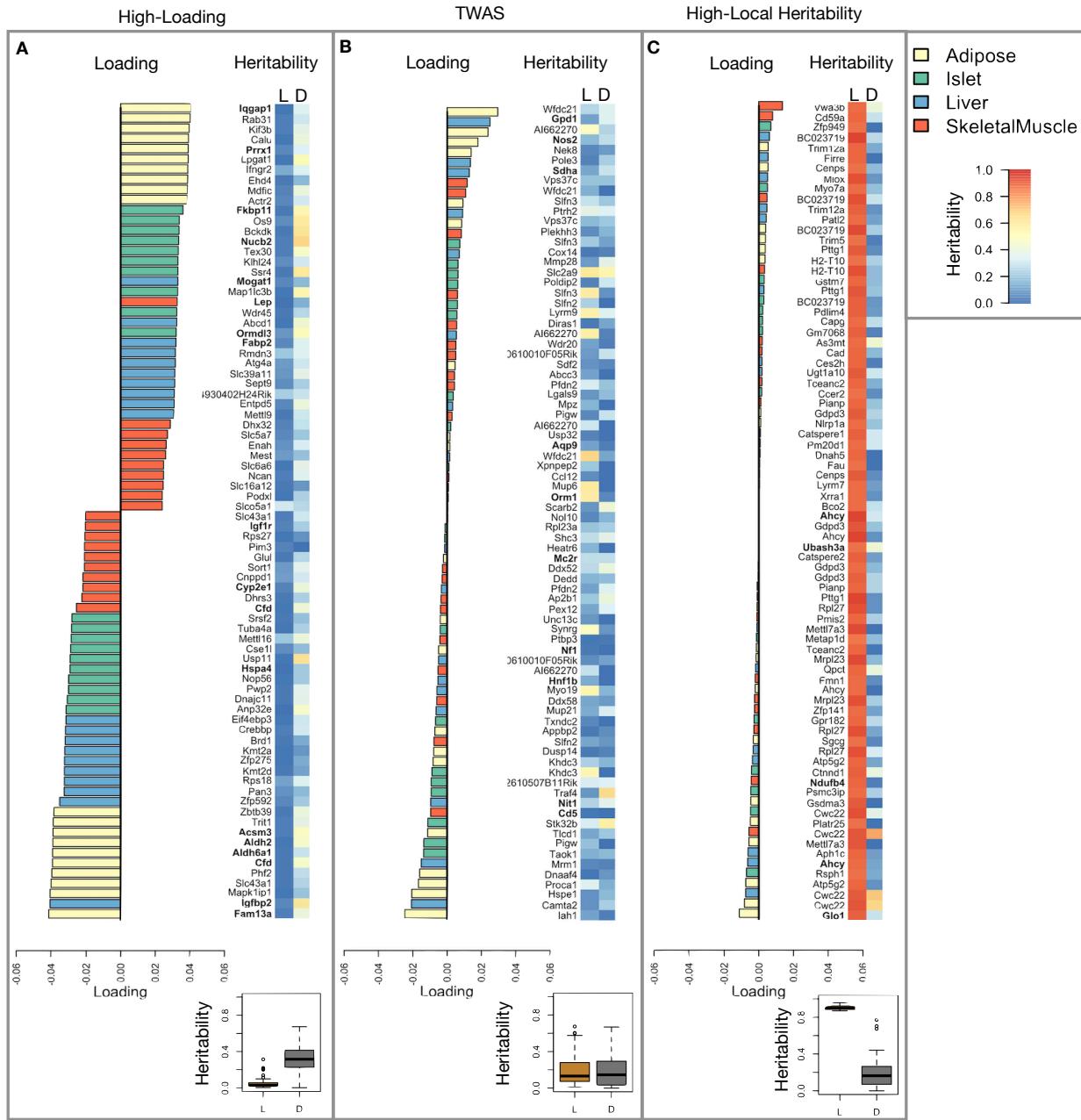
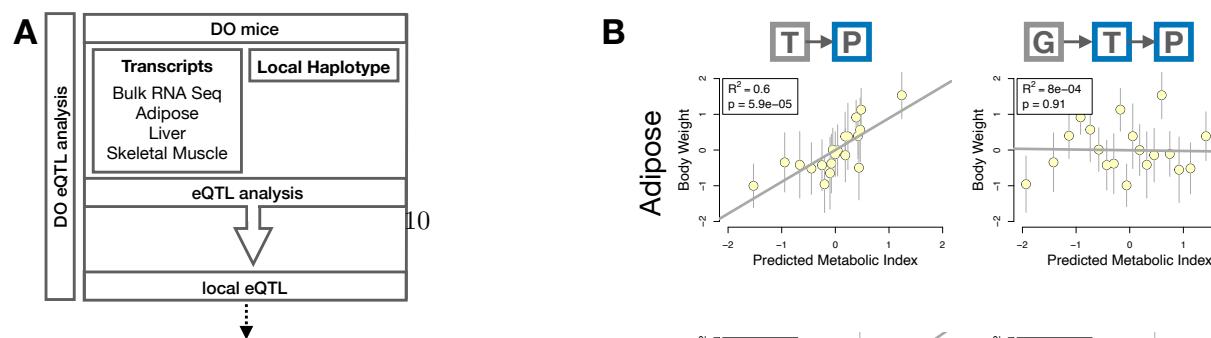


Figure 5: Transcripts with high loadings have high distal heritability and literature support. Each panel has a bar plot showing the loadings of transcripts selected by different criteria. Bar color indicates the tissue of origin. The heat map shows the local (L - left) and distal (D - right) heritability 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.



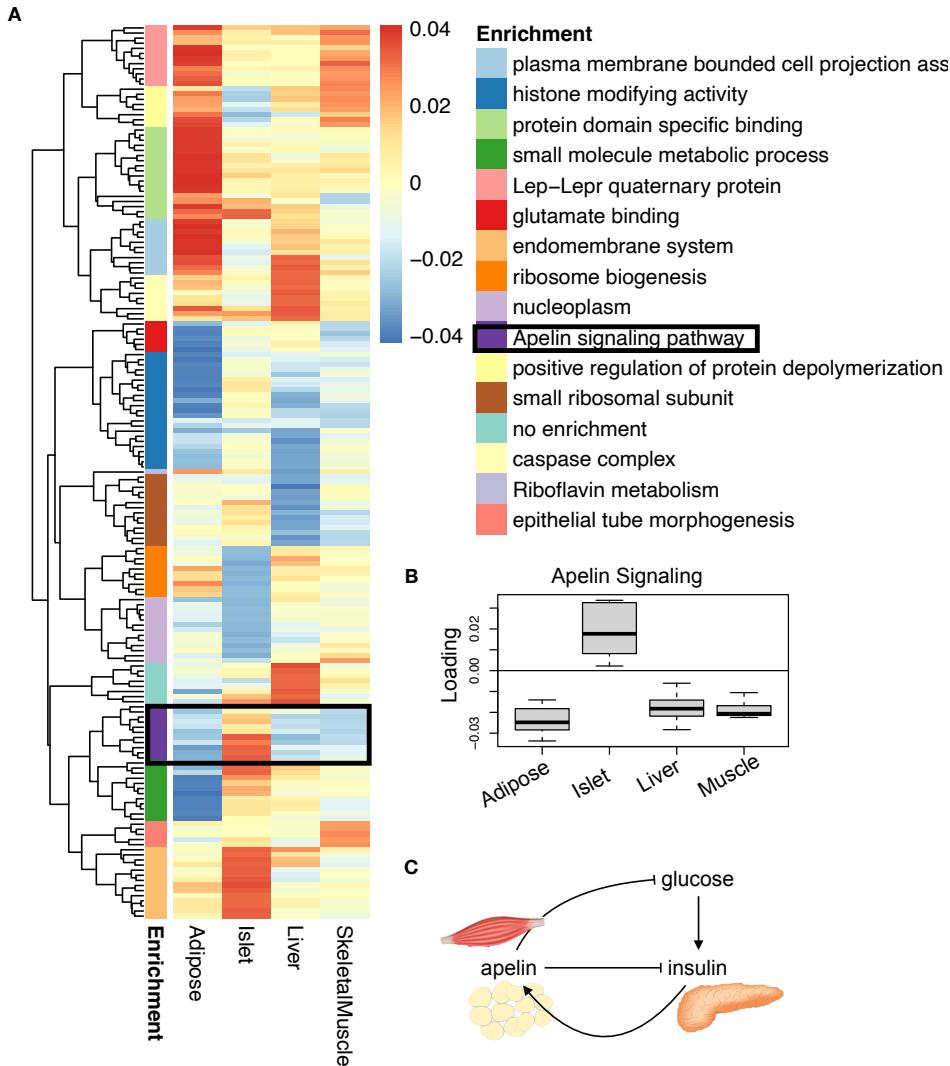


Figure 6: Tissue-specific transcriptional programs are associated with obesity and insulin resistance. **A.** Heat map showing the loadings of all transcripts with loadings greater than 2.5 standard deviations from the mean in any tissue. The heat map is hierarchically clustered. Functional enrichment of each cluster is indicated by color. An example cluster enriched for apelin signaling is highlighted. **B.** The distribution of the loadings of the apelin cluster across all tissues. These transcripts have positive loadings in the islet, and negative loadings in all other tissues. **C.** Model showing the interactions between glucose-stimulated insulin secretion by the pancreas and apelin signaling in skeletal and adipose tissue.

167 \caption{Transcription, but not local genotype, predicts phenotype in the CC-RIX. **A.** Workflow showing  
 168 procedure for translating HDM results to an independent population of mice. **B.** Relationships between the  
 169 predicted metabolic index and measured body weight. The left column shows the predictions using measured  
 170 transcripts. The right column shows the prediction using transcript levels imputed from local genotype. Gray  
 171 boxes indicate measured quantities, and blue boxes indicate calculated quantities. The dots in each panel  
 172 represent individual CC-RIX strains. The gray lines show the standard deviation on body weight for the

173 strain.

174 } \end{figure}

175 We first tested whether the transcript loadings derived from HDM in the DO were relevant to the relationship  
176 between the transcriptome and the phenotype in the CC-RIX. To do this, we multiplied the transcript loadings  
177 derived from HDM in the DO mice by transcript measurements in the CC-RIX standardized across individuals.  
178 This created a transcript vector weighted by importance to metabolic disease as determined in the DO.  
179 The mean of this vector was the predicted metabolic index for the animal based on its transcription in  
180 either adipose tissue, liver, or skeletal muscle. Across all three tissues, weighted transcription values were  
181 significantly correlated with metabolic index in the CC-RIX population measured as body weight (Fig. B left  
182 column). Adipose tissue transcription yielded the most accurate prediction (stats). This result confirms the  
183 validity and translatability of the transcript loadings determined in the DO population and their relationship  
184 to metabolic disease. It also supports the observation that transcription in adipose tissue is the strongest  
185 mediator of genetic effects on metabolic index.

186 We then tested whether this mediation signal was encoded by local genotype. To do this, we imputed gene  
187 expression in the CC-RIX using local genotype. We were able to estimate variation in gene transcription  
188 robustly. The correlation between measured gene expression and imputed gene expression across all tissues  
189 was close to  $R = 0.5$ , and the variance explained by local genotype was comparable in the DO and CC-RIX  
190 (Supp. Fig. 11). However, when weighted with the loadings derived from HDM in the DO population, these  
191 imputed transcripts across all tissues failed to predict metabolic index in the CC-RIX (Fig. B right column).  
192 Taken together, these results support the hypothesis that distal, rather than local genetic factors are primarily  
193 driving complex-trait related variation in gene expression.

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

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

203 by comparing the loadings of cell-type-specific transcripts (Methods). For adipose tissue we used a list of  
 204 cell-type specific genes identified in human adipose tissue

205 In adipose tissue, the mean loading of macrophage-specific genes was substantially greater than 0 (Fig. 7A),  
 206 indicating that obese mice were genetically predisposed to have high levels of macrophage infiltration in  
 207 adipose tissue in response to the high-fat, high-sugar diet.

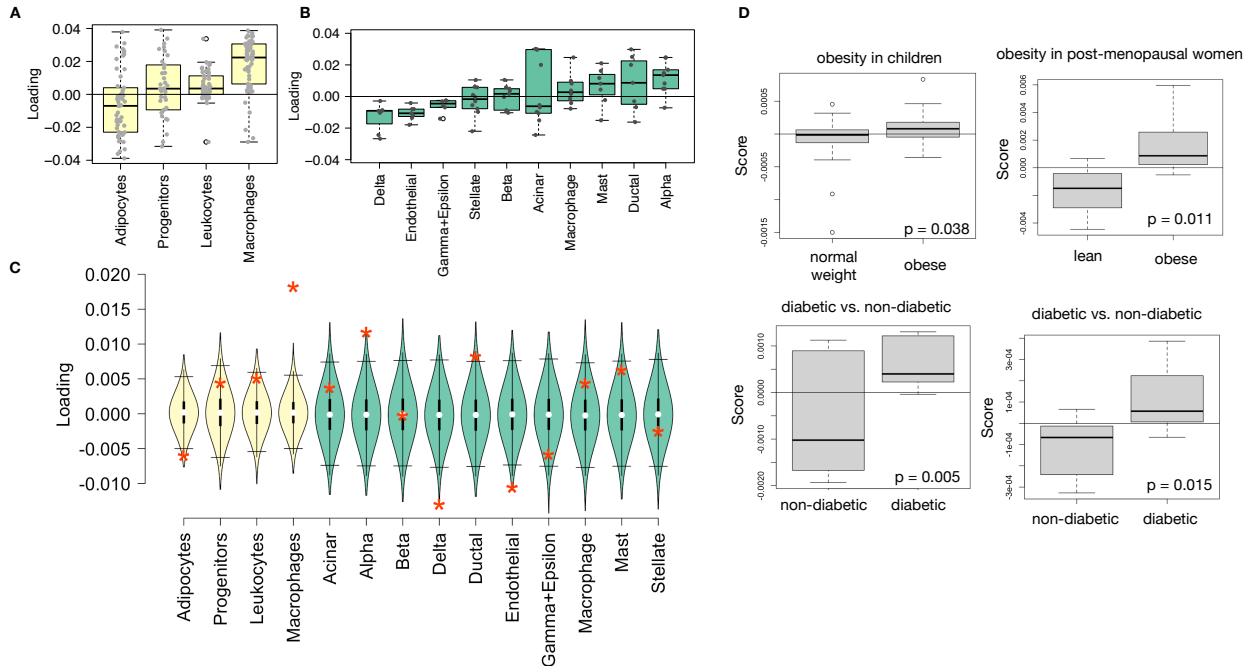


Figure 7: HDM results translate to humans. **A.** Distribution of loadings for cell-type-specific transcripts in adipose tissue. **B.** Distribution of loadings for cell-type-specific transcripts in pancreatic islets (green). **C.** Null distributions for the mean loading of randomly selected transcripts 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.

208 In islet, the mean loadings for alpha-cell specific transcripts were significantly positive, while the mean  
 209 loadings for delta- and endothelial-cell specific genes were significantly negative (Fig. 7B). These results  
 210 suggest that obese mice had inherited higher proportions of alpha cells, and lower proportions of endothelial  
 211 and delta cells in their pancreatic islets.

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

215 Biological interpretation of alpha, endothelial, delta cells??

**216 Distally heritable transcriptomic signatures translate to human disease**

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

222 We calculated a predicted obesity score for each individual in the human studies based on their adipose  
223 tissue gene expression (Methods) and compared the predicted scores for obese and non-obese groups as well  
224 as diabetic and non-diabetic groups. In all cases, the predicted obesity scores were higher on average for  
225 individuals in the obese and diabetic groups compared with the lean and non-diabetic groups, indicating that  
226 the distally heritable signature of obesity identified in DO mice is relevant to obesity and diabetes in human  
227 subjects.

**228 Targeting gene signatures**

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

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

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

- 242 • **Supplemental Table** results from CMAP

**243 Discussion**

- 244 • distal heritability correlates with phenotype relevance

<sup>245</sup> **Data Availability**

<sup>246</sup> Here we tell people where to find the data

<sup>247</sup> **Acknowledgements**

<sup>248</sup> Here we thank people

249 **Supplemental Figures**

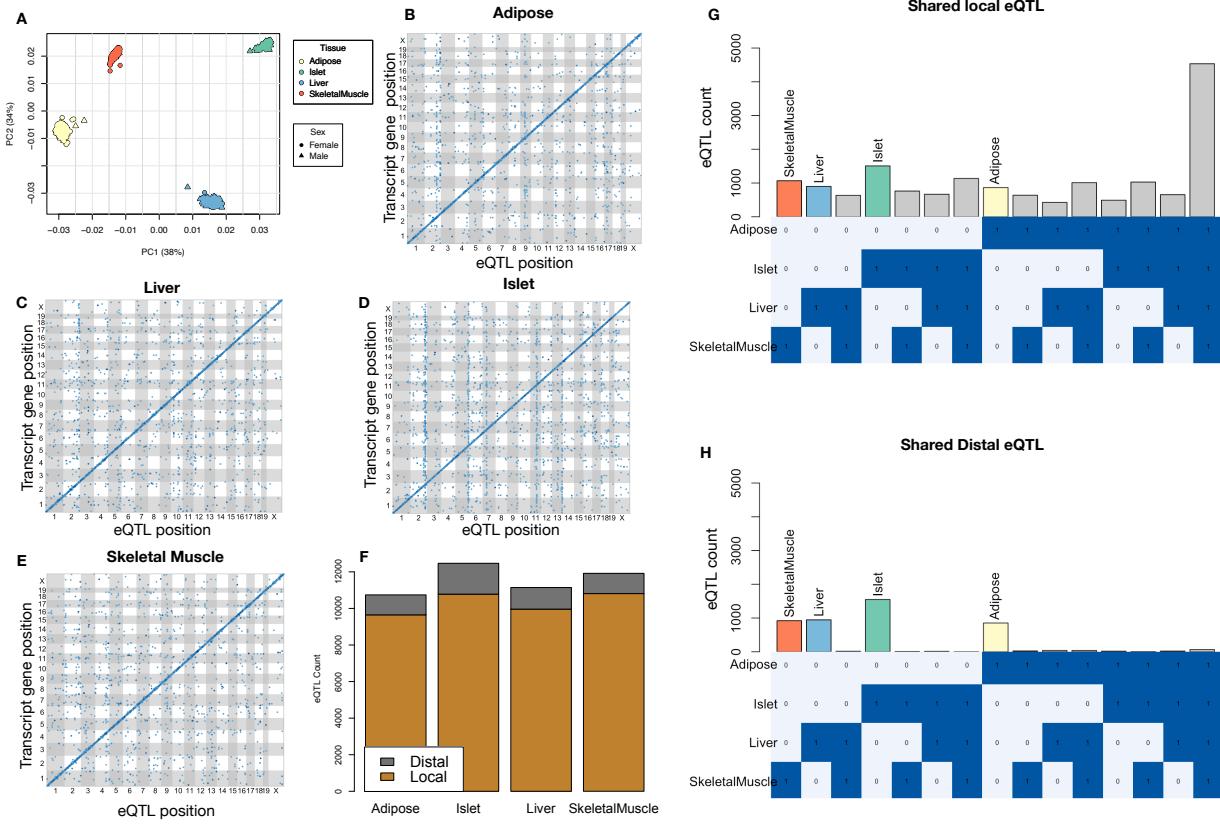


Figure 8: Overview of eQTL analysis in DO mice. **A.** RNA seq samples from the four different tissues clustered by tissue. **B.-E.** eQTL maps are shown for each tissue. The *x*-axis shows the position of the mapped eQTL, and the *y*-axis shows the physical position of the gene encoding each mapped 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.

## KEGG pathway enrichments by GSEA

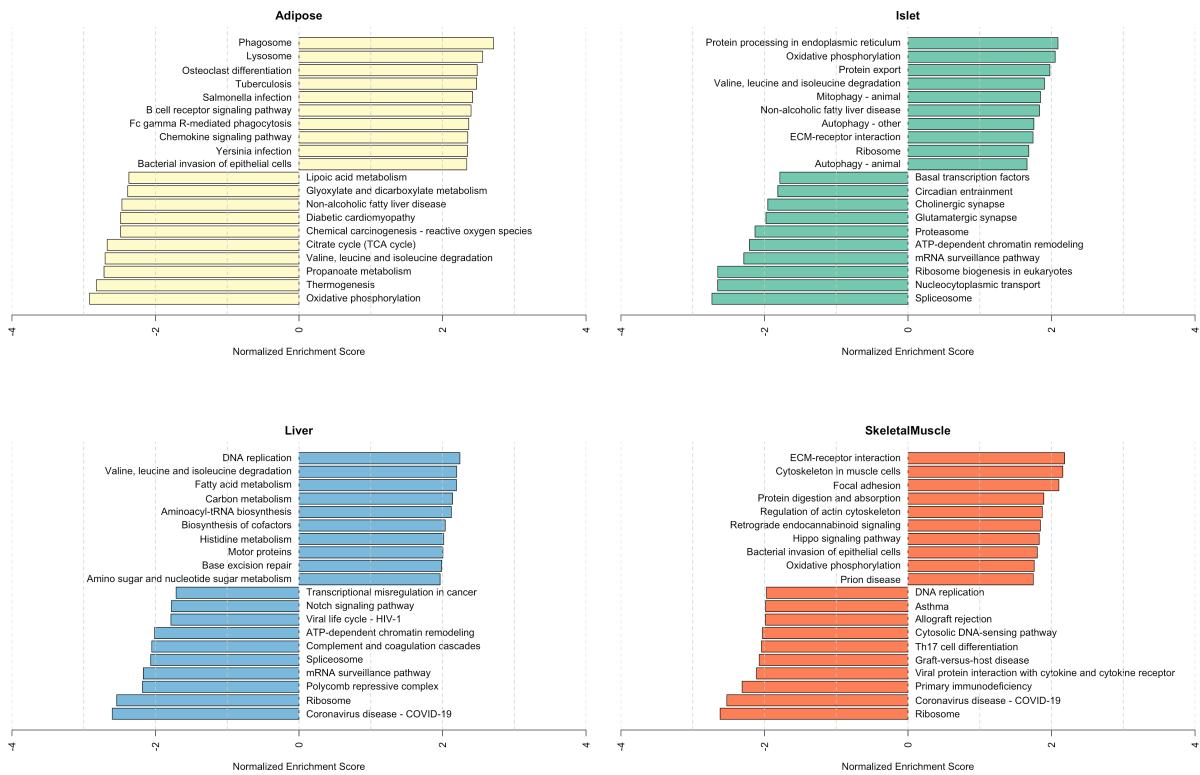


Figure 9: Bar plots showing normalized enrichment scores (NES) for KEGG pathways as determined by fast gene score enrichment analysis (fgsea). Only the top 10 positive and top 10 negative scores are shown. Colors indicate tissue. The name beside each bar shows the name of each enriched KEGG pathway.

## 250 References

- 251 [1] M. P. Keller, D. M. Gatti, K. L. Schueler, M. E. Rabaglia, D. S. Stapleton, P. Simecek, M. Vincent,  
252 S. Allen, A. T. Broman, R. Bacher, C. Kendziora, K. W. Broman, B. S. Yandell, G. A. Churchill, and  
253 A. D. Attie. Genetic Drivers of Pancreatic Islet Function. *Genetics*, 209(1):335–356, May 2018.
- 254 [2] J. H. Kemis, V. Linke, K. L. Barrett, F. J. Boehm, L. L. Traeger, M. P. Keller, M. E. Rabaglia, K. L.  
255 Schueler, D. S. Stapleton, D. M. Gatti, G. A. Churchill, D. Amador-Noguez, J. D. Russell, B. S. Yandell,  
256 K. W. Broman, J. J. Coon, A. D. Attie, and F. E. Rey. Genetic determinants of gut microbiota  
257 composition and bile acid profiles in mice. *PLoS Genet*, 15(8):e1008073, Aug 2019.
- 258 [3] J. M. Chick, S. C. Munger, P. Simecek, E. L. Huttlin, K. Choi, D. M. Gatti, N. Raghupathy, K. L.  
259 Svenson, G. A. Churchill, and S. P. Gygi. Author Correction: Defining the consequences of genetic

## Top GO term enrichments by GSEA

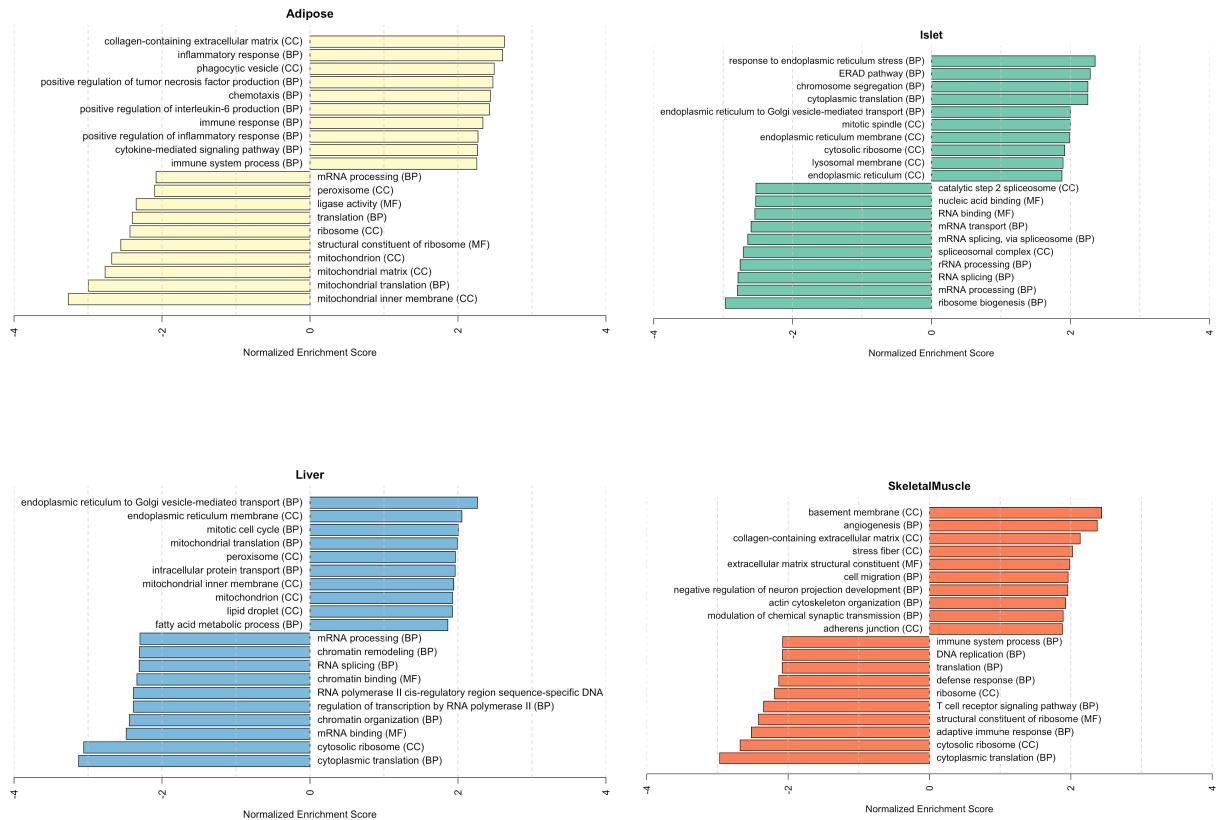


Figure 10: Bar plots showing normalized enrichment scores (NES) for GO terms as determined by fast gene score enrichment analysis (fgsea). Only the top 10 positive and top 10 negative scores are shown. Colors indicate tissue. The name beside each bar shows the name of each enriched GO term. The letters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

260 variation on a proteome-wide scale. *Nature*, 606(7915):E16, Jun 2022.

261 [4] W. L. Crouse, G. R. Keele, M. S. Gastonguay, G. A. Churchill, and W. Valdar. A Bayesian model  
262 selection approach to mediation analysis. *PLoS Genet*, 18(5):e1010184, May 2022.

263 [5] C. Yao, B. H. Chen, R. Joehanes, B. Otu, X. Zhang, C. Liu, T. Huan, O. Tastan, L. A. Cupples, J. B.  
264 Meigs, C. S. Fox, J. E. Freedman, P. Courchesne, C. J. O'Donnell, P. J. Munson, S. Keles, and D. Levy.  
265 Integromic analysis of genetic variation and gene expression identifies networks for cardiovascular disease  
266 phenotypes. *Circulation*, 131(6):536–549, Feb 2015.

267 [6] G. Jiang, A. Chakraborty, Z. Wang, M. Boustani, Y. Liu, T. Skaar, and L. Li. New aQTL SNPs for the  
268 CYP2D6 identified by a novel mediation analysis of genome-wide SNP arrays, gene expression arrays,

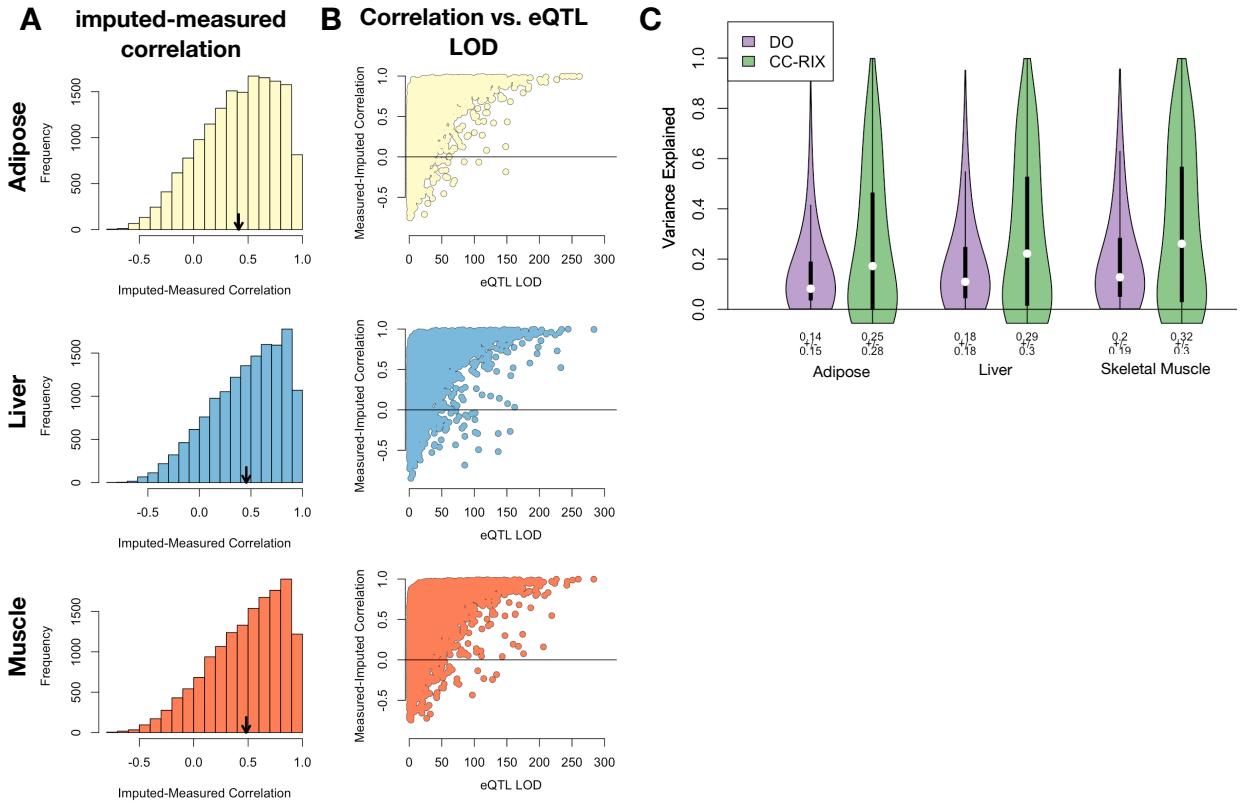


Figure 11: Validation of transcript imputation in the CC-RIX. **A.** Distributions of correlations between imputed and measured transcripts in the CC-RIX. The mean of each distribution is shown by the red line. All distributions were skewed toward positive correlations and had positive means near a Pearson 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.

269 and CYP2D6 activity. *Biomed Res Int*, 2013:493019, 2013.

- 270 [7] D. W. Yao, L. J. O'Connor, A. L. Price, and A. Gusev. Quantifying genetic effects on disease mediated  
271 by assayed gene expression levels. *Nat Genet*, 52(6):626–633, Jun 2020.
- 272 [8] H. J. Westra, M. J. Peters, T. Esko, H. Yaghootkar, C. Schurmann, J. Kettunen, M. W. Christiansen,  
273 B. P. Fairfax, K. Schramm, J. E. Powell, A. Zhernakova, D. V. Zhernakova, J. H. Veldink, L. H. Van den  
274 Berg, J. Karjalainen, S. Withoff, A. G. Uitterlinden, A. Hofman, F. Rivadeneira, P. A. C. ' Hoen,  
275 E. Reinmaa, K. Fischer, M. Nelis, L. Milani, D. Melzer, L. Ferrucci, A. B. Singleton, D. G. Hernandez,  
276 M. A. Nalls, G. Homuth, M. Nauck, D. Radke, U. Iker, M. Perola, V. Salomaa, J. Brody, A. Suchy-Dicey,  
277 S. A. Gharib, D. A. Enquobahrie, T. Lumley, G. W. Montgomery, S. Makino, H. Prokisch, C. Herder,  
278 M. Roden, H. Grallert, T. Meitinger, K. Strauch, Y. Li, R. C. Jansen, P. M. Visscher, J. C. Knight,  
279 B. M. Psaty, S. Ripatti, A. Teumer, T. M. Frayling, A. Metspalu, J. B. J. van Meurs, and L. Franke.

- 280 Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat Genet*,  
281 45(10):1238–1243, Oct 2013.
- 282 [9] S. M. Clee and A. D. Attie. The genetic landscape of type 2 diabetes in mice. *Endocr Rev*, 28(1):48–83,  
283 Feb 2007.
- 284 [10] C. R. Green, M. Wallace, A. S. Divakaruni, S. A. Phillips, A. N. Murphy, T. P. Ciaraldi, and C. M.  
285 Metallo. Branched-chain amino acid catabolism fuels adipocyte differentiation and lipogenesis. *Nat  
286 Chem Biol*, 12(1):15–21, Jan 2016.
- 287 [11] D. E. Lackey, C. J. Lynch, K. C. Olson, R. Mostaedi, M. Ali, W. H. Smith, F. Karpe, S. Humphreys,  
288 D. H. Bedinger, T. N. Dunn, A. P. Thomas, P. J. Oort, D. A. Kieffer, R. Amin, A. Bettaieb, F. G.  
289 Haj, P. Permana, T. G. Anthony, and S. H. Adams. Regulation of adipose branched-chain amino acid  
290 catabolism enzyme expression and cross-adipose amino acid flux in human obesity. *Am J Physiol  
291 Endocrinol Metab*, 304(11):E1175–1187, Jun 2013.