Bioinformatics, 37(3), 2021, 296–302
doi: 10.1093/bioinformatics/btaa715
Advance Access Publication Date: 13 August 2020
Original Paper

Genome analysis

Efficient and effective control of confounding in eQTL mapping studies through joint differential expression and Mendelian randomization analyses

Yue Fan[1,2], Huanhuan Zhu[2], Yanyi Song[2], Qinke Peng[3] and Xiang Zhou[2,4,*]

[1]Key Laboratory of Trace Elements and Endemic Diseases of National Health and Family Planning Commission, School of Public Health, Health Science Center, Xi'an Jiaotong University, Xi'an, Shaanxi 710061, China, [2]Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA, [3]Systems Engineering Institute, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China and [4]Center for Statistical Genetics, University of Michigan, Ann Arbor, MI 48109, USA

*To whom correspondence should be addressed.

Associate Editor: Peter Robinson
Received on February 26, 2020; revised on July 9, 2020; editorial decision on July 30, 2020; accepted on August 6, 2020

Abstract

Motivation: Identifying *cis*-acting genetic variants associated with gene expression levels—an analysis commonly referred to as expression quantitative trait loci (eQTLs) mapping—is an important first step toward understanding the genetic determinant of gene expression variation. Successful eQTL mapping requires effective control of confounding factors. A common method for confounding effects control in eQTL mapping studies is the probabilistic estimation of expression residual (PEER) analysis. PEER analysis extracts PEER factors to serve as surrogates for confounding factors, which is further included in the subsequent eQTL mapping analysis. However, it is computationally challenging to determine the optimal number of PEER factors used for eQTL mapping. In particular, the standard approach to determine the optimal number of PEER factors examines one number at a time and chooses a number that optimizes eQTLs discovery. Unfortunately, this standard approach involves multiple repetitive eQTL mapping procedures that are computationally expensive, restricting its use in large-scale eQTL mapping studies that being collected today.

Results: Here, we present a simple and computationally scalable alternative, Effect size Correlation for COnfounding determination (ECCO), to determine the optimal number of PEER factors used for eQTL mapping studies. Instead of performing repetitive eQTL mapping, ECCO jointly applies differential expression analysis and Mendelian randomization analysis, leading to substantial computational savings. In simulations and real data applications, we show that ECCO identifies a similar number of PEER factors required for eQTL mapping analysis as the standard approach but is two orders of magnitude faster. The computational scalability of ECCO allows for optimized eQTL discovery across 48 GTEx tissues for the first time, yielding an overall 5.89% power gain on the number of eQTL harboring genes (eGenes) discovered as compared to the previous GTEx recommendation that does not attempt to determine tissue-specific optimal number of PEER factors.

Availabilityand implementation: Our method is implemented in the ECCO software, which, along with its GTEx mapping results, is freely available at www.xzlab.org/software.html. All R scripts used in this study are also available at this site.

Contact: xzhousph@umich.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Identifying genetic variants associated with gene expression levels—an analysis commonly referred to as expression quantitative trait loci (eQTLs) mapping—is an important first step toward understanding the genetic determinant of gene expression variation (Cookson *et al.*, 2009). In the past decade, various eQTL mapping studies have identified many eQTLs that are associated with gene expression levels for a large number of genes (Battle *et al.*, 2014; Conesa *et al.*, 2016; Consortium, 2018; Lappalainen *et al.*, 2013; Pickrell *et al.*, 2010; Tung *et al.*, 2015). The identified eQTLs are enriched near transcription start sites (TSSs) and often reside nearby

their associated genes, thus likely influence the gene expression level of their associated genes in a *cis*-acting fashion (Bryois *et al.*, 2014; Parisien *et al.*, 2017). Importantly, the identified eQTLs are often colocalized with SNPs that are associated with common diseases or disease-related complex traits, thus representing an important molecular mechanism underlying SNP-disease associations (Davis *et al.*, 2013; Giambartolomei *et al.*, 2014; Hormozdiari *et al.*, 2016; Nanda *et al.*, 2018; Torres *et al.*, 2014). In addition, the estimated eQTL effect sizes on the gene expression level can be used to construct expression predictors in separate genome-wide association studies (GWAS), facilitating the subsequent transcriptome wide association studies (TWAS) and Mendelian randomization (MR) analyses that can reveal potentially causal genes underlying diseases (Gamazon *et al.*, 2015; Gusev *et al.*, 2016; Yuan *et al.*, 2019; Zeng and Zhou, 2017). Therefore, eQTL mapping can facilitate our understanding of the genetic basis of gene expression variation and reveal causal molecular mechanisms underlying diseases.

Successful eQTL mapping requires effective control of various measured and unmeasured confounding factors. These confounding factors, such as batch labels, environmental exposures as well as intracellular fluctuations, often have relatively large effects on the measured gene expression levels (Gibson, 2008; Stegle *et al.*, 2010). Subsequently, effective confounding effect control can often lead to substantially increased power and/or reduced false positives in eQTL mapping analysis (Leek and Storey, 2007). A standard procedure to control for confounding factors in eQTL mapping studies proceeds by extracting the top probabilistic estimation of expression residual (PEER) factors from the gene expression matrix (Consortium, 2018; Lappalainen *et al.*, 2013; Stegle *et al.*, 2010). These PEER factors are served as surrogate variables for confounding factors and are subsequently either removed from the gene expression levels before eQTL mapping or directly included as covariates in eQTL mapping (Liang *et al.*, 2013). In the standard procedure for confounding factor control, an important analytic step is to determine the optimal number of PEER factors, k, to be included in the eQTL mapping analysis (Parts *et al.*, 2011; Stegle *et al.*, 2010). A small k may be insufficient to capture all confounding effects while a large k may either introduce unnecessary noise or incorrectly include true genetic effects underlying expression. Consequently, failing to identify the optimal k can lead to a potential loss of eQTL mapping power. The standard procedure for confounding factor control determines the optimal k by examining one k at a time and choosing a k that maximizes the number of eQTLs or eQTL harboring genes (eGenes) discovered in the study (Degner *et al.*, 2012; Raj *et al.*, 2014; Tung *et al.*, 2015). However, examining one k at a time is computationally costly and such approach may not be applicable to even a moderate-sized eQTL mapping study with a few hundred samples. For example, due to the heavy computational cost, the Genotype-Tissue Expression (GTEx) study does not attempt to determine an optimal k for every tissue using the standard procedure; instead, a common k is used for a group of tissues with similar number of samples (Consortium, 2018). As we will show below, lack of a data-specific optimal k can lead to an appreciable power loss for eQTL mapping.

Here, we present a simple and computationally efficient alternative for determining the optimal number of PEER factors, k, for eQTL mapping studies. Our method requires the availability of an outcome phenotype in addition to the usual genotype and expression data required for eQTL mapping studies. With the outcome phenotype, we estimate the gene expression effect on the phenotype for one gene at a time through two different analyses: a differential expression regression analysis and a MR analysis. By computing and examining the correlation between the estimated effect sizes from the two different analyses, we can subsequently determine the optimal k for eQTL mapping. We refer to our method as Effect size Correlation for COnfounding determination (ECCO). In both simulations and an in-depth analysis of 48 GTEx tissues, ECCO determines a similar k as the previous standard approach but is two orders of magnitude faster. ECCO is freely available at www.xzlab.org/software.html.

2 Materials and methods

2.1 Method details

We focus on an eQTL mapping study performed on n samples and m genes. Our goal is to determine the number of PEER factors optimal for eQTL mapping. Our method requires that the eQTL study to also contain a phenotype in addition to the usual gene expression matrix and genotype matrix. Such phenotype should have a genetic determinant and should be associated with the expression level of at least a subset of genes. Exemplary phenotypes may include height, BMI, fasting glucose, blood pressure and so on (Porcu *et al.*, 2019). This phenotype is used to facilitate eQTL discovery but is not used for the discovery of phenotype-specific eQTLs. We denote $\boldsymbol{y}$ as an n-vector of such phenotypic measurements. For ease of presentation, we assume that the phenotype is quantitative. However, extension to a binary phenotype is straightforward, requiring only replacing the linear regression models with logistic regression models. With the available phenotype, we first perform a differentially expression analysis across all genes to estimate the effect size of each gene on the phenotype. Specifically, we examine one gene at a time and denote $\boldsymbol{x}_i$ as the n-vector of gene expression level for ith gene. We remove the effects from the first j PEER factors on $\boldsymbol{x}_i$ by fitting a regression model. In the regression model, $\boldsymbol{x}_i$ is treated as the outcome variable and the first j PEER factors are treated as covariates. Through the regression model, we obtained $\hat{\boldsymbol{x}}_{ij}$ as the gene expression residuals, which is now free of confounding effects captured by the first j PEER factors. Afterward, we fit the following linear regression to estimate the effect size of $\hat{\boldsymbol{x}}_{ij}$ on $\boldsymbol{y}$,

$$\boldsymbol{y} = 1_n\mu_{yj} + \hat{\boldsymbol{x}}_{ij}\beta_{ij} + \boldsymbol{\epsilon}_{ij}, \tag{1}$$

, where 1_n is an n-vector of 1 s; μ_{yj} is the intercept; β_{ij} is the effect size of gene expression on phenotype and is the focus of this study; $\boldsymbol{\epsilon}_i$ is an n-vector of residual errors that each is assumed to be independently and identically distributed from a normal distribution. The estimated effect size $\hat{\beta}_{ij}$ will change depending on j, which again is the number of PEER factors controlled for. With the ideal number of PEER factors, the estimated effect size $\hat{\beta}_{ij}$ would be close to the true gene effect size on the phenotype.

Next, in parallel, for each gene in turn, we extract its *cis*-SNPs that reside within 1 Mb of the TSS. Among the *cis*-SNPs, we select one that has the strongest association evidence with $\boldsymbol{x}_i$. We denote the n-vector genotype of such SNP as $\boldsymbol{g}_i$. We treat this selected SNP as the instrumental variable. To minimize estimation bias, we follow standard MR (Burgess *et al.*, 2011) and only retain the instrument if the F-statistics that measures the association evidence between $\boldsymbol{g}_i$ and $\boldsymbol{x}_i$ is above a threshold of 10. For genes that have a selected instrument, we consider the following MR model to estimate the effect size of the gene on the phenotype

$$\boldsymbol{x_i} = 1_n\mu_x + \boldsymbol{g}_i\alpha_i + \boldsymbol{\epsilon}_{xi}, \tag{2}$$

$$\boldsymbol{y} = 1_n\mu_y + \boldsymbol{g}_i\gamma_i + \boldsymbol{\epsilon}_{yi}, \tag{3}$$

where μ_x and μ_y are the intercepts; α_i is the SNP effect on gene expression; γ_i is the SNP effect on phenotype; $\boldsymbol{\epsilon}_{xi}$ and $\boldsymbol{\epsilon}_{yi}$ are both n-vectors of residual errors, each assumed to be independently and identically distributed from a normal distribution. The effect size of gene expression on phenotype can be estimated through the above two equations using the ratio method: $\sim\beta_i = \hat{\gamma}_i/\hat{\alpha}_i$ (Wald, 1940). Because MR is resilient to confounding factors, the estimated gene expression effect on phenotype $\sim\beta_i$ would be close to the true gene effect on the phenotype.

$\hat{\beta}_{ij}$ obtained in Equation (1) and $\sim\beta_i$ obtained in Equations (2) and (3) are both estimates for the true underlying gene expression effect on the phenotype. One would intuitively expect the two estimates to be correlated with each other across genes, and more so when the number of PEER factors included, j, is closer to the optimal number k. Therefore, for each j in turn, we can compute the correlation between $\hat{\beta}_{ij}$ and $\sim\beta_i$, and select a j that maximizes such correlation to serve as the optimal number of PEER factors needed for eQTL mapping. We can also visualize the selection procedure by

plotting the correlation values versus the number of PEER factors included.

While the above intuition is straightforward, we note that technically there are certain mathematical conditions we need to satisfy to ensure the validity of the above procedure. Specifically, for the MR model in Equations (2) and (3), the selected SNP for each gene needs to satisfy three standard MR conditions: (i) relevance condition, that the selected SNP is associated with the gene exposure $\boldsymbol{x_i}$; (ii) independence condition, that the selected SNP is independent of unmeasured confounders that affect both $\boldsymbol{x_i}$ and $\boldsymbol{y}$; (iii) exclusion restriction, that the select SNP only affects the outcome $\boldsymbol{y}$ through the exposure $\boldsymbol{x_i}$. For the relevance condition, we use *F*-statistics to measure the strong association strength between $\boldsymbol{g}_i$ and $\boldsymbol{x}_i$, and only use genes with the *F*-statistics above the usual threshold of 10 for analysis following (Bound *et al.*, 1995; Staiger and Stock, 1997). In addition, to further ensure the strong association between $\boldsymbol{g}_i$ and $\boldsymbol{x}_i$, we rank genes based on the *P*-values for testing α_i and use the top 1000 genes for computing correlations. For the independence condition, we note that SNP genotypes are generally measured accurately and are often measured in a separate genotyping study different from the gene expression study. Subsequently, SNP genotypes are unlikely affected by the same confounding factors such as batch effects in the gene expression study. For the exclusion restriction condition, we acknowledge that, as in any MR analysis, it is indeed impossible to validate such condition. When the exclusion restriction condition is not satisfied, the selected SNPs can affect the outcome $\boldsymbol{y}$ directly through pathways other than $\boldsymbol{x}_i$. Subsequently, the estimated gene effect $\sim\beta_i$ would deviate from the true gene effect β_i: $\sim\beta_i = \beta_i + \delta_i$, with the difference δ_i being related to the part contributed by the direct effect of SNP on the phenotype not mediated through the gene. However, as long as δ_i is not correlated with β_i across genes, then the violation of the exclusion restriction condition would not influence the validity of our method. Therefore, instead of the general exclusion restriction condition, our method effectively only requires the InSIDE assumption (Bowden *et al.*, 2015) to hold.

We refer to our method as Effect size Correlation for COnfounding determination (ECCO). While we have primarily focused on describing ECCO based on the MR method with a single instrumental variable, we note that ECCO can be paired with any other MR approaches. For example, ECCO can be paired with the Inverse-Variance Weighted (IVW) approach of MR (Jack Bowden *et al.*, 2015) to take advantage of multiple instrumental variables. ECCO can also be paired with MR-Egger regression (Bowden *et al.*, 2015) to control for potential horizontal pleiotropy (Yuan *et al.*, 2019). We examine these two additional variations of ECCO, ECCO-IVW and ECCO-Egger, in both simulations and real-data applications.

Finally, we note that our MR-based procedure to infer the optimal number k is computationally efficient, much more so than the previous standard procedure of choosing k by maximizing eQTL discovery. In particular, the previous standard procedure requires performing eQTL mapping across all r SNP–Gene pairs n different times to examine d different choices of PEER factors, all with l different permutations. Subsequently, the previous procedure has a computing time complexity of $O(nrdl)$). In contrast, our method only needs to perform eQTL mapping one time but fits a simple linear regression in Equation (1) d different times. Subsequently, our method has a computing time complexity of $O(nr + nd)$, approximately dl times faster than the previous standard procedure. Common eQTL studies use an l that equals 10 to 20 and a d that equals a few dozen. Therefore, we expect our method to be at least two orders of magnitude faster than the standard procedure.

2.2 Standard approach to determine *k*

The standard approach to determine the optimal number of PEER factors included for eQTL mapping analysis is through examining a range of PEER factors and identifying the number of PEER factors that maximize the eGenes discovery. Here, eGene represents eQTL harboring gene. To apply the standard approach, for each gene in turn, we identified the most significantly associated *cis*-SNP for the gene as the candidate eQTL. We treated the *P*-value from the candidate eQTL as the gene-level *P*-value. We permuted individual-label $l = 10$ times and applied the same procedure to obtain an empirical null distribution of the gene-level *P*-values. With the empirical null distribution, we calculate the number of eGenes detected based on 10% empirical false discovery rate (FDR). We repeat the above procedure across varied number of PEER factors and determine the optimal number of PEER factors as the one that maximized the number of discovered eGenes. Note that the empirical null *P*-values from permutation are computed based on the same set of covariates and PEER factors as used in the corresponding analysis. Therefore, there is no artificial selection of confounders to boost power. Instead, all power comparisons are carried out in a fair way based on the same empirical FDR. We used the computationally efficient MatrixeQTL package (Shabalin, 2012) to perform all these mapping analyses.

3 Results

Details of ECCO are provided in the Section 2. Briefly, ECCO aims to select the optimal number of expression PEER factors to control for in an eQTL mapping analysis. To do so, besides the usual SNP and expression data, we also require the presence of an additional phenotype. This additional phenotype should have a genetic determinant and is associated with the expression of at least a subset of genes. Some common exemplary phenotypes include height, BMI, fasting glucose, blood pressure, etc. This phenotype is used to facilitate eQTL discovery but is not used for the discovery of phenotype-specific eQTLs. With the phenotype, ECCO proceeds by fitting two regression models for one gene at a time to estimate the gene effect on the phenotype. In particular, for the ith gene, we first select a *cis*-SNP as the instrument for the expression level and perform a MR analysis to estimate the effect size of gene expression on the phenotype, $\sim\beta_{\mathrm{i}}$. For the same gene, we also perform a standard differential expression analysis to estimate the effect size of gene expression on the phenotype, $\hat{\beta}_{\mathrm{ij}}$, where the top j PEER factors are removed from the gene expression data (hence $\hat{\beta}_{\mathrm{ij}}$ depends on j). The two effect estimates from the two regression models, $\sim\beta_{\mathrm{i}}$ and $\hat{\beta}_{\mathrm{ij}}$, are both estimates for the true underlying gene expression effect on the phenotype. Subsequently, one would expect that the two estimates are correlated with each other across genes, more so when the correct number of PEER factors, j, is chosen. Therefore, for each j in turn, ECCO computes the correlation between $\sim\beta_{\mathrm{i}}$ and $\hat{\beta}_{\mathrm{ij}}$, and further selects a j that maximizes such correlation. The selected j is served as the optimal number of PEER factors to control for in an eQTL mapping analysis. In the process, we can also visualize the selection procedure by plotting the correlation value versus j across a range of j's examined.

3.1 Simulations

We performed simulations to validate the intuition underlying ECCO and examine its effectiveness. The simulation details are provided in the Supplementary Material. Briefly, we randomly selected 10 000 genes from 491 samples in GTEx (Consortium, 2018). We extracted *cis*-SNPs for each gene (median = 4818 *cis*-SNPs per gene) and randomly selected either one SNP or five SNPs among them to serve as the eQTLs. We also simulated ten confounding factors. Based on the eQTL genotype and confounding factors, we simulated the expression level for 10 000 genes. The expression level of each gene is contributed by the eQTL and the confounding factors. In particular, the eQTLs contribute to either 3% (one causal SNP case) or 10% (five causal SNP case) of the gene expression variation. For the confounding factor contribution, we explored two scenarios: a heterogeneous confounding scenario where each gene was affected by a randomly selected five confounding factors and a homogeneous confounding scenario where each gene was affected by all ten confounding factors. For each of those genes that are influenced by the confounding effects, the confounding factors in total contribute to 50% of expression variation, consistent with real-data estimates (Supplementary Fig. S1). The phenotype is contributed by either 1% of the genes (=100; sparse scenario) or all genes (polygenic

scenario), with the contributing genes explaining 25% of phenotypic variance. In the multiple causal SNP scenario, we also examined cases where the *cis*-SNP of a certain proportion of genes (set to be either 0, 1%, 10% or 100%) exhibit pleiotropic effects which equal to 10% of phenotypic variance. We then extracted the PEER factors from the gene expression data. As expected, the top 10 extracted PEER factors capture the majority of confounding effects (Supplementary Fig. S2). We then applied ECCO and compared it to the standard approach to determine the number of PEER factors needed. In total, we considered four different scenarios (heterogeneous versus homogeneous confounding; sparse versus polygenic gene effects). For each scenario, we performed 10 simulation replicates, with each consisting of 10 000 genes. We applied ECCO to all simulation replicates.

As a comparison, we also carried out the standard approach to determine the optimal number of PEER factors on the first simulation replicate in each scenario; we did not apply the standard approach to the other replicates due to its heavy computational burden. The standard approach explores different numbers of PEER factors in eQTL mapping and determines the optimal number based on the number of eGenes discovered. In the simulations, as one would expect, the standard approach correctly identified 10 as the optimal number of PEER factors across a range of simulation scenarios (Fig. 1A). For example, in scenario I, the number of eGenes is 4731 with zero PEER factor, increases to 9530 with 10 PEER factors, and becomes slightly reduced and stabilized afterward. While being effective, unfortunately, the standard approach for determining the optimal number of PEER factors requires a full-scale eQTL mapping analysis for each number of PEER factors considered and is thus computational expensive. Indeed, it took the standard method an average of 96.27 h to determine the correct number of PEER factors across simulation scenarios.

Next, we applied ECCO to determine the optimal number of PEER factors required. To do so, we estimated gene effect size on phenotype using either the MR analysis or the standard linear regression analysis. Afterward, we calculated the estimated effect size correlation between these two approaches across genes. Because the validity of MR analysis requires the presence of strong instruments, we sorted genes based on their instrument strength and calculated the effect size correlation using three different gene sets: either the top 1000 genes, top 5000 genes or all genes. Afterward, we plotted the calculated correlation values of effect size estimates against the number of PEER factors included (Fig. 1B; Supplementary Fig. S3). As expected, regardless of the number of genes included for computing the correlation, the effect size correlation is always the highest when the top 10 PEER factors are included. For example, in scenario II, the effect size correlation is 0.268 with zero PEER factor, increases to 0.305 with 10 PEER factors, and becomes stabilized afterward. The dependence of the effect size correlation on the number of PEER factors is general and holds in the polygenic phenotype setting where all genes have non-zero effects on the phenotype (scenario II and IV; Fig. 1B), in the sparse phenotype setting where only 1% of genes have non-zero effects on the phenotype (scenario I and III; Fig. 1B), in the heterogeneous confounding setting where all genes are affected by 5 of the 10 confounding factors (scenario III and IV; Fig. 1B), as well as the homogeneous confounding setting where all genes are affected by all 10 confounding factors (scenario I and II; Fig. 1B). With ECCO, the estimated number of optimal PEER factors is 10 across all four scenarios, with a relatively larger estimation variance in the two sparse settings than in the two polygenic settings (Fig. 1C). Besides the sparse and polygenic settings, we also explored the setting where none of the genes are associated with the phenotype. Because ECCO requires the phenotype to be associated with at least a few genes, ECCO fail to identify the optimal number of PEER factors in this setting as one would expect (Supplementary Fig. S4). Importantly, unlike the standard approach in the previous paragraph, ECCO is computationally efficient: ECCO took an average of 0.618 h to identify the optimal number of PEER factors across simulation scenarios, resulting in 156 times speed gain over the standard approach. The total computing time saving brought up by ECCO becomes more appreciable with increased number of SNP–gene pairs (Fig. 1D).

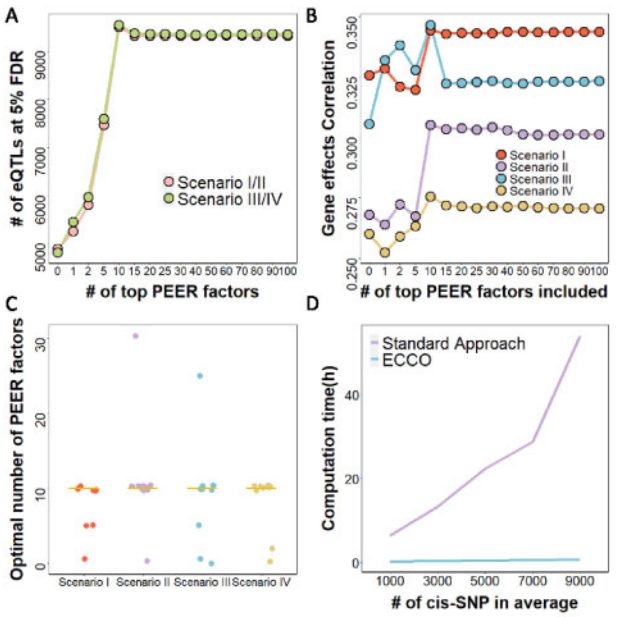

Fig. 1. ECCO identified a similar number of PEER factors as the standard approach in simulations. (**A**) In the standard approach, the number of eQTLs identified (*y*-axis) is plotted against the number of PEER factors removed in eQTL mapping analysis (*x*-axis). Results are based on 5% FDR and are shown for the homogeneous confounding simulation scenario (scenarios I–II; pink) and the heterogeneous confounding simulation scenario (scenarios III–IV; green). Removing 10 PEER factors results in the highest number of eQTLs detected. Note that the eQTL mapping results are almost identical between scenario I and II as well as between scenario III and IV. (**B**) In ECCO, correlation between the two types of effect size estimates across genes (*y*-axis) is plotted against the number of PEER factors controlled. Results are shown in the four scenarios (scenario I: red; scenario II: purple; scenario III: blue; scenario IV: brown) and based on correlation computed using the top 1000 genes with the strongest instrumental strength. ECCO achieves highest correlation when 10 PEER factors are included. (**C**) Estimated number of PEER factors by ECCO across 10 simulation replicates in each of the four simulation scenarios. The medium estimate is 10 across all scenarios, with larger estimation variance in the sparse settings (I and III) than in the polygenic scenarios (II and IV). (**D**) Computation time in hours (*y*-axis) for the two methods is plotted against the average number of *cis*-SNPs per gene. Specifically, we examined the computing time of ECCO and the standard approach with respect to the number of *cis*-SNPs per gene. Because different genes have a different number of *cis*-SNPs, we binned genes with similar number of *cis*-SNPs together. We then computed the average computing time (*y*-axis) and the average number of *cis*-SNPs (*x*-axis) in each bin and plotted them against each other. Time is recorded on 10 000 genes and on a single core of an Intel Xeon E5-2683 2.00 GHz processor

Finally, we examined the effectiveness of ECCO in the scenarios where multiple independent SNPs have effects on gene expression. Here, besides applying the standard ECCO where we select one SNP to serve as the instrumental variable, we also estimated other ECCO variations where we selected multiple independent SNPs to serve as instrumental variables and use either IVW approach or the MR-Egger regression to combine association evidence across multiple instruments. We term these two ECCO variations as ECCO-IVW and ECCO-Egger, respectively. Here, we again sorted genes based on their instrument strength and calculated the effect size correlation using the top 1000 genes. In the simulations, we found that the estimated number of optimal PEER factors from all three ECCO variants is centered around the truth across almost all scenarios, either in the absence of horizontal pleiotropic effects (Supplementary Fig. S5) or in the presence of horizontal pleiotropic effects in 1% (Supplementary Fig. S6), 10% (Supplementary Fig. S7) or 100% (Supplementary Fig. S8) of genes. Consistent with the early simulations, the estimated number of optimal PEER factors from all three

ECCO variants has a relatively large estimation variance in the two sparse settings as compared with the polygenic settings (Supplementary Figs S5–S8). Among these ECCO variations, ECCO-Egger often yields a larger estimation variance as compared to ECCO and ECCO-IVW. In addition, ECCO-Egger tends to underestimate the number of PEER factors in the sparse settings and does not show a clear advantage in the presence of horizontal pleiotropic effect presumably due to its power limitation. On the other hand, results from ECCO and ECCO-IVW are highly consistent with each other, with no noticeable differences between the two. Therefore, we recommend the use of either ECCO or ECCO-IVW to estimate the number of PEER factors.

3.2 Real-data application

We applied ECCO to perform eQTL mapping in GTEx (Consortium, 2018). We analyzed all 48 tissues and focused on five of them for in-depth comparison: three of the five tissues have the largest sample sizes in GTEx and the remaining two tissues have relatively small sample sizes. The five tissues are muscle skeletal ($n = 491$), thyroid ($n = 399$), artery tibial ($n = 388$), liver ($n = 153$) and artery coronary ($n = 152$). The number of PEER factors used in the original GTEx eQTL mapping study is 35 for tissues with ≥ 250 samples, 30 for tissues with ≥ 150 samples and < 250 samples and 15 for tissues with < 150 samples, respectively (Consortium, 2018). Given the heavy computational cost of eQTL mapping analysis in GTEx and the relatively small number of PEER recommended in (Consortium, 2018), we mostly focused on using up to 100 PEER factors (though we explored up to 250 PEER factors for the tissues that the optimal numbers of PEER factors were inferred to be larger than 100). Data and processing details are available in Supplementary Material.

We first applied the standard method to determine the number of optimal PEER factors in the five tissues. Here, we focused a total of 22 017 genes and 101 814 843 SNP–gene pairs, with a median number of 4571 *cis*-SNPs per gene across the five tissues. For all tissues, we found that the number of eGenes gradually increases with the increasing number of PEER factors included in the model and gradually decreases after reaching a plateau (Fig. 2A). The optimal number of PEER factors determined by the standard approach is 250 for muscle skeletal, 150 for thyroid, 200 for artery tibial, 70 for liver and 90 for artery coronary. Importantly, the peak number of eGenes detected with the optimal number of PEER factors detected by the standard method is higher than that with the GTEx recommended number of PEER factors included (35 for the first three tissues and 30 for the last two tissues). Specifically, the number of eGenes detected by the standard method is 9089 (muscle skeletal), 11 853 (thyroid), 9 796 (artery tibial), 3176 (liver) and 3454 (artery coronary). The number of eGenes detected by the GTEx recommended PEER factors is generally lower: 7755 (muscle skeletal), 10 557 (thyroid), 8712 (artery tibial), 3054 (liver) and 3367 (artery coronary). The increased number of eGenes discovered using the standard approach highlights the importance of identifying a tissue-specific optimal number of PEER factors for eQTL mapping.

Next, we applied ECCO to examine the optimal number of PEER factors in the five tissues. To do so, we obtained three quantitative phenotypes in GTEx. These three phenotypes include height, weight and body mass index (BMI). Regardless of the phenotypes we use, we found that the effect size correlation gradually increases with increasingly large number of PEER factors included and gradually decreases after reaching a plateau (Fig. 2B, Supplementary Figs S9 and S10). For example, for muscle skeletal tissue, with BMI as the phenotype, the effect size correlation is 0.291 with zero PEER factors included. The effect size correlation increases to 0.437 with 200 PEER factors included. The optimal number of PEER factors determined by ECCO is 200 for muscle skeletal, 150 for thyroid, 150 for artery tibial, 40 for liver and 40 for artery coronary, all close to that determined by the standard approach. The results based on height and weight are consistent (Supplementary Figs S9 and S10). We also applied ECCO-IVW and ECCO-Egger to analyze the muscle skeletal tissue. Consistent with simulations, we found that the optimal number of PEER factors identified by ECCO-IVW is similar with that form ECCO, while ECCO-Egger underestimates the optimal number of

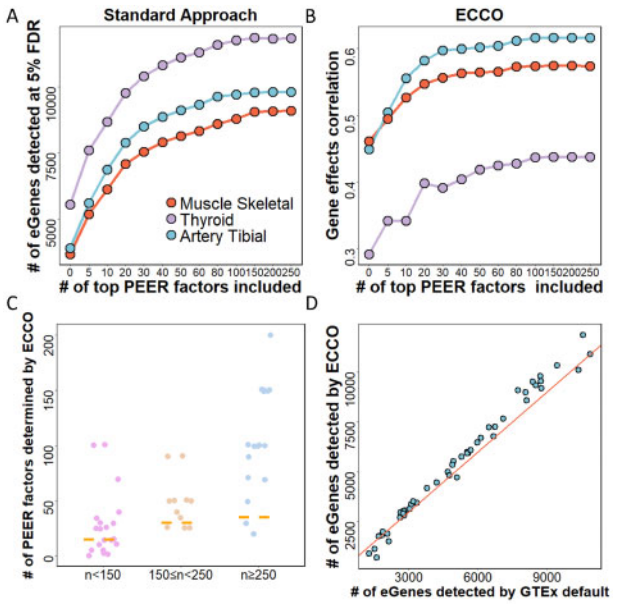

Fig. 2. ECCO identified a larger number of PEER factors as recommended in GTEx. (**A**) In the standard approach, the number of eGenes identified (*y*-axis) is plotted against the number of PEER factors removed in eQTL mapping analysis (*x*-axis) for three GTEx tissues: muscle skeletal (250), thyroid (150) and artery tibial (200). (**B**) In ECCO, correlation between the two types of effect size estimates across genes (*y*-axis) is plotted against the number of PEER factors controlled (*x*-axis) for three GTEx tissues: muscle skeletal (200), thyroid (150) and artery tibial (150). Results are based on correlation computed using the top 1000 genes with the strongest instrumental strength (i.e. *F*-statistics > 10) and are based on using the phenotype BMI. (**C**) Jittered point plot shows the estimated number of PEER factors by ECCO (*y*-axis) for tissues recommended using 15, 30 or 35 PEER factors in GTEx. The recommend number of PEER factors in GTEx is based on sample size (*x*-axis): 15 when the sample size $n < 150$ (20 tissues); 30 when the sample size $150 \leq n < 250$ (11 tissues); and 15 when the sample size $n \geq 250$ (17 tissues). The orange dashed line represents the recommended number of PEER factors in each bin. (**D**) Scatter plot shows the number of eGenes detected when one controls for the number of PEER factors determined by ECCO (*y*-axis) versus the number of eGenes detected when one controls for the number of PEER factors recommended in GTEx based on sample size (*x*-axis)

PEER factors (Supplementary Fig. S11). Importantly, in addition to achieving similar results as the standard approach, ECCO is 222 times faster (Table 1, Supplementary Fig. S12).

Finally, we applied our method to analyze the remaining 43 tissues in GTEx data. We did not apply the standard approach to infer the number of PEER factors in these tissues due to its heavy computational burden. The inferred optimal number of PEER factors by ECCO is shown in Supplementary Table S2, which are generally higher than that recommended by GTEx (Fig. 2C, Supplementary Fig. S13). Note that the optimal number of PEER factors is inferred to be 100 for two out of the twenty tissues that have less than 150 samples. Using a large number of PEER factors for eQTL mapping is often safe even for data with a small sample size: the lower-order PEER factors become increasingly sparse with a large fraction of zeros and low variance (Supplementary Fig. S14). Consequently, lower-order PEER factors tend to contain less information as compared to the higher-order PEER factors and using 100 PEER factors would not cause a substantial loss of degrees of freedom as would be expected by, for example, using 100 principal components. In addition, as expected, the effects of the selected instruments on gene expression are often strong: the median F-statistics of the selected SNP instrument for the top 1000 genes is 61.7 (mean = 79.7; min = 20.4; max = 949.0). Therefore, we would not expect weak instrument bias commonly observed in MR analysis when the F-statistics to measure instrumental strength is below 10 (Zeng *et al.*, 2019).

Table 1. Computation time of ECCO and the standard approach for determining the optimal number of PEER factors used in eQTL mapping studies

Data	Computation time (h)			
	ECCO	Standard approach	No. of individuals	No. of SNP–Gene pairs
Simulation	0.618	96.27	491	46 212 053
GTEx	0.92	204.16	491	91 610 671

Note: The average computing time (in hours) in the first simulation replicate across four scenarios is recoded for both approaches (top row). Computing time (in hours) is also recorded for eQTL mapping in the muscle skeletal tissue in GTEx (bottom row). Computing time is based on a single thread of a Xeon E5-2683 2.00 GHz processor.

Importantly, the overall number of eGenes detected with the number of PEER factors inferred by ECCO is 5.89% higher than that GTEx recommended numbers across all 48 tissues. The power gain brought by ECCO is higher for tissues with large sample sizes: the power gain by ECCO increases to 7.14% when we focus on the tissues with sample sizes > 150, and to 8.82% when we focus on the tissues with sample sizes > 250. In addition, we examined whether the eGenes detected only by ECCO are functionally important. To do so, we extracted gene association results from transcriptome-wide association studies (TWAS) and calculated the proportion of eGenes that are also a TWAS gene. Consistent with the higher power of ECCO, we found that the eGenes detected only by ECCO are more likely to be TWAS genes as compared to the eGenes detected by the standard approach (Supplementary Table S3). Among the 48 tissues we examined, we found that the eGenes detected by ECCO are more likely to be TWAS genes as compared to the eGenes detected by the standard approach in 46 tissues. The eQTL mapping results by ECCO highlight the importance of identifying the optimal number of PEER factors for eQTL mapping.

4 Discussion

We have presented ECCO, a new and simple approach for determining the optimal number of PEER factors used for confounding effects control in *cis*-eQTL mapping studies. ECCO estimates the gene expression effects on an outcome phenotype using two computationally efficient models: a standard differential expression analysis model and a MR analysis model. By examining the gene effect size correlation obtained from the two models, ECCO can be used to determine the optimal number of PEER factors used in eQTL mapping studies. In simulations and an in-depth analysis of 48 tissues in GTEx, ECCO determines a similar number of PEER factors for eQTL mapping as the previous standard approach, while being two orders of magnitude faster. Therefore, ECCO represents an efficient and effective alternative for controlling for confounding effects in eQTL mapping studies.

ECCO requires the availability of an outcome phenotype in addition to the usual genotype and expression data required for eQTL mapping studies. Because of this requirement, ECCO cannot be directly applied to eQTL studies that collect gene expression and genotypes only. Fortunately, most eQTL mapping studies do collect additional phenotype data. For example, among the seven relatively large eQTL mapping studies carried out previously [ABRP (Tung *et al.*, 2015), GEUVADIS (Lappalainen *et al.*, 2013), TCGA (Abeshouse *et al.*, 2015), METSIM (Stančáková *et al.*, 2012), DGN (Alexis Battle *et al.*, 2014), NTR (Wright *et al.*, 2014) and YFS (Raitakari *et al.*, 2008)], six of them have accompanying phenotype measurements and only one of them (GEUVADIS) does not. Therefore, ECCO can be applied to the majority of eQTL mapping studies. Importantly, as we have shown in the real-data application, the outcome phenotype does not have to be relevant to the expression tissue. For example, the number of eGenes detected in the Skin not sun exposed suprapubic tissue by the three different phenotypes are 7303, 8971 and 8971, respectively, with 7070 eGenes in common. After all, such outcome phenotype is used to facilitate eQTL discovery but is not used for the discovery of phenotype-specific eQTLs. Indeed, the only requirement for the outcome phenotype is that the phenotype should be associated with gene expression for at least some genes. In the simulations, we found that the optimal number of PEER factors inferred using different phenotypes are all centered around the truth. However, different phenotypes do influence estimation variance: in the sparse setting where 1% of genes are associated with the phenotype, the variance of the estimated number of PEER factors k across simulation replicates can be relatively large. In contrast, in the polygenic setting where all genes are associated with the phenotype, the variance of the estimated number of PEER factors k across simulation replicates is relatively small. Certainly, when no gene is associated with the phenotype, then ECCO would not work (Supplementary Fig. S4).

We have primarily focused on *cis*-eQTL mapping analysis and have not explored the utility of ECCO for *trans*-eQTL mapping. *Trans*-eQTL mapping aims to identify SNPs associated with genes that far away from the SNP or genes that reside on a different chromosome. The identified SNPs from *trans*-eQTL mapping may potentially influence the expression level of the targeted gene in a *trans*-fashion. While *cis*-eQTL mapping has been the primary task in most eQTL mapping studies, *trans*-eQTL mapping has become increasingly common, thanks to recent experimental and analytical advances. Experimentally, the sample size of eQTL mapping studies has been steadily increasing in the past years, leading to substantially improved power of *trans*-eQTL mapping. Analytically, the development of new approaches has mitigated many RNA-sequencing alignment errors, leading to substantially reduced false positives in *trans*-eQTL mapping (Liu *et al.*, 2018; Saha and Battle, 2018). The increased sample size and mitigation of sequencing alignment errors have altogether made *trans*-eQTL mapping feasible and reasonably effective (Aguet *et al.*, 2019). While practical *trans*-eQTL mapping often relies on the same procedure as in *cis*-eQTL mapping to deal with confounding effects (Aguet *et al.*, 2019), controlling for confounding effects in *trans*-eQTL mapping may face additional challenges. In particular, some of the extracted PEER factors from the gene expression data may represent the true genetic effects underlying the expression level of multiple genes (Consortium, 2018). Subsequently, controlling for these PEER factors may unintentionally reduce the power of *trans*-eQTL mapping (Rakitsch and Stegle, 2016). Therefore, investigating the utility of ECCO and exploring its extensions for confounding effects control in *trans*-eQTL mapping is an important future direction.

We have primarily focused on estimating the number of PEER factors. While PEER analysis is the most common approach for confounding effects control in eQTL mapping studies, many other confounding effects control approaches exist. For example, principal component analysis (PCA) extracts principal components to serve as surrogates for confounding factors (Degner *et al.*, 2012; Tung *et al.*, 2015). SVA (Leek and Storey, 2007) extracts sparse non-orthogonal components in the presence of covariates to serve as surrogate variables. RUV uses a set of housekeeping genes to serve as negative controls for effective extraction of confounding factors (Gagnon-Bartsch and Speed, 2012). scPLS further relies on the partial least squares to model both control genes and target genes jointly to effectively extract confounding factors(Chen and Zhou, 2017). While some of these methods have automatic ways for determining the number of confounding factors (e.g. SVA), many methods do not. Exploring the benefits of paring ECCO with these confounding factor analysis methods may have added benefits. For example, we have performed simulations that using PCA to control for confounding factors. In the simulations, we found that ECCO can also be used to identify the optimal number of PCs needed (Supplementary Figs S15 and S16).

ECCO is currently formulated as a two-step procedure that includes fitting two different regression models and subsequently examining the estimated effect size correlation. While intuitively

appealing and practically effective, the two-step procedure in ECCO does appear to be *ad hoc* in nature. It would be ideal in the future to both extend ECCO to a formal statistical model with an underlying data-generative process and place ECCO inference into a likelihood-based inference framework. For example, we could treat the gene effect sizes on the phenotype in the linear regression model as a summation of the gene effect sizes on the phenotype in the MR model and an additional noise term. The noise term effectively determines the correlation ρ between the two sets of effect sizes. We can then treat the number of included PEER factors k as another latent parameter and aim to optimize a target function that consists of both k and ρ. Certainly, despite the simple description, formalizing the above model and developing the corresponding algorithm remains a non-trivial task. Nevertheless, a model-based treatment of ECCO would help us further understand its pros and cons from a theoretical perspective.

Acknowledgements

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the NIH, and by the National Cancer Institute (NCI), the National Human Genome Research Institute (NHGRI), the National Heart, Lung, and Blood Institute (NHLBI), the National Institute on Drug Abuse (NIDA), the National Institute of Mental Health (NIMH), and the National Institute of Neurological Disorders and Stroke (NINDS). The GTEx v7 data used for the analyses described in the manuscript were obtained from dbGaP (phs000424) and the GTEx Portal.

Funding

This study was supported by the National Institutes of Health (NIH) [R01HG009124 and R01GM126553] and National Science Foundation (NSF) [DMS1712933]. Y.F. is also supported by a scholarship from the China Scholarship Council.

Conflict of Interest: We have no competing interests.

References

Abeshouse,A. *et al.* (2015) The molecular taxonomy of primary prostate cancer. *Cell*, **163**, 1011–1025.

Aguet,F. *et al.* (2019) The GTEx Consortium atlas of genetic regulatory effects across human tissues. bioRxiv, 787903.

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

Bound,J. *et al.* (1995) Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *J. Am. Stat. Assoc.*, **90**, 443–450.

Bowden,J. *et al.* (2015) Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.*, **44**, 512–525.

Bryois,J. *et al.* (2014) Cis and trans effects of human genomic variants on gene expression. *PLos Genet.*, **10**, e1004461.

Burgess,S. *et al.*; CRP CHD Genetics Collaboration. (2011) Avoiding bias from weak instruments in Mendelian randomization studies. *Int. J. Epidemiol.*, **40**, 755–764.

Chen,M.J. and Zhou,X. (2017) Controlling for confounding effects in single Cell RNA sequencing studies using both control and target genes. *Sci. Rep.*, **7, 1-14**.

Conesa,A. *et al.* (2016) A survey of best practices for RNA-seq data analysis. *Genome Biol.*, **17, 13**.

Consortium,G. (2018) Genetic effects on gene expression across human tissues. *Nature*, **553**, 530–530.

Cookson,W. *et al.* (2009) Mapping complex disease traits with global gene expression. *Nat. Rev. Genet.*, **10**, 184–194.

Davis,L.K. *et al.* (2013) Partitioning the heritability of tourette syndrome and obsessive compulsive disorder reveals differences in genetic architecture. *PLoS Genet.*, **9**, e1003864.

Degner,J.F. *et al.* (2012) DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature*, **482**, 390–394.

Gagnon-Bartsch,J.A. and Speed,T.P. (2012) Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, **13**, 539–552.

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

Giambartolomei,C. *et al.* (2014) Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.*, **10**, e1004383.

Gibson,G. (2008) The environmental contribution to gene expression profiles. *Nat. Rev. Genet.*, **9**, 575–581.

Gusev,A. *et al.* (2016) Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.*, **48**, 245–252.

Hormozdiari,F. *et al.* (2016) Colocalization of GWAS and eQTL signals detects target genes. *Am. J. Hum. Genet.*, **99**, 1245–1260.

Lappalainen,T. *et al.*; The Geuvadis Consortium. (2013) Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, **501**, 506–511.

Leek,J.T. and Storey,J.D. (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.*, **3**, e161.

Liang,L.M. *et al.* (2013) A cross-platform analysis of 14,177 expression quantitative trait loci derived from lymphoblastoid cell lines. *Genome Res.*, **23**, 716–726.

Liu,X. *et al.* (2018) GBAT: a gene-based association method for robust *trans*--gene regulation detection. bioRxiv, 395970.

Nanda,V. *et al.* (2018) Functional regulatory mechanism of smooth muscle cell-restricted LMOD1 coronary artery disease locus. *PLoS Genet.*, **14**, e1007755.

Parisien,M. *et al.* (2017) Effect of human genetic variability on gene expression in dorsal root ganglia and association with pain phenotypes. *Cell Rep.*, **19**, 1940–1952.

Parts,L. *et al.* (2011) Joint genetic analysis of gene expression data with inferred cellular phenotypes. *PLoS Genet.*, **7**, e1001276.

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

Porcu,E. *et al.*; eQTLGen Consortium. (2019) Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. *Nat. Commun.*, **10**, 1–12.

Raitakari,O.T. *et al.* (2008) Cohort profile: the cardiovascular risk in Young Finns Study. *Int. J. Epidemiol.*, **37**, 1220–1226.

Raj,T. *et al.* (2014) Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. *Science*, **344**, 519–523.

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**.

Saha,A. and Battle,A. (2018) False positives in trans-eQTL and co-expression analyses arising from RNA-sequencing alignment errors. *F1000Research*, **7**, 1860–1860.

Shabalin,A.A. (2012) Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, **28**, 1353–1358.

Staiger,D. and Stock,J.H. (1997) Instrumental variables regression with weak instruments. *Econometrica*, **65**, 557–586.

Stančáková,A. *et al.* (2012) Hyperglycemia and a common variant of GCKR are associated with the levels of eight amino acids in 9,369 Finnish men. *Diabetes*, **61**, 1895–1902.

Stegle,O. *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**, e1000770.

Torres,J.M. *et al.* (2014) Cross-tissue and tissue-specific eQTLs: partitioning the heritability of a complex trait. *Am. J. Hum. Genet.*, **95**, 521–534.

Tung,J. *et al.* (2015) The genetic architecture of gene expression levels in wild baboons. *Elife*, **4, e04729**.

Wald,A. (1940) The fitting of straight lines if both variables are subject to error. *Ann. Math. Stat.*, **11**, 284–300.

Wright,F.A. *et al.* (2014) Heritability and genomics of gene expression in peripheral blood. *Nat. Genet.*, **46**, 430–437.

Yuan,Z. *et al.* (2020) Testing and controlling for horizontal pleiotropy with the probabilistic Mendelian randomization in transcriptome-wide association studies. *Nat. Commun.*, **11**, 1–14.

Zeng,P. *et al.* (2019) Causal association of type 2 diabetes with amyotrophic lateral sclerosis: new evidence from Mendelian randomization using GWAS summary statistics. *BMC Med.*, **17**, 225.

Zeng,P. and Zhou,X. (2017) Non-parametric genetic prediction of complex traits with latent Dirichlet process regression models. *Nat. Commun.*, **8**, 1–11.

Zeng,P. and Zhou,X. (2019) Causal effects of blood lipids on amyotrophic lateral sclerosis: a Mendelian randomization study. *Hum. Mol. Genet.*, **28**, 688–697.