Imputed gene associations identify replicable *trans*-acting genes enriched in transcription pathways and complex traits

Heather E. Wheeler[1,2,3,✉], **Sally Ploch**[1], **Alvaro N. Barbeira**[4], **Rodrigo Bonazzola**[4], **Alireza Fotuhi Siahpirani**[5], **Ashis Saha**[6], **Alexis Battle**[6,7], **Sushmita Roy**[8], **and Hae Kyung Im**[4,✉]

[1]Department of Biology, Loyola University Chicago, Chicago, IL
[2]Department of Computer Science, Loyola University Chicago, Chicago, IL
[3]Department of Public Health Sciences, Stritch School of Medicine, Loyola University Chicago, Maywood, IL
[4]Section of Genetic Medicine, Department of Medicine, University of Chicago, Chicago, IL
[5]Department of Computer Sciences, University of Wisconsin-Madison, Madison, WI
[6]Department of Computer Science, Johns Hopkins University, Baltimore, MD
[7]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD
[8]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI

Regulation of gene expression is an important mechanism through which genetic variation can affect complex traits. A substantial portion of gene expression variation can be explained by both local (*cis*) and distal (*trans*) genetic variation. Much progress has been made in uncovering *cis*-acting expression quantitative trait loci (*cis*-eQTL), but *trans*-eQTL have been more difficult to identify and replicate. Here we take advantage of our ability to predict the *cis* component of gene expression coupled with gene mapping methods such as PrediXcan to identify high confidence candidate *trans*-acting genes and their targets. That is, we correlate the *cis* component of gene expression with observed expression of genes in different chromosomes. Leveraging the shared *cis*-acting regulation across tissues, we combine the evidence of association across all available GTEx tissues and find 2356 *trans*-acting/target gene pairs with high mappability scores. Reassuringly, *trans*-acting genes are enriched in transcription and nucleic acid binding pathways and target genes are enriched in known transcription factor binding sites. Interestingly, *trans*-acting genes are more significantly associated with height than target or background genes, consistent with percolating *trans* effects. Our scripts and summary statistics are publicly available for future studies of *trans*-acting gene regulation.

gene expression | trans-eQTLs | genetic prediction | complex trait genetics

Correspondence: *hwheeler1@luc.edu, haky@uchicago.edu*

Introduction

Transcription is modulated by both proximal genetic variation (*cis*-acting), which likely affects DNA regulatory elements near the target gene, and distal genetic variation (*trans*-acting), which likely affects regulation of a transcription factor (or coactivator) that goes on to regulate a target gene, often located on a different chromosome from the transcription factor gene. Expression quantitative trait loci (eQTL) mapping has been successful at identifying and replicating SNPs associated with gene expression in *cis*, typically meaning SNPs within 1 Mb of the target gene. Because effect sizes are large enough, around 100 samples in the early eQTL studies could detect replicable associations in the reduced multiple testing space of *cis*-eQTLs (1–3).

trans-eQTLs have been more difficult to replicate because their effect sizes are usually smaller and the multiple testing burden for testing all SNPs versus all genes can be too large to overcome. A few studies have had some success; one, with a discovery cohort of 5311 individuals and a replication cohort of 2775 individuals, identified and replicated 103 *trans*-eQTLs in whole blood (4). Unlike *cis*-eQTLs, *trans*-eQTLs are more likely to be tissue-specific, rather than shared across tissues (5). However, even when *trans*-eQTLs are discovered, the mechanism by which the SNP acts is not immediately clear.

Here, we apply PrediXcan (6) and MultiXcan to map *trans*-acting genes, rather than mapping *trans*-eQTLs (SNPs). Our method provides directionality, that is, whether the *trans*-acting gene activates or represses its target gene. We use genome-transcriptome data sets from the Framingham Heart Study (FHS) (7), Depression Genes and Networks (DGN) cohort (8), and the Genotype-Tissue expression (GTEx) Project (5). We show that our approach, called *trans*-PrediXcan, can identify replicable *trans*-acting regulator/target gene pairs. To leverage sharing of cis-eQTLs across tissues and improve our power to detect more *trans*-acting effects, we combine predicted expression across tissues in our *trans*-MulTiXcan model and show that it increases significant *trans*-acting gene pairs >10-fold.

Pathway analysis reveals the *trans*-acting genes are enriched in transcription and nucleic acid binding pathways and target genes are enriched in known transcription factor binding sites, indicating that our method identifies genes of expected function. We show that *trans*-acting genes are more strongly associated with height than target or background genes, demonstrating that *trans*-acting genes likely play a key role in the biology of complex traits.

Methods

Genome and transcriptome data.

Framingham Heart Study (FHS). We obtained genotype and exon expression array data (7, 9) through application to dbGaP accession phs000007.v29.p1. Genotype imputation and gene level quantification were performed by our group previously (10), leaving 4838 European ancestry individuals with both genotypes and observed gene expression levels for analysis. We used the Affymetrix power tools (APT) suite to perform the preprocessing and normalization steps. First the robust multi-array analysis (RMA) protocol was applied which consists of three steps: background correction, quantile normalization, and summarization (11). The summarized expression values were then annotated more fully using the annotation databases contained in the huex10stprobeset.db (exon-level annotations) and huex10sttranscriptcluster.db (gene-level annotations) R packages available from Bioconductor. The genotype data were then split by chromosome and pre-phased with SHAPEIT (12) using the 1000 Genomes phase 3 panel and converted to vcf format. These files were then submitted to the Michigan Imputation Server (https://imputationserver.sph.umich.edu/start.html) (13, 14) for imputation with the Haplotype Reference Consortium version 1 panel (15). Approximately 2.5M non-ambiguous strand SNPs with MAF > 0.05, imputation R^2 > 0.8 and, to match GTEx gene expression prediction models, inclusion in HapMap Phase II were retained for subsequent analyses.

Depression Genes and Networks (DGN). We obtained genotype and whole blood RNA-Seq data through application to the NIMH Repository and Genomics Resource, Study 88 (8). For all analyses, we used the HCP (hidden covariates with prior) normalized gene-level expression data used for the trans-eQTL analysis in Battle et al. (8) and downloaded from the NIMH repository. Quality control and genotype imputation were performed by our group previously (10), leaving 922 European ancestry individuals with both imputed genotypes and observed gene expression levels for analysis. Briefly, the 922 individuals were unrelated (all pairwise < 0.05) and thus all included in downstream analyses. Imputation of approximately 650K input SNPs (minor allele frequency [MAF] > 0.05, Hardy-Weinberg Equilibrium [P > 0.05], non-ambiguous strand [no A/T or C/G SNPs]) was performed on the Michigan Imputation Server (13, 14) with the following parameters: 1000G Phase 1 v3 ShapeIt2 (no singletons) reference panel, SHAPEIT phasing, and EUR population. Approximately 1.9M non-ambiguous strand SNPs with MAF > 0.05, imputation R^2 > 0.8 and, to match GTEx gene expression prediction models, inclusion in HapMap Phase II were retained for subsequent analyses.

Gene expression prediction models. Elastic net (alpha = 0.5) models built using GTEx V6p genome-transcriptome data from 44 tissues (16) were downloaded from `http://predictdb.org/` from the `GTEx-V6p-HapMap-2016-09-08.tar.gz` archive.

Mappability quality control. Genes with mappability scores less than 0.8 and gene pairs with a positive cross-mappability k-mer count were excluded from our analysis (17). Gene mappability is computed as the weighted average of its exon-mappability and untranslated region (UTR)-mappability, weights being proportional to the total length of exonic regions and UTRs, respectively. Mappability of a k-mer is computed as 1/(number of positions k-mer maps in the genome). For exonic regions, k = 75 and for UTRs, k = 36. Cross-mappability between two genes, A and B, is defined as the number of gene A k-mers (75-mers from exons and 36-mers from UTRs) whose alignment start within exonic or untranslated regions of gene B (17).
In addition, to further guard against false positives, we retrieved RefSeq Gene Summary descriptions from the UCSC hgFixed database on 2018-10-04 and removed genes from our analyses with a summary that contained one or more of the following strings: "paralog", "pseudogene", "retro".

***trans*-PrediXcan.** In order to map *trans*-acting regulators of gene expression, we implemented *trans*-PrediXcan, which consists of two steps. First, we predict gene expression levels from genotype dosages using models trained in independent cohorts and tissues, as in PrediXcan (6). This step gives us an estimate of genetic component of gene expression, $\widehat{GReX}$, for each gene. In the second step, for each $\widehat{GReX}$ estimate, we calculate the correlation between $\widehat{GReX}$ and the observed expression level of each gene located on a different chromosome. As in Matrix eQTL (18), variables were standardized to allow fast computation of the correlation and test statistic. In the discovery phase, we predicted gene expression in the FHS cohort using each of 44 tissue models from the GTEx Project. Significance was assessed via the Benjamini-Hochberg false discovery rate (FDR) method (19), with FDR < 0.05 in each individual tissue declared significant. We tested discovered *trans*-acting/target gene pairs for replication in the DGN cohort and declared those with P<0.05 validated. To estimate the expected true positive rate, we calculate π_1 statistics using the qvalue method (20, 21). π_1 is the expected true positive rate and was estimated by selecting the gene pairs with FDR < 0.05 in FHS and examining their P value distribution in DGN. π_0 is the proportion of false positives estimated by assuming a uniform distribution of null P values and $\pi_1 = 1 - \pi_0$ (21). For comparison to our *trans*-PrediXcan method, we performed traditional *trans*-eQTL analysis in FHS and DGN using Matrix eQTL (18), where *trans* is defined as genes on different chromosomes from each SNP.

***trans*-MulTiXcan.** To determine if jointly modeling the genetic component of gene expression across tissues would increase power to detect *trans*-acting regulators, we applied MulTiXcan (22) to our transcriptome cohorts. In our implementation of MulTiXcan, predicted expression from all available GTEx tissue models (up to 44) were used as explanatory variables. To avoid multicolinearity, we use the first k principal components of the predicted expression in our regression model for association with observed (target) gene expression. We keep the first k principal components out of i principal components estimated where

$$\frac{\lambda_{\max}}{\lambda_i} < 30$$

where λ_i is an eigenvalue in the predicted expression covariance matrix (22). We used an F-test to quantify the significance of the joint fit. We tested *trans*-acting/target gene pairs discovered in FHS (FDR < 0.05) for replication in the DGN cohort and declared those with P < 0.05 validated.

Pathway enrichment analysis. We used FUMA (Functional mapping and annotation of genetic associations) (23) to test for enrichment of biological functions in our top *trans*-acting and target genes. We limited our hypergeometric enrichment tests to Reactome (MSigDB v6.1 c2), Gene Ontology (GO) (MMSigDB v6.1 c5), transcription factor targets (MSigDB v6.1 c3), and GWAS Catalog (e91_r2018-02-06) pathways. We required at least 5 *trans*-acting or target genes to overlap with each tested pathway. For the *trans*-acting gene enrichment tests, there were 182 unique *trans*-acting genes at FDR < 0.05 in FHS and P < 0.05 in DGN (S2 Table) and the background gene set was the 16,185 genes with a MulTiXcan model. For the target gene enrichment tests, there were 211 unique target genes at FDR < 0.05 in FHS and P < 0.05 in DGN (S2 Table) and the background gene set was the 12,445 expressed genes. Pathways with Benjamini-Hochberg FDR < 0.05 were considered significant and reported.

We also tested the larger discovery gene sets from FHS (FDR < 0.05) for enrichment in known transcription factors and signaling proteins. The list of transcription factors were collected from Ravasi et al. (24) and signaling proteins were genes annotated as phosphatases and kinases in Uniprot (25, 26). There were 870 unique *trans*-acting and 1036 unique target genes discovered in FHS. We used the hypergeometric test (`hypergeom` function from `scipy.stats` Python library to determine the significance of enrichment. Given the size of the background gene set, M, number of genes with the property of interest in the background, K, and the size of the selected gene set, N, the hypergeometric test calculates the probability of observing x or more genes in the selected gene set with the property of interest. In our setting, K is the number of genes annotated as a TF or signaling protein and N is the size of the discovery gene sets.

***trans*-acting and target gene association studies with complex traits.** We used height as a representative complex trait because PrediXcan results with large samples sizes were available in two cohorts, UK Biobank (n=500,131) and the GIANT Consortium (n=253,288). We retrieved the PrediXcan results from the gene2pheno.org database (16). We compared the observed vs. expected P value distributions via QQ plots for three groups of genes: *trans*-acting genes discovered in FHS MulTiXcan (FDR < 0.05), target genes discovered in FHS MulTiXcan (FDR < 0.05), and background genes tested in MulTiXcan that were not significant. In GIANT, there were 565 *trans*-acting genes (FHS FDR < 0.05), 694 target genes (FHS FDR < 0.05), and 9894 background genes. In UK Biobank, there were 566 *trans*-acting, 697 target, and 9923 background genes.

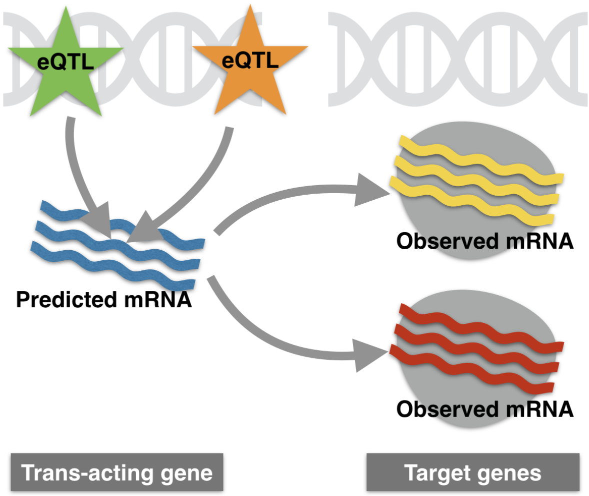


Fig. 1. *trans*-PrediXcan model. mRNA expression levels are predicted from *cis* region eQTLs as in PrediXcan. These predicted expression levels (*trans*-acting genes) are tested for association with observed expression on different chromosomes (target genes).

Results

***Trans*-acting gene discovery and validation with *trans*-PrediXcan.** We sought to map *trans*-acting and target gene pairs by applying the PrediXcan framework to observed expression as traits and term the approach *trans*-PrediXcan (Fig. 1). We excluded genes with poor genome mappability from our analyses (see Methods). We compared *trans*-PrediXcan results between the discovery FHS whole blood cohort (n = 4838) and the validation DGN whole blood cohort (n = 922). We first used PrediXcan (6) to generate a matrix of predicted gene expression from FHS genotypes using prediction models built in GTEx whole blood (16). Then, we calculated the correlation between predicted and observed FHS whole blood gene expression. Examining the correlations of gene pairs on different chromosomes, 55 pairs were significantly correlated in FHS, with an expected true positive rate (π_1) of 0.72 in DGN (Table 1). Gene pair information and summary statistics are shown in S1 Table. Note that the directions of effect of these 55 pairs discovered in FHS are largely consistent in DGN (Fig. 2).

To compare the performance of our *trans*-PrediXcan approach to traditional *trans*-eQTL analysis, we also examined the P-value distribution of top FHS *trans*-eQTLs (FDR < 0.05) in DGN to determine the expected true positive rate. In our SNP-level *trans*-eQTL analysis, π_1 was 0.46, 36% lower than the trans-PrediXcan π_1 of 0.72. We also compared our results to a traditional eQTL study in an independent cohort (4) of similar size to FHS. Of the unique target genes we iden-

Table 1. ***Trans*-acting and target gene pair counts and replication rates across GTEx tissue models.**

Model	FHS FDR < 0.05	FHS Tested	DGN P < 0.05	DGN Tested	DGN π_1
Multi-Tissue (MulTiXcan)	2356	2.0E+08	535	1902	0.49
Whole Blood (PrediXcan)	55	2.4E+07	26	54	0.72

FHS = Framingham Heart Study whole blood cohort, DGN = Depression Genes and Networks whole blood cohort, P = p-value, FDR = Benjamini-Hochberg false discovery rate, π_1 = expected true positive rate.

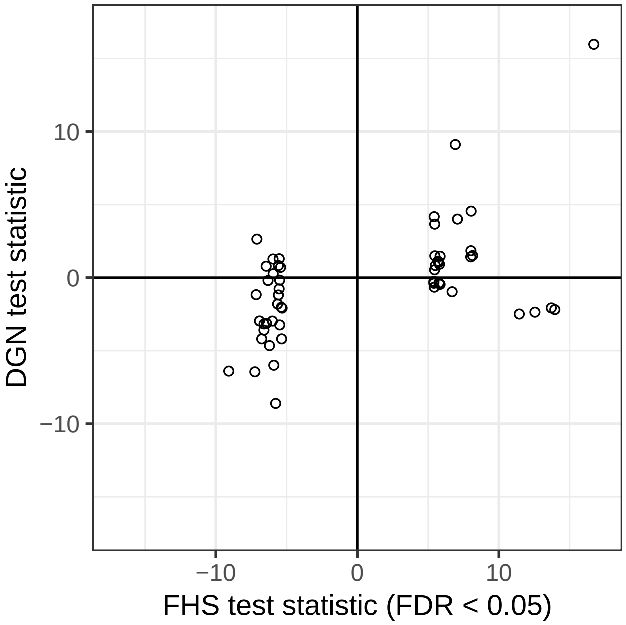


Fig. 2. Test statistic comparison between FHS and DGN results using the GTEx whole blood prediction models. Results of *trans*-acting gene pairs with FDR < 0.05 in the discovery cohort (FHS) are shown for both FHS (x-axis) and the validation cohort DGN (y-axis).

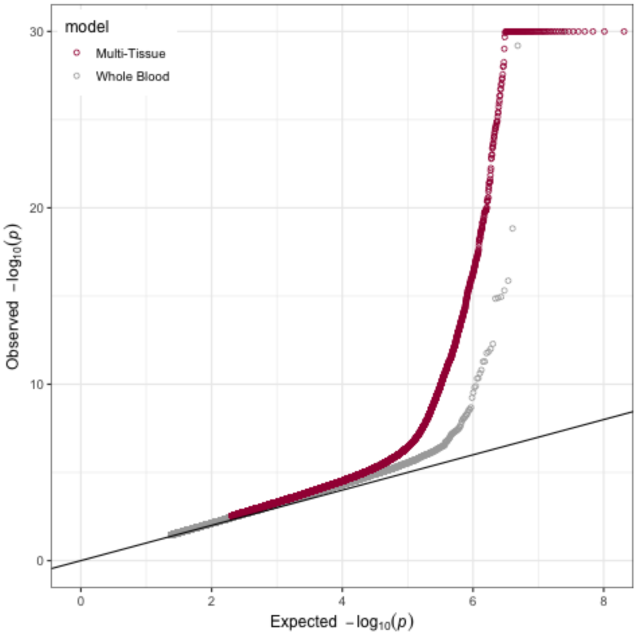


Fig. 3. Multi-Tissue *trans*-MulTiXcan finds more *trans*-acting gene pairs than a single tissue *trans*-PrediXcan (Whole Blood) model. Quantile-quantile plots show an increase in signal in the Multi-Tissue model compared to the Whole Blood model. -$\log_{10}$ p-values are capped at 30 for ease of viewing. The 1e6 most significant P values in each model are plotted to manage file size.

tified in FHS, 3 (7.7%) were also target genes of *trans*-eQTLs (FDR < 0.05) in Westra et al.(4). This is 2.8-fold more than expected by chance as just 2.8% of all genes tested in FHS were *trans*-eQTL targets in Westra et al.(4).

Multi-tissue prediction improves *trans*-acting gene discovery and validation. To leverage tissue sharing of *cis*-eQTLs, we used a multivariable regression approach called MulTiXcan, which accounts for correlation among predicted expression levels across 44 GTEx tissues (22). Notice that even though we seek to detect *trans* regulation, the instruments we are using, i.e. predicted expression, are based on *cis* regulation. Thus it makes sense to combine information across tissues to obtain the best local predictor of gene expression. To address multicolinearity issues, MulTiXcan uses principal component analysis to reduce the number of independent variables to those with the largest variation (22). When we applied *trans*-MulTiXcan to the FHS data, the number of *trans*-acting/target gene pairs increased dramatically (Fig. 3). At FDR < 0.05, there were 2,356 *trans*-acting gene pairs discovered in FHS using the multi-tissue method, while only 55 pairs were discovered with the GTEx whole blood predictors alone (Table 1). We could test 1,902 of these multi-tissue gene pairs for replication in DGN and found 535 of them were significant at P < 0.05 (blue in Fig. 4). Although the expected true positive rate was lower with the MulTiXcan model ($\pi_1 = 0.49$) than with the single tissue model ($\pi_1 = 0.72$), the absolute number of replicate gene pairs was much higher (Table 1). Thus, the number of genes that replicated in both cohorts was 20 times higher in the multi-tissue model compared to the whole blood model (Table 1). Summary statistics of the 2,356 gene pairs discovered in the *trans*-MulTiXcan are available in S2 Table. Of the unique target genes we identified in FHS using MulTiXcan, 93 (10%) were also target genes of *trans*-eQTLs in Westra et al. (FDR < 0.05), which represents 3.6-fold enrichment relative to chance. *trans*-eQTLs within 1Mb of our PrediXcan and MulTiXcan trans-acting genes with the same target genes are shown in S3 Table.

Master *trans*-acting genes associate with many targets. Points that form vertical lines in Figure 4 are indicative of potential master regulators, i.e. genes that regulate many downstream target genes. We defined master regulators as *trans*-acting genes that associate with 50 or more target genes. In our MulTiXcan analysis, we discovered three potential master regulator loci, which are labeled in Figure 4. The most likely master regulator we identified with MulTiX-

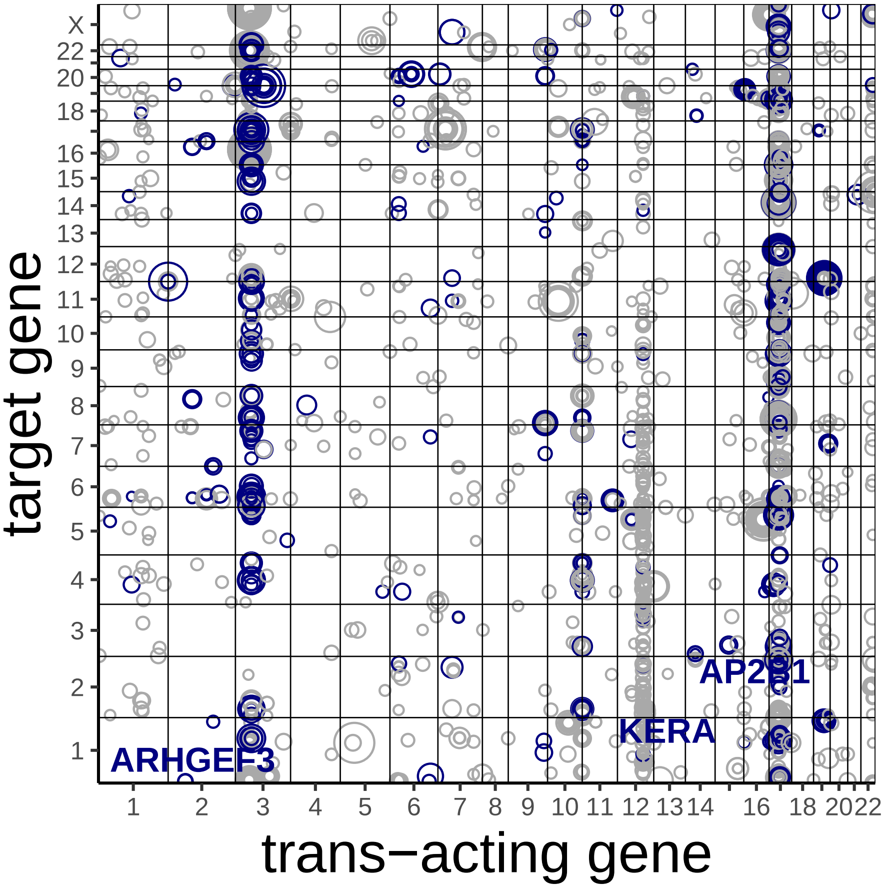


Fig. 4. *Trans*-acting/target gene pairs discovered using MultiXcan in FHS. Each point corresponds to one gene pair (FHS FDR < 0.05) positioned by chromosomal location of the *trans*-acting gene (x-axis) and target gene (y-axis). Size of the point is proportional to the -$\log_{10}$ p-value in FHS. Gene pairs that replicated in DGN MultiXcan (P < 0.05) are colored blue. Master *trans*-acting loci with greater than 50 target genes are labeled.

can is *ARHGEF3* on chromosome 3. *ARHGEF3* associated with 53 target genes in FHS (FDR < 0.05) and 45/51 tested replicated in DGN (P < 0.05). Also, SNPs in *ARHGEF3* have previously been identified as *trans*-eQTLs with multiple target genes. *ARHGEF3* encodes a ubiquitously expressed guanine nucleotide exchange factor. Multiple GWAS and functional studies in model organisms have implicated the gene in platelet formation (27–31). Similarly, SNPs at the chromosome 17 locus we identified have also been identified as *trans*-eQTLs (32) and one study showed the the *trans* effects are mediated by *cis* effects on *AP2B1* expression (30). *AP2B1* encodes a subunit of the adaptor protein complex 2 and GWAS have implicated it in red blood cell and platelet traits (31).

***trans*-acting genes are enriched in transcription factor pathways.** We tested validated *trans*-acting genes for enrichment in Reactome (MSigDB v6.1 c2), Gene Ontology (GO, MSigDB v6.1 c5), transcription factor targets (MSigDB v6.1 c3), and GWAS Catalog (e91_r2018-02-06) pathways using FUMA (23). In our MulTiXcan analysis, there were 174 unique *trans*-acting genes at FDR < 0.05 in FHS and P < 0.05 in DGN (S2 Table). We required at least 5 *trans*-acting genes to overlap with each tested pathway. The background gene set used in the enrichment test were the 15,432 genes with a MulTiXcan model. All pathways with FDR < 0.05 are shown in Table 2 and their gene overlap lists are available in S4 Table.

The top two most significant pathways were the GO nucleic acid binding transcription factor activity pathway and the reactome generic transcription pathway (Table 2). The *trans*-acting genes in each pathway are spread across multiple chromosomes as shown in S1 Figure. *PLAGL1*, which encodes a C2H2 zinc finger protein that functions as a suppressor of cell growth, is a notable *trans*-acting gene in the GO nucleic acid binding transcription factor activity pathway. Of the four *PLAGL1* target genes discovered in FHS, three replicated in DGN (S2 Table). One notable gene in the reactome generic transcription pathway is *MED24*. In our MulTiXcan analysis, *MED24* targeted 13 genes in FHS (FDR < 0.05) and 8/12 replicated in DGN (P < 0.05, S2 Table). *MED24* encodes mediator complex subunit 24. The mediator complex is a transcriptional coactivator complex required for the expression of almost all genes. The mediator complex is recruited by transcriptional activators or nuclear receptors to induce gene expression, possibly by interacting with RNA polymerase II and promoting the formation of a transcriptional pre-initiation complex (33).

We also found a significant enrichment of transcription factors from Ravasi et al. (24) in the 766 unique *trans*-acting genes discovered in FHS with FDR <0.05 (hypergeometric test P = 9.56×10^{-3}). However, the same *trans*-acting genes were not enriched in signaling proteins (P = 0.71).

Target genes are enriched in transcription factor binding sites. We tested MulTiXcan validated target genes for enrichment in the same pathways tested in the *trans*-acting gene analysis. There were 201 unique target genes at FDR < 0.05 in FHS and P < 0.05 in DGN (S2 Table). While just eight pathways were enriched in *trans*-acting genes, 118 pathways were enriched in the target genes (S4 Table). Two of these 118 target gene enriched pathways were transcription factor binding sites (Table 3). No binding motifs were enriched in the *trans*-acting genes. Additional pathways enriched in target genes included several platelet activation and immune response pathways (S4 Table). Target genes were spread across multiple chromosomes (S2 Figure). The target genes were not enriched for reactome generic transcription or GO nucleic acid binding transcription factor activity pathways. The 945 unique target genes discovered in FHS with FDR <0.05 were also not enriched for transcription factors (hypergeometric test P = 0.98) or signaling proteins (P = 0.46) from Ravasi et al. (24).

***Trans*-acting genes are more likely to associate with complex traits.** Using height as a representative complex trait, we compared PrediXcan results among three classes of gene: *trans*-acting, target, and background genes. *trans*-acting and target genes were those discovered in our FHS MultiXcan analysis (FDR < 0.05). Background genes are those tested in MulTiXcan, but not found significant. We examined QQ plots of PrediXcan results for each class in two large studies of height and found that *trans*-acting gene associations are more significant than background gene associations in both the GIANT Consortium and UK Biobank (Fig. 5). Though attenuated in comparison to *trans*-acting genes, target genes are also more significant than background genes for association with height (Fig. 5).

Discussion

We apply the PrediXcan framework to gene expression as a trait (*trans*-PrediXcan approach) to identify *trans*-acting genes that potentially regulate target genes on other chromosomes. We identify replicable predicted gene expression and observed gene expression correlations between genes on different chromosomes. Compared to *trans*-eQTL studies performed in the same cohorts, our *trans*-PrediXcan model shows a higher replication rate for discovered associations. For example, using the GTEx whole blood prediction model we show the expected true positive rate is 0.72 (Table 1). When we performed a traditional *trans*-eQTL study and examined the P-value distribution of top FHS eQTLs (FDR < 0.05) in DGN, the true positive rate was only 0.46. In an independent analysis of the same data, only 4% of eQTLs discovered in FHS replicated in DGN (7).

When predictive models built in 44 different tissues are combined with MulTiXcan, we increase the number of *trans*-acting gene pairs identified in FHS and replicated in DGN 20-fold compared to single-tissue models (Table 1). The SNPs used to predict expression of each gene are all within 1Mb of the gene, i.e. in *cis*. Previous work has shown that *cis*-eQTLs are often shared across many tissues (5). Thus, we show combining *cis*-acting effects across tissues as "replicate experiments" increases our power to detect *trans*-acting associations. Our approach can detect false positives due to linkage disequilibrium and thus colocalization and functional studies is required to reveal the causal *trans*-acting regulator of gene expression.

We found *trans*-acting genes discovered in our MulTiXcan analysis were enriched in transcription pathways and thus previously known to function in transcription regulation. Master regulators revealed by MulTiXcan, *ARHGEF3* and *AP2B1*, were also previously known (30, 32).

In contrast to our results, a recent study concluded trans-eQTLs have limited influence on complex trait biology (34). However, the authors mention limited power in their analyses and found most of the trans-eQTLs examined were not also cis-eQTLs for nearby genes (34). To combat lack of power, others have used *cis*-mediation analysis to identify *trans*-eQTLs (30, 35). Similar to our approach, a mechanism is built in to significant associations found via *cis*-mediation studies: the *cis*-acting locus causes variable expression of the local gene, which in turn leads to variable expression of its target gene on a different chromosome. Unlike *cis*-mediation analysis, our *trans*-PrediXcan approach allows multiple SNPs to work together to affect expression of the *trans*-acting gene and thus may reveal additional associations. A similar method, developed in parallel to ours, combines *cis*-region SNPs using a cross-validation BLUP to identify *trans*-acting genes within one eQTL cohort (36). Our findings have the advantage of discovery in a larger cohort, multiple tissue integration, and replication in an independent cohort.

Our transcriptome association scan presented here integrates gene expression prediction models from multiple tissues and validates results in an independent cohort. Encouragingly, the *trans*-acting and target genes we identify are enriched in transcription and transcription factor pathways.

Using height as a representative complex trait, *trans*-acting gene PrediXcan associations with height are more significant than target and background gene associations in both UK Biobank and the GIANT Consortium. This suggests percolating effects of *trans*-acting genes through target genes. We make our scripts and summary statistics available for future studies of *trans*-acting gene regulation at `https://github.com/WheelerLab/trans-PrediXcan`.

ACKNOWLEDGEMENTS

This work is supported by the NIH/NHGRI Academic Research Enhancement Award R15 HG009569 (Wheeler), NIH/NIMH R01 MH107666 (Im), and NIH/NIMH R01 MH109905 (Battle). The authors declare that there are no conflicts of interest.

Table 2. Validated *trans*-acting genes (MulTiXcan FDR < 0.05 in FHS and P < 0.05 in DGN) are enriched in transcription and GWAS pathways.

Source	GeneSet	N	n	P-value	adjusted P
GO molecular functions	GO nucleic acid binding transcription factor activity	1000	33	6.40e-09	5.76e-06
Reactome	Reactome generic transcription pathway	292	15	2.18e-07	1.47e-04
GWAS catalog	Reticulocyte count	138	10	4.80e-07	1.47e-04
GWAS catalog	Reticulocyte fraction of red cells	147	9	6.76e-06	2.75e-03
GWAS catalog	White blood cell count	152	8	5.90e-05	1.57e-02
GWAS catalog	Neuroticism	134	7	1.44e-04	2.20e-02
GWAS catalog	Platelet count	228	9	2.78e-04	3.40e-02
GWAS catalog	Crohn's disease	525	15	3.19e-04	3.54e-02

FHS = Framingham Heart Study, DGN = Depression Genes and Networks cohort, N = number of genes in GeneSet tested for *trans*-acting effects, n = number of validated genes in GeneSet, P-value = FUMA (23) enrichment P, adjusted P = enrichment Benjamini-Hochberg false discovery rate

Table 3. Validated target genes (MulTiXcan FDR < 0.05 in FHS and P < 0.05 in DGN) are enriched in transcription factor (TF) binding sites in the regions spanning up to 4 kb around their transcription starting sites (MSigDB v6.1 c3).

TF binding site	N	n	P-value	adjusted P
WGGAATGY_TEF1_Q6	247	14	2.23e-05	1.37e-02
PAX8_B	68	6	1.56e-04	4.79e-02

FHS = Framingham Heart Study, DGN = Depression Genes and Networks cohort, N = number of genes in GeneSet tested for target gene effects, n = number of validated genes in GeneSet, P-value = FUMA (23) enrichment P, adjusted P = enrichment Benjamini-Hochberg false discovery rate

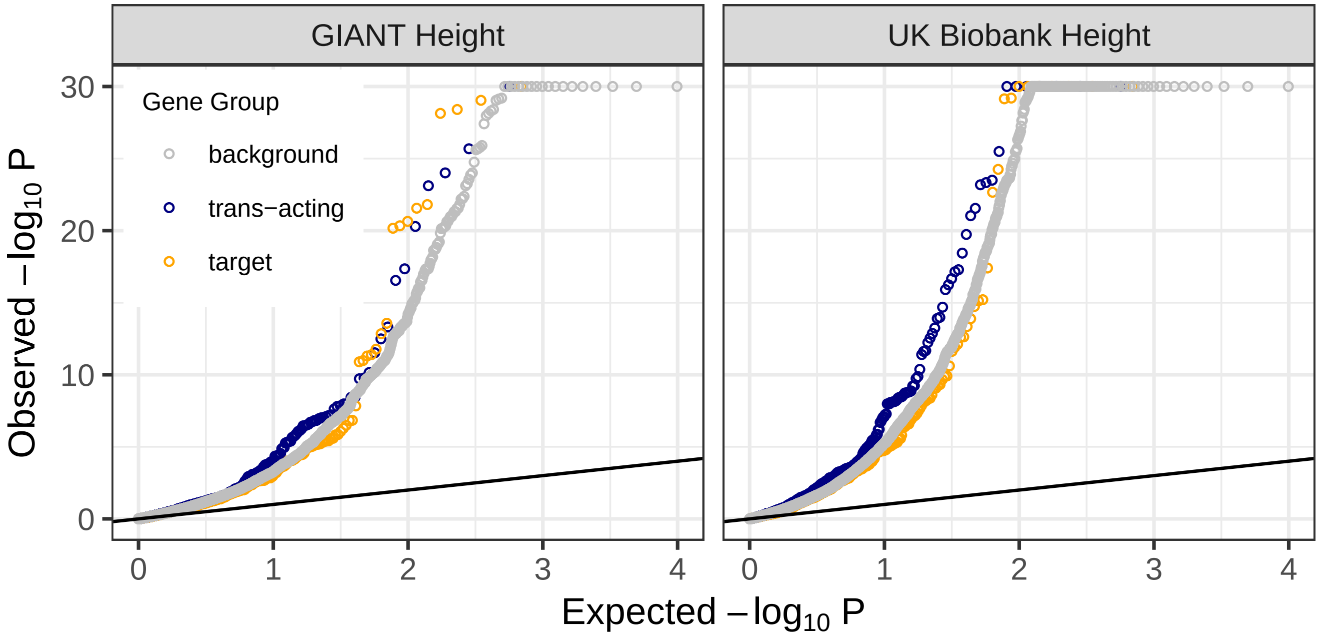


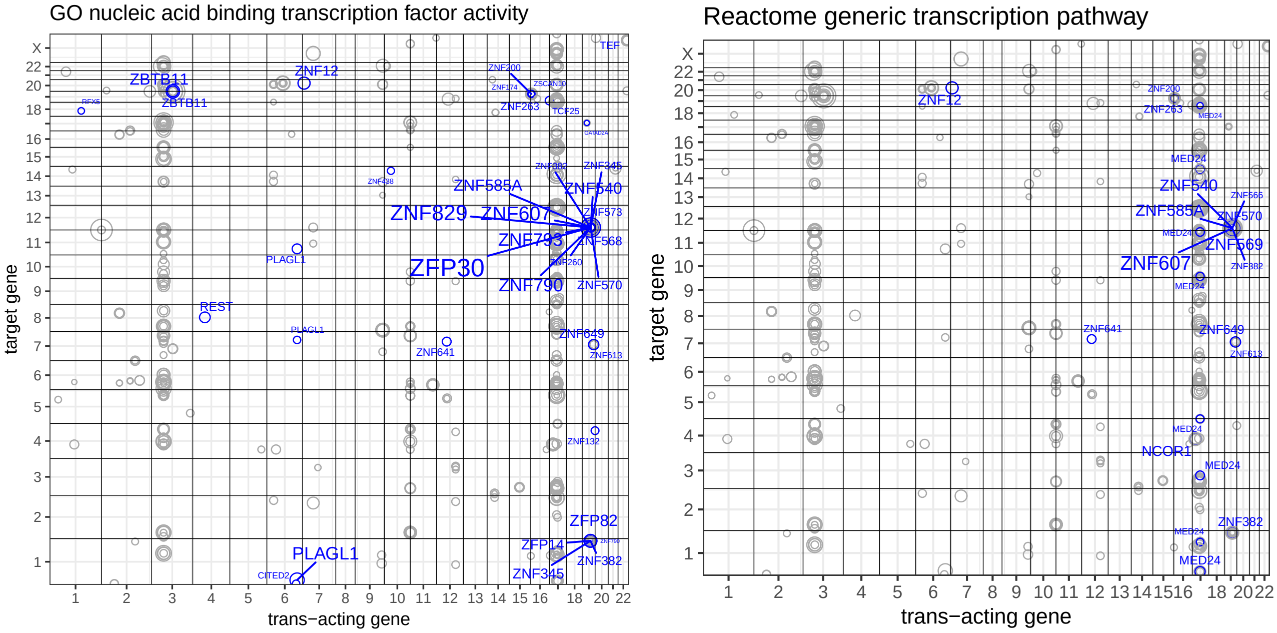
Fig. 5. Height associated genes are enriched for *trans*-acting genes. Quantile-quantile plots of S-PrediXcan results for height from the GIANT Consortium and the UK Biobank show an increase in signal for *trans*-acting genes (FHS MulTiXcan FDR < 0.05) compared to target genes (FHS MulTiXcan FDR < 0.05) and background (tested in MulTiXcan, but not significant) genes. $-\log_{10}$ p-values are capped at 30 for ease of viewing. In GIANT, there were 565 *trans*-acting genes (FHS FDR < 0.05), 694 target genes (FHS FDR < 0.05), and 9894 background genes (tested in MulTiXcan). In UK Biobank, there were 566 *trans*-acting, 697 target, and 9923 background genes.

References

1. Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, Burdick JT. Mapping determinants of human gene expression by regional and genome-wide association. Nature. 2005;437(7063):1365–1369. doi:10.1038/nature04244.
2. Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, et al. Population genomics of human gene expression. Nature genetics. 2007;39(10):1217–24. doi:10.1038/ng2142.
3. Myers AJ, Gibbs JR, Webster JA, Rohrer K, Zhao A, Marlowe L, et al. A survey of genetic human cortical gene expression. Nature Genetics. 2007;39(12):1494–1499. doi:10.1038/ng.2007.16.
4. Westra HJ, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nature genetics. 2013;45(10):1238–43. doi:10.1038/ng.2756.
5. Aguet F, Ardlie KG, Cummings BB, Gelfand ET, Getz G, Hadley K, et al. Genetic effects on gene expression across human tissues. Nature. 2017;550(7675):204–213. doi:10.1038/nature24277.
6. Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ, et al. A gene-based association method for mapping traits using reference transcriptome data. Nature genetics. 2015;47(9):1091–1098. doi:10.1038/ng.3367.
7. Joehanes R, Zhang X, Huan T, Yao C, Ying Sx, Nguyen QT, et al. Integrated genome-wide analysis of expression quantitative trait loci aids interpretation of genomic association studies. Genome Biology. 2017;18(1):16. doi:10.1186/s13059-016-1142-6.

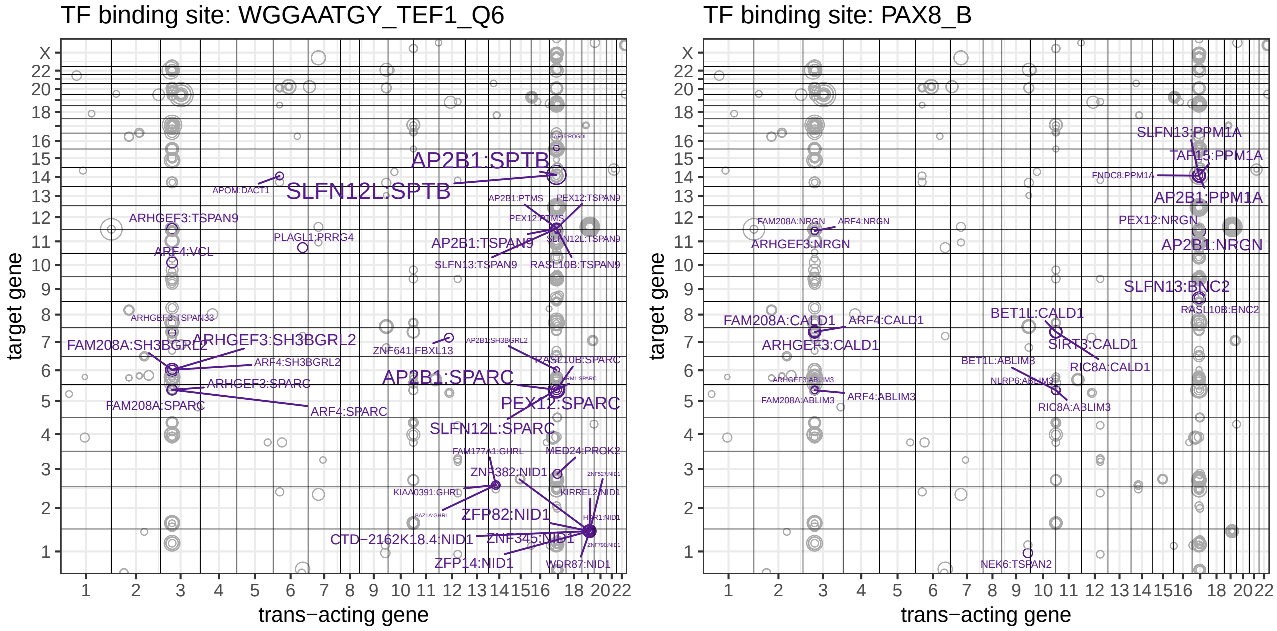
8. Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, McCormick C, et al. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Research. 2014;24(1):14–24. doi:10.1101/gr.155192.113.
9. Zhang X, Joehanes R, Chen BH, Huan T, Ying S, Munson PJ, et al. Identification of common genetic variants controlling transcript isoform variation in human whole blood. Nature Genetics. 2015;47(4):345–352. doi:10.1038/ng.3220.
10. Wheeler HE, Shah KP, Brenner J, Garcia T, Aquino-Michaels K, Cox NJ, et al. Survey of the Heritability and Sparse Architecture of Gene Expression Traits across Human Tissues. PLoS Genetics. 2016;12(11):e1006423. doi:10.1371/journal.pgen.1006423.
11. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. Nucleic acids research. 2003;31(4):e15. doi:10.1093/nar/gng015.
12. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. Nature Methods. 2012;9(2):179–181. doi:10.1038/nmeth.1785.
13. Fuchsberger C, Abecasis GR, Hinds DA. minimac2: faster genotype imputation. Bioinformatics. 2015;31(5):782–784. doi:10.1093/bioinformatics/btu704.
14. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nature Genetics. 2012;44(8):955–959. doi:10.1038/ng.2354.
15. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nature Genetics. 2016;48(10):1279–1283. doi:10.1038/ng.3643.
16. Barbeira A, Dickinson SP, Torres JM, Torstenson ES, Zheng J, Wheeler HE, et al. Integrating tissue specific mechanisms into GWAS summary results. bioRxiv. 2017; p. 045260. doi:10.1101/045260.
17. Saha A, Kim Y, Gewirtz ADH, Jo B, Gao C, McDowell IC, et al. Co-expression networks reveal the tissue-specific regulation of transcription and splicing. Genome Research. 2017;27(11):1843–1858. doi:10.1101/gr.216721.116.
18. Shabalin AA. Matrix eQTL: Ultra fast eQTL analysis via large matrix operations. Bioinformatics. 2012;28(10):1353–1358. doi:10.1093/bioinformatics/bts163.
19. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing; 1995.
20. Aguet F, Brown AA, Castel SE, Davis JR, He Y, Jo B, et al. Genetic effects on gene expression across human tissues. Nature. 2017;550(7675):204–213. doi:10.1038/nature24277.
21. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences. 2003;100(16):9440–9445. doi:10.1073/pnas.1530509100.
22. Barbeira AN, Pividori MD, Zheng J, Wheeler HE, Nicolae DL, Im HK. Integrating predicted transcriptome from multiple tissues improves association detection. bioRxiv. 2018; p. 292649. doi:10.1101/292649.
23. Watanabe K, Taskesen E, Van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. Nature Communications. 2017;8(1):1826. doi:10.1038/s41467-017-01261-5.
24. Ravasi T, Suzuki H, Cannistraci CV, Katayama S, Bajic VB, Tan K, et al. An Atlas of Combinatorial Transcriptional Regulation in Mouse and Man. Cell. 2010;140(5):744–752. doi:10.1016/j.cell.2010.01.044.
25. Roy S, Lagree S, Hou Z, Thomson JA, Stewart R, Gasch AP. Integrated Module and Gene-Specific Regulatory Inference Implicates Upstream Signaling Networks. PLoS Computational Biology. 2013;9(10). doi:10.1371/journal.pcbi.1003252.
26. Apweiler R, Martin MJ, O'Donovan C, Michele Magrane, Yasmin Alam-Faruque, Ricardo Antunes, et al. Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Research. 2012;40(D1):D71–D75. doi:10.1093/nar/gkr981.
27. Gieger C, Radhakrishnan A, Cvejic A, Tang W, Porcu E, Pistis G, et al. New gene functions in megakaryopoiesis and platelet formation. Nature. 2011;480(7376):201–208. doi:10.1038/nature10659.
28. Schramm K, Marzi C, Schurmann C, Carstensen M, Reinmaa E, Biffar R, et al. Mapping the Genetic Architecture of Gene Regulation in Whole Blood. PLoS ONE. 2014;9(4):e93844. doi:10.1371/journal.pone.0093844.
29. Zhang X, Gierman HJ, Levy D, Plump A, Dobrin R, Goring HH, et al. Synthesis of 53 tissue and cell line expression QTL datasets reveals master eQTLs. BMC Genomics. 2014;15(1):532. doi:10.1186/1471-2164-15-532.
30. Yao C, Joehanes R, Johnson AD, Huan T, Liu C, Freedman JE, et al. Dynamic Role of trans Regulation of Gene Expression in Relation to Complex Traits. The American Journal of Human Genetics. 2017;100(4):571–580. doi:10.1016/J.AJHG.2017.02.003.
31. Astle WJ, Elding H, Jiang T, Allen D, Ruklisa D, Mann AL, et al. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. Cell. 2016;167(5):1415–1429. doi:10.1016/J.CELL.2016.10.042.
32. Kirsten H, Al-Hasani H, Holdt L, Gross A, Beutner F, Krohn K, et al. Dissecting the genetics of the human transcriptome identifies novel trait-related <i>trans</i> -eQTLs and corroborates the regulatory relevance of non-protein coding loci. Human Molecular Genetics. 2015;24(16):4746–4763. doi:10.1093/hmg/ddv194.
33. Gustafsson CM, Samuelsson T. Mediator–a universal complex in transcriptional regulation. Molecular microbiology. 2001;41(1):1–8.
34. Yap CX, Lloyd-Jones L, Holloway A, Smartt P, Wray NR, Gratten J, et al. Trans-eQTLs identified in whole blood have limited influence on complex disease biology. European Journal of Human Genetics. 2018; p. 1. doi:10.1038/s41431-018-0174-7.
35. Yang F, Wang J, GTEx Consortium TG, Pierce BL, Chen LS, Aguet F, et al. Identifyingcis-mediators fortrans-eQTLs across many human tissues using genomic mediation analysis. Genome research. 2017;27(11):1859–1871. doi:10.1101/gr.216754.116.
36. Liu X, Mefford JA, Dahl A, Subramaniam M, Battle A, Price AL, et al. GBAT: a gene-based association method for robust trans-gene regulation detection. bioRxiv. 2018; p. 395970. doi:10.1101/395970.

Supplementary Information



A. S1 Figure.

trans-acting genes discovered and validated with MulTiXcan are enriched in transcription pathways. Shown are the MulTiXcan validated (FHS FDR <0.05 and DGN P <0.05) results plotted by chromosomal position. The size of the point is proportional to -$\log_{10}$ p-value. *Trans*-acting genes in the title pathway are labeled in blue.



B. S2 Figure.

Target genes discovered and validated with MulTiXcan are enriched in transcription factor binding site pathways. Shown are the MulTiXcan validated (FHS FDR <0.05 and DGN P <0.05) results plotted by chromosomal position. The size of the point is proportional to -$\log_{10}$ p-value. Gene pairs with the target gene containing the title binding site are labeled in purple (trans-acting gene:target gene).

C. S1 Table.

trans-PrediXcan whole blood model results in FHS (FDR < 0.05) and DGN.

D. S2 Table. *trans*-MulTiXcan results in FHS (FDR < 0.05) and DGN.

E. S3 Table. Westra et. al (4) replication of FHS *trans*-MulTiXcan and *trans*-PrediXcan whole blood model gene pairs (FDR < 0.05).

F. S4 Table. FUMA gene set enrichment results of validated (FHS FDR < 0.05 and DGN P < 0.05) *trans*-acting and target genes.