Distal mediator-enriched, placental transcriptome-wide analyses illustrate the Developmental Origins of Health and Disease

SUPPLEMENTAL METHODS

Data acquisition and quality control

Genotype data

Genomic DNA was isolated from umbilical cord blood and genotyping was performed using Illumina 1 Million Quad and Human OmniExpression-12 v1.0 arrays[1,2]. Prior to imputation, from the original set of 731,442 markers, we removed SNPs with call rate < 90% and MAF < 1%. We did not use deviation from Hardy-Weinberg equilibrium as an exclusion criterion since ELGAN is an admixed population. This resulted in 700,845 SNPs. We removed 4 individuals out of 733 with sample-level missingness > 10% using PLINK[3]. We first performed strand-flipping according to the TOPMed Freeze 5 reference panel and using eagle and minimac4 for phasing and imputation[4–6]. Genotypes were coded as dosages, representing 0, 1, and 2 copies of the minor allele. The minor allele was coded in accordance with the NCBI Database of Genetic Variation[7]. Overall, after QC and normalization, we considered a total of 6,567,190 SNPs. We obtained processed genetic data from the Rhode Island Children's Health Study, as described before[8].

Expression data

mRNA expression was determined using the Illumina QuantSeq 3' mRNA-Seq Library Prep Kit, a method with high strand specificity[9]. mRNA-sequencing libraries were pooled and sequenced (single-end 50 bp) on one lane of the Illumina HiSeq 2500. mRNA were quantified through pseudo-alignment with Salmon[10] mapped to the GENCODE Release 31 (GRCh37) reference transcriptome. miRNA expression profiles were assessed using the HTG EdgeSeq miRNA Whole Transcriptome Assay (HTG Molecular Diagnostics, Tucson, AZ). miRNA were aligned to probe sequences and quantified using the HTG EdgeSeq System[11].

Genes and miRNAs with less than 5 counts for each sample were filtered, resulting in 11,224 genes and 2,047 miRNAs for downstream analysis. Distributional differences between lanes were first upper-quartile normalized[12,13]. Unwanted technical and biological variation (e.g. tissue heterogeneity) was then estimated using RUVSeq[14], where we empirically defined transcripts not associated with outcomes of interest as negative control housekeeping probes[15]. One dimension of unwanted variation was removed from the variance-stabilized transformation of the gene expression data using the limma package[14–17]. We obtained processed RNA expression data from the Rhode Island Children's Health Study, as described before[8]. Overall, after QC and normalization, we considered 12,020 genes and 1,898 miRNAs.

Methylation data

Extracted DNA sequences were bisulfate-converted using the EZ DNA methylation kit (Zymo Research, Irvine, CA) and followed by quantification using the Infinium MethylationEPIC BeadChip (Illumina, San Diego, CA), which measures CpG loci at a single nucleotide resolution, as previously described[18–21]. Quality control and normalization were performed resulting in 856,832 CpG probes from downstream analysis, with methylation represented as the average methylation level at a single CpG site (β-value)[19,22–25]. DNA methylation data was imported into R for pre-processing using the minfi package[23,24]. Quality control was performed at the sample level, excluding samples that failed and technical duplicates; 411 samples were retained for subsequent analyses.

Functional normalization was performed with a preliminary step of normal-exponential out-of band (noob) correction method[26] for background subtraction and dye normalization, followed by the typical functional normalization method with the top two principal components of the control matrix[23,24]. Quality control was performed on individual probes by computing a detection P value and excluded 806 (0.09%) probes with

non-significant detection (P > 0.01) for 5% or more of the samples. A total of 856,832 CpG sites were included in the final analyses. Lastly, the ComBat function was used from the sva package to adjust for batch effects from sample plate[27]. The data were visualized using density distributions at all processing steps. Each probe measured the average methylation level at a single CpG site. Methylation levels were calculated and expressed as β values, with

$$\beta = \frac{M}{U + M + 100},$$

where $M$is the intensity of the methylated allele and U is the intensity of the unmethylated allele. β-values were logit transformed to M values for statistical analyses[28]. Overall, after QC and normalization, we considered 846,233 CpG sites.

QTL mapping

We conducted genome-wide eQTL mapping between all genotypes and all genes in the transcriptome using a standard linear regression in MatrixeQTL[29]. Here, we ran an additive model with gene expression as the outcome, SNP dosage as the primary predictor of interest, with covariate adjustments for 10 genotype PCs (for population stratification), sex, gestational duration, maternal age, maternal smoking status, and 20 expression PEER factors[30]. Mediators here are defined as RNA expression of genes that code for transcription factors, miRNAs, and CpG methylation sites. In sum, we call the expression or methylation of a mediator its intensity. We also conducted genome-wide mediator-QTL mapping with the intensity of mediators as the outcome with the same predictors as in the eQTL mapping. Lastly, we also assessed associations between mediators and gene expression using the same linear models, with mediator intensity as the main predictor. All intensities were scaled to zero mean and unit variance.

Predictive models of expression using MOSTWAS

Transcriptomic prediction using MeTWAS

Here, we present mediator-enriched TWAS, or MeTWAS, one of the two tools in the MOSTWAS R package[31]. Across n samples, consider the vector Y_G, the expression of a gene *G* of interest, the matrix $\boldsymbol{X}_G$ of local SNP dosages in a user-defined window around gene *G* (default of 1 Megabase), and m_G mediating biomarkers that we estimate to be significantly associated with the expression of gene *G* via a relevant one-way test of association. These mediating biomarkers can be DNA methylation sites, microRNAs, transcription factors (or genes that code for transcription factors), or any molecular profile that may be genetically heritable and affect transcription. Accordingly, let the matrix $\boldsymbol{X}_{M_j}$ by the local-genotype dosages in a 1 Mb around mediator $j, 1 \leq j \leq m_G$. Furthermore, let M_j be the intensity of mediator j (methylation M-value if j is a CpG site, expression if j is a miRNA or gene, etc). Prior to any modeling, we scale Y_G and all M_j to zero mean and unit variance. We also residualized M_j and Y_G with the covariates using limma[16] to account for population stratification using principal components of the global genotype matrix and relevant clinical covariates to obtain $\widetilde{M}_j$ and $\tilde{Y}_G$.

Transcriptome prediction in MeTWAS draws from two-step regression, as indicated in **Supplemental Figure S1**. First, in the training set for a given training-test split, for $1 \leq j \leq m_G$, we model the residualized intensity $\tilde{M}_j$ of training-set specific mediator j with the following additive model:

$$\widetilde{M}_j = \boldsymbol{X}_{M_j}^{train} w_j + \epsilon_m,$$

where w_j is the effect sizes of the SNPs on the mediator intensity in the training set. As in traditional, transcriptomic imputation models[32,33], we find $\widehat{w}_j$ using either (1) elastic net regression with mixing parameter $\alpha = 0.5$ and λ tuned over 5-fold cross validation using glmnet[34] or (2) linear mixed modeling assuming random effects for the SNPs using rrBLUP[35]. Only significantly heritable (default $P < 0.05$ for the likelihood ratio test[36]) and well-cross validated (default $R^2 \geq 0.01$) expression models are considered.

For all j, using these optimized predictive models for M_j, as denoted by $\widehat{w}_{M_j}$, we estimate the genetically regulated intensity (GRIn) of the mediator j, denoted $GRIn_{m_j}$, in the test set. Denote $\widehat{\boldsymbol{M}}$ as the $n \times m$ matrix of estimated GRIn, such that the $j$the column of $\widehat{\boldsymbol{M}}$ is $GRIn_{m_j}$ across all n samples.

Next, we consider the following additive model for the residualized expression of gene G:

$$\tilde{Y}_G = \widehat{\boldsymbol{M}}\beta_{\boldsymbol{M}} + \epsilon_{Y_{G_1}},$$

$$\tilde{Y}_G - \widehat{\boldsymbol{M}}\hat{\beta}_{\boldsymbol{M}} = X_G w_G + \epsilon_{Y_{G_2}},$$

where $\beta_{\boldsymbol{M}}$ is the fixed effect sizes of $GRIn_{m_j}$ on $\tilde{Y}_G$, $\widehat{\boldsymbol{M}}$ is the matrix of estimated GRIn for all m_j mediators, $\boldsymbol{X}_G$ are the local-SNPs to gene *G*, and w_G are the "random" or regularized effect sizes of the local-SNPs on expression of *G*. We first estimate $\hat{\beta}_M$ by traditional ordinary least squares. Next, using either elastic net or linear mixed modeling, we can estimate $\widehat{w}_G$.

Transcriptomic prediction using DePMA

Expression prediction in distal eQTL Prioritization via Mediation Analysis (DePMA) hinges on assessing distal-eSNPs for inclusion in the design matrix via mediation analysis, adopting methods from previous studies[37,38]. We first split the data for gene expression, SNP dosages, and potential mediators into k training-test splits (default 3).

In the training set, we identify mediation test triplets that consist of (1) a gene of interest *G* with expression Y_G (scaled to zero mean and unit variance), (2) a distal-eSNP s in association with *G* at $P < 10^{-6}$ with dosages X_s, and (3) a set of m biomarkers local to s that are associated with s at FDR-adjusted $P < 0.05$ with intensities in the m columns of $\boldsymbol{M}$. The columns of $\boldsymbol{M}$ are scaled to zero mean and unit variance. Consider the following mediation model:

$$Y_G = \boldsymbol{X}_s\beta_s + \boldsymbol{M}\beta_M + \boldsymbol{X}_C\beta_C + \epsilon_{Y_G},$$

$$M_j = \boldsymbol{X}_s\alpha_{M_j} + \boldsymbol{X}_C\alpha_{C,j} + \epsilon_{M_j}, 1 \le j \le m.$$

Here, β_M is the vector of effects of the mediators local to s on Y_G, adjusted for the effects from s and covariates, and $\alpha_{\boldsymbol{M}}$ as the effects of s on mediators $M_j, 1 \le j \le m$. We assume that $\epsilon_{Y_g} \sim N(0, \sigma^2)$ and $\epsilon_{\boldsymbol{M}} \sim \boldsymbol{N}_m(0, \boldsymbol{\Sigma_M})$, where $\boldsymbol{\Sigma_M}$ may have non-zero off-diagonal elements that represent non-zero covariance between mediator intensities. We assume that these error terms are independent. We define the total mediation effect (TME)[39] of SNP s as

$$TME = \alpha_{\boldsymbol{M}}^T\beta_{\boldsymbol{M}}.$$

We are interested in SNPs with large absolute TME, which we assess with a two-sided test of $H_0: TME = 0$. We assess this hypothesis with a permutation test to obtain a permutation P-value, as more direct methods of computing standard errors for the estimated TM are often biased[38,40]. We also provide an option to estimate an asymptotic approximation to the standard error of TME for a Wald-type test (see Bhattacharya *et al* for a discussion[31]). Corresponding to the t testing triplets identified, we obtain vectors of length t of TMEs and P-values for each distal-eSNP to G. We estimate the FDR-adjusted P-value for each test, and those SNPs with significant TMEs are prioritized. Final DePMA model weights are estimated from the local SNPs to gene *G* and all prioritized distal-eSNPs using elastic net or linear mixed models.

Tests of association

Overall TWAS test

In an external GWAS panel, if individual SNPs are available, model weights from either MeTWAS or DePMA can be multiplied by their corresponding SNP dosages to construct the Genetically Regulated eXpression (GReX) for a given gene. This value represents the portion of expression (in the given tissue) that is directly predicted or regulated by germline genetics. We run a linear model or test of association with phenotype using this GReX value for the eventual TWAS test of association.

If individual SNPs are not available, then the weighted burden Z-test, proposed by Gusev *et al*, can be employed using summary statistics[32]. Briefly, we compute

$$\tilde{Z} = \frac{w_G Z}{\left(w_G \mathbf{\Sigma}_{s,s} w_G^T\right)^{1/2}}.$$

Here, Z is the vector of Z-scores of SNP-trait associations for SNPs used in predicting expression. The vector w_G represents the vector of SNP-gene effects from MeTWAS or DePMA and $\mathbf{\Sigma}_{s,s}$ is the LD matrix between the SNPs represented in w_G. The test statistic $\tilde{Z}$ can be compared to the standard Normal distribution for inference.

Permutation test

We implement a permutation test, condition on the GWAS effect sizes, to assess whether the same distribution of SNP-gene effect sizes could yield a significant associations by chance[32]. We permute w_G 1,000 times without replacement and recompute the weighted burden test to generate a null distribution for $\tilde{Z}$. This permutation test is only conducted for overall associations at $P < 2.5 \times 10^{-6}$.

Distal-SNPs added-last test

Lastly, we also implement a test to assess the information added from distal-eSNPs in the weighted burden test beyond what we find from local SNPs. This test is analogous to a group added-last test in regression analysis, applied here to GWAS summary statistics. Let Z_l and Z_d be the vector of Z-scores from GWAS summary statistics from local and distal-SNPs identified by a MOSTWAS model. The local and distal-SNP effects from the MOSTWAS model are represented in w_l and w_d. Formally, we test whether the weighted Z-score $\tilde{Z}_d = w_d Z_d$ from distal-SNPs is significantly larger than 0 given the observed weighted Z-score from local SNPs $\tilde{Z}_l = w_l Z_l$. We draw from the assumption that $(\tilde{Z}_d, \tilde{Z}_l)$ follow a bivariate Normal distribution. Namely, we conduct a two-sided Wald-type test for the null hypothesis:

$H_0: w_d Z_d | w_l Z_l = \tilde{Z}_l = 0.$

We can derive a null distribution using conditional of bivariate Normal distributions (see Bhattacharya et al[31])

Genetic heritability and correlation estimation

At the genome-wide genetic level, we estimated the heritability of and genetic correlation between traits via summary statistics using LD score regression[41]. On the predicted expression level, we adopted approaches from Gusev *et al* and Mancuso *et al* to quantify the heritability (h^2_{GE}) of and genetic correlations (ρ_{GE}) between traits at the predicted placental expression level[32,42]. We assume that the expected χ^2 statistic under a complex trait is a linear function of the LD score[41]. The effect size of the LD score on the χ^2 is proportional to h^2_{GE}:

$$E[\chi^2] = 1 + \left(\frac{N_{Tl}}{M}\right) h^2_{GE} + N_T a,$$

where N_T is the GWAS sample size, M is the number of genes, l is the LD scores for genes, and a is the effect of population structure. We estimated the LD scores of each gene by predicting expression in European samples of 1000 Genomes and computing the sample correlations and inferred h^2_{GE} using ordinary least squares. We employed ROHGE to estimate and test for significant genetic correlations between traits at the predicted expression level (details in Mancuso *et al*[42]).

References

1. Ådén, U. *et al.* Candidate gene analysis: Severe intraventricular hemorrhage in inborn preterm neonates. *J. Pediatr.* **163**, (2013).
2. Yasuno, K. *et al.* Genome-wide association study of intracranial aneurysm identifies three new risk loci. *Nat. Genet.* **42**, 420–425 (2010).
3. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet* **81**, 559–575 (2007).
4. Kowalski, M. H. *et al.* Use of >100,000 NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations. *PLoS Genet.* **15**, e1008500 (2019).
5. Loh, P. R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).
6. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).
7. Sherry, S. T. *et al.* DbSNP: The NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).
8. Peng, S. *et al.* Expression quantitative trait loci (eQTLs) in human placentas suggest developmental origins of complex diseases. *Hum. Mol. Genet.* **26**, 3432–3441 (2017).
9. A Eaves, L. *et al.* A role for microRNAs in the epigenetic control of sexually dimorphic gene expression in the human placenta. *Epigenomics* **12**, 1543–1558 (2020).
10. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
11. Qi, Z. *et al.* Reliable Gene Expression Profiling from Small and Hematoxylin and Eosin–Stained Clinical Formalin-Fixed, Paraffin-Embedded Specimens Using the HTG EdgeSeq Platform. *J. Mol. Diagnostics* **21**, 796–807 (2019).
12. Bullard, J. H., Purdom, E., Hansen, K. D. & Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **11**, 94 (2010).
13. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).
14. Risso, D., Ngai, J., Speed, T. P. & Dudoit, S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.* **32**, 896–902 (2014).
15. Gagnon-Bartsch, J. A. & Speed, T. P. Using control genes to correct for unwanted variation in microarray data. *Biostatistics* **13**, 539–552 (2012).
16. Phipson, B., Lee, S., Majewski, I. J., Alexander, W. S. & Smyth, G. K. Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression. *Ann. Appl. Stat.* **10**, 946–963 (2016).
17. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
18. Addo, K. A. *et al.* Acetaminophen use during pregnancy and DNA methylation in the placenta of the extremely low gestational age newborn (ELGAN) cohort. *Environ. Epigenetics* **5**, (2019).
19. Santos, H. P. *et al.* Epigenome-wide DNA methylation in placentas from preterm infants: association with maternal socioeconomic status. *Epigenetics* **14**, 751–765 (2019).
20. Bulka, C. M. *et al.* Placental CpG methylation of inflammation, angiogenic, and neurotrophic genes and retinopathy of prematurity. *Investig. Ophthalmol. Vis. Sci.* **60**, 2888–2894 (2019).
21. Clark, J. *et al.* Associations between placental CpG methylation of metastable epialleles and childhood body mass index across ages one, two and ten in the Extremely Low Gestational Age Newborns (ELGAN) cohort. *Epigenetics* **14**, 1102–1111 (2019).

22. Aryee, M. J. *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363–1369 (2014).
23. Fortin, J.-P. *et al.* Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biol.* **15**, 503 (2014).
24. Fortin, J. P., Triche, T. J. & Hansen, K. D. Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi. *Bioinformatics* **33**, 558–560 (2017).
25. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
26. Triche, T. J., Weisenberger, D. J., Van Den Berg, D., Laird, P. W. & Siegmund, K. D. Low-level processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic Acids Res.* **41**, e90 (2013).
27. Leek, J. T. & Storey, J. D. Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. *PLoS Genet.* **3**, e161 (2007).
28. Du, P. *et al.* Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* **11**, 587 (2010).
29. Shabalin, A. A. Gene expression Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353–1358 (2012).
30. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* **7**, 500–507 (2012).
31. Bhattacharya, A., Li, Y. & Love, M. I. MOSTWAS: Multi-Omic Strategies for Transcriptome-Wide Association Studies. *PLOS Genet.* **17**, e1009398 (2021).
32. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).
33. Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**, 1091–1098 (2015).
34. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **33**, 1–22 (2010).
35. Endelman, J. B. Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *Plant Genome* **4**, 250–255 (2011).
36. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
37. Pierce, B. L. *et al.* Mediation Analysis Demonstrates That Trans-eQTLs Are Often Explained by Cis-Mediation: A Genome-Wide Analysis among 1,800 South Asians. *PLoS Genet.* **10**, (2014).
38. Shan, N., Wang, Z. & Hou, L. Identification of trans-eQTLs using mediation analysis with multiple mediators. *BMC Bioinformatics* **20**, (2019).
39. Sobel, M. E. Direct and Indirect Effects in Linear Structural Equation Models. *Sociol. Methods Res.* **16**, 155–176 (1987).
40. Mackinnon, D. P., Lockwood, C. M. & Williams, J. Confidence Limits for the Indirect Effect: Distribution of the Product and Resampling Methods. *Multivariate Behav. Res.* **39**, 99–128 (2004).
41. Bulik-Sullivan, B. *et al.* LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
42. Mancuso, N. *et al.* Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits. *Am. J. Hum. Genet.* **100**, 473–487 (2017).