PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

An independent component analysis confounding factor correction framework for identifying broad impact expression quantitative trait loci

Jin Hyun Ju[1,2], Sushila A. Shenoy[1], Ronald G. Crystal[1], Jason G. Mezey[1,2,3]*

1 Department of Genetic Medicine, Weill Cornell Medical College, New York, NY, United States of America, **2** Institute for Computational Biomedicine, Weill Cornell Medical College, New York, NY, United States of America, **3** Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY, United States of America

* jgm45@cornell.edu

OPEN ACCESS

Citation: Ju JH, Shenoy SA, Crystal RG, Mezey JG (2017) An independent component analysis confounding factor correction framework for identifying broad impact expression quantitative trait loci. PLoS Comput Biol 13(5): e1005537. https://doi.org/10.1371/journal.pcbi.1005537

Editor: Stephen B Montgomery, Stanford University, UNITED STATES

Received: November 22, 2016

Accepted: April 28, 2017

Published: May 15, 2017

Data Availability Statement: The genotype data for the MuTHER analysis is available from the TwinsUK resource executive committee for researchers who meet the criteria for access to confidential data, and the expression data is available for download from the ArrayExpress repository (https://www.ebi.ac.uk/arrayexpress/) accession E-TABM-1140. Expression data for the GTEx analysis is available through the GTEx portal (http://www.gtexportal.org/) release ID GTEx Analysis V6. Genotype data for the GTEx analysis are available through dbGaP (https://www.ncbi.

Abstract

Genome-wide expression Quantitative Trait Loci (eQTL) studies in humans have provided numerous insights into the genetics of both gene expression and complex diseases. While the majority of eQTL identified in genome-wide analyses impact a single gene, eQTL that impact many genes are particularly valuable for network modeling and disease analysis. To enable the identification of such broad impact eQTL, we introduce CONFETI: Confounding Factor Estimation Through Independent component analysis. CONFETI is designed to address two conflicting issues when searching for broad impact eQTL: the need to account for non-genetic confounding factors that can lower the power of the analysis or produce broad impact eQTL false positives, and the tendency of methods that account for confounding factors to model broad impact eQTL as non-genetic variation. The key advance of the CONFETI framework is the use of Independent Component Analysis (ICA) to identify variation likely caused by broad impact eQTL when constructing the sample covariance matrix used for the random effect in a mixed model. We show that CONFETI has better performance than other mixed model confounding factor methods when considering broad impact eQTL recovery from synthetic data. We also used the CONFETI framework and these same confounding factor methods to identify eQTL that replicate between matched twin pair datasets in the Multiple Tissue Human Expression Resource (MuTHER), the Depression Genes Networks study (DGN), the Netherlands Study of Depression and Anxiety (NESDA), and multiple tissue types in the Genotype-Tissue Expression (GTEx) consortium. These analyses identified both *cis*-eQTL and *trans*-eQTL impacting individual genes, and CONFETI had better or comparable performance to other mixed model confounding factor analysis methods when identifying such eQTL. In these analyses, we were able to identify and replicate a few broad impact eQTL although the overall number was small even when applying CONFETI. In light of these results, we discuss the broad impact eQTL that have been previously reported from the analysis of human data and suggest that

nlm.nih.gov/gap) accession phs000424.v6.p1, and expression and genotype data for the NESDA study are available via accession phs000486.v1.p1. For access to genotype and expression data from the DGN study the data access committee at NIMH Center for Collaborative Genetics Studies on Mental Disorders can be contacted (https://www.nimhgenetics.org/access_data_biomaterial.php).

Funding: JHJ was supported by National Institutes of Health (https://www.nih.gov/) grants HL113443, HL118541 and HL134549, and SAS was supported by National Institutes of Health (https://www.nih.gov/) grant HL094284 and Qatar National Research Fund (http://www.qnrf.org/en-us/) grant NPRP 7-1425-3-370. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

considerable caution should be exercised when making biological inferences based on these reported eQTL.

Author summary

The discovery of expression Quantitative Trait Loci (eQTL) from the analysis of genome-wide genotype and gene expression data has played an important role in the study of cellular processes and complex disease. Here, we introduce CONFETI: Confounding Factor Estimation Through Independent component analysis, an analysis framework that has been designed to identify eQTL with broad impacts on the expression levels of many genes. The CONFETI framework takes advantage of Independent Component Analysis (ICA) to separate putative genetic and non-genetic factors in a confounding factor mixed model analysis, such that broad impact eQTL are not corrected out of the analysis as confounding variation. We show that CONFETI has better performance for identifying broad impact eQTL compared to the most widely applied confounding factor correction methods when applied to simulated data. We also applied CONFETI and these same methods to identify eQTL that replicate between twin pairs from the MuTHER consortium, the Depression Genes Networks study (DGN), the Netherlands Study of Depression and Anxiety (NESDA), and common tissue type pairs in the Genotype-Tissue Expression (GTEx) consortium. Surprisingly, while CONFETI had comparable replication performance compared to other methods, we were able to identify and replicate a very small number of broad impact eQTL overall. We discuss reports of broad impact eQTL in humans and suggest that they should be interpreted with caution.

Introduction

The current genome-wide picture of the genetics of gene expression in humans has been driven by studies of expression Quantitative Trait Loci (eQTL) that analyze the statistical associations between genotypes and gene expression [1–14]. Such eQTL discovery approaches have lead to a number of generalizations about the genetics of gene expression and regulation at genome-wide scales [3, 15] including the observation that the majority of genes in the genome can be impacted by an eQTL [16], that *cis*-eQTL have significantly larger effect sizes than *trans*-eQTL [10, 17, 18], and that eQTL can have tissue specific impacts on an expressed gene [19, 20]. Genome-wide eQTL discovery has also provided a foundation for inferences about biological systems and disease. For example, eQTL are used within data aggregation methods to annotate the functional or fitness impacts of polymorphisms [21], which in turn is a main component of systems biology models of pathways and cellular processes [22–26]. Discovered eQTL are also used for network modeling, in large part because eQTL can be used to model a directed impact on gene expression, which in turn can be leveraged to infer other directed network relationships among expressed genes [27–30]. As a final example, eQTL are routinely leveraged to identify candidate disease risk loci within regions associated with complex diseases in genome-wide association studies (GWAS) by making the assumption that when an eQTL co-locates with a locus identified in a GWAS, the same allelic variants are impacting both gene expression and disease risk [18, 20, 31–51].

For studies that leverage eQTL as a foundation for network modeling or for identifying candidate disease risk loci, eQTL that are associated with multiple genes can be particularly

valuable. For directed network modeling, the value of such broad impact eQTL is clear, since the network inference depends on tracking the impact of eQTL through multiple genes [52–56]. When considering associations with complex diseases, eQTL that affect many genes have been hypothesized to have effects beyond the transcriptome and are therefore good candidates for affecting a downstream disease phenotype [57]. Such broad impact eQTL [13], variously referred to as eQTL hotspots [58], master regulators [59], *trans*-regulators [60], and *trans*-eQTL networks [61] could result from either hotspots of multiple co-located eQTL [58, 62] or from the pleiotropic effects of a single eQTL genotype [43]. Broad impact eQTL have regularly been observed in model organisms such as yeast [63–65] and mice [66], but have been reported less frequently and in smaller numbers in human eQTL studies [58].

Since broad impact eQTL are expected to primarily affect *trans*-genes, statistical power has been suggested as a possible reason for the relatively lower reporting of broad impact eQTL in humans, since *trans*-eQTL tend to have relatively weak associations in humans compared to model organisms [58]. Furthermore, due to the large number of possible genotype-expression variable pair comparisons in human eQTL studies, which can range from 10^9 to 10^{10} for array based studies [7] and 10^{11} to 10^{12} for data collected by next-generation sequencing technologies [12], it is common to reduce the multiple testing burden by only considering a subset of *trans*-pairs [18, 67, 68] or to not consider *trans*-associations at all [12, 14]. A consequence of such strategies is a significant undercount of the number of *trans*- compared to *cis*-eQTL genome-wide, making the identification of broad impact eQTL with multiple *trans*-gene effects almost impossible.

A promising analysis strategy that could partially alleviate the statistical difficulties in identifying broad impact eQTL is the use of confounding factor analysis [69]. Confounding factor methods account for non-genetic variation in eQTL studies by learning and modeling non-genetic effects or variation directly from the multivariate structure observed in gene expression data [62, 69–76]. When used in combination with corrections for population structure [70], confounding factor analysis can both increase power in eQTL studies and reduce false positives by accounting for non-genetic factors that impact many genes, such as technical variation caused by differences in laboratory procedures or distinct study environments [58, 77, 78]. While confounding factor analyses should increase the correct discovery of both *cis*- and *trans*-eQTL by increasing detection power [9, 77], a known problem of all confounding factor methods is the potential to model the effects of broad impact eQTL as confounding variation [72, 79]. Previous approaches to avoid the removal of broad impact eQTL as confounding factors include jointly estimating the error structure with genetic information [72], and using only a subset of genes to estimate the confounding structure [75]. However, such approaches do not explicitly identify individual confounding factors and could generate different results based on selected genes, which is a non-optimal strategy for avoiding the removal of variation produced by broad impact eQTL.

In this study, we describe a new framework that is designed to improve on the performance of confounding factor methods to identify broad impact eQTL. The CONFETI (Confounding Factor Estimation Through Independent component analysis) framework makes use of the machine learning method Independent Component Analysis (ICA) to separate genetic components from non-genetic components learned from multivariate gene expression variation. ICA is a widely used blind source separation method applied to problems such as voice and image separation, and more recently to high dimensional gene expression data to estimate non-Gaussian generative sources from an observed mixture [80–84]. CONFETI takes advantages of the key strength of ICA to estimate generative sources of variation from an observed mixture, which can be used to separate independent sources of variation, such as genetic versus non-genetic factors. After these generative sources have been estimated by ICA, CONFETI

automatically filters out those that are candidates for broad impact eQTL variation and retains the rest as a lower dimensional representation of the non-genetic confounding variation. By explicitly identifying clear candidate signals of broad impact eQTL, CONFETI prevents the explaining away of true genetic effects and increases the discovery potential of confounding factor analyses.

To show the potential of CONFETI for the discovery of broad impact eQTL, we evaluated performance using simulated genome-wide data. For these simulated datasets, we show that the CONFETI framework successfully corrects for the effects of the confounding factors without explaining away broad impact eQTL. We also show that CONFETI has considerably increased performance compared to the most commonly applied confounding factor analysis methods.

We then assessed the ability of the CONFETI framework and these other methods to identify and replicate eQTL between matched twin pair datasets in the Multiple Tissue Human Expression Resource (MuTHER) [10], between the whole blood samples of the Depression Genes Networks study (DGN) [13] and the Netherlands Study of Depression and Anxiety (NESDA) [85], and between tissues of the same broad type in the Genotype-Tissue Expression (GTEx) [14]. We found that confounding factor correction methods greatly increased the number of replicating eQTL in all of the analyzed datasets. In particular, linear mixed model based methods increased both the number of replicating *cis*- and *trans*-eQTL. While we found that CONFETI had better or comparable performance to other methods in the replication of both *cis*- and *trans*-eQTL with individual gene impacts, after careful modeling and consideration of population structure, confounding factors, annotation inconsistencies, read alignment artifacts, and visual inspection of false positive indicators, we were able to identify only a few replicating broad impact eQTL at a genome-wide significance threshold in the MuTHER lymphoblastoid cell line (LCL) dataset. Taken together, these results suggest that robustly identifiable broad impact eQTL in humans have considerably smaller effects per gene than the bulk of eQTL. We discuss the implications of these results when considering factors such as sample size that can impact broad impact eQTL discovery, as well as for the use of previously reported broad impact eQTL as a foundation for making biological inferences.

Overview of the CONFETI framework

The CONFETI framework is constructed to systematically avoid the tendency of other confounding factor analysis methods to model broad impact eQTL as confounding variation. This is accomplished by leveraging Independent Component Analysis (ICA) to identify generative sources of multivariate gene expression variation and then screening candidates based on component correlations with genotypes, which are then omitted from the confounding factor correction (Fig 1). ICA is widely used in machine learning for blind source separation problems to detect non-Gaussian signals from multivariate data and has been applied to a diverse set of problems including voice and image separation [86, 87]. The reason ICA is particularly well suited for identifying candidate broad impact eQTL is that the method is designed to separate independent sources of multivariate variation.

ICA assumes that the observed data for each sample is a linear combination of non-Gaussian statistically independent components. When applying ICA, the vector of expression values for an individual are modeled as weighted sum of independent components:

$$\vec{y}_i = a_{i1}\vec{s_1} + a_{i2}\vec{s_2} + \cdots + a_{ik}\vec{s_k} = \sum_{j=1}^{k} a_{ij}s_j \quad (1)$$

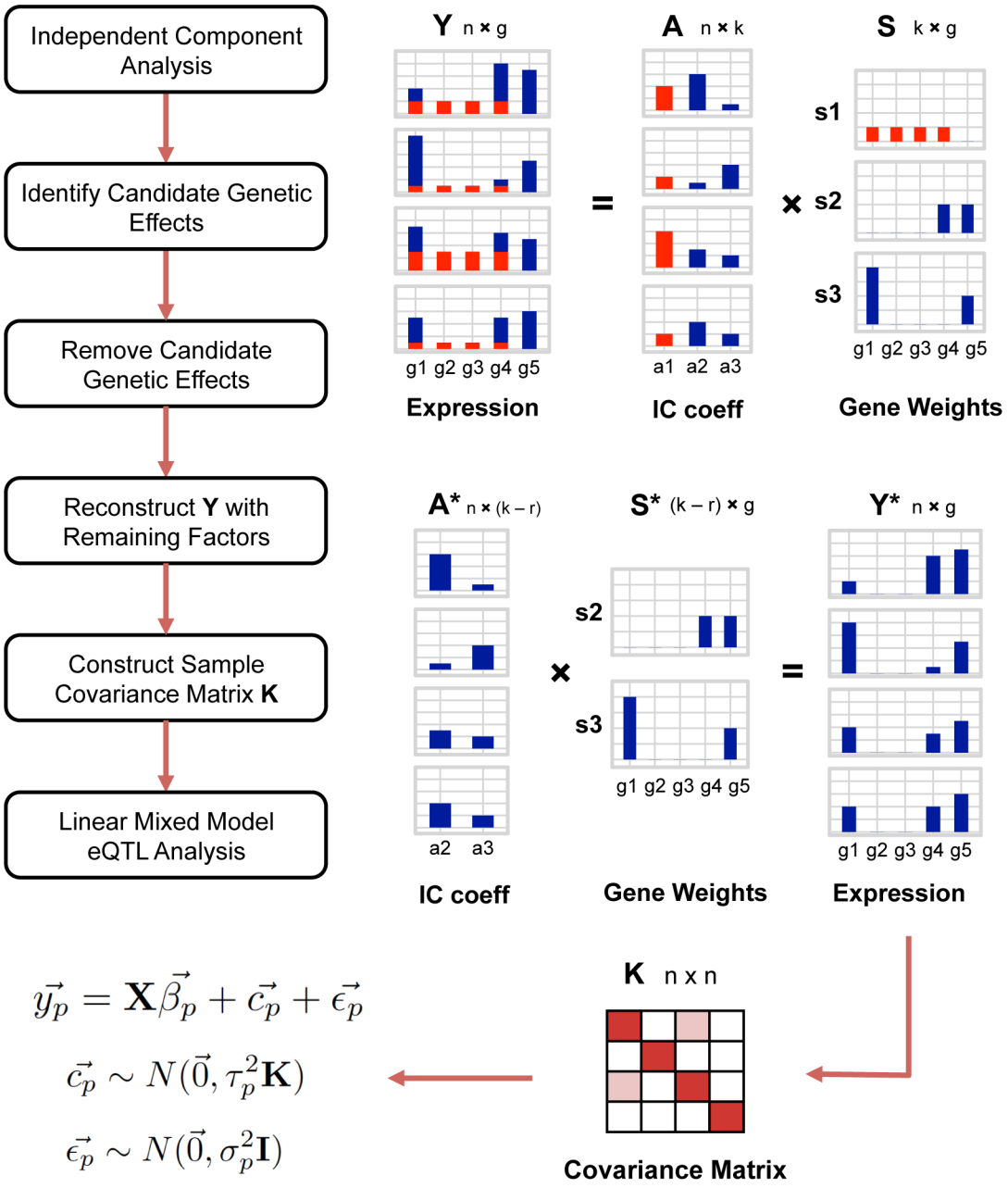

Fig 1. The CONFETI framework. ICA is used to decompose the gene expression matrix **Y** into an IC coefficient matrix **A** and a component matrix **S**. Associations between the genotypes and coefficients in matrix **A** are tested to label any candidate genetic effects to be removed from the correction. In the example above, the first IC, shown in red, is marked as a candidate genetic component and the corresponding columns of **A** and rows of **S** are removed. Using the lower rank **A*** and **S***, expression values originating from non-genetic components are reconstructed in **Y***. Finally, **K** is created by calculating the sample covariance matrix of **Y***, and included as a random effect in the mixed model for eQTL analysis.

https://doi.org/10.1371/journal.pcbi.1005537.g001

where $\vec{y}_i$ is a g-dimensional vector of gene expression values for a single sample, and independent components $\vec{s}_j$ are g-dimensional vectors of gene weights that are shared among all samples and the scalar component coefficients a_{ij} represent the contribution of each independent component $\vec{s}_j$ for sample i (Fig 1). When considering all samples together, the above can be

simply expressed as a matrix decomposition:

$$\mathbf{Y} = \mathbf{AS} \quad (2)$$

where $\mathbf{Y}$ is an $n \times g$ matrix with i^{th} row $\vec{y}_i$. $\mathbf{A}$ is the $n \times k$ mixing matrix with the j^{th} column holding component coefficients $\vec{a}_j$ for component j, and $\mathbf{S}$ is the $k \times g$ independent component matrix in which the j^{th} row is $\vec{s}_j$. $\mathbf{A}$ and $\mathbf{S}$ are estimated by finding a projection of $\mathbf{Y}$ that maximizes the non-Gaussianity of the gene weight distribution of each row in $\mathbf{S}$. In CONFETI these are identified by using the FastICA algorithm for reliable and fast computation [88].

Since ICA recovers factors by assessing non-Gaussianity and not the amount of variation explained as in methods such as Principal Component Analysis (PCA) or any other factor analysis method [86], ICA is able to more clearly resolve separate factors responsible for variation, while a PCA or factor analysis will tend to identify composite effects, which are likely to be mixtures of multiple factors (S1 Fig). The critical assumption for application of ICA in the CONFETI framework is that broad impact eQTL will have non-Gaussian impacts on the multivariate expression profile and that the effects of these eQTL will be relatively independent of other genetic and non-genetic factors. Complete independence is not necessary, since the framework only has to identify and retain enough of the expression variation due to a broad impact eQTL to make it detectable with an association test. The assumption that broad impact eQTL will tend to have non-Gaussian impacts is not particularly restrictive given that we expect eQTL with large enough effects to impact only a subset of the total number of genes and therefore be detectably non-Gaussian. The assumption that broad impact eQTL are relatively independent of each other is also not overly restrictive in humans given the low linkage disequilibrium observed among non-local genotypes throughout the genome. While the assumption that broad impact eQTL are largely independent of non-genetic factors is not always expected to hold, it seems likely in many cases unless there is a reason to expect broad impact eQTL to strongly interact with non-genetic factors such as sample-specific environmental effects or technical effects arising from differences between laboratories and procedures. Furthermore, in cases where broad impact eQTL are completely conflated with non-genetic factors, these broad impact eQTL will be indistinguishable from non-genetic contributions to the observed multivariate gene expression variation and will be modeled away by any confounding factor method. In summary, the only accurately detectable broad impact eQTL are those that have properties that are expected to make them identifiable by ICA.

The complete CONFETI framework involves running ICA on multivariate gene expression data, an automated detection step to identify candidate broad impact eQTL by assessing associations with genotypes, and omission of these factors for the construction of the random effect sample covariance matrix used in a mixed model confounding factor analysis (Fig 1). While this approach could be used in combination with confounding factor methods that use a fixed covariate approach [69, 74, 76, 89–92], the framework more naturally integrates with a mixed model approaches to confounding factor analysis, since the random effect modeling in these methods provides a high dimensional modeling of confounding variation. A covariance matrix constructed from the non-genetic independent components is used to model confounding factors as random effects in a linear mixed model eQTL approach.

We note that our framework differs from ICA methods for eQTL detection that treat the identified ICs as meta-genes, where these methods cannot reliably distinguish the specific gene effects of individual eQTL [83, 93]. The only method that we are aware of close to this framework is ISVA, which uses ICA within the Surrogate Variable Analysis (SVA) method for iteratively modeling pre-specified fixed effects and confounding variation [91]. ISVA is not appropriate for eQTL analysis since it begins the iterative approach by pre-specifying the fixed

effects and therefore pre-supposing the existence of a relationship, which would introduce a bias towards finding eQTL false positives. CONFETI on the other hand uses ICA to separate candidate broad impact eQTL without the need of pre-specifying the existence of the eQTL. We also note that in the mixed model based method PANAMA [72], the authors discuss a strategy for avoiding the over-correction of *trans*-eQTL by jointly estimating the covariance matrix with genotype effects to avoid including those effects in the correction [72]. However, this approach is not a feature of PANAMA included in the LIMIX package [94], which the authors have directed us to use. Moreover, the gene loadings in PANAMA are integrated out in the estimation step making it difficult to analyze the factors that are being corrected. In summary, the CONFETI framework utilizes the optimal properties of ICA to detect broad impact eQTL by excluding genetic effects from confounding variation accounted for in a mixed model, thereby taking advantage of the performance increases provided by mixed model confounding factor analysis without reducing the ability to identify broad impact eQTL.

Methods

Independent component analysis

To apply ICA to gene expression data and generate a sample covariance matrix, we developed a custom R package (https://github.com/jinhyunju/confeti). The independent component estimation features are using functions adopted from the fastICA R package [95] which implemented the computationally efficient and robust FastICA algorithm [88] based on a fixed-point algorithm to find directions maximizing the Negentropy to identify statistically independent components (ICs). The number of ICs that can be estimated is the smaller of the sample size or the number of features (genes), and the sign of any particular estimated component is arbitrary. As the estimated ICs do not have any particular order and have the potential to change based on the input of number of components to estimate [91, 96, 97], the package supports diagnostics for assessing optimal IC number such as functionality to estimate replicating ICs between multiple runs for ensemble ICA estimation. To provide a fair comparison between ICA and PANAMA [72], which both require as input the number of components to be considered prior to estimation, we set the number of ICs to be estimated in the fastICA algorithm to explain the same variance as for the set of principal components accounting for 95% of the variance in the data.

Removal of candidate broad impact eQTL

After decomposing the observed data $\mathbf{Y}$ into $\mathbf{A}$ and $\mathbf{S}$ we test for any significant associations between the component coefficients (columns of $\mathbf{A}$) and all genotypes. As in fixed effect eQTL models, we fit a linear regression model with the IC coefficient as the dependent variable and the genotype values as independent variables. After calculating p-values for each IC coefficient and genotype pair, we identified candidate broad impact eQTL using a global Bonferroni corrected p-value threshold of 0.05. Components with at least one significant association are marked as candidate genetic components. After filtering out r ($0 \leq r < k$) components with significant genotype association, we reconstruct expression matrix $\mathbf{Y}^*$ originating from non-genetic factors using the remaining $k - r$ components:

$$\mathbf{Y}^* = \mathbf{A}^*\mathbf{S}^* \tag{3}$$

where $\mathbf{Y}^*$ is an $n \times g$ matrix, $\mathbf{A}^*$ is a $n \times (k - r)$ matrix and $\mathbf{S}^*$ is a $(k - r) \times g$ matrix.

Given that the overall CONFETI method makes use of the phenotype and genotype data both in the filtering out of candidate genetic effects and in the identification of significant genotype-gene expression associations, using the full dataset could lead to model over-fitting impacts in the selection and removal of ICs. To assess this issue, we compared the approach of using CONFETI on the full dataset to a strategy where we split the genotype data into two random subsets. For the splitting strategy, we used one of the genotype subsets for filtering candidate genetic effects and the remaining genotypes for the eQTL analysis, we then repeated the analysis flipping the subsets that are used for filtering and eQTL analysis, and the combined the results. With this splitting strategy, genotypes used for the removal of candidate genetic effects do not overlap with the genotypes that are being tested for eQTL, such that each genotype is only accessed once in each subset.

From the analysis of multiple datasets, we found that the results obtained by using the full dataset and the splitting strategy largely overlapped with only minor differences (S2 Fig). A possible reason for this observation is that over-fitting issue in the CONFETI framework differs from more standard cases in machine learning applications in that the estimated independent components are not being directly used as features, but are rather included in the model to account for sample similarity structures that violate the independence assumption of the model, i.e., selected features are not being tested for associations. While we present the splitting strategy as an option for selecting and removing ICs for the users of CONFETI, given agreement with results when using the full dataset, and the additional complexity and computational costs in data splitting, separate analysis, and combining steps, we suggest applying CONFETI when considering the full dataset and adopt this approach in these analyses.

Construction of sample covariance matrices

We used two approaches to construct the sample covariance matrix $\mathbf{K}$ for the random effect part of the mixed model. Our first approach was to use a simple location-scale normalization of each gene of $\mathbf{Y}^*$:

$$Z^*_{ip} = (Y^*_{ip} - \mu_p)/\sigma_p \tag{4}$$

and then calculate sample covariance matrix:

$$\mathbf{K} = \mathrm{cov}(\mathbf{Z}^*) \tag{5}$$

We label this approach CONFETI-I since it can be thought of as a specific, lower dimensional approach to Intersample Correlation Emended (ICE), one of the first methods to estimate a sample structure for confounding factor analysis [62] by estimating the sample covariance matrix using the full dimensional observed expression data.

For our second approach, we couple CONFETI with PANAMA (Probabilistic ANAlysis of genoMic dAta) [72] that estimates the covariance structure using a maximum likelihood framework. Using this approach, the likelihood objective can be stated as:

$$p(\mathbf{Y}^*|\mathbf{K}_{\text{panama}}) = \prod_{p=1}^{g} \mathcal{N}(\vec{y^*_{\cdot p}}|\mathbf{K}_{\text{panama}} + \sigma_p^2\mathbf{I}) \tag{6}$$

$$(\hat{\boldsymbol{\theta}}, \hat{\mathbf{C}}) = \text{argmax}_{\boldsymbol{\theta},\mathbf{C}}\ p(\mathbf{Y}^*|\mathbf{C}, \boldsymbol{\theta}) \tag{7}$$

where $\mathbf{C}$ is an $n \times Q$ matrix initialized by projecting the observed data onto the first Q principal components explaining 95% of the variance and is further optimized in the process, and $\boldsymbol{\theta}$ is the set of hyperparameters consisting of $\{\{\alpha_q^2\}, \sigma_p^2\}$. Each α_q^2 then represents the optimized

weight of the q^{th} column of $\mathbf{C}$, $\mathbf{C}_{\cdot q}$ in constructing the sample covariance matrix:

$$\mathbf{K} = \sum_{q=1}^{Q} \hat{\alpha}_q^2 \hat{\mathbf{C}}_{\cdot q} \hat{\mathbf{C}}_{\cdot q}^{\mathrm{T}} \tag{8}$$

We label this approach CONFETI-P, where we use of the implementation of PANAMA included in the LIMIX package [94] for the estimation of $\mathbf{K}$.

Mixed model eQTL analysis

We model the genetic effects from SNPs and covariates as fixed effects and confounding factor effects as random effects, such that the expression levels for gene p in n individuals are:

$$\vec{y}_{\cdot p} = \mathbf{X}\vec{\beta}_p + \vec{c}_p + \vec{\epsilon}_p \tag{9}$$

$$\vec{c}_p \sim \mathcal{N}(\vec{0}, \tau_p^2 \mathbf{K}) \tag{10}$$

$$\vec{\epsilon}_p \sim \mathcal{N}(\vec{0}, \sigma_p^2 \mathbf{I}) \tag{11}$$

where n is the number of samples, g the number of genes, s the number of SNPs, and v the number of covariates. Each gene expression vector $\vec{y}_{\cdot p}$ has dimension $n \times 1$ and is mean centered. The $n \times (1 + v)$ genotype and covariate matrix $\mathbf{X}$ contains a single genotype as the number of minor alleles coded as 0,1,2 and any additional v number of covariates. $\vec{\beta}_p$ is the $(1 + v) \times 1$ dimensional coefficient vector representing the fixed effect of the SNPs and covariates on gene p. The confounding effect is included in the model as a $n \times 1$ random effect $\vec{c}_p$ sampled from a multivariate normal distribution with covariance $\tau_p^2\mathbf{K}$, where $\mathbf{K}$ is the $n \times n$ sample covariance matrix constructed the corresponding confounding correction method, τ_p^2 is a scalar weight for $\mathbf{K}$ in the random effect, and $\vec{\epsilon}_p$ is a $n \times 1$ vector representing the independent error for gene p with scalar weight σ_p^2.

Analysis methods compared

We compared CONFETI-I and CONFETI-P to a simple linear regression with no confounding factor correction (LINEAR), including PCA projections as fixed effects (PCA), probabilistic estimation of expression residuals (PEER) [92], and mixed model confounding factor methods ICE [62] and PANAMA [72]. For mixed model based confounding factor correction methods, we limited our comparison to methods that pre-calculate a sample covariance matrix ($\mathbf{K}$), which is kept constant when testing individual genotypes against phenotypes, to avoid the computational burden of recalculating $\mathbf{K}$ for every phenotype.

For each comparison of methods on simulated or real data, we ran each method to be as equivalent as possible, including the same covariates and using the same linear mixed model fitting function. For CONFETI-I, CONFETI-P, PANAMA, and ICE we used `lrgprApply()` function from the R package `lrgpr` [98] to fit the linear mixed model and calculate p-values for the genotype effects using a Wald test. Following the methodology of the GTEx analysis [14], the number of factors for PEER were decided based on the sample size. We used 30 factors for datasets with sample size between 150 and 250, and 35 factors for datasets with more than 250 samples. We used the same number of components for PCA correction. To fit the eQTL model using PEER, PCA, and LINEAR we used the `glmApply()` function from `lrgpr` and used a Wald test for significance testing.

Performance benchmarking on simulated data

To mirror real cases where a reasonable number of broad impact eQTL have been repeatedly identified, we used yeast as a model [63–65]. To create simulated datasets, we used 2956 yeast genotypes from the study of Smith et al. [99] and randomly sampled 3000 yeast gene annotations to simulate *cis*- and *trans*-eQTL relationships. To simulate eQTL, a matrix with a dimension of number of genotypes × number of expression phenotypes was first created that marks genotype and phenotype pairs *cis*- if the starting position of the gene and the genotype were within 100,000 base pairs distance and *trans*- if the distance was greater. From this matrix we sampled 2500 genotype and phenotype pairs which consisted of 80% *cis*- and 20% *trans*-genotypes. In total, for each simulated dataset, we included 2000 *cis*-eQTL, 500 *trans*-eQTL, and 10 broad impact eQTL. We simulated each broad impact eQTL to affect 10% of the expression phenotypes. Effect sizes for *cis*-eQTL were sampled from $\mathcal{N}(0.8, 1)$ and effect sizes for *trans*-eQTL and broad impact eQTL were sampled from $\mathcal{N}(0.48, 1)$ (70% attenuation of *trans*-effects) to reflect observed effect sizes in real data. After the eQTL effects were simulated, we added normally distributed random noise sampled from $\mathcal{N}(0, 1)$. For confounding factor effects, we simulated two types of confounding factors: sparse and dense. For sparse confounding factors 30% of phenotypes were affected with effect sizes drawn from $\mathcal{N}(1, 0.5)$, and for the dense confounding factors, the effect over all genes followed a standard normal distribution $\mathcal{N}(0, 1)$. We tested 2 scenarios, each with 30 confounding factors: sparse only, and mixed (15 sparse and 15 dense). We simulated and analyzed 50 datasets for each of these two scenarios, a total of 100 datasets.

We ran each of the methods CONFETI-I, CONFETI-P, PANAMA, ICE, PEER, PCA, and LINEAR on each of the 100 datasets using the method settings and parameters as described above. To evaluate performance for each method, we ranked the eQTL for each method according to their p-values and then calculated the True Positive Rate (TPR) and False Positive Rate (FPR) and generated Receiver Operating Characteristic (ROC) curves for each method, where we also calculated the area under the curve for each method across the simulation scenarios. True eQTL were further labeled as *cis*-, *trans*- or broad impact and the recovery rate for each category at different FDR thresholds was calculated by dividing the number of true genotype phenotype pairs that were called significant by the total number of true genotype phenotype pairs in each category. To provide an upper bound metric on how well methods could recover each of these eQTL types, we also simulated the same scenarios without any confounding factors and reported the ROC curves after running LINEAR. We labeled these results 'TMR' for 'Theoretical Maximum Recovery' since these represent the maximum recovery expected in theory if confounding factors were perfectly modeled by the confounding factor methods.

eQTL analysis in human datasets

We analyzed data from the Multiple Tissue Human Expression Resource (MuTHER) [10] project, the Depression Genes Networks study (DGN) [13], the Netherlands Study of Depression and Anxiety (NESDA) [85], and from the Genotype-Tissue Expression (GTEx) consortium [14] to compare the performance of the methods and to potentially identify broad impact eQTL in humans. Given that true eQTL are not known for human data we used replication as a metric for performance. While this is an imperfect metric and will tend to undercount true positives, replication does provide relative control over non-systematic false positives, such that a method that is overly liberal in calling of eQTL false positives will be appropriately assessed.

We ran eQTL analysis on the adipose, lymphoblastoid cell line (LCL), and skin datasets obtained through the MuTHER project [10]. Based on the matched twins information, there

Table 1. Sample size for each subset of MuTHER dataset analyzed.

Tissue	Subset Sample Size
Adipose	327
LCL	329
Skin	253

https://doi.org/10.1371/journal.pcbi.1005537.t001

were 161 monozygotic and 220 dizygotic twin pairs in the dataset. We only selected samples that had both genotype and gene expression measurements for both individuals in each twin pair for all three tissue types. To assess replication within a tissue type, we split each tissue specific dataset into two subsets separating each twin pair into different subsets. This created two subsets for each tissue type resulting in 327 samples for adipose, 329 for LCL, and 253 samples for skin (Table 1). For each subset there were 28,964 genes in Adipose, 28,894 genes in each LCL, and 28,893 genes in Skin. Genotype information was provided by the TwinsUK consortium, and we used only non-imputed genotypes from the downloaded data with minor allele frequencies higher than 5% (a total of 246,298 genotypes).

We also analyzed data from the DGN [13] and NESDA [85] studies. These independent studies analyzed blood samples and have large sample sizes. Normalized gene expression measurements and genotype files were obtained for DGN that were analyzed previously [13]. Genes which could not be unambiguously mapped to an Entrez Gene ID were excluded as well as SNPs which were not present in dbSNP. In the final DGN dataset there were 922 samples with 15,169 genes and 719,149 genotypes. Genotype data, gene expression data, and information regarding twin pairs for NESDA were downloaded via dbGaP (phs000486.v1.p1). SNPs with minor allele frequency less than 0.05 or which were not present in dbSNP were excluded. In the final NESDA datasets there were 641,753 genotypes and 45,137 genes with expression level measurements. To match the sample sizes in the two datasets to be within a similar size range for assessing replication, we split the NESDA dataset by available twin status information similar to the strategy used in the MuTHER analysis. This resulted in two subsets from the NESDA dataset with 636 samples in each subset (Table 2).

For the analysis of the GTEx datasets, we selected 4 pairs of tissues (Adipose, Artery, Heart, Skin) from GTEx release v6 (dbGaP Accession phs000424.v6.p1) with over 150 samples that have both RNA-seq gene expression and SNP array genotypes (Table 3). For gene expression, we included all genes which could be unambiguously mapped to Entrez Gene IDs (24,686 genes). Within each tissue, we excluded any genes which had zero measurements in more than 80% of samples as well as genes with highly skewed distributions, with more than 85% of measurements in the top or bottom 20%. After these filters were applied, the number of genes for each tissue was between 19,207 and 20,108. For genotypes, we excluded SNPs with missing genotypes and those with minor allele frequency <0.05. We also pruned SNPS within 10kb with pairwise $r^2 > 0.99$ and removed SNPs which were deprecated in dbSNP (1,270,565 SNPs remaining).

Table 2. Sample size for each blood dataset analyzed.

Dataset	Sample Size
DGN	922
NESDA subset1	636
NESDA subset2	636

https://doi.org/10.1371/journal.pcbi.1005537.t002

Table 3. Sample size for each GTEx dataset analyzed.

Tissue	Subtype	Sample Size
Adipose	Subcutaneous	298
Adipose	Visceral	185
Artery	Aorta	197
Artery	Tibial	285
Heart	Atrial Appendage	159
Heart	Left Ventricle	190
Skin	Leg	302
Skin	Suprapubic	196

https://doi.org/10.1371/journal.pcbi.1005537.t003

We fit CONFETI-I, CONFETI-P, PANAMA, ICE, PEER, PCA, and LINEAR for every phenotype and genotype pair in each of the datasets using the method settings and parameters as described above. To control for population structure, we included principal components derived from the genotypes, using the first five (DGN analysis), three (MuTHER analysis), and two (GTEx and NESDA analysis) principal components as covariates in each analysis. While a permutation approach is often applied to avoid any systematic inflation or deflation of the p-values, this was computationally infeasible for this study given the number of datasets analyzed and number of methods applied to each dataset. We therefore calculated the genomic inflation factor λ for each expression phenotype, a statistic which has been shown to provide a good metric for assessing model fit and appropriate p-value distributions [70, 72]. The λ statistic was calculated per gene using the median p-value m_p as

$$\lambda_p = \text{qchisq}(1 - m_p)/\text{qchisq}(0.5) \tag{12}$$

where qchisq is a quantile function for the chi-square distribution with 1 degree of freedom. For each method we assessed inflation using λ_p values for every gene to calculate $\lambda_{\text{diff},p} = 1 - \lambda_p$.

After calculating p-values for all phenotype and genotype pairs, we adjusted the p-values using Benjamini-Hochberg multiple hypothesis correction. The corrected p-values represent upper bounds on False Discovery Rate (FDR) [100]. We used a threshold of 0.01 on the adjusted p-values to mark significant eQTL. An eQTL (significant SNP gene pair) was labeled as *cis*- if the SNP and gene were located on the same chromosome within 1 Mb, and *trans*- otherwise.

To avoid potential artifacts caused by ambiguous RNA-seq alignment we screened *trans*-eQTL using two methods. First, we used annotated gene relationships available from NCBI (ftp://ftp.ncbi.nih.gov/gene/DATA/gene_group.gz) to identify trans-eQTL where the SNP was within 1Mb of a gene related to the eQTL gene (such as a pseudogene or functional gene 'parent' of a pseudogene). Because not all gene relationships were captured in the NCBI annotation, we searched for additional, potentially unannotated pseudogenes using the BLAT tool [101] to align all gene transcripts to the genome and identified all genomic regions matching at least 50% of each transcript. We omitted any trans-eQTL where the SNP was within 1Mb of a region matching the eQTL gene transcript. This "pseudo-*trans*" screening revealed that a number of the replicating *trans*-eQTL were artifacts arising due to incorrect/ambiguous mapping of RNA-seq reads that are in fact caused by *cis*-regulation of a gene, which shares sequence similarity with the eQTL gene. We also visually inspected eQTL for artifact or false positive indicators (e.g., individual genotype associations inconsistent with local linkage disequilibrium).

In order to avoid double-counting eQTL associated with multiple linked SNPs, we selected at most one significant *cis*- and *trans*- SNP per cytoband per gene. Using this criteria, we measured the replication of eQTL between and across different tissues counting the overlapping cytoband and gene pairs that were called significant in each dataset. We marked broad impact eQTL by searching for genotypes that showed more than a single *trans*-eQTL associations on different chromosomes that replicated between at least one twin or tissue pair.

Results

Simulation results

In our analysis of simulated data, we assessed the performance of the eQTL analysis methods CONFETI-I, CONFETI-P, PANAMA, ICE, PEER, PCA, and LINEAR on their ability to identify three types of eQTL, *cis*-, *trans*- and broad impact, in the presence of confounding factors. We also included the theoretical maximum recovery (TMR) as an upper limit of eQTL detection for each eQTL category, where the phenotype data has only normally distributed random noise added without any confounding factor effects. For both sparse and dense confounding factor effects, all methods showed significant improvements over LINEAR (linear regression without confounding factor correction), and CONFETI-I correctly identified the most eQTL at every FDR threshold (Fig 2). We found that linear mixed model based methods recovered individual *cis*- and *trans*-eQTL more accurately in comparison to linear fixed effect based correction methods PEER and PCA, where one explanation for this observation could be the lower power of fixed effect correction models by the increased number of parameters [102].

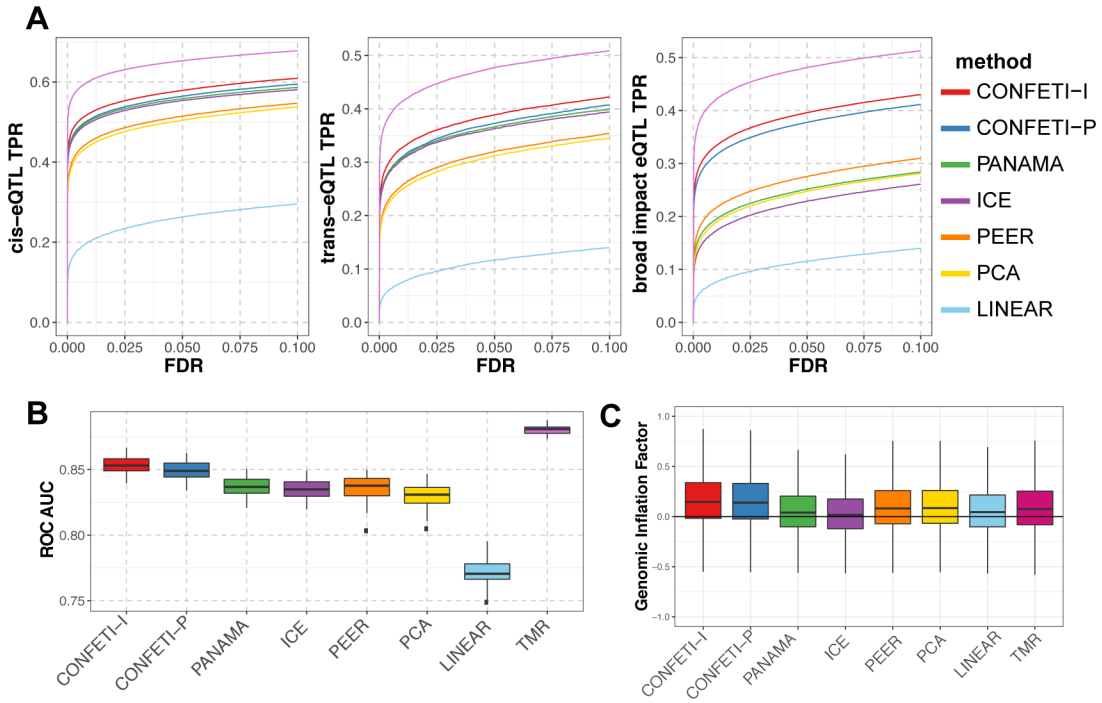

Fig 2. Comparison of method performance for simulated data in the presence of sparse confounding factors. **(A)** The recovery rate of simulated *cis*- (left), *trans*- (middle) and broad impact eQTL (right) for a range of FDR significance thresholds for each method averaging over the 50 simulated datasets with sparse confounding factors. The theoretical maximum recovery (TMR) shows the recovery when no confounding factors are included. **(B)** The Area Under the Curve (AUC) for the receiver operator characteristic (ROC) curves. **(C)** Box-plots of genomic inflation factors calculated for each method across the 50 simulated datasets with sparse factors.

https://doi.org/10.1371/journal.pcbi.1005537.g002

For broad impact eQTL in particular, CONFETI-I and CONFETI-P outperformed all other methods by a large margin illustrating the value of distinguishing genetic and non-genetic factors in the correction.

The difference between the confounding factor methods decreased with a combination of sparse and dense confounding factors compared to cases with just sparse confounding factors (S3 Fig), although the general trends remained consistent. This is likely due to the relative amount of total variance explained by each confounding factor and broad impact eQTL. In the dense confounding factor scenario, the confounding factors contribute a significantly higher proportion of the total variance compared to broad impact eQTL. In such a case, distinguishing genetic variance from non-genetic variance has less influence on the covariance matrix correction, since the majority of the variation in the data is originating from the confounding factors, and the resulting difference between methods in identifying true eQTL is expected to be smaller.

Overall, approaches such as PANAMA, ICE, PEER, and PCA which do not explicitly remove genetic effects from their correction, increased the accuracy in identifying individual *cis*- and *trans*-eQTL but incorrectly modeled broad impact eQTL as confounding factors. While the extent to which any simulated data will capture the true confounding factor conditions and genetic architectures of real eQTL datasets is unknown, these simulations demonstrate that the CONFETI framework can provide a considerable performance improvement compared to mixed model confounding factor methods in some situations, and performed at least as well as other methods overall.

Human data analysis results

We ran each of the eQTL analysis methods on the six datasets from MuTHER [10] (twin pairs in Adipose, LCL and Skin Tissues), the DGN [13] and NESDA [85] datasets (blood), and eight datasets from GTEx [14] (Adipose, Visceral vs. Subcutaneous; Artery, Aorta vs Tibial Artery; Heart, Atrial Appendage vs. Left Ventricle; Skin, Leg vs. Suprapubic). For each method applied to each dataset, we inspected the median λ genomic inflation factor [103] as a measure of appropriate model fit and control of false positives and false negatives rates. Linear mixed model based correction methods showed a slight inflation in comparison to linear fixed effect based methods with ICE showing the highest degree of inflation of p-values in every dataset. Overall, all methods were within acceptable fit levels of inflation or deflation when including genotype PCs as covariates (S4 Fig).

When considering different significance thresholds for individual datasets, we found that *cis*-eQTL discovery starts to asymptote while *trans*-eQTL discovery does not (Fig 3, S5 and S6 Figs). This is consistent with the overall smaller effect size of *trans*-eQTL, which makes them more difficult to detect. Confounding factor correction methods greatly increased the number of *cis*-eQTL identified in every dataset in comparison to LINEAR, demonstrating the increase of power by accounting for systematic variation. Linear mixed model correction methods CONFETI-I, CONFETI-P, PANAMA, and ICE identified comparable numbers of *cis*-eQTL in each dataset, followed by fixed effect correction methods PCA and PEER. Similarly, CONFETI-I, CONFETI-P, PANAMA, and ICE increased the number of identified *trans*-eQTL. However, the number of *trans*-eQTL identified by PEER and PCA were comparable or even lower than the results of LINEAR in some datasets.

While the DGN analysis yielded almost 3 to 4 fold increase for *cis* and *trans*-eQTL identification compared MuTHER and GTEx datasets, both subsets of NESDA found fewer *cis*-eQTL and similar numbers of *trans*-eQTL. While the decrease in NESDA sample size produced by splitting the datasets into subsets of twins could have affected the results, we would still expect

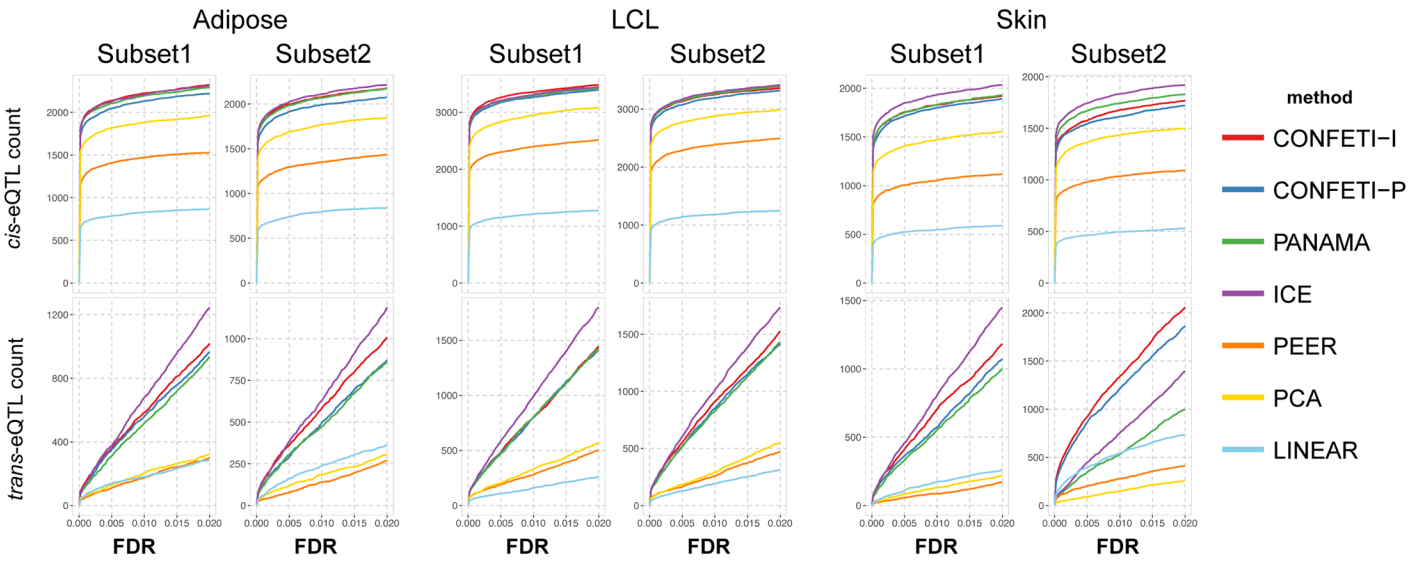

Fig 3. Significant eQTL discovered in MuTHER datasets for varying FDR thresholds. Plots showing the counts of *cis*- and *trans*-eQTL versus a range of FDR significance thresholds for each of the methods applied to every dataset.

https://doi.org/10.1371/journal.pcbi.1005537.g003

the number of *cis* and *trans*-eQTL discoveries to increase compared to the datasets analyzed in MuTHER and GTEx, which had roughly half the sample size. One potential factor influencing the results might be the higher multiple hypothesis testing correction burden in the subsets of NESDA mainly driven by the additional number of gene expression measurements. However, this alone could not explain the significantly lower number of eQTL found in the NESDA dataset, since we would expect to see a steeper increase of *cis*-eQTL discoveries at lower FDR thresholds based on the increased sample size compared to MuTHER and GTEx datasets.

We investigated the replication of eQTL found in each twin pair in the MuTHER dataset, across the DGN and NESDA datasets, and for each tissue pair in the GTEx datasets. Based on the results of individual tissues, we used a significance threshold of FDR < 0.01 to further investigate the replication of eQTL focusing on high confidence results. We found that similar to eQTL discovery in each dataset, confounding factor correction increased the number of replicating *cis* and *trans*-eQTL with linear mixed model based methods showing the most significant increase (S7 and S8 Figs). For MuTHER and GTEx, we observed a large number of replicating *cis*-eQTL in all twin pairs and tissue pairs, respectively, and a significantly lower number of replicating *trans*-eQTL, a result that was also observed in other studies [18, 104]. In each twin pair and tissue pair, CONFETI-I, CONFETI-P, PANAMA, and ICE identified similar numbers of replicating *cis*- and *trans*-eQTL, which were significantly higher than PCA and PEER. Between linear mixed model correction methods, the majority of eQTL were being found by multiple methods and only a few eQTL were unique to each method. This indicated that linear mixed model based correction increases the power of the model over linear fixed effect corrections, however that the differences between methods in constructing the sample covariance matrix lead to few novel discoveries per dataset (S9 and S10 Figs). Twin pairs showed a higher degree of replication compared to similar tissue pairs, which could be explained by the heterogeneity between tissue subtypes in the GTEx dataset (Fig 4). The replication ratio for *cis*-eQTL showed little difference between methods and was considerably higher than the replication ratio of *trans*-eQTL, which also showed higher variation between methods.

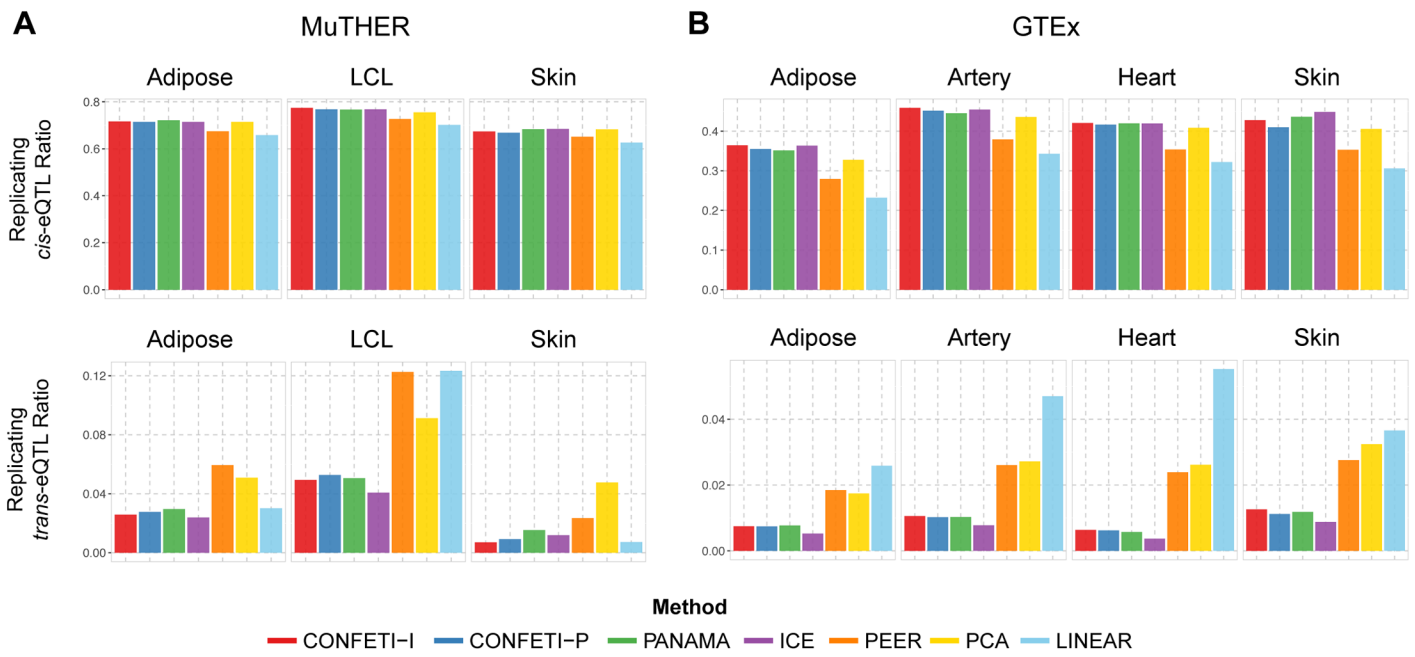

Fig 4. Replication ratio of *cis*- and *trans*-eQTL in MuTHER and GTEx dataset pairs by each method. The replication ratio calculated separately for *cis*- and *trans*-eQTL. The number of replicating eQTL are divided by the union of identified unique eQTL for each method in analyzing the **(A)** MuTHER twin pairs and **(B)** GTEx Tissue pairs.

https://doi.org/10.1371/journal.pcbi.1005537.g004

For the DGN and NESDA datasets the linear mixed model correction methods showed a higher increase in the number of replicating eQTL over linear fixed effect correction methods. The replication rate between the NESDA subsets were comparable to the results for MuTHER and GTEx, with most of the *cis*-eQTL identified in both datasets with few unique discoveries. However, due to the imbalance of identified eQTL between the DGN and NESDA datasets, the number of replicating eQTL were limited by the eQTL discovered in the NESDA subsets and resulted in lower replication rates with approximately 10% for *cis*-eQTL and below 1% for *trans*-eQTL.

We further investigated the results for replicating broad impact eQTL. Before the artifact correction protocol, we found replicating broad impact eQTL in GTEx datasets, but excluding pseudogenes from the replicating eQTLs effectively removed all replicating broad impact eQTL from the GTEx dataset. This is consistent with the findings in a study by Jo et al. [105], in which the authors state that they were unable to identify any individually significant genes with *trans*-eQTL after testing the associations between a single locus and all expressed genes in both subcutaneous and visceral subsets. This paper did report rs7037324 and rs1867277 on the 9q22 locus of being associated with TMEM253 and ARFGEF3 in the thyroid tissue, and rs2706381 and rs1012793 on the 5q31 locus to be associated with PSME1 and ARTD10 in skeletal muscle. However, these tissues had no replicates where we could assess eQTL replication across the same broad tissue type and were not included in our analysis. Both the DGN and NESDA studies reported broad impact eQTL separately [13, 85], but in our analysis we were unable to find any replicating broad impact eQTL among the datasets.

We were able to identify a few broad impact eQTL that replicated in the MuTHER LCL dataset (Fig 5). Most of these impacted only a few genes in *trans*, where rs3817963 impacted the highest number of genes (S1 Table), including a *cis* gene HLA-DRA, and six *trans* (CCDC28B, CSNK2A1, ERG, LIMS1, RPL34, XRCC6). The enrichment of regulatory signals

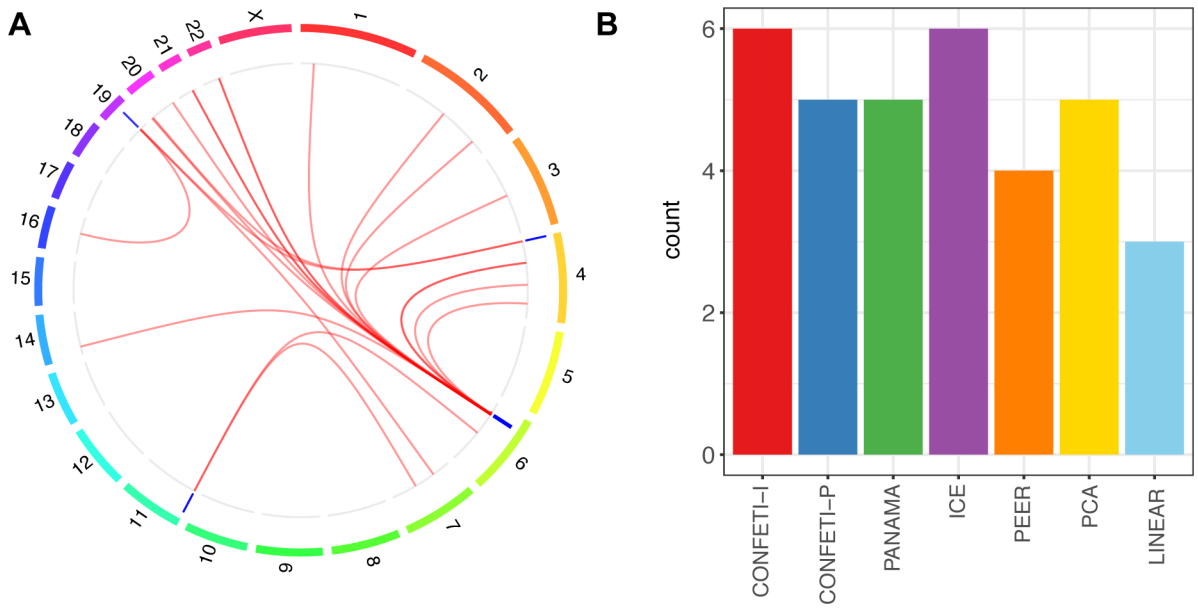

Fig 5. Replicating broad impact eQTL identified in the MuTHER LCL dataset. (A) Chromosomes are plotted in the outermost circles with replicating broad impact *trans*-eQTL as blue bands in the inner layer, where red lines connect each *trans*-eQTL to the associated gene. **(B)** The number of replicating broad impact eQTL found in the MuTHER LCL twin pair by each method.

https://doi.org/10.1371/journal.pcbi.1005537.g005

in the LCL dataset on the region of chromosome 6 which rs3817963 is located is proximal to the major histocompatibility complex (MHC) region, which is critical in immune cell function. A cluster of replicating eQTL on the same region of chromosome 6 was also found in the Adipose and Skin twin pair, however, we only found replicating eQTL associated with genes on different chromosomes in the LCL dataset (S11 Fig). We also found an individual case of a broad impact eQTL in the MuTHER Skin twin pair, which was found by ICE. Using PEER we did identify a genotype impacting two genes (C8orf82, MYL5) that were a subset of reported broad impact eQTL genes in the study by Small et al. [61] (S1 Table).

We did not find all of the broad impact eQTL reported by previous studies in the MuTHER Adipose dataset [10, 61], which might be a function of our conservative testing threshold. We therefore used the approach of considering the replicating broad impact eQTL we could identify by focusing only on genotypes with significant *cis*-eQTL as a strategy for adjusting the significance threshold. While using only a subset of genotypes effectively lowered the significance threshold for identifying eQTL overall and led to the identification of few additional replicating broad impact eQTL, it created little difference overall (Section 1 in S1 Text). We also investigated whether independent components significantly associated with genotypes could be used to identify broad impact eQTL. We found that a number of the components that were marked as candidate genetic effects resembled the significance level of individual eQTL with a small number of highly contributing genes. However, ICA does not have a stringent sparsity restriction in estimating the components, so distinguishing between genes, which are highly contributing to the component and noise is challenging (Section 2 in S1 Text). We note that methods enforcing sparsity in the estimation process of components [106, 107] could be an alternative to ICA in directly identifying broad impact eQTL from the data.

Discussion

We have introduced the confounding factor correction framework CONFETI, which uses Independent Component Analysis (ICA) to avoid over-correcting genetic effects in eQTL mixed model confounding factor analysis. CONFETI provides an easy to implement solution for a known problem with eQTL confounding factor methods: the tendency of these methods to model the effects of eQTL with broad impacts on many genes as confounding variation. In sum, the CONFETI approach provides a method for finding broad impact eQTL while leveraging the advantages of confounding factor analysis for eQTL discovery, a capability that has not been systematically implemented in currently available confounding factor analysis software.

In our real data evaluation of CONFETI and other methods, we found that confounding factor correction methods, especially linear mixed model based methods, increased the findings of replicating eQTL. This was also the case for identifying broad impact eQTL that replicated at a genome-wide significance level between datasets. While we did not find any replicating broad impact eQTL for the GTEx tissue pairs, we did find a number of broad impact eQTL when analyzing the MuTHER LCL dataset. Given that broad impact eQTL appear to have relatively small per gene impacts and the larger sample size of MuTHER compared to the GTEx datasets we analyzed, this supports power, and therefore sample size, as being a critical issue when detecting broad impact eQTL. However, this is clearly not the only critical factor, since only one broad impact eQTL was identified by PEER in the MuTHER Adipose dataset, only one was identified by ICE in the Skin dataset, and no broad impact eQTL were identified when comparing results for the considerably larger DGN and NESDA studies. Given LCL are likely to allow a more controlled and homogeneous measurement of gene expression variation compared to the mixed cell populations sampled *in vivo* for MuTHER adipose and skin datasets, and the even great heterogeneity across distinct studies of DGN and NESDA, it seems likely that different levels of sampling heterogeneity are also influencing broad impact eQTL discovery.

We were not able to replicate a small number of broad impact eQTLs reported by previous studies in the MuTHER Adipose dataset [10, 61]. One possible explanation could be the lower sample size of our analysis resulting from the splitting of twins in each dataset for replication. Another issue to consider is that both studies had less stringent thresholds for identifying significant *trans*-eQTL compared to the FDR of less than 1% threshold used in our study. Small et al. narrowed down the targets to investigate by testing a single genotype rs4731702, which significantly lowered the multiple testing burden [61], and both studies had a threshold of $P < 5 \times 10^{-8}$ which corresponded to an FDR threshold of less than 10% in Grundberg et al. [10]. Given that *trans* associations are the most prone to statistical false positives, it seems reasonable to view these previous reports of broad impact eQTL with caution.

In contrast to humans, broad impact eQTL have been easier to detect in model organisms and *trans*-eQTL seem less dispersed [58]. Given the landscape of broad impact eQTL in humans, the question is therefore what sample sizes and study conditions will be required to detect broad impact eQTL that are robust? Answering this question will require more genome-wide eQTL studies with larger sample sizes, more control over heterogeneity, and careful analysis with strategies designed to remove broad impact eQTL false positives.

Supporting information

S1 Fig. Example of multivariate gender effects recovered by Independent Component Analysis (ICA) and Principal Component Analysis (PCA). The components that had the strongest association with gender labels estimated by ICA (left) and PCA (right) for the Skin-

Leg GTEx dataset [14]. Gene weights for the independent component (IC) and principal component (PC) are shown on the top row, and the scatter plot and histogram graph pairs on the lower row show the coefficients of independent component (left) and projection of the samples onto the principal component (right). The scatter plots and histograms are colored based on the gender labels, female (orange) and male (green).
(PNG)

S2 Fig. Comparison of results when CONFETI is applied to a full dataset versus the genotype splitting strategy. This figure shows a typical result obtained when comparing CONFETI-I using the full dataset analysis to the splitting strategy. The MuTHER Adipose subset1 was analyzed using CONFETI-I with both strategies. A total of 2,633 hits were identified by both approaches and only 99 and 79 unique hits were identified for the splitting and full dataset analyses respectively. **(A)** The overlap of eQTL identified for the full dataset and splitting strategy. **(B)** Comparison of -log10 p-values for significant eQTL identified with the full dataset (x-axis) and splitting strategy (y-axis).
(PNG)

S3 Fig. Comparison of method performance for simulated data in the presence of a mix of sparse and dense confounding factors. (A) The recovery rate of simulated *cis*- (left), *trans*- (middle) and broad impact eQTL (right) for a range of FDR significance thresholds for each method averaging over the 50 simulated datasets with a mix of dense and sparse confounding factors. The theoretical maximum recovery (TMR) shows the recovery when no confounding factors are included. **(B)** The Area Under the Curve (AUC) for the receiver operator characteristic (ROC) curves. **(C)** Box-plots of genomic inflation factors calculated for each method for each method across the 50 simulated datasets with a mix of sparse and dense factors.
(PNG)

S4 Fig. Inflation factors for each MuTHER and GTEx dataset. Boxplots of the range of median λ genomic inflation factors calculated for each expression phenotype for each method in every dataset of **(A)** MuTHER and **(B)** GTEx.
(PNG)

S5 Fig. Significant eQTL discovered in GTEx datasets for varying FDR thresholds. Plots showing the counts of *cis*- and *trans*-eQTL versus FDR for each of the methods applied to every dataset.
(PNG)

S6 Fig. Number of identified *cis* and *trans*-eQTL in the DGN and NESDA datasets. The number of identified (Top) *cis*-eQTL and (Bottom) *trans*-eQTL in the DGN dataset and one of the two twin subsets of the NESDA study is shown for a range of FDR for all confounding factor correction methods.
(PNG)

S7 Fig. Replicating *cis*- and *trans*-eQTL discovered in the MuTHER and GTEx datasets for varying FDR thresholds. Plots showing the counts of replicating *cis*- and *trans*-eQTL versus FDR for each of the methods applied to every **(A)** MuTHER and **(B)** GTEx dataset.
(PNG)

S8 Fig. Number of replicating *cis* and *trans*-eQTL in datasets of human blood. The number of replicating *cis*-eQTL and *trans*-eQTL between the DGN dataset and two twin subsets of the NESDA study is shown for a range of FDR for all confounding factor correction methods.
(PNG)

S9 Fig. Replication of eQTL in each MuTHER tissue type. Replicating *cis*- and *trans*-eQTL found by each method in the respective tissue types are ordered on the x-axis by the amount of overlap between methods. Colored bars corresponding to their method indicate that the particular eQTL replicated. The total numbers of replicating eQTL for each method is shown on at the end of each bar. Results are shown for **(A)** Adipose, **(B)** LCL, and **(C)** Skin twin pairs.
(PNG)

S10 Fig. Replication of eQTL in each GTEx tissue type. Replicating *cis*- and *trans*-eQTL found by each method in the respective tissue types are ordered on the x-axis by the amount of overlap between methods. Colored bars corresponding to their method indicate that the particular eQTL replicated. The total numbers of replicating eQTL for each method is shown on at the end of each bar. Results are shown for **A.** Adipose, **B.** Artery, **C.** Heart, and **D.** Skin tissue pairs.
(PNG)

S11 Fig. Replication of eQTL in MuTHER twin pairs discovered by CONFETI-I. Chromosomes are plotted in the outermost circles with replicating *cis*-eQTL shown in gray bands within the next layer, and replicating *trans*-eQTL as blue bands in the innermost layer where red lines connect each *trans*-eQTL to the associated gene with gene annotations labeled in blue outside the circle. Replication shown for **(A)** Adipose, **(B)** LCL, **(C)** Skin twin pairs.
(PNG)

S12 Fig. Replication of eQTL in GTEx tissue pairs discovered by CONFETI-I. Chromosomes are plotted in the outermost circles with replicating *cis*-eQTL shown in gray bands within the next layer, and replicating *trans*-eQTL as blue bands in the innermost layer where red lines connect each *trans*-eQTL to the associated gene with gene annotations labeled in blue outside the circle. Replication after removal of pseudogenes are shown for **(A)** Adipose, **(B)** Artery, **(C)** Heart, and **(D)** Skin tissue pairs.
(PNG)

S1 Table. Table of replicating broad impact eQTL identified in MuTHER datasets. Detailed information about the dataset, significant SNP, method, *cis* and *trans*-eQTL count and genes are shown for replicating broad impact eQTL.
(TXT)

S1 Text.
(PDF)

Acknowledgments

We thank the Wellcome Trust and TwinsUK participants for providing the genotypes and twin information for the Multiple Tissue Human Expression Resource (MuTHER) datasets. TwinsUK is funded by the Wellcome Trust, Medical Research Council, European Union, the National Institute for Health Research (NIHR)-funded BioResource, Clinical Research Facility and Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust in partnership with King's College London. We would also like to thank the Genotype-Tissue Expression (GTEx) consortium, Depression Genes Networks (DGN) study, and the Netherlands Study of Depression and Anxiety (NESDA) for granting access to the data.

Author Contributions

Conceptualization: JHJ JGM.

Data curation: SAS JHJ.

Formal analysis: JHJ SAS.

Funding acquisition: RGC JGM.

Investigation: JHJ.

Methodology: JHJ JGM.

Project administration: JGM.

Resources: RGC JGM.

Software: JHJ SAS.

Supervision: JGM.

Visualization: JHJ SAS.

Writing – original draft: JHJ SAS.

Writing – review & editing: JHJ SAS RGC JGM.

References

1. Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, et al. Genetics of gene expression surveyed in maize, mouse and man. Nature. 2003; 422(6929):297–302. https://doi.org/10.1038/nature01434 PMID: 12646919
2. Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, et al. Genetic analysis of genome-wide variation in human gene expression. Nature. 2004; 430(7001):743–747. https://doi.org/10.1038/nature02797 PMID: 15269782
3. Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, Burdick JT. Mapping determinants of human gene expression by regional and genome-wide association. Nature. 2005; 437(7063):1365–1369. https://doi.org/10.1038/nature04244 PMID: 16251966
4. Doss S, Schadt EE, Drake TA, Lusis AJ. Cis-acting expression quantitative trait loci in mice. Genome research. 2005; 15(5):681–691. https://doi.org/10.1101/gr.3216905 PMID: 15837804
5. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science. 2007; 315(5813):848–853. https://doi.org/10.1126/science.1136678 PMID: 17289997
6. Göring HH, Curran JE, Johnson MP, Dyer TD, Charlesworth J, Cole SA, et al. Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. Nature genetics. 2007; 39(10):1208–1216. https://doi.org/10.1038/ng2119 PMID: 17873875
7. Veyrieras JB, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, Stephens M, et al. High-resolution mapping of expression-QTLs yields insight into human gene regulation. PLoS Genet. 2008; 4(10):e1000214. https://doi.org/10.1371/journal.pgen.1000214 PMID: 18846210
8. Heinzen EL, Ge D, Cronin KD, Maia JM, Shianna KV, Gabriel WN, et al. Tissue-specific genetic control of splicing: implications for the study of complex traits. PLoS Biol. 2008; 6(12):e1000001. https://doi.org/10.1371/journal.pbio.1000001
9. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature. 2010; 464(7289):768–772. https://doi.org/10.1038/nature08872 PMID: 20220758
10. Grundberg E, Small KS, Hedman ÅK, Nica AC, Buil A, Keildson S, et al. Mapping cis-and trans-regulatory effects across multiple tissues in twins. Nature genetics. 2012; 44(10):1084–1089. https://doi.org/10.1038/ng.2394 PMID: 22941192
11. Mehta D, Heim K, Herder C, Carstensen M, Eckstein G, Schurmann C, et al. Impact of common regulatory single-nucleotide variants on gene expression profiles in whole blood. European Journal of Human Genetics. 2013; 21(1):48–54. https://doi.org/10.1038/ejhg.2012.106 PMID: 22692066
12. Lappalainen T, Sammeth M, Friedländer MR, AC't Hoen P, Monlong J, Rivas MA, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013; 501(7468):506–511. https://doi.org/10.1038/nature12531 PMID: 24037378

13. Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, McCormick C, et al. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome research. 2014; 24(1):14–24. https://doi.org/10.1101/gr.155192.113 PMID: 24092820

14. Consortium G, et al. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science. 2015; 348(6235):648–660. https://doi.org/10.1126/science.1262110 PMID: 25954001

15. Stranger BE, Forrest MS, Clark AG, Minichiello MJ, Deutsch S, Lyle R, et al. Genome-wide associations of gene expression variation in humans. PLoS Genet. 2005; 1(6):e78. https://doi.org/10.1371/journal.pgen.0010078 PMID: 16362079

16. Pai AA, Pritchard JK, Gilad Y. The genetic and mechanistic basis for variation in gene regulation. PLoS Genet. 2015; 11(1):e1004857. https://doi.org/10.1371/journal.pgen.1004857 PMID: 25569255

17. Petretto E, Mangion J, Dickens NJ, Cook SA, Kumaran MK, Lu H, et al. Heritability and tissue specificity of expression quantitative trait loci. PLoS Genet. 2006; 2(10):e172. https://doi.org/10.1371/journal.pgen.0020172 PMID: 17054398

18. Westra HJ, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nature genetics. 2013; 45(10): 1238–1243. https://doi.org/10.1038/ng.2756 PMID: 24013639

19. Price AL, Helgason A, Thorleifsson G, McCarroll SA, Kong A, Stefansson K. Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. PLoS Genet. 2011; 7(2):e1001317. https://doi.org/10.1371/journal.pgen.1001317 PMID: 21383966

20. Raj T, Rothamel K, Mostafavi S, Ye C, Lee MN, Replogle JM, et al. Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. Science. 2014; 344(6183):519–523. https://doi.org/10.1126/science.1249547 PMID: 24786080

21. Gamazon ER, Zhang W, Konkashbaev A, Duan S, Kistner EO, Nicolae DL, et al. SCAN: SNP and copy number annotation. Bioinformatics. 2010; 26(2):259–262. https://doi.org/10.1093/bioinformatics/btp644 PMID: 19933162

22. Zhong H, Yang X, Kaplan LM, Molony C, Schadt EE. Integrating pathway analysis and genetics of gene expression for genome-wide association studies. The American Journal of Human Genetics. 2010; 86(4):581–591. https://doi.org/10.1016/j.ajhg.2010.02.020 PMID: 20346437

23. Civelek M, Lusis AJ. Systems genetics approaches to understand complex traits. Nature Reviews Genetics. 2014; 15(1):34–48. https://doi.org/10.1038/nrg3575 PMID: 24296534

24. Williams KA, Lee M, Hu Y, Andreas J, Patel SJ, Zhang S, et al. A systems genetics approach identifies CXCL14, ITGAX, and LPCAT2 as novel aggressive prostate cancer susceptibility genes. PLoS Genet. 2014; 10(11):e1004809. https://doi.org/10.1371/journal.pgen.1004809 PMID: 25411967

25. Johnson MR, Behmoaras J, Bottolo L, Krishnan ML, Pernhorst K, Santoscoy PLM, et al. Systems genetics identifies Sestrin 3 as a regulator of a proconvulsant gene network in human epileptic hippocampus. Nature communications. 2015; 6. https://doi.org/10.1038/ncomms7031 PMID: 25615886

26. Wang J, Ma MCJ, Mennie AK, Pettus JM, Xu Y, Lin L, et al. Systems biology with high-throughput sequencing reveals genetic mechanisms underlying the metabolic syndrome in the Lyon hypertensive rat. Circulation: Cardiovascular Genetics. 2015; 8(2):316–326. https://doi.org/10.1161/CIRCGENETICS.114.000520 PMID: 25573024

27. Zhu J, Zhang B, Smith EN, Drees B, Brem RB, Kruglyak L, et al. Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. Nature genetics. 2008; 40(7): 854–861. https://doi.org/10.1038/ng.167 PMID: 18552845

28. Blair RH, Kliebenstein DJ, Churchill GA. What can causal networks tell us about metabolic pathways? PLoS Comput Biol. 2012; 8(4):e1002458. https://doi.org/10.1371/journal.pcbi.1002458 PMID: 22496633

29. Mäkinen VP, Civelek M, Meng Q, Zhang B, Zhu J, Levian C, et al. Integrative genomics reveals novel molecular pathways and gene networks for coronary artery disease. PLoS Genet. 2014; 10(7): e1004502. https://doi.org/10.1371/journal.pgen.1004502 PMID: 25033284

30. Chick JM, Munger SC, Simecek P, Huttlin EL, Choi K, Gatti DM, et al. Defining the consequences of genetic variation on a proteome-wide scale. Nature. 2016; 534(7608):500–505. https://doi.org/10.1038/nature18270 PMID: 27309819

31. Moffatt MF, Kabesch M, Liang L, Dixon AL, Strachan D, Heath S, et al. Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. Nature. 2007; 448(7152):470–473. https://doi.org/10.1038/nature06014 PMID: 17611496

32. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet. 2010; 6(4):e1000888. https://doi.org/10.1371/journal.pgen.1000888 PMID: 20369019

33. Musunuru K, Strong A, Frank-Kamenetsky M, Lee NE, Ahfeldt T, Sachs KV, et al. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature. 2010; 466(7307):714–719. https://doi.org/10.1038/nature09266 PMID: 20686566

34. Nica AC, Montgomery SB, Dimas AS, Stranger BE, Beazley C, Barroso I, et al. Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. PLoS Genet. 2010; 6(4):e1000895. https://doi.org/10.1371/journal.pgen.1000895 PMID: 20369022

35. Dubois PC, Trynka G, Franke L, Hunt KA, Romanos J, Curtotti A, et al. Multiple common variants for celiac disease influencing immune gene expression. Nature genetics. 2010; 42(4):295–302. https://doi.org/10.1038/ng.543 PMID: 20190752

36. Nguyen HH, Takata R, Akamatsu S, Shigemizu D, Tsunoda T, Furihata M, et al. IRX4 at 5p15 suppresses prostate cancer growth through the interaction with vitamin D receptor, conferring prostate cancer susceptibility. Human molecular genetics. 2012; p. dds025. https://doi.org/10.1093/hmg/dds025 PMID: 22323358

37. Zou F, Chai HS, Younkin CS, Allen M, Crook J, Pankratz VS, et al. Brain expression genome-wide association study (eGWAS) identifies human disease-associated variants. PLoS Genet. 2012; 8(6): e1002707. https://doi.org/10.1371/journal.pgen.1002707 PMID: 22685416

38. Miller CL, Anderson DR, Kundu RK, Raiesdana A, Nürnberg ST, Diaz R, et al. Disease-related growth factor and embryonic signaling pathways modulate an enhancer of TCF21 expression at the 6q23. 2 coronary heart disease locus. PLoS Genet. 2013; 9(7):e1003652. https://doi.org/10.1371/journal.pgen.1003652 PMID: 23874238

39. Lamontagne M, Couture C, Postma DS, Timens W, Sin DD, Pare PD, et al. Refining susceptibility loci of chronic obstructive pulmonary disease with lung eqtls. PLoS One. 2013; 8(7):e70220. https://doi.org/10.1371/journal.pone.0070220 PMID: 23936167

40. Kumar V, Westra HJ, Karjalainen J, Zhernakova DV, Esko T, Hrdlickova B, et al. Human disease-associated genetic variation impacts large intergenic non-coding RNA expression. PLoS Genet. 2013; 9(1):e1003201. https://doi.org/10.1371/journal.pgen.1003201 PMID: 23341781

41. Singh T, Levine AP, Smith PJ, Smith AM, Segal AW, Barrett JC. Characterization of Expression Quantitative Trait Loci in the Human Colon. Inflammatory bowel diseases. 2015; 21(2):251. https://doi.org/10.1097/MIB.0000000000000265 PMID: 25569741

42. Dermitzakis ET. From gene expression to disease risk. Nature genetics. 2008; 40(5):492–493. https://doi.org/10.1038/ng0508-492 PMID: 18443581

43. Gilad Y, Rifkin SA, Pritchard JK. Revealing the architecture of gene regulation: the promise of eQTL studies. Trends in genetics. 2008; 24(8):408–415. https://doi.org/10.1016/j.tig.2008.06.001 PMID: 18597885

44. Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M. Mapping complex disease traits with global gene expression. Nature Reviews Genetics. 2009; 10(3):184–194. https://doi.org/10.1038/nrg2537 PMID: 19223927

45. Fransen K, Visschedijk MC, van Sommeren S, Fu JY, Franke L, Festen EA, et al. Analysis of SNPs with an effect on gene expression identifies UBE2L3 and BCL3 as potential new risk genes for Crohn's disease. Human molecular genetics. 2010; 19(17):3482–3488. https://doi.org/10.1093/hmg/ddq264 PMID: 20601676

46. Zhong H, Beaulaurier J, Lum PY, Molony C, Yang X, MacNeil DJ, et al. Liver and adipose expression associated SNPs are enriched for association to type 2 diabetes. PLoS Genet. 2010; 6(5):e1000932. https://doi.org/10.1371/journal.pgen.1000932 PMID: 20463879

47. Montgomery SB, Dermitzakis ET. From expression QTLs to personalized transcriptomics. Nature Reviews Genetics. 2011; 12(4):277–282. https://doi.org/10.1038/nrg2969 PMID: 21386863

48. Kang HP, Morgan AA, Chen R, Schadt EE, Butte AJ. Coanalysis of GWAS with eQTLs reveals disease-tissue associations. AMIA Summits on Translational Science proceedings. 2012; 2012:35. PMID: 22779046

49. Richards AL, Jones L, Moskvina V, Kirov G, Gejman PV, Levinson DF, et al. Schizophrenia susceptibility alleles are enriched for alleles that affect gene expression in adult human brain. Molecular psychiatry. 2012; 17(2):193–201. https://doi.org/10.1038/mp.2011.11 PMID: 21339752

50. Edwards SL, Beesley J, French JD, Dunning AM. Beyond GWASs: illuminating the dark road from association to function. The American Journal of Human Genetics. 2013; 93(5):779–797. https://doi.org/10.1016/j.ajhg.2013.10.012 PMID: 24210251

51. He X, Fuller CK, Song Y, Meng Q, Zhang B, Yang X, et al. Sherlock: detecting gene-disease associations by matching patterns of expression QTL and GWAS. The American Journal of Human Genetics. 2013; 92(5):667–680. https://doi.org/10.1016/j.ajhg.2013.03.022 PMID: 23643380

52. Ghazalpour A, Doss S, Zhang B, Wang S, Plaisier C, Castellanos R, et al. Integrating genetic and network analysis to characterize genes related to mouse weight. PLoS Genet. 2006; 2(8):e130. https://doi.org/10.1371/journal.pgen.0020130 PMID: 16934000

53. Wu C, Delano DL, Mitro N, Su SV, Janes J, McClurg P, et al. Gene set enrichment in eQTL data identifies novel annotations and pathway regulators. PLoS Genet. 2008; 4(5):e1000070. https://doi.org/10.1371/journal.pgen.1000070 PMID: 18464898

54. Logsdon BA, Mezey J. Gene expression network reconstruction by convex feature selection when incorporating genetic perturbations. PLoS Comput Biol. 2010; 6(12):e1001014. https://doi.org/10.1371/journal.pcbi.1001014 PMID: 21152011

55. Heinig M, Petretto E, Wallace C, Bottolo L, Rotival M, Lu H, et al. A trans-acting locus regulates an anti-viral expression network and type 1 diabetes risk. Nature. 2010; 467(7314):460–464. https://doi.org/10.1038/nature09386 PMID: 20827270

56. Aterido A, Palacio C, Marsal S, Ávila G, Julià A. Novel insights into the regulatory architecture of CD4+ T cells in rheumatoid arthritis. PloS one. 2014; 9(6):e100690. https://doi.org/10.1371/journal.pone.0100690 PMID: 24959711

57. Chen Y, Zhu J, Lum PY, Yang X, Pinto S, MacNeil DJ, et al. Variations in DNA elucidate molecular networks that cause disease. Nature. 2008; 452(7186):429–435. https://doi.org/10.1038/nature06757 PMID: 18344982

58. Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. Nature Reviews Genetics. 2015; 16(4):197–212. https://doi.org/10.1038/nrg3891 PMID: 25707927

59. Fairfax BP, Makino S, Radhakrishnan J, Plant K, Leslie S, Dilthey A, et al. Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. Nature genetics. 2012; 44(5):502–510. https://doi.org/10.1038/ng.2205 PMID: 22446964

60. Kirsten H, Al-Hasani H, Holdt L, Gross A, Beutner F, Krohn K, et al. Dissecting the genetics of the human transcriptome identifies novel trait-related trans-eQTLs and corroborates the regulatory relevance of non-protein coding loci. Human molecular genetics. 2015; 24(16):4746–4763. https://doi.org/10.1093/hmg/ddv194 PMID: 26019233

61. Consortium M, et al. Identification of an imprinted master trans regulator at the KLF14 locus related to multiple metabolic phenotypes. Nature genetics. 2011; 43(6):561–564. https://doi.org/10.1038/ng.833 PMID: 21572415

62. Kang HM, Ye C, Eskin E. Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. Genetics. 2008; 180(4):1909–1925. https://doi.org/10.1534/genetics.108.094201 PMID: 18791227

63. Brem RB, Yvert G, Clinton R, Kruglyak L. Genetic dissection of transcriptional regulation in budding yeast. Science. 2002; 296(5568):752–755. https://doi.org/10.1126/science.1069516 PMID: 11923494

64. Brem RB, Kruglyak L. The landscape of genetic complexity across 5,700 gene expression traits in yeast. Proceedings of the National Academy of Sciences of the United States of America. 2005; 102(5):1572–1577. https://doi.org/10.1073/pnas.0408709102 PMID: 15659551

65. Foss EJ, Radulovic D, Shaffer SA, Ruderfer DM, Bedalov A, Goodlett DR, et al. Genetic basis of proteome variation in yeast. Nature genetics. 2007; 39(11):1369–1375. https://doi.org/10.1038/ng.2007.22 PMID: 17952072

66. van Nas A, Ingram-Drake L, Sinsheimer JS, Wang SS, Schadt EE, Drake T, et al. Expression quantitative trait loci: replication, tissue-and sex-specificity in mice. Genetics. 2010; 185(3):1059–1068. https://doi.org/10.1534/genetics.110.116087 PMID: 20439777

67. Fehrmann RS, Jansen RC, Veldink JH, Westra HJ, Arends D, Bonder MJ, et al. Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. PLoS Genet. 2011; 7(8):e1002197. https://doi.org/10.1371/journal.pgen.1002197 PMID: 21829388

68. Lee MN, Ye C, Villani AC, Raj T, Li W, Eisenhaure TM, et al. Common genetic variants modulate pathogen-sensing responses in human dendritic cells. Science. 2014; 343(6175):1246980. https://doi.org/10.1126/science.1246980 PMID: 24604203

69. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. 2007; 3(9):e161. https://doi.org/10.1371/journal.pgen.0030161 PMID: 17907809

70. Listgarten J, Kadie C, Schadt EE, Heckerman D. Correction for hidden confounders in the genetic analysis of gene expression. Proceedings of the National Academy of Sciences. 2010; 107(38):16465–16470. https://doi.org/10.1073/pnas.1002425107 PMID: 20810919

71. Stegle O, Parts L, Durbin R, Winn J. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. PLoS Comput Biol. 2010; 6(5):e1000770. https://doi.org/10.1371/journal.pcbi.1000770 PMID: 20463871

72. Fusi N, Stegle O, Lawrence ND. Joint modelling of confounding factors and prominent genetic regulators provides increased accuracy in genetical genomics studies. PLoS Comput Biol. 2012; 8(1): e1002330. https://doi.org/10.1371/journal.pcbi.1002330 PMID: 22241974

73. Yang C, Wang L, Zhang S, Zhao H. Accounting for non-genetic factors by low-rank representation and sparse regression for eQTL mapping. Bioinformatics. 2013; 29(8):1026–1034. https://doi.org/10.1093/bioinformatics/btt075 PMID: 23419377

74. Gao C, Tignor NL, Salit J, Strulovici-Barel Y, Hackett NR, Crystal RG, et al. HEFT: eQTL analysis of many thousands of expressed genes while simultaneously controlling for hidden factors. Bioinformatics. 2013; p. btt690. https://doi.org/10.1093/bioinformatics/btt690 PMID: 24307700

75. Joo JWJ, Sul JH, Han B, Ye C, Eskin E. Effectively identifying regulatory hotspots while capturing expression heterogeneity in gene expression studies. Genome biology. 2014; 15(4):1. https://doi.org/10.1186/gb-2014-15-4-r61 PMID: 24708878

76. Mostafavi S, Battle A, Zhu X, Urban AE, Levinson D, Montgomery SB, et al. Normalizing RNA-sequencing data by modeling hidden covariates with prior knowledge. PLoS One. 2013; 8(7):e68141. https://doi.org/10.1371/journal.pone.0068141 PMID: 23874524

77. AC't Hoen P, Friedländer MR, Almlöf J, Sammeth M, Pulyakhina I, Anvar SY, et al. Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. Nature biotechnology. 2013; 31(11):1015–1022. https://doi.org/10.1038/nbt.2702 PMID: 24037425

78. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nature Reviews Genetics. 2010; 11(10): 733–739. https://doi.org/10.1038/nrg2825 PMID: 20838408

79. Goldinger A, Henders AK, McRae AF, Martin NG, Gibson G, Montgomery GW, et al. Genetic and non-genetic variation revealed for the principal components of human gene expression. Genetics. 2013; 195(3):1117–1128. https://doi.org/10.1534/genetics.113.153221 PMID: 24026092

80. Lee SI, Batzoglou S. Application of independent component analysis to microarrays. Genome biology. 2003; 4(11):1. https://doi.org/10.1186/gb-2003-4-11-r76 PMID: 14611662

81. Engreitz JM, Daigle BJ, Marshall JJ, Altman RB. Independent component analysis: mining microarray data for fundamental human gene expression modules. Journal of biomedical informatics. 2010; 43(6):932–944. https://doi.org/10.1016/j.jbi.2010.07.001 PMID: 20619355

82. Bang-Berthelsen CH, Pedersen L, Fløyel T, Hagedorn PH, Gylvin T, Pociot F. Independent component and pathway-based analysis of miRNA-regulated gene expression in a model of type 1 diabetes. BMC genomics. 2011; 12(1):97. https://doi.org/10.1186/1471-2164-12-97 PMID: 21294859

83. Rotival M, Zeller T, Wild PS, Maouche S, Szymczak S, Schillert A, et al. Integrating genome-wide genetic variations and monocyte expression data reveals trans-regulated gene modules in humans. PLoS Genet. 2011; 7(12):e1002367. https://doi.org/10.1371/journal.pgen.1002367 PMID: 22144904

84. Krumsiek J, Suhre K, Illig T, Adamski J, Theis FJ. Bayesian independent component analysis recovers pathway signatures from blood metabolomics data. Journal of proteome research. 2012; 11(8): 4120–4131. https://doi.org/10.1021/pr300231n PMID: 22713116

85. Wright FA, Sullivan PF, Brooks AI, Zou F, Sun W, Xia K, et al. Heritability and genomics of gene expression in peripheral blood. Nature genetics. 2014; 46(5):430–437. https://doi.org/10.1038/ng.2951 PMID: 24728292

86. Hyvärinen A, Karhunen J, Oja E. Independent component analysis. vol. 46. John Wiley & Sons; 2004.

87. Comon P, Jutten C. Handbook of Blind Source Separation: Independent component analysis and applications. Academic press; 2010.

88. Hyvarinen A. Fast and robust fixed-point algorithms for independent component analysis. IEEE transactions on Neural Networks. 1999; 10(3):626–634. https://doi.org/10.1109/72.761722 PMID: 18252563

89. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007; 8(1):118–127. https://doi.org/10.1093/biostatistics/kxj037 PMID: 16632515

90. Stegle O, Kannan A, Durbin R, Winn J. Accounting for non-genetic factors improves the power of eQTL studies. In: Annual International Conference on Research in Computational Molecular Biology. Springer; 2008. p. 411–422.

91. Teschendorff AE, Zhuang J, Widschwendter M. Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. Bioinformatics. 2011; 27(11): 1496–1505. https://doi.org/10.1093/bioinformatics/btr171 PMID: 21471010

92. Stegle O, Parts L, Piipari M, Winn J, Durbin R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nature protocols. 2012; 7(3):500–507. https://doi.org/10.1038/nprot.2011.457 PMID: 22343431

93. Biswas S, Storey JD, Akey JM. Mapping gene expression quantitative trait loci by singular value decomposition and independent component analysis. BMC bioinformatics. 2008; 9(1):1. https://doi.org/10.1186/1471-2105-9-244 PMID: 18492285

94. Lippert C, Casale FP, Rakitsch B, Stegle O. LIMIX: genetic analysis of multiple traits. BioRxiv. 2014; p. 003905.

95. Marchini JL, Heaton C, Ripley BD. fastICA: FastICA Algorithms to perform ICA and Projection Pursuit; 2013. Available from: http://CRAN.R-project.org/package=fastICA.

96. Frigyesi A, Veerla S, Lindgren D, Höglund M. Independent component analysis reveals new and biologically significant structures in micro array data. BMC bioinformatics. 2006; 7(1):1. https://doi.org/10.1186/1471-2105-7-290

97. Biton A, Bernard-Pierrot I, Lou Y, Krucker C, Chapeaublanc E, Rubio-Pérez C, et al. Independent component analysis uncovers the landscape of the bladder tumor transcriptome and reveals insights into luminal and basal subtypes. Cell reports. 2014; 9(4):1235–1245. https://doi.org/10.1016/j.celrep.2014.10.035 PMID: 25456126

98. Hoffman GE, Mezey JG, Schadt EE. lrgpr: interactive linear mixed model analysis of genome-wide association studies with composite hypothesis testing and regression diagnostics in R. Bioinformatics. 2014; p. btu435. https://doi.org/10.1093/bioinformatics/btu435 PMID: 25035399

99. Smith EN, Kruglyak L. Gene–environment interaction in yeast gene expression. PLoS Biol. 2008; 6(4): e83. https://doi.org/10.1371/journal.pbio.0060083 PMID: 18416601

100. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the royal statistical society Series B (Methodological). 1995; p. 289–300.

101. Kent WJ. BLAT—the BLAST-like alignment tool. Genome research. 2002; 12(4):656–664. https://doi.org/10.1101/gr.229202 PMID: 11932250

102. Hoffman GE. Correcting for population structure and kinship using the linear mixed model: theory and extensions. PLoS One. 2013; 8(10):e75707. https://doi.org/10.1371/journal.pone.0075707 PMID: 24204578

103. Balding DJ. A tutorial on statistical methods for population association studies. Nature Reviews Genetics. 2006; 7(10):781–791. https://doi.org/10.1038/nrg1916 PMID: 16983374

104. Fairfax BP, Humburg P, Makino S, Naranbhai V, Wong D, Lau E, et al. Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. Science. 2014; 343(6175): 1246949. https://doi.org/10.1126/science.1246949 PMID: 24604202

105. Jo B, He Y, Strober BJ, Parsana P, Aguet F, Brown AA, et al. Distant regulatory effects of genetic variation in multiple human tissues. bioRxiv. 2016; p. 074419.

106. Zhao S, Gao C, Mukherjee S, Engelhardt BE. Bayesian group factor analysis with structured sparsity. Journal of Machine Learning Research. 2016; 17(196):1–47.

107. Hore V, Viñuela A, Buil A, Knight J, McCarthy MI, Small K, et al. Tensor decomposition for multiple-tissue gene expression experiments. Nature Genetics. 2016; 48(9):1094–1100. https://doi.org/10.1038/ng.3624 PMID: 27479908