Adjusting for Principal Components of Molecular Phenotypes Induces Replicating False Positives

Andy Dahl,*,[1] Vincent Guillemot,[†] Joel Mefford,* Hugues Aschard,[†,‡] and Noah Zaitlen*,[1]
*Department of Medicine, University of California San Francisco, 94158 California, [†]Centre de Bioinformatique, Biostatistique et Biologie Intégrative, Institut Pasteur, Paris, 75015 France, and [‡]Department of Epidemiology, Harvard TH Chan School of Public Health, Boston, 02115 Massachusetts
ORCID IDs: 0000-0001-6520-4766 (A.D.); 0000-0002-7554-6783 (H.A.)

ABSTRACT High-throughput measurements of molecular phenotypes provide an unprecedented opportunity to model cellular processes and their impact on disease. These highly structured datasets are usually strongly confounded, creating false positives and reducing power. This has motivated many approaches based on principal components analysis (PCA) to estimate and correct for confounders, which have become indispensable elements of association tests between molecular phenotypes and both genetic and nongenetic factors. Here, we show that these correction approaches induce a bias, and that it persists for large sample sizes and replicates out-of-sample. We prove this theoretically for PCA by deriving an analytic, deterministic, and intuitive bias approximation. We assess other methods with realistic simulations, which show that perturbing any of several basic parameters can cause false positive rate (FPR) inflation. Our experiments show the bias depends on covariate and confounder sparsity, effect sizes, and their correlation. Surprisingly, when the covariate and confounder have $\rho^2 \approx 10\%$, standard two-step methods all have >10-fold FPR inflation. Our analysis informs best practices for confounder correction in genomic studies, and suggests many false discoveries have been made and replicated in some differential expression analyses.

KEYWORDS confounder; molecular trait; quantitative trait loci; eigenvector perturbation

ASSOCIATION studies of molecular phenotypes have helped characterize basic biological processes, including transcription, methylation, chromatin accessibility, translation, ribosomal occupancy, and expression response to stimuli. These tests can be performed on *cis* and *trans* genetic variants to search for functional quantitative trait loci [*QTL: eQTL (Montgomery *et al.* 2010; Pickrell *et al.* 2010), mQTL (Rakyan *et al.* 2011), caQTL (Degner *et al.* 2012), pQTL (Albert *et al.* 2014), rQTL (Battle *et al.* 2015), sQTL (Rivas *et al.* 2015; Li *et al.* 2016), reQTL (Fairfax *et al.* 2014; Lee *et al.* 2014), and iQTL (Barry *et al.* 2017)]. Functional measurements can also be tested against nongenetic covariates with broad genomic effects, including cell type composition (Houseman *et al.* 2012; Jaffe and Irizarry 2014; Rahmani *et al.* 2017; Yao *et al.* 2017), disease status [*e.g.*, cancer (van't Veer *et al.* 2002), autism (Parikshak *et al.* 2016), and obesity (Horvath *et al.* 2014)], fetal developmental stage (Colantuoni *et al.* 2011), and ancestry (Galanter *et al.* 2017). We mostly refer to gene expression for simplicity, but our arguments apply to any highly structured, high-dimensional measurements.

Unfortunately, unmeasured and unknown factors are common and often have large effects in functional genomic data, reducing power and skewing null transcriptome-wide *P*-values (Leek and Storey 2007; Gibson 2008). Conditioning on known confounders—like technical batch—is invaluable but incomplete. Because genetic effects are typically small, even modest confounders can induce spurious genetic associations that dwarf real signal (Leek and Storey 2007; Kang *et al.* 2008).

Fortunately, strong confounders induce large, low-dimensional structure in the transcriptome, which is exactly what principal components (PCs) aim to capture (as do their variants, which we collectively call CCs, for confounding components). This blessing of dimensionality motivates a two-step approach where CCs are first estimated and then conditioned

doi: https://doi.org/10.1534/genetics.118.301768
Manuscript received November 4, 2018; accepted for publication January 23, 2019; published Early Online January 28, 2019.
Supplemental material available at Figshare: https://doi.org/10.25386/genetics.7040186.
[1]Corresponding authors: Department of Medicine, University of California San Francisco, 1550 4th St., Bldg. 19B, San Francisco, CA 94158. E-mail: andywdahl@gmail.com; and noah.zaitlen@ucsf.edu

on downstream as surrogates for the confounders (Alter *et al.* 2000; Leek and Storey 2008). Domain-specific CC methods, like surrogate variable analysis (SVA) (Leek and Storey 2007) and PEER (Stegle *et al.* 2010), make different assumptions about the structure of the confounders, and often outperform PCA. Two-step CC correction is an essential element of thousands of functional genomics analysis pipelines (Leek *et al.* 2010; Rakyan *et al.* 2011; Stegle *et al.* 2012, 2015; Albert and Kruglyak 2015).

Acknowledging the substantial benefits of CCs in many settings, in this work, we explore their adverse impact on the false positive rate (FPR). Theoretically, we derive an unappreciated source of bias created by conditioning on two types of PCs. Our results suggest the two step approach is biased whenever CCs imperfectly partition the phenotype-covariate correlation. We formalize this in a unifying, unidentified likelihood (1) that the CC methods each resolve with distinct assumptions.

We also study the bias with a range of CCs and simulations using real expression data from the GEUVADIS consortium (Lappalainen *et al.* 2013). First, we find no nontrivial method that avoids bias even in the simple scenario where the covariate has a small effect on a single gene and is added to white noise. In more complex data, this bias can be negligible compared to confounder-induced miscalibration. Next, we perform a series of simulations varying the number and strength of covariate effects and see substantial inflation in all CC methods when their assumptions fail. Finally, we allow confounders to be correlated with the covariate—the ordinary meaning of a confounder (Leek *et al.* 2010)—and find that all CCs can be severely miscalibrated; further, we show these false positives replicate out-of-sample.

Confounder Estimation and Correction

We write P molecular phenotypes measured on N samples as $Y \in \mathbb{R}^{N\times P}$, and let $y_p = Y_{,p}$ be the p-th phenotype. The primary covariate of interest is $x \in \mathbb{R}^{N\times 1}$. We assume x and the y_p are standardized to mean 0 and variance 1.

We stylize the standard two-step confounder correction as:

1. Estimate a rank-K confounder U by solving

$$\hat{U} := \arg\min_{U} \min_{\alpha,V} \left|\left|Y - x\alpha^T - UV^T\right|\right|_F^2 + \mathcal{P}(\alpha, U) \qquad (1)$$

where $\|\cdot\|_F$ is the Frobenius norm and $\mathcal{P}$ is some penalty representing (potentially implicit) priors on the causal and confounding patterns. Here, α and V are dummy variables.

2. Estimate the α_p by regressing each gene y_p on x given $\hat{U}$:

$$\hat{\alpha}_p := \mathrm{OLS}\left(y_p \sim x \middle| \hat{U}\right) \qquad (2)$$

where OLS indicates the regression coefficient on x from ordinary linear regression of y_p on x, $\hat{U}$, and an intercept.

We call approaches "unsupervised" when $\mathcal{P}$ constrains $\alpha = 0$ and "supervised" otherwise.

Solving (1) with $\mathcal{P}(\cdot) = 0$ amounts to maximum likelihood (ML) estimation under an i.i.d. Gaussian noise model for errors in Y (Leek and Storey 2007, 2008; Stegle *et al.* 2010). Standard ML inference for $\hat{\alpha}$ would then (pseudo)-invert the information matrix for asymptotic standard errors that account for uncertainty in U and V. This exposes one problem with two-step confounder correction: step two conditions on a fixed $\hat{U}$ as if it were known without error. Theoretically, this difficulty can be resolved by appealing to assumptions that ensure $\hat{U}$ perfectly estimates U so that there is no uncertainty to propagate (Leek and Storey 2008; Wang *et al.* 2017).

Another difficulty for ML inference in (1) is that its solution is not unique (even when requiring, *e.g.*, $V^TV = I$), as

$$x\alpha^T + UV^T = x\left(\underbrace{\alpha + V\Delta}_{\alpha'}\right)^T + \left(\underbrace{U - x^T\Delta}_{U'}\right)V^T \quad \forall \Delta \in \mathbb{R}^{K\times 1}$$

Because U' satisfies the rank-K constraint, α' can obtain the same likelihood as α: adding any vector in span(V) to any α admits an equivalent solution. [We ignore the nonidentifiability of (U, V) from the product UV^T because only span(U) is used in (2).]

This nonidentifiability means all (well-defined) CCs must use nontrivial penalty functions $\mathcal{P}$; below, we describe the choices of $\mathcal{P}$ roughly made by several popular CC methods. Moreover, this nonidentification means CCs depend heavily on their chosen $\mathcal{P}$ capturing the true, unknown parameter structure. In particular, we can easily design simulations where any particular CC behaves badly. This means that choice of CC method is important in practice and should be dataset-specific.

We note that we do not claim to take significant steps toward solving this identification problem, which has been analyzed from various theoretical perspectives elsewhere, *e.g.* West 2003; Leek and Storey 2008; Gagnon-Bartsch and Speed 2012; Sun *et al.* 2012; Gerard and Stephens 2017; Wang *et al.* 2017.

Studied Confounder Estimation Methods

Unsupervised PCA takes $\hat{U}$ as the top eigenvectors of YY^T or, equivalently, the top left singular vectors of Y. By definition, PCA solves (1) if $\mathcal{P}$ solely constrains $\alpha = 0$. The key problem is that the effect of x on Y leads PCs to partially capture x, analogous to unshielded colliders in a directed graphical model (Figure 1). That is, conditioning on genomic PCs can cause, rather than remove, bias. Related concerns arise when conditioning on heritable covariates in other contexts (Aschard *et al.* 2015, 2017; Day *et al.* 2016). This bias creates test misspecification even marginally, for each gene; in contrast, previous theory assumed marginal tests were valid and

focused on correlations between tests of different genes (Leek and Storey 2008).

We approximate the PC conditioning bias for small causal effects with textbook eigenvector perturbation theory. Conditioning on one phenotypic PC, our bias approximation for gene p coincides with the bias that naively derived from Figure 1

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

where a is the causal effect, q is the causally affected gene, and V are the right singular vectors of Y.

A similar result can be derived for conditioning on genotypic PCs in genome-wide association studies. In Figure 1, this stylistically corresponds to replacing the phenotypes (y_p) with SNPs, the covariate (x) with the tested phenotype, and reversing the arrow from x to y_1. However, genotype matrices typically contain three orders of magnitude more variables than expression matrices, concomitantly reducing the bias [entries of V are $O(P^{-1/2})$]. Intuitively, causal SNPs have much lower leverage on genetic PCs than causal genes have on expression PCs.

We also test an approach we call supervised PCA, which aims to protect $\hat{U}$ from x by first residualizing Y on x. This solves (1) when $\mathcal{P}$ constrains α to its unconditional estimate, x^TY. We show this method simply amplifies biases in the unconditional estimate: unsupervised PCs are too correlated with x, but supervised PCs are too uncorrelated.

We study several other approaches through simulation. SVA penalizes (1) by assuming α is sparse, which is often plausible. We used the "two-step" and "irw" algorithms to implement unsupervised and supervised SVA, respectively ("irw" requires supervision, but "two-step" is infrequently used). The "irw" version learns which genes are determined by the signal α *vs.* the confounder V by testing with q-values. These association strengths are used in turn to weight the relative importance of each gene inside a singular value decomposition, effectively weighting PCA toward more-confounded genes to better estimate confounders. We also tested a recent reimplementation, SmartSVA (Chen *et al.* 2017); we found that it performs very similar to SVA with $B = 50$ iterations, but much faster.

PEER explicitly penalizes U and α through priors on their respective sizes. PEER uses automatic relevance determining priors for the factors U and V and fits parameters with variational Bayes. In simulations, its default hyperparameters perform well when x explains $\approx 1\%$ of transcriptome-wide variation but less well for larger α. PANAMA is a closely related approach that greedily adds the most relevant SNPs while learning latent factors. While PANAMA can improve performance over PEER for SNPs with large effects, it is more computationally expensive and is rarely used in practice for human datasets.

RUV and related methods estimate confounders by using only a submatrix of Y that is known, *a priori*, to be unaffected by the primary signal (Lucas *et al.* 2006; Gagnon-Bartsch and Speed 2012; Gerard and Stephens 2017; Wang *et al.* 2017). This prior information breaks the identifiability problem by constraining $\alpha_S = 0$ for some subset of genes S (corresponding to a barrier penalty function for $\mathcal{P}$). Latent factors can be safely identified by restricting to the negative control data, and then their effects on other genes can be extrapolated. We assume control genes or samples are unavailable, which is common, and do not study these approaches further.

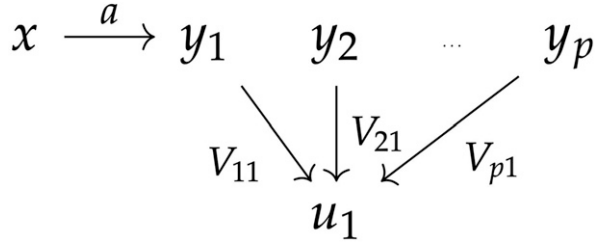

Figure 1 Graphical model suggesting CC conditioning causes bias. u_1 is the top PC, which is indirectly captures x. Although x affects only y_1, conditioning on the collider u_1 induces spurious correlation with all other y_p.

We also assess LEAPP (Sun *et al.* 2012), a recent method that uses sparsity assumptions to disentangle α and V. Unlike other methods we study, LEAPP provably obtains oracle performance, asymptotically and assuming that confounders are strong, signals are sparse, and noise is independent across genes modulo confounders (Wang *et al.* 2017).

Finally, we also test the linear mixed model-based method ICE, which uses a random effect with covariance kernel $\frac{1}{P}YY^T$ to capture confounding, and tests the fixed effects of x against each gene individually (Kang *et al.* 2008). Conceptually, ICE seeks a few genes that are highly correlated with x compared to typical genes. Like in RUV methods, control genes can be used to improve power, which we did not study (Joo *et al.* 2014).

The bias for one unsupervised PC

In this section we take Y as a *QTL plus some deterministic Y^0:

$$Y = x\alpha + Y^0 \tag{3}$$

We assume $x \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. We allow Y^0 to be fully general to capture all noise and confounding. This is closely related to the spiked covariance model (Johnstone 2001), though we use a general Y^0 in place of i.i.d. Gaussian noise. We assume $\alpha_q = a$ and $\alpha_p = 0$ for $p \neq q$, and we call x a local covariate for gene q.

Ideally, the OLS step (2) would condition exactly on the true, unknown confounders. We evaluate conditioning, instead, on the top PCs of Y (Y^0), which we call U (U^0) and define as the top left singular vectors of Y (Y^0). We compare conditioning on U to U^0 rather than the true confounders for two reasons. First, Y^0 is assumed independent of x, hence conditioning on U^0 does not cause bias. Second, this allows us not to assume any particular form for confounding, or even its existence.

We aim to quantify the error at gene $p \neq q$ from conditioning on the top feasible PC, u_1, instead of the top oracle PC, u_1^0:

$$\text{Error}_p := \hat{\alpha}_p - \hat{\alpha}_p^0$$

where $\hat{\alpha}$ ($\hat{\alpha}^0$) solves (2) given u_1 (u_1^0). Since $y_p = y_p^0$ for $p \neq q$, any error can only be caused by the effect of x on U.

The bias is the expected error over x, the only randomness:

$$\text{Bias}_p := \mathbb{E}(\text{Error}_p) = \mathbb{E}\left(\hat{\alpha}_p - \hat{\alpha}_p^0\right) = \mathbb{E}(\hat{\alpha}_p) - \alpha_p$$

We study Error_p rather than Bias_p to focus away from the ordinary regression error due to noise and the onto the error caused by x's perturbation of U. This enables stronger results—particularly, that $\text{Error}_p \approx \text{Bias}_p$ deterministically. Because of this, we often refer to this perturbed PC conditioning error as the bias, *i.e.*, the randomness in Error_p is negligible.

We assume α is small so that we can use a standard approximation to the perturbed eigenvector [*e.g.*, (Allez and Bouchaud 2012), Sec. II]: for any small E, the first eigenvector of $1/P, Y^TY + E$ is approximately

$$u_1 = u_1^0 + \sum_{j=2}^{N} \frac{u_1^{0^T} E u_j^0}{\lambda_1^0 - \lambda_j^0} u_j^0 + O\left(||E||^2\right) \tag{4}$$

where λ_j^0 is the j-th eigenvalue of $\frac{1}{P} Y^0 Y^{0^T}$.

Under our assumption on α, $E = a y_q x^T + a x y_q^T + a^2 x x^T$, giving the perturbation approximation

$$u_1 = u_1^0 + a \sum_{j>1} \frac{\tilde{y}_{1q}\tilde{x}_j + \tilde{y}_{jq}\tilde{x}_1}{\lambda_1^0 - \lambda_j^0} u_j^0 + O(a^2) \tag{5}$$

This uses simplifying definitions based on rotating with U:

$$\tilde{Y} := U^T Y; \;\; \tilde{x} := U^T x$$

Note that $\tilde{x}$ is still a spherical Gaussian random variable.

We show in Supplemental Material, Section S1.1 that this approximation can be combined with the standard two-step least squares expression for $\hat{\alpha}$ to give

$$\text{Error}_p \approx -2a\bar{c} \sum_{j=1}^{N} \tilde{Y}_{jp}\tilde{Y}_{jq} w_j^{\tilde{x}} \tag{6}$$

$\bar{c}$ is a condition number for Y^0 and $w^{\tilde{x}}$ are random weights:

$$\bar{c} := \frac{1}{N-1} \sum_{j>1} \frac{1}{\lambda_1^0 - \lambda_j^0}$$

$$w_j^{\tilde{x}} := \frac{1}{2(N-1)\bar{c}} \frac{\tilde{x}_1^2}{\lambda_1^0 - \lambda_j^0} \,(\text{for } j \neq 1)$$

$$w_i^{\tilde{x}} := \frac{1}{2(N-1)\bar{c}} \sum_{k>1} \frac{\tilde{x}_k^2}{\lambda_1^0 - \lambda_k^0}$$

The $w^{\tilde{x}}$ partition the perturbation among PCs and are proportional to the (random) squared correlations between x and the PCs (*i.e.*, $\tilde{x}^2$). The $w^{\tilde{x}}$ are nonnegative and sum to one in expectation. $\bar{c}$ is deterministic—depending only on the spectrum of Y^0—and quantifies the susceptibility of the first PC to perturbation.

These properties of $w^{\tilde{x}}$ mean the error in (6) is a (randomly) weighted correlation between the projections of genes p and q—the tested and the causal genes—onto the eigen-axes, *i.e.*

$$\text{Error}_p \approx -2a\bar{c}\rho^{w^{\tilde{x}}}(\tilde{y}_p, \tilde{y}_q) \tag{7}$$

where ρ^{π} is the correlation weighted by some π. In particular, if 1_N is a vector of 1s, $\rho^{1_N}(\tilde{y}_p, \tilde{y}_q) = \rho(y_p, y_q)$ is the ordinary correlation between the two genes. In contrast, $\rho^{w^{\tilde{x}}}$ randomly weights the eigen-axes, but with far greatest weight on axis 1 and successively less expected weight on subsequent axes.

$\rho^{w^{\tilde{x}}}$ is the only remaining randomness, so the error depends on x only through this (random) notion of correlation. And even this randomness is often negligible. First, $w_1^{\tilde{x}} \gg w_j^{\tilde{x}}$ for $j > 1$ (the former is the sum over the $N-1$ latter), and this gap grows for increasingly confounded data (Figure S1). Second, $w_1^{\tilde{x}}$ should be very well approximated by its expectation because it is an average over $N-1$ variables.

Together, this suggests the approximations $w_1^{\tilde{x}} \approx \frac{1}{2}$ and $w_j^{\tilde{x}} \approx 0$ for $j > 1$, giving a deterministic approximation to the random error. Because the error is approximately deterministic, it can immediately be recognized a bias approximation, as well:

$$\text{Bias}_p \approx \text{Error}_p \approx -aV_{p1}V_{q1} \tag{8}$$

This uses the approximation $\bar{c} \approx 1/\lambda_1^0$ (Equation S3). We find (8) is accurate in a realistic simulation (Figure S2).

While $\hat{\alpha}_p$ is biased conditional on q and p, this conditional bias itself has mean zero on average over q or p ($V_{,1}$ is mean zero). Nonetheless, our conditional definition of bias conveys the fact that p and q are biologically meaningful and replicable indices. Moreover, even random biases with mean zero introduce overdispersion that still causes false positive inflation.

We have not generalized these calculation to $K > 1$ PCs, though we suspect an analogous result will hold after appropriately modifying w. If correct, ρ^w will move toward ρ^{1_N} as K grows, suggesting the correlation between causal and tested traits is a good intuitive proxy for the bias.

The bias for supervised PCs

An apparent solution is to project out x before computing PCs, which we call supervised PCA. We show this is deeply flawed, even when x has no causal effect, supporting existing simulation results (Leek and Storey 2007).

First, after residualizing x from Y, x is the bottom supervised PC, hence orthogonal to the others. Thus the unconditional OLS estimates for α_p are unchanged by conditioning on supervised PCs. Similarly, the ratio of the conditional and unconditional SEs is just the ratio of the overall regression

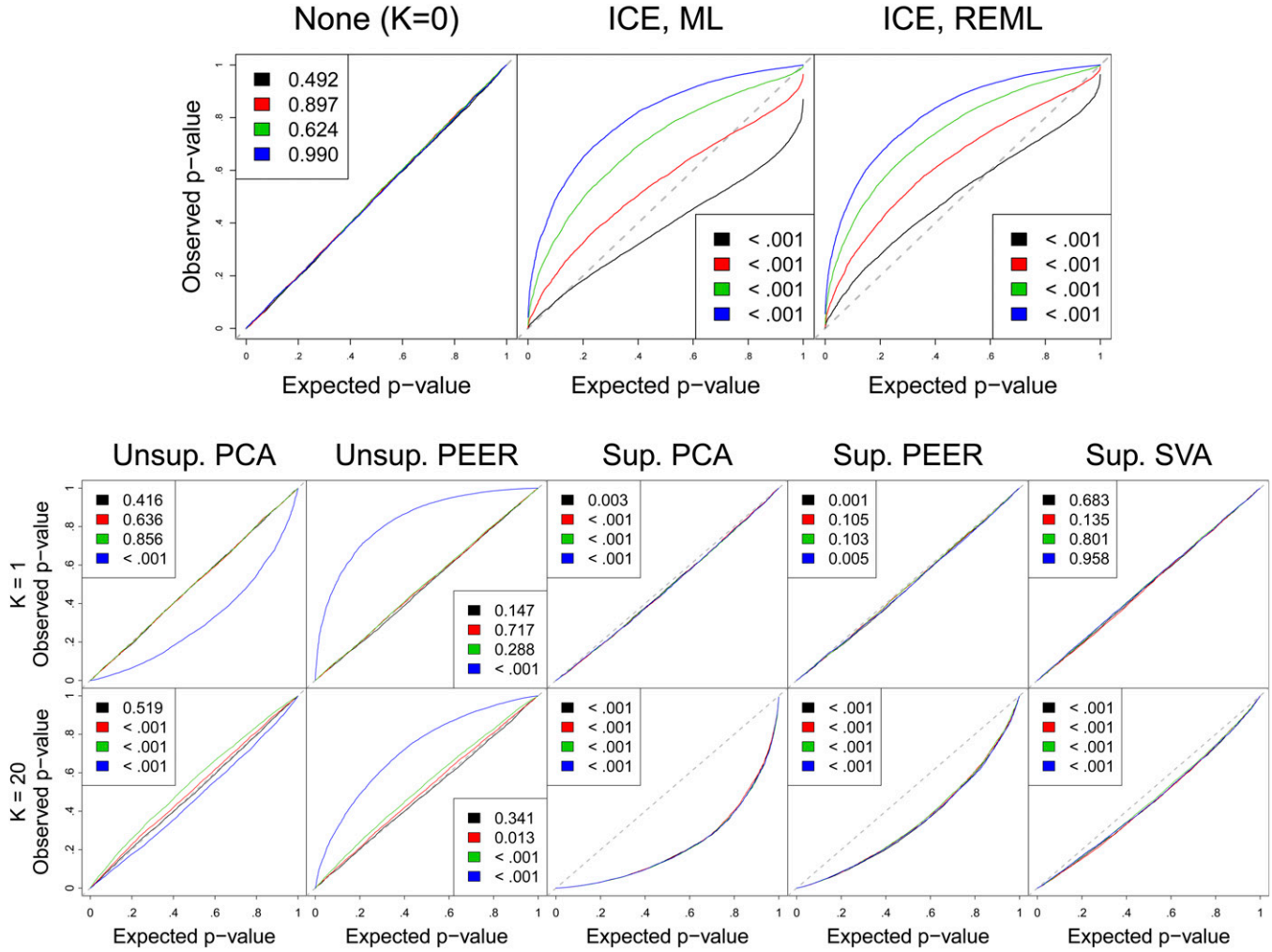

Figure 2 Confounder correction causes *P*-value miscalibration in simulations with a local effect and white noise. QQ plots show one-sided KS test *P*-values for the nominally null regression *P*-values in 5000 simulations. 2-sided KS tests of the KS *P*-values are in the legends. Variation explained in the causal gene is 0% (black), 30% (red), 60% (green), or 90% (blue).

error estimates, *i.e.*, $\hat{\sigma}^2_{\text{cond}}/\hat{\sigma}^2_{\text{uncond}}$. Together, the ratio of *t*-statistics testing $\alpha_p = 0$ is

$$\frac{t_{\text{cond}}}{t_{\text{uncond}}} = \frac{\hat{\alpha}_{\text{cond}}/\hat{\sigma}_{\text{cond}}}{\hat{\alpha}_{\text{uncond}}/\hat{\sigma}_{\text{uncond}}} = \frac{\hat{\sigma}_{\text{uncond}}}{\hat{\sigma}_{\text{cond}}} \tag{9}$$

By definition, U explains large amounts of variance in Y, making $\hat{\sigma}_{\text{cond}}$ smaller than $\hat{\sigma}_{\text{uncond}}$. Formally, the ratio (9) is inflated on average (over genes and x) and, in practice, is usually inflated (Figure S3 and Section S1.4).

Local covariate simulations with white noise

We now demonstrate the CC conditioning bias in a simplistic simulation using (3): the background expression Y^0 is drawn i.i.d. standard normal; the x_i are (independently) i.i.d. standard normal; and $\alpha = ae_q$, so that x is a local *QTL for gene q; finally, q is drawn, independently of x and Y, uniformly from $\{1, \ldots, P\}$, and its effect a is varied over $\{0, .3, .6, .9\}$. We chose $(N, P) = (375, 13120)$ to match the GEUVADIS data (see below for details).

After simulating Y and x, we test for $\alpha_p = 0$ using either PCA, SVA, PEER, or their supervised versions to estimate $\hat{U}$. We also test the mixed model implemented in ICE (which failed to converge in a few simulated datasets).

For 1000 independently simulated datasets, we perform one-sided Kolmogorov-Smirnov (KS) tests for deflation in the regression *P*-values at noncausally affected genes. A two-sided test should be used in the first step when testing for general miscalibration (Leek and Storey 2007, 2008); we, however, are testing for estimator bias, which decreases *P*-values, and discriminates inflation from deflation.

Figure 2 presents the QQ plots for the resulting KS *P*-values. Unsurprisingly, excluding all CCs (None) delivers well-calibrated *P*-values because we did not simulate confounders. Confirming our theory, unsupervised PCs cause noticeable bias for large a; however, no bias is detected for small a, emphasizing that the bias can be negligible for local covariates. Unsupervised PEER results are similar for small a, but for large a PEER becomes conservative (such observations require one-sided KS

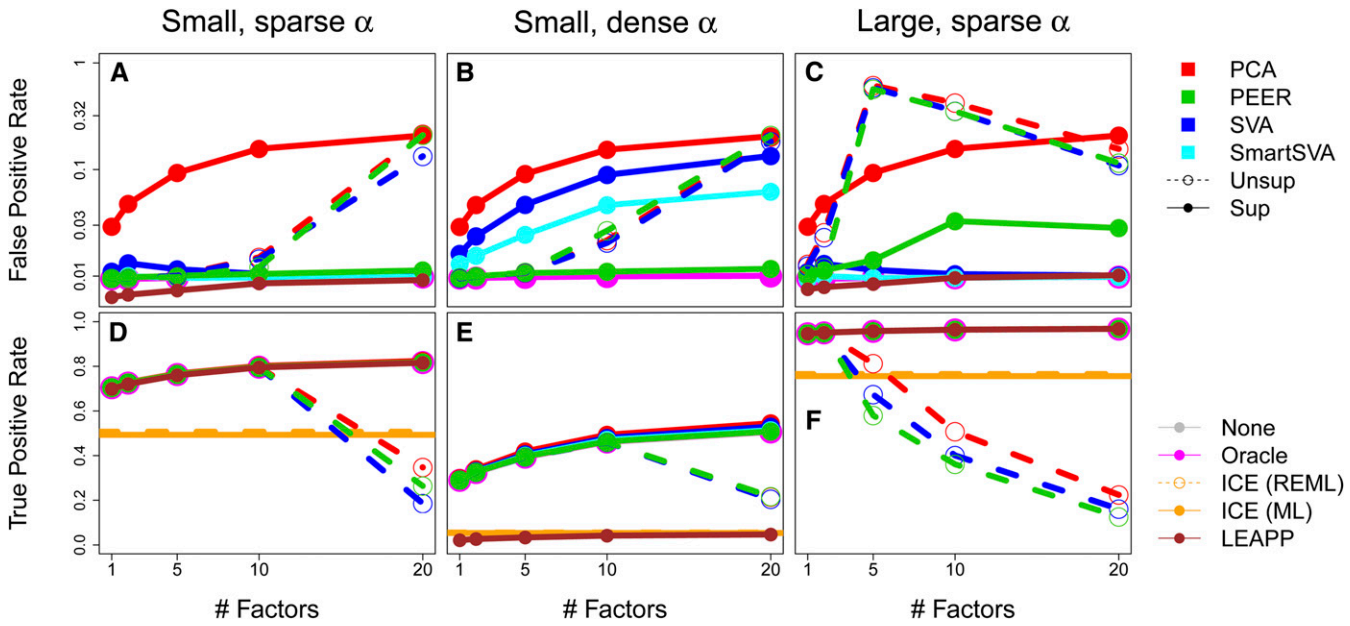

Figure 3 Mean FPR (a–c) and TPR (d–f) (on log scale) for a simulated global α added to GEUVADIS expression using three settings. The oracle is always calibrated and often covered by other lines. Two-step CC methods are at top-right; others are bottom-right. ICE has essentially 0 FPR and is omitted from the top plots.

tests for the regression *P*-values). ICE is mostly conservative, especially when using REML and $a > 0$. Unsupervised SVA correctly declares $K = 0$ [with the permutation test from (Buja and Eyuboglu 1992)] and is thus equivalent to "None"; although this is ideal behavior, any CC method could use this (or other) tests to choose K, and analysts in practice often turn to a different method in this situation (*e.g.*, Pierce *et al.* 2014).

The three supervised methods qualitatively share a different type of bias, growing with K and depending little on a. We theoretically characterized this for supervised PCA, but PEER and, especially, SVA seem less biased.

Overall, Figure 2 shows that all tested (nontrivial) CC methods create *P*-value miscalibration even in the complete absence of confounding, though the problem is small for small a.

Data availability

We used a high-quality RNA-sequencing dataset from the GEUVADIS consortium (Lappalainen *et al.* 2013) as a realistic simulation baseline. We aligned the raw transcript reads from the European individuals to the reference hg19 transcriptome using RSEM (Li and Dewey 2011). We removed perfectly correlated genes and quantile-normalized the rest to standard normal. The final matrix has $N = 375$ samples (rows) and $P = 13,120$ genes (columns) and column-means and -variances equal to 0 and 1. Its spectrum is shown in Figure S10. Supplemental material available at Figshare: https://doi.org/10.25386/genetics.7040186.

Global Covariate Simulations with Real Traits

We now simulate a global effect, meaning α is much denser and larger. We set 90% of its entries to 0 and the others to i.i.d. Gaussian with mean zero and variance such that x explains 1% of transcriptome-wide variation. Nongenetic x can easily have these sorts of effects, *e.g.*, even early studies with small sample sizes found broad expression profiles that still inform breast cancer treatment (Sparano *et al.* 2015; Cardoso *et al.* 2016). If x were genetic, it would be an extremely strong *trans*-*QTL.

We now use the GEUVADIS expression for the noise Y^0 to make the simulation more realistic, and let $x_i \overset{\text{iid}}{\sim}$ Binomial(2,20%). Finally, as we do not aim to match the perturbation theory here, we adopt standard practice and normalize x and columns of Y to mean 0, variance 1.

We assess empirical FPR and true positive rate (TPR) at the nominal $p = .01$ level and average over 250 independently simulated datasets (averaging before log-transforming, Figure 3, a and d). All unsupervised methods and supervised PCA are badly miscalibrated for $K \geq 10$, while other supervised CC methods, ICE, and LEAPP were calibrated. These calibrated methods had power similar to Oracle, which uses PCs of the pure noise term Y^0, except ICE.

We then decrease the sparsity of α from 90% zeros to 5% zeros, violating the sparsity assumptions of supervised (Smart)SVA and LEAPP. This leads (Smart)SVA to roughly 10-fold FPR inflation at $K = 10$ (Figure 3b) and LEAPP to lose essentially all power (Figure 3e). Supervised PEER, however, retains near-oracle power and calibration.

Next, we return to 90% sparsity in α but increase its variance explained from 1 to 25%. This apparently violates PEER's assumptions as it is ≈five-fold inflated at $K = 10$ (Figure 3c).

These conclusions qualitatively remain when using a $q = .01$ threshold (Storey 2003) or $p = .001$ (Figures S4 and S5).

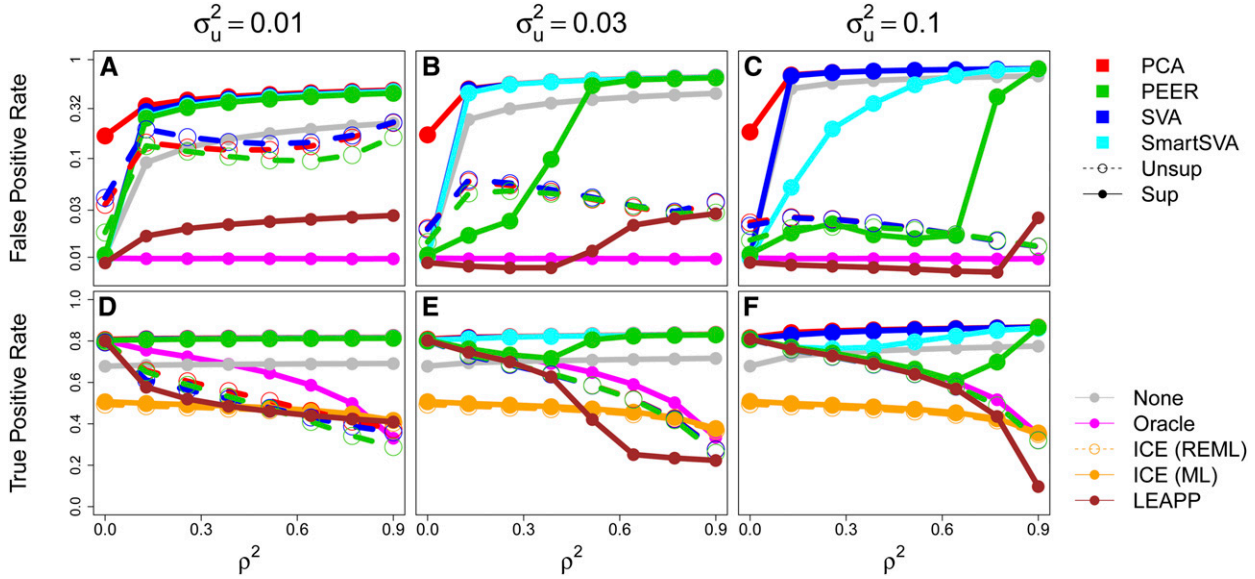

Figure 4 FPR and TPR at a nominal $p = .01$ level for testing a strong covariate x correlated with a confounder u. x and u have squared correlation ρ^2, and their respective variances explained are $\sigma_x^2 = 1\%$ and $\sigma_u^2 = 1\%$ (a and d), 3% (b and e), or 10% (c and f). Two-step CC methods are in the top legend; others are in the bottom legend.

Confounders Correlated with a Global Covariate

The previous simulations only added a causal x effect to some independent Y^0. We now add confounders correlated with x, which we feel is common in practice for nongenetic x (genetic x can only be confounded by population structure). For example, even within-tissue, PEER factors had ρ^2 ranging from 10% to 60% with known covariates (GTEx Consortium *et al.* 2015). Further, experimental procedures are often correlated with biological factors (Gilad and Mizrahi-Man 2015). And correlation between technical confounders and primary biological signal can be pernicious: tissue dissociation in quiescent muscle stem cells can resemble cellular activation from muscle injury (van den Brink *et al.* 2017).

We now add a confounder, u, that has correlation ρ with x:

$$Y = x\alpha + u\beta + Y^0$$

We draw each (x_i, u_i) pair i.i.d. from a bivariate Gaussian with mean zero, variances equal to 1 and correlation ρ. We repeated our above pipeline, simulating 250 independent datasets and plotting the FPR in Figure 4. Here, we always take $K = 12$ CCs, because $K = 10$ was used originally for Y^0 (Lappalainen *et al.* 2013) and we have added two rank-one effects.

We draw α as in Figure 3a: 10% of its entries are nonzero, drawn i.i.d.Gaussian with mean zero and variance such that $\sigma_x^2 = 1\%$ is the transcriptome-wide fraction of variance explained. Because u represents a confounder, we draw β i.i.d. Gaussian with variance such that $\sigma_u^2 = 10\%$. This is roughly in line with GTEx (Aguet *et al.* 2017), where the top 15–35 PEER factors collectively explained 59−78% of transcriptome-wide variation. We vary ρ^2 from 0 to 1 (ρ and $-\rho$ are equivalent).

The results in Figure 4 show that supervised PCA badly inflates FPR, even at $\rho = 0$. The other supervised methods are nearly as inflated for $\rho > 0$, except PEER for larger σ_u^2 and modest ρ. Unsupervised CCs also inflate FPR for $\sigma_u^2 = 1\%$, though this diminishes as σ_u^2 grows and u becomes near-perfectly captured. In particular, the apparently naive approach of simply ignoring confounding (None) can be less inflated than all CC methods, particularly for small ρ^2 or σ_u^2. This shows that, even when confounding exists, CC adjustment can cause more harm than good. Qualitatively similar patterns hold using $q = .01$ or $p = .001$ thresholds (Figures S6 and S7).

We found that ICE was always calibrated and LEAPP was always close to calibrated, with a maximum of roughly threefold inflation. The LEAPP inflation occurred for smaller σ_u^2 and larger ρ^2, scenarios excluded by assumptions in Wang *et al.* (2017). LEAPP was often more powerful than ICE, but these were typically settings where CC methods also performed well. For example, LEAPP was similar to supervised PEER/(Smart)SVA when $\rho^2 = 0$, and LEAPP was similar to unsupervised CCs and supervised PEER when $\sigma_u^2 = 10\%$ (though better calibrated). In the intermediate range (*e.g.*, $\sigma_u^2 = 3\%$ and $.1 < \rho^2 < .5$), however, LEAPP outperformed all competitors. LEAPP was far slower than the other methods, taking ≈2 hr on average over simulated datasets in Figure 4, compared to ≈1–5 min for two-step methods and ≈10 min for ICE (Figure S8) nonetheless, recent re-implementations of LEAPP are faster (Wang *et al.* 2017).

Similar simulations can be found in Figure S2 of Leek and Storey (2008), but they reached the opposite conclusion, *i.e.* that SVA is calibrated even when $\rho^2 > 0$. To test if this discrepancy is due to their smaller tested data dimensions, $(N, P) = (20, 1000)$, we repeated our simulations in Figure

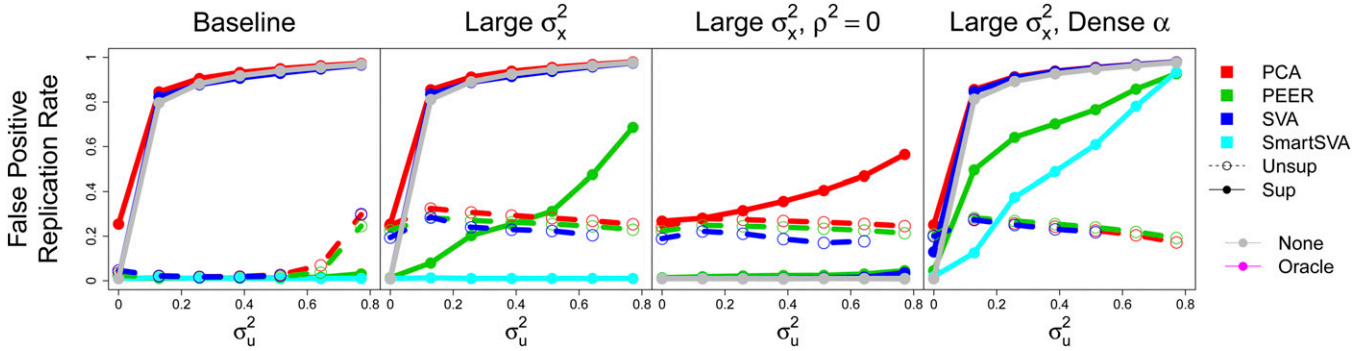

Figure 5 False positive replication rate at a nominal $p = .01$ level. σ_u^2 is the confounder strength. The signal strength σ_x^2 is either 1% (baseline) or 10% (others). The squared signal-confounder correlation ρ^2 is 25%, or 0 for the third panel.

4 after downsampling to 20 samples and 1000 genes (uniformly without replacement, and independently downsampling for each simulation). This reduced FPR for SVA, though SVA can still be inflated (*e.g.*, ≈10-fold for $\rho^2 \approx 50\%$ and $\sigma_u^2 = 10\%$, Figure S9). This remaining discrepancy is likely partially because there is, in fact, inflation in Figure S2 of Leek and Storey (2008): *e.g.*, row 2, column 4 visually seems miscalibrated.

False positive replication

Assuming a and q are biologically determined and universal, the PCA bias we derived depends only on the top PCs of Y. As N grows large, the bias remains, precision grows, and the null is rejected for every gene if x affects even one.

Analogously, biases will be similar between datasets with similar top PCs, which occurs in the presence of strong and similar confounders, *e.g.* batch effects (Leek *et al.* 2010), population structure (Listgarten *et al.* 2010), cell type composition (Jaffe and Irizarry 2014), or the signal x itself if it is strong.

We empirically assess replication rates by simulating as in Figure 4 with $\sigma_x^2 = 1\%$, $\rho^2 = .25$, and 10% of α's entries drawn i.i.d. Gaussian. We then split the data (x and Y) into halves, test each separately, compute the replication rate as the fraction of false positive discoveries from the first half that are deemed positive in the second half, and repeat after transposing the splits. We use a significance level of $p = .01$ in each split. We independently repeat the process 250 times, ignoring initial splits without false discoveries. Splitting an existing dataset simulates worst-case confounder sharing between discovery and replication cohorts.

The left panel of Figure 5 shows the average (positive) false replication rate. Unsupervised methods perform reasonably well and supervised SmartSVA and PEER are calibrated. Either performing no confounder correction or using supervised SVA or PCA creates severe spurious replication, with nearly all false positives replicating when σ_u^2 is large (*e.g.*, $> 20\%$). In this section, we do not assess ICE as it has low total positive rates.

We then increased the signal to $\sigma_x^2 = 10\%$. Unsupervised methods performed worse, as they more readily capture the larger x effect, and supervised PEER became miscalibrated, especially for larger σ_u^2. Next, we reduced the confounder correlation to 0: unsupervised methods performed slightly better, while supervised methods (except PCA) became roughly calibrated. Finally, we increased the density of α to 90%, causing supervised SmartSVA to suffer similarly to PEER.

Discussion

We have evaluated unappreciated sources of bias induced by conditioning on estimated confounders in functional genomic association tests. We used a combination of theory and simulation to cover different cases of interest. Overall, no two-step CC method we evaluated generally had calibrated FPR; moreover, most studies use PCA, one of the worst-performing methods in our simulations. Although all methods behave well when their assumptions hold, and these assumptions are often reasonable in practice, confounders even modestly correlated with the primary signal can cause substantial bias. We also showed these false positives can replicate at a high rate.

*QTL studies are often performed only within local genomic windows, which are called *cis*-*QTL studies. In this context, genetic effects are small and restricted to nearby genes, and our results suggest that the bias induced from CC conditioning is minimal. However, unlike in our Figure 2 simulations, *cis*-windows may contain many highly correlated genes, and *cis*-*QTL often causally affect genes other than the nearest one (Zhu *et al.* 2016), both of which serve to inflate FPR.

*QTL studies can also be performed genome-wide, which are called *trans*-*QTL studies. While such *QTL are biologically central, they are difficult to reliably uncover because the signals tend to be dispersed across the transcriptome and genome. Unlike *cis*-*QTL, *trans*-*QTL can have much larger effects on CCs, which has led modern studies to diametrically opposed methodology for testing *trans*-*QTL. For example, Brynedal *et al.* (2017) do not use CCs, despite the fact that "all [significant] gene sets were significantly correlated to ... the top 20 PEER factors." On the other hand, Yao *et al.* (2017) adjust for 20 PEER factors and cell type composition, and

many of the resulting "hotspot" signals are in fact loci known to affect cell type composition[1]. Similar to (Yao *et al.* 2017), (Aguet *et al.* 2017) adjust for PEER factors computed per tissue. We have proposed GBAT to address these and other limitations by testing gene-level *trans* associations (Liu *et al.* 2018). GBAT uses supervised SmartSVA, which performs best in our simulations without correlated confounders. This is feasible because we test only thousands of genes rather than millions of SNPs, though SNPs could be prescreened with unsupervised CCs. More generally, we recommend ICE or LEAPP when feasible: ICE was more reliably calibrated but had low power than LEAPP, while two-step CCs could perform very poorly when their assumptions are violated.

Our global covariate simulations are relevant for differential expression studies performed with linear regression and/or CC correction. Such tests have been broadly applied, including to differences between tissues, sexes, or ages. These factors can easily correlate with latent confounders when experiments are not randomized and can substantially affect the expression of many genes. This is the setting in our simulations underlying Figure 4, where no approach (except ICE) generally gives calibrated *P*-values and CC correction can be worse than a completely uncorrected analysis.

A key limitation of two-step approaches is that step 1 uncertainty is not propagated to the test in step 2. To address this, a multiple imputation-style approach can be used, performing step 2 on several draws from the step 1 CC posterior. We have not evaluated this concept as it is currently developed only within the RUV framework (Gerard and Stephens 2017). Related, the two steps can be integrated, which analytically conveys first-step uncertainty, though this has much greater computational cost when run genome-wide (Stegle *et al.* 2010; Fusi *et al.* 2012; Sun *et al.* 2012; Wang *et al.* 2017).

In the future, it may be useful to pursue other assumptions on Y^0 in the bias calculation we derived for unsupervised PCA. For example, we could use a spiked covariance model for Y^0, using results from Nadler (2008) to approximate the perturbed eigenvector (*e.g.*, replacing our Equation 5 with their Equation 2.15). An advantage of our approach, however, is that we allow general correlations between traits.

In more complex scenarios, the appropriate covariate and confounding model can be unclear. For example, coexpression studies learn complex and subtle graphical models from (partial) covariance (Horvath 2011; Shin *et al.* 2014). But uncorrected confounders (or biased corrections) will yield statistically significant, biologically meaningless networks. Latent variable graphical models may suit this problem (Chandrasekaran *et al.* 2012), and a related two-step approximation was recently proposed for genomics (Parsana *et al.* 2017). Finally, mixed models that learn specifically-genetic graphical models may be adaptable to adjust for low-dimensional confounding (Dahl *et al.* 2013).

[1]Thanks to Alexander Gusev for this observation.

Acknowledgments

We are grateful to Brunilda Balliu and Antonio Berlanga-Taylor for discussions about CC methods. This work was partially supported by National Institutes of Health grants 1U01HG009080-01, 5K25HL121295-03, and 1R03DE025665-01A1.

Literature Cited

Aguet, F., A. A. Brown, A. V. Segre, B. J. Strober, Z. Zappala *et al.*, 2017 Genetic effects on gene expression across human tissues. Nature 550: 204–213. https://doi.org/10.1038/nature24277

Albert, F. W., and L. Kruglyak, 2015 The role of regulatory variation in complex traits and disease. Nat. Rev. Genet. 16: 197–212. https://doi.org/10.1038/nrg3891

Albert, F. W., S. Treusch, A. H. Shockley, J. S. Bloom, and L. Kruglyak, 2014 Genetics of single-cell protein abundance variation in large yeast populations. Nature 506: 494–497. https://doi.org/10.1038/nature12904

Allez, R., and J. P. Bouchaud, 2012 Eigenvector dynamics: general theory and some applications. Phys. Rev. E Stat. Nonlin. Soft Matter Phys. 86: 046202. https://doi.org/10.1103/PhysRevE.86.046202

Alter, O., P. O. Brown, and D. Botstein, 2000 Singular value decomposition for genome-wide expression data processing and modeling. Proc. Natl. Acad. Sci. USA 97: 10101–10106. https://doi.org/10.1073/pnas.97.18.10101

Aschard, H., B. J. Vilhjálmsson, A. D. Joshi, A. L. Price, and P. Kraft, 2015 Adjusting for heritable covariates can bias effect estimates in genome-wide association studies. Am. J. Hum. Genet. 96: 329–339. https://doi.org/10.1016/j.ajhg.2014.12.021

Aschard, H., V. Guillemot, B. Vilhjalmsson, C. J. Patel, D. Skurnik *et al.*, 2017 Playing musical chairs in big data to reveal variables associations. bioRxiv. https://doi.org/10.1038/ng.3975

Barry, J. D., M. Fagny, J. N. Paulson, H. Aerts, J. Platig *et al.*, 2017 Histopathological image QTL discovery of immune infiltration variants. bioRxiv. https://www.biorxiv.org/content/10.1101/126730v3

Battle, A., Z. Khan, S. H. Wang, A. Mitrano, M. J. Ford *et al.*, 2015 Genomic variation. Impact of regulatory variation from RNA to protein. Science 347: 664–667. https://doi.org/10.1126/science.1260793

Brynedal, B., J. Choi, T. Raj, R. Bjornson, B. E. Stranger *et al.*, 2017 Large-scale trans -eQTLs affect hundreds of transcripts and mediate patterns of transcriptional Co-regulation. Am. J. Hum. Genet. 100: 581–591. https://doi.org/10.1016/j.ajhg.2017.02.004

Buja, A., and N. Eyuboglu, 1992 Remarks on parallel analysis. Multivariate Behav. Res. 27: 509–540. https://doi.org/10.1207/s15327906mbr2704_2

Cardoso, F., L. J. van't Veer, J. Bogaerts, L. Slaets, G. Viale *et al.*, 2016 70-Gene signature as an aid to treatment decisions in early-stage. Breast Cancer 375: 717–729.

Chandrasekaran, V., P. A. Parrilo, and A. S. Willsky, 2012 Latent variable graphical model selection via convex optimization. Ann. Stat. 40: 1935–1967. https://doi.org/10.1214/11-AOS949

Chen, J., E. Behnam, J. Huang, M. F. Moffatt, D. J. Schaid *et al.*, 2017 Fast and robust adjustment of cell mixtures in epigenome-wide association studies with SmartSVA. BMC Genomics 18: 413. https://doi.org/10.1186/s12864-017-3808-1

Colantuoni, C., B. K. Lipska, T. Ye, T. M. Hyde, R. Tao *et al.*, 2011 Temporal dynamics and genetic control of transcription in the human prefrontal cortex. Nature 478: 519–523. https://doi.org/10.1038/nature10524

Dahl, A., V. Hore, V. Iotchkova, and J. Marchini, 2013 Network inference in matrix-variate Gaussian models with non-independent noise. arxiv: 1312.1622v1.

Day, F. R., P. R. Loh, R. A. Scott, K. K. Ong, and J. R. Perry, 2016 A robust example of collider bias in a genetic association study. Am. J. Hum. Genet. 98: 392–393. https://doi.org/10.1016/j.ajhg.2015.12.019

Degner, J. F., A. A. Pai, R. Pique-Regi, J.-B. Veyrieras, D. J. Gaffney *et al.*, 2012 DNase I sensitivity QTLs are a major determinant of human expression variation. Nature 482: 390–394. https://doi.org/10.1038/nature10808

Fairfax, B. P., P. Humburg, S. Makino, V. Naranbhai, D. Wong *et al.*, 2014 Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. S 343: 1246949.

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

Gagnon-Bartsch, J. A., and T. P. Speed, 2012 Using control genes to correct for unwanted variation in microarray data. Biostatistics 13: 539–552. https://doi.org/10.1093/biostatistics/kxr034

Galanter, J. M., C. R. Gignoux, S. S. Oh, D. Torgerson, M. Pino-Yanes *et al.*, 2017 Differential methylation between ethnic sub-groups reflects the effect of genetic ancestry and environmental exposures. eLife 6: e20532. https://doi.org/10.7554/eLife.20532

Gerard, D., and M. Stephens, 2017 Unifying and generalizing methods for removing unwanted variation based on negative controls. arXiv: 1705.08393v1.

Gibson, G., 2008 The environmental contribution to gene expression profiles. Nat. Rev. Genet. 9: 575–581. https://doi.org/10.1038/nrg2383

Gilad, Y., and O. Mizrahi-Man, 2015 A reanalysis of mouse ENCODE comparative gene expression data. F1000 Res. 4: 121. https://doi.org/10.12688/f1000research.6536.1

GTEx Consortium, 2015 The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 348: 648–660. https://doi.org/10.1126/science.1262110

Horvath, S., 2011 *Weighted Network Analysis*. Springer, New York. https://doi.org/10.1007/978-1-4419-8819-5

Horvath, S., W. Erhart, M. Brosch, O. Ammerpohl, W. von Schönfels *et al.*, 2014 Obesity accelerates epigenetic aging of human liver. Proc. Natl. Acad. Sci. USA 111: 15538–15543. https://doi.org/10.1073/pnas.1412759111

Houseman, E. A., W. P. Accomando, D. C. Koestler, B. C. Christensen, C. J. Marsit *et al.*, 2012 DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics 13: 86. https://doi.org/10.1186/1471-2105-13-86

Jaffe, A. E., and R. A. Irizarry, 2014 Accounting for cellular heterogeneity is critical in epigenome-wide association studies. Genome Biol. 15: R31. https://doi.org/10.1186/gb-2014-15-2-r31

Johnstone, I., 2001 On the distribution of the largest eigenvalue in principal components analysis. Ann. Stat. 29: 295–327. https://doi.org/10.1214/aos/1009210544

Joo, J. W., J. H. Sul, B. Han, C. Ye, and E. Eskin, 2014 Effectively identifying regulatory hotspots while capturing expression heterogeneity in gene expression studies. Genome Biol. 15: r61. https://doi.org/10.1186/gb-2014-15-4-r61

Kang, H. M., C. Ye, and E. Eskin, 2008 Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. Genetics 180: 1909–1925. https://doi.org/10.1534/genetics.108.094201

Lappalainen, T., M. Sammeth, M. R. Friedländer, P. A. 't Hoen, J. Monlong *et al.*, 2013 Transcriptome and genome sequencing uncovers functional variation in humans. Nature 501: 506–511. https://doi.org/10.1038/nature12531

Lee, M. N., C. Ye, A. C. Villani, T. Raj, W. Li *et al.*, 2014 Common genetic variants modulate pathogen-sensing responses in human dendritic cells. Science 343: 1246980. https://doi.org/10.1126/science.1246980

Leek, J. T., and J. D. Storey, 2007 Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. 3: e161. https://doi.org/10.1371/journal.pgen.0030161

Leek, J. T., and J. D. Storey, 2008 A general framework for multiple testing dependence. Proc. Natl. Acad. Sci. USA 105: 18718–18723. https://doi.org/10.1073/pnas.0808709105

Leek, J. T., R. B. Scharpf, H. C. Bravo, D. Simcha, B. Langmead *et al.*, 2010 Tackling the widespread and critical impact of batch effects in high-throughput data. Nat. Rev. Genet. 11: 733–739. https://doi.org/10.1038/nrg2825

Li, B., and C. N. Dewey, 2011 RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12: 323. https://doi.org/10.1186/1471-2105-12-323

Li, Y. I., B. van de Geijn, A. Raj, D. A. Knowles, A. A. Petti *et al.*, 2016 RNA splicing is a primary link between genetic variation and disease. Science 352: 600–604. https://doi.org/10.1126/science.aad9417

Listgarten, J., C. Kadie, E. E. Schadt, and D. Heckerman, 2010 Correction for hidden confounders in the genetic analysis of gene expression. Proc. Natl. Acad. Sci. USA 107: 16465–16470. https://doi.org/10.1073/pnas.1002425107

Liu, X., J. A. Mefford, A. Dahl, M. Subramaniam, A. Battle *et al.*, 2018 GBAT: a gene-based association method for robust transgene regulation detection. bioRxiv. https://doi.org/10.1101/395970.

Lucas, J., C. Carvalho, Q. Wang, A. Bild, J. R. Nevins et al., 2006 Sparse statistical modelling in gene expression genomics, pp. 155–176 in *Bayesian Inference for Gene Expression and Proteomics*, edited by Do K.-A., and P. Muller. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511584589.009

Montgomery, S. B., M. Sammeth, M. Gutierrez-Arcelus, R. P. Lach, C. Ingle *et al.*, 2010 Transcriptome genetics using second generation sequencing in a Caucasian population. Nature 464: 773–777. https://doi.org/10.1038/nature08903

Nadler, B., 2008 Finite sample approximation results for principal component analysis: a matrix perturbation approach. Ann. Stat. 36: 2791–2817. https://doi.org/10.1214/08-AOS618

Parikshak, N. N., V. Swarup, T. G. Belgard, M. Irimia, G. Ramaswami *et al.*, 2016 Genome-wide changes in lncRNA, splicing, and regional gene expression patterns in autism. Nature 540: 423–427. https://doi.org/10.1038/nature20612

Parsana, P., C. Ruberman, A. E. Jaffe, M. C. Schatz, A. Battle *et al.*, 2017 Addressing confounding artifacts in reconstruction of gene co-expression networks. bioRxiv: https://doi.org/10.1101/202903.

Pickrell, J. K., J. C. Marioni, A. A. Pai, J. F. Degner, B. E. Engelhardt *et al.*, 2010 Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature 464: 768–772. https://doi.org/10.1038/nature08872

Pierce, B. L., L. Tong, L. S. Chen, R. Rahaman, M. Argos *et al.*, 2014 Mediation analysis demonstrates that trans-eQTLs are often explained by cis-mediation: a genome-wide analysis among 1,800 South Asians. PLoS Genet. 10: e1004818. https://doi.org/10.1371/journal.pgen.1004818

Rahmani, E., N. Zaitlen, Y. Baran, C. Eng, D. Hu *et al.*, 2017 Correcting for cell-type heterogeneity in DNA methylation: a comprehensive evaluation. Nat. Methods 14: 218–219. https://doi.org/10.1038/nmeth.4190

Rakyan, V. K., T. A. Down, D. J. Balding, and S. Beck, 2011 Epigenome-wide association studies for common human diseases. Nat. Rev. Genet. 12: 529–541. https://doi.org/10.1038/nrg3000

Rivas, M. A., M. Pirinen, D. F. Conrad, M. Lek, E. K. Tsang *et al.*, 2015 Effect of predicted protein-truncating genetic variants on the human transcriptome. Science 348: 666–669. https://doi.org/10.1126/science.1261877

Shin, S.-Y., E. B. Fauman, A.-K. Petersen, J. Krumsiek, R. Santos *et al.*, 2014 An atlas of genetic influences on human blood metabolites. Nat. Genet. 46: 543–550. https://doi.org/10.1038/ng.2982

Sparano, J. A., R. J. Gray, D. F. Makower, K. I. Pritchard, K. S. Albain *et al.*, 2015 Prospective validation of a 21-gene expression assay in breast cancer. N. Engl. J. Med. 373: 2005–2014. https://doi.org/10.1056/NEJMoa1510764

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

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

Stegle, O., S. A. Teichmann, and J. C. Marioni, 2015 Computational and analytical challenges in single-cell transcriptomics. Nat. Rev. Genet. 16: 133–145. https://doi.org/10.1038/nrg3833

Storey, J. D., 2003 The positive false discovery rate: a Bayesian interpretation and the q-value. Ann. Stat. 31: 2013–2035. https://doi.org/10.1214/aos/1074290335

Sun, Y., N. R. Zhang, and A. B. Owen, 2012 Multiple hypothesis testing adjusted for latent variables, with an application to the AGEMAP gene expression data. Ann. Appl. Stat. 6: 1664–1688. https://doi.org/10.1214/12-AOAS561

van den Brink, S. C., F. Sage, Á. Vértesy, B. Spanjaard, J. Peterson-Maduro *et al.*, 2017 Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. Nat. Methods 14: 935–936. https://doi.org/10.1038/nmeth.4437

van 't Veer, L. J., H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart *et al.*, 2002 Gene expression profiling predicts clinical outcome of breast cancer. Nature 415: 530–536. https://doi.org/10.1038/415530a

Wang, J., Q. Zhao, T. Hastie, and A. B. Owen, 2017 Confounder adjustment in multiple hypothesis testing. Ann. Stat. 45: 1863–1894. https://doi.org/10.1214/16-AOS1511

West, M., 2003 Bayesian factor regression models in the "large p, small n" paradigm, in *Bayesian Statistics*, Oxford University Press, Oxford.

Yao, C., R. Joehanes, A. D. Johnson, T. Huan, C. Liu *et al.*, 2017 Dynamic role of trans regulation of gene expression in relation to complex traits. Am. J. Hum. Genet. 100: 571–580. https://doi.org/10.1016/j.ajhg.2017.02.003

Zhu, Z., F. Zhang, H. Hu, A. Bakshi, M. R. Robinson *et al.*, 2016 Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat. Genet. 48: 481–487. https://doi.org/10.1038/ng.3538

Communicating editor: C. Sabatti