Transcriptome-wide association study of the plasma proteome reveals *cis* and *trans* regulatory mechanisms underlying complex traits

Authors

Henry Wittich, Kristin Ardlie, Kent D. Taylor, ...,
Ani Manichaikul, Hae Kyung Im,
Heather E. Wheeler

Correspondence

hwheeler1@luc.edu

We performed TWAS for thousands of plasma proteins, comparing same-gene, *cis*, and *trans* effects across tissues. We show the heritable component of gene expression more strongly correlates with protein levels than with total observed expression and is therefore more useful in uncovering the functions of SNPs associated with complex traits.

Wittich et al., 2024, The American Journal of Human Genetics *111*, 445–455
March 7, 2024 © 2024 American Society of Human Genetics.
https://doi.org/10.1016/j.ajhg.2024.01.006

Transcriptome-wide association study of the plasma proteome reveals *cis* and *trans* regulatory mechanisms underlying complex traits

Henry Wittich,[1] Kristin Ardlie,[2] Kent D. Taylor,[3] Peter Durda,[4] Yongmei Liu,[5] Anna Mikhaylova,[6] Chris R. Gignoux,[7] Michael H. Cho,[8] Stephen S. Rich,[9] Jerome I. Rotter,[3] NHLBI TOPMed Consortium, Ani Manichaikul,[9] Hae Kyung Im,[10] and Heather E. Wheeler[1,11,*]

Summary

Regulation of transcription and translation are mechanisms through which genetic variants affect complex traits. Expression quantitative trait locus (eQTL) studies have been more successful at identifying *cis*-eQTL (within 1 Mb of the transcription start site) than *trans*-eQTL. Here, we tested the *cis* component of gene expression for association with observed plasma protein levels to identify *cis*- and *trans*-acting genes that regulate protein levels. We used transcriptome prediction models from 49 Genotype-Tissue Expression (GTEx) Project tissues to predict the *cis* component of gene expression and tested the predicted expression of every gene in every tissue for association with the observed abundance of 3,622 plasma proteins measured in 3,301 individuals from the INTERVAL study. We tested significant results for replication in 971 individuals from the *Trans*-omics for Precision Medicine (TOPMed) Multi-Ethnic Study of Atherosclerosis (MESA). We found 1,168 and 1,210 *cis*- and *trans*-acting associations that replicated in TOPMed (FDR < 0.05) with a median expected true positive rate (π_1) across tissues of 0.806 and 0.390, respectively. The target proteins of *trans*-acting genes were enriched for transcription factor binding sites and autoimmune diseases in the GWAS catalog. Furthermore, we found a higher correlation between predicted expression and protein levels of the same underlying gene ($R = 0.17$) than observed expression ($R = 0.10$, $p = 7.50 \times 10^{-11}$). This indicates the *cis*-acting genetically regulated (heritable) component of gene expression is more consistent across tissues than total observed expression (genetics + environment) and is useful in uncovering the function of SNPs associated with complex traits.

Introduction

The regulation of gene expression and protein abundance are important mechanisms through which many noncoding genome-wide association study (GWAS) SNPs affect traits.[1] Expression quantitative trait locus (eQTL) mapping in multiple human tissues has discovered a variety of both distal (*trans*) and proximal (*cis*) variants associated with gene expression.[2–4] While *cis*-eQTLs tend to have larger effect sizes than *trans*-eQTLs,[5,6] studies have shown that *trans*-eQTLs account for a larger portion of the heritability of gene expression.[5,7] One study on twins found that, on average, *trans*-eQTLs explained over 3 times the variance in gene expression than *cis*-eQTLs.[8] Furthermore, *trans*-eQTLs tend to be more tissue specific, suggesting that they play a role in cell type differentiation.[4,6]

Despite the importance of *trans*-eQTLs in regulating gene expression, QTL mapping studies have been limited in their ability to detect *trans*-acting effects due in part to the high multiple testing burden, as well as their comparatively low effect sizes.[4,9,10] Methods that minimize the multiple testing burden by prioritizing subsets of variants or grouping *trans*-genes have proven more successful at identifying *trans*-eQTLs.[9–13] For example, one study that tested the *cis*-component of gene expression for association with the observed expression of distant genes identified more replicable *trans*-acting genes than a comparable *trans*-eQTL study.[9] This is because many *trans*-eQTLs colocalize with *cis*-eQTLs, and they affect the expression of distant genes through *cis*-mediators such as a nearby transcription factor (TF) gene.[4,5,10]

The advent of advanced assay technologies that capture and measure protein abundances has enabled protein quantitative trait locus (pQTL) mapping studies to identify variants associated with the abundance of proteins.[14,15] *Trans*-pQTLs, like eQTLs, tend to have lower effect sizes and be tissue specific.[16] Mirroring methods for detecting *trans*-eQTLs, we hypothesized that *cis*-prioritization will improve detection of *trans*-pQTLs.

Here, we applied a transcriptome-wide association study (TWAS) framework to proteomic data, testing the genetically predicted expression of genes for association with the observed abundance of plasma proteins.[17] We show that TWAS for protein levels is an effective method for identifying

[1] Program in Bioinformatics, Loyola University Chicago, Chicago, IL 60660, USA; [2] Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA; [3] The Institute for Translational Genomics and Population Sciences, Department of Pediatrics, The Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center, Torrance, CA 90502, USA; [4] Laboratory for Clinical Biochemistry Research, University of Vermont, Colchester, VT 05446, USA; [5] Department of Medicine, Duke University School of Medicine, Durham, NC 27710, USA; [6] Department of Biostatistics, University of Washington, Seattle, WA 98195, USA; [7] Division of Biomedical Informatics and Personalized Medicine, Department of Medicine, University of Colorado Denver Anschutz Medical Campus, Aurora, CO 80045, USA; [8] Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, MA 02115, USA; [9] Center for Public Health Genomics, Department of Public Health Sciences, University of Virginia, Charlottesville, VA 22908, USA; [10] Section of Genetic Medicine, The University of Chicago, Chicago, IL 60637, USA; [11] Department of Biology, Loyola University Chicago, Chicago, IL 60660, USA
*Correspondence: hwheeler1@luc.edu

https://doi.org/10.1016/j.ajhg.2024.01.006.

replicable *trans*-acting associations between predicted transcripts and proteins. We also found a high expected proportion of true positives for associations between the predicted transcripts and protein products of the same underlying gene. Furthermore, using RNA-sequencing data, we show that predicted gene expression better correlates with protein levels than observed gene expression.

Material and methods

This study was approved by the Loyola University Chicago institutional review board (IRB) project #2014. Appropriate informed consent was obtained from human subjects.

Genome and proteome data

Our discovery dataset was from the INTERVAL study, which was conducted on around 50,000 blood donors with European ancestry across England.[18] Here, we used data from the 3,301 individuals who had both a genotyping microarray performed (EGA: EGAD00010001544) and a targeted proteome assay run to measure their plasma proteome levels (EGA: EGAD00001004080).[14] Data generation and quality control have previously been described by the INTERVAL study.[14,18] Briefly, an Affymetrix Axiom UK Biobank array was used for genotyping, and imputation was performed on the Sanger imputation server using a combined 1000 Genomes phase 3-UK10K reference panel.[14,19] Genotypes were then filtered for minor allele frequency (MAF) > 0.01 and imputation $R^2 > 0.8$.[19] The SOMAscan assay used to collect the proteomic data targeted 3,622 plasma proteins.[20] The protein levels were log transformed and adjusted for age, sex, duration between blood draw and processing, and the first three genetic principal components (PCs).[14]

Our replication dataset was from the Trans-omics for Precision Medicine (TOPMed) Multi-Ethnic Study of Atherosclerosis (MESA) multi-omics pilot study. The TOPMed program is a research consortium that aims at improving personalized disease treatments through the study of genetics and other omics traits' effects on disease traits and drug responses.[21] MESA is a community-based cohort study designed to determine the prevalence, determinants, and progression of subclinical cardiovascular disease.[22] MESA recruited men and women aged 45–84 free of clinical cardiovascular disease at baseline from six different locations in the United States and from four major race/ethnicity groups, which included African American (AFA), Chinese (CHN), European (EUR), and Hispanic/Latino (HIS).[22] Individuals were genotyped as part of the MESA SHARe study (dbGaP: phs000420.v6.p3).[22] As previously described, an Affymetrix 6.0 array was used for genotyping, and imputation was performed on the Michigan imputation server using the 1000 Genomes phase 3 v5 reference panel.[19,23] Imputed genotypes were then filtered for MAF >0.01 and imputation $R^2 > 0.8$.[19] Additionally, individuals taking part in the MESA multi-omics pilot study had their plasma proteome measured with a SOMAscan HTS Assay that targeted 1,300 plasma proteins, 1,039 of which overlapped with the proteins tested in the INTERVAL study.[15] Protein levels were measured at two time points, exam 1 (2000–2002) and exam 5 (2010–2012). We log transformed each time point and then adjusted for age and sex. We then took the mean of the two time points (if a participant was not measured at both time points, we then used the single time point), performed rank inverse normalization, and adjusted for the first ten genotypic PCs.[19] In total, our replication cohort included 971 individuals with genotypes and plasma protein level measurements (AFA, $n = 183$; CHN, $n = 71$; EUR, $n = 416$; HIS, $n = 301$).

Transcriptome data

For our analysis comparing the genetically regulated transcriptome to the observed transcriptome, we used transcriptomic data from individuals in the MESA multi-omics pilot study. RNA sequencing was performed for individuals from all four populations (AFA, CHN, EUR, and HIS) in three different blood cell types: peripheral blood mononuclear cells (PBMC), CD14$^+$ monocytes, and CD4$^+$ T cells.[22] In total, 395 monocyte samples and 397 T cell samples were sequenced at one time point, exam 5, and 1,287 PBMC samples were sequenced over two time points, exam 1 and exam 5. Genes with average transcripts per million (TPM) values <0.1 were filtered out, leaving 18,193 genes with expression measurements in PBMC, monocytes, and T cells. After log transforming each TPM value and adjusting for age and sex as covariates using linear regression and extracting the residuals, we took the mean of the two time points (or the single adjusted log-transformed value if expression levels were only measured once), performed rank-based inverse normal transformation, and adjusted for the first 10 genotype and 10 expression PCs, as described previously.[24]

TWAS for protein levels

We performed TWAS with the software tool, PrediXcan, which leverages eQTL weights to predict genetically regulated expression (GReX) and performs a linear association analysis to correlate GReX with a measured trait.[17] We used gene expression prediction models from PredictDB, which were built using the Genotype-Tissue Expression (GTEx) Project's version 8 release, to impute GReX in 49 different human tissues.[6,17,25,26] The models were built using multi-variate adaptive shrinkage in R (MASHR)[27] and only include *cis*-eQTLs with MAF >0.01.[28] The number of genes included in each tissue's gene expression prediction model can be found in Table S1. The GTEx models collapse alternative transcripts into gene-level prediction models, meaning what we refer to as the predicted transcript levels for any one gene may include multiple different mRNA products. In each tissue, we tested genetically predicted transcript levels for association with the observed protein levels of all 3,622 plasma proteins measured in the INTERVAL study. We assessed significance via the Benjamini-Hochberg false discovery rate (FDR) method. Within each of the 49 tissues for which we predicted expression, we used the Qvalue R package to calculate q values for all predicted transcript-protein association tests conducted.[29] Transcript-protein pairs with a q value (FDR) <0.05 were considered statistically significant.

For every transcript-protein pair that we found significant (FDR < 0.05) in INTERVAL, we tested the association for replication using genotypes and protein levels from TOPMed MESA if the protein was measured in both studies. We assessed significance of replicating pairs via the Benjamini-Hochberg FDR method. Within each of the 49 tissues for which we predicted expression, we used the Qvalue R package to calculate q values for all predicted transcript-protein association tests conducted in TOPMed.[29] Transcript-protein pairs with a q value (FDR) <0.05 were considered statistically significant.

Calculating the proportion of true positives

The π_0 statistic is the estimated proportion of false positives from a distribution of p values, assuming a uniform distribution of null p values.[29] The q value function from the Qvalue R package calculates the π_0 statistic from a vector of p values.[29] Likewise, the π_1

statistic estimates the proportion of true positives given a distribution of p values and is derived from π_0 as defined below.[29]

$$\pi_1 = 1 - \pi_0$$

We divided the associations we tested in INTERVAL into four categories based on the genomic proximity of the predicted transcript and the target protein: *cis*-acting, *cis*-same, *cis*-different, and *trans*-acting. We defined *cis*-acting relationships as those where the transcription start site of the gene that encodes the predicted transcript was within 1 Mb of the transcription start site of the gene that encodes the target protein. Likewise, *trans*-acting transcript-protein pairs were greater than 1 Mb away from each other or on different chromosomes. We further divided *cis*-acting relationships into *cis*-same, where the gene that encodes the predicted transcript was the same as the gene that encodes the target protein, and *cis*-different, where the predicted transcript and target protein are encoded by different but nearby genes.

For each of these groups in every tissue, we pulled the p values for every tested association and calculated the π_0 statistic using the Qvalue R package. While we used the default q value function parameters in INTERVAL, we adjusted the q value parameters when replicating in TOPMed MESA. Because we only tested pairs that we already found significant in INTERVAL, most of the *cis*-same associations tested in TOPMed MESA returned significant p values, thus the p value distribution in most tissues did not extend all the way to 1. By default, the q value function calculates the average frequency of p values from 0.05 to 1.0 to determine the expected proportion of null p values, so there must be p values throughout this entire range for the function to work. These bounds are controlled by the lambda parameter, which we set from 0.05 to 0.75 instead of the default 0.05 to 1.0 when calculating π_0 in TOPMed MESA. With the estimated π_0 statistic, we calculated the π_1 value for every *cis*/*trans* group in every tissue.

When mapping *trans* effects, there is danger of false negatives when adjusting for potential confounders via PEER factor or PC correction.[30,31] However, failure to remove confounding factors could result in false positive *trans* associations. In a sensitivity analysis, we compared TWAS results without protein PC adjustment to TWAS results also adjusting the INTERVAL protein matrix by 5–40 PCs, which could control for unknown confounders. We observed consistent π_1 statistics across the protein PCs tested and observed consistent counts of significant transcript-protein pair hits (FDR <0.05) for *cis*-same and *cis*-different mechanisms with some variability and no clear trend in the *trans* results (Figure S1). Therefore, we kept the non-protein PC-adjusted results for downstream analyses.

Gene set enrichment analysis of target proteins

We used the web tool, Functional Mapping and Annotation of Genome-Wide Association Studies (FUMA GWAS) to perform a gene set enrichment analysis of all of the protein targets that replicated in TOPMed MESA.[32] We tested the targets involved in *cis*-acting and *trans*-acting associations separately. For both groups, we tested the target proteins for enrichment (FDR <0.05) of GWAS catalog associations[33] and motifs that are known targets of TFs annotated in the Molecular Signatures Database.[34,35]

Identifying pleiotropic regulatory loci

We defined a pleiotropic regulatory gene as one that is significantly associated with the abundance of more than 50 unique protein targets in INTERVAL. We counted the number of significant target proteins for each gene (FDR <0.05) across all 49 tissues in INTERVAL to identify pleiotropic regulatory genes. We grouped pleiotropic regulatory genes whose transcription start sites were within 200 kb of each other into pleiotropic regulatory loci. For each pleiotropic regulatory locus, we quantified the number of unique protein targets of the genes within that locus along with the number of these targets that were tested in TOPMed MESA and the number of these targets with replicated associations with any of the genes in that locus in TOPMed MESA.

Gene set enrichment analysis of pleiotropic regulators

We used FUMA GWAS to perform a gene set enrichment analysis of the pleiotropic regulatory genes as well as their protein targets.[32] We tested the protein targets of each pleiotropic regulatory locus that we discovered in INTERVAL for enrichment (FDR <0.05) of GWAS catalog associations and TF target motifs using all proteins measured in INTERVAL as background. For the pleiotropic regulatory loci with more than one gene, we tested the pleiotropic regulatory genes at that locus for enrichment of GWAS catalog associations and TF target motifs using the union of all genes in each tissue prediction model as background (22,133 genes total).

Cis-same observed expression association analysis

We performed a linear regression analysis to test observed expression levels for association with observed protein levels. RNA-sequencing data are not available in INTERVAL, but they are in TOPMed MESA. In each of these tissues, we leveraged PrediXcan's linear regression association script to test the observed gene expression of each gene measured in TOPMed MESA for association with the observed abundance of the protein product of that gene if it was also measured in TOPMed MESA. We compared these results to the association of the predicted gene expression of each gene with the observed abundance of the protein product of that gene. The number of genes that we tested for *cis*-same associations in each tissue are listed in Table S2.

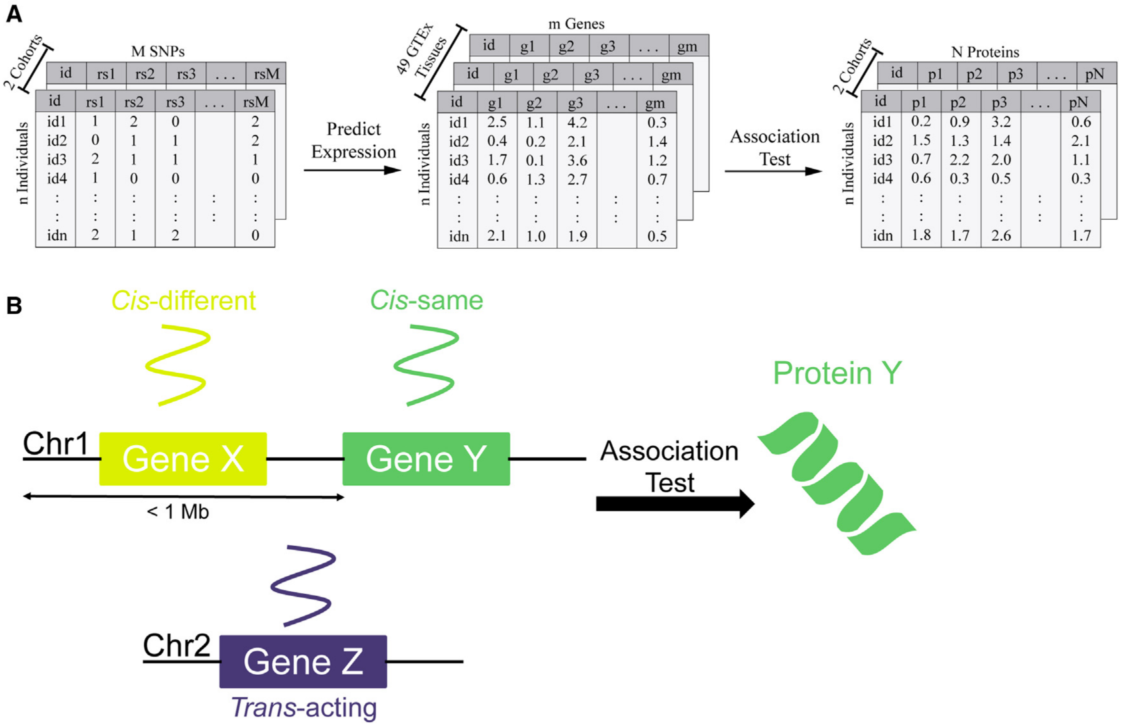
As above, we assessed significance via the Benjamini-Hochberg FDR method. Within each of the 49 tissues with gene expression prediction models, as well as the 3 tissues with observed gene expression data, we calculated q values for all the *cis*-same transcript-protein pairs tested using the Qvalue R package. We further calculated the π_1 statistic for the *cis*-same associations tested in every predicted and observed tissue using the Qvalue R package with a truncated lambda range (0.05–0.75), as described above for TOPMed MESA.

Finally, in every predicted and observed tissue, we calculated the Pearson correlation of gene expression with protein abundance for every gene with a significant *cis*-same association in any tissue. Because some genes were not included in every prediction model and a different set of genes were measured via RNA sequencing, we were not able to calculate the Pearson correlation of expression and protein levels for every gene in every tissue. To summarize results across tissues, we calculated the maximum correlation values between gene expression and protein levels for every gene across all the predicted tissues and across all the observed tissues.

Results

TWAS for proteins identifies replicable gene-protein associations

We sought to identify both *cis*- and *trans*-acting transcriptional regulators of plasma proteins by performing TWAS



id	rs1	rs2	rs3	. . .	rsM
id1	1	2	0		2
id2	0	1	1		2
id3	2	1	1		1
id4	1	0	0		0
:	:	:	:	:	:
:	:	:	:	:	:
idn	2	1	2		0

id	g1	g2	g3	. . .	gm
id1	2.5	1.1	4.2		0.3
id2	0.4	0.2	2.1		1.4
id3	1.7	0.1	3.6		1.2
id4	0.6	1.3	2.7		0.7
:	:	:	:	:	:
:	:	:	:	:	:
idn	2.1	1.0	1.9		0.5

id	p1	p2	p3	. . .	pN
id1	0.2	0.9	3.2		0.6
id2	1.5	1.3	1.4		2.1
id3	0.7	2.2	2.0		1.1
id4	0.6	0.3	0.5		0.3
:	:	:	:	:	:
:	:	:	:	:	:
idn	1.8	1.7	2.6		1.7

Figure 1. TWAS for protein levels

(A) Overview of TWAS analysis. Genotype data from both the INTERVAL and TOPMed MESA cohorts were used to impute genetically regulated expression levels (GReX) in 49 different GTEx tissues. GReX was tested for association with measured plasma protein levels for all proteins tested in both studies.

(B) Model for definition of *cis*-vs. *trans*-acting gene regulators of protein abundance. Here, the prediction model for expression of gene Y and protein Y have a *cis*-same relationship. The expression of gene X and the abundance of protein Y have a *cis*-different relationship because the genes that encode them are different, but their transcription start sites are within 1 Mb of each other. Finally, the expression of gene Z and the abundance of protein Y have a *trans*-acting relationship because the transcription start sites of the genes that encode them are greater than 1 Mb (in this case, the genes are on different chromosomes).

for protein levels. Using the PrediXcan software framework,[17] we tested the genetically regulated component of gene expression (GReX) for association with plasma protein levels. Our discovery set included individuals from the INTERVAL cohort (n = 3,301), and we sought to replicate our findings in the TOPMed MESA cohort (n = 971). For these individuals, we predicted gene expression using previously built prediction models in 49 tissues from the GTEx project (Figure 1A). Then, we calculated the correlation between predicted gene expression and observed protein levels for all 3,622 proteins measured in INTERVAL. We quantified significant transcript-protein pairs as *cis*- (within 1 Mb of each other) or *trans*-acting (greater than 1 Mb apart) relationships. We further divided the *cis*-acting pairs into *cis*-same, where a transcript is associated with the protein that it encodes, and *cis*-different, where a transcript is associated with the protein product of a nearby, different gene (Figure 1B).

We identified 3,699 significant (FDR <0.05) unique *cis*-acting associations for 482 unique proteins (240 *cis*-different and 242 *cis*-same) and 13,598 significant (FDR <0.05) unique *trans*-acting associations for 2,016 unique proteins in INTERVAL (Figures 2A–2D). The TOPMed MESA plasma proteome data included 1,039 proteins that were also measured in INTERVAL. Of the 17,297 significant transcript-protein pairs we discovered in INTERVAL, we tested 8,111 pairs for replication in TOPMed MESA and found 1,168 *cis*-acting pairs replicated (FDR <0.05) for 218 unique proteins (92 *cis*-different and 126 *cis*-same) and 1,210 *trans*-acting pairs replicated for 239 proteins (FDR <0.05, Figures 2B–2D). On average, the significant *cis*-acting relationships we discovered in INTERVAL were shared across more tissues than the significant *trans*-acting relationships we discovered in INTERVAL (Figure 3).

Of the transcript-protein pairs tested in INTERVAL, the *trans*-acting results had the lowest expected true positive rate (π_1) with a median π_1 of 0.004 across all 49 tissues, followed by the *cis*-different results with a median π_1 of 0.099, and the *cis*-same results with a median π_1 of 0.278 (Figure 4; Table S3). We have more confidence in the significant results from INTERVAL that were also tested in TOPMed. The median π_1 value across tissues increased to 0.390 for *trans*-acting relationships, 0.783 for *cis*-different pairs, and 0.888 for *cis*-same pairs (Figure 4; Table S4). Within mechanism categories, π_1 values were largely uncorrelated with the number of transcript-protein pairs tested in each tissue; only *cis*-same π_1 values in INTERVAL correlated with number of tests (Figure S2).

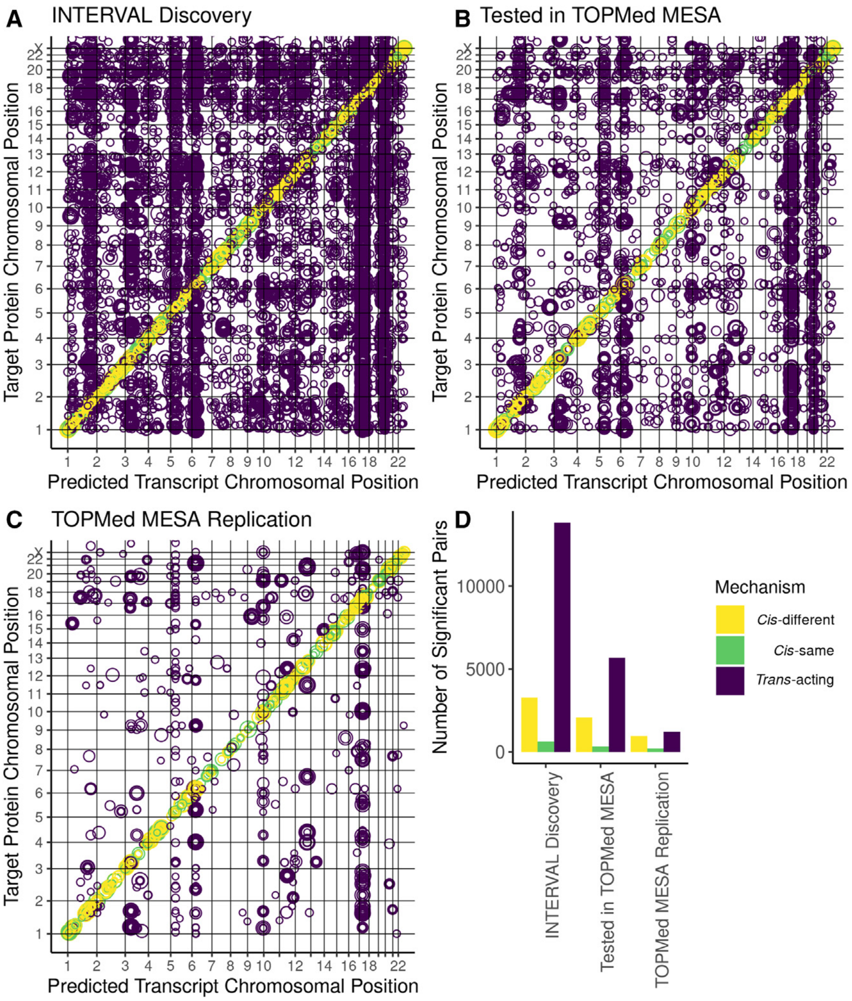


Figure 2. Overview of significant transcript-protein associations

(A–C) Tile plot shows relative genomic position of significantly (FDR <0.05) associated transcript-protein pairs. Each circle represents a uniquely associated predicted transcript and target protein pair. Gridlines delineate chromosomes, and the position along the x axis corresponds to the genomic position of the gene that encodes the predicted transcript, while the position along the y axis corresponds to the genomic position of the gene that encodes the target protein. The size of each circle corresponds to the number of tissues (out of all 49) in which the transcript-protein pair was significantly associated.

(A) Significantly (FDR <0.05) associated transcript-protein pairs discovered in INTERVAL.

(B) Significantly (FDR <0.05) associated transcript-protein pairs discovered in INTERVAL that were also tested in TOPMed MESA.

(C) Significantly (FDR <0.05) associated transcript-protein pairs discovered in INTERVAL that were also significant (FDR <0.05) in TOPMed MESA.

(D) Bar plot of the number of significant (FDR <0.05) associations discovered in INTERVAL, discovered in INTERVAL and tested in TOPMed MESA, and discovered in INTERVAL and significantly (FDR <0.05) replicated in TOPMed MESA.

Protein targets of *trans*-acting genes enriched for transcription factor target motifs and GWAS catalog phenotypes

We first tested the protein targets that replicated in TOPMed MESA, divided into targets of *cis*-acting genes and targets of *trans*-acting genes, for enrichment of motifs targeted by TFs. While the *cis*-targets were not enriched for TF targets, the *trans*-targets were enriched for motifs targeted by the TFs NFKB2, RELA, NFAT1C, FOXF2, AR, GATA1, and STAT1 (Figure 5; Table S5).

Furthermore, we tested the *cis*- and *trans*-targets for enrichment of GWAS catalog associations and found that the *trans*-targets were enriched for blood protein levels and inflammatory bowel disease, and the *cis*-targets were enriched for blood protein levels, ankylosing spondylitis, inflammatory bowel disease, and chronic inflammatory diseases (Table S6).

Pleiotropic regulatory regions enriched for TF target motifs and GWAS catalog phenotypes

By quantifying the number of target proteins that each transcript was significantly associated with, we identified several loci that may be involved in the regulation of many different proteins throughout the genome, which we have named "pleiotropic regulatory" loci. Here, we defined a pleiotropic gene as one with more than 50 unique protein targets in INTERVAL. We grouped pleiotropic genes whose transcription start sites are within 200 kb of each other into pleiotropic loci. These loci are represented through the vertical lines of dots in Figures 2A–2C. We discovered 11 distinct pleiotropic regulatory loci in INTERVAL (Table S7). While most of the loci did not have many targets that replicated in TOPMed MESA, there were a few that replicated well, including the *C7* locus on chromosome 5, the *SKIV2L* locus on chromosome 6, the *ABO* locus on chromosome 9, and the *SARM1* locus on chromosome 17 (Table S7). Only one of the 218 tested targets of the largest pleiotropic regulatory locus discovered in INTERVAL, the *MYADM* locus on chromosome 19, replicated in TOPMed MESA (Table S7).

We performed a gene set enrichment analysis of the protein targets in INTERVAL of each of these pleiotropic regulatory loci. For most of the loci, we found no significant enrichment of TF targets or GWAS catalog associations in the target proteins. However, we found that the target proteins of the *ABO* locus were enriched (FDR <0.05) for associations with blood protein levels in the GWAS catalog. Furthermore, we found that the target proteins of the *C7* locus were enriched (p value: 6.58e-5; adjusted p: 4.02e-2) for a motif (MSigDB: M18461) that is targeted

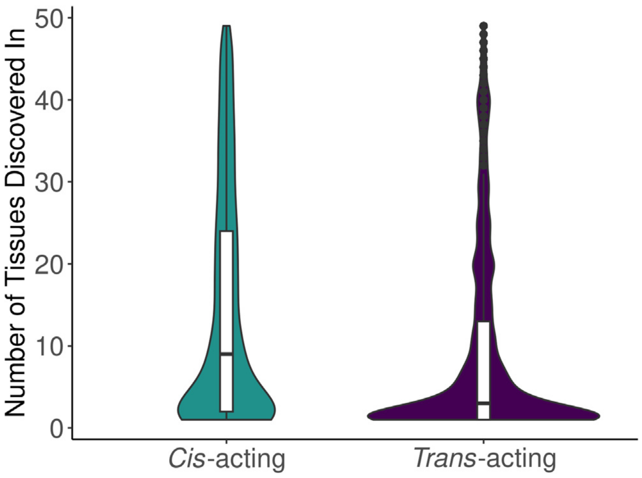


Figure 3. Sharing of *cis*- and *trans*-acting effects across tissues in INTERVAL
Distributions of the number of tissues in which each significant transcript-protein pair was discovered (FDR <0.05), divided into *cis*- and *trans*-acting associations.

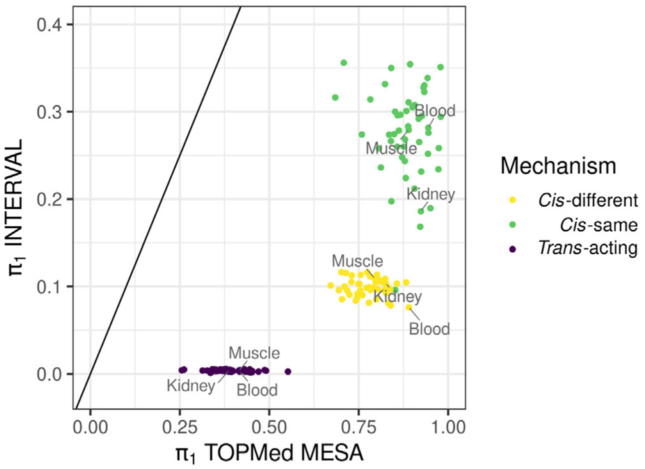


Figure 4. Expected true positive rates (π_1) for transcript-protein pairs across tissues
Discovery π_1 values across tissues of transcript-protein pairs tested in INTERVAL are compared to replication π_1 values in TOPMed MESA. Only significant (FDR <0.05) transcript-protein pairs in INTERVAL were tested in TOPMed MESA. Associations were divided into *cis*-same, *cis*-different, and *trans*-acting, and π_1 was calculated in every GTEx tissue separately. Tissues with the most samples in GTEx (muscle-skeletal, n = 706, and whole blood, n = 670) and the least samples in GTEx (kidney-cortex, n = 73) are labeled. The diagonal line is the identity line (y intercept = 0, slope = 1).

by the TF ARNT. Of the 271 genes in the gene set, we tested 42 in our TWAS, and 13 were targets of the *C7* locus. While *ARNT* had gene expression prediction models in many tissues, it was not significantly associated with any of the targets of the *C7* locus in our TWAS analysis.

Additionally, we performed a gene set enrichment analysis of the pleiotropic regulatory genes involved in each locus that comprised of more than one gene. Four of five loci tested were enriched for some GWAS catalog associations (Table S8). The *HLA* locus was enriched for 52 GWAS catalog associations, including a wide variety of immune-related diseases and conditions like neuromyelitis, lymphoma, pneumonia, and more. Only one locus was enriched (FDR <0.05) for TF targets; the *SARM1* locus on chromosome 17 was enriched (FDR <0.05) for a motif (MSigDB: M826) targeted by the TF, SREBF1. Of the 174 genes in this gene set, we tested 153 in our TWAS, and three were pleiotropic regulators at this locus: *POLDIP2*, *TMEM199*, and *SUPT6H*. While *SREBF1* had prediction models in many tissues, it was not significantly associated with any target proteins in our TWAS analysis.

Predicted gene expression correlates better with protein levels than observed gene expression

We used the RNA-sequencing data from TOPMed MESA to test how the correlation of observed gene expression with observed protein abundance compared to that of predicted gene expression with observed protein abundance. For each of the three tissues with observed gene expression data (PBMC, monocytes, and T cells), we tested the abundance of all 1,300 proteins measured for association with the observed expression of the genes that encode the proteins (*cis*-same gene-protein relationship). We compared these observed expression results to *cis*-same TWAS results using the GTEx prediction models. We discovered more genes with significant associations between predicted expression and observed protein levels (FDR <0.05) than genes with significant associations between observed gene expression and observed protein levels (FDR <0.05). In total, we discovered 407 genes with a significant *cis*-same association across all 49 predicted tissues and 121 genes with a significant *cis*-same association across all three measured tissues. We found a significant *cis*-same association with both predicted and observed expression for 89 genes, while the rest were unique associations (Figure 6A).

Furthermore, the proportion of true positive *cis*-same associations (π_1) was on average higher across predicted tissues than observed tissues (Figure 6B). The observed tissue with the highest π_1 value was PBMC at 0.239, followed by monocytes at 0.193, and T cells at 0.077. Likewise, all but one predicted tissue had a higher π_1 than the observed tissues (Table S9). Notably, whole blood, the closest predicted tissue to the observed tissues, had a higher π_1 than all three of the observed tissues at 0.331.

Finally, we wanted to see if the correlation of predicted expression and protein abundance was stronger than the correlation of observed gene expression and protein abundance. For the union of genes whose expression, predicted or observed, was significantly (FDR <0.05) associated with protein abundance, we calculated the Pearson correlation of expression and protein levels in every tissue where there was a measurement for both traits. When looking at the maximum correlation values across the predicted and observed tissues separately, we found that GReX on average had a stronger correlation with protein abundance

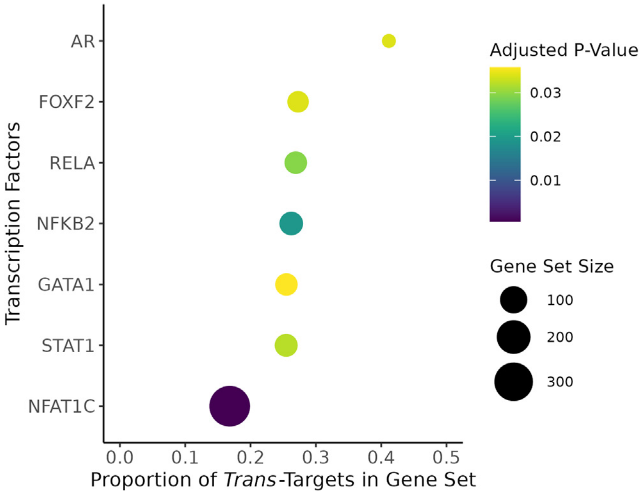


Figure 5. Enrichment of TF binding sites of target proteins of *trans*-acting genes

The target proteins of *trans*-acting genes were significantly enriched for binding motifs of the TFs listed on the y axis as annotated in the Molecular Signatures Database.[34,35] The size of each bubble corresponds to the number of genes annotated in the database that we tested in our TWAS analysis and the x axis represents the proportion of those genes whose protein products were significantly associated with a *trans*-acting gene in INTERVAL. The color of each bubble represents the adjusted p value (Benjamini-Hochberg) of the enrichment test.

than observed gene expression for significant *cis*-same genes (Figures 6C and 6D). We found that predicted tissues closely related to blood plasma, such as whole blood and liver, ranked high in terms of median correlation of expression levels and protein levels by gene, while most of the brain tissues had the lowest median correlation of expression levels and protein levels (Figure 7). While median correlation significantly associated with the number of *cis*-same genes tested ($R^2 = 0.27$, $p = 0.00011$), we note that whole blood and liver both had higher correlations than expected given the number of genes tested (Figure S3).

Discussion

Here, we applied the TWAS framework to test genetically regulated gene expression for association with measured plasma protein levels in order to discover gene regulatory relationships between both distant (*trans*-acting) and nearby (*cis*-acting) genes. Similar to a prior study, which applied *trans*-PrediXcan to test genetically regulated gene expression for association with observed expression levels, our approach proved more effective at identifying *trans*-acting effects than a typical QTL study.[9] Compared to a *trans*-pQTL study performed in our discovery cohort (INTERVAL), which found 1,104 proteins with *trans*-pQTL[14] ($p < 1.5 \times 10^{-11}$), our method discovered 2,016 protein targets of *trans*-acting genes, 239 of which replicated in the much smaller TOPMed MESA cohort. Methods like TWAS, which prioritize *cis*-eQTL, have been shown to be more effective at discovering *trans*-acting effects because often *trans*-eQTL act through *cis*-mediators like nearby TF genes.[10] We found that the protein targets of *trans*-acting genes were enriched for TF binding sites, while the *cis*-targets were not, supporting the idea that many *trans*-effects are driven by TF genes. Furthermore, we found that the *cis*-acting associations were shared across more tissues than the *trans*-acting effects, which tended to be more tissue specific, as has been shown in previous eQTL studies.[3,7]

We identified several loci throughout the genome with strong pleiotropic effects where one gene, or several in linkage disequilibrium, significantly (FDR <0.05) associated with many protein targets throughout the genome. Many of these loci have been identified before, including the *ABO*, *VTN*, *APOE*, *CFH*, and *BCHE* loci.[14,36–39] Here, we called these regions pleiotropic regulatory loci and discovered 11 in INTERVAL and 5 that replicated in TOPMed MESA. It has been shown previously that these *trans*-acting pleiotropic regulator genes are enriched for GWAS traits, suggesting that *trans*-protein regulation plays an important role in disease variation.[9,38] We performed a gene set enrichment analysis of all of the *trans*-acting genes in each of these pleiotropic regulatory loci as well as the target proteins of each of these pleiotropic regulatory loci. We found that the targets and pleiotropic regulatory genes of many of these loci were enriched for GWAS catalog associations including several autoimmune diseases and other disease phenotypes. Autoimmune disease enrichment is somewhat expected given the proteins in our TWAS were measured in blood plasma. For example, the genes at the *CFHR* locus were enriched for autoimmune diseases such as IgA nephropathy and age-related macular degeneration, as well as C3 and C4 levels. CFHR genes interact with proteins like C3 and C4 in the complement system, a cascade of proteins important to the immune response system, thus changes in expression of these pleiotropic regulatory genes could lead to the progression of autoimmune diseases.[40]

We found that our significant results discovered in INTERVAL had a low expected proportion of true positives (π_1) across all associations tested, though we have more confidence in the *cis*-acting results than *trans*-acting. This is a symptom of an ongoing issue with identifying *trans*-acting effects; the multiple testing burden is too high due to the high number of associations that must be tested combined with the observation that *trans*-acting effects are generally smaller than *cis*-acting effects.[3,6,41,42] Nevertheless, we replicated many of our significant associations discovered in INTERVAL in TOPMed MESA, where we found much higher proportions of true positives across all associations tested. In many tissues, we estimated a π_1 of nearly 1.0 for the *cis*-same results, indicating a strong correlation between genetically regulated gene expression levels and observed protein levels. This is in contrast with many studies that have shown a poor correlation between transcript and protein levels of the same underlying gene.[43–46] One of the main issues in correlating expression levels with protein levels is the high fluctuation in these

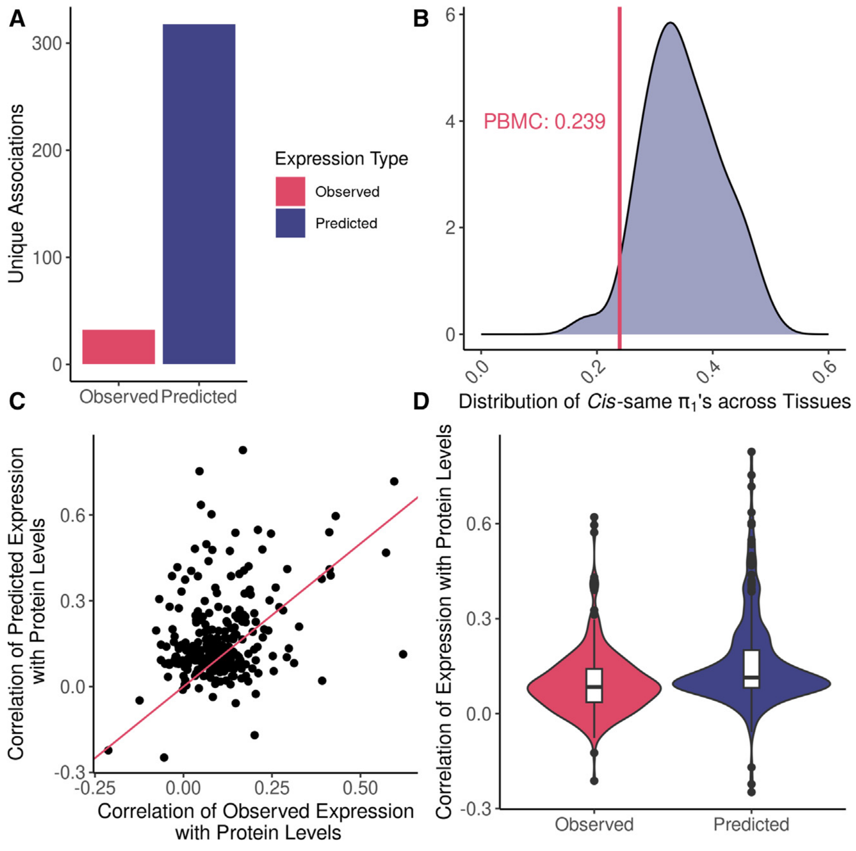


Figure 6. *Cis*-same associations using predicted expression vs. observed expression
(A) Number of unique genes with a *cis*-same correlation between expression levels (divided by predicted and observed) and protein abundance.
(B) Distribution of proportion of true positives (π_1 values) from tests conducted in all predicted tissues. The vertical red line indicates the tissue with observed gene expression that had the highest π_1; PBMC at 0.239.
(C) Scatterplot comparing the maximum correlation of predicted and observed expression with protein abundance by gene.
(D) Distribution of maximum Pearson correlation coefficients for correlating expression, predicted or observed, of significant *cis*-same genes with protein abundance.

traits due to environmental influence; it has been shown that proteins that can be more reproducibly measured, meaning they are less prone to environmental variation, have a stronger correlation with expression levels.[47] Furthermore, genetically predicted expression levels have been shown to strongly correlate with genetically predicted protein levels.[48]

Here, we show that genetically predicted expression levels correlate better with plasma protein abundance than observed expression levels. This indicates the genetically regulated (heritable) component of gene expression and protein abundance is more consistent across tissues than the non-genetic, i.e., environmental, components. We leveraged the TWAS framework to test both predicted expression in 49 tissues and observed expression in three tissues for association with plasma protein levels in individuals from the TOPMed MESA cohort. Most of the unique associations we discovered with observed expression were also significant when using predicted expression, and we found many unique associations with predicted expression that we did not with observed expression. Furthermore, we estimated a higher proportion of true positives for our predicted expression results. Even in a tissue-matched scenario (comparing predicted expression in whole blood to observed expression in PBMC), we found a higher proportion of true positive results for predicted expression. Additionally, we found that the Pearson correlation of expression levels with proteins levels of the same underlying gene was on average higher when working with predicted expression than observed expression. We found that tissues that are closely related to blood, like whole blood and liver, which is responsible for secreting many plasma proteins into the bloodstream, had a higher correlation of predicted expression levels and protein levels, which has been shown previously in another cohort.[48] Furthermore, the brain tissues tended to have the lowest correlation of expression levels and plasma protein levels, perhaps because of the blood-brain barrier, as has been suggested previously.[48]

A limitation of the study is that our discovery cohort is not ancestrally diverse, comprising entirely of individuals of European descent, while our replication cohort, which is diverse, has a small sample size. Another limitation of this study is the type of proteomic data we used. Our study was not truly proteome wide, as we could only test the proteins measured by the targeted proteome assay. As such, there are likely many regulatory relationships that we were not able to capture due to the limited number of proteins measured in both the INTERVAL and TOPMed study. Furthermore, we only have proteomic data for plasma proteins when, like gene expression levels, protein levels vary across tissues and cell types. Additionally, the aptamers on the SOMAscan assays used to target specific proteins are known to sometimes have multiple targets, so some of our protein level measurements may represent the abundance of multiple different proteins.[19] All protein assays that rely on binding could be affected by protein altering variants in the aptamer binding site. However, integrating proteomic data with RNA-sequencing transcriptome data alleviates some of these concerns. We note that just 120 of the 3,339 (3.6%) INTERVAL proteins and zero of the 1,335 (0%) TOPMed MESA proteins had protein-altering variants, defined by the Ensembl Variant Effect Predictor,[49] in their respective GTEx whole blood transcript prediction models.

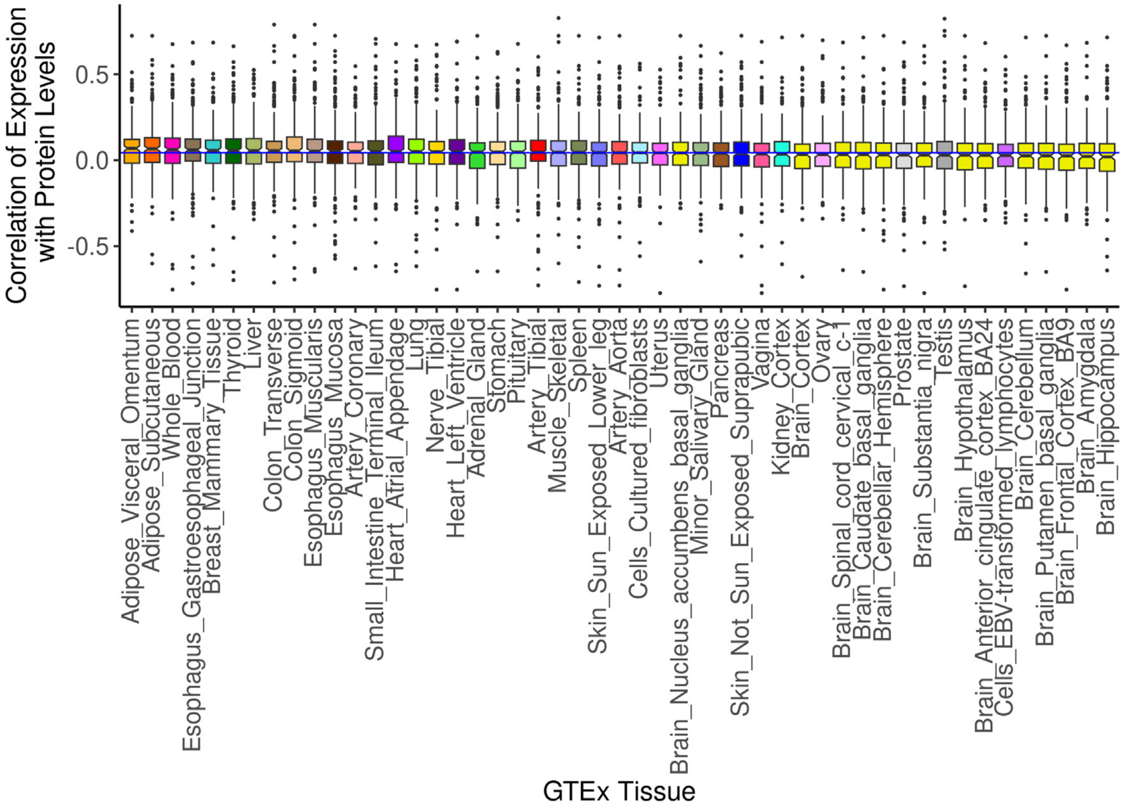


Figure 7. *Cis*-same correlation of predicted expression and protein levels by tissue
Distribution of Pearson correlation coefficients for correlating predicted expression of significant *cis*-same genes with protein abundance in every GTEx tissue. The horizontal blue line indicates the median correlation across all tissues.

Our results highlight the benefits of working with predicted expression over observed expression. First, it is easier to calculate predicted expression than it is to measure observed expression since many more studies have genome-wide genotypes than gene expression data. Also, using the *cis*-acting genetically regulated (heritable) component of gene expression to discover *trans*-acting gene effects on protein abundance finds more significant associations than traditional SNP-based pQTL studies. Most importantly, because this heritable component of gene expression more strongly correlates with protein levels than total observed expression, predicted expression is useful in uncovering the function of SNPs associated with complex traits.

Data and code availability

Full summary statistics for all association analyses performed and code for presented results are available at https://github.com/hwittich/TWAS_for_protein. Data from INTERVAL is under controlled access via the European Genome-phenome Archive at https://ega-archive.org/ for both genotypes (EGA: EGAD00010001544) and blood plasma aptamers levels as measured by a SOMAscan assay (EGA: EGAD00001004080). TOPMed MESA data are under controlled access in dbGaP at https://www.ncbi.nlm.nih.gov/gap/. Genotypes are available through accession dbGaP: phs000420.v6.p3 and RNA-sequencing and proteome data are available through accession dbGaP: phs001416.v2.p1.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2024.01.006.

Acknowledgments

This work is supported by the NIH National Human Genome Research Institute Academic Research Enhancement Award R15 HG009569 (HEW). INTERVAL and TOPMed MESA grants are detailed and acknowledged in the supplemental information.

Declaration of interests

The authors declare no competing interests.

Received: September 25, 2023
Accepted: January 12, 2024
Published: February 5, 2024

References

1. Hormozdiari, F., Gazal, S., van de Geijn, B., Finucane, H.K., Ju, C.J.-T., Loh, P.-R., Schoech, A., Reshef, Y., Liu, X., O'Connor, L., et al. (2018). Leveraging molecular quantitative trait loci

to understand the genetic architecture of diseases and complex traits. Nat. Genet. *50*, 1041–1047.

2. Strunz, T., Grassmann, F., Gayán, J., Nahkuri, S., Souza-Costa, D., Maugeais, C., Fauser, S., Nogoceke, E., and Weber, B.H.F. (2018). A mega-analysis of expression quantitative trait loci (eQTL) provides insight into the regulatory architecture of gene expression variation in liver. Sci. Rep. *8*, 5865.
3. Liu, X., Finucane, H.K., Gusev, A., Bhatia, G., Gazal, S., O'Connor, L., Bulik-Sullivan, B., Wright, F.A., Sullivan, P.F., Neale, B.M., and Price, A.L. (2017). Functional Architectures of Local and Distal Regulation of Gene Expression in Multiple Human Tissues. Am. J. Hum. Genet. *100*, 605–616.
4. Võsa, U., Claringbould, A., Westra, H.-J., Bonder, M.J., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Yazar, S., et al. (2021). Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. Nat. Genet. *53*, 1300–1310.
5. Liu, X., Li, Y.I., and Pritchard, J.K. (2019). Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. Cell *177*, 1022–1034.e6.
6. GTEx Consortium (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science *369*, 1318–1330.
7. Grundberg, E., Small, K.S., Hedman, Å.K., Nica, A.C., Buil, A., Keildson, S., Bell, J.T., Yang, T.-P., Meduri, E., Barrett, A., et al. (2012). Mapping cis- and trans-regulatory effects across multiple tissues in twins. Nat. Genet. *44*, 1084–1089.
8. Ouwens, K.G., Jansen, R., Nivard, M.G., van Dongen, J., Frieser, M.J., Hottenga, J.-J., Arindrarto, W., Claringbould, A., van Iterson, M., Mei, H., et al. (2020). A characterization of cis- and trans-heritability of RNA-Seq-based gene expression. Eur. J. Hum. Genet. *28*, 253–263.
9. Wheeler, H.E., Ploch, S., Barbeira, A.N., Bonazzola, R., Andaleon, A., Fotuhi Siahpirani, A., Saha, A., Battle, A., Roy, S., and Im, H.K. (2019). Imputed gene associations identify replicable trans-acting genes enriched in transcription pathways and complex traits. Genet. Epidemiol. *43*, 596–608.
10. Yang, F., Wang, J., Pierce, B.L., Chen, L.S.; and GTEx Consortium (2017). Identifying cis-mediators for trans-eQTLs across many human tissues using genomic mediation analysis. Genome Res. *27*, 1859–1871.
11. Dutta, D., He, Y., Saha, A., Arvanitis, M., Battle, A., and Chatterjee, N. (2022). Aggregative trans-eQTL analysis detects trait-specific target gene sets in whole blood. Nat. Commun. *13*, 4323.
12. Banerjee, S., Simonetti, F.L., Detrois, K.E., Kaphle, A., Mitra, R., Nagial, R., and Söding, J. (2021). Tejaas: reverse regression increases power for detecting trans-eQTLs. Genome Biol. *22*, 142.
13. Liu, X., Mefford, J.A., Dahl, A., He, Y., Subramaniam, M., Battle, A., Price, A.L., and Zaitlen, N. (2020). GBAT: a gene-based association test for robust detection of trans-gene regulation. Genome Biol. *21*, 211.
14. Sun, B.B., Maranville, J.C., Peters, J.E., Stacey, D., Staley, J.R., Blackshaw, J., Burgess, S., Jiang, T., Paige, E., Surendran, P., et al. (2018). Genomic atlas of the human plasma proteome. Nature *558*, 73–79.
15. Gold, L., Ayers, D., Bertino, J., Bock, C., Bock, A., Brody, E.N., Carter, J., Dalby, A.B., Eaton, B.E., Fitzwater, T., et al. (2010). Aptamer-Based Multiplexed Proteomic Technology for Biomarker Discovery. PLoS One *5*, e15004.
16. Yang, C., Farias, F.H.G., Ibanez, L., Suhy, A., Sadler, B., Fernandez, M.V., Wang, F., Bradley, J.L., Eiffert, B., Bahena, J.A., et al. (2021). Genomic atlas of the proteome from brain, CSF and plasma prioritizes proteins implicated in neurological disorders. Nat. Neurosci. *24*, 1302–1312.
17. Gamazon, E.R., Wheeler, H.E., Shah, K.P., Mozaffari, S.V., Aquino-Michaels, K., Carroll, R.J., Eyler, A.E., Denny, J.C., Nicolae, D.L., et al.; GTEx Consortium (2015). A gene-based association method for mapping traits using reference transcriptome data. Nat. Genet. *47*, 1091–1098.
18. Di Angelantonio, E., Thompson, S.G., Kaptoge, S., Moore, C., Walker, M., Armitage, J., Ouwehand, W.H., Roberts, D.J., Danesh, J., et al.; INTERVAL Trial Group (2017). Efficiency and safety of varying the frequency of whole blood donation (INTERVAL): a randomised trial of 45 000 donors. Lancet *390*, 2360–2371.
19. Schubert, R., Geoffroy, E., Gregga, I., Mulford, A.J., Aguet, F., Ardlie, K., Gerszten, R., Clish, C., Van Den Berg, D., Taylor, K.D., et al. (2022). Protein prediction for trait mapping in diverse populations. PLoS One *17*, e0264341.
20. Rohloff, J.C., Gelinas, A.D., Jarvis, T.C., Ochsner, U.A., Schneider, D.J., Gold, L., and Janjic, N. (2014). Nucleic Acid Ligands With Protein-like Side Chains: Modified Aptamers and Their Use as Diagnostic and Therapeutic Agents. Mol. Ther. Nucleic Acids *3*, e201.
21. Taliun, D., Harris, D.N., Kessler, M.D., Carlson, J., Szpiech, Z.A., Torres, R., Taliun, S.A.G., Corvelo, A., Gogarten, S.M., Kang, H.M., et al. (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature *590*, 290–299.
22. Bild, D.E., Bluemke, D.A., Burke, G.L., Detrano, R., Diez Roux, A.V., Folsom, A.R., Greenland, P., Jacob, D.R., Jr., Kronmal, R., Liu, K., et al. (2002). Multi-Ethnic Study of Atherosclerosis: Objectives and Design. Am. J. Epidemiol. *156*, 871–881.
23. Mogil, L.S., Andaleon, A., Badalamenti, A., Dickinson, S.P., Guo, X., Rotter, J.I., Johnson, W.C., Im, H.K., Liu, Y., and Wheeler, H.E. (2018). Genetic architecture of gene expression traits across diverse populations. PLoS Genet. *14*, e1007586.
24. Araujo, D.S., Nguyen, C., Hu, X., Mikhaylova, A.V., Gignoux, C., Ardlie, K., Taylor, K.D., Durda, P., Liu, Y., Papanicolaou, G., et al. (2023). Multivariate adaptive shrinkage improves cross-population transcriptome prediction and association studies in underrepresented populations. HGG Adv. *4*, 100216.
25. Barbeira, A.N., Bonazzola, R., Gamazon, E.R., Liang, Y., Park, Y., Kim-Hellmuth, S., Wang, G., Jiang, Z., Zhou, D., Hormozdiari, F., et al. (2021). Exploiting the GTEx resources to decipher the mechanisms at GWAS loci. Genome Biol. *22*, 49.
26. Barbeira, A.N., Dickinson, S.P., Bonazzola, R., Zheng, J., Wheeler, H.E., Torres, J.M., Torstenson, E.S., Shah, K.P., Garcia, T., Edwards, T.L., et al. (2018). Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nat. Commun. *9*, 1825.
27. Urbut, S.M., Wang, G., Carbonetto, P., and Stephens, M. (2019). Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. Nat. Genet. *51*, 187–195.
28. Barbeira, A.N., Melia, O.J., Liang, Y., Bonazzola, R., Wang, G., Wheeler, H.E., Aguet, F., Ardlie, K.G., Wen, X., and Im, H.K. (2020). Fine-mapping and QTL tissue-sharing information improves the reliability of causal gene identification. Genet. Epidemiol. *44*, 854–867.

29. Storey, J.D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. Proc. Natl. Acad. Sci. USA *100*, 9440–9445.
30. GTEx Consortium (2017). Genetic effects on gene expression across human tissues. Nature *550*, 204–213.
31. Kang, H.M., Ye, C., and Eskin, E. (2008). Accurate Discovery of Expression Quantitative Trait Loci Under Confounding From Spurious and Genuine Regulatory Hotspots. Genetics *180*, 1909–1925.
32. Watanabe, K., Taskesen, E., van Bochoven, A., and Posthuma, D. (2017). Functional mapping and annotation of genetic associations with FUMA. Nat. Commun. *8*, 1826.
33. Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. *47*, D1005–D1012.
34. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA *102*, 15545–15550.
35. Mootha, V.K., Lindgren, C.M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E., et al. (2003). PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat. Genet. *34*, 267–273.
36. Suhre, K., Arnold, M., Bhagwat, A.M., Cotton, R.J., Engelke, R., Raffler, J., Sarwath, H., Thareja, G., Wahl, A., DeLisle, R.K., et al. (2017). Connecting genetic risk to disease end points through the human blood plasma proteome. Nat. Commun. *8*, 14357.
37. Gudjonsson, A., Gudmundsdottir, V., Axelsson, G.T., Gudmundsson, E.F., Jonsson, B.G., Launer, L.J., Lamb, J.R., Jennings, L.L., Aspelund, T., Emilsson, V., and Gudnason, V. (2022). A genome-wide association study of serum proteins reveals shared loci with common diseases. Nat. Commun. *13*, 480.
38. Emilsson, V., Ilkov, M., Lamb, J.R., Finkel, N., Gudmundsson, E.F., Pitts, R., Hoover, H., Gudmundsdottir, V., Horman, S.R., Aspelund, T., et al. (2018). Co-regulatory networks of human serum proteins link genetics to disease. Science *361*, 769–773.
39. Pietzner, M., Wheeler, E., Carrasco-Zanini, J., Raffler, J., Kerrison, N.D., Oerton, E., Auyeung, V.P.W., Luan, J., Finan, C., Casas, J.P., et al. (2020). Genetic architecture of host proteins involved in SARS-CoV-2 infection. Nat. Commun. *11*, 6397.
40. Zipfel, P.F., Wiech, T., Stea, E.D., and Skerka, C. (2020). CFHR Gene Variations Provide Insights in the Pathogenesis of the Kidney Diseases Atypical Hemolytic Uremic Syndrome and C3 Glomerulopathy. J. Am. Soc. Nephrol. *31*, 241–256.
41. Fairfax, B.P., Makino, S., Radhakrishnan, J., Plant, K., Leslie, S., Dilthey, A., Ellis, P., Langford, C., Vannberg, F.O., and Knight, J.C. (2012). GENETICS OF GENE EXPRESSION IN PRIMARY IMMUNE CELLS IDENTIFIES CELL-SPECIFIC MASTER REGULATORS AND ROLES OF HLA ALLELES. Nat. Genet. *44*, 502–510.
42. Dimas, A.S., Deutsch, S., Stranger, B.E., Montgomery, S.B., Borel, C., Attar-Cohen, H., Ingle, C., Beazley, C., Gutierrez Arcelus, M., Sekowska, M., et al. (2009). Common regulatory variation impacts gene expression in a cell type dependent manner. Science *325*, 1246–1250.
43. Gygi, S.P., Rochon, Y., Franza, B.R., and Aebersold, R. (1999). Correlation between Protein and mRNA Abundance in Yeast. Mol. Cell Biol. *19*, 1720–1730.
44. Marguerat, S., Schmidt, A., Codlin, S., Chen, W., Aebersold, R., and Bähler, J. (2012). Quantitative Analysis of Fission Yeast Transcriptomes and Proteomes in Proliferating and Quiescent Cells. Cell *151*, 671–683.
45. Cheng, P., Zhao, X., Katsnelson, L., Camacho-Hernandez, E.M., Mermerian, A., Mays, J.C., Lippman, S.M., Rosales-Alvarez, R.E., Moya, R., Shwetar, J., et al. (2022). Proteogenomic analysis of cancer aneuploidy and normal tissues reveals divergent modes of gene regulation across cellular pathways. Elife *11*, e75227.
46. Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011). Global quantification of mammalian gene expression control. Nature *473*, 337–342.
47. Upadhya, S.R., and Ryan, C.J. (2022). Experimental reproducibility limits the correlation between mRNA and protein abundances in tumor proteomic profiles. Cell Rep. Methods *2*, 100288.
48. Zhang, J., Dutta, D., Köttgen, A., Tin, A., Schlosser, P., Grams, M.E., Harvey, B., Yu, B., Boerwinkle, E., et al.; CKDGen Consortium (2022). Plasma proteome analyses in individuals of European and African ancestry identify cis-pQTLs and models for proteome-wide association studies. Nat. Genet. *54*, 593–602.
49. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biol. *17*, 122.

The American Journal of Human Genetics, Volume *111*

Supplemental information

Transcriptome-wide association study of the plasma proteome reveals *cis* and *trans* regulatory mechanisms underlying complex traits

Henry Wittich, Kristin Ardlie, Kent D. Taylor, Peter Durda, Yongmei Liu, Anna Mikhaylova, Chris R. Gignoux, Michael H. Cho, Stephen S. Rich, Jerome I. Rotter, NHLBI TOPMed Consortium, Ani Manichaikul, Hae Kyung Im, and Heather E. Wheeler

Supplemental Figures

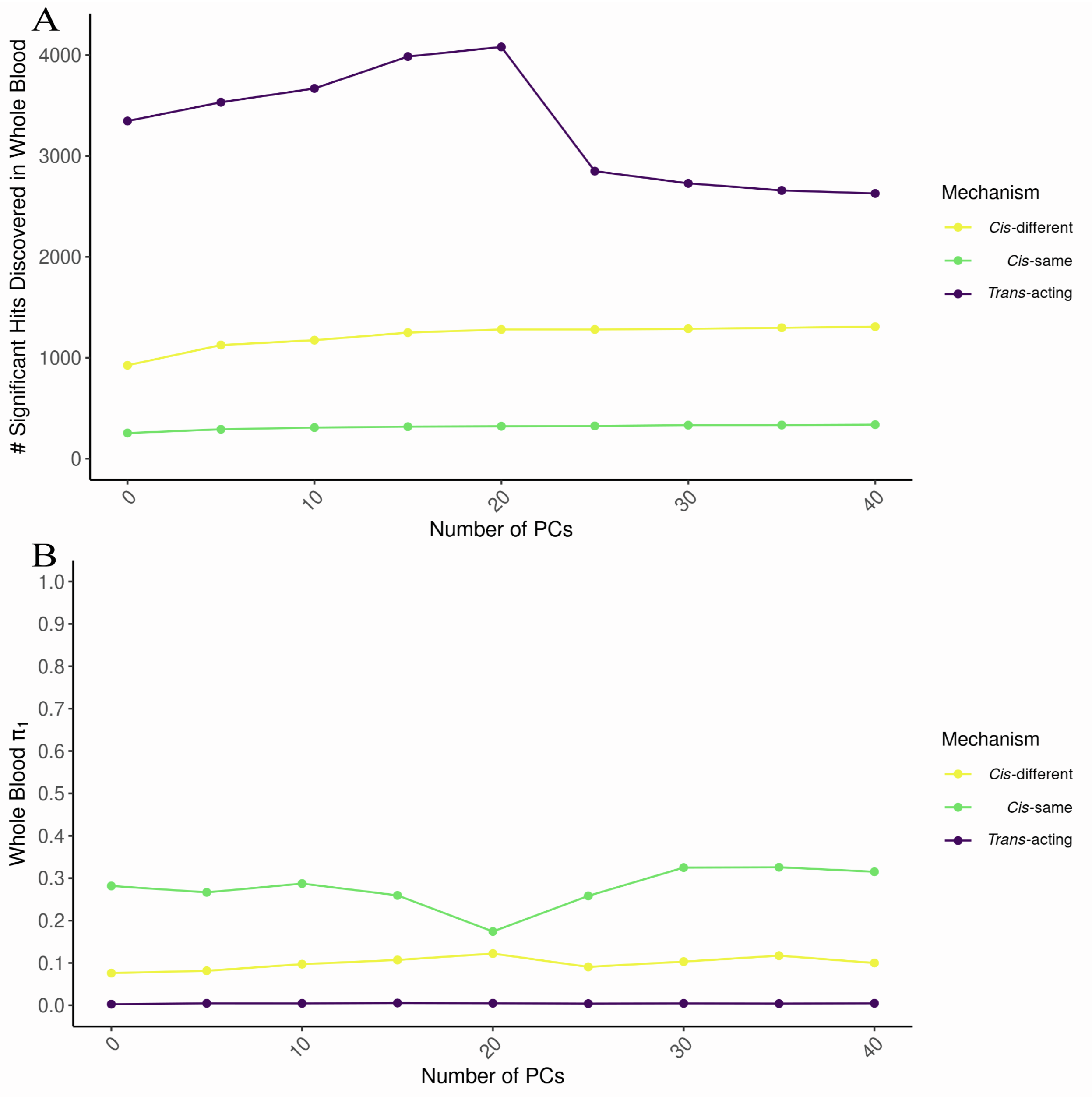


Figure S1. TWAS results in INTERVAL using GTEx whole blood transcriptome models across protein principal components (PCs). (A) Counts of transcript-protein pairs that associate in each mechanism category with FDR<0.05. The protein matrix was adjusted for 0-40 PCs in 5 PC increments, taking the residuals as the adjusted protein levels in the association tests. (B) Expected true positive rates (π_1) across protein PCs.

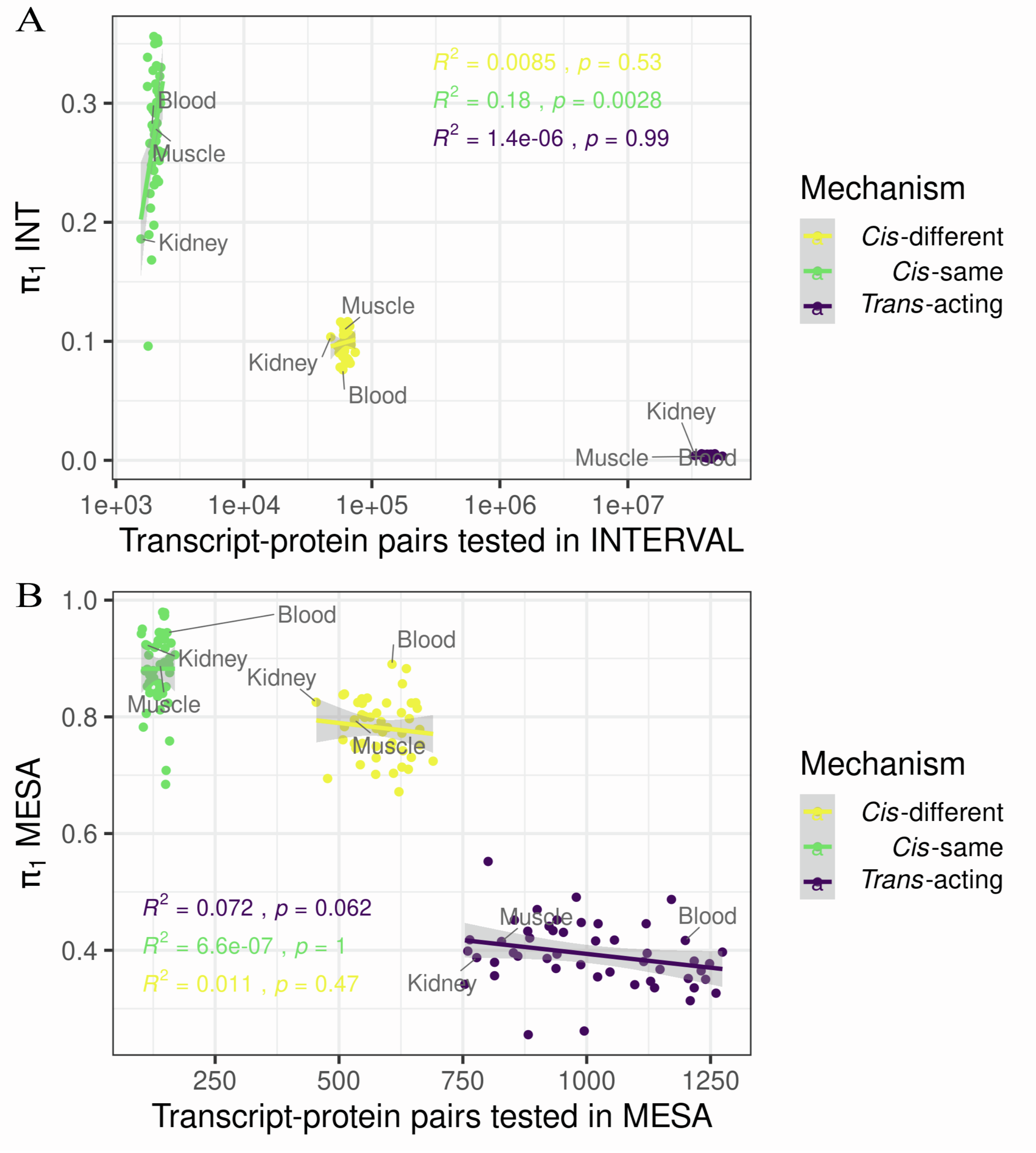


Figure S2. Comparison of π_1 expected true positive rates to the number of predicted transcript-protein pairs tested in each GTEx tissue trascriptome prediction model. (A) Linear regression results in INTERVAL colored by each mechanism class. (B) Linear regression results in MESA colored by each mechanism class. Note only transcript-protein pairs with FDR<0.05 in INTERVAL were tested in TOPMed MESA. Tissues with the most samples in GTEx (Muscle – Skeletal, n=706 and Whole Blood, n=670) and the least samples in GTEx (Kidney – Cortex, n=73) are labeled.

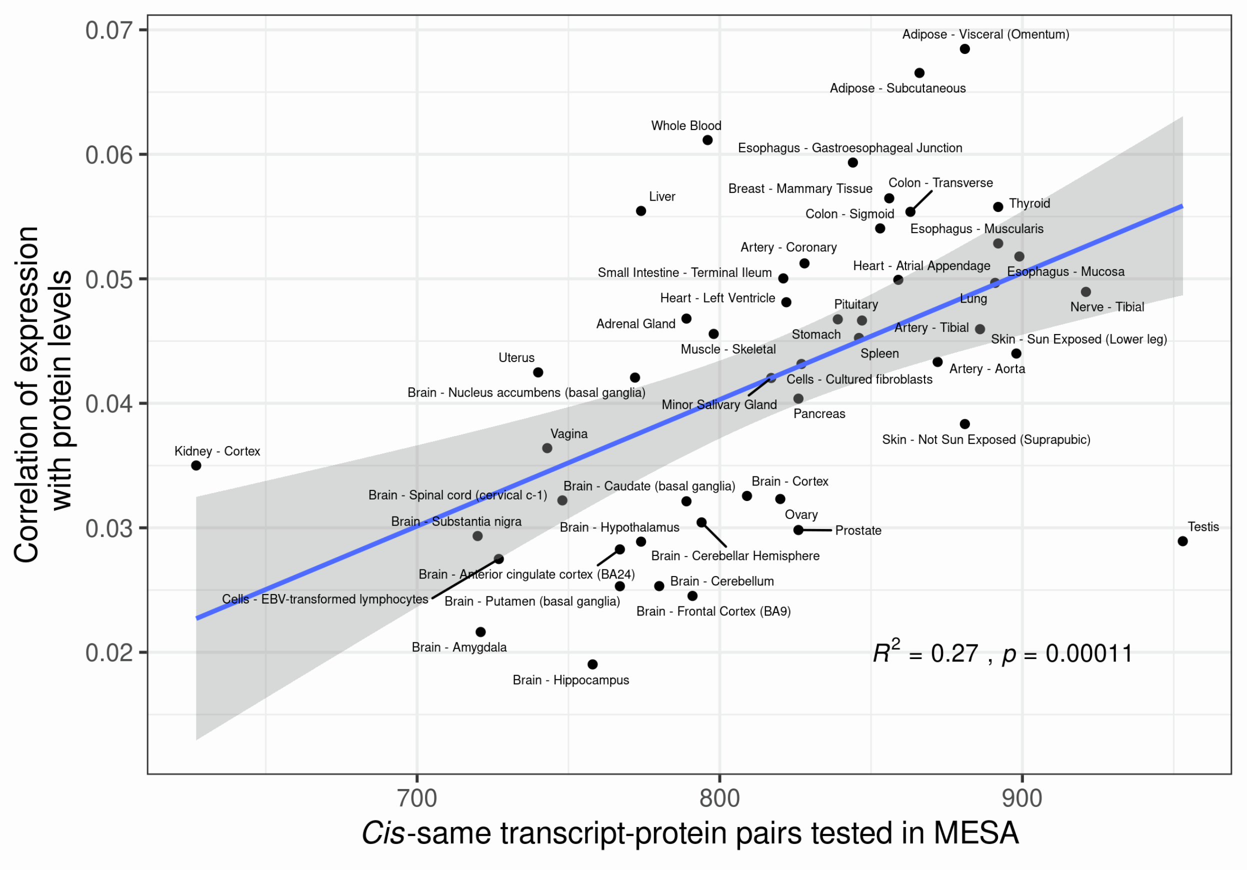


Figure S3. Comparison of median Pearson correlation between predicted expression and protein levels with the number of cis-same transcript-protein pairs tested. Points are labeled by GTEx transcriptome prediction model tissue. The blue line is the linear regression line and 95% confidence interval in gray.

Supplemental Tables

Table S1. Number of genes in the transcriptome prediction model per tissue.

Tissue	# of Genes	Tissue	# of Genes
Adipose Subcutaneous	14,732	Esophagus Mucosa	14,589
Adipose Visceral Omentum	14640	Esophagus Mucularis	14,603
Adrenal Gland	13,622	Heart Atrial Appendage	14,035
Artery Aorta	14,396	Heart Left Ventricle	13,200
Artery Coronary	13,878	Kidney Cortex	11,164
Artery Tibial	14,493	Liver	12,714
Brain Amygdala	12,814	Lung	15,058
Brain Anterior Cingulate Cortex BA24	13,528	Minor Salivary Gland	13,884
Brain Caudate Basal Ganglia	14,118	Muscle Skeletal	13,381
Brain Cerebellar Hemisphere	13,771	Nerve Tibial	15,373
Brain Cerebellum	13,992	Ovary	13,738
Brain Cortex	14,284	Pancreas	13,695
Brain Frontal Cortex BA9	14,091	Pituitary	14,647
Brain Hippocampus	13,526	Prostate	14,450
Brain Hypothalamus	13,741	Skin Not Sun Exposed Subrapubic	14,932
Brain Nucleus Accumbens Basal Ganglia	14,062	Skin Sun Exposed Lower Leg	15,204
Brain Putamen Basal Ganglia	13,694	Small Intestine Terminal Ileum	14,065
Brain Spinal Cord Cervical C-1	13,096	Spleen	14,073
Brain Substantia Nigra	12,637	Stomach	14,102
Breast Mammary Tissue	14,654	Testis	17,867
Cells Cultered Fibroblasts	13,976	Thyroid	15,308
Cells EBV-Transformed Lymphocytes	12,398	Uterus	13,199
Colon Sigmoid	14,363	Vagina	12,969
Colon Transverse	14,582	Whole Blood	12,623
Esophagus Gastroesophageal Junction	14,285		

Table S2. Genes with either predicted or observed expression and protein measurements for every tissue.

Tissue	Genes Tested	Tissue	Genes Tested
Adipose Subcutaneous	866	Esophagus Mucularis	892
Adipose Visceral Omentum	881	Heart Atrial Appendage	859
Adrenal Gland	789	Heart Left Ventricle	822
Artery Aorta	872	Kidney Cortex	627
Artery Coronary	828	Liver	774
Artery Tibial	886	Lung	891
Brain Amygdala	721	Minor Salivary Gland	817
Brain Anterior Cingulate Cortex BA24	767	Muscle Skeletal	798
Brain Caudate Basal Ganglia	789	Nerve Tibial	921
Brain Cerebellar Hemisphere	794	Ovary	820
Brain Cerebellum	780	Pancreas	826
Brain Cortex	809	Pituitary	847
Brain Frontal Cortex BA9	791	Prostate	826
Brain Hippocampus	758	Skin Not Sun Exposed Subrapubic	881
Brain Hypothalamus	774	Skin Sun Exposed Lower Leg	898
Brain Nucleus Accumbens Basal Ganglia	772	Small Intestine Terminal Ileum	821
Brain Putamen Basal Ganglia	767	Spleen	846
Brain Spinal Cord Cervical C-1	748	Stomach	839
Brain Substantia Nigra	720	Testis	953
Breast Mammary Tissue	856	Thyroid	892
Cells Cultered Fibroblasts	827	Uterus	740
Cells EBV-Transformed Lymphocytes	727	Vagina	743
Colon Sigmoid	853	Whole Blood	796
Colon Transverse	863	Monocytes – observed[a]	862
Esophagus Gastroesophageal Junction	844	PBMC – observed[a]	862
Esophagus Mucosa	899	T-cells – observed[a]	862

[a]Observed expression tissues are marked, the rest are predicted expression levels.

Table S3. INTERVAL TWAS for protein results for every tissue. π_1 is the expected true positive rate and the number of transcript-protein pairs tested is indicated for each mechanism class.

Tissue	*Trans-acting* π_1	*Pairs tested*	*Cis-acting* π_1	*Pairs tested*	*Cis-different* π_1	*Pairs tested*	*Cis-same* π_1	*Pairs tested*
Adipose Subcutaneous	0.0047	45519311	0.0905	68056	0.0839	65927	0.2950	2129
Adipose Visceral Omentum	0.0038	45112261	0.1214	67748	0.1154	65655	0.3107	2093
Adrenal Gland	0.0052	41622061	0.1056	62015	0.0995	60067	0.2943	1948
Artery Aorta	0.0052	44459436	0.1092	66129	0.1012	64000	0.3510	2129
Artery Coronary	0.0041	42425522	0.1185	63253	0.1133	61242	0.2743	2011
Artery Tibial	0.0040	45006116	0.1029	67045	0.0985	64890	0.2341	2155
Brain Amygdala	0.0043	38985756	0.0905	57171	0.0903	55386	0.0959	1785
Brain Anterior Cingulate Cortex BA24	0.0020	41159402	0.0977	60553	0.0954	58653	0.1683	1900
Brain Caudate Basal Ganglia	0.0015	43159564	0.1122	63791	0.1075	61796	0.2598	1995
Brain Cerebellar Hemisphere	0.0030	42161917	0.0953	63077	0.0920	61097	0.1975	1980
Brain Cerebellum	0.0033	42799470	0.1012	63273	0.0967	61296	0.2435	1977
Brain Cortex	0.0023	43762809	0.1091	64905	0.1030	62865	0.2958	2040
Brain Frontal Cortex BA9	0.0027	43036320	0.1023	63492	0.0970	61517	0.2653	1975
Brain Hippocampus	0.0036	41085591	0.0986	60906	0.0923	59018	0.2965	1888
Brain Hypothalamus	0.0026	41705870	0.1177	61681	0.1135	59784	0.2479	1897
Brain Nucleus Accumbens Basal Ganglia	0.0035	42906136	0.0906	63455	0.0853	61514	0.2580	1941
Brain Putamen Basal Ganglia	0.0036	41939367	0.1120	61914	0.1050	59964	0.3276	1950
Brain Spinal Cord Cervical C-1	0.0047	39565605	0.0825	58308	0.0782	56443	0.2119	1865
Brain Substantia Nigra	0.0032	38209132	0.1034	55808	0.0958	54041	0.3386	1767
Breast Mammary Tissue	0.0050	44906127	0.1239	66864	0.1165	64758	0.3543	2106
Cells Cultered Fibroblasts	0.0049	43559973	0.0872	64062	0.0812	62005	0.2683	2057
Cells EBV-Transformed Lymphocytes	0.0055	37649767	0.1021	57560	0.0954	55796	0.3141	1764
Colon Sigmoid	0.0035	44292598	0.1106	66017	0.1052	63910	0.2739	2107
Colon Transverse	0.0044	44792441	0.1154	66774	0.1115	64690	0.2361	2084

Esophagus Gastroesophageal Junction	0.0034	43875442	0.1134	65798	0.1072	63714	0.3008	2084
Esophagus Mucosa	0.0034	45175253	0.1093	68197	0.1046	66016	0.2518	2181
Esophagus Mucularis	0.0011	45189402	0.1100	67404	0.1034	65233	0.3076	2171
Heart Atrial Appendage	0.0035	43208691	0.1127	64749	0.1054	62675	0.3316	2074
Heart Left Ventricle	0.0026	40714721	0.0961	61147	0.0874	59168	0.3561	1979
Kidney Cortex	0.0039	33484277	0.1064	49300	0.1037	47733	0.1860	1567
Liver	0.0051	38860581	0.1210	58803	0.1163	56962	0.2663	1841
Lung	0.0031	46342757	0.1173	69343	0.1127	67168	0.2598	2175
Minor Salivary Gland	0.0048	42192031	0.1140	63014	0.1087	61039	0.2784	1975
Muscle Skeletal	0.0034	41571704	0.1147	62287	0.1092	60284	0.2792	2003
Nerve Tibial	0.0055	47830487	0.1102	70807	0.1030	68561	0.3301	2246
Ovary	0.0033	41915015	0.0972	62893	0.0914	60894	0.2737	1999
Pancreas	0.0044	42078288	0.1127	63231	0.1049	61220	0.3500	2011
Pituitary	0.0050	44752909	0.0916	66488	0.0854	64391	0.2833	2097
Prostate	0.0030	43999548	0.1039	65235	0.0984	63189	0.2759	2046
Skin Not Sun Exposed Subrapubic	0.0038	46342210	0.0891	69890	0.0815	67687	0.3227	2203
Skin Sun Exposed Lower Leg	0.0041	47192423	0.1037	71122	0.0972	68903	0.3051	2219
Small Intestine Terminal Ileum	0.0023	42745652	0.0995	63667	0.0952	61672	0.2314	1995
Spleen	0.0033	43155273	0.1053	64743	0.0993	62727	0.2919	2016
Stomach	0.0041	43048735	0.1069	64433	0.1000	62388	0.3163	2045
Testis	0.0035	54800000	0.0958	76465	0.0907	74134	0.2583	2331
Thyroid	0.0037	47563494	0.1112	70680	0.1051	68446	0.3000	2234
Uterus	0.0041	40141495	0.0998	60065	0.0959	58216	0.2241	1849
Vagina	0.0026	39167972	0.1004	58600	0.0975	56788	0.1895	1812
Whole Blood	0.0026	39095230	0.0825	61223	0.0761	59306	0.2816	1917

Table S4. TOPMed MESA TWAS for protein results for every tissue. π_1 is the expected true positive rate and the number of transcript-protein pairs tested is indicated for each mechanism class.

Tissue	*Trans*-Acting π_1	*Pairs tested*	*Cis*-acting π_1	*Pairs tested*	*Cis*-different π_1	*Pairs tested*	*Cis*-same π_1	*Pairs tested*
Adipose Subcutaneous	0.3517	1205	0.7811	790	0.7418	629	0.9264	161
Adipose Visceral Omentum	0.3265	1261	0.7529	784	0.7138	627	0.8893	157
Adrenal Gland	0.4307	953	0.8616	690	0.8309	546	0.9794	144
Artery Aorta	0.3410	1097	0.7311	769	0.6716	621	0.9786	148
Artery Coronary	0.2554	882	0.7736	745	0.7555	607	0.8493	138
Artery Tibial	0.3544	1022	0.8095	774	0.7720	627	0.9727	147
Brain Amygdala	0.4180	764	0.7772	621	0.7607	508	0.8522	113
Brain Anterior Cingulate Cortex BA24	0.4518	854	0.7556	657	0.7179	543	0.9218	114
Brain Caudate Basal Ganglia	0.4476	989	0.8169	712	0.8065	576	0.8575	136
Brain Cerebellar Hemisphere	0.3417	754	0.7715	649	0.7545	531	0.8412	118
Brain Cerebellum	0.4520	940	0.8446	617	0.8376	508	0.8787	109
Brain Cortex	0.3934	940	0.7664	692	0.7493	575	0.8585	117
Brain Frontal Cortex BA9	0.3860	920	0.8247	674	0.8037	546	0.9190	128
Brain Hippocampus	0.4210	885	0.8350	672	0.8267	546	0.8668	126
Brain Hypothalamus	0.5522	801	0.8129	670	0.8011	552	0.8714	118
Brain Nucleus Accumbens Basal Ganglia	0.4337	932	0.8005	674	0.7995	563	0.8059	111
Brain Putamen Basal Ganglia	0.4909	979	0.8254	687	0.7991	552	0.9309	135
Brain Spinal Cord Cervical C-1	0.3899	861	0.8511	627	0.8393	511	0.9059	116
Brain Substantia Nigra	0.3986	760	0.7339	578	0.6943	477	0.9424	101
Breast Mammary Tissue	0.3469	1129	0.7406	722	0.7014	574	0.8935	148
Cells Cultered Fibroblasts	0.2617	995	0.8408	682	0.8322	557	0.8793	125
Cells EBV-Transformed Lymphocytes	0.3793	814	0.7590	652	0.7545	547	0.7824	105
Colon Sigmoid	0.3949	1122	0.7361	733	0.7297	575	0.7586	158
Colon Transverse	0.4455	1120	0.7872	737	0.7816	598	0.8119	139

Esophagus Gastroesophageal Junction	0.3673	1148	0.8216	784	0.8071	626	0.8831	158
Esophagus Mucosa	0.3807	1115	0.8939	779	0.8826	636	0.9436	143
Esophagus Mucularis	0.3357	1137	0.8678	797	0.8566	628	0.9063	169
Heart Atrial Appendage	0.3969	1274	0.8026	797	0.7969	642	0.8235	155
Heart Left Ventricle	0.4327	881	0.7376	685	0.7459	534	0.7082	151
Kidney Cortex	0.3874	778	0.8443	564	0.8251	454	0.9237	110
Liver	0.4455	1023	0.7877	732	0.7741	588	0.8401	144
Lung	0.3502	1240	0.7596	804	0.7305	646	0.8757	158
Minor Salivary Gland	0.3688	938	0.8299	773	0.8237	646	0.8622	127
Muscle Skeletal	0.4151	828	0.8153	670	0.7949	531	0.8896	139
Nerve Tibial	0.3771	1248	0.7853	810	0.7520	664	0.9333	146
Ovary	0.3564	814	0.7627	739	0.7462	605	0.8350	134
Pancreas	0.4871	1171	0.8268	727	0.8238	596	0.8408	131
Pituitary	0.3628	1047	0.7373	755	0.7033	610	0.8881	145
Prostate	0.3753	988	0.8194	722	0.7918	585	0.9448	137
Skin Not Sun Exposed Subrapubic	0.3136	1209	0.8065	811	0.7784	663	0.9332	148
Skin Sun Exposed Lower Leg	0.3816	1217	0.8387	813	0.8243	655	0.9001	158
Small Intestine Terminal Ileum	0.4412	924	0.8430	681	0.8232	547	0.9246	134
Spleen	0.3356	1217	0.8058	724	0.7800	575	0.9175	149
Stomach	0.4160	1018	0.7056	790	0.7101	640	0.6843	150
Testis	0.4175	1056	0.7689	838	0.7241	690	0.9738	148
Thyroid	0.3648	1231	0.8217	809	0.8147	658	0.8516	151
Uterus	0.3957	852	0.8009	624	0.7831	511	0.8815	113
Vagina	0.4698	900	0.8451	641	0.8246	538	0.9504	103
Whole Blood	0.4168	1199	0.9013	760	0.8902	607	0.9443	153

Table S5. *Trans*-targets are enriched for transcription factor binding motifs.

Transcription Factor Gene	MSigDB Accession	# of Targets in Gene Set	# of Sig. Targets in Gene Set	Significant Target Genes	P-value	Adjusted P-value
NFKB	NFKB_C	66	16	*VCAM1, TNFSF18, BDNF, PTHLH, IL23A, CDH5, ICAM1, IL27RA, SIRT2, SNAP25, BCL2L1, IL2, TSLP, IRF1, TNFSF15, PAPPA*	2.6e-4	3.04e-2
NFKB	NFKB_Q6	54	16	*VCAM1, TNFSF18, BDNF, STIP1, CADM1, PTHLH, IL23A, GREM1, MED1, RNF43, ICAM1, TSLP, TNFSF15, PAPPA*	2.99e-4	3.04e-2
NFKB	NFKB_Q6_01	61	19	*VCAM1, TNFSF18, BDNF, CADM1, PTHLH, IL23A, GREM1, TP53, RNF43, EBI3, ICAM1, IL27RA, SIRT2, SNAP25, MMP9, IL6ST, TSLP, TNFSF15, PAPPA*	1.25e-6	5.43e-4
RELA	NFKAPPAB_01	61	16	*VCAM1, TNFSF18, BDNF, IL23A, GREM1, CXCL16, TP53, MED1, LAMA1, EBI3, ICAM1, IL27RA, MMP9, TSLP, IRF1, TNFSF15*	9.51e-5	1.93e-2
RELA	NFKAPPAB65_01	52	14	*TNFSF18, BDNF, PTHLH, GREM1, CXCL16, TP53, RNF43, LAMA1, ICAM1, SIRT2, MMP9, IL6ST, TSLP, TNFSF15*	1.94e-4	2.95e-2
NFAT1C	TGGAAA_NFAT_Q4_01	351	59	*ISG15, TNFRSF8, BCAR3, TSHB, NTRK1, FASLG, FGF8, ADM, BDNF, FTH1, STIP1, CADM1, ERBB3, IL23A, IGF1, TNFSF11, IL25, GREM1, SPINT1, CDH5, TP53, MAP2K3, RNF43, COLEC12, RETN, ICAM1,*	1.78e-6	5.43e-4

				POMC, LRP1B, SNAP25, JAG1, JAG1, USP25, PDGFB, CHL1, LTF, LSAMP, FSTL1, APOD, PDGFRA, IBSP, IL2, CCL28, IL6ST, PDE4D, CAST, SEMA6A, IL5, IL12B, APOM, EHMT2, IL17F, SYNCRIP, SMPDL3A, INHBA, HGF, SEMA3A, ANGPT1, TNFRSF11B, TPM2, PAK3		
FOXF2	FREAC2_01	44	12	*TXNDC12, BDNF, CADM1, IGF1, IL25, GREM1, AKT2, IL6ST, CD109, CGA, INHBA, PAPPA*	4.9e-4	3.4e-2
AR	AR_Q2	17	7	*BDNF, MSTN, SNAP25, AGER, SYNCRIP, INHBA, CD36*	5.02e-4	3.4e-2
GATA1	GATA1_04	51	13	*TSHB, ADM, IGF1, IL34, CEBPB, PDGFRA, CAST, IL5, EPO, MSR1, PRSS3, CCL27, PAPPA*	5.87e-4	3.40e-2
STAT1	STAT_01	55	14	*BDNF, TNFSF11, GZMB, MAP2K3, VTN, LPO, ICAM1, USP25, OSM, THPO, IL6ST, RASA1, IRF1, EHMT2*	3.68e-4	3.21e-2

Table S6. *Cis*- and *Trans*-targets are enriched for mapped GWAS catalog associations (FDR < 0.05).

Mechanism	Gene Set	# of Targets in Gene Set	# of Significant Targets in Gene Set	P-value	Adjusted P-value
Trans-acting	Blood Protein Levels	862	122	2.26e-8	4.10e-5
Trans-acting	Inflammatory Bowel Disease	190	38	2.64e-6	2.39e-3
Cis-acting	Blood Protein Levels	862	229	4.73e-125	8.58e-122
Cis-acting	Ankylosing Spondylitis	22	11	1.89e-7	1.62e-4
Cis-acting	Inflammatory Bowel Disease	190	36	2.68e-7	1.62e-4
Cis-acting	Chronic Inflammatory Diseases	48	13	5.21e-5	2.36e-2

Table S7. Pleiotropic regulatory loci discovered in INTERVAL.

Locus	Genes in Locus	Chromosome	Location (bp)	Unique Targets	Replicated Targets (significant / tested)
1	*CFHR3, CFHR1, CFHR4*	1	196,774,813 – 196,888,014	134	5/56
2	*BCHE*	3	165,772,904	56	3/12
3	*C7*	5	40,909,497	280	51/103
4	*C6*	5	41,142,116	81	9/42
5	*HLA-DQB2, HLA-DQA1*	6	32,628,179 – 32,756,098	82	10/31
6	*SKIV2L, CYP21A2, C4B*	6	31,959,117 – 32,038,327	86	18/34
7	*GSDMD*	8	143,553,207	54	2/20
8	*ABO*	9	133,233,278	55	27/33
9	*SARM1, TMEM199, POLDIP2, SUPT6H, TNFAIP1, TMEM97, IFT20, SLC46A1, ERVE-1, SLC13A2*	17	28,232,590 – 28,662,198	290	37/86
10	*MYADM, NLRP12, AC008753.3*	19	53,787,597 – 53,864,763	555	1/218
11	*APOE*	19	44,905,791	78	1/17

Table S8. Pleiotropic regulatory genes are enriched for mapped GWAS catalog associations (FDR < 0.05).

Locus	Genes in Locus	Associated GWAS Catalog Traits
1	*CFHR3, CFHR1, CFHR4*	Nephropathy, Age-related macular degeneration, Matrix metalloproteinase-8 levels, Complement C3 and C4 levels, IgA nephropathy, Advanced age-related macular degeneration
5	*HLA-DQB2, HLA-DQA1*	Immunoglobulin A vasculitis, Strep throat, Childhood steroid-sensitive nephrotic syndrome, Neuromyelitis optica, Pneumonia, Neuromyelitis optica (AQP4-IgG-positive), Chronic hepatitis C infection, Drug-induced liver injury (flucloxacillin), Plantar warts, Shingles, Myositis, Multiple sclerosis (OCB status), Late-onset myasthenia gravis, Lymphoma, PEG-asparaginase hypersensitivity without enzyme activity in childhood acute lymphoblastic leukemia, Peanut allergy, Response to hepatitis B vaccine, Nephropathy, Cervical cancer, Asthma (moderate or severe), Sarcoidosis (non-Lofgren's syndrome without extrapulmonary manifestations), IgA nephropathy, Allergy, Self-reported allergy, Allergic sensitization, Childhood ear infection, Sjögren's syndrome, Primary biliary cirrhosis, Tuberculosis, Systemic sclerosis, Tonsillectomy, Hypothyroidism, Takayasu arteritis, Chronic lymphocytic leukemia, Squamous cell lung carcinoma, Allergic rhinitis, Itch intensity from mosquito bite adjusted by bite size, Celiac disease, Asthma or allergic disease (pleiotropy), Lung cancer, Rheumatoid arthritis, Red blood cell count, Allergic disease (asthma, hay fever or eczema), Asthma, Prostate cancer, Systemic lupus erythematosus, Ulcerative colitis, Type 2 diabetes, Autism spectrum disorder or schizophrenia, Crohn's disease, Schizophrenia, Inflammatory bowel disease
6	*SKIV2L, CYP21A2, C4B*	Prostate cancer, Ulcerative colitis, Autism spectrum disorder or Schizophrenia, Inflammatory bowel disease
9	*SARM1, TMEM199, POLDIP2, SUPT6H, TNFAIP1, TMEM97, IFT20, SLC46A1, ERVE-1, SLC13A2*	Osteoprotegerin levels, Blood protein levels
10	*MYADM, NLRP12, AC008753.3*	None

Table S9. TOPMed MESA *cis*-same π_1 values for every predicted and observed tissue.

Tissue	*Cis*-same π_1	# Transcript-protein pairs tested
Brain Putamen basal ganglia	0.4912	767
Skin Sun Exposed Lower leg	0.4662	898
Small Intestine Terminal Ileum	0.4610	821
Kidney Cortex	0.4480	627
Esophagus Muscularis	0.4473	892
Uterus	0.4406	740
Liver	0.4317	774
Cells Cultured fibroblasts	0.4195	827
Adrenal Gland	0.4193	789
Minor Salivary Gland	0.4040	817
Esophagus Mucosa	0.3960	899
Testis	0.3945	953
Muscle Skeletal	0.3910	798
Brain Caudate basal ganglia	0.3896	789
Heart Atrial Appendage	0.3850	859
Pituitary	0.3845	847
Brain Hypothalamus	0.3726	774
Colon Transverse	0.3708	863
Thyroid	0.3651	892
Brain Frontal Cortex BA9	0.3633	791
Adipose Subcutaneous	0.3530	866
Brain Spinal cord cervical c-1	0.3472	748
Ovary	0.3440	820
Adipose Visceral Omentum	0.3427	881
Brain Anterior cingulate cortex BA24	0.3408	767
Heart Left Ventricle	0.3408	822
Brain Hippocampus	0.3366	758
Brain Substantia nigra	0.3319	720
Whole Blood	**0.3311**	796
Breast Mammary Tissue	0.3239	856
Prostate	0.3223	826
Skin Not Sun Exposed Suprapubic	0.3222	881
Artery Aorta	0.3183	872
Brain Cerebellum	0.3168	780
Cells EBV-transformed lymphocytes	0.3129	727
Brain Cortex	0.3048	809
Lung	0.3031	891

Vagina	0.3028	743
Esophagus Gastroesophageal Junction	0.2962	844
Brain Nucleus accumbens basal ganglia	0.2905	772
Colon Sigmoid	0.2872	853
Brain Amygdala	0.2848	721
Nerve Tibial	0.2764	921
Spleen	0.2747	846
Artery Coronary	0.2684	828
Artery Tibial	0.2678	886
Stomach	0.2662	839
Pancreas	0.2418	826
PBMC – observed	0.2393	862
Monocytes – observed	0.1928	862
Brain Cerebellar Hemisphere	0.1819	794
T-cells – observed	0.0768	862

Detailed Acknowledgements

This work is supported by the NIH National Human Genome Research Institute Academic Research Enhancement Award R15 HG009569 (HEW).

Participants in the INTERVAL randomised controlled trial were recruited with the active collaboration of NHS Blood and Transplant England (www.nhsbt.nhs.uk), which has supported field work and other elements of the trial. DNA extraction and genotyping was co-funded by the National Institute for Health Research (NIHR), the NIHR BioResource (http://bioresource.nihr.ac.uk) and the NIHR [Cambridge Biomedical Research Centre at the Cambridge University Hospitals NHS Foundation Trust]. The INTERVAL study was funded by NHSBT (11-01-GEN). The academic coordinating centre for INTERVAL was supported by core funding from: NIHR Blood and Transplant Research Unit in Donor Health and Genomics (NIHR BTRU-2014-10024), UK Medical Research Council (MR/L003120/1), British Heart Foundation (SP/09/002; RG/13/13/30194; RG/18/13/33946) and the NIHR [Cambridge Biomedical Research Centre at the Cambridge University Hospitals NHS Foundation Trust]. Proteomic assays were funded by the academic coordinating centre for INTERVAL and MRL, Merck & Co., Inc. A complete list of the investigators and contributors to the INTERVAL trial is provided in Di Angelantonio et al. The academic coordinating centre would like to thank blood donor centre staff and blood donors for participating in the INTERVAL trial. This work was supported by Health Data Research UK, which is funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation and Wellcome. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

MESA and the MESA SHARe project are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA is provided by contracts 75N92020D00001, HHSN268201500003I, N01-HC-95159, 75N92020D00005, N01-HC-95160, 75N92020D00002, N01-HC-95161, 75N92020D00003, N01-HC-95162, 75N92020D00006, N01-HC-95163, 75N92020D00004, N01-HC-95164, 75N92020D00007, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-TR-000040, UL1-TR-001079, UL1-TR-001420. Funding for SHARe genotyping was provided by NHLBI Contract N02-HL-64278. Genotyping was performed at Affymetrix (Santa Clara, California, USA) and the Broad Institute of Harvard and MIT (Boston, Massachusetts, USA) using the Affymetrix Genome-Wide Human SNP Array 6.0. MESA Family is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support is provided by grants and contracts R01HL071051, R01HL071205, R01HL071250, R01HL071251, R01HL071258, R01HL071259, by the National Center for Research Resources, Grant UL1RR033176. Also supported in part by the National Center for Advancing Translational Sciences, CTSI grant UL1TR001881, and the National Institute of Diabetes and Digestive and Kidney Disease Diabetes Research Center (DRC) grant DK063491 to the Southern California Diabetes Endocrinology Research Center.

The TOPMed MESA Multi-Omics project was conducted by the University of Washington and LABioMed (HHSN2682015000031/HHSN26800004). Molecular data for the

Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). SOMAscan proteomics for NHLBI TOPMed: Multi-Ethnic Study of Atherosclerosis (MESA) (phs001416.v1.p1) was performed at the Broad Institute and Beth Israel Proteomics Platform (HHSN268201600034I). Core support including centralized genomic read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN268201800002I). Core support including phenotype harmonization, data management, sample-identity QC, and general program coordination were provided by the TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract HHSN268201800001I). The MESA Epigenomics and Transcriptomics Studies were funded by National Institutes of Health grants 1R01HL101250, 1RF1AG054474, R01HL126477, R01DK101921, and R01HL135009. Support for analysis of the TOPMed MESA Multi-Omics data was provided by R01HL153248. The authors thank the other investigators, the staff, and the participants of the MESA study for their valuable contributions. A full list of participating MESA investigators and institutes can be found at http://www.mesa-nhlbi.org.