Manuscript Version: July 16, 2018

A LARGE-SCALE GENOME-WIDE ENRICHMENT ANALYSIS IDENTIFIES NEW TRAIT-ASSOCIATED GENES, PATHWAYS AND TISSUES ACROSS 31 HUMAN PHENOTYPES*

By Xiang Zhu and Matthew Stephens

Stanford University and University of Chicago

Genome-wide association studies (GWAS) aim to identify genetic factors that are associated with complex traits. Standard analyses test individual genetic variants, one at a time, for association with a trait. However, variant-level associations are hard to identify (because of small effects) and can be difficult to interpret biologically. "Enrichment analyses" help address both these problems by focusing on *sets of biologically-related variants*. Here we introduce a new model-based enrichment analysis method that requires only GWAS summary statistics, and has several advantages over existing methods. Applying this method to interrogate 3,913 biological pathways and 113 tissue-based gene sets in 31 human phenotypes identifies many previously-unreported enrichments. These include enrichments of the *endochondral ossification* pathway for adult height, the *NFAT-dependent transcription* pathway for rheumatoid arthritis, *brain-related* genes for coronary artery disease, and *liver-related* genes for late-onset Alzheimer's disease. A key feature of our method is that inferred enrichments automatically help identify new trait-associated genes. For example, accounting for enrichment in *lipid transport* genes yields strong evidence for association between *MTTP* and low-density lipoprotein levels, whereas conventional analyses of the same data found no significant variants near this gene.

Introduction

Genome-wide association studies (GWAS) have successfully identified many genetic variants – typically SNPs – underlying a wide range of complex traits [1–3]. GWAS are typically analyzed using "single-SNP" association tests, which assess the marginal correlation between the genotypes of each SNP and the trait of interest. This approach can work well for identifying common variants with sufficiently-large effects. However, for complex traits, most variants have small effects, making them difficult to identify even with large sample sizes [4]. Further, because many associated variants are non-coding it can be difficult to identify the biological mechanisms by which they may act.

Enrichment analysis – also referred to as "pathway" [5] or "gene set" [6] analysis – can help tackle both these problems. Instead of analyzing one

*Correspondence should be addressed to X.Z. (xiangzhu@stanford.edu) and M.S. (mstephens@uchicago.edu).

variant at a time, enrichment analysis assesses groups of related variants. The idea – borrowed from enrichment analysis of gene expression [7] – is to identify groups of biologically-related variants that are "enriched" for associations with the trait: that is, they contain a higher fraction of associated variants than would be expected by chance. By pooling information across many genetic variants this approach has the potential to detect enrichments even when individual genetic variants fail to reach a stringent significance threshold [5]. And because the sets of variants to be analyzed are often defined based on existing biological knowledge, an observed enrichment automatically suggests potentially relevant biological processes or mechanisms.

Although the idea of enrichment analysis is simple, there are many ways to implement it in practice, each with its own advantages and disadvantages. Here we build on a previous approach [8] that has several attractive features not shared by most methods. These features include: it accounts for linkage disequilibrium (LD) among associated SNPs; it assesses SNP sets for enrichment directly, without requiring intermediate steps like imposing a significance cut-off or assigning SNP-level associations to specific genes; and it can re-assess ("prioritize") variant-level associations in light of inferred enrichments to identify which genetic factors are driving the enrichment.

Despite these advantages, this approach has a major limitation: it requires individual-level genotypes and phenotypes, which are often difficult or impossible to obtain, especially for large GWAS meta analyses combining many studies. Our major contribution here is to overcome this limitation, and provide an implementation that requires only GWAS summary statistics (plus LD estimates from a suitable reference panel). This allows the method to be applied on a scale that would be otherwise impractical. Here we exploit this to perform enrichment analyses of 3,913 biological pathways and 113 tissue-based gene sets for 31 human phenotypes. Our results identify many novel pathways and tissues relevant to these phenotypes, as well as some that have been previously identified. By prioritizing variants within the enriched pathways we identify several trait-associated genes that do not reach genome-wide significance in conventional analyses of the same data. The results highlighted here demonstrate the potential for these enrichment analyses to yield novel insights from existing GWAS summary data.

RESULTS

Method overview. Figure 1 provides a schematic overview of the method. In brief, we combine the enrichment model from [8], with the multiple regression model for single-SNP association summary statistics from [9], to create a model-based enrichment method for GWAS summary data. We refer to the

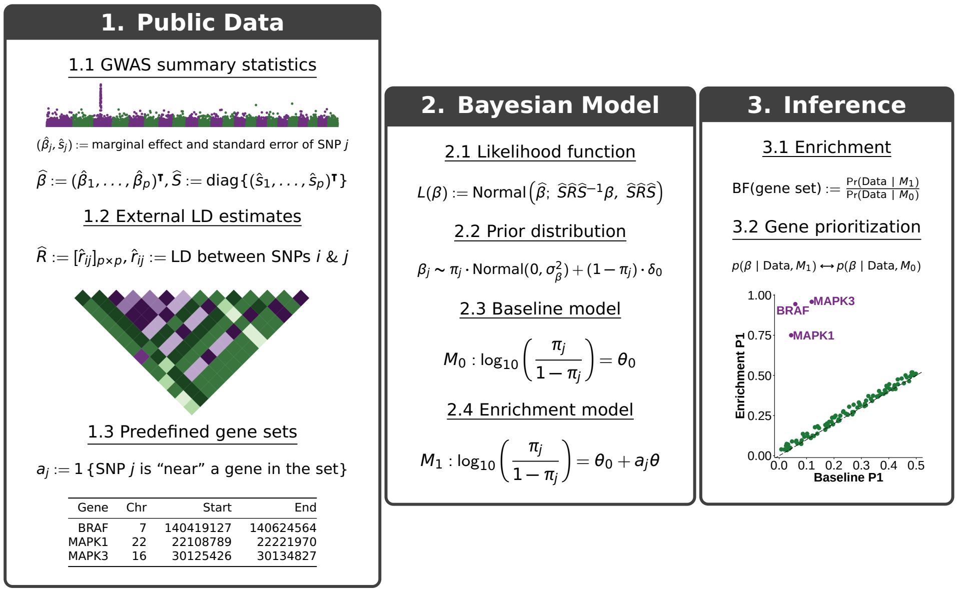

1. Public Data
1.1 GWAS summary statistics
$(\hat{\beta}_j, \hat{s}_j) :=$ marginal effect and standard error of SNP j
$\widehat{\beta} := (\hat{\beta}_1, \ldots, \hat{\beta}_p)^\intercal, \widehat{S} := \mathrm{diag}\{(\hat{s}_1, \ldots, \hat{s}_p)^\intercal\}$
1.2 External LD estimates
$\widehat{R} := [\hat{r}_{ij}]_{p \times p}, \hat{r}_{ij} :=$ LD between SNPs i & j
1.3 Predefined gene sets
$a_j := 1\,\{$SNP j is "near" a gene in the set$\}$
Gene Chr Start End
BRAF 7 140419127 140624564
MAPK1 22 22108789 22221970
MAPK3 16 30125426 30134827
2. Bayesian Model
2.1 Likelihood function
$L(\beta) := \mathrm{Normal}\left(\widehat{\beta};\ \widehat{S}\widehat{R}\widehat{S}^{-1}\beta,\ \widehat{S}\widehat{R}\widehat{S}\right)$
2.2 Prior distribution
$\beta_j \sim \pi_j \cdot \mathrm{Normal}(0, \sigma_\beta^2) + (1 - \pi_j) \cdot \delta_0$
2.3 Baseline model
$M_0 : \log_{10}\left(\frac{\pi_j}{1-\pi_j}\right) = \theta_0$
2.4 Enrichment model
$M_1 : \log_{10}\left(\frac{\pi_j}{1-\pi_j}\right) = \theta_0 + a_j\theta$
3. Inference
3.1 Enrichment
$\mathrm{BF(gene\ set)} := \frac{\Pr(\mathrm{Data} \mid M_1)}{\Pr(\mathrm{Data} \mid M_0)}$
3.2 Gene prioritization
$p(\beta \mid \mathrm{Data}, M_1) \longleftrightarrow p(\beta \mid \mathrm{Data}, M_0)$
MAPK3
BRAF
MAPK1
Enrichment P1
1.00
0.75
0.50
0.25
0.00
0.0
0.1
0.2
0.3
0.4
0.5
Baseline P1

Fig 1: **Schematic overview of RSS, a model-based enrichment analysis method for GWAS summary statistics.** RSS combines three types of public data: GWAS summary statistics (1.1), external LD estimates (1.2), and predefined SNP sets (1.3). GWAS summary statistics consist of a univariate effect size estimate ($\hat{\beta}_j$) and corresponding standard error ($\hat{s}_j$) for each SNP, which are routinely generated in GWAS. External LD estimates are obtained from an external reference panel with ancestry matching the population of GWAS cohorts. SNP sets here are derive from gene sets based on biological pathways or sequencing data. We combine these three types of data by fitting a Bayesian multiple regression (2.1-2.2) under two models about the enrichment parameter (θ): the "baseline model" (2.3) that each SNP has equal chance of being associated with the trait ($M_0 : \theta = 0$), and the "enrichment model" (2.4) that SNPs in the SNP set are more often associated with the trait ($M_1 : \theta > 0$). To test enrichment, RSS computes a Bayes factor (BF) comparing these two models (3.1). RSS also automatically prioritizes SNPs within an enriched set by comparing the posterior distributions of genetic effects (β) under M_0 and M_1, facilitating the discovery of new trait-associated genes (3.2). See Methods and Supplementary Note for additional details.

method as **R**egression with **S**ummary **S**tatistics, or RSS.

Specifically RSS requires single-SNP effect estimates and their standard errors from GWAS, and LD estimates from an external reference panel with similar ancestry to the GWAS cohort. Then, for any given set of SNPs, RSS estimates an “enrichment parameter”, θ, which measures the extent to which SNPs in the set are more often associated with the phenotype. This enrichment parameter is on a log-10 scale, so $\theta = 2$ means that the rate at which associations occur inside the set is ~ 100 times higher than the rate of associations outside the set, whereas $\theta = 0$ means that these rates are the same. When estimating θ RSS uses a multiple regression model to account for LD among SNPs. For example, RSS will (correctly) treat data from several SNPs that are in perfect LD as effectively a single observation, and not multiple independent observations. RSS ultimately summarizes the evidence for enrichment by a Bayes factor (BF) comparing the “enrichment model” ($M_1 : \theta > 0$) against the “baseline model” ($M_0 : \theta = 0$). RSS also provides posterior distributions of genetic effects under M_0 and M_1, and uses them to prioritize variants within enriched sets.

Although enrichment analysis could be applied to any SNP set, here we focus on SNP sets derived from “gene sets” such as biological pathways. Specifically, for a given gene set, we define a corresponding SNP set as the set of SNPs within ±100 kb of the transcribed region of any member gene; we refer to such SNPs as “inside" the gene set. If a gene set plays an important role in a trait then genetic associations may tend to occur more often near these genes than expected by chance; our method is designed to detect this signal.

To facilitate large-scale analyses, we designed an efficient, parallel algorithm implementing RSS. Our algorithm exploits variational inference [10], banded matrix approximation [11] and an expectation-maximization accelerator [12]. Software is available at `https://github.com/stephenslab/rss`.

Method comparison based on simulations. The novelty of RSS lies in its use of whole-genome association summary statistics to infer enrichments, and more importantly, its automatic prioritization of genes in light of the inferred enrichments. We are not aware of any published method with similar features. However, there are methods that can learn either enrichments or gene-level associations from GWAS summary statistics, *but not both*. We compared RSS to them through simulations using real genotypes [13].

To benchmark its enrichment component, we compared RSS with a suite of conventional pathway methods, Pascal [15], and a polygenic approach, LDSC [14]. We started with simulations without model mis-specification, where “baseline” and “enrichment” datasets were generated from correspond-

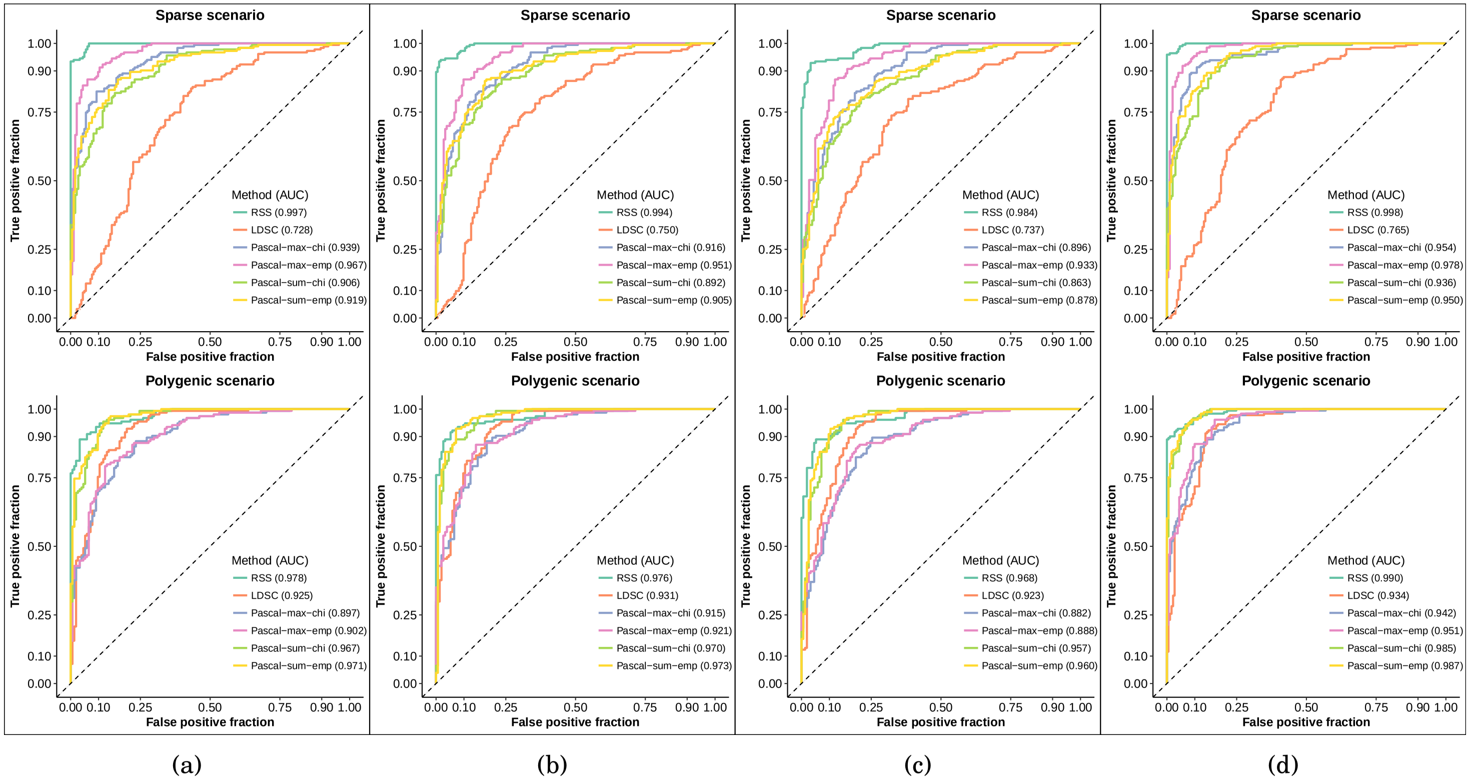


Fig 2: **Comparison of RSS to other methods for identifying enrichments from GWAS summary statistics.** We used real genotypes [13] to simulate individual-level data under two genetic architectures ("sparse" and "polygenic") with four baseline-enrichment patterns: **(a)** baseline and enrichment datasets followed baseline (M_0) and enrichment (M_1) models in RSS; **(b)** baseline datasets assumed that a random set of near-gene SNPs were enriched for genetic associations and enrichment datasets followed M_1; **(c)** baseline datasets assumed that a random set of coding SNPs were enriched for genetic associations and enrichment datasets followed M_1; **(d)** baseline datasets followed M_0 and enrichment datasets assumed that trait-associated SNPs were both more frequent, and had larger effects, inside than outside the target gene set. We computed the corresponding single-SNP summary statistics, and, on these summary data, we compared RSS with LDSC [14] and Pascal [15] using their default setups (URLs). Pascal includes two gene scoring options: maximum-of-χ^2 (`-max`) and sum-of-χ^2 (`-sum`), and two pathway scoring options: χ^2 approximation (`-chi`) and empirical sampling (`-emp`). For each simulated dataset, both Pascal and LDSC produced enrichment p-values, whereas RSS produced an enrichment BF; these statistics were used to rank the significance of enrichments. Each panel displays the trade-off between false and true enrichment discoveries for all methods in 200 baseline and 200 enrichment datasets of a given simulation scenario, and also reports the corresponding areas under the curve (AUCs), where a higher value indicates better performance. Simulation details and additional results are provided in Supplementary Figures 1-4.

ing models (M_0 and M_1). Figure 2a and Supplementary Figure 1 show the trade-off between false and true enrichment discoveries for each method. All methods are powerful when the true underlying genetic architecture is polygenic, whereas LDSC performs worse when the truth is sparse. In both polygenic and sparse scenarios RSS is the most powerful method.

Next, to assess its robustness to mis-specification, we performed three sets of simulations where either the baseline (M_0) or enrichment (M_1) model of RSS were mis-specified. Specifically, we considered scenarios where i) baseline data contained enrichments of random near-gene SNPs (Fig. 2b, Supplementary Fig. 2); ii) baseline data contained enrichments of random coding SNPs (Fig. 2c, Supplementary Fig. 3); and iii) enrichment data contained enrichments of effect sizes (Fig. 2d, Supplementary Fig. 4). The results show that RSS is highly robust to model mis-specifications, and still consistently outperforms Pascal and LDSC.

Recent analyses of GWAS summary statistics (e.g. [14]) often focus on the HapMap Project Phase 3 (HapMap3) SNPs [16], even though summary statistics of the 1000 Genomes Project SNPs [17] are available in many GWAS. Although not required, we used this "SNP subsetting" strategy in data analyses to reduce computation (Methods). To assess the impact of "SNP subsetting", we simulated data using all 1000 Genome SNPs, applied the enrichment methods to summary statistics of HapMap3 SNPs only, and compared HapMap3-based results with results of analyzing all 1000 Genome SNPs. For all methods, analysis with and without "SNP subsetting" produced similar results (Supplementary Fig. 5). The robustness to "SNP subsetting" is perhaps unsurprising, since all methods utilize LD information (in different ways) to capture the effects of potentially excluded causal variants. As above, RSS has the highest power in this set of simulations.

Finally, to benchmark its prioritization component, we compared RSS with four gene-based association methods [18–21]. Figure 3 and Supplementary Figures 6-7 show the power of each method to identify gene-level associations. RSS substantially outperforms existing methods even in the absence of enrichments, especially in the polygenic scenario. This is because RSS exploits a multiple regression framework [9] to learn the genetic architecture from data of all genes and assesses their effects jointly, whereas other methods implicitly assume a fixed, sparse architecture and only use data of a single gene to estimate its effect. When datasets contain enrichments, RSS automatically leverages them, which existing methods ignore, to further improve the power.

In conclusion, RSS outperforms existing methods in both enrichment and prioritization analysis, and is robust to a wide range of model mis-specification. To further investigate its real-world benefit, we applied RSS to analyze 31 complex traits and 4,026 gene sets.

Multiple regression on 1.1 million variants across 31 traits. The first step of our analysis is multiple regression of 1.1 million HapMap3 common SNPs for 31 traits, using GWAS summary statistics from 20,883-253,288 European ancestry individuals (Supplementary Table 1; Supplementary Fig. 8). This step essentially estimates, for each trait, a "baseline model" (M_0) against which "enrichment models" (M_1) can be compared. The fitted baseline model captures both the size and abundance ("polygenicity") of the genetic effects on each trait, effectively providing a two-dimensional summary of the genetic architecture of each trait (Fig. 4a; Supplementary Fig. 9).

The results emphasize that genetic architecture varies considerably among phenotypes: estimates of both polygenicity and effect sizes vary by several orders of magnitude. Height and schizophrenia stand out as being particularly polygenic, showing approximately 10 times as many estimated associated variants as any other phenotype. Along the other axis, fasting glucose, fasting insulin and haemoglobin show the highest estimates of effect sizes, with correspondingly lower estimates for the number of associated variants. Although not our main focus, these results highlight the potential for multiple regression models like ours to learn about effect size distributions and genetic architectures from GWAS summary statistics.

Fitting the baseline model also yields an estimate of effect size for each SNP. These can be used to identify trait-associated SNPs and loci. Reassuringly, these multiple-SNP results recapitulate many associations detected in single-SNP analyses of the same data (Supplementary Fig.s 10-12). For several traits, these results also identify additional associations (Supplementary Fig.s 13-14). These additional findings, while potentially interesting, may be difficult to validate and interpret. Enrichment analysis can help here: if the additional signals tend to be enriched in a plausible pathway, it may both increase confidence in the statistical results and provide some biological framework to interpret them.

Enrichment analyses of 3,913 pathways across 31 traits. We next performed enrichment analyses of SNP sets derived from 3,913 expert-curated pathways, ranging in size from 2 to 500 genes (Supplementary Fig.s 15-16). For each trait-pathway pair we computed a BF testing the enrichment model, and estimated the enrichment parameter θ.

Since these analyses involve large-scale computations that are subject to approximation error, we developed some "sanity checks" for confirming enrichments identified by RSS. Specifically these simple methods confirm that the z-scores for SNPs inside a putatively-enriched pathway have a different

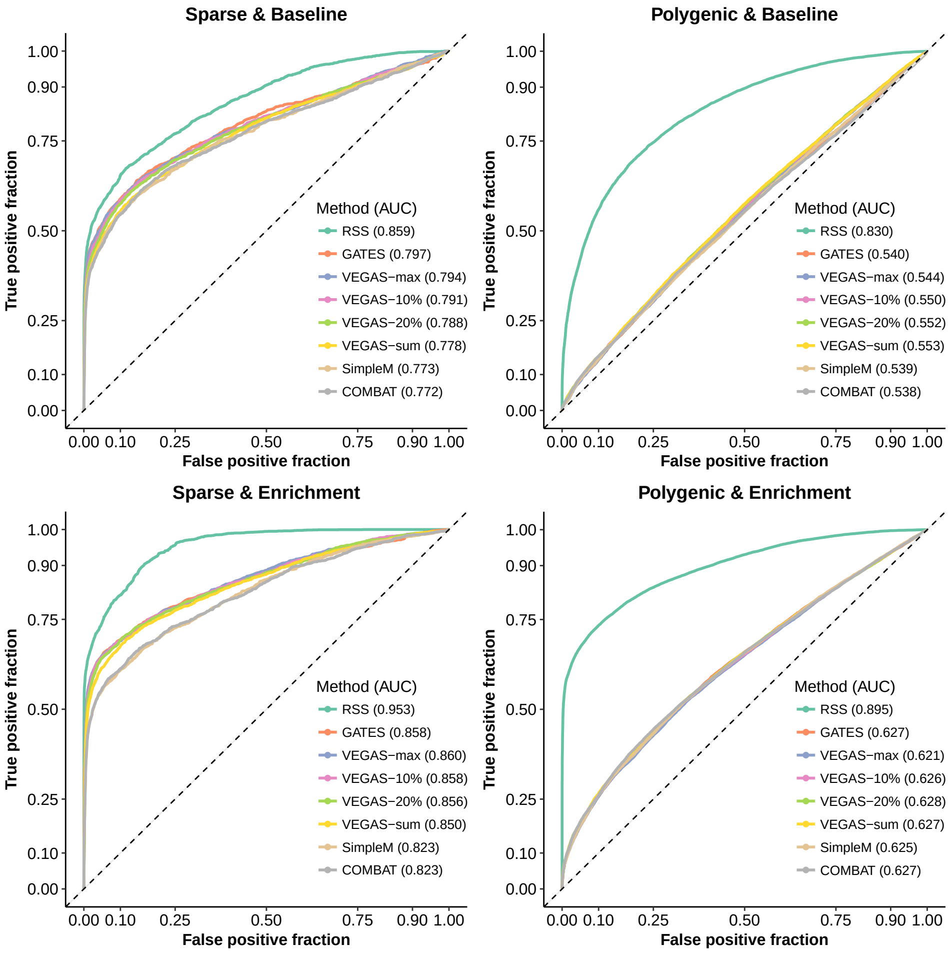
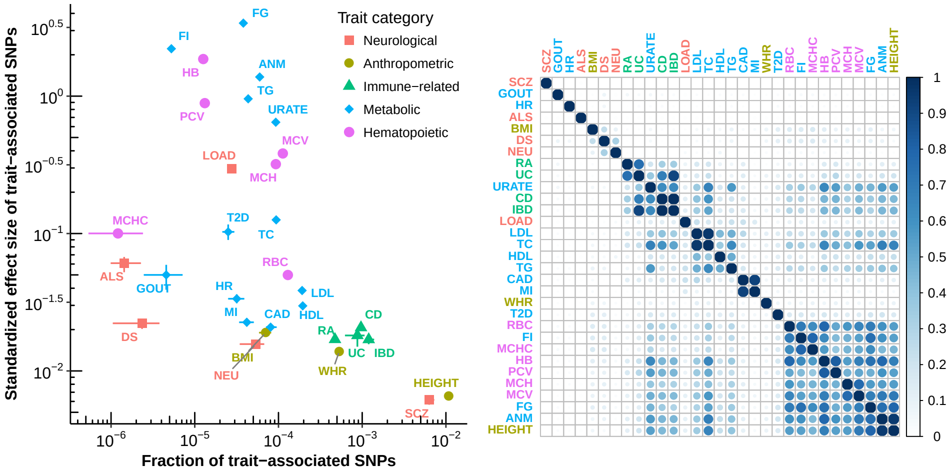


Fig 3: **Comparison of RSS to other methods for identifying trait-associated genes from GWAS summary statistics.** We used real genotypes [13] to simulate individual-level data under two genetic architectures ("Sparse" and "Polygenic"), with and without enrichment in the target gene set ("Enrichment" and "Baseline"), and then computed corresponding single-SNP summary statistics. On these summary data, we compared RSS with four other methods: SimpleM [18], VEGAS [19], GATES [20] and COMBAT [21]. We applied VEGAS to the full set of SNPs (`-sum`), to a specified percentage of the most significant SNPs (`-10%` and `-20%`), and to the single most significant SNP (`-max`), within ±100 kb of the transcribed region of each gene. All methods are available in the package `COMBAT` (URLs). For each simulated dataset, we defined a gene as "trait-associated" if at lease one SNP within ±100 kb of the transcribed region of this gene had non-zero effect. For each gene in each dataset, RSS produced the posterior probability that the gene was trait-associated. whereas the other methods produced association p-values; these statistics were used to rank the significance of gene-level associations. Each panel displays the trade-off between false and true gene-level associations for all methods in 100 datasets of a given simulation scenario, and reports the corresponding AUCs. Simulation details and additional results are provided in Supplementary Figures 6-7.



(a) (b)

Fig 4: **Baseline and enrichment analyses of 31 complex traits.** For both panels, traits are colored by categories and labeled by abbreviations. **(a)** Summary of inferred effect size distributions of 31 traits based on HapMap3 SNPs. Results are from fitting the baseline model (M_0) to 1.1 million common HapMap3 SNPs for each trait. We summarize effect size distribution using two numbers: the estimated fraction of trait-associated SNPs (average posterior probability of a HapMap3 SNP being associated with a trait; shown in x-axis) and the standardized effect size of trait-associated SNPs (average posterior mean effect size of all HapMap3 SNPs, normalized by the phenotypic standard deviation and the fraction of trait-associated SNPs; shown in y-axis). See Supplementary Note for details on these two quantities. Each dot represents a trait, with horizontal and vertical point ranges indicating posterior mean and 95% credible interval for each quantity. Note that some intervals are too small to be visible due to log-10 scales. See Supplementary Table 2 for numerical values of these intervals. **(b)** Pairwise sharing of 3,913 pathway enrichments among 31 traits. For each pair of traits, we estimated the proportion of pathways that are enriched in both traits, among pathways enriched in at least one of the traits (Methods). Traits are clustered by hierarchical clustering as implemented in the package `corrplot` (URLs). Darker color and larger shape represent higher sharing. The estimates of sharing are provided in Supplementary Table 3. ALS: amyotrophic lateral sclerosis [22]. DS: depressive symptoms [23]. LOAD: late-onset Alzheimer's disease [24]. NEU: neuroticism [23]. SCZ: schizophrenia [25]. BMI: body mass index [26]. HEIGHT: adult height [27]. WHR: waist-to-hip ratio [28]. CD: Crohn's disease [29]. IBD: inflammatory bowel disease [29]. RA: rheumatoid arthritis [30]. UC: ulcerative colitis [29]. ANM: age at natural menopause [31]. CAD: coronary artery disease [32]. FG: fasting glucose [33]. FI: fasting insulin [33]. GOUT: Gout [34]. HDL: high-density lipoprotein [35]. HR: heart rate [36]. LDL: low-density lipoprotein [35]. MI: myocardial infarction [32]. T2D: type 2 diabetes [37]. TC: total cholesterol [35]. TG: triglycerides [35]. URATE: serum urate [34]. HB: haemoglobin [38]. MCH: mean cell HB [38]. MCHC: MCH concentration [38]. MCV: mean cell volume [38]. PCV: packed cell volume [38]. RBC: red blood cell count [38].

Phenotype	Top enriched pathway	Database (Repository)	# of signals (# of genes)	$\log_{10}$ BF
Neurological traits				
Depressive symptoms	Eicosapentaenoate biosynthesis	HumanCyc (PC)	2 (12)	36.9
Alzheimer's disease	Golgi associated vesicle biogenesis	Reactome (PC)	3 (49)	83.7
Anthropometric traits				
Adult height	Endochondral ossification	WikiPathways (BS)	57 (65)	68.9
Immune-related traits				
Crohn's disease	Inflammatory bowel disease	KEGG (BS)	24 (61)	25.6
Inflammatory bowel disease	Inflammatory bowel disease	KEGG (BS)	26 (61)	24.2
Rheumatoid arthritis	CaN-regulated NFAT-dependent transcription in lymphocytes	PID (BS)	11 (45)	10.0
Ulcerative colitis	Inflammatory bowel disease	KEGG (BS)	16 (61)	11.8
Metabolic traits				
Age at natural menopause	IL-2Rβ in T cell activation	BioCarta	2 (37)	866.7
Coronary artery disease	p75(NTR)-mediated signaling	PID (BS)	4 (55)	16.0
Fasting glucose	Hexose transport	Reactome (BS)	4 (47)	1,898.4
Gout	Osteoblast signaling	WikiPathways (BS)	2 (13)	30.6
High-density lipoprotein	Statin pathway	WikiPathways (BS)	18 (30)	113.9
Low-density lipoprotein	Chylomicron-mediated lipid transport	Reactome (PC)	11 (17)	65.5
Myocardial infarction	Glutathione synthesis and recycling	Reactome (PC)	2 (11)	9.6
Total cholesterol	Glucose transport	Reactome (BS)	2 (41)	833.2
Triglycerides	Validated targets of C-MYC transcriptional activation	PID (BS)	3 (79)	604.9
Serum urate	Transport of glucose and others^{a}	Reactome (PC)	4 (95)	1,558.1
Hematopoietic traits				
Haemoglobin (HB)	RNA polymerase I transcription	Reactome (BS)	27 (107)	2,641.3
Mean cell HB (MCH)	Meiotic synapsis	Reactome (PC)	21 (72)	2,334.3
MCH concentration	SIRT1 negatively regulates ribosomal RNA expression	Reactome (PC)	3 (63)	700.8
Mean cell volume	DNA methylation	Reactome (PC)	28 (61)	2,077.3
Packed cell volume	RNA polymerase I promoter opening	Reactome (PC)	27 (59)	217.5
Red blood cell count	GSL biosynthesis (neolacto series)	KEGG (PC)	2 (21)	391.2

TABLE 1

Top-ranked pathways for enrichment of genetic associations in complex traits. For each trait here we report the most enriched pathway (if any) that i) has an enrichment Bayes factor (BF) greater than 10^8; ii) has at least 10 and at most 200 member genes; iii) has at least two member genes with $P_1 > 0.9$ (denoted as "signals") under the enrichment model; and iv) passes the visual and likelihood ratio sanity checks (Supplementary Fig. 17). All BFs reported here are larger than corresponding BFs that SNPs within ±100 kb of transcribed regions of all genes are enriched (Supplementary Fig. 19). The corresponding background and enrichment parameter estimates are provided in online results (URLs). P_1: the posterior probability that at least one SNP within ±100 kb of the transcribed region of a given gene has non-zero effect on the target trait. CaN: calcineurin. NFAT: nuclear factor of activated T cells. IL-2Rβ: interleukin-2 receptor beta chain. p75(NTR): p75 neurotrophin receptor. SIRT1: Sirtuin 1. GSL: glycosphingolipid. PC: Pathway Commons [39]. BS: NCBI BioSystems [40]. a: The full pathway name is "transport of glucose and other sugars, bile salts and organic acids, metal ions and amine compounds".

distribution from those outside the pathway (with more z-scores away from 0) – using both a visual check and a likelihood ratio statistic (Supplementary Fig. 17). Of note, these methods cannot replace RSS in the present study. The visual check requires human input, and thus is not suitable for large-scale analyses like ours. The likelihood ratio does not account for LD, and is expected to be less powerful (Supplementary Fig. 18).

Since genic regions may be generally enriched for associations compared with non-genic regions, we confirmed that top-ranked pathways often showed stronger evidence for enrichment than did the set containing all genes (Supplementary Fig. 19). We also created "null" (non-enriched) SNP sets by randomly drawing near-gene SNPs, and performed enrichment analyses of these "null" sets on real GWAS summary data. Enrichment signals of these simulated genic sets are substantially weaker than the actual top-ranked sets (Supplementary Fig. 20). Further, to check whether observed enrichments could be driven by other functional annotations (e.g. coding), we computed the correlation between enrichment BFs and proportions of gene-set SNPs falling into each of 52 functional categories [14]. Among 1,612 trait-category pairs, we did not observe any strong correlation (median 7.3×10^{-3}; 95% interval $[-0.08, 0.21]$; Supplementary Fig. 21). Together, these results suggest that observed enrichments are unlikely to be artifacts driven by model mis-specification.

For most traits our analyses identify many pathways with strong evidence for enrichment – for example, at a conservative threshold of BF $\geq 10^8$, 20 traits are enriched in more than 100 pathways per trait (Supplementary Fig. 22). Although the top enriched pathways for a given trait often substantially overlap (i.e. share many genes), several traits show enrichments with multiple non-overlapping or minimally-overlapping pathways (Supplementary Fig. 23). Table 1 gives examples of top enriched pathways, with full results available online (URLs).

Our results highlight many previously reported trait-pathway links. For example, the *Hedgehog pathway* is enriched for associations with adult height (BF=1.9×10^{40}), consistent with both pathway function [41] and previous GWAS [27]. Other examples include *interleukin-23 mediated signaling* with inflammatory bowel disease (BF=3.1×10^{23}; [42]), *T helper cell surface molecules* with rheumatoid arthritis (BF=3.2×10^{8}; [30]), *statin pathway* with levels of high-density lipoprotein cholesterol (BF=8.4×10^{113}; [43]), and *glucose transporters* with serum urate (BF=$1.2\times10^{1,558}$; [34]).

The results also highlight several pathway enrichments that were not reported in corresponding GWAS publications. For example, the top pathway for rheumatoid arthritis is *calcineurin-regulated nuclear factor of activated T cells (NFAT)-dependent transcription in lymphocytes* (BF=1.1×10^{10}). This result adds to the considerable existing evidence linking NFAT-regulated transcription to immune function [44] and bone pathology [45]. Other examples of novel pathway enrichments include *endochondral ossification* with adult height (BF=7.7×10^{68}; [46]), *p75 neurotrophin receptor-mediated signaling* with coronary artery disease (BF=9.6×10^{15}; [47]), and *osteoblast signaling* with gout (BF=3.8×10^{30}; [48]).

Overlapping pathway enrichments among related traits. Some pathways show enrichment in multiple traits. To gain a global picture of shared pathway enrichments among traits we estimated the proportions of shared pathway enrichments for all pairs of traits (Fig. 4b; Methods). Clustering these pairwise sharing results highlights four main clusters of traits: immune-related diseases, blood lipids, heart disorders and red blood cell phenotypes. Blood cholesterol shows strong pairwise sharing with serum urate (0.67), haemoglobin (0.66) and fasting glucose (0.53), which could be interpreted as a set of blood elements. Serum urate shows moderate to strong sharing with rheumatoid arthritis (0.19) and inflammatory bowel diseases (0.38-0.63), possibly due to the function of urate crystals in immune responses [49] Further, Alzheimer's disease shows moderate sharing with blood lipids (0.17-0.23), heart diseases (0.15-0.21) and inflammatory bowel diseases (0.10-0.13). This seems consistent with recent data linking Alzheimer's disease to lipid metabolism [50], vascular disorder [51] and immune activation [52]. The biologically relevant clustering of shared pathway enrichments helps demonstrate the potential of large-scale GWAS data to highlight similarities among traits, complementing other approaches such as clustering of shared genetic effects [53] and co-heritability analyses [54].

Novel trait-associated genes informed by enriched pathways. A key feature of RSS is that pathway enrichments, once identified, are automatically used to "prioritize" associations at variants near genes in the pathway. Specifically, RSS gives almost identical estimates of the background parameter (θ_0) in both baseline and enrichment analyses (Supplementary Fig. 24), and yields a positive estimate of the enrichment parameter (θ) if the pathway is identified as enriched (Supplementary Fig. 25). The positive estimate of θ increases the prior probability of association for SNPs in the pathway, which in turn increases the posterior probability of association for these SNPs.

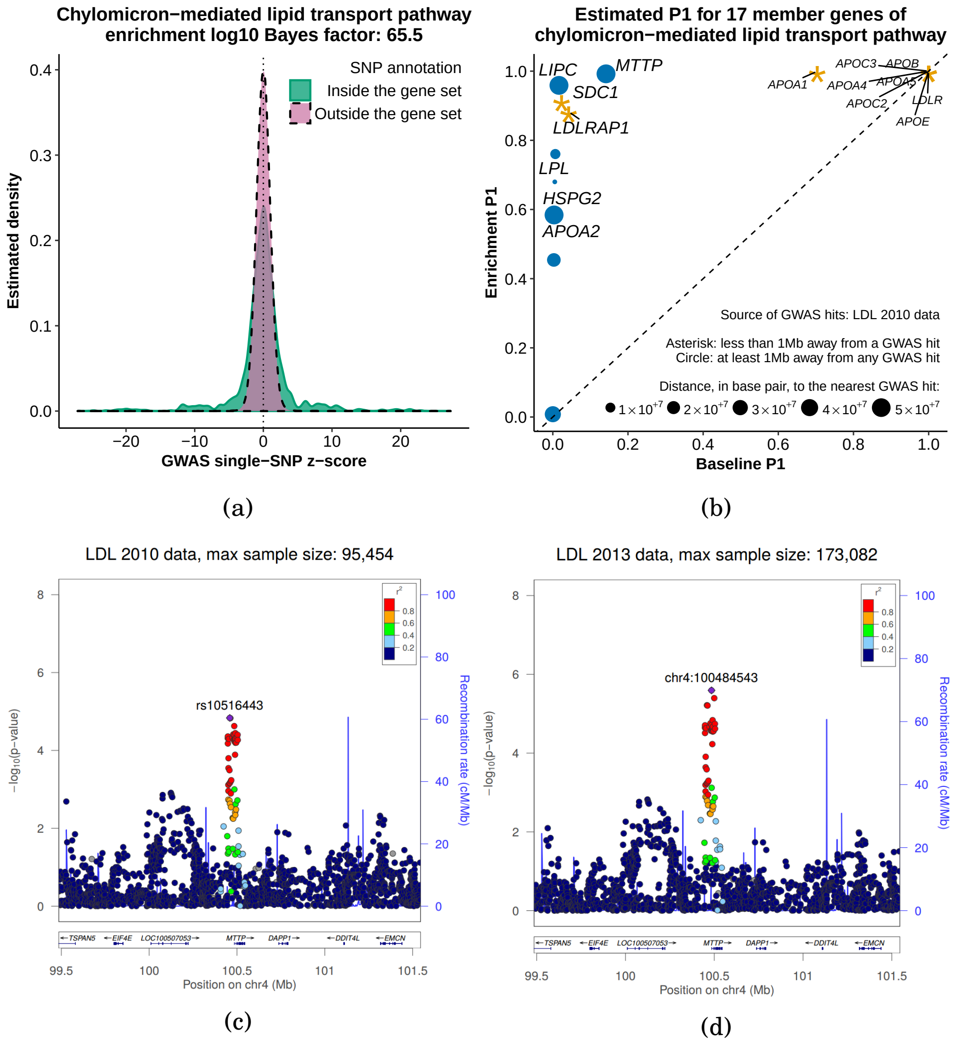
This ability to prioritize associations, which is not shared by most enrichment methods, has several important benefits. Most obviously, prioritization analyses can detect additional genetic associations that may otherwise be missed. Furthermore, prioritization facilitates the identification of genes influencing a phenotype in two ways. First, it helps identify genes that may explain individual variant associations, which is itself an important and challenging problem [55]. Second, prioritization helps identify genes that drive observed pathway enrichments. This can be useful to check whether a pathway enrichment may actually reflect signal from just a few key genes, and to understand enrichments of pathways with generic functions.

To illustrate, we performed prioritization analyses on the trait-pathway pairs showing strongest evidence for enrichment. Following previous Bayesian GWAS analyses [8, 56], here we evaluated genetic associations at the level of loci, rather than individual SNPs. Specifically, for each locus we compute P_1, the posterior probability that at least one SNP in the locus is associated with the trait, under both the baseline and enrichment hypothesis. Differences in these two P_1 estimates reflect the influence of enrichment on the locus.

The results show that prioritization analysis typically increases the inferred number of genetic associations (Supplementary Fig. 26), and uncovers putative associations that were not previously reported in GWAS. For example, enrichment in *chylomicron-mediated lipid transport* pathway (BF=3.4×10^{65}; Fig. 5a) informs a strong association between gene *MTTP* (baseline P_1: 0.14; enrichment P_1: 0.99) and levels of low-density lipoprotein (LDL) cholesterol (Fig. 5b). This gene is a strong candidate for harboring associations with LDL: *MTTP* encodes microsomal triglyceride transfer protein, which has been shown to involve in lipoprotein assembly; mutations in *MTTP* cause abetalipoproteinemia (OMIM: 200100), a rare disease characterized by permanently low levels of apolipoprotein B and LDL cholesterol; and *MTTP* is a potential pharmacological target for lowering LDL cholesterol levels [57]. However, no genome-wide significant SNPs near *MTTP* were reported in single-SNP analyses of either the same data [35] (Fig. 5c), or more recent data with larger sample size [58] (Fig. 5d).

Prioritization analysis of the same *chylomicron-mediated lipid transport* pathway also yields several additional plausible associations (Fig. 5b). These include *LIPC* (baseline P_1: 0.02; enrichment P_1: 0.96) and *LPL* (baseline P_1: 0.01; enrichment P_1: 0.76). These genes play important roles in lipid metabolism and both reach genome-wide significance in single-SNP analyses of HDL cholesterol and triglycerides [35] although not for LDL cholesterol (Supplementary Fig. 27); and a multiple-trait, single-SNP analysis [59] of the same data also did not detect associations of these genes with LDL.

Several other examples of putatively novel associations that arise from our gene prioritization analyses, together with related literature, are summarized in Box 1.



(a)

(b)

(c)

(d)

Fig 5: **Enrichment of *chylomicron-mediated lipid transport* pathway informs a strong association between a member gene *MTTP* and levels of low-density lipoprotein (LDL) cholesterol.** **(a)** Distribution of GWAS single-SNP z-scores from summary data published in 2010 [35], stratified by gene set annotations. The solid green curve is estimated from z-scores of SNPs within ± 100 kb of the transcribed region of genes in the *chylomicron-mediated lipid transport pathway* ("inside"), and the dashed reddish purple curve is estimated from z-scores of remaining SNPs ("outside"). This panel serves as a visual sanity check to confirm the observed enrichment. **(b)** Estimated posterior probability (P_1) that there is at least one associated SNP within ± 100 kb of the transcribed region of each pathway-member gene under the enrichment model (M_1) versus estimated P_1 under the baseline model (M_0). These gene-level P_1 estimates and corresponding SNP-level statistics are provided in Supplementary Table 4. **(c)** Regional association plot for *MTTP* based on summary data published in 2010 [35]. **(d)** Regional association plot for *MTTP* based on summary data published in 2013 [58].

Box 1 Select putatively novel associations from prioritization analyses

Adult height and *endochondral ossification* (65 genes, $\log_{10}$ BF = 68.9)

- *HDAC4* (baseline P_1: 0.98; enrichment P_1: 1.00)
  HDAC4 encodes a critical regulator of chondrocyte hypertrophy during skeletogenesis [60] and osteoclast differentiation [61]. Haploinsufficiency of *HDAC4* results in chromosome 2q37 deletion syndrome (OMIM: 600430) with highly variable clinical manifestations including developmental delay and skeletal malformations.
- *PTH1R* (baseline P_1: 0.94; enrichment P_1: 1.00)
  PTH1R encodes a receptor that regulates skeletal development, bone turnover and mineral ion homeostasis [62]. Mutations in *PTH1R* cause several rare skeletal disorders (OMIM: 215045, 600002, 156400).
- *FGFR1* (baseline P_1: 0.67; enrichment P_1: 0.97)
  FGFR1 encodes a receptor that regulates limb development, bone formation and phosphorus metabolism [63]. Mutations in *FGFR1* cause several skeletal disorders (OMIM: 101600, 123150, 190440, 166250).
- *MMP13* (baseline P_1: 0.45; enrichment P_1: 0.93)
  MMP13 encodes a protein that is required for osteocytic perilacunar remodeling and bone quality maintenance [64]. Mutations in *MMP13* cause a type of metaphyseal anadysplasia (OMIM: 602111) with reduced stature.

IBD and *cytokine-cytokine receptor interaction* (253 genes, $\log_{10}$ BF = 21.3)

- *TNFRSF14* (a.k.a. *HVEM*; baseline P_1: 0.98; enrichment P_1: 1.00)
  TNFRSF14 encodes a receptor that functions in signal transduction pathways activating inflammatory and inhibitory T-cell immune response. *TNFRSF14* expression plays a crucial role in preventing intestinal inflammation [65]. *TNFRSF14* is near a GWAS hit of celiac disease (rs3748816, $p = 3.3 \times 10^{-9}$) [66] and two hits of ulcerative colitis (rs734999, $p = 3.3 \times 10^{-9}$ [67]; rs10797432, $p = 3.0 \times 10^{-12}$ [68]).
- *FAS* (baseline P_1: 0.82; enrichment P_1: 0.99)
  FAS plays many important roles in the immune system [69]. Mutations in *FAS* cause autoimmune lymphoproliferative syndrome (OMIM: 601859).
- *IL6* (baseline P_1: 0.27; enrichment P_1: 0.87)
  IL6 encodes a cytokine that functions in inflammation and the maturation of B cells, and has been suggested as a potential therapeutic target in IBD [70].

CAD and *p75(NTR)-mediated signaling* (55 genes, $\log_{10}$ BF = 16.0)

- *FURIN* (baseline P_1: 0.69; enrichment P_1: 0.99)
  FURIN encodes the major processing enzyme of a cardiac-specific growth factor, which plays a critical role in heart development [71]. *FURIN* is near a GWAS hit (rs2521501 [72]) of both systolic blood pressure ($p = 5.2 \times 10^{-19}$) and hypertension ($p = 1.9 \times 10^{-15}$).
- *MMP3* (baseline P_1: 0.43; enrichment P_1: 0.97)
  A polymorphism in the promoter region of *MMP3* is associated with susceptibility to coronary heart disease-6 (OMIM: 614466). Inactivating *MMP3* in mice increases atherosclerotic plaque accumulation while reducing aneurysm [73].

HDL and *lipid digestion, mobilization and transport* (58 genes, $\log_{10}$ BF = 89.8)

- *CUBN* (baseline P_1: 0.24; enrichment P_1: 1.00)
  CUBN encodes a receptor for intrinsic factor-vitamin B12 complexes (cubilin) that maintains blood levels of HDL [74]. Mutations in *CUBN* cause a form of congenital megaloblastic anemia due to vitamin B12 deficiency (OMIM: 261100). *CUBN* is near a GWAS hit of total cholesterol (rs10904908, $p = 3.0 \times 10^{-11}$ [58]).
- *ABCG1* (baseline P_1: 0.01; enrichment P_1: 0.89)
  ABCG1 encodes an ATP-binding cassette transporter that plays a critical role in mediating efflux of cellular cholesterol to HDL [75].

RA and *lymphocyte NFAT-dependent transcription* (45 genes, $\log_{10}$ BF = 10.0)

- *PTGS2* (a.k.a. *COX2*; baseline P_1: 0.74; enrichment P_1: 0.98)
  PTGS2-specific inhibitors have shown efficacy in reducing joint inflammation in both mouse models [76] and clinical trials [77]. *PTGS2* is near a GWAS hit of Crohn's disease (rs10798069, $p = 4.3 \times 10^{-9}$ [29])
- *PPARG* (baseline P_1: 0.28; enrichment P_1: 0.98)
  PPARG has important roles in regulating inflammatory and immune responses with potential applications in treating chronic inflammatory diseases including RA [78, 79].

Enrichment analyses of 113 tissue-based gene sets across 31 traits. RSS is not restricted to pathways, and can be applied more generally. Here we use it to assess enrichment among tissue-based gene sets that we define based on gene expression data. Specifically we use RNA sequencing data from the Genotype-Tissue Expression project [80] to define sets of the most "relevant" genes in each tissue, based on expression patterns across tissues. The idea is that enrichment of GWAS signals near genes that are most relevant to a particular tissue may suggest an important role for that tissue in the trait.

A challenge here is how to define "relevant" genes. For example, are the highest expressed genes in a tissue the most relevant, even if the genes is ubiquitously expressed [81]? Or is a gene that is moderately expressed in that tissue, but less expressed in all other tissues, more relevant? To address this we considered three complementary methods to define tissue-relevant genes (Methods). The first method ("highly expressed", HE) uses the highest expressed genes in each tissue. The second method ("selectively expressed", SE) uses a tissue-selectivity score designed to identify genes that are much more strongly expressed in that tissue than in other tissues (S. Xi, personal communication). The third method ("distinctively expressed", DE) clusters the tissue samples and identifies genes that are most informative for distinguishing each cluster from others [82]. This last method yields a list of "relevant'' genes for each cluster, but most clusters are primarily associated with one tissue, and so we use this to assign genes to tissues. These methods often yield minimally overlapped gene sets for the same tissue (median overlap proportion: 4%; Supplementary Fig. 28)

Despite the small number of tissue-based gene sets relative to the pathway analyses above, this analysis identifies many strong enrichments. The top enriched tissues vary considerably among traits (Table 2), highlighting the benefits of analyzing a wide range of tissues. In addition, traits vary in which strategy for defining gene sets (HE, SE or DE) yields the strongest enrichment results. For example, HE genes in heart show strongest enrichment for heart rate; SE genes in liver show strongest enrichment for LDL. This highlights the benefits of considering multiple annotation strategies, and suggests

Phenotype	Tissue (Method)	$\log_{10}$ BF	Select top driving genes (# of genes with enrichment $P_1 > 0.9$)
Alzheimer's disease	Adrenal gland (SE)	45.6	*APOE*, *APOC1* (2)
Neuroticism	Brain (SE)	26.3	*LINGO1*, *KCNC2* (2)
Adult height	Nerve tibial (DE)	25.2^b	*PTCH1*, *SFRP4*, *FLNB* (59)
Crohn's disease	Cluster 1^a (DE)	15.4	*SMAD3*, *ZMIZ1*, *NUPR1* (6)
Inflammatory bowel disease	Cluster 1^a (DE)	15.8	*SMAD3*, *ZMIZ1*, *NUPR1* (10)
Ulcerative colitis	Heart (HE)	7.0	*PLA2G2A*, *TCAP*, *ALDOA* (4)
Age at natural menopause	Brain (DE)	1,053.2	*BRSK1*, *PPP1R1B*, *NPTXR* (6)
Coronary artery disease	Brain (DE)	8.5	*PSRC1*, *ZEB2*, *PTPN11* (3)
Fasting glucose	Pancreas (SE)	2,396.8	*G6PC2*, *PDX1*, *SLC30A8* (5)
Fasting insulin	Testis (SE)	866.7	*ABHD1*, *PRR30*, *C2orf16* (3)
Heart rate	Heart (HE)	4.1	*MYH6*, *PLN* (5)
High-density lipoprotein	Liver (HE)	20.2	*APOA1*, *APOE*, *MT1G*, *FTH1* (10)
Low-density lipoprotein	Liver (SE)	33.4	*ABCG5*, *LPA*, *ANGPTL3*, *HP* (13)
Total cholesterol	Liver (DE)	56.0	*APOA1*, *APOE*, *HP* (9)
Triglycerides	Liver (HE)	93.2	*APOA1*, *APOE*, *FTH1* (7)
Serum urate	Kidney (SE)	210.8^b	*SLC17A1*, *SLC22A11*, *PDZK1* (7)
Haemoglobin (HB)	Whole blood (DE)	2,078.1	*HIST1H1E*, *HIST1H1C* (4)
Mean cell HB	Whole blood (DE)	1,363.0	*NPRL3*, *FBXO7*, *UBXN6* (11)
Mean cell volume	Whole blood (DE)	$1{,}019.6^b$	*UBXN6*, *RBM38*, *NPRL3* (11)
Packed cell volume	Heart (HE)	945.4	*RPL19*, *TCAP* (2)
Red blood cell count	Breast (SE)	141.7	*OBP2B*, *STAC2* (2)

TABLE 2

Top enriched tissue-based gene sets in complex traits. Each tissue-based gene set contains 100 transcribed genes used in the Genotype-Tissue Expression project. For each trait we report the most enriched tissue-based gene set (if any) that has a Bayes factor (BF) greater than 1,000 and has more than two member genes with enrichment $P_1 > 0.9$. All trait-tissue pairs reported above pass the sanity checks (Supplementary Fig. 17). The corresponding background and enrichment parameter estimates are provided in online results (URLs). P_1: the posterior probability that at least one SNP within ±100 kb of the transcribed region of a given gene has non-zero effect on the target trait. HE: highly expressed. SE: selectively expressed. DE: distinctively expressed. a: Multiple tissues show partial membership in "Cluster 1", including ovary, thyroid, spleen, breast and stomach [82]. b: These three BFs are smaller than corresponding BFs that SNPs within ±100 kb of transcribed regions of all genes are enriched (Supplementary Fig. 19).

that, unsurprisingly, there is no single answer to the question of which genes are most "relevant" to a tissue.

For some traits, the top enriched results (Table 2) recapitulate previously known trait-tissue connections (e.g. lipids and liver, glucose and pancreas), supporting the potential for our approach to identify trait-relevant tissues. Further, many traits show enrichments in multiple tissues. For example, associations in coronary artery disease are strongly enriched in genes related to both *heart* (SE, BF = 6.6×10^7) and *brain* (DE, BF = 3.5×10^8). The multiple-tissue enrichments highlight the potential for our approach to also produce novel biological insights, which we illustrate through an in-depth analysis of late-onset Alzheimer's disease (LOAD).

Tissue-based analysis of LOAD identified three tissues with very strong evidence for enrichment (BF>10^{30}): liver, brain and adrenal gland. Because of the well-known connection between gene *APOE* and LOAD [83], and the fact that *APOE* is highly expressed in these three tissues (Supplementary Fig. 29), we hypothesized that *APOE* and related genes might be driving these results. To assess this we re-analyzed these strongly enriched gene sets after removing the entire apolipoproteins (APO) gene family from them. Of the three tissues, only liver remains (moderately) enriched after excluding APO genes (Fig. 6), suggesting a possible role for non-APO liver-related genes in the etiology of LOAD.

To identify additional genes underlying the liver enrichment, we performed prioritization analysis for non-APO liver-related genes. This highlighted an association of LOAD with gene *TTR* (baseline P_1: 0.64; enrichment P_1: 1.00; Supplementary Fig. 30). *TTR* encodes transthyretin, which has been shown to inhibit LOAD-related protein from forming harmful aggregation and toxicity [84, 85]. Indeed, transthyretin is considered a biomarker for LOAD: patients show reduced transthyretin levels in plasma [86] and cerebrospinal fluid [87]. Rare variants in *TTR* have recently been found to be associated with LOAD [88, 89]. By integrating GWAS with expression data our analysis identifies association of LOAD with *TTR* based on common variants.

Discussion

We have presented RSS, a new method for simultaneous enrichment and prioritization analysis of GWAS summary data, and illustrated its potential to yield novel insights by extensive analyses involving 31 phenotypes and 4,026 gene sets. We have space to highlight only select findings, and expect that researchers will find the full results (URLs) to contain further insights.

Enrichment tests, sometimes known as "competitive tests" [5, 6], have several advantages over alternative approaches – sometimes known as "self-

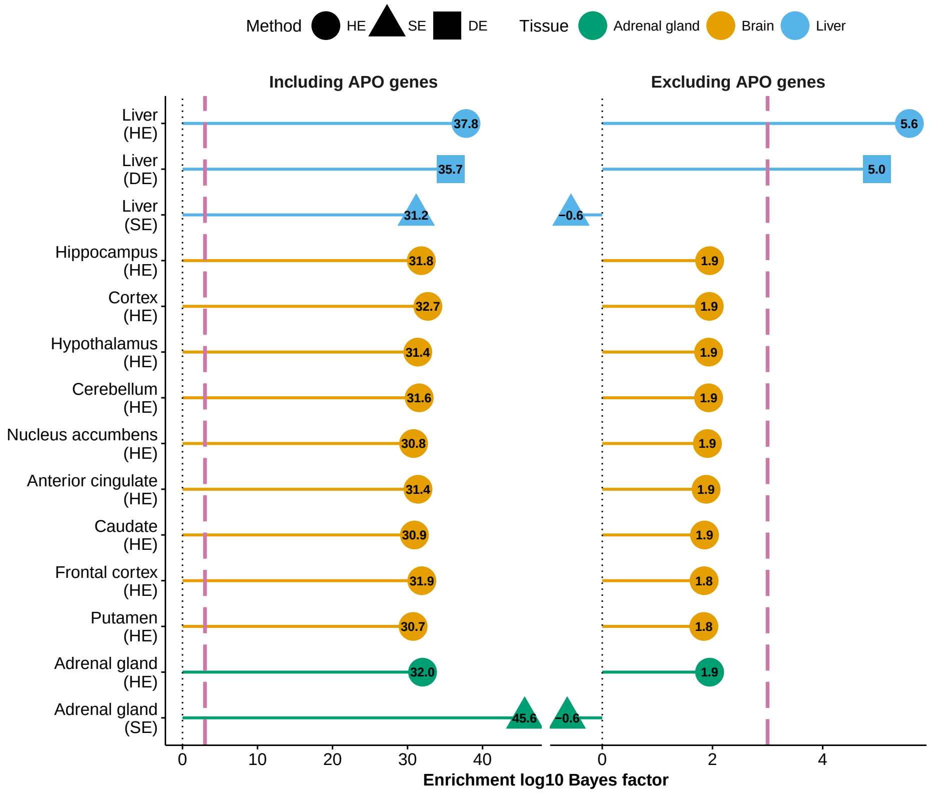


Fig 6: **Enrichment analyses of genes related to liver, brain and adrenal gland for Alzheimer's disease.** Shown are the tissue-based gene sets with the strongest enrichment signals for Alzheimer's disease. Each gene set was analyzed twice: the left panel corresponds to the analysis based on the original gene set; the right panel corresponds to the analysis where SNPs within ± 100 kb of the transcribed region of any gene in Apolipoproteins (APO) family (URLs) are excluded from the original gene set. Dashed lines in both panel denote the same Bayes factor threshold (1,000) used in our tissue-based analysis of all 31 traits (Table 2). HE: highly expressed. SE: selectively expressed. DE: distinctively expressed.

contained tests" (e.g. [90, 91]) – that simply test whether a SNP set contains at least one association. For example, for complex polygenic traits any large pathway will likely contain at least one association, making self-contained tests unappealing. Enrichment tests are also more robust to confounding effects such as population stratification, because confounders that affect the whole genome will generally not create artifactual enrichments. Indeed, in this sense enrichment results can be more robust than single-SNP results. (Nonetheless, most of the summary data analyzed here were corrected for confounding; see Supplementary Table 5.)

Compared with other enrichment approaches, RSS has several particularly attractive features. First, unlike many methods (e.g. [5, 92, 93]) RSS uses data from *all* variants, and not only those that pass some significance threshold. This increases the potential to identify subtle enrichments even in GWAS with few significant results. Second, RSS models enrichment directly as an increased rate of association of variants within a SNP set. This contrasts with alternative two-stage approaches (e.g. [15, 94, 95]) that first collapse SNP-level association statistics into gene-level statistics, and then assesses enrichment at the gene level. Our direct modeling approach has important advantages, most obviously that it avoids the difficult and error-prone steps of assigning SNP associations to individual genes, and collapsing SNP-level associations into gene-level statistics. For example, simply assigning SNP associations to the nearest gene may highlight the "wrong" gene and miss the "correct" gene [55]. Although our enrichment analyses of gene sets do involve assessing proximity of SNPs to genes in each gene set, they *avoid uniquely assigning each SNP to a single gene*, which is a subtle but important distinction. Finally, and perhaps most importantly, our model-based enrichment approach leads naturally to prioritization analyses that highlight which genes in an enriched pathways are most likely to be trait-associated. We know of only two published methods [8, 96] with similar features, but both require individual-level data and so could not perform the analyses presented here.

Although previous studies have noted potential benefits of integrating gene expression with GWAS data, our enrichment analyses of expression-based gene sets are different from, and complementary to, this previous work. For example, many studies have used expression quantitative trait loci (eQTL) data to help inform GWAS results (e.g. [97–104]). In contrast we bypass the issue of detecting (tissue-specific) eQTLs by focusing only on differences in gene expression levels among tissues. And, unlike methods that attempt to (indirectly) relate expression levels to phenotype (e.g. [105, 106]), our approach focuses firmly on genotype-phenotype associations. Nonetheless, as our results from different tissue-based annotations demonstrate, it can be extremely beneficial to consider multiple approaches, and we view these methods as complimentary rather than competing.

Like any method, RSS also has limitations that need to be considered when interpreting results. For example, annotating variants as being “inside a gene set” based on proximity to a relevant gene, while often effective, can occasionally give misleading results. We saw an example of this when our method identified an enrichment of SE genes in testis with both total cholesterol and triglycerides. Further prioritization analysis revealed that this enrichment was driven by a single gene, *C2orf16* which is a) uniquely expressed in testis, and b) physically close (53 kb) to another gene, *GCKR*, that is strongly associated with lipid traits (Supplementary Fig. 31). This highlights the need for careful examination of results, and also the utility of our prioritization analyses. Generally we view enrichments that are driven by a single gene as less reliable and useful than enrichments driven by multiple genes; indeed, enrichments driven by a single gene seem better represented as a gene association than as a gene set enrichment. Other problems that can affect enrichment methods (not only ours) include: a) an enrichment signal in one pathway can be caused by overlap with another pathway that is genuinely involved in the phenotype; and b) for some traits (e.g. height), genetic associations may be strongly enriched near all genes, which will cause many pathways to appear enriched.

Other limitations of RSS stem from its use of variational inference for approximate Bayesian calculations. Although these methods are computationally convenient in large datasets, and often produce reliable results (e.g. [8, 10, 107–115]), they also have features to be aware of. One feature is that when multiple SNPs in strong LD are associated with a trait, the variational approximation tends to select one of them and ignore the others. This feature will not greatly affect enrichment inference provided that SNPs that are in strong LD tend to have the same annotation (because then it will not matter which SNP is selected). And this holds for the gene-based annotations in the present study. However, it would not hold for “finer-scale” annotations (e.g. appearance in a DNase peak), and so in that setting the use of the variational approximation may need more care. More generally the accuracy of the variational approximation can be difficult to assess, especially since the underlying coordinate ascent algorithm only guarantees convergence to a local optimum. This said, the main alternative for making Bayesian calculations, Markov chain Monte Carlo, can experience similar difficulties.

Finally, the present study examines a single annotation (e.g. one gene set) at a time. Extending RSS to jointly analyze multiple annotations like [14] could further increase power to detect novel associations, and help distinguish between competing correlated annotations (e.g. overlapping pathways) when explaining observed enrichments.

URLs. Software, https://github.com/stephenslab/rss; Demonstration of software, http://stephenslab.github.io/rss/Example-5; Full results, https://xiangzhu.github.io/rss-gsea/; APO gene family: http://www.genenames.org/cgi-bin/genefamilies/set/405; Pascal, https://www2.unil.ch/cbg/index.php?title=Pascal; LDSC, https://github.com/bulik/ldsc; COMBAT, https://cran.r-project.org/web/packages/COMBAT; corrplot, https://cran.r-project.org/web/packages/corrplot.

Acknowledgments. This study is supported by the Grant GBMF # 4559 from the Gordon and Betty Moore Foundation and NIH Grant HG02585 to M.S. This study makes use of data generated by the Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. We thank all the GWAS consortia for making their summary statistics publicly available. A full list of acknowledgments is in Supplementary Note. We thank the GTEx consortium for making their RNA sequencing data publicly available. We thank P. Carbonetto and X. He for helpful comments on a draft manuscript. We thank C. A. Anderson, J. R. B. Perry, R. J. F. Loos and M. den Hoed for answering questions related to summary data of specific traits. We thank M. C. Turchin for downloading summary statistics of red blood cell phenotypes under accession number EGAS00000000132. We thank S. Xi and K. K. Dey for providing expression-based gene lists. This work was completed with resources provided by the University of Chicago Research Computing Center.

Author Contributions. X.Z. and M.S. conceived the study. X.Z. and M.S. developed the methods. X.Z. developed the algorithms, implemented the software and performed the analyses. X.Z. and M.S. wrote the manuscript.

Methods

GWAS summary statistics, LD estimates and SNP annotations. We analyze GWAS summary statistics of 31 traits, in particular, the estimated single-SNP effect and standard error for each SNP. Following [14], we use the same set of HapMap3 SNPs [16] for all 31 traits, even though some traits have summary statistics available on all 1000 Genomes SNPs [17]. We use this “SNP subsetting” strategy to reduce computation, since the computational complexity of RSS (per iteration) is linear with the total number of SNPs (Supplementary Notes).

Among the HapMap3 SNPs, we also exclude SNPs on sex chromosomes, SNPs with minor allele frequency less than 1%, SNPs in the major histocompatibility complex region, and SNPs measured on custom arrays (e.g. Metabochip, Immunochip) from analyses. The final set of analyzed variants consists of 1.1 million SNPs (Supplementary Table 1, Supplementary Fig. 8).

Since GWAS summary statistics used here were all generated from European ancestry cohorts, we use haplotypes of individuals with European ancestry from the 1000 Genomes Project, Phase 3 [17] to estimate LD [11].

To create SNP-level annotations for a given gene set, we use a distance-based approach from previous enrichment analyses [8, 94]. Specifically, we annotate each SNP as being "inside" a gene set if it is within ± 100 kb of the transcribed region of a gene in the gene set. The relatively broad region is chosen to capture signals from nearby regulatory variants, since the majority of GWAS hits are non-coding.

Biological pathways and genes. Biological pathway definitions are retrieved from nine databases (BioCarta, BioCyc, HumanCyc, KEGG, miRTarBase, PANTHER, PID, Reactome, WikiPathways) that are archived by four repositories: Pathway Commons (version 7) [39], NCBI Biosystems [40], PANTHER (version 3.3) [116] and BioCarta (used in [8]). Gene definitions are based on *Homo sapiens* reference genome GRCh37. Both pathway and gene data were downloaded on August 24, 2015. We use the same protocol described in [8] to compile a list of 3,913 pathways that contains 2-500 autosomal protein-coding genes for the present study. We summarize pathway and gene information in Supplementary Figures 15-16.

Tissue-based gene sets derived from transcriptome. Complex traits are often affected by multiple tissues, and it is not obvious *a priori* what the most relevant tissues are for the trait. Hence, it is necessary to examine a comprehensive set of tissues. The breadth of tissues in Genotype-Tissue Expression (GTEx) project [80] provides such an opportunity.

Here we use RNA sequencing data to create 113 tissue-based gene sets. Due to the complex nature of extracting tissue relevance from sequencing data, we consider three different methods to derive tissue-based gene sets.

The first method ("highly expressed") ranks the mean Reads Per Kilobase per Million mapped reads (RPKM) of all genes based on data of a given tissue, and then selects the top 100 genes with the largest mean RPKM values to represent the target tissue. We downloaded gene lists of 44 tissues with sample sizes greater than 70 from the GTEx Portal on November 21, 2016.

The second method ("selectively expressed") computes a tissue-selectivity score in a given tissue for each gene, which is essentially the average log ratio of expressions in the target tissue over other tissues, and then uses the top 100 genes with the largest tissue-selectivity scores to represent the target tissue. We obtained gene lists of 49 tissues from S. Xi on February 13, 2017.

The third method ("distinctively expressed") summarizes 53 tissues as 20 biologically-distinct clusters using admixture models, computes a cluster-distinctiveness score in a given cluster for each gene, and then uses the top 100 genes with the largest cluster-distinctiveness scores to represent the target cluster [82]. We extracted gene lists of 20 clusters from [82] on May 19, 2016.

Bayesian statistical models. Consider a GWAS with n unrelated individuals typed on p SNPs. For each SNP j, we denote its estimated single-SNP effect size and standard error as $\hat{\beta}_j$ and $\hat{s}_j$ respectively. To model $\{\hat{\beta}_j, \hat{s}_j\}$, we use the **R**egression with **S**ummary **S**tatistics (RSS) likelihood [9]:

$$L(\boldsymbol{\beta}) := \mathcal{N}(\widehat{\boldsymbol{\beta}}; \widehat{S}\widehat{R}\widehat{S}^{-1}\boldsymbol{\beta}, \widehat{S}\widehat{R}\widehat{S}) \tag{1}$$

where $\widehat{\boldsymbol{\beta}} := (\hat{\beta}_1, \ldots, \hat{\beta}_p)^\intercal$, $\widehat{S} := \text{diag}(\widehat{\mathbf{s}})$, $\widehat{\mathbf{s}} := (\hat{s}_1, \ldots, \hat{s}_p)^\intercal$, $\widehat{R}$ is the LD matrix estimated from an external reference panel with ancestry matching the GWAS cohort, $\boldsymbol{\beta} := (\beta_1, \ldots, \beta_p)^\intercal$ are the true effects of each SNP on phenotype, and $\mathcal{N}$ denotes the multivariate normal distribution.

To model enrichment of genetic associations within a given gene set, we borrow the idea from [8] and [56], to specify the following prior on $\boldsymbol{\beta}$:

$$\beta_j \sim \pi_j \mathcal{N}(0, \sigma_\beta^2) + (1 - \pi_j)\delta_0, \tag{2}$$

$$\sigma_\beta^2 = h \cdot \left(\textstyle\sum_{j=1}^p \pi_j n^{-1} \hat{s}_j^{-2}\right)^{-1}, \tag{3}$$

$$\pi_j = (1 + 10^{-(\theta_0 + a_j \theta)})^{-1}, \tag{4}$$

where δ_0 denotes point mass at zero, θ_0 reflects the background proportion of trait-associated SNPs, θ reflects the increase in probability, on the log10-odds scale, that a SNP inside the gene set has non-zero effect, h approximates the proportion of phenotypic variation explained by genotypes of all available SNPs, and a_j indicates whether SNP j is inside the gene set. Following [8], we place independent uniform grid priors on the hyper-parameters $\{\theta_0, \theta, h\}$ (Supplementary Tables 6-7). (If one had specific information about hyper-parameters in a given application then this could be incorporated here.)

Posterior computation. We combine the likelihood function and prior distribution above to perform Bayesian inference. The posterior computation procedures largely follow those developed in [10]. First, for each set of hyper-parameters $\{\theta_0, \theta, h\}$ from a predefined grid, we approximate the (conditional) posterior of $\boldsymbol{\beta}$ using a variational Bayes algorithm. Next, we approximate the posterior of $\{\theta_0, \theta, h\}$ by a discrete distribution on the predefined grid, using the variational lower bounds from the first step to compute the discrete probabilities. Finally, we integrate out the conditional posterior of $\boldsymbol{\beta}$ over the posterior of $\{\theta_0, \theta, h\}$ to obtain the full posterior of $\boldsymbol{\beta}$.

Following [8], we set random initialization as a default for the variational Bayes algorithm. Specifically, we randomly select an initialization, and then use this same initial value for all variational approximations over the grid of $\{\theta_0, \theta, h\}$. This simple approach was used in all simulations and data analyses for the present study, and yielded satisfying results in most cases.

To facilitate large-scale analyses, we employ several computational tricks. First, we use squared iterative methods [12] to accelerate the fixed point iterations in the variational approximation. Second, we exploit the banded LD matrix [11] to parallelize the algorithm. Third, we use a simplification in [8] that scales the enrichment analysis to thousands of gene sets by reusing expensive genome-wide calculations. See Supplementary Note for details.

For one trait, the total computational cost of our analyses is determined by the number of whole-genome SNPs, the number of gene sets and the grid size for hyper-parameters, all of which can vary considerably among studies. It is thus hard to make general statements about computational time. However, to give a specific example, we finished baseline and enrichment analyses of 1.1 million HapMap3 SNPs and 3,913 pathways for LDL within 36 hours in a standard computer cluster (48 nodes, 12-16 CPUs per node).

All computations in the present study were performed on a Linux system with multiple (4-22) Intel E5-2670 2.6GHz, Intel E5-2680 2.4GHz or AMD Opteron 6386 SE processors.

Assess gene set enrichment. To assess whether a gene set is enriched for genetic associations with a target trait, we evaluate a Bayes factor (BF):

$$\text{BF} := \frac{p(\widehat{\boldsymbol{\beta}} | \widehat{S}, \widehat{R}, \mathbf{a}, \theta > 0)}{p(\widehat{\boldsymbol{\beta}} | \widehat{S}, \widehat{R}, \mathbf{a}, \theta = 0)}, \tag{5}$$

where $\mathbf{a} := (a_1, \ldots, a_p)^\intercal$ and a_j indicates whether SNP j is inside the gene set. The observed data are BF times more likely under the enrichment model ($\theta > 0$) than under the baseline model ($\theta = 0$), and so the larger the BF, the stronger evidence for gene set enrichment. See Supplementary Note for details of computing enrichment BF.

Detect association between a locus and a trait. To identify trait-associated loci, we consider two statistics derived from the posterior distribution of $\boldsymbol{\beta}$. The first statistic is P_1, the posterior probability that at least 1

SNP in the locus is associated with the trait:

$$P_1 := 1 - \Pr(\beta_j = 0,\ \forall j \in \text{locus}|\mathbf{D}), \tag{6}$$

where **D** is a shorthand for the input data including GWAS summary statistics, LD estimates and SNP annotations (if applicable). The second statistic is ENS, the posterior expected number of associated SNPs in the locus:

$$\text{ENS} := \textstyle\sum_{j \in \text{locus}} \Pr(\beta_j \neq 0|\mathbf{D}). \tag{7}$$

See Supplementary Note for details of computing P_1 and ENS.

Estimate pairwise sharing of pathway enrichments. To capture the "sharing" of enrichments between two traits, we define $\pi = (\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11})$:

$$\pi_{ab} := \Pr(z_{1j} = a, z_{2j} = b),\ a \in \{0,1\},\ b \in \{0,1\}, \tag{8}$$

where z_{ij} equals one if pathway j is enriched in trait i and zero otherwise. Assuming independence among pathways and phenotypes, we estimate π by

$$\hat{\pi} := \arg\max_{\pi} \textstyle\prod_j (\pi_{00} + \pi_{01}\text{BF}_{2j} + \pi_{10}\text{BF}_{1j} + \pi_{11}\text{BF}_{1j}\text{BF}_{2j}), \tag{9}$$

where BF_{ij} is the enrichment BF for trait i and pathway j. We solve this optimization problem using an expectation-maximization algorithm implemented in the package ashr [117]. Finally, the conditional probability that a pathway is enriched in a pair of traits given that it is enriched in at least one trait, as plotted in Figure 4b, is estimated as $\hat{\pi}_{11}/(1 - \hat{\pi}_{00})$.

Connection with enrichment analysis of individual-level data. RSS has close connection with the method developed for individual-level data [8]. Under certain conditions [9], we can show that these two methods are mathematically equivalent, in the sense that they have the same fix point iteration scheme and lower bound in variational approximations. See Supplementary Note for proofs. In addition to their theoretical connections, we also empirically compared two methods through simulations, and observed similar inferential results (Supplementary Fig. 32).

Code and data availability. Links to source codes and full results of the present study are provided in URLs. Links to GWAS summary statistics are provided in Supplementary Note. HapMap3 SNP list: https://data.broadinstitute.org/alkesgroup/LDSCORE/w_hm3.snplist.bz2. 1000 Genomes Phase 3 data: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502. Pathway Commons: http://www.pathwaycommons.org/archives/PC2/v7. NCBI

Biosystems: ftp://ftp.ncbi.nih.gov/pub/biosystems. PANTHER: ftp://ftp.pantherdb.org/pathway. BioCarta: https://github.com/pcarbo/bmapathway/tree/master/data. GTEx Portal: https://www.gtexportal.org/home/. "Distinctively expressed" genes [82]: http://stephenslab.github.io/count-clustering. ashr: https://cran.r-project.org/web/packages/ashr.

References.

[1] Price, A. L., Spencer, C. C. & Donnelly, P. Progress and promise in understanding the genetic basis of common diseases. In *Proceedings of the Royal Society B*, vol. 282, 20151684 (The Royal Society, 2015).

[2] Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *The American Journal of Human Genetics* **90**, 7–24 (2012).

[3] McCarthy, M. I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics* **9**, 356–369 (2008).

[4] Sham, P. C. & Purcell, S. M. Statistical power and significance testing in large-scale genetic studies. *Nature Reviews Genetics* **15**, 335–346 (2014).

[5] Wang, K., Li, M. & Hakonarson, H. Analysing biological pathways in genome-wide association studies. *Nature Reviews Genetics* **11**, 843–854 (2010).

[6] de Leeuw, C. A., Neale, B. M., Heskes, T. & Posthuma, D. The statistical properties of gene-set analysis. *Nature Reviews Genetics* **17**, 353–364 (2016).

[7] Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102**, 15545–15550 (2005).

[8] Carbonetto, P. & Stephens, M. Integrated enrichment analysis of variants and pathways in genome-wide association studies indicates central role for IL-2 signaling genes in type 1 diabetes, and cytokine signaling genes in Crohn's disease. *PLoS Genetics* **9**, e1003770 (2013).

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

[10] Carbonetto, P. & Stephens, M. Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Analysis* **7**, 73–108 (2012).

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

[12] Varadhan, R. & Roland, C. Simple and globally convergent methods for accelerating the convergence of any EM algorithm. *Scandinavian Journal of Statistics* **35**, 335–353 (2008).

[13] Wellcome Trust Case Control Consortium . Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).

[14] Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics* **47**, 1228–1235 (2015).

[15] Lamparter, D., Marbach, D., Rueedi, R., Kutalik, Z. & Bergmann, S. Fast and rigorous computation of gene and pathway scores from SNP-based summary statistics. *PLoS Computational Biology* **12**, e1004714 (2016).

[16] International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).

[17] 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

[18] Gao, X., Starmer, J. & Martin, E. R. A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genetic Epidemiology* **32**, 361–369 (2008).

[19] Liu, J. Z. *et al.* A versatile gene-based test for genome-wide association studies. *The American Journal of Human Genetics* **87**, 139–145 (2010).

[20] Li, M.-X., Gui, H.-S., Kwan, J. S. & Sham, P. C. GATES: a rapid and powerful gene-based association test using extended Simes procedure. *The American Journal of Human Genetics* **88**, 283–293 (2011).

[21] Wang, M. *et al.* COMBAT: A combined association test for genes using summary statistics. *Genetics* **207**, 883–891 (2017).

[22] van Rheenen, W. *et al.* Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. *Nature Genetics* **48**, 1043–1048 (2016).

[23] Okbay, A. *et al.* Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nature Genetics* **48**, 624–633 (2016).

[24] Lambert, J.-C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature Genetics* **45**, 1452–1458 (2013).

[25] Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).

[26] Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).

[27] Wood, A. R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics* **46**, 1173–1186 (2014).

[28] Shungin, D. *et al.* New genetic loci link adipose and insulin biology to body fat distribution. *Nature* **518**, 187–196 (2015).

[29] Liu, J. Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nature Genetics* **47**, 979–986 (2015).

[30] Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376–381 (2014).

[31] Day, F. R. *et al.* Large-scale genomic analyses link reproductive aging to hypothalamic signaling, breast cancer susceptibility and BRCA1-mediated DNA repair. *Nature Genetics* **47**, 1294–1303 (2015).

[32] Nikpay, M. *et al.* A comprehensive 1000 genomes–based genome-wide association meta-analysis of coronary artery disease. *Nature Genetics* **47**, 1121–1130 (2015).

[33] Manning, A. K. *et al.* A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nature Genetics* **44**, 659–669 (2012).

[34] Köttgen, A. *et al.* Genome-wide association analyses identify 18 new loci associated with serum urate concentrations. *Nature Genetics* **45**, 145–154 (2013).

[35] Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).

[36] Den Hoed, M. *et al.* Identification of heart rate-associated loci and their effects on cardiac conduction and rhythm disorders. *Nature Genetics* **45**, 621–631 (2013).

[37] Morris, A. P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature Genetics* **44**, 981–990 (2012).

[38] van der Harst, P. *et al.* Seventy-five genetic loci influencing the human red blood cell. *Nature* **492**, 369–375 (2012).

[39] Cerami, E. G. *et al.* Pathway Commons, a web resource for biological pathway data.

Nucleic Acids Research **39**, D685–D690 (2011).

[40] Geer, L. Y. *et al.* The NCBI BioSystems database. *Nucleic Acids Research* **38**, D492–D496 (2010).

[41] Varjosalo, M. & Taipale, J. Hedgehog: functions and mechanisms. *Genes & Development* **22**, 2454–2472 (2008).

[42] Teng, M. W. *et al.* IL-12 and IL-23 cytokines: from discovery to targeted therapies for immune-mediated inflammatory diseases. *Nature Medicine* **21**, 719–729 (2015).

[43] Nicholls, S. J. *et al.* Statins, high-density lipoprotein cholesterol, and regression of coronary atherosclerosis. *Journal of the American Medical Association* **297**, 499–508 (2007).

[44] Macian, F. NFAT proteins: key regulators of T-cell development and function. *Nature Reviews Immunology* **5**, 472–484 (2005).

[45] Sitara, D. & Aliprantis, A. O. Transcriptional regulation of bone and joint remodeling by NFAT. *Immunological Reviews* **233**, 286–300 (2010).

[46] Mackie, E., Ahmed, Y., Tatarczuch, L., Chen, K.-S. & Mirams, M. Endochondral ossification: how cartilage is converted into bone in the developing skeleton. *The International Journal of Biochemistry & Cell Biology* **40**, 46–62 (2008).

[47] Elshaer, S. L. & El-Remessy, A. B. Implication of the neurotrophin receptor p75NTR in vascular diseases: beyond the eye. *Expert Review of Ophthalmology* **12**, 149–158 (2017).

[48] McQueen, F. M., Chhana, A. & Dalbeth, N. Mechanisms of joint damage in gout: evidence from cellular and imaging studies. *Nature Reviews Rheumatology* **8**, 173–181 (2012).

[49] Rock, K. L., Kataoka, H. & Lai, J.-J. Uric acid as a danger signal in gout and its comorbidities. *Nature Reviews Rheumatology* **9**, 13 (2013).

[50] Di Paolo, G. & Kim, T.-W. Linking lipids to Alzheimer's disease: cholesterol and beyond. *Nature Reviews Neuroscience* **12**, 284–296 (2011).

[51] Beeri, M. S. *et al.* Coronary artery disease is associated with Alzheimer disease neuropathology in APOE4 carriers. *Neurology* **66**, 1399–1404 (2006).

[52] Heppner, F. L., Ransohoff, R. M. & Becher, B. Immune attack: the role of inflammation in Alzheimer disease. *Nature Reviews Neuroscience* **16**, 358–372 (2015).

[53] Pickrell, J. K. *et al.* Detection and interpretation of shared genetic influences on 42 human traits. *Nature Genetics* **48**, 709–717 (2016).

[54] Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nature Genetics* **47**, 1236–1241 (2015).

[55] Smemo, S. *et al.* Obesity-associated variants within *FTO* form long-range functional connections with *IRX3*. *Nature* **507**, 371–375 (2014).

[56] Guan, Y. & Stephens, M. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *The Annals of Applied Statistics* **5**, 1780–1815 (2011).

[57] Rader, D. J. & Kastelein, J. J. Lomitapide and mipomersen: Two first-in-class drugs for reducing low-density lipoprotein cholesterol in patients with homozygous familial hypercholesterolemia. *Circulation* **129**, 1022–1032 (2014).

[58] Global Lipids Genetics Consortium. Discovery and refinement of loci associated with lipids levels. *Nature Genetics* **45**, 1274–1283 (2013).

[59] Stephens, M. A unified framework for association analysis with multiple related phenotypes. *PLoS ONE* **8**, e65245 (2013).

[60] Vega, R. B. *et al.* Histone deacetylase 4 controls chondrocyte hypertrophy during skeletogenesis. *Cell* **119**, 555–566 (2004).

[61] Obri, A., Makinistoglu, M. P., Zhang, H. & Karsenty, G. HDAC4 integrates PTH and sympathetic signaling in osteoblasts. *The Journal of Cell Biology* **205**, 771–780 (2014).

[62] Cheloha, R. W., Gellman, S. H., Vilardaga, J.-P. & Gardella, T. J. PTH receptor-1 sig-

nalling – mechanistic insights and therapeutic prospects. *Nature Reviews Endocrinology* **11**, 712–724 (2015).

[63] Su, N., Jin, M. & Chen, L. Role of FGF/FGFR signaling in skeletal development and homeostasis: learning from mouse models. *Bone Research* **2**, 14003 (2014).

[64] Tang, S. Y., Herber, R.-P., Ho, S. P. & Alliston, T. Matrix metalloproteinase–13 is required for osteocytic perilacunar remodeling and maintains bone fracture resistance. *Journal of Bone and Mineral Research* **27**, 1936–1950 (2012).

[65] Steinberg, M. W. *et al.* A crucial role for HVEM and BTLA in preventing intestinal inflammation. *Journal of Experimental Medicine* **205**, 1463–1476 (2008).

[66] Dubois, P. C. *et al.* Multiple common variants for celiac disease influencing immune gene expression. *Nature Genetics* **42**, 295–302 (2010).

[67] Anderson, C. A. *et al.* Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nature Genetics* **43**, 246–252 (2011).

[68] Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119–124 (2012).

[69] Strasser, A., Jost, P. J. & Nagata, S. The many roles of FAS receptor signaling in the immune system. *Immunity* **30**, 180–192 (2009).

[70] Neurath, M. F. Cytokines in inflammatory bowel disease. *Nature Reviews Immunology* **14**, 329–342 (2014).

[71] Susan-Resiga, D. *et al.* Furin is the major processing enzyme of the cardiac-specific growth factor bone morphogenetic protein 10. *Journal of Biological Chemistry* **286**, 22785–22794 (2011).

[72] International Consortium for Blood Pressure Genome-Wide Association Studies. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature* **478**, 103–109 (2011).

[73] Silence, J., Lupu, F., Collen, D. & Lijnen, H. Persistence of atherosclerotic plaque but reduced aneurysm formation in mice with stromelysin-1 (MMP-3) gene inactivation. *Arteriosclerosis, Thrombosis, and Vascular Biology* **21**, 1440–1445 (2001).

[74] Aseem, O. *et al.* Cubilin maintains blood levels of HDL and albumin. *Journal of the American Society of Nephrology* **25**, 1028–1036 (2014).

[75] Kennedy, M. A. *et al.* ABCG1 has a critical role in mediating cholesterol efflux to HDL and preventing cellular lipid accumulation. *Cell Metabolism* **1**, 121–131 (2005).

[76] Anderson, G. D. *et al.* Selective inhibition of cyclooxygenase (COX)-2 reverses inflammation and expression of COX-2 and interleukin 6 in rat adjuvant arthritis. *Journal of Clinical Investigation* **97**, 2672 (1996).

[77] Kivitz, A., Eisen, G. & Zhao, W. W. Randomized placebo-controlled trial comparing efficacy and safety of valdecoxib with naproxen in patients with osteoarthritis. *Journal of Family Practice* **51**, 530–537 (2002).

[78] Daynes, R. A. & Jones, D. C. Emerging roles of PPARs in inflammation and immunity. *Nature Reviews Immunology* **2**, 748–759 (2002).

[79] Széles, L., Töröcsik, D. & Nagy, L. PPARγ in immunity and inflammation: cell types and diseases. *Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids* **1771**, 1014–1030 (2007).

[80] The GTEx Consortium. The genotype-tissue expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).

[81] Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: From polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).

[82] Dey, K. K., Hsiao, C. J. & Stephens, M. Visualizing the structure of RNA-seq expression data using grade of membership models. *PLoS Genetics* **13**, e1006599 (2017).

[83] Liu, C.-C., Kanekiyo, T., Xu, H. & Bu, G. Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy. *Nature Reviews Neurology* **9**, 106–118 (2013).

[84] Schwarzman, A. L. *et al.* Transthyretin sequesters amyloid beta protein and prevents amyloid formation. *Proceedings of the National Academy of Sciences* **91**, 8368–8372 (1994).

[85] Buxbaum, J. N. *et al.* Transthyretin protects Alzheimer's mice from the behavioral and biochemical effects of Aβ toxicity. *Proceedings of the National Academy of Sciences* **105**, 2681–2686 (2008).

[86] Velayudhan, L. *et al.* Plasma transthyretin as a candidate marker for Alzheimer's disease. *Journal of Alzheimer's Disease* **28**, 369–375 (2012).

[87] Hansson, S. F. *et al.* Reduced levels of amyloid-β-binding proteins in cerebrospinal fluid from Alzheimer's disease patients. *Journal of Alzheimer's Disease* **16**, 389–397 (2009).

[88] Sassi, C. *et al.* Influence of coding variability in APP-Aβ metabolism genes in sporadic Alzheimer's Disease. *PLoS ONE* **11**, e0150079 (2016).

[89] Xiang, Q. *et al.* Rare genetic variants of the transthyretin gene are associated with Alzheimer's disease in Han Chinese. *Molecular Neurobiology* 1–9 (2016).

[90] Kwak, I.-Y. & Pan, W. Adaptive gene-and pathway-trait association testing with GWAS summary statistics. *Bioinformatics* **32**, 1178–1184 (2016).

[91] Zhang, H. *et al.* A powerful procedure for pathway-based meta-analysis using summary statistics identifies 43 pathways associated with type II diabetes in European populations. *PLoS Genetics* **12**, e1006122 (2016).

[92] Slowikowski, K., Hu, X. & Raychaudhuri, S. SNPsea: an algorithm to identify cell types, tissues and pathways affected by risk loci. *Bioinformatics* **30**, 2496–2497 (2014).

[93] Pers, T. H. *et al.* Biological interpretation of genome-wide association studies using predicted gene functions. *Nature Communications* **6** (2015).

[94] Segrè, A. V. *et al.* Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *PLoS Genetics* **6**, e1001058 (2010).

[95] de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Computational Biology* **11**, e1004219 (2015).

[96] Evangelou, M., Dudbridge, F. & Wernisch, L. Two novel pathway analysis methods based on a hierarchical model. *Bioinformatics* **30**, 690–697 (2014).

[97] Schadt, E. E. *et al.* An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genetics* **37**, 710–717 (2005).

[98] Nicolae, D. L. *et al.* Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genetics* **6**, e1000888 (2010).

[99] Nica, A. C. *et al.* Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genetics* **6**, e1000895 (2010).

[100] He, X. *et al.* Sherlock: detecting gene-disease associations by matching patterns of expression QTL and GWAS. *The American Journal of Human Genetics* **92**, 667–680 (2013).

[101] Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genetics* **10**, e1004383 (2014).

[102] Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature Genetics* **48**, 481–487 (2016).

[103] Hormozdiari, F. *et al.* Colocalization of GWAS and eQTL signals detects target genes. *The American Journal of Human Genetics* **99**, 1245–1260 (2016).

[104] Wen, X., Pique-Regi, R. & Luca, F. Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. *PLoS Genetics* **13**, e1006646 (2017).

[105] Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics* **47**, 1091–1098 (2015).

[106] Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics* **48**, 245–252 (2016).

[107] Logsdon, B. A., Hoffman, G. E. & Mezey, J. G. A variational Bayes algorithm for fast and accurate multiple locus genome-wide association analysis. *BMC Bioinformatics* **11**, 58 (2010).

[108] Stegle, O., Parts, L., Durbin, R. & Winn, J. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Computational Biology* **6**, e1000770 (2010).

[109] Li, Z. & Sillanpää, M. J. Estimation of quantitative trait locus effects with epistasis by variational Bayes algorithms. *Genetics* **190**, 231–249 (2012).

[110] Papastamoulis, P., Hensman, J., Glaus, P. & Rattray, M. Improved variational Bayes inference for transcript expression estimation. *Statistical Applications in Genetics and Molecular Biology* **13**, 203–216 (2014).

[111] Raj, A., Stephens, M. & Pritchard, J. K. fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics* **197**, 573–589 (2014).

[112] Logsdon, B. A. *et al.* A variational Bayes discrete mixture test for rare variant association. *Genetic Epidemiology* **38**, 21–30 (2014).

[113] Loh, P.-R. *et al.* Efficient Bayesian mixed model analysis increases association power in large cohorts. *Nature Genetics* **47**, 284–290 (2015).

[114] Gopalan, P., Hao, W., Blei, D. & Storey, J. Scaling probabilistic models of genetic variation to millions of humans. *Nature Genetics* **48**, 1587 (2016).

[115] Montesinos-López, O. A. *et al.* A variational Bayes genomic-enabled prediction model with genotype × environment interaction. *G3: Genes, Genomes, Genetics* (2017).

[116] Mi, H. & Thomas, P. PANTHER pathway: an ontology-based pathway database coupled with data analysis tools. *Protein Networks and Pathway Analysis* 123–140 (2009).

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

Xiang Zhu
Department of Statistics
Stanford University
E-mail: xiangzhu@stanford.edu

Matthew Stephens
Department of Statistics
University of Chicago
and
Department of Human Genetics
University of Chicago
E-mail: mstephens@uchicago.edu