Detailed Methods for High Dimensional Mediation for Massively Polygenic Traits

J. Matthew Mahoney and Anna L Tyler

August 30, 2024

Diversity Outbred Mice

A population of 500 diversity outbred mice (split evenly between male and female) from generates 18, 19, and 21, was placed on a high-fat (44.6% kcal fat), high-sugar (34% carbohydrate), adequate protein (17.3 % protein) diet from Envigo Teklad (catalog number TD.08811) starting at four weeks of age as described previously[1]. Each individual was assessed longitudinally for multiple metabolic measures including fasting glucose levels, glucose tolerance, insulin levels, body weight, and blood lipid levels.

Trait measurements

also described in[1] briefly describe: oral glucose tolerance tests

Genotyping

The mice were also genotyped using the Mouse Universal Genotyping Array (GigaMUGA)[1] .

briefly describe here.

RNA Sequencing

At the end of the experiment, we used RNASeq to measure gene expression in 384 mice in four tissues relevant to metabolic disease: adipose tissue, pancreatic islets, liver, and skeletal muscle.

also described in[1] describe islet picking, and how other organs were collected. mention details about GigaMUGA

Trait selection

We filtered the measured traits in this study to a set of relatively non-redundant measures that were well-represented in the population (having at least XXX individuals), and spanned multiple aspects of metabolic disease. A complete description of trait filtering can be found in File XXX (1b.Trait_Selection.Rmd).

We took two approaches for traits with multiple redundant measurements, for example logintudinal body weights. In the case of longitudinal measurements, we used the final measurement, as this was the closest physiological measurement to the measurement of gene expression, which was done at the end of the experiment. The labels for these traits are have the word "Final" appended to their name. For traits with multiple highly related measurements, such as cholesterol, we used the first principal component of the group of measurements. For example, we used the first principal component of all LDL measurements as the measurement of LDL. For each set of traits, we ensured the first principal component had the correct sign by correlating it with the average of the traits. For correlation coefficients (R) less than 0, we multiplied the principal component by -1. The labels for these traits have the term "PC1" appended to their name.

Processing of RNA sequencing data

We used the Expectation-Maximization algorithm for Allele Specific Expression (EMASE) [cite] to quantify multi-parent allele-specific and total expression from RNA-seq data for each tissue. EMASE was performed by the Genotype by RNA-seq (GBRS) software package (https://gbrs.readthedocs.io/en/latest/). In the process, R1 and R2 FASTQ files were combined and aligned to a hybridized (8-way) transcriptome generated for the 8 DO founder strains as single-ended reads. GBRS was also used to reconstruct the mouse genotype probabilities along ~69K markers, which was used for confirming genotypes in the quality control (QC) process. For the QC process, we used a Euclidean distances method (developed by Greg Keele - Churchill Lab) to compare the GBRS genotype probabilities between the tissues and the genotype probabilities array for all mice. The counts matrix for each tissue was processed to filter out transcripts with less than one read for at least half of the samples. RNA-seq batch effects were removed by regressing out batch as a random effect and considering sex and generation as fixed effects using lme4 R package. RNA-Seq counts were normalized relative to total read counts using the variance stabilizing transform (VST) as implemented in DESeq2 and using rank normal score.

We used R/qtl2 [cite] to perform eQTL analysis. We used the rank normal score data and used sex and DO generation as additive covariates. We also used kinship as a random effect. We used permutations to find a LOD threshold of 8 for significant QTLs which corresponded to a p value of 0.05.

To assess whether eQTL were shared across tissues, we compared eQTLs for each transcript across tissues. Significant eQTLs within 4Mb of each other were considered overlapping. We considered local and distal eQTLs separately.

To estimate local and distal heritability of each transcript, we scaled each normalized transcript to have a variance of 1. We then modeled this transcript with the local genotype using the fit1() function in R/qtl with a kinship correction. We used the resulting model to predict the transcript values. The variance of the predicted transcript its local heribatility. We then estimated the heritability of the residual of the model fit. The variance of the residual multiplied by its heritability is the distal heritability of the transcript.

We compared local and distal estimates of heritability to measures of trait relevance for each transcript. Trait relevance, was the Pearson correlation (R) between the transcript and the trait.

High-Dimensional Mediation Overview

The goal of high-dimensional mediation is to identify a composite trait that is perfectly mediated by a composite transcriptome. To accomplis this, we usied regularized and generalized canonical correlation analysis (RGCCA)[2], an extension of canonical correlation analysis (CCA)[3;4] that allows for more than two data sets with arbitrary relationships among them. Here we analyzed the correlations among three data sets, genotype, transcriptome, and phenotype explicitly modeling mediation in which the transcriptome (T) mediates the effect of the genome (G) on the phenome (P).

$$G \to T \to P$$

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

Kernelization

Before running high-dimensional mediation analysis, we kernelized the genotype, transcriptomic, and phenotype data sets to generate $n \times n$ matrices in where n is the number of individual mice. Each matrix described the relationships among individuals based on their genome, transcriptome, or phenome. Each matrix was generated as follows:

Kernelizing the genome

The kernel matrix of the genome is the overall kinship matrix as calculated by calc_kinship() in the R package qtl2 [cite]. We further mean-centered this matrix based on DO generation.

Kernelizing the transcriptome

Prior to kernelizing the transcriptome, we regressed out the effects of sex and DO generation. We then mean centered and standardized transcripts across individuals. The kernel matrix (K_t) for the transcriptome of each tissue was calculated as follows:

$$K_t = \frac{Tr \times Tr^T}{n_{Tx}}$$

where Tr is a matrix of transcript abundances with individuals in rows and transcripts in columns, Tr^T is Tr transpose, and n_{Tx} is the number of transcripts in the matrix. We kernelized each tissue's transcriptome and then averaged across all tissues to generate a single transcriptome kernel for all tissues.

Kernelizing the phenome

The phenome kernel was constructed the same way as the transcriptome kernel. We regressed out sex and DO generation and then mean centered and standardized the phenotypes. We used knn.impute() [cite] to impute missing values. We then used the above equation to generate the phenome kernel replacing the trancript matrix and number of transcripts with the phenotype matrix and number of phenotypes.

High-dimensional mediation Aanalysis

Matt has text for this section.

Calculation of loadings

Loadings onto transcripts and traits were calculated in the following way. calc_loadings()

Enrichment of biological terms

We performed gene set enrichment analysis (GSEA)[5] using the transcript loadings in each tissue as gene weights. GSEA determines enrichment of pathways based on where the contained genes appear in a ranked list of genes. If the genes in the pathway are more concentrated near the top (or the bottom) of the list than expected by chance, the pathway can be interpreted as being enriched with positively (negatively) loaded transcripts. We used the R package fgsea[6] to calculate normalized enrichment scores for all GO terms and all KEGG pathways.

We downloaded all KEGG[7] pathways for *Mus musculus* using the R package clusterProfiler[7]. We then used fgsea to calculate enrichment scores in each tissue using the transcript loadings in each tissue as our ranked list of genes. We reported the normalized enrichment score (NES) for the 10 pathways with the largest positive NES and the 10 pathways with the largest negative NES.

We used the R package pathview[8] to visualize the loadings from each tissue in interesting pathways. We scaled the loadings in each tissue by the maximum absolute value of loadings across all tissues to compare them across tissues.

We downloaded GO term annotations from Mouse Genome Informatics at the Jackson Laboratory[9] https://www.informatics.jax.org/downloads/reports/index.html We removed gene-annotation pairs labeled with NOT, indicating that these genes were known not to be involved in these GO terms. We also limited our search to GO terms with between 80 and 3000 genes. We used the R package annotate[10] to identify the ontology of each term and the R package pRoloc[11] to convert between GO terms and names. As with the KEGG pathways, we used fgsea to calculate a normalized enrichment score for each GO term and collected loadings for the transcripts in each term to compare across tissues.

TWAS in DO mice

We performed a transcriptome-wide analysis (TWAS) [cite] in the DO mice to compare to the results of high-dimensional mediation. To perform TWAS, we fit a linear model to explain variation in each transcript across the population using the genotype at the nearest marker to the gene transcription start site (TSS). We used kinship as a random effect and sex, diet, and DO generation as fixed effects. The predicted transcript from each of these models was the imputed transcript based only on the local genotype.

equation

We correlated each imputed transcript with each of the metabolic phenotypes after adjusting phenotypes for sex, diet, and DO generation. To calculate significance of these correlations, we performed permutation testing by shuffling labels of individual mice and recalculating correlation values. Significant correlations were those more extreme than any of the permuted values, corresponding to an empirical p value of 0. These are transcripts whose locally encoded expression level was significantly correlated with one of the metabolic traits. This suggests an association between the genetically encoded transcript level and the trait, but does not identify a direction of causation.

Literature support for genes

To determine whether each gene among those with large loadings or large heritability had a supported connection to obesity or diabetes in the literature, we used the R package easyPubMed[12]. We searched for the terms ("diabetes" OR "obesity") along with the tissue name (adipose, islet, liver, or muscle), and the gene name. We restricted the gene name to appear in the title or abstract as some short names appeared at random in contact information. We checked each gene with apparent literature support by hand to verify that support, and we removed spurious associations. For example, FAU is used as an acronym for fatty acid uptake and CAD is used as an acronym for coronary artery disease. Both terms co-occur with the terms diabetes and obesity in a manner independent of the genes *Fau* and *Cad*. Other genes that co-occurred with diabetes and obesity, but not as a functional connection were similarly removed. For example, the gene *Rpl27* is used as a reference gene for quantification of the expression of other genes, and co-occurrence with diabetes and obesity is a coincidence. We counted the abstracts associated with diabetes or obesity and each gene name, and determined that a gene had literature support when it had at least two abstracts linking it to the terms diabetes or obesity in the respective tissue.

Tissue-specific clusters

To compare the top loading genes across tissues, we selected genes with a loading at least 2.5 standard deviations from the mean across all tissues. We made a matrix consisting of the union of these sets populated with the tissue-specific loading for each gene. We used the pam() function in the R package cluster[13] to cluster the loading profiles around k medoids. We tested $k = 2$ through 20 and used silhouette andlysis to compare the separation of the clusters. The best separation was achieved with $k = 12$ clusters. For each cluster we used the R package gprofiler2[14] to enriched GO terms and KEGG pathways for the genes in each cluster.

CC-RIX mice

CC-RIX genotypes

We used the most recent common ancestor (MRCA) genotypes for the Collaborative Cross (CC) mice available on the University of North Carolina website: http://www.csbio.unc.edu/CCstatus/CCGenomes/

To generate CC-RIX genotypes, we averaged the haplotype probabilities for the two parental strains at each locus.

Imputation of gene expression in CC-RIX

To impute gene expression in the CC-RIX, we performed the following steps for each transcript in each tissue (adipose, liver, and skeletal muscle): 1. Calculate diploid CC-RIX genotype for all CC-RIX individuals at the marker nearest the transcription start site of the transcript. 2. Multiply the genotype probabilities by the eQTL coefficients identified in the DO population.

To check the accuracy of the imputation, we correlated the each imputed transcript with the measure transcript. The average Pearson correlation (r) was close to 0.5 for all three tissues (Supp. Fig. XXXA), and as expected, the correlation between the imputed transcript and the measured transcript was highly positively dependent on the local eQTL LOD score of the transcript (Supp. Fig. XXXB).

Prediction of CC-RIX traits

We used both measured expression and imputed expression combined with the results from HDM in the to predict phenotype in the CC-RIX. The traits measured in the DO and the CC-RIX were not identical, so we limited our prediction to body weight, which was measured in both populations, and was the largest contributor to the phenotype score in the DO.

For each CC-RIX individual, we multiplied the transcript abundaces across the transcriptome by the loadings derived from the HDM in the DO population (Fig. XXXA). This resulted in a vector with n elements, where n is the number of transcripts in the trancriptome. Each element was a weigted value that combined the relative abundance of the transcript with how that abundance affected the phenotype. We averaged the values in this vector to calculate an overall predicted phenotype score for the individual CC-RIX animal.

After calculating this predicted phenotype value across all CC-RIX animals, we correlated the predicted values from each tissue with measured body weight (Fig. XXXB).

Cell type specificity

We investigated whether the loadings derived from HDM reflected tissue composition changes in the DO mice prone to obesity on the high-fat diet. To do this, we acquired lists of cell-type specific transcripts from the literature. In adipose tissue, we looked at cell-type specific transcripts for macrophages, leukocytes, adipocyte progenitors, and adipocytes as defined in [29087381]. In pancreatic islets, we looked at cell-type specific transcripts for alpha cells, beta cells, delta cells, ductal cells, mast cells, macrophages, acinar cells, stellate cells, gamma and epsilon cells, and endothelial cells as defined by [36778506]. Both studies defined cell-type specific transcripts based on human cell types. We collected the loadings for each set of cell-type specific transcripts in the respective tissue and asked whether the mean loading for the cell type differed significantly from 0 (Figure XXX). A significant positive loading for the cell type would suggest a genetic predisposition to have a higher proportion of that cell type in the tissue. To determine whether each mean loading differed significantly from 0, we performed permutation tests. We randomly sampled n genes outside of the cell-type specific, where n was the number of genes in the set. We compared the distribution of loading means over 10,000 random draws to that seen in the observed data. We used a significance threshold of 0.01.

Comparison of transcriptomic signatures to human transcriptomic signatures

To compare the transcriptomic signatures identified in the DO mice to those seen in human patients, we downloaded human gene expression data from the Gene Expression Omnibus (GEO) [cite]. We focused on adipose tissue because this had the strongest relationship to obesity and insulin resistance in the DO. We downloaded the following human gene expression data sets:

- Accession number GSE152517 - Performed bulk RNA sequencing on visceral adipose tissue resected from seven diabetic and seven non-diabetic obese individuals.

- Accession number GSE44000 - Used Agilent-014850 4X44K human whole genome platform arrays (GPL6480) to measure gene expression in purified adipocytes derived from the subcutaneous adipose tissue of seven obese (BMI>30) and seven lean (BMI<25) post-menopausal women.

- Accession number GSE205668 - Subcutaneous adipose tissue was resected during elective surgery from 35 normal weight, and 26 obese children. Gene expression was measured by RNA sequencing with an Illumina HiSeq 2500.

- Accession number GSE29231 - Visceral adipose biopsies were taken from three female patients with type 2 diabetes, and three non-diabetic female patients. Expression was measured with Illumina HumanHT-12 v3 Expression BeadChip arrays.

We downloaded each data set from GEO using the R package GEOquery [cite]. In each case, we verified that gene expression was log transdormed, and performed the transformation ourselves if it hadn't already been done. When covariates such as age and sex were available in the meta data files we regressed out these variables. We mean cetered and standardized gene expression across transcripts.

We matched the human gene expression to the mouse gene expression by pairing orthologs as defined in The Jackson Laboratory's mouse genome informatics data base (MGI) [cite]. We multiplied each transcript in the human data by the adipose tissue loading of its ortholog in the DO mice. This resulted in a vector of weighted transcript values for each individual patient based on their own transcriptional profile and the obesity-related transcriptional signature from the DO analysis. The mean of this vector for an individual was the prediction of their obesity status. Higher values indicate a prediction of higher obesity or risk of metabolic disease based on adipose gene expression. We then compared the values across groups, either obese and non-obese, or diabetic and non-diabetic depending on the groups in each study.

Connectivity Map Queries

We queried the transcript loading signatures from adipose tissue and pancreatic islets with the CMAP database. These tissues are the most related to metabolic disease and diabetes respectively.

The gene expression profiles in the Connectivity Map database are derived from human cell lines and human primary cultures, and are indexed by Entrez gene IDs. To query the CMAP database, we identified the Entrez gene IDs for the human orthologs of the mouse genes expressed in each tissue. Each CMAP query takes the 150 most up-regulated and the 150 most down-regulated genes in a signature, however, not all human genes are included in their database. To ensure we had as many genes as possible in the query, we selected the top and bottom 200 genes with the most extreme positive and negative loadings respectively. We pasted these into the CLUE query application available at https://clue.io/query.

We filtered the results in two ways: First, we looked at the most significantly anti-correlated ($-log_{10}(\text{FDR}q) > 15$) hits across all cell types. Second, we looked at the most anti-correlated within the most related cell type to the query and considered hits regardless of $-log_{10}(\text{FDR}p)$. For adipose tissue we looked in normal adipocytes, abbreviated ASC in the CMAP database, and for pancreatic islets we looked in pancreatic cancer cells, abbreviated YAPC in the CMAP database.

References

[1] M. P. Keller, D. M. Gatti, K. L. Schueler, M. E. Rabaglia, D. S. Stapleton, P. Simecek, M. Vincent, S. Allen, A. T. Broman, R. Bacher, C. Kendziorski, K. W. Broman, B. S. Yandell, G. A. Churchill, and A. D. Attie. Genetic Drivers of Pancreatic Islet Function. *Genetics*, 209(1):335–356, May 2018.

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

[3] Thomas R Knapp. Canonical correlation analysis: A general parametric significance-testing system. *Psychological Bulletin*, 85(2):410, 1978.

[4] Harold Hotelling. Relations between two sets of variates. In *Breakthroughs in statistics: methodology and distribution*, pages 162–190. Springer, 1992.

[5] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102(43):15545–15550, Oct 2005.

[6] Gennady Korotkevich, Vladimir Sukhov, and Alexey Sergushichev. Fast gene set enrichment analysis. *bioRxiv*, 2019.

[7] M. Kanehisa, M. Furumichi, Y. Sato, M. Kawashima, and M. Ishiguro-Watanabe. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res*, 51(D1):D587–D592, Jan 2023.

[8] W. Luo and C. Brouwer. Pathview: an R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics*, 29(14):1830–1831, Jul 2013.

[9] J. A. Blake, R. Baldarelli, J. A. Kadin, J. E. Richardson, C. L. Smith, C. J. Bult, A. V. Anagnostopoulos, J. S. Beal, S. M. Bello, O. Blodgett, N. E. Butler, J. Campbell, K. R. Christie, L. E. Corbani, M. E. Dolan, H. J. Drabkin, M. Flores, S. L. Giannatto, A. Guerra, P. Hale, D. P. Hill, J. Judd, M. Law, M. McAndrews, D. Miers, C. Mitchell, H. Motenko, L. Ni, H. Onda, J. Ormsby, M. Perry, J. M. Recla, D. Shaw, D. Sitnikov, M. Tomczuk, L. Wilming, and Y. ' Zhu. Mouse Genome Database (MGD): Knowledgebase for mouse-human comparative biology. *Nucleic Acids Res*, 49(D1):D981–D987, Jan 2021.

[10] Jeff Gentry. *annotate: Annotation for microarrays*, 2024. R package version 1.82.0.

[11] L. Gatto, L. M. Breckels, S. Wieczorek, T. Burger, and K. S. Lilley. Mass-spectrometry-based spatial proteomics data analysis using pRoloc and pRolocdata. *Bioinformatics*, 30(9):1322–1324, May 2014.

[12] Damiano Fantini. *easyPubMed: Search and Retrieve Scientific Publication Records from PubMed*, 2019. R package version 2.13.

[13] Martin Maechler, Peter Rousseeuw, Anja Struyf, Mia Hubert, and Kurt Hornik. *cluster: Cluster Analysis Basics and Extensions*, 2023. R package version 2.1.6 — For new features, see the 'NEWS' and the 'Changelog' file in the package source).

[14] Liis Kolberg, Uku Raudvere, Ivan Kuzmin, Jaak Vilo, and Hedi Peterson. gprofiler2– an r package for gene list functional enrichment analysis and namespace conversion toolset g:profiler. *F1000Research*, 9 (ELIXIR)(709), 2020. R package version 0.2.3.