Transcripts with high distal heritability mediate genetic effects on complex traits

By lots of people

April 12, 2024

## Abstract

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

## Introduction

In the quest to understand genetic contributions to complex traits, gene expression is an important bridge between genotype and phenotype. The majority of variants identified in GWAS are in regulatory regions of the genome [cite], suggesting that they influence clinical phenotypes through regulation of gene expression. Consistent with this idea, powerful insights into disease mechanism can be made by linking genetic variants affecting gene expression (expression quantitative trait loci - eQTLs) to phenotypes. In particular, mediation analysis has been used to identify transcripts that mediate the effect of genetic variants on phenotypes. In mice… (mice: @pmid29567659,@pmid31465442,@pmid35672473) (bmediatR: @pmid35533209) In humans… (human - @pmid25533967, @pmid24232670)

Thus far, the primary focus of expression mediated traits has been on local genetic variation; that is genetic variation that influences the transcription of local genes, thereby causing variation in traits. However, there is evidence that the bulk of disease heritability is mediated by the distal component of gene expression, rather than the local component [@32424349]. Yao et al. [cite] observed that genes with low local heritability explain more expression-mediated disease heritability than genes with high local heritability. We have observed a similar pattern in mice, which we describe here. Thus, identifying heritable components of complex traits that are mediated through distally regulated variation in gene expression may provide important insights into mechanisms regulating complex traits.

Identification of distal factors influencing gene expression and traits is challenging, as the multiple testing corrections are much more severe for distal effects [@pmid24013639]. However, systems approaches that consider the entire transcriptome simultaneously and avoid univariate testing provide promising avenues for identification of broad transcriptomic patterns influencing complex traits that provide both biological insight and targets for therapeutics. Here we propose high-dimenaional mediation (HDM) as one such systems approach for identification of the heritable portion of the transcriptome that mediates the effect of the genome on phenome. HDM uses a regularized and generalized canonical correlation analysis (RGCCA) [cite], which is an extension of canonical correlation analysis (CCA) that allows for more than two data sets with an arbitrary relationship among them. Thus, we can identify linear combinations of the genome, transcriptome, and phenome, that describe the mediation of the genetic effects on the phenome through the transcriptome. Because of the central dogma of molecular biology, information flow is directed out of the genome, and not back into it. Thus, the otherwise undirected relationships between genome, transcriptome, and phenome can be inferred as a causal mediation by the transcriptome of the effects of the genome on the phenome.

Here we apply HDM

## Results

### Genetic variation contributes to wide phenotypic variation

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

Although the environment was consistent across animals, the genetic diversity present in this population resulted in widely varying distributions across physiological measurements (Fig. 1. and Fig. S1). For example, body weights of adult individuals varied from less than the average adult B6 body weight to several times the body weight of a B6 adult in both sexes (Fig. 1A). Body weight was strongly positively correlated with food consumption (Fig. 1B ) and fasting blood glucose (FBG) (Fig. 1D, ) suggesting a link between behavioral factors and metabolic disease. FBG levels varied widely across individuals (Fig. 1C), although few of the animals had FBG levels that would indicate pre-diabetes (XXX /%), or diabetes (XXX /%) according to previously developed cutoffs (pre-diabetes: FBG mg/dL, diabetes: FBG mg/dL) [cite]. However, the heritability of this trait and others (Fig. 1E) indicates that background genetics contribute substantially to correlates of metabolic disease in this population.

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

### 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 XXX). Significant local eQTLs far outnumbered distal eQTLs (Supp. Fig. XXX) and tended to be shared across tissues (Supp. Fig. XXX) whereas the few significant distal eQTL we identified tended to be tissue-specific (Supp. Fig. XXX)

The low number of significant distal eQTLs is driven in part by multiple testing corrections: Local eQTLs only require a single statitistical test to associate each transcript with the genotype at the nearest marker, whereas distal eQTLs require testing all markers except for the local marker. Additionally, distal effects can be spread out across multiple locations making each one more difficult to detect statistically. 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 XXX).

Local heritability of transcripts was negatively correlated with its trait relevance, defined as the maximum correlation of the transcript and all traits. This suggests that the more local genotype influenced transcript abundance, the less effect variation in transcript abundance was related to the measured traits. Conversely, distal heritability of transcripts was positively correlated with trait relevance. That is, transcripts that were more highly correlated with the measured traits tended to be distally heritable rather than locally heritable. That trait-correlated transcripts have low local heritability is consistent with previous observations that low-heritability transcripts explain more expression-mediated disease heritability than high-heritability transcripts [cite]. However, the positive relationship between trait correlation and distal heritability suggests that there are alternative mechanisms through which genetic regulation of transcripts may influence traits.

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

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

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

Standard CCA is prone to over-fitting because in any two large matrices it can be trivial to identify highly correlated composite vectors. To assess whether RGCCA was similarly prone to over-fitting in a high-dimensional space, we performed permutation testing. We permuted the individual labels on the transcriptome kernel matrix 1000 times and recalculated the path coefficient, which is the partial correlation of and multiplied by the partial correlation of and . This represents the path from to that is mediated through . The null distribution of the path coefficient is shown in Fig. XXXB, 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. XXXC illustrates this observation in more detail. Although we identified high correlations between and , and modest correlations between and in the null data (Fig XXXC), these two values could not be maximized simultaneously. The red dot shows that in the real data both the - correlation and the - correlation could be maximized simultaneously suggesting that that path from genotype to phenotype through transcriptome is highly non-trivial and identifiable in this case. These results suggest that these composite vectors represent genetically determined variation in phenotype that is mediated through genetically determined variation in transcription.

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

The loadings of each measured trait onto indicate how much each contributed to . Final body weight contributed the most to (Fig. XXXA), followed by homeostatic insulin resistance (HOMA\_IR) and fasting plasma insulin levels (Insulin\_Fasting). The high loadings of these traits indicate that these are the primary traits mediated by . Traits contributing the least to were measures of cholesterol and pancreas composition. The smaller contributions of these traits indicate a weaker relationship with the heritable transcriptomic signature described by . Thus, when we interpret the transcriptomic signature identified by HDM, we are explaining primarily transcriptional mediation of body weight and insulin resistance, as opposed to cholesterol measurements. Because higher composite trait scores have large, positive contributions from body weight and insulin resistance, larger positive scores for individual mice indicate greater metabolic disease (Fig. XXXB)

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

Transcripts that most strongly correlated with were the best mediators of effect of genetics on . Large positive loadings indicate that inheriting higher expression was associated with a higher (higher risk of obesity and metabolic disease on the high-fat diet) (Fig. XXXC). Conversely, large negative loadings indicate that inheriting lower expression of these transcripts was associated with a lower (lower risk of obesity and metabolic disease on the high-fat diet) (Fig. XXXC). Functional enrichments for the most highly correlated and anti-correlated transcripts are shown in Supp. Fig. XXX and represent known biology of obesity and diabetes. In adipose tissue, for example, the transcripts most strongly correlated with were enriched for immune system signaling and cell motility. It is well established that adipose tissue in obese individuals is highly inflamed [cite] and infiltrated by macrophages [cite]. The transcripts most strongly negatively correlated with were enriched for metabolism of the branched-chain amino acids (BCAA), valine, leuceine, and isoleucine. BCAA are used in adipose tissue in lipogenesis, and inhibiting BCAA catabolism inhibits adipogenesis [26571352]. BCAA levels are also related to insulin resistance and are elevated in insulin-resistant obese individuals relative to weight-matched non-insulin reisistand individuals [23512805]. In the DO mice studied here, inheriting reduced expression of genes involved in BCAA catabolism was associated with reduced body weight and insulin resistance.

Transcripts in the adipose tissue had the largest loadings, both positive and negative, of all tissues, suggesting that much of the effect of genetics on body weight and insulin reisistance is mediated through gene expression in adipose tissue (Fig. XXX). The loadings in liver and pancreas were comparable, and those in skeletal muscle were the weakest (Fig. XXX), suggesting that less of the genetic effects were mediated through transcription in skeletal muscle. Across all tissues, trahscripts with the largest loadings tended to have relatively high distal heritability compared with local heritability (Fig. XXXA). 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. XXXB, Methods). TWAS-nominated transcripts also had relatively weak loadings and high local heritabilty (Fig. XXXC). They were half as likely as transcripts with the highest loadings to be associated with diabetes and obesity in the literature (Fig. XXXC).

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

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

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

We used either measured transcription or transcription imputed from local eQTLs calculated in the DO population to predict body weight in the CC-RIX (Methods) (Fig. XXXB). Gene expression imputed from local eQTLs was completely uncorrelated with body weight in the CC-RIX regardless of tissue. In contrast, body weight predicterd from measured gene expression was highly correlated with meaasured body weight in the CC-RIX. The correlation was highest for expression measured in adipose tissue (, ), which is consistent with our previous observation that adipose transcription is the strongest mediator of the genetic effects on metabolic traits. Further consistent with our previous results, gene expression measured in skeletal muscle was the weakest predictor, although, it still produced a prediction that was significantly correlated with the measured value (, ).

Taken together, these results support the hypothesis that distal, rather than local genetic factors are primarily driving complex-trait related variation in gene expression.

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

Functional enrichments of high-loading genes in the adipose tissue, suggested that the obese mice in the population had a genetic predisposition toward elevated macrophage infiltration into the adipose tissue. We investigated this further bioinformatically by comparing the loadings of cell-type-specific transcripts (Methods). In adipose tissue, the mean loading of macrophage-specific genes was substantially above 0 (Fig. XXX), indicating that obese mice were genetically predisposed to have high levels of macrophage infiltration in adipose tissue in response to the high-fat, high-sugar diet.

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

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

Biological interpretation of alpha, endothelial, delta cells??

### Distally heritable transcriptomic signatures translate to human disease

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

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

### Targeting gene signatures

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

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

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

* **Supplemental Table** results from CMAP

## Discussion

* distal heritability correlates with phenotype relevance

## Data Availability

Here we tell people where to find the data

## Acknowledgements

Here we thank people