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

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

## Abstract

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

## Introduction

Evidence from genome-wide association studies (GWAS) suggests that most heritable variation in complex traits is mediated through regulation of gene expression. The majority of trait-associated variants lie in gene regulatory regions , suggesting a relatively simple causal model in which a variant alters the homeostatic expression level of a nearby (local) gene which, in turn, alters a trait. Statistical methods such as transcriptome-wide association studies (TWAS) and summary data-based Mendelian randomization (SMR) have used this idea to identify genes associated with multiple disease traits . However, despite the great promise of these methods, explaining trait effects with local gene regulation has been more difficult than initially assumed . Although trait-associated variants typically lie in non-coding, regulatory regions, these variants often have no detectable effects on gene expression and tend not to co-localize with expression quantitative trait loci (eQTLs) . These observations suggest that the relationship among genetic variants, gene expression, and organism-level traits is more complex than the simple, local model.

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

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

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

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

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

## Results

### Genetic variation contributed to wide phenotypic variation

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

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

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

### Distal Heritability Correlated with Phenotype Relevance

To comprehensively assess the genetic control of gene expression in metabolic disease we measured overall gene expression via bulk RNA-Seq in adipose, islet, liver, and skeletal muscle in the DO cohort (Supp. Fig. ). We performed eQTL analysis using R/qtl2 (Methods) and identified both local and distal eQTLs for transcripts in each of the four tissues (Supp. Fig. B-E). Significant local eQTLs far outnumbered distal eQTLs (Supp. Fig. F) and tended to be shared across tissues (Supp. Fig. G) whereas the few significant distal eQTLs we identified tended to be tissue-specific (Supp. Fig. H)

We estimated the heritability of each transcript in terms of local and all non-local (distal) genetic factors (Methods). Overall, local and distal genetic factors contributed approximately equally to transcript abundance. In all tissues, both local and distal factors explained between 8 and 17% of the variance in the median transcript (Fig, A).

The equal contribution of local and distal genetic variants to the heritability of transcript abundance contrasts with findings in humans in which local variants have been found to explain 20-30% of total heritability, while distal variants explain the remaining 70-80%. This discrepancy may arise due to the high degree of linkage disequilibrium in the DO mice compared to human populations. Each genetic marker in the mice captures information from a larger genomic region than each genetic marker in human populations, and thus may capture more local regulatory variants than SNPs capture in humans. It has been found that transcripts with muliple local eQTL have higher local heritability than transcripts with single local eQTL. Because of the high diversity in the DO and the high rates of linkage disequilibrium, it is possible that there are more local variants regulating transcription creating a proportionally larger effect of local regulation.

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

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

The above univariate analyses establish the importance of distal heritability for trait-relevant transcripts. However, the number of transcripts dramatically exceeds the number of phenotypes. Thus, we expect the heritable, trait-relevant transcripts to be highly correlated and organized according to coherent, biological processes representing the mediating endophenotypes driving clinical trait variation. To identify these endophenotypes in a theoretically principled way, we developed a novel dimension-reduction technique, high-dimension mediation analysis (HDMA), that uses the theory of causal graphical models to identify a transcriptomic signature that is simultaneously 1) highly heritable, 2) strongly correlated to the measured phenotypes, and 3) conforms to the causal mediation hypothesis (Fig. ). In HDMA, we first use a linear mapping called kernelization to dimension-reduce the genome, transcriptome, and phenome to kernel matrices , and respectively, which each have the dimensions by where is the number of individuals (Methods). These kernel matrices describe the relationships among the individual mice in genome space, transcriptome space, and phenome space and and ensure that these three omic spaces have the same dimensions, and thus the same weight in the analysis. If not dimension-reduced, the transcriptome would outweigh the phenome in the model. We then projected these -dimensional kernel matrices onto one-dimensional scores–a composite genome score (), a composite transcriptome score (), and a composite phenome score ()–and used the univariate theory of mediation to constrain these projections to satisfy the hypotheses of perfect mediation, namely that upon controlling for the transcriptomic score, the genome score is uncorrelated to the phenome score. A complete mathematical derivation and implementation details for HDMA are available in the Methods.

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

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

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

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

We interpreted large loadings onto transcripts as indicating strong mediation of the effect of genetics on MDI. Large positive loadings indicate that higher expression was associated with a higher MDI (i.e. higher risk of obesity and metabolic disease on the HFHS diet) (Fig. C). Conversely, large negative loadings indicate that high expression of these transcripts was associated with a lower MDI (i.e. lower risk of obesity and metabolic disease on the HFHS diet) (Fig. C). We used gene set enrichment analysis (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 KEGG pathway enrichments pointed to an axis of inflammation and metabolism (Figs. and ). GO terms and KEGG pathways associated with inflammation were positively associated with MDI, indicating that increased expression in inflammatory pathways was associated with a higher burden of disease. It is well established that adipose tissue in obese individuals is inflamed and infiltrated by macrophages , and the results here suggest that this may be a dominant heritable component of metabolic disease.

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

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

Looking at the 10 most positively and negatively loaded transcripts from each tissue, it is apparent that transcripts in the adipose tissue had the largest loadings, both positive and negative (Fig. A bar plot). This suggests that much of the effect of genetics on body weight and insulin reisistance is mediated through gene expression in adipose tissue. This finding does not speak to the relative importance of tissues not included in this study, such as brain, in which transcriptional variation may mediate a large portion of the genetic effect on obesity. The strongest loadings in liver and pancreas were comparable, and those in skeletal muscle were the weakest (Fig. A), suggesting that less of the genetic effects were mediated through transcription in skeletal muscle. As expected, heritability analysis showed that transcripts with the largest loadings had higher distal heritability than local heritability (Fig. A heat map and box plot). We also performed TWAS in this population by imputing transcript levels for each gene based on local genotype only and correlating the imputed transcript levels with each trait. In contrast to HDMA, the TWAS procedure tended to nominate transcripts with lower loadings (Fig. B), higher local heritability and lower distal heritability. Finally, we focused on transcripts with the highest local heritability in each tissue (Fig. C). This procedure selected transcripts with low loadings on average, consistent with our findings above (Fig. B).

We performed a literature search for the genes in each of these groups along with the terms “diabetes”, “obesity”, and the name of the expressing tissue to determine whether any of these genes had previous associations with metabolic disease in the literature (Methods). Multiple genes in each group had been previously associated with obesity and diabetes (Fig. bolded gene names). Genes with high loadings were most highly enriched for previous literature support. They were 2.1 times more likely than TWAS hits and 3.8 times more likely than genes with high local heritability to be previously associated with obesity or diabetes.

### Tissue-specific transriptional programs were associated with metabolic traits

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

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

To test whether the transcript loadings identified in the DO could be translated to another population, we tested whether they could predict metabolic phenotypes in an independent population of CC-RIX mice, which were F1 mice derived from multiple pairings of Collaborative Cross (CC) strains (Fig. ) (Methods). We tested two questions. First, we asked whether the loadings identified in the DO mice were relevant to the relationship between the transcriptome and the phenome in the CC-RIX. We predicted body weight (a surrogate for MDI) in each CC-RIX individual using measured gene expression in each tissue and the transcript loadings identified in the DO (Methods). The predicted body weight and acutal body weight were highly correlated (Fig. B left column). The best prediction was achieved for adipose tissue, which supports the observation in the DO that adipose expression was the strongest mediator of the genetic effect on MDI. This result also confirms the validity and translatability of the transcript loadings and their relationship to metabolic disease.

The second question related to the source of the relevant variation in gene expression. If local regulation was the predominant factor influencing trait-relevant gene expression, we should be able to predict phenotype in the CC-RIX using transcripts imputed from local genotype (Fig. A). The DO and the CC-RIX were derived from the same eight founder strains and so carry the same alleles throughout the genome. We imputed gene expression in the CC-RIX using local genotype and were able to estimate variation in gene transcription robustly (Supp. Fig. ). However, these imputed values failed to predict body weight in the CC-RIX when weighted with the loadings from HDMA. (Fig. B right column). This result suggests that local regulation of gene expression is not the primary factor driving heritability of complex traits. It is also consistent with our findings in the DO population that distal heritability was a major driver of trait-relevant gene expression and that high-loading transcripts had comparatively high distal and low local heritability.

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

The interpretation of global genetic influences on gene expression and phenotype is potentially more challenging than the interpretation and translation of local genetic influences, as genetic effects cannot be localized to individual gene variants or transcripts. However, there are global patterns across the loadings that can inform mechanism. For example, heritable variation in cell type composition can be inferred from transcript loadings. We observed above that immune activation in the adipose tissue was a highly enriched process correlating with obesity in the DO population. In humans, it has been extensively observed that macrophage infiltration in adipose tissue is a marker of obesity and metabolic disease . To determine whether the immune activation reflected a heritable change in cell composition in adipose tissue in DO mice, we compared loadings of cell-type specific genes in adipose tissue (Methods). The mean loading of macrophage-specific genes was significantly greater than 0 (Holm-adjusted two-sided empirical ) (Fig. A), indicating that obese mice were genetically predisposed to have high levels of macrophage infiltration in adipose tissue in response to the HFHS diet. Loadings for marker genes for other cell types were not statistically different from zero (Adipocytes: , Progenitors: , Leukocytes: ; all Holm-adjusted two-sided empirical p), indicating that changes in the abundance of those cell types was not a mediator of MDI.

We also compared loadings of cell-type specific transcripts in islet (Methods). The mean loadings for alpha-cell specific transcripts were significantly greater than 0 (), while the mean loadings for delta- (Holm-adjusted two-sided empirial ) and endothelial-cell (Holm-adjusted two-sided empirical ) specific genes were significantly less than 0 (Fig. B). These results suggest that mice with higher MDI inherited an altered cell composition that predisposed them to metabolic disease, or that these compositional changes were induced by the HFHS diet in a heritable way. In either case, these results support the hypothesis that alterations in islet composition drive variation in MDI. Notably, the mean loading for pancreatic beta cell marker transcripts was not significantly different from zero (Holm-adjusted two-sided empirical ). We stress that this is not necessarily reflective of the function of the beta cells in the obese mice, but rather suggests that any variation in the number of beta cells in these mice was unrelated to obesity and insulin resistance, the major contributors to MDI. This is further consistent with the islet composition traits having small loadings in the phenome score (Fig. ).

### Heritable transcriptomic signatures translated to human disease

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

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

### Existing therapies are predicted to target mediator gene signatures

Another application of the transcript loading landscape is in ranking potential drug candidates for the treatment of metabolic disease. Although high-loading transcripts may be good candidates for understanding specific biology related to obesity, the transcriptome overall is highly interconnected and redundant. The ConnectivityMap (CMAP) database developed by the Broad Institute allows querying thousands of compounds that reverse or enhance the extreme ends of transcriptomic signatures in multiple different cell types. By identifying drugs that reverse pathogenic transcriptomic signatures, we can potentially identify compounds that have favorable effects on gene expression. To test this hypothesis, we queried the CMAP database through the CLUE online query tool (, version 1.1.1.43) (Methods). We identified top anti-correlated hits across all cell types (Supp. Figs and ). To get more tissue-specific results, we also looked at top results in cell types that most closely resembled our tissues. We looked at results in adipocytes (ASC) as well as pancreatic tumor cells (YAPC) regardless of value (Supp. Figs and ).

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

## Discussion

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

Our result that distal regulation accounted for most trait-related gene expression differences is consistent with a complex model of genetic trait determination. It has frequently been assumed that gene regulation in is the primary driver of genetically associated trait variation, but attempts to use local gene regulation to explain phenotypic variation have had limited success . In recent years, evidence has mounted that distal gene regulation may be an important mediator of trait heritability . It has been observed that transcripts with high local heritability explain less expression-mediated disease heritability than those with low local heritability . Consistent with this observation, genes located near GWAS hits tend to be complexly regulated . They also tend to be enriched with functional annotations, in contrast to genes with simple local regulation, which tend to be depleted of functional annotations suggesting they are less likely to be directly involved in disease traits . These observations are consistent with principles of robustness in complex systems in which simple regulation of important elements leads to fragility of the system . Our results are consistent, instead, with a more complex picture where genes whose expression can drive trait variation are buffered from local genetic variation but are extensively influenced indirectly by genetic variation in the regulatory networks converging on those genes.

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

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

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

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

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

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

## Data and Code Availability

Genotypes, phenotypes, and pancreatic islet gene expression data were previously published . Gene expression for the other tissues can be found at the Gene Expression Omnibus with the following accession numbers: DO adipose tissue - GSE266549; DO liver tissue - GSE266569; DO skeletal muscle - GSE266567. Expression data with calculated eQTLs are available at Figshare DOI: 10.6084/m9.figshare.27066979

10.6084/m9.figshare.27066979.v1

: Gene expression can be found at the Gene Expression Omnibus with the following accession numbers: CC-RIX adipose tissue - GSE237737; CC-RIX liver tissue - GSE237743; CC-RIX skeletal muscle - GSE237747. Count matrices and phenotype data can be found at Figshare DOI: 10.6084/m9.figshare.27066979

: All code used to run the analyses reported here are available at Figshare: DOI: 10.6084/m9.figshare.27066979

## Acknowledgements

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

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

## Supplemental Figures