Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions

Sarah Margaret Urbut [1,2], Gao Wang [1], Matthew Stephens [1,3,‡]

1 Department of Human Genetics/ University of Chicago, Chicago, IL USA
2 Pritzker School of Medicine/Growth and Development Training Program/University of Chicago, Chicago, IL USA
3 Department of Statistics/ University of Chicago, Chicago, IL USA

Abstract

We introduce new statistical methods for analyzing genomic datasets that measure many effects in many conditions (e.g. gene expression changes under many treatments). These new methods improve on existing methods by allowing for arbitrary correlations in effects among conditions. This flexible approach increases power, improves effect-size estimates, and facilitates more quantitative assessments of effect-size heterogeneity than simple "shared/condition-specific" assessments. We illustrate these features through a detailed analysis of locally-acting ("cis") eQTLs in 44 human tissues. Our analysis identifies more eQTLs than existing approaches, consistent with improved power. More importantly, although eQTLs are often shared broadly among tissues, our more quantitative approach highlights that effect sizes can vary considerably among tissues: some shared eQTLs show stronger effects in a subset of biologically-related tissues (e.g. brain-related tissues), or in only a single tissue (e.g. testis). Our methods are widely applicable, computationally tractable for many conditions, and available at `https://github.com/stephenslab/mashr`.

Introduction

Genomic studies often involve estimating and comparing many effects across multiple conditions or outcomes. Examples include studying changes in expression of many genes under multiple treatments [1]; or differences in histone methylation at many genomic locations in multiple cell lines [2]; or the effects of many genetic variants on risk of multiple diseases [3]; or the impact of many eQTLs in multiple cell-types or tissues [4–6]. In these settings an initial goal is often to identify "significant" non-zero effects. Another important goal is to compare effects, and to identify differences in effect among conditions – sometimes referred to as "interactions". For example, in eQTL (expression Quantitative Trait Locus) studies, researchers are often interested in identifying tissue-specific effects, in the belief that they may have particular biological relevance.

The simplest, and perhaps most common, analysis strategy for such studies is to analyze the data in different conditions one at a time, and then compare the overlap of "significant" results in different conditions. Although appealingly simple, this "condition-by-condition" approach is unsatisfactory in several ways. For example, it can substantially under-represent sharing of effects among conditions, because many shared effects will be insignificant in some conditions just by chance. And when effects are shared among conditions it completely fails to exploit this, limiting its overall power [5].

To address these deficiencies of condition-by-condition analyses, several groups have developed methods for *joint* analysis of effects in multiple conditions (e.g. [2, 5–13]). The simplest of these methods build on traditional meta-analysis methodology [8, 9], and assume that the non-zero effects are present in every condition. Other methods are more flexible, allowing for condition-specific effects, for sharing of effects among subsets of conditions, and for heterogeneity in the shared effects [5, 6, 12]. Many of these methods also adapt themselves to the data at hand by learning patterns of sharing from the data, using a hierarchical model [5].

Nonetheless, existing methods remain limited in important ways. First, all of them make relatively restrictive assumptions about the correlations among non-zero effects. For example, [5] assumes correlations are non-negative, and that the non-zero effects are equally correlated among all conditions. In some applications correlations may be negative: for example, genetic variants that increase one trait may tend to decrease another. And, often, some subsets of conditions will be more correlated than others: for example, in our eQTL application (below) effects in brain tissues are more correlated with one another than with effects in non-brain tissues. Second, the most flexible methods are

computationally intractable for moderate numbers of conditions (e.g. 44 tissues in our eQTL application), and existing solutions to this problem substantially reduce flexibility. For example, [5] solves the computational problem by restricting effects to be shared in all conditions, or specific to a single condition. Alternatively, [12] allows for all possible patterns of sharing in an elegant computationally-tractable way, but only under the more restrictive assumption that the non-zero effects are uncorrelated among conditions, which will often not hold in practice. Third, existing methods typically focus only on *testing* for significant effects in each condition, and not on *estimating effect sizes*. As we illustrate here, estimating effect sizes can be essential to assessing heterogeneity of effects among conditions. Finally, software implementions of existing methods are often tailored to a specific application, making it harder to apply to other settings: for example, `eQTL-BMA` [5] is primarily designed for eQTL applications, whereas `corMotif` [12] is tailored to differential expression analyses. An exception here is the `metasoft` software [9, 14], which is designed to be *generic* in that it requires only effect estimates and their standard errors in multiple conditions, making it easily applicable to a wide range of settings.

Here we introduce more flexible statistical methods that combine the most attractive features of existing approaches, while overcoming their major limitations. The methods, which we refer to as "multivariate adaptive shrinkage" (`mash`), build on recent work in [15] for testing and estimation of effects in a *single* condition, and extend them to *multiple* conditions. Key features of `mash` include: i) It is *flexible*, allowing for both shared and condition-specific effects, and capable of capturing stronger correlations in effects among some conditions than others; ii) It is *computationally tractable* for hundreds of thousands of tests in (at least) dozens of conditions; iii) it provides not only measures of significance, but also *estimates of effect sizes*, together with measures of uncertainty; iv) It is *adaptive*, meaning that its behavior adapts to the patterns present in the particular data set being analyzed; and v) It is *generic*, requiring only a matrix containing the observed effects in each condition, and a matrix of their corresponding standard errors. (Indeed `mash` can work with just a matrix of Z scores, although that reduces the ability to estimate effect sizes.) Together these features make `mash` the most flexible and widely-applicable method available for estimating and testing multiple effects in multiple conditions.

As its name suggests, `mash` is built on the statistical concept of "shrinkage". Here shrinkage refers to modifying estimates towards some value – often towards zero – to improve accuracy. There are many good justifications for shrinkage, and it is widely viewed as a powerful statistical tool. However, it is seldom used in genomics applications. This may be due to the difficulty of deciding precisely *how much* to shrink. The "adaptive shrinkage" method in [15] solves this

problem in univariate settings by *learning from the data* how much to shrink. Here we extend this to multivariate settings. Shrinkage in the multivariate setting is more complex than in the univariate setting, but also potentially more useful. In particular, the multivariate setting provides the opportunity not only to shrink estimates towards zero (which improves accuracy if most effects are small), but also to shrink effects in related conditions towards one another (which improves accuracy when effects are similar among conditions). This focus on multivariate shrinkage estimation, and more generally on joint estimation of effects across multiple conditions, distinguishes `mash` from existing approaches that focus primarily on testing for non-zero effects. Estimation is particularly useful in settings where, as in our eQTL application here, there is considerable sharing of effects among conditions, but where effect sizes also vary considerably.

To demonstrate the potential for `mash` to provide novel insights we apply it here to analyse (*cis*) eQTL effects in 16,069 genes across 44 human tissues. Compared with previous analyses of human eQTLs among multiple tissues [4–6], our analysis involves many more tissues, and provides more insight into sharing of effects by examining variation in eQTL effect sizes among tissues. Focussing on the strongest "cis" eQTLs in each gene – which are the easiest to reliably assess – we find that the majority are shared among large numbers of tissues, in that their effects tend to be consistent in sign (positive or negative) across tissues. However, at the same time, effect sizes can vary considerably among tissues. Reassuringly, biologically-related tissues tend to show more correlated effects; for example, effects are often quite similar among the different brain tissues. Our analyses of variation in estimated effects among tissues suggest that assessments of "tissue-specific" vs "tissue-consistent" effects should pay attention to effect sizes, and not only to tests of significance.

Methods Overview

Multivariate adaptive shrinkage (`mash`)

Our method, `mash`, is designed to estimate the effects of many units in many conditions (n units in R conditions say). It takes as its input two $n \times R$ matrices, one containing "effect" estimates and the other containing their corresponding standard errors. For example, in the GTEx data analyzed here we consider the effects of hundreds of thousands of potential eQTLs (rows) in $R = 44$ tissues (columns). The method assumes that the true effects are centered on 0, and indeed allows that many effects – possibly the vast majority – may be at, or very near, zero. That is, the true effect matrix may be sparse. It also allows that some of the non-zero effects may be 'shared', being similar (though not necessarily identical) among conditions, while others may be 'specific' to only a subset of conditions. Although we illustrate `mash` on an eQTL application, it is sufficiently flexible to apply to most contexts involving many multivariate effects.

The `mash` method is an Empirical Bayes method with two steps: i) use all the observed data to learn typical patterns of sparsity, sharing and correlations among effects; ii) use these learned patterns to produce improved effect estimates, and corresponding measures of significance, for each unit in each condition. Step ii) is reasonably straightforward: it involves applying Bayes theorem to combine the background information (learned patterns of sharing from Step i)) with the observed data for each effect (the estimates and standard errors in every condition). Step i) is the difficult part, and where the primary innovations of our work lie. Specifically, we introduce a flexible model that allows for sparsity of effects and correlations among non-zero effects, and introduce a novel and efficient two-step approach to fitting this model.

Our flexible model uses a mixture of multivariate normal distributions that allows for a range of effect sizes and patterns of correlation. Specifically, each R-vector of effects across conditions, $\boldsymbol{b}$, is assumed to come from a mixture distribution,

$$p(\boldsymbol{b}; \boldsymbol{\pi}, \boldsymbol{U}) = \sum_{k=1}^{K} \sum_{l=1}^{L} \pi_{k,l} \; N_R(\boldsymbol{b}; \boldsymbol{0}, \omega_l U_k), \tag{1}$$

where $N_R(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the multivariate normal density in R dimensions with mean $\boldsymbol{\mu}$ and variance covariance matrix Σ; each U_k is a covariance matrix that captures some common "pattern" of (potentially-correlated) effects; each ω_l is a scalar scaling coefficient that corresponds to a different "size" of effect; and the mixture proportions $\pi_{k,l}$ determine the relative frequency of each pattern-size combination. The scaling coefficients ω_l take values on a fixed dense grid that

spans "very small" to "very large", to capture the full range of effects that could occur (the goal is that the grid is sufficiently large and dense that adding more values to it will not change results; see [15]).

To fit this model, we use a novel two-step procedure illustrated in Figure 1:

i-a) Generate a large list of candidate covariance matrices $U_k = (U_1, \ldots, U_K)$. This list includes both "data-driven" estimates, and "canonical" matrices that have simple interpretations. The data-driven estimates are obtained by applying covariance estimation methods [16], and dimension reduction techniques (e.g. Principal components analysis, and sparse factor analysis [17]) to a subset of the effects matrix, specifically the rows of the effect matrix that have the largest (univariate) effects. The canonical matrices we use include the identity matrix (representing independent effects across conditions); a matrix of all 1s (representing effects that are equal in all conditions); and R matrices that represent effects that are specific to condition r ($r = 1, \ldots, R$). See Detailed Methods for details.

i-b) Given this list, estimate $\boldsymbol{\pi}$ by maximum likelihood (using *all* observed effects, not only those used in Step i-a)).

The intuition is that Step i-a) can be relatively *ad hoc*, with the goal of producing a large list of matrices, only some of which may effectively capture key patterns in the data. Step i-b) is more formal, being based on the principle of maximum likelihood, and can rescue imperfections in Step i-a) by giving very low weight to covariance matrices that are not well supported by the data. Step i-b) is also the place where the overall sparsity of effects is taken account of: if most effects are zero, or very small, then this step will put most weight on very small effects (i.e. small scaling coefficients, ω). This modular approach has several attractive features. For example, Step i-b) is a convex optimization problem, and so can be solved efficiently and reliably for large problems. And if researchers have ideas for additional ways to generate candidate matrices in Step i-a), these are easily plugged into the procedure.

The model (1) is quite flexible, and includes many existing methods for this problem as special cases (Detailed Methods). One potential drawback of flexible models is the possibility of "overfitting". To address this we used a cross-validation procedure which trains the model on a random subset of the data (rows of the matrix) and then assesses its fit on the remaining data ("test data"). In practice we found overfitting not to be a major concern - that is, in general, we found that using more U_k typically improved, or at least did not harm, test set performance (e.g. Supplementary Figure 7). Thus, although `mash` is flexible, it is not *too* flexible. A still more flexible model could be obtained by

estimating the means of the multivariate normal distributions in (1), rather than setting them to 0, but this would substantially increase the potential for overfitting.

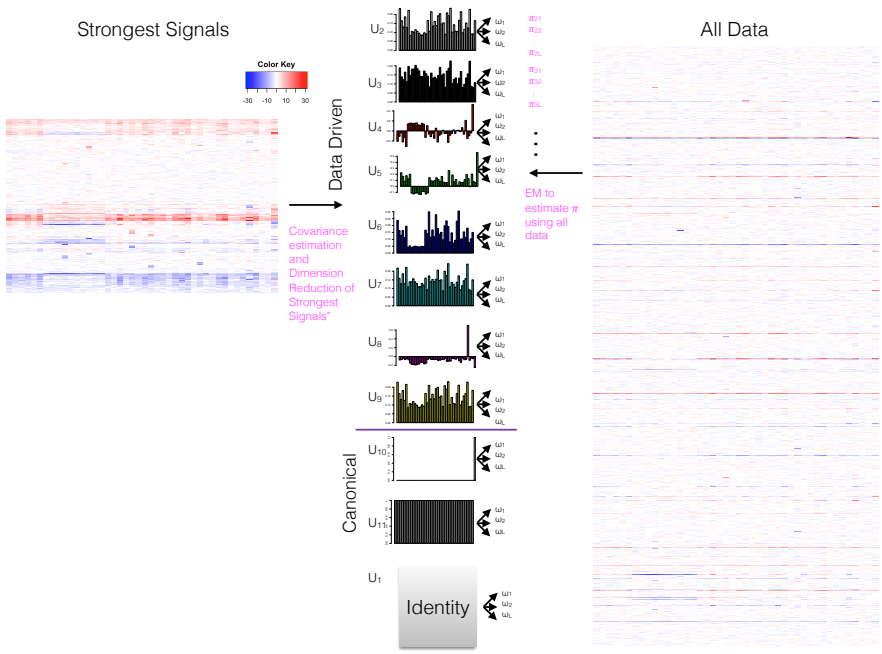


Figure 1. Overview of fitting procedure in `mash`, which estimates the multivariate distribution of effects present in the data. The data (**right**) consist of a matrix of effect size estimates for a large number of units (rows) in multiple conditions (columns), together with their corresponding standard errors (here assumed to be 1 for each effect for simplicity). Colors (red/blue) indicate the sign of the effects (positive/negative), with shading intensity indicating size of effect. First, using the rows containing the strongest signals (**left**), we apply covariance estimation and dimension-reduction methods to estimate candidate "data-driven" covariance matrices (here $U_2, \ldots, U_9$). To these we add several "canonical" covariance matrices, including the identity matrix, and matrices representing condition-specific effects. Each covariance matrix represents a "pattern" of effects that may occur in the data (summarized visually here by the first eigenvector, although each matrix is actually $R \times R$). We then scale each covariance matrix by a grid of scaling factors ω_l, varying from "very small" to "very large", which allow for effect sizes to range from very small to very large. Finally, using the whole data set (**right**), we use maximum likelihood estimation to estimate weights (relative frequencies) $\pi_{k,l}$ for each (ω_l, U_k) combination; this corresponds to estimating how commonly each pattern–effect size combination occurs.

Results

Improved effect size estimates

An important novelty of our method, `mash`, is its focus on *estimation of effect sizes*, in contrast with most existing multivariate analysis methods which focus only on *testing* for non-zero effects. Furthermore, `mash` is more than just an extension of existing methods to estimate effect sizes, because the underlying model (1) is more flexible than models underlying existing methods – and, indeed, includes existing models as special cases.

To illustrate the potential for multivariate analysis to improve accuracy of effect size estimates we performed simple simulations and compared three approaches to effect size estimation:

1. `mash`, the method we describe here.

2. A simpler version of our method, `mash-bmalite`, which represents an extension of existing methods to estimation of effect sizes. Specifically `mash-bmalite` performs effect size estimation based on the BMAlite models from [5], which include both the random effects models (RE and RE2) and fixed effects model (FE) used in the software `metasoft` [14]. These models allow for shared effects of equal size across all conditions (FE), shared effects of varying size across conditions (RE, RE2), and condition-specific effects (i.e. effects that occur in only one condition). Even though this, in itself, would represent a useful contribution, `mash` is more flexible than this. Specifically, `mash` can learn from the data that some subsets of conditions are more correlated than others, due to its use of data-driven covariance matrices in (1).

3. `ash` [15], which is a univariate analogue of `mash` designed to estimate effect sizes using results from a single condition. Results from `ash` are obtained by applying it separately to each condition, and so represent what can be achieved by a simple "condition-by-condition" analysis. This is included as a baseline against which to quantify the benefits of multivariate analysis.

We applied these three methods to estimate effect sizes under two scenarios:

1. "Shared, structured effects": data were simulated using the model (1), based on the fit of this model to the GTEx eQTL data below (see Methods for details). In this scenario effects tend to be shared among many conditions, and furthermore these shared effects are highly "structured", in that they are often similar in size (or at least sign), with

the similarity being greater among some subsets of conditions than others. For example, in the GTEx analysis later we see that effect sizes are often particularly similar among the subset of brain-derived tissues. This scenario will arise frequently in practice, and an important goal of our work is to provide methods that perform well here.

2. "Shared, unstructured effects": in this scenario effects are shared among all conditions (i.e. either every condition shows an effect, or no condition shows an effect), but the effect sizes and directions of the *non-zero* effects are independent across conditions. We aim to show that even in this unstructured setting `mash` provides improved effect estimates compared with an analogous univariate (condition-by-condition) approach, and in this case acts essentially as an extension of existing methods to estimate effect sizes.

In each case we simulate a 20,000 by 44 matrix of data $\hat{B}$ containing 20,000 estimated effects in each of 44 conditions (and their associated standard errors). We assume that non-null effects are rare: of the 20,000 effects, only 400 are non-null. Thus the matrix of effects is sparse, with non-zero values concentrated in a small number of rows.

Figure 2a (See also Supplementary Table 1) compares the accuracy of effect size estimates, as measured by the relative root mean squared error (RRMSE) (22), which is the RMSE of the estimates, divided by the RMSE achieved by simply using the original observed estimates $\hat{B}$ for the effects. Thus an RRMSE < 1 indicates that the method produces estimates that are more accurate than the original observations $\hat{B}$. As expected, the joint (multivariate) methods outperform the univariate method in both scenarios, due to their combining information across conditions. Furthermore, `mash` substantially outperforms the other methods in the "structured effects" scenario, and performs like `mash-bmalite` in the unstructured case. That is, the flexibility of `mash`, which is responsible for its improved performance in the structured setting, does not decrease performance in this simpler setting.

In all settings, all three methods have RRMSE< 1, indicating a substantial improvement in accuracy compared with the original observed effects $\hat{B}$. This improvement can come from two sources: i) the methods shrink estimated effects towards zero, which improves average accuracy because most effects are indeed null; ii) in the presence of "structured effects", the multivariate methods can share information across conditions to improve accuracy. For example, if a particular effect is shared, and similar in size, across a subset of conditions then averaging the observed effects in those conditions will improve estimation

accuracy. Both these factors help explain the strong performance of `mash` in the structured effects setting (Supplementary Table 1).

As a check on implementation we also applied the three methods to data simulated under an "Independent effects" scenario, in which *all* effects are entirely independent across conditions, with no greater sharing than expected by chance. (Note that this is very different from the "shared, unstructured" scenario, where only the non-zero effects are independent.) We used this to confirm the intuition that in such settings the univariate method that analyzes each condition independently should perform best, as indeed it does (Supplementary Table 1).

Improved detection of significant effects

In addition to effect estimates, `mash` also provides a measure of significance for each effect in each condition. Specifically `mash` estimates the "local false sign rate" (*lfsr*) [15], which is the probability that the estimated effect has the incorrect sign. The *lfsr* is analogous to the local false discovery rate [18], but more stringent in that it insists that effects be correctly signed to be considered "true discoveries". Similarly `mash-bmalite` can estimate the *lfsr*, but under its less flexible model; and `ash` can estimate the *lfsr* separately in each condition.

We used the same simulations as above to illustrate the gains in power to detect significant effects that come from the flexible multivariate model in `mash`. Figure 2b shows the trade-off between false positive and true positive discoveries for each method as the significance threshold is varied. The relative performance of the methods precisely mirrors the RRMSE results: multivariate methods perform best, and `mash` outperforms other methods for detecting shared structured effects. Further, in the "shared, structured" scenario `mash` is finding essentially all ($> 99\%$) of the signals that the other methods find, plus additional signals (Supplementary Table 2). (And in the "shared, unstructured" scenario `mash` and `mash-bmalite` not only have similar average performance, but are finding almost identical signals; Supplementary Table 2.)

Comparison with `metasoft`

Among existing software packages for this problem, `metasoft` [14] is in some ways the most comparable with `mash`. In particular, it is both generic – requiring only effect estimates and their standard errors – and computationally tractible for $R = 44$. The `metasoft` software implements several different multivariate tests for association analyses, each corresponding to different multivariate models for the effects. For example, the FE model assumes that the

effects in all conditions are equal; the RE2 model assumes that the effects are normally distributed about some common mean, with deviations from that mean being independent among conditions [19]; and the BE model is an extension of the RE2 model that allows that some effects are exactly zero [14]. These models are similar to the BMAlite models from [5], and none of them capture the kinds of structured effects that can be learned from the data by `mash`. Our comparisons above illustrate the benefits of the more flexible model in `mash`. However, because differences in software implementation sometimes lead to unanticipated differences in performance we also performed some simple direct benchmarks comparing `mash` and `mash-bmalite` with `metasoft`.

Specifically we compared these methods in the simplest type of multivariate test: separating the null from the non-null signals, where here null means zero effect in *all* conditions. Here, for each model (FE, RE2, and BE), `metasoft` produces a p value for each multivariate test, whereas `mash` and `mash-bmalite` produce a Bayes Factor (see Methods); in each case these can be used to rank the significance of the tests. Figure 2c) shows the trade-off between false positive and true positive discoveries for each method as the significance threshold is varied in the same simulation scenarios as above. In both cases `mash` is the most powerful method, again illustrating the benefits of its more flexible model.

Assessing heterogeneity and sharing in effects

In analyses of effects in multiple conditions, it is often desired to identify effects that are shared across many conditions, or, conversely, those that are specific to one or a few conditions. This turns out to be a particularly delicate task. For example, [5] emphasize that the simplest approach – first identifying significant signals separately in each condition, and then examining the overlap of the significant effects – can very substantially under-estimate sharing. This is due to incomplete power: by chance, a shared effect can easily be significant in one condition and not in another. To address this [5, 6] estimate sharing among conditions as a parameter in a joint hierarchical model, which takes account of incomplete power. However, these approaches are infeasible for $R = 44$. Furthermore, even for smaller values of R they have some drawbacks. In particular they are based on a "binary" notion of sharing, i.e. whether or not an effect is non-zero in each condition, and so do not capture differences in magnitude, or even signs, of effects among conditions. If effects that are shared among conditions actually differ greatly in magnitude – for example, being very strong in one condition and weak in all others – then this would seem important to know.

Here we address this problem with a new approach based on assessing

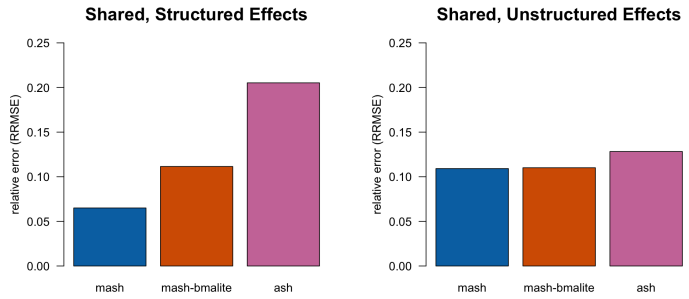
quantitative similarity of effects. Specifically, we assess sharing of effects in two ways: i) "sharing by sign" (estimates have the same sign); and ii) "sharing by magnitude" (effects are similar in magnitude). Here we define similar in magnitude to mean both the same sign and within a factor of 2 of one another (although other thresholds could be used, and in some settings – for example, where the "conditions" are different phenotypes – the requirement that effects have the same sign may best be dropped.) These measures of sharing can be computed for any pair of conditions, and an overall summary of sharing across conditions can be obtained by assessing how many conditions share with some reference condition (here, we use the condition with the largest estimated effect as the reference).

These measures of sharing could be naively estimated from the raw observed effect estimates from each condition; however, errors in these effect estimates will naturally lead to errors in assessed sharing. Because `mash` combines information across conditions to improve effect estimates (see above) it can provide more accurate estimates of sharing. To illustrate this we used the "shared structured" simulations to compare accuracy of overall estimates of sharing from `mash` with those from the raw effect estimates, as well as with `ash` and `mash-bmalite`. Table 1 summarizes these results, which confirm the improved accuracy of `mash`. For example, `mash` reduces the error in the estimated number of conditions sharing by sign from 4.7 for the raw estimates to 2.4 for `mash`.

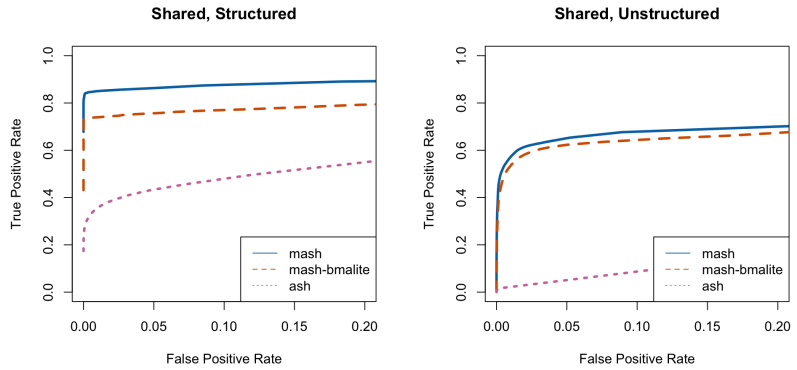
Errors in estimates of sharing for simulated data

Method	sign	magnitude
raw	4.7	4.3
`ash`	4.7	5.7
`mash-bmalite`	3.7	4.4
`mash`	2.4	2.3

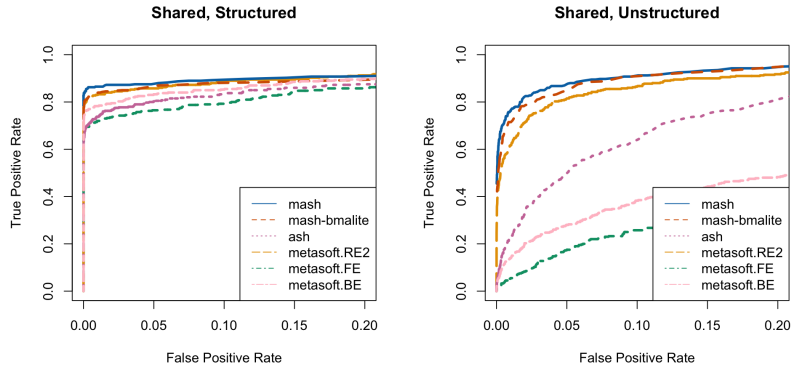
Table 1. For each method we computed the mean absolute error of the estimated number of conditions that share effects (by either sign or magnitude) with the condition with the largest estimated effects. Here "raw" indicates the performance of the raw estimates being input into each method. Each number is a mean error across the 400 non-null effects from the "shared structured effects" scenario.



(a) Accuracy of effect estimates (RRMSE).



(b) Detection of non-null effects in *each* condition (ROC curves).



(c) Detection of non-null effects in *any* condition (ROC curves).

Figure 2. Comparison of methods on simulated data. Results are shown for two simulation scenarios: "shared structured" effects, where the non-zero effects are shared among the 44 conditions in complex structured ways similar to patterns in the GTEx data; and "shared unstructured" effects, where the non-zero effects are shared among the 44 conditions with effect sizes that are independent among conditions. Panel (a) shows accuracy of effect estimates. Panels (b) and (c) show ROC curves for detecting significant effects. The primary difference between (b) and (c) is that in (b) each effect is treated as a separate discovery in each condition (which requires condition-specific measures of significance), whereas in (c) each effect is treated as a single discovery across all conditions (which requires only a single measure of significance, as in traditional meta-analyses). In (b) we require the estimated sign (+/-) of each significant effect to be correct to be considered a "true positive". Our new method (`mash`) outperforms other methods, particularly for "shared structured" effects, a scenario expected to be common in genomics applications.

GTEx cis-eQTL analysis

To illustrate the benefits and flexibility of `mash` in a substantive application we applied it to analyse expression Quantitative Trait Loci (eQTLs) across 44 human tissues/cell-types, using data from the Genotype Tissue Expression (GTEx) project [20]. The GTEx project aims to provide insights into the mechanisms of gene regulation by studying human gene expression and regulation in multiple tissues from health individuals. One fundamental question is which SNPs are eQTLs (i.e. associated with expression) in which tissues. Answering this could help distinguish regulatory regions and mechanisms that are specific to a few tissues vs shared among many tissues. It could also help with analyses that aim to integrate eQTL results with GWAS results to help identify the tissues that are most relevant to any specific complex disease (e.g. [20, 21]).

As input to `mash` we use a matrix of eQTL effect estimates $\hat{b}_{jr}$, and corresponding standard errors $\hat{s}_{ij}$, where the rows j index different SNP-gene pairs and the columns r index tissues (or cell types). We used the effect estimates and standard errors for candidate local ("cis") eQTLs for each gene, distributed by the GTEx project (v6 release). These were obtained by (univariate) single-SNP analyses in each tissue by applying `MatrixEQTL` [22] on expression levels that have been corrected for population structure (using genotype principal components [23]) and for other confounding factors affecting expression data (both measured factors such as age and sex, and unmeasured factors using factor analysis [24]), and then rank-transformed to the corresponding quantiles of a standard normal distribution. Thus the effect size estimates are in units of standard deviations on this transformed scale. Because, like most eQTL analyses, these estimates were obtained by single-SNP analysis, the estimated effects for each SNP actually reflect the effects of both the SNP itself and other SNPs in LD with it. Thus our analyses here do not distinguish causal eQTLs from SNPs that are in LD with the causal eQTLs; see Discussion.

We analysed the 16,069 genes for which univariate effect estimates were available for all 44 tissues we considered; the filtering criteria used ensure that these genes show at least some indication of expression in all 44 tissues.

Increased flexibility of `mash` improves model fit

Since the true effects are unknown we cannot compare models based on accuracy of effect estimates. Therefore, we instead illustrate the gains of the more flexible `mash` model using cross-validation: we fit each model to a random subset of the data ("training set") and assessed model fit by its log-likelihood computed on

the remaining data ("test set"). Comparing `mash` and `mash-bmalite` in this way we found that `mash` with a correlated residual framework improved the test set likelihood by 23,725 log-likelihood units, indicating a very substantial improvement in fit. Further, `mash` placed 79% of the mixture component weights on the data-driven covariance matrices, indicating that our methods for estimating these matrices are sufficiently effective that they capture most effects better than do the canonical matrices used by existing methods.

Identification of data-driven patterns of sharing

The increased flexibility of `mash` comes from its use of "data-driven" components to capture the main patterns of sharing (actually, covariance) of effects. This is illustrated in Figure 3, which shows the majority component that `mash` identifies in these data (relatively frequency 34%). The main patterns captured by this component are: i) effects are positively correlated among all tissues; ii) the brain tissues (and, to some extent, testis and pituitary) are particularly strongly correlated with one another, and less correlated with other tissues; iii) effects in whole blood tend to be somewhat less correlated with other tissues. Other components identifed by `mash` are shown in Supplementary Figure 2. Some of these components also have positive correlations among all tissues and/or highlight heterogeneity between brain tissues and other tissues, confirming these as very common features in these data. However, other components also capture rarer patterns, such as effects that are appreciably stronger in one tissue than others (Supplementary Figure 5).

Patterns of sharing inform effect size estimates

Having estimated patterns of sharing from the data, `mash` exploits these patterns to improve effect estimates at each putative eQTL. Although we cannot directly demonstrate improved average accuracy of effect estimates in the real data (for this, see simulations above), individual examples can provide helpful intuition into the way that `mash` achieves improved accuracy. In this vein, Figure 4 shows three illustrative examples, which we discuss in turn.

In the first example, the vast majority of effect estimates are positive in each tissue, with the strongest signals in a subset of brain tissues. Based on the patterns of sharing learned in the first step, `mash` estimates the effects in all tissues to be positive – even those with negative observed effects. This is because the few modest negative effects at this eQTL are outweighed by the strong background information that effects are highly correlated among tissues. Humans are notoriously bad at weighting background information against

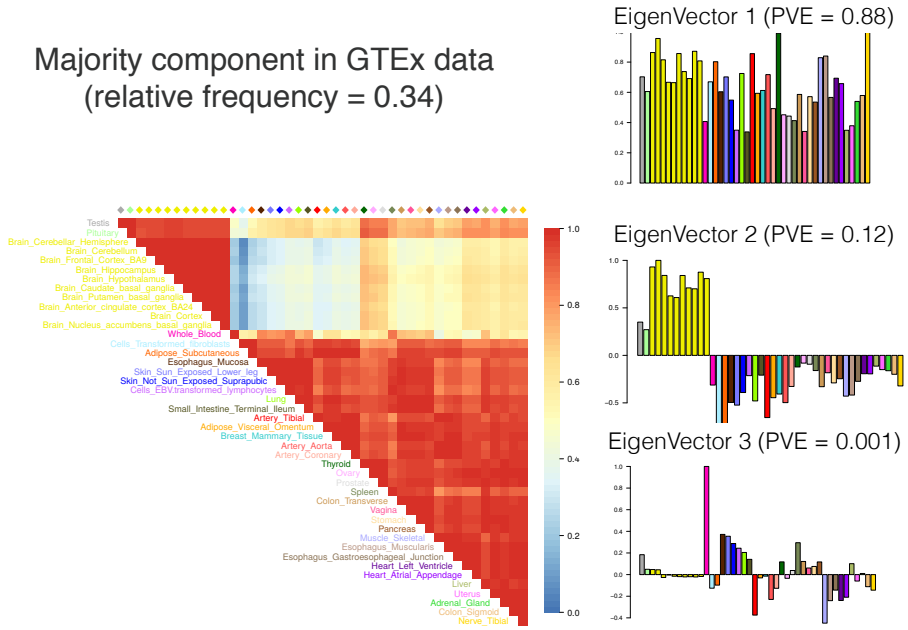


Figure 3. Summary of primary patterns identified by `mash` in GTEx data. Shown are the heatmap of the correlation matrix, and barplots of the first 3 eigenvectors, of the covariance matrix U_k corresponding to the dominant mixture component identified by `mash`. This component accounts for 34% of all weight in the GTEx data. In all cases, tissues are color-coded as indicated in the heatmap legend. The first eigenvector reflects broad sharing among tissues, with all effects in the same direction; the second eigenvector captures differences between brain (and, to a less extent, testis and pituitary) vs other tissues; the third eigenvector primarily captures effects that are stronger in whole blood than elsewhere.

specific instances [25] – they tend to underweight background information when presented with specific data – so this behavior may or may not be intuitive to the reader. But `mash` performs this weighting using Bayes rule, which is ideally suited to this job. The `mash` effect estimates are also appreciably larger in brain tissues than in other tissues. Again, this is the result of using Bayes rule to combine the effect estimates for *this* eQTL with the background information on heterogeneity among brain and non-brain effects learned from *all* eQTLs.

In the second example, the effect estimates in non-brain tissues are mostly (30/34) positive, but modest in size, and only one effect is, individually, nominally significant ($p < 0.05$). However, combining information among tissues, `mash` effect estimates in non-brain tissues are all positive, and mostly "significant" (*lfsr*< 0.05). In contrast the data in brain tissues are inconsistent, with a mix of both positive and negative effect estimates. `mash` concludes that we cannot be confident of the eQTL effect sign in brain tissues. This example illustrates how `mash` can learn from the data how to group conditions, rather than treating them equally. In this case `mash` has learned that effects in brain tissues are sometimes different from the other tissues, and hence avoids jumping to strong conclusions in the brain based on signal in other tissues.

In the final example, effect estimates vary in sign, and are modest except for a very strong signal in whole blood. While whole-blood-specific effects are estimated to be rare, `mash` (again, through Bayes theorem) recognizes that the strong data at this eQTL outweigh this background information, and estimates a strong effect in blood with insignificant effects in other tissues. This illustrates how `mash`, although focussed on combining information among tissues, can still recognize – and clarify – tissue-specific patterns when they exist.

Increased identification of significant effects

Our simulations demonstrated that the more flexible model behind `mash` can increase power to detect significant effects. To illustrate the effects of this here we compare the number of significant eQTLs detected by `mash` with those detected by our modified `mash-bmalite` and `ash`. To avoid double-counting of eQTLs in the same gene that are in LD with one another we assess the significance of only the "top SNP" in each gene, which we define to be the SNP with the largest (univariate) $|Z|$-statistic across all tissues. Thus we focus on 16,069 putative eQTLs, each with effect estimates in 44 tissues, for a total of 707,036 effects.

The vast majority of top SNPs show a very strong signal in at least one tissue (97% have a maximum $|Z|$ score exceeding 4), consistent with most of these genes containing at least one eQTL in at least one tissue. However, the

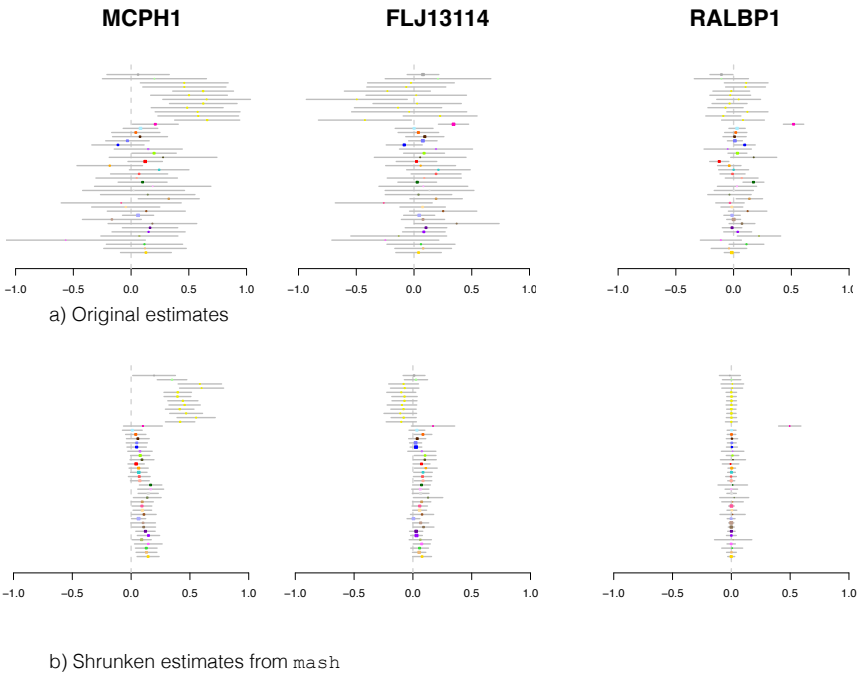


Figure 4. Examples illustrating of how `mash` uses patterns of sharing to inform effect estimates in the GTEx data. In panel a) each colored dot shows the raw effect estimate for a single tissue (color-coded as in Figure 3), with grey bar indicating $\pm$ 2 standard errors. These are the data input into `mash`. Panel b) shows the corresponding estimates output by `mash` (posterior mean, $\pm$ 2 posterior standard deviations). In each case `mash` combines information across all tissues, using the background information – patterns of sharing – it has learned from data on all eQTLs, to produce more precise estimates. Together, these three examples illustrate the flexibility of `mash` in combining information across different subsets of tissues for different eQTLs, depending on how their data match different patterns of sharing identified in the overall data. See main text for detailed discussion.

univariate tissue-by-tissue analysis (`ash`) identifies only 13% of these effects as "significant" at *lfsr*<0.05; that is, the univariate analysis is highly confident in the sign of the effect in only 13% of cases. In comparison `mash-bmalite` identifies 39% as significant at the same threshold, and `mash` identifies 47%. As in the simulations, the significant associations identified by `mash` include the vast majority (96%) of those found significant by either of the other methods (Supplementary Table 3). Thus, the multivariate methods identify the most significant effects, with `mash` identifying the most.

Overall, `mash` found 76% (12,189/16,069) of the top SNPs to be significant in at least one tissue. We refer to these as the "top eQTLs" in subsequent sections.

Sharing of effects among tissues

To investigate sharing and heterogeneity of the top eQTLs among tissues we used the quantitative measures of sharing introduced above: sharing of effects by sign and by magnitude. The results are summarized in Table 2 and Figure 5. Because a major feature of these data is that brain tissues generally show more similar effects than non-brain tissues we also show results separately for these subsets of tissues. The results confirm extensive eQTL sharing among tissues, particularly among the brain tissues. Sharing in sign exceeds 85% in all cases, and is as high as 98% among the brain tissues. (Furthermore, these numbers may underestimate the sharing in sign of actual causal effects, because of the potential effects of multiple eQTLs per gene in LD; see Supplementary Text.) Sharing in magnitude is inevitably lower, because sharing in magnitude implies sharing in sign. Overall, on average 37% of tissues show an effect within a factor of 2 of the strongest effect at each top eQTL. However, within brain tissues this number increases to 78%. That is, not only do eQTLs tend to be shared among the brain tissues, but the effect sizes tend to be quite homogeneous. Because these results are based on only the top eQTLs at each gene they reflect patterns of sharing among strong cis eQTLs; it is possible that weaker eQTLs may show different patterns of heterogeneity among tissues.

Of course, some tissues share eQTLs more than others. Figure 6 summarizes eQTL sharing by magnitude between all pairs of tissues (see Supplementary Figure 4 for sharing by sign). In addition to strong sharing among brain tissues, `mash` also identifies increased sharing among other biologically-related groups, including: arteries (tibial, coronary and aortal), two groups of gut tissues (one group containing esophagus and sigmoid colon; the other containing stomach, terminal ilium of the small intestine and transverse colon), skin (sun-exposed and non-exposed), adipose (Subcutaneous and Visceral-Omentum) and heart (left ventricle and atrial appendage). This figure also reveals that the main

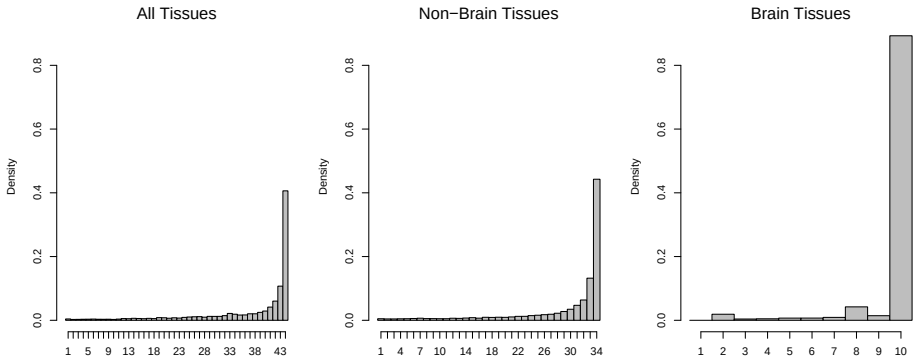
source of heterogeneity in effect sizes among brain tissues is in cerebellum vs non-cerebellum tissues, and also emphasizes sharing between the pituitary and brain tissues.

Different levels of effect sharing among tissues means that effect estimates in some tissues gain more precision than others from the joint analysis. To quantify this we computed an "effective sample size" (ESS) for each tissue that reflects the typical precision of its effect estimates (Supplementary Figure 1). The ESS values are smallest for tissue that show more "tissue-specific" behaviour (e.g. testis, whole blood; see below), and are largest for coronary artery, reflecting its stronger correlation with other tissues.

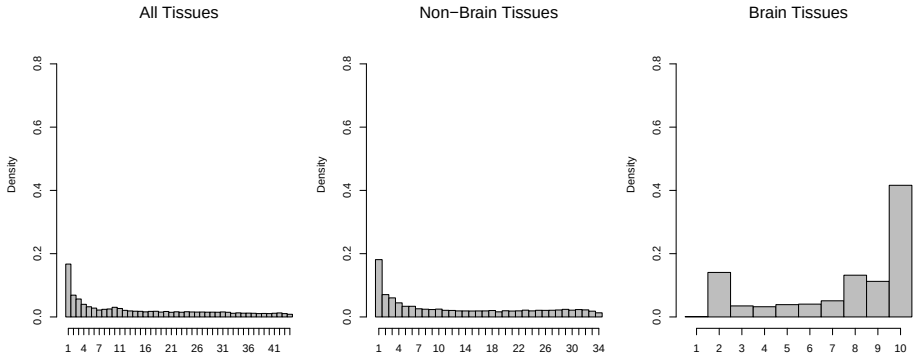
Tissue-specific eQTLs

Despite high average levels of sharing of eQTLs among tissues, `mash` also identifies eQTLs that are relatively "tissue-specific". Indeed, the distribution of the number of tissues in which an eQTL is shared by magnitude has a mode at 1 (Figure 5), representing a subset of eQTLs that have much stronger effect in one tissue than in any other (henceforth "tissue-specific" for brevity). Breaking down this group by tissue (Supplementary figure 5) identifies Testis as the tissue with the most tissue-specific effects. Testis also stands out, with whole blood, as having lower pairwise sharing of eQTLs with other tissues (Figure 6). Other tissues showing stronger-than-average tissue specificity (in either Supplementary Figure 5 or Figure 6) include skeletal muscle, thyroid, and transformed cell lines (fibroblasts and LCLs).

One possible explanation for tissue-specific eQTLs is tissue-specific expression. That is, if a gene is strongly expressed only in one tissue this could explain why an eQTL for that gene might show a strong effect only in that tissue. Whether or not a tissue-specific eQTL is due to tissue-specific expression could considerably impact biological interpretation. Thus we assessed whether tissue-specific eQTLs identified here could be explained by tissue-specific expression. Specifically, we took genes with tissue-specific eQTLs, and examined the distribution of expression in the eQTL-affected tissue relative to expression in other tissues. We found this distribution to be similar to genes without tissue-specific eQTLs (Supplementary Figure 6). Thus most tissue-specific eQTLs identified here are not simply reflecting tissue-specific expression.



(a) Number of Tissues Shared By Sign.



(b) Number of Tissues shared by Magnitude

Figure 5. Histogram showing estimated number of tissues in which top eQTLs are "shared" by two different definitions, a) sign and b) magnitude. Sharing by sign means that eQTL have the same sign of effect; Sharing by magnitude means that they also have similar effect size (within a factor of 2). Left: All tissues; Center: non-brain tissues; Right: brain tissues.

Table 2. Summary of sharing among top eQTLs

Data	All Tissues	Non-Brain	Brain
Shared by Sign ($\tilde{b} > 0$)	0.85	0.85 (0.88)	0.96 (0.98)
Shared by Magnitude: ($\tilde{b} > 0.5$)	0.36	0.40 (0.44)	0.76 (0.85)

Summary of sharing among top eQTLs. Numbers show the proportion of effects meeting a given sharing criterion. "Shared by sign" requires that the effect has the same sign as the strongest effect among tissues. "Shared by Magnitude" requires that the effect is also within a factor of 2 of the strongest effect. Numbers in parentheses are obtained by a secondary `mash` analysis of subsets of tissues.

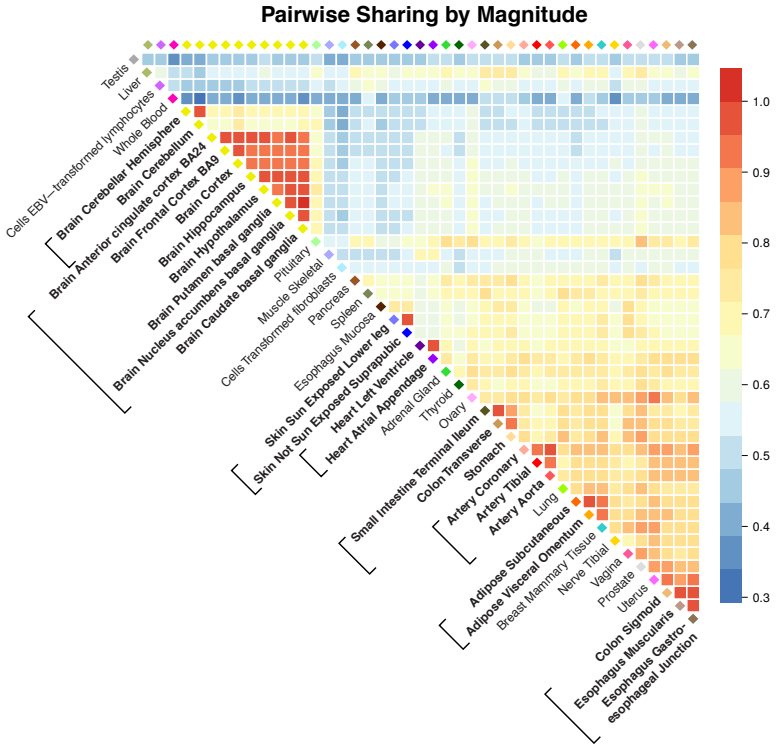


Figure 6. Pairwise sharing by magnitude of eQTL among tissues. For each pair of tissues we consider the top eQTLs that are significant in at least one of the two tissues, and plot the proportion of these that are "shared in magnitude" – that is, have effect estimates that are the same sign and within a factor of 2 of one another. Pink triangles highlight groups of biologically-related tissues mentioned in the text as showing particularly high levels of sharing.

Discussion

The statistical benefits of joint multivariate analyses compared with univariate analyses are well documented, and increasingly widely appreciated. But we believe this potential nonetheless remains under-exploited in practice. Our aim here is to provide a set of flexible and general tools to help in such analyses, and we designed `mash` with this aim in mind. In particular, `mash` is *generic* and *adaptive*. It is generic in that it can take as input any matrix of Z scores (or, better, a matrix of effect estimates and their corresponding standard errors) testing many effects in many conditions. For example, the effect estimates we used in our GTEx analysis came from simple linear regressions, but it would be perfectly possible to use `mash` with estimates from other approaches, such as generalized linear models or linear mixed models for example [14]. And `mash` is adaptive in that it learns patterns of sharing of multivariate effects from the data, allowing it to maximize power and precision for each setting. Consequently `mash` should be very widely applicable. Indeed, although genomics applications form our primary motivation, `mash` could be useful in any setting involving testing and estimation of multivariate effects.

At its core, `mash` uses an Emprical Bayes hierarchical model, and so is related to other methods that use this approach, including [5, 6, 12]. Indeed, the `mash` framework essentially includes these previous methods as special cases (as well as simpler methods such as "fixed effects" and "random effects" meta-analyses [9, 26]). However, one key feature that distinguishes `mash` from these previous methods is that `mash` puts greater focus on *quantitative* estimation and assessment of effects. More specifically, whereas previous methods have focussed on "binary" models for effects – that is, effects are either present or absent in each condition – `mash` focusses instead on allowing for and assessing quantitative variation among effects. This move away from binary-based models has at least two advantages. First, allowing for all possible binary configurations can create computational challenges. Second, in practice we have found that data often show widespread sharing of effects among many conditions, and that in such settings binary-based methods tend to conclude that effects are non-zero in most or all conditions, even when the signal is very modest in some conditions. This conclusion may not be technically incorrect – for example, in our GTEx analysis it is not impossible that all eQTLs are somewhat active in all tissues. However, as our analysis here illustrates, a more quantitative focus can reveal variation in effect sizes that may be of considerable biological importance.

One important limitation of our eQTL analysis is that it does not distinguish between SNPs that causally affect expression, and those that are merely

associated with expression due to being in LD with a causal SNP. This limitation also applies to most previous multi-tissue eQTL analyses, and indeed to most single-tissue eQTL analyses. This issue is particularly important to appreciate when cross-referencing, say, GWAS associations with eQTL effect estimates: a GWAS-associated SNP may be a "significant" eQTL simply because it is in LD with another causal SNP. For single-tissue eQTL mapping, this problem has been addressed in several ways. These include the development of (single-phenotype) fine-mapping methods that attempt to distinguish causal from non-causal effects [27–32], and also co-localization methods [33–35] that attempt to assess whether the same causal SNP may explain an observed association signal in two different phenotypes (e.g. GWAS and gene expression). For multi-tissue analysis, only more limited attempts exist to address this problem. For example, eQTLBMA [5] implements a Bayesian approach to fine-mapping under the simplifying assumption of at most causal SNP per gene [27, 28]. It would be straightforward to adapt `mash` to also perform fine-mapping under this assumption. However, although this simplifying assumption seems a reasonable starting point, it becomes decreasingly plausible in analyses that involve large numbers of tissues, and we view the development of more flexible fine-mapping multi-tissue eQTL methods as an important and challenging problem for future work.

One potentially powerful extension of `mash` would be to allow for the patterns of each effect to depend on covariates. For example, in an eQTL context, one might wish to allow functional annotations – such as the distance of the SNP from the transcription start site, or its coding/non-coding status – to affect the prior distributions on patterns of sharing or sizes of effects. Furthermore, one would want to estimate the effects of these covariates from the data [28, 36]. One possible way forward here would be to allow the mixture proportions π in `mash` to depend on covariates through a logistic link. However, this appears a challenging problem, and a fully satisfactory solution may require considerable further ingenuity.

Dealing with multiple tests is often described as a "burden". This description likely originates from the fact that controlling family-wise error rate (the probability of making even one false discovery) requires more and more stringent thresholds as the number of tests increases. However, most modern analyses prefer to control the false discovery rate (FDR) [37], which (under weak assumptions) does not depend on the number of tests [38]. Consequently the term "burden" is inaccurate and unhelpful. Indeed, we believe that the availability of results of many tests in many conditions should be viewed not as a burden, but an *opportunity*: specifically, an opportunity to learn about the relationships among underlying effects, and consequently to make data-driven

decisions that help improve both power to detect effects and precision of effect estimates. Approaches along these lines will inevitably, it seems, involve modelling assumptions, and the goal should be flexible models that are capable of dealing with a wide range of situations that can occur in practice. The methods presented here represent a substantial step towards this goal.

Software implementing our method is available at `http://github.com/stephenslab/mashr`. Scripts for generating results from the paper are at `https://github.com/surbut/gtexresults_mash`.

Materials and Methods

Model and Fitting

Let b_{jr} $(j = 1, \ldots, J; r = 1, \ldots, R)$ denote the true value of effect j in condition r. Further let $\hat{b}_{jr}$ denote the (observed) estimate of this effect, and $\hat{s}_{jr}$ the standard error of this estimate, so $z_{jr} := \hat{b}_{jr}/\hat{s}_{jr}$ is the usual z statistic for testing whether b_{jr} is zero. Let B, $\hat{B}, S$ and Z denote the corresponding $J \times R$ matrices, and let $\boldsymbol{b}_j$ (respectively $\hat{\boldsymbol{b}}_j, \boldsymbol{z}_j$) denote the jth row of B (respectively $\hat{B}, Z$).

We assume the vector $\hat{\boldsymbol{b}}_j$ is normally distributed about the true effects $\boldsymbol{b}_j$, with variance-covariance matrix V_j (defined below), and that the true effects follow (1). That is,

$$p(\hat{\boldsymbol{b}}_j | \boldsymbol{b}_j, V_j) = N_R(\hat{\boldsymbol{b}}_j; \boldsymbol{b}_j, V_j), \tag{2}$$

$$p(\boldsymbol{b}_j | \boldsymbol{\pi}, \boldsymbol{U}) = \sum_{k=1}^{K} \sum_{l=1}^{L} \pi_{k,l} \; N_R(\boldsymbol{b}_j; \boldsymbol{0}, \omega_l U_k). \tag{3}$$

where $N_R(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the density of the R-dimensional multivariate normal (MVN) distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and the scaling parameters $\omega_1, \ldots, \omega_L$ are fixed on a dense grid (detailed below). Combining these two implies that the marginal distribution of $\hat{\boldsymbol{b}}_j$, integrating out $\boldsymbol{b}_j$, is

$$p(\hat{\boldsymbol{b}}_j | \boldsymbol{\pi}, \boldsymbol{U}, V_j) = \sum_{k=1}^{K} \sum_{l=1}^{L} \pi_{k,l} \; N_R(\hat{\boldsymbol{b}}_j; \boldsymbol{0}, \omega_l U_k + V_j). \tag{4}$$

This last equation comes from the fact that the sum of two MVNs is MVN.

Here the covariance matrix V_j is given by $V_j = S_j C S_j$ where C is a correlation matrix that accounts for correlations among the measurements in the R conditions, and S_j is the $R \times R$ diagonal matrix with diagonal elements $(\hat{s}_{j1}, \ldots, \hat{s}_{jR})$. In settings where measurements in the R conditions are independent one would set $C = \mathbf{I}_R$, the $R \times R$ identity matrix, so $V_j = S_j^2$. However, in our GTEx analysis the measurements are correlated due to sample overlap (some individuals in common) among tissues; we estimate this correlation from the data (see Section "Estimating the correlation matrix C"). The methods implemented here can be applied for any specified matrices V_j.

The two steps of `mash` are:

i) Estimate $\boldsymbol{U}, \boldsymbol{\pi}$. This involves two substeps:

   a) Create a list of both data-driven and canonical covariance matrices, $\hat{\boldsymbol{U}}$.

b) Given $\hat{\boldsymbol{U}}$, estimate $\boldsymbol{\pi}$ by maximum likelihood. (A key idea here is that if some matrices generated in a) do not help capture patterns in the data then they will receive little weight.) Let $\hat{\boldsymbol{\pi}}$ denote this estimate.

ii) Compute, for each j, the posterior distribution $p(\boldsymbol{b}_j|\hat{\boldsymbol{b}}_j, \hat{\boldsymbol{U}}, \hat{\boldsymbol{\pi}}, V_j)$.

These steps are now detailed in turn.

Generate data-driven covariance matrices U_k

We first identify rows j of the matrix $\hat{B}$ that likely have an effect in at least one condition. For example, in the GTEx data we chose the rows corresponding to the "top" SNP for each gene, which we define to be the SNP with the highest value of $Z_j^{\max}$ where

$$Z_j^{\max} := \max_r \hat{b}_{jr}/\hat{s}_{jr}. \tag{5}$$

(We used max here, rather than, say, the sum, to try to include effects that are very strong in a single condition and not only effects that are shared among conditions.) For the simulated data we ran the univariate adaptive shrinkage method `ash` on the data in each condition r separately, and computed $lfsr_{jr}$ for each effect j. We then chose the rows j for which at least one of the conditions showed a significant effect in this univariate analyses ($\min_r lfsr_{jr} < 0.05$).

Next we fit a mixture of MVN distributions to these strongest effects, using methods from [16]. Specifically results in [16] provide an EM algorithm for fitting a model very similar to (3) − (2) with the crucial difference that there is no scaling parameters on the covariances. That is,

$$p(\boldsymbol{b}_j|\boldsymbol{\pi}, \boldsymbol{U}) = \sum_k \pi_k N_R(\boldsymbol{b}_j; \boldsymbol{0}, U_k). \tag{6}$$

The absence of the scaling factors ω_l means that, compared with `mash`, the model (6) is less well suited to capture effects that have similar patterns (relative sizes across conditions) but vary in magnitude. However, by applying it here to only the largest effects we seek to sidestep this issue. Estimates of U_k from this EM algorithm are sensitive to initialization. Furthermore, we noticed an interesting feature of the EM algorithm: each iteration preserves the rank of the matrices U_k, so the ranks of the estimated matrices are the same as the ranks of the matrices used to initialize the algorithm. We exploited this fact by including low-rank matrices in our initialization to ensure that some of the estimated U_k are low-rank matrices. This helps stabilize the estimates since rank-penalization is one way to regularize covariance matrix estimation.

To describe the initialization in detail, let $\tilde{J}$ denote the number of "strongest effects" selected above, and let $\tilde{Z}$ denote the column-centered $\tilde{J} \times R$ matrix of Z scores for these "strong effects". To attempt to extract the main patterns in $\tilde{Z}$ we perform dimension reduction on $\tilde{Z}$. Specifically we apply Principal Component Analysis (through Singular Value Decomposition, SVD) and Sparse Factor Analysis (SFA; [17]) to $\tilde{Z}$.

SVD yields a set of eigenvalues and eigenvectors of $\tilde{Z}$. Let λ_p, v_p denote the pth eigenvalue and corresponding (right) eigenvector. (So v_p is an R vector for $p = 1, \ldots, R$.)

SFA yields a representation

$$\tilde{Z} = LF + E \tag{7}$$

where L is a sparse $J \times Q$ matrix of loadings, and F is a $Q \times R$ matrix of factors. Here we used $Q = 5$.

Given this we initialized the EM with $K = 3$ and

- $\tilde{U}_1 = \frac{1}{\tilde{J}} \tilde{Z}' \tilde{Z}$, the empirical covariance matrix of $\tilde{Z}$.
- $\tilde{U}_2 = \frac{1}{\tilde{J}} \sum_{p=1}^{P} \lambda_p^2 v_p v_p'$, which is a rank P approximation of the covariance matrix of $\tilde{Z}$. Here we used $P = 3$.
- $\tilde{U}_3 = \frac{1}{\tilde{J}} (LF)'(LF)$ which is a rank Q approximation of the covariance matrix of $\tilde{Z}$.

In addition to the covariance matrices obtained from this EM algorithm, we added some more matrices based on the SFA results, specifically

- The 5 matrices $F_q' L_q' L_q F_q'$, which are each rank 1 matrices that reflect the effects captured by the qth factor in the SFA analysis ($q = 1, \ldots, 5$).

The rationale here is that the factors in the factor analysis may directly reflect effect patterns in the data, and if so then these matrices will be a helpful addition. (We view such additions as a low-risk, because If they are not helpful then they will receive little weight when we estimate $\boldsymbol{\pi}$).

In total this procedure produces 8 data-driven covariance matrices for our GTEx analyses.

Generate canonical covariance matrices U_k

To these "data-driven" covariance matrices we add the following "canonical" matrices:

1. The matrix $\mathbf{I}_R$. This represents the situation where the effects in different conditions are independent, which may be unlikely in some applications (like the GTEx application here), but seems useful to include if only to exclude it.

2. The R rank-1 matrices $\boldsymbol{e}_r\boldsymbol{e}_r'$ where $\boldsymbol{e}_r$ denotes the unit vector with 0s everywhere except for element r which is a 1. These represents effects that occur only in a single condition.

3. The rank-1 matrix $\mathbf{1}\mathbf{1}'$ where $\mathbf{1}$ denotes the R-vector of 1s. That is, the matrix of all 1s. This represents effects that are identical among all conditions.

The user can, if desired, add additional canonical matrices. For example, if R is moderate then one could consider adding the 2^R canonical matrices that correspond to shared (equal) effects in each of the 2^R subsets of conditions.

In total this procedure produces 46 canonical covariance matrices for our GTEx analyses.

Standardize covariance matrices

Since (3) uses the same grid of scaling factors ω we standardize the matrices U_k obtained above so that they are similar in scale. Specifically, for each k, we divide every element of U_k by the maximum diagonal element of U_k (so that the maximum diagonal element of the rescaled matrix is one). These rescaled matrices provide the $\hat{\boldsymbol{U}}$, completing step i)-a of `mash`.

Define grid of ω_l values

We choose a dense grid of ω_l ranging from "very small" to "very large". [15] provides a specific way to select suitable limits $(\omega_{\min}, \omega_{\max})$ for this grid in the univariate case; we simply apply this method to each condition r in turn and take the smallest $\omega_{\min}$ and the largest of the $\omega_{\max}$ as the grid limits. The internal points of the grid are then obtained as in the univariate case [15], by setting $\omega_l = \omega_{\max}/m^{l-1}$, for $l = 1, \ldots, L$, where $m > 1$ is a user-tunable parameter that affects the grid density and L is chosen to be just large enough so that $\omega_L < \omega_{\min}$. Our default choice of grid density is $m = \sqrt{2}$. In principle the grid should be made sufficiently dense that increasing its density would not change the answers obtained. In the GTEx data we found results with $m = \sqrt{2}$ provided similar results to $m = 2$, supporting this choice.

Estimate $\boldsymbol{\pi}$ by maximum likelihood

Given $\hat{\boldsymbol{U}}, \omega$, we estimate the mixture proportions $\boldsymbol{\pi}$ by maximum likelihood.

To simplify notation, let $\Sigma_{k,l} := \omega_l \hat{U}_k$, and replace the double index k, l with a single index p which ranges from 1 to $P := KL$. In this notation the prior (3) becomes

$$p(\boldsymbol{b}_j | \boldsymbol{\pi}, \boldsymbol{\Sigma}) = \sum_p \pi_p N_R(\boldsymbol{b}_j; \mathbf{0}, \Sigma_p), \tag{8}$$

and (4) becomes

$$p(\hat{\boldsymbol{b}}_j | \boldsymbol{\pi}, V, \boldsymbol{\Sigma}) = \sum_p^P \pi_p N_R(\hat{\boldsymbol{b}}_j; \mathbf{0}, \Sigma_p + V_j). \tag{9}$$

Assuming independence of rows of $\hat{B}$, the likelihood for $\boldsymbol{\pi}$ is given by

$$\begin{aligned} L(\boldsymbol{\pi}) &:= p(\hat{B} | \boldsymbol{\pi}, V, \boldsymbol{\Sigma}) \\ &= \prod_{j=1}^J p(\hat{\boldsymbol{b}}_j | \boldsymbol{\pi}, V, \boldsymbol{\Sigma}) \\ &= \prod_{j=1}^J \sum_p^P \pi_p N_R(\hat{\boldsymbol{b}}_j; \mathbf{0}, \Sigma_p + V_j). \end{aligned} \tag{10}$$

If the rows of $\hat{B}$ are not independent then this may be interpreted as a "composite likelihood" [39]. By conditioning on V here, rather than treating it as part of the data, we are using a multivariate analogue of the approximation in [40].

Maximising this likelihood over $\boldsymbol{\pi}$ is a convex optimization problem, which here we solve using an EM algorithm [41], accelerated using SQUAREM [42]. This optimization problem is identical to the optimization over $\boldsymbol{\pi}$ in the univariate setting ($R = 1$) in [15], but involves a much larger number of components. If the matrix $\hat{B}$ has many rows then to reduce computation time we can fit the model using a random subset of rows. For example, we used $20,000$ rows in our GTEx application. (It is important that this is a random subset, and not the $\tilde{J}$ rows of strong effects used to generate the data-driven $\hat{U}_k$; use of the strong effects in this step would be a mistake as it would bias estimates of $\boldsymbol{\pi}$ towards large effect sizes.)

Posterior Calculations

To specify the posterior distributions, recall the following standard result for Bayesian analysis of an R-dimensional MVN. If $\boldsymbol{b} \sim N_R(0, U)$, and

$\hat{\boldsymbol{b}}|\boldsymbol{b} \sim N_R(\boldsymbol{b}, V)$ then

$$\boldsymbol{b}|\hat{\boldsymbol{b}} \sim N_R(\tilde{\boldsymbol{\mu}}, \tilde{U}), \tag{11}$$

where:

$$\tilde{U} = \tilde{U}(U, V) := (U^{-1} + V^{-1})^{-1}, \tag{12}$$

$$\tilde{\boldsymbol{\mu}} = \tilde{\boldsymbol{\mu}}(U, V, \hat{\boldsymbol{b}}) := \tilde{U}(U, V)V^{-1}\hat{\boldsymbol{b}}. \tag{13}$$

This result is easily extended to the case where the prior on $\boldsymbol{b}$ is a mixture of MVNs (3). In this case the posterior distribution is simply a mixture of MVNs:

$$p(\boldsymbol{b}_j|\hat{\boldsymbol{b}}_j, \hat{V}_j, \hat{\boldsymbol{\pi}}) = \sum_p^P \tilde{\pi}_{jp} N_R(\boldsymbol{b}_j; \tilde{\boldsymbol{\mu}}_{jp}, \tilde{U}_{jp}) \tag{14}$$

where $\tilde{\boldsymbol{\mu}}_{jp} = \tilde{\boldsymbol{\mu}}(\Sigma_p, V_j, \hat{\boldsymbol{b}}_j)$ [equation (13)], $\tilde{U}_{jp} = \tilde{U}(\Sigma_p, V_j)$ [equation (12)], and

$$\tilde{\pi}_{jp} = \frac{\hat{\pi}_p N_R(\hat{\boldsymbol{b}}_j; \mathbf{0}, \Sigma_p + V_j)}{\sum_{p=1}^P \hat{\pi}_p N_R(\hat{\boldsymbol{b}}_j; \mathbf{0}, \Sigma_p + V_j)}. \tag{15}$$

From this is is straightforward to compute the posterior mean

$$\hat{\hat{\boldsymbol{b}}}_j := E(\boldsymbol{b}_j|\hat{\boldsymbol{b}}_j, \hat{V}_j, \hat{\boldsymbol{\pi}}) = \sum_p^P \tilde{\pi}_{jp} \tilde{\boldsymbol{\mu}}_{jp}, \tag{16}$$

and posterior variance

$$\mathrm{Var}(b_{jr}|\hat{\boldsymbol{b}}_j, \hat{V}_j, \hat{\boldsymbol{\pi}}) = \sum_{p=1}^P \tilde{\boldsymbol{\pi}}_p \, (\tilde{U}_{jp,rr} \; + \; \tilde{\boldsymbol{\mu}}_{jp,r}^2) - [\sum_p^P \tilde{\pi}_{jp} \tilde{\boldsymbol{\mu}}_{jp,r}]^2, \tag{17}$$

as well as the local false sign rate.

Local False Sign Rate

To measure "significance" of an effect b_{jr} we use the *local false sign rate* [15]:

$$\mathit{lfsr}_{jr} := \min[\Pr(b_{jr} \geq 0|D), \Pr(b_{jr} \leq 0|D)] \tag{18}$$

where D denotes all the available data. More intutively, lfsr_{jr} is the probability that we would get the sign of the effect b_{jr} incorrect if we were to use our best guess of the sign (positive or negative). Thus a small *lfsr* indicates high confidence in the sign of an effect. The *lfsr* is more conservative than its analogue, the local false discovery rate (*lfdr*) [18], because requiring confidence in the sign of an effect is more stringent than requiring confidence that it be non-zero. More importantly the *lfsr* is more robust to modelling assumptions than the *lfdr* [15], a particularly important issue in multivariate analyses where modelling assumptions inevitably play a larger role.

Bayes Factors testing Global Null

Although not our primary focus, it is straightforward to use the fitted model to compute Bayes Factors for the alternative model ($\boldsymbol{b} \neq 0$) vs the null model $\boldsymbol{b} = 0$. Specifically

$$\mathrm{BF}_j = p(\hat{\boldsymbol{b}}_j | \hat{\boldsymbol{U}}, \hat{\boldsymbol{\pi}}, V_j) / p(\hat{\boldsymbol{b}}_j | \boldsymbol{b} = 0, V_j) \tag{19}$$

where the numerator is given by (4) and the denominator by (2) with $\boldsymbol{b} = 0$.

The EZ model, and applying `mash` to Z scores

The model (3) assumes $\boldsymbol{b}_j$ are independent of their standard errors V_j. We refer to this as the "exchangeable effects" (EE) model [26]. An alternative assumption is to allow that the effects may scale with standard error, so that effects with larger standard error tend to be larger. That is:

$$S_j^{-1} \boldsymbol{b}_j | \boldsymbol{\pi}, \boldsymbol{U}, \boldsymbol{\omega}, V_j \sim g(), \tag{20}$$

where $g()$ represents the mixture of multivariate normal distributions in (3). We refer to (20) as the "Exchangeable Z" (EZ) model, because the left of this equation is the vector of Z scores for effect j.

As described in [15], this EZ model can be fit by applying exactly the same code as the EE model to the Z statistics, with the standard errors of the Z statistics set to be 1. (That is, set $\hat{B} = Z$ and $\hat{s}_{jr} = 1$.) One advantage of this model is that it can be fit using only the Z scores, and does not require access to both the estimates and their standard errors. The *lfsr* can also be computed using only the Z scores. However, the posterior mean estimates that arise from this model are estimates of $S_j^{-1} \boldsymbol{b}_j$; transforming these to estimates of effect sizes $\boldsymbol{b}_j$ requires knowledge of S_j.

We analyzed the GTEx data using both EE and EZ models. Results were qualitatively similar in terms of patterns of sharing, but the EZ model performed better in cross-validation tests of model fit (see below), and so we report results from that model.

Estimating the correlation matrix C

To estimate the correlation matrix C we exploit the fact that C is the correlation matrix of the Z scores $\boldsymbol{z}_j$ under the null ($\boldsymbol{b}_j = 0$). Specifically we estimate C using the empirical correlation matrix of the z scores for the effects j that are most consistent with the null, $\mathcal{N} = \{j : \max_r(|z_{jr}|) < 2\}$:

$$\hat{C} = \frac{1}{|\mathcal{N}|} \sum_{j \in \mathcal{N}} \boldsymbol{z}_j \boldsymbol{z}_j'. \tag{21}$$

For the GTEx data the measurements in different tissues are not very highly correlated: all elements of the estimated C were < 0.2 and 95% were < 0.1. However, in cross-validation tests (below) this estimated C produced better model fit than ignoring correlations ($C = I_R$).

Cross-validation of model fit

To compare the performance of different strategies for selecting the covariance matrices U_k we use a cross-validation-based approach to assess model fit. In brief, this involves first dividing the data matrix into two groups by selecting half the rows to form the "training data", with the remaining rows forming the "test data". We then apply `mash`, as above, to the training data: use the strongest effects to select candidate U_k, and then learn the weights $\pi_{k,l}$ from all the training data (or a random subset if the data are large; we used 20,000 effects in our analysis). This provides an estimate of the distribution of effects $\hat{g}$. We assess the "fit" of this estimated g by how well it predicts the test data. That is, by computing $p(\hat{B}|\hat{\boldsymbol{\pi}}, V, \hat{U})$, given by (2), for the test data.

This strategy facilitates experimentation with ways to estimate $\hat{U}$. In particular, if new ways to generate $\hat{U}$ are suggested then their effectiveness can be assessed using this strategy. Our current strategy described above was developed and refined using this framework. (However, performance of `mash` is relatively robust to the addition of poorly-estimated U_k because they are typically estimated to have small weight.)

When applying this strategy to the GTEx data we created the test and training data by randomly selecting half the *genes*, rather than half the rows (gene-SNP pairs). Specifically we used genes on even-numbered chromosomes as the training set, and genes on odd-numbered chromosomes as the test set. This ensures that rows in the test set are independent of rows in the training set.

Visualizing U_k

In our application to the GTEx data $R = 44$, so each U_k is a 44 by 44 covariance matrix, and each component of the mixture (1) is a distribution in 44 dimensions. Visualizing such a distribution is challenging, but we can get some insight from the first eigenvector of U_k, v_k say, which captures the principal direction of the effects in component k. If U_k is dominated by this principal direction then we can think of effects from that component as being of the form λv_k for some scalar λ. For example, if the elements of the vector v_k are approximately equal then component k captures effects that are approximately equal in all conditions. Or, if v_k has one large element, with other elements close

to 0, then component k corresponds to an effect that is strong in only one condition. See Figure 2 for illustration.

Relationship with existing methods

The `mash` method essentially includes many existing methods for joint analysis of multiple effects as special cases. Specifically, many existing methods correspond to making particular choices for the "canonical" covariance matrices $\boldsymbol{U}$ (and excluding the data-driven covariance matrices). For example, a simple "fixed effects" meta-analysis – which assumes equal effects in all conditions – corresponds to $K = 1$ with $U_1 = 11'$ (the matrix with all entries 1). (This covariance matrix is singular, but this is allowed within `mash`). A more flexible assumption is that effects in different conditions are normally distributed about some mean, and this also corresponds to a multivariate normal assumption if the mean is assumed to be normally distributed [26]. More flexible still are models that allow that effects may be exactly zero in some subset of conditions, as in [5, 6]. These models correspond to using (singular) covariances U_k with 0s in the rows and columns corresponding to the subset of conditions with zero effect.

However, `mash` also goes beyond these previous methods in two ways. First, `mash` includes a large number of scaling coefficients ω_l, which allows it to flexibly capture a range of effect distributions (see [15]). Second, and perhaps more important, `mash` includes data-driven covariance matrices (Step i-a)), making it more flexible and adaptive to patterns in the specific data being analyzed. This innovation is particularly helpful in settings with moderately large R (e.g., in our application here $R = 44$) where it becomes impractical to pre-specify canonical matrices for all patterns of sharing that might occur. For example, [5, 6] consider all 2^R different combinations of sparsity in the effects, which works for $R = 9$ [20], but is impractical for $R = 44$. While it is possible to restrict the number of combinations considered (e.g. BMAlite in [5]), this comes at an obvious cost in flexibility. The addition of data-driven covariance matrices helps rectify this problem, making `mash` both flexible and computationally tractable for moderately large R.

Definitions of various quantities

RRMSE (accuracy of estimates in simulation studies)

The RRMSEs for estimates $\hat{\hat{b}}_{jr}$ of b_{jr} reported in Figure 2a are computed as

$$\text{RRMSE} = \frac{\sqrt{E((b_{jr} - \hat{\hat{b}}_{jr}))^2)}}{\sqrt{E((b_{jr} - \hat{b}_{jr})^2)}}. \tag{22}$$

ROC curves

For the ROC curves in Figure 2b the True Positive Rate and False Positive Rate are computed at any given threshold t as

$$\text{True Positive Rate} := \frac{|CS \cap S|}{|T|} \tag{23}$$

$$\text{False Positive Rate} := \frac{|N \cap S|}{|N|} \tag{24}$$

where S is the set of significant results at threshold t, CS the set of correctly-signed results, T the set of true (non-zero) effects and N the set of null effects:

$$S := \{j, r : \mathit{lfsr}_{jr} \leq t\}, \tag{25}$$

$$CS := \{j, r : E(b_{jr}|D) \times b_{jr} > 0\}, \tag{26}$$

$$N := \{j, r : b_{jr} = 0\} \tag{27}$$

$$T := \{j, r : b_{jr} \neq 0\}. \tag{28}$$

(Thus, to be considered a true positive, we require that the effect be correctly signed and not only significant.)

For the ROC curves in Figure 2b the True Positive Rate and False Positive Rate are computed based on treating whole rows j as discoveries. For example, suppose a method produces a p value p_j for testing row j. Then at any threshold t the TPR and FPR are:

$$\text{True Positive Rate} := \frac{|\tilde{S}|}{|\tilde{T}|} \tag{29}$$

$$\text{False Positive Rate} := \frac{|\tilde{S}|}{|\tilde{N}|} \tag{30}$$

where $\tilde{S}$ is the set of significant results at threshold t, $\tilde{T}$ the set of true (non-zero) effects and $\tilde{N}$ the set of null effects:

$$\tilde{S} := \{j : p_j \leq t\}, \tag{31}$$

$$\tilde{N} := \{j : b_{jr} = 0 \quad \forall r\} \tag{32}$$

$$\tilde{T} := \{j : \exists r \text{ s.t. } b_{jr} \neq 0\}. \tag{33}$$

Effective sample size

We define the effect sample size for tissue r as

$$n_r^{\text{eff}} := n_r^{\text{orig}} \text{median}_j \frac{\hat{s}_{jr}^2}{\tilde{s}_{jr}^2} \tag{34}$$

where $\hat{s}_{jr}$ is the standard error and $\tilde{s}_{jr}$ is the posterior standard deviation for effect j in tissue r.

Normalized effects

We define the normalized effect $\tilde{b}$ in each condition as the ratio of its effect in that condition to the largest effect across all conditions:

$$\tilde{b}_{jr} = \frac{b_{jr}}{b_{jr_0}} \tag{35}$$

where

$$r_0 = \arg\max_r |b_{jr}| \tag{36}$$

For example, in our eQTL context, a normalized effect $\tilde{b}_{jr} = 0.5$ means that the effect of eQTL j in tissue r is half that of its effect in the strongest tissue.

Pairwise Sharing

To assess pairwise sharing in sign between tissues r and s (Supplementary Figure 4) we compute, for QTL that are significant ($lfsr < 0.05$) in at least one of r and s, the fraction that have effect estimates that are of the same sign.

To assess pairwise sharing in magnitude between tissues r and s (Figure 6) we compute, for QTL that are significant ($lfsr < 0.05$) in at least one of r and s, the fraction that have effect estimates that are within a factor of 2 of one another.

That is, let

$$\mathrm{QTL}_r := \{j : \mathit{lfsr}_{jr} < 0.05\} \tag{37}$$

$$\mathrm{SS}_{rs} := \{j : sign(\hat{\hat{b}}_{jr}) = sign(\hat{\hat{b}}_{js})\} \tag{38}$$

$$\mathrm{SM}_{rs} := \{j : 0.5 \leq \hat{\hat{b}}_{jr}/\hat{\hat{b}}_{js} \leq 2\}. \tag{39}$$

Then the sharing by sign between r and s is given by:

$$\frac{|SS_{rs} \cap (\mathrm{QTL}_r \cup \mathrm{QTL}_s)|}{|\mathrm{QTL}_r \cup \mathrm{QTL}_s|} \tag{40}$$

and sharing by magnitude between r and s is given by:

$$\frac{|SM_{rs} \cap (\mathrm{QTL}_r \cup \mathrm{QTL}_s)|}{|\mathrm{QTL}_r \cup \mathrm{QTL}_s|}. \tag{41}$$

These estimates of sharing are based on the point estimates $\hat{\hat{b}}$, which simplifies their calculation. However, we obtained similar estimates of sharing when we took account of uncertainty in $\hat{\hat{b}}$ by sampling from the posterior distributions of the effects.

`ash` analyses

For comparison with `mash` we also analyzed the GTEx data using the univariate shrinkage procedure `ash` [15]. We applied `ash` separately on each tissue using the same 20,000 randomly-selected gene-snp pairs as in the `mash` analysis. We then computed the posterior means and *lfsr* for the top SNPs.

`mash-bmalite` analyses

For comparison with `mash` we implemented a version on `mash-bmalite` ([5]) that outputs effect size estimates and *lfsr* values. This version of `mash-bmalite` can be thought of as a variation of `mash` but without the data driven covariance matrices, and with particular choices for the canonical covariance matrices, and with a smaller grid on ω than `mash` (consistent with the coarse grid used in [5]).

Specifically, the list of U_k for $\texttt{mash} - \texttt{bmalite}$ include the 44 singleton configurations ($U_k = e_k e_k'$), and matrices corresponding to the models in [5] with heterogeneity parameters $H = \{0.0, 0.25, 0.5, 1\}$ [5]. (When heterogeneity=0, effects are equal in all conditions; when heterogeneity = 1, effects are independent among conditions.) We use a grid of $\omega \in \{0.1, 0.40, 1.6, 6.4, 25.6\}$ consistent with the coarse grid in [5] and designed to capture the range of the GTEx Z-statistics.

Simulation Details

"Shared, Structured Effects"

We simulated $\boldsymbol{b}_j$ from model (3) with equal weights on 8 different covariance matrices learned from the GTEx data, but with the scaling factors ω simulated from a continuous distribution rather than using a fixed grid.

In detail:

1. Take the list of 8 "data-driven" covariance matrices learned from the GTEx data (see Section), standardized to have maximum diagonal element 1 (Section).

2. Simulate 400 "true effects": for each such effect j, a) choose U_j by selecting one of the eight U_k at random, all equally likely; b) simulate ω_j as the absolute value of an $N(0,1)$ random variable; c) simulate $\boldsymbol{b}_j \sim N_{44}(\mathbf{0}, \omega_j U_j)$.

3. For 19,600 "null effects" set $\boldsymbol{b}_j = \mathbf{0}$.

4. For all 20,000 effects, simulate $\hat{\boldsymbol{b}}_j \sim N(\boldsymbol{b}_j, V_j)$ where V_j is the diagonal matrix with diagonal elements 0.1^2. Here, all standard errors are approximately 0.10, consistent with the GTEx dataset.

"Shared, Unstructured Effects"

In these simulations the 400 true effects were all independent and identically distributed: $\boldsymbol{b}_j \sim N_{44}(\mathbf{0}, 0.1^2 \boldsymbol{I}_R)$. Other details are as for Shared, Structured Effects.

"Independent Effects"

We also simulated data where effects were entirely independent across conditions; These were simulated as follows:

1. Independently for each $r = 1, \ldots, 44$, choose a random set of 400 $j \in \{1, \ldots, 20,000\}$ to be the 'true' effects.

2. For the "true effects" simulate $b_{jr} \sim N(0, \boldsymbol{\Sigma}^2)$ where $\boldsymbol{\Sigma}^2$ is chosen with equal probability from the set $\{0.1, 0.5, 0.75, 1\}$ to represent small and large effects within each condition. (All other effects are set to be 0).

3. Simulate $\hat{\boldsymbol{b}}_j \sim N(\boldsymbol{b}_j, V_j)$ as in other simulations.

Analysis of simulated data

Each simulated dataset $(\hat{\boldsymbol{b}}_j, V_j)$ was analyzed using `mash` as detailed above. In particular we re-estimated the $U_k, \boldsymbol{\pi}$ from the data, without making use of the true values for U. We estimated effects by their posterior mean (16) and assessed significance by the *lfsr* (18). Analyses using `ash` and `mash-bmalite` were performed similarly to the applications on the GTEx data (see above).

Acknowledgments

This work was supported by NIH grants MH090951 and HG02585 to MS, and by a grant from the Gordon and Betty Moore Foundation (GBMF 4559). The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health. Additional funds were provided by the NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. Donors were enrolled at Biospecimen Source Sites funded by NCI-Frederick, Inc. (SAIC-F) subcontracts to the National Disease Research Interchange (10XS170), Roswell Park Cancer Institute (10XS171), and Science Care, Inc. (X10S172). The Laboratory, Data Analysis, and Coordinating Center (LDACC) was funded through a contract (HHSN268201000029C) to The Broad Institute, Inc. Biorepository operations were funded through an SAIC-F subcontract to Van Andel Institute (10ST1035). Additional data repository and project management were provided by SAIC-F (HHSN261200800001E). The Brain Bank was supported by a supplements to University of Miami grants DA006227 DA033684 and to contract N01MH000028. Statistical Methods development grants were made to the University of Geneva (MH090941 MH101814), the University of Chicago (MH090951, MH090937, MH101820, MH101825), the University of North Carolina - Chapel Hill (MH090936 MH101819), Harvard University (MH090948), Stanford University (MH101782), Washington University St Louis (MH101810), and the University of Pennsylvania (MH101822). The data used for the analyses described in this manuscript were obtained from the GTEx Portal on 10/17/2015.

We thank Peter Carbonetto, PhD for technical support and comments, and members of the Stephens labs for helpful discussions.

References

1. Blischak, J. D., Tailleux, L., Mitrano, A., Barreiro, L. B. & Gilad, Y. Mycobacterial infection induces a specific human innate immune response. *Scientific Reports* **5** (2015). URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4653619/`.

2. Ferguson, J. P., Cho, J. H. & Zhao, H. A New Approach for the Joint Analysis of Multiple Chip-Seq Libraries with Application to Histone Modification. *Statistical applications in genetics and molecular biology* **11**, Article–1 (2012). URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3770480/`.

3. Pickrell, J. K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768–772 (2010). URL `http://www.nature.com/nature/journal/v464/n7289/full/nature08872.html`.

4. Dimas, A. S. *et al.* Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science (New York, N.Y.)* **325**, 1246–1250 (2009).

5. Flutre, T., Wen, X., Pritchard, J. & Stephens, M. A Statistical Framework for Joint eQTL Analysis in Multiple Tissues. *PLoS Genet* **9**, e1003486 (2013). URL `http://dx.doi.org/10.1371/journal.pgen.1003486`.

6. Li, G., Shabalin, A. A., Rusyn, I., Wright, F. A. & Nobel, A. B. An Empirical Bayes Approach for Multiple Tissue eQTL Analysis. *arXiv:1311.2948 [stat]* (2013). URL `http://arxiv.org/abs/1311.2948`. ArXiv: 1311.2948.

7. Petretto, E. *et al.* New insights into the genetic control of gene expression using a bayesian multi-tissue approach. *PLOS Computational Biology* **6**, 1–13 (2010). URL `http://dx.doi.org/10.1371%2Fjournal.pcbi.1000737`.

8. Wen, X. & Stephens, M. Using Linear Predictors to Impute Allele Frequencies From Summary Of Pooled Genotype Data. *The annals of applied statistics* **4**, 1158–1182 (2010). URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3072818/`.

9. Han, B. & Eskin, E. Random-Effects Model Aimed at Discovering Associations in Meta-Analysis of Genome-wide Association Studies. *The American Journal of Human Genetics* **88**, 586–598 (2011). URL `http://www.cell.com/ajhg/abstract/S0002-9297(11)00155-8`.

10. Stephens, M. A Unified Framework for Association Analysis with Multiple Related Phenotypes. *PLoS ONE* **8**, e65245 (2013). URL `http://dx.doi.org/10.1371/journal.pone.0065245`.

11. Sul, J. H., Han, B., Ye, C., Choi, T. & Eskin, E. Effectively identifying eqtls from multiple tissues by combining mixed model and meta-analytic approaches. *PLOS Genetics* **9**, 1–13 (2013). URL `http://dx.doi.org/10.1371%2Fjournal.pgen.1003491`.

12. Wei, Y., Tenzen, T. & Ji, H. Joint analysis of differential gene expression in multiple studies using correlation motifs. *Biostatistics* **16**, 31–46 (2015). URL `http://biostatistics.oxfordjournals.org/content/16/1/31.abstract`. `http://biostatistics.oxfordjournals.org/content/16/1/31.full.pdf+html`.

13. Pickrell, J., Berisa, T., Segurel, L., Tung, J. Y. & Hinds, D. Detection and interpretation of shared genetic influences on 40 human traits. *bioRxiv* (2015). URL `http://biorxiv.org/content/early/2015/05/27/019885`. `http://biorxiv.org/content/early/2015/05/27/019885.full.pdf`.

14. Han, B. & Eskin, E. Interpreting Meta-Analyses of Genome-Wide Association Studies. *PLOS Genetics* **8**, e1002555 (2012). URL `http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1002555`.

15. Stephens, M. False Discovery Rates: A New Deal. *bioRxiv* 038216 (2016). URL `http://biorxiv.org/content/early/2016/01/29/038216`.

16. Bovy, J., Hogg, D. W. & Roweis, S. T. Extreme deconvolution: Inferring complete distribution functions from noisy, heterogeneous and incomplete observations. *The Annals of Applied Statistics* **5**, 1657–1677 (2011). URL `http://projecteuclid.org/euclid.aoas/1310562737`.

17. Engelhardt, B. E. & Stephens, M. Analysis of Population Structure: A Unifying Framework and Novel Methods Based on Sparse Factor Analysis.

PLOS Genet **6**, e1001117 (2010). URL http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1001117.

18. Efron, B. Local false discovery rates (2005).

19. Lebrec, J. J., Stijnen, T. & van, H. H. C. Dealing with Heterogeneity between Cohorts in Genomewide SNP Association Studies. *Statistical Applications in Genetics and Molecular Biology* **9** (2010). URL https://www.degruyter.com/view/j/sagmb.2010.9.1/sagmb.2010.9.1.1503/sagmb.2010.9.1.1503.xml.

20. Consortium, T. G. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348**, 648–660 (2015). URL http://science.sciencemag.org/content/348/6235/648.

21. Nicolae, D. L. *et al.* Trait-associated snps are more likely to be eqtls: Annotation to enhance discovery from gwas. *PLoS Genet* **6**, 1–10 (2010). URL http://dx.doi.org/10.1371%2Fjournal.pgen.1000888.

22. Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353–1358 (2012). URL http://bioinformatics.oxfordjournals.org/content/28/10/1353.

23. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38**, 904–909 (2006). URL http://www.nature.com.proxy.uchicago.edu/ng/journal/v38/n8/full/ng1847.html.

24. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nature protocols* **7**, 500–507 (2012). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3398141/.

25. Tversky, A. & Kahneman, D. Judgment under uncertainty: Heuristics and biases. *Science* **185**, 1124–1131 (1974). URL http://science.sciencemag.org/content/185/4157/1124. http://science.sciencemag.org/content/185/4157/1124.full.pdf.

26. Wen, X. & Stephens, M. Bayesian methods for genetic association analysis with heterogeneous subgroups: From meta-analyses to gene-environment interactions. *The Annals of Applied Statistics* **8**, 176–203 (2014). URL http://arxiv.org/abs/1111.1210. ArXiv:1111.1210 [stat].

27. Servin, B. & Stephens, M. Imputation-Based Analysis of Association Studies: Candidate Regions and Quantitative Traits. *PLoS Genet* **3**, e114 (2007). URL http://dx.plos.org/10.1371/journal.pgen.0030114.

28. Veyrieras, J.-B. *et al.* High-Resolution Mapping of Expression-QTLs Yields Insight into Human Gene Regulation. *PLoS Genet* **4**, e1000214 (2008). URL http://dx.doi.org/10.1371/journal.pgen.1000214.

29. Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B. & Eskin, E. Identifying Causal Variants at Loci with Multiple Signals of Association. *Genetics* genetics.114.167908 (2014). URL http://www.genetics.org/content/early/2014/08/06/genetics.114.167908.

30. Kichaev, G. *et al.* Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLOS Genetics* **10**, 1–16 (2014). URL http://dx.doi.org/10.1371%2Fjournal.pgen.1004722.

31. Chen, W. *et al.* Fine mapping causal variants with an approximate bayesian method using marginal test statistics. *Genetics* **200**, 719–736 (2015). URL http://www.genetics.org/content/200/3/719. http://www.genetics.org/content/200/3/719.full.pdf.

32. Moyerbrailean, G. A. *et al.* Which genetics variants in dnase-seq footprints are more likely to alter binding? *PLOS Genetics* **12**, 1–27 (2016). URL http://dx.doi.org/10.1371%2Fjournal.pgen.1005875.

33. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLOS Genetics* **10**, 1–15 (2014). URL http://dx.doi.org/10.1371%2Fjournal.pgen.1004383.

34. Fortune, M. D. *et al.* Statistical colocalization of genetic risk variants for related autoimmune diseases in the context of common controls. *Nature Genetics* **47**, 839–846 (2015). URL http://www.nature.com.proxy.uchicago.edu/ng/journal/v47/n7/abs/ng.3330.html.

35. Nica, A. C. *et al.* Candidate causal regulatory effects by integration of expression qtls with complex trait genetic associations. *PLOS Genetics* **6**, 1–11 (2010). URL http://dx.doi.org/10.1371%2Fjournal.pgen.1000895.

36. Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biology* **17**, 13 (2016). URL http://dx.doi.org/10.1186/s13059-016-0881-8.

37. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289–300 (1995). URL `http://www.jstor.org/stable/2346101`.

38. Storey, J. D. The positive false discovery rate: a Bayesian interpretation and the q-value. *The Annals of Statistics* **31**, 2013–2035 (2003). URL `http://projecteuclid.org/euclid.aos/1074290335`.

39. Larribe, F. & Fearnhead, P. Composite likelihood methods in statistical genetics. *Statistica Sinica* **21**, 43–69 (2011).

40. Wakefield, J. Bayes factors for genome-wide association studies: comparison with P-values. *Genetic Epidemiology* **33**, 79–86 (2009).

41. Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **39**, 1–38 (1977).

42. Varadhan, R. & Roland, C. Squared Extrapolation Methods (SQUAREM): A New Class of Simple and Efficient Numerical Schemes for Accelerating the Convergence of the EM Algorithm. *Johns Hopkins University, Dept. of Biostatistics Working Papers* (2004). URL `http://biostats.bepress.com/jhubiostat/paper63`.

43. Bulik-Sullivan, B. *et al.* LD Score Regression Distinguishes Confounding from Polygenicity in Genome-Wide Association Studies. *bioRxiv* (2014). URL `http://biorxiv.org/content/early/2014/02/21/002931`.

44. Zhu, X. & Stephens, M. Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. *bioRxiv* 042457 (2016). URL `http://biorxiv.org/content/early/2016/03/04/042457`.

Supporting Information

Supplementary Text

Effects of Linkage Disequilibrium

Linkage Disequilibrium (LD) between SNPs has two distinct effects.

First, LD causes correlations in the observations of effects for near-by SNPs in the same gene. This issue is likely minor here. Although, when estimating g, `mash` ignores correlations between rows of $\hat{B}$, this can be justified as a "composite likelihood" approach [39], and composite likelihood methods tend to perform well at point estimation.

Second, effect estimates we obtain for each SNP from single-SNP analysis are not actually the individual causal effects of that SNP; rather they are the *combined effects of all SNPs that are in LD with that SNP*, weighted by their LD [43], [44]. This issue is more important, because of the likely presence of multiple eQTLs in some or many genes. It also applies to all single-SNP eQTL analyses, which is the vast majority of all published eQTL analyses, and not just `mash`. Ideally one would develop multi-SNP multi-tissue methods for association analysis at each gene to avoid this issue. And indeed, we see `mash` as a first step towards this more ambitious goal. However, for now we limit ourselves to highlighting one specific feature of our results that we believe may be a consequence of the use of single-SNP effect estimates, and that may change in multi-SNP analyses that better account for LD.

Specifically, LD among multiple causal SNPs can cause single-SNP analyses to identify eQTL that appear to have strong effects of opposite sign in different tissues. One example is shown in Supplementary Figure 3: this eQTL has strong positive Z scores in brain tissues, and negative Z scores in most other tissues, initially suggesting that this eQTL might have causal effects in opposite directions in brain vs non-brain tissues However, the Z scores could also have a different explanation: there could be two eQTLs in LD with one another, one of which (A say) has a strong effect in brain tissues, and the other of which (B say) has a strong effect in other tissues. If the expression-increasing-allele at A is in negative LD with the expression-increasing-allele at B then the single SNP Z scores for either SNP will show opposite signs in brain vs non-brain. Indeed, closer examination of the data at this gene suggests that this explanation is likely correct in this case (Supplementary Figure 3). A similar example is discussed in [20] (their Supplementary Figure S14).

For this reason we believe that estimates of sharing in sign given above are likely to be underestimates of the sharing in sign of actual causal effects, and we

caution against over-interpreting eQTLs that show significant effects of different signs in different tissues.

Increase in effective sample size due to multivariate analysis

A particular emphasis of our work here is improved quantitative estimates of effect sizes in each condition. When estimating effects in a condition, `mash` uses the data not only from that condition but also from other "similar" conditions. In this way `mash` effectively increases the sample size available, and this improves both accuracy and precision of estimates. The improvement will be strongest for conditions that are similar to many other conditions, and weaker for conditions with more "condition-specific" effects.

To illustrate this effect in the GTEx data we compute an "effective sample size" (ESS) for each tissue based on the standard deviations of the `mash` estimates. The ESSs (Supplementary Figure 1) vary from 241 for testis to 1926 for coronary artery. Other tissues with relatively smaller ESS include liver, pancreas, spleen and brain cerebellum. Identifying tissues with smaller ESS could help guide prioritization of (effectively) under-represented tissues in future experimental efforts.

For testis the ESS of 241 represents only a small (1.4-fold) increase compared with actual sample size, reflecting that its effects are more "tissue specific", or, more precisely, that they are less correlated with other tissues. Other tissues showing a similarly small gain in ESS include transformed fibroblasts and whole blood, which are also highlighted as showing more "tissue specific" signals above. In contrast, the ESS for coronary artery represents a 14-fold increase compared with the actual sample size for this tissue, reflecting its stronger correlation with other tissues. On average, across all tissues, `mash` provides a 6-fold increase in ESS for estimating these (strongest) eQTL effects, reflecting the overall moderate to large correlation among effect sizes across tissues.

One caveat here is that ESS reflects *average* gains in precision for a tissue: in practice effects that are shared across many tissues will benefit more than effects that are tissue-specific. For example, if one were particularly interested in effects that are specific to uterus (which has the smallest actual sample size here), then the substantial ESS for uterus may not be as useful as it would first seem. More generally, detecting tissue-specific effects will inevitably benefit most from collecting more samples in that particular tissue.

Supplementary Tables

Method	Simulation Framework	$\text{RRMSE}^{\text{All}}$	$\text{RRMSE}^{\text{Non-null}}$	$\text{RRMSE}^{\text{Null}}$
`mash`	Shared, structured	0.06	0.44	0.015
`mash-bmalite`	Shared, structured	0.11	0.78	0.018
`ash`	Shared, structured	0.21	1.34	0.076
`mash`	Shared, unstructured	0.14	1.00	0.014
`mash-bmalite`	Shared, unstructured	0.15	1.03	0.014
`ash`	Shared, unstructured	0.21	1.37	0.078
`mash`	Independent	0.28	1.82	0.112
`mash-bmalite`	Independent	0.28	1.82	0.118
`ash`	Independent	0.21	1.37	0.076

Supplemental Table 1: Comparison of accuracy of effect size estimates for each method. Results show the RRMSE for all effects ($\text{RRMSE}^{\text{all}}$), and for the subsets of effects that are truly non-null ($\beta \neq 0; \text{RRMSE}^{\text{Non-null}}$) and truly null ($\beta = 0, \text{RRMSE}^{\text{Null}}$). Values of $\text{RRMSE}^{\text{Null}} < 1$ indicate how shrinkage towards zero is helping improve the estimates of null effects. Values of $\text{RRMSE}^{\text{Non-null}} < 1$ indicate how pooling information across conditions can improve accuracy of estimates of non-null effects. (In the Independent simulations the shrinkage of all methods improves overall performance, despite hurting performance for the non-null effects, because most effects are null.)

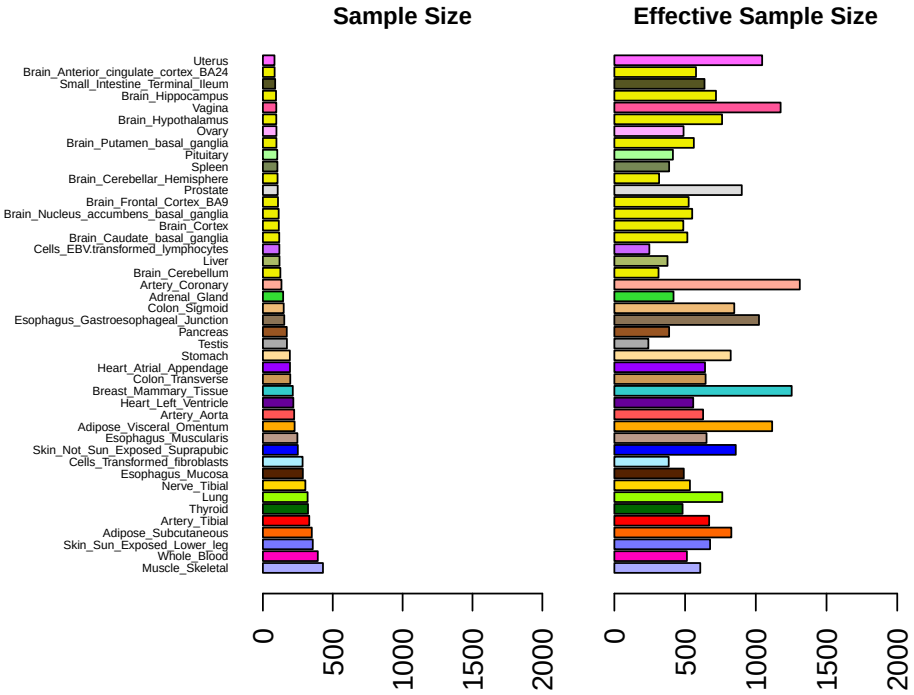
Method	Simulation Framework		
	Shared, structured	Shared, unstructured	Independent
`mash` only	3889	622	32
`ash` only	0	0	740
`mash-bmalite` only	37	9	79
`mash` & `ash`	7	0	44
`mash` & `mash-bmalite`	5777	336	70
`ash` & `mash-bmalite`	0	0	10
ALL	3477	2	5962

Supplemental Table 2: Comparison of Identified Associations in Simulations. Results show the overlap of associations identified by each combination of methods. In both "Shared" scenarios `mash` captures the vast majority of the associations identified by the other methods.

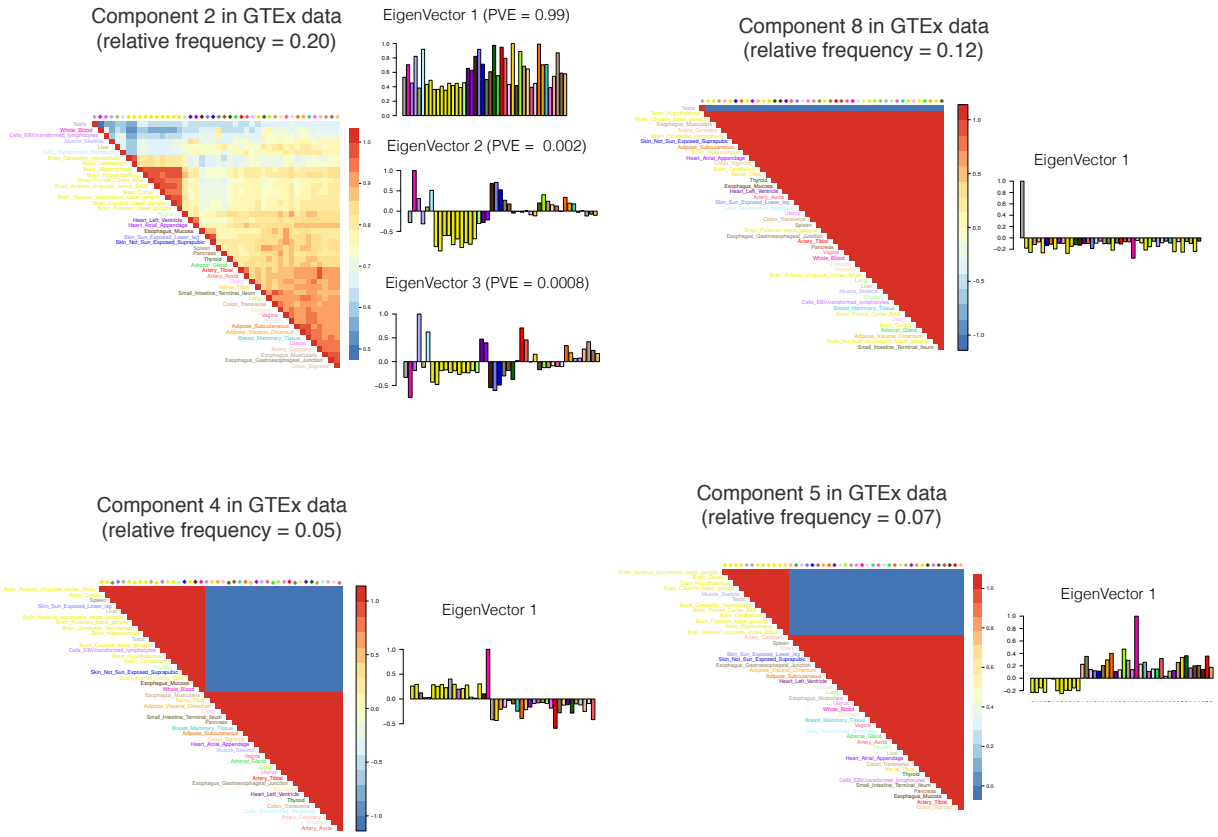
Method	Associations Identified
`mash` only	63956
`ash` only	2383
`mash-bmalite` only	11789
`mash` & `ash`	665
`mash` & `mash-bmalite`	176572
`ash` & `mash-bmalite`	248
ALL	88459

Supplemental Table 3: Comparison of Identified Associations in GTEx Data. Results show the overlap of identified associations with each method. `mash` captures the vast majority of the associations identified by the other methods, in addition to other associations.)

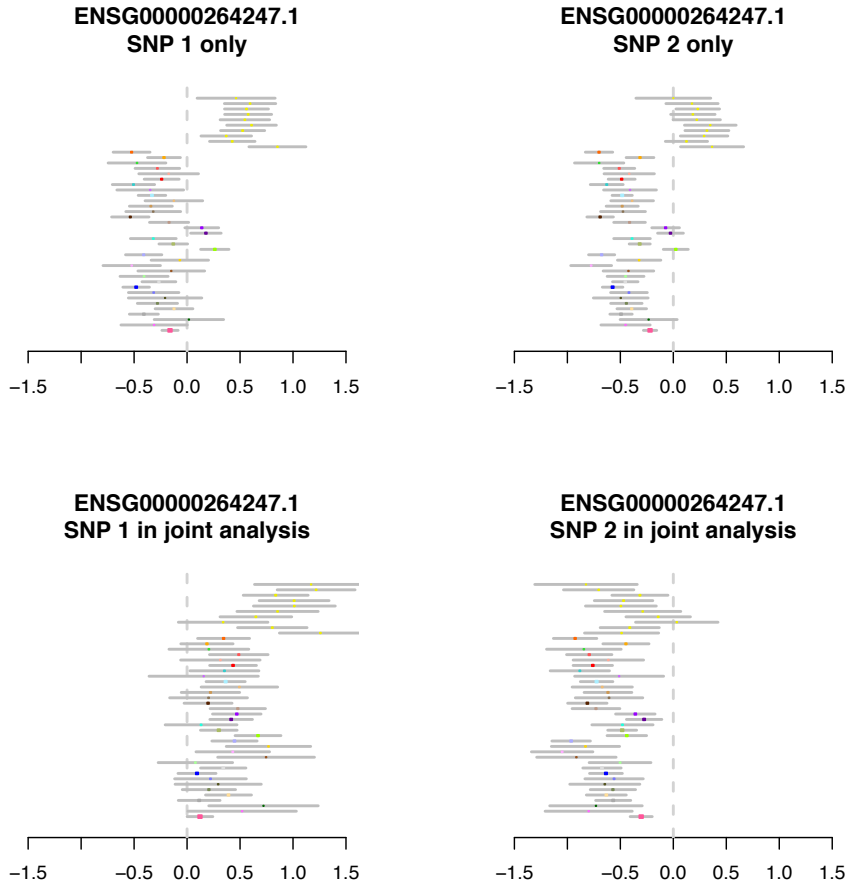
Supplementary Figures



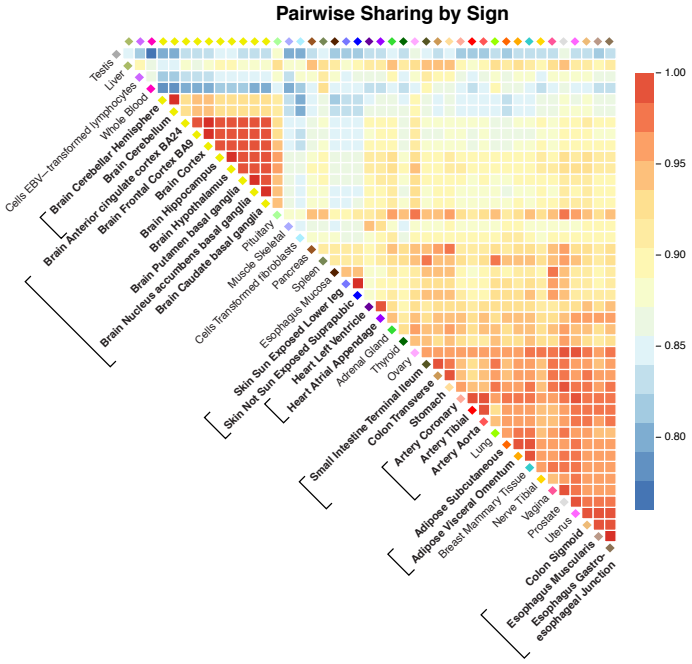
Supplementary Figure 1. Sample sizes and effective sample sizes from `mash` analysis across tissues. Left: sample size for each tissue; Right: median effective sample size for each tissue. Tissues are ordered by their original sample size. Effective sample sizes are consistently higher than actual sample sizes, primarily due to sharing of information among tissues.



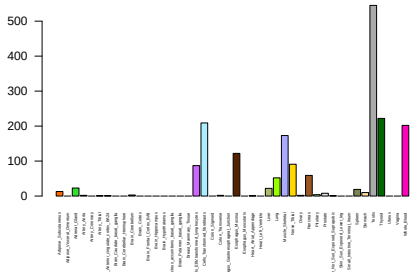
Supplementary Figure 2. Summary of covariance matrices U_k with largest estimated weight ($> 1\%$) in GTEx data. Component 2 largely captures qualitatively similar effects to the component highlighted in Figure 3, although with quantitative differences. Component 8 captures testis-specific effects. Components 4 and 5 primarily capture effects that are stronger in Whole Blood than other tissues.



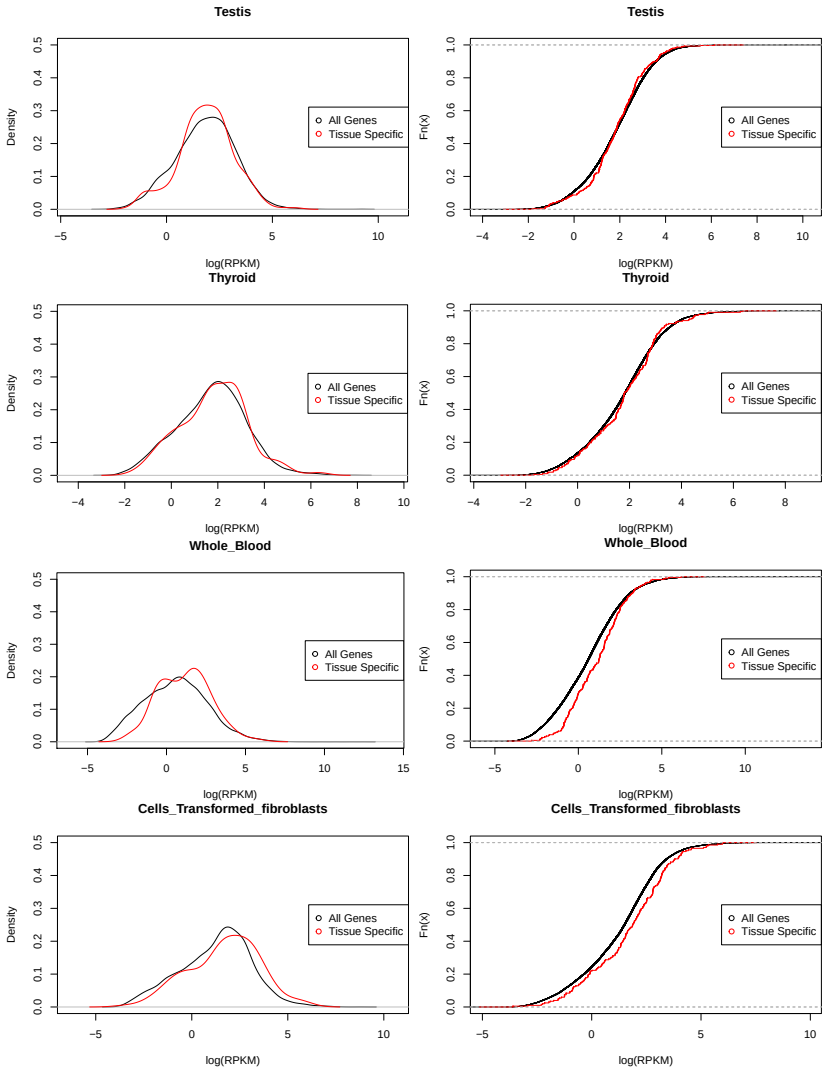
Supplementary Figure 3. Illustration of how Linkage Disequilibrium can impact effect estimates. This gene was chosen as an example where the effect estimates in the "top eQTL" were opposite in sign in brain vs non-brain tissues, and where further investigation suggested that this is likely due to multiple eQTLs in LD. Specifically, SNP1 and SNP2 are the SNPs that show the strongest eQTL association in brain and non-brain tissues respectively. The top panels show effect estimates for these SNPs from a simple (1-SNP) regression model in each tissue, $Y = \mu + \hat{B}_i g_i$ where $i \in \{1, 2\}$ indexes the two SNPs. The bottom panels show effects from a multiple (2-SNP) regression model in each tissue, $Y = \mu + \hat{B}_1 g_1 + \hat{B}_2 g_2$. The simple regression estimates show apparent opposite-sign effects in brain vs non-brain tissues (with testis and pituitary clustering with brain in one case). However, the multiple regression results suggest that in fact there are (at least) two eQTLs in this gene, because both SNPs show a significant effect that excludes 0 in most tissues. Furthermore, for both SNP1 and SNP2 the multiple regression effect estimates are consistent in sign across all tissues.



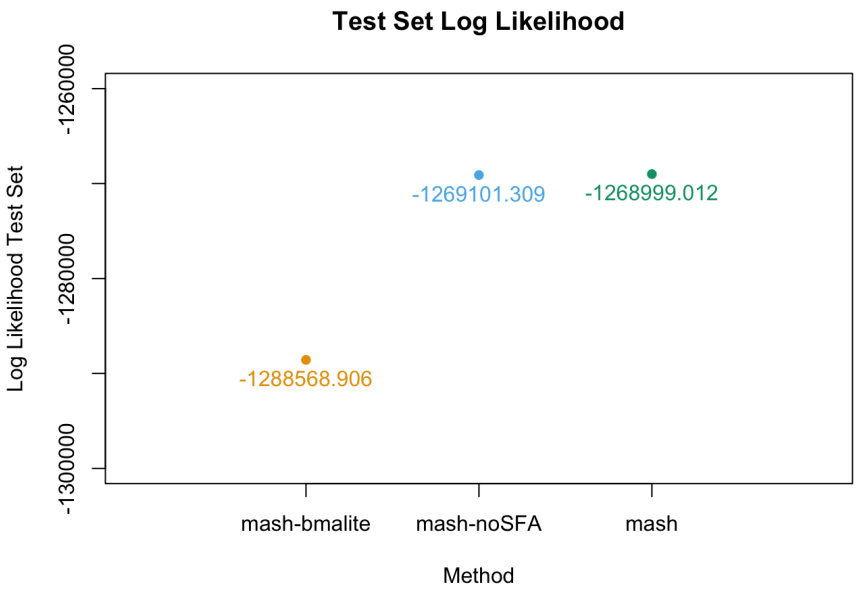
Supplementary Figure 4. Pairwise sharing by sign. For each pair of tissues we consider the top eQTLs that are significant in at least one of the tissues, and estimate the proportion that have effect sizes that are the same sign. These proportions are displayed in this heatmap.



Supplementary Figure 5. Number of "tissue-specific eQTLs" in each tissue. Here "tissue-specific" is defined to mean that the effect is at least 2-fold larger in one tissue than in any other (i.e. $\tilde{b}_{jr} > 0.5$ in only one tissue).



Supplementary Figure 6. Expression levels in genes with "tissue-specific eQTLs" are similar to those in other genes. The plots compare the densities (left) and cumulative distribution functions (right) of the expression level for all genes (black) and for genes identified as having a "tissue-specific" eQTL (red) in each of Testis, Thyroid, Whole Blood and Transformed Fibroblasts. In each case the distribution functions are reasonably similar, demonstrating that tissue-specific eQTLs are not simply reflecting tissue-specific expression. Expression is here defined as median across individuals of the log Reads per Kilobase Mapped (RPKM).



Supplementary Figure 7. Increase in log-likelihood on Test Set as new U_k are added. The figure shows the log-likelihood on the test set for different "models" (choices of U_k). From left to right the models are: `mash-bmalite` (no data-driven U_k); `mash-no-SFA` (the combination of canonical and data-driven covariances, excluding the rank-1 matrices derived from SFA); `mash` (the full combination of canonical and data-driven covariances described here). The result illustrates how, as more data-driven covariances are added, the log-likelihood on the test set typically increases. (Note that the point for `mash` is approximately 100 log-likelihood units higher than `mash-no-SFA`, although it is difficult to see the difference on this scale.)