

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. By identifying transcripts that mediate the effect of genetic loci on traits,  
11 we get one step closer to a mechanistic understanding of the influence of genetic variants on traits. There is  
12 evidence from genome-wide association studies (GWAS) that regulation of gene expression accounts for the  
13 bulk of the genetic effect on complex traits, as most trait-associated variants lie in gene regulatory regions  
14 [1, 2, 3, 4, 5, 6, 7]. It is widely assumed that these variants influence local transcription, and methods such  
15 as transcription-wide association studies (TWAS) [8, 9, 10] summary data-based Mendelian randomization  
16 (SMR) [9], and others [cite] have capitalized on this idea to identify genes associated with multiple disease  
17 traits [cite many]

18 Despite the great promise of these methods, however, they have not been as widely successful as it seemed  
19 they could have been, and the vast majority of complex trait heritability remains unexplained. Although  
20 trait-associated variants tend to lie in non-coding, putative regulatory regions, they often do not have  
21 detectable effects on gene expression [11] and do not tend to co-localize with eQTL [12, 13].

22 One possible explanation for these observations is that we are not measuring gene expression in the appropriate  
23 cell types and are thus unable to detect true eQTLs influencing traits [11]. An alternative explanation  
24 that has been discussed in recent years is that heritability of these variants is mediated not through local  
25 regulation of gene expression, but through distal regulation [13, 14]. Yao *et al.* [14] observed that genes

26 with low local heritability explain more expression-mediated disease heritability than genes with high local  
27 heritability. This observation is consistent with principles of robustness in complex systems. If a transcript  
28 were both important to a trait and subject to strong local regulation, a population would be susceptible to  
29 extremes in phenotype that could be defined as disease. Indeed, strong disruption of trait-important genes is  
30 the cause of Mendelian disease. Instead, what has been observed is that genes that are near GWAS hits and  
31 have obvious functional relevance to a trait tend to have highly complex regulatory landscapes under strong  
32 selection pressures [13]. In contrast, genes with strong local regulation tend to be depleted of functional  
33 annotations and are under looser selection constraints [13]. These observations and others led Liu et al. [15] to  
34 suggest that most heritability of complex traits is driven by weak trans-eQTLs. They proposed a framework  
35 of understanding heritability of complex traits in which massive polygenicity is distributed across common  
36 variants in both functional core genes, as well as more peripheral genes that may not seem obviously related  
37 to the trait.

38 If this model is correct, it should be the case that...

39 With dense enough data, it should be possible to test this model.

40 Here we investigated the roles of local and distal gene regulation on the heritability of complex metabolic  
41 traits in a population of diversity outbred (DO) mice. DO mice are derived from eight inbred founder mouse  
42 strains, five classical lab strains, and three strains more recently derived from wild mice [16]. They represent  
43 three subspecies of mouse *Mus musculus domesticus*, *Mus musculus musculus*, and *Mus musculus castaneus*,  
44 and capture 90% of the known variation in laboratory mice [cite]. We placed a population of 500 mice, both  
45 male and female, on a high-fat, high-sugar diet [17] to induce diet-associated obesity and metabolic disease.  
46 Over 18 weeks multiple metabolic traits were measured, including body weight, plasma levels of insulin and  
47 glucose, and plasma lipids. At the end of the experiment, we used RNASeq to measure gene expression in  
48 384 mice in four tissues relevant to metabolic disease: adipose tissue, pancreatic islets, liver, and skeletal  
49 muscle. The mice were also genotyped using the Mouse Universal Genotyping Array (GigaMUGA). The  
50 high dimensionality of phenotyping, genotyping, and transcriptome measurements in this large number of  
51 genetically diverse animals enabled thorough inquiry into the role of local and distal eQTL in the heritability  
52 of complex traits. Further, measuring gene expression in multiple tissues, enabled us to assess the importance  
53 of measuring gene expression in appropriate cell types.

54 To investigate the extent to which local and distal gene regulation across multiple tissues mediate genetic  
55 effects on complex traits, we propose a systems approach called high-dimensional mediation (HDM). HDM  
56 uses a regularized and generalized canonical correlation analysis (RGCCA) [cite], which is an extension of

57 canonical correlation analysis (CCA) that allows for more than two data sets with arbitrary relationships  
58 among them. Thus, we can identify linear combinations of the genome, transcriptome, and phenotype, that  
59 describe the mediation of the genetic effects on the phenotype through the transcriptome. This systems  
60 approach, as opposed to a univariate approach, puts local and distal regulation on the same statistical level

61 Because of the central dogma of molecular biology, information flow is directed out of the genome, and not  
62 back into it. Thus, the otherwise undirected relationships between genome, transcriptome, and phenotype can  
63 be inferred as a causal mediation by the transcriptome of the effects of the genome on the phenotype.

## 64 Results

### 65 Genetic variation contributes to wide phenotypic variation

66 A population of 500 diversity outbred mice (split evenly between male and female) from generates 18, 19,  
67 and 21, was placed on a high-fat (44.6% kcal fat), high-sugar (34% carbohydrate), adequate protein (17.3  
68 % protein) diet from Envigo Teklad (catalog number TD.08811) starting at four weeks of age as described  
69 previously [17].

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

72 Although the environment was consistent across animals, the genetic diversity present in this population  
73 resulted in widely varying distributions across physiological measurements (Fig. 1). For example, body  
74 weights of adult individuals varied from less than the average adult B6 body weight to several times the body  
75 weight of a B6 adult in both sexes (Fig. 1A). Fasting blood glucose (FBG) also varied considerably (Fig. 1B)  
76 although few of the animals had FBG levels that would indicate pre-diabetes ( animals, ), or diabetes (7  
77 animals, 1.4) according to previously developed cutoffs (pre-diabetes:  $FBG \geq 250$  mg/dL, diabetes:  $FBG \geq$   
78 300, mg/dL) [18]. Males had higher FBG than females on average (Fig. 1C) as has been observed before  
79 suggesting either that males were more susceptible to metabolic disease on the high-fat diet, or that males  
80 and females may require different thresholds for pre-diabetes and diabetes.

81 Body weight was strongly positively correlated with food consumption (Fig. 1D  $R^2 = 0.51, p = 1.5 \times 10^{-75}$ )  
82 and fasting blood glucose (FBG) (Fig. 1E,  $R^2 = 0.25, p = 2 \times 10^{-32}$ ) suggesting a link between behavioral  
83 factors and metabolic disease. However, the heritability of this trait and others (Fig. 1F) indicates that  
84 background genetics contribute substantially to correlates of metabolic disease in this population.

85 The landscape of trait correlations (Fig. 1G) shows that most of the metabolic trait pairs were relatively  
86 weakly correlated indicating complex relationships among the measured traits. This low level of redundancy

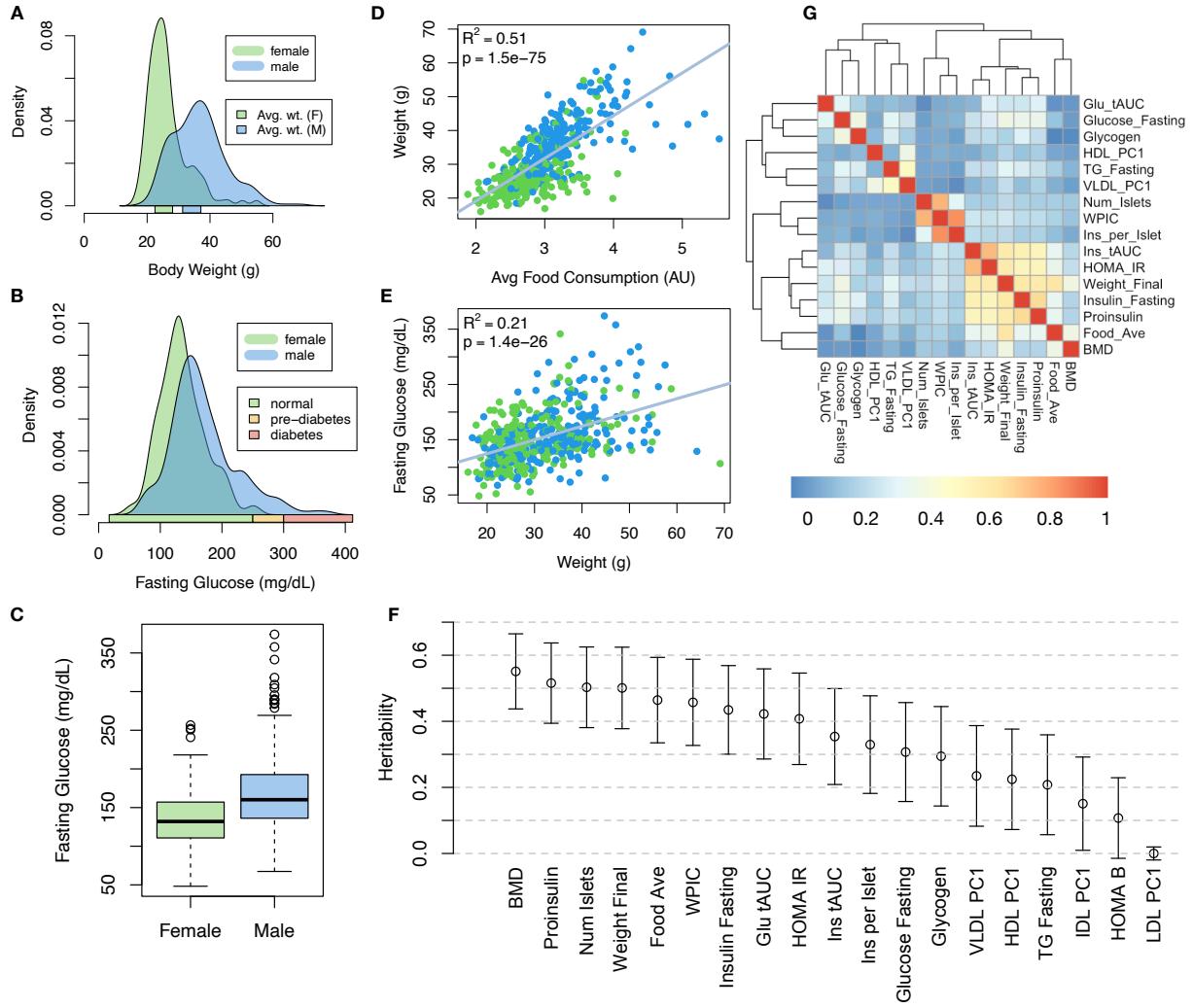


Figure 1: Clinical overview. **A.** Distributions of final body weight in the diversity outbred mice. Sex is indicated by color. The average B6 male and female adult weights at 24 weeks of age are indicated by blue and green bars on the x-axis. **B.** The distribution of final fasting glucose across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

suggests a broad sampling of multiple heritable aspects of metabolic disease including overall body weight, glucose homeostasis, pancreatic composition and liver function.

#### 89 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).

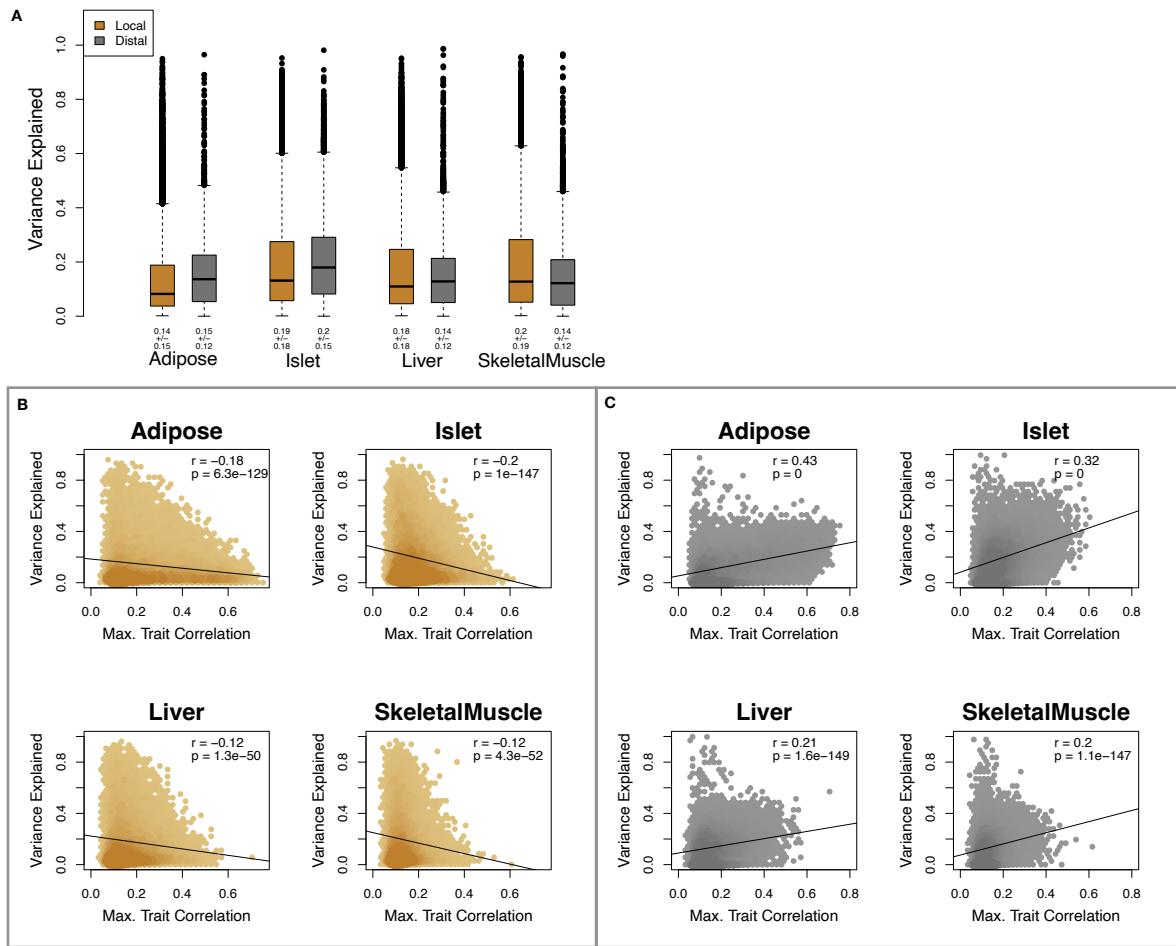


Figure 2: Transcript heritability and trait relevance. **A.** Distributions of distal and local heritability of transcripts across the four tissues. Overall local and distal factors contribute equally to transcript heritability. The relationship between **(B.)** local and **(C.)** distal heritability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

101 Local heritability of transcripts was negatively correlated with their trait relevance, defined as the maximum  
102 correlation of a transcript across all traits (Fig. 2B). This suggests that the more local genotype influenced  
103 transcript abundance, the less effect variation in transcript abundance was related to the measured traits.  
104 Conversely, distal heritability of transcripts was positively correlated with trait relevance (Fig. 2C). That is,  
105 transcripts that were more highly correlated with the measured traits tended to be distally, rather than locally,  
106 heritable. That trait-correlated transcripts have low local heritability is consistent with previous observations  
107 that low-heritability transcripts explain more expression-mediated disease heritability than high-heritability  
108 transcripts [14]. However, the positive relationship between trait correlation and distal heritability suggests  
109 that there are alternative mechanisms through which genetic regulation of transcripts may influence traits.

110 **High-Dimensional Mediation identifies composite transcript that perfectly mediates composite  
111 trait**

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

118 The result of this process is three vectors representing the composite genome ( $G_C$ ), composite transcriptome  
119 ( $T_C$ ) and the composite phenome ( $P_C$ ) where the composite transcriptome perfectly mediates the effect of the  
120 composite genome on the composite phenome. Each vector is of length  $n$  where  $n$  is the number of individual  
121 mice. Fig. 3A shows the partial correlations between all pairs of composite vectors. The partial correlation  $r$   
122 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  
123 transcriptome was taken into account, the partial correlation between  $G_S$  and  $P_S$  was effectively 0 (-0.01).

124 Standard CCA is prone to over-fitting because in any two large matrices it can be trivial to identify  
125 highly correlated composite vectors. To assess whether RGCCA was similarly prone to over-fitting in a  
126 high-dimensional space, we performed permutation testing. We permuted the individual labels on the  
127 transcriptome kernel matrix 1000 times and recalculated the path coefficient, which is the partial correlation  
128 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  
129  $P_C$  that is mediated through  $T_C$ . The null distribution of the path coefficient is shown in Fig. 3B, and the  
130 observed path coefficient from the original data is indicated by the red line. The observed path coefficient  
131 was well outside the null distribution generated by permutations. Fig. 3C illustrates this observation in more

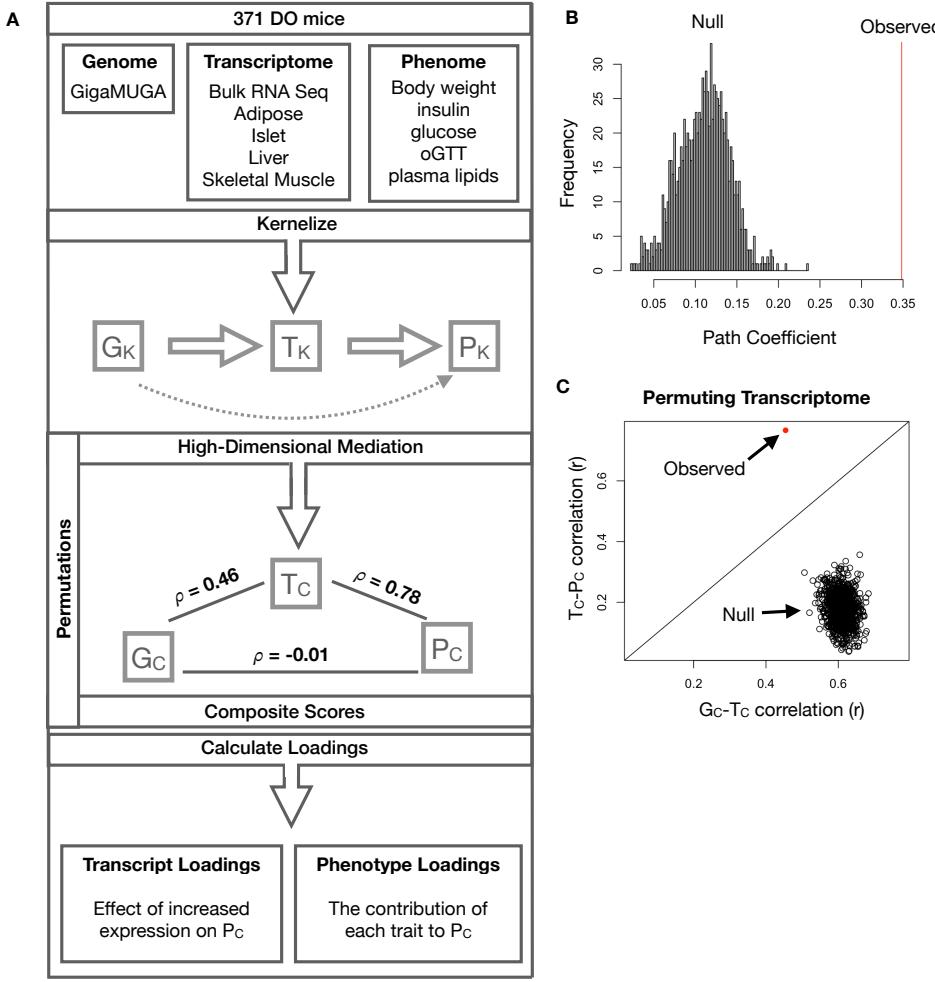


Figure 3: High-dimensional mediation. **A.** Workflow indicating major steps of high-dimensional mediation. The genotype, transcriptome, and phenotype matrices were kernelized to yield single matrices representing the relationships between all individuals for each data modality ( $G_K$  = genome kernel,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

132 detail. Although we identified high correlations between  $G_c$  and  $T_c$ , and modest correlations between  $T_c$  and  
 133  $P_c$  in the null data (Fig 3C), these two values could not be maximized simultaneously. The red dot shows that  
 134 in the real data both the  $G_c-T_c$  correlation and the  $T_c-P_c$  correlation could be maximized simultaneously  
 135 suggesting that that path from genotype to phenotype through transcriptome is highly non-trivial and  
 136 identifiable in this case. These results suggest that these composite vectors represent genetically determined  
 137 variation in phenotype that is mediated through genetically determined variation in transcription.

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

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

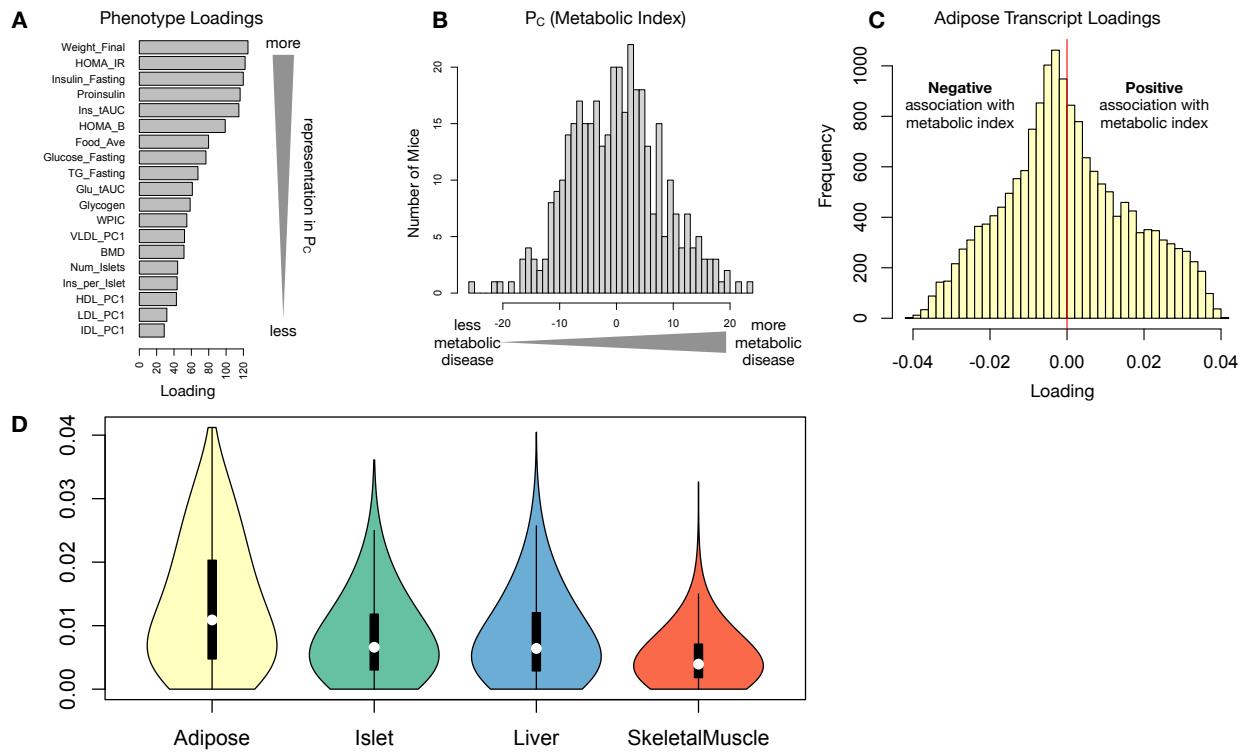


Figure 4: Interpretation of loadings. **A.** Loadings across traits. Body weight and insulin resistance contributed the most to the composite trait. **B.** Phenotype scores across individuals. Individuals with large positive phenotype scores had higher body weight and insulin resistance than average.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

<sup>148</sup> **High-loading transcripts have low local heritability, high distal heritability, and are linked**  
<sup>149</sup> **mechanistically to obesity**

<sup>150</sup> We interpreted large loadings onto transcripts as indicating strong mediation of the effect of genetics on  
<sup>151</sup> metabolic index. Large positive loadings indicate that inheriting higher expression was associated with a  
<sup>152</sup> higher metabolic index (i.e. higher risk of obesity and metabolic disease on the high-fat diet) (Fig. 4C).  
<sup>153</sup> Conversely, large negative loadings indicate that inheriting lower expression of these transcripts was associated  
<sup>154</sup> with a lower metabolic index (i.e. lower risk of obesity and metabolic disease on the high-fat diet) (Fig. 4C).  
<sup>155</sup> We used GSEA to look for biological processes and pathways that were enriched at the top and bottom of  
<sup>156</sup> this list (Methods).

<sup>157</sup> In adipose tissue, both GO processes and KEGG pathway enrichments pointed to an axis of inflammation  
<sup>158</sup> and metabolism (Supp. Fig. 10 and 11). Processes and pathways associated with inflammation, particularly  
<sup>159</sup> macrophage infiltration were positively associated with metabolic index, indicating that increased expression  
<sup>160</sup> in inflammatory pathways was associated with a higher metabolic index. It is well established that adipose  
<sup>161</sup> tissue in obese individuals is highly inflamed [cite] and infiltrated by macrophages [cite], and the results here  
<sup>162</sup> suggest that this may be a heritable component of metabolic disease.

<sup>163</sup> The strongest negative enrichments in adipose tissue were related to mitochondrial activity in general, and  
<sup>164</sup> thermogenesis in particular (Supp. Fig. 10 and 11). It has been shown mouse strains with greater thermogenic  
<sup>165</sup> potential are also less susceptible to obesity on a high-fat diet.

<sup>166</sup> Transcripts associated with the citric acid (TCA) cycle as well as the catabolism of branched-chain amino acids  
<sup>167</sup> (BCAA), valine, leucine, and isoleucine also had strong negative enrichment in the adipose tissue (Supp. Fig.  
<sup>168</sup> XXX). Expression of genes in both pathways (for which there is some overlap) has been previously associated  
<sup>169</sup> with insulin sensitivity [17, 19, 20], suggesting that impairment in these pathways may be associated with  
<sup>170</sup> insulin resistance. Selective PPAR $\gamma$  modulation by insulin-sensitizing thiazolidinedione drugs has further  
<sup>171</sup> been shown to influence both inflammation and BCAA metabolism in obese rats suggesting a relationship  
<sup>172</sup> between these pathways and insulin resistance [21]. BCAA levels are also related to insulin resistance in  
<sup>173</sup> human subjects and are elevated in insulin-resistant obese individuals relative to weight-matched non-insulin  
<sup>174</sup> resistant individuals [22]. In the DO mice studied here, inheriting increased expression of genes involved in  
<sup>175</sup> BCAA catabolism was associated with reduced body weight and insulin resistance.

<sup>176</sup> Transcripts in the adipose tissue had the largest loadings, both positive and negative, of all tissues, suggesting  
<sup>177</sup> that much of the effect of genetics on body weight and insulin resistance is mediated through gene expression  
<sup>178</sup> 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, transcripts 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 heritability (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).

#### **186    Tissue-specific transcriptional programs are associated with metabolic traits**

187    Clustering of transcripts with top loadings in each tissue shows tissue-specific functional modules associated  
188    with obesity and insulin resistance in the DO population (Fig. 6A). In this figure, the importance of immune  
189    activation specifically in the adipose tissue is apparent. There are also other tissue-specific processes. Positive  
190    loadings on lipid metabolism in liver suggest that inheriting high liver expression of genes in this cluster is  
191    positively associated with metabolic disease. This cluster included the gene *Pparg*, whose primary role is in  
192    the adipose tissue where it is considered a master regulator of adipogenesis [23]. Agonists of *Pparg*, such  
193    as Thiazolidinediones, which are FDA-approved to treat type II diabetes, reduce inflammation and adipose  
194    hypertrophy [23]. Consistent with this role, the loading for *Pparg* in adipose tissue is slightly negative,  
195    suggesting that upregulation is associated with leaner mice (Fig. 6B). In contrast, *Pparg* has a large positive  
196    loading in liver, where it plays a role in the development of hepatic steatosis, or fatty liver. Mice that lack  
197    *Pparg* specifically in the liver, are protected from developing steatosis and show reduced expression of lipogenic  
198    genes [24, 25]. Overexpression of *Pparg* in the livers of mice with a *Ppara* knockout, causes upregulation of  
199    genes involved in adipogenesis [26]. In the livers of both mice and humans [27, 28] High *Pparg* expression is  
200    associated with hepatocytes that accumulate large lipid droplets and have gene expression profiles similar to  
201    adipocytes.

202    The local and distal heritability of *Pparg* is low in adipose tissue suggesting its expression in this tissue is  
203    highly constrained in the population (Fig. 6B). However, the distal heritability of *Pparg* in liver is relatively  
204    high suggesting it is complexly regulated and has sufficient variation in this population to drive variation  
205    in phenotype. Both local and distal heritability of *Pparg* in the islet are fairly high, but the loading is  
206    low, suggesting that variability of expression in the islet does not drive phenotypic variation. These results  
207    highlight the importance of tissue context when investigating the role of heritable transcript variability in  
208    driving phenotype.

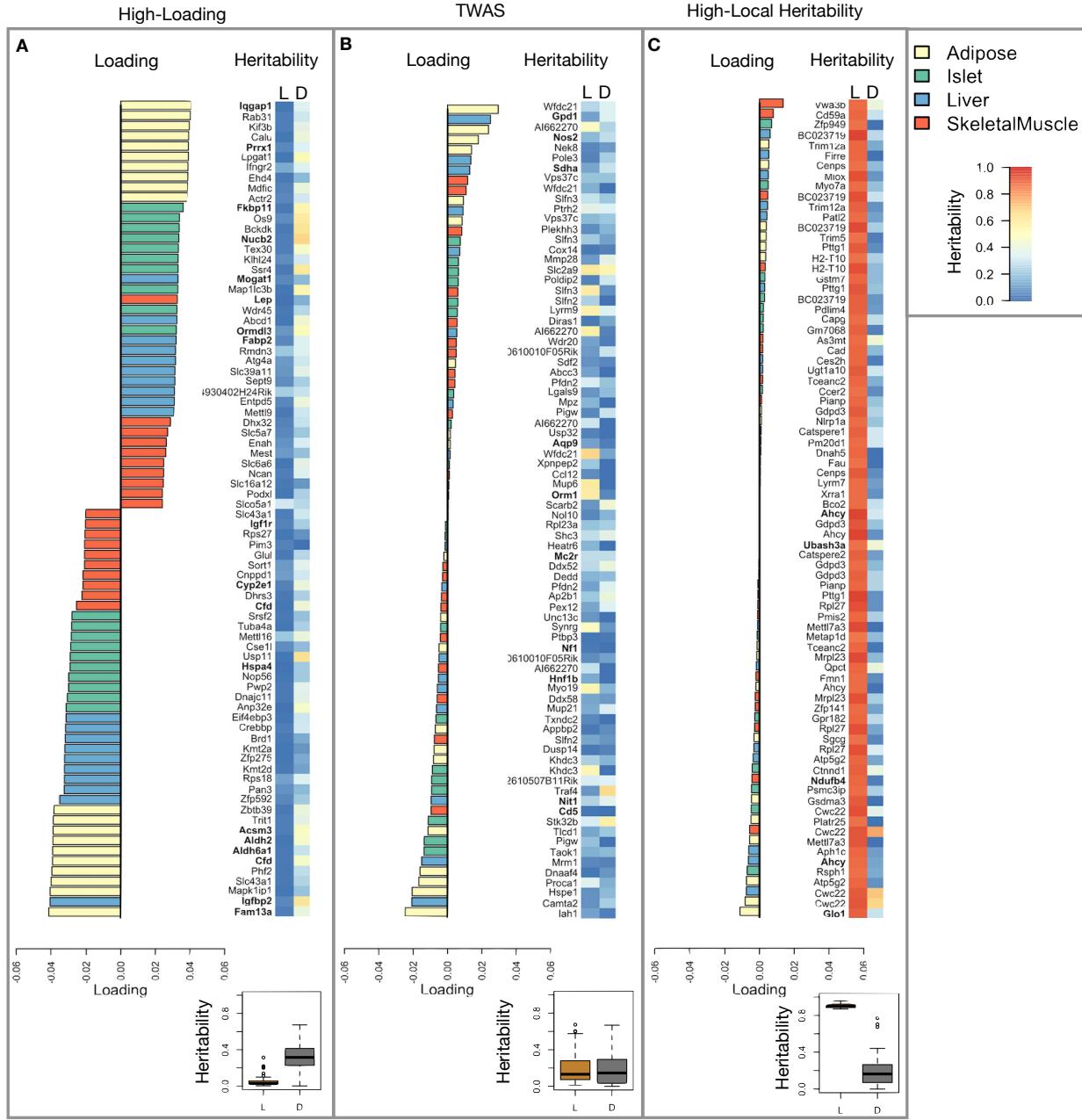


Figure 5: Transcripts with high loadings have high distal heritability and literature support. Each panel has a bar plot showing the loadings of transcripts selected by different criteria. Bar color indicates the tissue of origin. The heat map shows the local (L - left) and distal (D - right) heritability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

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

210 The loading of each transcript indicates how inherited expression levels influence metabolic phenotypes.

211 If local regulation is the predominant factor influencing gene expression, we should be able to predict an

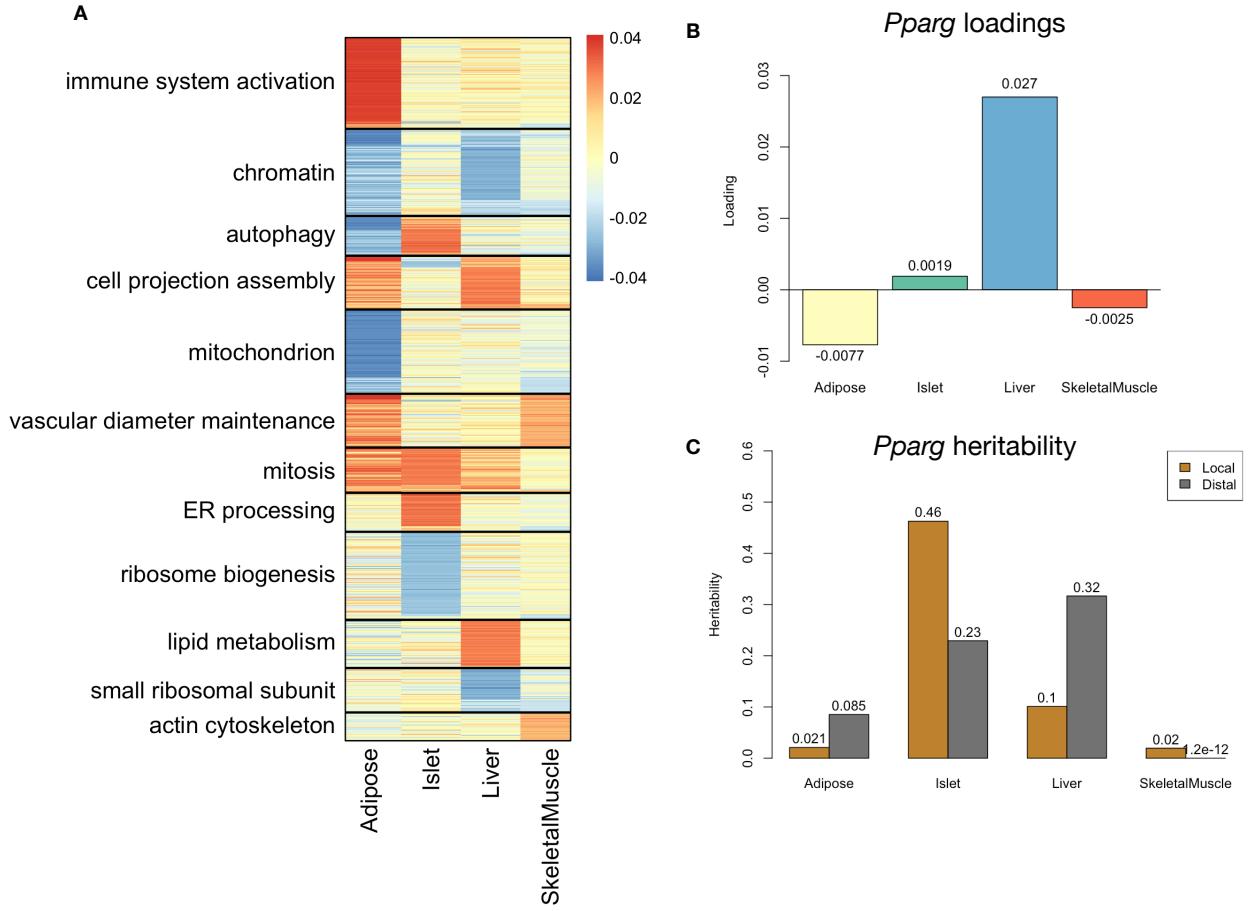


Figure 6: Tissue-specific transcriptional programs are associated with obesity and insulin resistance. **A** Heat map showing the loadings of all transcripts with loadings greater than 2.5 standard deviations from the mean in any tissue. The heat map was clustered using k medoid clustering. Functional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

212 individual's phenotype based on their genotypes across all local eQTLs. We tested this hypothesis in an  
 213 independent population of F1 mice generated through multiple pairings of Collaborative Cross (CC) [cite]  
 214 strains (Fig. 7A) (Methods).  
 215 We first tested whether the transcript loadings derived from HDM in the DO were relevant to the relationship  
 216 between the transcriptome and the phenotype in the CC-RIX. To do this, we multiplied the transcript loadings  
 217 derived from HDM in the DO mice by transcript measurements in the CC-RIX standardized across individuals.  
 218 This created a transcript vector weighted by importance to metabolic disease as determined in the DO.  
 219 The mean of this vector was the predicted metabolic index for the animal based on its transcription in  
 220 either adipose tissue, liver, or skeletal muscle. Across all three tissues, weighted transcription values were  
 221 significantly correlated with metabolic index in the CC-RIX population measured as body weight (Fig. 7B left

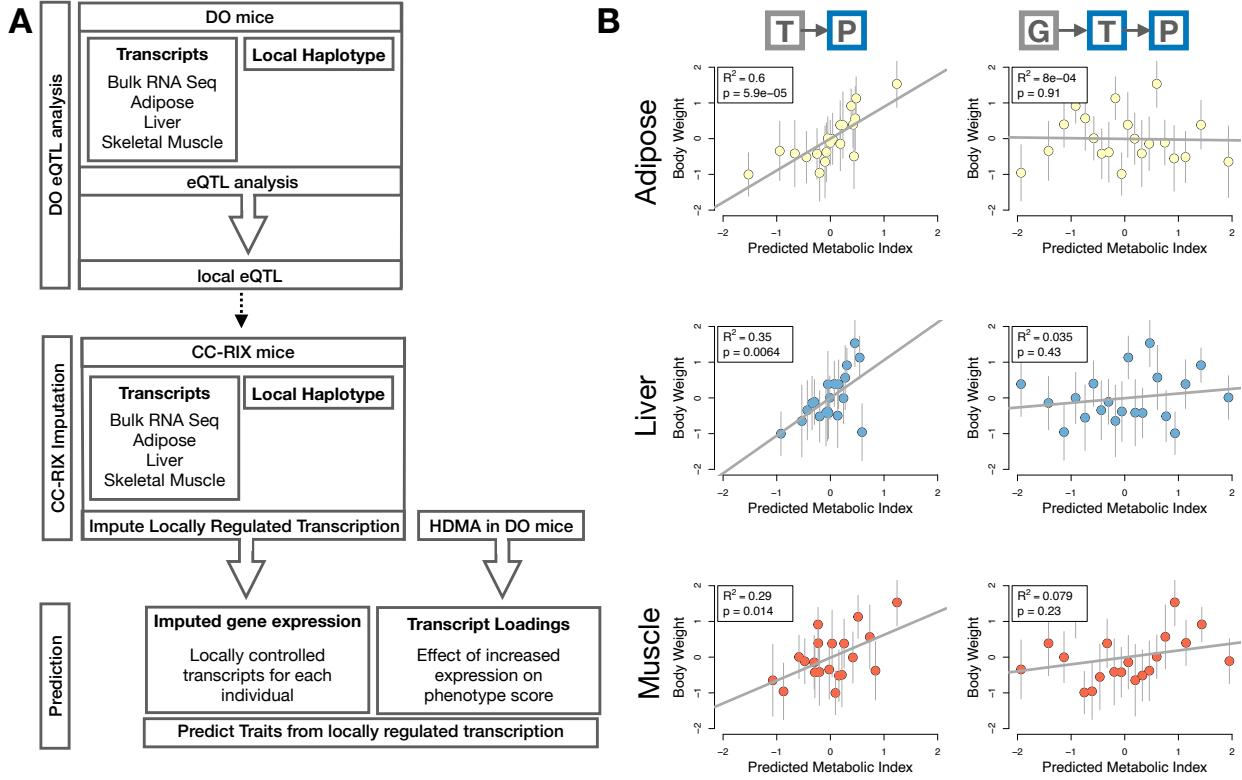


Figure 7: Transcription, but not local genotype, predicts phenotype in the CC-RIX. **A.** Workflow showing procedure for translating HDM results to an independent population of mice. **B.** Relationships between the predicted metabolic index and measured body weight. The left column shows the predictions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

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

226 We then tested whether this mediation signal was encoded by local genotype. To do this, we imputed gene  
 227 expression in the CC-RIX using local genotype. We were able to estimate variation in gene transcription  
 228 robustly. The correlation between measured gene expression and imputed gene expression across all tissues  
 229 was close to  $R = 0.5$ , and the variance explained by local genotype was comparable in the DO and CC-RIX  
 230 (Supp. Fig. 12). However, when weighted with the loadings derived from HDM in the DO population, these  
 231 imputed transcripts across all tissues failed to predict metabolic index in the CC-RIX (Fig. 7B right column).  
 232 Taken together, these results support the hypothesis that distal, rather than local genetic factors are primarily

233 driving complex-trait related variation in gene expression.

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

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

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

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

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

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

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

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

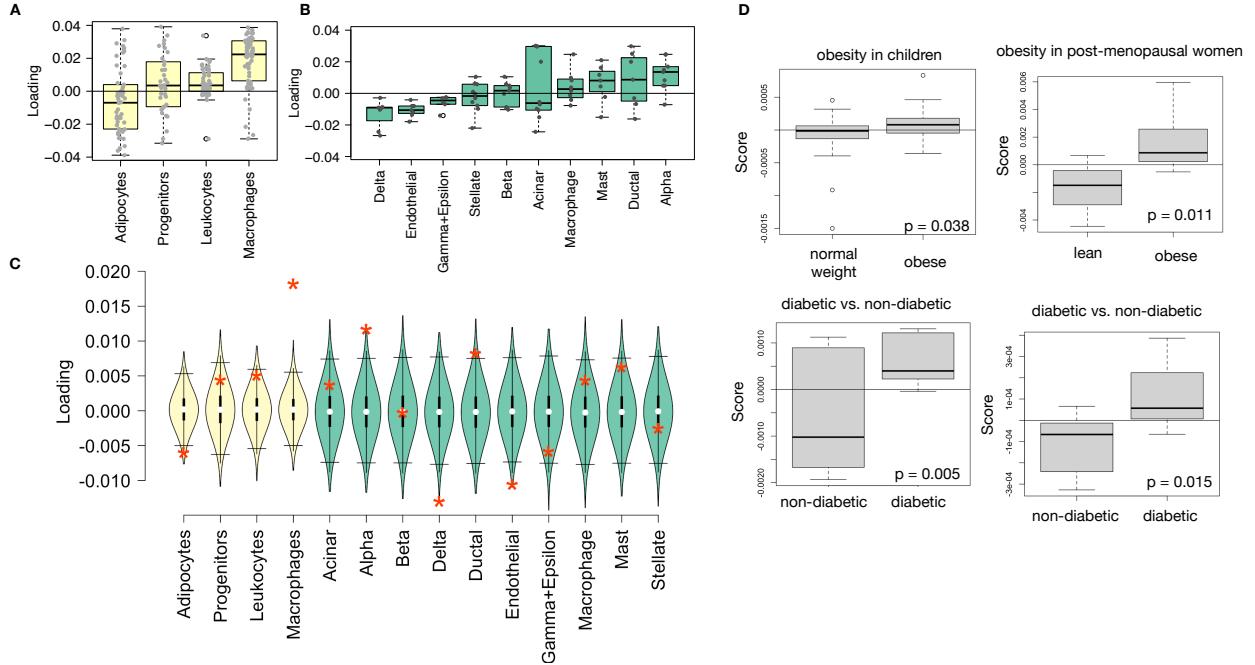


Figure 8: HDM results translate to humans. **A.** Distribution of loadings for cell-type-specific transcripts in adipose tissue. **B.** Distribution of loadings for cell-type-specific transcripts in pancreatic islets (green). **C.** Null distributions for the mean loading of randomly selected transcripts 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.

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

#### 268 Targeting gene signatures

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

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

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

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

282 • **Supplemental Table** results from CMAP

## 283 Discussion

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

## 289 Data Availability

290 Here we tell people where to find the data

## 291 Acknowledgements

292 Here we thank people

293 **Supplemental Figures**

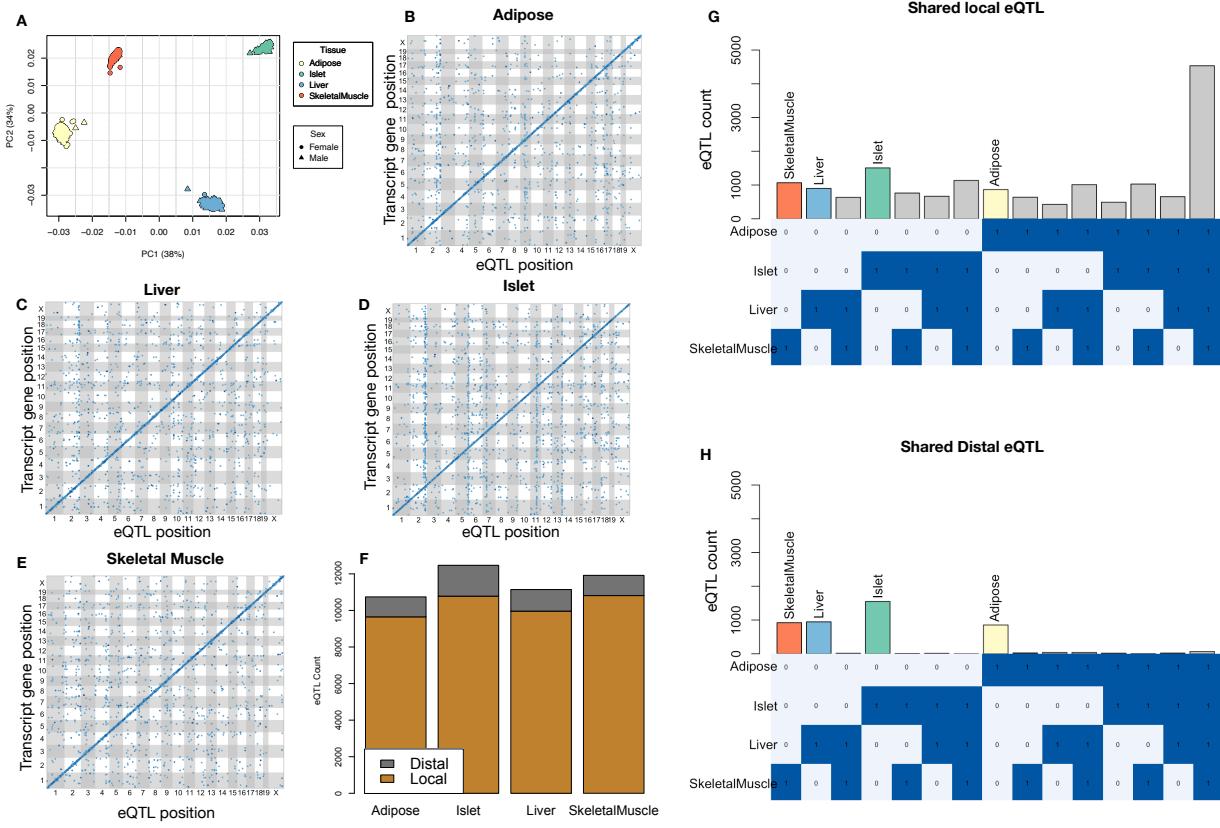


Figure 9: Overview of eQTL analysis in DO mice. **A.** RNA seq samples from the four different tissues clustered by tissue. **B.-E.** eQTL maps are shown for each tissue. The *x*-axis shows the position of the mapped eQTL, and the *y*-axis shows the physical position of the gene encoding each mapped 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.

## KEGG pathway enrichments by GSEA

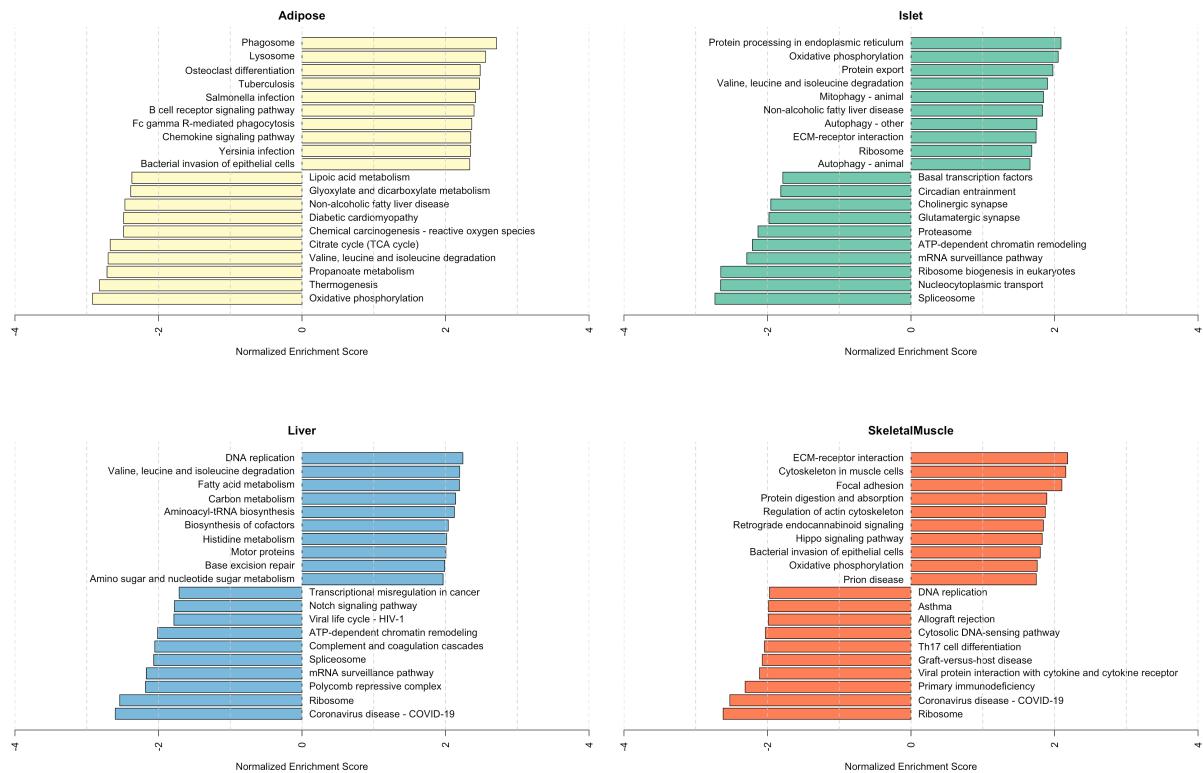


Figure 10: Bar plots showing normalized enrichment scores (NES) for KEGG pathways as determined by fast gene score enrichment analysis (fgsea). Only the top 10 positive and top 10 negative scores are shown. Colors indicate tissue. The name beside each bar shows the name of each enriched KEGG pathway.

## References

- 294 [1] M. T. Maurano, R. Humbert, E. Rynes, R. E. Thurman, E. Haugen, H. Wang, A. P. Reynolds,  
 295 R. Sandstrom, H. Qu, J. Brody, A. Shafer, F. Neri, K. Lee, T. Kutyavin, S. Stehling-Sun, A. K.  
 296 Johnson, T. K. Canfield, E. Giste, M. Diegel, D. Bates, R. S. Hansen, S. Neph, P. J. Sabo, S. Heimfeld,  
 297 A. Raubitschek, S. Ziegler, C. Cotsapas, N. Sotoodehnia, I. Glass, S. R. Sunyaev, R. Kaul, and J. A.  
 298 Stamatoyannopoulos. Systematic localization of common disease-associated variation in regulatory DNA.  
 299 *Science*, 337(6099):1190–1195, Sep 2012.
- 300  
 301 [2] K. K. Farh, A. Marson, J. Zhu, M. Kleinewietfeld, W. J. Housley, S. Beik, N. Shores, H. Whitton, R. J.  
 302 Ryan, A. A. Shishkin, M. Hatan, M. J. Carrasco-Alfonso, D. Mayer, C. J. Luckey, N. A. Patsopoulos,  
 303 P. L. De Jager, V. K. Kuchroo, C. B. Epstein, M. J. Daly, D. A. Hafler, and B. E. Bernstein. Genetic

## Top GO term enrichments by GSEA

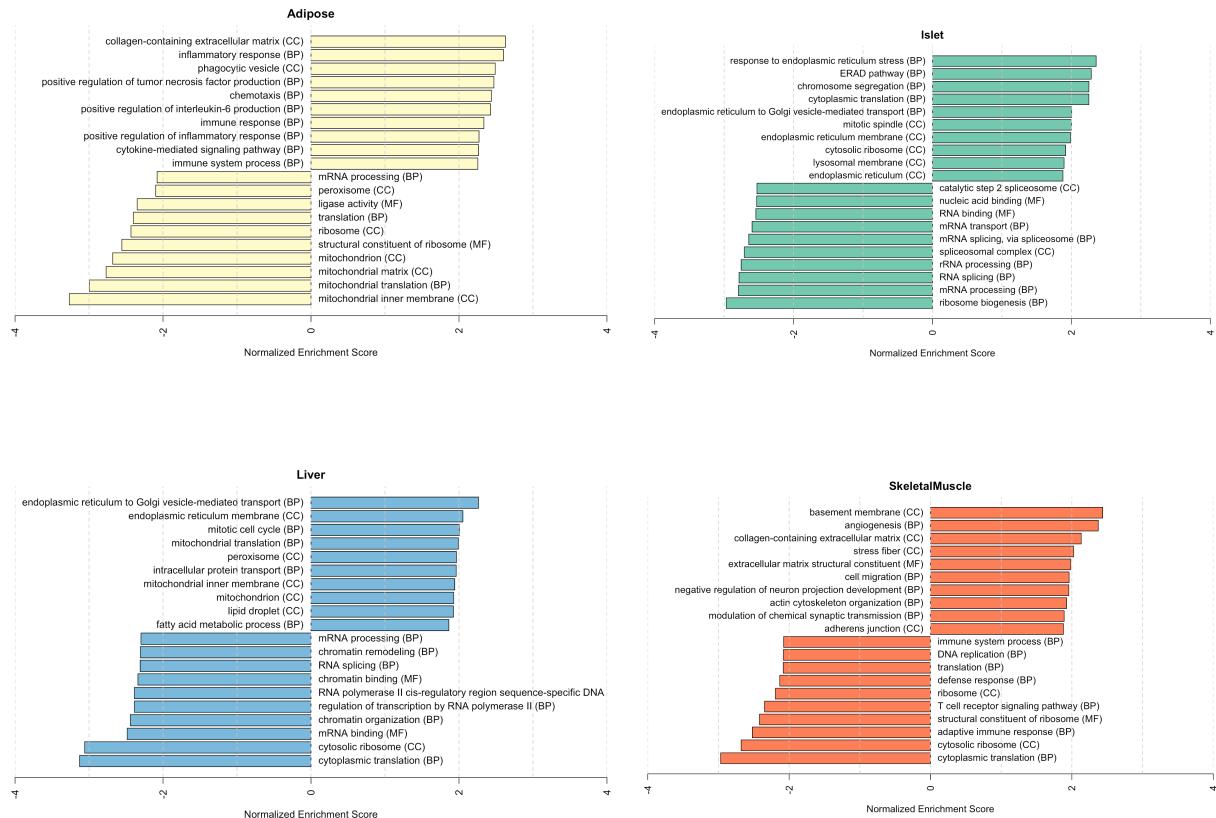


Figure 11: Bar plots showing normalized enrichment scores (NES) for GO terms as determined by fast gene score enrichment analysis (fgsea). Only the top 10 positive and top 10 negative scores are shown. Colors indicate tissue. The name beside each bar shows the name of each enriched GO term. The letters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

- 304 and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, 518(7539):337–343, Feb  
 305 2015.
- 306 [3] E. Pennisi. The Biology of Genomes. Disease risk links to gene regulation. *Science*, 332(6033):1031, May  
 307 2011.
- 308 [4] L. A. Hindorff, P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins, and T. A. Manolio.  
 309 Potential etiologic and functional implications of genome-wide association loci for human diseases and  
 310 traits. *Proc Natl Acad Sci*, 106(23):9362–9367, Jun 2009.
- 311 [5] J. K. Pickrell. Joint analysis of functional genomic data and genome-wide association studies of 18  
 312 human traits. *Am J Hum Genet*, 94(4):559–573, Apr 2014.

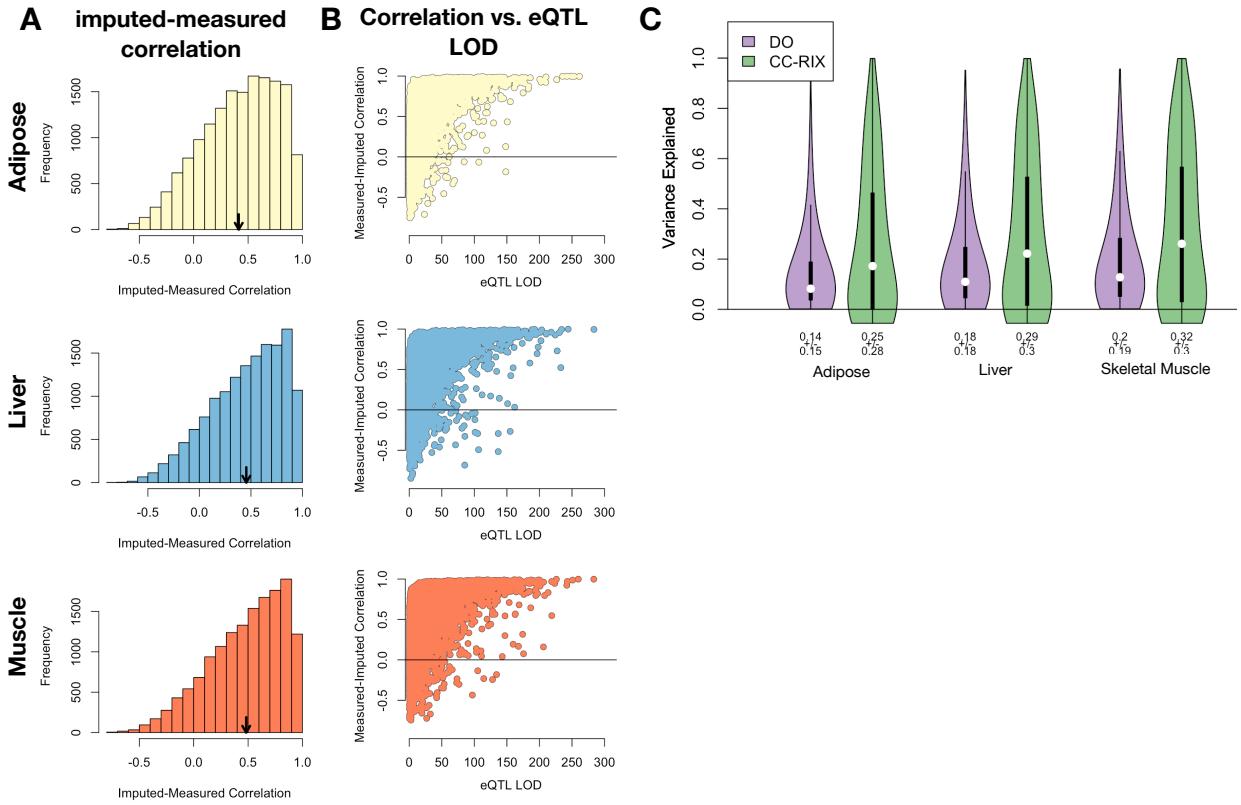


Figure 12: Validation of transcript imputation in the CC-RIX. **A.** Distributions of correlations between imputed and measured transcripts in the CC-RIX. The mean of each distribution is shown by the red line. All distributions were skewed toward positive correlations and had positive means near a Pearson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

- 313 [6] D. Welter, J. MacArthur, J. Morales, T. Burdett, P. Hall, H. Junkins, A. Klemm, P. Flicek, T. Manolio,  
314 L. Hindorff, and H. Parkinson. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations.  
315 *Nucleic Acids Res*, 42(Database issue):D1001–1006, Jan 2014.
- 316 [7] Y. I. Li, B. van de Geijn, A. Raj, D. A. Knowles, A. A. Petti, D. Golan, Y. Gilad, and J. K. Pritchard.  
317 RNA splicing is a primary link between genetic variation and disease. *Science*, 352(6285):600–604, Apr  
318 2016.
- 319 [8] E. R. Gamazon, H. E. Wheeler, K. P. Shah, S. V. Mozaffari, K. Aquino-Michaels, R. J. Carroll, A. E.  
320 Eyler, J. C. Denny, D. L. Nicolae, N. J. Cox, and H. K. Im. A gene-based association method for  
321 mapping traits using reference transcriptome data. *Nat Genet*, 47(9):1091–1098, Sep 2015.
- 322 [9] Z. Zhu, F. Zhang, H. Hu, A. Bakshi, M. R. Robinson, J. E. Powell, G. W. Montgomery, M. E. Goddard,  
323 N. R. Wray, P. M. Visscher, and J. Yang. Integration of summary data from GWAS and eQTL studies

- 324 predicts complex trait gene targets. *Nat Genet*, 48(5):481–487, May 2016.
- 325 [10] A. Gusev, A. Ko, H. Shi, G. Bhatia, W. Chung, B. W. Penninx, R. Jansen, E. J. de Geus, D. I. Boomsma,  
326 F. A. Wright, P. F. Sullivan, E. Nikkola, M. Alvarez, M. Civelek, A. J. Lusis, T. ki, E. Raitoharju,  
327 M. nen, I. ä, O. T. Raitakari, J. Kuusisto, M. Laakso, A. L. Price, P. Pajukanta, and B. Pasaniuc.  
328 Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet*, 48(3):245–252,  
329 Mar 2016.
- 330 [11] B. D. Umans, A. Battle, and Y. Gilad. Where Are the Disease-Associated eQTLs? *Trends Genet*,  
331 37(2):109–124, Feb 2021.
- 332 [12] N. J. Connally, S. Nazeen, D. Lee, H. Shi, J. Stamatoyannopoulos, S. Chun, C. Cotsapas, C. A. Cassa,  
333 and S. R. Sunyaev. The missing link between genetic association and regulatory function. *Elife*, 11, Dec  
334 2022.
- 335 [13] H. Mostafavi, J. P. Spence, S. Naqvi, and J. K. Pritchard. Systematic differences in discovery of genetic  
336 effects on gene expression and complex traits. *Nat Genet*, 55(11):1866–1875, Nov 2023.
- 337 [14] D. W. Yao, L. J. O’Connor, A. L. Price, and A. Gusev. Quantifying genetic effects on disease mediated  
338 by assayed gene expression levels. *Nat Genet*, 52(6):626–633, Jun 2020.
- 339 [15] X. Liu, Y. I. Li, and J. K. Pritchard. Trans Effects on Gene Expression Can Drive Omnipotent Inheritance.  
340 *Cell*, 177(4):1022–1034, May 2019.
- 341 [16] G. A. Churchill, D. M. Gatti, S. C. Munger, and K. L. Svenson. The Diversity Outbred mouse population.  
342 *Mamm Genome*, 23(9-10):713–718, Oct 2012.
- 343 [17] M. P. Keller, D. M. Gatti, K. L. Schueler, M. E. Rabaglia, D. S. Stapleton, P. Simecek, M. Vincent,  
344 S. Allen, A. T. Broman, R. Bacher, C. Kendziorski, K. W. Broman, B. S. Yandell, G. A. Churchill, and  
345 A. D. Attie. Genetic Drivers of Pancreatic Islet Function. *Genetics*, 209(1):335–356, May 2018.
- 346 [18] S. M. Clee and A. D. Attie. The genetic landscape of type 2 diabetes in mice. *Endocr Rev*, 28(1):48–83,  
347 Feb 2007.
- 348 [19] C. B. Newgard. Interplay between lipids and branched-chain amino acids in development of insulin  
349 resistance. *Cell Metab*, 15(5):606–614, May 2012.
- 350 [20] D. D. Sears, G. Hsiao, A. Hsiao, J. G. Yu, C. H. Courtney, J. M. Ofrecio, J. Chapman, and S. Subramaniam.  
351 Mechanisms of human insulin resistance and thiazolidinedione-mediated insulin sensitization. *Proc Natl  
352 Acad Sci U S A*, 106(44):18745–18750, Nov 2009.

- 353 [21] G. Hsiao, J. Chapman, J. M. Ofrecio, J. Wilkes, J. L. Resnik, D. Thapar, S. Subramaniam, and D. D.  
354 Sears. modulation of insulin sensitivity and metabolic pathways in obese rats. *Am J Physiol Endocrinol*  
355 *Metab*, 300(1):E164–174, Jan 2011.
- 356 [22] D. E. Lackey, C. J. Lynch, K. C. Olson, R. Mostaedi, M. Ali, W. H. Smith, F. Karpe, S. Humphreys,  
357 D. H. Bedinger, T. N. Dunn, A. P. Thomas, P. J. Oort, D. A. Kieffer, R. Amin, A. Bettaieb, F. G.  
358 Haj, P. Permana, T. G. Anthony, and S. H. Adams. Regulation of adipose branched-chain amino acid  
359 catabolism enzyme expression and cross-adipose amino acid flux in human obesity. *Am J Physiol*  
360 *Endocrinol Metab*, 304(11):E1175–1187, Jun 2013.
- 361 [23] R. Stienstra, C. Duval, M. ller, and S. Kersten. PPARs, Obesity, and Inflammation. *PPAR Res*,  
362 2007:95974, 2007.
- 363 [24] O. Gavrilova, M. Haluzik, K. Matsusue, J. J. Cutson, L. Johnson, K. R. Dietz, C. J. Nicol, C. Vinson,  
364 F. J. Gonzalez, and M. L. Reitman. Liver peroxisome proliferator-activated receptor gamma contributes  
365 to hepatic steatosis, triglyceride clearance, and regulation of body fat mass. *J Biol Chem*, 278(36):34268–  
366 34276, Sep 2003.
- 367 [25] K. Matsusue, M. Haluzik, G. Lambert, S. H. Yim, O. Gavrilova, J. M. Ward, B. Brewer, M. L. Reitman,  
368 and F. J. Gonzalez. Liver-specific disruption of PPARgamma in leptin-deficient mice improves fatty  
369 liver but aggravates diabetic phenotypes. *J Clin Invest*, 111(5):737–747, Mar 2003.
- 370 [26] D. Patsouris, J. K. Reddy, M. ller, and S. Kersten. Peroxisome proliferator-activated receptor alpha  
371 mediates the effects of high-fat diet on hepatic gene expression. *Endocrinology*, 147(3):1508–1516, Mar  
372 2006.
- 373 [27] S. E. Schadinger, N. L. Bucher, B. M. Schreiber, and S. R. Farmer. PPARgamma2 regulates lipogenesis  
374 and lipid accumulation in steatotic hepatocytes. *Am J Physiol Endocrinol Metab*, 288(6):E1195–1205,  
375 Jun 2005.
- 376 [28] W. Motomura, M. Inoue, T. Ohtake, N. Takahashi, M. Nagamine, S. Tanno, Y. Kohgo, and T. Okumura.  
377 Up-regulation of ADRP in fatty liver in human and liver steatosis in mice fed with high fat diet. *Biochem*  
378 *Biophys Res Commun*, 340(4):1111–1118, Feb 2006.