Trans-eQTL mapping in gene sets identifies network effects of genetic variants

Lili Wang[1,2], Nikita Babushkin[2], Zhonghua Liu[3], Xuanyao Liu[1,2,4*]

[1]The Committee on Genetics, Genomics and Systems Biology, University of Chicago, Chicago, IL, USA
[2]Department of Medicine, Section of Genetic Medicine, University of Chicago, Chicago, IL, USA
[3]Department of Biostatistics, Columbia University, New York, NY, USA
[4]Department of Human Genetics, University of Chicago, Chicago, IL, USA

*Correspondence to xuanyao@uchicago.edu

Abstract

Nearly all trait-associated variants identified in GWAS are non-coding. The *cis* regulatory effects of these variants have been extensively characterized, but how they impact gene regulation in *trans* has been the subject of much fewer studies. Mapping *trans* genetic effects is very challenging because their effect sizes tend to be small and a large multiple testing burden reduces the power to detect them. In addition, read mapping biases can lead to many false positives. To reduce mapping biases and substantially improve power to map trans-eQTLs, we developed a pipeline called trans-PCO, which combines careful read and gene filters with a principal component (PC)-based multivariate association test. Our simulations demonstrate that trans-PCO substantially outperforms existing trans-eQTL mapping methods, including univariate and primary PC-based methods. We applied trans-PCO to two gene expression datasets from whole blood, DGN (N = 913) and eQTLGen (N = 31,684), to identify *trans*-eQTLs associated with gene co-expression networks and hallmark gene sets representing well-defined biological processes. In total, we identified 14,985 high-quality *trans*-eQTLs associated with 197 co-expression gene modules and biological processes. To better understand the effects of trait-associated variants on gene regulatory networks, we performed colocalization analyses between GWAS loci of 46 complex traits and *trans*-eQTLs identified in DGN. We highlight several examples where our map of *trans* effects helps us understand how trait-associated variants impact gene regulatory networks and biological pathways. For example, we found that a locus associated with platelet traits near *ARHGEF3 trans*-regulates a set of co-expressed genes significantly enriched in the platelet activation pathway. Additionally, six red blood cell trait-associated loci *trans*-regulate a gene set representing heme metabolism, a crucial process in erythropoiesis. In conclusion, trans-PCO is a powerful and reliable tool that detects *trans* regulators of cellular pathways and networks, which opens up new opportunities to learn the impact of trait-associated loci on gene regulatory networks.

Main

Genome-wide association studies (GWAS) have identified tens of thousands of genetic loci associated with a large number of complex traits and diseases. More than 90% of GWAS loci are located in non-coding regions of the genome and are thought to affect human traits by regulating gene expression[1–5]. Nearly all studies to date have focused on understanding the effects of trait-associated variants on gene expression in *cis*, which only include effects on genes that are near the associated loci. However, multiple lines of evidence suggest *cis*-regulatory effects only capture a small proportion of the heritability of complex traits and diseases. For example, Yao et al.[6] estimated that only an average of 11% of trait heritability is explained by *cis*-genetic effects on gene expression levels. We previously hypothesized that *trans*-eQTLs, despite having very small effects on each individual gene, may cumulatively account for a large proportion of trait variance[7]. Indeed, our modeling indicates that *trans*-eQTL effects account for twice as much genetic variance in complex traits as *cis*-eQTL effects[7]. Thus, establishing a representative map of genetic variants and their *trans* effects is a critical step toward understanding complex trait and disease genetics.

Two major challenges have precluded *trans*-eQTL discovery to date. First, *trans*-eQTL mapping is extremely prone to false positives due to mapping errors that cause short sequences to map to homologous regions of the genome[8]. This causes spurious associations between the mapping coverages at multiple homologous regions, which result in strong but artificial *trans*-eQTL signals if unaccounted for. The second challenge is by far more difficult to overcome: *trans*-eQTLs are challenging to detect compared to *cis*-eQTLs because (i) they have much smaller effect sizes than *cis*-eQTLs[7], and (ii) a genome-wide search of *trans*-eQTLs involves a huge number of statistical tests ($\sim 10^6$ SNPs$\times \sim 20k$ genes) resulting in a heavy burden of multiple testing corrections.

Previous work in yeast and human cells suggests that *trans*-eQTLs generally affect the expression levels of multiple genes. In particular, Albert et al.[9] found that the 90% of *trans*-eQTLs in their yeast segregant system could be mapped to just 102 hotspot loci that regulate a median of 425 genes each. In fact, three hotspots were found to affect over half of the 5600 expressed *trans*-regulated genes, indicating that some *trans*-eQTLs may have significant genome-wide effects. In humans, the largest *trans*-eQTL study to date from eQTLGen[10] (n = 31,684) identified 59,786 *trans*-eQTL signals for 3,853 SNPs, indicating that each locus affects at least 15 genes on average. Thus, compared to the traditional approach of testing the association between genetic variants and the expression level of a single gene[11,12], testing *trans* associations between genetic variants and the expression levels of a group of genes can greatly improve the power to identify *trans*-eQTLs.

Indeed, many disease-associated loci are “peripheral master regulators”, which regulate multiple genes in the core disease-relevant pathways[7,13]. For example, *KLF14*, which is significantly associated with type 2 diabetes, is a peripheral master regulator that modulates the expression of 385 genes involved in lipid metabolism[14]. Other examples include the *FTO/IRX3/IRX5* locus in obesity[15] or the p53 tumor-suppressor gene in cancers[16]. A method that can detect a large number

of *trans*-eQTLs associated with multiple genes in gene networks would allow functional interpretation of more disease associated loci and shed light on the underlying mechanisms.

A naïve way to detect *trans*-eQTLs associated with multiple genes is to use the primary gene expression principal component (PC) to represent the genes[17] and use it as the response variable in association tests. However, while the primary PC captures the largest amount of total variance in gene expression levels, it can be less powerful or even powerless in detecting significant associations than higher-order PCs, because the direction of the genetic effects on the genes may not align with the primary PC[18,19]. It is also difficult to predict which higher-order PC has the highest power[19]. To combat this, we propose trans-PCO, a flexible approach that uses the PCA-based omnibus test[19] (PCO) to combine multiple PCs and improve power to detect *trans*-eQTLs. Trans-PCO tackles both major challenges in *trans*-eQTL mapping. First, trans-PCO carefully filters sequencing reads and genes based on mappability across different regions of the genome to avoid false positives due to mis-mapping. Second, trans-PCO uses a novel multivariate association test[19] to detect genetic variants with effects on multiple genes in predefined sets and captures genetic effects on multiple PCs. By default, trans-PCO defines sets of genes based on co-expression gene modules as identified by WGCNA[20]. It also accepts user-defined sets; for example, genes that belong to the same gene ontology[21], KEGG pathway[22], or protein complex[23].

We applied trans-PCO to RNA-sequencing data from the Depression Genes and Networks study[24] (DGN, sample size N = 913) and the summary-level statistics from the eQTLGen study[10] (sample size N = 31,684) to identify *trans*-eQTLs associated with co-expression gene modules and well-defined biological processes in whole blood. In total, trans-PCO identified 14,985 high-quality *trans*-eQTLs associated with 197 co-expression gene modules and biological processes. We also performed colocalization analysis[25] of GWAS loci of 46 complex traits and *trans*-eQTLs, in order to explore how trait-associated variants impact gene regulatory networks and pathways in *trans*. All *trans*-eQTLs that are associated with gene co-expression networks and biological pathways can be found in www.networks-liulab.org/transPCO. In summary, trans-PCO is a powerful method for mapping *trans*-eQTLs associated with multiple genes and facilitates functional interpretation of trans-signals. Ultimately, these *trans*-eQTLs should elucidate the *trans*-regulatory mechanisms underlying complex traits and diseases.

Results

Overview of the method

We developed the trans-PCO pipeline to detect *trans*-eQTLs that are associated with the expression levels of a group of genes (gene module) by using a PC-based multivariate association test[19] that combines multiple gene expression PCs. The trans-PCO method consists of three main steps (Figure 1).

First, trans-PCO pre-processes RNA-seq data to reduce false positive *trans*-eQTL associations due to read multimapping errors. Specifically, trans-PCO removes all sequencing reads mapped to low mappability regions of the genome (mappability score <1, Methods) before profiling gene expression levels. These procedures substantially reduce the occurrence of false positive *trans*-eQTLs due to sequencing alignment errors[8,26]. When only summary-level data are available (e.g., eQTLGen dataset[10]), trans-PCO dynamically excludes from the module any genes that are cross-mappable to genes within 100 kb of the tested SNP.

Second, trans-PCO groups genes into clusters, which alleviates the burden of multiple testing by reducing the number of statistical tests and thus increases the statistical power. By default, trans-PCO determines the gene groupings by using WGCNA[20] to identify co-expression modules from gene expression levels (see Methods). We remove covariates and confounders, such as batch effects, gene expression PCs, and cell type proportions (e.g., DGN dataset[24]), from gene expression levels before grouping gene modules. This step is necessary to ensure that the gene modules are not primarily driven by confounding factors. Trans-PCO also allows customization of the gene groups or sets, i.e., genes in the same pathway or protein-protein interaction network[21–23] can be grouped into user-defined gene modules.

Lastly, trans-PCO tests for association between each SNP and the expression levels of the genes in each gene module by adapting the PCO method, which combines multiple gene expression PCs by using six PC-based statistical tests: PCMinP, PCFisher, PCLC, WI, Wald and VC (see Methods for details). Each PC-based test combines multiple PCs uniquely, which allow signals under various genetic architectures to be captured. PCO evaluates the six PC-based tests and takes the minimum p-value as the final test statistic (see Methods)[19]. Only PCs with eigenvalues $\lambda_k > 0.1$ are used in trans-PCO (See Supplementary Note). To avoid identifying associations driven by *cis*-effects, we excluded from the module all genes on the same chromosome as the test SNP. To correct for multiple testing, we performed 10 permutations to establish an empirical null distribution of p-values (See Supplementary Note).

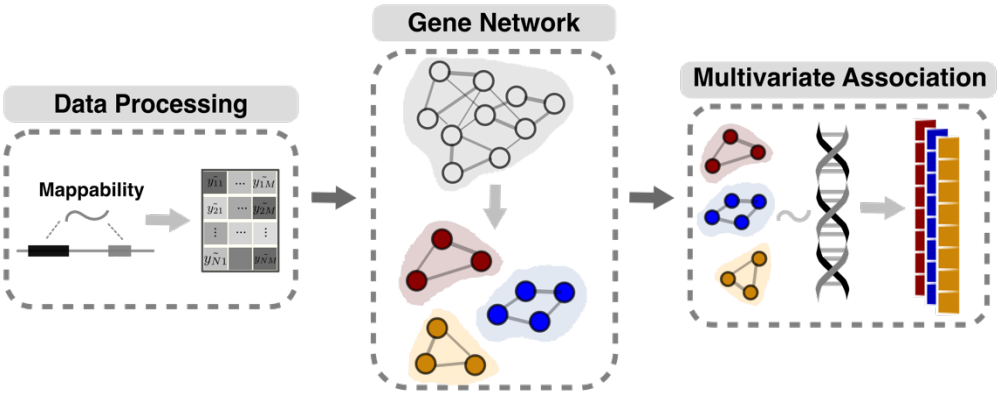

Figure 1. Three main steps in trans-PCO pipeline. The first step of trans-PCO pre-processes RNA-seq data to reduce false positive *trans*-eQTL associations due to read alignment errors. The second step involves grouping genes into gene sets, such as co-expression modules or biological pathways. The last step tests for *trans*-eQTLs of each gene set by a PC-based multivariate association test[17].

Trans-PCO outperforms existing methods in simulations

We performed simulations to evaluate the power of trans-PCO in detecting *trans*-eQTLs associated with multiple genes. We compared the power to (i) the standard univariate test ("MinP"), and (ii) the primary PC-based test ("PC1") (see Methods). We used a co-expression gene module consisting of 101 genes from the DGN dataset (module 29). In null simulations, we simulated the z-scores between a SNP and $K = 101$ genes in a gene module following the null distribution, $N_K(0, \Sigma_{K\times K})$, where $\Sigma_{K\times K}$is the residualized expression correlation matrix. In power simulations, we generated z-scores from the distribution $N_K(\sqrt{n}\beta, \Sigma_{K\times K})$, where n is sample size (N = 200,400,600, and 800) and β is a vector representing the true effect sizes of the SNP on K genes. Among K genes, a proportion γ of them are causal with non-zero effects. Therefore, we generated β_k from a point normal distribution, where $\beta_k \sim N(0, \sigma^2{}_b)$ for proportion γ ($\gamma = 1\%, 5\%, 10\%, 30\%\ and\ 50\%$), and $\beta_k = 0$, otherwise. The *trans*-genetic variance is $\sigma^2{}_b$ ($\sigma^2{}_b = 0.001$), which is a low and realistic per SNP heritability for *trans* effects. We simulated 10,000 SNPs and performed 1000 simulations.

To compare the power of different approaches to identify *trans*-eQTLs, we defined significant univariate tests for a gene module based on the minimum p-values across K genes. For the primary PC-based method, we used PC1 of the gene module co-expression matrix to represent the module and tested it for associations. In power simulations, we set significance levels at 10% FDR to be consistent with real data analysis. We computed power as the average proportion of significant tests out of 10,000 simulated SNPs across 1000 simulations.

Trans-PCO significantly outperformed the univariate test and the primary PC method across different sample sizes and proportions of causal genes **(Figure 2)**. Specifically, the power of trans-PCO increases rapidly with increasing sample sizes. At the sample size of 800, assuming 30% of

genes have causal effects in the gene module, the power of trans-PCO is 74%, compared to 15% for the univariate test and 0.0018% for the primary PC method **(Figure 2A)**.

We also compared the power of each method across various causal gene proportions using a fixed sample size (n=500). All three methods have little power in detecting *trans*-eQTLs when the proportion of causal genes is below 10%. However, above this threshold, the power of trans-PCO increases dramatically: 36% at 30% causal genes and 86% at 50% causal genes. In contrast, the univariate and the primary PC methods remain almost powerless for nearly all simulated scenarios (**Figure 2B**). We note that the primary PC method appears to be almost powerless across the scenarios, which agrees with the previous observation that the primary PC can be less powerful than higher-order PCs in GWAS[11]. Simulation results at various genetic variances can be found in the supplement (Figure S1). Null simulations demonstrated all three methods are well-controlled for false-positive inflations (Figure S2).

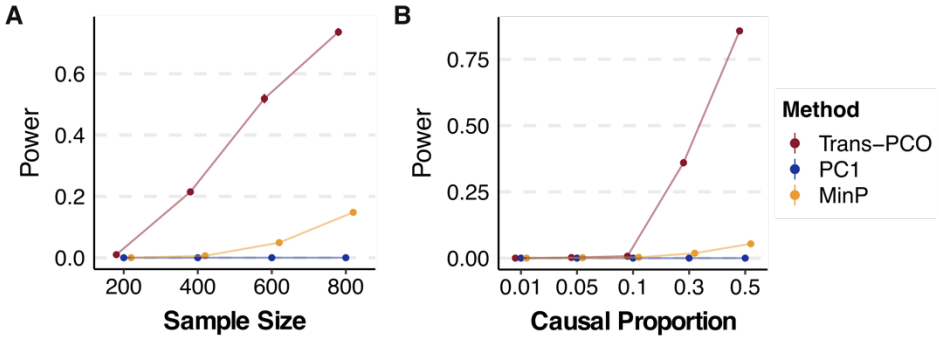

Figure 2. Power of trans-PCO across different sample sizes and causal gene proportions, in comparison to PC1 and univariate (MinP) methods. (A) Power comparison across various sample sizes (*N* = 200, 400, 600, and 800). *Trans*-genetic variance $\sigma^2{}_b$ was simulated to be 0.001 and the proportion of causal genes in the gene module was 30%. Error bars representing 95% confidence intervals are plotted, but many are too small to be visible. See numerical results in Table S2. **(B) Power comparison across different proportions of causal genes in the gene module.** The simulated sample size was 500. Points show average power across 1000 simulations. Error bars representing 95% confidence intervals are plotted, but many are too small to be visible. See numerical results in Table S2.

Trans-PCO identifies 3899 *trans*-eQTL SNP–module pairs associated with co-expression gene modules in the DGN dataset

We used trans-PCO to identify *trans*-eQTLs associated with co-expression gene modules in RNA-seq data from whole blood samples of the DGN cohort (*N* = 913)[24]. WGCNA[20] identified 166 co-expression gene modules, with the number of genes in each module ranging between 625 (module 1, M1) and 10 (module 166, M166) (Supplementary Table S1). We then performed genome-wide scans of *trans*-eQTLs for each gene module. At 10% FDR, trans-PCO identified significant *trans*-eQTLs for 102 out of 166 gene modules, corresponding to 3899 significant *trans*-

215 eQTL SNP–module pairs and 160 linkage-disequilibrium (LD)-independent loci (**Figure 3A**,
216 Supplementary Table S3).
217
218 We compared *trans*-eQTL signals detected in DGN by trans-PCO to signals identified by the
219 univariate method in Battle et al.[24]. Out of 12,132 genes analyzed by trans-PCO, the univariate
220 method detected 326 significant *trans*-eQTL SNP–gene pairs for 128 genes, which correspond to
221 62 LD-independent loci at 5% FDR[24]. At the same FDR level, trans-PCO identified 3031 significant
222 *trans*-eQTL SNP–gene module pairs, which correspond to 114 LD-independent loci. We
223 compared the magnitude of the significant *trans*-eQTL effects detected by trans-PCO and the
224 univariate method. More specifically, we compared the maximum univariate z-scores of SNPs
225 and each gene in significant *trans*-eQTL SNP–module pairs identified by trans-PCO to the z-
226 scores of significant *trans*-eQTL SNP–gene pairs by the univariate method. We found that the
227 maximum z-scores of trans-PCO signals are much smaller than z-scores of the univariate method
228 signals (**Figure 3B**), indicating that our multivariate approach can detect much smaller *trans*
229 effects than univariate methods. We also applied the primary PC method to DGN, and found
230 trans-PCO is much more powerful than the primary PC method (see Supplementary Note).

231 *Trans*-eQTLs are enriched in variants with *cis*-regulatory effects on 232 transcription factors

233 We found that only 31 *trans*-eQTL SNPs (1%) are in coding regions, suggesting that a very small
234 proportion of *trans*-eQTLs impact gene expression levels in *trans* by altering protein coding
235 sequences. Several studies have shown that *trans*-eQTL SNPs have *cis*-regulatory effects,
236 impacting the expression levels or splicing of nearby genes[10,24]; thus, we evaluated our identified
237 *trans*-eQTLs for concomitant *cis*-regulatory activity. We first overlapped *trans*-eQTL SNPs with
238 *cis*-eQTLs and *cis*-splicing QTLs (*cis*-sQTLs) in DGN[27]. Of the 2955 *trans*-eQTL SNPs, we found
239 that 71% are significant *cis*-eQTLs in DGN, and 46% are significant *cis*-sQTLs, together
240 accounting for 73% of all *trans*-eQTL SNPs. To further examine whether the *cis* and *trans* effects
241 are driven by the same variant, we performed colocalization analysis of *trans*-eQTL loci with *cis*-
242 eQTLs and *cis*-sQTLs using coloc[25] (see Methods). Specifically, we grouped *trans*-eQTL SNP-
243 gene module pairs into 179 *trans*-eQTL-gene module pairs (see Methods) and performed
244 colocalization analyses between *trans*-eQTLs and *cis*-eQTLs/*cis*-sQTLs. 51 out of 179 *trans*-
245 eQTLs colocalized with a *cis*-eQTL (PP4>0.75, **Figure 3C** and Figure S3). 41 *trans*-eQTLs
246 colocalized with a *cis*-sQTL. Overall, 60 *trans*-eQTLs share causal variants with at least one *cis*-
247 eQTL or *cis*-sQTL (**Figure 3C**, Table S4), confirming that *trans*-eQTL effects are generally
248 mediated through *cis*-gene regulation. Additionally, a large fraction of *trans*-eQTLs (66%) do not
249 colocalize with *cis*-eQTLs or *cis*-sQTLs. While power may have limited our ability to detect
250 colocalization of some *trans*-eQTLs and *cis*-eQTLs, there might also exist unknown *trans*-
251 regulatory mechanisms, independent of *cis* gene expression or splicing, which is subject to future
252 studies.
253
254 We also investigated the types and functions of genes that are likely to mediate *trans*-eQTL
255 effects. We found that the genes nearest *trans*-eQTL loci are highly enriched in "RNA polymerase
256 II transcription regulatory region sequence-specific DNA binding" (adjusted $P = 1.26 \times 10^{-3}$) and

"DNA-binding transcription factor activity" (adjusted $P = 1.39 \times 10^{-3}$, Table S5), suggesting that transcription factors are important mediators of *trans*-eQTL effects. Indeed, trans-PCO identified and replicated several well-known master *trans* regulators in blood, such as *IKZF1*[26,28,29], *NFKBIA*[26], *NFE2*[10,26,30], and *PLAGL1*[26,29] (**Figure 3A**). We also found colocalization of these *trans*-eQTLs with *cis*-eQTLs at the *NFKBIA*, *NFE2* and *PLAGL1* loci (**Figure 3D**, Figure S3), supporting the conclusion that these genes are likely the *cis*-mediating genes.

High quality map of *trans*-eQTL SNP to gene module associations improves functional interpretation

Most of the gene modules used in trans-PCO have functional annotations, which allows us to interpret the functional roles of the *trans*-eQTLs identified by the method. We first functionally annotated the 166 co-expression modules using g:Profiler[31], which performs functional enrichment analysis on gene sets using predefined gene ontology and pathway annotations. This allowed us to annotate 131 of the 166 modules with at least one significantly enriched gene ontology or pathway (Table S6).

These annotations helped us interpret the function of identified *trans* effects. For example, the *trans*-eQTL signal near *IKZF1* (on chromosome 7) is significantly associated with 27 gene modules. *IKZF1* encodes a transcription factor IKAROS that belongs to the family of zinc finger DNA binding proteins[32]. The *IKZF1* (IKAROS) *trans*-target gene module 159 (M159) is significantly enriched in the "positive regulation of transcription of Notch receptor target" (adjusted $P = 6.82 \times 10^{-3}$, **Figure 3E**). Reassuringly, it was previously found that IKAROS is a repressor of many Notch targets, and our *trans*-eQTL signal further supports the *trans* regulation of Notch signaling pathway by IKAROS[33]. *IKZF1 trans*-target module 3 (M3) is significantly enriched in the gene ontology term "defense response to virus" (Figure S4, adjusted $P = 8.7 \times 10^{-31}$) and M35 is significantly enriched in the innate immune system (adjusted $P = 4.09 \times 10^{-17}$). This data supports the conclusion that the *IKZF1* locus plays a *trans*-regulatory role in immune responses (**Figure 3E**). The *trans*-eQTL SNPs near *NFKBIA*, which encode NF-kappa-B inhibitor subunit A, are significantly associated with module 66 (M66) (adjusted $P < 1.8 \times 10^{-7}$, adjusted $P < 9.9 \times 10^{-2}$). Interestingly, we found that M66 is highly enriched in NF-kappa-B signaling pathway (adjusted $P = 8.35 \times 10^{-5}$, **Figure 3E**), which supports the *trans* regulation of the NF-kappa-B signaling pathway by *NFKBIA*. The complete list of *trans*-eQTLs signals and functional annotations of *trans*-target gene modules can be found in Supplementary Table S3 and Table S6.

Trans-PCO identifies 965 *trans*-eQTLs associated with well-defined biological processes

To further demonstrate the utility of trans-PCO, we applied trans-PCO to 50 MSigDB hallmark gene sets, which represent well-defined biological processes[21], including DNA repair, coagulation, heme metabolism, Notch signaling etc. (Table S14). Each gene set contains between 32 and 200 genes. In DGN, we identified 965 significant *trans*-eQTL SNP–module pairs, corresponding to 41 gene sets, at 10% FDR level (Figure S5, Table S15).

296
297 *Trans*-eQTLs associated with well-defined biological processes facilitate interpretation of the
298 *trans*-eQTL signals. For example, we identified several *trans*-eQTL signals at the *NLRC5* locus
299 (Table S15). The *trans* target gene set is the "interferon alpha response" gene set, suggesting
300 *trans* regulation from *NLRC5* to the interferon signaling pathway. Reassuringly, earlier studies
301 have confirmed that *NLRC5* is a master regulator for MHC class II genes and negatively regulates
302 the interferon signaling pathway[34,35]. The *trans*-eQTL signals also validated our previous
303 interpretations of *trans*-eQTLs associated with co-expression gene modules. For example, in
304 agreement with our analysis of co-expression modules, we found that the *IKZF1* locus is
305 significantly associated with several immune-related biological processes, such as interferon
306 gamma response (Table S15, **Figure 3E**).
307

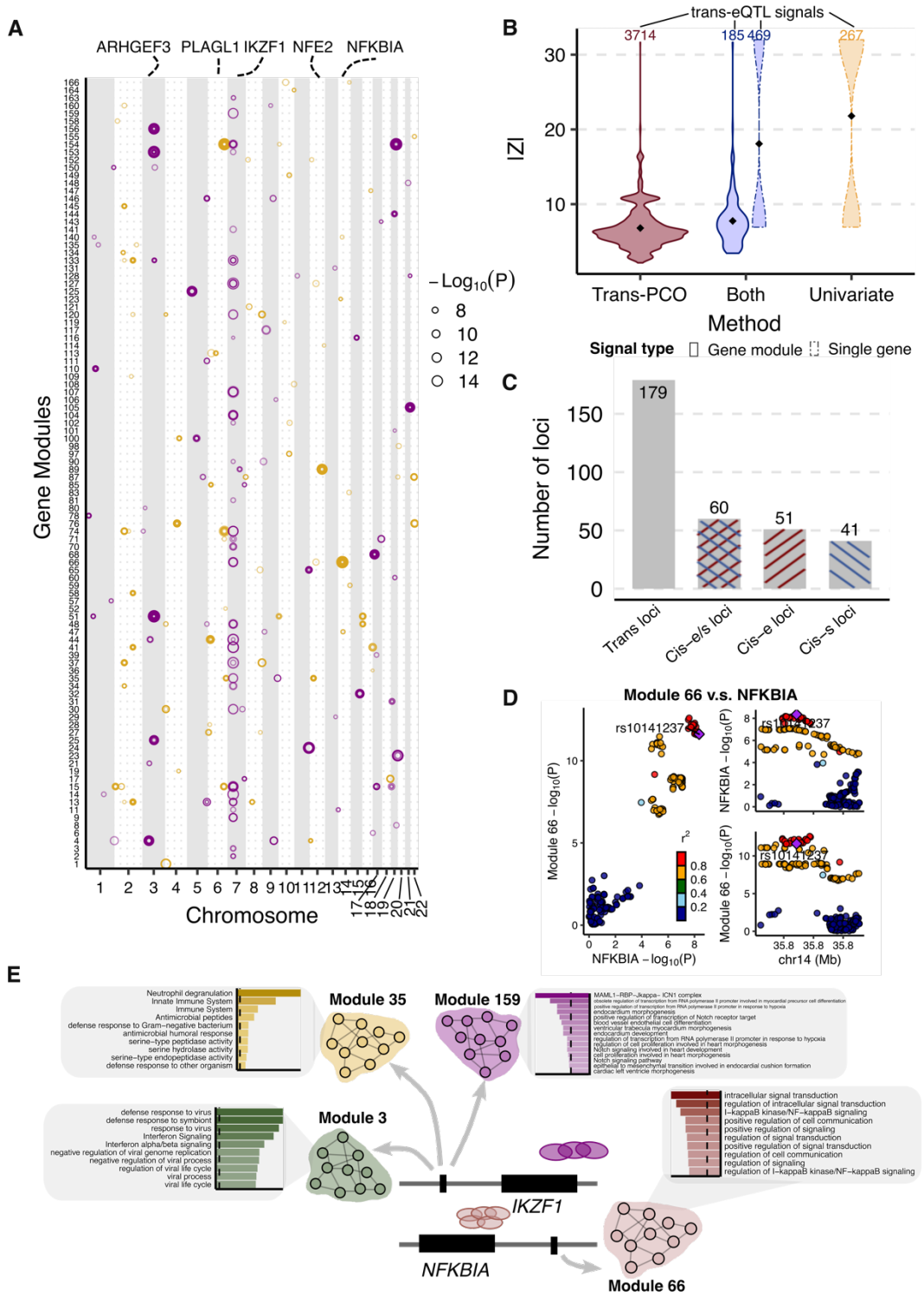

Figure 3. Trans-PCO identifies *trans*-eQTLs associated with co-expression gene modules in DGN. (A) Significant *trans*-eQTL signals associated with 166 co-expression modules in DGN. Chromosomal positions of *trans*-eQTL SNPs are on the x-axis, and gene modules are on the y-axis. Point sizes are -$\text{Log}_{10}(P)$ values of significant *trans*-eQTLs. Purple and orange represent odd and even chromosomes, respectively. (**B**) Comparison of the magnitude of

significant *trans*-eQTLs effects detected by trans-PCO and the univariate method. X-axis shows signal categories: trans-PCO specific signals (Trans-PCO), univariate test specific signals (Univariate), and signals identified by both methods (Both). The maximum z-scores of each SNP and each gene in a gene module is used to represent the SNP-module pair. The numbers on top are the number of signals in each category. Line type represents the target type of signals (gene module vs single gene). Y-axis is the absolute value of the z-scores of the signals. **(C) Colocalization of *trans*-eQTLs and *cis*-e/sQTLs.** The gray bar represents the *trans* loci used for colocalization analyses. The bar highlighted in blue represents *trans*-eQTL loci colocalized with *cis*-sQTLs, red for *cis*-eQTLs, and mixed color for either *cis*-eQTLs or *cis*-sQTLs. **(D) Colocalization of *trans*-eQTLs of Module 66 and and *cis*-eQTLs of *NFKBIA*. (E) Functional annotations of gene sets facilitate functional interpretation of *trans*-eQTL signals.** The *trans*-eQTLs near *NFKBIA* and *IKZF1* are associated with several gene modules. The bar plots show the functional enrichments in modules. The numerical values of enrichments are in Table S6.

Trans-PCO improves understanding of *trans* regulatory effects of disease-associated loci

To understand *trans* regulatory effects of genetic variants associated with complex traits, we performed colocalization analysis of *trans*-eQTL signals with GWAS loci of 46 complex traits and diseases, including 29 blood traits and 8 other common complex traits (such as height and BMI) from the UK Biobank[30,36] and 9 autoimmune diseases[27,37–43] (Methods and Table S7).

We grouped the *trans*-eQTL SNPs into 200kb regions (loci) for colocalization analyses (see Methods). The 3899 *trans*-eQTLs associated with co-expression gene modules were grouped into 179 *trans*-eQTL locus–module pairs. 42 out of 46 complex traits have at least one GWAS loci colocalized with the 179 *trans*-eQTL locus–module pairs. On average across all traits, 8.8% of *trans*-eQTL loci colocalize with GWAS loci (**Figure 4A**, Table S8). We observed a higher proportion of colocalization with blood traits (mean proportion 12.0%) than non-blood traits (mean proportion 1.5%). Although we expect some higher proportions of colocalization with blood traits to occur in a whole blood sample, our results may also indicate some residual effects due to cell composition--despite corrections for cell composition using both gene expression PCs and estimated cell-type proportions[24], such that some *trans*-eQTLs may regulate the abundance of cell proportions and therefore are associated genes that are specifically expressed in certain cell types. Our results are consistent with a recent study by the eQTLGen consortium, which has shown that *trans*-eQTLs in whole blood reflect a combination of cell-type composition and intracellular effects[10].

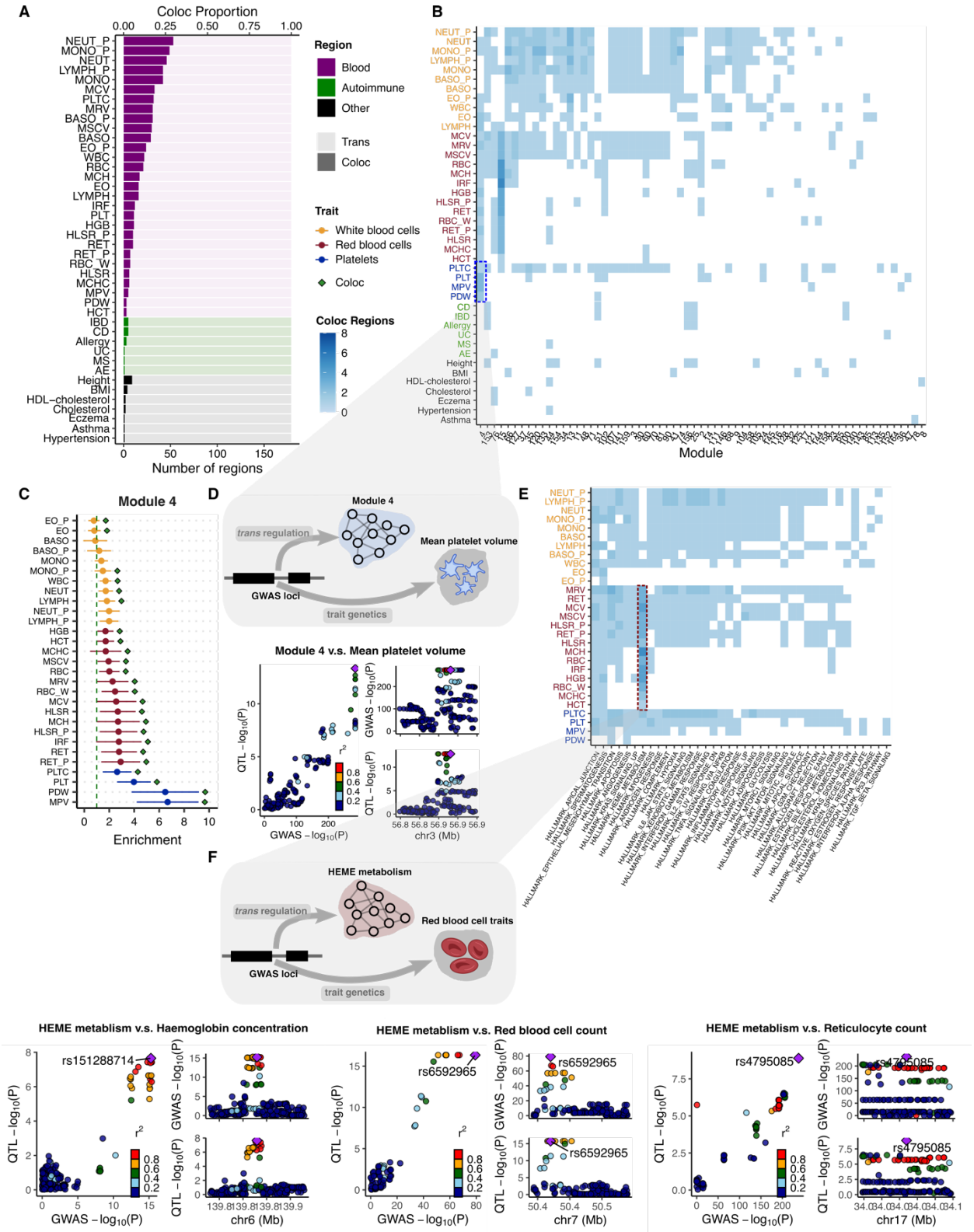

Figure 4. Colocalization of *trans*-eQTLs with GWAS loci of 42 complex traits with at least one colocalization region. (A) The number of colocalized *trans* loci associated with co-expression gene modules with GWAS loci. The traits are first ordered by broad categories:

blood traits, autoimmune diseases, and other traits in UKBB. The traits within each category are then ordered by the total number of colocalized regions. **(B) Heatmap of the number of colocalized *trans* loci associated with co-expression gene modules with GWAS loci between each module and trait.** The traits are first ordered by broad categories: white blood cells (Orange), red blood cells (Red), platelet cells (Blue), autoimmune diseases (Green) and other traits in UKBB (Black). The traits within each category are then ordered by the number of colocalized gene modules. The blue shades represent the number of colocalized regions. **(C) Heritability enrichment of Module 4 (M4) in blood traits.** Heritability enrichment was estimated by using S-LDSC. Error bars are 95% confidence intervals. **(D) Colocalization of mean platelet volume associated locus near *ARHGEF3* and *trans*-eQTL of M4. (E) Heatmap of the number of colocalized *trans* loci associated with MSigDB hallmark gene sets with GWAS loci across modules and blood traits.** The blue shades represent the number of colocalized regions. **(F) Colocalization of GWAS loci associated red blood cell traits and *trans*-eQTLs associated with heme metabolism.** Six loci associated with red blood cell traits are associated with heme metabolism in *trans*. Numerical results can be found in Table S16. Colocalization plots of the other loci are in Figure S6.

Nevertheless, we found several *trans*-eQTLs that colocalized with GWAS loci, which revealed specific interpretable pathways or functional gene sets (**Figure 4B**, Table S9). For example, *trans*-eQTLs associated with co-expression module 4 (M4) colocalized with 24 out of 29 blood traits (**Figure 4B**). M4 is highly enriched for genes involved in platelet activation (adjusted P = 1.12×10^{-12}, Figure S4, Table S6). One of the colocalized *trans*-eQTL SNPs associated with M4 is in the introns of the *ARHGEF3* gene (**Figure 4D**), which has been shown to play a significant role in platelet size in mice[44]. To further support the interpretation of colocalized signals, we estimated heritability enrichment of M4 in blood traits using stratified LDscore regression[45] (S-LDSC, **Figure 4C**). We reasoned that an enrichment of trait heritability near genes in a module would strongly support the involvement of a module in the genetic etiology of a trait. Strikingly, we found that M4 is significantly enriched in the heritability of multiple blood traits, and the enrichment was especially strong for platelet traits such as platelet distribution width (enrichment = $6.5 \times$, P = 7.0×10^{-5}) and mean platelet volume (enrichment = $6.7 \times$, P = 1.2×10^{-5}, **Figure 4C**, Table S10). Additionally, we identified that the *ARHGEF3* locus is significantly associated with the MSigDB coagulation hallmark gene set (Table S15). These findings strengthen the model where genetic variation near *ARHGEF3* impacts the expression levels of multiple genes that are involved in platelet biology and that also harbor nearby genetic variation associated with platelet traits.

We also performed colocalization analysis of *trans*-eQTLs associated MSigDB hall mark gene sets **(Figure 4E**, Table S16**)**. One of the gene sets represents heme metabolism, which is an essential process underlying erythroblast differentiation and red blood cell counts. We found that six *trans*-eQTL loci of heme metabolism significantly colocalized with GWAS loci associated with red blood cell traits, such as hemoglobin concentration, red blood cell count, and reticulocyte count (PP4=0.76-1.00, **Figure 4F**, Figure S6, Table S16). Our results provide evidence that these six loci regulate heme metabolism in *trans,* which is an essential process underlying erythroblast differentiation and red blood cell counts.

In another example, we found a *trans*-eQTL near *IKZF1* for M3 that colocalizes with 11 blood traits, seven of which are related to white blood cells (Table S9). As mentioned previously, M3 is significantly enriched for gene ontology terms including "defense response to virus" (adjusted $P = 8.7 \times 10^{-31}$) and "negative regulation of viral processes" (adjusted $P = 1.07 \times 10^{-17}$, Table S6). The enrichments are driven by many genes related to interferon (e.g., *IFI6*, *IFI16*, *IRF7*), which are proteins released by host cells in response to the presence of viruses and indicate immune related functions (Table S1, Table S6). Additionally, our heritability analysis of genes in M3 identified enrichments for multiple traits associated with blood cell-type count including neutrophil count (enrichment = 2.3×, $P = 1.7 \times 10^{-4}$) and white blood cell count (enrichment = 2.1×, $P = 1.3 \times 10^{-4}$, Figure S7). Our analyses support that the while blood cell associated locus *IKZF1* regulates immune response pathways in *trans*.

Taken together, our functional map of *trans*-eQTLs revealed concrete examples where genetic variants associated with complex traits also influence a biological pathway or a coherent set of genes with similar functions. Thus, *trans*-eQTL of gene sets have the potential to reveal *trans*-regulatory mechanisms underlying complex traits and diseases. The complete list of colocalization signals for each trait can be found in Supplementary Table S9.

Summary-statistics–based trans-PCO identified 10,167 *trans*-eQTL SNP–module pairs in eQTLGen

We developed summary-statistics–based trans-PCO to increase its applicability to gene expression datasets of large sample sizes, such as eQTLGen (N=31,684, whole blood). To ensure that summary-statistics–based trans-PCO signals are well-controlled for test statistics inflation and false positives, we added two steps to the original pipeline. First, we carefully select gene sets to minimize the noise when approximating the gene correlation matrices. When only summary statistics are available, the correlation matrix of each gene set is approximated with the correlations of z-scores of the insignificantly associated SNPs of each gene. A low ratio of SNPs to genes (<50) results in a noisy approximation of correlation matrices and test statistics inflation (Figure 6A, Methods, and Supplementary Note). Therefore, we only use gene modules with ratios greater than 50 to test for *trans*-eQTLs, which we show are well-controlled for inflation (**Figure 5A**, Figure S8). Second, we remove genes that are cross-mappable to the test SNP loci (see details in Methods) in the association test to reduce false positives caused by multi-mapping reads.

The eQTLGen study performed the standard univariate *trans*-eQTL mapping on 10,317 GWAS SNPs and the summary statistics of these *trans*-eQTLs are available. We applied the summary-statistics–based trans-PCO to these summary statistics to identify *trans*-eQTLs associated co-expression gene modules and MSigDB hallmark gene sets.

Of the 166 co-expression gene modules identified in DGN, we used 129 modules with reliable correlation matrix approximations to ensure the *trans*-eQTL signals are well-controlled for inflation (**Figure 5A**, Figure S8, Methods, and Supplementary Note). Similarly, Of the 50 MSigDB hallmark gene sets, we only used 11 gene sets with accurate correlation matrix approximations (Figure

S17). In total, there were 4533 genes in the tested co-expression gene modules and hallmark gene sets. We identified 8116 *trans*-eQTL SNP-gene coexpression module pairs, corresponding to 2161 eQTLGen test SNPs and 122 gene modules (**Figure 5B**, Table S11); and 2051 significant *trans*-eQTLs associated corresponding to 1018 SNPs and all 11 hallmark gene sets, using Bonferroni correction (Table S17). Notably, the univariate method used in eQTLGen[12] identified 1050 hub SNPs targeting more than 10 genes at 5% FDR, 89% of which could also be identified by trans-PCO (**Figure 5C**).

The large sample size in eQTLGen improves the power of *trans*-eQTL detection. Of the 3899 significant *trans*-eQTL SNP–co-expression module pairs in DGN, 38 pairs were also tested in eQTLGen. Reassuringly, we found that all 38 *trans*-eQTL signals are replicated in eQTLGen (under a replication *P* value cutoff of 0.1/38, Table S12) and all association p-values are highly significant ($P < 10^{-12}$, Figure S9). In contrast, most of the *trans*-eQTL signals in eQTLGen were not found in DGN. For example, of the 7577 SNP-module pairs analyzed in both datasets, there were 7291 pairs (96%) that were uniquely identified in eQTLGen (which is defined as at least 1 MB away from *trans*-eQTL SNPs in DGN). This is not surprising, because the association p-values are much smaller in the eQTLGen dataset due to the larger sample size (Figure S10).

The nearest genes of eQTLGen *trans*-eQTL loci are significantly enriched in DNA binding activity (adjusted $P = 3.73 \times 10^{-4}$) and transcription factor binding (adjusted $P = 1.74 \times 10^{-7}$) , as well as immune responses, such as cytokine receptor activity (adjusted $P = 7.27 \times 10^{-7}$) or MHC class II receptor activity (adjusted $P = 9.93 \times 10^{-5}$, **Figure 5B**, Table S13). We found that the enrichment of immune responses was driven by *trans*-eQTL loci in the HLA region on chromosome 6 (such as *HLA-DRA*, *HLA-DRB1 etc*, Table S11) or near cytokine receptor genes (such as *IL23R, IL1R1, CXCR4 and* genes on the chemokine receptor gene cluster region*: CCR2, CCR3, CCR5 etc*). These *trans*-eQTL SNPs are associated with several autoimmune diseases, such as type 1 diabetes, autoimmune thyroid diseases, cutaneous lupus erythematosus and inflammatory bowel disease (Table S11). The trans-PCO signals help us understand the *trans* regulatory mechanism of these loci. For example, we found that the *trans* target gene modules of the HLA loci are enriched in immune related functions, such as cytokine production (M44), B cell differentiation (M54), IgE binding (M60), TNF signaling pathway (M62), T cell activation (M63 and M87), and cytokine signaling pathway (M62 and M76, **Figure 5D**). The *IL23R* locus is associated with cytokine signaling pathway (M76) in *trans*. The chemokine receptor genes were associated with several gene modules including cytokine production (M44), IgE binding (M60) and T cell activation (M87). These *trans*-eQTL signals support the conclusion that genetic loci associated with autoimmune disease regulate immune related pathways in *trans*.

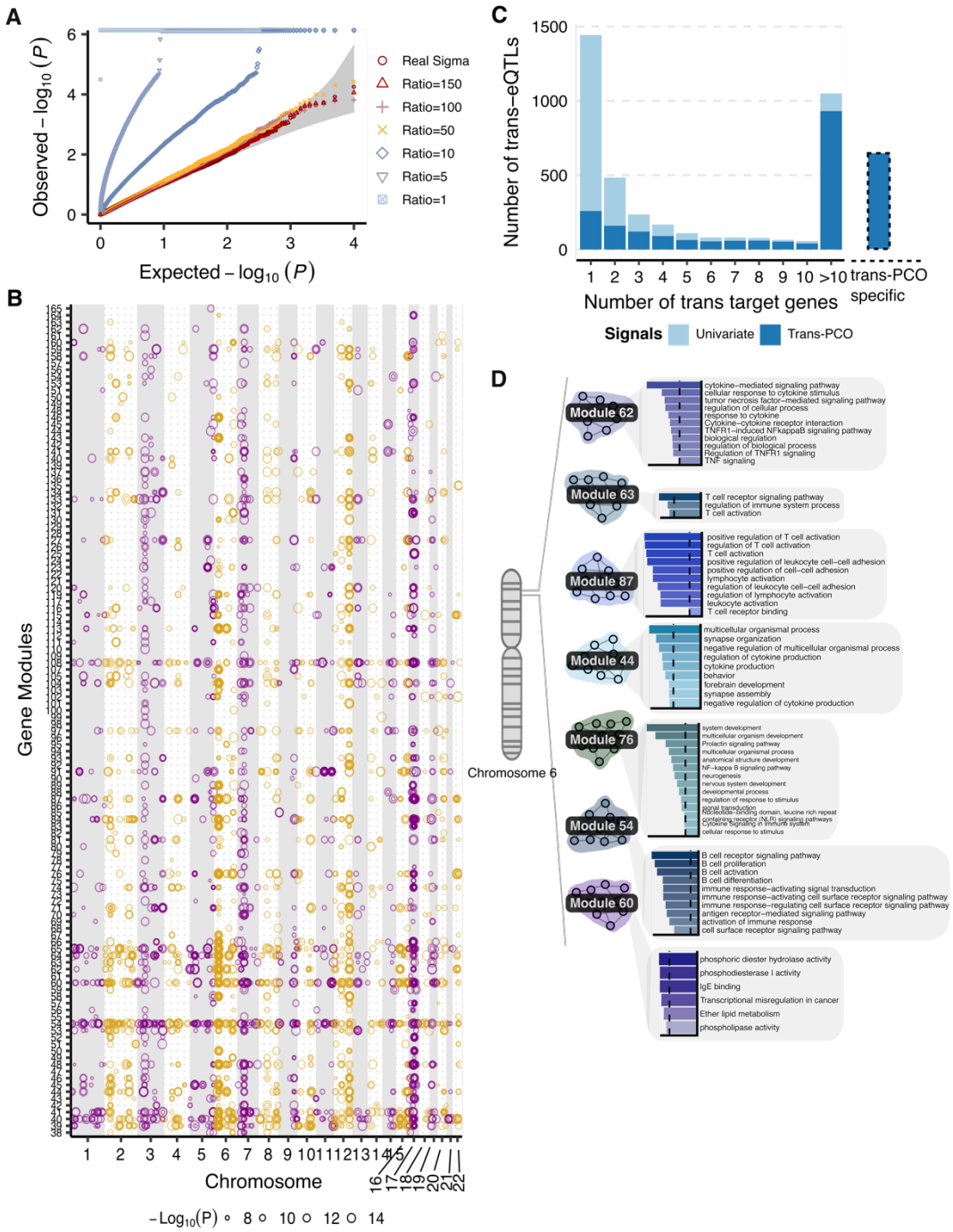

Figure 5. Trans-PCO identifies *trans*-eQTLs associated with co-expression gene modules and MSigDB hallmark gene sets in eQTLGen. (A) Summary-statistics–based trans-PCO is well controlled for test statistics inflations. We show gene module 1 (size 625) as an example. SNP to gene ratios used for correlation matrix estimation are in different shapes and colors. Red-yellow shades represent higher ratios (>=50) and blue shades represent lower ratios. Gray area shows 95% CI. Trans-PCO used a minimum ratio of 50. **(B) 8199 significant *trans*-eQTL signals associated with co-expression modules in eQTLGen.** Chromosomal positions of *trans*-eQTL SNPs are on the x-axis and gene modules are on the y-axis. Point sizes are -Log$_{10}$(P) values of

significant *trans*-eQTLs. **(C) The majority of hub SNPs targeting more than 10 genes in the original eQTLGen study are identified by trans-PCO.** Light blue bar represents the total number of *trans*-eQTLs in the original eQTLGen study at 5% FDR level. Dark blue bar represents the *trans*-eQTLs also detected by trans-PCO under Bonferroni correction that are associated with co-expression modules or MSigDB gene sets. The bar on the right shows the *trans*-eQTLs detected only by trans-PCO. **(D) The HLA locus is associated with several immune related gene modules in *trans*.** The bar plots show the functional enrichment of co-expression gene modules.

Discussion

In summary, we developed a powerful method, trans-PCO, to detect *trans*-eQTLs associated with expression levels of co-expressed or co-regulated genes. The multivariate approach of trans-PCO can detect much smaller *trans* effects (**Figure 3B,** Figure S13) and is substantially more powerful than existing methods. (**Figure 2**). Trans-PCO is also flexible. It can be applied to both RNA-seq data with genotypes or summary statistics, and the user can employ various definitions of gene sets. Applying trans-PCO to both the DGN and the eQTLGen datasets, we identified nearly 15,000 *trans*-eQTLs associated with co-expression modules and well-defined biological processes. *Trans*-eQTLs with annotated gene modules facilitate our understanding of the *trans*-eQTL signals. These *trans*-eQTLs also improve our understanding of the *trans* regulatory effects of disease associated loci. We highlight multiple examples where our map of *trans* effects helps us identify how trait-associated variants impact gene regulatory networks and pathways. For example, we found six genetic loci associated with red blood cell traits to have significant *trans*-associations with the heme metabolism gene set. It is possible that these genetic loci are "peripheral master regulators" that regulate core processes of red blood cell production.

A limitation of multivariate association tests, including trans-PCO, is that they do not explicitly identify which genes in the gene sets are significantly associated with the test SNP. Furthermore, while functional annotations of gene sets facilitate our understanding of the *trans*-eQTL signals, it is possible that the genes driving *trans*-associations are different from the genes driving functional enrichment of the gene sets. Therefore, the biological interpretation of *trans*-eQTL signals should be supported with other evidence before it is considered definitive. While the exact molecular mechanism requires further validation, the large number of *trans*-eQTLs identified by trans-PCO in our study opens up new opportunities to understand complex traits-associated loci and underlying mechanisms.

Trans-eQTLs identified in bulk tissues can be a combination of cell composition *trans*-eQTLs, which are driven by cell type proportions, and intracellular *trans*-eQTLs, which capture *trans* regulatory effects in a single cell type. To get higher proportions of intracellular eQTLs, the common approach is to correct for cell type proportions in association tests. For example, the eQTLGen study[12] corrected for cell proportion effect by using gene expression PCs. They validated some trans-eQTLs using single-cell RNA sequencing data, indicating that these *trans*-eQTLs were intracellular *trans*-eQTLs. In our analysis of DGN dataset, we included the estimated cell proportions as covariates, in addition to gene expression PCs. This strategy might have given

528 rise to higher proportions of intracellular *trans*-eQTLs. Co-expression gene modules could also
529 capture cell proportion effects. In our study, we removed cell proportions from gene expression
530 levels before clustering genes into co-expression modules. While this can correct for cell
531 proportion effects in the co-expression modules to some extent, we note that it does not guarantee
532 their complete removal.
533
534 While many studies consider cell composition QTLs to be artifacts and seek to avoid them, we
535 think cell composition *trans*-eQTLs can be biologically interesting too. For example, the *IKZF1*
536 locus is significantly associated with several gene modules enriched with viral defense and other
537 immune related functions in *trans*. The locus is also significantly associated with white blood cell
538 proportions. Given the general function of white blood cells in fighting infections, these
539 observations raise the possibility that the *trans*-eQTLs near *IKZF1* regulate antiviral activity by
540 affecting white blood cell-type proportion. Supporting this hypothesis, we found earlier that genetic
541 variants near *IKZF1* are also associated with expression levels of genes in M159, which are
542 enriched in genes involved in the Notch signaling pathway. The Notch signaling pathway plays a
543 central role in cell proliferation, cell fate, and cell differentiation[46]; thus, our analyses reveal a
544 plausible mode of action whereby genetic variants near *IKZF1* impact multiple immune-related
545 functions by influencing white blood cell-type proportions.
546
547 All the *trans*-eQTL signals, with functional annotation of the gene sets, can be downloaded and
548 browsed in www.networks-liulab.org/transPCO, which will facilitate interpretation of *trans*
549 regulatory effects of disease associated loci.
550
551

Methods

Trans-PCO pipeline

We test if a genetic variant is associated with genes in a module through *trans* regulations using the multivariate model as follows,

$$[y_1 \cdots y_K] = G[\beta_1 \cdots \beta_K] + covariates + e,$$

where G is the dosage of a reference allele representing the genotype of a SNP, β_k is the effect of the SNP on k-th gene in the module with K genes, and y_k is the expression level of the k-th gene. To test if a SNP of interest is significantly associated with the module, we test the null hypothesis

$$H_0: \beta_1 = \cdots = \beta_K = 0.$$

We use a PC-based omnibus test (PCO)[19], which is a powerful and robust PC-based approach aiming at testing genetic association with multiple genes with no prior knowledge of the true effects.

Specifically, PCO combines multiple single PC-based tests in linear and non-linear ways, corresponding to a range of causal relationships between the genetic variant and genes, to achieve higher power and better robustness. A single PC-based test (most commonly the first primary PC_1) is,

$$T_{PC_k} = {\mu_k}^T Z \sim N({\mu_k}^T \beta, \lambda_k), 1 \leq k \leq K,$$

where Z is a $K \times 1$ vector of univariate summary statistic z-scores of the SNP for K genes in a module, μ_k is the k-th eigenvector of the covariance matrix $\Sigma_{K \times K}$ of Z, λ_k is the corresponding eigenvalue, and β represents the true causal effect. PCO combines six PC-based tests, including,

$$PCMinP = min_{1 \leq k \leq K} p_k, \text{ and } PCFisher = -2 \sum_{k=1}^{K} log(p_k),$$

where p_k is the p-value of T_{PC_k}. These two tests take the best p-value of single PC-based tests and combine multiple PC p-values as the test statistic. Other tests include,

$$PCLC = \sum_{k=1}^{K} \frac{T_{PC_k}}{\lambda_k}, WI = \sum_{k=1}^{K} {T_{PC_k}}^2, Wald = \sum_{k=1}^{K} \frac{{T_{PC_k}}^2}{\lambda_k}, VC = \sum_{k=1}^{K} \frac{{T_{PC_k}}^2}{\lambda_k^2},$$

which are linear and quadratic combinations of each single PC-based test weighted by eigenvalues. The six tests achieve best power in specific genetic settings with different true causal effects[19]. PCO takes the best p-value of the PC-based tests as the final test statistic,

$$T_{PCO} = min\, p_{\{PCMinP, PCFisher, PCLC, WI, Wald, VC\}},$$

to achieve robustness under unknown genetic architectures while maintaining a high power.

To prevent cis-regulatory effects from driving the identified *trans* associations between a SNP and module, we removed genes in the module that are on the same chromosome as the tested variant. In addition, to avoid false positive signals in *trans* associations due to alignment errors, we discarded RNA-seq reads that are mapped to multiple locations or poorly mapped genomic regions (mappability score <1)[24,26] before quantifying gene expression levels.

589 # Simulation

590 To evaluate the power of trans-PCO, we performed a series of simulations with various parameter
591 settings corresponding to different genetic architectures. We applied trans-PCO to the simulated
592 datasets and assessed the false positive rate and statistical power. We also ran two additional
593 statistical tests, a univariate test ("MinP") and a primary PC-based test ("PC1"), to compare their
594 performances with trans-PCO and to obtain more in-depth insights on trans-PCO,

595 $$T_{PC_1} = {\mu_1}^T Z,\ T_{MinP} = min\{p_1^g, \cdots, p_K^g\}.$$

596 The PC1 based test takes only the first PC as the proxy of a gene module and uses it as the
597 response variable to test for genetic variants with significant associations. We also compared
598 trans-PCO with a non-PC based statistical test, MinP, which takes the minimum p-value across
599 genes in the module as the test statistics.
600
601 To implement trans-PCO, PC1, and MinP tests, there are two main pieces of information that are
602 required as input, i.e. correlation matrix of the gene module and summary statistics (z-scores) of
603 SNPs with genes in the module. We used a gene module from the RNA-seq data (see Genotype
604 and RNA-seq QC) consisting of 101 genes ($K = 101$) and the corresponding correlation matrix to
605 make the settings more realistic. We sampled z-scores of 10^7 SNPs from the null distribution,

606 $$Z_{NULL} \sim N(0, \Sigma_{101}).$$

607 We tested the associations between each SNP and the gene module using the above three tests
608 and evaluated the p-values against the uniform distribution to validate if the statistical tests are
609 well calibrated.
610
611 We also simulated 10k z-scores of SNPs from the alternative distribution,

612 $$Z_{Alt} \sim N(\sqrt{n}[\beta_{101\gamma}, 0]^T, \Sigma_{101}),$$

613 where n is the sample size, β is a 101γ-long vector representing the causal effect of a SNP on
614 101 genes, and γ is the proportion of true target genes in the module. Each component of β
615 follows $N(0, \sigma_b^2)$, where σ_b^2 is the genetic variance. By default, we set the sample size n to be 500,
616 30% genes (30) in the module are true *trans* target genes, and σ_b^2 to be 0.001. To evaluate how
617 three tests perform across different genetic architectures, we simulated multiple scenarios across
618 varying sample sizes, target gene proportions, and genetic variances. Specifically, we looked at
619 the cases where sample size is 200, 400, 600, and 800, causal genes proportion is 1%, 5%, 10%,
620 30%, and 50%, and genetic variance is 0.002, 0.003, 0.004, 0.005, and 0.006. We simulated 10k
621 SNPs and performed 1000 simulations. To control the false discovery rate, we corrected the p-
622 values for multiple testing based on the simulated empirical null distribution of p-values, to keep
623 it consistent with the method used in the RNA-seq dataset (see Supplementary Note). An
624 association is significant if its adjusted p-value is lower than 0.1. The power is calculated as the
625 proportion of SNPs that were identified to be significant among 10k SNPs.

626 # Genotype and RNA-seq QC

627 We analyzed an RNA-seq dataset from whole blood[24]. We performed a series of QC on
628 individuals, genotypes, RNA-seq reads, and genes before quantifying gene expression profiles.
629 Specifically, we referred to the procedures in Liu et al.[26]. For individual-level QC, we removed

630 related individuals from 922 samples with RNA-seq reads available and kept 913 individuals in
631 total for further analysis. For genotype-level QC, we used SNPs with genotyping rate > 99%, minor
632 allele frequency > 5%, and Hardy-Weinberg equilibrium < 10^{-6}.
633
634 RNA-seq reads can be falsely aligned to genomic regions with high sequence similarity. The
635 misalignment onto these regions can lead to false positive signals in *trans*-eQTL analysis and
636 spurious correlations in gene co-expression networks[8]. To help address this problem, we
637 performed RNA-seq read-level QC to remove the reads with alignment issues. To be more
638 specific, we filtered out the reads that were mapped to multiple genomic regions and reads with >2
639 mismatches. We also removed the reads aligned to regions with low mappability.
640
641 On the gene level, we quantified the gene expression levels as Transcript Per Million (TPM). We
642 first normalized the expression levels across samples to the normal distribution by quantile
643 normalization, and then we normalized the expression levels across genes. We also filtered out
644 genes that are not protein-coding, lincRNA genes, or genes on sex chromosomes. As a result,
645 there are 12,132 genes left for follow up analysis. Finally, to control for potential confounding
646 factors and capture the co-expressed gene modules only driven by genetic effects, we regressed
647 out covariates from the expression profiles. We used biological and technical covariates, including
648 genotype PCs, expression PCs, and blood cell type proportions[24,26].

649 Identification of the gene co-expression network

650 We are interested in jointly testing co-regulated genes in a multivariate association test. To this
651 end, we first used WGCNA[20] to construct a gene co-expression network, where genes are
652 connected through correlations among their residualized expression levels. WGCNA uses
653 hierarchical clustering to cut the network into separate gene modules with highly correlated
654 expression levels. We used the default parameter settings, except that we specified the minimum
655 module size parameter ('minModuleSize') to 10 to obtain small gene modules.

656 Colocalization of *trans*-eQTLs and GWAS loci

657 To explore the role of *trans*-eQTLs in understanding complex traits and diseases, we performed
658 colocalization between *trans*-eQTLs of a gene module and GWAS loci of 46 complex traits and
659 diseases. Specifically, we used GWAS summary statistics of 29 blood-related traits[30] and 8 other
660 traits from UK Biobank, provided by Neale Lab (http://www.nealelab.is/uk-biobank/), and 9
661 autoimmune diseases collected in Mu et al.[27] (Table S7).
662
663 To define a region to perform colocalization, we first selected the *trans*-eQTL SNP with the most
664 significant p-value and expanded a 200 kb flanking genomic region around the lead SNP as a
665 region to perform colocalization analysis. We stopped searching for lead SNPs when all *trans*-
666 eQTLs were included in one region. This resulted in 255 *trans* regions. We ran colocalization
667 analysis between each *trans* region and GWAS loci of 46 complex traits using the R package
668 coloc[25], assuming there is at most one causal variant for each trait. We used the default priors
669 and 0.75 as the PP4 cutoff for significant colocalizations. We then merged regions into a single

region if their lead SNPs are within 200kb base pairs, generating 179 merged regions in total. We defined a merged region as being colocalized with a trait if any of its sub-regions has significant colocalization with the trait. We visualized the colocalized regions using LocusCompareR[47].

Colocalization of *trans*-eQTLs and *cis*-e/sQTLs

We performed colocalization analysis between *trans*-eQTLs and *cis*-eQTLs (*cis*-sQTLs) of genes near the *trans*-eQTLs. We used the same 179 *trans* loci defined in the colocalization analysis of GWAS loci. For a *trans* loci, we searched for the genes within 500 kb around the lead *trans*-eQTLs of the loci, and used these genes to perform colocalization. We used summary statistics of *cis*-eQTLs and *cis*-sQTLs in the DGN dataset from Mu et al.[27]. We ran coloc[25] with default priors and 0.75 as PP4 cutoff.

Trait heritability enrichment in gene modules

To investigate whether a gene module is enriched for trait heritability, we applied stratified LD score regression[45] (S-LDSC) to 166 co-expression gene modules and 46 complex traits and diseases. Specifically, for each module we defined the annotation set as the SNPs within genomic regions of genes in the module and also a 500 base-pair window around the genes. We also included 97 annotations from the baseline model. Partitioned heritability enrichment was calculated as the proportion of trait heritability contributed by SNPs in the module annotation over the proportion of SNPs in that annotation.

Summary-statistics–based trans-PCO applied to eQTLGen

The eQTLGen Consortium[10] has conducted the largest *cis*- and *trans*-eQTLs association analyses in blood to date. Specifically, 31,684 samples were tested for over 11 million SNPs across 37 cohorts. The summary statistics of *trans*-eQTLs are available for 10,317 trait-associated SNPs on 19,942 genes.

We applied our pipeline trans-PCO to eQTLGen summary statistics, using the same 166 co-expression gene modules defined in DGN dataset. We searched for *trans*-eQTLs among 10,317 SNPs.

The eQTLGen summary statistics are marginal z-scores meta-weighted across multiple cohorts. Most z-scores are from studies where the RNA-seq reads with mappability issues were not filtered out before quantifying gene expression profiles. Therefore, directly applying trans-PCO to the summary statistics could lead to false positive signals. To avoid that, we removed the genes in a module that are cross-mappable with any *cis*-genes (genes within 100 kb) of the test SNP. We further removed genes on the same chromosome as the test SNP to prevent the detected *trans*-effects from being dominated by *cis* regulations.

The gene expression profiles are not available in eQTLGen. Therefore, to estimate the gene correlation Σ of a module, we searched among eQTLGen SNPs for SNPs insignificantly

associated with the module (null SNPs) (see Supplementary Note for details). We observed that there are less null SNPs that can be found for large modules. And simulations show that the low ratio of the number of null SNPs used for Σ estimation to the module size leads to false positive signals (Supplementary Note). Therefore, we removed 37 gene modules with ratios lower than 50. Finally, we performed trans-PCO on the remaining 129 gene modules.

Data availability

All *trans*-eQTL signals, with functional annotation of the gene sets, can be browsed and downloaded at www.networks-liulab.org/transPCO. The genotype and gene expression data of DGN were downloaded by application through the NIMH Center for Collaborative Genomic Studies on Mental Disorders, under the "Depression Genes and Networks study (D. Levinson, PI)". The eQTLGen summary statistics are publicly available at https://www.eqtlgen.org/. The MSigDB hallmark gene sets are publicly available at http://www.gsea-msigdb.org/gsea/msigdb/human/genesets.jsp?collection=H. GWAS summary statistics of traits in the UK Biobank are available at Neale Lab (http://www.nealelab.is/uk-biobank/).

Code availability

The trans-PCO pipeline and code to reproduce analyses is available at https://github.com/liliw-w/Trans. This work also uses TensorQTL (https://github.com/broadinstitute/tensorqtl) to perform QTL mapping between genotypes and single genes, R package coloc (https://github.com/chr1swallace/coloc) to perform colocalization analyses between trans-eQTLs and cis-eQTLs, cis-sQTLs and GWAS loci, and S-LDSC software (https://github.com/bulik/ldsc) to estimate trait heritability enrichment in gene modules.

References

1. Maurano, M. T. *et al.* Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science* **337**, 1190–1195 (2012).
2. Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832–838 (2010).
3. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci.* **106**, 9362–9367 (2009).
4. Watanabe, K. *et al.* A global overview of pleiotropy and genetic architecture in complex traits. *Nat. Genet.* **51**, 1339–1348 (2019).
5. Nicolae, D. L. *et al.* Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.* **6**, e1000888 (2010).
6. Yao, D. W., O'Connor, L. J., Price, A. L. & Gusev, A. Quantifying genetic effects on disease mediated by assayed gene expression levels. *Nat. Genet.* **52**, 626–633 (2020).
7. Liu, X., Li, Y. I. & Pritchard, J. K. Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. *Cell* **177**, 1022-1034.e6 (2019).
8. Saha, A. & Battle, A. False positives in trans-eQTL and co-expression analyses arising from RNA-sequencing alignment errors. *F1000Research* **7**, 1860 (2019).
9. Albert, F. W., Bloom, J. S., Siegel, J., Day, L. & Kruglyak, L. Genetics of trans-regulatory variation in gene expression. *eLife* **7**, e35471 (2018).
10. 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).
11. Kim, S. & Xing, E. P. Statistical Estimation of Correlated Genome Associations to a Quantitative Trait Network. *PLOS Genet.* **5**, e1000587 (2009).
12. Ferreira, M. A. R. & Purcell, S. M. A multivariate test of association. *Bioinformatics* **25**, 132–

758 133 (2009).
759 13. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From
760 Polygenic to Omnigenic. *Cell* **169**, 1177–1186 (2017).
761 14. Small, K. S. *et al.* Regulatory variants at KLF14 influence type 2 diabetes risk via a female-
762 specific effect on adipocyte size and body composition. *Nat. Genet.* **50**, 572–580 (2018).
763 15. Smemo, S. *et al.* Obesity-associated variants within FTO form long-range functional
764 connections with IRX3. *Nature* **507**, 371–375 (2014).
765 16. Vogelstein, B., Lane, D. & Levine, A. J. Surfing the p53 network. *Nature* **408**, 307–310
766 (2000).
767 17. Kolberg, L., Kerimov, N., Peterson, H. & Alasoo, K. Co-expression analysis reveals
768 interpretable gene modules controlled by trans-acting genetic variants. *eLife* **9**, e58705
769 (2020).
770 18. Aschard, H. *et al.* Maximizing the Power of Principal-Component Analysis of Correlated
771 Phenotypes in Genome-wide Association Studies. *Am. J. Hum. Genet.* **94**, 662–676 (2014).
772 19. Liu, Z. & Lin, X. A Geometric Perspective on the Power of Principal Component Association
773 Tests in Multiple Phenotype Studies. *J. Am. Stat. Assoc.* **114**, 975–990 (2019).
774 20. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network
775 analysis. *BMC Bioinformatics* **9**, 559 (2008).
776 21. Liberzon, A. *et al.* The Molecular Signatures Database Hallmark Gene Set Collection. *Cell*
777 *Syst.* **1**, 417–425 (2015).
778 22. Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M. & Tanabe, M. KEGG:
779 integrating viruses and cellular organisms. *Nucleic Acids Res.* **49**, D545–D551 (2021).
780 23. Szklarczyk, D. *et al.* The STRING database in 2021: customizable protein–protein networks,
781 and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.*
782 **49**, D605–D612 (2021).
783 24. Battle, A. *et al.* Characterizing the genetic basis of transcriptome diversity through RNA-

784 sequencing of 922 individuals. *Genome Res.* **24**, 14–24 (2014).
785 25. Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic
786 Association Studies Using Summary Statistics. *PLOS Genet.* **10**, e1004383 (2014).
787 26. Liu, X. *et al.* GBAT: a gene-based association test for robust detection of trans-gene
788 regulation. *Genome Biol.* **21**, 211 (2020).
789 27. Mu, Z. *et al.* The impact of cell type and context-dependent regulatory variants on human
790 immune traits. *Genome Biol.* **22**, 122 (2021).
791 28. Westra, H.-J. *et al.* Systematic identification of trans eQTLs as putative drivers of known
792 disease associations. *Nat. Genet.* **45**, 1238–1243 (2013).
793 29. Luijk, R. *et al.* Genome-wide identification of directed gene networks using large-scale
794 population genomics data. *Nat. Commun.* **9**, 3097 (2018).
795 30. Morris, J. A. *et al.* Discovery of target genes and pathways of blood trait loci using pooled
796 CRISPR screens and single cell RNA sequencing. 2021.04.07.438882 Preprint at
797 https://doi.org/10.1101/2021.04.07.438882 (2021).
798 31. Raudvere, U. *et al.* g:Profiler: a web server for functional enrichment analysis and
799 conversions of gene lists (2019 update). *Nucleic Acids Res.* **47**, W191–W198 (2019).
800 32. Schwickert, T. A. *et al.* Stage-specific control of early B cell development by the transcription
801 factor Ikaros. *Nat. Immunol.* **15**, 283–293 (2014).
802 33. Lemarié, M., Bottardi, S., Mavoungou, L., Pak, H. & Milot, E. IKAROS is required for the
803 measured response of NOTCH target genes upon external NOTCH signaling. *PLOS Genet.*
804 **17**, e1009478 (2021).
805 34. Cui, J. *et al.* NLRC5 Negatively Regulates the NF-κB and Type I Interferon Signaling
806 Pathways. *Cell* **141**, 483–496 (2010).
807 35. Kobayashi, K. S. & van den Elsen, P. J. NLRC5: a key regulator of MHC class I-dependent
808 immune responses. *Nat. Rev. Immunol.* **12**, 813–820 (2012).
809 36. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data.

810 *Nature* **562**, 203–209 (2018).
811 37. Liu, J. Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel
812 disease and highlight shared genetic risk across populations. *Nat. Genet.* **47**, 979–986
813 (2015).
814 38. de Lange, K. M. *et al.* Genome-wide association study implicates immune activation of
815 multiple integrin genes in inflammatory bowel disease. *Nat. Genet.* **49**, 256–261 (2017).
816 39. Ferreira, M. A. *et al.* Shared genetic origin of asthma, hay fever and eczema elucidates
817 allergic disease biology. *Nat. Genet.* **49**, 1752–1757 (2017).
818 40. Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed-model association
819 for biobank-scale datasets. *Nat. Genet.* **50**, 906–908 (2018).
820 41. INTERNATIONAL MULTIPLE SCLEROSIS GENETICS CONSORTIUM. Multiple sclerosis
821 genomic map implicates peripheral immune cells and microglia in susceptibility. *Science* **365**,
822 eaav7188 (2019).
823 42. Ferreira, M. A. R. *et al.* Genetic Architectures of Childhood- and Adult-Onset Asthma Are
824 Partly Distinct. *Am. J. Hum. Genet.* **104**, 665–684 (2019).
825 43. Bentham, J. *et al.* Genetic association analyses implicate aberrant regulation of innate and
826 adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. *Nat. Genet.*
827 **47**, 1457–1464 (2015).
828 44. Zou, S. *et al.* SNP in human ARHGEF3 promoter is associated with DNase hypersensitivity,
829 transcript level and platelet function, and Arhgef3 KO mice have increased mean platelet
830 volume. *PLOS ONE* **12**, e0178095 (2017).
831 45. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide
832 association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
833 46. Artavanis-Tsakonas, S., Rand, M. D. & Lake, R. J. Notch Signaling: Cell Fate Control and
834 Signal Integration in Development. *Science* **284**, 770–776 (1999).
835 47. Liu, B., Gloudemans, M. J., Rao, A. S., Ingelsson, E. & Montgomery, S. B. Abundant

associations with gene expression complicate GWAS follow-up. *Nat. Genet.* **51**, 768–769 (2019).

Acknowledgements

We thank Y. Li, A. Dahl, Y. Gilad and Z. Mu for useful scientific discussions. We thank N. Gonzales, C. Jones and S. Sumner for editing the manuscript. This work was funded through the NIGMS Maximizing Investigators' Research Award (R35GM138084).

Supplementary information

Supplementary Text and Figures

Supplementary Figures 1-18, Supplementary Tables 1-17 and Supplementary Note.