A flexible empirical Bayes approach to multivariate multiple regression, and its improved accuracy in predicting multi-tissue gene expression from genotypes

Fabio Morgante[1,2,3*], Peter Carbonetto[4,5], Gao Wang[4,6,7], Yuxin Zou[8,9], Abhishek Sarkar[4,¤], Matthew Stephens[4,8*]

1 Center for Human Genetics, Clemson University, Greenwood, SC, USA
2 Department of Genetics and Biochemistry, Clemson University, Clemson, SC, USA
3 Section of Genetic Medicine, Department of Medicine, University of Chicago, Chicago, IL, USA
4 Department of Human Genetics, University of Chicago, Chicago, IL, USA
5 Research Computing Center, University of Chicago, Chicago, IL, USA
6 Department of Neurology, Columbia University, New York, NY, USA
7 Gertrude. H. Sergievsky Center, Columbia University, New York, NY, USA
8 Department of Statistics, University of Chicago, Chicago, IL, USA
9 Regeneron Genetics Center, Regeneron Pharmaceuticals Inc., Tarrytown, NY, USA
¤Current address: Vesalius Therapeutics, Cambridge, MA, USA

* fabiom@clemson.edu, mstephens@uchicago.edu

Abstract

Predicting phenotypes from genotypes is a fundamental task in quantitative genetics. With technological advances, it is now possible to measure multiple phenotypes in large samples. Multiple phenotypes can share their genetic component; therefore, modeling these phenotypes jointly may improve prediction accuracy by leveraging *effects that are shared across phenotypes.* However, effects can be shared across phenotypes in a variety of ways, so computationally efficient statistical methods are needed that can accurately and flexibly capture patterns of effect sharing. Here, we describe new Bayesian multivariate, multiple regression methods that, by using flexible priors, are able to model and adapt to different patterns of effect sharing and specificity across phenotypes. Simulation results show that these new methods are fast and improve prediction accuracy compared with existing methods in a wide range of settings where effects are shared. Further, in settings where effects are not shared our methods still perform competitively with state-of-the-art methods. In real data analyses of expression data in the Genotype Tissue Expression (GTEx) project our methods improve prediction performance on average for all tissues, with the greatest gains in tissues where effects are strongly shared, and in the tissues with smaller sample sizes. While we use gene expression prediction to illustrate our methods, the methods are generally applicable to any multi-phenotype applications, including prediction of polygenic scores and breeding values. Thus, our methods have the potential to provide improvements across fields and organisms.

Author summary

Predicting phenotypes from genotypes is a fundamental problem in quantitative genetics. Thanks to recent advances, it is increasingly feasible to collect data on many

phenotypes and genome-wide genotypes in large samples. Here, we tackle the problem of predicting *multiple phenotypes* from genotypes using a new method based on a multivariate, multiple linear regression model. Although the use of a multivariate, multiple linear regression model is not new, in this paper we introduce a flexible and computationally efficient *empirical Bayes* approach based on this model. This approach uses a prior that captures how the effects of genotypes on phenotypes are shared across the different phenotypes, and then the prior is adapted to the data in order to capture the most prominent sharing patterns present in the data. We assess the benefits of this flexible Bayesian approach in simulated genetic data sets, and we illustrate its application in predicting gene expression measured in multiple human tissues. We show that our methods can outperform competing methods in terms of prediction accuracy, and the computational effort involved in fitting the model and making the predictions scales well to large data sets.

Introduction

Multiple regression has been an important tool in genetics for different tasks relating genotypes and phenotypes, including discovery, inference, and prediction. For discovery, multiple regression has been used to fine-map genetic variants discovered by Genome-Wide Association Study (GWAS) [1, 2]. For inference, multiple regression has been used to estimate the proportion of phenotypic variance explained by genetic variants—*i.e.*, "genomic heritability" or "SNP heritability"— [3–5]. For prediction, multiple regression has been used extensively to predict yet-to-be-observed phenotypes from genotypes. This task is relevant to the prediction of breeding values for selection purposes in agriculture [6, 7], the prediction of "polygenic scores" for disease risk and medically relevant phenotypes in human genetics [8, 9], and the prediction of gene expression as an intermediate step in transcriptome-wide association studies (TWAS) [10, 11]. Traditionally, frequentist multiple regression methods such as penalized regression and linear mixed models [12–15] have been used for these tasks. However, Bayesian methods have received particular attention in genetic applications because they provide a natural way to incorporate prior information about and cope with different genetic architectures. This attractive feature has spurred the development and application of many Bayesian methods that differ in their prior distribution on the effect sizes and their approach to computing posterior distributions [6, 16–26].

Most multiple regression methods in widespread use are "univariate" in that they model one outcome (phenotype) at a time. However, many studies involve multiple outcomes that may share genetic effects [27]. Examples of this include organismal-level phenotypes measured in multiple environments or populations, such as those available in UK Biobank [28] or BioBank Japan [29], and multiple molecular phenotypes such as the expression levels of multiple genes in multiple tissues, available in reference data sets like the Genotype Tissue Expression (GTEx) Project [30]. In such cases, joint ("multivariate") modeling of multiple phenotypes can improve performance over separate univariate analyses that consider one phenotype at a time. Indeed, multivariate analysis can improve performance even when phenotypes are not genetically correlated, provided that phenotypes are phenotypically correlated [31]. Multivariate analysis of multiple phenotypes has been shown to improve power to discover associations [32–35] and accuracy of phenotype prediction [36–39].

However, currently available multivariate multiple regression methods have important limitations. The multivariate versions of popular penalized regression methods (e.g., ridge regression, the Elastic Net, the Lasso implemented in the popular R package `glmnet` [40]) do not allow for missing phenotype values and, more importantly, do not exploit patterns of effect sharing. Urbut *et al* [34] showed the benefits of

multivariate methods that learn effect sharing from the data. Multivariate linear mixed models (MLMM) [41] can also learn effect sharing from the data, but they lack flexibility—these models make the "infinitesimal architecture" assumption that every variant has an effect on all phenotypes which is not appropriate for phenotypes with sparse architectures [42]. Bayesian methods are a natural way to achieve flexibility in terms of sparsity of the signal and can learn patterns of effect sharing from the data. These methods include multivariate versions of of the "Bayesian alphabet" methods such as BayesB, BayesCΠ, and the Bayesian Lasso [43, 44]. However, despite the added flexibility compared to the MLMM model, the prior families used in existing multivariate Bayesian methods make them relatively inflexible to cope with the complex distribution of effect sizes that many complex traits have. In fact, most of those methods either have a single distribution or a spike-and-slab type of prior, with only one non-point mass component. In addition, the use of computationally intensive Markov Chain Monte Carlo (MCMC) algorithms for model fitting makes the multivariate Bayesian alphabet methods impractical in many genome-scale settings, even with a moderate number of phenotypes.

To overcome these limitations, we introduce a new method, "Multiple Regression with Multivariate Adaptive Shrinkage" or "*mr.mash*". *mr.mash* is a Bayesian multivariate, multiple regression method that is able to learn complex patterns of effect sharing from the data, while also being computationally efficient. We achieve this by combining three powerful ideas: (1) flexible prior distributions that allow for complex patterns of effect sharing across phenotypes; (2) empirical Bayes for adapting the priors to the data; and (3) variational inference for fast Bayesian computations. In practice, this work aims to integrate previous work by Urbut *et al* [34] (ideas 1 and 2) with previous work by Carbonetto and Stephens [19] (idea 3) into a single framework, and extends the methods of Kim *et al* [26] to the multivariate setting. We show via extensive simulations of multi-tissue gene expression prediction from genotypes that *mr.mash* can adapt to complex patterns of effect sharing and specificity and outperforms competing methods. These results are confirmed in analyses of real data from the Genotype Tissue Expression (GTEx) project [30], demonstrating the potential for our method to more accurately impute expression levels, as is required for TWAS [10, 11]. Although this work was primarily motivated by our interest in improving predictions of gene expression, *mr.mash* can be applied to any setting where predictions from multivariate multiple regression are desired, such as computing polygenic scores or breeding values.

Description of the method

We consider the multivariate multiple regression model of outcomes $\boldsymbol{Y}$ on predictors $\boldsymbol{X}$,

$$\begin{aligned} \boldsymbol{Y} &= \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E} \\ \boldsymbol{E} &\sim MN_{n\times r}(\boldsymbol{0}, \boldsymbol{I}_n, \boldsymbol{V}), \end{aligned} \tag{1}$$

where $\boldsymbol{Y}$ is an $n \times r$ matrix of r outcomes observed in n samples (possibly containing missing values), $\boldsymbol{X}$ is an $n \times p$ matrix of p predictors observed in the same n samples, $\boldsymbol{B}$ is the $p \times r$ matrix of effects, $\boldsymbol{E}$ is an $n \times r$ matrix of residuals, and $MN_{n\times r}(\boldsymbol{M}, \boldsymbol{U}, \boldsymbol{V})$ is the matrix normal distribution with mean $\boldsymbol{M} \in \mathrm{R}^{n\times r}$ and covariance matrices $\boldsymbol{U} \in \mathbb{R}^{n\times n}$, $\boldsymbol{V} \in \mathbb{R}^{r\times r}$ [45, 46]. For example, in our application later we aim to predict gene expression in multiple tissues from genetic variant genotypes, so y_{is} is the observed gene expression in individual i and tissue s, and x_{ij} is the genotype of individual i at genetic variant j.

Let $\boldsymbol{b}_j$ denote the jth row of $\boldsymbol{B}$; thus, $\boldsymbol{b}_j$ is an r-vector reflecting the effects of variable j on the r outcomes. To capture the potential similarity of the effects among

the different outcomes, we use a mixture of multivariate normals prior on $\boldsymbol{b}_j$:

$$\boldsymbol{b}_j \mid \boldsymbol{w}_0, \mathscr{S}_0 \sim \sum_{k=1}^{K} w_{0,k} N_r(\boldsymbol{0}, \boldsymbol{S}_{0,k}), \tag{2}$$

where $N_r(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the multivariate normal distribution on $\mathbb{R}^r$ with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, $\boldsymbol{w}_0 := (w_{0,1}, \ldots, w_{0,K})$ is a set of mixture weights (non-negative and summing to one), and $\mathscr{S}_0 := \{\boldsymbol{S}_{0,1}, \ldots, \boldsymbol{S}_{0,K}\}$ denotes a collection of $r \times r$ covariance matrices. Following [34], we assume that the covariance matrices $\mathscr{S}_0$ are pre-specified, and treat the mixture weights $\boldsymbol{w}_0$ as parameters to be estimated from the data. The idea is that the collection of matrices $\mathscr{S}_0$ should be chosen to include a wide variety of potential effect sharing patterns; the estimated $\boldsymbol{w}_0$ should then assign most weight to sharing patterns that are present in the data, and little or no weight on patterns that are inconsistent with the data. We discuss selection of suitable covariance matrices $\mathscr{S}_0$ in S1 Text.

Since our approach combines the multiple regression model (1) with multivariate adaptive shrinkage priors from (2), we call our approach "*mr.mash*", which is short for "*Multiple Regression* with *Multivariate Adaptive Shrinkage*".

Variational empirical Bayes for *mr.mash*

To fit the *mr.mash* model we use variational inference methods [47, 48] which have been successfully applied to fit univariate multiple regressions [19, 23, 24, 26, 49]. Variational inference recasts the posterior computation as an optimization problem. Specifically, we seek a distribution $q(\boldsymbol{B})$ which approximates the true posterior distribution, $p(\boldsymbol{B} \mid \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{V}, \boldsymbol{w}_0, \mathscr{S}_0)$. By imposing simple conditional independence assumptions on the approximate posterior distribution, $q(\boldsymbol{B})$, the posterior computations and optimization of $q(\boldsymbol{B})$ become tractable.

In addition to approximating the posterior distribution of $\boldsymbol{B}$, the variational approach also provides a way to estimate the model parameters, $\boldsymbol{w}_0$ and $\boldsymbol{V}$, by maximizing an approximation to the marginal likelihood, $p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{V}, \boldsymbol{w}_0, \mathscr{S}_0)$, which is known as the "evidence lower bound" (ELBO) [47]. This approach was called "variational empirical Bayes" in [50], although this idea of fitting the model parameters by maximizing an approximate marginal likelihood dates back to earlier work [51, 52].

A key output of *mr.mash* is the (approximate) posterior mean of the regression coefficients, $\hat{\boldsymbol{B}} := E_q(\boldsymbol{B})$. This point estimate can be used to predict unobserved outcomes for new samples from their predictor values. Specifically, given predictor values stored as an $n^{\text{new}} \times p$ matrix $\boldsymbol{X}^{\text{new}}$, we can predict the outcomes as

$$\hat{\boldsymbol{Y}}^{\text{new}} := \boldsymbol{X}^{\text{new}} \hat{\boldsymbol{B}}. \tag{3}$$

The variational empirical Bayes approach accomplishes the twin goals of (a) computing posterior effect estimates and (b) adapting the priors to the data, while making the underlying computations fast and scalable to large data sets, especially compared with alternative strategies like MCMC [19]. The trade-off is that the approximate posterior distribution obtained with our variational methods will tend to understate uncertainty compared with the true posterior distribution [47], and so its use for inference (as opposed to prediction) requires particular care [19]. In this regard, one might consider *mr.mash* more directly comparable to penalized regression methods like the *Elastic Net*, which are also more naturally applied to prediction than inference.

Handling missing data

When analyzing multivariate data, it is commmon for a large fraction of the $\boldsymbol{Y}$ values to be unavailable, or "missing." For example, in the GTEx expression data [30] (see Applications), the average missing rate is about 60% (after removing a few tissues that are mostly missing). Thus, for broad applicability, it is important for multivariate methods to be able to cope with missing values.

To deal with missing values we extend the variational approximation to approximate the posterior distribution of the missing entries. Computationally, this extension ends up adding an additional step to the iterative algorithm that *imputes* the missing values. Specifically, denoting $\boldsymbol{Y}_{\text{obs}}$ as the set of observed expression levels and $\boldsymbol{Y}_{\text{miss}}$ as the set of unobserved (missing) expression levels, the approach imputes the missing values $\boldsymbol{Y}_{\text{miss}}$ by computing an approximate posterior distribution for $\boldsymbol{Y}_{\text{miss}}$ given $\boldsymbol{Y}_{\text{obs}}$, the effects $\boldsymbol{B}$, and the residual covariance $\boldsymbol{V}$. A similar approach was taken in [53].

Software availability

The methods introduced in this paper are implemented as an R package [54] which is available for download at https://github.com/stephenslab/mr.mash.alpha.

Verification and comparison

Simulations using GTEx genotypes

We compared, *mr.mash* and other methods based on multiple regression (1), in the task of predicting gene expression in multiple tissues from genetic variant genotypes. To perform systematic evaluations of the methods in realistic settings, we simulated multi-tissue gene expression data using genotypes from the GTEx project [30]. Specifically, we used the 838 genotype samples generated by whole-genome sequencing. (The GTEx project also collected extensive gene expression data via RNA sequencing but we did not use these data in our simulations.)

We performed simulations under several scenarios; the scenarios differed in the way the effects of the causal variants were simulated. (We use "causal variant" as shorthand for "genetic variant having true non-zero effect in the linear regression for at least one tissue"; that is, $\boldsymbol{b}_j \neq \boldsymbol{0}$.)

First, we considered three simple scenarios that captured the extreme settings one might encounter in a multivariate analysis:

A. "Equal Effects," in which each causal variant affects all tissues, with *the same effect in every tissue*.

B. "Independent Effects," in which each causal variant affects all tissues, and *the effects are independent across tissues* (more precisely, the effects are independent *conditioned on the genetic variant being a causal variant*).

C. "Mostly null," in which each causal variant affects only the first tissue, and therefore the remaining tissues are unaffected by genotype. This represents a scenario in which the genetic effects on gene expression are *tissue-specific*.

In these scenarios, the cuasal variants explained 20% of the variance of each tissue. We also considered two more complex scenarios designed to capture a combination of factors that one might encounter in more realistic settings:

D. "Equal Effects + Null," in which the effects on tissues 1 through 3 were equal and explained 20% of the variance of each tissue, and there were no effects in tissues 4

through 10. This represents a scenario where effects are shared only within a subset of tissues.

E. "Shared Sffects in Subgroups," in which effects were drawn from a mixture of effect sharing patterns: half of the time, the effects were shared (unequally) across tissues 1 through 3 and explained 20% of the variance of each tissue; otherwise, the effects were shared (unequally) in tissues 4 through 10 and explained only 5% of the variance of each tissue. This scenario was intended to reflect the patterns of effect sharing estimated from previous multivariate analyses of GTEx Project data (see for example Fig. 3a in [34]).

In each Scenario A through E, we simulated 20 gene expression data sets for 20 randomly selected genes.

We summarized the accuracy of predicted expression levels in test set samples using the commonly used "root mean squared error" (RMSE) metric, defined as

$$\text{RMSE} := \sqrt{\frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} (y_i - \hat{y}_i)^2}, \tag{4}$$

where y_i is the true expression value in the ith test sample, $\hat{y}_i$ is the estimated expression value in the ith test sample, and n_{test} is the number of samples in the test set (which in these experiments was always 168). To make the RMSE more comparable across tissues with different variances we always standardized the RMSE by dividing it by the standard deviation of the true expression measurements in the test set.

See S1 Text for more details about the simulations.

Methods compared

We compared *mr.mash* with existing multivariate multiple regression methods: the *Group Lasso* [55] and the *Sparse Multi-task Lasso* [38, 56], both of which use *penalties* to stabilize and improve accuracy of the fitted models; and a univariate, penalty-based method, the *Elastic Net* [14], applied independently to each tissue. For the *Elastic Net*, the mixing parameter α was set to 0.5 as is common in gene expression prediction [10]. In the results, these three comparison methods are labelled "g-lasso","smt-lasso" and "e-net", respectively.

We also assessed the impact of the choice of prior covariance matrices on the performance of *mr.mash*. To do so, we compared three variants of *mr.mash*: (1) *mr.mash* with only "canonical" prior covariance matrices; (2) *mr.mash* with only "data-driven" prior covariance matrices; and (3) *mr.mash* with both types of prior covariance matrices. (See S1 Text for details on these matrices.) We expected that the third variant would adapt well to the widest range of scenarios, and therefore would be the most competitive method overall, with the disadvantage being that it would require more computational effort. However, we found that *mr.mash* with only data-driven matrices was competitive in terms of prediction accuracy in all the simulated scenarios and was also faster than the other two variants (S1 Text, S1 Fig, S2 Fig). Thus, in the comparisons with other methods, we ran *mr.mash* with the data-driven matrices only.

See S1 Text for more details on how the methods were applied to the simulated data sets.

Results with full data

We begin with the results on the simulations in the "Equal Effects," "Independent Effects" and "Mostly Null" scenarios. Although these scenarios are not the most

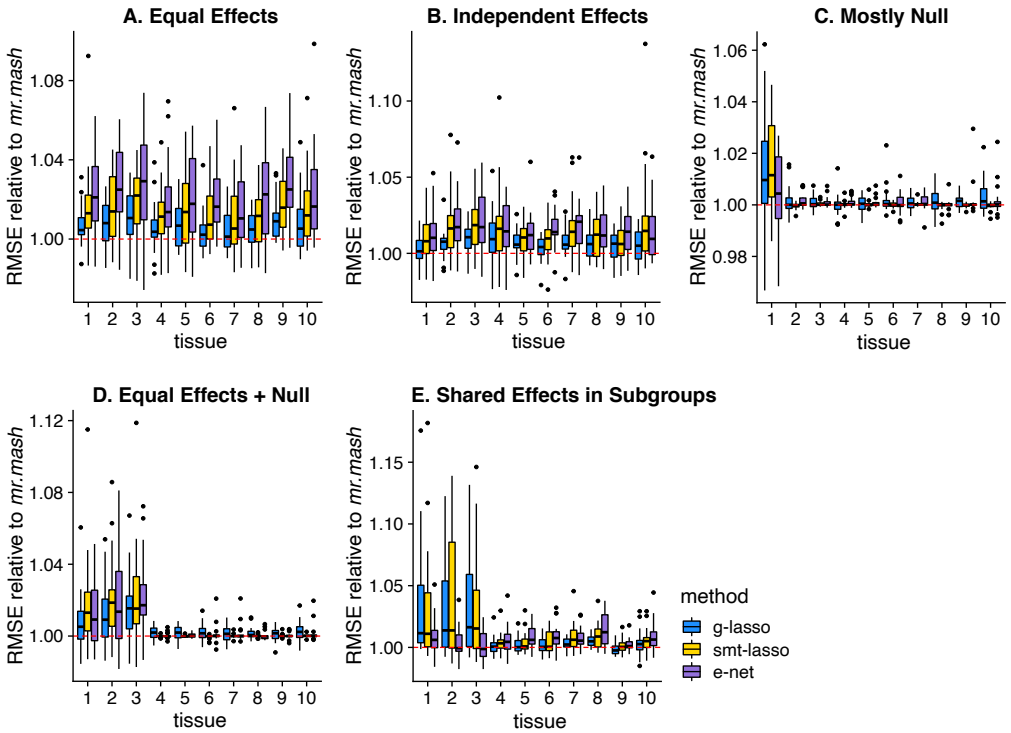

Fig 1. Prediction accuracy in simulations with full data. Each plot summarizes the accuracy of the test set predictions in 20 simulations. The thick, black line in each box gives the median RMSE relative to the *mr.mash* RMSE. Note that RMSE is a summary of *prediction error*, so lower values of RMSE indicate better prediction accuracy. Also note that the range of relative RMSE values shown in the y-axis is different in each panel.

realistic, they are simpler to understand, and help clarify the behaviour of different approaches.

In the Equal Effects scenario (Fig 1A), *mr.mash* substantially outperformed the other methods. Under the Equal Effects scenario, the effects of each causal variant were the same in all tissues, and among the methods compared here *mr.mash* is unique in its ability to adapt to this scenario; in particular, by adapting the prior to the data, *mr.mash* learned that most of the effects were shared equally or nearly equally across tissues. To illustrate, in one simulation *mr.mash* assigned 81% of the non-null prior weight to matrices capturing equal effects or very similar effects. By contrast, the penalty terms in the penalty-based methods are not able to capture this scenario. Unsurprisingly, the *Elastic Net* performed worst in this scenario because it implicitly assumes that the effects are independent, whereas in fact they are highly dependent. Also, *Group Lasso* performed substantially better than the *Sparse Multi-task Lasso* in this scenario; however, this may reflect differences in the way these methods were applied (see S1 Text), rather than a fundamental advantage of the *Group Lasso* over the *Sparse Multi-task Lasso*.

In the Independent Effects scenario (Fig 1B), performance was more similar among the methods. In this scenario there is less to be gained from multivariate regression methods because, once the causal variants are identified, knowing the effect size in one tissue does not help with estimating the effect size in another tissue. Nonetheless, multivariate methods do still have some benefits because they can more accurately identify the casual variants. Specifically, the effects for a given genetic variant are either

all zero or all nonzero, and all three multivariate methods we consider (*Group Lasso*, *Sparse Multi-task Lasso* and *mr.mash*) can take advantage of this situation. Consequently, the qualitative differences between methods are somewhat similar to the Equal Effects scenario, although the quantitative differences are smaller.

In the Mostly Null scenario (Fig 1C), there is much less benefit to multivariate methods because tissues 2–10 are uncorrelated with the genotypes. In fact, all the methods performed similarly in tissues 2–10. In tissue 1—the one tissue that is partly explained by genotype—the *Group Lasso* and *Sparse Multi-task Lasso* methods performed *worse* than the *Elastic Net*. For the *Group Lasso*, its penalty is poorly suited to this setting—the penalty effectively assumes that effects are either all zero or all nonzero—and because 9 out of the 10 tissues had no genetic effects, the *Group Lasso* penalty strongly encouraged the non-zero effects in tissue 1 toward zero. More surprisingly, the *Sparse Multi-task Lasso* also did not adapt to this scenario, despite having an additional penalty that in principle allows for sparsity across tissues. In contrast to the *Group Lasso* and *Sparse Multi-task Lasso*, *mr.mash*'s prior could adapt to this setting thanks to covariance matrices that allow for tissue-specific effects. Although the prediction accuracy of *mr.mash* in tissue 1 was essentially the same as *Elastic Net*'s, it is nonetheless reassuring that, in contrast to the other multivariate methods, *mr.mash* was no worse than *Elastic Net*.

We now describe the results from the two more complex scenarios, "Equal Effects + Null" and "Shared Effects in Subgroups."

The Equal Effects + Null scenario is a hybrid of the Equal Effects and Mostly Null scenarios, and so the results in Fig 1D reflect those in Panels A and C. As expected, all methods performed similiarly in tissues 4–10 (which were uncorrelated with the genotypes), whereas in tissues 1–3 the performance differences were similiar to those observed in the Equal Effects scenario, although smaller because here these effects were shared across fewer tissues. As in the Mostly Null scenario, the *Group Lasso* and *Sparse Multi-task Lasso* overshrank the effects in tissues 1–3, whereas *mr.mash* learned to shrink the effects in tissues 1–3 differently from the effects in tissues 4–10, with prior covariance matrices that allowed for strong correlations among tissues 1–3 only. For example, in one simulation *mr.mash* assigned 79% of the non-null prior weight to matrices capturing equal effects or very similar effects in tissues 1–3 and no effects or small effects in the remaining tissues.

The Shared Effects in Subgroups scenario (Fig 1E) is designed to be reflective of actual gene expression studies, and is therefore the most complex of the simulation scenarios we consider. Here all methods performed similarly in tissues 4–10, where the genetic effects explain only a small proportion of phenotypic variance (5%). In tissues 1–3 this scenario includes shared effects (explaining 20% of the phenotypic variance), but the sharing was not quite as strong as in the Equal Effects simulations. As a result, performance gains from conducting a multivariate analysis should be similar to, but not as strong as, the Equal Effects + Null scenario, and the results confirm this. The benefit of *mr.mash* over the *Elastic Net* is more modest in this more complex scenario, possibly reflecting the challenge of adapting *mr.mash*'s flexible prior to the complex patterns of effect sharing. Like the Mostly Null and Equal Effects + Null scenarios, the relatively inflexible penalty in the *Group Lasso* cannot capture the complex patterns of sharing, and this explains its inferior performance in tissues 1–3.

We also compared the computational time of the different methods (Fig 2). The runtime of *mr.mash* (with data-driven matrices only) was typically only slightly higher than *Elastic Net* or *Group Lasso*, usually within a factor of 2. Although the *Elastic Net* and *Group Lasso* solved a much simpler optimization problem, they required a more intensive cross-validation step to tune the strength of the penalty term; in contrast, the analogous step in *mr.mash* involved tuning the prior, and was achieved by an empirical

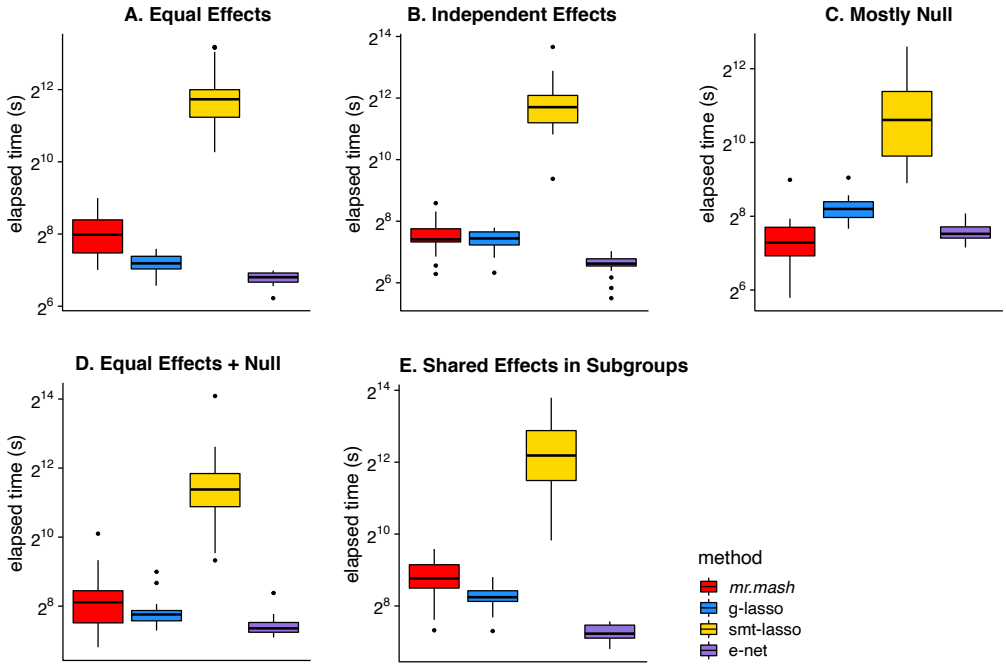

Fig 2. Runtimes in simulations with full data. Each plot summarizes the distribution of model fitting runtimes for the 20 simulations in that scenario. Note that the *mr.mash* runtimes did not include the initialization step, which was performed using the *Group Lasso*. Once the model fitting was completed, computing the predictions was very fast for all methods, so we did not include the prediction step in these runtimes. See S1 Text for details on the computing environment used to run the simulations. The thick, black line in each box gives the median runtime. Also note that the range of runtimes shown in the y-axis is different in some panels.

Bayes approach that was integrated into the overall model fitting, thereby reducing the effort of model fitting. The *Sparse Multi-task Lasso* took the longest to run in part because it tuned two parameters by cross-validation, in contrast to the one parameter in the *Elastic Net* and *Group Lasso*. One caveat is that the computational complexity of *mr.mash* scales either quadratically or cubically in r, the number of tissues, so for much larger numbers of tissues *mr.mash* may be much slower than the *Elastic Net* or *Group Lasso* (which scale linearly in r).

Results with missing data

We also compared the methods in settings where some measurements were missing by repeating the simulations described above, but randomly assigned missing values to 70% of the training samples, $\boldsymbol{Y}$, before running the methods. For motivation, note that in the actual GTEx gene expression data about 62% of the entries in $\boldsymbol{Y}$ are missing (unmeasured), so being able to deal with high levels of missing data is important here. Since the R package `glmnet` implementing the *Group Lasso* does not allow for missing values, here we compared *mr.mash* to the *Elastic Net* and the *Sparse Multi-task Lasso* only.

As in the simulations without missing data, *mr.mash* outperformed both the *Elastic Net* and the *Sparse Multi-task Lasso* in most scenarios (Fig 3). The greatest gains in performance were in the Equal Effects and Independent Effects scenarios, and these

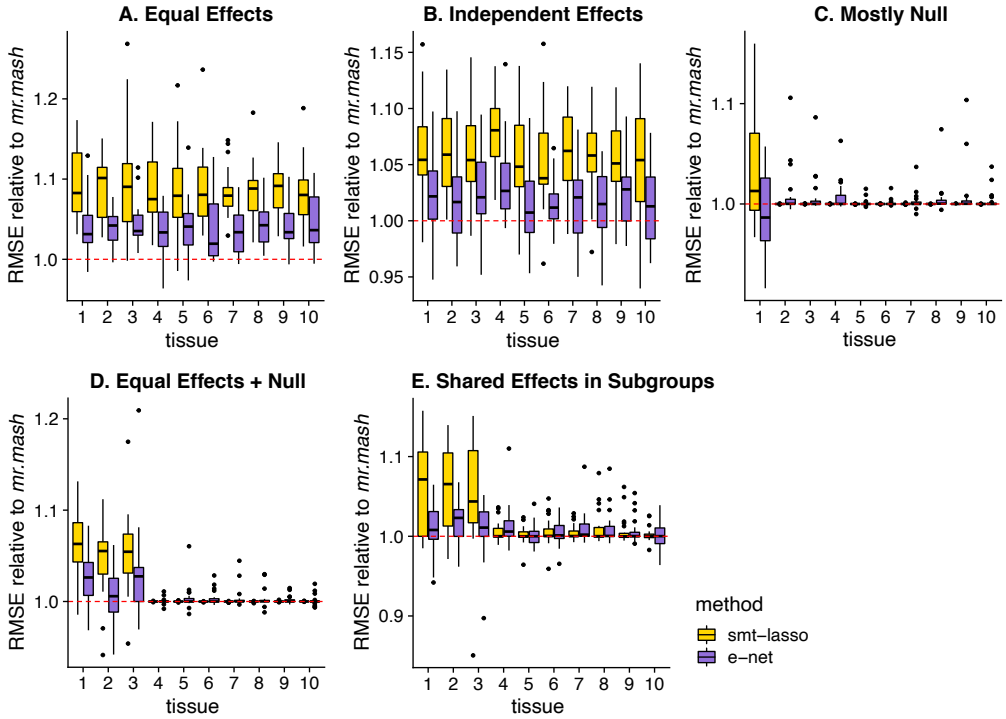

Fig 3. Prediction accuracy in simulations with missing data. Each plot summarizes the accuracy of the test set predictions in 20 simulations. The thick, black line in each box gives the median RMSE relative to the *mr.mash* RMSE. Note that RMSE is a summary of *prediction error*, so lower values of RMSE indicate better prediction accuracy. Also note that the range of RMSE values shown in the y-axis is different in each panel.

gains were greater than in the simulations without missing data (compare to Fig 1). We attribute these greater gains to the fact that the effective sample sizes were smaller in these simulations, and therefore there was more potential benefit to estimating effects jointly when the effects were shared across tissues. Only in the Mostly Null scenario did *mr.mash* perform (slightly) worse than *Elastic Net*. This is not unexpected because there was little benefit to analyzing the tissues jointly in this scenario.

We found the *Sparse Multi-task Lasso* performed poorly in all simulations with missing data, even in scenarios such as the Equal Effects and Independent Effects that favor multivariate regression approaches. This was unexpected and suggests that the implementation of this method for missing data may need improving to be applied in practice.

The introduction of missingness into the simulations increased differences in computation time; in particular *Elastic Net* was faster than with full data, whereas *mr.mash* was slower (Fig 4). This was because *Elastic Net* was applied to each tissue separately, and the missing data simply reduced the size of the data sets, whereas *mr.mash* iteratively imputed the missing data, so the expected computational effort was as if *mr.mash* were run on a full data set. Further, the missing data increased the typical number of iterations needed for the *mr.mash* algorithm to converge to a solution, increasing the overall time involved in model fitting. Like the full-data simulations, the *Sparse Multi-task Lasso* was much slower than the other methods.

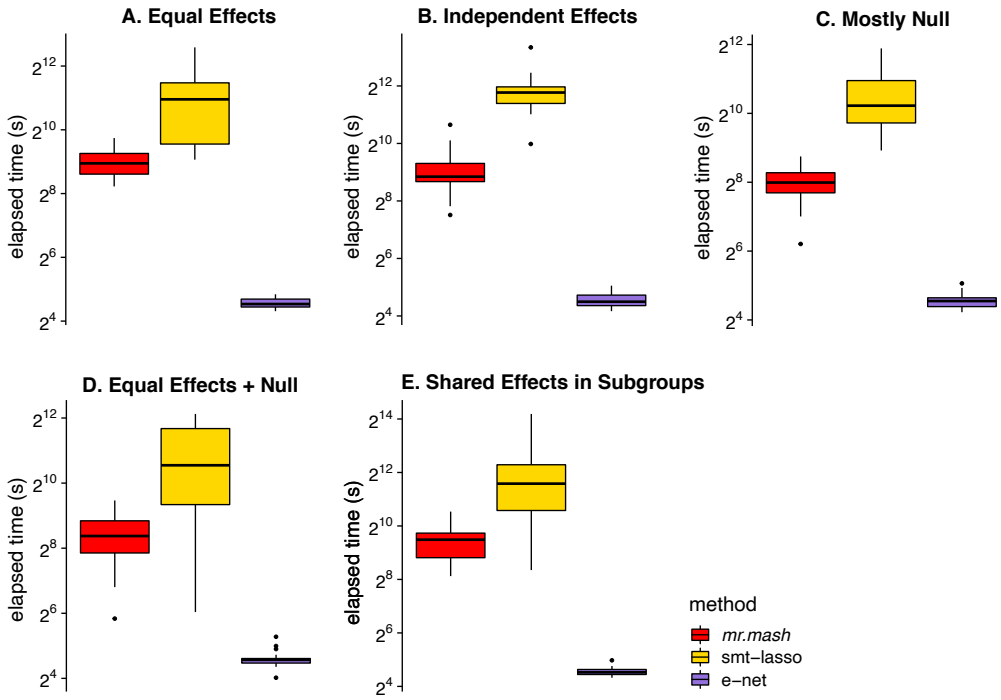

Fig 4. Runtimes in simulations with missing data. Each plot summarizes the distribution of model fitting runtimes for the 20 simulations in that scenario. Note that the *mr.mash* runtimes did not include the initialization step, which was performed using the *Group Lasso*. Once the model fitting was completed, computing the predictions was very fast for all methods, so we did not include the prediction step in these runtimes. See S1 Text for the details on the computing environment used to run the simulations. The thick, black line in each box gives the median runtime. Also note that the range of runtimes shown in the y-axis is different in each panel.

Applications

Case study: predicting gene expression from GTEx data

Finally we considered an application with real data: using genotypes to predict gene expression in 48 tissues, using data from the GTEx Project. The GTEx data includes *post mortem* gene expression measurements obtained by RNA sequencing and genotypes obtained by whole-genome sequencing for 838 human donors [30]. Since expression measurements were not always available in all 48 tissues, it was important for the multivariate analysis to be able to handle missing data. The tissues varied greatly in the number of available gene expression measurements: among the 48 tissues, skeletal muscle had the most measurements available (706), whereas substantia nigra had the fewest (114) (Fig 5).

Using these data, we compared *mr.mash* and the *Elastic Net* for predicting expression from unseen (test) genotypes. (We also performed a more limited comparison with the *Sparse Multi-task Lasso*; see below.) We analyzed 1,000 genes chosen at random, and for each gene we used all genetic variants within 1 Mb of the gene's transcription start site (also removing genetic variants not satisfying certain criteria for inclusion; see S1 Text). To assess the prediction accuracy of each method, we randomly split the 838 GTEx samples into 5 subsets and performed 5-fold cross-validation; that is, we fit the model using a training set composed of 4 out of 5 subsets, then we assessed

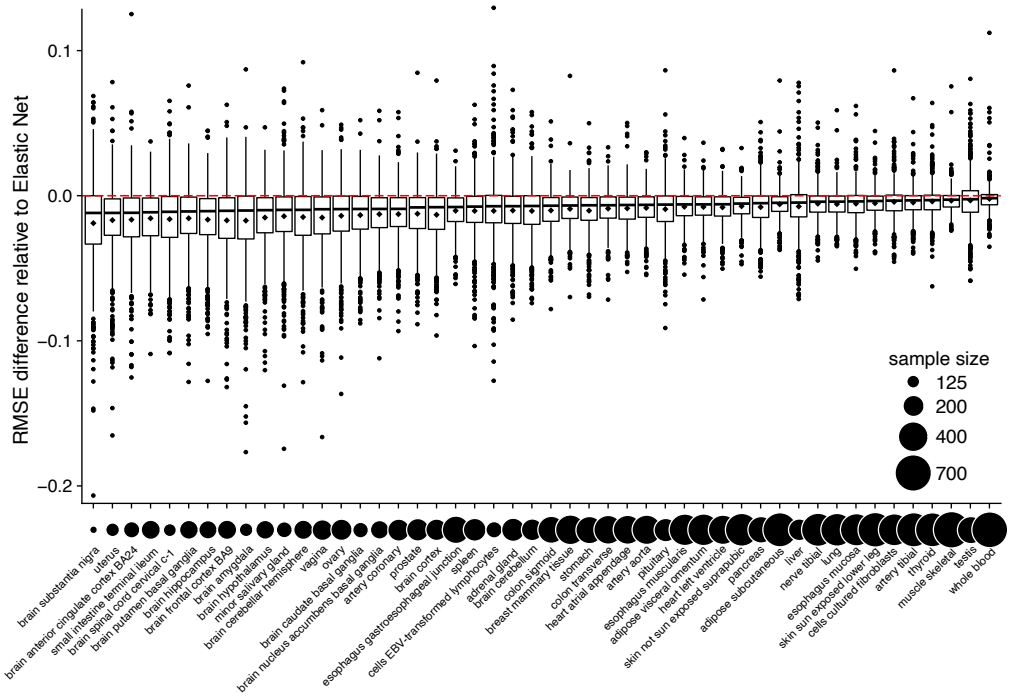

Fig 5. Accuracy of gene expression predictions in GTEx data. Relative RMSE differences between the *Elastic Net* predictions and the *mr.mash* predictions in GTEx test samples are plotted along the y-axis as $\frac{\text{RMSE}(mr.mash)-\text{RMSE}(\text{e-net})}{\text{RMSE}(\text{e-net})}$. Each box in the box plot summarizes the relative RMSE differences from predictions for 1,000 genes. Note that since RMSE is a measure of *prediction error*, lower RMSE differences mean better prediction accuracy. Below the boxes in the box plot, the circles are scaled in area linearly by the number of available gene expression measurements in each tissue.

prediction accuracy in the fifth subset. We repeated this 5 times for each of the 5 splits and summarized prediction accuracy as the average RMSE in the 5 test sets. Prediction accuracy varied considerably with gene and tissue because some genes in some tissues were more strongly predicted by genetic variant genotypes. Therefore, to make results more comparable across genes, we reported *relative* performance accuracy—specifically, the relative difference in RMSE between the two methods, using the *Elastic Net* as a reference point, $\frac{\text{RMSE}(mr.mash)-\text{RMSE}(\text{e-net})}{\text{RMSE}(\text{e-net})}$.

These comparisons are summarized in Fig 5. Overall, *mr.mash* produced substantially more accurate gene expression predictions, although the improvement varied considerably from gene to gene and from tissue to tissue. Interestingly, the improvements tended to be greatest for tissues with more sharing of effects, and for tissues with smaller sample sizes (Fig 5, S3 Fig). In such tissues, the improvement in accuracy was more reflective of the Equal Effects or Independent Effects simulations. For example, the substantia nigra brain tissue had the fewest measurements and benefited from strong sharing of effects with other brain tissues. This strong sharing among the brain tissues is illustrated by the top covariance matrix in the *mr.mash* prior (Fig 6).

In contrast, tissues with the largest sample sizes and more tissue-specific eQTLs tended to show less improvement with multivariate analysis. For example, testis, whole blood and skelelal muscle had weaker sharing of effects (Fig 6), consistent with earlier analyses [30, 34]. In such tissues, there was still some benefit to *mr.mash*, but the gains

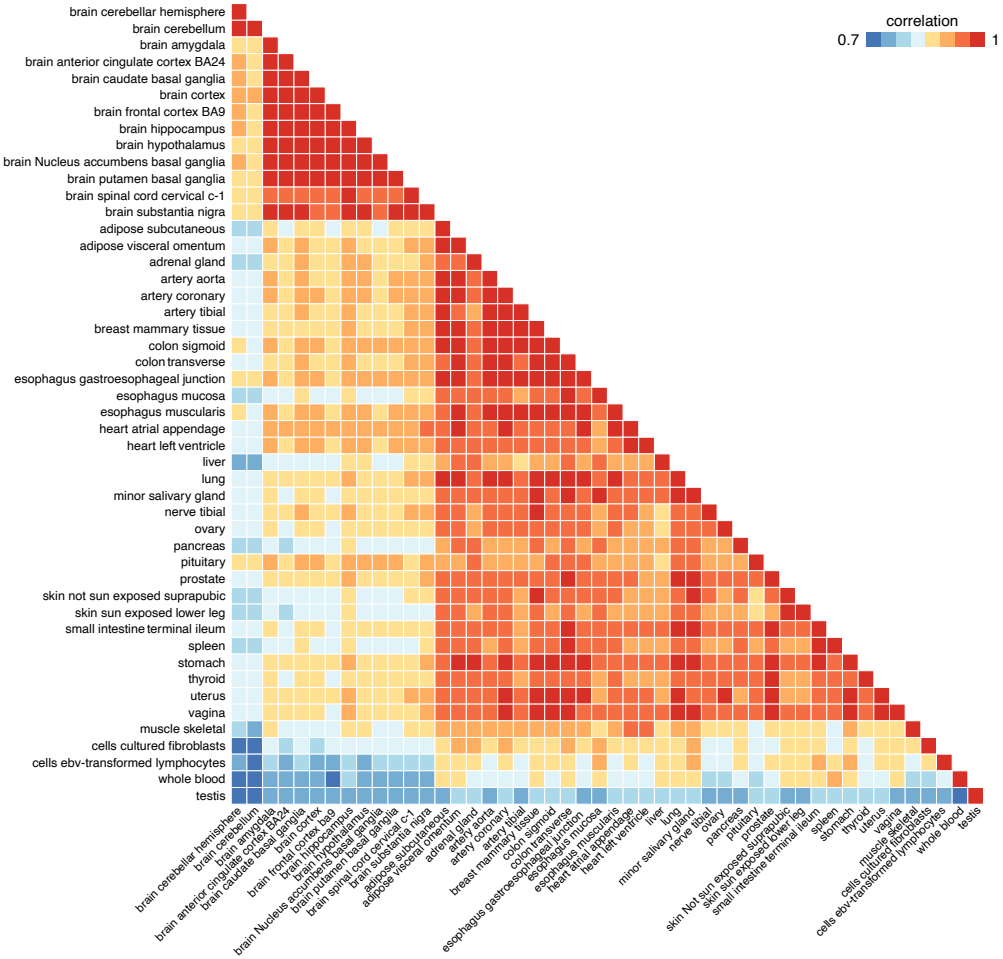

Fig 6. Top pattern of eQTL sharing in GTEx data identified by *mr.mash*. This heatmap shows the the prior covariance matrix that was assigned the greatest weight by *mr.mash*. This covariance matrix was scaled to obtain a correlation matrix shown above.

were more reflective of the Shared Effects in Subgroups or Mostly Null simulations.

We also compared *mr.mash* to the *Sparse Multi-task Lasso*. However, due to the very long running time of the *Sparse Multi-task Lasso* software in these data sets, we performed a more limited comparison on only 10 randomly chosen genes. We fit the *Sparse Multi-task Lasso* to the training set, increasing the size of the grid of the two penalty parameters to 50 in an attempt to improve its performance (in simulations we used a smaller grid of 10 points to reduce computation). The results of this comparison (S4 Fig) illustrate the tendency of the *Sparse Multi-task Lasso* to overshrink effect size estimates, to the point that, in many cases, the scaled RMSE was 1, implying that all the estimated coefficients were exactly zero. *mr.mash* achieved a lower RMSE than the *Sparse Multi-task Lasso* in most cases.

Discussion

We have introduced *mr.mash*, a Bayesian multiple regression framework for modeling multiple (e.g., several dozen) responses jointly, with accurate prediction being the main

goal. A key feature of our approach is that it can learn patterns of effect sharing across responses from the data and use the learned structure to improve prediction accuracy. This feature makes our method flexible and adaptive, which are advantages of particular importance for analyzing large, complex data sets. Our method is also fast and computationally scalable thanks to the use of variational inference (rather than MCMC) for model fitting.

Although we focussed on a specific application—predicting gene expression from genotypes—*mr.mash* is a general method that could be applied to any problem calling for multivariate, multiple regression. This includes, for example, breeding value prediction for multiple related phenotypes in agricultural settings and polygenic score computation for multiple populations in human genetics. Indeed, recent work, performed indepently but using a similar approach, showed improved accuracy in cross-ancestry prediction [25]. In these applications, the number of causal variants is typically much larger (*i.e.*, in the hundreds or even thousands) than for gene expression phenotypes, which could lead to larger improvements in prediction accuracy. While we expect *mr.mash* to be slower in these whole-genome regression applications, it is scalable in that its (per iteration) computational complexity is linear in the number of samples and in the number of predictors. Thus, our method has the potential to have an impact across different fields.

A limitation of *mr.mash* is that it is not ideally suited for selecting among highly correlated variables (which has, for example, been the emphasis of statistical fine-mapping methods [1, 2, 23, 57, 58]). This is because the variational approximation used in *mr.mash* cannot capture the strong dependence in the posterior distribution for the effects of highly correlated variables. Indeed, if two variables are perfectly correlated, and one is causal, *mr.mash* will select one at random and exclude the other [19]. This behavior is also displayed by the Lasso [40]. Therefore, in settings where variable selection is the main goal, alternative approaches (e.g., [23]) may be preferred. Since selecting randomly among correlated variables does not diminish prediction accuracy [19], *mr.mash* can perform well for prediction problems, even when highly correlated variables are present.

Acknowledgments

We thank the University of Chicago Research Computing Center for providing high-performance computing resources used to run the numerical experiments. We also thank Jeff Spence and Jonathan Pritchard for useful discussions. Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Numbers P20GM139769 and R35GM146868 to FM. MS acknowledges support from National Human Genome Research Institute grant R01HG002585. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Schaid DJ, Chen W, Larson NB. From genome-wide associations to candidate causal variants by statistical fine-mapping. Nature Reviews Genetics. 2018;19(8):491–504.

2. Hutchinson A, Asimit J, Wallace C. Fine-mapping genetic associations. Human Molecular Genetics. 2020;29(R1):R81–R88.

3. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. Nature Genetics. 2010;42(7):565–569.

4. de los Campos G, Sorensen D, Gianola D. Genomic heritability: what is it? PLOS Genetics. 2015;11(5):e1005048.

5. Yang J, Zeng J, Goddard ME, Wray NR, Visscher PM. Concepts, estimation and interpretation of SNP-based heritability. Nature Genetics. 2017;49(9):1304–1310.

6. Meuwissen TH, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157(4):1819–1829.

7. Hickey JM, Chiurugwi T, Mackay I, Powell W. Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery. Nature Genetics. 2017;49(9):1297–1303.

8. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nature Genetics. 2018;50(9):1219–1224.

9. Lewis CM, Vassos E. Polygenic risk scores: From research tools to clinical instruments. Genome Medicine. 2020;12:44.

10. Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ, et al. A gene-based association method for mapping traits using reference transcriptome data. Nature Genetics. 2015;47(9):1091–1098.

11. Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BWJH, et al. Integrative approaches for large-scale transcriptome-wide association studies. Nature Genetics. 2016;48(3):245–252.

12. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. Technometrics. 1970;12(1):55–67.

13. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B. 1996;58(1):267–288.

14. Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, Series B. 2005;67(2):301–320.

15. de los Campos G, Vazquez AI, Fernando R, Klimentidis YC, Sorensen D. Prediction of complex human traits using the genomic best linear unbiased predictor. PLoS Genetics. 2013;9(7):e1003608.

16. George EI, McCulloch RE. Variable selection via Gibbs sampling. Journal of the American Statistical Association. 1993;88(423):881–889.

17. Park T, Casella G. The Bayesian Lasso. Journal of the American Statistical Association. 2008;103(482):681–686.

18. Habier D, Fernando RL, Kizilkaya K, Garrick DJ. Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics. 2011;12.

19. Carbonetto P, Stephens M. Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. Bayesian Analysis. 2012;7(1):73–108.

20. Gianola D. Priors in whole-genome regression: the Bayesian alphabet returns. Genetics. 2013;194(3):573–596.

21. Zhou X, Carbonetto P, Stephens M. Polygenic modeling with Bayesian sparse linear mixed models. PLoS Genetics. 2013;9(2):e1003264.

22. Moser G, Lee SH, Hayes BJ, Goddard ME, Wray NR, Visscher PM. Simultaneous discovery, estimation and prediction analysis of complex traits Using a Bayesian mixture model. PLoS Genetics. 2015;11(4):e1004969.

23. Wang G, Sarkar A, Carbonetto P, Stephens M. A simple new approach to variable selection in regression, with application to genetic fine mapping. Journal of the Royal Statistical Society, Series B. 2020;82(5):1273–1300.

24. Zabad S, Gravel S, Li Y. Fast and accurate Bayesian polygenic risk modeling with variational inference. bioRxiv. 2022;doi:10.1101/2022.05.10.491396.

25. Spence JP, Sinnott-Armstrong N, Assimes TL, Pritchard JK. A flexible modeling and inference framework for estimating variant effect sizes from GWAS summary statistics. bioRxiv. 2022;doi:10.1101/2022.04.18.488696.

26. Kim Y, Wang W, Carbonetto P, Stephens M. A flexible empirical Bayes approach to multiple linear regression and connections with penalized regression. arXiv. 2022;2208.10910.

27. Falconer DS, Mackay TFC. Introduction to quantitative genetics. 4th ed. Essex: Harlow, Longman; 1996.

28. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018;562(7726):203–209.

29. Kanai M, Akiyama M, Takahashi A, Matoba N, Momozawa Y, Ikeda M, et al. Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. Nature Genetics. 2018;50(3):390–400.

30. Aguet F, Barbeira AN, Bonazzola R, Jo B, Kasela S, Liang Y, et al. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 2020;369(6509):1318–1330.

31. Stephens M. A unified framework for association analysis with multiple related phenotypes. PLoS ONE. 2013;8(7):e65245.

32. Inouye M, Ripatti S, Kettunen J, Lyytikäinen LP, Oksala N, Laurila PP, et al. Novel Loci for Metabolic Networks and Multi-Tissue Expression Studies Reveal Genes for Atherosclerosis. PLoS Genetics. 2012;8(8).

33. O'Reilly PF, Hoggart CJ, Pomyen Y, Calboli FCF, Elliott P, Jarvelin MR, et al. MultiPhen: Joint model of multiple phenotypes can increase discovery in GWAS. PLoS ONE. 2012;7(5).

34. Urbut SM, Wang G, Carbonetto P, Stephens M. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. Nature Genetics. 2019;51(1):187–195.

35. Turchin MC, Stephens M. Bayesian multivariate reanalysis of large genetic studies identifies many new associations. PLoS Genetics. 2019;15(10):e1008431.

36. Jia Y, Jannink JL. Multiple-trait genomic selection methods increase genetic value prediction accuracy. Genetics. 2012;192(4):1513–1522.

37. Maier RM, Zhu Z, Lee SH, Trzaskowski M, Ruderfer DM, Stahl EA, et al. Improving genetic prediction by leveraging genetic correlations among human diseases and traits. Nature Communications. 2018;9:989.

38. Hu Y, Li M, Lu Q, Weng H, Wang J, Zekavat SM, et al. A statistical framework for cross-tissue transcriptome-wide association analysis. Nature Genetics. 2019;51(3):568–576.

39. Grinberg NF, Wallace C. Multi-tissue transcriptome-wide association studies. Genetic Epidemiology. 2021;45(3):324–337.

40. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software. 2010;33(1):1–22.

41. Henderson CR, Quaas RL. Multiple trait evaluation using relatives' records. Journal of Animal Science. 1976;43(6):1188–1197.

42. Calus MPL, Veerkamp RF. Accuracy of multi-trait genomic selection using different methods. Genetics Selection Evolution. 2011;43:26.

43. Cheng H, Kizilkaya K, Zeng J, Garrick D, Fernando R. Genomic prediction from multiple-trait Bayesian regression methods using mixture priors. Genetics. 2018;209(1):89–103.

44. Gianola D, Fernando RL. A multiple-trait Bayesian lasso for genome-enabled analysis and prediction of complex traits. Genetics. 2020;214(2):305–331.

45. Dawid AP. Some matrix-variate distribution theory: notational considerations and a Bayesian application. Biometrika. 1981;68(1):265–274.

46. Gupta AK, Nagar DK. Matrix variate distributions. Boca Raton, FL: Chapman & Hall; 2000.

47. Blei DM, Kucukelbir A, McAuliffe JD. Variational inference: a review for statisticians. Journal of the American Statistical Association. 2017;112(518):859–877.

48. Jordan MI, Ghahramani Z, Jaakkola TS, Saul LK. An introduction to variational methods for graphical models. Machine Learning. 1999;37(2):183–233.

49. Logsdon BA, Hoffman GE, Mezey JG. A variational Bayes algorithm for fast and accurate multiple locus genome-wide association analysis. BMC Bioinformatics. 2010;11.

50. Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. Journal of Machine Learning Research. 2003;3:993–1022.

51. Saul LK, Jordan MI. Exploiting tractable substructures in intractable networks. In: Touretzky DS, Mozer MC, Hasselmo ME, editors. Advances in Neural Information Processing Systems. vol. 8; 1996. p. 486–492.

52. Ghahramani Z, Hinton GE. Variational learning for switching state-space models. Neural Computation. 2000;12(4):831–864.

53. Hayashi T, Iwata H. A Bayesian method and its variational approximation for prediction of genomic breeding values in multiple traits. BMC Bioinformatics. 2013;14:34.

54. R Core Team. R: a language and environment for statistical computing; 2020. Available from: https://www.R-project.org/.

55. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society, Series B. 2006;68(1):49–67.

56. Lee S, Zhu J, Xing EP. Adaptive multi-task lasso: with application to eQTL detection. In: Lafferty JD, Williams CKI, Shawe-Taylor J, Zemel RS, Culotta A, editors. Advances in Neural Information Processing Systems. vol. 23; 2010. p. 1306–1314.

57. Benner C, Spencer CCA, Havulinna AS, Salomaa V, Ripatti S, Pirinen M. FINEMAP: efficient variable selection using summary data from genome-wide association studies. Bioinformatics. 2016;32(10):1493–1501.

58. Zou Y, Carbonetto P, Wang G, Stephens M. Fine-mapping from summary data with the "Sum of Single Effects" model. PLOS Genetics. 2022;18(7):e1010299.

Supporting information

S1 Fig. Prediction accuracy of *mr.mash* variants in simulations with full data. Each plot summarizes the accuracy of the test set predictions in the 20 simulations for that scenario. The three methods compared were: (1) *mr.mash* with only "canonical" prior covariance matrices; (2) *mr.mash* with only "data-driven" prior covariance matrices; and (3) *mr.mash* with both types of prior covariance matrices. The thick, black line in each box gives the median RMSE relative to the "data-driven" *mr.mash* RMSE. Note that RMSE is a summary of *prediction error*, so lower values of RMSE indicate better prediction accuracy. Also note that the range of RMSE values shown in the y-axis is different in each panel.

S2 Fig. Runtimes for *mr.mash* variants in simulations with full data. Each plot summarizes the distribution of model-fitting runtimes for the 20 simulations in that scenario. Note the runtimes did not include the initialization step, which was implemented by running the *Group Lasso* on the same data set. Once the model fitting was completed, computing the predictions was very fast, so we did not include the prediction step in these runtimes. See S1 Text for the details on the computing environment used to run the simulations. Note that the range of runtimes shown in the y-axis is different in each panel.

S3 Fig. Relationship between improvement in prediction acccuracy and GTEx tissue sample size. Tissues are plotted along the x-axis by the number of available gene expression measurements and along the y-axis by the improvement in RMSE relative to the *Elastic Net*; that is,
$(\text{RMSE}(mr.mash) - \text{RMSE}(\text{e-net}))/\text{RMSE}(\text{e-net})$.

S4 Fig. Comparison of *mr.mash* vs. *Sparse Multi-task Lasso* for 10 randomly chosen genes in GTEx data. Each plot compares the accuracy of the *mr.mash* and *Sparse Multi-task Lasso* gene expression predictions in test samples for a single gene, separately for each tissue. The prediction accuracy is summarized as the

RMSE relative to the RMSE that would be obtained by the "naive" predictor in which the genotype has no effect on expression (the naive predictor is therefore simply the mean of the expression measurements in the training data); that is, the x-axis shows RMSE(smt-lasso)/RMSE(naive) and the y-axis shows RMSE(mr.mash)/RMSE(naive). Note that some genes are not expressed in all tissues and so some plots have fewer than 48 points.

S1 Text. Detailed methods. Detailed description of the methods, including: preparation of GTEx data; simulations with GTEx genotypes; methods compared in the simulations; derivations of *mr.mash* variational algorithms with full data; and derivations of algorithms with missing data.