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

Sarah M. Urbut[1,2], Gao Wang[2], Peter Carbonetto[2,4] & Matthew Stephens[2,3]*

[1]Pritzker School of Medicine, Growth & Development Training Program, University of Chicago, Chicago, IL, USA.

[2]Department of Human Genetics, University of Chicago, Chicago, IL, USA.

[3]Department of Statistics, University of Chicago, Chicago, IL, USA.

[4]Research Computing Center, University of Chicago, Chicago, IL, USA.

*e-mail: mstephens@uchicago.edu

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 allows for more quantitative assessments of effect-size heterogeneity than simple "shared/condition-specific" assessments. We illustrate these features through a detailed analysis of locally-acting variants associated with gene expression ("*cis* eQTLs") in 44 human tissues. Our analysis identifies more eQTLs than existing approaches, consistent with improved power. We show that genetic effects on gene expression are extensively shared among tissues, but that effect sizes can nonetheless vary widely among tissues. Some shared eQTLs show stronger effects in subsets 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 online.

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], differences in histone methylation at many genomic locations in multiple cell lines[2], effects of many genetic variants on risk of multiple diseases[3]; and the impact of expression quantitative trait loci (eQTLs) in multiple cell-types or tissues[4–6]. In these settings, an initial goal is often to identify the "significant" non-zero effects. Another important goal is to compare effects, and to identify differences in effects among conditions—sometimes referred to as "interactions". For example, in eQTL 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, then compare the overlap of "significant" results in different conditions. Although appealingly simple, this "condition-by-condition" approach is unsatisfactory in several ways. It can substantially under-represent sharing of effects among

conditions, because shared effects will be insignificant in some conditions by chance. When effects are shared among conditions, a condition-by-condition approach fails to exploit this, limiting overall power[5].

To address 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 builds on traditional meta-analysis methodology[8,9], and assumes 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 adapt to the data 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, Flutre *et al.*[5] assume correlations are non-negative, and that the non-zero effects are equally correlated among all conditions. Correlations may be negative in some applications; for example, genetic variants that increase one trait may decrease another. And, often, 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, Flutre *et al.*[5] solve the computational problem by restricting effects to be shared in all conditions, or specific to a single condition. Alternatively, Wei *et al.*[12] allow for all possible patterns of sharing in an elegant and 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 is important for accurately assessing heterogeneity of effects among conditions.

Finally, software implementations 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 intended for eQTL applications, whereas `corMotif`[12] is tailored to differential expression analyses. An exception is the `metasoft` software[9,14], which is more generic because it requires only effect estimates and their standard errors in multiple conditions, making it easy to apply 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 by Stephens[15] for testing and estimating 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, including effects that are stronger in some conditions than others; (ii) it is *computationally tractable* for hundreds of thousands of tests in dozens of conditions, or more; (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 data set being analyzed; and (v) it is *generic*, requiring only a matrix containing the observed effects in each condition, and a matrix of the 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 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[15] solves this problem in univariate settings by *learning from the data* how much to shrink. Here we extend this idea 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 such as our eQTL application where there is considerable sharing of effects among conditions, and the sizes of shared effects vary considerably.

To demonstrate the potential for `mash` to provide novel insights, we apply it here to analyze *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. 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 more similar among brain tissues. Our analysis of variation in estimated effects among tissues suggest that assessments of "tissue-specific" versus "tissue-consistent" effects should pay attention to effect sizes, not only tests of significance.

Methods overview

Our method, `mash`, is designed to estimate effects of many units (n) in many conditions (R). It takes as input two $n \times R$ matrices, one containing effect estimates $\hat{B}$, and the other containing the corresponding standard errors. For example, in the GTEx data analyzed here, we consider the effects of hundreds of thousands of potential eQTLs in $R = 44$ tissues.

The method assumes that the true effects are centered on zero, and indeed allows that many effects—possibly the vast majority—may be at or 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. We illustrate `mash` in an eQTL application, though it is sufficiently flexible to apply to other contexts involving many multivariate effects.

The `mash` method is an Empirical Bayes method with two steps: (i) use all the 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 the first step) 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 we introduce a novel and efficient two-step approach for 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 $N_R(\cdot\,; \boldsymbol{\mu}, \Sigma)$ 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 a common pattern of (potentially correlated) effects; each ω_l is a scaling coefficient that corresponds to a different size of effect; and the mixture proportions $\pi_{k,l}$ determine the relative frequency of each covariance-scale combination. The scaling coefficients ω_l take values on a fixed dense grid that spans "very small" to "very large" so as to capture the full range of effects that could occur. (The aim is that the grid is sufficiently large and dense that adding more values to it will not change the results; see Stephens[15].)

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

(i–a) Generate a large list of candidate covariance matrices $\boldsymbol{U} = (U_1,\ldots, U_K)$. 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, 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 1's (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 also Methods.

(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 low weight to covariance matrices that are not well supported by the data. Step i–b is also the place where sparsity is accounted for; if most effects are zero, or small, this step will put most weight on small effects (*i.e.*, small scaling coefficients, ω).

This modular approach has several attractive features. The model (eq. 1) is flexible, and includes many existing methods for analysis of multiple conditions as special cases (see Methods). 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 included.

One potential drawback of a more flexible approach 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 find that overfitting is not a major concern—that is, we found that using more U_k typically improved, or at least did not harm, test-set performance (e.g., Supplementary Fig. 1). 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 eq. 1, rather than setting them to zero, but this would increase risk of overfitting.

Results

Improved effect size estimates. An important novel feature of our method, `mash`, is its focus on *estimation of effect sizes*, in contrast with most existing multivariate analysis methods that focus only on *testing* for non-zero effects. Further, *mash* is more than just an extension of existing methods to estimate effect sizes because the underlying model (eq. 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 Flutre *et al.*[5], which include both the random effects models (RE, RE2), and the 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 `mash-bmalite`, in itself, would represent a useful contribution, `mash` is more flexible—for example, `mash` can learn that some subsets of conditions are more correlated than others due to its use of data-driven covariance matrices (eq. 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` were obtained by applying the method separately to each condition. This represents what can be achieved by a simple "condition-by-condition" analysis. This serves as a baseline to quantify the benefits of the multivariate analysis.

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

1. "Shared, structured effects": Data were simulated using eq. 1, based on the fit of this model to the GTEx eQTL data (see Methods). In this scenario, effects tend to be shared among many conditions. Further, these shared effects are highly "structured"; they are often similar in size or direction, with greater similarity among some subsets of conditions. For example, in the GTEx analysis below, we saw that effect sizes are often similar among brain-derived tissues. Shared, structured effects will arise frequently in practice, and an important goal of our work is to provide methods that perform well in this setting.
2. "Shared, unstructured effects": Effects are shared among all conditions—either every condition shows an effect or no condition shows an effect—but effect sizes and directions are independent across conditions. We aimed 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 for estimating effect sizes.

In each case, we simulated a 20,000 × 44 data matrix $\hat{B}$ containing 20,000 estimated effects (and associated standard errors) in each of the 44 conditions. We assumed that "non-null" (*i.e.*, non-zero) 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–b (see also Supplementary Table 1) compares the accuracy of effect size estimates, as measured by the relative root mean squared error (RRMSE), which is the RMSE of the estimates divided by the RMSE from the original effect estimates, $\hat{B}$. Thus, an RRMSE < 1 indicates an improvement in accuracy over the original estimates, $\hat{B}$. As expected, the joint (multivariate) methods outperformed the univariate method in both scenarios because they combined information across conditions. Furthermore, `mash` substantially outperformed other methods in the "shared, structured effects" scenario (Fig. 2a), and performed like `mash-bmalite` in the unstructured case (Fig. 2b). This last observation is explained by the fact that `mash` put negligible weight on the data-driven covariance matrices in this case. Thus, the flexibility of `mash`, which was responsible for its improved performance in the structured setting, did not decrease performance in the unstructured setting.

In all settings, the three methods showed a substantial improvement in accuracy (RRMSE < 1) compared with the original estimates, $\hat{B}$. This improvement could have come from two sources: (i) shrinkage of estimated effects towards zero, which improves average accuracy because most effects are null; (ii) in the presence of "structured effects", the multivariate methods share information across conditions to improve accuracy. For example, if a particular effect is shared and is similar in size across a subset of conditions, then averaging the observed effects in those conditions will improve accuracy. Both these factors help explain the strong performance of `mash` in the "shared, structured effects" setting (Supplementary Table 1).

For further validation, we also applied the three methods to data simulated under an "Independent effects" scenario in which *all* effects are independent across conditions, with sharing no greater than expected by chance. (This is unlike the "shared, unstructured" scenario, in which only non-zero effects are independent.) We used this to confirm the intuition that the univariate method that analyzes each condition independently should perform best in this setting (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 is more stringent in that it insists that effects be correctly signed to be considered "true discoveries". Similarly, `mash-bmalite` can estimate the *lfsr* (under its less flexible model), and `ash` can estimate the *lfsr* separately for each condition.

In simulations, the flexible multivariate model in `mash` improved power to detect significant effects. Figure 2c–d 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 mirrors the RRMSE results; the multivariate methods performed best, and `mash` outperformed other methods for detecting shared, structured effects. Further, in the "shared, structured effects" scenario, `mash` found essentially all (>99%) of the signals that the other methods found,

plus additional signals (Supplementary Table 2). (In the “shared, unstructured effects” scenario, `mash` and `mash-bmalite` not only had similar average performance, but also found almost identical signals; see Supplementary Table 2.)

Comparison with `metasoft`. Among existing software packages for this problem, `metasoft`[14] is in some respects the most comparable to `mash`. In particular, it is both generic—requiring only effect estimates and their standard errors—and computationally tractable 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 allowing that some effects are exactly zero[14]. These models are similar to the BMAlite models from Flutre *et al.*[5], and none 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 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 "null” here means zero effect in *all* conditions. Here, for each model (FE, RE2, BE), `metasoft` produced a *p* value for each multivariate test, whereas `mash` and `mash-bmalite` produced a Bayes Factor (see Methods); in each case, these can be used to rank the significance of the tests. Figure 2e–f shows the trade-off between false positive and true positive discoveries for each method as the significance threshold is varied. In both cases, `mash` was 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 desirable to identify effects that are shared across many conditions, or, conversely, those that are specific to one or a few conditions. This is a particularly delicate task. For example, Flutre *et al.*[5] emphasize that the simplest approach—first identifying significant signals separately in each condition, then examining the overlap of significant effects—can substantially underestimate 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, Flutre *et al.*[5] and Li *et al.*[6] estimated 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 an effect is non-zero in each condition—so they do not capture differences in magnitude or directions of effects among conditions. If effects shared among conditions differ greatly in magnitude—for example, being very strong in one condition and weak in all others—then this would seem useful to know.

We addressed this limitation by taking a new approach to quantify similarity of effects. Specifically, we assessed sharing of effects in two ways: (i) "sharing by sign" (estimates have the same direction); and (ii) "sharing by magnitude" (effects are similar in magnitude). We defined "similar in magnitude" to mean both the same sign and within a factor of 2 of one another. (Other thresholds could be used, and in some settings—e.g., when "conditions" are different phenotypes—the requirement that effects have the same sign could 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. (We used the condition with the largest estimated effect as reference.)

These measures of sharing could be naively estimated from the original effect estimates in 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 effects" simulations to compare sharing estimates from `mash` to those from the original effect estimates, as well as to `ash` and `mash-bmalite`. These results, summarized in Table 1, confirm the improved accuracy of `mash`; for example, `mash` reduced the error in the estimated number of conditions shared by sign from 4.70 to 2.44.

GTEx *cis*-eQTL analysis. To illustrate the benefits and flexibility of `mash` in a substantive application, we applied it to analyze expression 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 healthy 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 or shared among many tissues. It could also help with analyses that aim to integrate eQTL results with results of genome-wide association studies (GWAS) to help identify the tissues that are most relevant to a given complex disease (e.g., [20,21]).

As input to `mash`, we used a matrix of eQTL effect estimates $\hat{\beta}_{jr}$ and corresponding standard errors $\hat{s}_{jr}$, where rows *j* index gene-SNP pairs, and columns *r* index tissues (or cell types). We used the effect estimates and standard errors for candidate local ("*cis*") eQTLs for each gene, as distributed by the GTEx project (v6 release). These were obtained from (univariate) single-SNP analyses in each tissue; we applied Matrix eQTL[22] to gene expression levels that were corrected for population structure (using genotype principal components[23]) and other confounding factors affecting expression data (both measured factors such as age and sex, and unmeasured factors estimated using factor analysis[23]), then rank-transformed to the corresponding quantiles of a standard normal distribution. Therefore, the effect size estimates were in units of standard deviations on this transformed scale. Because these estimates were obtained by single-SNP analysis, similar to most eQTL analyses, the estimated effects for each SNP reflect the effects of the SNP itself as well as 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 and Supplementary Text.

We analyzed the 16,069 genes for which univariate effect estimates were available for all 44 tissues considered; the filtering criteria we used ensured 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 illustrated the gains of the more flexible `mash` model using cross-validation; we fitted 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 (Supplementary Fig. 1), we found that `mash` with a correlated residual framework improved the test set likelihood by 23,796 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 better capture most effects 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 Fig. 3, which shows the majority component that `mash` identified in these data (relative frequency of 34%). The main patterns captured by this component are: (i) effects are positively correlated among all tissues; (ii) the brain tissues—and, to a lesser 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 less correlated with other tissues. Other components identified by `mash` are shown in Supplementary Fig. 2. Some of these other components also have positive correlations among all tissues and/or highlight heterogeneity between brain tissues and other tissues, confirming that these are common features of these data. Other components capture less prevalent patterns, such as effects that are appreciably stronger in one tissue (e.g., Supplementary Fig. 2b).

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 candidate eQTL. Although we cannot directly demonstrate improved accuracy of effect estimates in the GTEx data (for this, see simulations above), individual examples can provide insight into how `mash` achieves improved accuracy. 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` estimated 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 were outweighed by the strong background information that effects are highly correlated among

tissues. This weighing of evidence was done using Bayes rule. Humans are notoriously bad at weighting background information against 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. The `mash` effect estimates are also appreciably larger in brain tissues than in other tissues. Again, this resulted from using Bayes rule to combine the effect estimates for this eQTL with the background information on heterogeneity of brain and non-brain effects that was learned from *all* eQTLs.

In the second example, most of the effect estimates in non-brain tissues are positive (30 out of 34), but modest in size, and only one effect is, individually, nominally significant $p < 0.05$. However, by combining information among tissues, `mash` estimated that all effects in non-brain tissues are positive, and mostly "significant" (*lfsr* < 0.05). By contrast, the data in brain tissues are inconsistent, with a mix of positive and negative estimated effects. Therefore, `mash` assigned low confidence to the direction of the eQTL effects in brain tissues. This example illustrates that `mash` can learn from the data to treat subsets of conditions differently. In this case, `mash` learned that effects in brain tissues are occasionally different from effects in other tissues, and therefore did not make strong inferences in the brain based on the signal in other tissues.

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

In all three examples, the `mash` estimates are noticeably different from those obtained by `ash` (Supplementary Fig. 3) and `mash-bmalite` (Fig. 4c). The univariate `ash` method simply shrank the non-significant effects towards zero, ignoring the information from other tissues. (Given data from only one tissue, this is a reasonable strategy, since most effects are small; however, given data from many tissues one can do better by combining information across tissues.) The `mash-bmalite` method fit each example into a "canonical" configuration; the first example was assessed as shared across tissues, whereas the other examples were assessed to be specific to a single tissue, as reflected in the effect estimates. The decreased flexibility of `mash-bmalite` compared to `mash` is most evident in the first example; the `mash-bmalite` estimates are almost constant across tissues, hiding the stronger effects in brain-related tissues that are clearly visible in the original data, and maintained by `mash`.

Increased identification of significant effects. Our simulations demonstrated that the more flexible model behind `mash` can increase power to detect significant effects. To demonstrate the impact of this increased power, we compared the number of significant eQTLs detected by `mash` with those detected by `mash-bmalite` and `ash`. To avoid double-counting eQTLs in the same gene that are in LD with one another, we assessed the significance of only the "top" SNP

in each gene, which we defined to be the SNP with the largest (univariate) $|Z|$-statistic across all tissues. Thus, our results are based on 16,069 candidate eQTLs, each with effect estimates in 44 tissues, for a total of 44 × 16,069 = 707,036 effects.

The vast majority of top SNPs show a strong signal in at least one tissue (97% have a maximum $|Z|$ score exceeding 4), consistent with most genes containing at least one eQTL in at least one tissue. However, the univariate tissue-by-tissue analysis (`ash`) identified only 13% of these effects as "significant" at *lfsr* < 0.05; that is, the univariate analysis was highly confident in the direction of the effect in only 13% of cases. By comparison, `mash-bmalite` identified 39% as significant at the same threshold, and `mash` identified 47%. As in our simulations above, the significant associations identified by `mash` include the vast majority (95%) of those found significant by either of the other methods (Supplementary Table 3). In summary, the multivariate methods determined more effects to be significant, and `mash` identified the most.

Overall, `mash` found 76% (12,171 out of 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 magnitude. The results are summarized in Table 2 and Fig. 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 96% among brain tissues. (Furthermore, these numbers may underestimate the sharing by sign of actual causal effects due to the impact of multiple eQTLs in LD; see Supplementary Text and Supplementary Fig. 5.) Sharing by magnitude is necessarily lower because sharing by magnitude implies sharing by sign. On average, 36% 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 76%; not only do eQTLs tend to be shared among the brain tissues, but the effect sizes tend to be homogeneous. Because our analyses were based on the top eQTLs, our results reflect patterns of sharing only among strong *cis* eQTLs; it is possible that weaker eQTLs may show different patterns of heterogeneity among tissues.

Some tissues share eQTLs more than others; Fig. 6 summarizes eQTL sharing by magnitude between all pairs of tissues (see Supplementary Fig. 6 for sharing by sign). In addition to strong sharing among brain tissues, `mash` identified 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 versus non-cerebellum tissues, and shows higher sharing between 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 typical precision of its effect estimates (Supplementary Fig. 7, Supplementary Text). The ESS values are smallest for tissues 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` identified 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 (Fig. 5), representing a subset of eQTLs that have a much stronger effect in one tissue than in any other (henceforth, "tissue-specific" for brevity). Breaking down this group by tissue identified testis as the tissue with the most tissue-specific effects (Supplementary Fig. 4). Testis and whole blood stand out as having lower pairwise sharing of eQTLs with other tissues (Fig. 6). Other tissues showing stronger-than-average tissue specificity (Fig. 5, Supplementary Fig. 4) 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 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 Fig. 8). Therefore, most of the tissue-specific eQTLs we identified do not solely reflect tissue-specific expression.

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. Our aim was to provide a set of flexible and general tools to facilitate such analyses. 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, preferably, a matrix of effect estimates and a matrix of corresponding standard errors) that test many effects in many conditions. For example, the effect estimates used in our GTEx analysis came from simple linear regressions, but it would be equally possible to use `mash` with estimates from other approaches, such as generalized linear models or linear mixed models[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 widely applicable; although we were primarily motivated by genomics applications, `mash` could be useful in any setting involving testing and estimation of multivariate effects.

The `mash` method is based on an Emprical Bayes hierarchical model, so it is related to other methods that use this approach[5,6,12]. Indeed, the `mash` framework essentially includes these methods as special cases (as well as simpler methods such as "fixed effects" and "random effects" meta-analyses[9,26]. One key feature that distinguishes `mash` from previous methods is that `mash` puts greater emphasis 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 on allowing for and assessing quantitative variation among effects. Moving 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 effects are often shared broadly among many conditions, and in such cases binary-based methods tend to infer that effects are non-zero in most or all conditions, even when the signal is modest in some conditions. This conclusion may be technically correct—for example, in our GTEx analysis it is not impossible that all eQTLs are somewhat active in all tissues. However, as our analysis has illustrated, 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 variant. This limitation applies to most multi-tissue (and single-tissue) eQTL analyses. This issue is particularly important to appreciate when, for example, cross-referencing 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 co-localization methods[33–35] that attempt to assess whether the same causal SNP explains two different association signals (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 that there is at most one causal SNP per gene[27,28]. It would be straightforward to adapt `mash` to also perform fine-mapping under this assumption. However, this simplifying assumption becomes less plausible in analyses that involve many tissues, and we view the development of more flexible multi-tissue eQTL methods for fine-mapping as an important and challenging problem for future work.

One potentially powerful extension of `mash` would be to allow for sharing patterns to depend on covariates. For example, in an eQTL mapping 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 parameterize the prior distributions on patterns of sharing or sizes of effects,

and estimate the these parameters from the data[28,36]. One possible approach would be to allow the mixture proportions $\boldsymbol{\pi}$ in `mash` to depend on covariates through a logistic link. However, this appears to be a challenging problem, and a satisfactory solution may require some ingenuity.

Dealing with multiple tests is often described as a "burden". This interpretation likely originates from the fact that controlling family-wise error rate (the probability of making one false discovery) requires 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. We believe that the availability of results of many tests in many conditions should not be viewed as a burden, rather as an *opportunity*—specifically, an opportunity to learn about relationships among underlying effects, and consequently to make data-driven decisions that help improve both power to detect effects and precision of effect estimates. This direction will inevitably, it seems, involve modelling assumptions, and the goal should be to design flexible models that are capable of dealing with a wide range of situations that can occur in practice. The methods presented here represent a major step towards this goal.

URLs. multivariate adaptive shrinkage ("mash") software, https://github.com/stephenslab/mashr; code and data resources for GTEx analysis, https://github.com/stephenslab/gtexresults, https://doi.org/10.5281/zenodo.1296399; GTEx project, http://gtexportal.org; adaptive shrinkage ("ash") software, https://github.com/stephens999/ashr; Sparse Factor Analysis (SFA) software, http://stephenslab.uchicago.edu/software.html; Extreme Deconvolution software, https://github.com/jobovy/extreme-deconvolution; METASOFT software, http://genetics.cs.ucla.edu/meta; Matrix eQTL software, http://www.bios.unc.edu/research/genomic_software/Matrix_eQTL.

Methods

Model and fitting. Let b_{jr} (j = 1,…, J; r = 1,…, R) denote the true value of effect j in condition r. Further, let $\hat{b}_{jr}$ denote the "observed" estimate of this effect, and let $\hat{s}_{jr}$ be the standard error of this estimate, so $z_{jr} := \hat{b}_{jr}/\hat{s}_{jr}$ is the standard Z statistic used to test 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 $\hat{\boldsymbol{b}}_j$ is normally distributed about $\boldsymbol{b}_j$, with variance-covariance matrix V_j (defined below), and that the $\boldsymbol{b}_j$ follows eq. 1:

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

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

Combining these two implies that the marginal distribution of $\hat{\boldsymbol{b}}_j$, integrating out $\boldsymbol{b}_j$, is

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

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

Each covariance matrix V_j is specified as $V_j = S_j C S_j$, where *C* is a correlation matrix that accounts for correlations among the measurements in the *R* conditions, and S_j is the *R* × *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 $C = I_R$, the *R* × *R* identity matrix, so $V_j = S_j^2$. However, in our GTEx analysis the measurements are correlated due to sample overlap (individuals in common) among tissues; we estimate this correlation from the data (see "Estimating the correlation matrix *C*" below). The methods place no restriction on V_j.

The two steps of `mash` are:

i. Estimate $\boldsymbol{U}$ and $\boldsymbol{\pi}$. This involves two sub-steps:
   a. Define a set of (normalized) data-driven and canonical covariance matrices, $\hat{\boldsymbol{U}}$.
   b. Given $\hat{\boldsymbol{U}}$, estimate $\boldsymbol{\pi}$ by maximum likelihood. (A key idea is that if some matrices generated in Step i–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 $p(\boldsymbol{b}_j \mid \hat{\boldsymbol{b}}_j, \hat{\boldsymbol{\pi}}, \hat{\boldsymbol{U}}, V_j)$.

These steps are now detailed in turn.

Generate data-driven covariance matrices U_k. We first identify rows *j* of matrix $\hat{B}$ that likely have an effect in at least one condition. For example, in the GTEx data we chose rows corresponding to the "top" SNP for each gene, which we define to be the SNP with the highest value of $Z_j^{\max} := \max_r \hat{b}_{jr}/\hat{s}_{jr}$. (We use the maximum, rather than the sum because we would like to include effects that are strong in a single condition, not just effects that are shared among conditions.) For the simulated data, we run `ash` separately for each condition *r*, compute $lfsr_{jr}$

for each effect *j*, then choose rows *j* for which at least one condition shows a significant effect; *i.e.*, $\min_r lfsr_{jr} < 0.05$.

Next, we fit a mixture of MVN distributions to these strongest effects using methods from Bovy *et al.*[16]. Results in Bovy *et al.*[16] describe an EM algorithm for fitting a model similar to eqs. 2 & 3, with the crucial difference that there are no scaling parameters on the covariances;

$$p(\boldsymbol{b}_j \mid \boldsymbol{\pi}, \boldsymbol{U}) = \sum_{k=1}^{K} \pi_k N_R(\boldsymbol{b}_j; \boldsymbol{0}, U_k). \tag{5}$$

The absence of the scaling factors ω_l means that, compared with `mash`, their model is less well suited to capturing 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 that ranks of the estimated matrices are the same as the ranks of the matrices used to initialize the algorithm. We exploited this fact to ensure that some of the estimated U_k have low rank. 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 extract the main patterns in $\tilde{Z}$, we perform dimension reduction on $\tilde{Z}$; specifically, we apply principal components analysis (through a singular value decomposition, or 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 *p*th eigenvalue and corresponding (right) eigenvector. SFA yields matrix factorization

$$\tilde{Z} = LF + E,$$

where *L* is a sparse *J* × *Q* matrix of loadings and *F* is a *Q* × *R* matrix of factors. We use *Q* = 5.

Given this matrix factorization, we use the EM algorithm to fit a mixture model (eq. 5) with *K* = 3, in which the three matrices U_k are initialized as follows:

- $U_1 = \tilde{Z}^T \tilde{Z} / \tilde{J}$, the empirical covariance matrix of $\tilde{Z}$.
- $U_2 = \sum_{p=1}^{P} \lambda_p v_p v_p^T / \tilde{J}$, which is a rank-*P* approximation of the covariance matrix of Z, with *P* < *Q*. We use *P* = 3.

- $U_3 = F^T L^T LF/\tilde{J}$, which is a rank-Q approximation of the covariance matrix of Z.

The output of the EM algorithm defines U_1, U_2 and U_3 in the `mash` mixture (eq. 1).

In addition to the covariance matrices obtained from this EM algorithm, we define more covariance matrices from the SFA results; specifically, the five rank-1 matrices $F_q^T L_q^T L_q F_q$, with q = 1,…,5. The rationale for adding these rank-1 covariance matrices is that the SFA factors may directly reflect effect patterns in the data, and if so these matrices will be a helpful addition. (If they are not helpful, they will receive little weight when estimating $\boldsymbol{\pi}$).

In summary, SFA and the EM algorithm produce 8 data-driven covariance matrices for our analysis of the GTEx and simulated data sets.

Generate canonical covariance matrices U_k. To these "data-driven" covariance matrices, we add the following "canonical" covariance matrices:

- The identity matrix, 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 show that the data assign low weight to it.
- R rank-1 matrices $\boldsymbol{e}_r \boldsymbol{e}_r^T$, where $\boldsymbol{e}_r$ denotes the unit vector with zeros everywhere except for element r, which is set to 1. These matrices represent effects that occur in only one condition.
- The rank-1 matrix $\mathbf{1}\mathbf{1}^T$, where $\mathbf{1}$ denotes the R-vector of ones—that is, this matrix is a matrix of all ones. This represents effects that are identical among all conditions.

The user can, if desired, add other canonical matrices. For example, if R is moderate then one could consider adding the 2^{R-1} canonical matrices that correspond to shared (equal) effects in each of the 2^{R-1} subsets.

In summary, this procedure produces 46 canonical covariance matrices for our analyses of the GTEx and simulated data sets.

Standardize the covariance matrices. Since eq. 3 uses the same grid of scaling factors for each covariance matrix U_k, we standardize the matrices U_k so that they are similar in scale. Specifically, for each k, we divide all elements 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 define $\hat{\boldsymbol{U}}$ in Step i–a of `mash`.

Define grid of ω_l values. We define a dense grid of scaling factors, ranging from "very small" to "very large". Stephens[15] suggests a way to select suitable limits $(\omega_{\min}, \omega_{\max})$ for this grid in the univariate case; we 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 obtained, as in the univariate case[15], by setting $\omega_l = \omega_{\max}/m^{l-1}$, for l = 1,…,L, where m > 1 is a user-tunable parameter that controls 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 $m = \sqrt{2}$. The grid should be made sufficiently dense that increasing its density would not change the answers obtained. In the GTEx data, results with $m = \sqrt{2}$ were similar to results with a much finer grid (m = 2), supporting this choice.

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

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

$$p\left(\boldsymbol{b}_j \mid \boldsymbol{\pi}, \boldsymbol{U}\right) = \sum_{p=1}^{P} \pi_p N_R(\boldsymbol{b}_j; \mathbf{0}, \Sigma_p),$$

and the marginal distribution (eq. 4) becomes

$$p\left(\widehat{\boldsymbol{b}}_j \mid \boldsymbol{\pi}, \boldsymbol{U}, V_j\right) = \sum_{p=1}^{P} \pi_p N_R(\widehat{\boldsymbol{b}}_j; \mathbf{0}, \Sigma_p + V_j).$$

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

$$L(\boldsymbol{\pi}) \coloneqq p\left(\hat{B} \mid \boldsymbol{\pi}, \boldsymbol{U}, \boldsymbol{V}\right) = \prod_{j=1}^{J} p\left(\widehat{\boldsymbol{b}}_j \mid \boldsymbol{\pi}, \boldsymbol{U}, V_j\right) = \prod_{j=1}^{J} \sum_{p=1}^{P} \pi_p N_R(\widehat{\boldsymbol{b}}_j; \mathbf{0}, \Sigma_p + V_j)$$

If the rows of $\hat{B}$ are not independent, this may be interpreted as a "composite likelihood"[39], which generally yields consistent point estimates. (We also empirically confirmed that our GTEx results were robust to LD pruning of SNPs; see Supplementary Fig. 9.) By conditioning on $\boldsymbol{V}$, rather than treating it as part of the data, we are using a multivariate analogue of the approximation used by Wakefield[40].

Maximizing this likelihood over $\boldsymbol{\pi}$ is a convex optimization problem. We solve this optimization problem using EM[41], accelerated using SQUAREM[42]. This problem is identical to the optimization over $\boldsymbol{\pi}$ in the univariate setting (R = 1)[15], but involves a much larger number of components. If $\hat{B}$ has a large number of rows, one way to reduce computational effort is to take a random subset of rows. In the GTEx analysis, we used a random subset of 20,000 rows. (It is important that this is a random subset, and not the $\tilde{J}$ rows of strong effects used to generate the

data-driven covariance matrices; 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 $\boldsymbol{b} \sim N_R(\mathbf{0}, U)$ and $\hat{\boldsymbol{b}} \mid \boldsymbol{b} \sim N_R(\boldsymbol{b}, V)$, then

$$\hat{\boldsymbol{b}} \mid \boldsymbol{b} \sim N_R\left(\tilde{\boldsymbol{\mu}}, \tilde{U}\right),$$

where

$$\tilde{U} = \tilde{U}(U, V) \coloneqq (U^{-1} + V^{-1})^{-1} \tag{6}$$

$$\tilde{\boldsymbol{\mu}} = \tilde{\boldsymbol{\mu}}\left(U, V, \hat{\boldsymbol{b}}\right) \coloneqq \tilde{U}(U, V)V^{-1}\hat{\boldsymbol{b}} \tag{7}$$

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

$$p\left(\boldsymbol{b}_j \mid \hat{\boldsymbol{b}}_j, V_j, \boldsymbol{\pi}\right) = \sum_{p=1}^{P} \tilde{\pi}_{jp} N_R\left(\boldsymbol{b}_j;\ \tilde{\boldsymbol{\mu}}_{jp}, \tilde{U}_{jp}\right),$$

where $\tilde{U}_{jp} \coloneqq \tilde{U}\left(\Sigma_p, V_j\right)$ (eq. 6) and $\tilde{\boldsymbol{\mu}}_{jp} \coloneqq \tilde{\boldsymbol{\mu}}\left(\Sigma_p, V_j, \hat{\boldsymbol{b}}_j\right)$ (eq. 7), and

$$\tilde{\pi}_{jp} \coloneqq \frac{\pi_p N_R(\hat{\boldsymbol{b}}_j; 0, \Sigma_p + V_j)}{\sum_{p'=1}^{P} \pi_{p'} N_R(\hat{\boldsymbol{b}}_j; 0, \Sigma_{p'} + V_j)}.$$

From these results, it is straightforward to compute the posterior mean

$$\hat{\hat{\boldsymbol{b}}}_j \coloneqq E\left(\boldsymbol{b}_j \mid \hat{\boldsymbol{b}}_j, V_j, \boldsymbol{\pi}\right) = \sum_{p=1}^{P} \tilde{\pi}_{jp} \tilde{\boldsymbol{\mu}}_{jp} \tag{8}$$

and posterior variance

$$\mathrm{Var}\left(b_{jr} \mid \hat{\boldsymbol{b}}_j, V_j, \boldsymbol{\pi}\right) = \sum_{p=1}^{P} \tilde{\pi}_{jp}\left(\tilde{U}_{jp,rr} + \tilde{\boldsymbol{\mu}}_{jp,r}^2\right) - \left(\sum_{p'=1}^{P} \tilde{\pi}_{jp'} \tilde{\boldsymbol{\mu}}_{jp',r}\right)^2,$$

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],

$$lfsr_{jr} := \min\{\Pr(b_{jr} \geq 0 \mid D), \Pr(b_{jr} \leq 0 \mid D)\},$$

where *D* denotes the available data. More intuitively, $lfsr_{jr}$ is the probability that we would incorrectly predict the sign of the effect 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 high confidence in the sign is more stringent than requiring high confidence that the effect 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 ($\boldsymbol{b} \neq 0$) against the null model ($\boldsymbol{b} = 0$),

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

The numerator is given by eq. 4 and the denominator is given by eq. 2, with $\boldsymbol{b} = 0$.

The EZ model, and applying `mash` to *Z* scores. The model (eq. 3) assumes the $\boldsymbol{b}_j$ 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 the effects to scale with standard error, so that effects with larger standard error tend to be larger,

$$S_j^{-1}\boldsymbol{b}_j \mid \boldsymbol{\pi}, \boldsymbol{U}, V_j \sim g,$$

where *g* is shorthand for the mixture of MVNs (eq. 3). We refer to this as the "Exchangeable *Z*" (EZ) model because the left of this equation is the vector of *Z* scores for effect *j*.

As explained in Stephens[15], the EZ model can be fit by applying exactly the same code as the EE model to the *Z* statistics, in which the standard errors of the *Z* statistics are set to 1; in other words, set $\hat{b}_{jr} = z_{jr}$ and $\hat{s}_{jr} = 1$. One advantage of the EZ model is that it can be fit using only the *Z* scores; it does not require access to both the effect 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), so we used the results from the EZ model.

Estimating the correlation matrix *C*. *C* is an *R* × *R* correlation matrix that accounts for correlations among measurements in the *R* conditions. To estimate this matrix, we exploit the fact that *C* is the correlation matrix of the *Z* scores under the null ($\boldsymbol{b}_j = 0$). Specifically, we estimate *C* from the empirical correlation matrix of the *Z* scores for the effects *j* that are most consistent with the null,

$$C = \frac{1}{|\mathcal{N}|} \sum\nolimits_{j \in \mathcal{N}} \boldsymbol{z}_j \boldsymbol{z}_j^T,$$

where $\mathcal{N} \coloneqq \left\{ j : \max_r |z_{jr}| < 2 \right\}$.

For the GTEx data, we found that the measurements in different tissues were not highly correlated; all elements of the estimated *C* had magnitude less than 0.2, and 95% had magnitude less than 0.1. However, in cross-validation tests (below), this estimated *C* produced better model fit than ignoring correlations (*i.e.*, $C = I_R$).

Cross-validation of model fit. We use a cross-validation approach to assess different strategies for selecting the covariance matrices U_k. In brief, this involves first dividing the rows of the data matrix into two subsets—the first subset is the "training set", and the remaining rows form the "test set". We then apply `mash`, as above, to the training set; as before, we use the strongest effects to select candidate U_k, then the mixture weights $\pi_{k,l}$ are fitted to the training data (or a random subset if the training set is too large; in practice, 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, we compute $p(\hat{B} \mid \hat{\boldsymbol{\pi}}, \boldsymbol{V}, \hat{\boldsymbol{U}})$, given by eq. 2, for the test data.

This strategy facilitates experimentation with different approaches to estimating $\hat{\boldsymbol{U}}$. In particular, if new approaches to estimate $\hat{\boldsymbol{U}}$ are suggested, then their effectiveness can be assessed using this strategy. Our current strategy, described above, was developed and refined using this framework. (Note that performance of `mash` is still relatively robust to the addition of poorly estimated U_k because they are typically assigned small weights after the model fitting procedure.)

When applying this strategy to the GTEx data, we created the test and training sets 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 ensured that rows in the test set were independent of rows in the training set.

Visualizing U_k. In the simulations, and in our application to the GTEx data, $R = 44$, so each U_k is a 44 × 44 covariance matrix, and each component of the mixture (eq. 1) is a distribution in 44 dimensions. Visualizing such a distribution is challenging. However, we can get some insight from the first eigenvector of U_k, denoted by v_k, which captures the principal direction of the effects. 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; alternatively, if v_k has one large element, with other elements close to zero, then component k corresponds to an effect that is strong in only one condition. See Fig. 3 and Supplementary Fig. 2 for illustration.

Relationship with existing methods. The `mash` method includes many existing methods for joint analysis of multiple effects as special cases. Specifically, many existing methods correspond to making particular choices for the set of "canonical" covariance matrices (with no data-driven matrices). For example, a simple "fixed effects" meta-analysis—which assumes equal effects in all conditions—corresponds to $K = 1$ with $U_1 = \mathbf{1}\mathbf{1}^T$ (a matrix of all ones). (This covariance matrix is singular, but it is still allowed in `mash`.) A more flexible assumption is that effects in different conditions are normally distributed about some mean—this corresponds to the multivariate normal assumption made in `mash` if the mean is assumed to be normally distributed as in Wen & Stephens[26]. More flexible still are models that allow effects to be exactly zero in some subset of conditions, as in Flutre *et al.*[5] and Li *et al.*[6]. These models correspond to using (singular) covariances U_k with zeros in the rows and columns corresponding to the subset of conditions with no effect.

`mash` extends the capabilities of 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 importantly, `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—as in the GTEx data, with $R = 44$—where it becomes impractical to pre-specify canonical matrices for all patterns of sharing that might occur. For example, Flutre *et al.*[5] and Li *et al.*[6] consider all 2^R combinations of sparsity in the effects, which is feasible for $R = 9$ (see [20]), but impractical at $R = 44$. While it is possible to restrict the number of combinations considered (e.g., BMAlite[5]), this comes at a cost to flexibility. The addition of data-driven covariance matrices helps to address this issue, making `mash` both flexible and computationally tractable for moderately large R.

RRMSE (accuracy of estimates in simulation studies). Since the ground-truth effects are known in the simulations, we can assess error in the effect estimates from each method. Let b^*_{jr} denote the estimate of effect j in condition r (for mash, $b^*_{jr} = \hat{\hat{b}}_{jr}$). Then the RRMSE (see Fig. 2) is computed as

$$\mathrm{RRMSE} = \frac{\sqrt{\frac{1}{JR}\sum_{j=1}^{J}\sum_{r=1}^{R}(b_{jr} - b_{jr}^{*})^2}}{\sqrt{\frac{1}{JR}\sum_{j=1}^{J}\sum_{r=1}^{R}(b_{jr} - \hat{b}_{jr})^2}}.$$

ROC curves. For the ROC curves in Fig. 2c–d, the True Positive Rate (TPR) and False Positive Rate (FPR) are computed at any given threshold *t* as

$$\mathrm{TRP} := \frac{|CS \cap S|}{|T|}$$
$$\mathrm{FPR} := \frac{|N \cap S|}{|N|},$$

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:

$$S := \{j, r : \ lfsr_{jr} < t\}$$
$$CS := \{j, r : \ E(b_{jr} \mid D) \times b_{jr} > 0\}$$
$$N := \{j, r : \ b_{jr} = 0\}$$
$$T := \{j, r : \ b_{jr} \neq 0\}.$$

Thus, to be considered a true positive, we require that the effect not only be "significant", but also be correctly signed.

For the ROC curves in Fig. 2e–f, the TPR and FPR are computed by treating entire 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

$$TPR := |\tilde{S}|/|\tilde{T}|$$
$$FPR := |\tilde{S}|/|\tilde{N}|,$$

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:

$$\tilde{S} := \{j : \ p_j \leq t\}$$
$$\tilde{N} := \{j : \ b_{jr} = 0, \forall r\}$$
$$\tilde{T} := \{j : \ \exists\, r \text{ s.t. } b_{jr} \neq 0\}.$$

Effective sample size. We define the "effective sample size" (ESS) for tissue *r* to be

$$n_r^{\text{eff}} := n_r^{\text{orig}} \times \text{median}_j\left(\hat{s}_{jr}^2 / \tilde{s}_{jr}^2\right)$$

where n_r^{orig} is the number of available samples in tissue *r*, and $\hat{s}_{jr}$ and $\tilde{s}_{jr}$ are the standard error and posterior standard deviation, respectively, for effect *j* in tissue *r*.

Sharing by sign and magnitude. To assess pairwise sharing by sign between pair of tissues *r* and *s* (Supplementary Fig. 6), we compute, for QTLs that are "significant" (*lfsr* < 0.05) in at least one of *r* and *s*, the fraction that have effect estimates of the same sign.

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

More formally, we define

$$\begin{aligned}
\text{QTL}_r &:= \left\{j : \ lfsr_{jr} < 0.05\right\} \\
\text{SS}_{rs} &:= \left\{j : \text{sign}\left(\hat{\hat{b}}_{jr}\right) = \text{sign}(\hat{\hat{b}}_{js})\right\} \\
\text{SM}_{rs} &:= \left\{j : \tfrac{1}{2} \leq \hat{\hat{b}}_{jr} / \hat{\hat{b}}_{js} \leq 2\right\}.
\end{aligned}$$

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

$$\frac{|\text{SS}_{rs} \cap (\text{QTL}_r \cup \text{QTL}_s)|}{|\text{QTL}_r \cup \text{QTL}_s|},$$

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

$$\frac{|\text{SM}_{rs} \cap (\text{QTL}_r \cup \text{QTL}_s)|}{|\text{QTL}_r \cup \text{QTL}_s|}.$$

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

`ash` analyses. For comparison, we applied the univariate shrinkage method `ash`[15] separately to 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.

`bmalite` analyses. For comparison with `mash`, we implemented a version of `mash-bmalite`[5] that outputs effect size estimates and *lfsr* values. This version of `mash-bmalite` can be considered a variation of `mash` without the data-driven covariance matrices, and with particular

choices for the canonical covariance matrices, and with a grid of scaling factors that is smaller than the one used in mash (consistent with the coarser grid used in Flutre *et al.*[5]).

Specifically, the list of U_k for `mash-bmalite` includes the 44 singleton configurations ($U_k = e_k e_k^T$), and matrices corresponding to the models in Flutre *et al.*[5], with heterogeneity parameters H = {0, 0.25, 0.5, 1}. (When H = 0, effects are equal in all conditions; when H = 1, effects are independent among conditions.) We use grid of scaling factors $\boldsymbol{\omega}$ = {0.1, 0.4, 1.6, 6.4, 25.6}, consistent with the coarse grid used in Flutre *et al.*[5], and chosen to capture the full range of the GTEx (real and simulated) Z statistics.

"Shared, structured effects" simulations. For these scenarios, we simulated $\boldsymbol{b}_j$ from the model given by eq. 3, with equal weights on the 8 covariance matrices learned from the GTEx data, but with the scaling factors $\boldsymbol{\omega}$ simulated from a continuous distribution rather than a fixed grid.

In greater detail, the data simulation procedure was as follows:

1. Take the set of 8 "data-driven" covariance matrices learned from the GTEx data, standardized as described above.
2. Simulate 400 "true effects": independently for each effect j, (a) choose a component k at random from 1,….,8, with equal probability, (b) simulate a scaling factor ω as the absolute value of an $N(0,1)$ random variable, and (c) simulate $\boldsymbol{b}_j \sim N_{44}(\mathbf{0}, \omega U_k)$.
3. For the 19,600 "null effects", set $\boldsymbol{b}_j = 0$.
4. For all 20,000 effects, simulate $\widehat{\boldsymbol{b}}_j \sim N_{44}(\boldsymbol{b}_j, V_j)$, where V_j is the diagonal matrix with diagonal elements 0.1^2. Here, all standard errors are approximately 0.1, similar to the GTEx dataset.

"Shared, unstructured effects" simulations. In these simulations, the 400 true effects were all *i.i.d.* $\boldsymbol{b}_j \sim N_{44}(\mathbf{0}, 0.1^2 I)$. Other details remain unchanged from the "Shared, Structured Effects" simulation procedure.

"Independent effects" simulations. We also simulated data in which effects were entirely independent across conditions. This was accomplished as follows:

1. Independently for each r = 1,…,44, choose a random set of 400 $j \in \{1, \ldots, 2 \times 10^4\}$ to be the "true" effects.
2. For the "true effects", simulate $b_{jr} \sim N(0, \Sigma^2)$, where Σ^2 is chosen with equal probability from the set {0.1, 0.5, 0.75, 1}. In this way, the true effects capture small and large effects for each condition. All other effects are set to zero.
3. Simulate $\widehat{\boldsymbol{b}}_j \sim N(\boldsymbol{b}_j, V_j)$ as in other simulations.

Analysis of simulated data. Each simulated dataset $(\hat{\boldsymbol{b}}_j, V_j)$ was analyzed using `mash`, following the procedure described above. Note that we re-estimated the U_k and $\boldsymbol{\pi}$ from the data, without making use of the $\boldsymbol{U}$ that was used to simulated the data. We estimated effects by their posterior mean (eq. 8), and assessed significance by the *lfsr*. The `ash` and `mash-bmalite` methods were applied to the simulated data in the same way that they were applied to the GTEx data.

Data availability. The GTEx study data are available through dbGaP under accession phs000424.v6.p1. The GTEx summary statistics used in the mash analysis have been deposited in Zenodo (doi:10.5281/zenodo.1296399).

Code availability. Source code implementing our methods, including processing and analysis of the GTEx data, adaptive shrinkage ("ash") and multivariate adaptive shrinkage ("mash", "mash-bmalite"), is available online, and is distributed under GNU and BSD open source licenses (see URLs).

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**, 16882 (2015).

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**, 1 (2012).

3. Pickrell, J. K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768–772 (2010).

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

5. Flutre, T., Wen, X., Pritchard, J. & Stephens, M. A statistical framework for joint eQTL analysis in multiple tissues. *PLoS Genetics* **9**, e1003486 (2013).

6. Li, G., Shabalin, A. A., Rusyn, I., Wright, F. A. & Nobel, A. B. An Empirical Bayes approach for multiple tissue eQTL Analysis. *Biostatistics* **19**, 391–406 (2017).

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).

8. Wen, X. & Stephens, M. Using linear predictors to impute allele frequencies from summary of pooled genotype data. *Annals of Applied Statistics* **4**, 1158–1182 (2010).

9. Han, B. & Eskin, E. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *American Journal of Human Genetics* **88**, 586–598 (2011).

10. Stephens, M. Unified framework for association analysis with multiple related phenotypes. *PLoS ONE* **8**, e65245 (2013).

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).

12. Wei, Y., Tenzen, T. & Ji, H. Joint analysis of differential gene expression in multiple studies using correlation motifs. *Biostatistics* **16**, 31–46 (2015).

13. Pickrell, J., Berisa, T., Ségurel, L., Tung, J. Y. & Hinds, D. Detection and interpretation of shared genetic influences on 40 human traits. *Nature Genetics* **48**, 709–717.

14. Han, B. & Eskin, E. Interpreting meta-analyses of genome-wide association studies. *PLoS Genetics* **8**, e1002555 (2012).

15. Stephens, M. False discovery rates: a new deal. *Biostatistics* 18: 275–294 (2017).

16. Bovy, J., Hogg, D. W. & Roweis, S. T. Extreme Deconvolution: inferring complete distribution functions from noisy, heterogeneous and incomplete observations. *Annals of Applied Statistics* **5**, 1657–1677 (2011).

17. Engelhardt, B. E. & Stephens, M. Analysis of population structure: a unifying framework and novel methods based on sparse factor analysis. *PLoS Genetics* **6**, e1001117 (2010).

18. Efron, B. Local false discovery rates. http://statweb.stanford.edu/~ckirby/brad/papers/2005LocalFDR.pdf (2005).

19. Lebrec, J. J., Stijnen, T. & van Houwelingen, H. C. Dealing with heterogeneity between cohorts in genomewide SNP association studies. *Statistical Applications in Genetics and Molecular Biology* **9**, 8 (2010).

20. GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).

21. Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M.E. & Cox, N.J. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genetics* **6**, e1000888 (2010).

22. Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353–1358 (2012).

23. Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A. & Reich, D. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38**, 904–909 (2006).

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).

25. Tversky, A. & Kahneman, D. Judgment under uncertainty: heuristics and biases. *Science* **185**, 1124–1131 (1974).

26. Wen, X. & Stephens, M. Bayesian methods for genetic association analysis with heterogeneous subgroups: from meta-analyses to gene-environment interactions. *Annals of Applied Statistics* **8**, 176–203 (2014).

27. Servin, B. & Stephens, M. Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genetics* **3**, e114 (2007).

28. Veyrieras, J.-B., Kudaravalli, S., Kim, S. Y., Dermitzakis, E. T., Gilad, Y., Stephens, M. & Pritchard, J. K. High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genetics* **4**, e1000214 (2008).

29. Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B. & Eskin, E. Identifying causal variants at loci with multiple signals of association. *Genetics* **198**, 497–508 (2014).

30. Kichaev, G., Yang, W.-Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A. L., Kraft, P. & Pasanuic, B. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genetics* **10**, 1–16 (2014).

31. Chen, W., Larrabee, B. R., Ovsyannikova, I. G., Kennedy, R. B., Haralambieva, I. H, Poland, G.A. & Schaid, D. J. Fine mapping causal variants with an approximate Bayesian method using marginal test statistics. *Genetics* **200**, 719–736 (2015).

32. Moyerbrailean, G.A., Kalita, C.A., Harvey, C.T., Wen, X., Luca, F. & Pique-Regi, R. Which genetic variants in DNase-seq footprints are more likely to alter binding? *PLOS Genetics* **12**, 1–27 (2016).

33. Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace, C. & Plagnol, V. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genetics* **10**, 1–15 (2014).

34. Fortune, M. D., Guo, H., Burren, O., Schofield, E., Walker, N. M., Ban, M., Sawcer, S. J., Bowes, J., Worthington, J., Barton, A., Eyre, S., Todd, J. A. & Wallace, C. Statistical colocalization of genetic risk variants for related autoimmune diseases in the context of common controls. *Nature Genetics* **47**, 839–846 (2015).

35. Nica, A.C., Montgomery, S.B., Dimas, A.S., Stranger, B.E., Beazley, C., Barroso, I. & Dermitzakis, E. T. Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genetics* **6**, 1–11 (2010).

36. Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X. & Mortazavi, A. A survey of best practices for RNA-seq data analysis. *Genome Biology* **17**, 13 (2016).

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* **57**, 289–300 (1995).

38. Storey, J. D. The positive false discovery rate: a Bayesian interpretation and the *q*-value. *Annals of Statistics* **31**, 2013–2035 (2003).

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. Simple and globally convergent methods for accelerating the convergence of any EM algorithm. *Scandinavian Journal of Statistics* **35**, 335–353 (2008).

Acknowledgements This work was supported by NIH grants MH090951 and HG02585 to M.S., and by a grant from the Gordon and Betty Moore Foundation (GBMF #4559). S.M.U. was supported by NIH grant T32HD007009. 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/SAIC-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 supplement 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 October 17, 2015.

Author contributions S.M.U. and M.S. conceived of the project and developed the statistical methods. S.M.U. implemented the comparisons with simulated data. S.M.U. and G.W. performed the analyses of the GTEx data, and additional analyses. S.M.U., G.W. and M.S. implemented the software, with contributions from P.C. S.M.U. and M.S. wrote the manuscript, with input from G.W. and P.C. P.C. and G.W. prepared the online code and data resources.

Competing interests The authors declare no competing financial interests.

Figure legends

Figure 1 | Overview of fitting procedure in mash, which estimates the multivariate distribution of effects present in the data. The data consist of a matrix of summary data (e.g., *Z* scores) for a large number of units (e.g., gene-SNP pairs) in multiple conditions (e.g., tissues), and, optionally, their standard errors (not shown). Color (yellow, blue) indicates the sign of an effect or covariance (positive, negative), with shading intensity indicating size. After selecting rows containing the strongest signals (**1**)—in this example, the top 6 rows—we apply covariance estimation and dimension-reduction techniques to estimate candidate "data-driven" covariance matrices U_k (**2**). To these, we add "canonical" covariance matrices U_k, including the identity matrix, and matrices representing condition-specific effects. Each covariance matrix represents a "pattern" of effects that may occur in the data. We scale each covariance matrix by a grid of scaling factors, ω_l, varying from "very small" to "very large", which allow for *a priori* effect sizes to range from very small to very large. Using the whole data set (**3**), we compute maximum-likelihood estimates of the weights (relative frequencies) $\pi_{k,l}$ for each (U_k, ω_l) combination; this corresponds to estimating how commonly each pattern-effect size combination occurs in the data. Finally (**4**), we compute posterior statistics using the fitted model; the posterior mean estimates shown (bottom-right) illustrate that effect estimates are "shrunk" adaptively using the fitted `mash` model.

Figure 2 | Comparison of methods on simulated data. Results are shown for two simulation scenarios: "shared, structured effects", in which 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", in which the non-zero effects are shared among the 44 conditions with effect sizes and directions that are independent among conditions. Panels **a** and **b** summarize

the error in the estimated effects in these two simulation scenarios (RRMSE obtained from n = 20,000 observed effects in 44 simulated tissues). Panels **c**–**f** show ROC curves for detecting significant effects; in **c**–**d,** each effect in each condition is treated as a separate discovery (n = 44 × 20,000 = 880,000 discoveries), which requires condition-specific measures of significance, whereas in **e**–**f** each effect is treated as a single discovery across all conditions (n = 20,000 discoveries), which requires only a single measure of significance, as in traditional meta-analyses. In **c**–**d**, we also require the estimated sign (+/–) of each significant effect to be correct to be considered a "true positive". Our new method (`mash`) outperformed other methods, particularly for the "shared, structured effects" scenario that is expected to reflect genomics applications.

Figure 3 | Summary of primary patterns identified by `mash` in GTEx data. Shown are the heatmap of the correlation matrix (**a**) and bar plots of the first three eigenvectors (**b**, **c**, **d**) of the covariance matrix U_k corresponding to the dominant mixture component identified by `mash` (n = 16,069 independent gene-SNP pairs). This component accounts for 34% of all weight in the GTEx data. In all cases, tissues are color-coded as indicated by the tissue labels in the heatmap. The first eigenvector (**b**) reflects broad sharing among all tissues, with all effects in the same direction; the second eigenvector (**c**) captures differences between brain (and, to a lesser extent, testis and pituitary) and other tissues; the third eigenvector (**d**) primarily captures effects that are stronger in whole blood.

Figure 4 | Examples illustrating that `mash` uses learned patterns of sharing to inform effect estimates in the GTEx data. In panel **a**, each colored dot shows the original ("raw") effect estimate for a single tissue (color-coded as in Fig. 3), with grey bars indicating ±2 standard errors. These are the data provided to `mash`. Panel **b** shows the corresponding estimates outputted by `mash`. In each case, `mash` combines information across all tissues, using the background information—patterns of sharing—learned from data on all eQTLs, to produce more precise estimates. Panel **c** shows, for contrast, the corresponding estimates from `mash-bmalite`, which—due to its more restricted model—fails to capture features clearly apparent in the original data (e.g., strong brain effects in *MCPH1*). In **b** and **c**, colored dots are posterior means, and error bars depict ±2 posterior standard deviations. For all estimates, n = 83–430 individuals, depending on the tissue (refer to Supplementary Fig. 7). See the main text for a detailed discussion.

Figure 5 | Number of tissues shared by sign and magnitude. Histograms show estimated number of tissues in which top eQTLs are "shared," considering all tissues (n = 12,171 gene-SNP pairs with a significant eQTL in at least one tissue), non-brain tissues (n = 12,117), and brain tissues only (n = 8,474), and using two different sharing definitions: by sign (**a**) and by magnitude (**b**). Sharing by sign means that the eQTLs have the same sign in the estimated effect; sharing by magnitude means that they also have similar effect size (within a factor of 2).

Figure 6 | Pairwise sharing by magnitude of eQTLs among tissues. For each pair of tissues, we considered the top eQTLs that were significant (*lfsr* < 0.05) in at least one of the two

tissues, and plotted 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 (n = 5,605–9,811 gene-SNP pairs, depending on pair of tissues compared). Brackets surrounding tissue labels highlight groups of biologically related tissues mentioned in the text as showing particularly high levels of sharing.

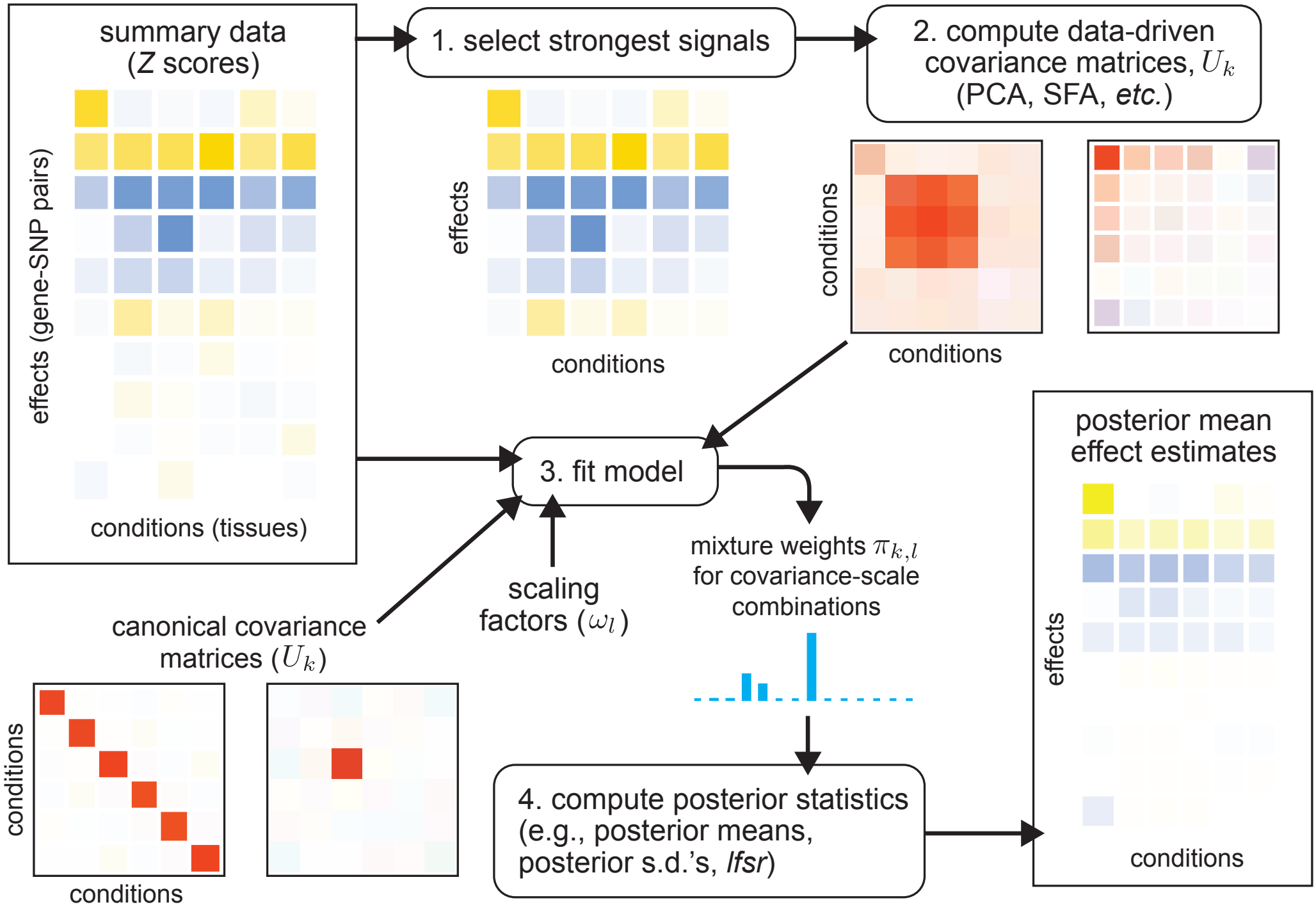
Tables

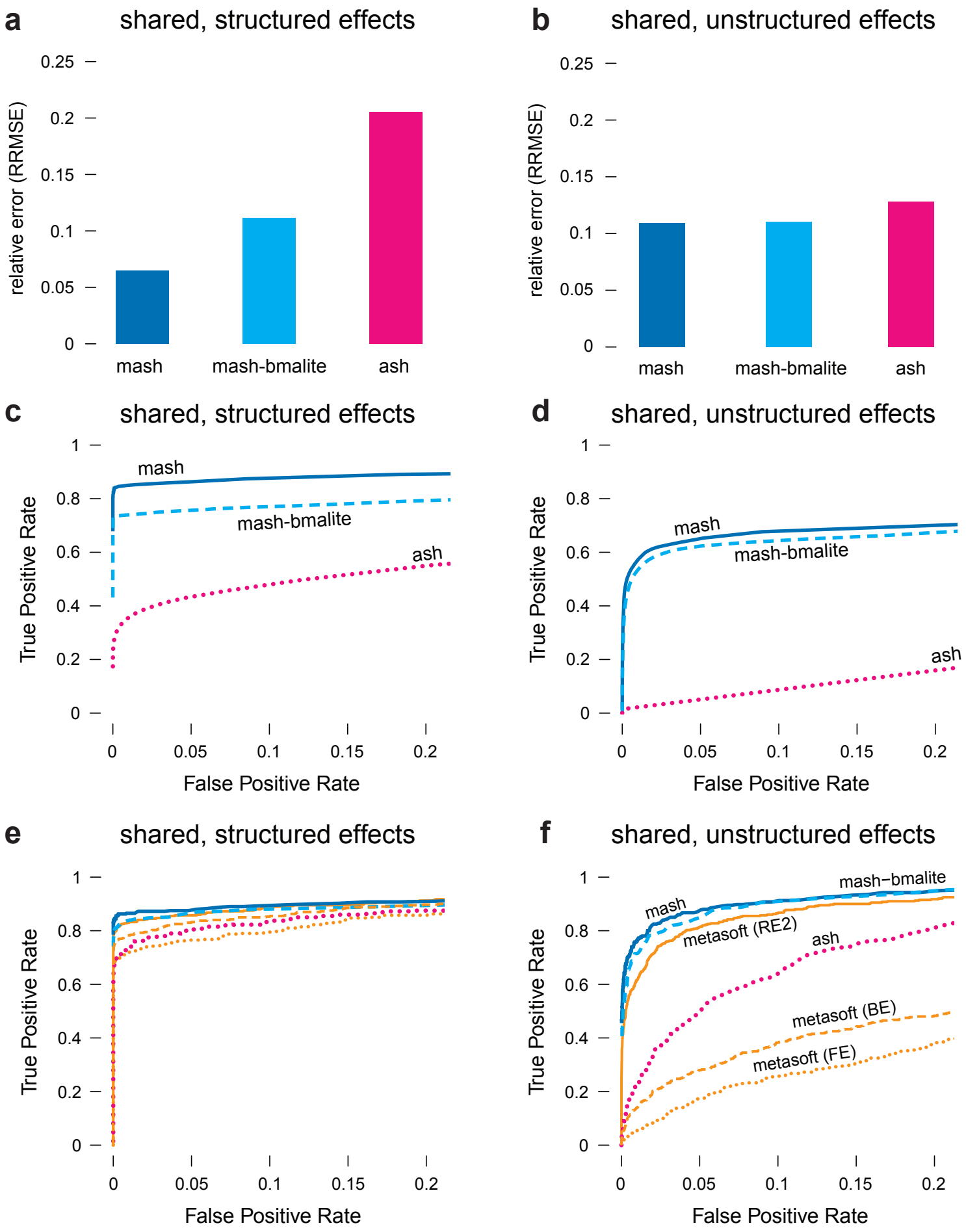
	mean absolute error	
method	**sharing by sign**	**sharing by magnitude**
raw	4.70	4.32
ash	4.75	5.67
mash-bmalite	3.67	4.39
mash	2.44	2.29

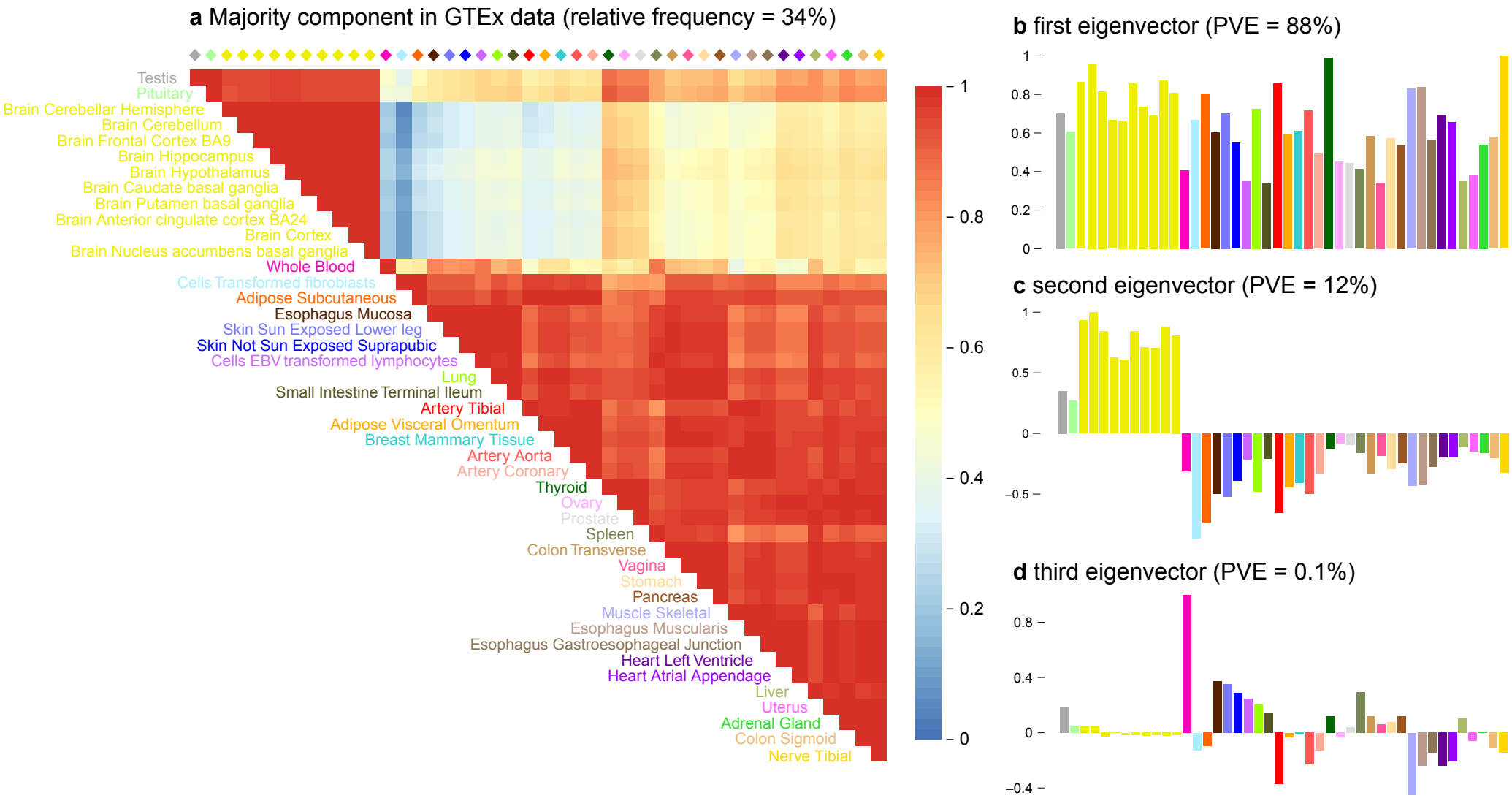
Table 1 | Performance of sharing estimates in simulated data. For each method, the table shows the mean absolute error in the estimated number of conditions that share effects (by sign or magnitude) with the condition estimated to have the largest effect. Here, "raw" indicates performance of the original estimates provided as input to all methods. Each error value is taken as a mean across the 400 non-null effects simulated in the "shared, structured effects" scenario.

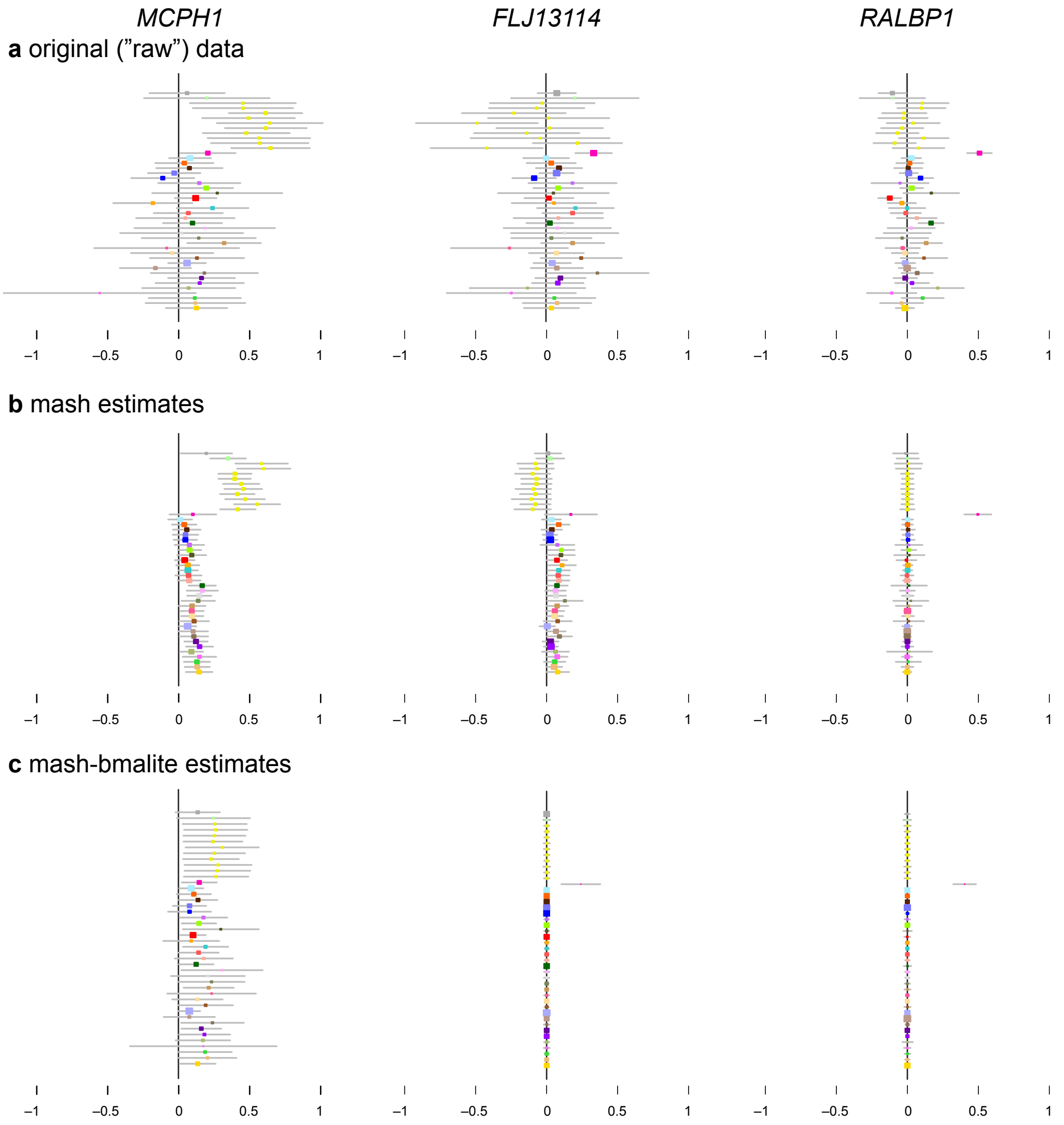
	proportion of effects		
sharing criterion	**all tissues**	**non-brain**	**brain**
shared by sign ($\tilde{b} > 0$)	0.85	0.85 (0.88)	0.96 (0.98)
shared by magnitude ($\tilde{b} > 0.5$)	0.36	0.40 (0.44)	0.76 (0.85)

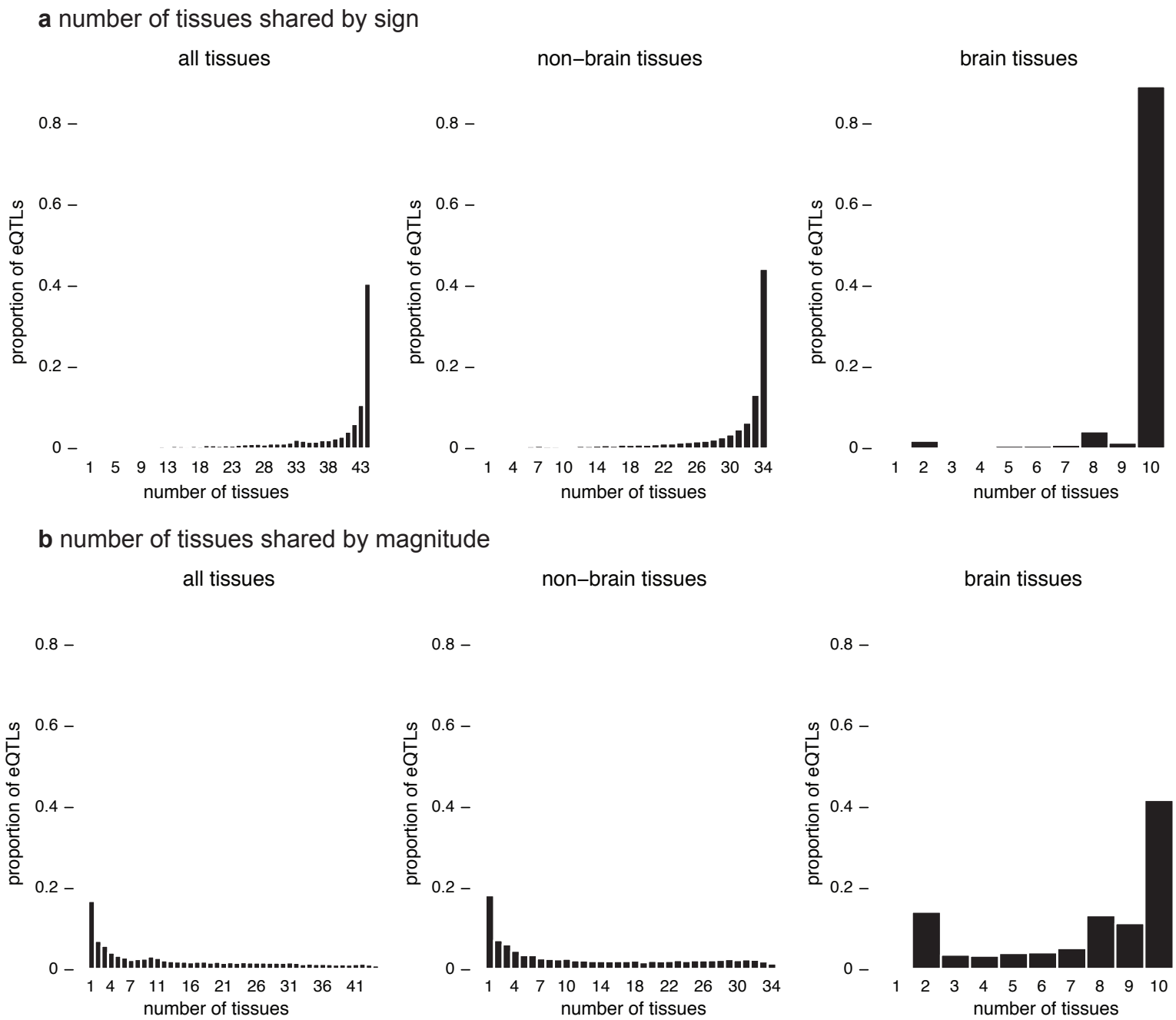
Table 2 | Summary of sharing among top eQTLs. Numbers give the proportion of effects meeting a given sharing criterion. "Shared by sign" requires that the effect has the same sign as the strongest effect among tissues. "Shared by magnitude" requires that the effect is also within a factor of 2 of the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

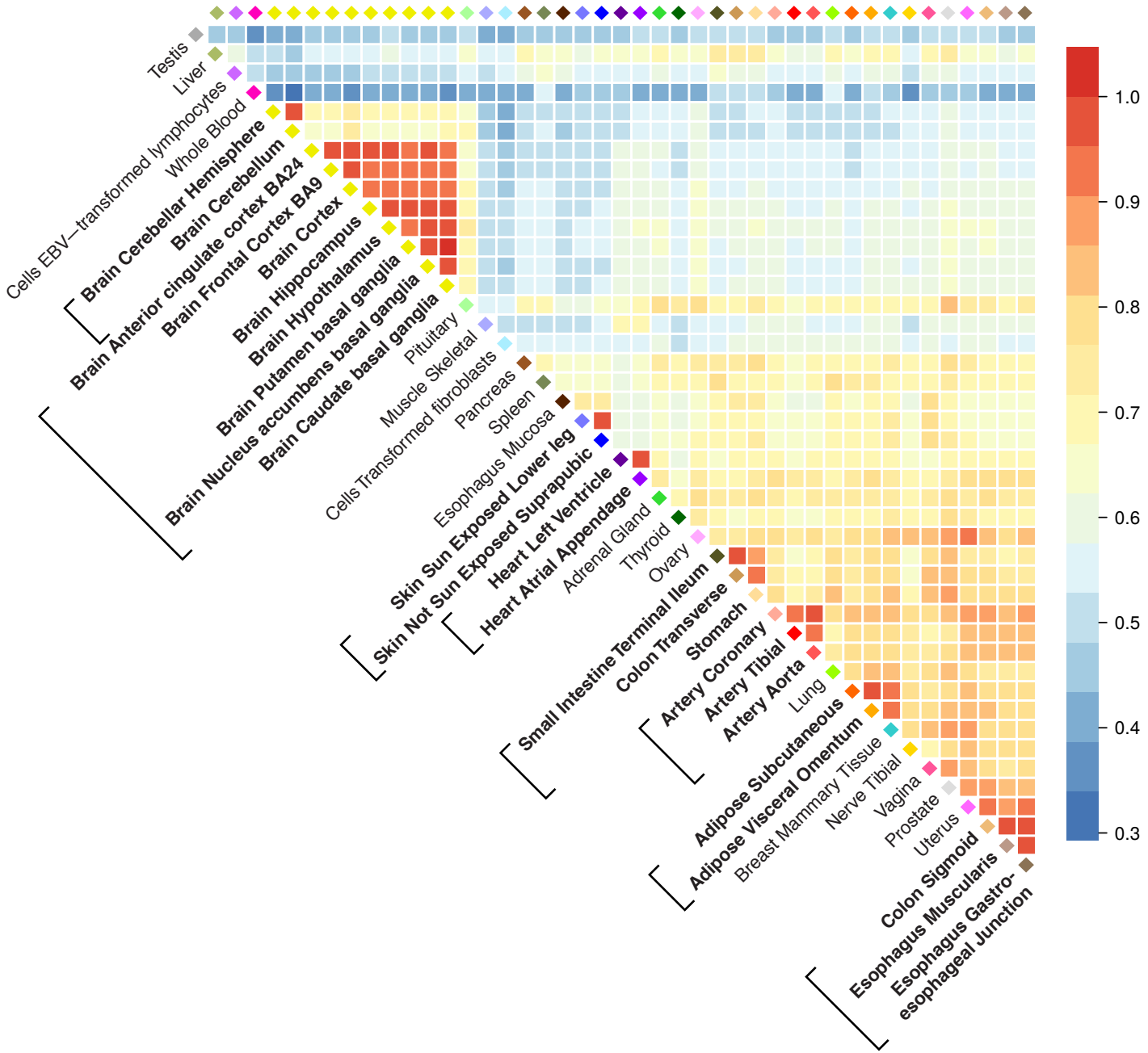

summary data
(Z scores)
effects (gene-SNP pairs)
conditions (tissues)
1. select strongest signals
effects
conditions
2. compute data-driven
covariance matrices, U_k
(PCA, SFA, etc.)
conditions
conditions
3. fit model
scaling
factors (ω_l)
mixture weights $\pi_{k,l}$
for covariance-scale
combinations
canonical covariance
matrices (U_k)
conditions
conditions
4. compute posterior statistics
(e.g., posterior means,
posterior s.d.'s, lfsr)
posterior mean
effect estimates
effects
conditions


a
shared, structured effects
relative error (RRMSE)
0.25
0.2
0.15
0.1
0.05
0
mash
mash-bmalite
ash
b
shared, unstructured effects
relative error (RRMSE)
0.25
0.2
0.15
0.1
0.05
0
mash
mash-bmalite
ash
c
shared, structured effects
True Positive Rate
1
0.8
0.6
0.4
0.2
0
mash
mash-bmalite
ash
0
0.05
0.1
0.15
0.2
False Positive Rate
d
shared, unstructured effects
True Positive Rate
1
0.8
0.6
0.4
0.2
0
mash
mash-bmalite
ash
0
0.05
0.1
0.15
0.2
False Positive Rate
e
shared, structured effects
True Positive Rate
1
0.8
0.6
0.4
0.2
0
0
0.05
0.1
0.15
0.2
False Positive Rate
f
shared, unstructured effects
True Positive Rate
1
0.8
0.6
0.4
0.2
0
mash−bmalite
mash
metasoft (RE2)
ash
metasoft (BE)
metasoft (FE)
0
0.05
0.1
0.15
0.2
False Positive Rate


a Majority component in GTEx data (relative frequency = 34%)
Testis
Pituitary
Brain Cerebellar Hemisphere
Brain Cerebellum
Brain Frontal Cortex BA9
Brain Hippocampus
Brain Hypothalamus
Brain Caudate basal ganglia
Brain Putamen basal ganglia
Brain Anterior cingulate cortex BA24
Brain Cortex
Brain Nucleus accumbens basal ganglia
Whole Blood
Cells Transformed fibroblasts
Adipose Subcutaneous
Esophagus Mucosa
Skin Sun Exposed Lower leg
Skin Not Sun Exposed Suprapubic
Cells EBV transformed lymphocytes
Lung
Small Intestine Terminal Ileum
Artery Tibial
Adipose Visceral Omentum
Breast Mammary Tissue
Artery Aorta
Artery Coronary
Thyroid
Ovary
Prostate
Spleen
Colon Transverse
Vagina
Stomach
Pancreas
Muscle Skeletal
Esophagus Muscularis
Esophagus Gastroesophageal Junction
Heart Left Ventricle
Heart Atrial Appendage
Liver
Uterus
Adrenal Gland
Colon Sigmoid
Nerve Tibial
1
0.8
0.6
0.4
0.2
0
b first eigenvector (PVE = 88%)
1
0.8
0.6
0.4
0.2
0
c second eigenvector (PVE = 12%)
1
0.5
0
−0.5
d third eigenvector (PVE = 0.1%)
0.8
0.4
0
−0.4


MCPH1
FLJ13114
RALBP1
a original ("raw") data
b mash estimates
c mash-bmalite estimates
–1
–0.5
0
0.5
1


a number of tissues shared by sign
all tissues
non−brain tissues
brain tissues
proportion of eQTLs
0.8
0.6
0.4
0.2
0
1 5 9 13 18 23 28 33 38 43
1 4 7 10 14 18 22 26 30 34
1 2 3 4 5 6 7 8 9 10
number of tissues
b number of tissues shared by magnitude
all tissues
non−brain tissues
brain tissues
1 4 7 11 16 21 26 31 36 41
1 4 7 10 14 18 22 26 30 34
1 2 3 4 5 6 7 8 9 10


Testis
Liver
Cells EBV–transformed lymphocytes
Whole Blood
Brain Cerebellar Hemisphere
Brain Cerebellum
Brain Anterior cingulate cortex BA24
Brain Frontal Cortex BA9
Brain Cortex
Brain Hippocampus
Brain Hypothalamus
Brain Putamen basal ganglia
Brain Nucleus accumbens basal ganglia
Brain Caudate basal ganglia
Pituitary
Muscle Skeletal
Cells Transformed fibroblasts
Pancreas
Spleen
Esophagus Mucosa
Skin Sun Exposed Lower leg
Skin Not Sun Exposed Suprapubic
Heart Left Ventricle
Heart Atrial Appendage
Adrenal Gland
Thyroid
Ovary
Small Intestine Terminal Ileum
Colon Transverse
Stomach
Artery Coronary
Artery Tibial
Artery Aorta
Lung
Adipose Subcutaneous
Adipose Visceral Omentum
Breast Mammary Tissue
Nerve Tibial
Vagina
Prostate
Uterus
Colon Sigmoid
Esophagus Muscularis
Esophagus Gastro-esophageal Junction
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3

Supplementary Results

Effects of linkage disequilibrium. Linkage disequilibrium (LD) between SNPs has two distinct effects.

First, LD causes correlations in the observed effects of nearby SNPs for the same gene. This issue is likely minor here; `mash` ignores correlations between rows of $\hat{B}$ when estimating the prior density *g*, and this can be justified as a "composite likelihood" approach[1], which can perform well for computing point estimates of model parameters.

Second, the effect estimates we obtained 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[2,3]. This issue more likely has a greater impact 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 a multi-SNP, multi-tissue method 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 have limited this analysis to highlighting one specific feature of our results that we believe may be a consequence of the use of single-SNP effect estimates, and which will hopefully be better addressed as multi-SNP analyses are developed to better account for LD.

Specifically, we found that LD among multiple causal SNPs can cause single-SNP analyses to identify eQTLs that appear to have strong effects of opposite sign in different tissues. One example is shown in Supplementary Fig. 5; this eQTL has strong, positive *Z* scores in brain tissues, and negative *Z* scores in most other tissues. Initially, this suggests that this eQTL might have causal effects in opposite directions in brain versus non-brain tissues. However, there is another way to explain this result: there could be two eQTLs in LD with one another, one of which (e.g., eQTL A) has a strong effect in brain tissues, and the other of which (e.g., eQTL B) has a strong effect in other tissues. If the expression-increasing allele at eQTL A is in negative LD with the expression-increasing allele at eQTL B, then the single-SNP *Z* scores for both SNPs will show opposite signs in brain versus non-brain. Indeed, closer examination of the data at this particular gene suggests that this explanation is likely correct in this case (Supplementary Fig. 5). A similar example is discussed in the GTEx pilot study[4] (their Supplementary Fig. 14).

Based on this reasoning, we believe that estimates of sharing in sign from single-SNP analyses such as ours are likely to be underestimates of the sharing in sign of actual causal effects. Therefore, we caution against misinterpreting an eQTL in multiple tissues that shows significant effects in different directions.

Increase in effective sample size due to multivariate analysis. A feature of our multivariate analysis approach is improved quantitative estimates of effect sizes in each condition. When estimating effects in a single condition, `mash` uses the data not only from that condition but also from other "similar" conditions. In this way, `mash` effectively increases the available sample size,

and this improves both accuracy and precision of estimates. The improvement will be greatest for conditions that are similar to many other conditions, and weakest for conditions with many "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 ESS estimates (Supplementary Fig. 7) vary from 240 for testis to 1,392 for coronary artery. Other tissues with smaller ESS include liver, pancreas, spleen and brain cerebellum. Identifying tissues with smaller ESS could help guide prioritization of "under-represented" tissues in future experimental efforts.

For testis, the ESS of 240 represents only a small (1.4-fold) increase compared with actual sample size, reflecting that its effects are more "tissue specific"; that is, they are less correlated with other tissues. Other tissues showing a small gain in ESS include transformed fibroblasts and whole blood, which we also highlight for having more "tissue specific" signals. By contrast, the ESS for coronary artery represents a 10-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 4-fold increase in ESS for estimating the top eQTL effects, reflecting an overall moderate-to-large correlation in effect sizes across tissues.

One caveat of this analysis 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 in our study), then the high reported ESS for uterus may not be as useful. In the end, detection of tissue-specific effects will be most benefitted by collecting more samples in the tissue of interest.

Supplementary References

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

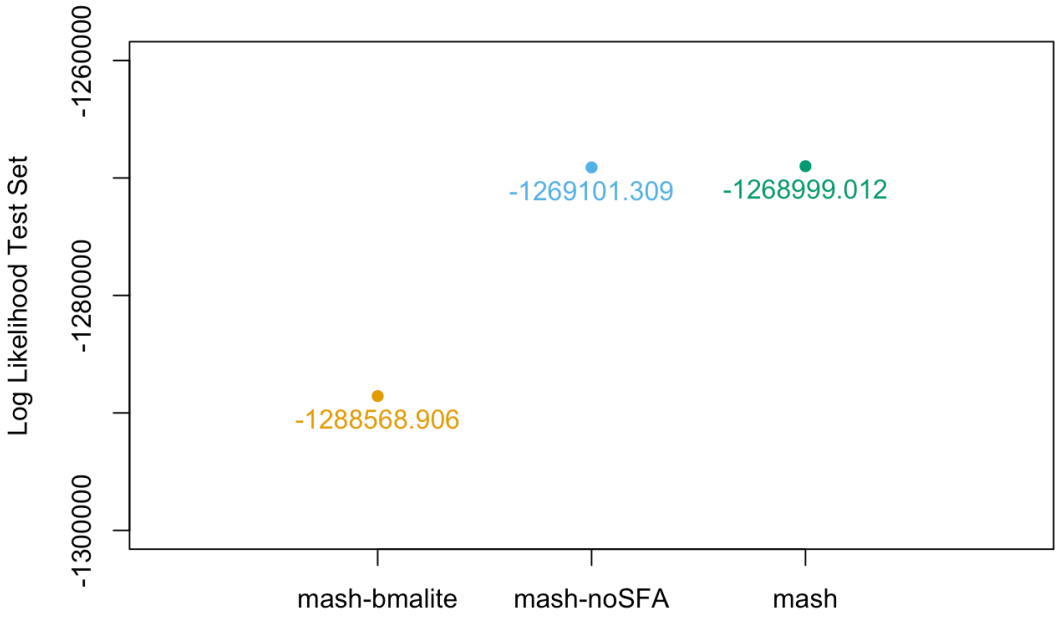
2. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics* **47**, 291–295 (2015).

3. Zhu, X. & Stephens, M. Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. *Annals of Applied Statistics* **11**, 1561–1592 (2017).

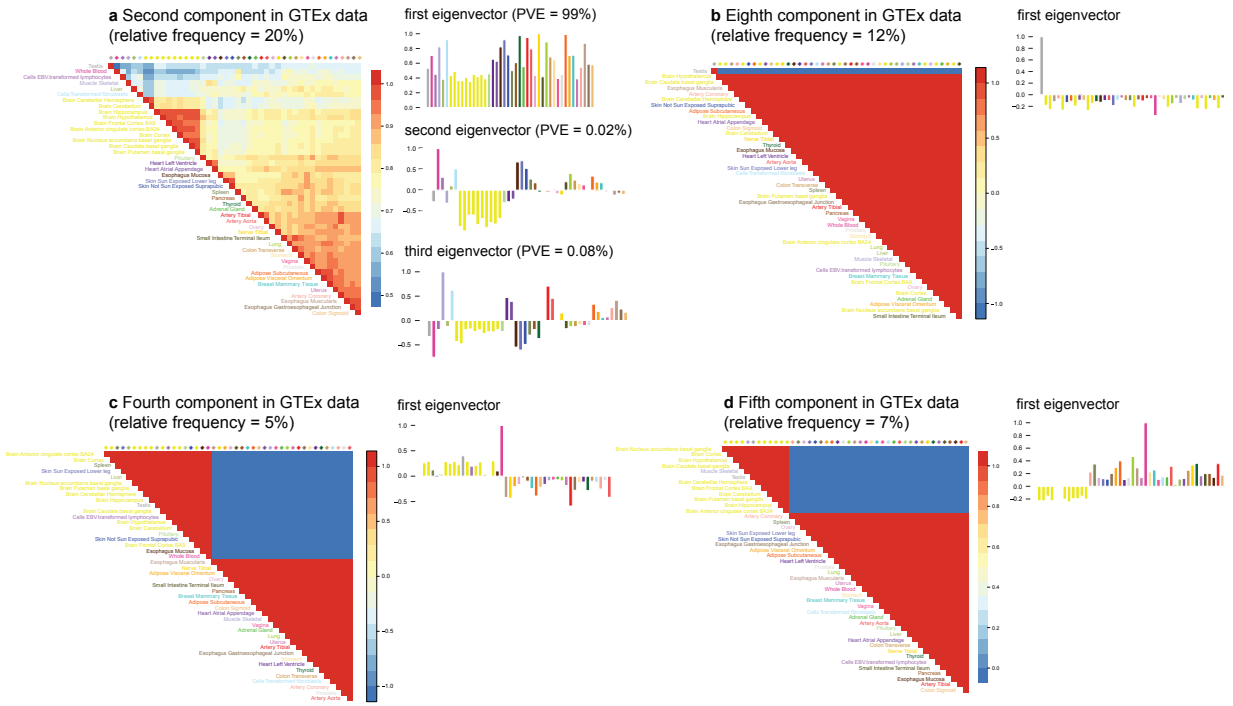
4. GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).

5. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* **81**, 559–575 (2007).

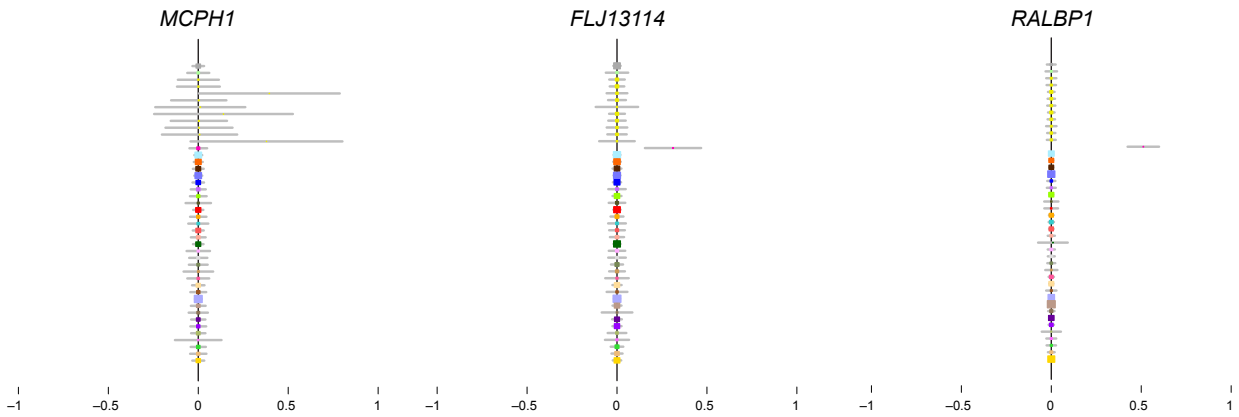
Supplementary Figures



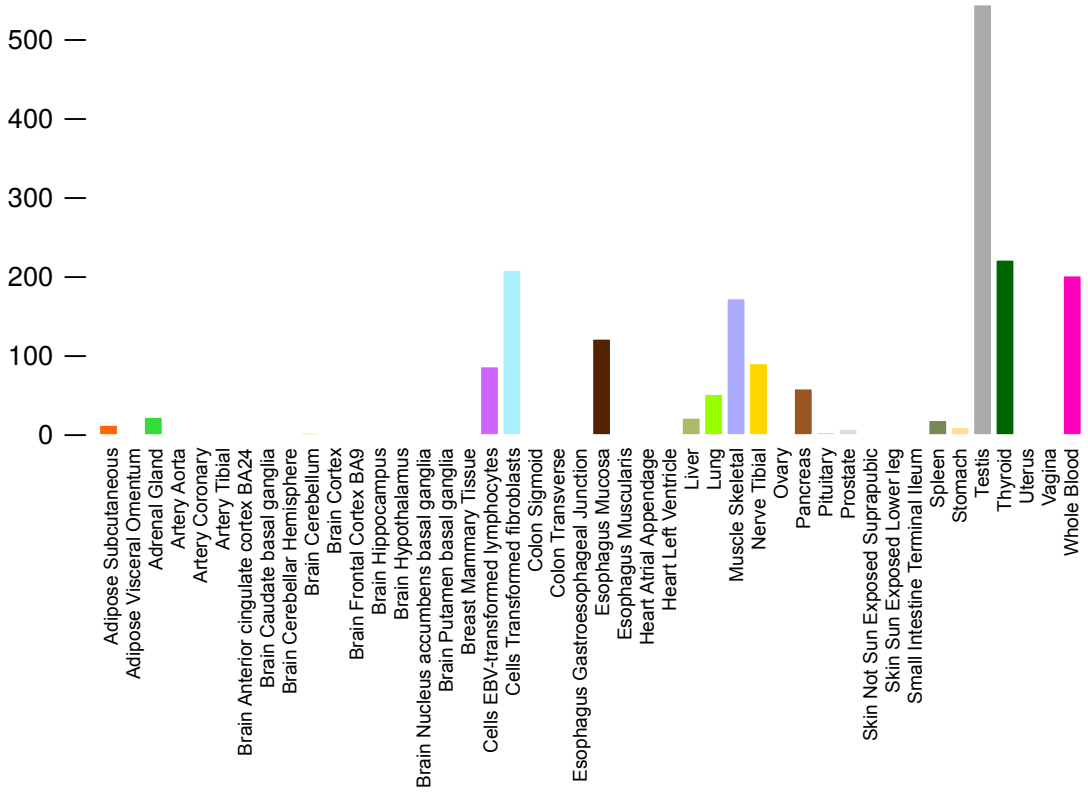
Supplementary Figure 1 | Increase in log-likelihood on test set as new U_k are added. The figure shows the log-likelihood on the test set for different "models" (choices of U_k). From left to right, the models are: `mash-bmalite` (no data-driven U_k); `mash-no-SFA` (the combination of canonical and data-driven covariances, excluding the rank-one 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 increases. Note that the difference in likelihood between the `mash` and `mash-no-SFA` is large—`mash` is approximately 100 log-likelihood units higher than `mash-no-SFA`—although this is difficult to see at this scale. Test-set log-likelihoods are based on n = 28,198 randomly selected gene-SNP pairs.



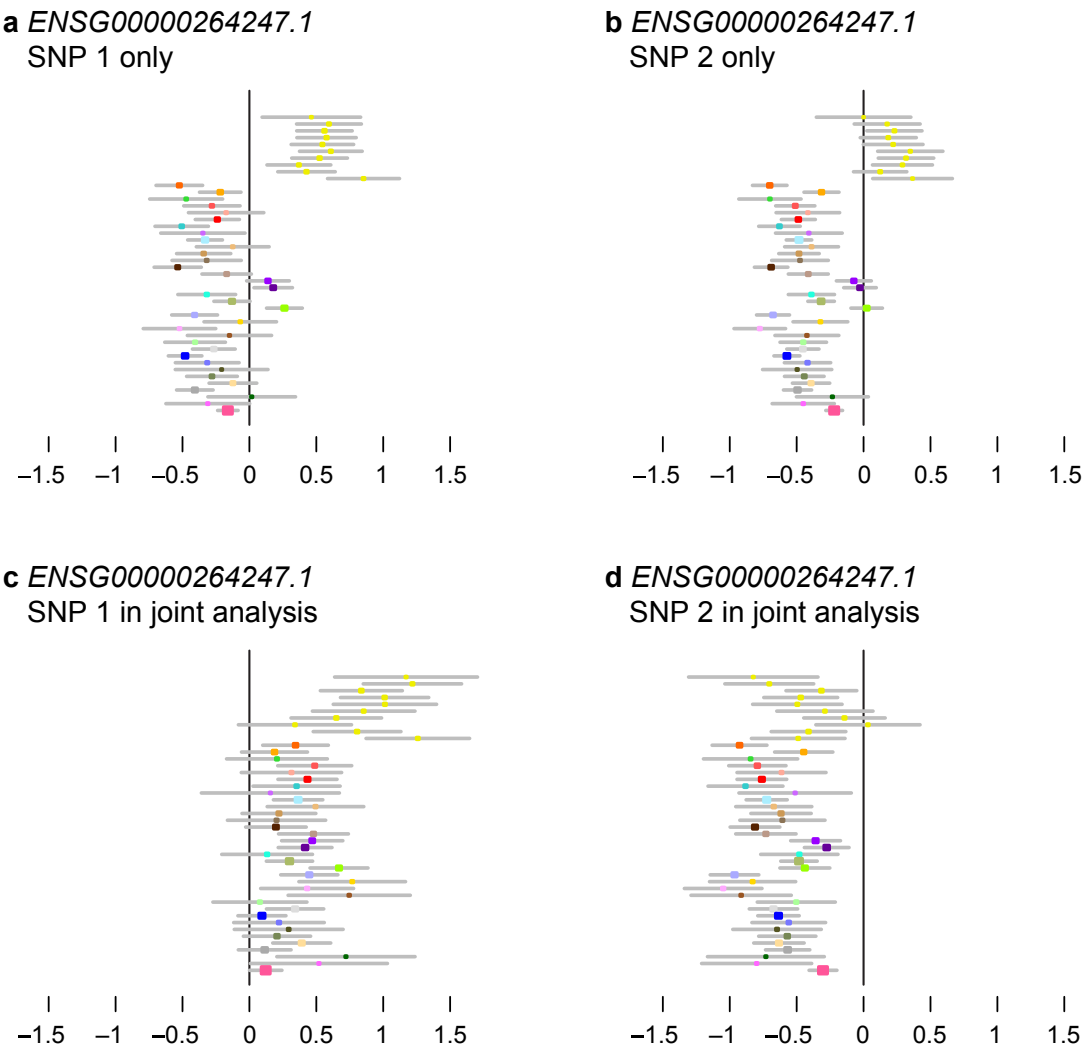
Supplementary Figure 2 | Summary of covariance matrices U_k with greatest estimated weights (>1%) in GTEx data. For each covariance matrix U_k, the figure shows the heatmap of the corresponding correlation matrix, and bar plots of the top eigenvectors of U_k (n = 16,069 independent gene-SNP pairs). Component 2 (**a**) captures qualitatively similar effects to the component highlighted in Fig. 3. Component 8 (**b**) captures testis-specific effects. Components 4 (**c**) and 5 (**d**) primarily capture effects that are stronger in whole blood than in other tissues.



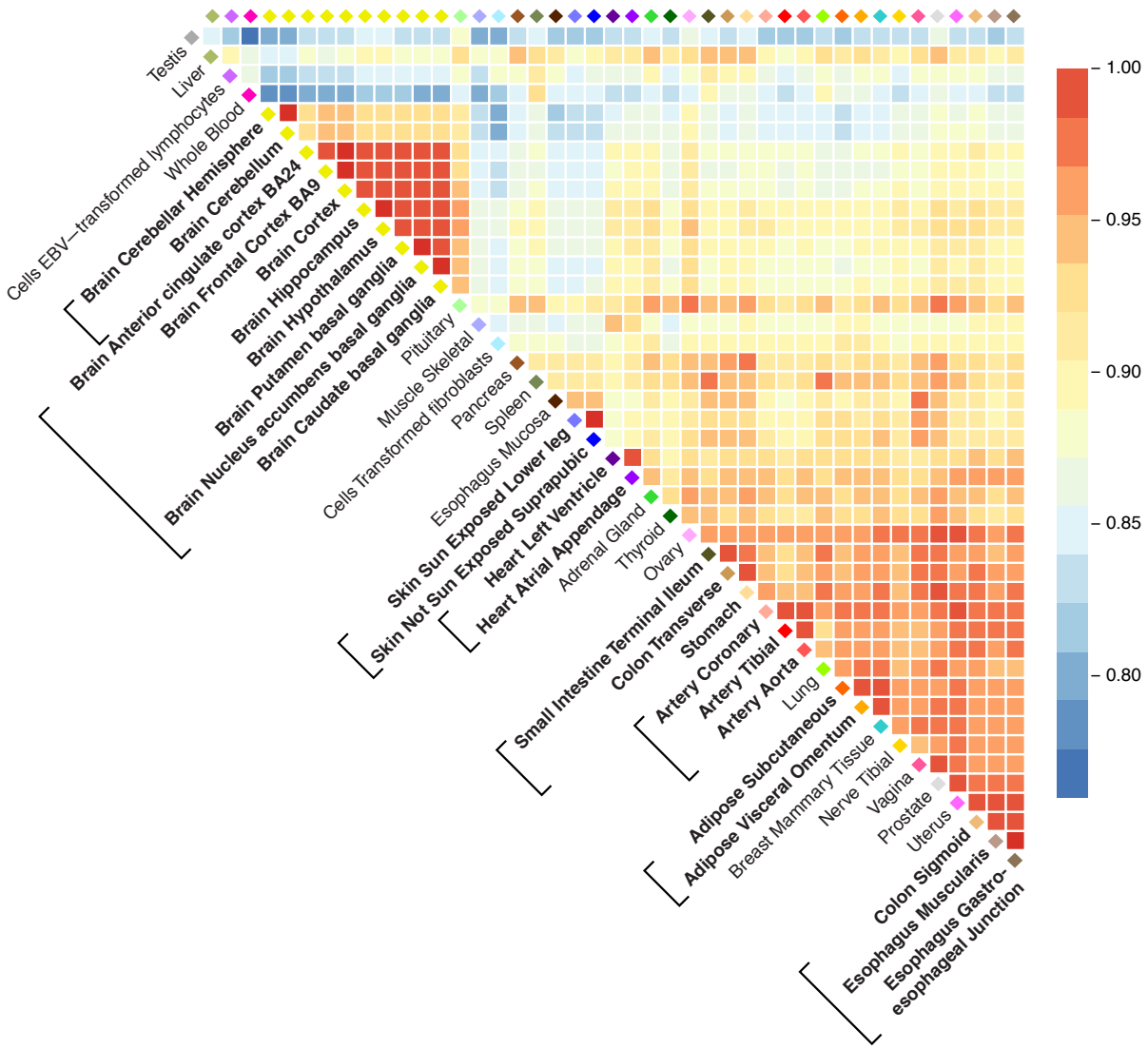
Supplementary Figure 3 | Estimates from the univariate method `ash` for the examples shown in Fig. 4. Each colored dot (color-coded as in Fig. 3) shows the effect estimates (posterior means) from `ash`, with horizontal gray bars indicating ±2 posterior standard deviations. For all estimates, *n* = 83–430 individuals, depending on the tissue (see Supplementary Fig. 7).



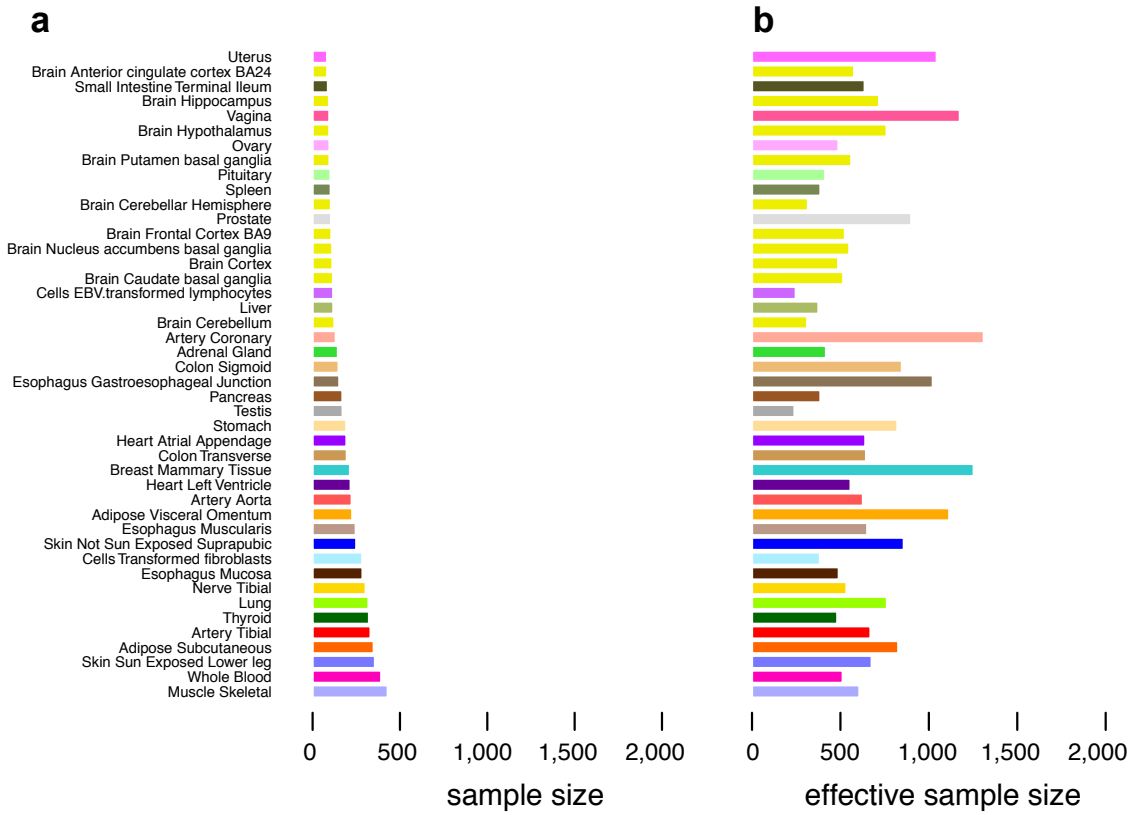
Supplementary Figure 4 | Number of "tissue-specific eQTLs" in each tissue. Here, "tissue-specific" is defined to mean that the effect is at least 2-fold larger in one tissue than in any other (*i.e.*, $\tilde{b}_{jr} > 0.5$ in only one tissue).



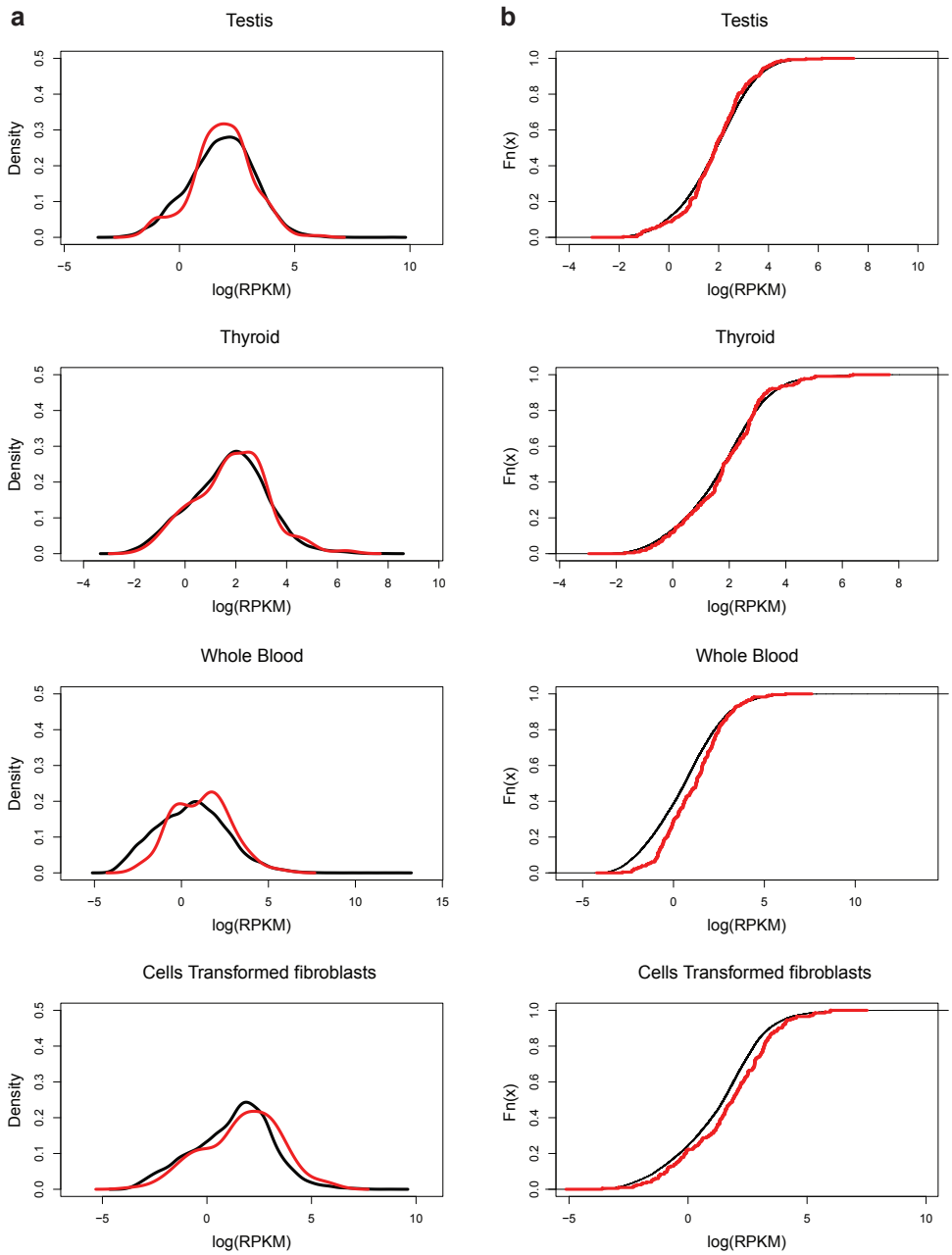
Supplementary Figure 5 | Illustration of how linkage disequilibrium (LD) can impact effect estimates. This gene was chosen as an example where the effect estimates in the "top eQTL" were opposite in sign in brain compared to non-brain tissues, and where further investigation suggested that this difference in effect directions could be explained by multiple eQTLs in LD. In this example, we define SNP1 and SNP2 as the SNPs that show the strongest eQTL associations 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, $Y = \mu + \widehat{B_i} g_i$ where *i* in {1, 2} indexes the two SNPs. The bottom panels show effects from a multiple (2-SNP) regression model in each tissue, $Y = \mu + \widehat{B_1} g_1 + \widehat{B_2} g_2$. Each colored dot shows the effect estimate for a single tissue (color-coded as in Fig. 3), with grey bars indicating ±2 standard errors. For all estimates, *n* = 83–430 individuals, depending on the tissue (see Supplementary Fig. 7). The simple regression estimates show opposite-direction effects in brain versus 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, as SNP1 and SNP2 show a significant effect that excludes zero in most tissues. Furthermore, for both SNP1 and SNP2 the multiple regression effect estimates are consistent in sign across all tissues.



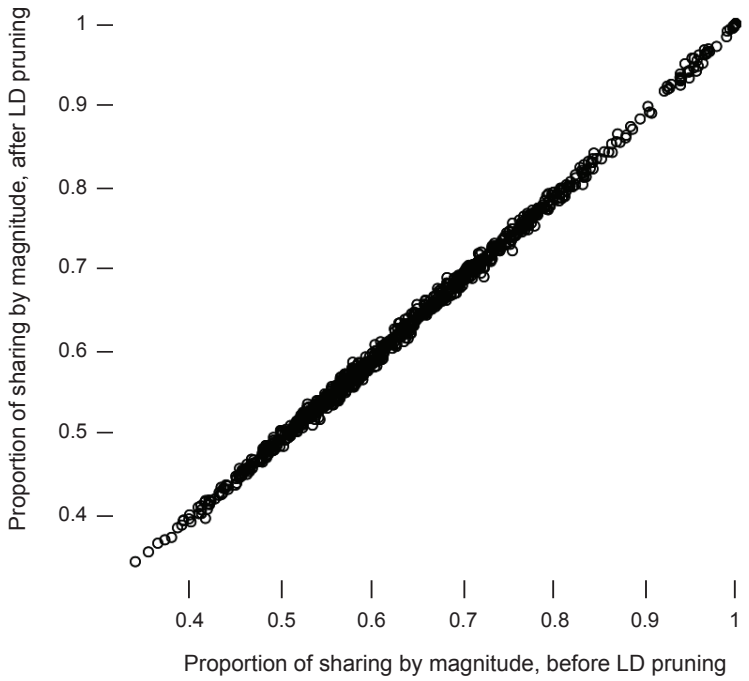
Supplementary Figure 6 | Pairwise sharing by sign. For each pair of tissues, we considered the top eQTLs that were significant (*lfsr* < 0.05) in at least one of the tissues, and estimated the proportion that have effect sizes with the same sign. These proportions are shown in this heatmap. *n* = 5,605–9,811 gene-SNP pairs, depending on pair of tissues compared.



Supplementary Figure 7 | Sample sizes and effective sample sizes from `mash` analysis across tissues. Sample size (**a**) and median effective sample size (ESS) for each tissue (**b**). 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 8 | Expression levels in genes with "tissue-specific eQTLs" are similar to those in other genes. The plots compare the densities (**a**) and cumulative distribution functions (**b**) of the average expression level for genes identified as having a "tissue-specific" eQTL (red), and remaining genes (black), separately in four tissues—testis, thyroid, whole blood and transformed fibroblasts. In each case, the distribution functions are reasonably similar, showing that tissue-specific eQTLs mostly do not reflect tissue-specific expression. Expression is defined as the median of log-Reads per Kilobase Mapped (log-RPKM) across individuals. Densities for genes having tissue-specific eQTLs (red) are estimated using average expression levels from n = 201–301 genes, depending on the tissue, and densities for remaining genes (black) are based on n ≥ 15,768 average gene expression levels.



Supplementary Figure 9 | Comparison of pairwise sharing by magnitude for top eQTLs, without and with LD pruning. Each point corresponds to a pair of tissues. The horizontal axis gives results from the original `mash` analysis reported in the main paper; the vertical axis shows results from an "LD-pruned'" analysis, where training data and top eQTLs were first pruned (using PLINK) to avoid any pair of SNPs being in LD ($r^2 > 0.2$) before `mash` was applied. The strong similarity of the results illustrates the robustness of `mash` to LD pruning.

Supplementary Tables

method	simulation framework	$RRMSE^{All}$	$RRMSE^{Non\text{-}null}$	$RRMSE^{Null}$
`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

Supplementary Table 1 | Accuracy of effect size estimates for each method. Table shows relative root mean squared error (RRMSE) for all effects ($RRMSE^{All}$), for the subsets of effects that are truly non-null ($\beta \neq 0$; $RRMSE^{Non\text{-}null}$) and truly null ($\beta \neq 0$, $RRMSE^{Null}$). RRMSE values less than 1 indicate improvements in accuracy over the original, observed estimates. Values of $RRMSE^{Null} < 1$ indicate that shrinkage toward zero helps improve estimates of null effects. Values of $RRMSE^{Non\text{-}null} < 1$ indicate that pooling information across conditions can improve accuracy of estimates of non-null effects. Note that, in the "independent" simulations, most effects are null, so the shrinkage of all methods improved overall performance compared to no shrinkage ($RRMSE^{All} < 1$), at the expense of lowering accuracy in the non-null effects ($RRMSE^{Non\text{-}null} > 1$). $RRMSE^{Non\text{-}null}$, $RRMSE^{Null}$ and $RRMSE^{All}$ values were calculated from n = 400, 19,600 and 20,000 and observed effects, respectively, in 44 simulated tissues.

	—simulation framework—		
associations	**shared, structured**	**shared, unstructured**	**independent**
`mash`, not `ash`, not `mash-bmalite`	3,889	622	32
`ash`, not `mash`, not `mash-bmalite`	0	0	740
`mash-bmalite`, not `mash`, not `ash`	37	9	79
`mash`, `ash`, not `mash-bmalite`	7	0	44
`mash`, `mash-bmalite`, not `ash`	5,777	336	70
`ash`, `mash-bmalite`, not `mash`	0	0	10
all	3,477	2	5,962

Supplementary Table 2 | Overlap in associations identified from simulated data sets. Table summarizes the overlap in significant associations (*lfsr* < 0.05) identified among all methods compared. In both "shared effects" scenarios, `mash` captures the vast majority of the associations identified by the other methods. All association counts in the table are a subset of n = 20,000 × 44 = 880,000 simulated gene-SNP effects (most of which are zero).

associations	count
`mash`, not `ash`, not `mash-bmalite`	63,956
`ash`, not `mash`, not `mash-bmalite`	2,383
`mash-bmalite`, not `mash`, not `ash`	11,789
`mash`, `ash`, not `mash-bmalite`	665
`mash`, `mash-bmalite`, not `ash`	176,572
`ash`, `mash-bmalite`, not `mash`	248
all	88,459

Supplementary Table 3 | Overlap in associations identified from GTEx data. Table summarizes the overlap in significant associations (*lfsr* < 0.05) identified among all methods compared. The `mash` method captures the vast majority of the associations identified by the other methods—only 248 associations identified by `ash` or `mash-bmalite` are not identified by `mash`—in addition to many other associations that are not identified by either `ash` or `mash-bmalite` (63,956). All association counts in the table are a subset of n = 16,069 × 44 = 707,036 gene-SNP effects considered.