nature genetics

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41588-018-0268-8

In the format provided by the authors and unedited.

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

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

[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]Research Computing Center, University of Chicago, Chicago, IL, USA. [4]Department of Statistics, University of Chicago, Chicago, IL, USA. *e-mail: mstephens@uchicago.edu

Supplementary Note

Multivariate adaptive shrinkage (`mash`), and its relationship with existing methods. An important novel feature of our method, `mash`, is its focus on *estimation of effect sizes*, in contrast to 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 is more flexible than models underlying existing methods—and, indeed, includes existing models as special cases.

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[1]. More flexible still are models that allow effects to be exactly zero in subsets of conditions, as in Flutre *et al.*[2] and Li *et al.*[3]. 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 $\boldsymbol{\omega}$, which allows `mash` to flexibly capture a range of effect distributions[4]; second, and perhaps more importantly, `mash` includes data-driven covariance matrices, making it more flexible and adaptive to patterns in the data. 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.*[2] and Li *et al.*[3] consider all 2^R combinations of sparsity in the effects, which is feasible for $R = 9$ (see Flutre *et al.*[2]), but impractical at $R = 44$. While it is possible to restrict the number of combinations considered (e.g., BMAlite[2]), 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.

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*)[4], which is the probability that the effet is estimated with the incorrect sign. The *lfsr* is analogous to the local false discovery rate[6], 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.

Comparison with `metasoft`. Among existing software packages for this problem, `metasoft`[7] 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 analysis, 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[8]; and the BE model is an extension of the RE2 model allowing that some effects are exactly zero[7]. These models are similar to the BMAlite models from Flutre *et al.*[2], and none capture the kinds of structured effects that can be learned from the data by `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`. For each model (FE, RE2, BE), `metasoft` produces a *p* value for each multivariate test, whereas `mash` and `mash-bmalite` produced a Bayes Factor (see Online Methods); in each case, these can be used to rank the significance of the tests.

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.*[2] 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.*[2] and Li *et al.*[3] 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, it can also provide more accurate estimates of sharing.

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 to be 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[9], which can perform well for computing joint 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*[10,11]. This issue more likely has an impact because of the presence of multiple eQTLs in some or many genes. It also applies to all single-SNP eQTL analyses, which are 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. 4; this eQTL has strong, positive *Z* scores in brain tissues, and negative *Z* scores in most other tissues. Initially, this suggested 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. 4). A similar example is discussed in the GTEx pilot study[5] (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 urge careful interpretation of 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, improving 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 computed an "effective sample size'" (ESS) for each tissue based on the standard deviations of the `mash` estimates. The ESS estimates (Supplementary Fig. 6) 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 prioritize "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 strong 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 benefit most from collecting more samples in the tissue of interest.

Supplementary References

1. 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).

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

3. 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).

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

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

6. Efron, B. Microarrays, empirical Bayes and the two-groups model. *Statistical Science* **23**, 1–22 (2008).

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

8. 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** (2010).

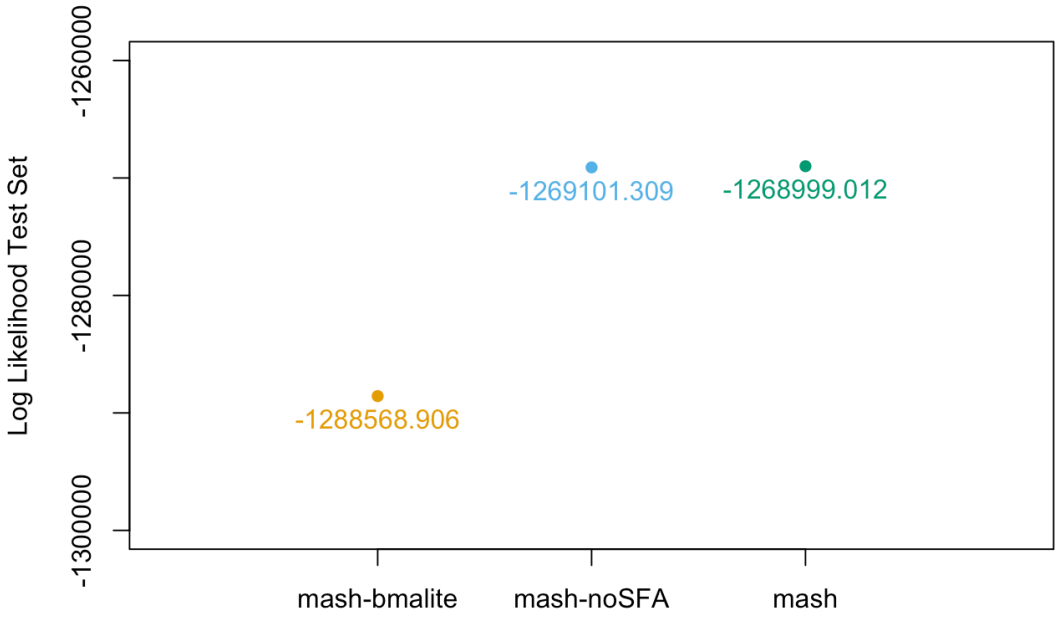
9. Larribe, F. & Fearnhead, P. Composite likelihood methods in statistical genetics. *Statistica Sinica* **21**, 43–69 (2011).

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

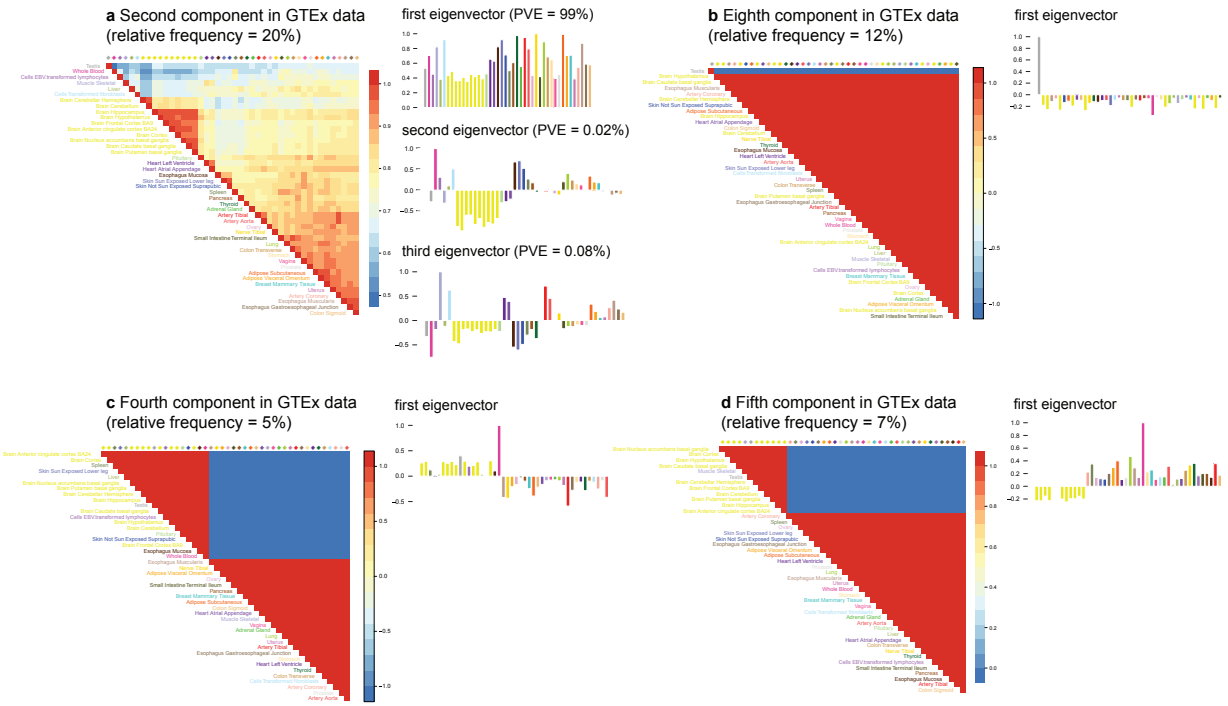
11. Zhu, X. & Stephens, M. Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. *Annals of Applied Statistics* **11**, 1561–1592 (2017).

12. 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).

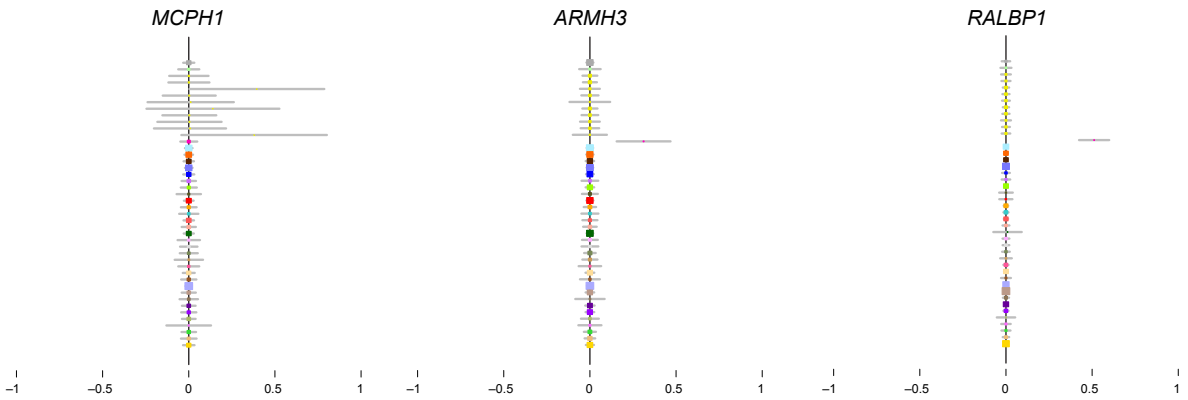
Supplementary Figures



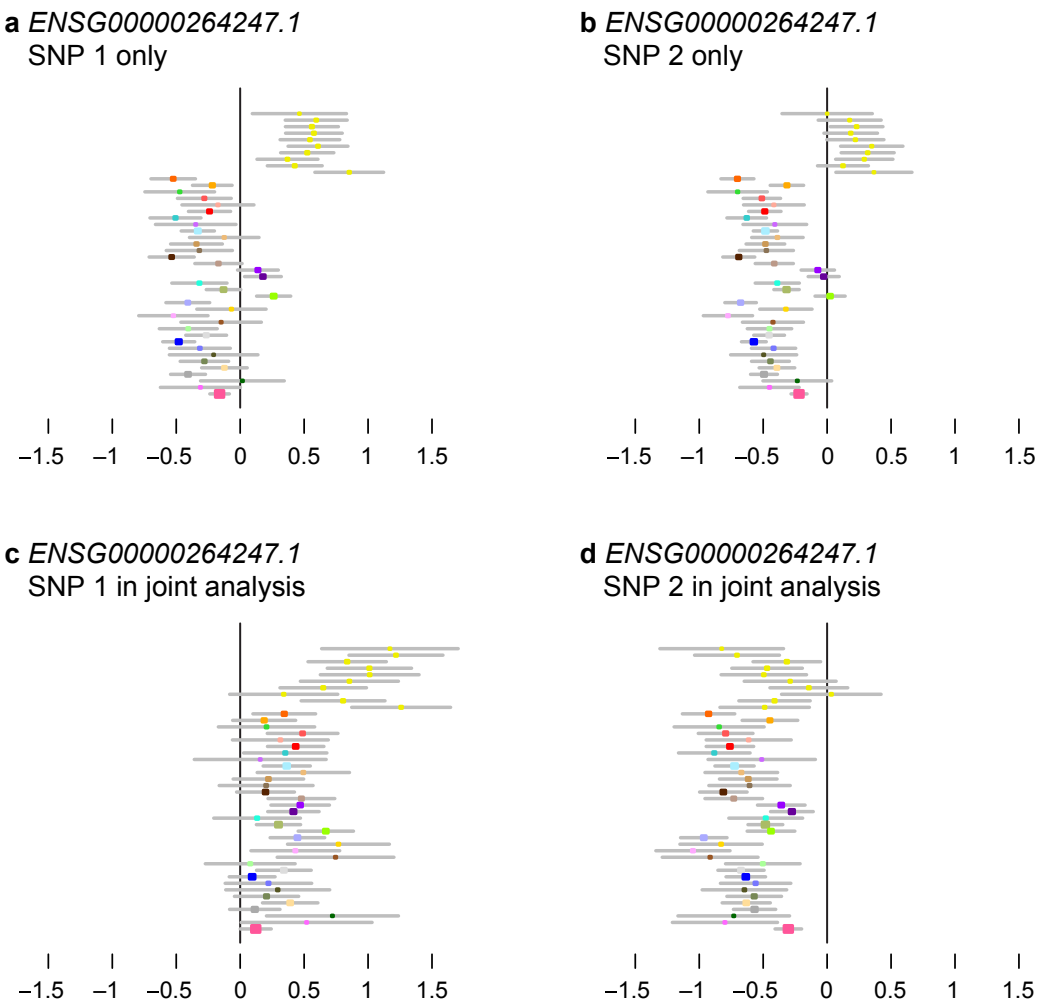
Supplementary Figure 1 | Increase in log-likelihood on test set as new U_k are added. The figure shows the log-likelihood on the test set for different "models" (choices of U_k). From left to right, the models are: `mash-bmalite` (no data-driven U_k); `mash-no-SFA` (the combination of canonical and data-driven 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.



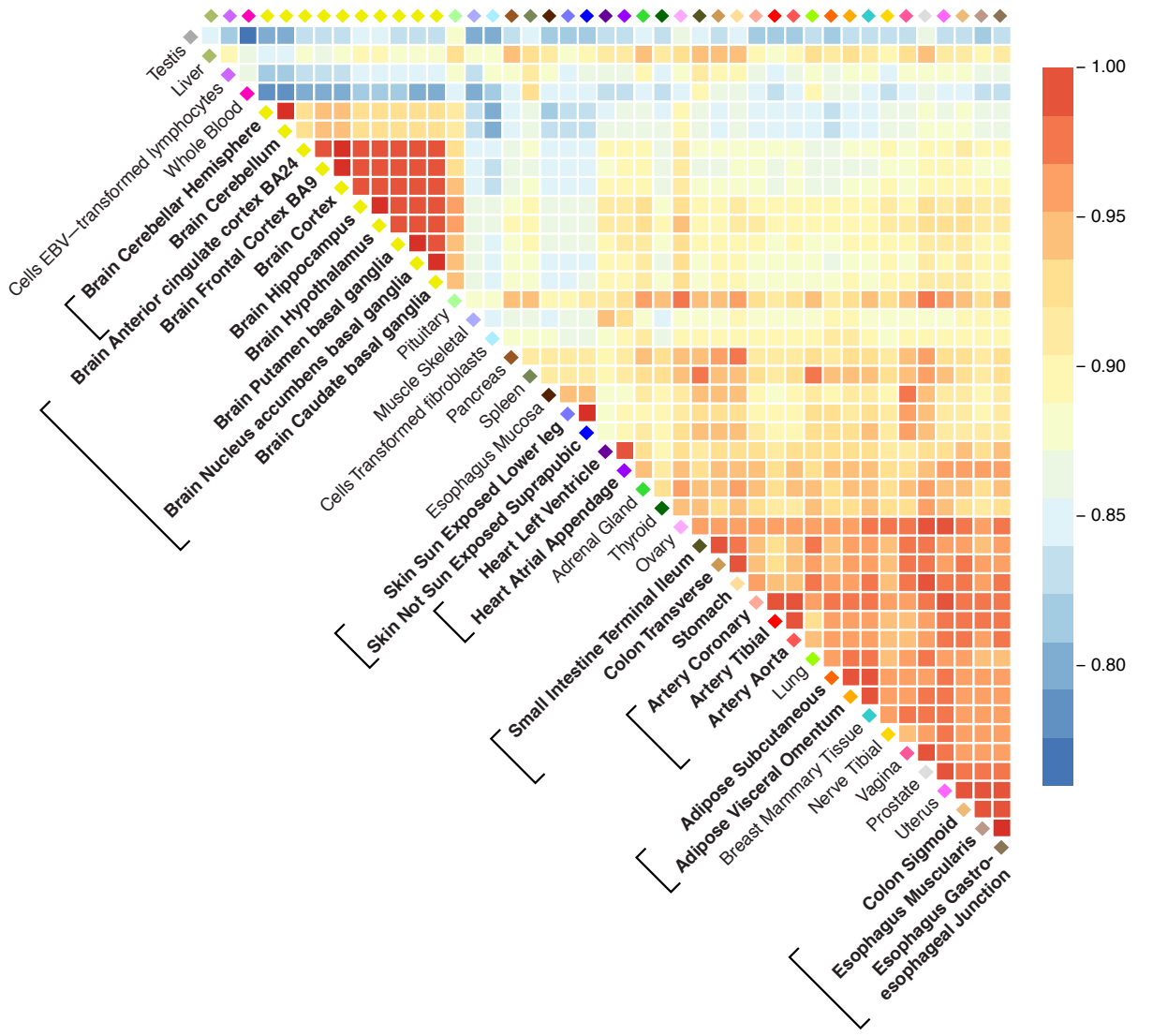
Supplementary Figure 2 | Summary of covariance matrices U_k with the largest estimated weights (>1%) in the GTEx data. For each covariance matrix U_k, the figure shows the heatmap of the corresponding correlation matrix, and bar plots of the top eigenvectors of U_k (n = 16,069 independent gene-SNP pairs). Component 2 (**a**) captures qualitatively similar effects to the component shown 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.



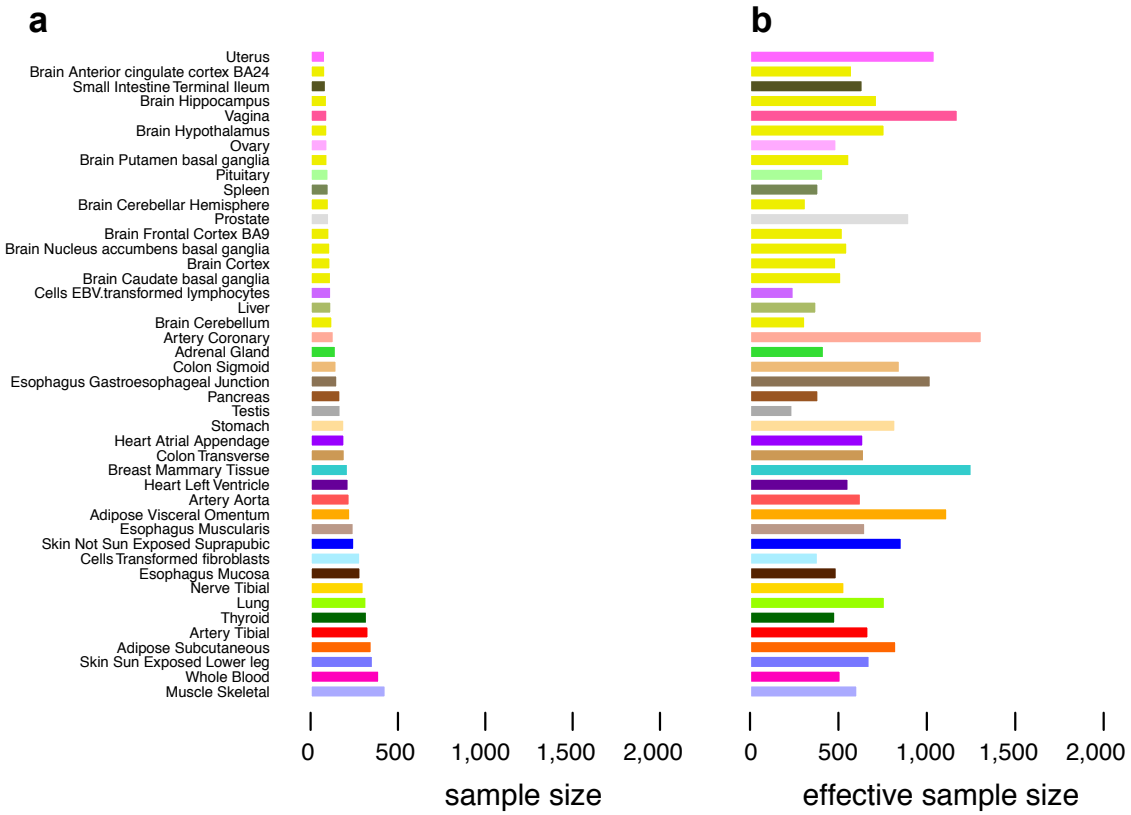
Supplementary Figure 3 | Estimates from the univariate method `ash` for the examples shown in Fig. 4. Each dot (color-coded as in Fig. 3) shows the effect estimate (posterior mean) from `ash`, with horizontal gray bars indicating ±2 posterior standard deviations. For all estimates, *n* = 83–430 individuals,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



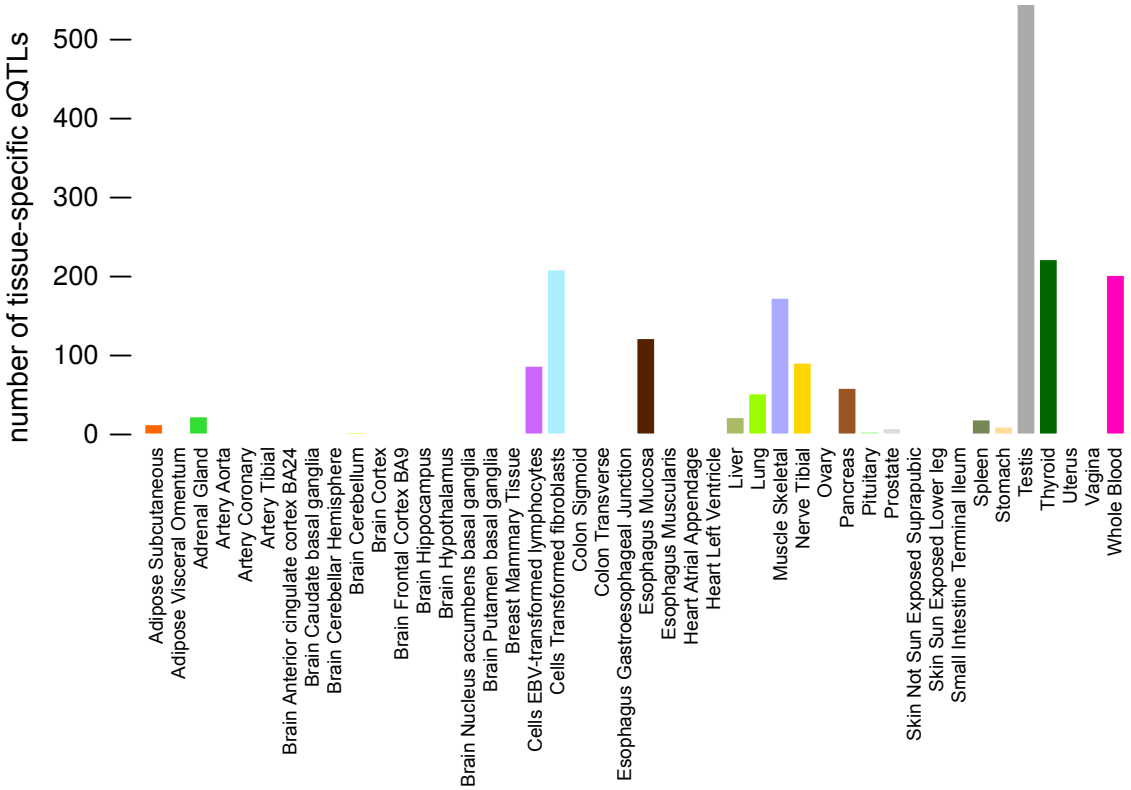
Supplementary Figure 4 | Illustration of how linkage disequilibrium (LD) can impact effect estimates. This gene was chosen as an example where the effect estimates in the "top eQTL" were opposite in sign in brain compared to non-brain tissues, and where further investigation suggested that this difference 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, $Y = \mu + \hat{B}_i g_i$ where *i* in {1, 2} indexes the two SNPs. The bottom panels show effects from a multiple (2-SNP) regression model in each tissue, $Y = \mu + \hat{B}_1 g_1 + \hat{B}_2 g_2$. Each 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 (Supplementary Table 3). The simple regression estimates (**a**, **b**) 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 (**c**, **d**) 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.



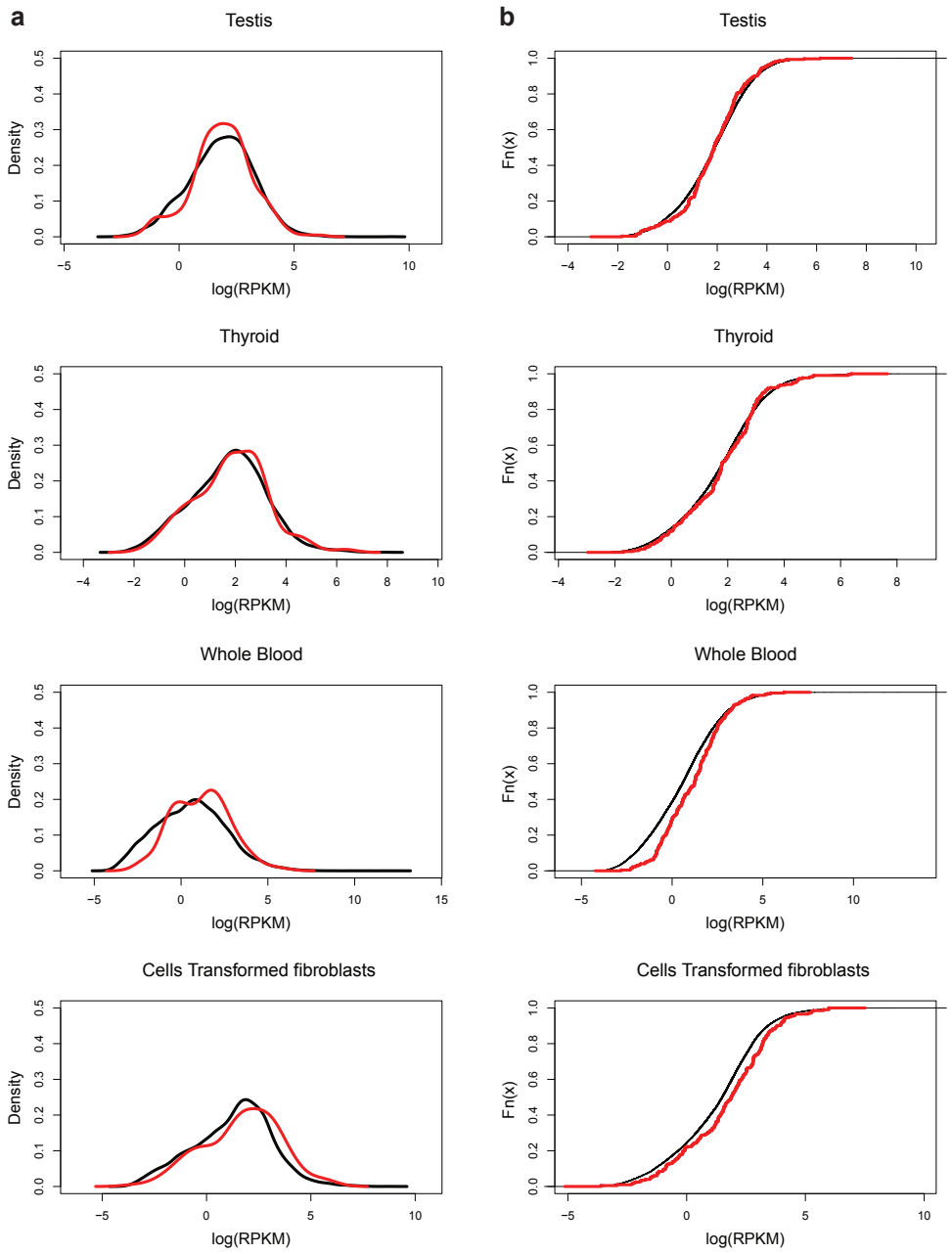
Supplementary Figure 5 | Pairwise sharing by sign. For each pair of tissues, we considered the top eQTLs that were significant (*lfsr* < 0.05) in at least one of the tissues, and calculated the proportion that have effect sizes with the same sign. These proportions are shown in this heatmap. *n* = 5,605–9,811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



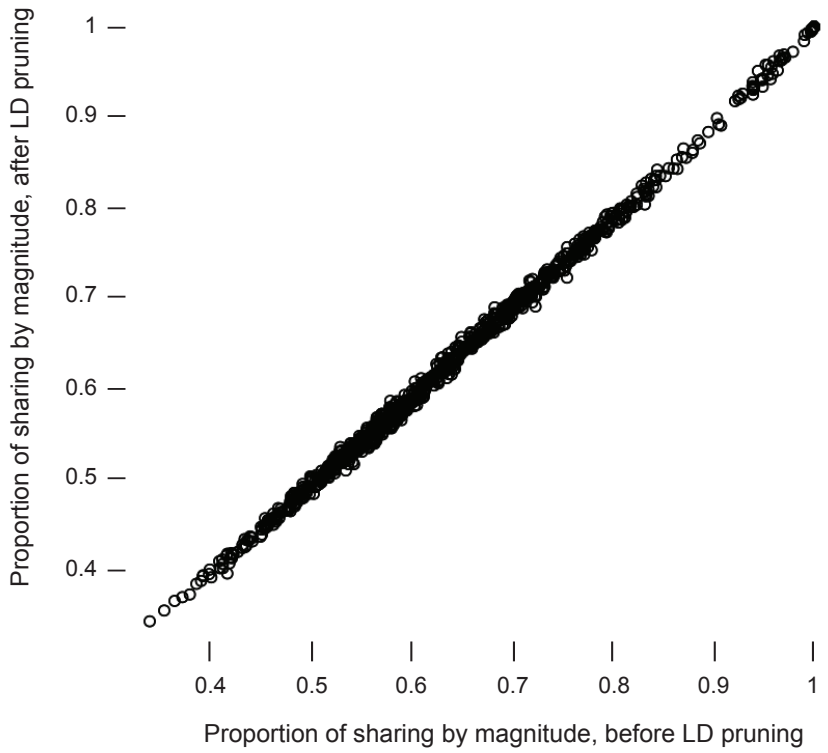
Supplementary Figure 6 | Sample sizes and effective sample sizes from `mash` analysis across tissues. Sample size (**a**) and median "effective sample size" (ESS) for each tissue (**b**). Tissues are ordered by their (original) sample size (Supplementary Table 3). Effective sample sizes are consistently higher than actual sample sizes, primarily due to sharing of information among tissues.



Supplementary Figure 7 | 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 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 at least 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[12]) 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 estimates. Values of RRMSE^{Null} < 1 indicate that shrinkage toward zero helped 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 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 that were compared. In both "shared effects" scenarios, mash captured 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).

Tissue	sample size
adipose visceral omentum	227
adrenal gland	145
artery aorta	224
artery coronary	133
artery tibial	332
brain anterior cingulate cortex BA24	84
brain caudate basal ganglia	117
brain cerebellar hemisphere	105
brain cerebellum	125
brain cortex	114
brain frontal cortex BA9	108
brain hippocampus	94
brain hypothalamus	96
brain nucleus accumbens basal ganglia	113
brain putamen basal ganglia	97
breast mammary tissue	214
cells EBV-transformed lymphocytes	118
cells transformed fibroblasts	284
colon sigmoid	149
colon transverse	196
esophagus gastroesophageal junction	153
esophagus mucosa	286
esophagus muscularis	247
heart atrial appendage	194
heart left ventricle	218
liver	119
lung	320
muscle skeletal	430
nerve tibial	304
ovary	97
pancreas	171
pituitary	103
prostate	106
skin not sun exposed suprapubic	250
skin sun exposed lower leg	357
small intestine terminal ileum	88
Spleen	104
Stomach	193
Testis	172
thyroid	323
uterus	83
vagina	96
whole blood	393

Supplementary Table 3 | Tissue sample sizes. Right-hand column gives the sample size (*n*) for each tissue in the GTEx data set.

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 4 | 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 the $n = 16{,}069 \times 44 = 707{,}036$ gene-SNP effects considered.