Manuscript Version: July 7, 2017

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

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.

[*] Correspondence should be addressed to X.Z. (xiangzhu@uchicago.edu) or M.S. (mstephens@uchicago.edu).

Enrichment analysis – also referred to as "pathway analysis" [5] or "gene set analysis" [6] – can help tackle both these problems. Instead of analyzing one 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 testing for enrichment is itself simple, there are many ways to implement it in practice, each with its own advantages and disadvantages. Here we build on a previous model-based 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 initial 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 model-based 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. A major contribution of our work here is to overcome this limitation, and provide an implementation [9] that requires only GWAS summary statistics (plus data on patterns of LD in 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, including several involving large GWAS meta-analyses. 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 data. Full searchable and browse-able results are available at `http://xiangzhu.github.io/rss-gsea/results`.

Results

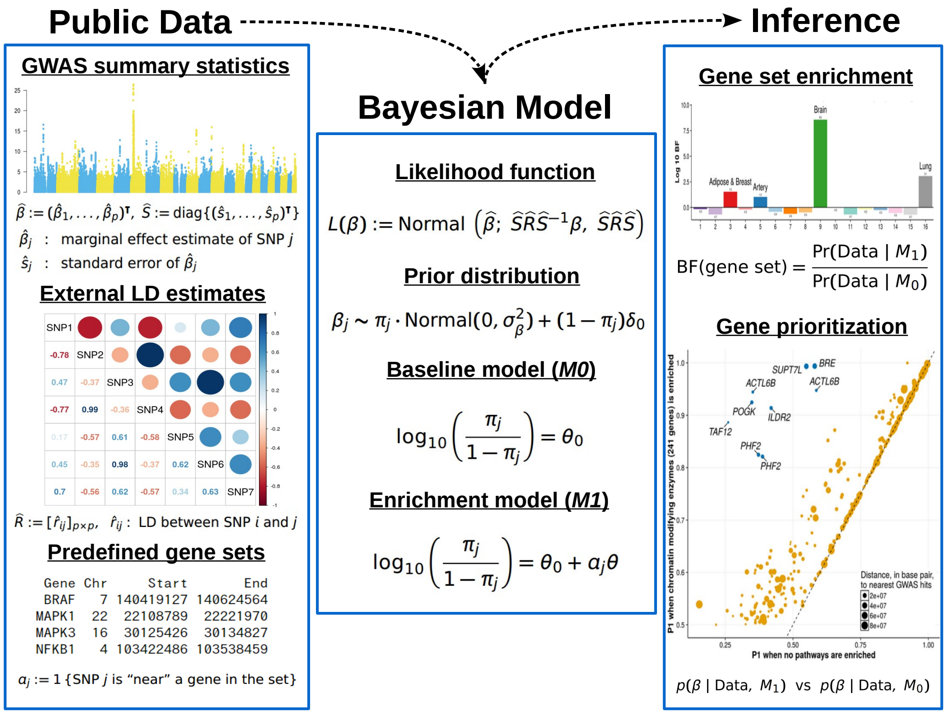
Method overview. Figure 1 provides a schematic overview of the method. In brief, the method combines the enrichment model from [8], with the multiple regression model for single-SNP association summary statistics from [9], to create a model-based enrichment method for GWAS summary data.

Specifically the method 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 ("SNP set"), the method estimates a (log10) "enrichment parameter", θ, which measures the extent to which SNPs in the set are more often associated with the phenotype. For example, $\theta = 2$ 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 the method uses a multiple regression model to account for LD among SNPs. For example, the method will (correctly) treat data from several SNPs that are in perfect LD as effectively a single observation, and not multiple independent observations. The method ultimately summarizes the evidence for enrichment by a Bayes factor (BF) comparing the *enrichment model* ($\theta > 0$) against the *baseline model* ($\theta = 0$). It also provides posterior distributions of genetic effects (β) to identify significant variants within enriched sets. See Methods for details.

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 this method. Our algorithm exploits variational inference [10], banded matrix approximation [11] and an expectation-maximization algorithm accelerator [12] (Methods; Supplementary Note). Software is available at `https://github.com/stephenslab/rss`.

Multiple regression on 1.1 million variants across 31 traits. The first step of our analysis is a multiple regression analysis of 1.1 million common SNPs for 31 phenotypes, using publicly available GWAS summary statistics from 20,883-253,288 European ancestry individuals (Supplementary Table 1; Supplementary Fig. 1). This step essentially estimates, for each trait, a "baseline model" against which enrichment hypotheses can be compared. The


Public Data
Inference
Bayesian Model
GWAS summary statistics
$\widehat{\beta} := (\widehat{\beta}_1, \ldots, \widehat{\beta}_p)^\intercal,\ \widehat{S} := \mathrm{diag}\{(\widehat{s}_1, \ldots, \widehat{s}_p)^\intercal\}$
$\widehat{\beta}_j$: marginal effect estimate of SNP j
$\widehat{s}_j$: standard error of $\widehat{\beta}_j$
External LD estimates
SNP1
-0.78 SNP2
0.47 -0.37 SNP3
-0.77 0.99 -0.36 SNP4
0.17 -0.57 0.61 -0.58 SNP5
0.45 -0.35 0.98 -0.37 0.62 SNP6
0.7 -0.56 0.62 -0.57 0.34 0.63 SNP7
$\widehat{R} := [\widehat{r}_{ij}]_{p \times p},\ \widehat{r}_{ij}$: LD between SNP i and j
Predefined gene sets
$a_j := 1\,\{\text{SNP } j \text{ is "near" a gene in the set}\}$
Likelihood function
$L(\beta) := \mathrm{Normal}\left(\widehat{\beta};\ \widehat{S}\widehat{R}\widehat{S}^{-1}\beta,\ \widehat{S}\widehat{R}\widehat{S}\right)$
Prior distribution
$\beta_j \sim \pi_j \cdot \mathrm{Normal}(0, \sigma_\beta^2) + (1-\pi_j)\delta_0$
Baseline model (M0)
$\log_{10}\left(\frac{\pi_j}{1-\pi_j}\right) = \theta_0$
Enrichment model (M1)
$\log_{10}\left(\frac{\pi_j}{1-\pi_j}\right) = \theta_0 + a_j\theta$
Gene set enrichment
Log 10 BF
Adipose & Breast
Artery
Brain
Lung
$\mathrm{BF(gene\ set)} = \frac{\Pr(\mathrm{Data} \mid M_1)}{\Pr(\mathrm{Data} \mid M_0)}$
Gene prioritization
SUPT7L
BRE
ACTL6B
ACTL6B
POGK
ILDR2
TAF12
PHF2
PHF2
P1 when chromatin modifying enzymes (241 genes) is enriched
P1 when no pathways are enriched
Distance, in base pair, to nearest GWAS hits
2e+07
4e+07
6e+07
8e+07
$p(\beta \mid \mathrm{Data},\ M_1)$ vs $p(\beta \mid \mathrm{Data},\ M_0)$

Gene	Chr	Start	End
BRAF	7	140419127	140624564
MAPK1	22	22108789	22221970
MAPK3	16	30125426	30134827
NFKB1	4	103422486	103538459

Fig 1: **Schematic overview of model-based enrichment analysis method for GWAS summary statistics.** The method combines three types of data: GWAS summary statistics, external LD estimates, and predefined SNP sets, which here we derive from gene sets based on biological pathways or other sources (Methods). 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 by conventional analyses in GWAS. External LD estimates are obtained from an external reference panel with ancestry matching the population of GWAS cohorts. We combine these three types of data by fitting a Bayesian multiple regression model under two hypotheses about the enrichment parameter (θ): the *baseline hypothesis* that each SNP has equal chance of being associated with the trait ($\theta = 0$), and the *enrichment hypothesis* that SNPs in the SNP set are more often associated with the trait ($\theta > 0$). To test for enrichment, the method computes a Bayes factor (BF) comparing these two hypotheses. The method also automatically prioritizes SNPs within an enriched set, facilitating the discovery of new trait-associated SNPs and genes.

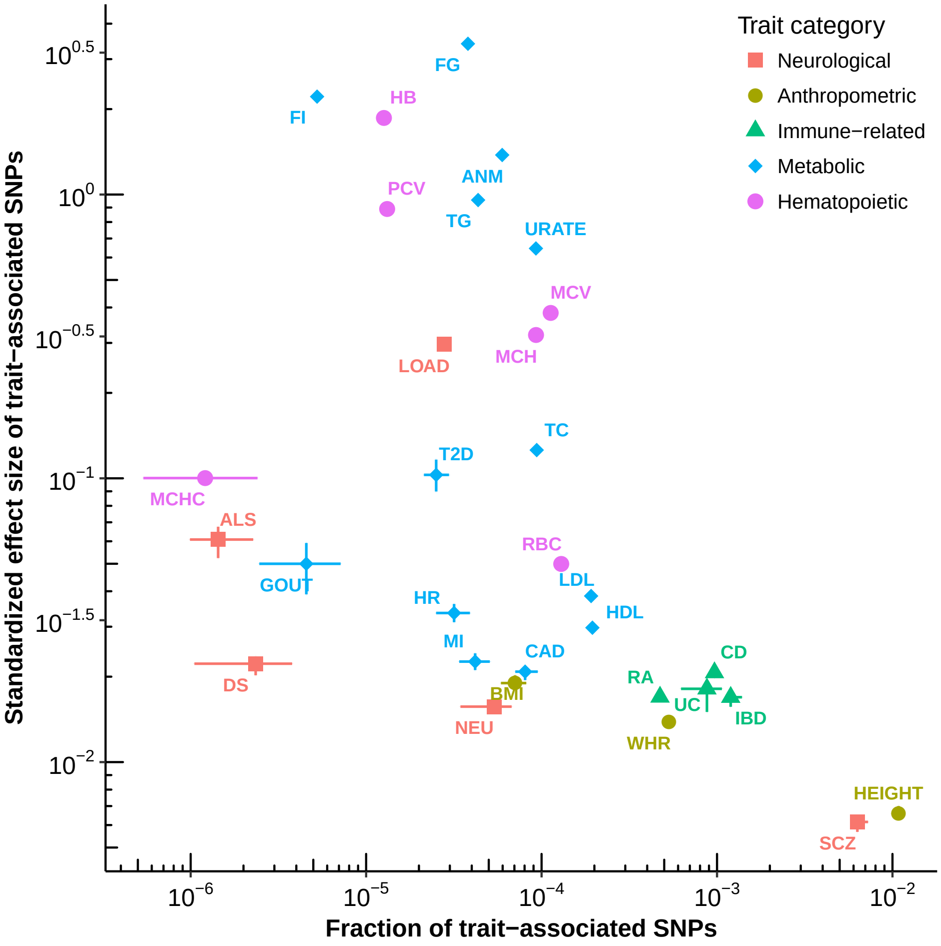


Fig 2: **Summary of inferred genetic architecture of 31 phenotypes.** Results are from fitting the baseline model ($\theta = 0$) to 1.1 million common SNPs for each trait. We summarize genetic architecture using two numbers: the estimated fraction of trait-associated SNPs (a measure of “polygenicity"; x-axis) and the standardized effect size of trait-associated SNPs (y-axis). See Supplementary Note for details on computing these two quantities. Each dot represents a trait, with horizontal and vertical point ranges indicating posterior mean and 95% credible interval (C.I.) for each estimate. ALS: amyotrophic lateral sclerosis [13]. DS: depressive symptoms [14]. LOAD: late-onset Alzheimer’s disease [15]. NEU: neuroticism [14]. SCZ: schizophrenia [16]. BMI: body mass index [17]. HEIGHT: adult height [18]. WHR: waist-to-hip ratio [19]. CD: Crohn’s disease [20]. IBD: inflammatory bowel disease [20]. RA: rheumatoid arthritis [21]. UC: ulcerative colitis [20]. ANM: age at natural menopause [22]. CAD: coronary artery disease [23]. FG: fasting glucose [24]. FI: fasting insulin [24]. GOUT: Gout [25]. HDL: high-density lipoprotein [26]. HR: heart rate [27]. LDL: low-density lipoprotein [26]. MI: myocardial infarction [23]. T2D: type 2 diabetes [28]. TC: total cholesterol [26]. TG: triglycerides [26]. URATE: serum urate [25]. HB: haemoglobin [29]. MCH: mean cell HB [29]. MCHC: MCH concentration [29]. MCV: mean cell volume [29]. PCV: packed cell volume [29]. RBC: red blood cell count [29].

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 phenotype (Fig. 2; Supplementary Fig. 2).

The results emphasize that genetic architecture varies considerably among phenotypes: estimates of both polygenicity and effect sizes vary by several orders of magnitude (Fig. 2). 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 architecture from GWAS summary statistics.

Fitting the baseline model also yields an estimate of the effect size (specifically, the multiple regression coefficient β) for each SNP. These can be used to identify trait-associated SNPs and loci. Reassuringly, these multiple-SNP results recapitulate many associations detected in previous single-SNP analyses of the same data (Supplementary Fig.s 3-5). For several traits, these results also identify additional putative associations (Supplementary Fig.s 6-7). 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, retrieved from nine databases (BioCarta, BioCyc, HumanCyc, KEGG, miRTarBase, PANTHER, PID, Reactome, WikiPathways); see Supplementary Figures 8-9. For each trait-pathway pair we compute a BF testing the enrichment hypothesis, and estimate the enrichment parameter θ.

Since these analyses involve large-scale computations that are subject to approximation error, we also developed some simpler methods for confirming enrichments identified by this approach. Specifically these simpler methods confirm that the z-scores for SNPs inside a putatively-enriched pathway have a different distribution from those outside the pathway (with more z-scores away from 0) – using both a likelihood ratio statistic and a visual check (Fig. 4a; Supplementary Fig. 10). We also filtered out enrichments that were most

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 enriched $P_1 > 0.9$ (denoted as "signals"); and iv) passes the sanity checks (Supplementary Fig. 10). All BFs reported here are larger than the corresponding BFs that SNPs near a gene are enriched (Supplementary Fig. 11). 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 [30]. BS: NCBI BioSystems [31]. a: The full pathway name is "transport of glucose and other sugars, bile salts and organic acids, metal ions and amine compounds".

likely driven by a single gene, both because these seem better represented as a gene association than a pathway enrichment, and because we found these to be more prone to artifacts (Discussion). Finally, since genic regions may be generally enriched for associations compared with non-genic regions, we checked that top-ranked pathways often showed stronger evidence for enrichment than did the set containing all genes (Supplementary Fig. 11).

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. 12). 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. 13). 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 [32] and previous analyses [18]. Other examples include *interleukin-23 mediated signaling* with inflammatory bowel disease (BF=3.1×10^{23}; [33]), *T helper cell surface molecules* with rheumatoid arthritis (BF=3.2×10^8; [21]), *statin pathway* with levels of high-density lipoprotein cholesterol (BF=8.4×10^{113}; [34]), and *glucose transporters* with serum urate (BF=$1.2 \times 10^{1,558}$; [25]).

The results also highlight several pathway enrichments that were not reported in the 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 [35] and bone pathology [36]. Other examples of novel pathway enrichments include *endochondral ossification* with adult height (BF=7.7×10^{68}; [37]), *p75 neurotrophin receptor-mediated signaling* with coronary artery disease (BF=9.6×10^{15}; [38]), and *osteoblast signaling* with gout (BF=3.8×10^{30}; [39]).

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. 3; 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. 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 [40], vascular disorder [41] and immune activation [42]. 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 [43] and co-heritability analyses [44].

Novel trait-associated genes informed by enriched pathways. A key feature of our method is that once an enriched pathway is identified this information can be used to improve association detection, and "prioritize" associations at variants near genes in the pathway. Specifically, the estimated enrichment parameter (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 [45]. 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, 46], 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. 14), 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. 4a) informs a strong association between gene *MTTP* (baseline P_1: 0.14; enriched P_1: 0.99) and levels of low-density lipoprotein (LDL) choles-

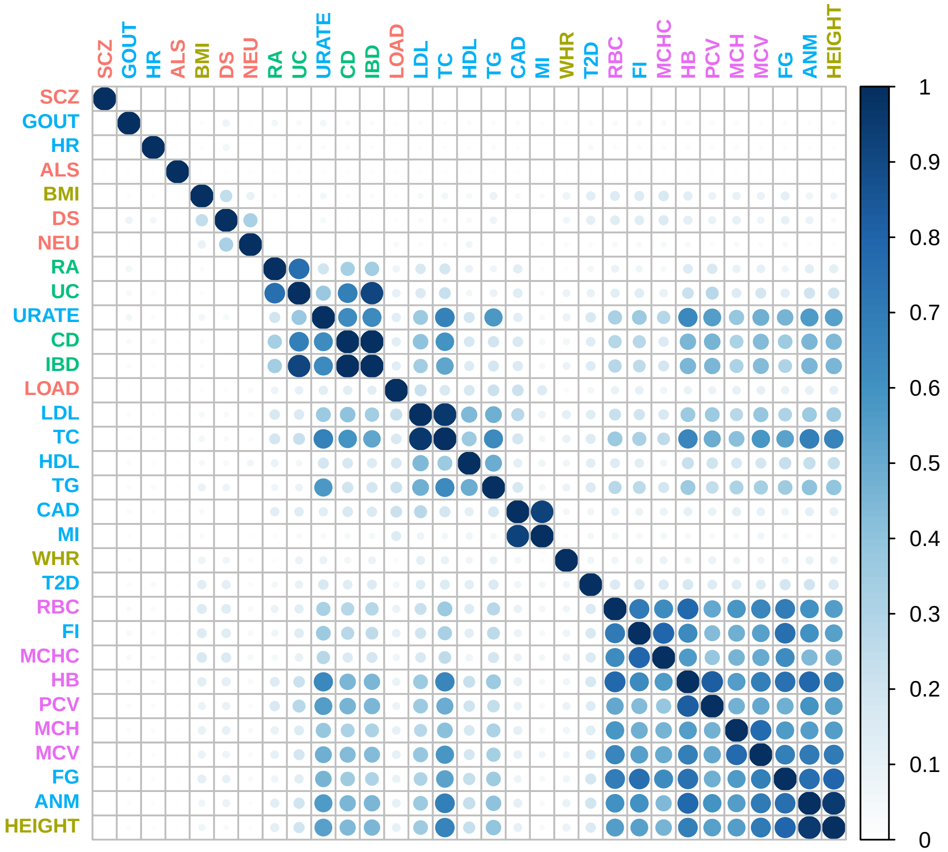


Fig 3: **Pairwise sharing of pathway enrichments among 31 traits.** For each pair of traits, we estimated the proportion of pathways that are enriched in both traits, among pathways enriched in at least one of the traits (Methods). Darker color and larger shape represent higher sharing. Traits are colored by categories and labeled by abbreviations (Fig. 2), and clustered by hierarchical clustering as implemented in `R` package `corrplot`.

terol (Fig. 4b). 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 [47]. However, no genome-wide significant SNPs near *MTTP* were reported in single-SNP analyses of either the same data [26] (Fig. 4c), or more recent data with larger sample size [48] (Fig. 4d).

Prioritization analysis of this same *chylomicron-mediated lipid transport* pathway also yields several additional plausible associations (Fig. 4b). These include *LIPC* (baseline P_1: 0.02; enriched P_1: 0.96) and *LPL* (baseline P_1: 0.01; enriched P_1: 0.76). These genes play important roles in lipid metabolism and both reach genome-wide significance in single-SNP analyses of blood lipids [26] although not for LDL cholesterol (Supplementary Fig. 15); and a multiple-trait, single-SNP analysis [49] 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.

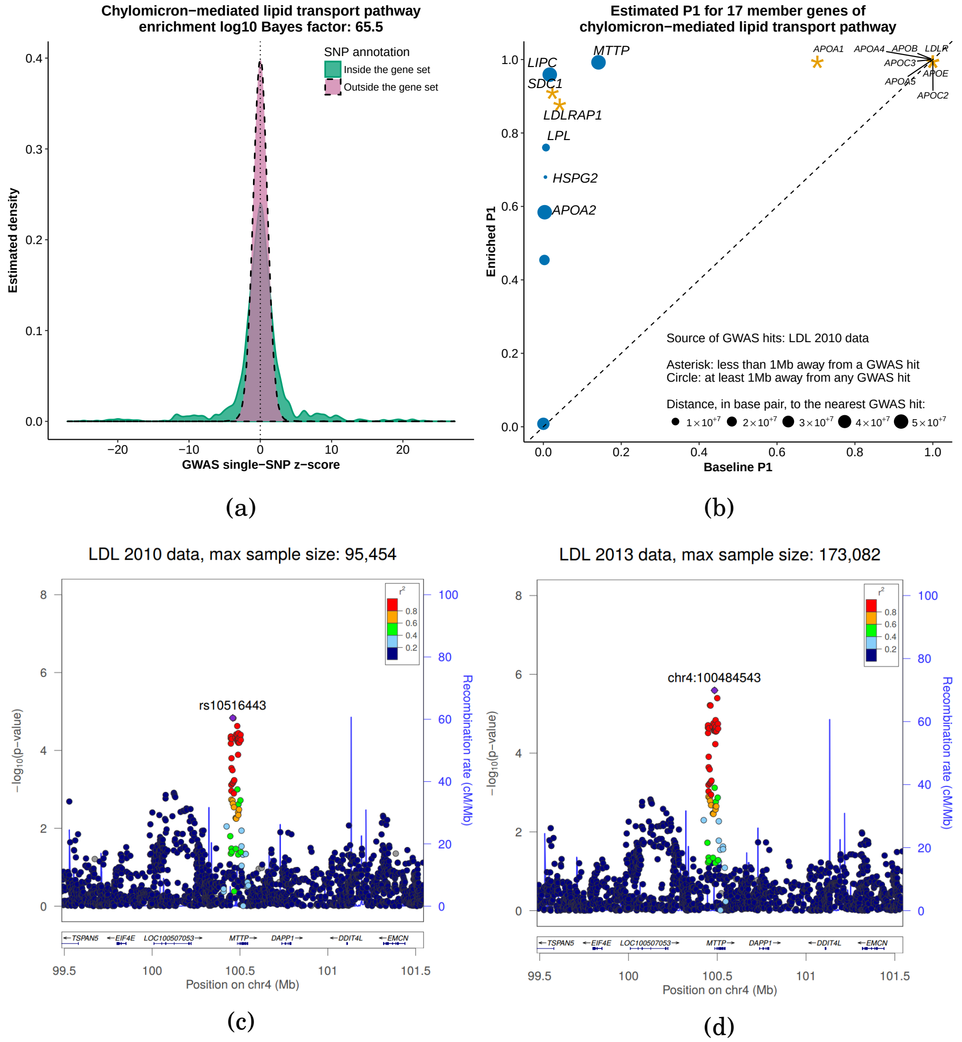
Box 1 Select putatively novel associations from prioritization analyses

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

- *HDAC4* (baseline P_1: 0.98; enriched P_1: 1.00)
  HDAC4 encodes a critical regulator of chondrocyte hypertrophy during skeletogenesis [50] and osteoclast differentiation [51]. 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; enriched P_1: 1.00)
  PTH1R encodes a receptor that regulates skeletal development, bone turnover and mineral ion homeostasis [52]. Mutations in *PTH1R* cause several rare skeletal disorders (OMIM: 215045, 600002, 156400).
- *FGFR1* (baseline P_1: 0.67; enriched P_1: 0.97)
  FGFR1 encodes a receptor that regulates limb development, bone formation and phosphorus metabolism [53]. Mutations in *FGFR1* cause several skeletal disorders (OMIM: 101600, 123150, 190440, 166250).
- *MMP13* (baseline P_1: 0.45; enriched P_1: 0.93)
  MMP13 encodes a protein that is required for osteocytic perilacunar remodeling and bone quality maintenance [54]. 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; enriched 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 [55]. *TNFRSF14* is near a GWAS



(a)

(b)

(c)

(d)

Fig 4: **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 [26], 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"). **(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 hypothesis versus estimated P_1 under the null hypothesis. **(c)** Regional association plot for *MTTP* based on summary data published in 2010 [26]. **(d)** Regional association plot for *MTTP* based on summary data published in 2013 [48].

hit of celiac disease (rs3748816, $p = 3.3 \times 10^{-9}$) [56] and two hits of ulcerative colitis (rs734999, $p = 3.3 \times 10^{-9}$ [57]; rs10797432, $p = 3.0 \times 10^{-12}$ [58]).

- *FAS* (baseline P_1: 0.82; enriched P_1: 0.99)
  FAS plays many important roles in the immune system [59]. Mutations in *FAS* cause autoimmune lymphoproliferative syndrome (OMIM: 601859).
- *IL6* (baseline P_1: 0.27; enriched 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 [60].

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

- *FURIN* (baseline P_1: 0.69; enriched P_1: 0.99)
  FURIN encodes the major processing enzyme of a cardiac-specific growth factor, which plays a critical role in heart development [61]. *FURIN* is near a GWAS hit (rs2521501 [62]) 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; enriched 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 [63].

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

- *CUBN* (baseline P_1: 0.24; enriched P_1: 1.00)
  CUBN encodes a receptor for intrinsic factor-vitamin B12 complexes (cubilin) that maintains blood levels of HDL [64]. 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, $p = 3.0 \times 10^{-11}$ [48]).
- *ABCG1* (baseline P_1: 0.01; enriched P_1: 0.89)
  ABCG1 encodes an ATP-binding cassette transporter that plays a critical role in mediating efflux of cellular cholesterol to HDL [65].

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

- *PTGS2* (a.k.a. *COX2*; baseline P_1: 0.74; enriched P_1: 0.98)
  PTGS2-specific inhibitors have shown efficacy in reducing joint inflammation in both mouse models [66] and clinical trials [67]. *PTGS2* is near a GWAS hit of Crohn's disease (rs10798069, $p = 4.3 \times 10^{-9}$ [20])
- *PPARG* (baseline P_1: 0.28; enriched P_1: 0.98)
  PPARG has important roles in regulating inflammatory and immune responses with potential applications in treating chronic inflammatory diseases including RA [68, 69].

Enrichment analysis of 113 tissue-based gene sets across 31 traits. Our enrichment method 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 (GTEx) project [70] to define sets of the most "relevant" genes in each tissue, based on expression patterns across tissues (Methods). 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 [71] ? 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 approaches to defining tissue-relevant genes (Methods). The first approach ("highly expressed", HE) uses the highest expressed genes in each tissue. The second approach ("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 approach ("distinctively expressed", DE) clusters the tissue samples and identifies genes that are most informative for distinguishing each cluster from others [72]. This last approach 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 gene sets to tissues.

Phenotype	Tissue (Annotation rule)	$\log_{10}$ BF	Select top driving genes (# of genes with $P_1 > 0.9$)
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)	$1,020.0^b$	*UBXN6*, *RBM38*, *NPRL3* (11)
Red blood cell count	Breast (SE)	141.7	*OBP2B*, *STAC2* (2)
Packed cell volume	Heart (HE)	945.4	*RPL19*, *TCAP* (2)

TABLE 2

Top enriched tissue-based gene sets in complex traits. Each tissue-based gene set contains 100 transcribed genes used in GTEx project. For each trait here 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 enriched $P_1 > 0.9$. All trait-tissue pairs reported here pass the sanity checks (Supplementary Fig. 10). HE: highly expressed. SE: selectively expressed. DE: distinctively expressed. a: Multiple tissues show partial membership in Cluster 1 [72]. b: These three BFs are smaller than the corresponding BFs that SNPs near a gene are enriched (Supplementary Fig. 11).

Despite the small number of tissue-based gene sets relative to the pathway analyses above, this analysis identifies many strong enrichments (URLs). 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, genes *highly* expressed in heart show strongest enrichment for heart rate; genes *selectively* expressed in liver show strongest enrichment for LDL. This highlights the benefits of considering multiple annotation strategies, and suggests 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 (URLs). For example, associations in coronary artery disease are strongly enriched in both *heart*-related ($\mathrm{BF} = 6.6 \times 10^7$) and *brain*-related ($\mathrm{BF} = 3.5 \times 10^8$) genes. 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 ($\mathrm{BF}>10^{30}$): liver, brain and adrenal gland. Because of the well-known connection between gene *APOE* and LOAD [73], and the fact that *APOE* is highly expressed in these three tissues (Supplementary Fig. 16), we hypothesized that *APOE* and related genes might be driving these results. To assess this we re-ran the enrichment analyses after removing the entire apolipoproteins (APO) gene family from the gene sets. Of the three tissues, only liver remains (moderately) enriched after excluding APO genes (Fig. 5), 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 (Fig. 5). This highlighted an association of LOAD with gene *TTR* (baseline P_1: 0.64; enriched P_1: 1.00; Supplementary Fig. 17). *TTR* encodes transthyretin, which has been shown to inhibit LOAD-related protein from forming harmful aggregation and toxicity, *in vitro* [74] and *in vivo* [75]. Indeed, transthyretin is considered a biomarker for LOAD: patients show reduced transthyretin levels in plasma [76] and cerebrospinal fluid [77]. And rare variants in *TTR* have recently been found to be associated with LOAD [78, 79]. By integrating GWAS with expression data our analysis identifies association of *TTR* based on common variants.

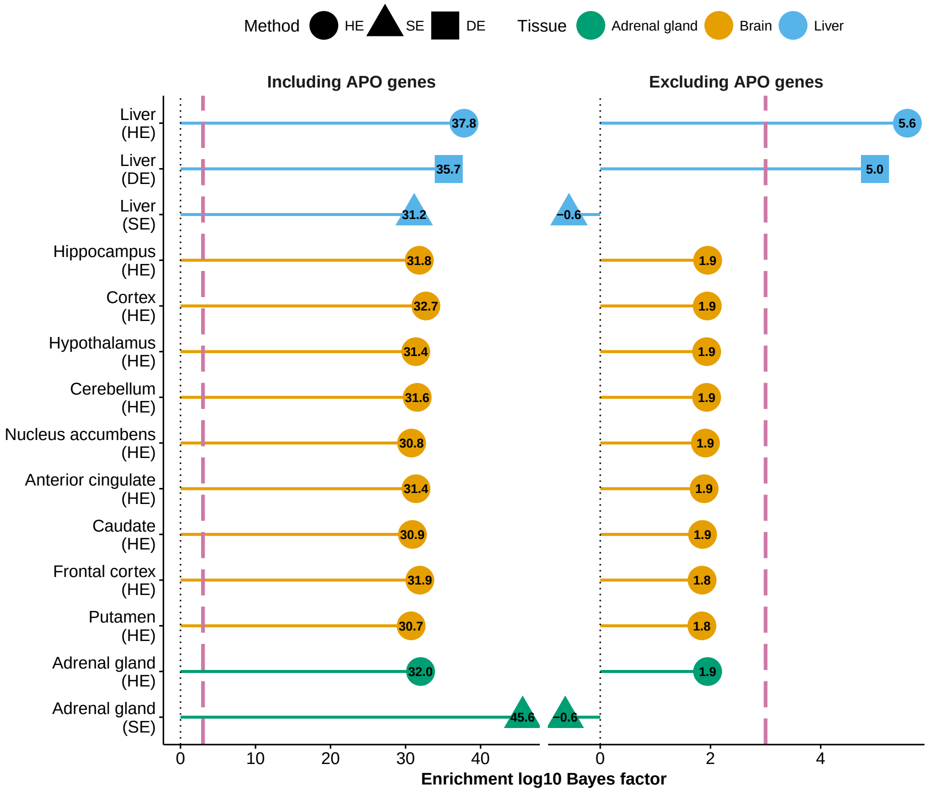


Fig 5: **Enrichment analyses of genes related to liver, brain and adrenal gland for Alzheimer's disease.** Shown are the tissue-based gene sets with the strongest enrichment signals for Alzheimer's disease. Each gene set was analyzed twice: the left panel corresponds to the analysis based on the original 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. HE: highly expressed. SE: selectively expressed. DE: distinctively expressed.

Discussion

We have presented a new method for 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 useful insights.

Enrichment tests, sometimes known as "competitive tests" [5, 6], have several advantages over alternative approaches – sometimes known as "self-contained tests" (e.g. [80, 81]) – 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 2.)

Compared with other enrichment approaches, our method has several particularly attractive features. First, unlike many methods (e.g. [5, 82, 83]) our method 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, our method models enrichment directly as an increased rate of association of variants within a SNP set. This contrasts with alternative two-stage approaches (e.g. [84–86]) 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 [45]. 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, 87] 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. [88–95]). 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. [96, 97]), our approach focuses firmly on genotype-phenotype associations. Nonetheless, as our results from different annotations demonstrate, it can be extremely beneficial to consider multiple approaches, and we view these methods as complimentary rather than competing.

Like any method, our approach 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 genes that are "selectively expressed" 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) highly expressed in testis, and b) physically close (53 kb) to another gene, *GCKR*, that is strongly associated with lipid traits (Supplementary Fig. 18). This highlights the need for careful examination of results, and also the utility of prioritization analyses. Generally we view enrichments that are driven by a single gene as less reliable and useful than enrichments driven by multiple genes. 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 (Supplementary Fig. 13); and b) for some traits (e.g. height), genetic associations may be strongly enriched near all genes (Supplementary Fig. 11), which will cause many pathways to appear enriched.

Other limitations of our method 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, 98–106]), 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, and indeed we occasionally observed convergence to what appeared to be unreliable estimates (Methods). This said, the main alternative for making Bayesian calculations, Markov chain Monte Carlo, can experience similar difficulties.

Finally, the present study focuses on testing a single annotation (e.g. one gene set) at a time. Extending the method to jointly analyze multiple annotations (e.g. [107–109]) 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; Full results, https://xiangzhu.github.io/rss-gsea/results; 1000 Genomes, http://www.internationalgenome.org; OMIM, https://www.omim.org; GTEx Portal, https://www.gtexportal.org; APO gene family: http://www.genenames.org/cgi-bin/genefamilies/set/405; ggplot2, http://ggplot2.tidyverse.org; corrplot, https://cran.r-project.org/web/packages/corrplot.

Acknowledgments. This research is supported by the Grant GBMF # 4559 from the Gordon and Betty Moore Foundation and NIH Grant HG02585 to M.S. We thank all the GWAS consortia for making their summary statistics publicly available. A full list of acknowledgments appears 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 analysis results of GTEx gene expression data. This work was completed in part with resources provided by the University of Chicago Research Computing Center.

Author Contributions. X.Z. and M.S. conceived the idea and designed the study. X.Z. and M.S. developed and refined 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 phenotypes (Supplementary Note;

Supplementary Table 1), in particular, the estimated single-SNP effect size and its standard error for each SNP.

For all 31 traits, we analyze the same set of SNPs in the HapMap 3 reference panel [110], since LD among these SNPs can be reliably estimated from existing panels [111]. To further ensure the quality of LD estimates, we also *exclude* SNPs with minor allele frequency less than 1%, SNPs in the major histocompatibility complex region, and SNPs measured on custom arrays from our analyses. The final set of variants retained for analyses consists of ~ 1.1 million SNPs (Supplementary Fig. 1).

Since the analyzed GWAS summary statistics were all generated from European ancestry individuals, we use phased haplotypes of 503 Europeans from the 1000 Genomes Project, Phase 3 [112] to estimate LD [11].

To create SNP-level annotations for a given gene set, we use a distance-based approach [8, 84]. 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 regions, 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) [30], NCBI Biosystems [31], PANTHER (version 3.3) [113] 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 8-9.

Tissue-based gene sets derived from GTEx transcriptomic data. Complex traits are often affected by multiple tissues, and it is not obvious *a priori* what the most relevant tissues are for any given human phenotype. Hence, it is necessary to examine a comprehensive set of tissues. The breadth of tissues in GTEx project [70] 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 approach ("highly expressed" or HE) 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. Here we focus on 44 tissues with sample sizes greater than 70. We downloaded these 44 gene lists from the GTEx Portal on November 21, 2016.

The second approach ("selectively expressed" or SE) 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 unpublished gene lists of 49 tissues from Dr. Simon Xi on February 13, 2017.

The third approach ("distinctively expressed" or DE) summarizes 53 tissues as 20 biologically-distinct clusters using grade of membership 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 [72]. We downloaded these 20 gene lists from `http://stephenslab.github.io/count-clustering` 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 regression with summary statistics (RSS) likelihood [9]:

$$(1) \qquad L_{\text{rss}}(\boldsymbol{\beta}) := \mathcal{N}(\widehat{\boldsymbol{\beta}}; \widehat{S}\widehat{R}\widehat{S}^{-1}\boldsymbol{\beta}, \widehat{S}\widehat{R}\widehat{S})$$

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 SNPs under the multiple-SNP model, 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 [46], to specify the following prior on $\boldsymbol{\beta}$:

$$\begin{array}{llcl}
(2) & \beta_j & \sim & \pi_j \mathcal{N}(0, \sigma_\beta^2) + (1 - \pi_j)\delta_0, \\
(3) & \sigma_\beta^2 & = & h \cdot \left(\sum_{j=1}^p \pi_j n^{-1} \hat{s}_j^{-2}\right)^{-1}, \\
(4) & \pi_j & = & (1 + 10^{-(\theta_0 + a_j \theta)})^{-1},
\end{array}$$

where δ_0 denotes point mass at zero, θ_0 reflects the background proportion of trait-associated SNPs under the multiple-SNP model, θ reflects the increase in probability, on the log10-odds scale, that a SNP inside the gene set has nonzero genetic 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. We place independent uniform grid priors on the hyper-parameters $\{\theta_0, \theta, h\}$ (Supplementary Tables 3-4).

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]. Firstly, 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 posterior probabilities. Finally, we integrate out the conditional posterior of $\boldsymbol{\beta}$ over the posterior of $\{\theta_0, \theta, h\}$ to obtain the posterior of $\boldsymbol{\beta}$.

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

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.

Initialize the variational Bayes algorithm. Following previous work [10], we use a coordinate ascent algorithm in the variational Bayes approximation step, which only guarantees convergence to a local optimum, and thus is potentially sensitive to initialization. By default, we randomly select an initialization, and then use the same initial value for all variational approximations over the grid of $\{\theta_0, \theta, h\}$. The random initialization seems to work well enough in most of our analyses.

For a few traits (e.g. triglyceride), the default random start approach produces inconsistent posterior results (Supplementary Fig. 19), which may be due to convergence to a local maximum of variational lower bound. To address this issue, we first fit the model using a modified variational algorithm that jointly estimates $\boldsymbol{\beta}$ and θ_0 (Supplementary Note), and then use the solution from the modified algorithm to initialize future variational approximations. We test this initialization strategy on triglyceride [26], and obtain improved posterior results (Supplementary Fig. 20).

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):

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

where $\mathbf{a} := (a_1, \dots, a_p)^\intercal$ and a_j indicates whether SNP j is inside the gene set. The observed data are BF times more likely under the enrichment hypothesis ($\theta > 0$) than under the baseline hypothesis ($\theta = 0$), and so the larger the BF, the stronger evidence for gene set enrichment. See Supplementary Note for details of computing enrichment BF. The BF threshold is 10^8 for the analyses of 3,913 pathways, and the threshold is 10^3 for 113 tissue-based gene sets.

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 phenotype:

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

where $\mathbf{D}$ 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 pathway enrichments between two traits (Fig. 3), we define a parameter $\pi := (\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11})$ as follows:

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

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 `R` package `ashr` [114]. Finally, the conditional probability that a pathway is enriched in a pair of phenotypes given that it is enriched in at least one phenotype, as plotted in Figure 3, is estimated as $\hat{\pi}_{11}/(1-\hat{\pi}_{00})$.

Connection with enrichment analysis of individual-level data. Our method has close connection with the method developed for individual-level data, which we temporarily refer to as the CS (Carbonetto-Stephens) method [8]. The key difference is that CS method uses a multiple-SNP likelihood based on individual-level genotypes and phenotypes, whereas our method uses a multiple-SNP likelihood based on GWAS summary statistics [9]. These two likelihoods are mathematically equivalent under certain conditions [9]. Under the same conditions, here we further show that our method and CS method are also mathematically equivalent, in the sense that they have the same fix point iteration scheme and lower bound used in variational Bayes approximations. See Supplementary Note for details.

In addition to their theoretical connections, we also empirically compare two methods through a wide range of simulations. Both methods produce similar inferential results, including parameter estimation (θ_0 and θ), type 1 error and power of detecting enrichment (Supplementary Fig.s 21-22).

Code and data availability. Code: https://github.com/stephenslab/rss. Full results: http://xiangzhu.github.io/rss-gsea/results. Demonstration: http://stephenslab.github.io/rss/Example-5. 1000 Genomes Phase 3: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502. GTEx RNA-seq: https://gtexportal.org/home/datasets. 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. Links to GWAS summary statistics are provided in Supplementary Note.

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 (2017). *The Annals of Applied Statistics*, To appear.

[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] 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).

[14] 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).

[15] 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).

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

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

[18] 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).

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

[20] 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).

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

[22] 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).

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

[24] 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).

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

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

[27] 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).

[28] 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).

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

[30] Cerami, E. G. *et al.* Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Research* **39**, D685–D690 (2011).

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

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

[33] 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).

[34] 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).

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

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

[37] 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).

[38] 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).

[39] 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).

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

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

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

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

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

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

[46] 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).

[47] 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).

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

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

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

[51] 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).

[52] Cheloha, R. W., Gellman, S. H., Vilardaga, J.-P. & Gardella, T. J. PTH receptor-1 signalling – mechanistic insights and therapeutic prospects. *Nature Reviews Endocrinol-*

ogy **11**, 712–724 (2015).

[53] 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).

[54] 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).

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

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

[57] 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).

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

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

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

[61] 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).

[62] 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).

[63] 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).

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

[65] 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).

[66] 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).

[67] 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).

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

[69] 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).

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

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

[72] 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).

[73] 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).

[74] 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).

[75] 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).

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

[77] 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).

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

[79] 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).

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

[81] 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).

[82] 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).

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

[84] 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).

[85] 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).

[86] 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).

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

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

[89] 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).

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

[91] 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).

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

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

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

[95] 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).

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

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

[98] 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).

[99] 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).

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

[101] 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).

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

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

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

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

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

[107] Pickrell, J. K. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *The American Journal of Human Genetics* **94**, 559–573 (2014).

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

[109] Li, Y. & Kellis, M. Joint bayesian inference of risk variants and tissue-specific epigenomic enrichments across multiple complex human diseases. *Nucleic Acids Research* **44**, e144 (2016).

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

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

[112] 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

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

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

XIANG ZHU
DEPARTMENT OF STATISTICS
UNIVERSITY OF CHICAGO
5747 S. ELLIS AVENUE
CHICAGO, ILLINOIS 60637
USA
E-MAIL: xiangzhu@uchicago.edu

MATTHEW STEPHENS
DEPARTMENT OF STATISTICS
UNIVERSITY OF CHICAGO
5747 S. ELLIS AVENUE
AND
DEPARTMENT OF HUMAN GENETICS
UNIVERSITY OF CHICAGO
920 E. 58TH STREET
CHICAGO, ILLINOIS 60637
USA
E-MAIL: mstephens@uchicago.edu