9-6-2024

Transcriptome-Wide Association Study of the Plasma Proteme Reveals Cis and Trans Regulatory Mechanisms Underlying Complex Traits

Henry Wittich
Loyola University of Chicago Graduate School

LOYOLA UNIVERSITY CHICAGO

TRANSCRIPTOME-WIDE ASSOCIATION STUDY OF THE PLASMA PROTEOME
REVEALS CIS AND TRANS REGULATORY
MECHANISMS UNDERLYING COMPLEX TRAITS

A THESIS SUBMITTED TO
THE FACULTY OF THE GRADUATE SCHOOL
IN CANDIDACY FOR THE DEGREE OF
MASTER OF SCIENCE

PROGRAM IN BIOINFORMATICS

BY
HENRY WITTICH
CHICAGO, IL
AUGUST 2023

ACKNOWLEDGMENTS

Above all, I would like to thank Dr. Wheeler for her invaluable mentorship and guidance during my time working in her research lab. She has helped me realize the potential in myself and helped me accomplish things I didn't know I could. I have learned so much working on my thesis with her, and I am extremely grateful to her for creating that space for me to grow. I leave her lab feeling energized and prepared to apply my skills to furthering the field of human genetics and the study and treatment of human diseases.

I'm grateful to all the figures at Loyola that provided mentorship or support to me during my academic career here, including Dr. Putonti, Dr. Laten, and the members of the Wheeler Lab. I especially want to thank Daniel Araujo, who navigated the MS program alongside me, for being great partner in our many group projects, a useful sounding board whenever I got new results or was trying out a new analysis for my thesis, and for being a great friend outside of the lab. I also want to thank Emily Radz and Evelyn Chavez for being with me throughout my entire BS/MS and being my cheerleaders the whole way.

Lastly, I want to thank all of my close friends and family outside of Loyola who have supported me throughout my life. My mom has always seen the potential in me and the way she has been an academic advocate for me has given me the confidence and skills to advocate for myself. I'm so grateful for her endless support and I'm proud to make her proud through my work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AFA	African American population from the TOPMed MESA cohort
BMI	Body mass index
CHN	Chinese population from the TOPMed MESA cohort
DNA	Deoxyribonucleic acid
eQTL	Expression quantitative trait locus
EUR	European population from the TOPMed MESA cohort
FDR	False discovery rate
FUMA	Functional Mapping and Annotation
GReX	Genetically-regulated expression
GTEx	Genotype-Tissue Expression Project
GWAS	Genome-wide association study
HIS	Hispanic population from the TOPMed MESA cohort
HTS	High throughput system
LD	Linkage disequilibrium
MAF	Minor allele frequency
MASHR	Multivariate adaptive shrinkage in R
MESA	Multi-Ethnic Study of Atherosclerosis
PBMC	Peripheral blood mononuclear cell
PWAS	Proteome-wide association study
RNA	Ribonucleic acid
SNP	Single nucleotide polymorphism
TF	Transcription factor
TOPMed	NHLBI Trans-Omics for Precision Medicine consortium

TSS	Transcription start site
TWAS	Transcriptome-wide association study

ABSTRACT

Most genetic variants associated with complex human traits exist in non-coding regions, and thus the mechanism by which they affect a trait can be unclear. Genetic regulation of transcription and translation are key mechanisms through which genetic variants impact traits. Quantitative trait locus (QTL) mapping studies leverage data produced by advanced sequencing and assay technology to identify variants associated with the abundance of a molecular trait like RNA expression (eQTL) or protein levels (pQTL). While proximal genetic variants (*cis*-acting), like those in promoter or enhancer regions of genes, tend to have the largest effect sizes on RNA transcript and protein levels, distal genetic variation (*trans*-acting) still contributes considerably to regulating transcription and translation. However, *trans*-QTL can be difficult to discover due to the high multiple testing burden, their comparatively low effect sizes, and their tendency to have tissue- or cell type-specific effects. Methods that prioritize testing *cis*-eQTL for *trans*-acting associations have proven effective because they reduce the multiple testing burden and many *trans*-eQTL colocalize with *cis*-eQTL. For example, a transcriptome-wide association study (TWAS) that used observed gene expression as trait found more *trans*-acting genes than a comparable eQTL study. We hypothesized that performing a TWAS using protein levels as trait would be effective at identifying *trans*-pQTL because it prioritized *cis*-eQTL for testing *trans*-acting genes. While gene expression and protein levels have previously been shown to have a low correlation, we hypothesized that GReX would have a higher correlation with observed protein levels because it excludes variation in expression due to environmental factors. We used genotype and plasma protein measurements from individuals participating in the INTERVAL study for our TWAS and replicated these results with genotype and plasma protein measurements from individuals in the TOPMed MESA cohort. We used transcriptome prediction models from 49 tissues trained with GTEx Project genotype and RNA-Seq data. Furthermore, we used RNA-Seq data from the TOPMed MESA cohort to compare the correlation of observed expression levels with protein levels to the correlation of predicted expression levels with protein levels. We discovered many replicable *cis*- and *trans*-acting gene-protein relationships and found that predicted expression had a higher correlation and true positive rate than observed expression for significant association with protein levels. These results indicate that predicted gene expression may better uncover the genetic mechanisms underlying complex traits than observed expression.

INTRODUCTION

Genetic Regulation of Gene Expression

Since the publication of the first human genome in 2001, the time and cost efficiency of genome sequencing has improved greatly.[1,2] The proliferation of genomic data has allowed for extensive study of human genetic variation. Comparison of thousands of human genomes has revealed that, while greater than 99.9% of the human genome is shared between individuals, there are also millions of DNA variants that play a substantial role in the phenotypic differences observed between people.[3,4] The basic unit of variation is the single nucleotide polymorphism (SNP) – a location in the genome where individuals have a different nucleotide base pair (i.e. a different allele) from others. Genome-wide association studies (GWAS) leverage genotype data across large populations to identify specific variants that are associated with complex human traits by testing the genotypes of millions of variants for association with a measurable trait of interest like BMI or disease presentation.[5] As of 2023, more than 500,000 significant GWAS associations for thousands of traits have been reported in the GWAS catalog, an online database for storing and reporting GWAS results.[6] While some GWAS SNPs are found within exons, most GWAS SNPs are located in noncoding regions and thus likely play a role in affecting complex traits through gene regulation.[7–9]

Transcriptional regulation of gene expression is an important mechanism through which genetic variation influences trait presentation. On top of environmental influence in the form of epigenetic modification, transcription is carefully controlled by a variety of both nearby genetic variation (*cis*-SNPs), in the form of promoter, insulator, and enhancer regions responsible for recruiting or blocking proteins necessary for gene expression, and distant genetic variation (*trans*-SNPs), which largely involves the regulation of expression of transcription factor genes. Technological advances like genotyping microarrays, which contain thousands to millions of probes that target and classify known genetic variants in DNA, and RNA-Seq, which uses techniques like reverse transcription, short read next-generation sequencing, and DNA mapping to identify and quantify RNA transcripts in a sample, have enabled the computational study of gene regulation.[10,11]

Expression quantitative trait locus (eQTL) mapping leverages genotype and expression level data produced by expression microarrays, which measure the expression of specific, target genes, or next-generation sequencing methods to identify genetic variants that influence gene expression. Like performing a GWAS with expression levels as a trait, eQTL studies test genotypes for association with the expression levels of thousands of genes. Due to the difficulties in collecting samples for RNA-Seq from most human tissues, the first eQTL studies were performed in blood.[12–15] However, gene expression is variable across different cell and tissue types, so eQTL studies conducted in blood were limited in their ability to detect eQTL effects not present in blood. The Gentoype-Tissue Expression (GTEx) Project was established to provide a resource for genetic variation's impact on expression across a wide variety of human tissues.[16] The latest, version 8 release of the GTEx Project performed eQTL mapping in 49 different human tissues.[16] GTEx and other early eQTL studies found that the strength of effect of eQTL decayed with distance from the transcription start site (TSS) of a gene.[16,17] Furthermore, cross tissue analysis revealed that eQTL closer to the TSS of a gene (*cis*-acting) were more shared across tissues while distant eQTL (*trans*-acting) had more cell-specific effects.[16–19] Despite *trans*-eQTL having low effect sizes, they clearly play an important role in cell type differentiation, and they still contribute a large amount to the heritability of gene expression, highlighting how most eQTL studies have been underpowered at identifying *trans*-acting effects.[20]

The smaller effect sizes of *trans*-eQTL, combined with the high multiple testing burden for mapping *trans*-eQTL have limited their identification and confirmation. However, many *cis*-eQTL also participate in *trans*-acting effects, thus some studies have been successful at prioritizing known *cis*-eQTL for discovering *trans*-acting effects.[21,22] One method that has proven successful for identifying *trans*-acting effects is *trans*-PrediXcan.[23,24] PrediXcan is a framework for performing transcriptome-wide association Studies (TWAS), in which the regularized effect sizes of *cis*-eQTL are used to impute the genetically-regulated expression (GReX) of genes from genotype data and test each gene for association with a measured trait.[23,25] *Trans*-PrediXcan aimed at identifying *trans*-acting genes by testing the genetically regulated expression of a gene for association with the observed expression of all distant genes.[24] This method had more power for identifying *trans*-acting effects than *trans*-eQTL studies because it combined the effects of multiple *cis*-eQTL to predict expression for *trans*-acting genes.[24]

Incorporating Proteomic Analysis

There are limitations to taking a transcriptomic approach to studying complex human traits. Translation takes place after transcription, as do additional posttranslational modification and processing steps, which could affect the mechanism by which genetic variation influences complex traits.[26] Incorporating multiple molecular quantitative trait loci is essential for picking apart the pathways through which noncoding GWAS SNPs affect trait presentation.[8,27] Advanced proteomic assay technology has allowed for the incorporation of proteomic data into complex trait genetics approaches.[21,28] Protein quantification platforms like SomaScan use aptamers that specifically bind to one (or sometimes a few) proteins to capture and quantify target proteins from a biological sample.[29,30] Large-scale quantification of protein levels like this has powered protein quantitative trait locus (pQTL) studies, which identify genetic variants associated with protein levels.[26,31,32] As with early eQTL studies, most pQTL mapping has been performed in blood tissues due to the ease of sampling. Similar to TWAS, proteome-wide association studies (PWAS) predict genetically regulated protein levels using effect sizes from *cis*-pQTL and test genetically regulated protein expression for association with a measured trait in order to better understand the genetic mechanisms underlying the complex trait.[33,34]

TWAS for Proteins Approach

Like eQTL studies, pQTL studies have been underpowered for discovering *trans*-pQTL.[31,32] Furthermore, many studies have found large overlap between eQTL and pQTL, especially in blood and liver cells.[26,31,32] In contrast, many studies have reported that observed RNA expression levels do not correlate well with the observed protein levels of the same underlying gene.[26,35–37] Here, we leveraged the PrediXcan framework to examine how gene regulation across many human tissues influences plasma protein level in *cis* and *trans*.[23] By prioritizing *cis*-eQTL for testing *trans*-acting genes, we hypothesized that this method would have more power for detecting *trans*-pQTL. Furthermore, we hypothesized that the genetically regulated component of expression (GReX) is more correlated with protein levels than observed expression of the encoding gene because GReX excludes variability of expression caused by nongenetic factors. Our discovery cohort included individuals from the INTERVAL study.[31] We used genotype data from these individuals and gene expression prediction models from 49 GTEx tissues to impute GReX, which we tested for association with their measured plasma protein levels.[16] We replicated these results using genotype and plasma protein data taken from individuals in the TOPMed MESA multi-omics pilot study.[38] Finally, we utilized RNA-Seq for individuals in the TOPMed MESA cohort to test observed expression for association with plasma protein levels and compared the correlation of observed and predicted expression with plasma protein levels.[39] We found both *cis*- and *trans*-acting relationships that replicate between transcripts and plasma protein levels, with the highest true positive rates for transcripts that encode the measured protein. We also show predicted expression has a higher true positive rate than observed expression for significant association with protein levels.

METHODS

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.[40] 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).[31] Data generation and quality control have previously been described by the INTERVAL study.[31,40] An Affymetrix Axiom UK Biobank array was used for genotyping and imputation was performed on the Sanger imputation server using a combined 1000 Genomes Phase3-UK10K reference panel.[5,7] Genotypes were then filtered for minor allele frequency (MAF) > 0.01 and $R^2 > 0.8$.[39] The SOMAscan assay used to collect the proteomic data targeted 3,622 plasma proteins.[30] The protein levels were log-transformed and adjusted for age, sex, duration between blood draw and processing, and the first three genetic principal components.[31]

Our replication dataset was from the Trans Omics for Precision Medicine (TOPMed) MESA (Multiethnic Study of Atherosclerosis) 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' effect on disease traits and drug responses.[38] The MESA study, under TOPMed, includes individuals from multiple genetic ancestries. The four study-defined groups included in MESA are African American (AFA), Chinese (CHN), European (EUR), and Hispanic/Latino (HIS). Individuals in the study were genotyped as part of the MESA SHARe study (dbGaP: phs000420.v6.p3).[41] Genotype quality control has been previously described.[39,42] Genotypes were imputed on the Michigan imputation server using a reference panel from 1000 Genomes.[42] They were then filtered for MAF > 0.01 and R^2 > 0.8.[39] 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.[29] The protein levels were log-transformed and adjusted for age, sex, time-point, and the first ten genetic principal components.[39] 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-Seq was performed for individuals from all four different populations (AFA, CHN, EUR, and HIS) in three different blood cell types: peripheral blood mononuclear cells (PBMC), CD16+ monocytes (Mono), and CD4+ T-cells.[41] In total, 1,287 PBMC samples, 395 Mono samples, and 397 T-Cell samples were sequenced. Quality control for the RNA-Seq data has been previously described.[43] 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. Finally, expression levels were log-transformed, adjusted for age, sex, time point, the first ten genetic principal components, and the first ten expression components.[43] .

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.[23] We used gene expression prediction models from

PredictDB, which were built off the Genotype-Tissue Expression (GTEx) Project's version 8 release, to impute GReX in 49 different human tissues.[16,23,44,45] The models were built using MASHR (Multi-Variate Adaptive shrinkage in R)[46] and only include *cis*-eQTL with MAF > 0.01.[47] The number of genes included in each tissue's gene expression prediction model can be found in Table 1. 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 that we predicted expression in, we used the Qvalue R package to calculate qvalues for all predicted transcript-protein association tests conducted.[48] Transcript-protein pairs with a qvalue (FDR) < 0.05 were considered significant.

For every transcript-protein pair that we found significant (FDR < 0.05) in INTERVAL, we tested that association using genotypes and protein levels from TOPMed MESA if the protein was measured in both studies (Figure 1a).

Table 1. 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		

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.[48] The qvalue function from the Qvalue R package calculates the π_0 statistic from a vector of pvalues.[48] 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.[48]

$$\pi_1 = 1 - \pi_0$$

We divided the associations we tested in INTERVAL into 4 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 entirely. 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 (Figure 1b).

For each of these groups in every tissue, we pulled the pvalues for every tested association and calculated the π_0 statistic using the qvalue R package. While we used the default qvalue function parameters in INTERVAL, we adjusted the qvalue 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 pvalues, thus the pvalue distribution in most tissues did not extend all the way to 1. By default, the qvalue function calculates the average frequency of pvalues from 0.05 to 1.0 to determine the expected proportion of null pvalues, so there must be pvalues throughout this entire range for the function to work. These bounds are controlled by the lambda parameter, which we truncated from 0.05 to 1.0 to 0.05 to 0.75 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.

Gene Set Enrichment Analysis of Protein Targets

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.[49] We tested the targets involved in *cis*-acting and *trans*-acting associations separately. For both groups, we tested the protein targets for enrichment (FDR < 0.05) of GWAS catalog associations and motifs that are known targets of transcription factors (TFs).

Identifying Master Regulatory Loci

We defined a master 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 master regulatory genes. We further defined a master regulatory locus as a set of master regulatory genes whose transcription start sites are within 200kb of the transcription start site of the nearest gene in the locus. For each master 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 replicated associations with any of the genes in that locus in TOPMed MESA.

Gene Set Enrichment Analysis of Master Regulators

We used FUMA GWAS to perform a gene set enrichment analysis of the master regulatory genes as well as their protein targets.[49] We tested the protein targets of each master 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 master regulatory loci with more than one gene, we tested the master 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-Seq 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. Furthermore, we performed a TWAS with PrediXcan to test the predicted gene expression of each gene included in each of the 49 prediction models from GTEx for association with the observed abundance of the protein product of that gene if it was also measured in TOPMed MESA. The number of genes that we tested for *cis*-same associations in each tissue are listed in Table 2.

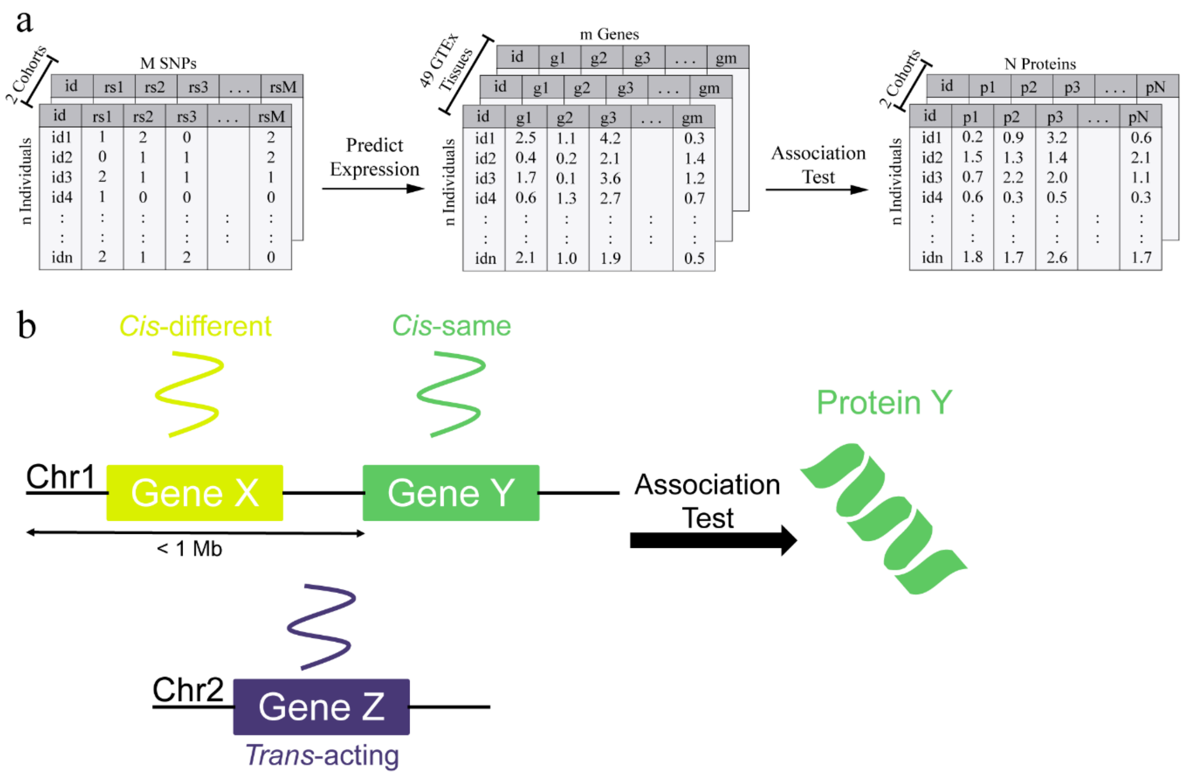
Table 2. Genes with either predicted or observed expression and protein measurements for every tissue.

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	862
Esophagus Gastroesophageal Junction	844	PBMC – observed	862
Esophagus Mucosa	899	T-cells – observed	862

Note: Observed expression tissues are marked, the rest are predicted expression levels.

As above, we assessed significance via the Benjamini-Hochberg false discovery rate (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 qvalues for all of the *cis*-same transcript-protein pairs tested using the qvalue R. 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 to 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-Seq, 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.



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. Experiment design.

(a) Overview of TWAS analysis. Genotype data from both the INTERVAL and TOPMed MESA cohorts was used to impute genetically regulated transcript levels in 49 different GTEx tissues. GReX was linearly associated 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 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. Likewise, the expression of gene Y and the abundance of protein Y have a *cis*-same relationship because they are encoded by the same gene. 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).

RESULTS

TWAS for Protein Identifies Replicable Gene-Protein Interactions

We sought to identify both *cis*- and *trans*-acting transcriptional regulators of plasma proteins by performing TWAS for protein levels. Using the PrediXcan software framework, 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 prediction models built 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 (Figure 2a,d). 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 for 218 unique proteins (92 *cis*-different and 126 *cis*-same) and 1,210 *trans*-acting pairs replicated for 239

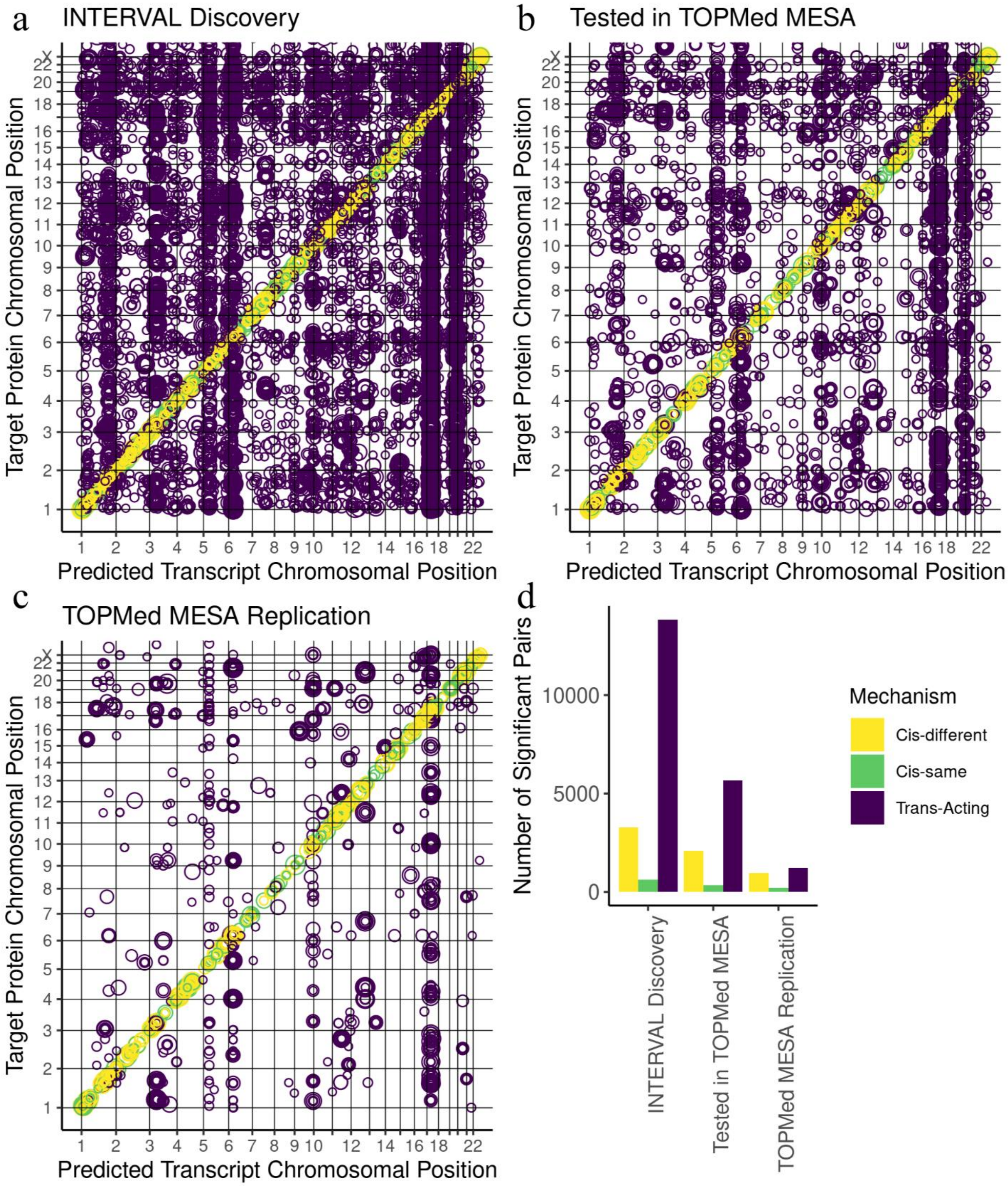


Figure 2. Overview of significant transcript-protein associations.

(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 location in bp of the gene that encodes the predicted transcript while the position along the y-axis corresponds to the genomic location in bp of the gene that encodes the target protein. The size of each circle corresponds to the number of tissues (out of all 49) where the pair was discovered significantly associated. (a) This plot shows all significantly (FDR < 0.05) associated transcript-protein pairs discovered in INTERVAL. (b) This plot shows all significantly (FDR < 0.05) associated transcript-protein pairs discovered in INTERVAL that were also tested in TOPMed MESA. (c) This plot shows all 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.

proteins (FDR < 0.05, Figure 2b-d). 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).

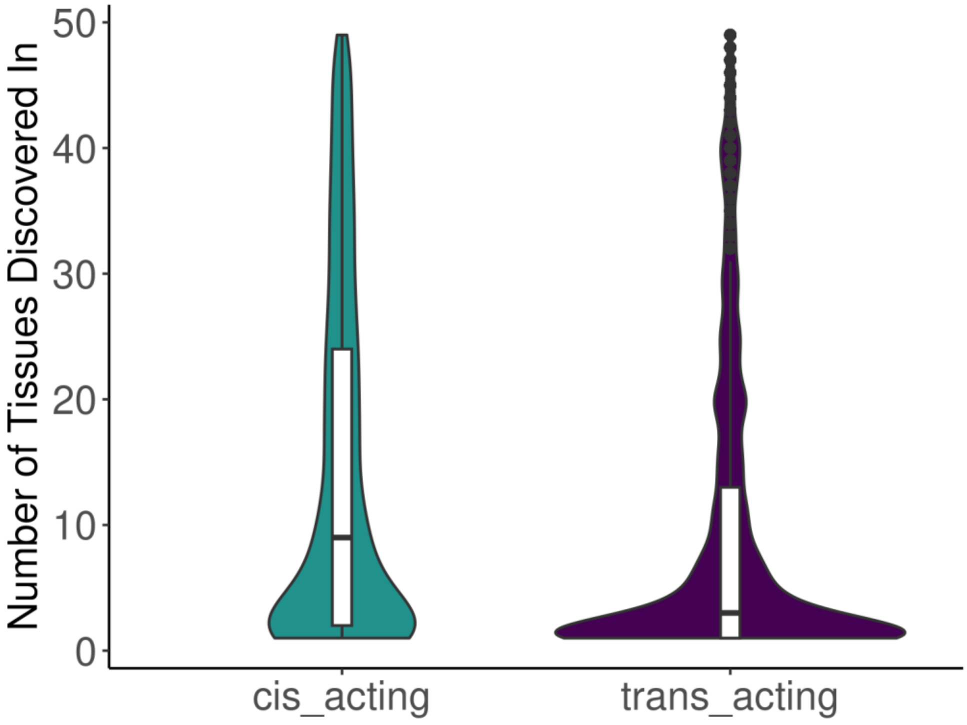


Figure 3. Sharing of *cis*- and *trans*-acting effects across tissues in INTERVAL.

Violin plots depicting the number of tissues in which each significant transcript-protein pair was discovered (FDR < 0.05), divided into *cis*- and *trans*-acting associations.

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 4a, Table 3). 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 4b, Table 4).

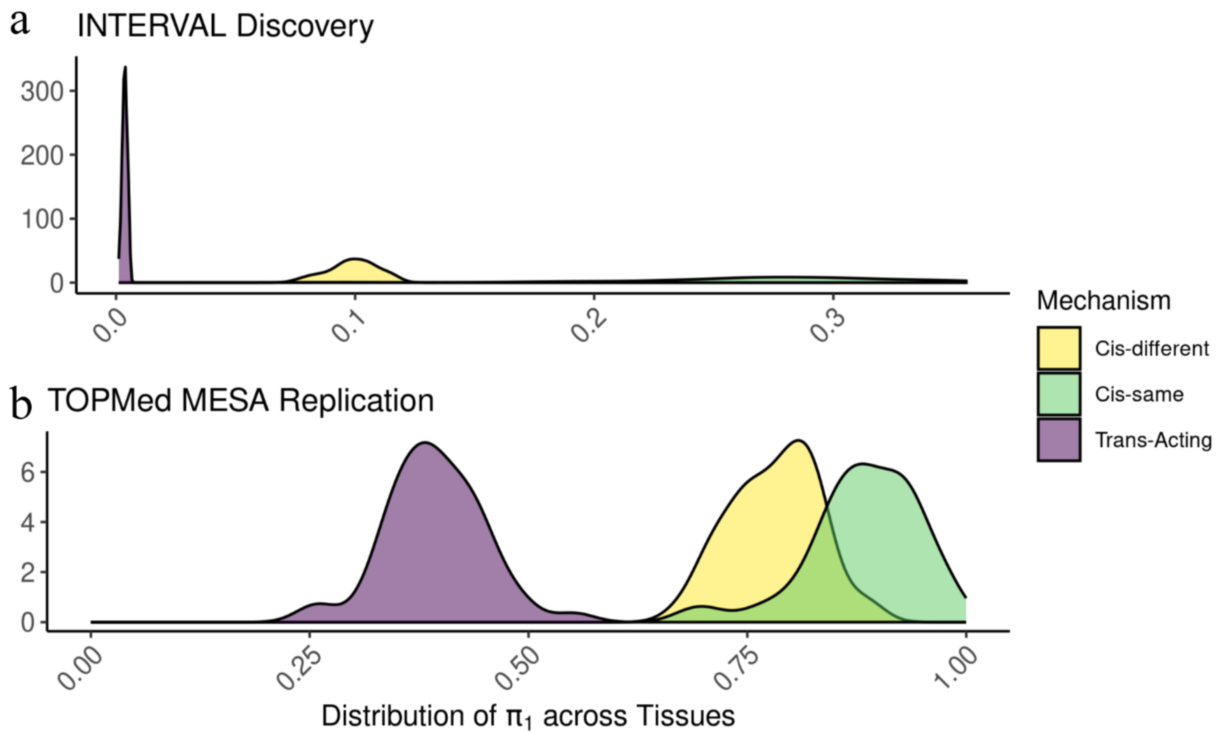


Figure 4. Expected true positive rates (π_1) for transcript-protein pairs across tissues.

(a) π_1 values of transcript-protein pairs tested in INTERVAL. Associations were divided into *cis*-same, *cis*-different, and *trans*-acting and π_1 was calculated in every GTEx tissue. (b) π_1 values of significant (FDR < 0.05) transcript-protein pairs discovered in INTERVAL that were also tested in TOPMed MESA. Associations were divided into *cis*-same, *cis*-different, and *trans*-acting and π_1 was calculated in each GTEx tissue separately.

Table 3. INTERVAL π_1 values for every tissue

Tissue	*Trans*-Acting	*Cis*-acting	*Cis*-different	*Cis*-same
Adipose Subcutaneous	0.0047	0.0905	0.0839	0.2950
Adipose Visceral Omentum	0.0038	0.1214	0.1154	0.3107
Adrenal Gland	0.0052	0.1056	0.0995	0.2943
Artery Aorta	0.0052	0.1092	0.1012	0.3510
Artery Coronary	0.0041	0.1185	0.1133	0.2743
Artery Tibial	0.0040	0.1029	0.0985	0.2341
Brain Amygdala	0.0043	0.0905	0.0903	0.0959
Brain Anterior Cingulate Cortex BA24	0.0020	0.0977	0.0954	0.1683
Brain Caudate Basal Ganglia	0.0015	0.1122	0.1075	0.2598
Brain Cerebellar Hemisphere	0.0030	0.0953	0.0920	0.1975
Brain Cerebellum	0.0033	0.1012	0.0967	0.2435
Brain Cortex	0.0023	0.1091	0.1030	0.2958
Brain Frontal Cortex BA9	0.0027	0.1023	0.0970	0.2653
Brain Hippocampus	0.0036	0.0986	0.0923	0.2965
Brain Hypothalamus	0.0026	0.1177	0.1135	0.2479
Brain Nucleus Accumbens Basal Ganglia	0.0035	0.0906	0.0853	0.2580
Brain Putamen Basal Ganglia	0.0036	0.1120	0.1050	0.3276
Brain Spinal Cord Cervical C-1	0.0047	0.0825	0.0782	0.2119
Brain Substantia Nigra	0.0032	0.1034	0.0958	0.3386
Breast Mammary Tissue	0.0050	0.1239	0.1165	0.3543
Cells Cultered Fibroblasts	0.0049	0.0872	0.0812	0.2683
Cells EBV-Transformed Lymphocytes	0.0055	0.1021	0.0954	0.3141
Colon Sigmoid	0.0035	0.1106	0.1052	0.2739
Colon Transverse	0.0044	0.1154	0.1115	0.2361

Esophagus Gastroesophageal Junction	0.0034	0.1134	0.1072	0.3008
Esophagus Mucosa	0.0034	0.1093	0.1046	0.2518
Esophagus Mucularis	0.0011	0.1100	0.1034	0.3076
Heart Atrial Appendage	0.0035	0.1127	0.1054	0.3316
Heart Left Ventricle	0.0026	0.0961	0.0874	0.3561
Kidney Cortex	0.0039	0.1064	0.1037	0.1860
Liver	0.0051	0.1210	0.1163	0.2663
Lung	0.0031	0.1173	0.1127	0.2598
Minor Salivary Gland	0.0048	0.1140	0.1087	0.2784
Muscle Skeletal	0.0034	0.1147	0.1092	0.2792
Nerve Tibial	0.0055	0.1102	0.1030	0.3301
Ovary	0.0033	0.0972	0.0914	0.2737
Pancreas	0.0044	0.1127	0.1049	0.3500
Pituitary	0.0050	0.0916	0.0854	0.2833
Prostate	0.0030	0.1039	0.0984	0.2759
Skin Not Sun Exposed Subrapubic	0.0038	0.0891	0.0815	0.3227
Skin Sun Exposed Lower Leg	0.0041	0.1037	0.0972	0.3051
Small Intestine Terminal Ileum	0.0023	0.0995	0.0952	0.2314
Spleen	0.0033	0.1053	0.0993	0.2919
Stomach	0.0041	0.1069	0.1000	0.3163
Testis	0.0035	0.0958	0.0907	0.2583
Thyroid	0.0037	0.1112	0.1051	0.3000
Uterus	0.0041	0.0998	0.0959	0.2241
Vagina	0.0026	0.1004	0.0975	0.1895
Whole Blood	0.0026	0.0825	0.0761	0.2816

Table 4. TOPMed MESA π_1 values for every tissue

Tissue	*Trans*-Acting	*Cis*-acting	*Cis*-different	*Cis*-same
Adipose Subcutaneous	0.3517	0.7811	0.7418	0.9264
Adipose Visceral Omentum	0.3265	0.7529	0.7138	0.8893
Adrenal Gland	0.4307	0.8616	0.8309	0.9794
Artery Aorta	0.3410	0.7311	0.6716	0.9786
Artery Coronary	0.2554	0.7736	0.7555	0.8493
Artery Tibial	0.3544	0.8095	0.7720	0.9727
Brain Amygdala	0.4180	0.7772	0.7607	0.8522
Brain Anterior Cingulate Cortex BA24	0.4518	0.7556	0.7179	0.9218
Brain Caudate Basal Ganglia	0.4476	0.8169	0.8065	0.8575
Brain Cerebellar Hemisphere	0.3417	0.7715	0.7545	0.8412
Brain Cerebellum	0.4520	0.8446	0.8376	0.8787
Brain Cortex	0.3934	0.7664	0.7493	0.8585
Brain Frontal Cortex BA9	0.3860	0.8247	0.8037	0.9190
Brain Hippocampus	0.4210	0.8350	0.8267	0.8668
Brain Hypothalamus	0.5522	0.8129	0.8011	0.8714
Brain Nucleus Accumbens Basal Ganglia	0.4337	0.8005	0.7995	0.8059
Brain Putamen Basal Ganglia	0.4909	0.8254	0.7991	0.9309
Brain Spinal Cord Cervical C-1	0.3899	0.8511	0.8393	0.9059
Brain Substantia Nigra	0.3986	0.7339	0.6943	0.9424
Breast Mammary Tissue	0.3469	0.7406	0.7014	0.8935
Cells Cultered Fibroblasts	0.2617	0.8408	0.8322	0.8793
Cells EBV-Transformed Lymphocytes	0.3793	0.7590	0.7545	0.7824
Colon Sigmoid	0.3949	0.7361	0.7297	0.7586
Colon Transverse	0.4455	0.7872	0.7816	0.8119

Esophagus Gastroesophageal Junction	0.3673	0.8216	0.8071	0.8831
Esophagus Mucosa	0.3807	0.8939	0.8826	0.9436
Esophagus Mucularis	0.3357	0.8678	0.8566	0.9063
Heart Atrial Appendage	0.3969	0.8026	0.7969	0.8235
Heart Left Ventricle	0.4327	0.7376	0.7459	0.7082
Kidney Cortex	0.3874	0.8443	0.8251	0.9237
Liver	0.4455	0.7877	0.7741	0.8401
Lung	0.3502	0.7596	0.7305	0.8757
Minor Salivary Gland	0.3688	0.8299	0.8237	0.8622
Muscle Skeletal	0.4151	0.8153	0.7949	0.8896
Nerve Tibial	0.3771	0.7853	0.7520	0.9333
Ovary	0.3564	0.7627	0.7462	0.8350
Pancreas	0.4871	0.8268	0.8238	0.8408
Pituitary	0.3628	0.7373	0.7033	0.8881
Prostate	0.3753	0.8194	0.7918	0.9448
Skin Not Sun Exposed Subrapubic	0.3136	0.8065	0.7784	0.9332
Skin Sun Exposed Lower Leg	0.3816	0.8387	0.8243	0.9001
Small Intestine Terminal Ileum	0.4412	0.8430	0.8232	0.9246
Spleen	0.3356	0.8058	0.7800	0.9175
Stomach	0.4160	0.7056	0.7101	0.6843
Testis	0.4175	0.7689	0.7241	0.9738
Thyroid	0.3648	0.8217	0.8147	0.8516
Uterus	0.3957	0.8009	0.7831	0.8815
Vagina	0.4698	0.8451	0.8246	0.9504
Whole Blood	0.4168	0.9013	0.8902	0.9443

Protein Targets of *Trans*-acting Genes Enriched for TF 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 transcription factors. While the *cis*-targets were not enriched for transcription factor targets, the *trans*-targets were enriched for motifs targeted by transcription factors like *NFKB2*, *RELA*, *NFAT1C*, *FOXF2*, *AR*, *GATA1*, and *STAT1* (Figure 5).

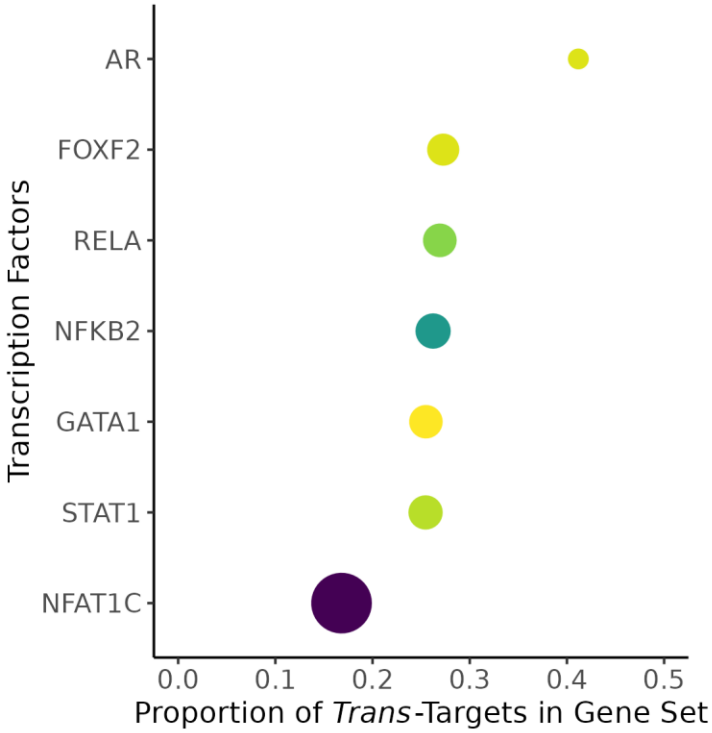


Figure 5. Enrichment of transcription factor binding sites of target proteins of *trans*-acting genes.

The target proteins of *trans*-acting genes were significantly enriched for binding motifs of the transcription factors listed on the y-axis as annotated in the Molecular Signatures Database. 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 p-value of the enrichment test.

While we had prediction models for all of these genes in some or all tissues, only *RELA*, *NFATC*, and *NFKB2* were significantly associated with any target proteins in our TWAS analysis. *RELA* was significantly associated with one *trans*-target, *SHISA3*, which was not annotated in FUMA as having the motif targeted by *RELA*. *NFAT1C* was associated with two *trans*-targets, *PLAG2G5* and *IFNGR2*, both of which were not annotated as targets of *NFAT1C* by FUMA. Finally, *NFKB2* was significantly associated with one *trans*-target, *RRM1*, which was not annotated as a target of *NFKB2* by FUMA.

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 5).

Table 5. *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

Master 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 “master regulatory” loci. These loci are represented through the vertical lines of dots in Figures 2a-c. We discovered 11 distinct master regulatory loci in INTERVAL (Table 6). 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 6). Interestingly, almost none of the targets of the largest master regulatory locus discovered in INTERVAL, the *MYADM* locus on chromosome 19 replicated in TOPMed MESA (Table 6).

We performed a gene set enrichment analysis of the protein targets in INTERVAL of each of these master 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 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 prediction models in many tissues, it was not significantly associated with any of the targets of the *C7* locus in our TWAS analysis.

Table 6. Master 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

Additionally, we performed a gene set enrichment analysis of the master 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 7). Notably, 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 1 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 transcription factor, *SREBF1*. Of the 174 genes in this gene set, we tested 153 in our TWAS and 3 were master 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.

Table 7. Master regulatory genes are enriched for mapped GWAS catalog associations (FDR < 0.5).

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

Predicted Gene Expression Correlates Better with Protein Levels than Observed Gene Expression

We used the RNA-Seq data from TOPMed MESA to test how the correlation of observed gene expression with protein abundance compared to that of predicted gene expression with observed protein abundance. For each of the 3 tissues with observed gene expression data (PBMC, monocytes, T cells), we used PrediXcan to test the *cis*-same associations between the abundance of all 1,300 proteins measured in the study and the observed expression of the genes that encode the proteins. 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 protein levels (FDR < 0.05) associated. 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 3 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, the highest π_1 in a predicted tissue was 0.491 in Brain Putamen Basal Ganglia and the only predicted tissue with a π_1 lower than that of PBMC was the Brain Cerebellar Hemisphere at 0.182 (Table 8). Notably, Whole Blood, the only tissue for which we have both predicted and observed expression levels, albeit in different blood cell types, had a higher π_1 than all three of the observed tissues at 0.331.

Table 8. TOPMed MESA *cis*-same π_1 values for every predicted and observed tissue.

Tissue	*Cis*-same π_1	Tissue	*Cis*-same π_1
Brain Putamen basal ganglia	0.4912	Brain Hippocampus	0.3366
Skin Sun Exposed Lower leg	0.4662	Brain Substantia nigra	0.3319
Small Intestine Terminal Ileum	0.4610	**Whole Blood**	**0.3311**
Kidney Cortex	0.4480	Breast Mammary Tissue	0.3239
Esophagus Muscularis	0.4473	Prostate	0.3223
Uterus	0.4406	Skin Not Sun Exposed Suprapubic	0.3222
Liver	0.4317	Artery Aorta	0.3183
Cells Cultured fibroblasts	0.4195	Brain Cerebellum	0.3168
Adrenal Gland	0.4193	Cells EBV-transformed lymphocytes	0.3129
Minor Salivary Gland	0.4040	Brain Cortex	0.3048
Esophagus Mucosa	0.3960	Lung	0.3031
Testis	0.3945	Vagina	0.3028
Muscle Skeletal	0.3910	Esophagus Gastroesophageal Junction	0.2962
Brain Caudate basal ganglia	0.3896	Brain Nucleus accumbens basal ganglia	0.2905
Heart Atrial Appendage	0.3850	Colon Sigmoid	0.2872
Pituitary	0.3845	Brain Amygdala	0.2848
Brain Hypothalamus	0.3726	Nerve Tibial	0.2764
Colon Transverse	0.3708	Spleen	0.2747
Thyroid	0.3651	Artery Coronary	0.2684
Brain Frontal Cortex BA9	0.3633	Artery Tibial	0.2678
Adipose Subcutaneous	0.3530	Stomach	0.2662
Brain Spinal cord cervical c-1	0.3472	Pancreas	0.2418
Ovary	0.3440	PBMC – observed	0.2393
Adipose Visceral Omentum	0.3427	Monocytes – observed	0.1928
Brain Anterior cingulate cortex BA24	0.3408	Brain Cerebellar Hemisphere	0.1819
Heart Left Ventricle	0.3408	T-cells – observed	0.0768

Note: Observed expression tissues are marked, the rest are predicted expression levels.

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 than observed gene expression for significant *cis*-same genes (Figure 6c-d). We found that predicted tissues closely related to blood, 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).

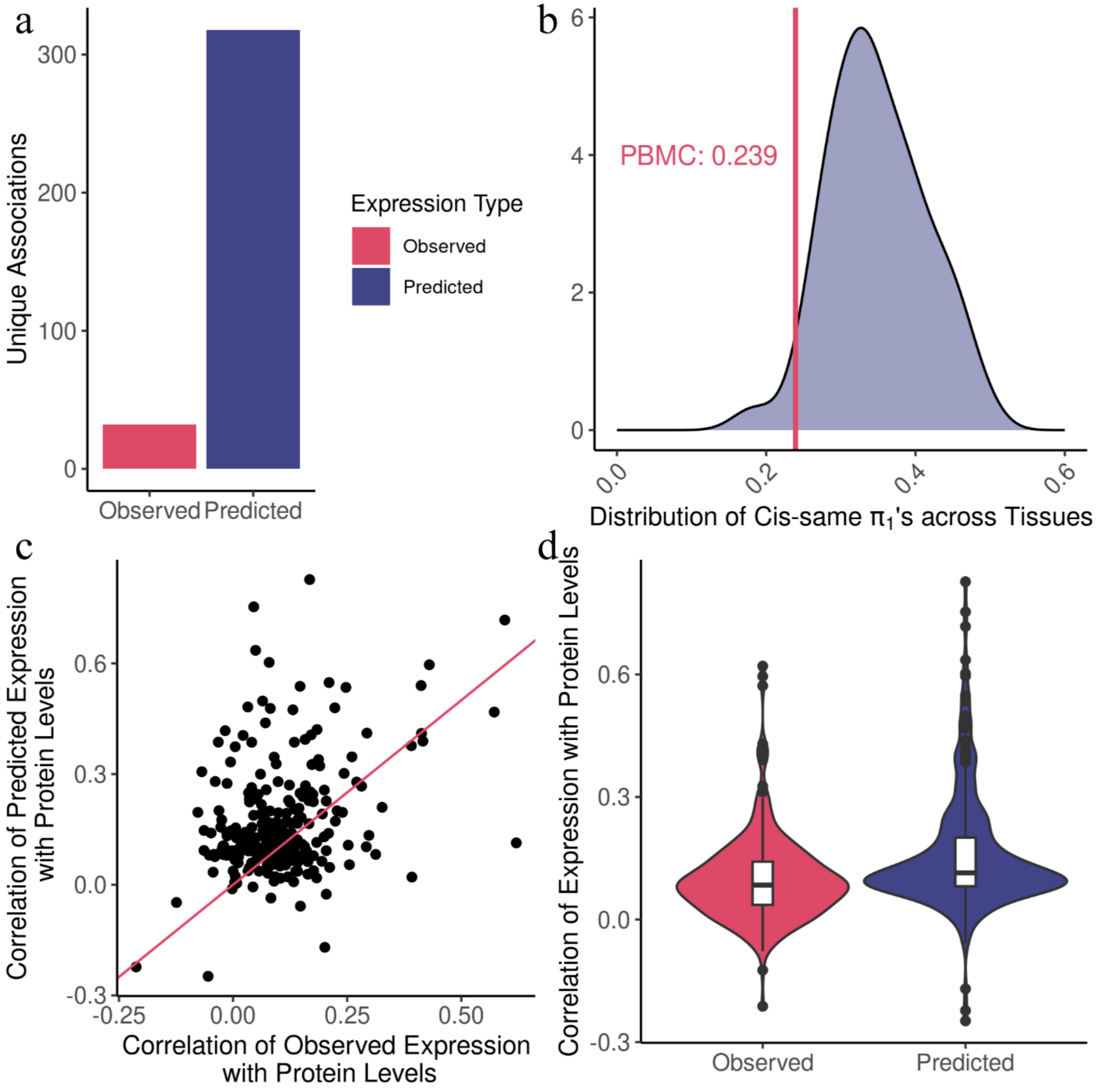


Figure 6. *Cis*-same associations using predicted expression vs. observed expression.

(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) Scatter plot 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.

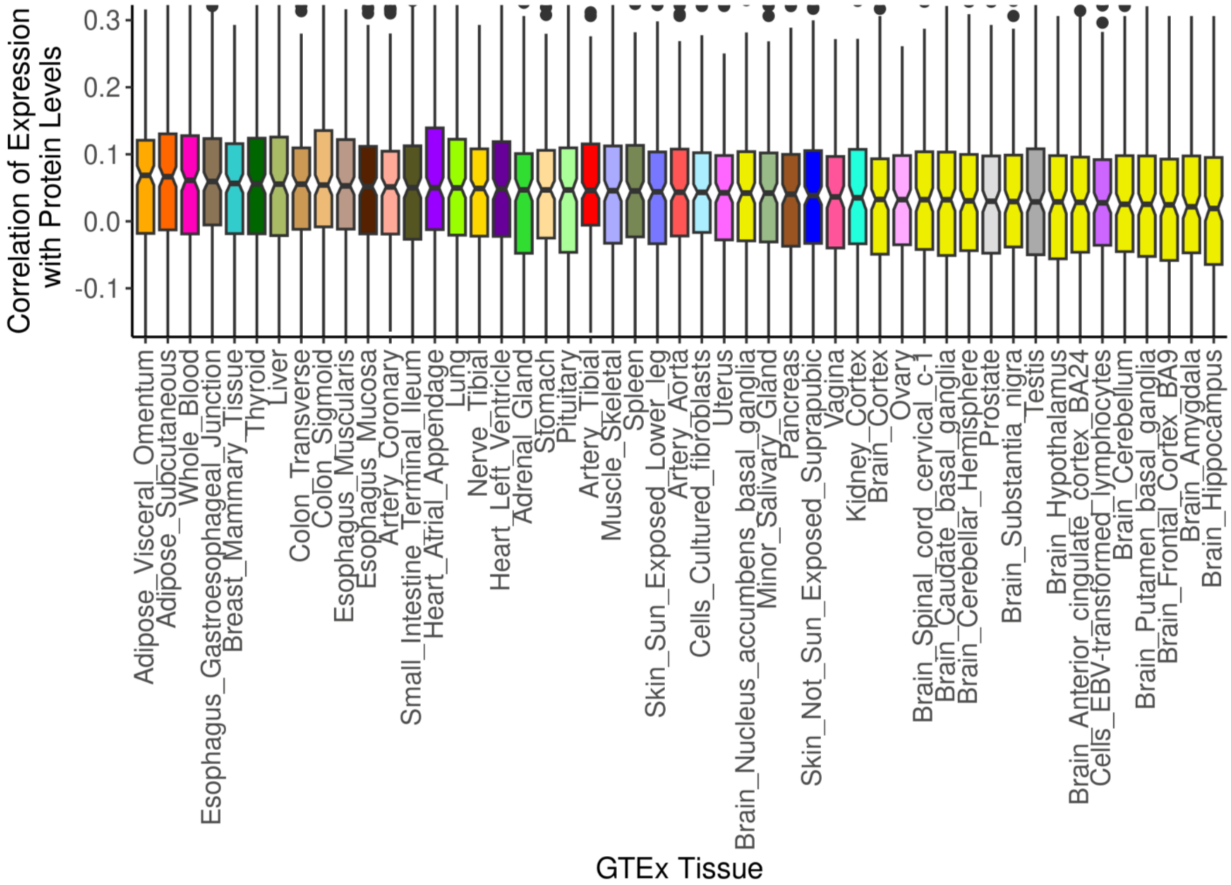


Figure 7. *Cis*-same correlation of predicted expression and protein levels by tissue.

Distribution of Pearson correlation coefficients for correlating predicted expression of significant *cis*-same genes with protein abundance in every GTEx tissue. Figure is truncated in the y-axis at correlation = -0.15 and 0.3.

DISCUSSION AND CONCLUSION

Here, we applied the PrediXcan 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 TWAS method proved more effective at identifying *trans*-acting effects than a regular QTL study.[24] Compared to a *trans*-pQTL study performed in our discovery cohort (INTERVAL) which found 1,104 proteins with *trans*-pQTL[31] ($P < 1.5 \times 10^{-11}$), 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 transcription factor genes.[22] We found that the protein targets of *trans*-acting genes were enriched for transcription factor binding sites, while the *cis*-targets were not, supporting the idea that many *trans*-effects are driven by transcription factor genes. Furthermore, we found that the *cis*-acting associations were shared across more tissue than the *trans*-acting effects, which tended to be more tissue-specific, as has been shown in previous eQTL studies.[19,20]

We identified several loci throughout the genome with strong pleiotropic effects where one gene, or several in LD, 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.[28,31,50–52] Here, we called these regions master regulatory loci and discovered 11 in INTERVAL and 5 that replicated well in TOPMed MESA. It has been shown previously that these *trans*-acting master regulator genes are enriched for GWAS traits, suggesting that *trans*- protein regulation plays an important role in disease variation.[24,51] We performed a gene set enrichment analysis of all of the *trans*-acting genes in each of these master regulatory loci as well as the target proteins of each of these master regulatory loci. We found that the targets and master regulatory genes of many of these loci were enriched for GWAS catalog associations including several autoimmune diseases and other disease phenotypes. For example, the CFHR genes were enriched for autoimmune diseases such as IgA nephropathy and age-related macular degeneration as well as C3 and C4 levels. Studies have shown that the 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 master regulatory genes could lead to the progression of autoimmune diseases.[53]

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.[16–19] Nevertheless, we were able replicate 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.[35–37,54] One of the main issues in correlating expression levels with protein levels is the high fluctuation in these 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.[55] Furthermore, genetically predicted expression levels have been shown to strongly correlate with genetically predicted protein levels.[26]

Here, we show that genetically predicted expression levels correlate better with plasma protein abundance than observed expression levels. We leveraged the PrediXcan framework to test both predicted expression in 49 tissues and observed expression in 3 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. 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 the 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.[26] 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.[26]

One major 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 are inaccurate in that they represent the abundance of multiple different proteins.[39] Another limitation of the study is that our discovery cohort is not diverse, comprising entirely of individuals of European descent, while our replication cohort, which is diverse, has a very small sample size. Both these issues limit our ability to discover and replicate the transcript-protein regulatory relationships.

All of these results highlight the benefits of working with predicted expression over observed expression; it is easier to calculate predicted expression than it is to measure observed expression, by leveraging information from multiple tissues, you can find more significant associations with predicted expression, and by reducing the variation due to environmental influence, the expected true positive rate of significant associations discovered with predicted expression is much higher.

REFERENCE LIST

1. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

2. Shendure, J. *et al.* DNA sequencing at 40: past, present and future. *Nature* **550**, 345–353 (2017).

3. Fine, M. J., Ibrahim, S. A. & Thomas, S. B. The Role of Race and Genetics in Health Disparities Research. *Am. J. Public Health* **95**, 2125–2128 (2005).

4. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

5. Uffelmann, E. *et al.* Genome-wide association studies. *Nat. Rev. Methods Primer* **1**, 1–21 (2021).

6. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).

7. Freedman, M. L. *et al.* Principles for the post-GWAS functional characterization of cancer risk loci. *Nat. Genet.* **43**, 513–518 (2011).

8. Tak, Y. G. & Farnham, P. J. Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. *Epigenetics Chromatin* **8**, 57 (2015).

9. Zhang, F. & Lupski, J. R. Non-coding genetic variants in human disease. *Hum. Mol. Genet.* **24**, R102–R110 (2015).

10. Gresham, D., Dunham, M. J. & Botstein, D. Comparing whole genomes using DNA microarrays. *Nat. Rev. Genet.* **9**, 291–302 (2008).

11. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 (2009).

12. Montgomery, S. B. *et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**, 10.1038/nature08903 (2010).

13. Stranger, B. E. *et al.* Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**, 848–853 (2007).

14. Pickrell, J. K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768–772 (2010).

15. Battle, A. *et al.* Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res.* **24**, 14–24 (2014).

16. THE GTEX CONSORTIUM. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).

17. Fairfax, B. P. *et al.* Genetics of gene expression in primary immune cells identifies cell-specific master regulators and roles of *HLA* alleles. *Nat. Genet.* **44**, 502–510 (2012).

18. Dimas, A. S. *et al.* Common regulatory variation impacts gene expression in a cell type dependent manner. *Science* **325**, 1246–1250 (2009).

19. Liu, X. *et al.* Functional Architectures of Local and Distal Regulation of Gene Expression in Multiple Human Tissues. *Am. J. Hum. Genet.* **100**, 605–616 (2017).

20. Grundberg, E. *et al.* Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat. Genet.* **44**, 1084–1089 (2012).

21. Võsa, U. *et al.* 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 (2021).

22. Yang, F., Wang, J., Pierce, B. L. & Chen, L. S. Identifying cis-mediators for trans-eQTLs across many human tissues using genomic mediation analysis. *Genome Res.* **27**, 1859–1871 (2017).

23. Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**, 1091–1098 (2015).

24. Wheeler, H. E. *et al.* Imputed gene associations identify replicable trans-acting genes enriched in transcription pathways and complex traits. *Genet. Epidemiol.* **43**, 596–608 (2019).

25. Li, B. & Ritchie, M. D. From GWAS to Gene: Transcriptome-Wide Association Studies and Other Methods to Functionally Understand GWAS Discoveries. *Front. Genet.* **12**, (2021).

26. Zhang, J. *et al.* 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 (2022).

27. Hormozdiari, F. *et al.* Leveraging molecular quantitative trait loci to understand the genetic architecture of diseases and complex traits. *Nat. Genet.* **50**, 1041–1047 (2018).

28. Suhre, K. *et al.* Connecting genetic risk to disease end points through the human blood plasma proteome. *Nat. Commun.* **8**, 14357 (2017).

29. Gold, L. *et al.* Aptamer-Based Multiplexed Proteomic Technology for Biomarker Discovery. *PLOS ONE* **5**, e15004 (2010).

30. Rohloff, J. C. *et al.* Nucleic Acid Ligands With Protein-like Side Chains: Modified Aptamers and Their Use as Diagnostic and Therapeutic Agents. *Mol. Ther. - Nucleic Acids* **3**, (2014).

31. Sun, B. B. *et al.* Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018).

32. Ferkingstad, E. *et al.* Large-scale integration of the plasma proteome with genetics and disease. *Nat. Genet.* **53**, 1712–1721 (2021).

33. Brandes, N., Linial, N. & Linial, M. PWAS: proteome-wide association study—linking genes and phenotypes by functional variation in proteins. *Genome Biol.* **21**, 173 (2020).

34. Pietzner, M. *et al.* Mapping the proteo-genomic convergence of human diseases. *Science* **374**, eabj1541 (2021).

35. Gygi, S. P., Rochon, Y., Franza, B. R. & Aebersold, R. Correlation between Protein and mRNA Abundance in Yeast. *Mol. Cell. Biol.* **19**, 1720–1730 (1999).

36. Marguerat, S. *et al.* Quantitative Analysis of Fission Yeast Transcriptomes and Proteomes in Proliferating and Quiescent Cells. *Cell* **151**, 671–683 (2012).

37. Cheng, P. *et al.* Proteogenomic analysis of cancer aneuploidy and normal tissues reveals divergent modes of gene regulation across cellular pathways. *eLife* **11**, e75227 (2022).

38. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).

39. Schubert, R. *et al.* Protein prediction for trait mapping in diverse populations. *PLOS ONE* **17**, e0264341 (2022).

40. Angelantonio, E. D. *et al.* Efficiency and safety of varying the frequency of whole blood donation (INTERVAL): a randomised trial of 45 000 donors. *The Lancet* **390**, 2360–2371 (2017).

41. Bild, D. E. *et al.* Multi-Ethnic Study of Atherosclerosis: Objectives and Design. *Am. J. Epidemiol.* **156**, 871–881 (2002).

42. Mogil, L. S. *et al.* Genetic architecture of gene expression traits across diverse populations. *PLOS Genet.* **14**, e1007586 (2018).

43. Araujo, D. S. *et al.* Multivariate adaptive shrinkage improves cross-population transcriptome prediction for transcriptome-wide association studies in underrepresented populations. *bioRxiv* 2023.02.09.527747 (2023) doi:10.1101/2023.02.09.527747.

44. Barbeira, A. N. *et al.* Exploiting the GTEx resources to decipher the mechanisms at GWAS loci. *Genome Biol.* **22**, 49 (2021).

45. Barbeira, A. N. *et al.* Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat. Commun.* **9**, 1825 (2018).

46. Urbut, S. M., Wang, G., Carbonetto, P. & Stephens, M. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nat. Genet.* **51**, 187–195 (2019).

47. Barbeira, A. N. *et al.* Fine-mapping and QTL tissue-sharing information improves the reliability of causal gene identification. *Genet. Epidemiol.* **44**, 854–867 (2020).

48. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci.* **100**, 9440–9445 (2003).

49. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).

50. Gudjonsson, A. *et al.* A genome-wide association study of serum proteins reveals shared loci with common diseases. *Nat. Commun.* **13**, 480 (2022).

51. Emilsson, V. *et al.* Co-regulatory networks of human serum proteins link genetics to disease. *Science* **361**, 769–773 (2018).

52. Pietzner, M. *et al.* Genetic architecture of host proteins involved in SARS-CoV-2 infection. *Nat. Commun.* **11**, 6397 (2020).

53. Zipfel, P. *et al. CFHR* gene variations provide insights in the pathogenesis of the kidney diseases atypical hemoltyic uremic syndrome and C3 glomerulopathy. *JASN*. **31**, 241-256 (2020).

54. Schwanhäusser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473**, 337–342 (2011).

55. Upadhya, S. R. & Ryan, C. J. Experimental reproducibility limits the correlation between mRNA and protein abundances in tumor proteomic profiles. *Cell Rep. Methods* **2**, 100288 (2022).

VITA

Henry Wittich grew up in the Chicagoland area. He started attending Loyola University Chicago in August 2018 and earned his Bachelor of Science in Bioinformatics, *cum laude* with departmental honors, in 2021. He began working as an undergraduate research assistant in Dr. Wheeler's lab in August 2020 and joined the BS/MS Bioinformatics program at Loyola to continue on with his research through his master's degree. Wittich will graduate with his Master of Science in Bioinformatics degree this August. Looking ahead, Wittich is excited to join the Bioinformatics team at NorthShore University Health System, where he will help develop software to support the personalized medicine program at the hospital.