Reverse regression increases power for detecting trans-eQTLs

Saikat Banerjee[†,1], Franco L. Simonetti[†,1], Kira E. Detrois[1], Anubhav Kaphle[1,2], Raktim Mitra[3], Rahul Nagial[3], and Johannes Söding[1,*]

[1]Max-Planck Institute for Biophysical Chemistry, 37077 Göttingen, Germany
[2]Georg-Augustät University, 37075 Göttingen, Germany
[3]Indian Institute of Technology, Kanpur, India
[*]soeding@mpibpc.mpg.de
[†]These authors contributed equally.

Abstract

Knowledge of trans-acting expression quantitative trait loci (trans-eQTLs) regulating distant target genes can reveal biological mechanisms that link single nucleotide polymorphisms (SNPs) with complex traits. However, identifying trans-eQTLs is challenging because their effect sizes are typically small and simple regression of millions of SNPs against each gene expression imposes a severe multiple testing burden. Here we present Tejaas, an efficient method to discover trans-eQTLs using L_2-regularized 'reverse' multiple regression of the gene expressions against each SNP. Tejaas aggregates evidence of small trans-effects from all distant target genes simultaneously while being robust against the strong correlation of the gene expressions. Tejaas, coupled with a novel k-nearest neighbors algorithm for unsupervised confounder correction, discovers 18 851 unique trans-eQTLs across 49 tissues from the GTEx (v8) data. They are enriched in several functional signatures, including mediation via proximal genes, chromatin accessibility and occurrence in enhancer and promoter regions. Several trans-eQTLs overlap with disease-associated SNPs and reveal underlying transcriptional regulation mechanism. Tejaas is available at https://github.com/soedinglab/tejaas

Introduction

Over the last decade, genome-wide association studies (GWASs) have identified over 100 000 unique associations between single nucleotide polymorphisms (SNPs) and human traits [1, 2]. However, our understanding of the underlying mechanism through which SNPs influence the risk of complex, non-infectious diseases has not grown in proportion because more than 90% of the SNPs identified by GWAS do not reside in coding regions [3].

Several lines of evidence suggest the involvement of these SNPs in regulation of intermediate cellular phenotypes [4], including gene expression levels [5], chromatin accessibility [6], chromatin state [7] and protein abundance [8]. SNPs that are associated with the gene expression levels are called expression quantitative trait loci (eQTL). For example, non-coding SNPs lying in cell-type specific enhancer regions can alter the expression of target genes [9], which can then increase or decrease disease risk [10].

The eQTLs, which are proximal ($<$ 1Mb) to the regulated genes are called cis-eQTLs and the eQTLs, which regulate distal genes located elsewhere in the genome, are called trans-eQTLs. Heritability estimates from 856 female twins suggest that, on average, cis-eQTLs explain $< 40\%$ of the heritable variation in the gene expression of adipose tissue, lymphoblastoid cell line and skin tissue [11]. For African Americans, cis-eQTLs explain only $12 \pm 3\%$ of the heritable gene expression variation in the lymphoblastoid cell line [12]. The remaining heritability of gene expression levels is generally attributed to trans-eQTLs [11, 12].

Discovering trans-eQTLs is important not only for explaining the observed gene expression variations, but also for understanding the transcriptional regulation mechanisms, which can then shed light on the aetiology of complex diseases. For example, a recent ATAC-Seq study [13] identified a single SNP that alters the chromatin accessibility across multiple genomic loci including the BLK

region, which is associated with multiple autoimmune diseases. In spite of such growing evidence of long-range regulation, the systematic discovery of trans-eQTLs [14, 15] has hardly advanced due to the enormous statistical challenges involved.

Discovery of eQTLs *in silico* is possible by analyzing paired genotype and gene expression data collected in parallel from many individuals. Simple linear regression is commonly employed on each SNP-gene pair to test for associations. Cis-eQTLs often have a large effect size and the number of association tests are limited to genes in the vicinity (generally < 1Mb) of each SNP. Therefore, relatively modest sample sizes enable their detection using a simple regression method. In contrast, identification of trans-eQTLs remains a major challenge because they (1) tend to have a smaller effect size, (2) impose a severe multiple testing burden due to the need to examine possible association between each gene and all SNPs across the genome, and (3) are frequently tissue- and context-specific. Therefore, several subjective constraints are imposed to reduce false positives while scanning for trans-eQTLs, for instance, working with a reduced set of SNPs that are associated with disease traits or that have known cis effects. Subjective constraints might sacrifice the discovery of many true trans-eQTLs.

Scientists are now trying to use known biological signatures of trans-eQTLs to boost the power to detect them. For example, Rakitsch and Stegle [16] developed a two-stage gene-network linear mixed model (GNetLMM), which implicitly assumed that a trans-eQTL is linked to a target trans-eGene via an intermediate cis-associated gene. This property of trans-eQTLs allowed them to construct local, directed gene-regulatory networks and identify exogenous genes that account for hidden variation in the target trans-eGene. Conditioning the trans-eGene on the exogenous gene improved the power for the trans-eQTL association test. Hore *et al.* [17] assumed another biological signature: Trans-eQTLs create variation in the expression levels of gene networks across tissues. They decomposed the three-dimensional (individual, genes and tissues) array or tensor using a Variational Bayes (VB) approach with sparsity enforced by a spike-and-slab prior to obtain latent components that represent the major modes of variation in the data. They tested each latent component against genetic variation across the genome to discover underlying QTL effects. The VB optimization results in different latent components in separate runs and the authors ensured robustness by only considering latent components that are persistently found across multiple runs.

In this work, we rely on another commonly considered aspect of trans-eQTLs, that they regulate multiple genes simultaneously [18]. Instead of looking at each SNP-gene pair, we try to find SNPs which regulate tens to hundreds of genes. Earlier, Brynedal *et al.* [19] used cross-phenotype meta-analysis (CPMA) to find trans-eQTLs based on the same property. They evaluated the p-values for the pairwise linear association of a candidate SNP with all available gene expression levels. For the *null* SNPs with no trans effect, p-values will follow a uniform distribution and the $-\log p$-values will follow a chi-square distribution. A *trans*-eQTL will be associated with more genes than expected by chance and the distribution of $-\log p$-values will be overdispersed near zero. The CPMA statistic estimates the overdispersion near zero. However a major limitation of this approach arises due to strong correlations among the gene expression levels, which induce strong correlations among the p-values. This leads to overdispersion near zero by chance, increasing the false positive rate and diminishing the power of the method significantly. For example, let us consider a SNP that changes the expression of tens or hundreds of genes. With increasing strength of the gene expression correlation, the probability of finding similar associations to null SNPs by chance increases and the significance of the truly causal SNP decreases.

In order to circumvent the problem of correlation among the gene expression levels, we use multiple regression in the reverse direction by explaining the minor allele counts using the gene expression levels. Since available eQTL data generally has significantly lower number of samples than the number of expressed genes, we used an L_2 regularizer (equivalent to a Gaussian prior) to limit model complexity. Our motivation is that multiple regression using a regularizer should help find the causal genes, with directly affected genes explaining away the effect of the genes which are merely correlated. There are two major benefits of our approach: (1) Although the effect sizes of the trans-eQTLs on individual genes are small, the signal is accumulated over many genes, making them easier to discover. (2) The multiple testing problem is reduced significantly because each SNP is tested only once instead of being tested against every gene separately. We note that the work of Brynedal *et al.* [19] also has the same benefits but suffers from the correlation among the single SNP-gene p-values.

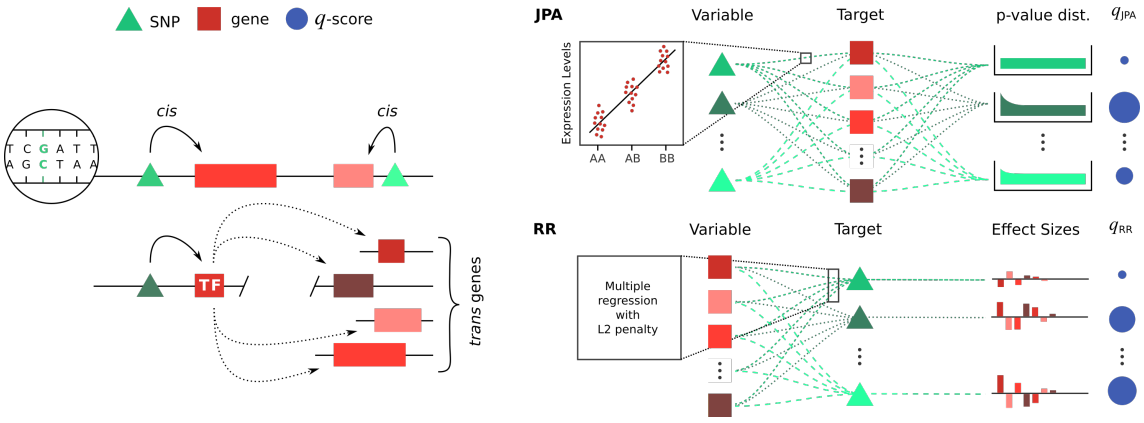

Figure 1: **Summary of methods implemented in Tejaas**. We assume that *trans*-eQTLs affect multiple genes simultaneously via some transcription factors (TFs), as shown in the left. In Joint P-value Analysis (JPA) we analyze the distribution of p-values for the association of a candidate SNP with all available gene expression levels. The JPA-score (q_{jpa}) estimates whether this distribution of p-values is enriched near zero. In Reverse Regression (RR), we perform a L_2 regularized multiple linear regression using the genotype as target, and gene expression levels as explanatory variables. The RR-score (q_{rr}) estimates whether there are 'significant' number of genes with non-zero effect sizes.

Results

Methods overview. Tejaas (see URLs) is a new tool for discovering trans-eQTLs. It implements the Reverse Regression (**RR-score**, q_{rr}) for ranking trans-eQTLs and a non-linear **KNN correction** for removing confounding effects from the gene expression. We wanted to compare Tejaas with CPMA statistic of Brynedal *et al.* [19] because we use the same underlying assumption. As there are no currently available software for CPMA, we also implemented the Joint P-value Analysis (**JPA-score**, q_{jpa}) within Tejaas as an alternative. Both the JPA-score and RR-score are summarized in Fig. 1 and briefly introduced in the ensuing paragraphs. For a detailed discussion, along with explanation on software usage and choosing model parameters, please refer to Supplementary Sec. 2.

JPA evaluates the distribution of p-values of the pairwise linear association of a candidate SNP with all available gene expression levels. The *null* SNPs (no trans-effect) will have a uniform distribution of p-values, while *trans*-eQTLs will be associated with more genes than expected by chance, leading to overdispersion near zero. We defined the JPA-score (q_{jpa}) as a statistic which estimates whether the distribution of p-values is significantly overdispersed near zero.

Reverse Regression (RR) performs a multiple linear regression using expression levels of all genes to explain the genotype of a candidate SNP. In contrast to conventional methods, the direction of the regression is reversed, with the gene expressions as explanatory variables. In brief, let $\mathbf{x}$ denote the vector of scaled and centered minor allele counts of a SNP for N samples and $\mathbf{Y}$ be the $G \times N$ matrix of preprocessed expression levels for G genes. We model $\mathbf{x}$ with a normal distribution whose mean depends linearly on the gene expression through a vector of regression coefficients $\boldsymbol{\beta}$:

$$p\left(\mathbf{x} \mid \mathbf{Y}\right) \propto \mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\beta}^{\mathsf{T}}\mathbf{Y}, \mathbb{I}\sigma^2\right) . \tag{1}$$

Generally, the number of explanatory variables (genes) is much larger than the number of samples ($G \gg N$) in currently available eQTL data sets. To avoid overtraining, we introduce a normal prior on $\boldsymbol{\beta}$, with mean 0 and variance γ^2,

$$p(\boldsymbol{\beta}) = \mathcal{N}\left(\boldsymbol{\beta} \mid 0, \gamma^2\right) . \tag{2}$$

This L_2 regularization pushes the effect size of non-target genes towards zero. Ideally, a spike-and-slab prior should work better than the current model but is analytically intractable and is too slow to approximate. We calculated the significance of the trans-eQTL model ($\boldsymbol{\beta} \neq \mathbf{0}$) compared to the

null model ($\boldsymbol{\beta} = \mathbf{0}$) using Bayes theorem to define the RR-score (q_{rr}),

$$\ln\left(\frac{P\left(\boldsymbol{\beta} \neq \mathbf{0} \mid \mathbf{x}, \mathbf{Y}\right)}{P\left(\boldsymbol{\beta} = \mathbf{0} \mid \mathbf{x}, \mathbf{Y}\right)}\right) = q_{\text{rr}} + \text{const.}, \text{ where}$$

$$q_{\text{rr}} := \mathbf{x}^{\mathsf{T}}\mathbf{W}\mathbf{x} \tag{3}$$

$$\text{and, } \mathbf{W} := \frac{1}{\sigma^2}\mathbf{Y}^{\mathsf{T}}\left(\mathbf{Y}\mathbf{Y}^{\mathsf{T}} + \frac{\sigma^2}{\gamma^2}\mathbb{I}_G\right)^{-1}\mathbf{Y}\,. \tag{4}$$

For each SNP, the null distribution of q_{rr} can be obtained by randomly permuting the sample labels of the genotype multiple times. Note that this null distribution will depend on the minor allele frequency and preprocessing of the SNPs but it is computationally infeasible to obtain the null distribution empirically for each SNP independently. We could, however, analytically obtain the expectation and variance of q_{rr} under this permuted null model. Assuming that the null distribution is Gaussian, we calculated a p-value to get the significance of any observed q_{rr}.

Our method requires the gene expression matrix $\mathbf{Y}$ to have full column rank. Any covariate correction method involving linear regression would also reduce the column rank of $\mathbf{Y}$ and cannot be used for preprocessing the gene expression for calculating q_{rr}. Therefore, we developed an unsupervised non-linear correction using k-nearest neighbors, which we call KNN correction (Supplementary Sec. 3.2) to remove confounding effects.

Simulation studies. We ran simulations to benchmark Tejaas against existing methods, to compare different preprocessing methods for removing confounders and to estimate the model parameters. Several software packages exist for finding trans-eQTLs using single SNP-gene regression and we used MatrixEQTL [20] as a representative of these methods. As an alternative for CPMA, we used our JPA implementation in Tejaas, henceforth referred to as JPA (∼CPMA).

For the simulations, we used the strategy of Hore *et al.* [17], the details of which are discussed in the Supplementary Sec. 4. In brief, we sampled $I = 12\,639$ SNPs from the real genotype of the Genotype Tissue Expression (GTEx) project to retain the complexity of real data. We simulated the expression levels for $G = 12\,639$ genes, containing non-genetic signals (background correlation and confounding factors) and genetic signals (*cis* and *trans* effects). The background correlation of the gene expression was obtained with same covariance structure as that of the artery-aorta tissue of the GTEx project. The strength for confounder effects, cis effects and trans effects were obtained from Hore *et al.* , while we additionally introduced genotype principal components as confounders to simulate population substructure.

For every simulation, we randomly selected 800 SNPs to be cis-eQTLs, out of which 30 SNPs were also trans-eQTLs [17]. The trans-eQTLs regulated the nearest gene via cis effect. This cis target gene was considered a transcription factor (TF) and regulated multiple target genes downstream (excluding other TFs). Let M_{trans} be the number of target genes regulated by each TF and $|\beta_{gj}| \sim \text{Gamma}\left(\psi^{\text{trans}}, 0.02\right)$ be the effect size of the j^{th} TF on the g^{th} target gene.

In Fig. 2a, we show the results for different covariate correction strategies: (1) without any covariate correction (denoted as 'None'), (2) the most commonly used confounder correction method using residuals after linear regression of the gene expression with known covariates (denoted as 'CCLM'), and (3) KNN correction with 30 nearest neighbors. The GTEx consortium recommended using inverse normal transformation of the gene expression data before applying covariate correction. Hence, the CCLM correction was done on inverse normal transformed gene expression data. The KNN correction was applied directly on the gene expression data because we found that trans-eQTL signals are removed if KNN correction is applied on inverse normal transformed data (Supplementary Fig. S4). We then applied MatrixEQTL, JPA (∼CPMA) and Tejaas (q_{rr}) to find trans-eQTLs from the corrected gene expressions. The ranking with q_{rr} depends on the parameter γ and we set it empirically at $\gamma = 0.2$ (Supplementary Fig. S3). For Tejaas, we used the cis-masking option (Supplementary Sec. 2.8) in our software, *i.e.,* removed all genes located within ±1Mb of each SNP to avoid the strong cis-eQTL signals. For each preprocessing option, we performed 20 simulation replicates. We compared the ranking of trans-eQTLs using the partial area under the ROC curve (pAUC) where the false positive rate (FPR) ≤ 0.1. This is because we are only interested in the top predictions.

Our results show that the KNN correction is the most effective covariate correction for Tejaas. Unlike simulations, in real data we do not have exact knowledge of the confounders. Hence, it is

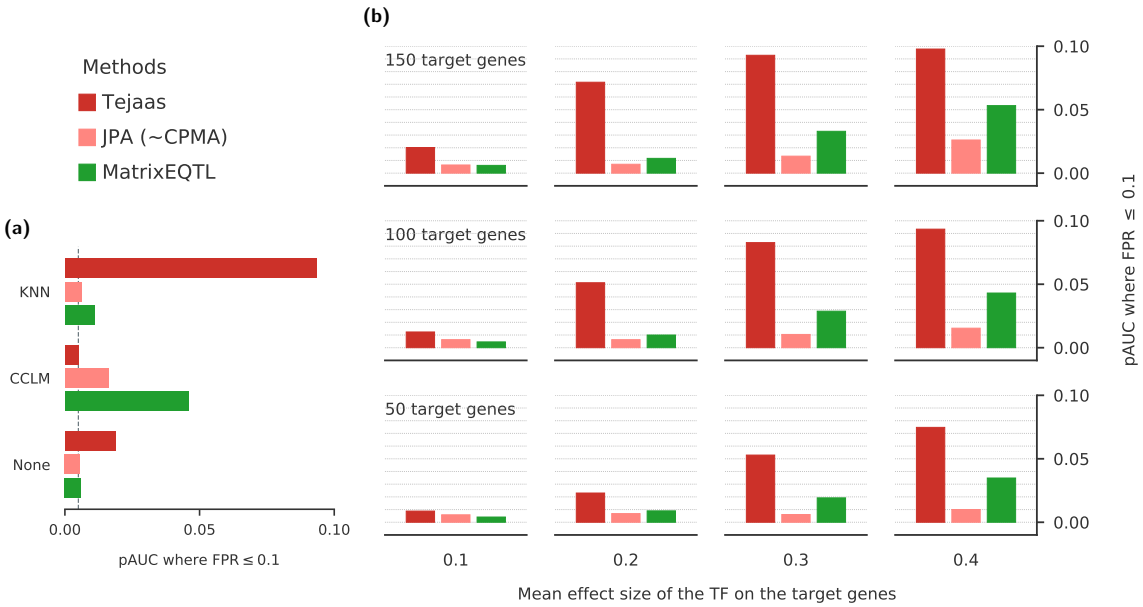

Figure 2: **Comparison of methods on simulated data.** We compared the ranking performance of finding trans-eQTLs using Tejaas (q_{rr} with $\gamma = 0.2$), JPA (~CPMA) and MatrixEQTL, as measured with the partial area under the ROC curve (pAUC) where the false positive rate (FPR) ≤ 0.1. The maximum possible pAUC for a perfect method is 0.1 and the expected pAUC for a random method is 0.005. The mean pAUC for each method was obtained by averaging over 20 simulation replicates. Panel (a) shows different confounder correction options: no covariate correction (None), covariate correction using linear regression of known confounders (CCLM) on inverse normal transformed gene expression, and KNN correction (30 nearest neighbors) applied directly on the gene expression data. Panel (b) shows different strength of trans-eQTL signals obtained by using (1) different number of target genes for the TF linked with the trans-eQTL, $M_{\text{trans}} = 150$, 100 and 50 from top to bottom panels and (2) different distribution for sampling the effect sizes of the TF on the target genes, $\langle \beta_{gj} \rangle = 0.1$, 0.2, 0.3 and 0.4 from left to right (corresponding to $\psi^{\text{trans}} = 5$, 10, 15 and 20 respectively).

encouraging to note that the KNN correction can remove the background noise in an unsupervised fashion. Covariate correction using linear regression (CCLM) is effective for traditional SNP-gene pair analysis (if the true covariates are known) but unfortunately it reduces the rank of the gene expression matrix and breaks down the Tejaas ranking (Supplementary Sec. 2.5 and Fig. S2).

In Fig. 2b, we compared different methods for discovering trans-eQTLs at different signal strengths. The varying signal strength was simulated by tuning (1) the number of target genes (M_{trans}) of the TF linked to the trans-eQTL and (2) the effect size of the TF on the target genes, which is sampled from a Gamma$(\psi^{\text{trans}}, 0.02)$ distribution with mean $\langle \beta_{gj} \rangle = 0.02\psi^{\text{trans}}$. For discovering trans-eQTLs, Tejaas used KNN correction with $K = 30$ directly on the gene expression and q_{rr} with $\gamma = 0.2$. For MatrixEQTL and JPA (~CPMA), all known covariates introduced in the previous simulation steps were corrected out using CCLM on the inverse normal transformed gene expression. We compared the accuracy of the methods using the partial area under the ROC curve (pAUC) where the false positive rate (FPR) is ≤ 0.1. We find that JPA (~CPMA) has slightly lower pAUC than MatrixEQTL, while Tejaas performs best with significantly higher pAUC at all values of M_{trans} and $\langle \beta_{gj} \rangle$, even without exact knowledge of covariates. At very low signals, for example with mean effect size of 0.1 and 50 target genes, the ranking performance of all methods are significantly reduced and we would need a larger sample size for efficient trans-eQTL discovery. However, Tejaas improves more rapidly compared to JPA (~CPMA) or MatrixEQTL with increasing signal strength of the trans-eQTLs.

Genotype Tissue Expression trans-eQTL analysis. To illustrate Tejaas in a relevant data set, we analyzed trans-eQTLs across 49 human tissues using data from the Genotype Tissue

Expression (GTEx) project [21–23]. The GTEx project aims to provide insights into mechanisms of gene regulation by collecting gene expression measurements from multiple tissues in human donors. The latest analysis on the GTEx v8 release yielded 143 trans-eQTLs across all tissues, with 121 linked to protein coding genes and 22 linked to lincRNA [24]. Of these trans-eQTLs, 47 trans-eQTLs were observed in testis alone.

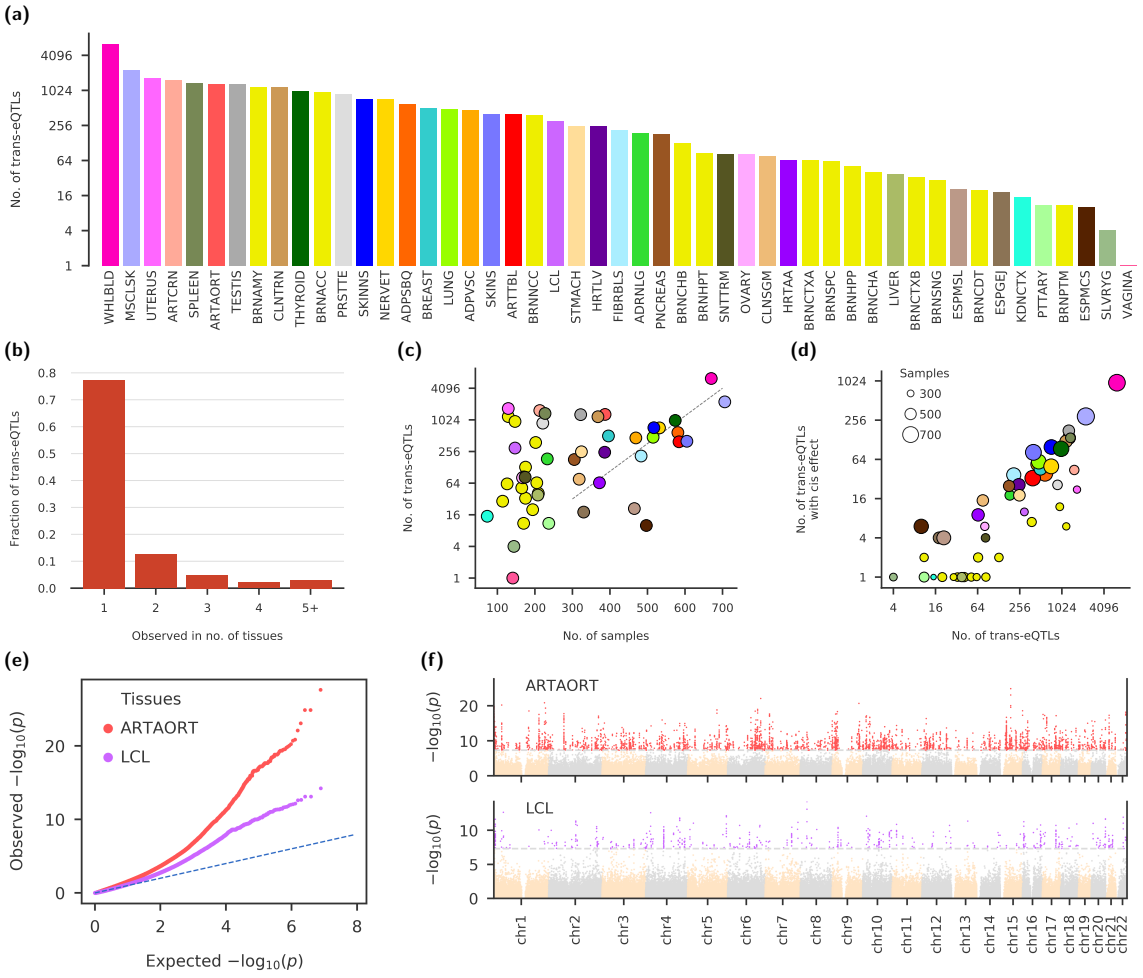

Figure 3: **Summary of trans-eQTLs identified by Tejaas in GTEx.** Using Tejaas, we calculated genome-wide q_{rr} and corresponding p-values for each SNP in 49 GTEx tissues, using KNN corrected gene expression. The tissues were broadly classified into two groups: 45 tissues analyzed with prior $\gamma = 0.1$, and 4 tissues analyzed with prior $\gamma = 0.006$. For each SNP, we excluded the proximal genes ($\pm$1Mb) from analysis. We predicted all SNPs with $p < 5 \times 10^{-8}$ as trans-eQTLs and pruned the list to obtain lead trans-eQTLs in each independent LD region. (a) The number of lead trans-eQTLs discovered per tissue (note the logarithmic scale on y-axis). For tissue abbreviations, please refer to Appendix 2 of Supplementary. (b) The proportion of trans-eQTLs (y-axis) which are found in a given number of tissues (x-axis). More than 75% of the trans-eQTLs are found uniquely in single tissues. (c) The number of lead trans-eQTLs as a function of the number of samples in each tissue. Each point is a tissue. For tissues with more than 250 samples, we observe an exponential increase in the number of trans-eQTLs with increasing sample size (regression line shown in gray). (d) Trans-eQTLs act via cis-eGenes. On the x-axis, we show the number of lead trans-eQTLs and on the y-axis, we show the number of lead trans-eQTLs which also happen to be cis-eQTLs in the GTEx analysis. (e) Representative example of quantile-quantile plot in two tissues: artery aorta (ARTAORT) and EBV-transformed lymphocytes (LCL). (f) Corresponding Manhattan plot for the above two tissues, showing the $-\log_{10}(p)$-values for genome-wide variants.

We used the GTEx genotype and gene expression data provided in dbGaP (accession phs000424) for our analysis. Details of the preprocessing steps are discussed in Supplementary Sec. 5. In brief, we converted the gene expression read counts obtained from phASER to standardized TPMs (Transcripts per Millions) for all the 49 tissues and used KNN correction with 30 nearest neighbors to remove confounders. We then estimated the optimal values of γ for each tissue. and broadly classified the tissues into two groups: (a) 45 tissues analyzed with $\gamma = 0.1$ and (b) 4 tissues analyzed with $\gamma = 0.006$ (Supplementary Fig. S8). For each SNP, we removed all corresponding genes from the vicinity ($\pm$1Mb) to avoid the relatively stronger cis-eQTL signals inflating q_{rr}. We predicted all SNPs with $p < 5 \times 10^{-8}$ as trans-eQTLs for further analyses. To avoid double-counting trans-eQTLs that are in LD with one another, we pruned the list of trans-eQTLs by retaining only the best trans-eQTL (with lowest p-value) in each independent LD region defined by $r^2 > 0.5$ in 200kb windows.

We discovered 16 929 unique lead trans-eQTLs across all GTEx tissues except brain (Fig. 3a) and 1 922 unique lead trans-eQTLs in brain tissues. Consistent with our simulation results, Tejaas is able to discover more trans-eQTLs than traditional methods in GTEx. We find that the trans-eQTLs are tissue-specific, with 77.3% of the trans-eQTLs being discovered in single tissues (Fig. 3b). The number of trans-eQTLs discovered increases exponentially with the number of samples (Fig. 3c) for $N > 250$, indicating that larger studies would be able to discover more trans-eQTLs. In Fig. 3d, we show that the fraction of trans-eQTLs with a cis-effect vary proportionally with the total number of trans-eQTLs in each tissue, implying that a significant proportion of trans-eQTLs act via cis-eGenes. To note the results of Tejaas at single-tissue level, we show the quantile-quantile plot (Fig. 3e) and Manhattan plot (Fig. 3f) for two representative tissues, namely artery-aorta (ARTAORT) and EBV-transformed lymphocytes (LCL).

Functional enrichment analyses of trans-eQTLs. Enrichment of the newly discovered trans-eQTLs in functionally relevant regulatory annotation of the genome provides insight into the underlying biological mechanisms of the trans-eQTLs. Given the lack of experimental validation, the biological relevance of the trans-eQTLs suggested by their functional enrichment in several diverse, independent experiments is indicative of them being true positives. The enrichment of the functional features were measured in comparison to a random set of SNPs obtained by sampling from the GTEx genotype (Supplementary Sec. 5.6).

A possible mechanism of trans-eQTLs involves mediation via cis-eQTLs, where the cis-eGene (for example, some known transcription factor) might regulate distant genes. Indeed, we observed a significant enrichment of trans-eQTLs being also cis-eQTLs to proximal genes in the same tissue (Fig. 4a), although our trans-eQTLs were discovered excluding all genes in the vicinity of that SNP. We also observed that the cis-mediator genes have a higher proportion of being protein-coding than the background distribution of GTEx cis-eGenes (Fig. 4d). For this analysis, the cis-eQTLs and their target genes (mediator genes for trans-eQTLs) were obtained from the GTEx portal. Although we rarely found significant enrichment of transcription factors (TFs) among the cis-mediator genes, trans-eQTLs are enriched in proximal locations ($<$ 100Kb) of TFs (Fig. 4a).

Reporter assay QTLs (raQTLs) are SNPs that alter the activity of putative regulatory elements (enhancers and promoters), partially in a cell-type-specific manner. In Fig. 4a, we show the enrichment of the trans-eQTLs in two sets of raQTLs for two cell types, K562 and HepG2. The raQTL data was obtained from the survey of regulatory elements (SuRE) [26]. K562 is an erythroleukemia cell line with strong similarities to whole blood tissue and HepG2 cells are derived from hepatocellular carcinoma with similarities to liver tissue.

DNase I hypersensitive sites (DHSs) are accessible regions of the chromatin, often considered as markers in the genome for regulatory elements (promoters, enhancers, insulators and other control regions) and are functionally associated with transcriptional activity. We found that the trans-eQTLs occur within these regions more often than expected by chance, showing significant DHS enrichment in most tissues (4b).

With well-powered trans-eQTL mapping by Tejaas, it also becomes possible to describe and disentangle tissue-specific enrichments. Using chromatin state predictions from a set of tissues from the Roadmap Epigenomics project [28], we show that the trans-eQTLs are enriched in enhancer, bivalent and repressed polycomb regions of their matched tissues (Fig. 4c). They are depleted in the inaccessible heterochromatin regions for most of the tissues while they show no enrichment or

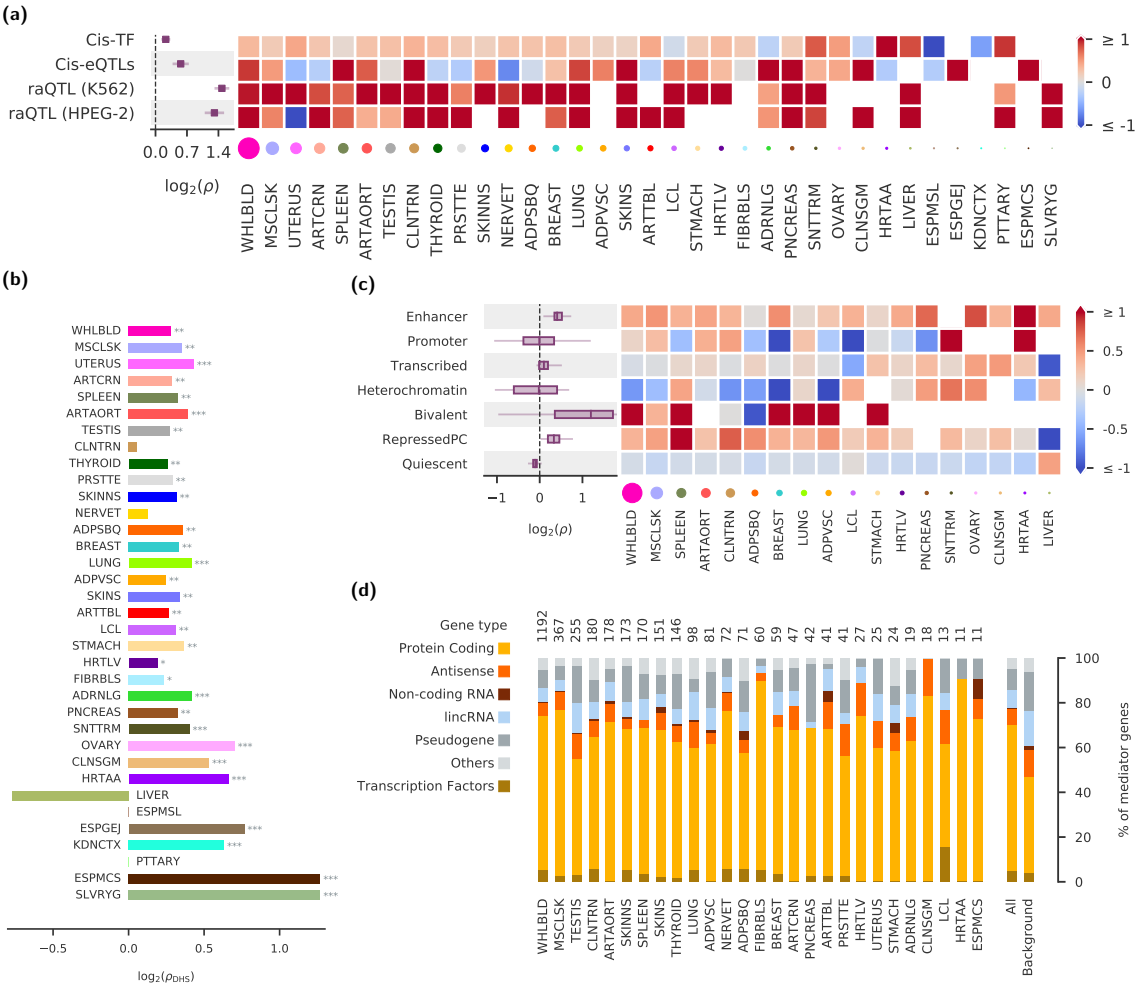

Figure 4: **Functional mechanisms of genetic regulatory effects of trans-eQTLs.** We calculated the enrichment (ρ) of the trans-eQTLs in functionally relevant regulatory annotations of the genome. (a) Log_2 enrichment of lead trans-eQTLs in all non-brain tissues to occur within $\pm$ 100Kb distance from transcription factors reported in [25] (cis-TF), to mediate via known cis-eQTLs reported in the GTEx v8 analysis (Cis-eQTLs) and to occur in reporter assay QTLs (raQTLs) showing activity on enhancers and promoter [26] in K562 and HepG2 cells. In the heatmap, the color signifies the $\log_2(\rho)$ as shown by the scale on the right. On the x-axis, the tissues are labelled with corresponding abbreviations, and the area of the colored circles represents the number of lead trans-eQTLs discovered in that tissue. On the left panel, we show the mean $\log_2(\rho)$ across all tissues. (b) Log_2 enrichment of the trans-eQTLs in all non-brain tissues in DHS regions reported in [27]. Each colored bar shows the $\log_2(\rho)$ for the corresponding tissue, and their p-value for significance is denoted by the stars ($p \leq 0.05$ denoted by *, $p \leq 0.01$ denoted by ** and $p \leq 0.001$ denoted by ***). (c) Tissue-matched $\log_2$ enrichments for cis-regulatory elements (labels on the y-axis). GTEx tissues were matched to their corresponding tissue annotation in the Roadmap Epigenomics Project [28]. Shown here are tissues, which had a corresponding matching tissue in Roadmap and had at least 10 trans-eQTLs. The x-axis is the same as in panel (a). On the left, we show the distribution of $\log_2(\rho)$ across all tissues with bar plots. (e) Gene type composition for target genes of cis-eQTLs with trans-eQTL effects. Cis-eQTLs and their target genes were obtained from the GTEx portal (v8). Here we only show tissues which had at least 10 cis-eQTL mediators. Each bar correspond to a tissue, with colors proportional to the composition of the mediator genes. The number of mediator genes in each tissue is mentioned at the top of each bar.

depletion in inactive quiescent regions.

Association with complex diseases. We investigated the overlap of the novel trans-eQTLs discovered by Tejaas with GWAS variants of complex traits to find transcriptional regulatory mechanisms through which SNPs affect complex diseases. We used the GWAS summary statistics from 87 traits harmonized and imputed to GTEx v8 variants with MAF > 0.01 using only European samples by Barbeira *et al.* [29]. These 87 traits were broadly classified into a range of disease categories. For example, the category "Immune" contained all studies related to immune diseases such as asthma or psoriasis.

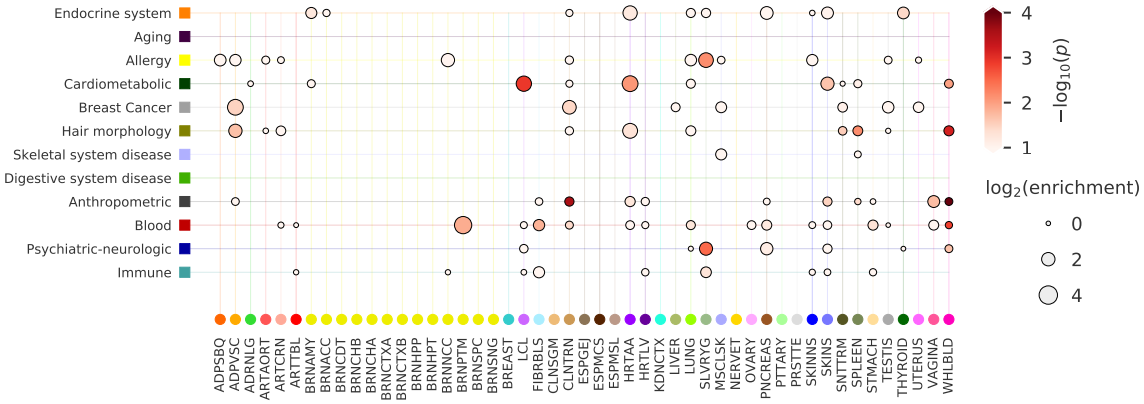

Figure 5: **Trans-eQTLs are enriched in GWAS SNPs for complex diseases.** We calculated the enrichment of lead trans-eQTLs predicted by Tejaas across all tissues (x-axis) in GWAS hits of multiple disease categories (y-axis). Each point in the plot represents a pair of tissue and disease category, the size of the point scales with the $\log_2$ enrichment, and the color scales with the significance $(-\log_{10}(p))$ of the enrichment.

We calculated enrichments for each tissue in each individual trait (Supplementary Sec. 6 and Fig. S13) and each disease category (Fig. 5). We considered all SNPs with imputed $p < 10^{-7}$ to be a significant GWAS hit for the corresponding study or disease category. There are several tissue - disease category pairs that have a clear biological relationship. For example, trans-eQTLs discovered in whole blood are 1.3-fold enriched ($p = 0.0014$) in the disease category of "Blood", which contains different studies investigating varying blood cell counts such as those of red blood cells and lymphocytes. Trans-eQTLs in whole blood and heart atrial appendage are 1.7 and 7-fold enriched in cardiometabolic traits, with $p = 0.01$ and $p = 0.008$ respectively. The cardiometabolic disease category includes studies on cholesterol levels, blood pressure and coronary artery disease, among others.

GWAS-associated trans-eQTLs can provide insight to previously unknown disease pathways. For example, three of our trans-eQTLs rs7864322, rs4297160 and rs10983975 (all in chr9q22) discovered in thyroid tissue were found to be associated with hypothyroidism. These trans-eQTLs control the expression of nearby PTCSC2 lncRNA, a thyroid-specific regulator, and FOXE1 gene, which is known to play an important role in thyroid development. The distant target genes were enriched in the 'thyroid hormone signaling' pathway, indicating possible disease mechanism. For instance, the DIO1 gene in chr1 targeted by rs4297160 plays an important role in the production of T3, which is the main mediator of thyroid action.

Discussion

We developed Tejaas to increase the power for detecting trans-eQTLs by using two key innovations: the reverse regression and the KNN correction. We created a fast, parallel open-source software using these concepts, validated the method in a semi-realistic synthetic data and demonstrated its usefulness on a substantive real data set from the GTEx consortium to discover trans-eQTLs with clear biological and statistical significance. A marginal analysis of single SNP-gene pair or a method like CPMA would not have discovered those trans-eQTLs because of the low effect size of the trans-eQTLs on each single target gene and the strong correlated noise of the gene expression levels.

Tejaas complements other eQTL pipelines that focus on analyzing single SNP-gene pairs. Tejaas excels in discovering trans-eQTLs with multiple small effects by accumulating signals from many genes, which are regulated by that trans-eQTL, while other methods excel in discovering trans-eQTLs with a single large effect on a distant gene. Hence, we expect Tejaas and other existing methods to be complementary rather than overlapping.

Distinguishing causation from correlation is a long-standing and well-studied problem in statistics. In human genetics, the low number of samples compared to the explanatory variables (in our case, the number of genes) additionally requires controlling for sparsity. One widely accepted Bayesian approach is to use multiple regression with a sparsity-enforcing prior, for example the spike-and-slab prior, which has been previously used with success in different contexts such as fine-mapping in GWAS [30, 31]. Reverse regression controls for the correlation among the gene expression levels by using them as explanatory variables in a multiple regression setting. However, due to computational limitations, we had to use a normal prior which reduces model complexity but cannot enforce sparsity. In Tejaas, the standard deviation γ of the normal prior is not learnt from the data, but is set empirically. As expected, a high value of γ (> 0.2) would be too wide to reduce the model complexity and lead to overtraining. A low value of γ (< 0.001), on the other hand, will be too restrictive for the model and lead to false signals even with chance correlations of a single gene with a genotype. We encourage future users to make informed decision on the choice of γ for every gene expression profile, for example by first simulating a null set of q_{rr} on a simulated genotype and calculating the non-Gaussian parameter as explained in the supplementary text for the GTEx gene expressions.

The current method can be improved by introducing sparsity-enforcing priors on the effect size of the genes. It will not only improve accuracy for finding trans-eQTLs but also remove the dependency on γ. Additionally, it will enable robust variable selection, giving a more refined selection of trans-eQTL target genes. This could replace the current two-stage procedure for finding target genes with single SNP-gene pairwise regression after the trans-eQTLs are discovered by reverse regression. In spite of such multiple anticipated benefits, it remains technically challenging to implement such a method for large data sets.

Although reverse regression proved to be a powerful approach for finding trans-eQTLs, a major impediment was that the gene expression could not be corrected for confounders with the standard approach of regressing the known covariates or hidden PEER factors [32] (Supplementary Sec. 3.1). Hence, we proposed the KNN correction, a simple but powerful method for unsupervised confounder correction. Indeed, it corrected for most of the known covariates in GTEx (Supplementary Fig. S7). We expect the KNN correction to become an important tool for confounder correction in future eQTL pipelines.

Robust identification of trans-eQTLs and underlying disease pathways is crucial to further our understanding of genetics and its implication in complex diseases. Alongside larger studies with more samples, this will inevitably require more powerful methods for analyses. Tejaas represents a major step towards this goal.

Acknowledgments

This work was supported by the German Federal Ministry of Education and Research (BMBF) within the framework of the e:Med research and funding concept (grants e:AtheroSysMed 01ZX1313A-2014). We thank Markus Scholz for helpful communications and alerting us to the problem of strong cis-eQTLs, which led us to use cis-masking. We thank Hae Kyung Im and Alvaro Barbeira for email communications and kindly providing us the GWAS summary statistics from 87 traits harmonized and imputed to GTEx v8 variants. We thank our colleagues, especially Eli Levy Karin, Christian Roth, Wanwan Ge, Salma Sohrabi-Jahromi, Milot Mirdita and Ruoshi Zhang for helpful discussions and feedback. We used the data generated by the GTEx Consortium for the trans-eQTL analysis (accession phs000424). We thank the participants of the GTEx Consortium as well as all the research staff who worked on the data collection.

Author Contributions

J.S. conceptualized the problem, acquired funding, and supervised the project. J.S. designed the reverse regression with comments from S.B., F.L.S., A.K. and R.M.; S.B. and F.L.S. wrote the software with assistance from A.K. (JPA and KNN) and R.M. (RR). S.B. designed and performed the simulations. F.L.S. performed the GTEx preprocessing. F.L.S. and S.B. analyzed the GTEx data for discovering trans-eQTLs and assessing their functional enrichments. F.L.S. checked the contribution of known covariates in KNN and the effect of cross-mappable genes. K.E.D. analyzed the GWAS data. R.N. contributed to the initial phase of the project setup. S.B. wrote the original manuscript with assistance from F.L.S.; J.S., S.B. and F.L.S. reviewed and edited the manuscript.

Competing Interests

The authors declare no competing interests.

URLs

Tejaas: https://github.com/soedinglab/tejaas
GTEx trans-eQTLs: http://wwwuser.gwdg.de/~compbiol/tejaas/2020_03/gtex_v8_trans_eqtls_summary

References

[1] MacArthur, J., Bowler, E., Cerezo, M., et al. (2016). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res 45*, 896–901. DOI: 10.1093/nar/gkw1133.

[2] NHGRI-EBI (2019). GWAS Catalog. `https://www.ebi.ac.uk/gwas/`. [Online; accessed 24-February-2019].

[3] Maurano, M. T., Humbert, R., Rynes, E., et al. (2012). Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Sci 337*, 1190–1195. DOI: 10.1126/science.1222794.

[4] Visscher, P. M., Wray, N. R., Zhang, Q., et al. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *The Am J Hum Genet 101*, 5–22. DOI: 10.1016/j.ajhg.2017.06.005.

[5] Stranger, B. E., Nica, A. C., Forrest, M. S., et al. (2007). Population genomics of human gene expression. *Nat Genet 39*, 1217–1224. DOI: 10.1038/ng2142.

[6] Degner, J. F., Pai, A. A., Pique-Regi, R., et al. (2012). DNase I sensitivity QTLs are a major determinant of human expression variation. *Nat 482*, 390–394. DOI: 10.1038/nature10808.

[7] Kasowski, M., Kyriazopoulou-Panagiotopoulou, S., Grubert, F., et al. (2013). Extensive Variation in Chromatin States Across Humans. *Sci 342*, 750–752. DOI: 10.1126/science.1242510.

[8] Battle, A., Khan, Z., Wang, S. H., et al. (2015). Impact of regulatory variation from RNA to protein. *Sci 347*, 664–667. DOI: 10.1126/science.1260793.

[9] Corradin, O. and Scacheri, P. C. (2014). Enhancer variants: evaluating functions in common disease. *Genome Medicine 6*, 85. DOI: 10.1186/s13073-014-0085-3.

[10] Gupta, R. M., Hadaya, J., Trehan, A., et al. (2017). A Genetic Variant Associated with Five Vascular Diseases Is a Distal Regulator of Endothelin-1 Gene Expression. *Cell 170*, 522–533. DOI: 10.1016/j.cell.2017.06.049.

[11] Grundberg, E., Small, K. S., Hedman, Å. K., et al. (2012). Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat Genet 44*, 1084–1089. DOI: 10.1038/ng.2394.

[12] Price, A. L., Patterson, N., Hancks, D. C., et al. (2008). Effects of cis and trans Genetic Ancestry on Gene Expression in African Americans. *PLOS Genet 4*, 1–7. DOI: 10.1371/journal.pgen.1000294.

[13] Kumasaka, N., Knights, A. J., and Gaffney, D. J. (2019). High-resolution genetic mapping of putative causal interactions between regions of open chromatin. *Nat Genet 51*, 128–137. DOI: 10.1038/s41588-018-0278-6.

[14] Joehanes, R., Zhang, X., Huan, T., et al. (2017). Integrated genome-wide analysis of expression quantitative trait loci aids interpretation of genomic association studies. *Genome Biol 18*, 16. DOI: 10.1186/s13059-016-1142-6.

[15] Võsa, U., Claringbould, A., Westra, H.-J., et al. (2018). Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. *bioRxiv*. DOI: 10.1101/447367.

[16] Rakitsch, B. and Stegle, O. (2016). Modelling local gene networks increases power to detect trans-acting genetic effects on gene expression. *Genome Biol 17*, 33. DOI: 10.1186/s13059-016-0895-2.

[17] Hore, V., Viñuela, A., Buil, A., et al. (2016). Tensor decomposition for multiple-tissue gene expression experiments. *Nat Genet 48*, 1094–1100. DOI: 10.1038/ng.3624.

[18] Li, Q., Seo, J.-H., Stranger, B., et al. (2013). Integrative eQTL-Based Analyses Reveal the Biology of Breast Cancer Risk Loci. *Cell 152*, 633–641. DOI: 10.1016/j.cell.2012.12.034.

[19] Brynedal, B., Choi, J., Raj, T., et al. (2017). Large-Scale trans-eQTLs Affect Hundreds of Transcripts and Mediate Patterns of Transcriptional Co-regulation. *The Am J Hum Genet 100*, 581–591. DOI: 10.1016/j.ajhg.2017.02.004.

[20] Shabalin, A. A. (2012). Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinforma (Oxford, England) 28*, 1353–8. DOI: 10.1093/bioinformatics/bts163.

[21] Lonsdale, J., Thomas, J., Salvatore, M., et al. (2013). The Genotype-Tissue Expression (GTEx) project. *Nat Genet 45*, 580–585. DOI: 10.1038/ng.2653.

[22] GTEx Consortium. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Sci 348*, 648–660. DOI: 10.1126/science.1262110.

[23] Aguet, F., Brown, A. A., Castel, S. E., et al. (2017). Genetic effects on gene expression across human tissues. *Nat 550*, 204–213. DOI: 10.1038/nature24277.

[24] Aguet, F., Barbeira, A. N., Bonazzola, R., et al. (2019). The GTEx Consortium atlas of genetic regulatory effects across human tissues. *bioRxiv*. DOI: 10.1101/787903.

[25] Lambert, S. A., Jolma, A., Campitelli, L. F., et al. (2018). The Human Transcription Factors. *Cell 172*, 650–665. DOI: 10.1016/j.cell.2018.01.029.

[26] van Arensbergen, J., Pagie, L., FitzPatrick, V. D., et al. (2019). High-throughput identification of human SNPs affecting regulatory element activity. *Nat Genet 51*. DOI: 10.1038/s41588-019-0455-2.

[27] Thurman, R. E., Rynes, E., Humbert, R., et al. (2012). The accessible chromatin landscape of the human genome. *Nat 489*, 75–82. DOI: 10.1038/nature11232.

[28] Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., et al. (2015). Integrative analysis of 111 reference human epigenomes. *Nat 518*, 317–329. DOI: 10.1038/nature14248.

[29] Barbeira, A. N., Bonazzola, R., Gamazon, E. R., et al. (2019). Widespread dose-dependent effects of RNA expression and splicing on complex diseases and traits. *bioRxiv*. DOI: 10.1101/814350.

[30] Guan, Y. and Stephens, M. (2011). Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *Ann Appl Stat 5*, 1780–1815. DOI: 10.1214/11-AOAS455.

[31] Banerjee, S., Lingyao, Z., Heribert, S., et al. (2019). Bayesian multiple logistic regression for case-control GWAS. *PLOS Genet 14*, 1–27. DOI: 10.1371/journal.pgen.1007856.

[32] Stegle, O., Leopold, P., Richard, D., et al. (2010). A Bayesian Framework to Account for Complex Non-Genetic Factors in Gene Expression Levels Greatly Increases Power in eQTL Studies. *PLOS Comput Biol 6*, 1–11. DOI: 10.1371/journal.pcbi.1000770.

Additional Information

Supplementary Text and Figures. Supporting information available for download.