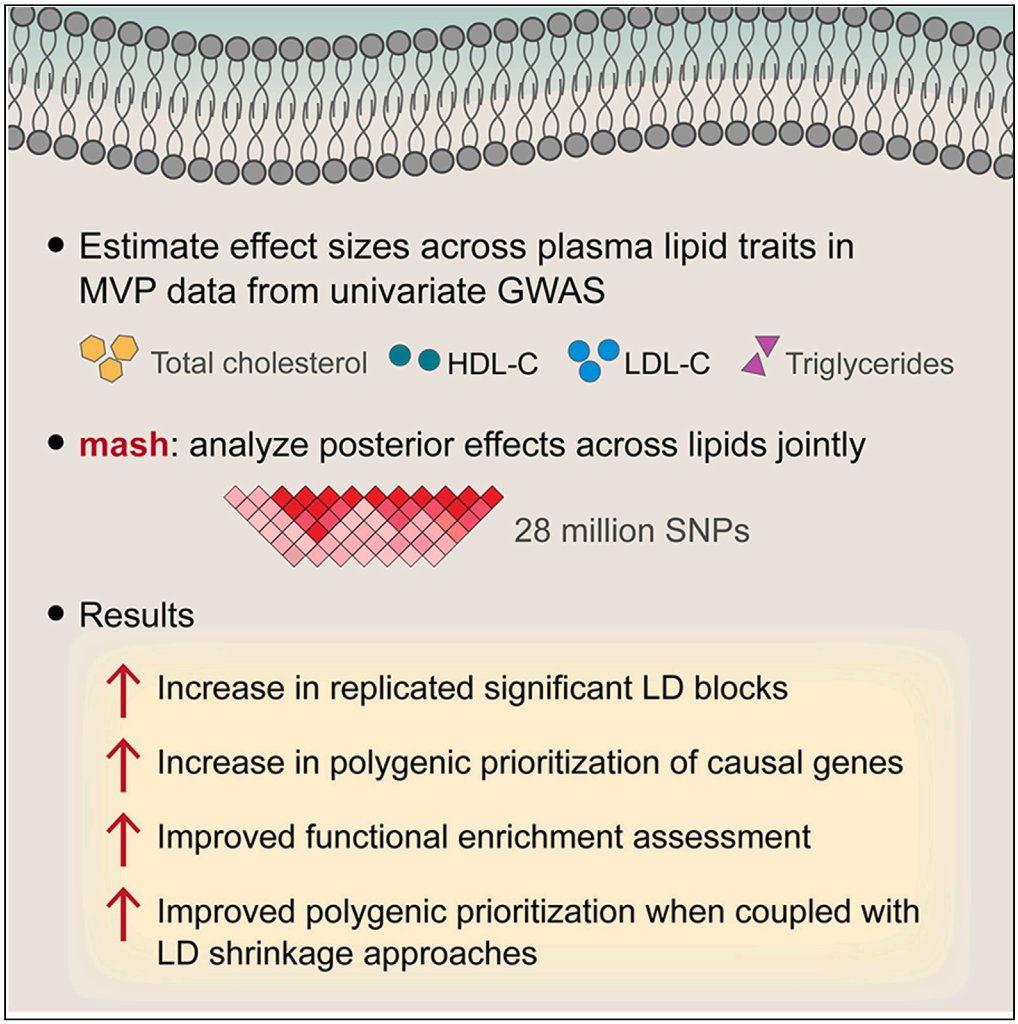
iScience

CellPress
OPEN ACCESS

Article

Bayesian multivariate genetic analysis improves translational insights



Sarah M. Urbut, Satoshi Koyama, Whitney Hornsby, ..., Christopher J. O'Donnell, Gina M. Peloso, Pradeep Natarajan

pnatarajan@mgh.harvard.edu

Highlights

Bayesian shrinkage tool, *mash*, improves genomic understanding of plasma lipids

Controlling local false discovery identifies non-null effects associated with lipids

Coupling a multivariate approach with existing polygenic scoring improves prediction

Improved enrichment for annotations and prioritization of causal genes for lipids

Urbut et al., iScience *26*, 107854
October 20, 2023
https://doi.org/10.1016/j.isci.2023.107854

iScience

CellPress
OPEN ACCESS

Article

Bayesian multivariate genetic analysis improves translational insights

Sarah M. Urbut,[1,2] Satoshi Koyama,[1,2,3] Whitney Hornsby,[1,2,3] Rohan Bhukar,[1,2,3] Sumeet Kheterpal,[1,2] Buu Truong,[1,2,3] Margaret S. Selvaraj,[1,2,3] Benjamin Neale,[2,3,4] Christopher J. O'Donnell,[3,5] Gina M. Peloso,[6] and Pradeep Natarajan[1,2,3,7,8,*]

SUMMARY

While lipid traits are known essential mediators of cardiovascular disease, few approaches have taken advantage of their shared genetic effects. We apply a Bayesian multivariate size estimator, mash, to GWAS of four lipid traits in the Million Veterans Program (MVP) and provide posterior mean and local false sign rates for all effects. These estimates borrow information across traits to improve effect size accuracy. We show that controlling local false sign rates accurately and powerfully identifies replicable genetic associations and that multivariate control furthers the ability to explain complex diseases. Our application yields high concordance between independent datasets, more accurately prioritizes causal genes, and significantly improves polygenic prediction beyond state-of-the-art methods by up to 59% for lipid traits. The use of Bayesian multivariate genetic shrinkage has yet to be applied to human quantitative trait GWAS results, and we present a staged approach to prediction on a polygenic scale.

INTRODUCTION

A principal goal of genome-wide association studies (GWAS) is to accurately identify genetic variants that influence the risk of developing a trait. While the number of significantly associated variants increases with a larger sample size, most estimated heritability remains unexplained.[1,2] Novel Bayesian methods leveraging genetic pleiotropy applied to existing samples may improve power for genetic discovery beyond widespread univariate approaches.[3,4] As many phenotypes in biology exist on quantitative and continuous spectra, describing multivariate continuous effects is an essential step toward a better understanding of complex phenotypes.

Multivariate adaptive shrinkage (mash[4]) is a Bayesian adaptive shrinkage tool designed to estimate genetic variants associated with multiple phenotypes. Mash takes advantage of any correlation in genetic signals that might increase the power to detect associations and improve the precision of effect size estimates.[4,5] Mash uses empirical estimates of the overall covariance structure of phenotypes to model the genetic effect at any single nucleotide polymorphism (SNP) as a mixture of multivariate normal distributions. Each mixture component defines a 'pattern of sharing' of effects among conditions from which the SNP might arise (Figure 1) and is readily available using the accompanying software mashR.[4] Using empirical Bayesian methods, the covariance patterns are estimated from the strongest signals in the dataset, scaled to unit variance, and 'stretched' by a grid of magnitudes to reflect an abundance of shape–scale combinations. The relative frequency of each shape and scale combination is then estimated from a sampling of the overall dataset. Given this prior information, the likelihood for each SNP is then calculated for each mixture component. If the effects are strongly correlated, such estimates are enhanced by incorporating additional data — that is, by using all traits jointly. Importantly, if there is no sharing, multivariate estimates have been shown to do no worse than univariate estimates.[4,6] To date, no existing methods for Bayesian multivariate effect size shrinkage have been applied to human GWAS and staged into a predictive framework to estimate polygenic scores. The proposed approach leverages the information borrowed across quantitative clinical phenotypes to improve effect size estimates and later across genetic markers to improve polygenic prediction.

Our work is an application of the existing multivariate Bayesian modeling approach, mash, to multivariate GWAS to establish improved power, prediction, and prioritization using a joint approach. Bayesian methods allow one to combine information across traits to better estimate individual SNP effects. In addition, they provide a rational and quantitative way to incorporate biological data, and they can allow for a

[1]Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA 02114, USA
[2]Program in Medical and Population Genetics, Broad Institute, Cambridge, MA 02142, USA
[3]Department of Medicine Harvard Medical School, Boston, MA 02115, USA
[4]Analytic Translational and Genetics Unit, Massachusetts General Hospital, Boston, MA 02114, USA
[5]VA Boston Department of Veterans Affairs, Boston, MA 02130, USA
[6]Department of Biostatistics, Boston University School of Public Health, Boston, MA 02218, USA
[7]X (formerly Twitter): @pnatarajanmd
[8]Lead contact
*Correspondence: pnatarajan@mgh.harvard.edu
https://doi.org/10.1016/j.isci.2023.107854

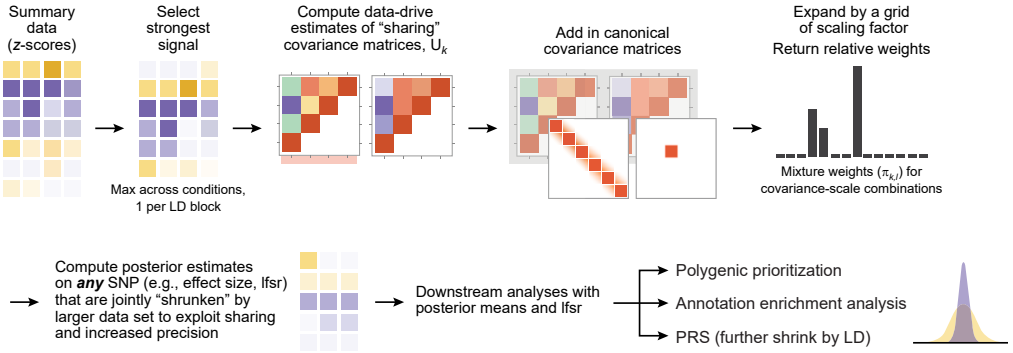


Figure 1. Mash estimates data-drive covariance patterns of true genetic effects as the multivariate prior to improve posterior estimates for downstream analyses

Mash[35] estimates the covariance of the effects in an empirical Bayes fashion, thus estimating patterns of sharing among conditions (here, lipid traits) from the strongest signals in the data, and estimating the relative abundance of such patterns from a random set of all data. This allows us to provide the posterior estimate of the effect and its associated local false sign rate, or posterior probability of incorrectly identifying the sign of the effect, for each SNP and use these posterior estimates to improve performance in polygenic prioritization, enrichment analyses, on polygenic risk scoring. mash, multivariate adaptive shrinkage; SNP, single nucleotide polymorphism; lfsr, local false sign rate; PRS, polygenic risk score; LD, linkage disequilibrium.

range of possible genetic models in a single analysis.[7] We then combine information between traits using a Bayesian framework for shrinkage by patterns of linkage disequilibrium.[8]

The use of Bayesian multivariate tools has broad implications for GWAS and polygenic risk score construction. Here we present a staged approach to prediction on a polygenic scale, which has yet to be applied to human quantitative trait GWAS results.

As a case study, we apply mash to GWAS for blood lipid concentrations. These studies have advanced our understanding of causal relationships for diverse cardio-metabolic conditions, including coronary artery disease (CAD), the leading cause of death worldwide.[9] Since currently identified variants explain only a tiny fraction of the estimated overall heritability of plasma lipids,[10,11] improved efficiency for lipid genetic associations may yield new insights using existing genetic association data.

RESULTS

Multivariate genetic discovery for plasma lipids

Using summary-level lipid GWAS data generated with conventional methods from the Million Veterans Project (MVP) across 291,746 individuals (210,967; 72.3% European ancestry[10]), we sought first to jointly estimate effect sizes across the four plasma lipid traits (i.e., total cholesterol [TC], low-density lipoprotein [LDL-C] cholesterol, high-density lipoprotein [HDL-C] cholesterol, and triglycerides [TG]) for approximately 28 million SNPs using mash. We identify 5583 500-kb linkage disequilibrium (LD) blocks (from 1000 Genomes European Samples[12]) in the MVP data containing at least one variant with a local false sign rate <0.05 for at least one lipid trait. Here, the local false sign rate (lfsr) is defined as the posterior probability of incorrectly identifying the sign of the effect, which has been used to help provide a bridge between FDR (false discovery rates) and effect size estimation.[5] We use an lfsr threshold of 0.05 in line with existing published resources to define an effect with a low posterior probability of being null or incorrectly signed.[3,4,7]

We emphasize that a local false sign rate is distinct from a family-wise error rate. A local false sign rate reflects only the probability of an effect being non-zero and incorrectly signed. By contrast, a family-wise error rate corrected p value controls the global level of at least one false positive association without accounting for the inherent significance (or nullness) of the dataset.

Omnigenic model

We next used mashR to analyze the distribution of regression coefficients from the set of all SNPs.[5] Mash models the GWAS results as a mixture of SNPs that have a true effect size of precisely zero, with SNPs that have a true effect size that is not zero across multiple traits. The additional multivariate layer captures information about patterns of sharing across traits when compared to univariate shrinkage approaches.[5] Critically, our approach uses *adaptive shrinkage* in that patterns of sharing are learned from the data in a hierarchical fashion and then used to 'nudge' noisy likelihood measurements toward patterns consistent with the overall signal. Using this approach, we estimated that 84% of all SNPs present in the MVP dataset are associated with non-zero effects on LDL-C, including both causal SNPs and nearby SNPs in linkage disequilibrium (LD). This does not imply that most variants are causal, given that the typical extent of LD was around 10 kb–100 kb.[13] Instead, this suggests that most 100-kb windows in the genome include variants that affect lipid levels. When stratifying by the LD score[14] for each SNP, we see a clear effect that SNPs with more LD partners are more likely to be associated with each lipid trait (Figure S13). This is directly consistent with the results of Bulik–Sullivan[14] et al. who showed that in settings of polygenic architecture, the LD score plot is indeed a straight line. Indeed, the LD score for a given SNP **i** is the sum of the squared correlations[14,15] with SNP_j: $LD_i = \left(\sum_j \rho_{ij}\right)^2$. We used LD scores

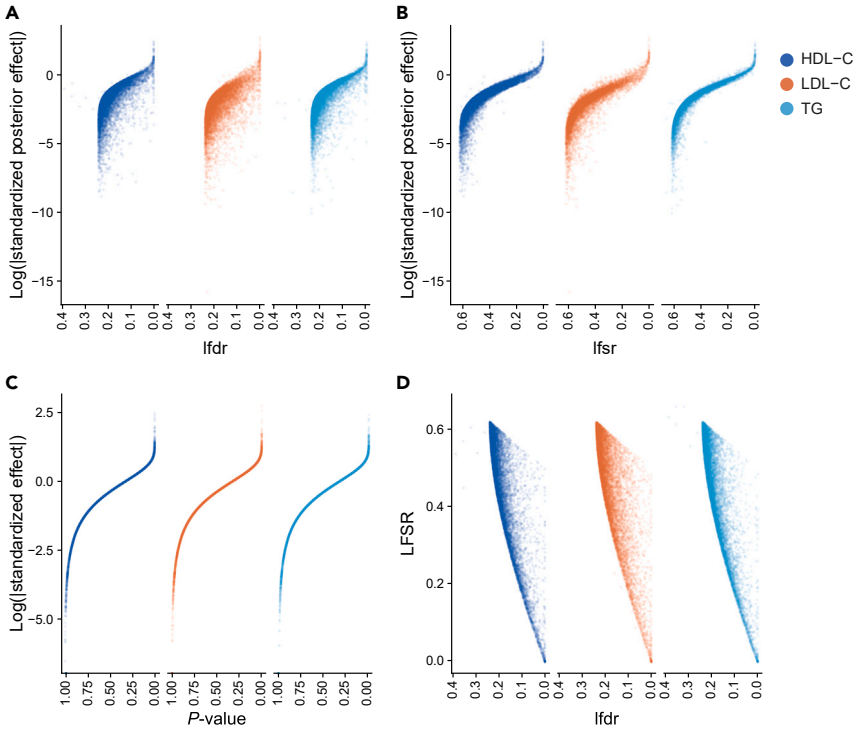


Figure 2. The utility of controlling for false discovery

(A and B) (A) A multivariate approach allows that for a given probability of being null (lfdr)[36] or for a given local false sign rate (lfsr) (B) there can be a variety of effect sizes depending on the relative strength of evidence in alternative subgroups.

(C) We demonstrate the relationship between effect size and p value.

(D) Finally, (D) a given non-null rate can lead to greater resolutions in the range of possible local false sign rates as reflected in a variety of Local false sign rates for a given non-null rate. HDL-C, high-density lipoprotein cholesterol; lfsr, local false sign rate; lfdr, local false discovery ratel; LDL-C, low-density lipoprotein cholesterol; LDSC, linkage disequilibrium score; TG, triglycerides.

from the European ancestry samples in the 1000 Genomes Project (EUR) with an unbiased estimator of r2 with 1 centiMorgan (cM) windows.[14] Importantly, this finding is also consistent with the conclusions of Boyle et al.[16] who introduced the idea of an 'omnigenic' hypothesis, in which many quantitative traits are influenced by both the majority of genomic SNPs and a vast number of causal variants, each with tiny effect sizes on quantitative traits.

Boyle et al.[16] found clear enrichment of a shared directional signal for most SNPs, even for SNPs with p values as large as 0.5. This led us to consider the information contained in those with non-null but potentially not genome-wide significant effects. After applying mashR, we found that the median absolute effect size for an SNP satisfying an lfsr threshold of 0.05 was 0.15 while the median absolute effect size for an SNP satisfying a local false discovery rate (lfdr) of 0.05 was 0.15. By contrast, an SNP satisfying a genome-wide significance threshold of 5×10^{-8} has a standardized effect of 6.54, consistent with the results of Boyle et al.[16] who found that the median effect of non-null SNPs was approximately 10% of genome-wide significant SNPs (Figures 2A–2C). Importantly, both the resolution and stringency using lfsr thresholds is greater than lfdr for a given level of evidence (Figure 2D). For example, an effect may have a very low estimated local false discovery rate if there exists sufficient evidence it is significantly different from 0. However, the local false sign rate is in direct proportion to both the size and the precision of the effect: for small, precise effects, the probability of incorrectly signing the effect may still be substantial for a given local false discovery rate.

Given this observation, we hypothesized that utilizing methods capable of incorporating refined joint posterior effect size estimates and quantifying posterior probabilities of being non-zero (or, even more stringently, incorrectly signed) would add to the ability to explain the heritability of complex disease in a polygenic risk score. We sought to compare our estimates directly to a univariate approach for adaptive Bayesian shrinkage,[5] as published GWAS have used a 5% FDR threshold to replicate GWAS targets.[17] We show that a univariate shrinkage approach that controls for local false discovery (ash) replicates everything in previously published MVP data,[10] which results in a consistent increase in power across phenotypes (Figures 3A and 3C, green to blue). This is explained by the adaptive control of local false discovery when compared with Bonferroni-corrected GWAS. We depict the number of LD blocks containing an effective variant in at least one block, defined by p value $<5 \times 10^{-8}$ in traditional analyses or lfsr $<5 \times 10^{-2}$ in adaptive shrinkage analyses. We note that given the large size of the conservatively chosen LD blocks, the probability of containing at least one associated SNP is increased and therefore there is a non-null probability of replication by chance. However, the number of blocks replicated using mash for identification is greater than the number of blocks replicated by alternative methods.

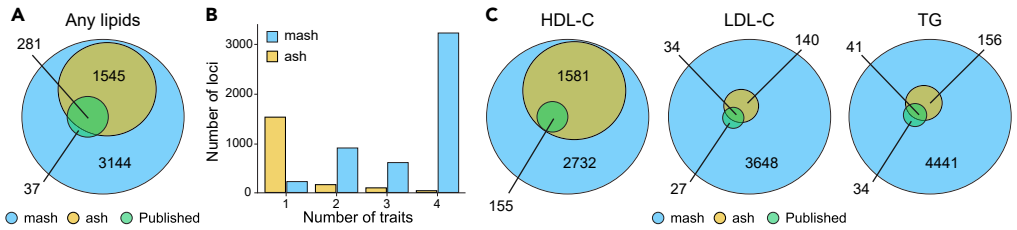


Figure 3. Control of false discovery improves power to detect over control of family-wise error rate

(A and B) (A) Univariate measure of local false sign rate control using ash[5] replicates essentially all existing associations and dramatically increases power to detect. Multivariate adaptive shrinkage adds an additional layer of local false sign rate control by incorporating information across phenotypes. We plot the number of LD blocks containing at least one significant variant across traits in (B) joint approach results in most significant associations being shared in at least 2 subgroups, whereas a univariate approach does not capture the tendency to share effects across conditions.

(C) HDL-C, LDL-C, and TG. Of note, there are 5583 500-kb blocks present in our dataset. Ash, univariate adaptive shrinkage; mash, multivariate adaptive shrinkage; HDL-C, high-density lipoprotein cholesterol; LDL-C, low-density lipoprotein cholesterol; TG, triglycerides.

We ascertained the improvement in multivariate control of false discovery when comparing multivariate adaptive shrinkage to univariate adaptive shrinkage (Figure 3C, blue to red). Most effects were shared across conditions, owing to the assessment of effects jointly, in comparison to univariate shrinkage effects (Figure 3B).

Improved polygenic risk scoring across ancestries

Mash-derived weights show improvements up to 58% in the proportion of variance explained by the polygenic score compared to weights derived from traditional univariate maximum likelihood association mapping. We compared with the same approach using the results from an MTAG analysis of these summary statistics as inputs to the Ldpred2 model. In most cases, using MTAG results as input performed no better than using univariate (maximum likelihood estimate, MLE) analysis (Figures 4A–4C). These improvements hold when subdividing the testing population into European and non-European individuals on all traits excluding triglycerides, a phenomenon which has been previously observed[18] (Figures 4A–4C; Table S1). Of note, the baseline performance of models excluding genetic variables from association with

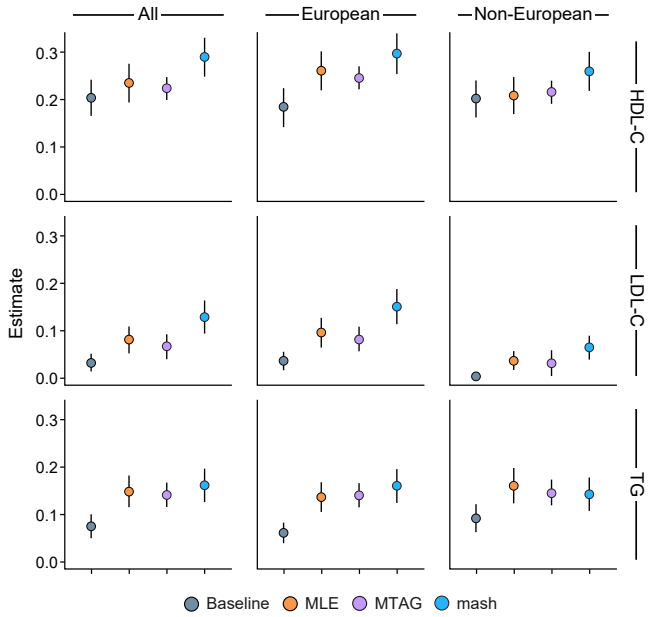


Figure 4. Mash improves polygenic prediction

We consider the improvement in proportion of variation explained by LDpred2[8] on prediction of lipid traits across ethnicities using mash derived posteriors and univariate GWAS estimates as weight inputs over a model including only baseline covariates. Here we display the estimate of R^2 and corresponding 95% CI. We compare the performance of the infinitesimal model using maximum likelihood estimates (MLE), multivariate (mash) or multivariate trait association for GWAS (MTAG) output for all (global), European ancestry, or non-European ancestry (See STAR methods for details; Table S4 for results in tabular form) to a baseline model using only baseline covariates of age and sex in each model. GWAS, genome-wide association study; Ash, univariate adaptive shrinkage; mash, multivariate adaptive shrinkage; HDL-C, high-density lipoprotein cholesterol; LDL-C, low-density lipoprotein cholesterol; TG, triglycerides.

LDL-C is poor owing to the phenotypic heterogeneity inherent in statin dosing and duration present in the UKBB population. Despite this, mash inputs still improve the proportion of variation explained by 53%. Significantly, we shrink the input 'weights' derived from our multivariate discovery set using LD score regression from the 1000 Genomes reference panel.[12] Next, we computed scores on an independent set of individuals in the UKBB, without the need for parameter tuning in an additional collection of data. This is due to the analytic solution innate in the infinitesimal model introduced by Privé[8] (details in Materials and Methods). Most importantly, we compare our results using mash to existing multivariate effect size estimation for GWAS, MTAG.[19] Mash is superior to MTAG across ancestries and lipid traits. Furthermore, MTAG appears to perform similarly to MLEs. MTAG estimates the covariance of the errors using LD score regression, but mash's additional power comes from its ability to boost (or shrink) effect sizes from the estimated covariance of the true effects rather than the residual error covariance matrix alone. We also compared using PRS-CS and showed that mash often improves predictive power even without additional tuning parameter selection (as in the infinitesimal model of LDpred2) when compared to raw MLE summary statistic input (Table S7). However, in a previously published head-to-head comparison, LDpred2 performance was superior to all Bayesian shrinkage approaches in polygenic prediction.[8] We display the results using LDpred2-infinitesimal model. We demonstrate with additional cross-validation, it is possible to achieve even greater performance as shown using the LDPred2-Grid model (Table S8). Importantly, while the exchangeable effects are necessary for the estimation of the mash prior, it is **not** a condition of the prediction algorithm.

Furthermore, mash allowed us to consider which hierarchical patterns received the most 'weight' in the mixture model. As expected, the components that received most of the hierarchical weight showed effects that were shared in sign and magnitude among LDL-C, TC and TG and strongly inversely correlated with effects in HDL-C (Figure S1, online workflow).

Sources of improved heritability explained

To understand the improved heritability explained, we evaluated the extent of cross-replication between UK Biobank (UKBB) and MVP results. After computing mash posterior means for all overlapping SNPs from models fit separately using MVP and UKBB (Figures S2 and S3), we aimed to identify replication on a per-trait and across-trait basis. Using UKBB lipid genetic summary statistics for 315,133 individuals (100% European ancestry)[20] overlapping the same boundaries used above, we identified 3,935 500-kb LD blocks containing at least one lfsr <0.05 across traits. 3761 of these 3935 of these identified UKBB blocks were replicated in the MVP discovery dataset. We choose 500-kb blocks as this is a conservative estimate of human LD using data from the 1000 Genomes Project.[21,22] We note that given the large size of the conservatively chosen LD blocks, the probability of containing at least one associated SNP is increased and therefore there is a non-null probability of replication by chance. However, the number of blocks replicated using mash for identification is greater than the number of blocks replicated by alternative methods: 95.6% and 75.1% of the discoveries in UKBB and MVP, respectively, cross-replicate (Figure 5A). Using a hypergeometric test for the probability of observing replication by chance given the number of significant blocks, we find $p = 1.27 \times 10^{-81}$ in HDL-C, 4.98×10^{-60} in LDL-C, and 6.96×10^{-75} in TG. Additionally, we reproduced all associations captured by the prevailing multivariate GWAS package MTAG,[19] and we replicated 12-fold (Figure 5A; Figure S3) more LD blocks when comparing non-zero blocks containing at least one significant effect across traits between MTAG[19] and mash-posterior results without incurring additional false positives.

In the per-trait analysis, 68% and 88% of MVP and UKBB discoveries cross-replicate for HDL-C, 42% and 82% of MVP and UKBB discoveries cross-replicate for LDL-C, 63% and 91% of MVP and UKBB cross-replicate for triglycerides, and 50% and 81% of MVP and UKBB cross-replicate for total cholesterol (Figure S3). Again, this increase in power holds across traits when compared to existing multivariate approaches for common diseases using the software MTAG.[19] We identify substantially more blocks containing a variant significant in at least one trait (Figure 3A) when compared to a traditional univariate assessment using a genome-wide family-wise error rate of 5×10^{-8}. Our Bayesian multivariate effect size estimate using mash improved the sensitivity to detect and replicate associations between datasets.

We find that this sizable improvement over univariate association is driven largely by gains in precision as well as control of false discovery versus family-wise error rate. This concept is shown to safely increase the power to detect associated non-zero effects The effective sample size can be determined by considering the ratio of the original standard error to the posterior standard error for an individual trait (Equation 1). We found that effects with small sample sizes (and accordingly large standard errors) benefit from an increase in posterior sample size over initial sample size.

$$n_{effective} = \frac{\hat{s}^2_{original}}{\tilde{s}^2_{posterior}} n_{original} \quad \text{(Equation 1)}$$

We found a median effective sample boost of 4.4-fold (IQR 3.54–5.64) using the relationship between the original standard error and the posterior marginal variance, consistent with the robust sharing among lipid subgroups (Figure S4).

Improved causal gene prioritization

We investigated how mash would prioritize known Mendelian lipid targets compared to univariate methods. We used the Polygenic Priority Score (PoPS), which is a gene prioritization method[23] that leverages genome-wide signal from GWAS summary statistics. PoPS incorporates data from extensive public bulk and single-cell expression datasets, curated biological pathways, and predicted protein–protein interactions.

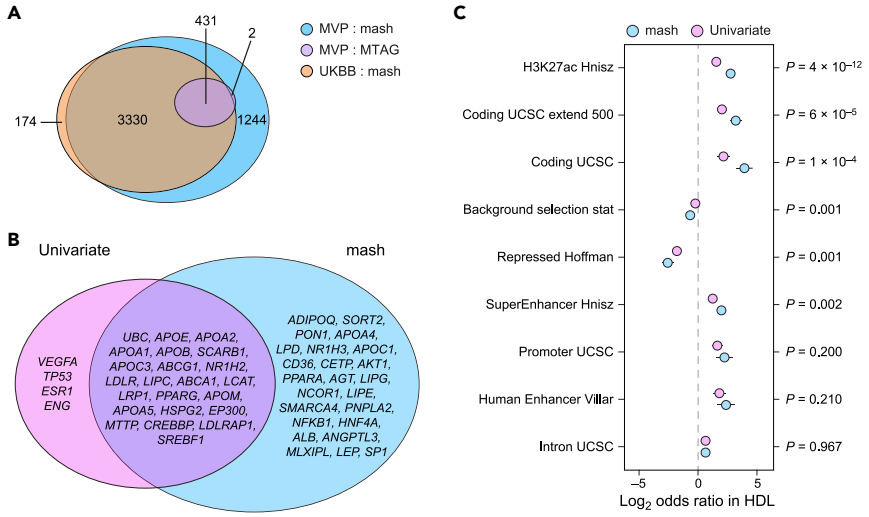


Figure 5. Bayesian multivariate method improves discovery and improves polygenic prioritization consistency of known lipid targets while enhancing known annotation estimation

(A) MVP and UKB were fit using mash separately. MTAG was fit on the MVP dataset. We delimited identical 500-kb LD blocks and computed all blocks containing at *least one variant* at an lfsr <0.05 across traits. There are 5583 blocks present in total. Hypergeometric p = 1 × 10^{-83} for replication between mash and UKBB.

(B) Mash consistently prioritizes 47 genes among LDL-C, HDL-C, and TG, while univariate methods prioritize 23. Of these, 24 are found consistently by mash but not by univariate (MLE) approach, while only 4 are found consistently by univariate approach but not mash. We use polygenic prioritization framework detailed in.[37]

(C) Using TORUS[25] we consider enrichment in 27 of the 52 classes examined by Finucane et al.[14] and see that mash versus univariate estimates tend to increase features known to be enriched in GWAS hits and decrease those known to be depleted (p values for difference in the plot). We display for HDL-cholesterol (LDL-C, **TG**, and **TC** in Figures S5–S7; Table S5).GWAS = genome-wide association study, HDL-C, high-density lipoprotein cholesterol; LDL-C, low-density lipoprotein cholesterol; mash, multivariate adaptive shrinkage; TG, triglycerides; TC, total cholesterol; MVP:mash, Million Veterans Program data analyzed using mash; MVP:uni, Million Veterans Program Data analyzed using traditional GWAS univariate analysis; UKB:mash, UK Biobank data analyzed using mash; UKBB:uni, UK Biobank data analyzed using traditional GWAS univariate analysis; MVP:MTAG, Million Veterans Program Data analyzed using MTAG.[19]

Compared to univariate summary statistics, we found that marginal estimates from posterior means supplied by mash better prioritized known lipid candidate genes more consistently between traits. Among the top 100 prioritized genes for each trait, there were 47 intersections among HDL-C, LDL-C, and TG using mash and only 27 using raw univariate estimates. Of the 47 genes, only 48% (23) were shared with those prioritized by the univariate approach (Figure 5B; Table S2). Consequently, 24 genes consistently identified by mash across all four lipid traits — but not by univariate methods — including *LPL* (lipoprotein lipase), *CETP* (cholesterol ester transfer protein), *APOA4* (apolipoprotein A4), *LIPG* (endothelial lipase) and *ADIPOQ* (adiponectin precursor). These genes have established relevance to lipids in model systems and human studies. However, the four genes consistently identified by univariate methods but not by multivariate methods were *VEGFA*, *TP53*, *ESR1*, and *ENG*. These genes are not currently known to robustly influence lipids in model systems or human studies in an interrogation of the Jackson lab phenotypic mutation database. When comparing a list of known Mendelian lipid genes,[24] the median rank assigned using mash results was consistently higher than when using univariate association summary statistics (Tables S3 and S4). We also demonstrate boxplots of these genes showing in general higher prioritization of known lipid candidates using mash over univariate approaches (Figures S11 and S12). Recall that lower ranks indicate improved prioritization. Thus, while critics may argue an overzealous estimate of associated features, the shrinkage of error within mash refutes erroneous univariate estimates and strengthens biological conclusions. The consistency of mash and univariate methods reinforce the reliability of our multivariate approach.

We further compare polygenic prioritization with mash to one using MTAG inputs (Figure 6). When comparing the top 100 genes prioritized by each model per trait (Figures 6A–6C), we show that mash and MTAG share many of the same lipid-specific candidates. However, mash prioritizes some known lipid genes in LDL (*APOA4*), HDL (*LIPG*), and TG (*LIPC*) that MTAG does not, and mash consistently prioritizes 47 among all three while MTAG prioritizes only 37 (Figure S8).

Improved enrichment for functional annotations

Finally, we sought to determine whether mash discoveries prioritize biologically relevant associations better than those observed from conventional univariate methods. Using TORUS[25] and conservatively defined LD blocks,[22] we showed that using mash compared to univariate statistics for this analysis improved the estimation of expected enrichment parameters and depletion when applying a list of previously described annotations.[14] For example, we demonstrated that the areas of the genome known to be enriched for transcriptional activity (i.e., super-enhancers) and promoters showed stronger log odds ratio (base 2) of enrichment using mash-derived summary statistics compared to the estimation from using univariate association statistics to assess annotation enrichment parameters

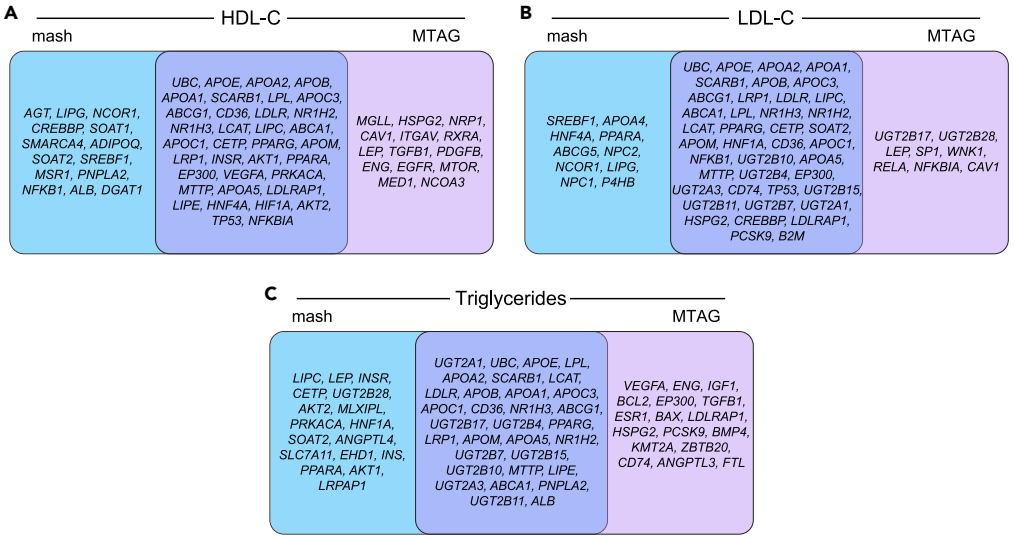


Figure 6. Performance of polygenic prioritization using MTAG and mash

(A–C) Above, we use mash or MTAG summary effect sizes for 11.8 M variants from the Millions Veterans Project (N = 330K) as inputs to PoPS polygenic prioritization and return the top 50 ranked genes in HDL, LDL and TG (A,B,C). HDL-C, HDL cholesterol; LDL-C, LDL-cholesterol; TG, Triglycerides. Full list available in Table S2B mash, multivariate adaptive shrinkage; MLE, maximum likelihood estimate; MTAG, multi-trait analysis of GWAS.

(Figure 5C; Figures S5–S7; Table S5 and S6). Similarly, annotations with prior evidence for depletion (i.e., repressors, background selection) showed a greater degree of depletion using multivariate summary statistics, as evidenced by background selection and repressors, which are repressed to a greater degree ($p = 0.001$), and coding regions which are enriched to a greater degree using mash ($p = 6 \times 10^{-5}$; 1×10^{-4}, see Figure 5C) compared to using univariate association statistics. The direction of enrichment or depletion was preserved between univariate and mash multivariate estimates. We also compare to results using MTAG as inputs to the same framework and show consistency in direction with slight though non-significant improvements using mash as inputs (Figure S10).

Mash exceeds MTAG in power and accuracy

Detailed methodological work established mash superior to both univariate shrinkage and multivariate configuration based[6] shrinkage approaches. We design two simulations: in the first, we simulate true effects using the empirical covariance matrix from our lipid analysis and only 1% of effects are non-null. We use a shared structured approach according to these empirical covariance matrices (Figure S9). We use a conservatively specified residual (error) correlation coefficient of 0.8 to emulate a noisy GWAS setting where participants may be shared among conditions. Under these conditions, mash exceeds univariate (ash) and joint (eQTL BMA) Bayesian shrinkage approaches. In this setting, effects are broadly shared (non-null units have an effect in many conditions), and "structured"— that is, similar in size and direction, with greater similarity among some subsets of conditions. The empirical covariance matrix from the model estimation generates these effects as in[4] but with additional sparsity and stronger correlation of residuals. As discussed in we note that mash uses two distinct strategies to improve the accuracy of effect estimates by shrinking estimates toward zero, which improves average accuracy because most effects are null; and in the presence of "structured effects", it shares information across conditions to improve accuracy. In our setting, we borrow information across lipid traits which are known to have high between trait heritability.[10,20,26]

Next, we simulated a GWAS case in which the 1.3 M HapMap3 SNPS are assessed in association with four simulated traits as described in our detailed STAR methods section (see STAR methods). The simulation is based on the simulated EUR genotype data provide by Zhang et al.[27] In brief, we randomly sampled 1000 of the 1.3 million HapMap3 SNPs as causal SNPs to conservatively emulate the lipid setting. We assumed that the causal SNPs were shared across 5 traits and their per-allele effect size followed multivariate normal distribution while the non-causal SNPs had zero effect. In this setting, all 1000 (<0.01% of total) true effects are correlated with the master trait ('trait 1') with ρ_g of 0.70. Furthermore, the cross-trait heritability is specified at 0.60, and there are 1000 causal SNPS to conservatively emulate the lipid setting. Under these conditions, mash exceeds MTAG in both power as measured by AUC (Figure 7A) and accuracy as measured by Root Mean Squared Error (Figure 7B) and when compared to maximum likelihood estimates. Interestingly, MTAG shows little improvement over maximum likelihood estimates.

DISCUSSION

Understanding the genetic basis of common disease is a crucial paradigm toward meeting modern genomic medicine goals. The principal strategies for improving power for discovery hinge on large sample sizes and the inclusion of diverse ancestries to analyze new alleles. Here, we introduce the application of a Bayesian multivariate approach, 'mash', for enabling improved discovery and effect size estimation of

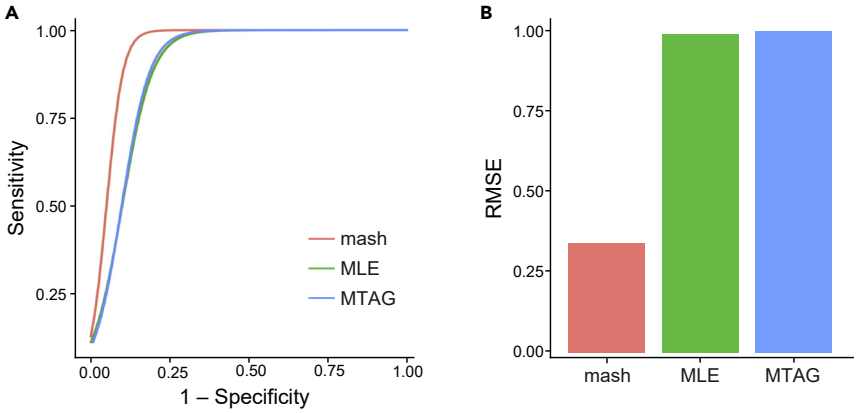


Figure 7. Mash exceeds existing multivariate method MTAG in simulated framework

(A) Here, we simulate 1.3 million HapMap3 SNPS with genome-wide heritability of 0.6 across four traits. In this setting, the 1000 causal SNPS are shared identically by all traits, while the effect sizes have a between trait correlation of 0.7 with the main trait. Under these conditions, we estimate the tradeoff in True Positives versus False Positives for a given threshold. The empirical True Positive (sensitivity) and False Positive (1- specificity) are plotted along the x axis in (A).

(B) We display the root mean squared error for all effects, defined as $\boldsymbol{RMSE} = \sqrt{(\theta - \hat{\boldsymbol{\theta}})^2}$ **where here** $\boldsymbol{\theta}$ **represents the true effect**. The simulation is intentionally sparse to replace a GWAS instance with less than 0.001% causal effects. Please see detailed STAR Methods section for further details. mash, multivariate adaptive shrinkage; MLE, maximum likelihood estimate; MTAG, multi-trait analysis of GWAS.

genomic variants without increasing the sample size. Not only do such joint approaches share information across traits to improve power, but critically, we show that robust effect size estimates enable more precise prioritization of genomic targets, enhance assessment of enrichment parameters, and improve prediction on a polygenic basis when coupled with methods to shrink across LD blocks. There has been much work on the utility of summary statistics in both fine-mapping and prediction.[5,28,29] Namely, working with two numbers $\widehat{B}$ and $\widehat{se}$ rather than simply *p* or *Z*, can yield substantial gains in functionality while providing a convenient estimation framework instead of only testing.

False discovery rates offer a flexible way of capturing inherent differences in the relative signal between populations while controlling the proportion of discoveries that are false. Local false discovery rates are thus obviate arbitrarily stringent p values by allowing one to include the prior probability of absence of signal (often termed p_0) in computing the posterior probability of being null and are widely accepted in the genomics community.[17,30,31] The local false sign rate is analogous to the "local false discovery rate"(lfdr), but measures confidence in the *sign* of each and confidence in each effect being non-zero, and is thus more stringent.[4]

Importantly, we use the local false sign rate to characterize effects, a convenient method of controlling for multiple hypotheses while also incorporating the consistency in sign and magnitude of effects. It is more stringent than the lfdr because it requires actual discoveries to be not only nonzero, but also correctly signed. The local false discovery rate is well known to conservatively control for false discoveries in genomic applications.[5,32,33] Furthermore, in some settings with many discoveries, the lfsr and lfdr can be quite different and emphasize the benefits of the lfsr, particularly its increased robustness to modeling assumptions.[5] This allows for a multiple hypotheses correction that incorporates the precision of the effect size estimate[5] beyond confidence in the binary classification of a variant as associated or unassociated.

While assembling the evidence, we found that considering the non-null distribution of SNPs provided a much broader understanding of the collective contribution of genetic variation to quantitative phenotype mapping. Moreover, the inclusion of local false sign rates broadened this resolution as a non-null SNP can have varying effect sizes (Figure 2C). Our multivariate mapping enhanced the resolution, showing that two SNPs with the same local false sign rate can have different posterior effect estimates arising from the mixture of multivariate normal distributions that depend on the 'boost' the SNP receives from the effects in correlated phenotypes. More generally, the heritability of complex traits and diseases is spread broadly across the genome,[34] implying that a significant fraction of all genes contribute to variation in quantitative traits. Then, we combined this increased power from false discovery control with the precision and adaptive nature of estimating effects *across* conditions, presumably adding to the increase in power. We show that these effects are biologically believable. To summarize, we improved our power to both predict and detect through i) control of false discovery instead of family-wise error rate, and ii) incorporation of additional information captured by between-trait sharing of effect-size information.

While well-suited to genetic settings in which effects are additive, multivariate normal methods are limited by settings in which the effects are roughly normally distributed in each trait. Furthermore, the benefit of such a method is stronger when the effects are more strongly correlated than the errors. Perhaps most notably, after estimating effects, correction still must be done for LD as such a method does not currently consider the correlation between SNPs. However, such work must also be done after univariate GWAS estimation. Herein, we provide a framework to do so with available LD tools.[8]

Limitations of the study

Here, we demonstrate that the use of joint estimation of effects can enhance power for prediction, prioritization, and discovery. However, the limits of such an approach are as follows: First, the use of local false discovery cannot be interpreted in the same context as family-wise error

rate control and should be thought of in the context of describing effect size estimation rather than limiting the probability of observing at least one false positive. Second, in reporting replication, we note that we chose conservatively estimated 500-kb blocks to define an area of approximate linkage disequilibrium; however, the choice of large blocks means some amount of replication is expected by chance. To overcome those challenges, we use a hypergeometric test to assess the significance of observing a defined number of replicated variants in both the UK Biobank and MVP using mash. Finally, in this paper we report the results using LDpred2 as a method for PRS estimation and focus particularly on the results under their infinitesimal model. However, this approach can be extended to a setting with alternative approaches to PRS development, and we hope this will motivate future work to explore the use of joint methods for discovery in combination with methods of LD shrinkage for prediction.

In conclusion, applying these methods demonstrates the significant promise of multivariate approaches for GWAS of complex traits and the improvement gained in both prediction and prioritization through control of false discovery of multivariate adaptively shrunk effect size estimates.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Study participants
  - Input use as summary statistics
  - Model and fitting of mash model
  - Gene enrichment
  - Gene prioritization
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Polygenic score prediction
  - General workflow as follows
  - Simulations
  - "Shared, structured effects" simulations
- ADDITIONAL RESOURCES

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.107854.

ACKNOWLEDGMENTS

The authors would like to thank the participants and staff of the Million Veterans Program and the UK Biobank. The UK Biobank analyses were supported by application 7089.

Funding: This work was primarily supported by grants to GMP and PN from the National Heart, Lung, and Blood Institute (R01HL142711, R01HL127564). PN is also supported by additional grants from the National Heart, Lung, and Blood Institute (R01HL148050, R01HL151283, R01HL148565, R01HL151152), National Institute of Diabetes and Digestive and Kidney Diseases (R01DK125782), Foundation Leducq (TNE-18CVD04), and Massachusetts General Hospital. SMU is supported by NIH NHGRI T32 (#1T32HG010464).

AUTHOR CONTRIBUTIONS

Conceptualization: S.M.U., P.N., G.M.P., and C.J.O.
Methodology: S.M.U., G.M.P., and P.N.
Software: S.M.U.
Investigation: S.M.U., G.M.P., and P.N.
Validation: S.M.U., B.T., and R.B.
Data Curation: S.M.U., B.T., R.B., and M.S.
Visualization: S.M.U., S.K., R.H., B.T., P.N., and S.K.
Funding acquisition: P.N. and C.J.O.
Project administration: P.N.
Supervision: P.N., C.J.O., G.M.P., and B.N.
Writing – original draft: S.M.U. and P.N.

Writing – review & editing: S.M.U., W.H., P.N., C.J.O., and G.M.P.
Funding Acquisition: P.N. and W.H.

DECLARATION OF INTERESTS

C.J.O. is an employee of Novartis. P.N. reports research grants from Allelica, Apple, Amgen, Boston Scientific, Genentech/Roche, and Novartis, personal fees from Allelica, Apple, AstraZeneca, Blackstone Life Sciences, Eli Lilly & Co, Foresite Labs, Genentech/Roche, GV, HeartFlow, Magnet Biomedicine, and Novartis, scientific advisory board membership of Esperion Therapeutics, Preciseli, and TenSixteen Bio, scientific co-founder of TenSixteen Bio, equity in MyOme, Preciseli, and TenSixteen Bio, and spousal employment at Vertex Pharmaceuticals, all unrelated to the present work.

INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research.

Received: January 10, 2023
Revised: May 15, 2023
Accepted: September 5, 2023
Published: September 9, 2023

REFERENCES

1. Zuk, O., Hechter, E., Sunyaev, S.R., and Lander, E.S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. Proc. Natl. Acad. Sci. USA *109*, 1193–1198. https://doi.org/10.1073/pnas.1119675109.
2. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature *461*, 747–753.
3. Zhu, X., and Stephens, M. (2016). Bayesian Large-Scale Multiple Regression with Summary Statistics from Genome-wide Association Studies. Preprint at bioRxiv. https://doi.org/10.1101/042457.
4. Urbut, S.M., Wang, G., Carbonetto, P., and Stephens, M. (2019). Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. Nat. Genet. *51*, 187–195. https://doi.org/10.1038/s41588-018-0268-8.
5. Stephens, M. (2017). False discovery rates: a new deal. Biostatistics *18*, 275–294. kxw041. https://doi.org/10.1093/biostatistics/kxw041.
6. Flutre, T., Wen, X., Pritchard, J., and Stephens, M. (2013). A Statistical Framework for Joint eQTL Analysis in Multiple Tissues. PLoS Genet. *9*, e1003486. https://doi.org/10.1371/journal.pgen.1003486.
7. Stephens, M., and Balding, D.J. (2009). Bayesian statistical methods for genetic association studies. Nat. Rev. Genet. *10*, 681–690. https://doi.org/10.1038/nrg2615.
8. Privé, F., Arbel, J., and Vilhjálmsson, B.J. (2021). LDpred2: better, faster, stronger. Bioinformatics *36*, 5424–5431. https://doi.org/10.1093/bioinformatics/btaa1029.
9. Nowbar, A.N., Gitto, M., Howard, J.P., Francis, D.P., and Al-Lamee, R. (2019). Mortality From Ischemic Heart Disease: Analysis of Data From the World Health Organization and Coronary Artery Disease Risk Factors From NCD Risk Factor Collaboration. Circ. Cardiovasc. Qual. Outcomes *12*, e005375. https://doi.org/10.1161/CIRCOUTCOMES.118.005375.
10. Klarin, D., Damrauer, S.M., Cho, K., Sun, Y.V., Teslovich, T.M., Honerlaw, J., Gagnon, D.R., DuVall, S.L., Li, J., Peloso, G.M., et al. (2018). Genetics of Blood Lipids Among ~300,000 Multi-Ethnic Participants of the Million Veteran Program. Nat. Genet. *50*, 1514–1523. https://doi.org/10.1038/s41588-018-0222-9.
11. Peloso, G.M., and Natarajan, P. (2018). Insights from population-based analyses of plasma lipids across the allele frequency spectrum. Curr. Opin. Genet. Dev. *50*, 1–6. https://doi.org/10.1016/j.gde.2018.01.003.
12. Clarke, L., Fairley, S., Zheng-Bradley, X., Streeter, I., Perry, E., Lowy, E., Tassé, A.M., and Flicek, P. (2017). The international Genome sample resource (IGSR): A worldwide collection of genome variation incorporating the 1000 Genomes Project data. Nucleic Acids Res. *45*, D854–D859. https://doi.org/10.1093/nar/gkw829.
13. International HapMap Consortium (2005). A haplotype map of the human genome. Nature *437*, 1299–1320. https://doi.org/10.1038/nature04226.
14. Bulik-Sullivan, B.K., Loh, P.-R., Finucane, H.K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics Consortium, Patterson, N., Daly, M.J., Price, A.L., and Neale, B.M. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet. *47*, 291–295. https://doi.org/10.1038/ng.3211.
15. Ni, G., Moser, G., Wray, N.R., and Lee, S.H., Schizophrenia Working Group of the Psychiatric Genomics Consortium (2018). Estimation of genetic correlation via linkage disequilibrium score regression and genomic restricted maximum likelihood. Am. J. Hum. Genet. *102*, 1185–1194. https://doi.org/10.1016/j.ajhg.2018.03.021.
16. Boyle, E.A., Li, Y.I., and Pritchard, J.K. (2017). An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell *169*, 1177–1186. https://doi.org/10.1016/j.cell.2017.05.038.
17. EPIC-CVD Consortium, CARDIoGRAMplusC4D, The UK Biobank CardioMetabolic Consortium CHD working group, Nelson, C.P., Goel, A., Butterworth, A.S., Kanoni, S., Webb, T.R., Marouli, E., Zeng, L., et al. (2017). Association analyses based on false discovery rate implicate new loci for coronary artery disease. Nat. Genet. *49*, 1385–1391. https://doi.org/10.1038/ng.3913.
18. Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C.A., and Smoller, J.W. (2019). Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nat. Commun. *10*, 1776. https://doi.org/10.1038/s41467-019-09718-5.
19. Turley, P., Walters, R.K., Maghzian, O., Okbay, A., Lee, J.J., Fontana, M.A., Nguyen-Viet, T.A., Wedow, R., Zacher, M., Furlotte, N.A., et al. (2018). Multi-trait analysis of genome-wide association summary statistics using MTAG. Nat. Genet. *50*, 229–237. https://doi.org/10.1038/s41588-017-0009-4.
20. Ye, Y., Chen, X., Han, J., Jiang, W., Natarajan, P., and Zhao, H. (2021). Interactions Between Enhanced Polygenic Risk Scores and Lifestyle for Cardiovascular Disease, Diabetes, and Lipid Levels. Circ. Genom. Precis. Med. *14*, e003128. https://doi.org/10.1161/CIRCGEN.120.003128.
21. Pickrell, J.K. (2014). Joint Analysis of Functional Genomic Data and Genome-wide Association Studies of 18 Human Traits. Am. J. Hum. Genet. *94*, 559–573. https://doi.org/10.1016/j.ajhg.2014.03.004.
22. Berisa, T., and Pickrell, J.K. (2016). Approximately independent linkage disequilibrium blocks in human populations. Bioinformatics *32*, 283–285. btv546. https://doi.org/10.1093/bioinformatics/btv546.
23. Weeks, E.M., Ulirsch, J.C., Cheng, N.Y., Trippe, B.L., Fine, R.S., Miao, J., Patwardhan, T.A., Kanai, M., Nasser, J., Fulco, C.P., et al. (2023). Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases. Nat. Genet. *55*, 1267–1276. https://doi.org/10.1038/s41588-023-01443-6.
24. Natarajan, P., Peloso, G.M., Zekavat, S.M., Montasser, M., Ganna, A., Chaffin, M., Khera, A.V., Zhou, W., Bloom, J.M., Engreitz, J.M., et al. (2018). Deep-coverage whole genome sequences and blood lipids among 16,324 individuals. Nat. Commun. *9*, 3391. https://doi.org/10.1038/s41467-018-05747-8.
25. Wen, X. (2016). Molecular QTL discovery incorporating genomic annotations using Bayesian false discovery rate control. Ann.

Appl. Stat. *10*. https://doi.org/10.1214/16-AOAS952.

26. Goode, E.L., Cherny, S.S., Christian, J.C., Jarvik, G.P., and de Andrade, M. (2007). Heritability of longitudinal measures of body mass index and lipid and lipoprotein levels in aging twins. Twin Res. Hum. Genet. *10*, 703–711. https://doi.org/10.1375/twin.10.5.703.
27. Zhang, H., Zhan, J., Jin, J., Zhang, J., Lu, W., Zhao, R., Ahearn, T.U., Yu, Z., O'Connell, J., Jiang, Y., et al. (2023). Novel Methods for Multi-Ancestry Polygenic Prediction and Their Evaluations in 5.1 Million Individuals of Diverse Ancestry. Preprint at bioRxiv. https://doi.org/10.1101/2022.03.24.485519.
28. Wang, G., Sarkar, A., Carbonetto, P., and Stephens, M. (2020). A simple new approach to variable selection in regression, with application to genetic fine mapping. J. R. Stat. Soc. Series B Stat. Methodol. *82*, 1273–1300. https://doi.org/10.1111/rssb.12388.
29. Zhu, X., and Stephens, M. (2017). Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. Ann. Appl. Stat. *11*, 1561–1592. https://doi.org/10.1214/17-AOAS1046.
30. Storey, J.D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. Proc. Natl. Acad. Sci. USA *100*, 9440–9445. https://doi.org/10.1073/pnas.1530509100.
31. Storey, J.D. (2003). The positive false discovery rate: a Bayesian interpretation and the q-value. Ann. Stat. *31*, 2013–2035. https://doi.org/10.1214/aos/1074290335.
32. Brzyski, D., Peterson, C.B., Sobczyk, P., Candès, E.J., Bogdan, M., and Sabatti, C. (2017). Controlling the Rate of GWAS False Discoveries. Genetics *205*, 61–75. https://doi.org/10.1534/genetics.116.193987.
33. Sesia, M., Bates, S., Candès, E., Marchini, J., and Sabatti, C. (2021). False discovery rate control in genome-wide association studies with population structure. Proc. Natl. Acad. Sci. USA *118*, e2105841118. https://doi.org/10.1073/pnas.2105841118.
34. Loh, P.R., Bhatia, G., Gusev, A., Finucane, H.K., Bulik-Sullivan, B.K., Pollack, S.J., Schizophrenia Working Group of Psychiatric Genomics Consortium, de Candia, T.R., Lee, S.H., Wray, N.R., et al. (2015). Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. Nat. Genet. *47*, 1385–1392. https://doi.org/10.1038/ng.3431.
35. Urbut, S.M., Wang, G., Carbonetto, P., and Stephens, M. (2019). Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. Nat. Genet. *51*, 187–195.
36. Efron, B. (2007). Size, power and false discovery rates. Ann. Stat. *35*, 1351–1377. https://doi.org/10.1214/009053606000001460.
37. Weeks, E.M., Ulirsch, J.C., Cheng, N.Y., Trippe, B.L., Fine, R.S., Miao, J., Patwardhan, T.A., Kanai, M., Nasser, J., Fulco, C.P., et al. (2020). Leveraging Polygenic Enrichments of Gene Features to Predict Genes Underlying Complex Traits and Diseases (Genetic and Genomic Medicine). Nat. Genet. *55*, 1267–1276. https://doi.org/10.1101/2020.09.08.20190561.
38. Gaziano, J.M., Concato, J., Brophy, M., Fiore, L., Pyarajan, S., Breeling, J., Whitbourne, S., Deen, J., Shannon, C., Humphries, D., et al. (2016). Million Veteran Program: A mega-biobank to study genetic influences on health and disease. J. Clin. Epidemiol. *70*, 214–223. https://doi.org/10.1016/j.jclinepi.2015.09.016.
39. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209. https://doi.org/10.1038/s41586-018-0579-z.
40. Peloso, G.M., Auer, P.L., Bis, J.C., Voorman, A., Morrison, A.C., Stitziel, N.O., Brody, J.A., Khetarpal, S.A., Crosby, J.R., Fornage, M., et al. (2014). Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. Am. J. Hum. Genet. *94*, 223–232. https://doi.org/10.1016/j.ajhg.2014.01.009.
41. Bovy, J., Hogg, D.W., and Roweis, S.T. (2011). Extreme deconvolution: Inferring complete distribution functions from noisy, heterogeneous and incomplete observations. Ann. Appl. Stat. *5*, 1657–1677. https://doi.org/10.1214/10-AOAS439.
42. Wen, X., and Stephens, M. (2014). Bayesian methods for genetic association analysis with heterogeneous subgroups: From meta-analyses to gene–environment interactions. Ann. Appl. Stat. *8*, 176–203. https://doi.org/10.1214/13-AOAS695.
43. Finucane, H.K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.R., Anttila, V., Xu, H., Zang, C., Farh, K., et al. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet. *47*, 1228–1235. https://doi.org/10.1038/ng.3404.
44. Qiao, M., and Wang. (2020). TORUS Workflow (Version 2.0) [Source code]. Gao. https://github.com/cumc/bioworkflows/blob/master/fine-mapping/gwas_enrichment.ipynb.
45. Weeks, Elle and Finucane, Hilary (2020). Polygenic Priority Score (v0.2). https://github.com/FinucaneLab/pops.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
MVP Summary Statistics[38]	dbGAP	phs002453.v1.p1.c999
UK Biobank[39]	UK Biobank	Application 7089, Basket 2008463
Software and algorithms		
mashR[4]	Urbut et al.[4]	https://cran.r-project.org/web/packages/mashr/vignettes/intro_mash.html
Polygenic Prioritization Software	Weeks et al.[22]	https://github.com/FinucaneLab/pops
LDPred2	Prive et al.[8]	https://privefl.github.io/bigsnpr/articles/LDpred2.html
R		https://www.r-project.org
TORUS	Wen et al.[25]	https://github.com/xqwen/torus
Other		
Online methods for rerunning mash with available summary statistics and displaying all covariance matrices.	N/A	https://broadinstitute.github.io/natarajanlab_wiki/MVP-mfit.html
Worklow for running LDpred2 with MVP summary statistics using SoS pipeline for simple univariate case (template for adding analogous mash summary stats for additional traits)	N/A	" \o "https://broadinstitute.github.io/natarajanlab_wiki/hdl_univariate.htmlhttps://broadinstitute.github.io/natarajanlab_wiki/hdl_univariate.html https://broadinstitute.github.io/natarajanlab_wiki/hdl_mash.html
Online methods for producing Torus metaplots:	N/A	https://broadinstitute.github.io/natarajanlab_wiki/metaplot_strat_torus.html
Online methods for reproducing Venn diagrams:	N/A	https://broadinstitute.github.io/natarajanlab_wiki/venn_diagrams.html
Ranking of all genes by mashR and MLE via shiny app:	N/A	https://surbut.shinyapps.io/testgenelist/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Pradeep Natarajan (pnatarajan@mgh.harvard.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Data

The input summary statistics and corresponding posterior means effect size estimates generated in the analyses for the Million Veterans Project and UK Biobank reported in this study cannot be deposited in a public repository because it is available with appropriate IRB access. To request access, contact dbGAP and the UK Biobank for details. Accession numbers or DOIs are listed in the key resources table. The polygenic risk score weights have been deposited in the PGS catalog and are available upon request from the lead author.

- Code

All original code has been deposited at broadinstitute.github.io/natarajan_lab/index.html and is publicly available as of the date of publication.

- Other items.

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Study participants

Association testing was performed in up to 297,626 white (European ancestry), black (African ancestry), and Hispanic Million Veterans Program (MVP) participants with blood lipids stratified by ethnicity followed by a meta-analysis of results across all three groups as previously described.[10] Samples were imputed to the 1000 Genomes project p3v5 reference panel (b37), and ancestry specific Hardy-Weinberg equilibrium P < 1x10-20, posterior call probability <0.9, imputation quality/INFO <0.3, minor allele frequency (MAF) < 0.0003, call rate <97.5% for common variants (MAF >1%), and call rate <99% for rare variants (MAF <1%) were used for QC. Variants were also excluded if they deviated >10% from their expected allele frequency based on reference data from the 1000 Genomes Project. *Trans*-ethnic meta-analysis of white (European ancestry), black (African ancestry), and Hispanic MVP participants for 291,933 and 297,626 people was produced for inverse normal transformed HDL cholesterol, Inverse normal transformed LDL cholesterol, Inverse normal transformed triglyceride levels and inverse normal transformed total cholesterol.[38]

UK Biobank samples[39] were genotyped on either the UK BiLEVE or UK Biobank Axiom arrays and imputed into the Haplotype Reference Consortium panel and the UK10K+1000 Genomes panel. Variant positions were keyed to the GRCh37 human genome reference. Genotyped variants with genotyping call rate <0.95 and imputed variants with INFO score <0.3 or minor allele frequency $\leq$ 0.005 in the analyzed samples[36] were excluded. After variant-level quality control, 11,622,901 imputed variants remained for analysis Lipid levels were collected on the Beckman Coulter AU5800 Platform and were adjusted for cholesterol medication.[40] Participants without imputed genetic data, or with a genotyping call rate <0.98, mismatch between self-reported sex and sex chromosome count, sex chromosome aneuploidy, excessive third-degree relatives, or outliers for heterozygosity were excluded from genetic analysis.[39] IRB approval was obtained by an institutional review committee in accordance with the principles outlined by *iScience*.

Input use as summary statistics

There has been much work on the utility of summary statistics in both fine-mapping and prediction.[5,28,29] Namely, working with two numbers $\widehat{B}$ and $\widehat{se}$ rather than simply p or Z, can yield substantial gains in functionality while providing a convenient framework for estimation as opposed to only testing. It is not the goal of this paper to summarize the substantial body of literature on summary statistics.

Model and fitting of mash model

Using the procedure outlined in Urbut et al.,[4] we first generated data-driven covariance matrices U_k. We first identified the rows of the **M** SNPs by **R** traits matrix $\widehat{\boldsymbol{Z}}$ of Z statistics across traits that were likely have an effect in at least one condition. In the MVP data, we chose rows corresponding to the "top" SNP for each of the 1703 conservatively defined LD blocks specified in[22] as the SNP with the strongest absolute observed Z statistic across traits from the matrix of Z statistics, which we defined to be the SNP with the highest value of $Z_j^{max} := \max_r \widehat{b}_{jr}/\widehat{s}_{jr}$. We used the maximum, rather than the sum because we wanted to include effects that were strong in a single lipid trait rather than effects that were shared among all lipids. The matrix of residual errors, $\widehat{V}$, as articulated here[4]: namely, by estimating the empirical covariance (correlation) matrix of the smallest (by absolute value) z statisticis present in the dataset, namely those with an absolute value of |Z-statistic| <2. The intutition is that:

$$\widehat{B} = B + E$$

$$\widehat{B} \sim N(0, \widehat{V} + U_k)$$

Hence truly null effects in which B = 0 arise solely from the covariance of errors. In our simulation based on the initial framework of mash[4] (Figure S9), we liberally simulate a setting in which the covariance of residuals is 0.8 between traits. Next, we fitted a mixture of MVN distributions to these strongest effects using methods from Bovy et al.[41] We used a list of 16 matrices (i.e., K = 16) in this application that incorporated 7 data driven matrices in addition to the identity and four canonical 'trait specific' matrices as well as matrix for equal effects and three matrices with varying levels of fixed heterogeneity.[4,6,42] Next, we expanded by a grid of 21 scaling factors (i.e., L = 21) ranging from $\omega^2 = 0.07$ to $\omega^2 = 72.43$ as specified by the range of observed 'noisy' effects in a training set of 40,000 SNPS.

Then we repeated this model fitting procedure with the UKBB data, choosing a new set of 'maxes' and computing weights on a training set from UKBB.

We computed posterior means and local false sign rates for 28,686,877 SNPs in the MVP databases and 13,788,619 million SNPs in the UKBB which overlapped at 11,1577,790 loci. Detailed calculations are described in (4).

Gene enrichment

TORUS[25] uses a hierarchical Bayes model to estimate the enrichment of a variety of genomic annotations in QTL and in turn assigns a posterior probability of inclusion to the 'causal' nature of the corresponding loci. We used the assigned posterior mean expected Z scores from mash

and corresponding univariate Z statistics as inputs to the TORUS model and observed the estimation of log OR of an annotation parameter being enriched (depleted) in QTL. This package considers the strongest loci per block, and we used the 1703 blocks as conservatively defined by Berisa et al.[22] Furthermore, we used the set of annotation parameters available through the Price lab.[43] We provide a link to the commands to run appropriate pipeline[44] and corresponding annotation file.

Gene prioritization

In order to rank genes based on their strength of effects as demonstrated by associated SNPs and consequent feature enrichment, The Polygenic Gene Prioritization (PoPs) algorithm works in three steps.[37] Recall that lower ranks indicate higher priority.

1. Use magma to assign gene association statistics based on significance of associated SNPs.
2. Select marginally associated features by performing enrichment analysis for each gene feature separately.
3. Estimate Polygenic Priority Scores[23] (PoPs) by fitting a joint model for the enrichment of all selected features using 'leave one out' regression.

Here, we used the local false sign rate (lfsr) and p value for all SNPs considered by the MVP project for our mash and univariate analyses, respectively. We used the 1000 Genomes human reference to specify LD among variants per the PoPS protocol, the MAGMA list of genetic annotations to map genes to respective loci, and control and PoPS features from the available packages as detailed within the PoPS package and accompanying software.[45] We then ranked genes in order of prioritization score and compared mash estimates to univariate assessments. Recall that lower ranks indicate higher priority.

QUANTIFICATION AND STATISTICAL ANALYSIS

Polygenic score prediction

We used the LDpred2[8] infinitesimal sites model to control for this genetic correlation. In this model, all markers are causal (p = 1), and effects are drawn from a Gaussian distribution, i.e.,

$$B_{jr} \sim N\left(0, \frac{h_r^2}{M}\right) \quad \text{(Equation 2)}$$

where h^2 represents the heritability of the trait and M the number of markers. The posterior mean can be derived analytically.

$$E\left(B_{jr} \middle| \widehat{B}\right\}_{jr}, D \approx \left(\frac{M}{Nh^2} I + D\right)^{-1} \widehat{B}_{jr} \quad \text{(Equation 3)}$$

Here, $\widehat{B}_{jr}$ is the ordinary least square estimate from a univariate GWAS analysis. Here, D denotes the LD matrix between the markers obtained from an outside reference panel, in our case the 1000 Genomes EUR dataset.

In a typical analysis, one uses $\widehat{B}_{jr}$ that are estimated from a univariate GWAS analysis in each trait separately. However, we replace this with the posterior means on a per trait basis from mash. The β_{jr} are each the marginal output of the posterior mean arising from the mixture normal cited in[4] which weight the importance of each prior distribution shape and scale by the probability density of the observed effect under that distribution with the observed mean and standard error. The posterior mean is thus a weighted average of the observed 'noisy' estimate and the prior mean, in this case a mixture of unimodal distributions centered at 0.

As noted we use the exchangeable Z-statistic model in which we assume the standard errors $\widehat{s^2}$ are known as in,[4] these estimates are then rescaled by their standard error to compute the input summary statistic weights for rescaling in LDpred2.[8] As noted we use the exchangeable 'Z' statistic model which assumes that the effects scale with standard errors (alpha = 1 in [5]), which we find is most appropriate in genomic contexts. We use the posterior estimates of $E(\beta|(\widehat{se}\}, \widehat{\beta})$ in all downstream analyses, and given that our posterior estimates are of $E(\beta/\widehat{se}|\widehat{se}, \widehat{\beta})$ and V $(\beta/\widehat{se}|\widehat{se}, \widehat{\beta})$ we rescale these posterior estimates by the original standard error to produce posterior means for both β and its corresponding posterior standard deviation (a hyperparameter). We report the marginal 'diagonal' of the posterior covariance matrix and rescale by each conditions corresponding standard error.

General workflow as follows

1) We perform QC on reference panel (1000 genomes[12])
2) We intersect SNPs common to the reference panel (1000 genomes[12]), discovery dataset (MVP) and scoring dataset (UKBB). This left us with approximately 400K SNPs.
3) Harmonize alleles for shared direction
4) We calculate the LD matrix and fit the LDSC per LDpred2.[8]
5) We use this LD matrix and to compute the posterior weights from initial summary statistics (either arising from GWAS summary statistics or from the posterior means of mash output).
6) We compute the PRS using the infinitesimal model[8] for all individuals in UKBB.

7) We associated these scores with the phenotype of interest in a linear model that includes age and sex as additional (baseline) covariates.
8) We then divide into EUR and non-EUR individuals to assess population-specific performance. A sample vignette running it on HDL univariate summary statistics is available on github as detailed in the key resources table. We use both the inf and grid models and cross-validate on a holdout subset of 400 individuals for the grid model.

Simulations

In Figure 7, we simulate 1.3M HapMap SNPs with N_c 1000 causal SNPS (0.01%), 0.6 H^2 such that each causal SNP has an effect correlated at 0.7 with the master trait. The simulation is based on the simulated EUR genotype data provide by Zhang et al.[27] In brief, we randomly sampled 1000 of the 1.3 million HapMap3 SNPs as causal SNPs and assumed that the causal SNPs were shared across four traits and their per-allele effect size followed multivariate normal distribution and the non-causal SNPs had zero effect. The intertrait correlation was set at 0.70. For all the traits, the phenotype was generated using $y = X * \beta + e$, where X was the genotype matrix, β is the simulated per-allele effect size, e was the non-genetic component following a normal distribution of mean = 0. The standard deviation of e was set so that the heritability for all the traits was fixed at 60%. The GWAS was performed with 100k of the simulated individuals. mashR was fit according to the empirical covariance matrix and with additional dimensional reductions as above, and mtag was run as in[19] with n_min = 0.0

In Figure S9, we simulate a setting of shared structured effects in which 100 causal SNPs out of 10,000 (1%) are active with generative model arising from the empirical covariance matrix generated in Equation 1. We simulate with residual correlation matrix $\widehat{V}$ generated such that the correlation is 0.8 between traits.

"Shared, structured effects" simulations

We simulated each bj from the mixture model in Equation 1 with mixture parameters based on fit to the MVP lipid data. This is the simulation procedure in more detail.

1. We took the 8 "data-driven" covariance matrices U_1, …,U_8 learned from the MVP data, standardized as described above.
2. We simulated 100 "non-null units": independently for j = 1, …,1 00, we (a) chose a component k uniformly at random from 1, ….,8, (b) simulated a scaling factor ω as the absolute value of an N(0,1) random variable, and (c) simulated $bj \sim N4(0, \omega Uk)$.
3. For the 9,900 "null units", we set bj = 0.
4. For all 10,000 units, we simulated $\widehat{b_4} \sim N(b_j, \widehat{V_j})$ where $\widehat{Vj}$ was the diagonal matrix with diagonal elements 0.1^2 and correlation $\rho = 0.8$.

ADDITIONAL RESOURCES

Additional References for this manuscript including link to workflow code for completing analyses and producing plots available at https://broadinstitute.github.io/natarajanlab_wiki/index.html.