Correction for hidden confounders in the genetic analysis of gene expression

Jennifer Listgarten[a,1], Carl Kadie[b], Eric E. Schadt[c], and David Heckerman[a,1]

[a]Microsoft Research, 1100 Glendon Avenue, Suite PH1, Los Angeles, CA; [b]Microsoft Research, 1 Microsoft Way, Redmond, WA; and [c]Pacific Biosciences, 1505 Adams Drive, Menlo Park, CA

Edited* by David Haussler, University of California, Santa Cruz, CA, and approved July 21, 2010 (received for review February 26, 2010)

Understanding the genetic underpinnings of disease is important for screening, treatment, drug development, and basic biological insight. One way of getting at such an understanding is to find out which parts of our DNA, such as single-nucleotide polymorphisms, affect particular intermediary processes such as gene expression. Naively, such associations can be identified using a simple statistical test on all paired combinations of genetic variants and gene transcripts. However, a wide variety of confounders lie hidden in the data, leading to both spurious associations and missed associations if not properly addressed. We present a statistical model that jointly corrects for two particular kinds of hidden structure—population structure (e.g., race, family-relatedness), and microarray expression artifacts (e.g., batch effects), when these confounders are unknown. Applying our method to both real and synthetic, human and mouse data, we demonstrate the need for such a joint correction of confounders, and also the disadvantages of other possible approaches based on those in the current literature. In particular, we show that our class of models has maximum power to detect eQTL on synthetic data, and has the best performance on a bronze standard applied to real data. Lastly, our software and the associations we found with it are available at http://www.microsoft.com/science.

differential expression | genome wide association | microarray | population structure | expression heterogeneity

Just as it has become fruitful to systematically scan the genome for genetic markers of disease (i.e., Genome Wide Association Studies—GWAS), so too has it become fruitful to systematically scan the genome for genetic markers of change in gene-expression levels. For example, one might scan individual SNPs across the entire genome to see with which, if any, gene-expression levels a SNP is associated. The motivation for conducting such eQTL (for "expression Quantitative Trait Locus" (1)) studies is wide ranging (e.g., see recent reviews (2–5)). These studies are also referred to as Genetical Genomic or Genetics of Gene Expression—GOGE studies (3). From a basic biology perspective, such eQTL studies can shed light on how genes are regulated without the need for prior knowledge of particular regulatory mechanisms (3, 4). From the perspective of dissecting the genetics of a complex disease, eQTL studies can provide insight into how a disease-associated locus might be contributing to the disease of interest by providing information about which gene-expression pathway(s) the locus is affecting (2, 4). From a statistical perspective, one can use the results of an eQTL study to prioritize a list of disease-associated loci to follow up on; that is, one can use the existence of a SNP–gene-expression association as prior evidence that variation at the locus is more likely to have disease consequences (2, 6). Furthermore, eQTL studies can infuse causal information into gene-gene and protein-protein correlation networks by making use of the fact that DNA can affect gene expression, but not the other way around (1, 7–9). Finally, the utility of eQTL studies is likely to increase as larger and more diverse datasets are amassed, and with the advent of new technologies such as RNA sequencing and exon arrays (2).

Along with the potential pay offs of eQTL studies come major statistical challenges. In particular, we have to contend with challenges from two formerly distinct areas of statistical analysis: (*i*) GWAS and (*ii*) differential expression (i.e., searching for gene-expression levels that are associated with some variable of interest). In eQTL studies, each time we scan the genome for association with one gene's expression level, we are conducting a traditional GWAS scan, and hence the GWAS statistical pitfalls such as confounding by population structure must be dealt with (10–12, 13). In addition each time we ask if a particular genomic locus is associated with a gene's expression level in an eQTL scan we are in effect asking if that gene is differentially expressed between individuals with different alleles at that locus. Thus the statistical pitfalls of differential expression analysis must be dealt with, including confounding due to expression artifacts such as batch effects, and more generally, *expression heterogeneity* (4, 14–17). To our knowledge, no one has yet provided a coherent eQTL statistical framework to jointly tackle the issues of confounding across both of these formerly distinct areas. In this paper, we introduce such a framework and show the utility of it on both real and synthetic datasets. Additionally, within our framework, a method for correction of expression artifacts alone, in situations where the confounders are unknown, is introduced and shown to be superior to other methods in use today, such as Inter-sample Correlation Emended (ICE) (14) and Surrogate Variable Analysis (SVA) (15). Our software is available at http://www.microsoft.com/science.

To date, the utility of eQTL studies has been limited by the small number of individuals in the study (5). Thus, efforts are being ramped up to create much larger datasets, and so confounding factors will play an even larger (negative) role if not properly accounted for—both because larger datasets are likely to be more heterogeneous, and because they contain more power to reveal confounding structure (18, 19). Therefore, tackling these confounders in a rigorous way will help to pave the way for further discoveries in this burgeoning area.

Results

Our approach, which we call *LMM-EH-PS*, builds on a class of previous approaches for modeling hidden confounders in association studies called linear mixed-effects models (20). These models have been used previously either to correct for *population structure* (PS) such as race, family, or cryptic relatedness in GWAS studies (12, 13, 21), or, to correct for hidden expression artifacts

Author contributions: J.L., E.E.S., and D.H. designed research; J.L. performed research; J.L. and C.K. contributed new reagents/analytic tools; J.L. analyzed data; and J.L. and D.H. wrote the paper.

Conflict of interest statement: Eric E. Schadt is Chief Scientific Officer of Pacific Biosciences and owns stock in the company.

*This Direct Submission article had a prearranged editor.

Freely available online through the PNAS open access option.

[1]To whom correspondence may be addressed. E-mail: jennl@microsoft.com and heckerma@microsoft.com.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1002425107/-/DCSupplemental.

arising from technical, demographic, genetic, and environmental factors (14) (which we henceforth refer to as *expression heterogeneity* (15) or EH). Nonetheless, no one has used a mixed model approach, or any approach for that matter, to simultaneously correct for both types of confounders. Doing so requires combining and extending the current linear mixed modeling approaches, yielding a method well suited to analysis of eQTL studies as demonstrated shortly.

Roughly speaking, mixed-effects models work to tackle confounders in a two-step process. First, a set of similarities between every two individuals is computed, either in "SNP space" (for GWAS), or in "expression space" (for expression studies). Encoded in these similarities are the very confounders we seek to model—respectively PS and EH. For example, SNP-based similarity might encode information about race and relatedness between individuals, whereas expression-based similarity might encode information about technical artifacts. Second, these sets of similarities are then used in a regression model to tease apart the true eQTL associations from the spurious ones. Without such correction, these types of confounders are known to cause many false positives and also loss of power (10–13, 17, 18, 21–24).

Mixed-effects models are complemented in the association literature by principal components based approaches such as Eigenstrat, SVA, and similar algorithms (11, 15, 17, 22, 23, 25). Still coarser methods for PS correction include Structured Association (24), which clusters individuals and then uses the cluster labels as covariates in the association model of interest, and Genomic Control (18), which rescales the test statistic by a single factor to alleviate spurious associations from confounding. Because mixed-effects models are fully specified probabilistic generative models (26–28), adapting and extending them to new problems such as the joint correction tackled in this work is natural and intuitive.

We performed two main sets of experiments, both focused on datasets containing SNPs and microarray gene expression for the same individuals. The first set of experiments analyzed data from the human liver cohort dataset (29), of which we used the 378 individuals of only Caucasian descent, containing no detectable confounding population/relatedness structure. We used this dataset to demonstrate the utility of our method for correction of EH alone, that is, when PS is not present. This dataset contained 39,296 probes and 571,229 SNPs, although we worked with a subset of the data for evaluation purposes, as described with each experiment. In the second set of experiments, we focused on a mouse dataset (30) consisting of 188 male individuals across 19 strains of mice (16 inbred and 3 wild, with 7–11 individuals per strain) with 40,639 probes and 48,186 SNPs. We detected both PS and EH in this mouse data. Thus we applied our full model to this dataset, demonstrating its utility and also the weaknesses of other approaches. We evaluated the following types of models:

- *LMM-EH-PS*, our model (corrects for EH and PS)
- *LMM-EH*, our model (corrects for EH)
- *LMM-PS*, a linear mixed-effects model (corrects for PS)
- *ICE*, the model reported in ref. 14 (corrects for EH)
- *ICE-PS*, the model reported in ref. 14 with an additional PS component added into the linear mixed model (corrects for EH and PS)
- *SVA*, linear regression with SVA covariates computed as in ref. 15 (corrects for EH)
- *SVA-PS*, a linear mixed-effects model with SVA covariates computed as in ref. 15 (corrects for EH and PS)
- *LINREG* linear regression (no corrections)

We also compared a modification of SVA that outperformed SVA, as shown in the *SI Appendix*. Finally, Genomic Control was used as a postprocessing step to the models listed above, with little apparent benefit, also as shown in the *SI Appendix*.

As is common in the GWAS community, we summarized the departure of an observed p-value distribution from the theoretical null distribution by use of the so-called λ statistic. This statistic represents how much smaller the observed median p-value is compared to that expected in the theoretical null distribution (10, 18). Therefore, on data containing no (or very few) eQTL associations, $\lambda > 1$ suggests that the p-value distribution is *inflated* (too many small p-values), which can happen when a confounder such as PS is not properly addressed. In contrast, $\lambda < 1$, a much less common phenomenon, represents *deflated* p-values (too few small p-values). Of course small variations from $\lambda = 1$ occur even in synthetically generated datasets with no associations because of sampling error (i.e., finite data).

To more fully capture how an observed p-value distribution departs from the theoretical null distribution, beyond comparing just the median p-values, we also used the Kolmogorov-Smirnov (KS) test which better takes into account the full distribution of p-values.

Modeling of Expression Heterogeneity Alone. For experiments on human data, we randomly selected 6,000 SNP-probe hypotheses with which to work. Under the assumption that most of these pairs do not contain true associations, one would expect to observe a distribution of p-values similar to a theoretical null, that is, uniformly distributed on [0, 1] with $\lambda = 1$. Because eQTL analysis can be viewed as many GWAS problems analyzed together (one per gene-probe), this assumption is reasonable. Additionally, this assumption is further supported by the fact that results on synthetic null data were strikingly similar to those on the real data, as will be shown next.

The left column in Fig. 1 shows SNP-probe hypothesis p-value histograms for the real human data. Of interest is that ICE p-values were deflated as summarized by $\lambda = 0.93$, suggesting a problem with the method. This deflation can occur when an inconsistent estimator is used for model parameter K_{EH} (see *Materials and Methods*) in the mixed model and is discussed in detail in the *SI Appendix*. In contrast, when we used our model, LMM-EH, the deflation essentially disappeared, yielding $\lambda = 0.99$.

The right column in Fig. 1 shows the analogous p-value histograms on synthetic data containing no associations. The resemblance between the left and right columns in Fig. 1 is striking, suggesting that our synthetic data was representative of the real data. For example, the ICE model was still deflated, whereas our model (LMM-EH) was not. Of particular interest is that SVA appeared to have a comparable number of associations as on the real data. The SVA algorithm found 31 Surrogate Variables (SVs) on the synthetic data vs. 50 on the real data, using 100

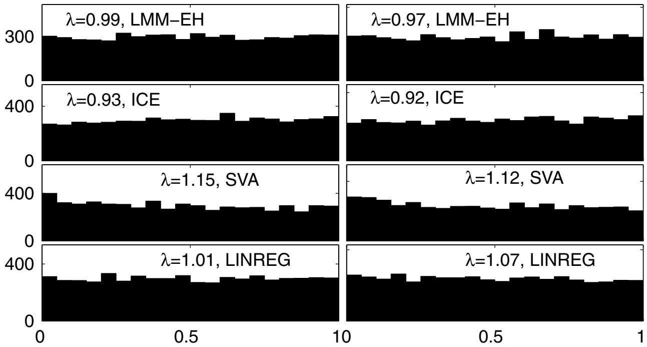

Fig. 1. *P*-value histograms for human data. Left column shows results on real-data: ICE p-values were deflated as indicated by $\lambda = 0.93$, whereas our model, LMM-EH, corrected this to $\lambda = 0.99$. The right column shows results on synthetic data: ICE p-values were deflated as indicated by $\lambda = 0.92$, whereas our model corrected this to $\lambda = 0.97$. Linear regression with SVA covariates was more inflated than linear regression alone.

permutations over a range of eigengene-significance from 0.01 to 0.5. Because this was synthetic data, however, we now know that these were spurious associations, suggesting that the associations SVA found on the real data were also spurious. Such behavior likely resulted from overfitting of the model parameters (see "*Overview of Synthetic Experiments*" and the *SI Appendix*). On the null synthetic data, a two-sided, one-sample KS test for uniformity indicated that the p-value distribution from our model (LMM-EH) was not significantly different from the theoretical null ($p = 0.26$), but that of ICE was ($p = 0.02$). These results suggest that our approach to correcting for EH was the only one among those tested that yielded calibrated p-values.

Fig. 2 shows the results of power experiments on the synthetic data containing 5% associations, at strengths found in the real data. The ICE model had power comparable to our model. However, in a setting where one does not know the true False Discovery Rate (FDR), one must estimate it and pick a p-value cut-off based on this estimated FDR. When we did so for the ICE method, this method underestimated its power because of its deflated p-values, resulting in a loss of hits for a given estimated FDR level (e.g., 60 for ICE rather than 80 with our model at an estimated FDR = 0.15, or 80 instead of 110 with our model at estimated FDR = 0.2). Our model achieved the highest number of true associations for a given estimated FDR cut-off.

Because there is no gold standard with which to evaluate eQTL analyses of real data, some researchers have advocated the use of a "cis-enrichment" (31) bronze standard. The rationale behind this score is as follows. When searching for eQTL associations between a SNP and a gene-probe level, there are two primary types of associations that one can find: (*i*) a local (or cis) association in which the SNP is close to the gene probe and thus likely to be acting *in cis*, and (*ii*) a distant (or trans) association in which the SNP is not close to the gene probe and thus likely acting *in trans* (4). It is commonly believed that cis-acting SNPs have stronger associations than trans-acting SNPs (3) because the mode of action is more direct. Consequently, it is sometimes assumed that cis associations should tend to get lower p-values than trans associations, and that the extent to which this is so is a reflection of the quality of the analysis (31).

We applied a cis-enrichment test (see *Materials and Methods*) to the real human data. In order to have enough power to detect differences between different models, we required enough potential cis eQTL (here defined as within 500 Kb of the gene, as was done in refs. 14, 30, 31) in the set of hypotheses tested and so focused on SNPs and gene probes in chromosome 1 rather than a random set of hypotheses across the genome. We selected every sixth SNP and all gene probes on chromosome 1. This procedure resulted in respectively 3,674 probe × 7,266 SNPs. Although each model had lower p-values for cis vs. trans hypotheses ($p < 1e-16$ by Mann Whitney), our model, LMM-EH, had significantly better cis enrichment than the other models (see *Materials and Methods*) as shown in Table 1. Additionally, the ranking of models provided by this evaluation matched the ranking implied by our power experiments on synthetic data, and also the ranking of models suggested by deviation from uniform of p-value histograms shown earlier.

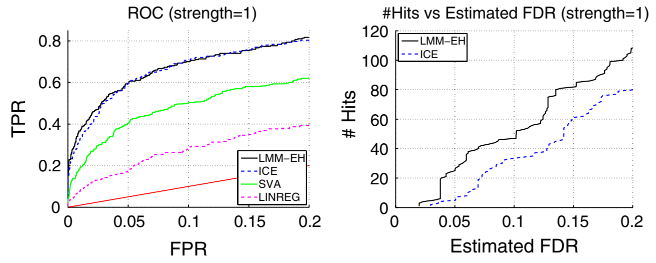

Fig. 2. Power curves for synthetic human data. The left plot shows the Receiver Operating Characteristic (ROC) curve, which displays the true positive rate (TPR) as a function of the false positive rate (FPR). This plot demonstrates that our model and ICE achieved similar power, surpassing linear regression with or without SVA covariates. The red line denotes what random guessing would have achieved. The right plot shows the number of associations called significant for each estimated FDR level (estimated as in ref. 15), demonstrating that in a real setting, ICE would be penalized for its deflated p-values ($\lambda = 0.93$) because they result in overly conservative FDR estimates.

Table 1. Human cis-enrichment, a starred *p*-value indicates that the starred model outperforms the nonstarred model

	ICE	LINREG	SVA
LMM-EH*	8.31e–3*	<1e–16*	9.85e–6*
ICE*		<1e–16*	1.01e–4*
LINREG*			8.27E–11

Modeling of Expression Heterogeneity and Population Structure. For experiments on mouse data, we again randomly selected 6,000 SNP-probe hypotheses which with to work. The left column of Fig. 3 shows p-value histograms on the real mouse data. We see that ICE-based models (ICE, ICE-PS) lead to dramatically deflated p-values ($\lambda = 0.71$ $\lambda = 0.63$) again pointing to a problem with the method. In contrast, when we used our model (LMM-EH-PS) the deflation disappeared, giving $\lambda = 1.02$.

The right column of Fig. 3 shows the analogous p-value histograms on synthetic data containing no true associations. The resemblance between the histograms on the real and synthetic data was again striking and suggests that our synthetic data was representative of the real data. Because the right column in Fig. 3 shows results on data with no associations, it is apparent that all models other than our joint model either produced serious inflation or deflation of p-values. Of particular interest is that when we added SVA-based covariates (with 24 SVs found on the synthetic datasets, and 29 on the real data, using 100 permutations and for a range of eigengene-significance from 0.01 to 0.5) to a PS-correcting mixed-effects model (SVA-PS), there was a large amount of inflation ($\lambda = 3.13$), more so than when we used

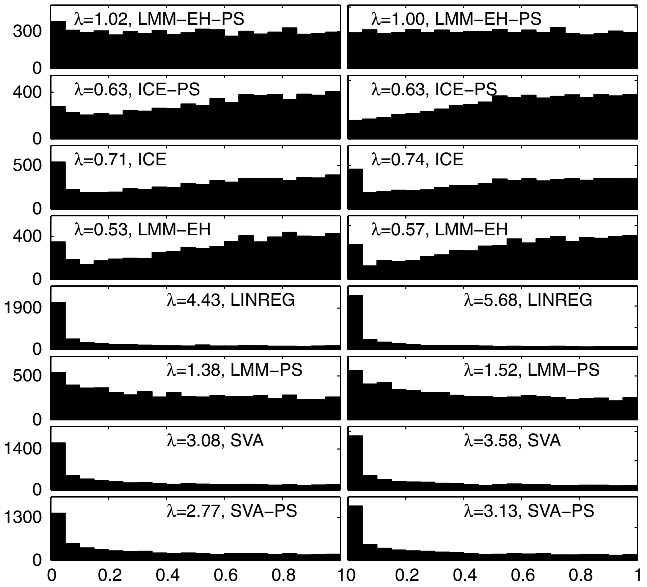

Fig. 3. *P*-value histograms for mouse data. Left column shows results on real-data: ICE-based p-values were deflated, as indicated by $\lambda \ll 1$. SVA-based models, LMM-PS, LMM-EH, and LINREG were inflated. Only our model (LMM-EH-PS) appeared to be calibrated, with $\lambda = 1.02$. Our model also indicated a small number of true associations (roughly 100 of 6,000 tests). The right column shows results on synthetic data: ICE-based p-values were deflated as indicated by $\lambda \ll 1$. Other models were inflated, and our model (LMM-EH-PS) appeared calibrated with $\lambda = 1.00$.

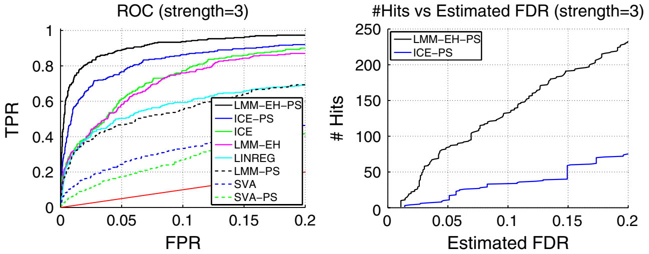

Fig. 4. Power curves for synthetic mouse data. The left plot shows the ROC curve, where our model LMM-EH-PS achieved maximum power. The red line denotes what random guessing would have achieved. The right plot illustrates how the best ICE-based model (ICE-PS), which yielded deflated p-values, penalized itself because of its overly conservative estimated FDR.

SVA with only linear regression (SVA). These results suggest that SVA can be lead further astray in the presence of PS.

Fig. 4 shows the results of power experiments on synthetic data which contained 5% associations, at strengths found in the real data multiplied by a factor of 3 (factors 1 and 5 are shown in the *SI Appendix*). At a strength multiplier of 1, all models recovered only a few associations. At a strength multiplier of 3, our model performed best ($p = 0.002$, based on our permutation test). At a strength multiplier of 5, this statistically significant win remained. Additionally, the plot showing *#Hits* vs. *Estimated FDR* indicates that ICE-PS was further crippled when forced to use its estimated FDR, which is highly conservative because of the deflated p-values. In particular, at an estimated FDR of 0.1, our model found around 130 hits, whereas ICE-PS found only about 40. Again, our model achieved the highest number of true associations for a given estimated FDR cut-off, and also for a given actual FDR.

We also applied our cis-enrichment test to the real mouse data, restricting the hypotheses tested to chromosome 1, as with the human data (using all SNPs and all gene probes on chromosome 1, resulting in 2,751 probe × 4,343 SNP hypotheses tested). Similarly to the cis-enrichment tests on the human data, although each model had lower p-values for cis vs. trans hypotheses ($p < 1e-16$ by Mann Whitney), our model, LMM-EH-PS, had significantly better cis enrichment than the other models, as shown in Table 2. Additionally, the ranking of models provided by this evaluation matched the ranking implied by our power experiments on synthetic data, and also the ranking of models suggested by deviation of p-value histograms shown earlier.

One other point worth mentioning is that upon a full analysis of the mouse data (the cross-product of all 40,639 gene probes× 48,186 SNPs), there was not a single common hypothesis among the top 5,000 ranked gene probes between LMM-EH-PS and linear regression. Between the two most similarly and best performing models, LMM-EH-PS and ICE-PS, there were 15 hypotheses in common among the top 5,000. Of these 15 hypotheses, the SNP fixed effect was always in the same direction and the correlation coefficient between them was 0.98. A list of the top 10,000 SNP–gene-probe pairs found by our model can be found on the same web page as our software.

eQTL Hotspots/Trans-Bands. It has been previously observed that eQTL analyses sometimes show eQTL hotspots/trans-bands, that is, SNPs that are associated with a large number of gene probes. It is believed that some of these may be spurious due to confounding effects, but that some may also be real—resulting from, for example, a SNP that affects expression of a transcription factor which in turn acts on many genes (e.g., (14, 16)). We looked for trans-bands using the same data as for the cis-enrichment tests. As shown in the *SI Appendix*, we found that in the human data, there were visually apparent trans-bands when linear regression was used, and subtle ones when SVA was used. Neither ICE nor LMM-EH, however, showed any trans-bands. On the mouse data, trans-bands were apparent in linear regression and LMM-PS, and more subtly in SVA and PS-SVA, but not in the other models (those that correct for EH). Under the assumption that our other evaluations of the models are correct, most if not all of the trans-bands observed here are likely spurious, because they appeared in the worst-performing models.

Discussion

We have introduced a statistical framework for joint correction of population structure and EH in eQTL studies showing that such a correction is needed and that other models that might naturally be applied to this problem do not perform well. Our evaluations suggest that our models, in comparison to other approaches, provide better calibrated p-values and maximum power of eQTL detection, both on synthetic data where ground truth is known, and on real data when using a cis-enrichment bronze standard.

Future work that would naturally follow from that presented here would be to extend beyond the search for pair wise SNP–gene associations to that of identifying multivariate relationships among genes, among SNPs, and among SNPs and genes, for example in the form of modules (32, 33), and also incorporating auxiliary information (33). Additionally, applying these kinds of ideas to eQTL studies involving copy number variation, or RNA sequencing data, or to combinations of data types would likely to be a productive endeavor.

Materials and Methods

Linear mixed-effects models (20) can be understood on an intuitive level from a variety of viewpoints. We view the linear mixed-effects model as a probabilistic generative model (26–28). Probabilistic generative models are a class of statistical models in which the semantics literally describe how one would generate observed data from the model for a fixed set of model parameters. Of course in practice we have the opposite situation—we have the observed data, but not the parameters of the model that generated it. Our task of parameter fitting is then to find the model parameters that, for example, make the observed data most likely (i.e., maximum likelihood). Once the parameters of the model have been fit, we can then use the fit of the data under different models—for example, those with and without a proposed SNP effect on one gene probe's expression—to generate p-values for particular SNP–probe eQTL hypotheses, using, for example, a likelihood ratio test (LRT) (13).

Our LMM-EH-PS model is set up to generate microarray expression data, whereas the SNP data is assumed to be fixed and is not generated by the model. This decision is motivated by the fact that SNP data can affect changes in expression data but not the other way around. In the generative model view of the mixed-effects model, the central idea is that each individual in the dataset is assigned a single number, u^i (for the ith individual) which represents where that individual lies in "confounder space." As a concrete example, if one had a dataset consisting of individuals with varying degrees of admixture between Caucasian and African, this single number might

Table 2. Mouse cis-enrichment, a starred p-value indicates that the starred model outperforms the nonstarred model

mouse	ICE-PS	LMM-EH	ICE	LMM-PS	LINREG	SVA	PS-SVA
LMM-EH-PS*	1.80e−014*	<1e−16*	<1e−16*	1.89e−008*	2.22e−015*	<1e−16*	<1e−16*
ICE-PS*		4.02e−14*	<1e−16*	0.163*	4.97e−5*	<1e−16*	<1e−16*
LMM-EH*			0.497	0.403*	0.1959*	<1e−16*	<1e−16*
ICE*				0.106	0.390*	<1e−16*	2.22e−16*
LMM-PS*					8.88e−5*	<1e−16*	<1e−16*
LINREG*						<1e−16*	2.44e−14*
SVA*							<1e−16

represent the fraction of say Caucasian ancestry in each person. For more complex types of PS, the space becomes more complex but the general idea remains the same. In order to generate "observed" gene-expression data for one gene probe, g (total of G gene probes), and all individuals, $\vec{y}_g$ (dimension $N \times 1$ for N individuals), from the model, one first samples these "confounder coefficients," $\vec{u}_g = [u_g^1,\ldots,u_g^N]$, and then treats them as covariates in a linear regression model (Eq. **1**) in order to generate the gene-expression values. One generates the confounder coefficients by way of a matrix, K (dimension $N \times N$), representing similarities between individuals. K may contain the similarity between all pairs of individuals as measured in "SNP space" if correcting for PS, or the similarity between all pairs of individuals in "expression space" if correcting for EH. The confounder coefficients are drawn from a zero-mean Gaussian with covariance K. This model is pictorially represented in the *SI Appendix* and is fully specified as follows,

$$\vec{u}_g|K \sim N(\vec{u}_g|\vec{0},K) \text{ generate the confounding coefficients}$$

$$\vec{e}_g \sim N(\vec{e}_g|\vec{0},I) \text{ generate the noise}$$

$$\vec{y}_g = X\vec{\beta}_g + \tau_g\vec{u}_g + \sigma_g\vec{e}_g \text{ compute gene expression data,} \qquad [1]$$

where ~ denotes that a variable is stochastically generated from the distribution to the right of this symbol; $N(\vec{r}|\vec{m},\Sigma)$ denotes a Gaussian distribution in $\vec{r}$ with mean $\vec{m}$ and covariance matrix Σ ; I denotes the identity matrix; σ_g (scalar) is the magnitude of the Gaussian residual noise; τ_g (scalar) is the magnitude of the confounding PS (or EH); $\vec{\beta}_g$ (dimension $Q \times 1$) is the effect of S SNPs and $Q - S$ other effects (e.g., bias/offset term and covariates such as gender and age) on the gene-expression level for gene-probe g, and X (dimension $Q \times N$) is the corresponding design matrix. See the *SI Appendix* for a way to explicitly include strain relationships in this model. Readers familiar with the conventional mixed model presentation will recognize $\vec{u}_g$ as a random effect X, as the fixed effects, and the likelihood function for this model, $p(\vec{y}_g|X,\vec{\beta}_g,\tau_g,\sigma_g,K) = N(\vec{y}_g|X\vec{\beta}_g,\tau_g^2K + \sigma_g^2I)$, which arises from integrating out $\vec{u}_g$. The likelihood of all of the gene-expression data, $Y = [\vec{y}_1,\vec{y}_2,\ldots,\vec{y}_G]$ (dimension $N \times G$), is given by $p(Y|\{\vec{\beta}_g,\tau_g,\sigma_g\},K) = \prod_g p(\vec{y}_g|X,\vec{\beta}_g,\tau_g,\sigma_g,K)$, where $\{\vec{\beta}_g,\tau_g,\sigma_g\}$ denotes the parameters over all gene probes. Because we only test for one SNP at a time, we do not actually incorporate all SNPs into the model jointly, instead restricting the fixed effects to include just one SNP at a time. Details of this model and on how to learn its parameters, $\{\vec{\beta}_g,\tau_g,\sigma_g\}$, are provided in the *SI Appendix*. We use maximum likelihood (ML) parameter fitting, and an LRT test to generate p-values from two models—one with a single SNP effect, and one without it. Our experiments have shown that use of REML (Restricted Maximum Likelihood) provides comparable results in this setting.

When one seeks to account for both PS and EH simultaneously, one need simply generate two sets of confounder coefficients, $\vec{u}_g|K_{\text{PS}} \sim N(\vec{u}_g|\vec{0},K_{\text{PS}})$ and $\vec{v}_g|K_{\text{EH}} \sim N(\vec{v}_g|\vec{0},K_{\text{EH}})$, independently, using the similarities between individuals in respectively SNP space (K_{PS}), or expression space (K_{EH}), and then add these into the regression model, yielding a likelihood, for one gene, g, of

$$\begin{aligned} p(\vec{y}_g|X,\vec{\beta}_g,\sigma_g,w_g,K_{\text{EH}},K_{\text{PS}}) \\ = N(\vec{y}_g|X\vec{\beta}_g,\tau_g^2[w_gK_{\text{EH}} + (1-w_g)K_{\text{PS}}] + I\sigma_g^2), \end{aligned}$$

where w_g in [0,1] is the relative weight of K_{EH} to K_{PS}. The log likelihood for all gene probes can be written in a similar manner to that shown earlier.

When correcting for both PS and EH simultaneously, we use a K_{PS} determined from the SNP data just as we do when correcting for PS alone. An Identity-By-Descent, Identity-By-State, and covariance matrix (11) have been used for K_{PS} (e.g., (12, 13, 21)). In separate experiments, we found all three yield comparable results. Thus, we used only the covariance matrix in this study.

When modeling EH, Kang et al. use the covariance matrix of the gene-expression data in their ICE model (14). This estimate of K_{EH} is inconsistent (see *SI Appendix*); and we have found that its use leads to deflated p-values, as seen in the *Results* section. Therefore we have developed an approach to correcting for EH with mixed-effects models in which we treat K_{EH} as a parameter to be learned in the model and use ML to fit it. This approach alleviates the problem of deflated p-values seen with ICE. We develop a new algorithm to estimate K_{EH} that combines coordinate-ascent and expectation-maximization (34).

We estimate parameters, $\{\vec{\beta}_g,\tau_g,w_g,\sigma_g\}$ and K_{EH} for all gene-probe models simultaneously by iterating between two steps. First, we identify the ML values of $\{\vec{\beta}_g,\tau_g,w_g,\sigma_g\}$ conditioned on a fixed value of K_{EH}. Then we identify the ML value of K_{EH} conditioned on fixed values of $\{\vec{\beta}_g,\tau_g,w_g,\sigma_g\}$. In each step the likelihood either increases or remains the same. Note that when learning K_{EH}, we omit use of the SNPs in X as is also done in ref. 15. We then incorporate use of the learned K_{EH}, and relearn the other parameters in the context of a now known K_{EH}, to evaluate SNP–gene-probe hypotheses. For our mouse dataset, estimation of K_{EH} took 10 h when parallelized across 1,100 processors. For our human dataset, estimation of K_{EH} took 5 h when parallelized across 1,100 processors. Further details about the joint EH and PS model, including time complexity, and also details about parameter fitting including K_{EH} are provided in the *SI Appendix*.

Because the goal of this paper is to demonstrate the strengths and weaknesses of different models that correct for confounding structure in our data, we have concentrated on additive SNP effects, encoding the pair of each SNPs an individual has by the number of wild-type alleles (as defined by the data itself). Our conclusions are likely to be insensitive to these restrictions. We also impute any missing SNPs as in ref. 11, and any missing gene-probe values by the median value for that probe.

Overview of Synthetic Experiments. At present there are no benchmark eQTL datasets with which to evaluate the success of different analyses. Following the work of others, we therefore worked with synthetic datasets (11–13). To generate synthetic expression data, we first fitted our model to the real data. Then, because our model is a generative model, we used the estimated parameters from the fitted models to generate gene-probe data. To generate a SNP-probe association, we used the SNP regression weight ($\vec{\beta}_g$) estimated for our model on the real data for the top 5% of SNP-probe hypotheses, multiplied by some strength factor to obtain a variety of strengths (e.g., strength = 1, 3, 5). Thus, our synthetic datasets contained 5% true associations.

Whereas one can never be absolutely certain of the relevance of synthetically-generated data to a real problem on hand, there are actions one can take to achieve relevance, and empirical assessments that can be made to gauge relevance. Forcing data generation to use only parameters estimated on real data provides some reassurance that the amount of structure (e.g., EH or PS) the model is generating is on the order of that found in the real data. Also, if results from analysis of synthetic data are similar to those achieved on the real data for a variety of experimental conditions, then one has reason to believe that the generative model has captured the important properties of the data well. As already mentioned, we restricted the synthetic data generation in this way—that is, to contain only the amount of PS and EH that our model inferred from the real data. Additionally, as seen in our experiments, the p-value distributions for real and synthetic data across all models used were strikingly similar. To further assess robustness of our model, we also generated null data from a linear mixed model with a PS correction and SVA covariates (PS-SVA) fitted to the real data, as reported in the *SI Appendix*. Interestingly, when we used that same model (PS-SVA) to analyze the data, we observed statistically significant inflation, which likely resulted from overfitting. In contrast, our model was able to successfully capture the confounding structure generated by PS-SVA, having produced a p-value distribution not significantly different from the null distribution.

Methods of Empirical Assessment Used in Results Section. The calibration (inflation/deflation) of p-values was assessed using p-value histograms, the one-sample KS test for uniformity, and the p-value distribution summary statistic λ. The λ statistic, pervasive in GWAS studies and used to help judge inflation/deflation of p-value distributions, is defined as the ratio of the median observed to median theoretical p-value, after conversion from p-value space to log-likelihood space by way of an inverse chi-square mapping (10, 18).

We evaluated power of our models applied to synthetic data in several complementary ways. One, we plotted Receiver Operating Characteristic (ROC) curves, which show the true positive rate (TPR) as a function of the false positive rate (FPR). Two, we plotted the number of associations found vs. the estimated FDR (computed using ref. 35). Here, we do not use the actual FDR which is possible to compute in our synthetic experiments, because we specifically want to show how methods with deflated p-values hurt themselves by causing their estimated FDR to be conservative—in real experiments one can only compute the estimated FDR, not the actual FDR. Three, we use a nonparametric permutation test equivalent to the one reported in ref. 36, using 1,000 randomizations, to determine whether the Area Under the Curve was different between pairs of models. For all three of these assessment techniques, we focused on the regime of interest for GWAS problems—that is, a relatively small FPR or FDR. In particular, we use only the portion of ROC curves where FPR < 20%; in plots of number of associations found vs. FDR, we show only the portion where estimated FDR < 20%. Note that we do not filter our inferred associations by whether or not they are deemed to be potentially cis rather than trans, nor do we select one among several SNPs in linkage disequilibrium (LD). Rather, we use all 6,000 p-values resulting from the 6,000 tests performed by each model in each experiment. Among the 6,000 SNPs chosen at random, we did not see much LD (e.g., mean Pearson correlation between all pairs of 6,000 mouse SNPs was 0.07 and standard deviation was 0.3).

To test which of two models was better able to assign lower *p*-values to cis eQTL hypotheses over trans hypotheses (i.e., our cis-enrichment test), we used a two-step procedure: (1) for each model, for each SNP, we used a one-tailed Mann-Whitney test to test the null hypothesis that the model ranked cis hypotheses no better than trans hypotheses, for hypotheses involving that SNP, (2) for each pair of models compared we used a two-tailed, paired Wilcoxon sign-rank test, on the *p*-values for all SNPs from Step 1, to test the null hypothesis that the median difference in the Mann-Whitney test *p*-values for each SNP is zero.

Data Sets. Our experiments were based on two datasets. The first was the Caucasian subset of the tissue-specific human liver cohort eQTL dataset reported and released in ref. 29. DNA samples were genotyped on the Affymetrix 500K SNP and Illumina 650Y SNP genotyping arrays. RNA samples were profiled on a custom Agilent 44,000 feature microarray. Expression data was processed as in ref. 29, with an additional step of median-imputing the gene-expression data, and filtering out SNPs with call rates less than 90%. Thus in total, our dataset contained 39,296 probes in the expression data and 571,229 SNPs. We used only those samples predicted as Caucasian in ref. 29, resulting in 378 individuals. All microarray data associated with the human liver cohort were previously deposited into the Gene-Expression Ominbus database under accession number GSE9588.

The second dataset was an eQTL dataset from 16 classical and 3 wild-derived inbred strains and reported and released in ref. 30. We used a total of 188 male individuals, with on average, 10 individuals per strain, and never fewer than 7 nor more than 11. Strains selected represented the distinct genealogies of inbred mice and includes eight Castle's mice (129S1/SvImJ, A/J, AKR/J, BALB/cByJ, C3H/HeJ, DBA/2J, NZB/BlNJ, and SM/J), three C57-related strains (C57BL/6J, C57BLKS/J, and C57L/J), four Swiss mice (FVB/NJ, NOD/LtJ, and SJL/J, SWR/J), three wild-derived inbred strains (CAST/EiJ, CZECHII/EiJ, and PERA/EiJ,) and one other inbred strain (LG/J). All microarray data are available in the NCBI GEO database under accession number GSE13870 and contains data for 40,639 probes. SNP data was obtained from the Broad Institute. SNPs with a minor allele frequency of less than 15% in the 16 classical inbred strains were removed. The resulting genotype dataset consists of 48,186 markers.

ACKNOWLEDGMENTS. We thank the anonymous reviewers for constructive suggestions, Jonathan Carlson for useful discussion and technical help, Oliver Stegle and Jimmie Chun for providing user-friendly HapMap data, Wan-Lin Su for help in accessing data, Ashish Kapoor and Chris Meek for useful discussion about Gaussian Process Regression, Hyun Min Kang for answering questions about his papers, Mac Manson and the High Performance Computing team for cluster support, Leigh Johnston for Matlab help, and Nico Pfeifer and Tony Gitter for catching typos.

1. Schadt EE, et al. (2003) Genetics of gene expression surveyed in maize, mouse, and man. *Nature* 422:297–302.
2. Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M (2009) Mapping complex disease traits with global gene expression. *Nat Rev Genet* 10:184–194.
3. Cheung VG, Spielman RS (2009) Genetics of human gene expression: mapping DNA variants that influence gene expression. *Nat Rev Genet* 10:595–604.
4. Gilad Y, Rifkin SA, Pritchard JK (2008) Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends Genet* 24:408–415.
5. Nica AC, Dermitzakis ET (2008) Using gene expression to investigate the genetic basis of complex disorders. *Hum Mol Genet* 17:R129–R134.
6. Grundberg E, et al. (2009) Population genomics in a disease targeted primary cell model. *Genome Res* 19:1942–1952.
7. Zhu J, et al. (2007) Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations. *PLoS Comput Biol* 3:e69.
8. Schadt EE, et al. (2005) An integrative genomics approach to infer causal associations between gene expression and disease. *Nat Genet* 37:710–717.
9. Suthram S, Beyer A, Karp RM, Eldar Y, Ideker T (2008) eQED: an efficient method for interpreting eQTL associations using protein networks. *Mol Syst Biol* 4:162.
10. Balding DJ (2006) A tutorial on statistical methods for population association studies. *Nat Rev Genet* 7:781–791.
11. Price AL, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38:904–909.
12. Yu J, et al. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* 38:203–208.
13. Kang HM, et al. (2008) Efficient control of population structure in model organism association mapping. *Genetics* 178:1709–1723.
14. Kang HM, Ye C, Eskin E (2008) Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. *Genetics* 180:1909–1925.
15. Leek JT, Storey JD (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet* 3:e161.
16. Michaelson JJ, Loguercio S, Beyer A (2009) Detection and interpretation of expression quantitative trait loci (eQTL). *Methods* 48:265–276.
17. Stegle O, Kannan A, Durbin R, Winn J (2008) Accounting for non-genetic factors improves the power of eQTL studies. . *Lecture Notes in Computer Science*, eds M Vingron and L Wong (Springer, Singapore), RECOMB, 4955, pp 411–422.
18. Devlin B, Roeder K (1999) Genomic control for association studies. *Biometrics* 55:997–1004.
19. Reich DE, Goldstein DB (2001) Detecting association in a case-control study while correcting for population stratification. *Genet Epidemiol* 20:4–16.
20. Demidenko E (2004) *Mixed models: theory and applications* (John Wiley and Sons, Inc., Hoboken, New Jersey).
21. Zhao K, et al. (2007) An Arabidopsis example of association mapping in structured samples. *PLoS Genet* 3:e4.
22. Li Q, Yu K (2007) Improved correction for population stratification in genome-wide association studies by identifying hidden population structures. *Genet Epidemiol* 32:215–226.
23. Lee AB, Luca D, Klei L, Devlin B, Roeder K (2009) Discovering genetic ancestry using spectral graph theory. *Genet Epidemiol* 34:51–59.
24. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000) Association mapping in structured populations. *Am J Hum Genet* 67:170–181.
25. Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLoS Genet* 2:2074–2093.
26. Heckerman D (1998) A tutorial on learning with Bayesian networks. *Learning in graphical models* (Kluwer, Cambridge, MA), pp 301–354.
27. Pearl J (1988) *Probabilistic reasoning in intelligent systems: networks of plausible inference* (Morgan Kaufmann, San Mateo, CA).
28. Bishop CM (2006) *Pattern recognition and machine learning (Information Science and Statistics)* (Springer).
29. Schadt EE, et al. (2008) Mapping the genetic architecture of gene expression in human liver. *PLoS Biol* 6:e107.
30. Su W-L, et al. (2010) Assessing the prospects of genome-wide association studies performed in inbred mice. *Mamm Genome* 21:143–152.
31. McClurg P, et al. (2007) Genomewide association analysis in diverse inbred mice: power and population structure. *Genetics* 176:675–683.
32. Chen Y, et al. (2008) Variations in DNA elucidate molecular networks that cause disease. *Nature* 452:429–435.
33. Lee SI, et al. (2009) Learning a prior on regulatory potential from eQTL data. *PLoS Genet* 5:e1000358.
34. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the em algorithm. *J Roy Stat Soc B Met* 39:1–38.
35. Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. *Proc Natl Acad Sci* 100:9440–9445.
36. Carlson JM, et al. (2008) Phylogenetic dependency networks: inferring patterns of CTL escape and codon covariation in HIV-1 Gag. *PLoS Comput Biol* 4:e1000225.