Adjusting for principal components of molecular phenotypes induces replicating false positives

Andy Dahl
Department of Medicine, University of California San Francisco
Vincent Guillemot
Centre de Bioinformatique, Biostatistique et Biologie Intgrative, Institut Pasteur
Joel Mefford
Department of Medicine, University of California San Francisco
Hugues Aschard
Centre de Bioinformatique, Biostatistique et Biologie Intgrative, Institut Pasteur
and Department of Epidemiology, Harvard TH Chan School of Public Health
and
Noah Zaitlen
Department of Medicine, University of California San Francisco

March 26, 2017

Abstract

High-throughput measurements of molecular phenotypes provide an unprecedented opportunity to model cellular processes and their impact on disease. Such highly-structured data is strongly confounded, and principal components and their variants reliably estimate latent confounders. Conditioning on PCs in downstream analyses is known to improve power and reduce multiple-testing miscalibration and is an

indispensable element of thousands of published functional genomic analyses. Further clarifying this approach is of fundamental interest to the genomics and statistics communities. We uncover a novel bias induced by PC conditioning and provide an analytic, deterministic and intuitive approximation. The bias exists because PCs are, roughly, unshielded colliders on a causal path: because PCs partially incorporate a causal genotype effect on one phenotype, the genotype becomes correlated with every phenotype conditional on PCs. We empirically quantify this bias in realistic simulations. For small genetic effects, a nearly negligible bias is observed for all tested PC variants. For large genetic effects, or other differential covariates, dramatic false positives can arise. Though one PC variant (supervised SVA) largely avoids this bias, it is computationally prohibitive genome-wide; further, its immunity to this bias is novel. Our analysis informs best practices for confounder correction in genomic studies.

Keywords: Confounding; Eigenvector perturbation; Quantitative trait loci; Molecular phenotype

1 Introduction

Association studies of molecular phenotypes have helped uncover genetic regulatory programs underlying a variety of processes including transcription, methylation, chromatin accessibility, translation, ribosomal occupancy, and response to cellular stress. These functional quantitative trait loci (*QTL: eQTL [30, 32], mQTL [33, 29, 7], caQTL [13, 22], pQTL [2], rQTL [8], and reQTL [24, 14]) studies test the associations between genetic variants and molecular phenotypes in a cohort of individuals. For parsimony, we refer only to gene expression phenotypes going forward; however, we emphasize our theoretical arguments and much of our simulations apply to all molecular phenotypes, as well as to other domains with high-dimensional and highly structured data.

These genetic association tests are both performed in local genomic windows (called *cis*) and genome-wide (called *trans*). In *cis*, genetic signals are stronger and simpler, often involving disruption to the physical process of transcription. In *trans*, by contrast, genetic associations are presumably mediated by some complex cellular regulatory process, and so *trans* associations are both biologically central and leave only a subtle and weak signal.

Molecular phenotypes are highly sensitive to structured environmental noise, and this confounding both reduces power and skews the joint distribution of p-values across genes [17, 26]. To partially address these shortcomings, known confounders–such as batch effects–are typically included as covariates in *QTL analyses. But unmeasured confounders–such as subtle experimental variations or cell cycle state–or entirely unexpected sources of confounding often have substantial effect. Fortunately, these confounders intuitively introduce large but low-dimensional variation in the high-dimensional molecular phenotype measurements, and PCs and their variants (which we collectively call CCs, for confounding components) often accurately estimate these unknown factors in practice [26, 36, 31]. As

CCs estimate unknown confounders, it is natural to include them as covariates alongside known confounders, and this approach has been shown to decrease power and attenuate joint p-value miscalibration in a wide variety of settings. Domain-specific CCs, like surrogate variables (SVs) [26] or PEER factors [36], have been shown to outperform PCs in eQTL studies, and these methods have become an essential element of thousands of *QTL analysis pipelines [25, 37, 11, 1, 38, 12].

Acknowledging the substantial benefits of CCs, we seek to explore their adverse impact on the false positive rate. The key observation is that CCs are constructed from gene expression measurements that are themselves partially determined by genetic variants and so, inevitably, some causal genetic signal will be captured by CCs. Thus conditioning on CCs is analogous to conditioning on an unshielded collider in a directed graphical model (Figure 1), which generally induces spurious correlations [16, 34]. This unshielded collider bias has become increasingly relevant in modern genetic studies [39, 18, 5, 4]. We show that the bias created by conditioning on PCs results in tests that are misspecified even marginally, are asymptotically inconsistent, and replicate out-of-sample; in contrast, previous discussion of p-value miscalibration focused only on the joint distribution of null p-values, which are not independent in the presence of confounders [26].

The PC conditioning bias can be formalized for small genetic effects using standard eigenvector perturbation theory. The approximation to the (suitably defined) bias for testing the p-th gene conditional on the first phenotypic PC takes an extremely simple form:

$$\text{Bias} \approx -aV_{p1}V_{q1}$$

where a is the causal effect size, q is the gene that is truly causally affected, and V are the right singular vectors of the molecular phenotype matrix. Tighter approximations are given

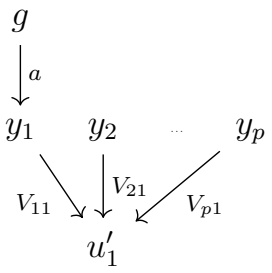

Figure 1: Graphical illustration of the novel bias problem induced by conditioning on CCs. g is the genotype affecting one gene expression measurement (y_1). u'_1 is the first expression PC, which is determined by all y and, indirectly, g. Conditioning on u'_1 will induce spurious correlation between g and all expression measures y_j for $j \neq 1$.

below, but the form of this approximation suggests taking the graphical representation in Figure 1 somewhat seriously, as the approximate bias coincides with the bias derived from the graph. A similar, but far smaller, bias can be shown to arise when conditioning on genetic PCs in GWAS, and we suspect similar results may be applicable in other fields.

We then study the practical relevance of this bias using a range of simulations and CC approaches. First, we find that no tested method can correct for this bias even in the extremely simplistic scenario where expression measurements are i.i.d. Gaussian with a small added *cis*-eQTL; this includes PCA, SVA, PEER and linear mixed model approaches [21]. In particular, we show that supervising the CC decompositions with the genotype g does not eliminate the bias. We theoretically support this claim for a naive approach that incorporates g into PCA by performing PCA on regression residuals. In more complex data with strong confounding, this bias may be drowned out by joint p-value miscalibration and can easily be missed.

In a different regime, where a *trans*-eQTL strongly affects many genes, the bias problem

can be massive, even relative to the value added by CC methods. We show this using simulated genotypes and real gene expression data from the GEUVADIS consortium [23] to simultaneously obtain realistic phenotype properties and known false/true positive patterns of genetic association. We show that supervised variants of CC methods, primarily used to improve power, largely solve this bias problem. Identical results hold when the *trans*-eQTL is instead any other known covariate of primary interest; in these differential expression studies, large, transcriptome-wide effects are common. One particularly important feature we identify is that the false positive rate is concave in the number of conditioning CCs, which severely undermines the near-ubiquitous practice of choosing this number to maximize the total number of positive tests.

2 The bias due to PC conditioning

This section uses a stylized model of a causal genetic effect added to a background matrix of gene expression values. We imagine a causal genotype vector $g \in \mathbb{R}^{N\times 1}$ has been measured on N individuals and we have measured some baseline gene expression matrix $Y \in \mathbb{R}^{N\times P}$ at P genes and the same N individuals. Our observed gene expression matrix $Y' \in \mathbb{R}^{N\times P}$ is created by adding a small linear effect of g on Y:

$$Y' = Y + g\alpha \tag{1}$$

Y is considered deterministic but, in this section, fully general, while g and the causal effect size $\alpha \in \mathbb{R}^{1\times P}$ are random. We assume g is spherical Gaussian and, for simplicity, that it causally affects only gene q. We write this as $\alpha = ae_q$, where a is the causal effect on gene q and e_q is the vector of all zeros except a one in the q-th coordinate.

Let u'_j (respectively, u_j) be the j-th eigenvector of Y' (respectively, Y), and let $\hat{\alpha}'_p$ be the regression coefficient on g in the regression of $Y'_{,p}$ on g and u'_1, where $Y'_{,p}$ denotes the

p-th column of Y' (and define $\hat{\alpha}_p$ analagously). We aim to assess the impact of conditioning on u_1' in a regression on g. If u_1 were used instead, estimates of α will be unbiased as g is independent of Y.

More specifically, our goal is to quantify the bias induced from g's influence on u_1':

$$\text{Bias} := \hat{\alpha'}_p - \hat{\alpha}_p$$

For all $p \neq q$, $Y'_{,p} = Y_{,p}$, and so the discrepancy between these regressions arises only due to g's effect on u_1'.

We assume that α is small so that we can use a standard approximation to the perturbed eigenvector (e.g. Section II of [3]): for a generic small perturbation E, the first eigenvector of $E + Y^T Y$ is approximately

$$u_1' \approx u_1 + \sum_{j=2}^{N} \frac{u_1^T E u_j}{d_1^2 - d_j^2} u_j \tag{2}$$

where d_j is the j-th singular value of $\frac{1}{\sqrt{P}}Y$. In our case, $E = a y_q g^T + a g y_q^T + a^2 g g^T$, using $y_q := Y_{,q}$ as shorthand for the q-th column of Y. Plugging this in to the perturbation approximation gives

$$\begin{aligned} u_1' &\approx u_1 + \sum_{j>1} \frac{1}{d_1^2 - d_j^2}(a(y_q^T u_1)(g^T u_j) + a(y_q^T u_j)(g^T u_1) + a^2(g^T u_1)(g^T u_j))u_j \\ &\approx u_1 + a\sum_{j>1} c_j(z_{1q}x_j + z_{jq}x_1)u_j \end{aligned}$$

The second line drops $O(a^2)$ terms and uses the following simplifying definitions based on the SVD $\frac{1}{\sqrt{P}}Y = UDV^T$:

$$\begin{aligned} Z &:= U^T Y \\ x &:= U^T g \\ c_j &:= \frac{1}{d_1^2 - d_j^2} \end{aligned}$$

Note that x is still a spherical Gaussian random variable.

As an aside, the perturbation approximation is first order in E which, itself, depends on a^2. However, a proper second order expansion in a would also require incorporating $O(a^2)$ terms from a second-order expansion in E, hence we drop $O(a^2)$ terms from our approximations.

Before turning to the regression estimates, we evaluate a few helpful terms:

$$\begin{aligned} g^T u_1' &\approx x_1 + a\sum_{j>1} c_j x_j (z_{1q}x_j + z_{jq}x_1) \\ y_p^T u_1' &\approx z_{1p} + a\sum_{j>1} c_j z_{jp} (z_{1q}x_j + z_{jq}x_1) \\ g^T u_1' y_p^T u_1' &\approx x_1 z_{1p} + a\sum_{j>1} c_j (x_1^2 z_{jp}z_{jq} + x_j^2 z_{1p}z_{1q} + x_1 x_j (z_{jp}z_{1q} + z_{1p}z_{jq})) \end{aligned}$$

These terms are useful in expanding the bias, which we do using standard two-stage least squares expressions for the regression coefficients:

$$\begin{aligned} \hat{\alpha'}_p &= \frac{y_p^T g - y_p^T u_1' g^T u_1'}{N - (g^T u_1')^2} \\ &= \frac{(y_p^T g - y_p^T u_1 g^T u_1) + (y_p^T u_1 g^T u_1 - y_p^T u_1' g^T u_1')}{N - (g^T u_1')^2} \\ &\approx \frac{N - (g^T u_1)^2}{N - (g^T u_1')^2} \left(\hat{\alpha}_p - \frac{a\sum_{j>1} c_j (x_1^2 z_{jp}z_{jq} + x_j^2 z_{1p}z_{1q} + x_1 x_j (z_{jp}z_{1q} + z_{1p}z_{jq}))}{N - (g^T u_1)^2} \right) \\ &\approx \hat{\alpha}_p (1+\gamma) - \frac{a(1+\gamma)}{N - (g^T u_1)^2} \left(x_1^2 \sum_{j>1} c_j z_{jq}z_{jp} + x_1 \sum_{j>1} c_j x_j (z_{jq}z_{1p} + z_{1q}z_{jp}) + z_{1q}z_{1p} \sum_{j>1} c_j x_j^2 \right) \end{aligned} \tag{3}$$

where the approximations are correct to first order in a. The last line introduced γ, which is defined as

$$\gamma := \frac{(g^T u_1')^2 - (g^T u_1)^2}{N - (g^T u_1')}$$

This quantity is written in simple terms and shown to be negligible in the Appendix and ignored going forward.

The bias then becomes

$$\hat{\alpha'}_p - \hat{\alpha}_p \approx -\frac{a}{N-(g^T u_1)^2}\left(x_1^2 \sum_{j>1} c_j z_{jq} z_{jp} + x_1 \sum_{j>1} c_j x_j (z_{jq} z_{1p} + z_{1q} z_{jp}) + z_{1q} z_{1p} \sum_{j>1} c_j x_j^2\right)$$

We now drop middle term inside the parentheses that sums terms proportional to $x_1 x_j$. These summands are each products of normal random variables with mean zero and only have standard deviation $c_j(z_{jq} z_{1p} + z_{1q} z_{jp})$. In contrast, the summands in the third term in the parentheses have mean $c_j z_{1p} z_{1q}$ and variance $2c_j^2 z_{1p}^2 z_{q1}^2$. By the central limit theorem, for large N the comparison simplifies to comparing a $\mathcal{N}\left(0, \sum_{j>1} c_j^2 \frac{(z_{jq} z_{1p} + z_{1q} z_{jp})^2}{z_{1q}^2 z_{1p}^2}\right)$ to a $\mathcal{N}\left(\sum_{j>1} c_j, \sum_{j>1} c_j^2\right)$. We say the former is negligible because its standard deviation is smaller than the mean of the latter:

$$\sqrt{\sum_{j>1} c_j^2 \frac{(z_{jq} z_{1p} + z_{1q} z_{jp})^2}{z_{1q}^2 z_{1p}^2}} \approx \|c_{-1}\|_2 \leq \|c_{-1}\|_1$$

The above approximation assumes $\frac{z_{jq} z_{1p} + z_{1q} z_{jp}}{z_{1q} z_{1p}} = \frac{d_j}{d_1} \frac{V_{jq} V_{1p} + V_{1q} V_{jp}}{V_{1q} V_{1p}} \leq 1$, which is reasonable so long as V is not too sparse. The inequality is fully general and in our case holds loosely: in the GEUVADIS data (described below) and Marchenko-Pastur spectra with the same aspect ratio, $\frac{\|c_{-1}\|_1}{\|c_{-1}\|_2}$ is 19.3 and 9.1, respectively.

Dropping the middle summand simplifies the expression considerably, giving

$$\begin{aligned}\hat{\alpha'}_p - \hat{\alpha}_p &\approx -\frac{a}{N-(g^T u_1)^2}\left(x_1^2 \sum_{j>1} c_j z_{jq} z_{jp} + z_{1q} z_{1p} \sum_{j>1} c_j x_j^2\right) \\ &\approx -2a\bar{c} \sum_{j=1}^{N} z_{jp} z_{jq} w_j^x \qquad (4)\end{aligned}$$

The last line defines the weights w on each PC and the overall magnitude term $\bar{c}$ by

$$w_1^x := \frac{\sum_{k>1} c_k x_k^2}{2(N-1)\bar{c}}; \quad w_j^x := \frac{c_j x_1^2}{2(N-1)\bar{c}}$$
$$\bar{c} := \frac{1}{N-1}\sum_{j>1} c_j \approx \frac{1}{N-(g^T u_1)^2}\sum_{j>1} c_j$$

The approximation for $\bar{c}$ uses

$$\frac{1}{N-(g^T u_1)^2} \approx \frac{1}{N-1}\left(1+\frac{(g^T u_1)^2-1}{N-1}\right) \approx \frac{1}{N-1}$$

which is correct to first order in the random variable $\frac{(g^T u_1)^2-1}{N-1}$, which has mean zero and variance $\frac{2}{N-1}$.

The w^x partition the perturbation effect among PCs. They are random and depend on the correlation between the random genotype g and each PC (i.e. x). They are nonnegative and, while they do not sum to one, they do sum to one in expectation. $\bar{c}$, on the other hand, is deterministic and depends only on the spectrum of the baseline expression matrix Y. $\bar{c}$ is a condition number and measures the susceptibility of the first eigenvector to perturbations, which is a natural scale factor for the bias induced by perturbing the top PC.

Because of these properties of w^x, the bias in (4) is a weighted correlation between the projections of p and q–the tested and the causal genes–onto the eigen-axes, i.e.

$$\hat{\alpha}'_p - \hat{\alpha}_p \approx -2a\bar{c}\rho^{w^x}(z_p, z_q) \tag{5}$$

where ρ^π is the correlation between vectors weighting enries by π. In particular, ρ^{1_N} gives the ordinary correlation between genes p and q (with or without rotating to eigen-axes). In contrast, ρ^{w^x} randomly weights the eigen-axes, but with far greatest weight on component 1 and successively less weight to subsequent axes. The very large weight

on component 1 is natural given the regression we consider conditions only on the first component. The successively lower weight on subsequent components is also expected, as perturbation theory argues that eigenvectors with distant eigenvalues are unlikely to interact upon subtle system modifications.

ρ^{w^x} is the only remaining randomness in the system because nothing else depends on g. This means that, to first order in a, choice of g affects the bias for testing gene p only by defining the relevant notion of similarity between gene p and the causally affected gene q. In expectation over g,

$$\mathbb{E}\left(\hat{\alpha}'_p - \hat{\alpha}_p\right) = -2a\bar{c}\rho^w(z_p, z_q)$$
$$w_i := \mathbb{E}\left(w_i^x\right) \implies w_1 = \frac{1}{2};\ w_j = \frac{c_j}{2(N-1)\bar{c}}$$

In general, w_1^x should be very well approximated by its expectation because it is an average over $N-1$ variables. The other entries of w^x, however, scale with only one (common) random variable, x_1^2; fortunately, this excess variance is counteracted by the fact that the w_{-1}^x are small (they are collectively as large as w_1^x). In the GEUVADIS gene expression data (described below), $w_1 = \frac{1}{2}$, while $w_2 = 0.003$. In Marchenko-Pastur data with (asymptotic) aspect ratio equal to the GEUVADIS aspect ratio, however, $w_1 = \frac{1}{2}$ and $w_2 = 0.022$, confirming that this approximation is better for increasing data confounding.

A final approximation can be made by dropping w_j terms because $w_1 \gg w_j$ (ignoring the fact that, together, the w_j are as large as w_1):

$$\begin{aligned}\hat{\alpha}'_p - \hat{\alpha}_p &\approx -2w_1 a\bar{c}z_{1p}z_{1q} \\ &= -a\bar{c}d_1^2 V_{p1}V_{q1} \\ &\approx -aV_{p1}V_{q1} \end{aligned} \tag{6}$$

The final line uses the relatively loose approximation that $\bar{c} \approx \frac{1}{d_1^2}$, which is derived in equation (7) in the Appendix.

While $\hat{\alpha}$ is biased conditional on α, it is worth noting that this conditional bias itself has mean zero after averaging over α. Equivalently, when $\alpha = ae_q$, the bias is zero after averaging over q because $V_{,1}$ is mean zero For the same reason, the bias is zero on average over p. There is no transcriptome-wide average bias or expected bias without knowing the causal gene. Nonetheless, we feel our definition of bias that is conditional on p and q properly conveys the fact that p and q are not meaningless indices but biologically determined and replicable out-of-sample.

We have not attempted to generalize the bias calculation to conditioning on more than one PC. However, we suspect an analogous result will hold after modifying w, presumably retaining the properties that entries decrease and that the first K entries are qualitatively larger than the rest. If this is correct, the weighted correlation will increasingly look like the ordinary correlation as K grows. This suggests that the (ordinary) correlation between causal and tested traits is a good intuitive proxy for the extent (and direction) of the regression bias.

2.1 Accuracy of the approximations

To empirically assess the quality of the above approximations, we used a prominent and high-quality RNA-sequencing dataset from the GEUVADIS consortium [23]. We first obtained the raw transcript reads from the European individuals in the GEUVADIS consortium. These were then mapped to gene transcripts by aligning the raw reads to the reference hg19 transcriptome using RSEM [28]. We then removed perfectly correlated genes and quantile-normalized genes to standard normal. The resulting matrix has $N = 375$ samples (rows) and $P = 13,120$ genes (columns). This matrix of gene expression values was used as the baseline expression matrix Y both for the simulations in this section and the *trans* simulations described in section 4.

Given the baseline Y matrix, we simulated 1,000 independent datasets from the model in (1) with $g \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ and $\alpha = ae_q$, where the causal gene q was chosen uniformly at random from $\{1,\ldots,P\}$ and the effect size a was varied. For each dataset, the bias $\hat{\alpha}_p - \hat{\alpha}'_p$ is then computed for each gene $p \neq q$, along with our theoretical approximations to this bias. Finally, the observed biases are regressed on the theoretical biases for each dataset, and we store the regression coefficient and R^2 values.

For both the "full" approximation given in (3) and the simplest approximation given in (6), the median R^2 is greater than .9999 for all a. The empirical distribution of these regression coefficients is shown in Figure 2. The median coefficients are always negligibly far from 1, though the empirical 95% confidence intervals are nontrivial. These observations suggest our bias approximation is off by a scale factor near one, possibly because higher-order terms also predominantly scale in $V_{p1}V_{q1}$. We have not pursued more accurate approximations as the goal is only to demonstrate the existence and qualitative behavior of the bias; moreover, where the perturbation is truly small, we show in simulations below that the resulting bias is essentially negligible.

Overall, the final approximation given in (6) appears to be a very good estimator, despite depending only on a, q and $V_{,1}$ (when d_1 is large).

2.2 Inconsistency and spurious replication

As the approximations developed above depend on the data only through the top right singular vector when $d_1 \gg d_2$, datasets with similar $V_{,1}$ will have similar regression bias. This is because a and q are determined by nature.

One implication is that as N grows large, the bias stays constant but the regression standard error shrinks, resulting asymptotically in rejecting the null hypothesis transciptome-wide for any g causally affecting even one gene. This argument implicitly assumes the

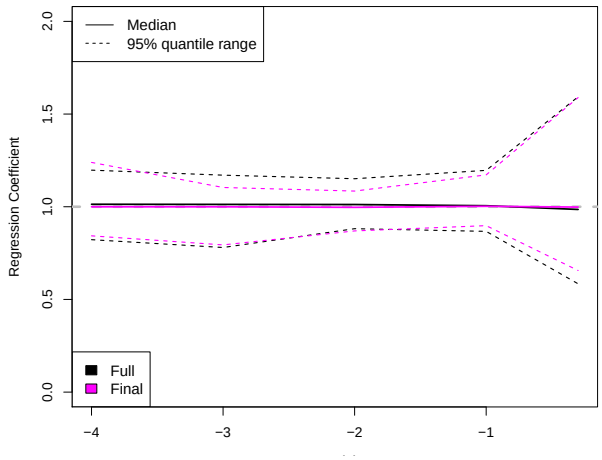

Figure 2: Results from regressing the observed bias in $\hat{\alpha}$ on two novel approximations to the bias. The "final" approximation simplifies several negligible terms in the "full" approximation. The distribution of the regression coefficient is shown, which is identically 1 for perfect approximations.

large-N limit for $V_{,1}$ (or at least its distribution) is defined and non-sparse (respectively, a.s. non-sparse). Though we avoid formalizing this large-N limit, it seems reasonable that if one large confounder is present and that N grows large, then $V_{1,}$ will converge to the confounder and d_{-1} will only shrink.

A related implication is that the regression bias will replicate in a new dataset with the same top right singular vectors. There is no requirement on the two sets of left singular vectors to obtain this spurious replication. Again avoiding formality, the top right singular vectors will be similar in datasets with similar confounding structure, and such shared structure is common in molecular phenotypic assays. Further, only sign consistency is assessed in genomic replication analyses, which reduces the burden to prove false positive

replication: the datasets need not have equal $V_{p1}V_{q1}$ terms, only sign-consistency in these terms.

Finally, we note that if the causal signal due to g is sufficiently strong, it will dominate leading right singular vectors, which intuitively satisfies the necessary condition for bias replication without any other assumptions on shared confounding across-datasets.

2.3 PCA on residuals

An apparent solution to this problem of PC conditioning is to project out g before computing PCs. We call this approach supervised PCA (though it is unrelated to [6] and other published definitions). This approach has been studied before through simulations that show supervised SVA provides superior joint p-value recalibration in the presence of confounders [26].

To prove theoretically the flaw in this approach, even for marginal tests when g has no causal effect, we first make an assumption: that in realistic data the supervised PCs approximate an approach that projects out g from the ordinary PCs, which we call residual PCA. In the GEUVADIS data, described below, we find that the top residual and supervised PC pairs each have squared correlation greater than .99 for independently simulated i.i.d. Gaussian genotypes, suggesting our assumption is reasonable. However, we have not investigated this issue theoretically.

It is easy to prove that conditioning on the residualized PCs spuriously inflates the t-statistic for g. First, by construction, g and the residualized PCs are orthogonal, hence the regression coefficient for g is unmodified by conditioning on residualized PCs. The same is true for g's regression coefficient's standard error, except that the overall regression error $\hat{\sigma}^2$ may change. In particular, as PCs (or residual PCs) explain significant variation in y_p, $\hat{\sigma}^2$ will decrease after conditioning, thus inflating the t statistic for g's effect.

The change in g's regression statistics can be directly computed. Let U be the first k left singular vectors of Y and assume columns of Y are mean zero and variance 1. We consider the regression equation

$$y_p = U\beta + \epsilon$$

Conditional on only the first k PCs, entries of ϵ have mean 0 and variance $\sigma^2 := 1 - \frac{1}{N}\|U\beta\|^2$; we also assume entries of ϵ are uncorrelated for illustrative purposes. Now let $\hat{\sigma}^2$ be the ordinary regression estimate of σ^2 conditional on g and U. If g is normalized to length 1, then

$$\begin{aligned}
\mathbb{E}\left(\hat{\sigma}^2\right) &= \frac{1}{N-(k+1)}\mathbb{E}\left(\|(I - \left(gg^T + UU^T\right))y_p\|^2\right) \\
&= \frac{1}{N-(k+1)}\mathrm{tr}\left(\left(\sigma^2 I + U\beta\beta^T U^T\right)\left(I - gg^T - UU^T\right)\right) \\
&= \sigma^2 - \frac{1}{N-(k+1)}(g^T U\beta)^2
\end{aligned}$$

We ignore the second term, which is $O(N^{-2})$. As g is assumed independent of Y, the expected value for $\hat{\sigma}^2$ when regressing solely on g is just $\mathbb{V}\left(y_p\right) = 1$. Therefore, conditioning on residual PCs simply inflates g's t statistic by roughly a factor of $1/\sigma$.

More broadly, this suggests that confounder estimates must strike a delicate balance. Previous sections show how excess correlation with g can cause bias, and previous work has shown similar excess correlation leads to loss in power [36]. However, it is now also clear that excess uncorrelation–here, in the form of residualizing g from PCs–can lead to different bias problems. We refer to this balance in later sections when discussing the relative performance of different methods, each of which learns different forms of relationships between g and estimated confounders.

3 *cis*-eQTL simulations with white noise

Having shown that conditioning on expression PCs induces a replicating bias for genetic association tests, we turn now to the practical significance of the problem. We begin with one of the simplest possible simulations based on equation (1): gene expression measurements in Y are i.i.d. standard normal; g is (indepedently) i.i.d. standard normal; and $\alpha = ae_q$, dictating that g is a *cis*-eQTL affecting only gene q; finally, q is drawn, independently of g and Y, uniformly from $\{1, \ldots, P\}$, and the effect size a is varied. The dimensions of Y, $N = 375$ and $P = 13,120$, were chosen to match the dimensions of the GEUVADIS data. More realistic simulations are presented in the next section; the focus here is on demonstrating the existence of conditioning bias, which is made dramatically easier by eliminating other sources of null miscalibration.

For each simulated dataset, we perform a 1-sided Kolmogorov-Smirnov (KS) test for deflation in the null regression p-values (for regressing g on genes in Y' other than q). We also include the (2-sided) KS test for these KS p-values in the plots; this is the "nested" or "double" KS test previously used to detect null p-value miscalibration [26, 27], except we use one 1-sided KS test as the bias we define is expected to decrease null p-values.

After simulating Y' and g, we test the association between the genotype and each column of Y'–other than the causal gene q–with ordinary linear regression conditioned on various CCs. The first set are the PCs of Y'; the second set are PEER factors [36]; and the third set are surrogate variables [26]. In this simulation setting, however, SVA declares 0 components significant (through internal permutation tests) and stops–this is arguably ideal behavior. (Also, when no covariate was included, we were unable to implement the "irw" algorithm, which is the suggested default and was used for all of the supervised analyses, discussed below.) We also regress g on genes in Y' using linear mixed models

that include a random effect with covariance $Y'Y'^T$, as implemented in ICE [21].

We also consider CCs that are computed with knowledge of the tested genotype g. Both SVA and PEER have such supervised options to include a covariate alongside the decomposition. We also compare the supervised PCA approach, discussed in section 2.3, that projects g out of Y' before PCs are computed. The primary motivation for supervising these decompositions is to avoid factors that are too correlated with g, absorbing the signal and attenuating power [26, 36]. A secondary, and countervailing, goal in supervising these decompositions is to avoid factors that are too uncorrelated with g; this balance is described in [26] and in more detail in section 2.3.

Figure 3 presents the quantile-quantile plots for the resulting KS p-values. Unsurprisingly, "None"–including no CCs of any kind–delivers well-calibrated p-values. Also unsurprisngly, the PC approach (for $K = 1$ and $K = 20$) causes noticeable bias for the high-heritability simulation in blue, where the genotype explains 90% of the gene's variation (equivalently, $a = 9$). However, no statistically significant problem is detected for smaller a, suggesting the induced bias from PC conditioning can often be negligible. PEER results are also essentially unperturbed for small a, but give highly conservative p-values for large a. The linear mixed model (LMM) approach–using either REML or ML (not shown)–yields highly conservative p-values.

A different type of bias also appears qualitatively shared among the three supervised methods. The bias grows with K (even beyond 20, not shown) and appears independent of the causal effect a, persisting even for the global null where $a = 0$. In the case of supervised PCA, this is provably due to supervised PCs being too uncorrelated with the genotype, as noted above. For supervised PEER and SVA, qualitatively similar but attenuated biases are present, suggesting these factors strike a better balance between explaining too much and too little of the signal in g. SVA seemed less susceptible to this bias. Some variants

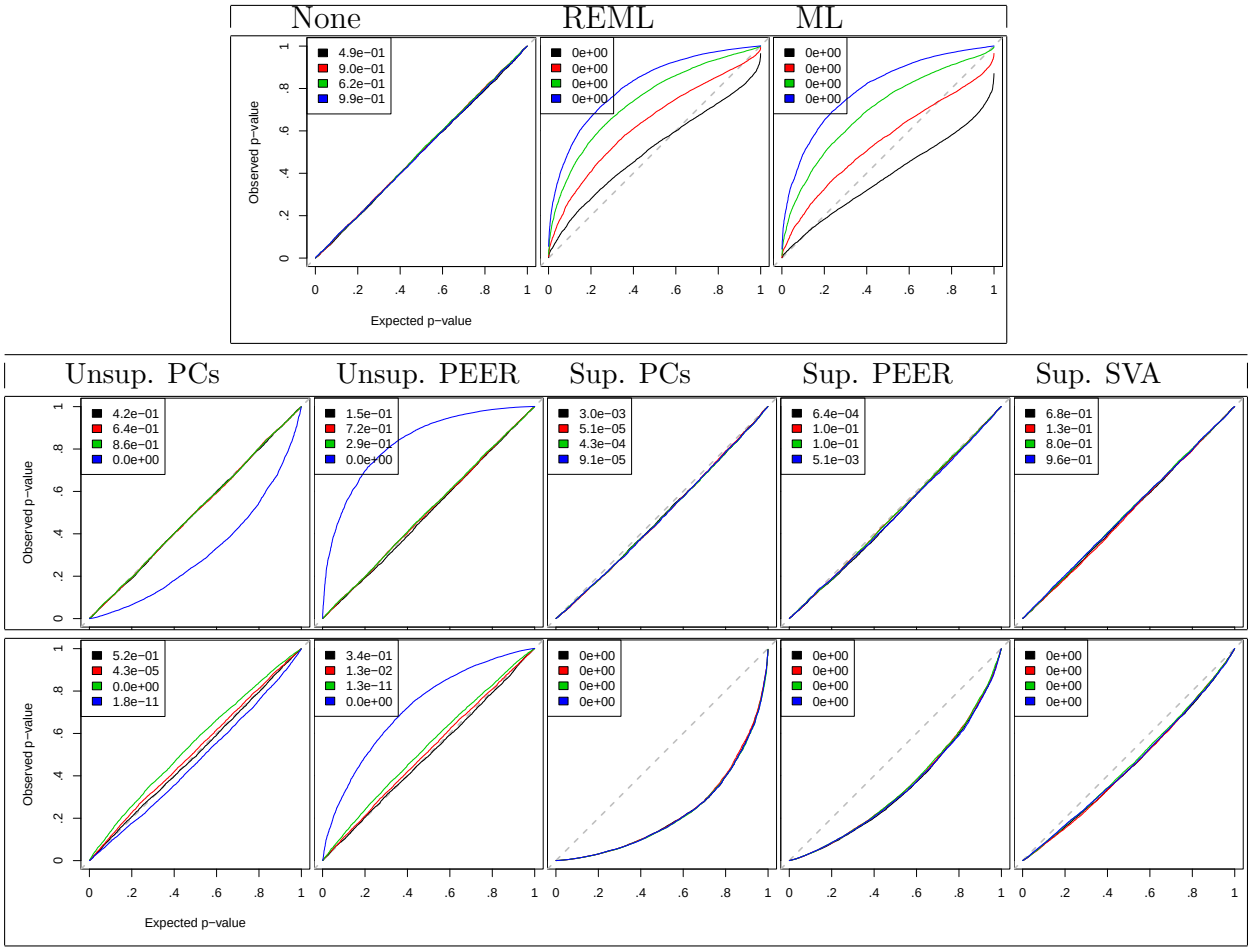

Figure 3: p-value miscalibration for a variety of confounder corrections applied to simulated datasets with a weak genetic effect added to Gaussian white noise. QQ plots for KS tests applied to nominally null p-values are presented, aggregated over 5,000 simulated datasets. KS tests of these KS p-values are presented in the legends. Colors index the strength of the *cis*-eQTL: black corresponds to 0% variation explained by the genotype in the causal gene (the global null); red to 30%; green to 60%; and blue to 90%. The bottom set uses $K = 1$ CC on the top row, $K = 20$ CCs on bottom; the top set uses methods with no K.

of PEER will likely perform better, perhaps including the joint modelling approach in iVBQTL [36] or random effect approach in PANAMA [15]; however, we did not investigate these approaches, which for computational reasons have not been applied to genome-wide scale human data.

Finally, we observed that as K grows larger, unsupervised PC adjustment becomes slightly, but significantly, conservative. This effect is only detectable when $a > 0$, i.e. when g is truly causal. Surprisingly, the conservativeness it is non-monotonic in a: for $K = 20$, 60% PVE (green) gives conservative p-values while 90% (blue) PVE gives anticonservative p-values. This is consistent with some conservative force that emerges for $K > 1$ and scales at a faster rate in a than the countervailing anticonservative force we identified theoretically and empirically for $K = 1$. Unsupervised PEER has a related problem, though its conservativeness simply increases in a.

Overall, Figure 3 shows that no tested CC method can correct p-value miscalibration in even an extremely simple setting. However, it also shows this problem is very small, especially for small causal effect sizes.

4 *trans*-eQTL simulations with real phenotypes

We now present simulations that extend the above *cis*-eQTL simulations in two ways. First, we modified g to be a *trans*-eQTL, as g now causally explains 30% of the variation in each of 5% of all genes. This means g explains 1.5% of variation transcriptome-wide. Null simulations where g affects nothing are also included for comparison.

Second, we make the simulation considerably more realistic. We replace the i.i.d. standard normal Y matrix by the real GEUVADIS gene expression matrix, described above. Also, as effect sizes are now larger and matching the theoretical bias result is no longer a

direct goal, we adopt the standard practice of normalizing columns of Y' to mean zero and variance one. Finally, we simulated g to be a binomial SNP genotype with minor allele frequency 20% and then normalized it to mean zero and variance one.

Having established p-value miscalibration in the *cis*-eQTL simulation, we now plot empirical false positive rates in Figure 4. Unlike the KS tests, which assess correlation among gene p-values within-dataset, the FPR assesses a type of mean among the gene p-values. This means the KS miscalibration identified in [26] will only manifest as high variance in the FPR between-datasets. We also show $-\log_{10}(p)$-values averaged over the causally affected genes within each simulated dataset to assess power.

The clearest implication from 4 is that the unsupervised approaches deliver highly inflated false positive rates (top, middle panel) and attenuated power (top, right panel). While the LMM approaches we studied are well-calibrated, they offer even lower power than simply regressing on g; however, other LMM approaches may deliver higher power [20]. The obvious conclusion is that one should supervise CCs with all large-effect covariates, roughly yielding well-calibrated FPRs for SVA and slightly inflated FPRs for PEER. However, as is expected from our above theory and is discernible from previous empirical results [26], supervised PCA gives dramatically inflated FPR.

The obvious caveat is that the large-effect covariates must be known in advance. This does not hold when searching for a causal effect among a large set of potential covariates, often outnumbering the sample size, as is standard in genetic studies. While one could refit confounder estimates for each tested g in turn, this is typically computationally infeasible; however, pre-screening loci, perhaps by using unsupervised CCs, may dramatically reduce the computational burden. Moreover, practitioners often fail to include known large effect covariates when creating CCs. Finally, to our knowledge, the observation that this supervision eliminates false positives–rather than merely recalibrating the across-gene

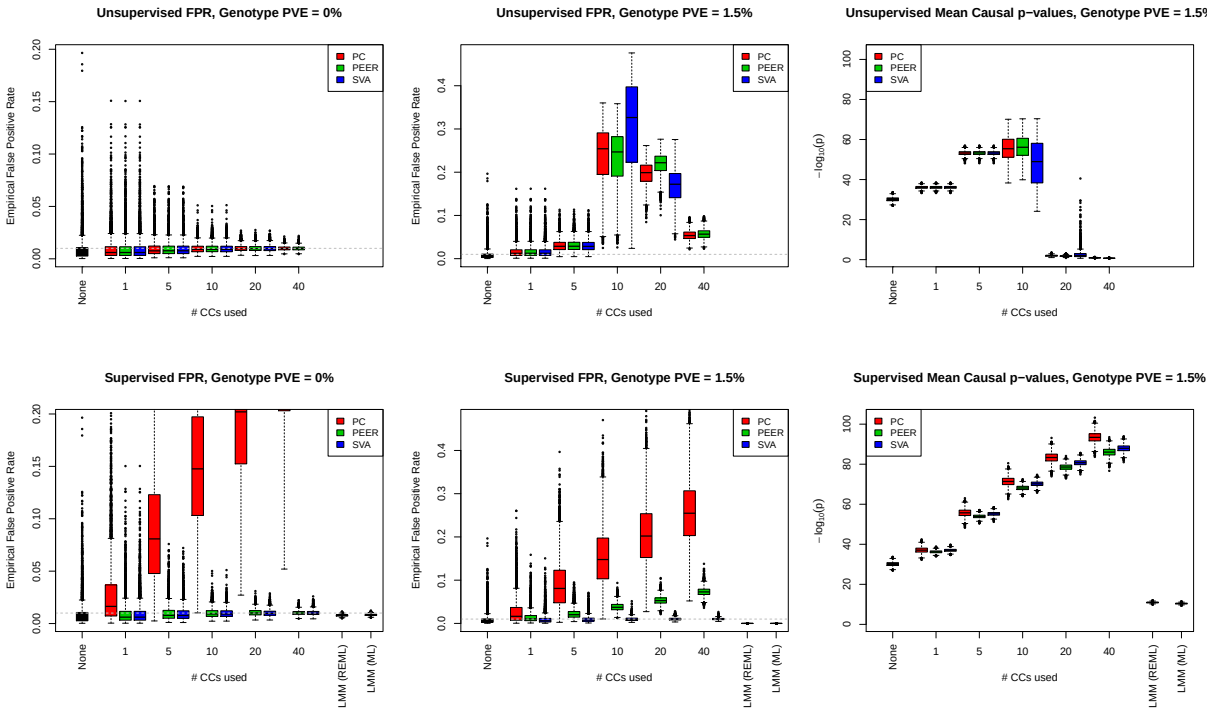

Figure 4: False positive rates and power for a simulated, strong, *trans*-effect added to real gene expression data from GEUVADIS. Results are presented for non-causally affected loci when no (left column) or a large (center column) genetic effect is present; also shown are average $-\log_{10}(p)$-values for the causally affected genes (right column). Methods that are unsupervised (top) or supervised with respect to the causal genotype (bottom) are both tested. As SVA always deems less than 40 components significant, the corresponding results are absent.

independence of p-values [26] or increasing power [26, 36]–is novel.

4.1 False positive replication

We slightly modified the *trans* simulation to assess the out-of-sample replication rate for false positives. For each simulated dataset, generated exactly as in the above *trans*-simulation, we split the data (g and Y') into halves and repeat the above analyses on each half. We compute the replication rate as the fraction of false positive discoveries from the first half that are deemed positive in the second half, using a significance level of $\alpha = .001$ for both splits; we then compute the replication rate after transposing the roles of the two splits. Finally, we average over false replication rates for 5,000 independently simulated datasets and two choices of initial split per dataset (10,000 scenarios in total), discarding scenarios where there are no false discoveries in the initial split.

Results are shown in Table 1 after converting to the replication rate inflation, defined as the ratio of the replication rate to its null expectation (i.e. .001). First, no significant inflation is obtained if confounder corrections are entirely omitted, where significance is defined by the mean inflation being more than 1.65 standard errors greater greater than 1. But both PC and PEER correction substantially inflate the false positive replication rate. For both methods, the using larger K exacerbates problem, though supervision improves PEER's performance and hurts PCA. Unsupervised SVA has inflation problems similar to unsupervised PCA and PEER, but after supervision no statistically significant inflation is detected.

5 Discussion

We have shown a novel source of bias induced by conditioning on estimated confounders in association tests. We derived an analytical approximation in the simple case of conditioning on one PC and a small causal effect. We assessed the magnitude of the bias empirically

	No CCs	1 PC	10 PCs	1 PEER	10 PEER	1 SV	10 SVs
Unsupervised	0.84	2.59*	151.6*	2.70*	148.38*	2.67*	181.15*
Supervised	N/A	32.47*	168.69*	2.02*	8.22*	0.68	1.09

Table 1: False positive replication rate inflation after confounder correction. Shown are the false positive replication rates at a significance level of $\alpha = .001$ divided by their expectation. Inflations significantly greater than 1 are indicated with an asterisk.

through simulations, with a small effect simulation showing the existence of bias and a large effect simulation demonstrating the impact on the FPR in a realistic setting. This bias will replicate in datasets with similar confounding structures, which is often obtained for high-throughput molecular phenotypes.

This bias problem is related to, but distinct from, known biases caused by confounding. Previous theoretical results have focused on confounding's impact on the joint null distribution of p-values, showing that an idealized form of correction–i.e. where confounders are perfectly estimated–yields independent p-values for the tests of association between one genotype and each gene [27]. Our results undermine the basic assumptions of this theory, though, by showing that causal effects naturally bias confounder estimates. This novel bias that we show arises from confounder correcting, unlike known biases from uncorrected confounders, results in tests that are misspecified even marginally and that will replicate in similar datasets.

Overall, we find that the supervised versions of SVA and, to a lesser extent, PEER give well-calibrated FPRs and the greatest power among approaches considered. Though no confounder correction method delivers truly independent tests, this effect is generally small and, in the presence of strong confounders, the biases induced by supervised methods are typically negligible compared to the eliminated biases. Further, to avoid computing

supervised factors for each SNP genome-wide, *QTL analyses using unsupervised factors can be corrected *post-hoc* by repeating the analysis at significant (or suggestive) genotypes with the appropriately supervised decompositions.

Nonetheless, many high-profile studies fail to supervise their SVA or PEER analyses with respect to strong covariates. This is partially due to a common misconception amongst practitioners that SVA and PEER corrections solely improve power and, thus, that suboptimal implementations are at worst conservative.

There are also more complex scenarios where the appropriate model for covariates is unclear. For example, in gene coexpression studies, covariance or partial covariance can be used to learn complex and subtle graphical models [19, 35]: failing to correct for confounding will yield biologically meaningless confounder networks, while a correction approach with subtle biases may induce a different biologically meaningless network. Latent variable graphical models present a possible solution to this problem [9]. More generally, low-rank-plus-sparse decompositions may be able to simultaneously learn latent confounders and eQTLs transcriptome-wide [10].

We note that qualitatively similar replicating false positive associations can arise when conditioning on unshielded colliders in multiphenotype association tests [4]. This can be expressed in our motivating graphical model in Figure 1 as an environmental node e causally affected by several of the phenotype nodes, y_p. As y_p's contribution to a generic e is arbitrary–unlike its contribution to a PC, which is $O(P^{-1/2})$–the replicating false positive problem can be substantially larger than in our context. However, that confounding is entirely dataset-dependent, while our bias is in a sense universal–existing, for example, even when all noise in the y_p is i.i.d. Gaussian. Nonetheless, our false positives will only regularly replicate when the y_p are correlated.

A theoretical question we leave open is whether it is possible to deliver tests that are

truly independent across genes in the presence of confounding. The difficulty is clear, as any errors in the estimated confounders will, upon conditioning, induce a fresh source of confounding (albeit one that is potentially better behaved and smaller). Another difficult question is whether the decomposition methods studied can be supervised with respect to a large set of tested covariates, most of which will be entirely null, in a computationally feasible manner. If possible, efficiently updating SVA decompositions for each new added covariate would largely achieve this.

Appendix: γ

Defining and simplifying γ gives

$$\begin{aligned}
\gamma &:= \frac{N - x_1^2}{N - (g^T u_1')^2} - 1 \\
&= \frac{1}{1 - 2ax_1 \sum_{j>1} c_j x_j (z_{1q} x_j + z_{jq} x_1)(N - x_1^2)^{-1} + O(a^2)} - 1 \\
&= 2ax_1 \sum_{j>1} c_j x_j (z_{1q} x_j + z_{jq} x_1)(N - x_1^2)^{-1} + O(a^2) \\
&= \frac{2a}{N - x_1^2} \left(x_1^2 \sum_{j>1} c_j z_{jq} x_j + x_1 z_{1q} \sum_{j>1} c_j x_j^2 \right) + O(a^2) \\
&\approx 2ax_1 z_{1q} \bar{c}
\end{aligned}$$

The last line uses the approximations that $\frac{N-1}{N-x_1^2} \approx 1$ and replaces the sums over j by their expectations, which is reasonable as

$$\sum_{j>1} c_j x_j^2 \sim \mathcal{N}\left((N-1)\bar{c}, 2 \sum_{j>1} c_j^2 \right) \qquad \text{(by CLT)}$$

$$\sum_{j>1} c_j z_{jq} x_j \sim \mathcal{N}\left(0, \sum_{j>1} c_j^2 z_{jq}^2 \right)$$

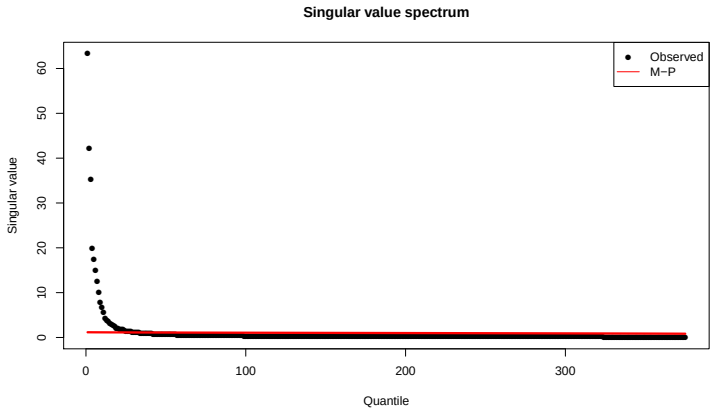

Figure 5: Singular-value spectrum of our process gene expression matrix from GEUVADIS. M-P refers to the Marchenko-Pastur distribution with asymptotic aspect ratio equal matching the GEUVADIS data.

The CLT approximation should be good unless a very small number of eigenvalues (other than the first) are far larger than the rest.

Unfortunately, K eigenvalues will be large compared to all other if there are K strong confounders, so this approximation will be worse for more realistic data. In Figure 5, for example, the top singular value is roughly 1.5 times larger than the second. Presumably a regression conditioning on K PCs would require the last $N - K$ to be small, which would be more reasonable.

We now derive a loose approximation to $\bar{c}$ that is also used in the main text:

$$\begin{aligned}
\bar{c} &:= \frac{1}{N-1}\sum_{j>1}\frac{1}{d_1^2-d_j^2} \approx \frac{1}{(N-1)\,d_1^2}\sum_{j>1}\left(1+\frac{d_j^2}{d_1^2}\right) \\
&= \frac{1}{(N-1)\,d_1^2}\left(N-1+\frac{\left(\sum_{j=1}^N d_j^2\right)-d_1^2}{d_1^2}\right) \\
&= \frac{1}{d_1^2}\left(1+\frac{1}{d_1^2}+\frac{1-d_1^{-2}}{(N-1)\,d_1^2}\right) \\
&\approx \frac{1}{d_1^2} \qquad (7)
\end{aligned}$$

This used the fact that $\sum_{j=1}^N d_j^2 = N$, which holds because columns of Y have been centered and scaled:

$$\sum_{j=1}^N d_j^2 = \mathrm{tr}\left(\frac{1}{P}Y^TY\right) = \frac{1}{P}\sum_{i,p}Y_{ip}^2 = \frac{1}{P}\sum_p \|Y_{\cdot p}\|_2^2 = N$$

Using (7) to simplify $\bar{c}$, γ can then be approximated by

$$\gamma \approx 2ax_1V_{q1}\frac{1+d_1^{-2}}{d_1} \qquad (8)$$

In practice, this term is negligible: a is assumed small, V_{q1} is on the order of $P^{-1/2}$, and d_1 tends to be large in real data, e.g. 8 in GEUVADIS.

References

[1] Frank W Albert and Leonid Kruglyak. The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics*, 16(4):197–212, February 2015.

[2] Frank W Albert, Sebastian Treusch, Arthur H Shockley, Joshua S Bloom, and Leonid Kruglyak. Genetics of single-cell protein abundance variation in large yeast populations. *Nature*, 506(7489):494–497, 2014.

[3] Romain Allez and Jean-Philippe Bouchaud. Eigenvector dynamics: General theory and some applications. *Physical Review E*, 86(4):046202, October 2012.

[4] Hugues Aschard, Bjarni Vilhjalmsson, Chirag Patel, David Skurnik, Jimmy Yu, Brian Wolpin, Peter Kraft, and Noah Zaitlen. Playing Musical Chairs in Big Data to Reveal Variables Associations. *BioRxiv*, 2016.

[5] Hugues Aschard, Bjarni J Vilhjálmsson, Amit D Joshi, Alkes L Price, and Peter Kraft. Adjusting for heritable covariates can bias effect estimates in genome-wide association studies. *American journal of human genetics*, 96(2):329–339, February 2015.

[6] Eric Bair, Trevor Hastie, Debashis Paul, and Robert Tibshirani. Prediction by Supervised Principal Components. *Journal of the American Statistical Association*, 101(473):119–137, March 2006.

[7] Nicholas E Banovich, Xun Lan, Graham McVicker, Bryce van de Geijn, Jacob F Degner, John D Blischak, Julien Roux, Jonathan K Pritchard, and Yoav Gilad. Methylation QTLs Are Associated with Coordinated Changes in Transcription Factor Binding, Histone Modifications, and Gene Expression Levels. *PLoS Genetics*, 10(9):e1004663, September 2014.

[8] A Battle, Z Khan, S H Wang, A Mitrano, M J Ford, J K Pritchard, and Y Gilad. Impact of regulatory variation from RNA to protein. *Science*, 347(6222):664–667, February 2015.

[9] Venkat Chandrasekaran, Pablo A Parrilo, and Alan S Willsky. Latent variable graphical model selection via convex optimization. *The Annals of Statistics*, 40(4):1935–1967, August 2012.

[10] Venkat Chandrasekaran, Sujay Sanghavi, Pablo A Parrilo, and Alan S Willsky. Sparse and low-rank matrix decompositions. In *2009 47th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 962–967. IEEE, 2009.

[11] Carlo Colantuoni, Barbara K Lipska, Tianzhang Ye, Thomas M Hyde, Ran Tao, Jeffrey T Leek, Elizabeth A Colantuoni, Abdel G Elkahloun, Mary M Herman, Daniel R Weinberger, and Joel E Kleinman. Temporal dynamics and genetic control of transcription in the human prefrontal cortex. *Nature*, 478(7370):519–523, October 2011.

[12] The GTEx Consortium, Kristin G Ardlie, David S Deluca, Ayellet V Segre, Timothy J Sullivan, Taylor R Young, Ellen T Gelfand, Casandra A Trowbridge, Julian B Maller, Taru Tukiainen, Monkol Lek, Lucas D Ward, Pouya Kheradpour, Benjamin Iriarte, Yan Meng, Cameron D Palmer, Tõnu Esko, Wendy Winckler, Joel N Hirschhorn, Manolis Kellis, Daniel G MacArthur, Gad Getz, Andrey A Shabalin, Gen Li, Yi-Hui Zhou, Andrew B Nobel, Ivan Rusyn, Fred A Wright, Tuuli Lappalainen, Pedro G Ferreira, Halit Ongen, Manuel A Rivas, Alexis Battle, Sara Mostafavi, Jean Monlong, Michael Sammeth, Marta Mele, Ferran Reverter, Jakob M Goldmann, Daphne Koller, Roderic Guigó, Mark I McCarthy, Emmanouil T Dermitzakis, Eric R Gamazon, Hae Kyung Im, Anuar Konkashbaev, Dan L Nicolae, Nancy J Cox, Timothée Flutre,

Xiaoquan Wen, Matthew Stephens, Jonathan K Pritchard, Zhidong Tu, Bin Zhang, Tao Huang, Quan Long, Luan Lin, Jialiang Yang, Jun Zhu, Jun Liu, Amanda Brown, Bernadette Mestichelli, Denee Tidwell, Edmund Lo, Mike Salvatore, Saboor Shad, Jeffrey A Thomas, John T Lonsdale, Michael T Moser, Bryan M Gillard, Ellen Karasik, Kimberly Ramsey, Christopher Choi, Barbara A Foster, John Syron, Johnell Fleming, Harold Magazine, Rick Hasz, Gary D Walters, Jason P Bridge, Mark Miklos, Susan Sullivan, Laura K Barker, Heather M Traino, Maghboeba Mosavel, Laura A Siminoff, Dana R Valley, Daniel C Rohrer, Scott D Jewell, Philip A Branton, Leslie H Sobin, Mary Barcus, Liqun Qi, Jeffrey McLean, Pushpa Hariharan, Ki Sung Um, Shenpei Wu, David Tabor, Charles Shive, Anna M Smith, Stephen A Buia, Anita H Undale, Karna L Robinson, Nancy Roche, Kimberly M Valentino, Angela Britton, Robin Burges, Debra Bradbury, Kenneth W Hambright, John Seleski, Greg E Korzeniewski, Kenyon Erickson, Yvonne Marcus, Jorge Tejada, Mehran Taherian, Chunrong Lu, Margaret Basile, Deborah C Mash, Simona Volpi, Jeffery P Struewing, Gary F Temple, Joy Boyer, Deborah Colantuoni, Roger Little, Susan Koester, Latarsha J Carithers, Helen M Moore, Ping Guan, Carolyn Compton, Sherilyn J Sawyer, Joanne P Demchok, Jimmie B Vaught, Chana A Rabiner, Nicole C Lockhart, Kristin G Ardlie, Gad Getz, Manolis Kellis, and Simona Volpi. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235):648–660, May 2015.

[13] Jacob F Degner, Athma A Pai, Roger Pique-Regi, Jean-Baptiste Veyrieras, Daniel J Gaffney, Joseph K Pickrell, Sherryl De Leon, Katelyn Michelini, Noah Lewellen, Gregory E Crawford, Matthew Stephens, Yoav Gilad, and Jonathan K Pritchard. DNaseI sensitivity QTLs are a major determinant of human expression variation. *Nature*, 482(7385):390–394, February 2012.

[14] B P Fairfax, P Humburg, S Makino, V Naranbhai, D Wong, E Lau, L Jostins, K Plant, R Andrews, C McGee, and J C Knight. Innate Immune Activity Conditions the Effect of Regulatory Variants upon Monocyte Gene Expression. *Science*, 343(6175):1246949–1246949, March 2014.

[15] Nicolo Fusi, Oliver Stegle, and Neil D Lawrence. Joint modelling of confounding factors and prominent genetic regulators provides increased accuracy in genetical genomics studies. 8(1):e1002330, January 2012.

[16] M H Gail, S Wieand, and S Piantadosi. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika*, 71(3):431–444, 1984.

[17] Greg Gibson. The environmental contribution to gene expression profiles. *Nature Reviews Genetics*, 9(8):575–581, August 2008.

[18] Alexander Gusev, Gaurav Bhatia, Noah Zaitlen, Bjarni J Vilhjálmsson, Dorothée Diogo, Eli A Stahl, Peter K Gregersen, Jane Worthington, Lars Klareskog, Soumya Raychaudhuri, Robert M Plenge, Bogdan Pasaniuc, and Alkes L Price. Quantifying Missing Heritability at Known GWAS Loci. *PLoS Genetics*, 9(12):e1003993–19, December 2013.

[19] Steve Horvath. *Weighted Network Analysis*. Springer New York, New York, NY, 2011.

[20] Jong J Joo, Jae Sul, Buhm Han, Chun Ye, and Eleazar Eskin. Effectively identifying regulatory hotspots while capturing expression heterogeneity in gene expression studies. *Genome biology*, 15(4):r61, 2014.

[21] H M Kang, C Ye, and E Eskin. Accurate Discovery of Expression Quantitative Trait Loci Under Confounding From Spurious and Genuine Regulatory Hotspots. *Genetics*, 180(4):1909–1925, October 2008.

[22] Natsuhiko Kumasaka, Andrew J Knights, and Daniel J Gaffney. Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nature Genetics*, 48(2):206–213, February 2016.

[23] Tuuli Lappalainen, Michael Sammeth, Marc R Friedländer, Peter A C t Hoen, Jean Monlong, Manuel A Rivas, Mar Gonzàlez-Porta, Natalja Kurbatova, Thasso Griebel, Pedro G Ferreira, Matthias Barann, Thomas Wieland, Liliana Greger, Maarten van Iterson, Jonas Almlöf, Paolo Ribeca, Irina Pulyakhina, Daniela Esser, Thomas Giger, Andrew Tikhonov, Marc Sultan, Gabrielle Bertier, Daniel G MacArthur, Monkol Lek, Esther Lizano, Henk P J Buermans, Ismael Padioleau, Thomas Schwarzmayr, Olof Karlberg, Halit Ongen, Helena Kilpinen, Sergi Beltran, Marta Gut, Katja Kahlem, Vyacheslav Amstislavskiy, Oliver Stegle, Matti Pirinen, Stephen B Montgomery, Peter Donnelly, Mark I McCarthy, Paul Flicek, Tim M Strom, The Geuvadis Consortium, Hans Lehrach, Stefan Schreiber, Ralf Sudbrak, Ángel Carracedo, Stylianos E Antonarakis, Robert Häsler, Ann-Christine Syvänen, Gert-Jan van Ommen, Alvis Brazma, Thomas Meitinger, Philip Rosenstiel, Roderic Guigó, Ivo G Gut, Xavier Estivill, and Emmanouil T Dermitzakis. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511, September 2013.

[24] M N Lee, C Ye, A C Villani, T Raj, W Li, T M Eisenhaure, S H Imboywa, P I Chipendo, F A Ran, K Slowikowski, L D Ward, K Raddassi, C McCabe, M H Lee, I Y Frohlich, D A Hafler, M Kellis, S Raychaudhuri, F Zhang, B E Stranger, C O Benoist, P L De Jager, A Regev, and N Hacohen. Common Genetic Variants Modulate

Pathogen-Sensing Responses in Human Dendritic Cells. *Science*, 343(6175):1246980–1246980, March 2014.

[25] Jeffrey T Leek, Robert B Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733–739, October 2010.

[26] Jeffrey T Leek and John D Storey. Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. *PLoS Genetics*, 3(9):e161, 2007.

[27] Jeffrey T Leek and John D Storey. A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences of the United States of America*, 105(48):18718–18723, December 2008.

[28] Bo Li and Colin N Dewey. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12(1):323, 2011.

[29] Yun Liu, Martin J Aryee, Leonid Padyukov, M Daniele Fallin, Espen Hesselberg, Arni Runarsson, Lovisa Reinius, Nathalie Acevedo, Margaret Taub, Marcus Ronninger, Klementy Shchetynsky, Annika Scheynius, Juha Kere, Lars Alfredsson, Lars Klareskog, Tomas J Ekström, and Andrew P Feinberg. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nature Biotechnology*, 31(2):142–147, February 2013.

[30] Stephen B Montgomery, Micha Sammeth, Maria Gutierrez-Arcelus, Radoslaw P Lach, Catherine Ingle, James Nisbett, Roderic Guigó, and Emmanouil T Dermitzakis. Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature*, 464(7289):773–777, 2010.

[31] Leopold Parts, Oliver Stegle, John Winn, and Richard Durbin. Joint Genetic Analysis of Gene Expression Data with Inferred Cellular Phenotypes. *PLoS Genetics*, 7(1):e1001276, January 2011.

[32] Joseph K Pickrell, John C Marioni, Athma A Pai, Jacob F Degner, Barbara E Engelhardt, Everlyne Nkadori, Jean-Baptiste Veyrieras, Matthew Stephens, Yoav Gilad, and Jonathan K Pritchard. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, 464(7289):768–772, January 2010.

[33] Vardhman K Rakyan, Thomas A Down, David J Balding, and Stephan Beck. Epigenome-wide association studies for common human diseases. *Nature Reviews Genetics*, 12(8):529–541, August 2011.

[34] Laurence D Robinson and Nicholas P Jewell. Some Surprising Results about Covariate Adjustment in Logistic Regression Models. *International Statistical Review / Revue Internationale de Statistique*, 59(2):227, August 1991.

[35] So-Youn Shin, Eric B Fauman, Ann-Kristin Petersen, Jan Krumsiek, Rita Santos, Jie Huang, Matthias Arnold, Idil Erte, Vincenzo Forgetta, Tsun-Po Yang, Klaudia Walter, Cristina Menni, Lu Chen, Louella Vasquez, Ana M Valdes, Craig L Hyde, Vicky Wang, Daniel Ziemek, Phoebe Roberts, Li Xi, Elin Grundberg, The Multiple Tissue Human Expression Resource MuTHER Consortium, Melanie Waldenberger, J Brent Richards, Robert P Mohney, Michael V Milburn, Sally L John, Jeff Trimmer, Fabian J Theis, John P Overington, Karsten Suhre, M Julia Brosnan, Christian Gieger, Gabi Kastenmüller, Tim D Spector, and Nicole Soranzo. An atlas of genetic influences on human blood metabolites. *Nature Genetics*, 46(6):543–550, June 2014.

[36] Oliver Stegle, Leopold Parts, Richard Durbin, and John Winn. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. 6(5):e1000770, May 2010.

[37] Oliver Stegle, Leopold Parts, Matias Piipari, John Winn, and Richard Durbin. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nature Protocols*, 7(3):500–507, 2012.

[38] Oliver Stegle, Sarah A Teichmann, and John C Marioni. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics*, 16(3):133–145, January 2015.

[39] Noah Zaitlen, N Zaitlen, Bogdan Pasaniuc, B Pasaniuc, N Patterson, Nick Patterson, Samuela Pollack, S Pollack, Benjamin Voight, B Voight, L Groop, Leif Groop, D Altshuler, David Altshuler, B E Henderson, Brian E Henderson, Laurence N Kolonel, L N Kolonel, L L Marchand, Loic Le Marchand, Kevin Waters, K Waters, C A Haiman, Christopher A Haiman, B E Stranger, Barbara E Stranger, E T Dermitzakis, Emmanouil T Dermitzakis, Peter Kraft, P Kraft, Alkes L Price, and Alkes L Price. Analysis of case-control association studies with known risk variants. *Bioinformatics*, 28(13):1729–1737, 2012.