THE UNIVERSITY OF CHICAGO

A BAYESIAN LARGE-SCALE MULTIPLE REGRESSION MODEL FOR GENOME-WIDE ASSOCIATION SUMMARY STATISTICS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS

BY
XIANG ZHU

CHICAGO, ILLINOIS
AUGUST 2017

To my family

“Truth is much too complicated to allow anything but approximations.”

– John von Neumann

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I would like to express my most sincere gratitude to my dissertation adviser, Matthew Stephens, for his superb guidance, continuous support and limitless patience during my doctoral training. I have benefited a great deal from all the profound insights and intellectual conversations that he has generously shared over the years. I am very fortunate and truly honored to work with such an world-class statistician and geneticist.

I am also deeply grateful to the other members of my dissertation committee, Rina Foygel Barber and Xin He for their precious time, insightful advice and constructive feedback. During my doctoral studies, I have been always inspired by their excellent work in high-dimensional statistics and integrative genetics.

I would also like to thank other faculty members in the Department of Statistics: Mihai Anitescu, Steven Lalley, Lek-Heng Lim, Peter McCullagh, Mary Sara McPeek, Dan Nicolae, Michael Stein, Wei-Biao Wu and others. Without their generous help and valuable advice, my transition from an undergraduate to a graduate student would have been much harder.

I am also very grateful to all the past and current members of Matthew Stephens and Xin He Laboratories, including Nicholas Knoblauch, John Blischak, Peter Carbonetto, Kushal Dey, David Gerard, Joyce Hsiao, Michael Turchin, Gao Wang, Siming Zhao, Yongtao Guan, Xiaoquan Wen and Xiang Zhou. It is a great pleasure to work in a friendly, collaborative and productive environment created by these talented people.

The last but not the least, I especially wish to acknowledge the everlasting love of my family. They have made tremendous sacrifices to help me pursue my dream over the years. I am forever in their debt, and I dedicate this dissertation to them.

ABSTRACT

Bayesian methods for large-scale multiple regression provide attractive approaches to the analysis of genome-wide association studies (GWAS). For example, they can estimate heritability of complex traits, allowing for both polygenic and sparse models; and by incorporating external genomic information into the priors they can increase statistical power and yield new biological insights. However, these methods require access to individual genotypes and phenotypes, which are often not easily available for large studies.

Here we provide a framework for performing these analyses without individual-level data. Specifically, we introduce a "Regression with Summary Statistics" (RSS) likelihood, which relates the multiple regression coefficients to univariate regression results that are often easily available. The RSS likelihood requires estimates of correlations among covariates (SNPs), which also can be obtained from public databases. We perform Bayesian multiple regression analysis by combining the RSS likelihood with previously-proposed prior distributions, and then using Markov chain Monte Carlo or Variational Bayes algorithms to compute posteriors. In a wide range of simulations RSS performs similarly to analyses using individual-level data, including SNP heritability estimation, genetic association detection and gene set enrichment analysis.

We apply RSS methods to analyze published GWAS summary statistics of 1.1 millions common variants from 31 human phenotypes, 3,913 biological pathways retrieved from nine public databases, and 113 tissue-associated gene sets derived from gene expression profiles of 53 human tissues. We identify many previously-unreported genes, pathways and tissues that show strong evidence for association with complex traits in our large-scale integrated analyses. Software is available at `https://github.com/stephenslab/rss`.

CHAPTER 1

INTRODUCTION

Consider the multiple linear regression model:

$$\mathbf{y} = X\beta + \epsilon, \tag{1.1}$$

where $\mathbf{y}$ is an $n \times 1$ (centered) vector, X is an $n \times p$ (column-centered) matrix, β is the $p \times 1$ vector of multiple regression coefficients, and ϵ is the error term. Assuming the "individual-level" data $\{X, \mathbf{y}\}$ are available, many methods exist to infer β. Here, motivated by applications in genetics, we assume that individual-level data are not available, but instead the summary statistics $\{\hat{\beta}_j, \hat{\sigma}_j^2\}$ from p simple linear regression are provided:

$$\hat{\beta}_j := (X_j^\intercal X_j)^{-1} X_j^\intercal \mathbf{y} \tag{1.2}$$

$$\hat{\sigma}_j^2 := (nX_j^\intercal X_j)^{-1} (\mathbf{y} - X_j\hat{\beta}_j)^\intercal (\mathbf{y} - X_j\hat{\beta}_j) \tag{1.3}$$

where X_j is the jth column of X, $j \in [p] := \{1, \ldots, p\}$. We also assume that information on the correlation structure among $\{X_j\}$ is available. With this in hand, we address the following question: how do we infer *multiple regression coefficients* β using *univariate regression summary statistics* $\{\hat{\beta}_j, \hat{\sigma}_j^2\}$? Specifically, we derive a likelihood for β given $\{\hat{\beta}_j, \hat{\sigma}_j^2\}$ and estimated correlations among $\{X_j\}$, and combine it with suitable priors to perform Bayesian inference for β.

This work is motivated by applications in genome-wide association studies (GWAS), which over the last decade have helped elucidate the genetics of dozens of complex traits and diseases [e.g. Donnelly (2008); McCarthy et al. (2008); Hindorff et al. (2009); Stranger et al. (2011); Visscher et al. (2012); Welter et al. (2014); Price et al. (2015)]. GWAS come in various flavors – and can involve, for example, case-control data and/or related individuals – but here we focus on the simplest case of a quantitative trait (e.g. height) measured on

random samples from a population. Model (1.1) applies naturally to this setting: the covariates X are the (centered) genotypes of n individuals at p genetic variants (typically Single Nucleotide Polymorphisms, or SNPs) in a study cohort; the response $\mathbf{y}$ is the quantitative trait whose relationship with genotype is being studied; and the coefficients β are the effects of each SNP on phenotype, estimation of which is a key inferential goal.

In GWAS individual-level data can be difficult to obtain. Indeed, for many publications no author had access to all the individual-level data. This is because many GWAS analyses involve multiple research groups pooling results across many cohorts to maximize sample size, and sharing individual-level data across groups is made difficult by many factors, including consent and privacy issues, and the substantial technical burden of data transfer, storage, management and harmonization. In contrast, summary data like $\{\hat{\beta}_j, \hat{\sigma}_j^2\}$ are much easier to obtain: collaborating research groups often share such data to perform simple (though useful) "single-SNP" meta-analyses on a very large total sample size [e.g. Zeggini and Ioannidis (2009); Begum et al. (2012); Evangelou and Ioannidis (2013)]. Furthermore these summary data are often made freely available in public domains (Nature Genetics, 2012). In addition, information on the correlations among SNPs [referred to in population genetics as "linkage disequilibrium", or LD; see Pritchard and Przeworski (2001) and Slatkin (2008)] is also available through public databases [e.g. Frazer et al. (2007); International HapMap 3 Consortium (2010); 1000 Genomes Project Consortium (2010, 2015); Sudlow et al. (2015)]. Thus, by providing methods for fitting the model (1.1) using only summary data and LD information, our work greatly facilitates the "multiple-SNP" analysis of GWAS data. For example, as we describe later (Chapter 6), we performed multiple-SNP analyses of GWAS data on 31 human complex traits collected from 20,883-253,288 European ancestry individuals typed at ~ 1.1 million SNPs [Supplementary Table 1 of Zhu and Stephens (2017b)]. Doing this for the individual-level data appears impractical.

Multiple-SNP analyses of GWAS compliment the standard single-SNP analyses in several ways. Multiple-SNP analyses are particularly helpful in fine-mapping causal loci, al-

lowing for multiple causal variants in a region [e.g. Servin and Stephens (2007); Yang et al. (2012); Wen et al. (2016)]. In addition, they can increase power to identify associations [e.g. Hoggart et al. (2008); Guan and Stephens (2011); Moser et al. (2015)]; and can help estimate the overall proportion of phenotypic variation explained by genotyped SNPs (PVE; or "SNP heritability") [e.g. Yang et al. (2010); Zhou et al. (2013); Janson et al. (2016)]. See Sabatti (2013) and Guan and Wang (2013) for more extensive discussion. Despite these benefits, few GWAS are analyzed with multiple-SNP methods, presumably, at least in part, because existing methods require individual-level data that can be difficult to obtain. In addition, most multiple-SNP methods are computationally challenging for large studies [see Bottolo et al. (2013); Peise et al. (2015); Loh et al. (2015) for examples of recent progress in developing computationally tractable approaches]. Our methods help with both these issues, allowing inference to be performed with summary-level data, and reducing computation by exploiting matrix bandedness (Wen and Stephens, 2010).

Because of the importance of this problem for GWAS, many recent publications have described analysis methods based on summary statistics. These include methods for estimation of effect size distribution [e.g. Park et al. (2010); Thompson et al. (2015); Holland et al. (2016)], multiple-SNP association detection [e.g. Yang et al. (2012); Ehret et al. (2012); Newcombe et al. (2016)], single-SNP analysis with correlated phenotypes [e.g. Stephens (2013); Zhu et al. (2015); Pickrell et al. (2016)], gene-level association testing [e.g. Liu et al. (2010); Lee et al. (2015); Gusev et al. (2016)], joint analysis of functional genomic data and GWAS [e.g. He et al. (2013); Pickrell (2014); Finucane et al. (2015)], imputation of allele frequencies [e.g. Wen and Stephens (2010); Chen and Schaid (2014)] and single-SNP association statistics [e.g. Lee et al. (2013); Pasaniuc et al. (2014); Xu et al. (2015)], fine mapping of causal variants [e.g. Hormozdiari et al. (2014); Kichaev et al. (2014); Chen et al. (2015); Farh et al. (2015); Benner et al. (2016)], estimation of SNP heritability [e.g. Bulik-Sullivan et al. (2015b); Palla and Dudbridge (2015); Shi et al. (2016a)], and prediction of polygenic risk scores [e.g. Vilhjalmsson et al. (2015); Shi et al. (2016b); So and Sham (2017); Mak et al.

(2017)]. See Pasaniuc and Price (2017) for more extensive discussion. Together these methods adopt a variety of approaches, many of them tailored to their specific applications. Our approach, being based on a likelihood for the multiple regression coefficients β (Chapters 2-3), provides the foundations for more generally-applicable methods. Having a likelihood opens the door to a wide range of statistical machinery for inference; here we illustrate this by using it to perform Bayesian inference for β, and specifically to estimate SNP heritability (Chapter 4), detect genome-wide multiple-SNP association (Chapter 5) and assess gene set enrichment (Chapter 6).

Our work has close connections with recent Bayesian approaches to this problem, notably Hormozdiari et al. (2014) and Chen et al. (2015). These methods posit a model relating the observed z-scores $\{\hat{\beta}_j/\hat{\sigma}_j\}$ to "non-centrality" parameters, and perform Bayesian inference on the non-centrality parameters. Here, we instead derive a likelihood for the regression coefficients β in (1.1), and perform Bayesian inference for β. These approaches are closely related, but working directly with β seems preferable to us. For example, the non-centrality parameters depend on sample size, which means that appropriate prior distributions may vary among studies depending on their sample size. In contrast, β maintains a consistent interpretation across studies. And working with β allows us to exploit previous work developing prior distributions for β for multiple-SNP analysis [e.g. Guan and Stephens (2011); Zhou et al. (2013); Carbonetto and Stephens (2013)]. We also give a more rigorous statement and derivation of the likelihood being used (Chapter 2), which provides insight into what approximations are being made and when they may be valid (Chapter 3). Finally, this previous work focused only on small genomic regions, whereas here we analyze whole chromosomes (Chapters 4-5) and entire human genome (Chapter 6).

CHAPTER 2

REGRESSION WITH SUMMARY STATISTICS (RSS) LIKELIHOOD

In this chapter, we derive a novel likelihood for the multiple regression coefficients β (1.1), based on the univariate regression summary statistics (1.2, 1.3) and estimated correlation matrix of covariates. Chapters 2-5 are largely based on a manuscript entitled "Bayesian large-scale multiple regression with summary statistics from genome-wide association studies" (Zhu and Stephens, 2017a).

2.1 Introduction

We first introduce some notation. For any vector $\mathbf{v}$, diag($\mathbf{v}$) denotes the diagonal matrix with diagonal elements $\mathbf{v}$. Let $\widehat{\beta} := (\hat{\beta}_1, \ldots, \hat{\beta}_p)^\top$, $\widehat{S} := \text{diag}(\widehat{\mathbf{s}})$, and $\widehat{\mathbf{s}} := (\hat{s}_1, \ldots, \hat{s}_p)^\top$, where

$$\hat{s}_j^2 := \hat{\sigma}_j^2 + n^{-1}\hat{\beta}_j^2 \tag{2.1}$$

and $\{\hat{\beta}_j, \hat{\sigma}_j^2\}$ are the single-SNP summary statistics (1.2, 1.3). We denote probability densities as $p(\cdot)$, and rely on the arguments to distinguish different distributions. Let $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ denote the multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ, and $\mathcal{N}(\boldsymbol{\xi}; \boldsymbol{\mu}, \Sigma)$ denote its density at $\boldsymbol{\xi}$.

In addition to the summary data $\{\hat{\beta}_j, \hat{\sigma}_j^2\}$, we assume that we have an estimate, $\widehat{R}$, of the matrix R of LD (correlations) among SNPs in the population from which the genotypes were sampled. Typically $\widehat{R}$ will come from some public database of genotypes in a suitable reference population; here, we use the shrinkage method from Wen and Stephens (2010) to obtain $\widehat{R}$ from such a reference. The shrinkage method produces more accurate results than the sample correlation matrix (Section 2.7), and has the advantage that it produces a sparse, banded matrix $\widehat{R}$, which speeds computation for large-scale genomic data analyses (Chapters 4-6). For our likelihood to be well-defined, $\widehat{R}$ must be positive definite, and the shrinkage method also ensures this.

With this in place, the likelihood we propose for β is as follows.

Definition 2.1.1.

$$L_{\mathsf{rss}}(\beta;\widehat{\beta},\widehat{S},\widehat{R}) := \mathcal{N}(\widehat{\beta};\widehat{S}\widehat{R}\widehat{S}^{-1}\beta,\widehat{S}\widehat{R}\widehat{S}). \tag{2.2}$$

We refer to (2.2) as the "Regression with Summary Statistics" (RSS) likelihood. We provide a formal derivation in Section 2.6, but informally the derivation assumes that i) the correlation of $\mathbf{y}$ with any single covariate (SNP) X_j is small; and ii) the matrix $\widehat{R}$ accurately reflects the correlation of the covariates (SNPs) in the population from which they were drawn.

The derivation of (2.2) also makes other assumptions that may not hold in practice: that all summary statistics are computed from the same samples, that there is no confounding due to population stratification (or that this has been adequately controlled for), and that genotypes used to computed summary statistics are accurate (so it ignores imputation error in imputed genotypes). Indeed, most analyses of individual-level data also make these last two assumptions.These assumptions can be relaxed, and generalizations of (2.2) derived; see Chapter 3. However these generalizations require additional information – beyond the basic single-SNP summary data (1.2, 1.3) – that is often not easily available. It is therefore tempting to apply (2.2) even when these assumptions may not hold (Chapters 4-6). This is straightforward to do, but results in model misspecification and care is required; see Chapter 3.

2.2 Variations on RSS likelihood

We define $\widehat{S}$ by (2.1). In a GWAS context the sample sizes are often large and $\hat{\beta}_j^2$ are typically small (Table 2.1), and so $\hat{s}_j \approx \hat{\sigma}_j$. Consequently, replacing $\hat{s}_j$ in (2.2) with $\hat{\sigma}_j$ produces a minor variation on the RSS likelihood that, for GWAS applications, differs negligibly from our definition [Supplementary Figure 4 of Zhu and Stephens (2017a)]. This variation has slightly closer connections with existing work (Section 2.5).

Another variation comes from noting that the mean term in (2.2) does not change if we multiply $\widehat{S}$ by any non-zero scalar constant: any constant will cancel out due to the presence of both $\widehat{S}$ and $\widehat{S}^{-1}$. Note further that $\hat{s}_j = \hat{\sigma}_y/(\sqrt{n}\hat{\sigma}_{x,j})$ where $\hat{\sigma}_y^2$ is the sample variance of $\mathbf{y}$ (phenotype), and $\hat{\sigma}_{x,j}^2$ the sample variance of X_j (genotype at SNP j). Since n and $\hat{\sigma}_y$ are constants, the RSS likelihood is unchanged if we replaced $\widehat{S}$ in the mean term with the matrix $\text{diag}^{-1}(\widehat{\boldsymbol{\sigma}}_x)$, where $\widehat{\boldsymbol{\sigma}}_x := (\hat{\sigma}_{x,1}, \ldots, \hat{\sigma}_{x,p})^\intercal$. That is:

$$L^*_{\text{rss}}(\beta) := \mathcal{N}(\widehat{\beta}; \text{diag}^{-1}(\widehat{\boldsymbol{\sigma}}_x)\widehat{R}\,\text{diag}(\widehat{\boldsymbol{\sigma}}_x)\beta, \widehat{S}\widehat{R}\widehat{S}). \tag{2.3}$$

This variation on RSS helps emphasize the role of $\widehat{S}$ in the mean term of (2.2): it is simply a convenience that exploits the fact that $\hat{s}_j \propto 1/\hat{\sigma}_{x,j}$. The form (2.2) is more convenient in practice than (2.3), both because $\widehat{S}$ is easily computed from commonly-used summary data and because the appearance of the same matrix $\widehat{S}$ in the mean and variance terms of (2.2) produces algebraic simplifications that we exploit in our implementation. However, this convenient approach – which is also used in previous work (Section 2.5) – can contribute to model misspecification when, for example, different SNPs are typed on different samples; see Chapter 3.

2.3 Intuition behind RSS likelihood

The RSS likelihood (2.2) is obtained by first deriving an approximation for $p(\widehat{\beta}|S,R,\beta)$, where S is the diagonal matrix with the jth diagonal entry $s_j \approx \text{SD}(\hat{\beta}_j)$, of which $\widehat{S}$ is an estimate (see Section 2.6 for details). Specifically, we have

$$\widehat{\beta}|S,R,\beta \dot{\sim} \mathcal{N}(SRS^{-1}\beta, SRS), \tag{2.4}$$

from which the RSS likelihood (2.2) is derived by plugging in the estimates $\{\widehat{S},\widehat{R}\}$ for $\{S,R\}$.

The distribution (2.4) captures three key features of the single-SNP association test

statistics in GWAS. First, the mean of the single-SNP effect size estimate $\hat{\beta}_j$ depends on both its own effect and the effects of all SNPs that it "tags" (i.e. is highly correlated with):

$$\mathrm{E}(\hat{\beta}_j|S,R,\boldsymbol{\beta}) = s_j \cdot \sum_{i=1}^{p} r_{ij} s_i^{-1} \beta_i, \tag{2.5}$$

where r_{ij} is the (i,j)-th entry of R. Second, the likelihood incorporates the fact that the estimated single-SNP effects are heteroscedastic:

$$\mathrm{Var}(\hat{\beta}_j|S,R,\boldsymbol{\beta}) = s_j^2 \approx \hat{s}_j^2 = (nX_j^\intercal X_j)^{-1}\mathbf{y}^\intercal\mathbf{y}. \tag{2.6}$$

Since s_j^2 is roughly proportional to $(X_j^\intercal X_j)^{-1}$, the likelihood takes account of differences in the informativeness of SNPs due to their variation in allele frequency and imputation quality (Guan and Stephens, 2008). Third, single-SNP test statistics at SNPs in LD are correlated:

$$\mathrm{Corr}(\hat{\beta}_j, \hat{\beta}_k|S,R,\boldsymbol{\beta}) = r_{jk}, \tag{2.7}$$

for any pair of SNP j and k.

Note that SNPs in LD with one another have "correlated" test statistics $\{\hat{\beta}_j\}$ for two distinct reasons. First, they share "signal", which is captured in the mean term (2.5). This shared signal becomes a correlation if the true effects $\boldsymbol{\beta}$ are assumed to arise from some distribution and are then integrated out. Second, they share "noise", which is captured in the correlation term (2.7). This latter correlation occurs even in the absence of signal ($\boldsymbol{\beta} = \mathbf{0}$) and is due to the fact that the summary data are computed on the same samples. If the summary data were computed on independent sets of individuals, then this latter correlation would disappear (Chapter 3).

2.4 Connection with the full-data likelihood

When individual-level data are available the multiple regression model is

$$\mathbf{y}|X,\beta,\tau \sim \mathcal{N}(X\beta,\tau^{-1}I). \tag{2.8}$$

If we further assume the residual variance τ^{-1} is *known*, model (2.8) specifies a likelihood for β, which we denote $L_{\mathsf{mvn}}(\beta;\mathbf{y},X,\tau)$. The following proposition gives conditions under which this full-data likelihood and RSS likelihood are equivalent.

Proposition 2.4.1. Let $\widehat{R}^{\mathsf{sam}}$ denote the sample LD matrix computed from the genotypes X of the study cohort, $\widehat{R}^{\mathsf{sam}} := D^{-1}X^{\mathsf{T}}XD^{-1}$ where $D := \mathrm{diag}(\mathbf{d})$, $\mathbf{d} := (||X_1||,\ldots,||X_p||)^{\mathsf{T}}$, $||X_j|| := (X_j^{\mathsf{T}}X_j)^{1/2}$. If $n > p$, $\tau^{-1} = n^{-1}\mathbf{y}^{\mathsf{T}}\mathbf{y}$ and $\widehat{R} = \widehat{R}^{\mathsf{sam}}$ then

$$\log L_{\mathsf{rss}}(\beta;\widehat{\beta},\widehat{S},\widehat{R}) - \log L_{\mathsf{mvn}}(\beta;\mathbf{y},X,\tau) = C \tag{2.9}$$

where C is some constant that does not depend on β.

Proof. Notice that $\widehat{\beta} = D^{-2}X^{\mathsf{T}}\mathbf{y}$ and $\widehat{S} = \sqrt{n^{-1}\mathbf{y}^{\mathsf{T}}\mathbf{y}}\cdot D^{-1}$. If $\tau^{-1} = n^{-1}\mathbf{y}^{\mathsf{T}}\mathbf{y}$ and $\widehat{R} = \widehat{R}^{\mathsf{sam}}$, then $\widehat{S}^{-2}\widehat{\beta} = \tau X^{\mathsf{T}}\mathbf{y}$ and $\widehat{S}^{-1}\widehat{R}\widehat{S}^{-1} = \tau X^{\mathsf{T}}X$. When $n > p$, the matrix X is full column rank and thus $\widehat{R} = \widehat{R}^{\mathsf{sam}}$ is non-singular, the full data and summary data likelihood are given by

$$\begin{aligned}
-2\log L_{\mathsf{mvn}}(\beta;\mathbf{y},X,\tau) &= p\log(2\pi\tau^{-1}) + \tau\mathbf{y}^{\mathsf{T}}\mathbf{y} - 2\tau\mathbf{y}^{\mathsf{T}}X\beta + \tau\beta^{\mathsf{T}}X^{\mathsf{T}}X\beta, \\
-2\log L_{\mathsf{rss}}(\beta;\widehat{\beta},\widehat{S},\widehat{R}) &= p\log(2\pi) + \log|\widehat{S}\widehat{R}\widehat{S}| + \widehat{\beta}^{\mathsf{T}}(\widehat{S}\widehat{R}\widehat{S})^{-1}\widehat{\beta} - 2\widehat{\beta}^{\mathsf{T}}\widehat{S}^{-2}\beta + \beta^{\mathsf{T}}\widehat{S}^{-1}\widehat{R}\widehat{S}^{-1}\beta,
\end{aligned}$$

respectively, and their difference does not depend on the parameter of interest β,

$$-2[\log L_{\mathsf{rss}}(\beta;\widehat{\beta},S,R) - \log L_{\mathsf{mvn}}(\beta;\mathbf{y},X)] = \log|D^{-1}\widehat{R}D^{-1}| - \tau\mathbf{y}^{\mathsf{T}}[I - X(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}]\mathbf{y}, \tag{2.10}$$

implying that these two likelihoods of β are equivalent. □

The assumption $n > p$ in Proposition 2.4.1 could possibly be relaxed, but certainly simplifies the proof. The key assumption then is $\tau^{-1} = n^{-1}\mathbf{y}^\intercal\mathbf{y}$; that is, that the total variance in $\mathbf{y}$ explained by X is negligible. This will typically not hold in a genome-wide context, but might hold, approximately, when fine mapping a small genomic region since SNPs in a small region typically explain a very small proportion of phenotypic variation[1]. Hence, provided that $\widehat{R} = \widehat{R}^{\mathsf{sam}}$, RSS and its full-data counterpart will produce approximately the same inferential results in small regions. This is illustrated through simulations in Section 2.7 (Figure 2.1); see also Chen et al. (2015).

2.5 Connection with previous work

The RSS likelihood is connected to several previous approaches to inference from summary data, as we now describe. [These connections are precise for the variation on the RSS likelihood with $\hat{s}_j = \hat{\sigma}_j$ (Section 2.2), which differs negligibly in practice from (2.2).]

In the simplest case, if $\widehat{R}$ is an identity matrix, then

$$\widehat{\beta}|\beta,\widehat{S} \sim \mathcal{N}(\beta,\widehat{S}^2), \tag{2.11}$$

which is the implied likelihood based on the standard confidence interval (Efron, 1993). Wakefield (2009) has recently popularized this likelihood for calculation of approximate Bayes factors; see also Stephens (2017).

If we let $\mathbf{z}$ denote the vector of single-SNP z-scores, $\mathbf{z} := \widehat{S}^{-1}\widehat{\beta}$, and plug $\{\widehat{S},\widehat{R}\}$ into (2.4), then

$$\mathbf{z}|\widehat{S},\widehat{R},\beta \sim \mathcal{N}(\widehat{R}\widehat{S}^{-1}\beta,\widehat{R}). \tag{2.12}$$

This is analogous to the likelihood proposed in Hormozdiari et al. (2014), $\mathbf{z} \sim \mathcal{N}(\widehat{R}\nu,\widehat{R})$, where they refer to ν as the "non-centrality parameter". If further $\beta = \mathbf{0}$, then $\mathbf{z} \sim \mathcal{N}(\mathbf{0},\widehat{R})$, a

1. There are exceptions; for example the human leukocyte antigen region is estimated to explain 11-37% of the heritability of rheumatoid arthritis (Kurkó et al., 2013).

GWAS phenotype	# of SNPs	$\log_{10}(\hat{c}^2)$						$\log_{10}(n)$		
	(million)	Min	Q1	Median	Q3	Max	Histogram	Median	Mean	SD
Height (GIANT, 2010)	2.82	−12.64	−6.25	−5.60	−5.12	−2.90		5.26	5.26	NA
Height (GIANT, 2014)	2.53	−10.74	−6.06	−5.41	−4.93	−2.54		5.40	5.37	0.09
BMI (GIANT, 2015)	2.54	−10.89	−6.30	−5.65	−5.18	−2.66		5.37	5.34	0.09
WHRadjBMI (GIANT, 2015)	2.53	−10.81	−6.11	−5.46	−4.99	−2.81		5.15	5.13	0.08
HDL (GLGC, 2010)	2.68	−10.78	−5.90	−5.25	−4.77	−1.23		5.00	4.89	0.33
HDL (GLGC, 2013)	2.43	−10.16	−5.89	−5.25	−4.78	−1.59		4.97	4.97	0.06
LDL (GLGC, 2010)	2.68	−10.72	−5.89	−5.23	−4.75	−1.44		4.98	4.87	0.33
LDL (GLGC, 2013)	2.42	−9.95	−5.89	−5.24	−4.78	−1.40		4.95	4.95	0.06
TC (GLGC, 2010)	2.68	−10.33	−5.91	−5.25	−4.77	−1.38		5.00	4.89	0.33
TC (GLGC, 2013)	2.43	−10.28	−5.91	−5.26	−4.79	−1.79		4.98	4.97	0.06
TG (GLGC, 2010)	2.68	−10.55	−5.89	−5.24	−4.76	−1.17		4.98	4.87	0.33
TG (GLGC, 2013)	2.42	−10.07	−5.88	−5.24	−4.78	−1.90		4.96	4.96	0.06
Cigarettes per day (TAG, 2010)	2.46	−14.16	−5.84	−5.19	−4.73	−2.69		4.87	4.87	NA
Smoking age of onset (TAG, 2010)	2.43	−11.27	−5.82	−5.18	−4.73	−3.44		4.87	4.87	NA
Ever smoked (TAG, 2010)	2.45	−11.89	−5.82	−5.17	−4.71	−3.44		4.87	4.87	NA
Former smoker (TAG, 2010)	2.45	−12.68	−5.83	−5.19	−4.73	−3.40		4.87	4.87	NA
Years of education (SSGAC, 2013)	2.14	−7.10	−5.70	−5.30	−4.85	−3.51		5.10	5.10	NA
College or not (SSGAC, 2013)	2.25	−8.37	−5.93	−5.33	−4.88	−3.39		5.10	5.10	NA
Depressive (SSGAC, 2016)	6.03	−7.44	−6.00	−5.46	−5.01	−3.60		5.21	5.21	NA
Neuroticism (SSGAC, 2016)	6.04	−7.88	−5.92	−5.45	−4.98	−3.29		5.23	5.23	NA
Schizophrenia (PGC, 2014)	9.43	−12.50	−6.01	−5.35	−4.88	−3.04		5.18	5.18	NA
Alzheimer (IGAP, 2013)	7.04	−11.20	−5.69	−5.04	−4.57	−1.33		4.73	4.73	NA
CAD (CARDIoGRAM, 2011)	2.42	−17.43	−5.84	−5.18	−4.71	−2.74		4.91	4.88	0.08
T2D (DIAGRAM, 2012)	2.09	−7.83	−6.00	−5.49	−5.13	−2.93		4.80	4.78	0.10
Hb (HaemGen, 2012)	2.58	−9.79	−5.64	−4.99	−4.52	−2.47		4.74	4.68	0.15
MCHC (HaemGen, 2012)	2.57	−9.72	−5.62	−4.98	−4.51	−2.50		4.70	4.65	0.15
MCH (HaemGen, 2012)	2.58	−10.24	−5.56	−4.91	−4.44	−2.02		4.67	4.62	0.14
MCV (HaemGen, 2012)	2.59	−11.02	−5.61	−4.96	−4.48	−2.09		4.71	4.66	0.15
PCV (HaemGen, 2012)	2.59	−10.67	−5.59	−4.94	−4.47	−2.70		4.69	4.63	0.14
RBC (HaemGen, 2012)	2.56	−8.92	−5.55	−4.91	−4.45	−2.11		4.69	4.63	0.15
FGadjBMI (MAGIC, 2012)	2.61	−11.63	−5.70	−5.07	−4.61	−2.10		4.76	4.76	NA
FIadjBMI (MAGIC, 2012)	2.60	−11.54	−5.65	−5.02	−4.56	−2.96		4.71	4.71	NA
Heart rate (HRgene, 2013)	2.52	−12.08	−5.88	−5.23	−4.76	−2.88		4.95	4.93	0.07
Serum urate (GUGC, 2013)	2.44	−10.36	−5.95	−5.30	−4.83	−1.49		5.04	5.03	0.02
Gout (GUGC, 2013)	2.54	−12.17	−5.80	−5.15	−4.69	−2.68		4.84	4.84	0.01
RA (Okada et al, 2014)	7.70	−8.16	−5.44	−4.97	−4.55	−1.09		4.77	4.77	NA
IBD (IIBDGC, 2015)	12.70	−13.05	−5.49	−4.84	−4.38	−2.07		4.54	4.54	NA
CD (IIBDGC, 2015)	12.27	−12.97	−5.28	−4.62	−4.16	−1.89		4.32	4.32	NA
UC (IIBDGC, 2015)	12.24	−12.69	−5.40	−4.75	−4.28	−2.07		4.44	4.44	NA
CAD (CARDIoGRAM+C4D, 2015)	9.46	−15.00	−6.26	−5.61	−5.14	−2.62		5.27	5.27	NA
MI (CARDIoGRAM+C4D, 2015)	9.29	−18.89	−6.22	−5.56	−5.09	−2.69		5.22	5.22	NA
ANM (ReproGen, 2015)	2.09	−7.40	−5.44	−4.84	−4.56	−2.16		4.84	4.84	NA

Table 2.1: Summary of per-SNP sample squared correlation $\{\hat{c}_j^2\}$ and sample size $\{n_j\}$ for 42 large GWAS performed in European-ancestry individuals. The full names of phenotypes and corresponding references are provided in Supplementary Table 1 of Zhu and Stephens (2017a). The five-number summaries and histograms are across SNPs. The sample correlation $\hat{c}_j$ between phenotype and SNP j is defined as $\hat{c}_j := (||\mathbf{y}|| \cdot ||X_j||)^{-1}(X_j^\top \mathbf{y})$. Note that $\hat{c}_j^2 = (n_j\hat{\sigma}_j^2 + \hat{\beta}_j^2)^{-1}\hat{\beta}_j^2 = (n_j\hat{s}_j^2)^{-1}\hat{\beta}_j^2$, and $\hat{c}_j \xrightarrow{p} c_j$. The SD of sample sizes per SNP $\{n_j\}$ is NA when $\{n_j\}$ are not publicly available.

result that has been used for multiple testing adjustment [e.g. Seaman and Müller-Myhsok (2005); Lin (2005)], gene-based association detection [e.g. Liu et al. (2010); Lee et al. (2015)] and single-SNP association z-score imputation [e.g. Lee et al. (2013); Pasaniuc et al. (2014)].

If $\boldsymbol{\beta}$ is given a prior distribution that assumes zero mean and independence across all j, that is, $p(\boldsymbol{\beta}|\widehat{S},\widehat{R}) = \prod_j p(\beta_j|\widehat{S},\widehat{R})$, $\mathrm{E}(\beta_j|\widehat{S},\widehat{R}) = 0$, then integrating $\boldsymbol{\beta}$ out in (2.12) yields

$$\mathrm{E}(z_j^2|\widehat{S},\widehat{R}) = 1 + \sum_{i=1}^{p} \hat{r}_{ij}^2 \hat{s}_i^{-2} \mathrm{E}(\beta_i^2|\widehat{S},\widehat{R}). \tag{2.13}$$

This is a key element of LD score regression (Bulik-Sullivan et al., 2015b); see Chapter 3 for further details and discussion.

2.6 Derivations and proofs

We treat the (unobserved) genotypes of each individual, $\boldsymbol{x}_i$ (the ith row of X), as being independent and identically distributed (i.i.d.) draws from some population $\boldsymbol{x}$. Without loss of generality, assume these have been centered, by subtracting the mean, so that $\mathrm{E}(\boldsymbol{x}_i) = \mathbf{0}$. Let $\sigma_{x,j} > 0$ denote the population standard deviation (SD) of x_{ij}, and R denote the $p \times p$ positive definite population correlation matrix, so $\mathrm{Var}(\boldsymbol{x}_i) := \Sigma_x := \mathrm{diag}(\boldsymbol{\sigma}_x) \cdot R \cdot \mathrm{diag}(\boldsymbol{\sigma}_x)$, where $\boldsymbol{\sigma}_x := (\sigma_{x,1},\ldots,\sigma_{x,p})^\intercal$.

We assume that the phenotypes $\mathbf{y} := (y_1,\ldots,y_n)^\intercal$ are generated from the multiple-SNP model (1.1), where $\mathrm{E}(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{\epsilon}) = \tau^{-1} I_n$, $\tau \in (0,\infty)$. We also assume that X, $\boldsymbol{\epsilon}$ and $\boldsymbol{\beta}$ are mutually independent.

Let $\mathbf{c} := (c_1,\ldots,c_p)^\intercal$ denote the vector of (population) marginal correlations between the phenotype and genotype of each SNP:

$$\mathbf{c} := \sigma_y^{-1} \mathrm{diag}^{-1}(\boldsymbol{\sigma}_x) \boldsymbol{\mu}_{xy} \tag{2.14}$$

where $\boldsymbol{\mu}_{xy} := \mathrm{E}(\boldsymbol{x}_i y_i)$ and $\sigma_y^2 := \mathrm{Var}(y_i)$.

We first derive the asymptotic distribution of $\widehat{\boldsymbol{\beta}}$ (with $n \to \infty$ and p fixed), using the Multivariate Central Limit Theorem and Delta Method [e.g. van der Vaart (1998)].

Proposition 2.6.1. Let $\Sigma := \sigma_y^2 \text{diag}^{-1}(\boldsymbol{\sigma}_x)(R + \Delta(\mathbf{c}))\text{diag}^{-1}(\boldsymbol{\sigma}_x)$, where $\Delta(\mathbf{c}) \in \mathbb{R}^{p \times p}$ is a continuous function of $\mathbf{c}$ and $\Delta(\mathbf{c}) = \mathcal{O}(\max_j c_j^2)$.

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \text{diag}^{-1}(\boldsymbol{\sigma}_x) R \text{diag}(\boldsymbol{\sigma}_x)\boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma). \tag{2.15}$$

Proof. First define the statistic $T_n \in \mathbb{R}^{2p \times 1}$,

$$T_n := n^{-1}\left(\textstyle\sum_{i=1}^n x_{i1}y_i, \ldots, \sum_{i=1}^n x_{ip}y_i, \sum_{i=1}^n x_{i1}^2, \ldots, \sum_{i=1}^n x_{ip}^2\right)^\top. \tag{2.16}$$

The asymptotic distribution of T_n is given by the Multivariate Central Limit Theorem

$$\sqrt{n}(T_n - \boldsymbol{\mu}_T) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma_T), \tag{2.17}$$

where $\boldsymbol{\mu}_T := \text{E}(\mathbf{t})$, $\Sigma_T := \text{Var}(\mathbf{t})$ and $\mathbf{t} := (x_1 y, \ldots, x_p y, x_1^2, \ldots, x_p^2)^\top$. Note that Σ_T has finite entries because τ^{-1} is finite and $\boldsymbol{x}$ are genotypes.

Next, for any $\boldsymbol{\xi} \in \mathbb{R}^{2p \times 1}$, define the following function $g(\boldsymbol{\xi}) \in \mathbb{R}^{p \times 1}$:

$$g(\boldsymbol{\xi}) := (\xi_1/\xi_{p+1}, \ldots, \xi_p/\xi_{2p})^\top. \tag{2.18}$$

Note that $g(T_n) = \widehat{\boldsymbol{\beta}}$ and $g(\boldsymbol{\mu}_T) = \text{diag}^{-2}(\boldsymbol{\sigma}_x)\boldsymbol{\mu}_{xy} = \text{diag}^{-1}(\boldsymbol{\sigma}_x) R \text{diag}(\boldsymbol{\sigma}_x)\boldsymbol{\beta}$.

Use the Multivariate Delta Method to show that

$$\sqrt{n}(g(T_n) - g(\boldsymbol{\mu}_T)) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \nabla^\top g(\boldsymbol{\mu}_T)\Sigma_T \nabla g(\boldsymbol{\mu}_T)) \tag{2.19}$$

where $\nabla g(\boldsymbol{\mu}_T) \in \mathbb{R}^{2p \times p}$ is the gradient matrix of g at $\boldsymbol{\mu}_T$. A straightforward calculation yields that

$$\nabla^\top g(\boldsymbol{\mu}_T)\Sigma_T \nabla g(\boldsymbol{\mu}_T) = \sigma_y^2 \text{diag}^{-1}(\boldsymbol{\sigma}_x)(R + \Delta(\mathbf{c}))\text{diag}^{-1}(\boldsymbol{\sigma}_x). \tag{2.20}$$

The explicit form of $\Delta(\mathbf{c})$ is given by

$$\Delta(\mathbf{c}) := \text{diag}^{-1}(\boldsymbol{\sigma}_x) \cdot \Big[G_1(\mathbf{c}) + G_2(\mathbf{c}) + G_2^\top(\mathbf{c}) + G_3(\mathbf{c}) \Big] \cdot \text{diag}^{-1}(\boldsymbol{\sigma}_x), \tag{2.21}$$

where functions $G_i(\mathbf{c}) : \mathbb{R}^{p\times 1} \mapsto \mathbb{R}^{p\times p}$ are defined as follows:

$$\begin{aligned}
G_1(\mathbf{c}) &:= -(\mathbf{c}^\top R^{-1}\mathbf{c})\Sigma_x - \text{diag}(\boldsymbol{\sigma}_x)\mathbf{c}\mathbf{c}^\top \text{diag}(\boldsymbol{\sigma}_x) + \text{E}[(\boldsymbol{x}^\top \text{diag}^{-1}(\boldsymbol{\sigma}_x) R^{-1}\mathbf{c})^2 \boldsymbol{x}\boldsymbol{x}^\top], \\
G_2(\mathbf{c}) &:= \text{diag}^{-1}(\boldsymbol{\sigma}_x)\text{diag}(\mathbf{c}) W(\mathbf{c}),\ \ [W(\mathbf{c})]_{ij} := \sigma_{x,i}\sigma_{x,j}^2 c_i - \mathbf{c}^\top R^{-1}\text{diag}^{-1}(\boldsymbol{\sigma}_x)\text{E}(x_i x_j^2 \boldsymbol{x}), \\
G_3(\mathbf{c}) &:= \text{diag}^{-1}(\boldsymbol{\sigma}_x)\ \text{diag}(\mathbf{c})\ \Sigma_{xx}\ \text{diag}\ (\mathbf{c})\ \text{diag}^{-1}(\boldsymbol{\sigma}_x),\ \ [\Sigma_{xx}]_{ij} := \text{Cov}(x_i^2, x_j^2).
\end{aligned}$$

Note that $G_i(\mathbf{c})$ are continuous functions of $\mathbf{c}$, $G_i(\mathbf{0}) = \mathbf{0}$, and $G_i(\mathbf{c}) = \mathcal{O}(\max_j c_j^2)$ for $i = 1,2,3$. □

Proposition 2.6.1 suggests $\mathcal{N}(\text{diag}^{-1}(\boldsymbol{\sigma}_x) R \text{diag}(\boldsymbol{\sigma}_x)\boldsymbol{\beta}, n^{-1}\Sigma)$ approximates the sampling distribution of $\widehat{\boldsymbol{\beta}}$ for large n. Without additional assumptions, this may be the best[2] probability statement that can be used to infer $\boldsymbol{\beta}$. It is difficult to work with this asymptotic distribution, mainly because of the complicated form of $\Delta(\mathbf{c})$ (2.21). However, we can justify ignoring this term in a typical GWAS by the fact that $\{c_j^2\}$ are typically small in GWAS (Table 2.1), and the following proposition:

Proposition 2.6.2. Let $S := n^{-\frac{1}{2}}\sigma_y \text{diag}^{-1}(\boldsymbol{\sigma}_x)$. For each $\boldsymbol{\beta} \in \mathbb{R}^p$,

$$\log \mathcal{N}(\widehat{\boldsymbol{\beta}}; SRS^{-1}\boldsymbol{\beta}, SRS) - \log \mathcal{N}(\widehat{\boldsymbol{\beta}}; \text{diag}^{-1}(\boldsymbol{\sigma}_x) R \text{diag}(\boldsymbol{\sigma}_x)\boldsymbol{\beta}, n^{-1}\Sigma) = \mathcal{O}_p(\max_j c_j^2).$$

2. A more rigorous approximation of likelihood based on the convergence in distribution requires additional technical assumptions; see Boos (1985) and Sweeting (1986).

Proof. First note that $SRS^{-1} = \text{diag}^{-1}(\boldsymbol{\sigma}_x) R \text{diag}(\boldsymbol{\sigma}_x)$. Hence,

$$\begin{aligned} & \log \mathcal{N}(\widehat{\boldsymbol{\beta}}; SRS^{-1}\boldsymbol{\beta}, SRS) - \log \mathcal{N}(\widehat{\boldsymbol{\beta}}; \text{diag}^{-1}(\boldsymbol{\sigma}_x) R \text{diag}(\boldsymbol{\sigma}_x)\boldsymbol{\beta}, n^{-1}\Sigma) \\ = \; & \frac{1}{2}\left\{ \log|R + \Delta(\mathbf{c})| - \log|R| + \sigma_y^{-2} \boldsymbol{\lambda}^\intercal \text{diag}(\boldsymbol{\sigma}_x)[(R + \Delta(\mathbf{c}))^{-1} - R^{-1}]\text{diag}(\boldsymbol{\sigma}_x)\boldsymbol{\lambda} \right\}, \quad (2.22) \end{aligned}$$

where $\boldsymbol{\lambda} := \sqrt{n}(\widehat{\boldsymbol{\beta}} - SRS^{-1}\boldsymbol{\beta})$. Since the determinant and inverse of a matrix are both continuous, we invoke $\boldsymbol{\lambda} = \mathcal{O}_p(1)$ and $\Delta(\mathbf{c}) = \mathcal{O}(\max_j c_j^2)$ (shown in Proposition 2.6.1), to complete the proof. □

These propositions justify the approximate asymptotic distribution of $\widehat{\boldsymbol{\beta}}$ given in (2.4), provided n is large and $\{c_j^2\}$ close to zero, yielding

$$L_{\text{rss}}(\boldsymbol{\beta}; \widehat{\boldsymbol{\beta}}, S, R) := \mathcal{N}(\widehat{\boldsymbol{\beta}}; SRS^{-1}\boldsymbol{\beta}, SRS). \quad (2.23)$$

Finally, the RSS likelihood (2.2) is obtained by replacing the nuisance parameters $\{S, R\}$ with their estimates $\{\widehat{S}, \widehat{R}\}$. There remains obvious potential for errors in the estimates $\{\widehat{S}, \widehat{R}\}$ to impact inference, and we assess this impact empirically through simulations and data analyses (Chapters 4-6).

2.7 Choice of correlation matrix

The estimated correlation matrix among covariates $\widehat{R}$ (i.e. LD matrix in the context of GWAS) plays a key role in the RSS likelihood, as well as in previous work using summary data [e.g. Yang et al. (2012); Hormozdiari et al. (2014); Bulik-Sullivan et al. (2015b)]. One simple choice for $\widehat{R}$, commonly used in previous work, is the sample LD matrix computed from a suitable "reference panel" that is deemed similar to the study population. This is a viable choice if the number of SNPs p is smaller than the number of individuals m in the reference panel, as the sample LD matrix is then invertible. However, for large-scale genomic applications $p \gg m$, and the sample LD matrix is not invertible. Our proposed

solution is to use the shrinkage estimator from Wen and Stephens (2010), which shrinks the off-diagonal entries of the sample LD matrix towards zero, resulting in an invertible matrix.

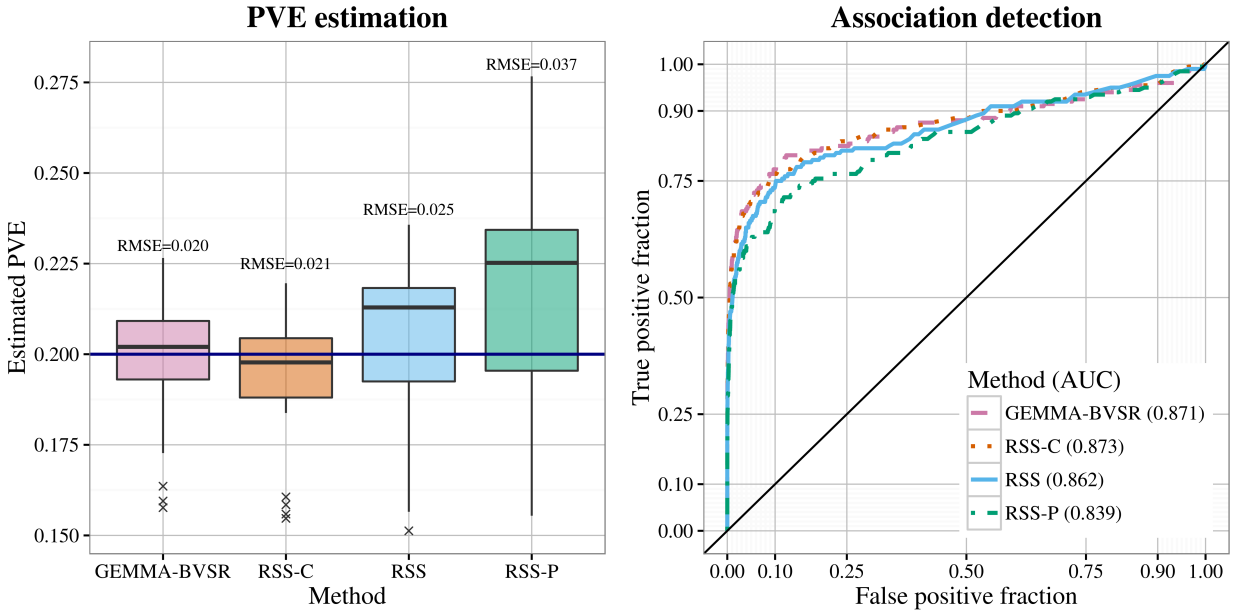


Figure 2.1: Comparison of PVE estimation and association detection on three types of LD matrix $\widehat{R}$: cohort sample LD (RSS-C), shrinkage panel sample LD (RSS) and panel sample LD (RSS-P). Performance of estimating PVE is measured by the root of mean square error (RMSE), where a lower value indicates better performance. Performance of detecting associations is measured by the area under the curve (AUC), where a higher value indicates better performance. See Chapter 4 and Chapter 5 for the details of PVE estimation and association detection respectively.

The shrinkage-based estimate of R can result in improved inference even if $p < m$. To illustrate this, we performed a small simulation study, with 982 SNPs within the ±5 Mb region surrounding the gene *IL27* (Wellcome Trust Case Control Consortium , 2007). We simulated 20 independent datasets, each with 10 causal SNPs and true PVE 0.2. [We also performed simulations with the true PVE being 0.02 and 0.002; see Supplementary Figure 2 of Zhu and Stephens (2017a).] For each dataset, we ran RSS-BVSR with two strategies for computing $\widehat{R}$ from a reference panel (here, the 1,480 control individuals in the WTCCC 1958 British Birth Cohort): the sample LD matrix (RSS-P), and the shrinkage-based estimate (RSS). We compared results with analyses using the full data [GEMMA-BVSR; Zhou

et al. (2013); Guan and Stephens (2011)], and also with our RSS approach using the *cohort* LD matrix (RSS-C), which by Proposition 2.4.1 should produce results similar to the full data analysis. The results (Figure 2.1) show that using the shrinkage-based estimate for R produces consistently more accurate inferences – both for estimating PVE and detecting associations – than using the reference sample LD matrix, and indeed provides similar accuracy to the full data analysis.

CHAPTER 3

EXTENSIONS OF RSS LIKELIHOOD AND PRACTICAL ISSUES

The RSS likelihood (2.2) is based on the multiple linear regression model (1.1), with the following assumptions being made (by default):

- Assumption 1: all covariates X and responses $\mathbf{y}$ come from the same sample;
- Assumption 2: all covariates X are observed without measurement error;
- Assumption 3: covariates X and errors ϵ are independent (or at least uncorrelated).

These canonical assumptions in regression, however, are often violated in large-scale GWAS. Assumption 1 does not hold when GWAS consist of multiple cohorts, each of which collects genotypes on a different set of SNPs (Sections 3.1.1 and 3.2.1). Assumption 2 does not hold when genotypes of some SNPs are not directly assayed but probabilistically imputed (Sections 3.1.2 and 3.2.2). Assumption 3 does not hold when confounding effects are not fully adjusted for prior to regression analysis (Sections 3.1.3 and 3.2.3). In this section, we theoretically extend the RSS framework to capture these important features of GWAS summary data (Section 3.1), and provide practical suggestions when the theoretical extensions cannot be easily implemented (Section 3.2)

3.1 Theoretical extensions

3.1.1 Data on different individuals

The RSS likelihood (2.2) assumes that the univariate summary data are computed from the same set of individuals, but this assumption is often violated in practice (see Section 3.2.1 below). Here we derive a variant of RSS likelihood for summary data generated from different individuals.

Suppose that for each SNP j, its single-SNP summary statistics $\{\hat{\beta}_j, \hat{\sigma}_j^2\}$ are computed on a *predefined, nonempty* subset of individuals $\mathscr{I}_j \subseteq [n]$:

$$\hat{\beta}_j(\mathscr{I}_j; X_j, \mathbf{y}) := \left(\sum_{i\in\mathscr{I}_j} x_{ij}^2\right)^{-1} \left(\sum_{i\in\mathscr{I}_j} x_{ij} y_i\right), \tag{3.1}$$

$$\hat{\sigma}_j^2(\mathscr{I}_j; X_j, \mathbf{y}) := \left(|\mathscr{I}_j| \cdot \sum_{i\in\mathscr{I}_j} x_{ij}^2\right)^{-1} \left[\sum_{i\in\mathscr{I}_j} \left(y_i - x_{ij}\hat{\beta}_j\right)^2\right], \tag{3.2}$$

where $|\cdot|$ denotes the cardinality of a set. Let $\mathscr{I} := \{\mathscr{I}_1, \ldots, \mathscr{I}_p\}$ and $\widehat{\boldsymbol{\beta}}(\mathscr{I}; X, \mathbf{y}) \in \mathbb{R}^p$, whose jth element is $\hat{\beta}_j(\mathscr{I}_j; X_j, \mathbf{y})$. Let $N := \mathrm{diag}(\mathbf{n})$, $\mathbf{n} := (|\mathscr{I}_1|, \ldots, |\mathscr{I}_p|)^\intercal$.

Let $\widehat{F}(\mathscr{I}; X, \mathbf{y}) := \mathrm{diag}\left(\widehat{\mathbf{f}}(\mathscr{I}; X, \mathbf{y})\right)$, $\widehat{\mathbf{f}} \in \mathbb{R}^p$, whose jth element is

$$\hat{f}_j(\mathscr{I}_j; X_j, \mathbf{y}) := \sqrt{\left(\sum_{i\in\mathscr{I}_j} x_{ij}^2\right)^{-1} \left(\sum_{i\in\mathscr{I}_j} y_i^2\right)}. \tag{3.3}$$

Note that $\hat{f}_j^2$ is the estimated ratio of phenotypic over genotypic variance at SNP j (i.e. $\sigma_y^2/\sigma_{x,j}^2$), and it can be computed from the single-SNP summary statistics (3.1, 3.2),

$$\hat{f}_j^2(\mathscr{I}_j; X_j, \mathbf{y}) = |\mathscr{I}_j| \cdot \hat{\sigma}_j^2(\mathscr{I}_j; X_j, \mathbf{y}) + \hat{\beta}_j^2(\mathscr{I}_j; X_j, \mathbf{y}). \tag{3.4}$$

We omit the index $\mathscr{I}$ labeling subsets in the following discussion.

We introduce a matrix H to reflect the proportions of sample overlap among different SNPs. Specifically, the (i,j)-entry of H is defined as $H_{ij} := (|\mathscr{I}_i| \cdot |\mathscr{I}_j|)^{-\frac{1}{2}} |\mathscr{I}_i \cap \mathscr{I}_j|$. Note that the diagonals of H are all 1; the other entries are between 0 and 1. For any pair of SNPs (i,j), $H_{ij} = 1$ if and only if $\mathscr{I}_i = \mathscr{I}_j$ (the same set of individuals); $H_{ij} = 0$ if and only if $\mathscr{I}_i \cap \mathscr{I}_j = \emptyset$ (two disjoint sets of individuals).

With this in place, the modified RSS likelihood of $\boldsymbol{\beta}$ accounting for sample difference is given as follows.

Definition 3.1.1.

$$L_{\text{rss}}^{\text{subset}}(\beta) := \mathcal{N}(\widehat{\beta};\, \widehat{F}\widehat{R}\widehat{F}^{-1}\beta,\, N^{-\frac{1}{2}}\widehat{F}(H\circ\widehat{R})\widehat{F}N^{-\frac{1}{2}}), \tag{3.5}$$

where $H\circ\widehat{R}$ is the Hadamard product of H and $\widehat{R}$.

Note that the modified likelihood (3.5) includes the original RSS likelihood (2.2) as a special case. To see this, when $\mathcal{I}_j = [n]$ for each SNP j, we have $\widehat{F} = \sqrt{n}\widehat{S}$, $N = nI_p$ and H is an all-one matrix, further yielding that

general form simple form

$$\widehat{F}\widehat{R}\widehat{F}^{-1}\beta = \widehat{S}\widehat{R}\widehat{S}^{-1}\beta \qquad \text{(mean vector)}, \tag{3.6}$$

$$N^{-\frac{1}{2}}\widehat{F}(H\circ\widehat{R})\widehat{F}N^{-\frac{1}{2}} = \widehat{S}\widehat{R}\widehat{S} \qquad \text{(covariance matrix)}. \tag{3.7}$$

However, the relations (3.6, 3.7) do not hold when the summary data are not generated from the same sample. These differences, especially in the mean (3.6), are important but often omitted by previous work.

The modified likelihood (3.5) is derived from the Propositions 3.1.1 and 3.1.2 below. Similar to the RSS likelihood 2.2, the final form of (3.5) is obtained by replacing the nuisance parameters $\{F,R\}$ with the estimates $\{\widehat{F},\widehat{R}\}$.

Proposition 3.1.1. Let $\Pi := \text{diag}(\pi)$, $\pi := (\pi_1,\ldots,\pi_p)^\top$, where $\pi_j := |\mathcal{I}_j|/n$. Assume that both H and π are non-random and do not depend on n. For any predefined, nonempty subsets $\mathcal{I} := \{\mathcal{I}_1,\ldots,\mathcal{I}_p\}$,

$$\sqrt{n}(\widehat{\beta}(\mathcal{I};X,\mathbf{y}) - FRF^{-1}\beta) \xrightarrow{d} \mathcal{N}(\mathbf{0},\Sigma^*). \tag{3.8}$$

where $\Sigma^* := (\Pi^{-\frac{1}{2}}F)\cdot[H\circ(R+\Delta(\mathbf{c}))]\cdot(\Pi^{-\frac{1}{2}}F)^\top$, $F := \sigma_y\text{diag}^{-1}(\boldsymbol{\sigma}_x)$ and $\Delta(\mathbf{c})$ is defined by (2.21).

Proof. First define the statistic $T_n^* \in \mathbb{R}^{2p\times 1}$,

$$T_n^* := n^{-1}\left(\sum_{i=1}^n m_{i1}x_{i1}y_i,\ldots,\sum_{i=1}^n m_{ip}x_{ip}y_i,\sum_{i=1}^n m_{i1}x_{i1}^2,\ldots,\sum_{i=1}^n m_{ip}x_{ip}^2\right)^\top, \tag{3.9}$$

where $m_{ij} := \mathbf{1}\{i \in \mathscr{I}_j\}$, indicating whether the genotype and phenotype data of individual i are used to compute the summary statistics of SNP j. Here we assume that the subsets $\{\mathscr{I}_j\}$ are *pre-defined* so that the indicators $\{m_{ij}\}$ are *non-random* constants.

Notice that $T_n^* = n^{-1}\sum_{i=1}^n \mathbf{t}_i^*$, where

$$\mathbf{t}_i^* := (m_{i1},\ldots,m_{ip},m_{i1},\ldots,m_{ip})^\top \circ \mathbf{t}_i, \tag{3.10}$$

$$\mathbf{t}_i := (x_{i1}y_i,\ldots,x_{ip}y_i,x_{i1}^2,\ldots,x_{ip}^2)^\top. \tag{3.11}$$

From the proof of Proposition 2.6.1 (Chapter 2), we know that $\mathbf{t}_i$'s are i.i.d. draws from $\mathbf{t}$ with mean $\boldsymbol{\mu}_T$ and covariance matrix Σ_T. Hence, T_n^* is a sum of independent but non-identical random vectors, and its asymptotic distribution is given by the Multivariate Lindeberg-Feller Central Limit Theorem [e.g. Appendix D, Greene (2012)]

$$\sqrt{n}(T_n^* - \boldsymbol{\mu}_T^*) \xrightarrow{d} \mathscr{N}(\mathbf{0},\Sigma_T^*), \tag{3.12}$$

where the asymptotic mean and covariance matrix are given by

$$\boldsymbol{\mu}_T^* := (I_2 \otimes \Pi)\cdot\boldsymbol{\mu}_T,\ \Sigma_T^* := \left(J_2 \otimes \left(\Pi^{\frac{1}{2}}\cdot H\cdot\Pi^{\frac{1}{2}}\right)\right)\circ\Sigma_T, \tag{3.13}$$

with I_2 and J_2 denoting the 2×2 identity and all-ones matrix respectively, and $\otimes$ denoting the Kronecker product.

Next, use the Multivariate Delta Method and to show that

$$\sqrt{n}(g(T_n^*) - g(\boldsymbol{\mu}_T^*)) \xrightarrow{d} \mathscr{N}(\mathbf{0},\nabla^\top g(\boldsymbol{\mu}_T^*)\Sigma_T^*\nabla g(\boldsymbol{\mu}_T^*)), \tag{3.14}$$

where the function $g(\cdot)$ is defined by (2.18) and $\nabla g(\boldsymbol{\mu}_T^*)$ is the gradient of g at $\boldsymbol{\mu}_T^*$. A straightforward calculation yields that

$$
\begin{aligned}
g(T_n^*) &= \widehat{\boldsymbol{\beta}}(\mathscr{S}; X, \mathbf{y}), && (3.15)\\
g(\boldsymbol{\mu}_T^*) &= FRF^{-1}\boldsymbol{\beta}, && (3.16)\\
\nabla^\intercal g(\boldsymbol{\mu}_T^*)\Sigma_T^*\nabla g(\boldsymbol{\mu}_T^*) &= (\Pi^{-\frac{1}{2}}F)\cdot[H\circ(R+\Delta(\mathbf{c}))]\cdot(\Pi^{-\frac{1}{2}}F)^\intercal, && (3.17)
\end{aligned}
$$

where $\Delta(\mathbf{c})$ is defined by (2.21). □

Proposition 3.1.2. For each $\boldsymbol{\beta}\in\mathbb{R}^p$,

$$\log\mathscr{N}(\widehat{\boldsymbol{\beta}}; FRF^{-1}\boldsymbol{\beta}, N^{-\frac{1}{2}}F(H\circ R)FN^{-\frac{1}{2}}) - \log\mathscr{N}(\widehat{\boldsymbol{\beta}}; FRF^{-1}\boldsymbol{\beta}, n^{-1}\Sigma^*) = \mathscr{O}_p(\max_j c_j^2).$$

Proof. A straightforward calculation yields that

$$
\begin{aligned}
& \log\mathscr{N}(\widehat{\boldsymbol{\beta}}; FRF^{-1}\boldsymbol{\beta}, N^{-\frac{1}{2}}F(H\circ R)FN^{-\frac{1}{2}}) - \log\mathscr{N}(\widehat{\boldsymbol{\beta}}; FRF^{-1}\boldsymbol{\beta}, n^{-1}\Sigma^*)\\
= \;& \frac{1}{2}\left\{\log|H\circ(R+\Delta(\mathbf{c}))| - \log|H\circ R| + \boldsymbol{\lambda}^\intercal\Pi^{\frac{1}{2}}F^{-1}[(H\circ(R+\Delta(\mathbf{c})))^{-1} - (H\circ R)^{-1}]F^{-1}\Pi^{\frac{1}{2}}\boldsymbol{\lambda}\right\},
\end{aligned}
$$

where $\boldsymbol{\lambda} := \sqrt{n}(\widehat{\boldsymbol{\beta}} - FRF^{-1}\boldsymbol{\beta})$. Since the determinant and inverse of a matrix are both continuous, we invoke $\boldsymbol{\lambda} = \mathscr{O}_p(1)$ and $\Delta(\mathbf{c}) = \mathscr{O}(\max_j c_j^2)$ (shown in Proposition 3.1.1), to complete the proof. □

3.1.2 Imputation quality

The RSS likelihood (2.2) assumes that the single-SNP summary data are computed at fully observed genotypes. In typical GWAS, however, not all SNPs are directly assayed, and the (missing) genotypes of untyped SNPs are obtained by probabilistic prediction (a.k.a "genotype imputation"; see Section 3.2.2). Here we modify the RSS likelihood for the summary data generated from imputed genotypes.

We first outline the assumptions used in later derivations.

- The true centered genotypes of n individuals $\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_n^* \overset{\text{i.i.d.}}{\sim} \boldsymbol{x}^*$, where $\mathrm{E}(\boldsymbol{x}^*) = \mathbf{0}$, $\mathrm{Var}(\boldsymbol{x}^*) = \Sigma_x^* = \mathrm{diag}(\boldsymbol{\sigma}_x^*) R^* \mathrm{diag}(\boldsymbol{\sigma}_x^*)$, and $\boldsymbol{\sigma}_x^* := (\sigma_{x,1}^*, \ldots, \sigma_{x,p}^*)^\top$.

- The imputed centered genotypes of n individuals $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \overset{\text{i.i.d.}}{\sim} \boldsymbol{x}$, where $\mathrm{E}(\boldsymbol{x}) = \mathbf{0}$, $\mathrm{Var}(\boldsymbol{x}) = \Sigma_x = \mathrm{diag}(\boldsymbol{\sigma}_x) R \mathrm{diag}(\boldsymbol{\sigma}_x)$, and $\boldsymbol{\sigma}_x := (\sigma_{x,1}, \ldots, \sigma_{x,p})^\top$.

- The imputed and true genotypes follow the *measurement error model*:

$$\boldsymbol{x} = \boldsymbol{x}^* + \boldsymbol{\eta} \tag{3.18}$$

  where $\mathrm{E}(\boldsymbol{\eta}) = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{\eta}) = \Sigma_\eta$. Note that the diagonal elements of Σ_η (i.e. variances of $\boldsymbol{\eta}$) reflect the imputation quality of each SNP: large variance indicates that the SNP is poorly imputed.

- The centered phenotypes of n individuals $y_1, \ldots, y_n \overset{\text{i.i.d.}}{\sim} y$, where

$$y = (\boldsymbol{x}^*)^\top \boldsymbol{\beta} + \epsilon, \tag{3.19}$$

  with the error term satisfying $\mathrm{E}(\epsilon) = 0$ and $\mathrm{Var}(\epsilon) = \tau^{-1}$. Note that the coefficients $\boldsymbol{\beta}$ are the effects of each SNP on phenotype based on the *true* genotypes, rather than the *imputed* genotypes.

- The true genotype $\boldsymbol{x}^*$, measurement error $\boldsymbol{\eta}$ and residual error ϵ are mutually independent.

- The single-SNP summary statistics $\{\widehat{\boldsymbol{\beta}}, \widehat{S}\}$ are computed from the phenotypes and imputed genotypes $\{y_i, \boldsymbol{x}_i\}$.

With this in place, the modified RSS likelihood of $\boldsymbol{\beta}$ accounting for imputation quality is given as follows.

Definition 3.1.2.

$$L_{\text{rss}}^{\text{impute}}(\boldsymbol{\beta}) := \mathcal{N}(\widehat{\boldsymbol{\beta}}; (\widehat{S}\widehat{R}\widehat{S}^{-1} - \text{diag}^{-2}(\widehat{\boldsymbol{\sigma}}_x)\Sigma_\eta)\boldsymbol{\beta}, \widehat{S}\widehat{R}\widehat{S}). \tag{3.20}$$

Note that the modified likelihood (3.20) includes the original RSS likelihood (2.2) as a special case. This is because when all SNPs are directly genotyped, the measurement error $\boldsymbol{\eta}$ is zero and Σ_η becomes an all-zero matrix.

The modified likelihood (3.20) is derived from the Propositions 3.1.3 and 3.1.4 below. Similar to the RSS likelihood (2.2), the final form of (3.20) is obtained by replacing the nuisance parameters $\{S, R, \boldsymbol{\sigma}_x\}$ with their estimates $\{\widehat{S}, \widehat{R}, \widehat{\boldsymbol{\sigma}}_x\}$.

Proposition 3.1.3. Let $\widetilde{\Sigma} := \sigma_y^2 \text{diag}^{-1}(\boldsymbol{\sigma}_x)(R + \widetilde{\Delta}(\mathbf{c}))\text{diag}^{-1}(\boldsymbol{\sigma}_x)$.

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \text{diag}^{-2}(\boldsymbol{\sigma}_x)(\Sigma_x - \Sigma_\eta)\boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \widetilde{\Sigma}), \tag{3.21}$$

where $\widetilde{\Delta}(\mathbf{c}) \in \mathbb{R}^{p \times p}$ is a continuous function of $\mathbf{c}$ and $\widetilde{\Delta}(\mathbf{c}) = \mathcal{O}(\max_j c_j^2)$.

Proof. The proof is almost identical to the proof of Proposition 2.6.1 (Chapter 2). Here we only highlight the differences.

First, $g(\boldsymbol{\mu}_T)$ is different from Proposition 2.2. Specifically,

$$g(\boldsymbol{\mu}_T) = \text{diag}^{-2}(\boldsymbol{\sigma}_x)\Sigma_x^*\boldsymbol{\beta} = \text{diag}^{-2}(\boldsymbol{\sigma}_x)(\Sigma_x - \Sigma_\eta)\boldsymbol{\beta}, \tag{3.22}$$

where the last equation holds because $\boldsymbol{x}^*$ and $\boldsymbol{\eta}$ are mutually independent.

Second, $\nabla^\top g(\boldsymbol{\mu}_T)\Sigma_T \nabla g(\boldsymbol{\mu}_T)$ also has a different analytic form:

$$\nabla^\top g(\boldsymbol{\mu}_T)\Sigma_T \nabla g(\boldsymbol{\mu}_T) = \sigma_y^2 \text{diag}^{-1}(\boldsymbol{\sigma}_x)(R + \widetilde{\Delta}(\mathbf{c}))\text{diag}^{-1}(\boldsymbol{\sigma}_x). \tag{3.23}$$

The explicit form of $\widetilde{\Delta}(\mathbf{c})$ is given by

$$\widetilde{\Delta}(\mathbf{c}) := \mathrm{diag}^{-1}(\boldsymbol{\sigma}_x) \cdot \left[\widetilde{G}_1(\mathbf{c}) + \widetilde{G}_2(\mathbf{c}) + \widetilde{G}_2^{\intercal}(\mathbf{c}) + \widetilde{G}_3(\mathbf{c})\right] \cdot \mathrm{diag}^{-1}(\boldsymbol{\sigma}_x), \tag{3.24}$$

where functions $\widetilde{G}_i(\mathbf{c}) : \mathbb{R}^{p\times 1} \mapsto \mathbb{R}^{p\times p}$ are defined as follows:

$$\begin{aligned}
\widetilde{G}_1(\mathbf{c}) &:= -(\mathbf{c}^{\intercal}\mathrm{diag}(\boldsymbol{\sigma}_x)(\Sigma_x^*)^{-1}\mathrm{diag}(\boldsymbol{\sigma}_x)\mathbf{c})\Sigma_x - \mathrm{diag}(\boldsymbol{\sigma}_x)\mathbf{c}\mathbf{c}^{\intercal}\mathrm{diag}(\boldsymbol{\sigma}_x) \\
&\quad +\mathrm{E}[((\boldsymbol{x}^*)^{\intercal}(\Sigma_x^*)^{-1}\mathrm{diag}(\boldsymbol{\sigma}_x)\mathbf{c})^2\boldsymbol{x}\boldsymbol{x}^{\intercal}], \\
\widetilde{G}_2(\mathbf{c}) &:= \mathrm{diag}^{-1}(\boldsymbol{\sigma}_x)\mathrm{diag}(\mathbf{c})\widetilde{W}(\mathbf{c}),\ \ [\widetilde{W}(\mathbf{c})]_{ij} := \sigma_{x,i}\sigma_{x,j}^2 c_i - \mathbf{c}^{\intercal}\mathrm{diag}(\boldsymbol{\sigma}_x)(\Sigma_x^*)^{-1}\mathrm{E}(x_i x_j^2 \boldsymbol{x}^*), \\
\widetilde{G}_3(\mathbf{c}) &:= \mathrm{diag}^{-1}(\boldsymbol{\sigma}_x)\ \mathrm{diag}(\mathbf{c})\ \Sigma_{xx}\ \mathrm{diag}\ (\mathbf{c})\ \mathrm{diag}^{-1}(\boldsymbol{\sigma}_x),\ \ [\Sigma_{xx}]_{ij} := \mathrm{Cov}(x_i^2, x_j^2).
\end{aligned}$$

Notice that $\widetilde{G}_i(\mathbf{c})$ are continuous functions of $\mathbf{c}$, $\widetilde{G}_i(\mathbf{0}) = \mathbf{0}$, and $\widetilde{G}_i(\mathbf{c}) = \mathcal{O}(\max_j c_j^2)$ for $i = 1,2,3$. □

Proposition 3.1.4. Let $S := n^{-\frac{1}{2}}\sigma_y \mathrm{diag}^{-1}(\boldsymbol{\sigma}_x)$. For each $\boldsymbol{\beta} \in \mathbb{R}^p$,

$$\begin{aligned}
&\log \mathcal{N}(\widehat{\boldsymbol{\beta}}; (SRS^{-1} - \mathrm{diag}^{-2}(\boldsymbol{\sigma}_x)\Sigma_\eta)\boldsymbol{\beta}, SRS) \\
=\ &\log \mathcal{N}(\widehat{\boldsymbol{\beta}}; \mathrm{diag}^{-2}(\boldsymbol{\sigma}_x)(\Sigma_x - \Sigma_\eta)\boldsymbol{\beta}, n^{-1}\widetilde{\Sigma}) + \mathcal{O}_p(\max_j c_j^2).
\end{aligned}$$

Proof. The proof is the same as the proof of Proposition 2.6.2; see Chapter 2. □

3.1.3 Uncorrected confounding

The RSS likelihood assumes that the GWAS summary data are generated in the absence of confounding effects. Although peer-reviewed and published GWAS summary data were often carefully corrected for confounding (Section 3.2.3), there may be still some uncorrected confounding left. Here we propose a variant of RSS likelihood to address this issue.

Before introducing our modification, we first review a recent and innovative approach to explicitly modeling confounding biases in GWAS summary data: "LD score regression

model" (Bulik-Sullivan et al., 2015b). The statistical model is given by,

$$\mathrm{E}(\chi_j^2|\ell_j) = nh^2\ell_j/p + na + 1, \tag{3.25}$$

where n is the sample size, p is the number of SNPs, h^2/p is the heritability per SNP, a is the contribution of confounding biases per individual, $\chi_j^2 := (\widehat{\beta}_j/s_j)^2$ is the single-SNP association χ^2 statistic and $\ell_j := \sum_{k=1}^p r_{jk}^2$ is the "LD score" of SNP j (r_{jk} is the pairwise LD between SNP j and k).

Motivated by the LD score regression model (3.25), we modify the original RSS likelihood (2.2) to incorporate confounding by introducing an additional dispersion parameter a:

$$\widehat{\boldsymbol{\beta}}|S,R,\boldsymbol{\beta} \sim \mathcal{N}(SRS^{-1}\boldsymbol{\beta}, SRS + na\cdot S^2), \tag{3.26}$$

where the parameter a is the same a in the LD score regression (3.25). Note that when $a = 0$, the modified likelihood (3.26) becomes the original RSS likelihood.

The LD score regression (3.25) and the modified RSS likelihood (3.26) are closely related, as formalized by the following proposition:

Proposition 3.1.5. If the prior distribution of $\boldsymbol{\beta}$ satisfies:

$$p(\boldsymbol{\beta}|S,R) = \textstyle\prod_{j=1}^p p(\beta_j|S,R),\ \ \mathrm{E}(\beta_j|S,R) = 0,\ \ \mathrm{Var}(\beta_j|S,R) = (p\sigma_{x,j}^2)^{-1}(h^2\sigma_y^2), \tag{3.27}$$

the LD score regression (3.25) can be derived from the modified RSS likelihood (3.26).

Proof. Let $\mathbf{z} = (\mathrm{z}_1,\ldots,\mathrm{z}_p)^\intercal$, where $\mathrm{z}_j := \widehat{\beta}_j/s_j$ is the single-SNP z-score of SNP j and $\mathrm{z}_j^2 = \chi_j^2$. Noting that $\mathbf{z} = S^{-1}\widehat{\boldsymbol{\beta}}$, we rewrite (3.26) in terms of z-scores,

$$\mathbf{z}|S,R,\boldsymbol{\beta} \sim \mathcal{N}(RS^{-1}\boldsymbol{\beta}, R + na\cdot I_p). \tag{3.28}$$

Integrating out β under prior (3.27), we obtain the LD score regression model:

$$\begin{aligned} \mathrm{E}(\mathrm{z}_j^2|S,R) &= \mathrm{E}(\mathrm{E}(\mathrm{z}_j^2|S,R,\beta)) = \mathrm{E}(\mathrm{Var}(\mathrm{z}_j|S,R,\beta)) + \mathrm{E}(\mathrm{E}^2(\mathrm{z}_j|S,R,\beta)) \\ &= 1 + na + \textstyle\sum_{k=1}^{p} r_{jk}^2 s_k^{-2} \mathrm{E}(\beta_k^2|S,R) + \sum_{k\neq\ell} r_{jk} r_{j\ell} s_k^{-1} s_\ell^{-1} \mathrm{E}(\beta_k\beta_\ell|S,R) \\ &= 1 + na + (nh^2/p)\textstyle\sum_{k=1}^{p} r_{jk}^2, \end{aligned} \tag{3.29}$$

where the last equality holds because $s_j := (\sqrt{n}\sigma_{x,j})^{-1}\sigma_y$ and $\mathrm{E}(\beta_k\beta_\ell|S,R) = 0,\ \forall k \neq \ell$. □

3.2 Practical issues

Theoretical derivations in Section 3.1 highlight a few key assumptions underlying the original RSS likelihood (2.2). Specifically, summary data should be computed from a single set of individuals (Section 3.1.1) at fully observed genotypes (Section 3.1.2), with confounding effects completely removed (Section 3.1.3). In practical applications summary data may deviate from this ideal (Chapters 4-6). In this section we consider these issues from a practical perspective, and make suggestions for how to deal with them - both when generating summary dataset for distribution and when analyzing it.

3.2.1 Data on different individuals

In many studies data are available on different individuals at different SNPs [Table 2.1 and Supplementary Figure 7 of Zhu and Stephens (2017a)]. This can happen for many reasons. For example, it can happen when combining information across individuals that are typed on different genotyping platforms [e.g. GWAS arrays, Immunochip (Cortes and Brown, 2011), Metabochip (Voight et al., 2012), TxArray (Li et al., 2015)]. Or it can happen when combining data across multiple cohorts if quality control filters remove SNPs in some cohorts and not others (Winkler et al., 2014).

It is important to note that the derivation of the RSS likelihood assumes that the summary statistics are generated from the same individuals at each SNP. Specifically, the co-

variances in likelihoods (2.2) and (2.3) depend on this assumption. [In contrast, the mean in likelihood (2.3) holds even if different individuals are used at each SNP; see Section 3.1.1 for details.] To take an extreme example, if entirely different individuals are used to compute summary data for two SNPs then the correlation in their $\hat{\beta}$ values (given β) will be 0, even if the SNPs are in complete LD.

While RSS can be modified to allow for the use of different individuals when computing summary data at different SNPs [Propositions 3.1.1 and 3.1.2; see also Zhang et al. (2016)], in practice this modification is unattractive because it requires considerable additional information in addition to the usual summary data – specifically, specification of sample overlaps for many pairs of SNPs. Instead, we recommend that genotype imputation [e.g. Servin and Stephens (2007); Marchini et al. (2007); Li et al. (2009); Marchini and Howie (2010)] be used when generating GWAS summary data for public release, so that summary statistics are computed on the same individuals for each SNP.

When distributing summary data that are *not* computed on the same individuals, we recommend that at least the sample size used to compute data at each SNP also be made available, since these may be helpful both in modeling and in assessing the likely scope of the problem (Section 3.2.5). (Absent this, analysts may be able to estimate the number of individuals used at each SNP from $\{\hat{s}_j\}$ and information on allele frequency of the SNP.)

3.2.2 Imputation quality

Many GWAS make use of genotype imputation to estimate genotypes that were not actually observed. Like almost all GWAS analysis methods that are used in practice, the RSS likelihood (2.2) does not formally incorporate the potential for error in the imputed genotypes.

In principle the RSS likelihood can be extended to account for imputation errors (Propositions 3.1.3 and 3.1.4). However, this extension requires extra information – the imputation quality for each SNP – that is not always available. Fortunately, however, applying RSS to imputed genotypes, ignoring imputation quality, seems likely to provide sensible (if con-

servative) inferences in most cases. This is because imputation errors will tend to reduce estimated effects compared with what would have been obtained if all SNPs were typed: for example, if a SNP is poorly imputed then its estimated coefficient in the multiple regression model will be shrunk towards zero, and some of that SNP's contribution to heritability will be lost. This issue is not restricted to RSS: indeed, it will also occur in analyses of individual-level data that use imputed genotypes.

A complimentary approach is to compile a list of SNPs that are expected, *a priori*, to be "well imputed" (Bulik-Sullivan et al., 2015b), and to apply RSS only to these SNPs. This cannot remedy the loss of poorly-imputed SNPs' contributions to heritability, but it may help avoid poorly-imputed SNPs undesirably influencing estimates of model hyper-parameters.

3.2.3 Uncorrected confounding

Another important issue that can impact many association studies is "confounding" due to population stratification (Devlin and Roeder, 1999; Price et al., 2010), which can cause over-estimation of genetic effects and heritability if not appropriately corrected for. A standard approach to dealing with this problem is to use methods such as principal components analysis [e.g. Price et al. (2006); Patterson et al. (2006); Tucker et al. (2014)] and/or linear mixed models [e.g. Kang et al. (2010); Lippert et al. (2011); Zhou and Stephens (2012); Yang et al. (2014)] to correct for stratification. These methods require access to the individual-level genotype data, and so cannot be used directly by analysts with access only to summary data. Instead they must be used by analysts who are computing the summary data for public distribution: doing so should substantially reduce the effects of confounding on summary data analyses, including RSS.

A complementary approach to dealing with population stratification is to directly model its effects on the summary data. One recent and innovative approach to this is LD score regression (Bulik-Sullivan et al., 2015b), which uses the intercept of a regression of signal versus "LD score" to assess the effects of confounding. Along similar lines, we could modify

the RSS likelihood to incorporate the effects of confounding by introducing an additional dispersion parameter (3.26); see Proposition 3.1.5. This modification would not require extra information, and may have an additional benefit of improving robustness of RSS to other model misspecification issues (e.g. genotyping error, mismatches between LD in the reference panel and sample). However, this modification requires additional computation [some linear algebra simplifications we use when implementing (2.2) do not hold for (3.26)], and we have not yet implemented it.

3.2.4 Filtering and diagnostics

Some of the recommendations above can only be implemented when the summary data are being computed from individual data for public distribution, and not at a later stage when only the summary data are available. This raises the question, what can analysts with only access to summary data do to check that their results are likely reliable? This may be the trickiest part of summary data analysis: even with access to the full individual-level data it can be hard to assess all sources of bias and error. Recognizing that there is no universal approach that will guarantee reliable results we nonetheless hope to provide some useful suggestions.

Since the RSS likelihood (2.2) defines a statistical model, it is possible to perform a model fit diagnostic check. A generic approach to model checking (e.g. common in linear regression) is to first fit the model, compute residuals that measure deviations of observations from expected values, and then discard outlying observations before refitting the model. We have implemented an approach along these lines for identifying outlying SNPs, as follows. First, after fitting the model, we compute the residual (the difference between the observed $\hat{\beta}$ and its fitted expected value) at each SNP. We then perform a "leave-one-out" (LOO) check on each residual: we compute its conditional expectation and variance given the residuals at all other SNPs, and compute a diagnostic z-score based on how the observed residual compares with this expectation and variance. See Box 1 below for details. This approach

targets SNPs whose summary data are most inconsistent with data at other nearby SNPs in LD. If the model is correctly specified for a given SNP then its diagnostic z-score approximately follows a standard normal distribution, from which a large deviation indicates potential misspecification. To assess robustness of RSS fit one can filter out SNPs with large diagnostic z-scores, and refit the RSS model on the remaining SNPs.

Box 1: "Leave-one-out" (LOO) residual diagnostic

We define the marginally standardized error of $\widehat{\beta}$ as $\mathbf{e} := S^{-1}(\widehat{\beta} - SRS^{-1}\beta)$. When the model is correctly specified (2.23), $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, R)$. For each $i \in [p]$, the univariate complete conditional distribution of the ith entry of $\mathbf{e}$ is also normal:

$$e_i|\mathbf{e}_{-i} \sim \mathcal{N}\left(-v_{ii}^{-1}\textstyle\sum_{i\neq j} v_{ij}e_j, v_{ii}^{-1}\right), \tag{3.30}$$

where v_{ij} is the (i,j)-entry of matrix V, $V := R^{-1}$. The conditional distribution (3.30) provides us a way to impute the error of SNP i based on the errors of other SNPs. Furthermore, we can evaluate the quality of imputation using the following z-score:

$$z_i(\mathbf{e}) := \frac{e_i - \mathrm{E}(e_i|\mathbf{e}_{-i})}{\sqrt{\mathrm{Var}(e_i|\mathbf{e}_{-i})}} = \sqrt{v_{ii}}\left(e_i + v_{ii}^{-1}\textstyle\sum_{i\neq j} v_{ij}e_j\right) \sim \mathcal{N}(0,1). \tag{3.31}$$

The error $\mathbf{e}$ is not observed because of the unknown true effect β. Instead, we can only calculate the marginally standardized residual of $\widehat{\beta}$, $\check{\mathbf{e}} := S^{-1}(\widehat{\beta} - SRS^{-1}\check{\beta})$, where $\check{\beta}$ is the posterior estimate of β obtained from the MCMC. We perform the LOO imputation (3.30) on the residual $\check{\mathbf{e}}$. The corresponding z-scores $\{z_i(\check{\mathbf{e}})\}$ empirically measure the goodness of fit, and thus can be used to filter out SNPs that may be misspecified in the RSS likelihood.

Other simpler filters are of course possible, and multiple filters can be used together. One widely-used filter simply discards SNPs with sample sizes lower than a certain cut-off (Pickrell, 2014). This can reduce problems caused by SNPs being typed on different subsets

of individuals discussed above (Section 3.2.1). Another possibility is to filter out SNPs that are in very strong LD with one another, since these have potential for producing severe misspecification (Section 3.2.5). Some advantages of the model-based LOO diagnostic include that it could detect model misspecification problems from several sources – including genotyping error or misspecification of the LD matrix R – and not only those caused by typing of different individuals at different SNPs. Also, the sample size filter cannot be used unless the sample size for each SNP is made available, which is not always the case (Table 2.1). Finally, choice of threshold for the diagnostic z-score can be guided by the standard normal distribution; in contrast, selecting principled thresholds for sample sizes seems less straightforward [and a stringent threshold can yield conservative results; see Supplementary Figure 8 of Zhu and Stephens (2017a)] On the other hand, the LOO diagnostic may tend to filter out SNPs that show particularly strong signal (if they are not in LD with other SNPs), an undesirable property that should be remembered when interpreting results post-filtering see Supplementary Figure 9 of Zhu and Stephens (2017a).

3.2.5 Extreme example

One way to help avoid problems with model misspecification is to be aware of the most severe ways in which things can go wrong. In this vein, we offer one illustrative example that we encountered when applying RSS to the summary data of a blood lipid GWAS (Global Lipids Genetics Consortium, 2013).

Table 3.1 shows summary statistics for high-density lipoprotein (HDL) cholesterol for seven SNPs in the gene *ADH5* that are in complete LD with one another in the reference panel (1000 Genomes European $r^2 = 1$). If summary data were computed on the same set of individuals at each SNP, then they would be expected to vary very little among SNPs that are in such strong LD. And indeed, the RSS likelihood captures this expectation. However, in this case we see that the summary data actually vary considerably at some SNPs. The differences between one SNP (rs7683704) and the others are likely explained by the fact

SNP	n_j	$\hat{\beta}_j$	$\hat{\sigma}_j$	1-SNP $\log_{10}$ BF	2-SNP $\log_{10}$ BF	r^2
rs7683704	187,124	0.0096	0.0058	-0.676	NA	1.0
rs13125919	94,311	0.0038	0.0079	-1.084	172.638	1.0
rs4699701	94,311	0.0054	0.0081	-1.028	88.364	1.0
rs17595424	94,274	0.0055	0.0081	-1.024	83.925	1.0
rs11547772	94,311	0.0056	0.0081	-1.021	79.756	1.0
rs7683802	94,311	0.0056	0.0081	-1.021	79.756	1.0
rs4699699	94,311	0.0058	0.0081	-1.013	71.580	1.0

Table 3.1: Example of problems that can arise due to severe model misspecification. The table reports the sample sizes, single-SNP effect size estimates, SEs, and 1-SNP BFs of seven SNPs that are in complete LD in the reference panel (1000 Genomes, European ancestry). The 2-SNP BFs reported are for rs7683704 with each of the other SNPs. These unreasonably large 2-SNP BFs are due to model misspecification.

that this SNP was typed on more individuals: data at this SNP come from both GWAS (up to 94,595 individuals) and Metabochip arrays (up to 93,982 individuals). Thus this is an example of model misspecification due to SNPs being typed on different individuals. However, another SNP, rs13125919, also shows notable differences in summary data from the other SNPs, for reasons that are unclear to us. (This highlights a challenge of working with summary data – it is difficult to investigate the source of such anomalies without access to individual data.)

Whatever the reasons, applying RSS to these data results in severe model misspecification: based on their LD patterns RSS expects data at these SNPs to be almost identical, but they are not. This severe model misspecification can lead to unreliable results. For example, we used the RSS likelihood (2.2) to compute the 1-SNP and 2-SNP Bayes factors (BFs) [as in Servin and Stephens (2007); see also Chen et al. (2015)]. None of the SNPs shows evidence for marginal association with HDL ($\log_{10}$ 1-SNP BF are all negative, indicating evidence for the null). However, the 2-SNP BFs for rs7683704 together with any of the other SNPs are unreasonably large, due to the severe model misspecification.

We emphasize that this is an extreme example, chosen to highlight the worst things that can go wrong. For simulations illustrating the effects of less extreme model misspecification, see Supplementary Figure 6 of Zhu and Stephens (2017a).

CHAPTER 4

ESTIMATE SNP HERITABILITY USING GWAS SUMMARY DATA

Our first application of RSS is estimating SNP heritability from GWAS summary statistics. SNP heritability, or PVE, is the fraction of total variation in phenotypes $\mathbf{y}$ that is explained by genotypes X of all SNPs available in the study. Learning PVE from genetic data can help improve our understanding of complex human traits (Visscher et al., 2008; Manolio et al., 2009; de los Campos et al., 2015). PVE is also closely related to the "coefficient of determination" (R^2), an important concept in regression analysis. However, similar to R^2, traditional definitions of PVE often depend on individual-level data $\{X, \mathbf{y}\}$, which are not suitable for GWAS summary statistics. To address this question, we introduce a new definition of PVE based on summary data (Section 4.1). Based on the new definition, we estimate PVE using the posterior distribution of the multiple regression coefficients β (Section 4.3), which is obtained by combining the RSS likelihood (2.2) with previously-prosed prior distributions (Section 4.2). Finally, we test our method through simulations based on real genotypes (Section 4.4), and apply the method to estimate PVE of adult human height (Section 4.5) from summary statistics of a recent large study (Wood et al., 2014).

4.1 Define PVE based on summary data

Given the full data $\{X, \mathbf{y}\}$ and the true value of $\{\beta, \tau\}$ in model (2.8), Guan and Stephens (2011) define the PVE as

$$\mathrm{PVE}(\beta, \tau) := \frac{V(X\beta)}{\tau^{-1} + V(X\beta)}. \tag{4.1}$$

By this definition, PVE reflects the proportion of total variation in phenotypes $\mathbf{y}$ that is explained by available genotypes X. Guan and Stephens (2011) then estimate PVE using the posterior sample of $\{\beta, \tau\}$.

Because X is unknown here, we cannot compute PVE as defined above even if β and τ were known. Moreover, τ does not appear in our inference procedure. For these reasons we

introduce the "Summary PVE" (SPVE) as an analogue of PVE for our setting:

$$\text{SPVE}(\boldsymbol{\beta}) := \sum_{i,j} \frac{\hat{r}_{ij}\beta_i\beta_j}{\sqrt{(n\hat{\sigma}_i^2+\hat{\beta}_i^2)(n\hat{\sigma}_j^2+\hat{\beta}_j^2)}}. \tag{4.2}$$

This definition is motivated by noting that PVE can be approximated by replacing τ^{-1} with $\text{V}(\mathbf{y}) - V(X\boldsymbol{\beta})$:

$$\text{PVE} \approx \frac{V(X\boldsymbol{\beta})}{V(\mathbf{y})} = \sum_{i,j} \frac{X_i^\top X_j}{\mathbf{y}^\top\mathbf{y}}\beta_i\beta_j = \sum_{i,j} \frac{\hat{r}_{ij}^{\text{sam}}\beta_i\beta_j}{\sqrt{(n\hat{\sigma}_i^2+\hat{\beta}_i^2)(n\hat{\sigma}_j^2+\hat{\beta}_j^2)}}, \tag{4.3}$$

where $\hat{r}_{ij}^{\text{sam}}$ is the (i,j)-entry of (unknown) sample LD matrix of the study cohort ($\widehat{R}^{\text{sam}}$), which we approximate in SPVE by $\hat{r}_{ij}$, and the last equation in (4.3) holds because of (1.2-1.3). Simulations (Figure 4.1) using both synthetic and real genotypes verify that SPVE is a highly accurate approximation to PVE, given the true value of $\boldsymbol{\beta}$.

We infer PVE using the posterior draws of SPVE, which are obtained by computing $\text{SPVE}(\boldsymbol{\beta}^{(i)})$ for each sampled value $\boldsymbol{\beta}^{(i)}$ from our MCMC algorithms [Supplementary Appendix B of Zhu and Stephens (2017a)]. Unlike the original PVE (4.1), the definition of SPVE (4.2) is not bounded above by 1. Although we have not seen any estimates above 1 in our simulations or data analyses, we expect this could occur if the posterior of $\boldsymbol{\beta}$ is poorly simulated and/or $\widehat{R}$ is severely misspecified.

4.2 Prior specification

Using the RSS likelihood (2.2), we perform Bayesian inference for the multiple regression coefficients $\boldsymbol{\beta}$. If $\{S,R\}$ were known, then one could perform Bayesian inference by specifying a prior on $\boldsymbol{\beta}$:

$$\underbrace{p(\boldsymbol{\beta}|\widehat{\boldsymbol{\beta}},S,R)}_{\text{Posterior}} \propto \underbrace{p(\widehat{\boldsymbol{\beta}}|S,R,\boldsymbol{\beta})}_{\text{Likelihood}} \cdot \underbrace{p(\boldsymbol{\beta}|S,R)}_{\text{Prior}}. \tag{4.4}$$

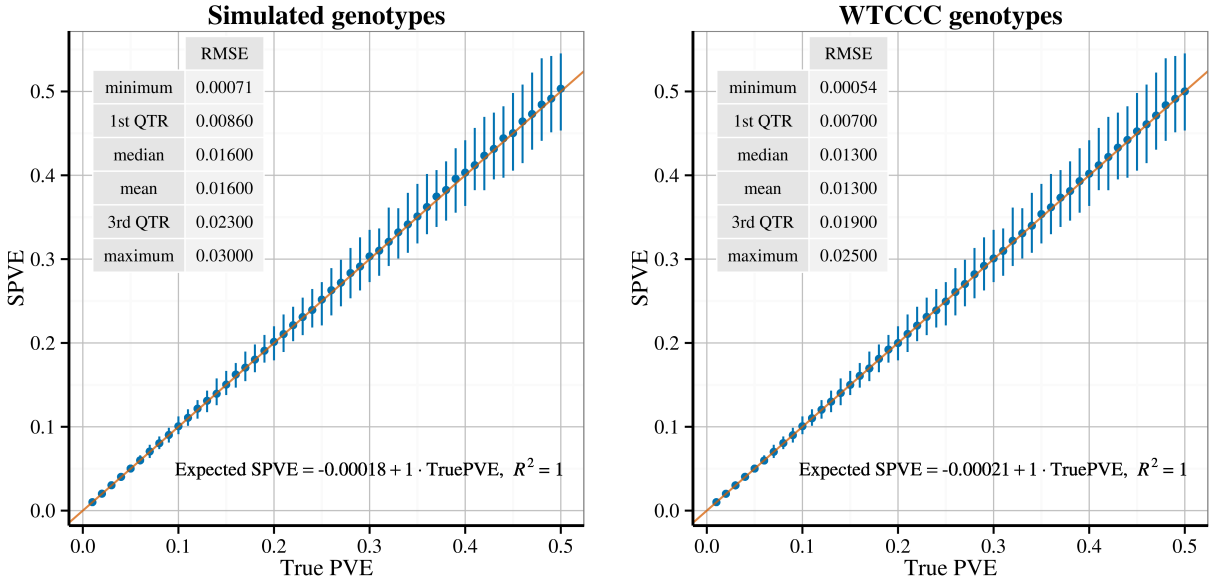


Figure 4.1: Comparison of true PVE and Summary PVE (SPVE) given the true β. The true PVE is computed from the true values of $\{\beta, \tau\}$ and the individual-level data $\{X, \mathbf{y}\}$. The SPVE is computed from the true β, the summary-level data $\{\hat{\beta}_j, \hat{\sigma}_j^2\}$ and the estimated LD matrix $\widehat{R}$. The simulated genotypes consist of 10,000 independent SNPs from 1,000 individuals, so $\widehat{R}$ is set as identity matrix; The real genotypes are 10,000 correlated SNPs randomly drawn from chromosome 16 (WTCCC UK Blood Service control group, 1,458 individuals), and $\widehat{R}$ is estimated from WTCCC 1958 British Birth Cohort (1,480 individuals) and HapMap CEU genetic maps using the shrinkage method in Wen and Stephens (2010). Solid dots indicate sample means of 200 replicates; vertical bars indicate symmetric 95% intervals; orange line indicates the reference line with intercept 0 and slope 1. The tables summarize the RMSEs between SPVE and true PVE.

To deal with unknown $\{S,R\}$ the RSS likelihood (2.2) approximates the likelihood in (4.4) by replacing $\{S,R\}$ with their estimates $\{\widehat{S},\widehat{R}\}$. We take a similar approach to prior specification: we specify a prior $p(\beta|S,R)$ and replace $\{S,R\}$ with $\{\widehat{S},\widehat{R}\}$.

Instead of developing new prior for the regression coefficients β, here we modify existing prior specifications from previous work (Zhou et al., 2013). A major benefit of using existing prior distribution is that we can "fairly" our methods with previous work, since the major difference only lies in the underlying likelihood function.

Our prior specification is based on the prior from Zhou et al. (2013) which was designed for the analysis of individual-level GWAS data. This prior assumes that β is independent of R *a priori*, with the prior on β_j being a mixture of two normal distributions

$$\beta_j \sim \pi\mathcal{N}(0,\sigma_B^2+\sigma_P^2)+(1-\pi)\mathcal{N}(0,\sigma_P^2). \tag{4.5}$$

The motivation is that the first ("sparse") component can capture rare "large" effects, while the second ("polygenic") component can capture large numbers of very small effects. To specify priors on the variances $\{\sigma_B^2,\sigma_P^2\}$ Zhou et al. (2013) introduce two free parameters $h,\rho\in[0,1]$ where h[1] represents, roughly, the proportion of variance in $\mathbf{y}$ explained by X, and ρ represents the proportion of genetic variance explained by the sparse component. They write σ_B^2 and σ_P^2 as functions of π,h,ρ and place independent priors on the hyper-parameters (π,h,ρ):

$$\log\pi\sim\mathcal{U}(\log(1/p),\log 1),\quad h\sim\mathcal{U}(0,1),\quad \rho\sim\mathcal{U}(0,1). \tag{4.6}$$

See Zhou et al. (2013) for details.

Here we must modify this prior slightly because the original definitions of σ_B and σ_P depend on the genotypes X (which here are unknown) and the residual variance τ^{-1} (which

1. Parameter h is related to heritability (Visscher et al., 2008), which is often denoted as h^2 in genetics literature. We use h here to keep notation consistent with previous closely-related work (Zhou et al., 2013; Guan and Stephens, 2011).

does not appear in our likelihood). Specifically we define

$$\sigma_B^2(S) := h\rho\left(\pi\sum_{j=1}^{p} n^{-1}s_j^{-2}\right)^{-1},\ \ \sigma_P^2(S) := h(1-\rho)\left(\sum_{j=1}^{p} n^{-1}s_j^{-2}\right)^{-1}, \tag{4.7}$$

where s_j is the jth diagonal entry of S. Because $ns_j^2 = \sigma_y^2\sigma_{x,j}^{-2}$, definitions (4.7) ensure that the effect sizes of both components do not depend on n, and have the same measurement unit as the phenotype $\mathbf{y}$. Further, with these definitions, ρ and h have interpretations similar to those in previous work. Specifically, $\rho = (\pi\sigma_B^2)/(\pi\sigma_B^2+\sigma_P^2)$, so it represents the expected proportion of total genetic variation explained by the sparse components. Parameter h represents, roughly, the proportion of the total variation in $\mathbf{y}$ explained by X, as formalized by the following proposition:

Proposition 4.2.1. If $\beta|S$ is distributed as (4.5), with (4.7), then

$$\mathrm{E}[V(X\beta)] = h\cdot\mathrm{E}[V(\mathbf{y})], \tag{4.8}$$

where $V(X\beta)$ and $V(\mathbf{y})$ are the sample variance of $X\beta$ and $\mathbf{y}$ respectively.

Proof. Since the matrix X is column-centered,

$$V(X\beta) = n^{-1}\textstyle\sum_{i=1}^{n}(x_i^\intercal\beta)^2 = n^{-1}\mathrm{trace}[(X\beta)(X\beta)^\intercal] = n^{-1}\beta^\intercal X^\intercal X\beta, \tag{4.9}$$

and therefore,

$$\mathrm{E}[V(X\beta)|S,X] = \mu_\beta^\intercal\cdot(n^{-1}X^\intercal X)\cdot\mu_\beta + \mathrm{trace}[(n^{-1}X^\intercal X)\cdot\Sigma_\beta], \tag{4.10}$$

where $\boldsymbol{\mu}_\beta := \mathrm{E}(\boldsymbol{\beta}|S) = \mathbf{0}$ and $\Sigma_\beta := \mathrm{Var}(\boldsymbol{\beta}|S) = (\pi\sigma_B^2 + \sigma_P^2)\cdot I_p$. Hence,

$$\begin{aligned}\mathrm{E}[V(X\boldsymbol{\beta})] = \mathrm{E}[\mathrm{E}[V(X\boldsymbol{\beta})|S,X]] &= (\pi\sigma_B^2+\sigma_P^2)\cdot\sum_{j=1}^{p}\mathrm{E}[V(X_j)] = \frac{h}{\sum_{j=1}^{p} n^{-1}s_j^{-2}}\cdot\sum_{j=1}^{p}\mathrm{E}[V(X_j)] \\ &= \frac{h}{\sum_{j=1}^{p} n^{-1}s_j^{-2}}\cdot\sum_{j=1}^{p} n^{-1}s_j^{-2}\mathrm{E}[V(\mathbf{y})] = h\cdot\mathrm{E}[V(\mathbf{y})],\end{aligned}$$

where the last line holds because $ns_j^2 = \sigma_y^2\sigma_{x,j}^{-2}$ and $\mathrm{E}[V(X_j)] = n^{-1}s_j^{-2}\mathrm{E}[V(\mathbf{y})]$. □

Because of its similarity with the prior from "Bayesian sparse linear mixed model" [BSLMM, Zhou et al. (2013)], we refer to our modified prior as BSLMM. We also implement a version of this prior where $\rho = 1$. This sets the polygenic variance $\sigma_P^2 = 0$, making the prior on $\boldsymbol{\beta}$ sparse, and corresponds closely to the prior from "Bayesian variable selection regression" [BVSR, Guan and Stephens (2011)]. We therefore refer to this special case as BVSR here.

4.3 Posterior computation

We use Markov chain Monte Carlo (MCMC) to sample from the posterior distribution of $\boldsymbol{\beta}$ under RSS-BSLMM and RSS-BVSR models; see Supplementary Appendix B of Zhu and Stephens (2017a) for details.

To fit the RSS-BSLMM model, we implement a new algorithm that is different from previous work (Zhou et al., 2013). Instead of integrating out $\boldsymbol{\beta}$ analytically, we perform MCMC sampling on $\boldsymbol{\beta}$ directly. Most of MCMC updates in this algorithm have linear complexity, with only a few "expensive" exceptions. The costs of these "expensive" updates are further reduced from being cubic in the total number of SNPs to being quadratic, by leveraging the banded structure of LD matrix $\widehat{R}$ (Wen and Stephens, 2010).

To fit the RSS-BVSR model, we largely follow the algorithm developed in Guan and Stephens (2011), which exploit sparsity. Specifically, computation time per iteration scales cubically with the number of SNPs with non-zero effects, which is much smaller than the

total number of SNPs under sparse assumptions. Setting a fixed maximum number of non-zero effects, and/or, using the banded LD structure to guide variable selection, can further improve computational performance, but we do not use these strategies here.

All computations in this section were performed on a Linux system with a single Intel E5-2670 2.6GHz or AMD Opteron 6386 SE processor. Computation times for simulation studies and data analyses are shown in Supplementary Figure 5 and Supplementary Table 6 of Zhu and Stephens (2017a) respectively. Software implementing the methods is available at `https://github.com/stephenslab/rss`.

4.4 Simulations

Here we use simulations to assess the performance of RSS for estimating PVE. We first use real genotypes from Wellcome Trust Case Control Consortium (2007) (specifically, the 1,458 individuals from the UK Blood Service Control Group) and simulated phenotypes, and then perform single-SNP analysis on them to generate summary statistics. To mimic the behavior of real data, we use LD matrix estimated from a different set of individuals (the 1,480 individuals from 1958 British Birth Cohort). To reduce computation the simulations use genotypes from a single chromosome (12,758 SNPs on chromosome 16). One consequence of this is that the simulated effect sizes per SNP in some scenarios are often larger than would be expected in a typical GWAS [Table 2.1 and Supplementary Figure 3 of Zhu and Stephens (2017a)]. This is, in some ways, not an ideal case for RSS, because the likelihood derivation assumes that effect sizes are small (Proposition 2.6.2). We use the simulations to demonstrate that inferences from RSS agree well with both the simulation ground truth, and with results from methods based on the full data [specifically, BVSR and BSLMM implemented in the software package GEMMA (Zhou and Stephens, 2012)].

We first simulated phenotypes under two genetic architectures:

- Scenario 1.1 (sparse): randomly select 50 “causal” SNPs, with effects coming from

$\mathcal{N}(0,1)$; effects of remaining SNPs are zero.

- Scenario 1.2 (polygenic): randomly select 50 "causal" SNPs, with effects coming from $\mathcal{N}(0,1)$; effects of remaining SNPs come from $\mathcal{N}(0,0.001^2)$.

For each scenario we simulated datasets with true PVE ranging from 0.05 to 0.5 (in steps of 0.05, with 50 independent replicates for each PVE). We ran RSS-BVSR on Scenario 1.1, and RSS-BSLMM on Scenario 1.2. Figure 4.2 summarizes the resulting PVE estimates. The estimated PVEs generally correspond well with the true values, but with a noticeable upward bias when the true PVE is large. We speculate that this upward bias is due to deviations from the assumption of small effects underlying RSS in Proposition 2.6.2. (Note that with 50 causal SNPs and PVE=0.5, on average each causal SNP explains 1% of the phenotypic variance, which is substantially higher than in typical GWAS; thus the upward bias in a typical GWAS may be less than in these simulations.)

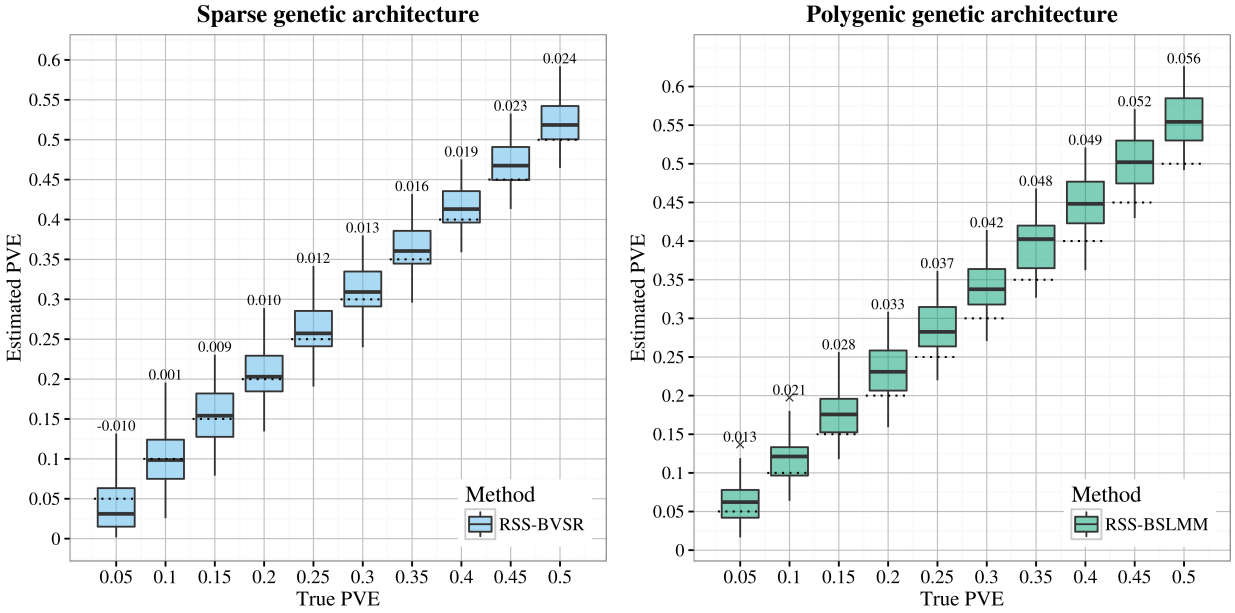


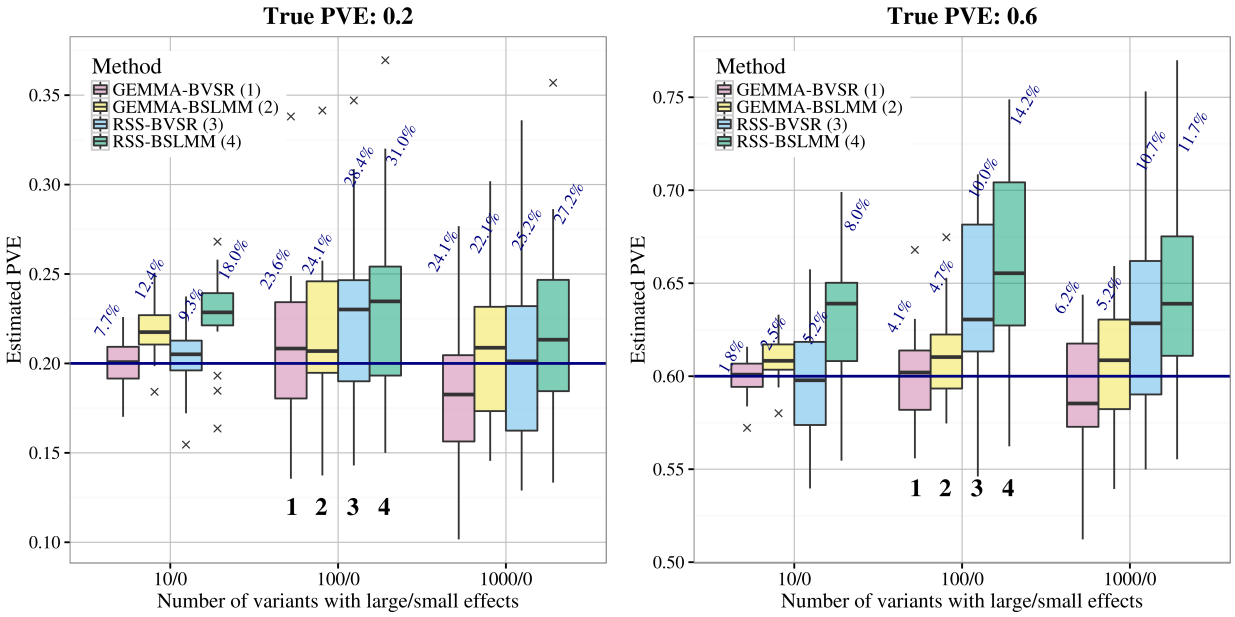
Figure 4.2: Comparison of true PVE with estimated PVE (posterior median) in Scenarios 1.1 (sparse) and 1.2 (polygenic). The dotted lines indicate the true PVEs, and the bias of estimates is reported on top of each box plot. Each box plot summarizes results from 50 replicates.

Next, we compare accuracy of PVE estimation using summary versus full data. With the genotype data as above we consider two scenarios:

- Scenario 2.1 (sparse): simulate a fixed number T of causal SNPs ($T = 10, 100, 1000$), with effect sizes coming from $\mathcal{N}(0, 1)$, and the effect sizes of the remaining SNPs are zero;

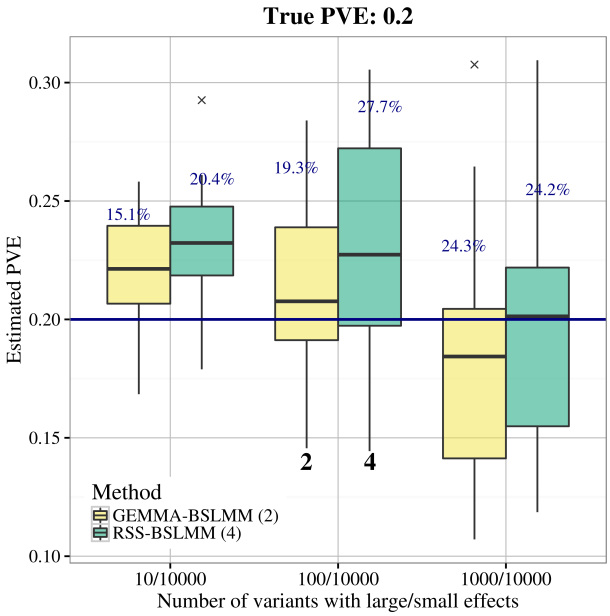
- Scenario 2.2 (polygenic): simulate two groups of causal SNPs, the first group containing a small number T of large-effect SNPs ($T = 10, 100, 1000$), plus another larger group of $10,000$ small-effect SNPs; the large effects are drawn from $\mathcal{N}(0, 1)$, the small effects are drawn from $\mathcal{N}(0, 0.001^2)$, and the effects of the remaining SNPs are zero.

For each scenario we created datasets with true PVE 0.2 and 0.6 (20 independent replicates for each parameter combination). For Scenario 2.1 we compared results from the summary data methods (RSS-BVSR and RSS-BSLMM) with the corresponding full data methods (GEMMA-BVSR and GEMMA-BSLMM). For Scenario 2.2 we compared only the BSLMM methods, since the BVSR-based methods, which assume effects are sparse, are not well suited to this setting, in terms of both computation and accuracy (Zhou et al., 2013); see also Supplementary Appendix B of Zhu and Stephens (2017a). Figure 4.3 summarizes the results. With modest true PVE (0.2), GEMMA-BVSR and RSS-BVSR perform better than other methods when the true model is very sparse (e.g. Scenario 2.1, $T = 10$), whereas GEMMA-BSLMM and RSS-BSLMM perform better when the true model is highly polygenic (e.g. Scenario 2.2, $T = 1000$). When the true PVE is large (0.6), the summary-based methods show an upward bias (Figure 4.3b and 4.3d), consistent with Figure 4.2. This bias is less severe when the true signals are more "diluted" (e.g. $T = 1000$), consistent with our speculation above that the bias is due to deviations from the "small effects" assumption. Overall, as expected, the summary data methods perform slightly less accurately than the full data methods. However, using different modeling assumptions (BVSR versus BSLMM) has a bigger impact on the results than using summary versus full data.

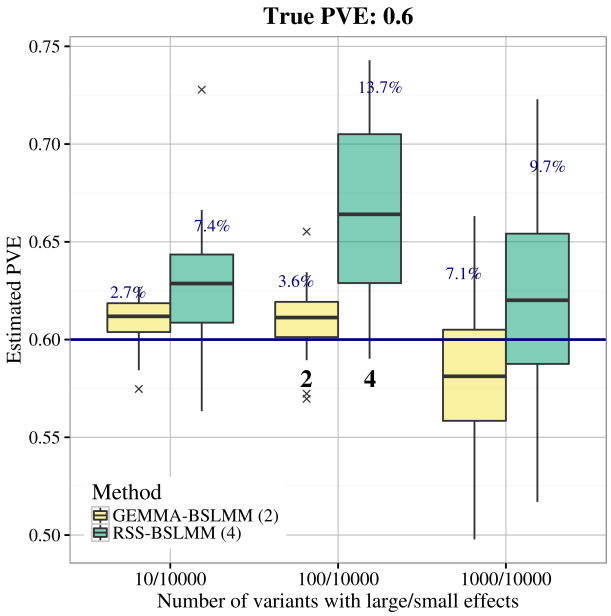


(a) Scenario 2.1

(b) Scenario 2.1



(c) Scenario 2.2



(d) Scenario 2.2

Figure 4.3: Comparison of PVE estimates (posterior median) from GEMMA and RSS in Scenario 2.1 and 2.2. The accuracy of estimation is measured by the relative RMSE, which is defined as the RMSE between the ratio of estimated over true PVEs and 1. Relative RMSE for each method is reported (percentages on top of box plots). The true PVEs are shown as the solid horizontal lines. Each box plot summarizes results from 20 replicates.

4.5 Example: human height (Wood et al., 2014)

We applied RSS to summary statistics from a GWAS of human adult height, involving 253,288 individuals of European ancestry typed at ~ 1.06 million SNPs (Wood et al., 2014). Accessing the individual-level data would be a considerable undertaking; in contrast the summary data are easily and freely available.

Following the protocol from Bulik-Sullivan et al. (2015b), we filtered out poorly imputed SNPs and then removed SNPs absent from the genetic map of HapMap European-ancestry population Release 24 (Frazer et al., 2007). To avoid negative recombination rate estimates, we excluded SNPs in regions where the genome assembly had been rearranged. We also removed triallelic sites by manual inspection in BioMart (Smedley et al., 2015). This left 1,064,575 SNPs retained for analysis. We estimated the LD matrix R using phased haplotypes from 379 European-ancestry individuals in 1000 Genomes Project Consortium (2010).

Although the summary data were generated after genotype imputation to the same reference panel [Section 1.1.2, Supplementary Note of Wood et al. (2014)], only 65% of the 1,064,575 analyzed SNPs were computed from the total sample [Supplementary Figure 7 of Zhu and Stephens (2017a)]. This is because SNP filters applied by the consortium separately in each cohort often filtered out SNPs from a subset of cohorts [Section 1.1.4, Supplementary Note of Wood et al. (2014)]. As shown in Chapter 3, properly accounting for the sample difference would require sample overlap information that is not publicly available. Instead, we directly applied the original RSS likelihood (2.2) to the summary data. As discussed in Chapter 3, this simplification results in model misspecification. To assess the impact of this, in addition to the primary analysis using all the summary data, we also performed secondary analyses after applying the LOO residual diagnostic described in Chapter 3 to filter out SNPs whose diagnostic z-scores exceeded a threshold (2 or 3).

To reduce computation time and hardware requirement, we separately analyzed each of the 22 autosomal chromosomes so that all chromosomes were run in parallel in a computer cluster. In our analysis, each chromosome used a single CPU core. To assess convergence of

the MCMC algorithm, we ran the algorithm on each dataset multiple times; results agreed well among runs (results not shown), suggesting no substantial problems with convergence. Here we report results from a single run on each chromosome with 2 million iterations. The CPU time of RSS-BVSR ranged from 1 to 36 hours, and the time of RSS-BSLMM ranged from 4 to 36 hours [Supplementary Table 6 of Zhu and Stephens (2017a)].

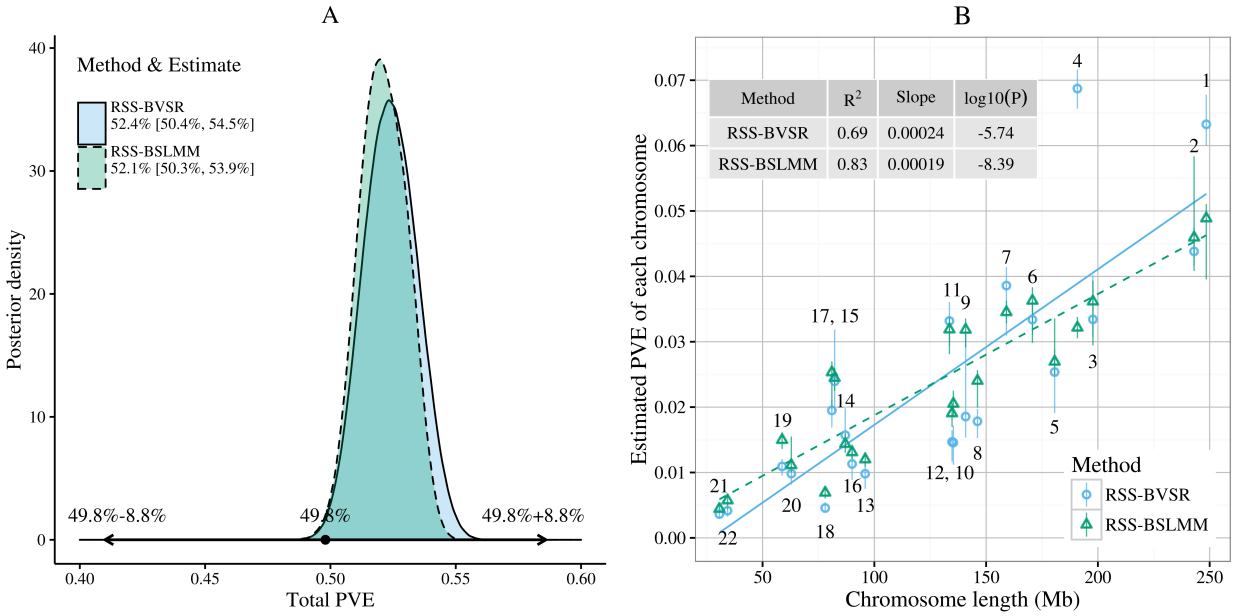


Figure 4.4: Posterior inference of PVE (SNP heritability) for adult human height. Panel A: posterior distributions of the total PVE, where the interval spanned by the arrows is the 95% confidence interval from Wood et al. (2014). Panel B: posterior median and 95% credible interval for PVE of each chromosome against the chromosome length, where each dot is labeled with chromosome number and the lines are fitted by simple linear regression (solid: RSS-BVSR; dash: RSS-BSLMM). The simple linear regression output is shown in Supplementary Table 2 of Zhu and Stephens (2017a). The data to reproduce Panel B are provided in Supplementary Table 3 of Zhu and Stephens (2017a).

Figure 4.4 shows the estimated total and per-chromosome PVEs based on RSS-BVSR and RSS-BSLMM. For both methods, we can see an approximately linear relationship between PVE and chromosome length, consistent with a genetic architecture where many causal SNPs each contribute a small amount to PVE (a.k.a. "polygenicity"), and consistent with previous results using a mixed linear model (Yang et al., 2011) on three smaller individual-level datasets (number of SNPs: 593,521-687,398; sample size: 6,293-15,792). By summing

PVE estimates across all 22 chromosomes, we estimated the total autosomal PVE to be 52.4%, with 95% credible interval [50.4%, 54.5%] using RSS-BVSR, and 52.1%, with 95% credible interval [50.3%, 53.9%] using RSS-BSLMM. Our estimates are consistent with, but more precise than, previous estimates based on individual-level data from subsets of this GWAS. Specifically, Wood et al. (2014) estimated PVE as 49.8%, with standard error 4.4%, from individual-level data of five cohorts (number of SNPs: 0.97-1.12 million; sample size: 1,145-5,668). The increased precision of the PVE estimates illustrates one benefit of being able to analyze summary data with a large sample size.

One caveat to these results is that the RSS likelihood (2.2) ignores confounding such as population stratification (Section 3.2.3). Here the summary data were generated using genomic control, principal components and linear mixed effects, to control for population stratification within each cohort [Section 1.1.3, Supplementary Note of Wood et al. (2014)]. Thus, we might hope that confounding has limited impact on PVE estimation. However, it is difficult to be sure that all confounding has been completely removed, and any remaining confounding could upwardly bias our estimated PVE. (Unremoved confounding could similarly bias estimates based on individual-level data.)

Finally, to check for misspecification we performed the LOO residual-based diagnostic. Specifically, we ran the LOO residual imputation using the RSS-BVSR output, and then refitted the models on the filtered SNPs (absolute LOO z-score ≤ 2). This resulted in a substantial reduction in PVE estimates (RSS-BVSR: 34.0%, [32.9%, 35.0%]; RSS-BSLMM: 45.3%, [44.7%, 46.0%]). However, this may reflect the fact that the filter removed 12% of SNPs, possibly biased towards SNPs showing association signal [Supplementary Figure 9 of Zhu and Stephens (2017a)]. Results are similar based on a less stringent threshold (3); see Supplementary Table 4 of Zhu and Stephens (2017a).

CHAPTER 5

DETECT GENETIC ASSOCIATION USING GWAS SUMMARY DATA

Compared with existing summary-based methods, an important practical advantage of RSS is that multiple tasks can be performed using the same posterior sample of β. To illustrate this feature, we show in this chapter that posterior samples of β from PVE estimation (Chapter 4) can be directly use for our second application: detecting genetic association.

5.1 Introduction

Under the BVSR prior a natural summary of the evidence for a SNP being associated with phenotype is the "posterior inclusion probability" (PIP), $\Pr(\beta_j \neq 0|\mathbf{y}, X)$. Similarly, we define the PIP of SNP j based on summary data

$$\text{SPIP}(j) := \Pr(\beta_j \neq 0|\widehat{\beta}, \widehat{S}, \widehat{R}). \tag{5.1}$$

Here we estimate SPIP(j) by the proportion of MCMC draws for which $\beta_j \neq 0$. [We also provide a Rao-Blackwellised estimate (Casella and Robert, 1996; Guan and Stephens, 2011) in Supplementary Appendix B of Zhu and Stephens (2017a).]

In some cases it is useful to assess genetic associations at the level of *regions* rather than at the level of individual SNPs. Here we define "Expected Number of included SNPs" (ENS) in a locus L to capture region-level association:

$$\text{ENS}(L) := \text{E}\left(\#\left\{j \in L:\ \beta_j \neq 0\right\}|\widehat{\beta}, \widehat{S}, \widehat{R}\right) = \sum_{j \in L} \text{SPIP}(j). \tag{5.2}$$

5.2 Simulations

Previous studies using individual-level data have shown that multiple-SNP model can have higher power to detect genetic associations than single-SNP analyses [e.g. Servin and

Stephens (2007); Hoggart et al. (2008); Guan and Stephens (2011); Peltola et al. (2012); Moser et al. (2015)]. Here we compare the power of multiple-SNP analyses based on summary data with those based on individual-level data. Specifically, we focus on comparing RSS-BVSR with GEMMA-BVSR, because the BVSR-based methods naturally select the associated SNPs (whereas BSLMM assumes that all SNPs are associated).

To compare associations detected by RSS-BVSR and GEMMA-BVSR, we use the simulated data under Scenario 2.1 in Chapter 4. With BVSR analyses, associations are most robustly assessed at the level of *regions* rather than at the level of individual SNPs (Guan and Stephens, 2011), so we compare the association signals from the two methods in sliding 200-kb windows (sliding each window 100kb at a time). Specifically, for each 200-kb region, and each method, we sum the PIPs of SNPs in the region to obtain the corresponding ENS, which summarizes the strength of association in that region. Results (Figure 5.1) show a strong correlation between the ENS values from the summary and individual data, across different numbers of causal variants and PVE values. Consequently, the summary data analyses have similar power to detect associations as the full data analyses (Figure 5.2). Similar to PVE estimation (Chapter 4), the agreement of RSS-BVSR with GEMMA-BVSR is highest when underlying PVE is diluted among many SNPs (e.g. $T = 1000$).

5.3 Example: human height (Wood et al., 2014)

Here we applied RSS-BVSR to the summary data of adult height to detect multiple-SNP associations, and compared results with previous analyses of these summary data. Using a stepwise selection strategy proposed by Yang et al. (2012), Wood et al. (2014) reported a total of 697 genome-wide significant SNPs (GWAS hits). Among them, 531 SNPs were within the $\pm$40-kb regions with estimated ENS ≥ 1. Since only 384 GWAS hits were included in our filtered set of SNPs, we expected a higher replication rate for these included GWAS hits. Taking a region of $\pm$40-kb around each of these 384 SNPs, our analysis identified almost all of these regions (371/384) as showing strong signal for association (estimated ENS ≥ 1).

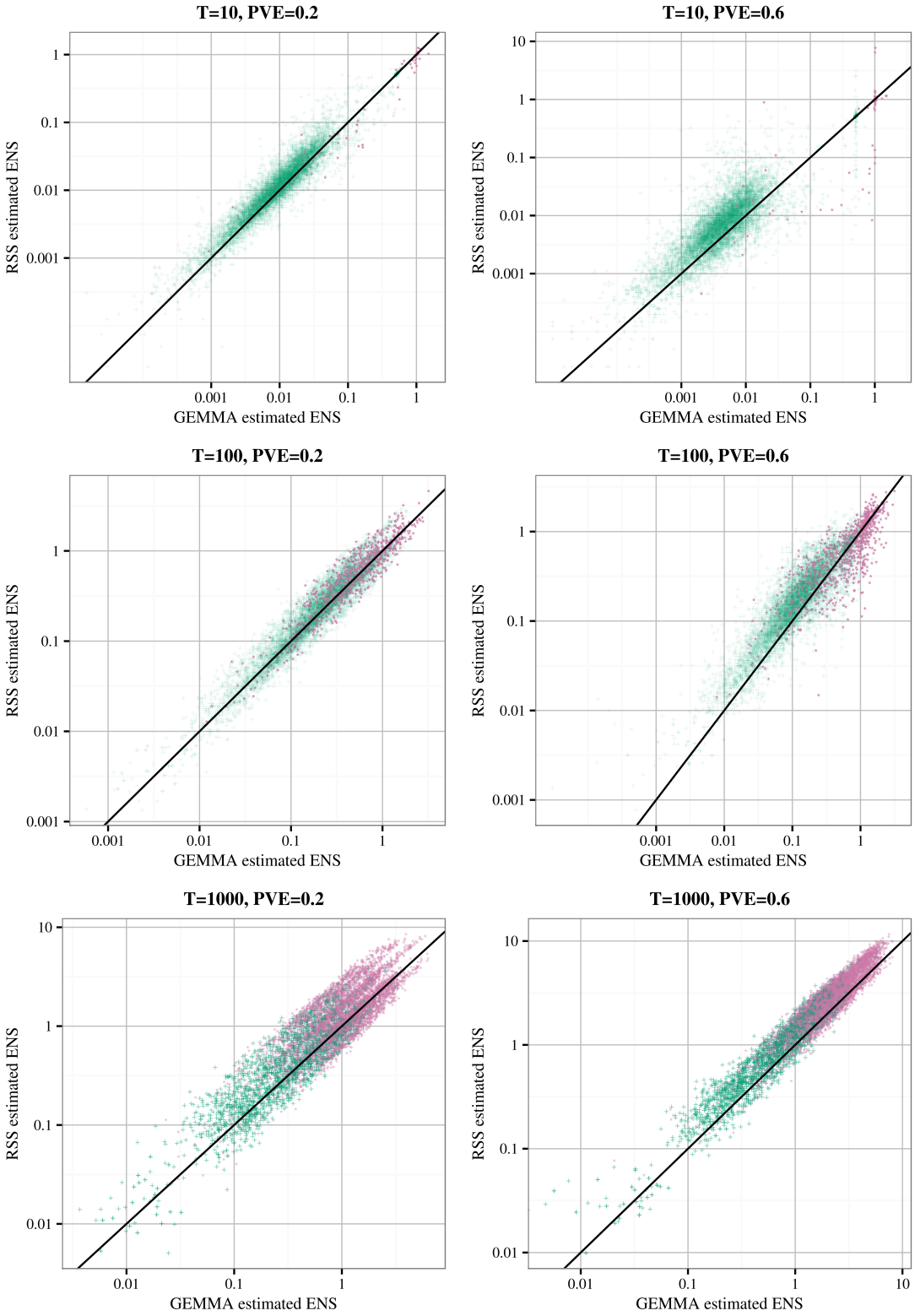


Figure 5.1: Comparison of the 200-kb region posterior expected numbers of included SNPs (ENS) for GEMMA-BVSR (x-axis) and RSS-BVSR (y-axis), based on the simulation study of Scenario 2.1 in Chapter 4. Each point is a 200-kb genomic region, colored according to whether it contains at least one causal SNP (reddish purple "*") or not (bluish green "+").

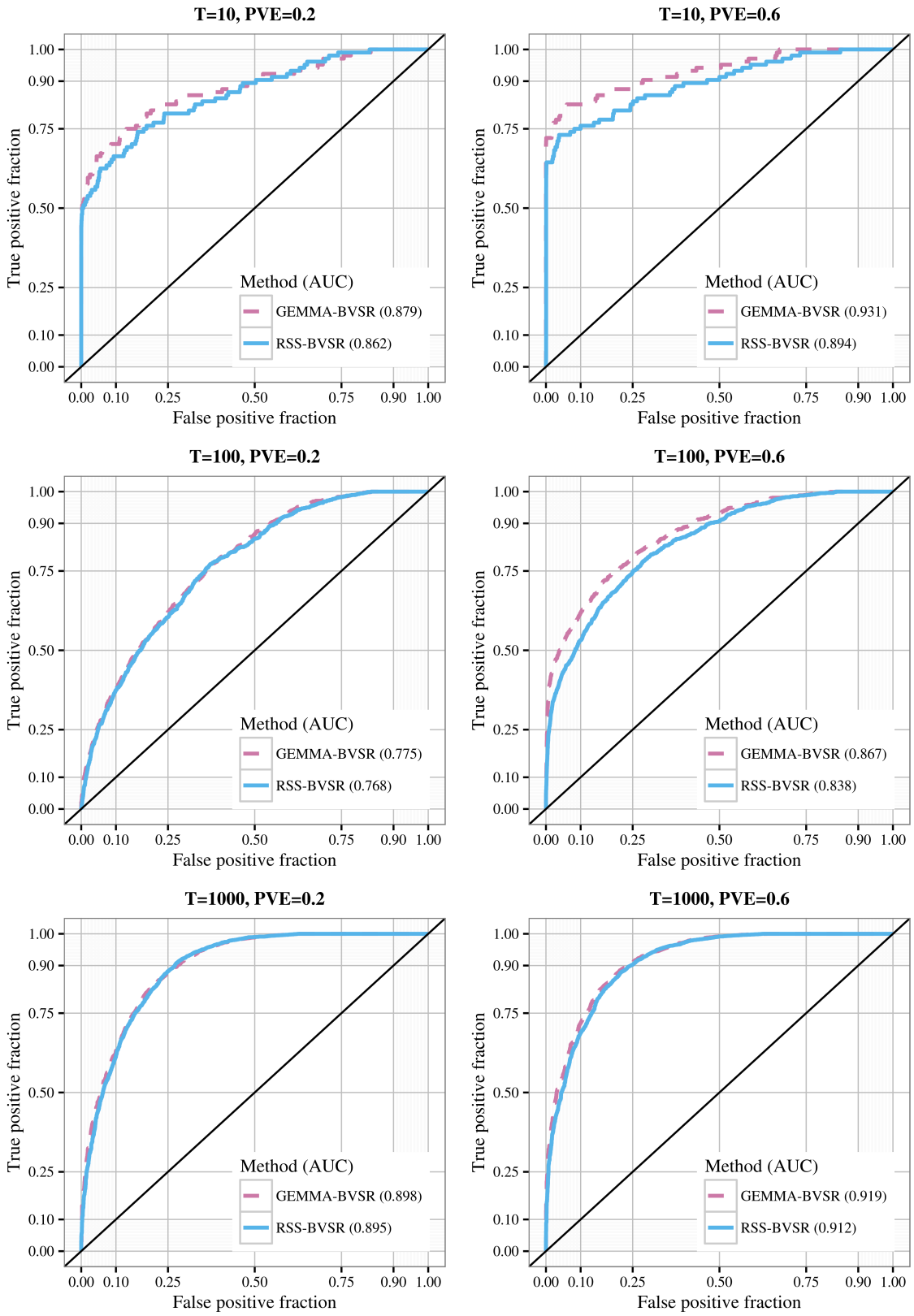


Figure 5.2: Trade-off between true and false positives for GEMMA-BVSR (dash) and RSS-BVSR (solid) in simulations of Scenario 2.1 in Chapter 4.

Only 125 of the 384 SNPs showed, individually, strong evidence for inclusion (estimated SPIP > 0.9). This suggests that, perhaps unsurprisingly, many of the reported associations are likely driven by a SNP in LD with the one identified in the original analysis.

To assess the potential for RSS to identify novel putative loci associated with human height, we estimated the ENS for ±40-kb windows across the whole genome. We identified 5,194 regions with ENS ≥ 1, of which 2,138 are putatively novel in that they are not near any of the previous 697 GWAS hits (distance > 1 Mb). Some of these 2,138 regions are overlapping, but this nonetheless represents a large number of potential novel associations for further investigation. We manually examined the putatively novel regions with highest ENS, and identified several loci harboring genes that seem plausibly related to height. These include the gene *SCUBE1*, which is critical in promoting bone morphogenetic protein signaling (Liao et al., 2016), the gene *WWOX*, which is linked to skeletal system morphogenesis (Del Mare et al., 2011; Aqeilan et al., 2008), the gene *IRX5*, which is essential for proximal and anterior skeletal formation (Li et al., 2014), and the gene *ALX1* (a.k.a. *CART1*), which is involved in bone development (Iioka et al., 2003). See Supplementary Table 5 for the full list of putatively new loci (ENS > 3).

Finally, to check for misspecification we performed the LOO residual-based diagnostic. Specifically, we ran the LOO residual imputation using the RSS-BVSR output, and then refitted the models on the filtered SNPs (absolute LOO z-score ≤ 2). Association results were more robust to SNP filtering than PVE estimation (Chapter 4). Among the ±40-kb regions around the previous GWAS hits, our reanalysis identified 532 of the 697 total hits, and 373 of the 384 included hits. Moving the ±40-kb window across the genome, we identified 6,426 regions with ENS ≥ 1, of which 2,798 were at least 1 Mb away from the 697 GWAS hits. Results are similar based on a less stringent threshold (3); see Supplementary Table 4 of Zhu and Stephens (2017a).

CHAPTER 6

ASSESS GENE SET ENRICHMENT USING GWAS SUMMARY DATA

Our last application of RSS in this dissertation is gene set enrichment analysis of GWAS summary statistics. This chapter is largely based on a manuscript entitled "A large-scale genome-wide enrichment analysis identifies new trait-associated genes, pathways and tissues across 31 human phenotypes" (Zhu and Stephens, 2017b).

6.1 Introduction

Genome-wide association studies (GWAS) have successfully identified many genetic variants – typically SNPs – underlying a wide range of complex traits (Price et al., 2015; Visscher et al., 2012; McCarthy et al., 2008). 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 (Sham and Purcell, 2014). 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 analysis" (Wang et al., 2010) or "gene set analysis" (de Leeuw et al., 2016) – 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 (Subramanian et al., 2005) – 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 (Wang et al., 2010). 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 (Carbonetto and Stephens, 2013) 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 (Zhu and Stephens, 2017a) 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 search-able and browse-able results are available at `http://xiangzhu.github.io/rss-gsea/results`.

6.2 Method overview

Figure 6.1 provides a schematic overview of the method. In brief, the method combines the enrichment model from (Carbonetto and Stephens, 2013), with the multiple regression model for single-SNP association summary statistics from (Zhu and Stephens, 2017a), 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 Detailed 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 (Carbonetto and Stephens,

2012), banded matrix approximation (Wen and Stephens, 2010) and an expectation maximization algorithm accelerator (Varadhan and Roland, 2008) [Detailed methods; Supplementary Note of Zhu and Stephens (2017b)]. Software implementing the method is available at `https://github.com/stephenslab/rss`.

6.3 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 and Supplementary Figure 1 of Zhu and Stephens (2017b)]. This step essentially estimates, for each trait, a "baseline model" against which enrichment hypotheses 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 phenotype [Figure 6.2; Supplementary Figure 2 of Zhu and Stephens (2017b)].

The results emphasize that genetic architecture varies considerably among phenotypes: estimates of both polygenicity and effect sizes vary by several orders of magnitude (Figure 6.2). 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 Figures 3-5 of Zhu

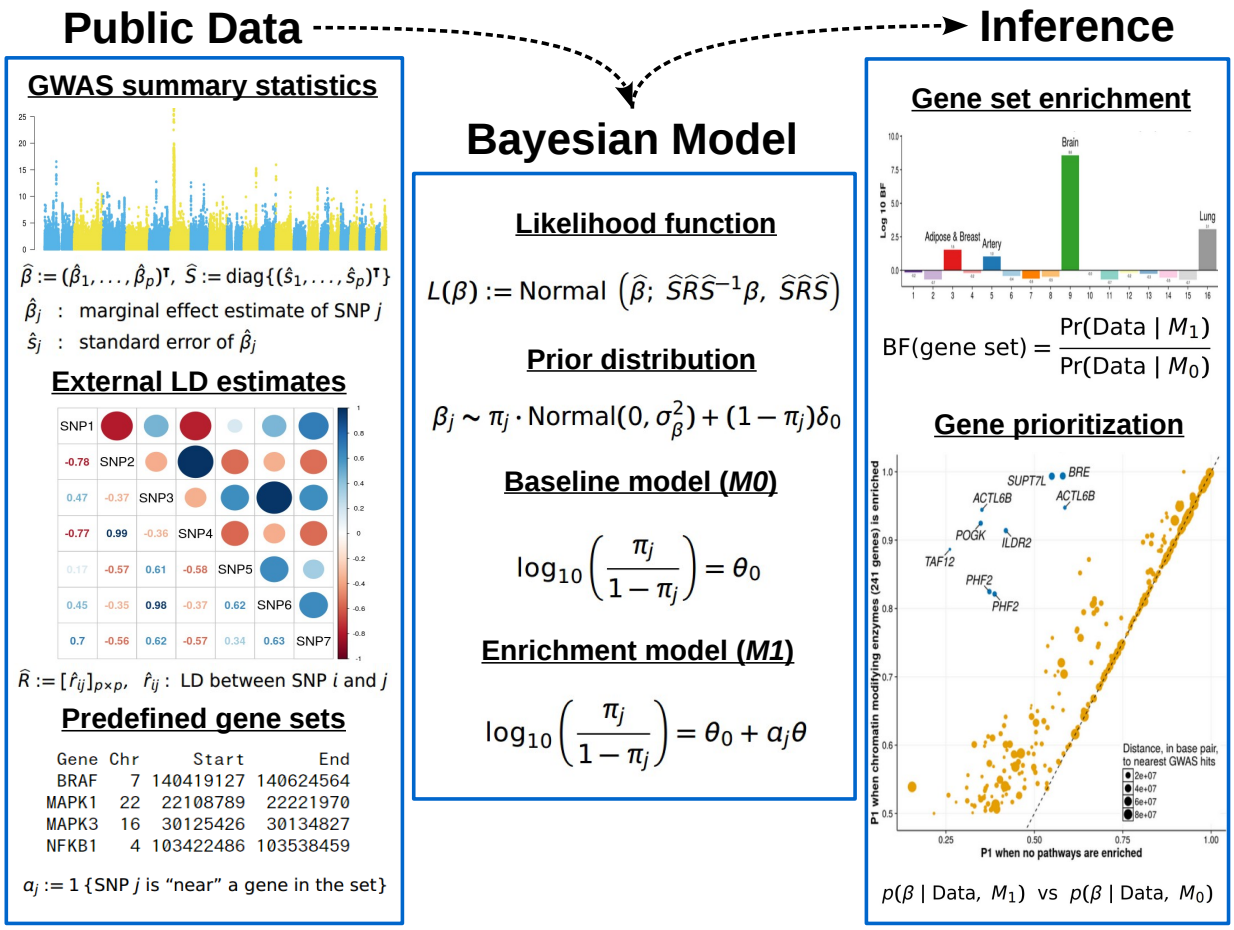


Figure 6.1: Schematic overview of model-based enrichment analysis method for GWAS summary statistics. The method combines three types of data: GWAS summary statistics, external LD estimates, and predefined SNP sets, which here we derive from gene sets based on biological pathways or other sources (Detailed 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.

and Stephens (2017b)]. For several traits, these results also identify additional putative associations [Supplementary Figures 6-7 of Zhu and Stephens (2017b)]. 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.

6.4 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 of Zhu and Stephens (2017b). 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 [Figure 6.4a; Supplementary Figure 10 of Zhu and Stephens (2017b)]. We also filtered out enrichments that were most 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 Figure 11 of Zhu and Stephens (2017b)].

For most traits our analyses identify many pathways with strong evidence for enrich-

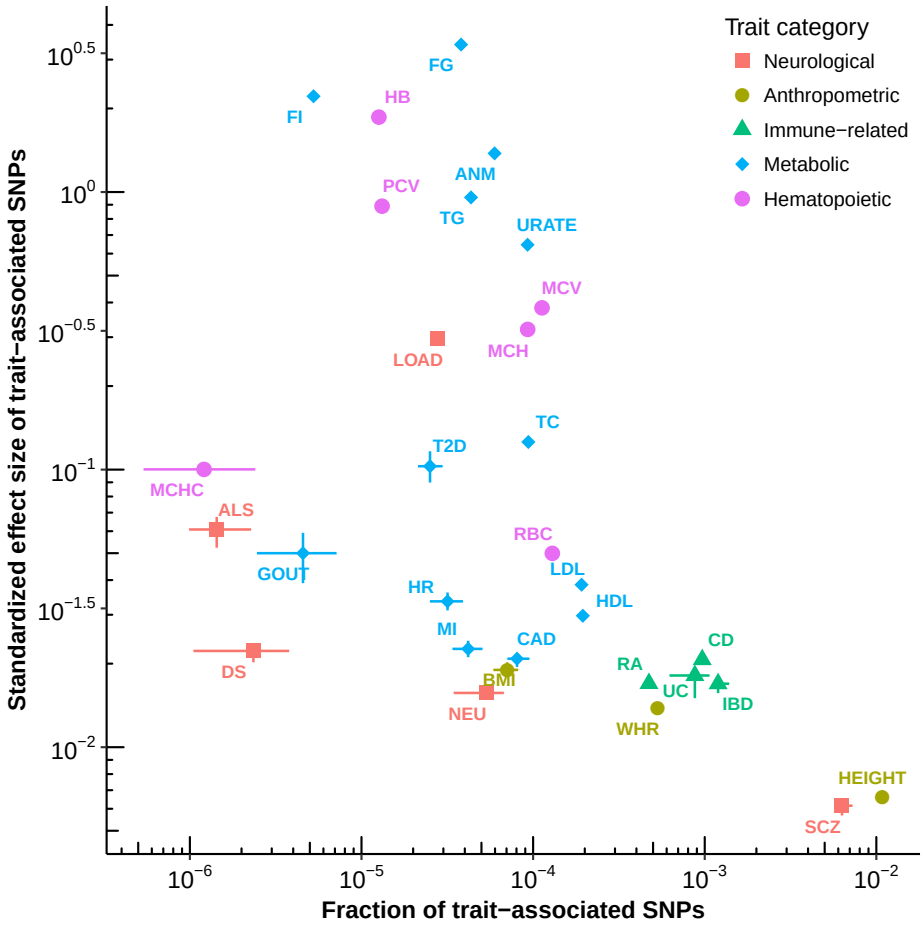


Figure 6.2: Summary of inferred genetic architecture of 31 phenotypes. Results are from fitting the baseline model ($\theta = 0$) to 1.1 million common SNPs for each trait. We summarize genetic architecture using two numbers: the estimated fraction of trait-associated SNPs (a measure of "polygenicity"; x-axis) and the standardized effect size of trait-associated SNPs (y-axis). See Supplementary Note of Zhu and Stephens (2017b) 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 (van Rheenen et al., 2016). DS: depressive symptoms (Okbay et al., 2016). LOAD: late-onset Alzheimer's disease (Lambert et al., 2013). NEU: neuroticism (Okbay et al., 2016). SCZ: schizophrenia (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014). BMI: body mass index (Locke et al., 2015). HEIGHT: adult height (Wood et al., 2014). WHR: waist-to-hip ratio (Shungin et al., 2015). CD: Crohn's disease (Liu et al., 2015). IBD: inflammatory bowel disease (Liu et al., 2015). RA: rheumatoid arthritis (Okada et al., 2014). UC: ulcerative colitis (Liu et al., 2015). ANM: age at natural menopause (Day et al., 2015). CAD: coronary artery disease (Nikpay et al., 2015). FG: fasting glucose (Manning et al., 2012). FI: fasting insulin (Manning et al., 2012). GOUT: Gout (Köttgen et al., 2013). HDL: high-density lipoprotein (Teslovich et al., 2010). HR: heart rate (Den Hoed et al., 2013). LDL: low-density lipoprotein (Teslovich et al., 2010). MI: myocardial infarction (Nikpay et al., 2015). T2D: type 2 diabetes (Morris et al., 2012). TC: total cholesterol (Teslovich et al., 2010). TG: triglycerides (Teslovich et al., 2010). URATE: serum urate (Köttgen et al., 2013). HB: haemoglobin (van der Harst et al., 2012). MCH: mean cell HB (van der Harst et al., 2012). MCHC: MCH concentration (van der Harst et al., 2012). MCV: mean cell volume (van der Harst et al., 2012). PCV: packed cell volume (van der Harst et al., 2012). RBC: red blood cell count (van der Harst et al., 2012).

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 6.1: 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 Figure 10 of Zhu and Stephens (2017b)]. All BFs reported here are larger than the corresponding BFs that SNPs near a gene are enriched [Supplementary Figure 11 of Zhu and Stephens (2017b)]. 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 (Cerami et al., 2011). BS: NCBI BioSystems (Geer et al., 2010). *a*: The full pathway name is "transport of glucose and other sugars, bile salts and organic acids, metal ions and amine compounds".

ment – for example, at a conservative threshold of BF $\geq 10^8$, 20 traits are enriched in more than 100 pathways per trait [Supplementary Figure 12 of Zhu and Stephens (2017b)]. 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 Figure 13 of Zhu and Stephens (2017b)]. Table 6.1 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 (Varjosalo and Taipale, 2008) and previous analyses (Wood et al., 2014). Other examples include *interleukin-23 mediated signaling* with inflammatory bowel disease [BF=3.1×10^{23}; Teng et al. (2015)], *T helper cell surface molecules* with rheumatoid arthritis [BF=3.2×10^8; Okada et al. (2014)], *statin pathway* with levels of high-density lipoprotein cholesterol [BF=8.4×10^{113}; Nicholls et al. (2007)], and *glucose transporters* with serum urate [BF=$1.2 \times 10^{1,558}$; Köttgen et al. (2013)].

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 (Macian, 2005) and bone pathology (Sitara and Aliprantis, 2010). Other examples of novel pathway enrichments include *endochondral ossification* with adult height [BF=7.7×10^{68}; Mackie et al. (2008)], *p75 neurotrophin receptor-mediated signaling* with coronary artery disease [BF=9.6×10^{15}; Elshaer and El-Remessy (2017)], and *osteoblast signaling* with gout [BF=3.8×10^{30}; McQueen et al. (2012)].

6.5 Overlapping pathway enrichment profiles

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 (Figure 6.3; Detailed 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 (Di Paolo and Kim, 2011), vascular disorder (Beeri et al., 2006) and immune activation (Heppner et al., 2015). 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 (Pickrell et al., 2016) and co-heritability analyses (Bulik-Sullivan et al., 2015a).

6.6 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

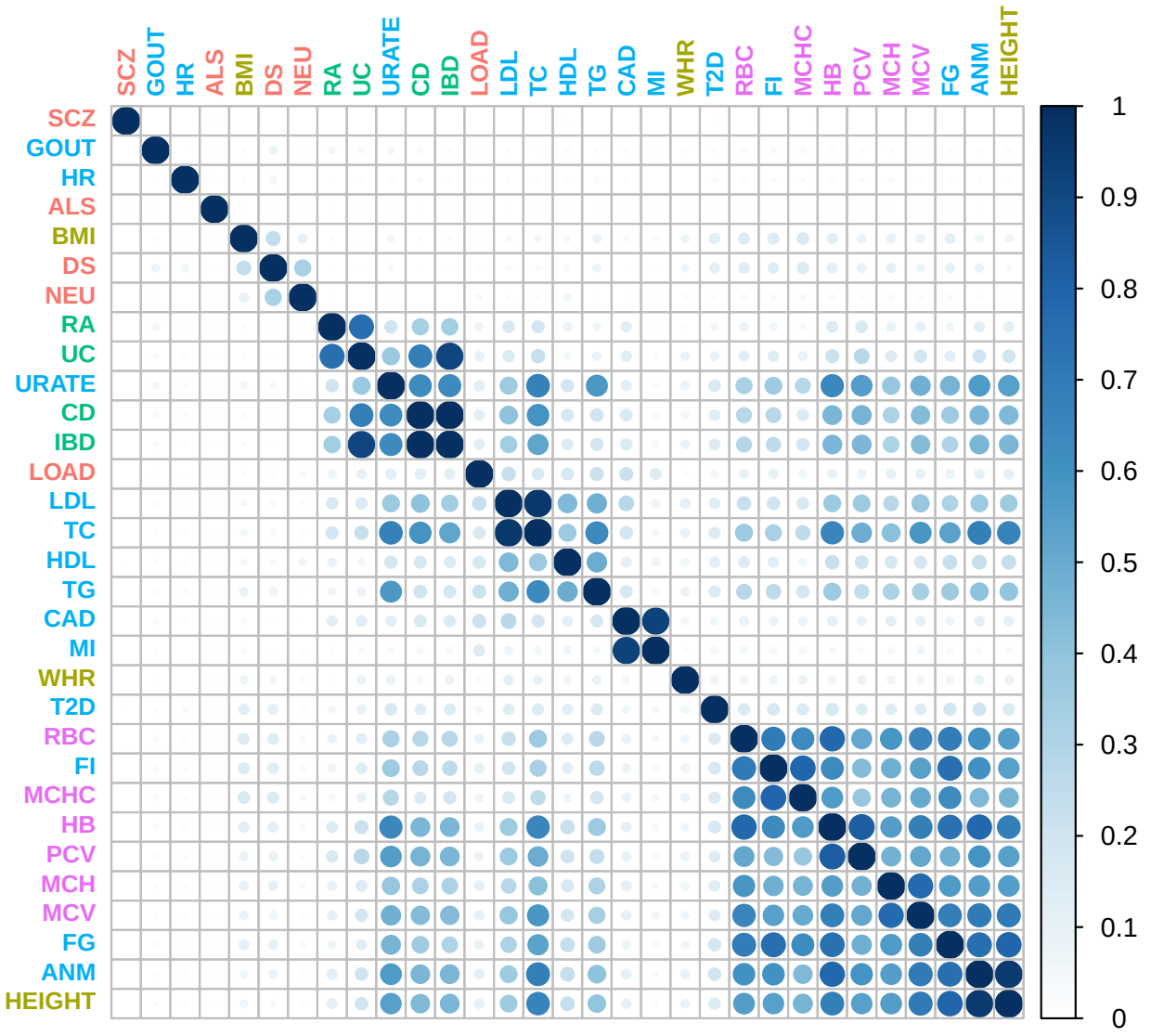


Figure 6.3: Pairwise sharing of pathway enrichments among 31 traits. For each pair of traits, we estimated the proportion of pathways that are enriched in both traits, among pathways enriched in at least one of the traits (Detailed methods). Darker color and larger shape represent higher sharing. Traits are colored by categories and labeled by abbreviations (Figure 6.2), and clustered by hierarchical clustering as implemented in `R` package `corrplot`.

that may explain individual variant associations, which is itself an important and challenging problem (Smemo et al., 2014). 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 (Guan and Stephens, 2011; Carbonetto and Stephens, 2013), 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 Figure 14 of Zhu and Stephens (2017b)], 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}; Figure 6.4a) informs a strong association between gene *MTTP* (baseline P_1: 0.14; enriched P_1: 0.99) and levels of low-density lipoprotein (LDL) cholesterol (Figure 6.4b). 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 (Rader and Kastelein, 2014). However, no genome-wide significant SNPs near *MTTP* were reported in single-SNP analyses of either the same data [Figure 6.4c; Teslovich et al. (2010)], or more recent data with larger sample size [Figure 6.4d); Global Lipids Genetics Consortium (2013)].

Prioritization analysis of this same *chylomicron-mediated lipid transport* pathway also

yields several additional plausible associations (Figure 6.4b). 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 (Teslovich et al., 2010) although not for LDL cholesterol [Supplementary Figure 15 of Zhu and Stephens (2017b)]; and a multiple-trait, single-SNP analysis (Stephens, 2013) 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 2.

Box 2 Select putatively novel associations from prioritization analyses

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

- *HDAC4* (baseline P_1: 0.98; enriched P_1: 1.00)

  HDAC4 encodes a critical regulator of chondrocyte hypertrophy during skeletogenesis (Vega et al., 2004) and osteoclast differentiation (Obri et al., 2014). 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 (Cheloha et al., 2015). 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 (Su et al., 2014). Mutations in *FGFR1* cause several skeletal disorders (OMIM: 101600, 123150, 190440, 166250).

- *MMP13* (baseline P_1: 0.45; enriched P_1: 0.93)

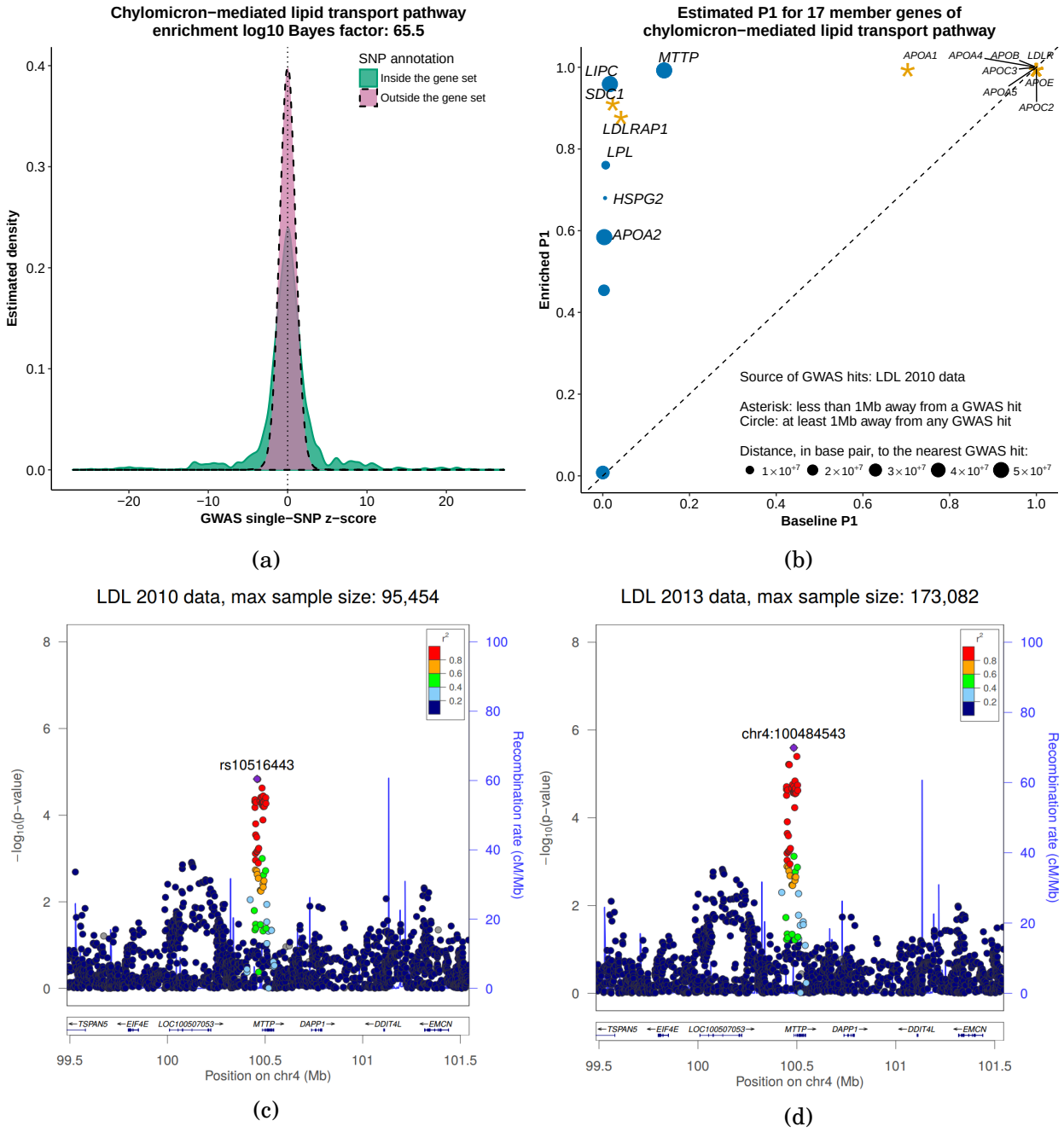


Figure 6.4: Enrichment of *chylomicron-mediated lipid transport* pathway informs a strong association between a member gene *MTTP* and levels of low-density lipoprotein (LDL) cholesterol. **(a)** Distribution of GWAS single-SNP z-scores from summary data published in 2010 (Teslovich et al., 2010), stratified 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 (Teslovich et al., 2010). **(d)** Regional association plot for *MTTP* based on summary data published in 2013 (Global Lipids Genetics Consortium, 2013).

MMP13 encodes a protein that is required for osteocytic perilacunar remodeling and bone quality maintenance (Tang et al., 2012). Mutations in *MMP13* cause a type of metaphyseal anadysplasia (OMIM: 602111) with reduced stature.

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

- *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 (Steinberg et al., 2008). *TNFRSF14* is near a GWAS hit of celiac disease [rs3748816, $p = 3.3 \times 10^{-9}$), Dubois et al. (2010)] and two hits of ulcerative colitis [rs734999, $p = 3.3 \times 10^{-9}$ Anderson et al. (2011); rs10797432, $p = 3.0 \times 10^{-12}$ Jostins et al. (2012)].

- *FAS* (baseline P_1: 0.82; enriched P_1: 0.99)

  FAS plays many important roles in the immune system (Strasser et al., 2009). 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 (Neurath, 2014).

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

- *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 (Susan-Resiga et al., 2011). *FU-*

RIN is near a GWAS hit [rs2521501 International Consortium for Blood Pressure Genome-Wide Association Studies (2011)] 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 (Silence et al., 2001).

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

- *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 (Aseem et al., 2014). 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, $p = 3.0 \times 10^{-11}$ Global Lipids Genetics Consortium (2013)].

- *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 (Kennedy et al., 2005).

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

- *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 (Anderson et al., 1996) and clinical trials (Kivitz et al., 2002). *PTGS2* is near a GWAS hit of Crohn's disease [rs10798069, $p = 4.3 \times 10^{-9}$, Liu et al. (2015)]

- *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 (Daynes and Jones, 2002; Széles et al., 2007).

6.7 Enrichment analysis of 113 tissue-related gene sets

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 (The GTEx Consortium, 2015) to define sets of the most "relevant" genes in each tissue, based on expression patterns across tissues (Detailed 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 (Boyle et al., 2017) ? 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 (Detailed 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 (Dey et al., 2017). 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 6.2: 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 Figure 10 of Zhu and Stephens (2017b)]. HE: highly expressed. SE: selectively expressed. DE: distinctively expressed. a: Multiple tissues show partial membership in Cluster 1 (Dey et al., 2017). b: These three BFs are smaller than the corresponding BFs that SNPs near a gene are enriched [Supplementary Figure 11 of Zhu and Stephens (2017b)].

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 6.2), 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 6.2) 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 (BF = 6.6×10^7) and *brain*-related (BF = 3.5×10^8) 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 (BF>10^{30}): liver, brain and adrenal gland. Because of the well-known connection between gene *APOE* and LOAD (Liu et al., 2013), and the fact that *APOE* is highly expressed in these three tissues [Supplementary Figure 16 of Zhu and Stephens (2017b)], 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 (Figure 6.5), 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 prioritiza-

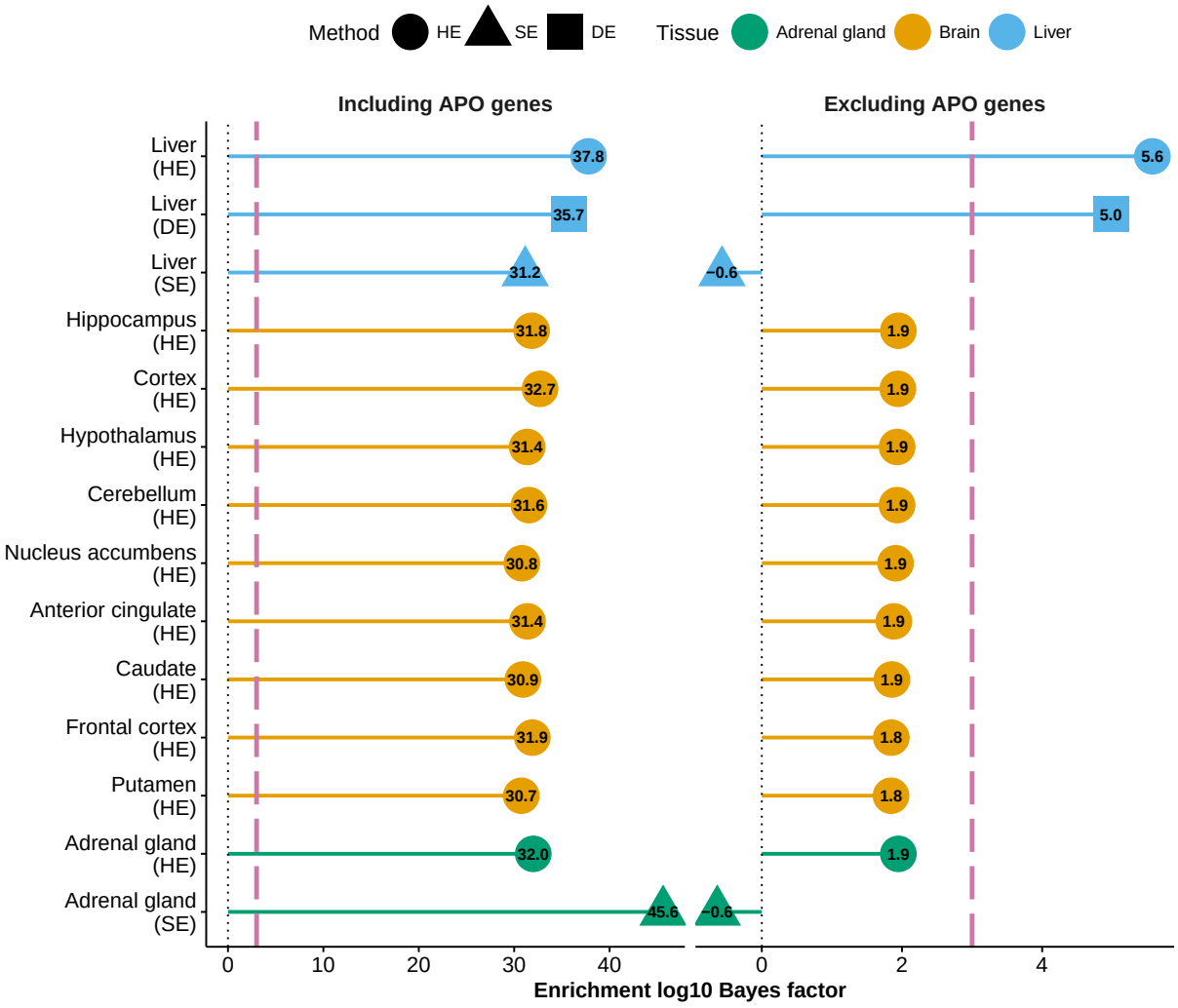


Figure 6.5: Enrichment analyses of genes related to liver, brain and adrenal gland for Alzheimer's disease. Shown are the tissue-based gene sets with the strongest enrichment signals for Alzheimer's disease. Each gene set was analyzed twice: the left panel corresponds to the analysis based on the original 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.

tion analysis for non-APO liver-related genes (Figure 6.5). This highlighted an association of LOAD with gene *TTR* [baseline P_1: 0.64; enriched P_1: 1.00; Supplementary Figure 17 of Zhu and Stephens (2017b)]. *TTR* encodes transthyretin, which has been shown to inhibit LOAD-related protein from forming harmful aggregation and toxicity, *in vitro* (Schwarzman et al., 1994) and *in vivo* (Buxbaum et al., 2008). Indeed, transthyretin is considered a biomarker for LOAD: patients show reduced transthyretin levels in plasma (Velayudhan et al., 2012) and cerebrospinal fluid (Hansson et al., 2009). And rare variants in *TTR* have recently been found to be associated with LOAD (Sassi et al., 2016; Xiang et al., 2016). By integrating GWAS with expression data our analysis identifies association of *TTR* based on common variants.

6.8 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" (Wang et al., 2010; de Leeuw et al., 2016), have several advantages over alternative approaches – sometimes known as "self-contained tests" [e.g. Kwak and Pan (2016); Zhang et al. (2016)] – 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 of Zhu and Stephens (2017b).]

Compared with other enrichment approaches, our method has several particularly at-

tractive features. First, unlike many methods [e.g. Wang et al. (2010); Slowikowski et al. (2014); Pers et al. (2015)] 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. Segrè et al. (2010); de Leeuw et al. (2015); Lamparter et al. (2016)] 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 (Smemo et al., 2014). 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 (Carbonetto and Stephens, 2013; Evangelou et al., 2014) 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. Schadt et al. (2005); Nicolae et al. (2010); Nica et al. (2010); He et al. (2013); Giambartolomei et al. (2014); Zhu et al. (2016); Hormozdiari et al. (2016); Wen et al. (2017)]. 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. Gamazon et al. (2015); Gusev et al. (2016)], 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 Figure 18 of Zhu and Stephens (2017b)]. 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 Figure 13 of Zhu and Stephens (2017b)]; and b) for some traits (e.g. height), genetic associations may be strongly enriched near all genes [Supplementary Figure 11 of Zhu and Stephens (2017b)], 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. Logsdon et al. (2010); Stegle et al. (2010); Carbonetto and Stephens (2012); Li and Sillanpää (2012); Carbonetto and Stephens (2013); Papastamoulis et al. (2014); Raj et al. (2014); Logsdon et al. (2014); Loh et al. (2015); Gopalan et al. (2016); Montesinos-López et al. (2017)] 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 (Detailed 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. Pickrell (2014); Finucane et al. (2015); Li and Kellis (2016)] could further increase power to detect novel associations, and help distinguish between competing correlated annotations (e.g. overlapping pathways) when explaining observed enrichments.

6.9 Detailed methods

GWAS summary statistics, LD estimates and SNP annotations

We analyze GWAS summary statistics of 31 phenotypes [Supplementary Note and Supplementary Table 1 of Zhu and Stephens (2017b)], 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 (International HapMap 3 Consortium, 2010), since LD among these SNPs can be reliably estimated from existing panels (Bulik-Sullivan et al., 2015b). 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 Figure 1 of Zhu and Stephens (2017b)].

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 (1000 Genomes Project Consortium, 2015) to estimate LD (Wen and Stephens, 2010).

To create SNP-level annotations for a given gene set, we use a distance-based approach (Segrè et al., 2010; Carbonetto and Stephens, 2013). 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, Cerami et al. (2011)], NCBI Biosystems (Geer et al., 2010), PANTHER [version 3.3, Mi and Thomas (2009)] and BioCarta [used in Carbonetto and Stephens (2013)]. 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 Carbonetto and Stephens (2013) 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 of Zhu and Stephens (2017b).

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 (The GTEx Consortium, 2015) 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 (Dey et al., 2017). 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 [Chapter 2, or Zhu

and Stephens (2017a)]:

$$L_{\text{rss}}(\beta) := \mathcal{N}(\widehat{\beta}; \widehat{S}\widehat{R}\widehat{S}^{-1}\beta, \widehat{S}\widehat{R}\widehat{S}) \tag{6.1}$$

where $\widehat{\beta} := (\hat{\beta}_1, \ldots, \hat{\beta}_p)^\intercal$, $\widehat{S} := \text{diag}(\widehat{\mathbf{s}})$, $\widehat{\mathbf{s}} := (\hat{s}_1, \ldots, \hat{s}_p)^\intercal$, $\widehat{R}$ is the LD matrix estimated from an external reference panel with ancestry matching the GWAS cohort, $\beta := (\beta_1, \ldots, \beta_p)^\intercal$ 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 Carbonetto and Stephens (2013) and Guan and Stephens (2011), to specify the following prior on β:

$$\begin{aligned}
\beta_j &\sim \pi_j \mathcal{N}(0, \sigma_\beta^2) + (1 - \pi_j)\delta_0, && (6.2)\\
\sigma_\beta^2 &= h \cdot \left(\textstyle\sum_{j=1}^{p} \pi_j n^{-1} \hat{s}_j^{-2} \right)^{-1}, && (6.3)\\
\pi_j &= (1 + 10^{-(\theta_0 + a_j \theta)})^{-1}, && (6.4)
\end{aligned}$$

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 of Zhu and Stephens (2017b)].

Posterior computation

We combine the likelihood function and prior distribution above to perform Bayesian inference. The posterior computation procedures largely follow those developed in Carbonetto and Stephens (2012). Firstly, for each set of hyper-parameters $\{\theta_0, \theta, h\}$ from a predefined grid, we approximate the (conditional) posterior of β 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 β over the posterior of $\{\theta_0, \theta, h\}$ to obtain the posterior of β.

To facilitate large-scale analyses, we further employ several computational tricks. First, we use squared iterative methods (Varadhan and Roland, 2008) to accelerate the fixed point iterations in the variational Bayes approximation step. Second, we exploit the banded LD matrix (Wen and Stephens, 2010) to parallelize the algorithm. Third, we use a simplification introduced in Carbonetto and Stephens (2013) that scales the enrichment analysis to thousands of gene sets by reusing expensive genome-wide calculations. See Supplementary Note of Zhu and Stephens (2017b) 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 (Carbonetto and Stephens, 2012), 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 Figure 19 of Zhu and Stephens (2017b)], 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 β and θ_0 [Supplementary Note of Zhu and Stephens (2017b)], and then use the solution from the modified algorithm to initialize future variational approximations. We test this initialization strategy on triglyceride (Teslovich et al., 2010), and obtain improved posterior results

[Supplementary Figure 20 of Zhu and Stephens (2017b)].

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):

$$\text{BF} := \frac{p(\widehat{\beta}|\widehat{S},\widehat{R},\mathbf{a},\theta>0)}{p(\widehat{\beta}|\widehat{S},\widehat{R},\mathbf{a},\theta=0)}, \tag{6.5}$$

where $\mathbf{a} := (a_1,\ldots,a_p)^\top$ 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 of Zhu and Stephens (2017b) 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 β. The first statistic is P_1, the posterior probability that at least 1 SNP in the locus is associated with the phenotype:

$$P_1 := 1 - \Pr(\beta_j = 0,\ \forall j \in \text{locus}|\mathbf{D}), \tag{6.6}$$

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:

$$\text{ENS} := \textstyle\sum_{j\in\text{locus}} \Pr(\beta_j \neq 0|\mathbf{D}). \tag{6.7}$$

See Supplementary Note of Zhu and Stephens (2017b) for details of computing P_1 and ENS.

Estimate pairwise sharing of pathway enrichments

To capture the "sharing" of pathway enrichments between two traits (Figure 6.3), we define a parameter $\pi := (\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11})$ as follows:

$$\pi_{ab} := \Pr(z_{1j} = a, z_{2j} = b),\ a \in \{0,1\},\ b \in \{0,1\}, \tag{6.8}$$

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

$$\hat{\pi} := \arg\max_{\pi} \prod_j (\pi_{00} + \pi_{01}\mathrm{BF}_{2j} + \pi_{10}\mathrm{BF}_{1j} + \pi_{11}\mathrm{BF}_{1j}\mathrm{BF}_{2j}), \tag{6.9}$$

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` (Stephens, 2017). 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 6.3, 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 method (Carbonetto and Stephens, 2013). 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 (Zhu and Stephens, 2017a). These two likelihoods are mathematically equivalent under certain conditions (Zhu and Stephens, 2017a). 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 of Zhu and Stephens (2017b) 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 Figures 21-22 of Zhu and Stephens (2017b)].

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` package, `http://ggplot2.tidyverse.org`;
- `corrplot` package, `https://cran.r-project.org/web/packages/corrplot`;

CHAPTER 7

CONCLUDING REMARKS

We have presented a novel Bayesian approach to inferring multiple linear regression coefficients using simple linear regression summary statistics, and demonstrated its application in GWAS. On both simulated and real data our method produces results comparable to methods based on individual-level data. Compared with existing summary-based methods, our approach takes advantage of an explicit likelihood for the multiple regression coefficients (Chapter 2), and thus provides a unified framework for various genome-wide analyses (Chapters 4-6). We also theoretically extend this framework to capture certain features of GWAS summary data, and provide practical suggestions when the theoretical extensions cannot be easily implemented (Chapter 3).

In this chapter, we conclude with some summary key points and future research directions in methodology (Section 7.1), computation (Section 7.2) and application (Section 7.3).

7.1 Methodology of modeling summary statistics

The RSS likelihood is closely related to "pivot" or "pivotal quantity", a function of data and parameters whose probability distribution does not have unknown parameters. In particular, the RSS likelihood (2.2) can be viewed as the probability density of the following asymptotic pivot:

$$\widehat{R}^{-\frac{1}{2}}\widehat{S}^{-1}(\widehat{\beta}-\widehat{S}\widehat{R}\widehat{S}^{-1}\beta) \dot{\sim} \mathcal{N}(\mathbf{0}, I_p). \tag{7.1}$$

The idea of obtaining likelihood from pivotal quantities can be traced back at least to Fisher's fiducial inference (Fisher, 1930). Kalbfleisch and Sprott (1970) suggest the use of pivot density as in effect a likelihood function, and Cox (1993) derives the conditions under which this approach would yield unbiased estimates. Sprott (1990) studies in detail the likelihood induced from linear pivot; see also Chamberlin (1989). Schweder and Hjort (2002) show how to obtain a single scalar parameter likelihood reduced of all nuisance parameters

from exact or approximate pivots, and discuss its connection with confidence distributions (Efron, 1998). Hoff and Wakefield (2013) suggest using the asymptotic "sandwich" distribution of a pivotal quantity to construct a likelihood, and illustrate how Bayesian methods based on this likelihood improved performance over the standard non-Bayesian "sandwich" procedure (Huber, 1967); see also Chapter 4 of Harmon (2015) for a detailed discussion of pivot-based likelihood in Bayesian analysis. Despite these advances, theoretical justifications for pivot-based likelihood are still very limitedly available, or even do not exist. From a practical perspective, however, our empirical results show that pivot may be an useful tool for likelihood construction. Hence, our applied work raises the following questions in theoretical statistics: under which conditions does "pivot likelihood" produce "valid" inferences, and how can it be connected with other approaches to statistical modeling (Liu and Meng, 2016)?

Although it was not our focus here, RSS methods may be easily modified to analyze summary data of discrete phenotypes, such as disease status in a typical case-control study. For example, one can directly apply RSS methods to the estimated log odds ratios and standard errors generated from univariate logistic regression analyses, and then interpret the resulting joint effect size estimates on a log odds ratio scale. A more principled approach would be to derive a variant of RSS likelihood based on generalized linear models (McCullagh and Nelder, 1989), using asymptotic theories of maximum likelihood estimator or, more generally, the M-estimator (van der Vaart, 1998).

In this work, our prior specification for β largely follows previous Bayesian "spike-and-slab" approaches (Mitchell and Beauchamp, 1988; George and McCulloch, 1993, 1997; Ishwaran and Rao, 2005). Specifically, we use either a point-normal mixture distribution [e.g. Guan and Stephens (2011); Carbonetto and Stephens (2013)] or a two-normal mixture distribution [e.g. Zhou et al. (2013)] in our applications. In the meantime, there are many other effect size distributions for β emerging from recently-developed penalized likelihood and Bayesian approaches to regression analysis of individual-level data; see Hesterberg et al.

(2008); O'Hara and Sillanpää (2009); Tibshirani (2011); Mallick and Yi (2013); Bühlmann et al. (2014) for literature reviews. These include g-prior (Zellner, 1986), mixture of g-prior (Liang et al., 2008), Laplace prior [a.k.a. Bayesian Lasso, Park and Casella (2008)], horseshoe prior (Carvalho et al., 2010), Ising prior (Li and Zhang, 2010), non-local prior (Johnson and Rossell, 2010, 2012), and mixture of normal prior (Erbe et al., 2012; Moser et al., 2015; Stephens, 2017). It is definitely possible that the spike-and-slab approaches might not fully capture the actual effect size distribution in some cases, and thus one might consider alternative distributional assumptions. As we have shown in the spike-and-slab case, incorporating an existing prior of β into RSS, at least in principle, is not hard. Moreover, combining these prior distributions with the RSS likelihood provides a unified approach to extending these Bayesian methods to summary data analysis.

There is also a large body of literature describing innovative methods that greatly improve inference of regression analysis based on individual-level data. These include methods for selective inference (Taylor and Tibshirani, 2015), false discovery rate control [e.g. Barber and Candès (2015); Brzyski et al. (2017)], and confounding adjustment [e.g. Leek and Storey (2007); Sun et al. (2012); Risso et al. (2014)]. It may be useful to embed these modern techniques into a regression-based framework for summary data like RSS.

7.2 Computation for increasingly large datasets

Large-scale statistical analyses often has a non-trivial computation component. As described in Chapter 6, we apply RSS methods to an integrated analysis of 31 human traits, 4,026 biological annotations and 1.1 million genetic variants. Given a trait, each annotation has a corresponding parameter of interest. Given a trait-annotation pair, each SNP also has a specific model parameter. Hence, in this example, the total number of parameters is $31 \times 4,026 \times 1.1$ million ≈ 137 billion. Solving problems of this size requires efficient, scalable algorithms and implementations.

We use relatively straightforward algorithms and implementations in this work, which

seem to be able to handle most of the large datasets to date. Further computational innovation, however, may be required to apply RSS methods to the very large datasets that are currently being generated [e.g. UK Biobank (Sudlow et al., 2015), China Kadoorie Biobank (Du et al., 2017)]. One possibility to improve computation is to use more sophisticated and potentially more efficient algorithms, including hybrid Monte Carlo [a.k.a. Hamiltonian Monte Carlo; Duane et al. (1987); Neal (2011)], evolutionary Monte Carlo [e.g. Liang and Wong (2000); Wilson et al. (2010); Bottolo and Richardson (2010)], expectation propagation (Minka, 2001), integrated nested Laplace approximations (Rue et al., 2009), stochastic variational inference [e.g. Hoffman et al. (2013); Gopalan et al. (2016)], etc. Another important extension is to consider emerging technologies for "Big Data" analytics in industry such as Hadoop MapReduce and Apache Spark^{TM}.

7.3 Application in human complex trait genetics

Recently, there seems to be a gap in human complex trait genetics: massive amounts, multiple types of genomic (a.k.a. "multi-omics") data are generated, but deep understanding of many complex traits is still very limited. This gap, presumably, at least in part, is because most complex trait analyses to date are still performed on a single type of data. As an attempt to bridge this gap, many quantitative methods are recently developed for integrated analysis of multiple types of data, or, "data integration" (Ritchie et al., 2015; Yang et al., 2016; Richardson et al., 2016; Hasin et al., 2017).

Data integration is conceptually easy to implement in the RSS framework (Chapter 6). This is mainly because of its "modularity" feature embedded in Bayesian hierarchical modeling. Specifically, each component of the hierarchical framework (e.g. likelihood, prior, hyper-prior) models one type of data, and data integration occurs automatically in the end by Bayes' Theorem. Further, if there is an update in one type of data (e.g. better technology, larger sample), one can simply modify the corresponding model component while keep other components the same.

Here our work has facilitated incorporating GWAS summary statistics, or, more generally, genotype-phenotype correlations into the integrated framework, by developing the RSS likelihood (2.2). Future work thus focuses on developing new prior distributions to incorporate different types of biological information.

Finally, it is important to note that the RSS likelihood is based on generic linear models (Chapter 2) and thus is not restricted to GWAS summary statistics of "observable" traits. The RSS likelihood (2.2) can be readily applied to univariate summary results from large-scale genetic association studies of many quantitative molecular and cellular traits, including gene expression levels [i.e. eQTL; Nica and Dermitzakis (2013); Sun and Hu (2013); Albert and Kruglyak (2015)], methylation patterns [i.e. meQTL; Gutierrez-Arcelus et al. (2013); Banovich et al. (2014)] metabolites [i.e. mQTL; Suhre and Gieger (2012); Shin et al. (2014)] and proteins [i.e. pQTL; Stark et al. (2014); Portelli et al. (2014)].

APPENDIX A

LINKS TO SUPPLEMENTARY MATERIALS

Below are links to supplementary materials of Zhu and Stephens (2017a) and Zhu and Stephens (2017b) that are referenced in this dissertation. Copies of these supplementary documents are also included here as appendices.

- **Supplementary Note, Figures and Tables of Zhu and Stephens (2017a)**
  http://www.biorxiv.org/content/biorxiv/suppl/2016/12/01/042457.DC2/042457-1.pdf

- **Supplementary Note of Zhu and Stephens (2017b)**
  http://www.biorxiv.org/content/biorxiv/suppl/2017/07/08/160770.DC1/160770-2.pdf

- **Supplementary Figures of Zhu and Stephens (2017b)**
  http://www.biorxiv.org/content/biorxiv/suppl/2017/07/08/160770.DC1/160770-1.pdf

- **Supplementary Tables of Zhu and Stephens (2017b)**
  http://www.biorxiv.org/content/biorxiv/suppl/2017/07/08/160770.DC1/160770-3.pdf

APPENDIX B

SUPPLEMENTARY NOTE OF ZHU AND STEPHENS (2017A)

This appendix is largely based on the "Appendix B: Details of posterior sampling scheme" for Zhu and Stephens (2017a). Here we describe the Markov chain Monte Carlo (MCMC) algorithms in terms of $\{S,R\}$, and then replace the unknown $\{S,R\}$ with their estimates $\{\widehat{S},\widehat{R}\}$ in practice. This is similar to the likelihood derivation and prior specification in main text of Zhu and Stephens (2017a).

B.1 Rank-based strategy

When locally updating the SNP-specific parameters (e.g. genetic effect β_j and inclusion indicator $\gamma_j := \mathbf{1}\{\beta_j \neq 0\}$ for each SNP j) in the MCMC algorithms, we allocate more computational resources to SNPs with larger marginal association signals, using the rank-based strategy (Guan and Stephens, 2011). In particular, we first rank all the variants based on the single-SNP p-values and draw one SNP to update according to some probability distributions with decreasing probability. Currently, we use a mixture distribution $q_p = 0.3u_p + 0.7g_p$, where u_p is a discrete uniform distribution and g_p is a geometric distribution truncated to $1,\ldots,p$ with its parameter chosen to give a mean of 2000.

Based on q_p, we introduce $Q(\cdot|\gamma)$, a rank-based proposal for the indicator vector $\gamma := (\gamma_1,\ldots,\gamma_p)^\intercal$. To propose a new value γ^* given the current value γ, we start by setting $\gamma^* = \gamma$ and then randomly choose one of the following:

1. With probability P_a, draw SNP r according to q_p until $\gamma_r = 0$ and set $\gamma_r^* = 1$.

2. With probability P_r, draw SNP r uniformly from $\{j : \gamma_j = 1\}$ and set $\gamma_r^* = 0$.

3. With probability P_e, sample two SNPs by the above two steps and switch their indicators.

The default setting in our software is $P_a = P_r = 0.4, P_e = 0.2$.

B.2 BVSR prior

For RSS with BVSR prior, we use Metropolis-Hastings (MH) algorithm to obtain posterior samples of (γ, π, h) on the product space of $\{0,1\}^p \times (0,1) \times (0,1)$,

$$p(\gamma, \pi, h | \widehat{\beta}, S, R) \propto p(\widehat{\beta} | S, R, \gamma, \pi, h) p(\gamma|\pi) p(\pi) p(h). \tag{B.1}$$

Here we exploit the fact that β can be integrated out analytically to compute $p(\widehat{\beta}|S,R,\gamma,\pi,h)$:

$$\widehat{\beta} | S, R, \gamma, \pi, h \sim \mathcal{N}(\mathbf{0}, SRS + \sigma_B^2 M_\gamma M_\gamma^\intercal), \tag{B.2}$$

where $M := SRS^{-1}$ and M_γ denotes the sub-matrix of M restricted to those columns j for which $\gamma_j = 1$. We update γ using the rank-based proposal $Q(\cdot|\gamma)$. We update $\log\pi$ by adding a random number from $\mathcal{U}(-0.05, 0.05)$ to the current value, and update h by adding a random number from $\mathcal{U}(-0.1, 0.1)$ to the current value. New values of $\log\pi$ and h outside boundaries are reflected back.

After drawing a posterior sample of (γ, π, h), we then sample β according to its conditional distribution given (γ, π, h) and $(\widehat{\beta}, S, R)$:

$$\beta_\gamma | \widehat{\beta}, S, R, \gamma, \pi, h \sim \mathcal{N}(\mu, \Omega^{-1}), \tag{B.3}$$

$$\beta_{-\gamma} | \widehat{\beta}, S, R, \gamma, \pi, h \sim \delta_0, \tag{B.4}$$

where β_γ and $\beta_{-\gamma}$ denote the subsets of β corresponding to the entries that $\gamma_j = 1$ and 0 respectively, δ_0 denotes the point mass at zero and,

$$\Omega := M_\gamma^\intercal (SRS)^{-1} M_\gamma + \sigma_B^{-2}(\gamma, \pi, h) I_{|\gamma|}, \tag{B.5}$$

$$\mu := \Omega^{-1} M_\gamma^\intercal (SRS)^{-1} \widehat{\beta}. \tag{B.6}$$

The marginal likelihood (B.2), up to some constant, can be written in terms of $(\Omega, \boldsymbol{\mu})$,

$$p(\widehat{\boldsymbol{\beta}}|S,R,\boldsymbol{\gamma},\pi,h) \propto \sigma_B^{-|\boldsymbol{\gamma}|}|\Omega|^{-1/2}\exp\{\boldsymbol{\mu}^\top \mathbf{q}_{\boldsymbol{\gamma}}/2\}, \tag{B.7}$$

where $\mathbf{q}_{\boldsymbol{\gamma}}$ denotes the subset of $\mathbf{q} := S^{-1}\boldsymbol{\beta}$ corresponding to the entries that $\gamma_j = 1$. The matrix computation in a single step of the MCMC algorithm above involves one Cholesky decomposition of Ω and three triangular linear systems. Hence, the computational cost for each iteration of MCMC is $\mathcal{O}(|\boldsymbol{\gamma}|^3 + 3|\boldsymbol{\gamma}|^2)$, where $|\boldsymbol{\gamma}|$ denotes the number of non-zero entries in $\boldsymbol{\gamma}$.

To improve precision, we can use Rao-Blackwellized estimates (Casella and Robert, 1996; Guan and Stephens, 2011). For SPIP, we have

$$\Pr(\gamma_j = 1|\widehat{\boldsymbol{\beta}},S,R) = \mathrm{E}(\Pr(\gamma_j = 1|\widehat{\boldsymbol{\beta}},S,R,\boldsymbol{\xi}_{-j})) \approx M^{-1}\textstyle\sum_{i=1}^{M}\Pr(\gamma_j = 1|\widehat{\boldsymbol{\beta}},S,R,\boldsymbol{\xi}_{-j}^{(i)})$$

where $\boldsymbol{\xi}_{-j}$ stands for $\{\boldsymbol{\beta}_{-j},\boldsymbol{\gamma}_{-j},\pi,h\}$, $\boldsymbol{\gamma}_{-j}$ and $\boldsymbol{\beta}_{-j}$ denote the vectors $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ excluding the jth coordinate and $\boldsymbol{\xi}_{-j}^{(i)}$ denotes the ith MCMC sample from the posterior distribution of $\boldsymbol{\xi}_{-j}$. For the posterior mean of the multiple-SNP effect at SNP j, we have

$$\mathrm{E}(\beta_j|\widehat{\boldsymbol{\beta}},S,R) = \mathrm{E}(\mathrm{E}(\beta_j|\widehat{\boldsymbol{\beta}},S,R,\boldsymbol{\xi}_{-j})) \approx M^{-1}\textstyle\sum_{i=1}^{M}\mathrm{E}(\beta_j|\widehat{\boldsymbol{\beta}},S,R,\gamma_j = 1,\boldsymbol{\xi}_{-j}^{(i)})\Pr(\gamma_j = 1|\widehat{\boldsymbol{\beta}},S,R,\boldsymbol{\xi}_{-j}^{(i)}).$$

To obtain the Rao-Blackwellized estimates, we need $p(\gamma_j|\widehat{\boldsymbol{\beta}},S,R,\boldsymbol{\xi}_{-j})$ and $p(\beta_j|\widehat{\boldsymbol{\beta}},S,R,\gamma_j,\boldsymbol{\xi}_{-j})$:

$$\begin{aligned}
\frac{\Pr(\gamma_j = 1|\widehat{\boldsymbol{\beta}},S,R,\boldsymbol{\xi}_{-j})}{\Pr(\gamma_j = 0|\widehat{\boldsymbol{\beta}},S,R,\boldsymbol{\xi}_{-j})} &= \frac{\pi}{1-\pi}\sqrt{\frac{s_j^2}{s_j^2+\sigma_B^2}}\exp\left\{\frac{1}{2(\sigma_B^{-2}+s_j^{-2})}\left(\frac{\hat{\beta}_j}{s_j^2} - \sum_{i\neq j}\frac{r_{ij}\beta_i}{s_i s_j}\right)^2\right\} \\
\beta_j|\widehat{\boldsymbol{\beta}},S,R,\gamma_j = 1,\boldsymbol{\xi}_{-j} &\sim \mathcal{N}\left(\frac{1}{\sigma_B^{-2}+s_j^{-2}}\left(\frac{\hat{\beta}_j}{s_j^2} - \sum_{i\neq j}\frac{r_{ij}\beta_i}{s_i s_j}\right), \frac{1}{\sigma_B^{-2}+s_j^{-2}}\right) \\
\beta_j|\widehat{\boldsymbol{\beta}},S,R,\gamma_j = 0,\boldsymbol{\xi}_{-j} &\sim \delta_0
\end{aligned}$$

where r_{ij} is the (i,j)-th entry of R.

B.3 BSLMM prior

We propose a component-wise MCMC algorithm for RSS with BSLMM prior. First, we re-parameterize the multiple-SNP effect β_j as follows

$$\beta_j|\gamma_j = 1,\pi,h,\rho,S = \sqrt{\sigma_B^2+\sigma_P^2}\cdot\tilde{\beta}_j \tag{B.8}$$

$$\beta_j|\gamma_j = 0,\pi,h,\rho,S = \sigma_P\cdot\tilde{\beta}_j \tag{B.9}$$

where the standardized effect $\tilde{\beta}_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, for $j\in\{1,\ldots,p\}$. Equivalently,

$$\boldsymbol{\beta} = B\widetilde{\boldsymbol{\beta}},\ \ \widetilde{\boldsymbol{\beta}}\sim\mathcal{N}(\mathbf{0},I_p) \tag{B.10}$$

where the scaling matrix B is diagonal with the jth diagonal b_j defined as

$$b_j = \sigma_P\mathbf{1}\{\gamma_j=0\}+\sqrt{\sigma_B^2+\sigma_P^2}\mathbf{1}\{\gamma_j=1\}. \tag{B.11}$$

The new parameterization may help speed up the convergence of MCMC, since $\widetilde{\boldsymbol{\beta}}$ is independent with $(\boldsymbol{\gamma},\pi,h,\rho)$ *a priori*. We then draw posterior samples of $(\tilde{\boldsymbol{\beta}},\boldsymbol{\gamma},\pi,h,\rho)$ iteratively.

- Given $(\widetilde{\boldsymbol{\beta}},\pi,h,\rho)$, we update $\boldsymbol{\gamma}$ by a standard MH algorithm, where the proposal is $Q(\cdot|\boldsymbol{\gamma})$.

- Given $(\boldsymbol{\gamma},\pi,h,\rho)$, we update $\widetilde{\boldsymbol{\beta}}$ by a mixture of global and local moves. With probability P_{g}, we draw a new value of $\widetilde{\boldsymbol{\beta}}$ from its full conditional ("global move"),

$$\widetilde{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}},S,R,\boldsymbol{\gamma},\pi,h,\rho\sim\mathcal{N}((BS^{-1}RS^{-1}B+I)^{-1}BS^{-2}\widehat{\boldsymbol{\beta}},(BS^{-1}RS^{-1}B+I)^{-1}). \tag{B.12}$$

  With probability $1-P_{\text{g}}$, we randomly pick a SNP j according to the distribution q_p

and draw $\tilde{\beta}_j$ from its full conditional ("local move")

$$\tilde{\beta}_j|\widehat{\beta},S,R,\widetilde{\beta}_{-j},\gamma,\pi,h,\rho \sim \mathcal{N}\left(\frac{b_j s_j \ell_j}{s_j^2+b_j^2}, \frac{s_j^2}{s_j^2+b_j^2}\right),\ \ell_j := \frac{\hat{\beta}_j}{s_j} - \sum_{i\neq j}\frac{r_{ij}b_i\tilde{\beta}_i}{s_i}. \tag{B.13}$$

- Given $(\widetilde{\beta},\gamma,h,\rho)$, we update π by a Metropolis algorithm, where the proposal is a symmetric Gaussian random walk on $\log((\pi-p^{-1})/(1-\pi))$.
- Given $(\widetilde{\beta},\gamma,\pi,\rho)$, we update h by a Metropolis algorithm, where the proposal is a symmetric Gaussian random walk on $\log(h/(1-h))$.
- Given $(\widetilde{\beta},\gamma,\pi,h)$, we update ρ by a Metropolis algorithm, where the proposal is a symmetric Gaussian random walk on $\log(\rho/(1-\rho))$.

The most computationally intensive step is drawing $\widetilde{\beta}$ from a p-dimensional multivariate normal distribution (B.12). For each draw, one Cholesky decomposition of $BS^{-1}RS^{-1}B+I$ and two triangular linear systems are required. Since matrix R is banded with some bandwidth w (Wen and Stephens, 2010), the matrix $BS^{-1}RS^{-1}B+I$ also has the same bandwidth and therefore, the per-iteration cost of the algorithm above is at most $\mathcal{O}(pw^2+2p^2)$. For all the simulations, we set $P_{\mathrm{g}}=0.05$[1]. For the analysis of height data, we set $P_{\mathrm{g}}=0.001$ (default in our software).

B.4 Small world proposal

To improve the convergence rate of the MCMC schemes, we use the "small-world" proposal (Guan and Krone, 2007) as an add-on for every Metropolis step in our main algorithms above. Specifically, with probability 0.3 in each iteration, a long-range move is made by compounding a random number (from 2 to 20) of local proposals.

1. The large value of P_{g} in simulations increases the computation time of RSS-BSLMM; see Supplementary Figure 5 of Zhu and Stephens (2017a).

APPENDIX C

SUPPLEMENTARY FIGURES OF ZHU AND STEPHENS (2017A)

Supplementary Figure 1

Comparison of true PVE and Summary PVE (SPVE) given the true β. The true PVE is computed from the true values of $\{\beta, \tau\}$ and the individual-level data $\{X, \mathbf{y}\}$. The SPVE is computed from the true β, the summary-level data $\{\hat{\beta}_j, \hat{\sigma}_j^2\}$ and the estimated LD matrix $\widehat{R}$. The simulated genotypes consist of 10,000 independent SNPs from 1000 individuals, so $\widehat{R}$ is set as identity matrix; The real genotypes are 10,000 correlated SNPs randomly drawn from chromosome 16 (WTCCC UK Blood Service control group, 1458 individuals), and $\widehat{R}$ is estimated from WTCCC 1958 British Birth Cohort (1480 individuals) and HapMap CEU genetic maps using the shrinkage method in Wen and Stephens (2010). Solid dots indicate sample means of 200 replicates; vertical bars indicate symmetric 95% intervals; orange line indicates the reference line with intercept 0 and slope 1. The tables summarize the RMSEs between SPVE and true PVE.

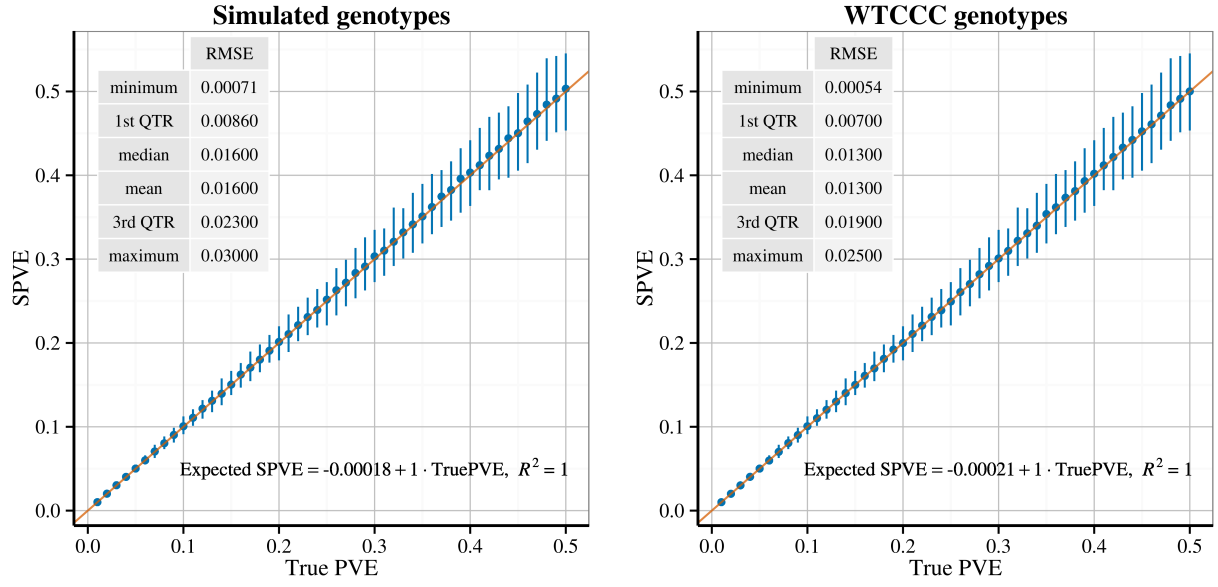
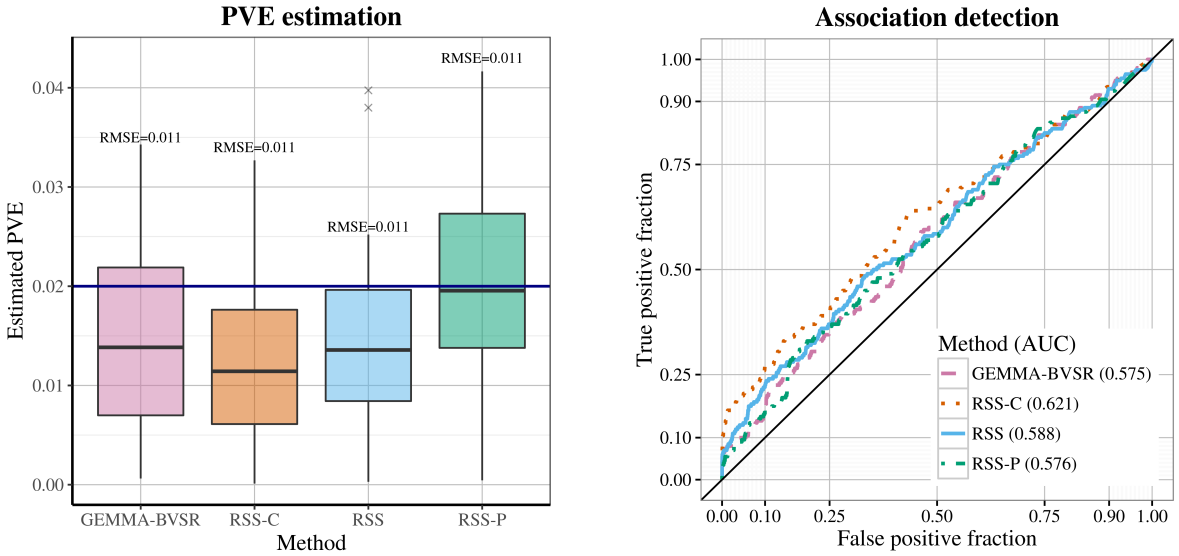


Figure C.1: Comparison of true PVE and Summary PVE (SPVE) given the true β.

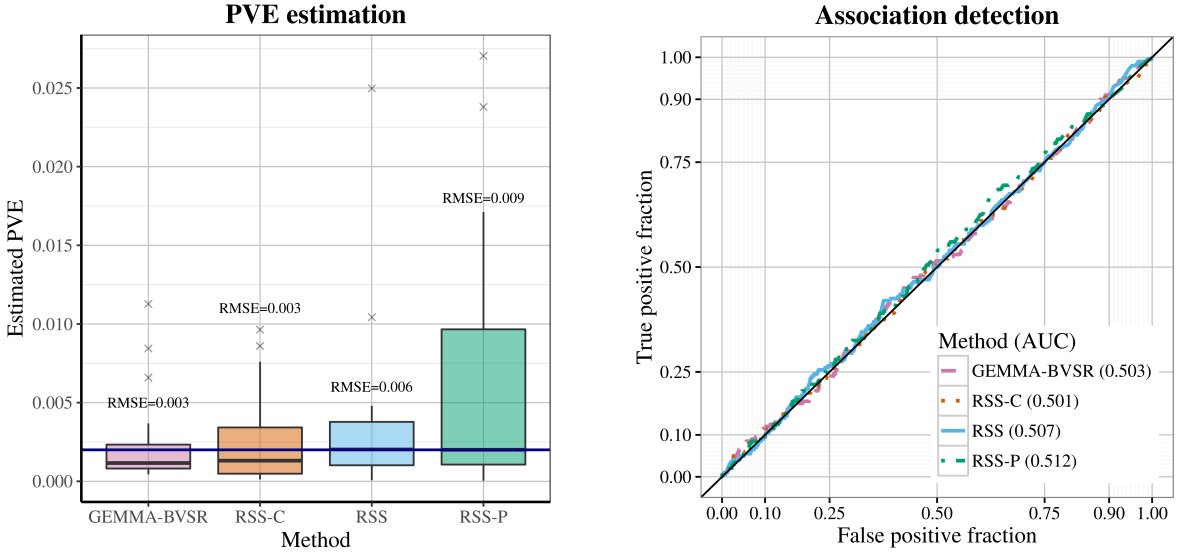
Supplementary Figure 2

Comparison of PVE estimation and association detection on three types of LD matrix: cohort sample LD (RSS-C), shrinkage panel sample LD (RSS) and panel sample LD (RSS-P). The simulation schemes and statistical methods are the same as Figure 1 of Zhu and Stephens (2017a), except that the true PVE is 0.02 and 0.002 respectively.



(a) True PVE: 0.02

(b) True PVE: 0.02



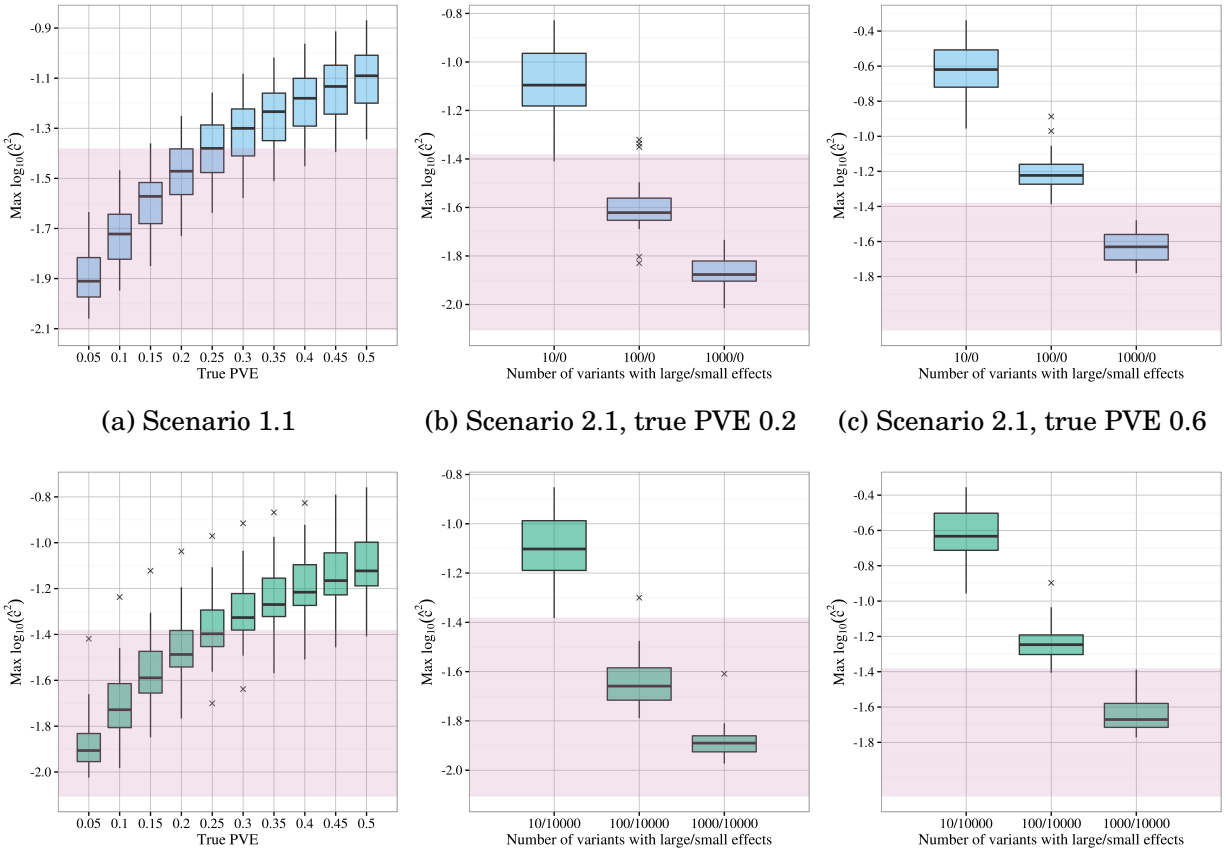
(c) True PVE: 0.002

(d) True PVE: 0.002

Figure C.2: Comparison of PVE estimation and association detection on three types of LD matrix.

Supplementary Figure 3

Distribution of $\max_j \log_{10}(\hat{c}_j^2)$ in all the simulated datasets used in main text of Zhu and Stephens (2017a). For each SNP $j \in [p]$, $\hat{c}_j := (||\mathbf{y}|| \cdot ||X_j||)^{-1}(X_j^\top \mathbf{y})$ is the sample marginal correlation between phenotype ($\mathbf{y}$) and genotype of SNP j (X_j), and it can be computed from the single-SNP summary data, $\hat{c}_j^2 = (n\hat{\sigma}_j^2 + \hat{\beta}_j^2)^{-1}\hat{\beta}_j^2$. The simulations use the real genotypes of 12,758 (p) SNPs on chromosome 16 from 1,458 (n) individuals. The shaded area in the following plots corresponds to the 60%-90% quantile of $\max_j \log_{10}(\hat{c}_j^2)$ across 42 complex traits listed in main text [Table 1 of Zhu and Stephens (2017a)]. This helps us identify which simulations have "realistic" $\max_j \log_{10}(\hat{c}_j^2)$ values that are close to real GWAS datasets.



(a) Scenario 1.1 (b) Scenario 2.1, true PVE 0.2 (c) Scenario 2.1, true PVE 0.6

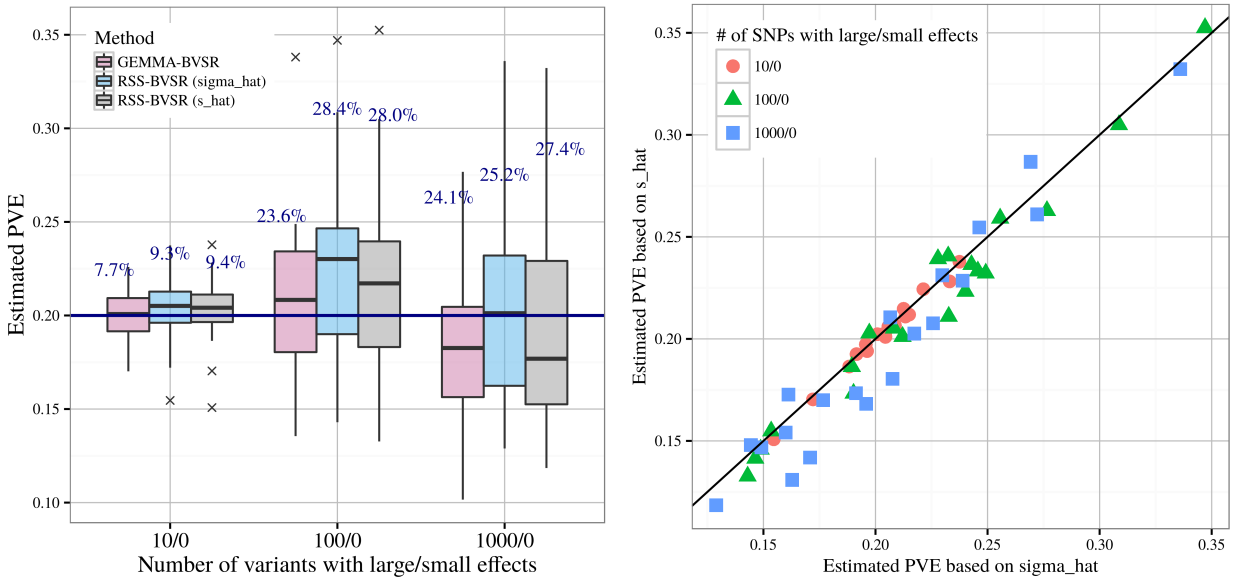
(d) Scenario 1.2 (e) Scenario 2.2, true PVE 0.2 (f) Scenario 2.2, true PVE 0.6

Figure C.3: Distribution of $\max_j \log_{10}(\hat{c}_j^2)$ in all the simulated datasets used in main text of Zhu and Stephens (2017a).

Supplementary Figure 4

Comparison of PVE estimation and association detection based on $\{\hat{\sigma}_j^2\}$ and $\{\hat{s}_j^2\}$ respectively. The RSS-BVSR models are fitted on the Scenario 2.1 simulated datasets in main text of Zhu and Stephens (2017a), with $\widehat{S}$ defined by $\{\hat{\sigma}_j^2\}$ and $\{\hat{s}_j^2\}$ respectively.

(a) Comparison of PVE estimation. Left panel: Relative RMSE for each method is reported (percentages on top of box plots). The true PVE are shown as the solid horizontal line. Each box plot summarizes results from 20 replicates. Right panel: Each point corresponds to one simulated dataset. The reference line has intercept 0 and slope 1.



(b) Comparison of association detection. The associations are evaluated at the 200-kb region level. A region is causal if and only if it contains at least one causal SNP.

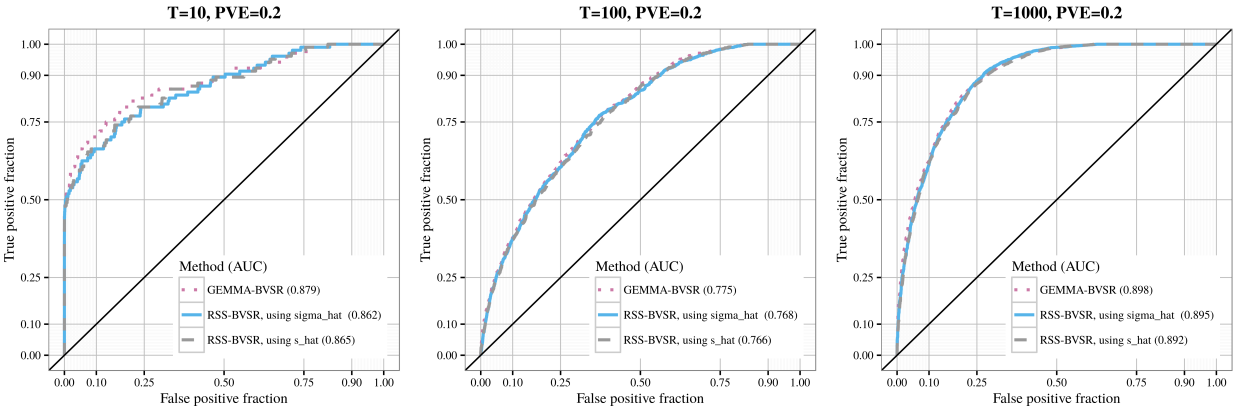
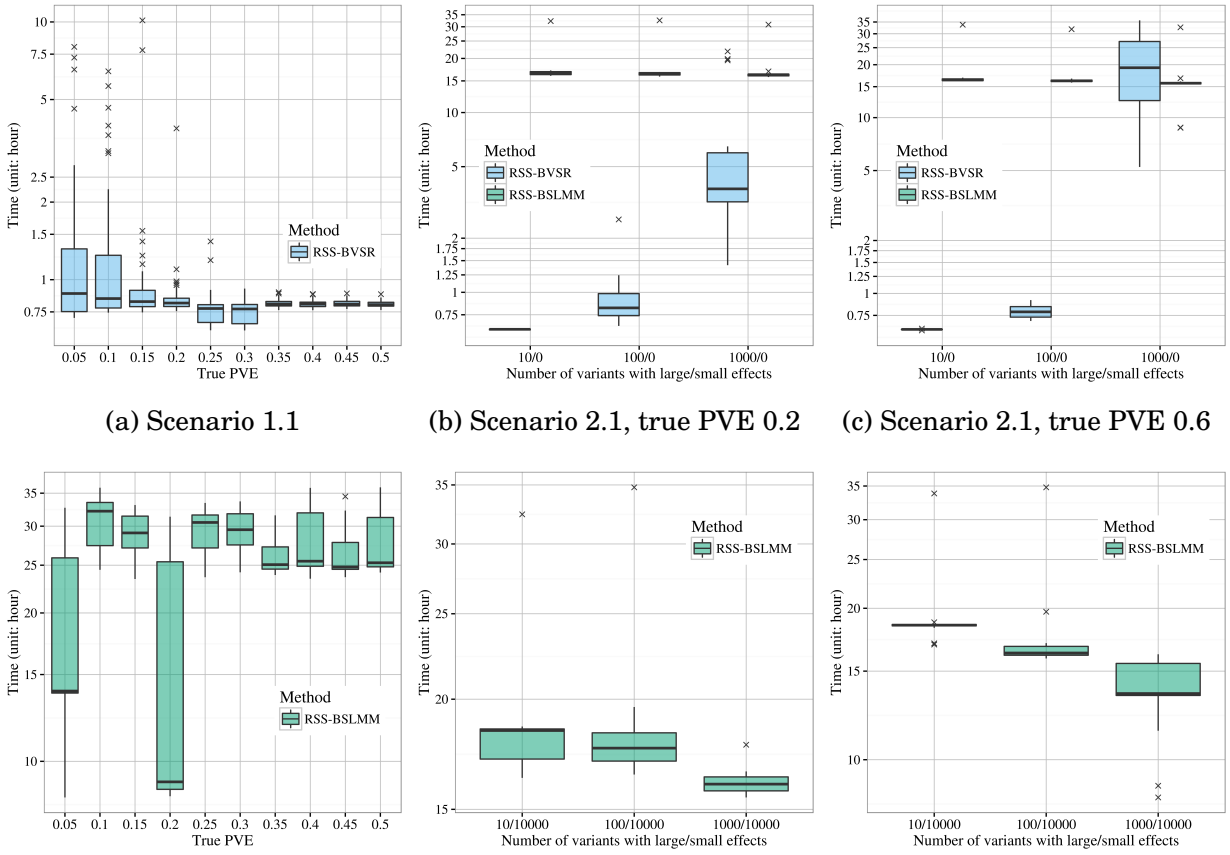


Figure C.4: Comparison of PVE estimation and association detection based on $\{\hat{\sigma}_j^2\}$ and $\{\hat{s}_j^2\}$ respectively.

Supplementary Figure 5

Computation time, in hours, of RSS-BVSR and RSS-BSLMM in the simulation studies in main text of Zhu and Stephens (2017a). For each simulated dataset and method, the computation was performed on a single core of Intel E5-2670 2.6GHz, with 2 million MCMC iterations. There are 50 replicates in Scenario 1.1 and 1.2, and 20 replicates in Scenario 2.1 and 2.2. The computation time of RSS-BSLMM in simulations is longer than height data analyses [Supplementary Table 6 of Zhu and Stephens (2017a)] because a larger P_g was used.



(a) Scenario 1.1 (b) Scenario 2.1, true PVE 0.2 (c) Scenario 2.1, true PVE 0.6

(d) Scenario 1.2 (e) Scenario 2.2, true PVE 0.2 (f) Scenario 2.2, true PVE 0.6

Figure C.5: Computation time, in hours, of RSS-BVSR and RSS-BSLMM in the simulation studies in main text of Zhu and Stephens (2017a).

Supplementary Figure 6

Simulations show that PVE estimation can be biased when RSS methods are applied to summary data that are *not* generated from the same set of individuals. Here the summary data are generated as follows. For each simulated individual-level dataset (Scenario 2.1, true PVE = 0.2 and $T = 1000$), we first randomly draw 10% of SNPs. For each of these SNPs, we randomly draw 50% of individuals and use their data to compute the single-SNP summary statistics. For the remaining SNPs, we compute their summary statistics from all individuals.

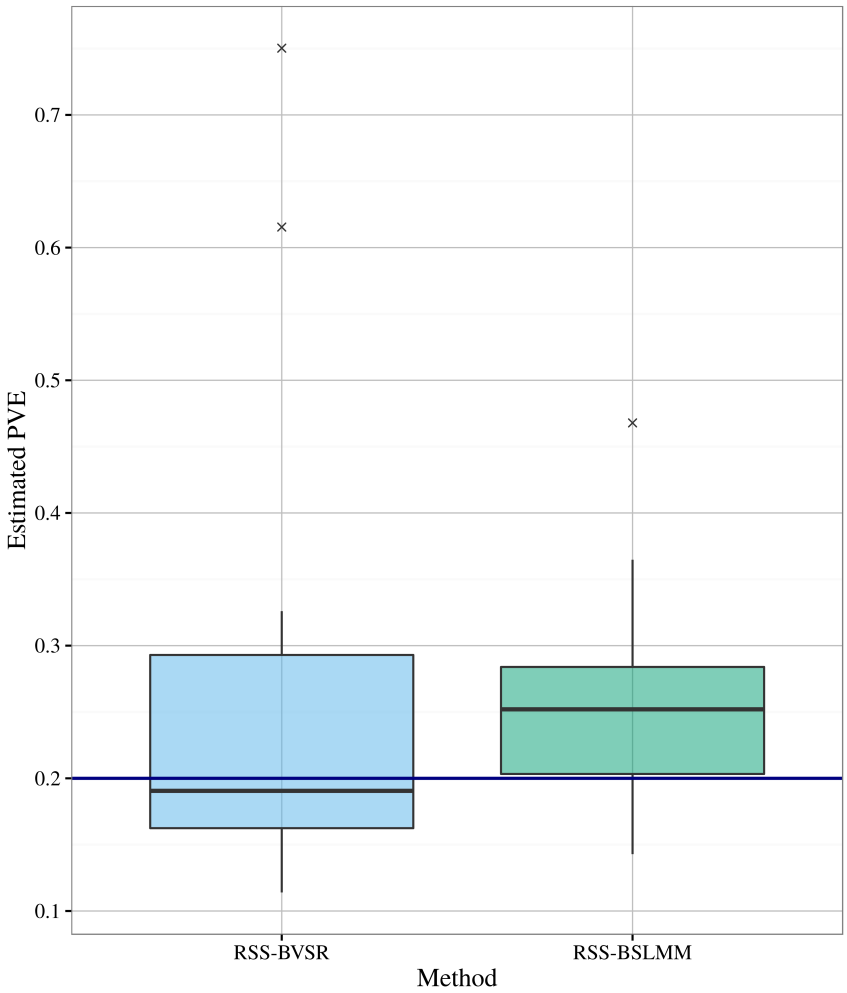


Figure C.6: Simulations show that PVE estimation can be biased when RSS methods are applied to summary data that are *not* generated from the same set of individuals.

Supplementary Figure 7

Summary of sample sizes and maximum squared correlations (r^2) for the 1,064,575 analyzed SNPs from the human height summary dataset (Wood et al., 2014).

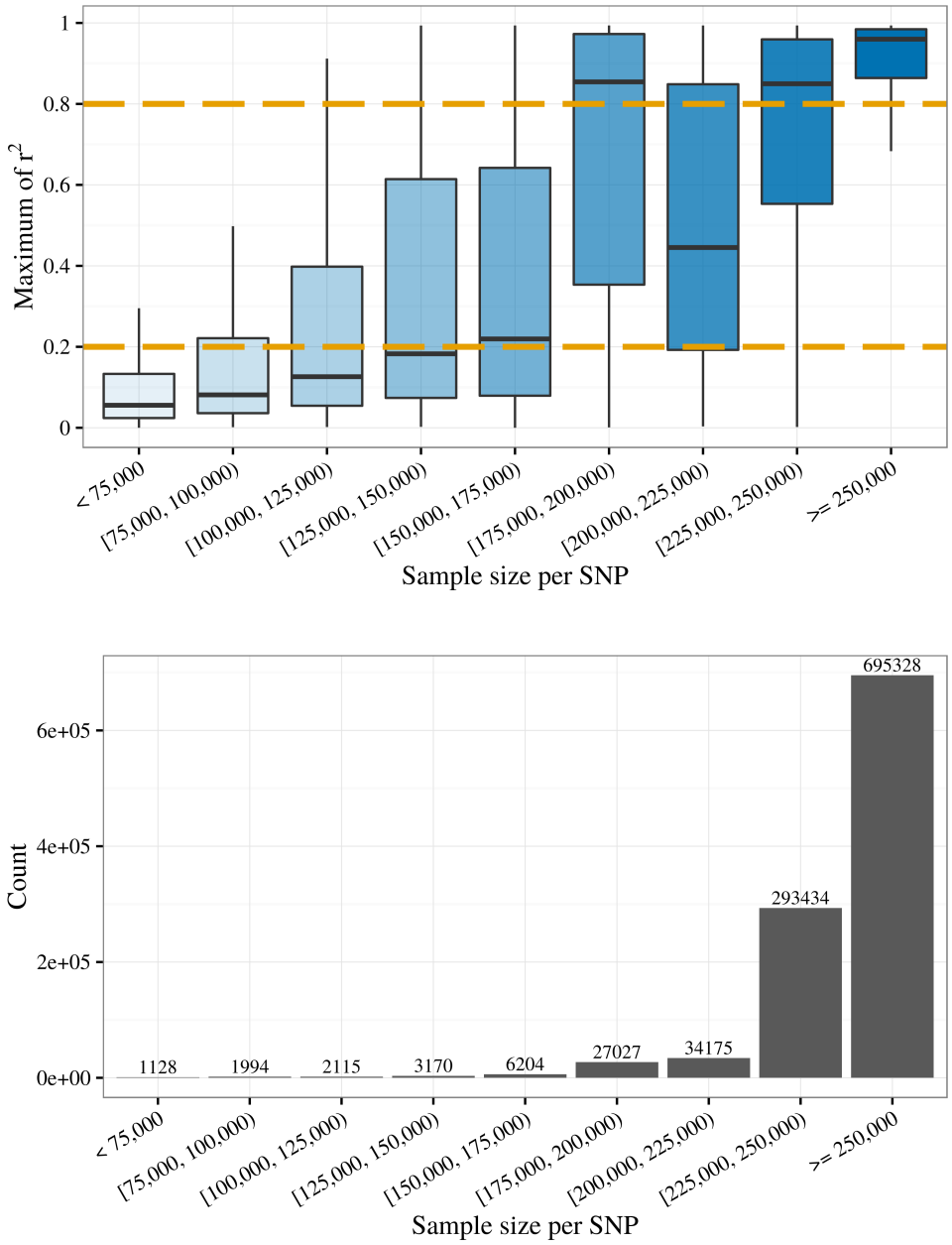


Figure C.7: Summary of sample sizes and maximum squared correlations (r^2) for the 1,064,575 analyzed SNPs from the human height summary dataset (Wood et al., 2014).

Supplementary Figure 8

SNP filtering based on sample sizes can lead to conservative results if the sample size cut-off is too high. Below are the results of fitting RSS-BVSR to the human height summary data (Wood et al., 2014) on Chromosome 16, using all 32,260 SNPs and the 17,721 SNPs with sample size greater than or equal to 250,000, respectively. The cut-off 250,000 may ensure that the summary data of the filtered SNPs are approximately generated from the same sample, but it removes almost *half* of SNPs on Chromosome 16, which further reduces the PVE estimates and association signals.

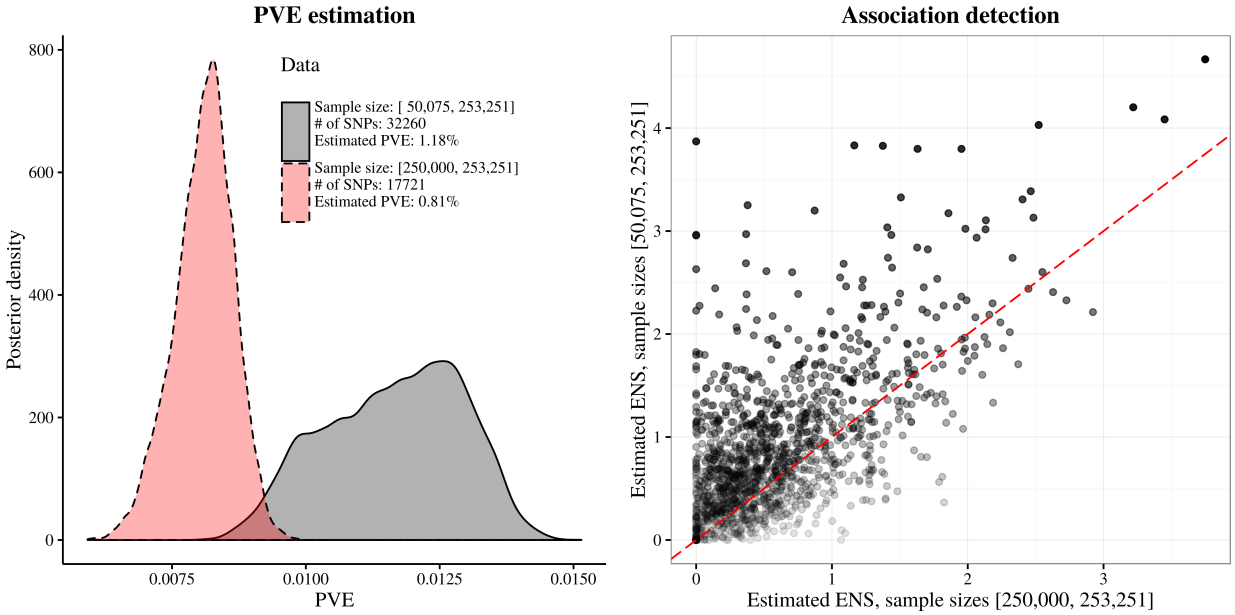


Figure C.8: SNP filtering based on sample sizes can lead to conservative results if the sample size cut-off is too high.

Supplementary Figure 9

Distributions of single-SNP z-scores from the human height GWAS (Wood et al., 2014). Each panel below contains the GWAS z-score distribution of SNPs that pass the leave-one-out (LOO) residual diagnostic filter (red solid curve), and the z-score distribution of SNPs that do *not* pass the filter (green dash curve).

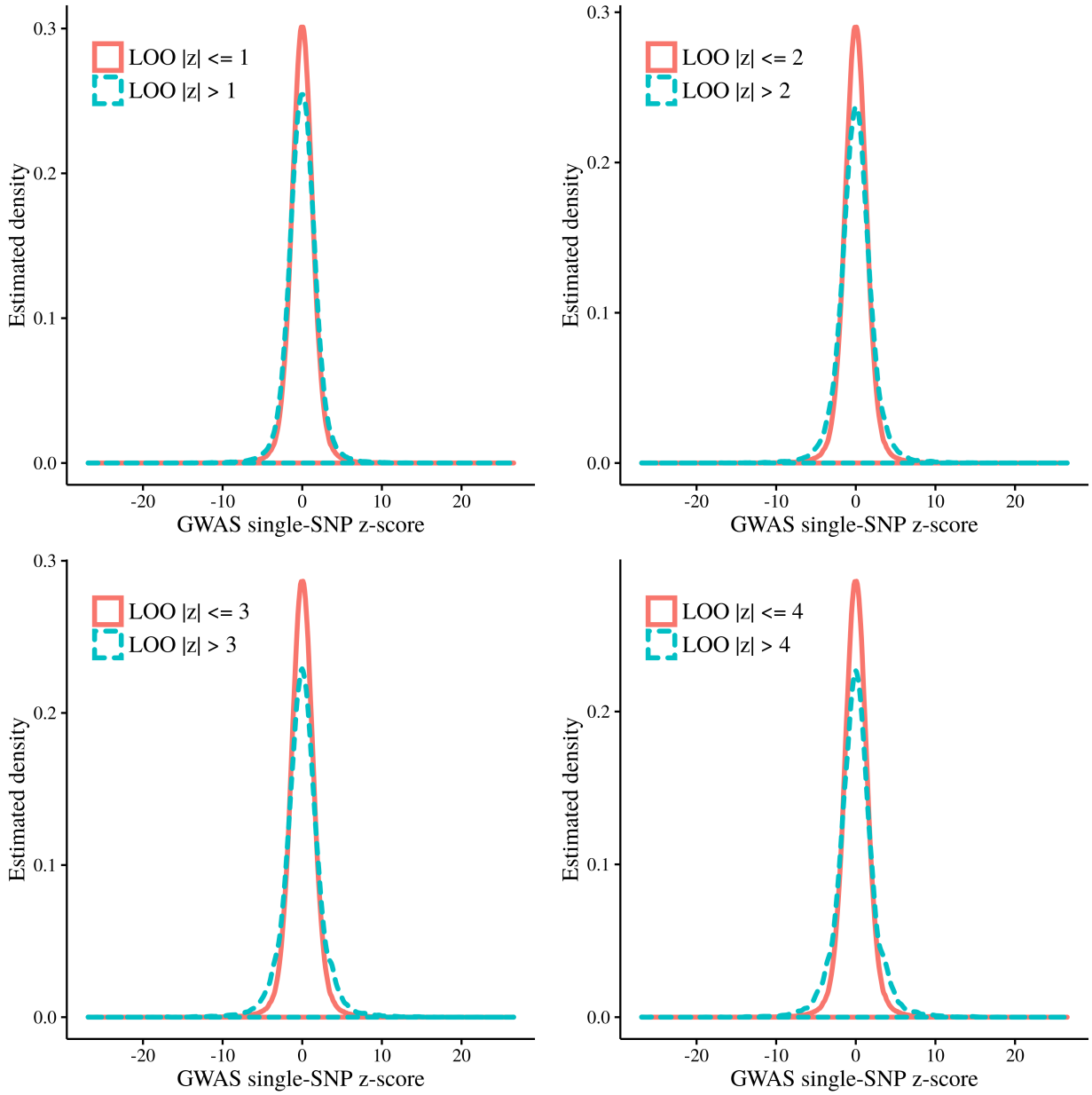


Figure C.9: Distributions of single-SNP z-scores from the human height GWAS (Wood et al., 2014).

APPENDIX D

SUPPLEMENTARY TABLES OF ZHU AND STEPHENS (2017A)

Supplementary Table 1

Phenotype (abbreviation)	Reference
Adult human height	Lango Allen et al. (2010)
Adult human height	Wood et al. (2014)
Body mass index (BMI)	Locke et al. (2015)
Waist-to-hip ratio adjusted for BMI (WHRadjBMI)	Shungin et al. (2015)
High-density lipoprotein (HDL)	Teslovich et al. (2010)
HDL	Global Lipids Genetics Consortium (2013)
Low-density lipoprotein (LDL)	Teslovich et al. (2010)
LDL	Global Lipids Genetics Consortium (2013)
Total cholesterol (TC)	Teslovich et al. (2010)
TC	Global Lipids Genetics Consortium (2013)
Triglycerides (TG)	Teslovich et al. (2010)
TG	Global Lipids Genetics Consortium (2013)
Cigarettes per day	Tobacco and Genetics Consortium (2010)
Smoking age of onset	Tobacco and Genetics Consortium (2010)
Ever versus never smoked	Tobacco and Genetics Consortium (2010)
Current versus former smoker	Tobacco and Genetics Consortium (2010)
Years of educational attainment	Rietveld et al. (2013)
College completion or not	Rietveld et al. (2013)
Depressive	Okbay et al. (2016)
Neuroticism	Okbay et al. (2016)
Schizophrenia	Ripke et al. (2014)
Alzheimer disease	Lambert et al. (2013)
Coronary artery disease (CAD)	Schunkert et al. (2011)

Continued on next page

Table D.1 – continued from previous page

Phenotype (abbreviation)	Reference
Type 2 diabetes (T2D)	Morris et al. (2012)
Haemoglobin	van der Harst et al. (2012)
Mean cell haemoglobin (MCH)	van der Harst et al. (2012)
Mean cell haemoglobin concentration (MCHC)	van der Harst et al. (2012)
Mean cell volume (MCV)	van der Harst et al. (2012)
Packed cell volume (PCV)	van der Harst et al. (2012)
Red blood cell count (RBC)	van der Harst et al. (2012)
Fasting glucose adjusted for BMI (FGadjBMI)	Manning et al. (2012)
Fasting insulin adjusted for BMI (FIadjBMI)	Manning et al. (2012)
Heart rate	Den Hoed et al. (2013)
Serum urate	Köttgen et al. (2013)
Gout	Köttgen et al. (2013)
Rheumatoid arthritis (RA)	Okada et al. (2014)
Inflammatory bowel disease (IBD)	Liu et al. (2015)
Crohn's disease (CD)	Liu et al. (2015)
Ulcerative colitis (UC)	Liu et al. (2015)
CAD	Nikpay et al. (2015)
Myocardial infarction (MI)	Nikpay et al. (2015)
Age at natural menopause (ANM)	Day et al. (2015)

Table D.1: Full names, abbreviations and corresponding references of the GWAS phenotypes that are listed in Table 1 of Zhu and Stephens (2017a).

Supplementary Table 2

Linear relationship between the estimated PVE (SNP heritability) of each chromosome and the chromosome length (unit: Mb) for adult human height (Wood et al., 2014). Shown are the simple linear regression analyses with and without intercept.

	Estimate	**Std. Error**	t **value**	p **value**
Intercept	-6.505×10^{-3}	5.022×10^{-3}	-1.295	0.21
Length	2.379×10^{-4}	3.581×10^{-5}	6.644	1.81×10^{-6}

(a) RSS-BVSR

	Estimate	**Std. Error**	t **value**	p **value**
Intercept	2.189×10^{-4}	2.639×10^{-3}	0.083	0.94
Length	1.854×10^{-4}	1.882×10^{-5}	9.853	4.06×10^{-9}

(b) RSS-BSLMM

	Estimate	**Std. Error**	t **value**	p **value**
Length	1.961×10^{-4}	1.574×10^{-5}	12.460	3.62×10^{-11}

(c) RSS-BVSR

	Estimate	**Std. Error**	t **value**	p **value**
Length	1.868×10^{-4}	7.943×10^{-6}	23.520	$< 2 \times 10^{-16}$

(d) RSS-BSLMM

Table D.2: Linear relationship between the estimated PVE (SNP heritability) of each chromosome and the chromosome length (unit: Mb) for adult human height (Wood et al., 2014).

Supplementary Table 3

Estimated PVE (SNP heritability) of each chromosome for human adult height (Wood et al., 2014). The chromosome length is defined as the distance between the first and the last analyzed SNPs on each chromosome, in Megabases (Mb). The restricted maximum likelihood (REML) estimates h^2_C are obtained from the individual-level data of three GWAS of height (number of SNPs: 593,521-687,398; sample size: 6,293-15,792); see Supplementary Table 2 of Yang et al. (2011). The RSS results are summarized as posterior median and 95% credible interval (C.I.).

Chr.	Length (Mb)	REML		RSS-BVSR		RSS-BSLMM	
		h^2_C	se(h^2_C)	Median	95% C.I.	Median	95% C.I.
1	246.42	0.0377	0.0088	0.0633	[0.0600, 0.0678]	0.0489	[0.0395, 0.0511]
2	242.56	0.0513	0.0094	0.0438	[0.0417, 0.0475]	0.0459	[0.0408, 0.0583]
3	199.30	0.0354	0.0084	0.0334	[0.0308, 0.0402]	0.0362	[0.0294, 0.0394]
4	191.11	0.0310	0.0079	0.0687	[0.0656, 0.0716]	0.0322	[0.0305, 0.0338]
5	180.54	0.0233	0.0078	0.0254	[0.0191, 0.0289]	0.0270	[0.0249, 0.0336]
6	170.64	0.0314	0.0079	0.0334	[0.0311, 0.0361]	0.0363	[0.0298, 0.0383]
7	158.67	0.0147	0.0069	0.0386	[0.0309, 0.0414]	0.0345	[0.0328, 0.0363]
8	146.11	0.0166	0.0068	0.0178	[0.0153, 0.0197]	0.0240	[0.0199, 0.0257]
9	140.15	0.0160	0.0067	0.0186	[0.0153, 0.0312]	0.0318	[0.0292, 0.0336]
10	135.19	0.0196	0.0071	0.0146	[0.0112, 0.0172]	0.0205	[0.0185, 0.0225]
11	134.25	0.0181	0.0064	0.0147	[0.0117, 0.0165]	0.0191	[0.0170, 0.0207]
12	132.26	0.0199	0.0067	0.0332	[0.0294, 0.0361]	0.0319	[0.0281, 0.0339]
13	96.18	0.0139	0.0061	0.0098	[0.0075, 0.0112]	0.0120	[0.0109, 0.0131]
14	87.01	0.0183	0.0060	0.0157	[0.0141, 0.0198]	0.0144	[0.0130, 0.0160]
15	81.88	0.0284	0.0064	0.0239	[0.0194, 0.0319]	0.0245	[0.0225, 0.0260]
16	88.66	0.0129	0.0058	0.0113	[0.0089, 0.0132]	0.0131	[0.0120, 0.0143]
17	78.61	0.0190	0.0060	0.0195	[0.0169, 0.0211]	0.0253	[0.0198, 0.0270]
18	76.11	0.0080	0.0054	0.0046	[0.0039, 0.0055]	0.0069	[0.0060, 0.0079]
19	63.57	0.0067	0.0045	0.0109	[0.0095, 0.0120]	0.0150	[0.0136, 0.0162]
20	62.37	0.0185	0.0058	0.0098	[0.0082, 0.0109]	0.0111	[0.0100, 0.0155]
21	36.88	0.0000	0.0037	0.0036	[0.0029, 0.0045]	0.0044	[0.0038, 0.0051]
22	35.13	0.0080	0.0040	0.0042	[0.0033, 0.0049]	0.0057	[0.0044, 0.0067]
Total		0.4487	0.0290	0.5238	[0.5035, 0.5449]	0.5209	[0.5027, 0.5390]

Table D.3: Estimated PVE (SNP heritability) of each chromosome for human adult height (Wood et al., 2014).

Supplementary Table 4

(a) Total PVE (SNP heritability) estimates and 95% credible intervals.

	RSS-BVSR	**RSS-BSLMM**
All SNPs	52.4%, [50.4%, 54.5%]	52.1%, [50.3%, 53.9%]
Filtered SNPs, LOO $\|z\|$-score ≤ 2	34.0%, [32.9%, 35.0%]	45.3%, [44.7%, 46.0%]
Filtered SNPs, LOO $\|z\|$-score ≤ 3	35.3%, [34.2%, 36.3%]	48.2%, [47.5%, 48.9%]

(b) The number of genome-wide significant SNPs (GWAS hits) reported in Wood et al. (2014) that are identified by RSS-BVSR (i.e. covered by a $\pm$40-kb region with estimated ENS ≥ 1).

	All 697 GWAS hits	**Included 384 GWAS hits**
All SNPs	531	371
Filtered SNPs, LOO $\|z\|$-score ≤ 2	532	373
Filtered SNPs, LOO $\|z\|$-score ≤ 3	540	370

(c) The number of $\pm$40-kb regions in the whole genome that are identified by RSS-BVSR (estimated ENS ≥ 1), and the number of putatively new regions (estimated ENS ≥ 1, and at least 1 Mb away from the 697 previously reported GWAS hits).

	All regions	**Putatively new regions**
All SNPs	5194	2138
Filtered SNPs, LOO $\|z\|$-score ≤ 2	6426	2798
Filtered SNPs, LOO $\|z\|$-score ≤ 3	6848	3079

Table D.4: Summary of RSS analyses of human height data (Wood et al., 2014).

Supplementary Table 5

Putatively new loci identified by RSS-BVSR analyses that are associated with adult human height (estimated ENS > 3). Table columns from left to right are: (1) chromosome number; (2) starting position of the ±40-kb region; (3) ending position of the region; (4) estimated ENS; (5) the nearest genome-wide significant SNP reported by Wood et al. (2014); (6) the physical distance to the nearest GWAS hit, in Megabases (Mb); (6) the nearest neighbor gene; (7) the relationship between the region and the nearest gene. The nearest genes to genomic regions are found and annotated by the function `matchGenes` in the package `bumphunter` (Jaffe et al., 2012). All SNP information and genomic positions are based on Human Genome Assembly 19 (Genome Reference Consortium GRCh37).

(a) Using summary data of all SNPs (1,064,575).

Chr.	Start	End	ENS	Nearest Hit	Distance (Mb)	Nearest Gene	Annotation
5	86116344	86196344	5.22	rs6894139	2.13	*COX7C*	downstream
5	86156344	86236344	4.74	rs6894139	2.09	*MIR4280*	downstream
16	10715041	10795041	4.03	rs1659127	3.59	*TEKT5*	covers
16	78795041	78875041	3.83	rs4243206	2.71	*WWOX*	inside intron
16	78835041	78915041	3.83	rs4243206	2.67	*WWOX*	inside intron
22	43637135	43717135	3.78	rs11090631	2.13	*SCUBE1*	covers exon(s)
22	43597135	43677135	3.71	rs11090631	2.17	*SCUBE1*	overlaps 3’
12	85911619	85991619	3.67	rs17783015	4.24	*RASSF9*	downstream
19	57923127	58003127	3.55	rs2059877	9.73	*ZNF419*	overlaps 5’
8	6364984	6444984	3.54	rs4875421	1.54	*MCPH1*	inside intron
19	15723127	15803127	3.50	rs8103068	1.72	*CYP4F12*	overlaps 5’
20	821795	901795	3.41	rs7273787	3.20	*FAM110A*	covers
16	73755041	73835041	3.39	rs11640018	1.49	*LINC01568*	downstream
16	19275041	19355041	3.33	rs2023693	1.52	*CLEC19A*	covers
16	80315041	80395041	3.31	rs4243206	1.19	*DYNLRB2*	upstream
12	85871619	85951619	3.31	rs17783015	4.28	*ALX1*	downstream
20	50181795	50261795	3.31	rs6020202	1.55	*ATP9A*	overlaps 3’
17	14612467	14692467	3.25	rs8069300	2.63	*CDRT7*	upstream
19	52003127	52083127	3.16	rs2059877	3.81	*SIGLEC6*	covers
19	57963127	58043127	3.10	rs2059877	9.77	*ZNF419*	covers
12	30791619	30871619	3.05	rs10843390	1.29	*CAPRIN2*	overlaps 3’
16	19235041	19315041	3.04	rs2023693	1.56	*CLEC19A*	overlaps 5’
16	55075041	55155041	3.02	rs8058684	1.56	*IRX5*	downstream

(b) Using summary data of 1,018,617 filtered SNPs with LOO imputation $|z|$-score ≤ 3.

Chr.	Start	End	ENS	Nearest Hit	Distance (Mb)	Nearest Gene	Annotation
22	43597135	43677135	4.57	rs11090631	2.17	*SCUBE1*	overlaps 3'
22	43637135	43717135	4.44	rs11090631	2.13	*SCUBE1*	covers exon(s)
16	73755041	73835041	4.08	rs11640018	1.49	*LINC01568*	downstream
19	52123127	52203127	4.02	rs2059877	3.93	*SIGLEC5*	overlaps 5'
17	75492467	75572467	3.97	rs1552173	1.15	*SEPT9*	overlaps 3'
19	1043127	1123127	3.97	rs11880992	1.05	*POLR2E*	covers
19	15723127	15803127	3.93	rs8103068	1.72	*CYP4F12*	overlaps 5'
19	54683127	54763127	3.85	rs2059877	6.49	*LILRA6*	overlaps 3'
16	78835041	78915041	3.82	rs4243206	2.67	*WWOX*	inside intron
19	52163127	52243127	3.80	rs2059877	3.97	*MIR99B*	covers
16	78795041	78875041	3.75	rs4243206	2.71	*WWOX*	inside intron
8	3564984	3644984	3.70	rs4875421	1.18	*CSMD1*	inside intron
17	1612467	1692467	3.59	rs870183	1.01	*MIR22HG*	covers
21	41246282	41326282	3.58	rs2211866	1.56	*PCP4*	overlaps 3'
10	5978481	6058481	3.57	rs4332428	1.01	*FBXO18*	overlaps 3'
14	94785431	94865431	3.45	rs7154721	2.36	*SERPINA6*	overlaps 5'
17	9092467	9172467	3.42	rs8067165	1.06	*STX8*	overlaps 3'
19	14923127	15003127	3.42	rs8103068	2.52	*OR7A10*	covers
16	79035041	79115041	3.40	rs4243206	2.47	*WWOX*	inside intron
17	3932467	4012467	3.40	rs870183	3.33	*ZZEF1*	overlaps 3'
19	51283127	51363127	3.39	rs2059877	3.09	*ACPT*	covers
19	1083127	1163127	3.37	rs11880992	1.01	*POLR2E*	covers
17	5692467	5772467	3.34	rs9217	1.59	*LOC339166*	covers exon(s)
15	96121372	96201372	3.30	rs7181724	1.57	*LINC00924*	downstream
19	15763127	15843127	3.25	rs8103068	1.68	*CYP4F12*	covers
16	12635041	12715041	3.22	rs1659127	1.67	*SNX29*	overlaps 3'
16	12675041	12755041	3.18	rs1659127	1.63	*CPPED1*	overlaps 3'
8	3524984	3604984	3.16	rs4875421	1.22	*CSMD1*	covers exon(s)
21	41206282	41286282	3.15	rs2211866	1.52	*PCP4*	overlaps 5'
17	35172467	35252467	3.15	rs2338115	1.68	*LHX1*	upstream
17	6252467	6332467	3.13	rs9217	1.03	*AIPL1*	overlaps 3'
17	1652467	1732467	3.10	rs870183	1.05	*SERPINF2*	overlaps 3'
22	34197135	34277135	3.09	rs2413143	1.14	*LARGE*	covers exon(s)
17	75532467	75612467	3.08	rs1552173	1.11	*LOC100507351*	covers
17	14612467	14692467	3.06	rs8069300	2.63	*CDRT7*	upstream
17	75012467	75092467	3.05	rs1552173	1.63	*SCARNA16*	covers
16	65875041	65955041	3.04	rs1966913	1.43	*LINC00922*	upstream
17	52932467	53012467	3.04	rs11867943	1.22	*TOM1L1*	overlaps 5'
17	55852467	55932467	3.00	rs1401795	1.01	*MRPS23*	covers

(c) Using summary data of 938,798 filtered SNPs with LOO imputation $|z|$-score ≤ 2.

Chr.	**Start**	**End**	**ENS**	**Nearest Hit**	**Distance (Mb)**	**Nearest Gene**	**Annotation**
22	43637135	43717135	4.23	rs11090631	2.13	*SCUBE1*	covers exon(s)
16	73755041	73835041	3.93	rs11640018	1.49	*LINC01568*	downstream
22	43597135	43677135	3.91	rs11090631	2.17	*SCUBE1*	overlaps 3'
19	54683127	54763127	3.61	rs2059877	6.49	*LILRA6*	overlaps 3'
19	723127	803127	3.58	rs11880992	1.37	*PTBP1*	overlaps 5'
15	91561372	91641372	3.45	rs2238300	1.71	*VPS33B*	overlaps 5'
16	79275041	79355041	3.43	rs4243206	2.23	*WWOX*	downstream
17	10852467	10932467	3.13	rs8069300	1.05	*PIRT*	upstream
21	41206282	41286282	3.09	rs2211866	1.52	*PCP4*	overlaps 5'
17	50252467	50332467	3.09	rs4605213	1.01	*CA10*	upstream
16	79315041	79395041	3.08	rs4243206	2.19	*WWOX*	downstream
17	71092467	71172467	3.07	rs10083886	1.17	*SSTR2*	covers
19	15723127	15803127	3.07	rs8103068	1.72	*CYP4F12*	overlaps 5'
16	55075041	55155041	3.04	rs8058684	1.56	*IRX5*	downstream
17	17052467	17132467	3.02	rs4640244	4.15	*MPRIP*	covers

Table D.5: Putatively new loci identified by RSS-BVSR analyses that are associated with adult human height (estimated ENS > 3).

Supplementary Table 6

Computation time (hour:minute:second) of RSS-BVSR and RSS-BSLMM in the analyses of adult human height data (Wood et al., 2014). Computations were performed on a single core of Intel E5-2670 2.6GHz or AMD Opteron 6386 SE, with 2 million MCMC iterations per chromosome.

(a) All SNPs (1,064,575).

Chr.	# of SNPs	RSS-BVSR	RSS-BSLMM
1	86924	08:50:41	18:49:15
2	94042	16:27:26	31:33:12
3	76481	01:34:58	34:30:41
4	67627	05:42:59	15:02:51
5	67452	15:01:51	29:39:41
6	60268	03:39:18	24:32:09
7	59740	02:59:43	17:06:43
8	58361	14:59:04	28:19:17
9	52633	11:28:05	20:57:16
10	58236	28:16:28	24:40:29
11	52180	21:12:16	21:43:08
12	51123	02:02:10	18:35:34
13	43464	07:45:23	20:33:17
14	37540	01:02:52	16:32:27
15	34726	08:31:56	15:47:45
16	32260	08:43:07	10:44:12
17	25533	15:33:12	09:04:38
18	31596	05:24:35	13:50:00
19	17507	16:50:13	04:18:35
20	25983	05:58:52	08:31:22
21	15300	01:51:53	04:30:42
22	15599	02:04:01	05:55:55

(b) Filtered SNPs (LOO imputation $|z|$-score ≤ 2).

Chr.	# of SNPs	RSS-BVSR	RSS-BSLMM
1	75746	05:07:03	19:26:49
2	83175	05:54:35	24:53:23
3	67258	05:42:02	24:44:39
4	59391	18:44:43	15:51:33
5	59886	04:35:35	18:27:03
6	52539	05:00:59	15:41:28
7	52739	27:20:38	13:40:14
8	52067	18:32:13	35:38:15
9	46720	05:49:21	14:34:32
10	51038	25:59:36	35:41:12
11	46036	23:03:39	35:40:52
12	44721	03:59:24	28:02:50
13	38644	04:31:00	13:58:33
14	33118	18:32:42	12:22:30
15	30644	28:49:42	10:03:44
16	28770	16:10:05	13:06:48
17	25533	16:50:20	05:57:31
18	22337	04:40:37	08:58:16
19	15267	05:49:31	03:00:28
20	23086	03:31:40	05:51:34
21	13663	02:52:47	04:40:02
22	13674	02:14:44	05:20:14

(c) Filtered SNPs (LOO imputation $|z|$-score ≤ 3).

Chr.	**# of SNPs**	**RSS-BVSR**	**RSS-BSLMM**
1	82625	05:20:16	26:37:00
2	90263	03:40:14	29:57:33
3	73042	04:58:29	35:41:18
4	64605	20:10:46	19:51:18
5	64869	05:22:55	27:45:49
6	57241	03:05:23	18:44:27
7	57243	04:43:04	16:08:54
8	56139	33:10:26	35:35:08
9	50555	05:53:44	15:32:51
10	55544	28:07:26	35:36:17
11	49893	24:21:42	35:36:44
12	48770	05:26:07	35:36:09
13	41685	15:52:05	35:36:03
14	36012	22:36:37	27:39:27
15	33203	18:27:45	29:00:33
16	31008	22:47:40	29:01:19
17	24322	19:59:12	07:01:32
18	30449	03:54:09	21:05:19
19	16595	10:26:02	03:26:06
20	24956	04:17:32	06:15:45
21	14755	02:13:41	05:24:09
22	14843	03:00:16	06:36:39

Table D.6: Computation time (hour:minute:second) of RSS-BVSR and RSS-BSLMM in the analyses of adult human height data (Wood et al., 2014).

APPENDIX E

SUPPLEMENTARY NOTE OF ZHU AND STEPHENS (2017B)

E.1 Posterior computation

We use variational inference to estimate the posterior distribution of multiple regression coefficients β based on the input dataset $\mathbf{D} := \{\widehat{\beta}, S, R, \mathbf{a}\}$, which includes GWAS summary statistics $(\widehat{\beta}, S)$, LD estimates (R) and SNP-level annotations $(\mathbf{a})$.

We first introduce a binary vector $\gamma := (\gamma_1, \ldots, \gamma_p)^\intercal \in \{0,1\}^p$, where $\gamma_j = 1$ when β_j is drawn from the non-zero component $\mathcal{N}(0, \sigma_\beta^2)$ *a priori*, and $\gamma_j = 0$ when β_j is drawn from point mass zero δ_0.

The posterior computation procedures in Zhu and Stephens (2017b) largely follow those developed by Carbonetto and Stephens (2012). Firstly, for each set of hyper-parameters $\{\theta_0, \theta, h\}$ from a predefined grid, we approximate $p(\beta, \gamma | \mathbf{D}, \theta_0, \theta, h)$ using a mean-filed variational Bayes algorithm (Section E.1). Next, we approximate $p(\theta_0, \theta, h | \mathbf{D})$ by a discrete distribution on the predefined grid, using the variational solutions from the first step to compute the posterior probabilities (Section E.1). Finally, we integrate out $p(\beta, \gamma | \mathbf{D}, \theta_0, \theta, h)$ over the posterior of $p(\theta_0, \theta, h | \mathbf{D})$ to obtain the posterior of $\{\beta, \gamma\}$:

$$p(\beta, \gamma | \mathbf{D}) = \int p(\beta, \gamma | \mathbf{D}, \theta_0, \theta, h) p(\theta_0, \theta, h | \mathbf{D}) \mathrm{d}\theta_0 \mathrm{d}\theta \mathrm{d}h. \tag{E.1}$$

Estimate $p(\beta, \gamma | \mathbf{D}, \theta_0, \theta, h)$

The aim of the first step is to search for a distribution $q(\beta, \gamma)$ that minimize the Kullback-Leibler (KL) divergence between $q(\beta, \gamma)$ and $p(\beta, \gamma | \mathbf{D}, \theta_0, \theta, h)$.

For any distribution $q(\boldsymbol{\beta},\boldsymbol{\gamma})$, we have the following decomposition:

$$\log p(\widehat{\boldsymbol{\beta}}|S,R,\mathbf{a},\theta_0,\theta,h) = \underbrace{\mathrm{E}_q \log\left[\frac{q(\boldsymbol{\beta},\boldsymbol{\gamma})}{p(\boldsymbol{\beta},\boldsymbol{\gamma}|\mathbf{D},\theta_0,\theta,h)}\right]}_{\text{Kullback-Leibler divergence}} + \underbrace{\mathrm{E}_q \log\left[\frac{p(\widehat{\boldsymbol{\beta}},\boldsymbol{\beta},\boldsymbol{\gamma}|S,R,\mathbf{a},\theta_0,\theta,h)}{q(\boldsymbol{\beta},\boldsymbol{\gamma})}\right]}_{\text{Variational lower bound}}. \tag{E.2}$$

Because the left-hand side of Equation (E.2) does not dependent on $\{\boldsymbol{\beta},\boldsymbol{\gamma}\}$, minimizing KL divergence is equivalent to maximizing the variational lower bound. In the present study, we only restrict the family of $q(\boldsymbol{\beta},\boldsymbol{\gamma})$ to be of fully-factorized form (a.k.a. mean-field approximation):

$$q(\boldsymbol{\beta},\boldsymbol{\gamma}) = \prod_{j=1}^{p} q_j(\beta_j,\gamma_j), \tag{E.3}$$

and do not make any additional assumption for q.

Straightforward algebra shows that for each q_j, the optimal variational solution $q_j^\star$ is given by

$$q_j^\star(\beta_j,\gamma_j) = \left[\alpha_j^\star \mathcal{N}(\beta_j;\mu_j^\star,(\sigma_j^\star)^2)\right]^{\gamma_j}\left[(1-\alpha_j^\star)\delta_0(\beta_j)\right]^{1-\gamma_j}, \tag{E.4}$$

implying that with probability $\alpha_j^\star$, β_j is normally distributed with mean $\mu_j^\star$ and variance $(\sigma_j^\star)^2$, and with probability $1-\alpha_j^\star$, $\beta_j = 0$. Following Carbonetto and Stephens (2012), we use a coordinate descent algorithm to estimate $\left\{\alpha_j^\star,\mu_j^\star,\sigma_j^\star\right\}$:

$$(\sigma_j^\star)^2 = (s_j^{-2}+\sigma_\beta^{-2})^{-1} \tag{E.5}$$

$$\mu_j^\star = (\sigma_j^\star)^2\left(\frac{\hat{\beta}_j}{s_j^2} - \sum_{i\neq j}\frac{R_{ij}\alpha_i^\star\mu_i^\star}{s_i s_j}\right) \tag{E.6}$$

$$\frac{\alpha_j^\star}{1-\alpha_j^\star} = \frac{\pi_j}{1-\pi_j}\cdot\frac{\sigma_j^\star}{\sigma_\beta}\cdot\exp\left\{\frac{(\mu_j^\star)^2}{2(\sigma_j^\star)^2}\right\} \tag{E.7}$$

Although it is not explicitly shown in notation here, the optimal solution $q^\star$ depends on the values of hyper-parameters $\{\theta_0,\theta,h\}$, because π_j is a function of $\{\theta_0,\theta,h\}$.

Estimate $p(\theta_0,\theta,h|\mathbf{D})$

Since independent uniform grid priors are used [Supplementary Tables 3-4 of Zhu and Stephens (2017b)], the posterior distribution of $\{\theta_0,\theta,h\}$ is proportional to marginal likelihood:

$$p(\theta_0,\theta,h|\mathbf{D}) = p(\theta_0,\theta,h|\widehat{\beta},S,R,\mathbf{a}) \propto p(\widehat{\beta}|S,R,\mathbf{a},\theta_0,\theta,h). \tag{E.8}$$

Noting that the marginal likelihood $p(\widehat{\beta}|S,R,\mathbf{a},\theta_0,\theta,h)$ is analytically intractable, we thus use variational lower bound as an approximation.

Using Jensen's inequality, we can see that the marginal log likelihood of (θ_0,θ,h) is bounded from below by the variational lower bound,

$$\log p(\widehat{\beta}|S,R,\mathbf{a},\theta_0,\theta,h) \geq \mathrm{E}_q \log\left[\frac{p(\widehat{\beta},\beta,\gamma|S,R,\mathbf{a},\theta_0,\theta,h)}{q(\beta,\gamma)}\right]. \tag{E.9}$$

Furthermore, if the distribution $q(\beta,\gamma)$ takes the form (E.4)[1], then the variational lower bound is analytically available:

$$\mathrm{E}_q \log\left[\frac{p(\widehat{\beta},\beta,\gamma|S,R,\mathbf{a},\theta_0,\theta,h)}{q(\beta,\gamma)}\right] = F_0(\mathbf{D}) + F(\mathbf{D},\theta_0,\theta,h), \tag{E.10}$$

where $F_0(\mathbf{D})$ is a constant term with respect to $\{\theta_0,\theta,h\}$,

$$F_0(\mathbf{D}) = -\frac{1}{2}\log|2\pi SRS| - \frac{1}{2}\widehat{\beta}^\top (SRS)^{-1}\widehat{\beta}, \tag{E.11}$$

and

$$\begin{aligned} F(\mathbf{D},\theta_0,\theta,h) &= \widehat{\beta}^\top S^{-2}\mathrm{E}_q(\beta) - \frac{1}{2}\mathrm{E}_q^\top(\beta)S^{-1}RS^{-1}\mathrm{E}_q(\beta) - \frac{1}{2}\sum_{j=1}^{p}\frac{\mathrm{Var}_q(\beta_j)}{s_j^2} - \sum_{j=1}^{p}\alpha_j\log\left(\frac{\alpha_j}{\pi_j}\right) \\ &\quad - \sum_{j=1}^{p}(1-\alpha_j)\log\left(\frac{1-\alpha_j}{1-\pi_j}\right) + \sum_{j=1}^{p}\frac{\alpha_j}{2}\left[1+\log\left(\frac{\sigma_j^2}{\sigma_\beta^2}\right) - \frac{\sigma_j^2+\mu_j^2}{\sigma_\beta^2}\right], \end{aligned} \tag{E.12}$$

1. Note that here the parameters $\{\alpha_j,\mu_j,\sigma_j\}$ do *not* have the constraints specified by (E.5)-(E.7).

$\mathrm{E}_q(\boldsymbol{\beta}) = (\mathrm{E}_q(\beta_1), \ldots, \mathrm{E}_q(\beta_p))^\top$, $\mathrm{E}_q(\beta_j) = \alpha_j \mu_j$ and $\mathrm{Var}_q(\beta_j) = \alpha_j(\sigma_j^2 + \mu_j^2) - (\alpha_j \mu_j)^2$.

Finally, $p(\theta_0, \theta, h|\mathbf{D})$ is estimated as

$$p(\theta_0, \theta, h|\mathbf{D}) \approx \widetilde{w}(\theta_0, \theta, h) \propto \exp\{F(\mathbf{D}, \theta_0, \theta, h)\}. \tag{E.13}$$

Note that $p(\theta_0, \theta, h|\mathbf{D})$ is approximated by a discrete distribution $\widetilde{w}(\theta_0, \theta, h)$, since the support of $\{\theta_0, \theta, h\}$ is a predefined grid.

Squared iterative method

When estimating $p(\boldsymbol{\beta}, \boldsymbol{\gamma}|\mathbf{D}, \theta_0, \theta, h)$, the coordinate descent updates (E.5)-(E.7) essentially define a fixed-point mapping. To improve the convergence, we embed an "off-the-shelf" accelerator, squared iterative methods [SQUAREM, Varadhan and Roland (2008)], in this fixed-point mapping.

E.2 Bayes factor for gene set enrichment

To measure the evidence for the enrichment hypothesis that a candidate gene set is enriched ($\theta > 0$) for phenotype-genotype associations against the baseline hypothesis ($\theta = 0$), we evaluate the following Bayes factor (BF):

$$\mathrm{BF} = \frac{p(\widehat{\boldsymbol{\beta}}|S, R, \mathbf{a}, \theta > 0)}{p(\widehat{\boldsymbol{\beta}}|S, R, \mathbf{a}, \theta = 0)}. \tag{E.14}$$

To compute BF (E.14), we approximate the intractable marginal likelihoods by the corresponding variational lower bounds (E.12):

$$\begin{aligned}
\text{BF} &= \frac{\int p(\widehat{\beta}|S,R,\mathbf{a},\theta_0,\theta,h)p(\theta_0)p(\theta)p(h)\mathrm{d}\theta\mathrm{d}\theta_0\mathrm{d}h}{\int p(\widehat{\beta}|S,R,\mathbf{a},\theta_0,\theta=0,h)p(\theta_0)p(h)\mathrm{d}\theta_0\mathrm{d}h} \\
&\approx \frac{\int \exp\{F_0(\mathbf{D})+F(\mathbf{D},\theta_0,\theta,h)\}p(\theta_0)p(\theta)p(h)\mathrm{d}\theta\mathrm{d}\theta_0\mathrm{d}h}{\int \exp\{F_0(\mathbf{D})+F(\mathbf{D},\theta_0,\theta=0,h)\}p(\theta_0)p(h)\mathrm{d}\theta_0\mathrm{d}h} \\
&\approx \frac{n_1^{-1}\sum_{s=1}^{n_1}\exp\{F(\mathbf{D},\theta_0^{(s)},\theta^{(s)},h^{(s)})\}}{n_0^{-1}\sum_{t=1}^{n_0}\exp\{F(\mathbf{D},\theta_0^{(t)},\theta=0,h^{(t)})\}},
\end{aligned} \tag{E.15}$$

where $\left\{\theta_0^{(s)},\theta^{(s)},h^{(s)}\right\}$ and $\left\{\theta_0^{(t)},h^{(t)}\right\}$ are evenly spaced points on a regular grid over finite intervals.

E.3 Posterior statistics of genetic associations

To identify loci associated with a given phenotype, we consider two posterior statistics derived from the variational inference.

The first statistic is P_1, the posterior probability that at least one SNP in a given locus is associated with the phenotype:

$$P_1 := 1-\Pr(\beta_j=0,\ \forall j\in\text{locus}|\mathbf{D}). \tag{E.16}$$

Given the grid $\left\{\theta_0^{(s)},\theta^{(s)},h^{(s)}\right\}$, the numerical estimate of P_1 is given by

$$\begin{aligned}
P_1 &= 1-\int \Pr(\beta_j=0,\ \forall j\in\text{locus}|\mathbf{D},\theta_0,\theta,h)p(\theta_0,\theta,h|\mathbf{D})\mathrm{d}\theta_0\mathrm{d}\theta\mathrm{d}h \\
&\approx 1-\sum_{s=1}^{n_1}\Pr(\beta_j=0,\ \forall j\in\text{locus}|\mathbf{D},\theta_0^{(s)},\theta^{(s)},h^{(s)})\cdot\widetilde{w}(\theta_0^{(s)},\theta^{(s)},h^{(s)}).
\end{aligned}$$

Since the posterior of β is approximated by a fully-factorized distribution $q^\star$, for any (θ_0,θ,h),

$$\Pr(\beta_j=0,\ \forall j\in\text{locus}|\mathbf{D},\theta_0,\theta,h)\approx\prod_{j\in\text{locus}}q_j^\star(\beta_j=0)=\prod_{j\in\text{locus}}\left[1-\alpha_j^\star(\theta_0,\theta,h)\right].$$

Hence, the final estimate of the posterior statistic P_1 averaged over the grid $\left\{\theta_0^{(s)}, \theta^{(s)}, h^{(s)}\right\}$ is

$$P_1 \approx 1 - \sum_{s=1}^{n_1} \prod_{j \in \text{locus}} \left[1 - \alpha_j^\star(\theta_0^{(s)}, \theta^{(s)}, h^{(s)})\right] \cdot \widetilde{w}(\theta_0^{(s)}, \theta^{(s)}, h^{(s)}). \tag{E.17}$$

The second statistic is ENS, the posterior expected number of associated SNPs in the locus:

$$\text{ENS} := \sum_{j \in \text{locus}} \Pr(\beta_j \neq 0 | \mathbf{D}). \tag{E.18}$$

Given the grid $\left\{\theta_0^{(s)}, \theta^{(s)}, h^{(s)}\right\}$ and the corresponding variational estimates, ENS is estimated as

$$\text{ENS} \approx \sum_{j \in \text{locus}} \sum_{s=1}^{n_1} \alpha_j^\star(\theta_0^{(s)}, \theta^{(s)}, h^{(s)}) \cdot \widetilde{w}(\theta_0^{(s)}, \theta^{(s)}, h^{(s)}). \tag{E.19}$$

Note that in contrast to P_1, the estimate of ENS does not require the fully-factorized assumption about $q^\star$. When estimating whole genome ENS in Supplementary Figure 6 of Zhu and Stephens (2017b), we treat the entire genome as a "locus" in the calculation.

Notice that we compute P_1 and ENS above under the enrichment hypothesis that the candidate gene set is enriched ($\theta > 0$). Similarly, under the baseline hypothesis that no gene set is enriched ($\theta = 0$), the numerical estimates of P_1 and ENS are given by:

$$P_1 \approx 1 - \sum_{t=1}^{n_0} \prod_{j \in \text{locus}} \left[1 - \alpha_j^\star(\theta_0^{(t)}, \theta = 0, h^{(t)})\right] \cdot \widetilde{w}(\theta_0^{(t)}, \theta = 0, h^{(t)}), \tag{E.20}$$

$$\text{ENS} \approx \sum_{j \in \text{locus}} \sum_{t=1}^{n_0} \alpha_j^\star(\theta_0^{(t)}, \theta = 0, h^{(t)}) \cdot \widetilde{w}(\theta_0^{(t)}, \theta = 0, h^{(t)}). \tag{E.21}$$

E.4 Estimate the fraction of trait-associated SNPs

The fraction of trait-associated SNPs is one of the two quantities that we use to summarize the effect size distribution of a trait [Figure 2 and Supplementary Figure 2 of Zhu and Stephens (2017b)]. Here we consider two methods to estimate this quantity.

The first approach uses both estimated variational lower bounds and variational param-

eters:

$$\frac{1}{p}\sum_{j=1}^{p}\sum_{t=1}^{n_0}\alpha_j^\star(\theta_0^{(t)},\theta=0,h^{(t)})\cdot\widetilde{w}(\theta_0^{(t)},\theta=0,h^{(t)}). \tag{E.22}$$

We use this approach to generate Figure 2, Supplementary Figure 2 and the y-axis of Supplementary Figure 19 of Zhu and Stephens (2017b).

The second approach only uses estimated variation lower bounds:

$$\sum_{t=1}^{n_0}\left(1+10^{-\theta_0^{(t)}}\right)^{-1}\cdot\widetilde{w}(\theta_0^{(t)},\theta=0,h^{(t)}). \tag{E.23}$$

We use this approach to generate the x-axis of Supplementary Figure 19 of Zhu and Stephens (2017b).

E.5 Estimate the standardized effect size of trait-associated SNPs

The standardized effect size of trait-associated SNPs is another quantity that we use to summarize the effect size distribution of a trait [Figure 2 and Supplementary Figure 2 of Zhu and Stephens (2017b)]. For any given variational approximation $q^\star$, we estimate this quantity by

$$\frac{\sum_{j=1}^{p}\alpha_j^\star\mu_j^\star}{\hat{\sigma}_y\cdot\sum_{j=1}^{p}\alpha_j^\star}. \tag{E.24}$$

Here $\hat{\sigma}_y^2$ is the sample variance of phenotype measurements, which is often not publicly available but can be estimated as follows:

$$\hat{\sigma}_y^2\approx 2n_jf_j(1-f_j)s_j^2, \tag{E.25}$$

where f_j and n_j are the minor allele frequency and sample size of SNP j. Although the approximated values of $\hat{\sigma}_y^2$ sometimes differ across SNPs because of different $\{n_j,f_j,s_j\}$, they often fall into a small range, and thus we use the median of these values as an final estimate.

E.6 Compute credible intervals

Following Carbonetto and Stephens (2013), we use the variational estimate of $p(\theta_0, \theta, h|\mathbf{D})$, $\widetilde{w}(\theta_0, \theta, h)$, to compute a credible interval for any quantity that depends on $\{\theta_0, \theta, h\}$. Specifically, for a predefined grid $\left\{\theta_0^{(s)}, \theta^{(s)}, h^{(s)}\right\}$ and a quantity $Q(\theta_0, \theta, h)$, we add up the variational estimates $\widetilde{w}(\theta_0^{(s)}, \theta^{(s)}, h^{(s)})$ over successively wider intervals of $Q(\theta_0, \theta, h)$, beginning at the posterior mean, until the sum of $\widetilde{w}(\theta_0^{(s)}, \theta^{(s)}, h^{(s)})$ reaches a given interval coverage (e.g. 0.95).

E.7 A modified variational algorithm that estimates θ_0

When performing the genome-wide multiple-SNP analysis under the baseline hypothesis ($\theta = 0$), we find that the default variational algorithm occasionally produces inconsistent posterior results [Supplementary Figure 19 of Zhu and Stephens (2017b)] on a few datasets [e.g. triglyceride data (Teslovich et al., 2010)]. This is most likely due to relatively poor estimates of variational lower bound $F(\mathbf{D}, \theta_0, \theta = 0, h)$ and hyper-parameter θ_0.

To address this issue, we develop a modified variational algorithm that jointly estimates β and θ_0. The modified algorithm is almost the same as the default algorithm, except that there is an additional step of updating θ_0 after the coordinates descent updates for $q^\star$:

$$\theta_0^\star = \arg\max_{\theta_0} F(\mathbf{D}, \theta_0, \theta = 0, h). \tag{E.26}$$

Straightforward algebra yields that

$$\theta_0^\star = \log_{10}\left(\sum_{j=1}^{p} \alpha_j^\star\right) - \log_{10}\left(p - \sum_{j=1}^{p} \alpha_j^\star\right). \tag{E.27}$$

Instead of using the fixed value θ_0 (as the default algorithm), we now use the updated value $\theta_0^\star$ in the next iteration of updating $q^\star$.

We test this approach on the triglyceride data (Teslovich et al., 2010), and obtain im-

proved estimates of variational lower bound and θ_0; see Supplementary Figure 20 of Zhu and Stephens (2017b).

E.8 Scaling computation to many gene sets

When performing genome-wide enrichment analysis of thousands of gene sets in the present study, we exploit a simplification introduced by Carbonetto and Stephens (2013), which allows us to reuse "expensive" whole-genome calculations and thus reduce computing time. Specifically, we assume that SNPs that are not near any member gene of the enriched gene set (i.e. "outside") are unaffected by the inferred enrichment *a posteriori*:

$$q^\star(\beta_{\bar{A}};\mathbf{D},\theta_0,\theta,h) = q^\star(\beta_{\bar{A}};\mathbf{D},\theta_0,\theta=0,h), \tag{E.28}$$

where $q^\star$ is the estimated variational posterior distribution of β, A is the set of SNPs assigned to the enriched gene set (i.e. "inside"), and $\bar{A}$ is complement of A. Since the outside set $\bar{A}$ typically contains most of SNPs in the whole genome, we only need to re-estimate the variational posterior distribution for a relatively small number of "inside" SNPs under assumption (E.28).

Following Carbonetto and Stephens (2013), we approximate variational lower bound as follows:

$$F(\mathbf{D},\theta_0,\theta,h) \approx F(\mathbf{D},\theta_0,\theta=0,h) + F_A(\mathbf{D}_A,\theta_0,\theta,h) - F_A(\mathbf{D}_A,\theta_0,\theta=0,h), \tag{E.29}$$

where for each set $I \in \{A, \bar{A}\}$,

$$\begin{aligned} F_I(\mathbf{D}_I, \theta_0, \theta, h) &= \widehat{\beta}_I^\intercal S_I^{-2} \mathrm{E}_q(\beta_I) - \frac{1}{2}(\mathrm{E}_q(\beta_I))^\intercal S_I^{-1} R_I S_I^{-1} \mathrm{E}_q(\beta_I) - \frac{1}{2}\sum_{j\in I}\frac{\mathrm{Var}_q(\beta_j)}{s_j^2} \\ &\quad - \sum_{j\in I}\alpha_j \log\left(\frac{\alpha_j}{\pi_j}\right) - \sum_{j\in I}(1-\alpha_j)\log\left(\frac{1-\alpha_j}{1-\pi_j}\right) \\ &\quad + \sum_{j\in I}\frac{\alpha_j}{2}\left[1 + \log\left(\frac{\sigma_j^2}{\sigma_\beta^2}\right) - \frac{\sigma_j^2+\mu_j^2}{\sigma_\beta^2}\right], \end{aligned} \tag{E.30}$$

$\mathbf{D}_I := \{\widehat{\beta}_I, S_I, R_I, \mathbf{a}_I\}$, $S_I := \mathrm{diag}(\mathbf{s}_I)$, $\widehat{\beta}_I$ and $\mathbf{s}_I$ are the vectors of single-SNP effect size estimates and corresponding standard errors that are restricted to SNPs in the set I, and R_I is the LD matrix of SNPs in the set I. Note that $F(\mathbf{D}, \theta_0, \theta = 0, h)$ has been previously computed when fitting the baseline model on whole-genome summary data ($\theta = 0$). Hence, calculation of (E.29) only requires re-fitting the baseline and enrichment model on a relatively small dataset $\mathbf{D}_A$ to obtain the corresponding lower bounds $F_A(\mathbf{D}_A, \theta_0, \theta = 0, h)$ and $F_A(\mathbf{D}_A, \theta_0, \theta, h)$, respectively.

E.9 Parallel implementation

To speed up the whole genome analysis, we implement the step of estimating $p(\beta, \gamma | \mathbf{D}, \theta_0, \theta, h)$ in parallel across multiple threads.

Our parallel implementation is built on the key property that $R_{ij} = 0$ if SNPs i and j are on different chromosomes. As a result, the coordinate descent updates (E.5)-(E.7) for the variational parameters $\{\alpha_j, \mu_j, \sigma_j\}$ of SNP j on Chromosome c only requires $\mathbf{D}_c$, where $\mathbf{D}_c := \{\widehat{\beta}_c, S_c, R_c, \mathbf{a}_c\}$ denotes the input data from Chromosome c. Further, the variational lower bound $F(\mathbf{D}, \theta_0, \theta, h)$ based on whole genome data has the following decomposition:

$$F(\mathbf{D}, \theta_0, \theta, h) = \sum_{c=1}^{22} F_c(\mathbf{D}_c, \theta_0, \theta, h), \tag{E.31}$$

where F_c is defined in (E.30).

Algorithm 1 outlines the parallel implementation. First, we partition the whole-genome input data $\mathbf{D}$ into 22 sub-data $\{\mathbf{D}_c\}$ by chromosomes. We then request 22 threads from a single computer, each of which is responsible for updating the variational parameters $\{\alpha_j, \mu_j, \sigma_j\}$ and computing the variational lower bound $F_c(\mathbf{D}_c, \theta_0, \theta, h)$ on each chromosome c; see Step 6 of Algorithm 1. Finally, we aggregate the per-chromosome updated variational parameters and lower bounds in the "reduce" step; see Step 8 of Algorithm 1.

Algorithm 1 Parallel implementation

for $s = 1$ to N **do** ▷ outer loop

  initialize $\boldsymbol{\alpha}^{(s)}$ and $\boldsymbol{\mu}^{(s)}$ randomly

  compute $\boldsymbol{\sigma}^{(s)}$ by (E.5)

  repeat ▷ inner loop

    for Chromosome $c = 1$ to 22 **do** ▷ parallel step

      use sub-data $\mathbf{D}_c$ to update $\boldsymbol{\alpha}^{(c,s)}$, $\boldsymbol{\mu}^{(c,s)}$ and $F_c(\mathbf{D}_c, \theta_0^{(s)}, \theta^{(s)}, h^{(s)})$ by (E.7), (E.6) and (E.30)

    end for

    aggregate chromosome-level results: ▷ reduce step

$$\boldsymbol{\alpha}^{(s)} = [\boldsymbol{\alpha}^{(1,s)}; \ldots; \boldsymbol{\alpha}^{(22,s)}] \tag{E.32}$$

$$\boldsymbol{\mu}^{(s)} = [\boldsymbol{\mu}^{(1,s)}; \ldots; \boldsymbol{\mu}^{(22,s)}] \tag{E.33}$$

$$F(\mathbf{D}, \theta_0^{(s)}, \theta^{(s)}, h^{(s)}) = \sum_{c=1}^{22} F_c(\mathbf{D}_c, \theta_0^{(s)}, \theta^{(s)}, h^{(s)}) \tag{E.34}$$

  until convergence criteria are met

  compute the posterior weight $\widetilde{w}(\theta_0^{(s)}, \theta^{(s)}, h^{(s)})$ by (E.13)

end for

integrate out hyper-parameters $\{\theta_0, \theta, h\}$:

$$\widetilde{\boldsymbol{\alpha}} = \sum_{s=1}^{N} \widetilde{w}(\theta_0^{(s)}, \theta^{(s)}, h^{(s)}) \cdot \boldsymbol{\alpha}^{(s)} \tag{E.35}$$

$$\widetilde{\boldsymbol{\mu}} = \sum_{s=1}^{N} \widetilde{w}(\theta_0^{(s)}, \theta^{(s)}, h^{(s)}) \cdot \boldsymbol{\mu}^{(s)} \tag{E.36}$$

$$\widetilde{\boldsymbol{\sigma}} = \sum_{s=1}^{N} \widetilde{w}(\theta_0^{(s)}, \theta^{(s)}, h^{(s)}) \cdot \boldsymbol{\sigma}^{(s)} \tag{E.37}$$

E.10 Connection with variational inference based on full data

Zhu and Stephens (2017a) derived the conditions under which multiple regression likeli-

hood based on individual-level data is equivalent to multiple regression likelihood based on summary-level data. Under the same conditions, here we show that variational inferences based on individual-level data (Carbonetto and Stephens, 2012) and summary-level data (Zhu and Stephens, 2017a) are also equivalent.

Proposition E.10.1. Let $\hat{\sigma}_y^2$ denote the sample variance of individual-level phenotypes $\mathbf{y}$, $\widehat{R}^{\mathsf{sam}}$ denote the sample correlation matrix of individual-level genotypes X, and σ^2 denote the residual variance in the multiple linear regression model. If $\sigma^2 = \hat{\sigma}_y^2$ and $R = \widehat{R}^{\mathsf{sam}}$, the coordinate descent equations (E.5)-(E.7) based on summary-level data yields the *same* solution of $\{\alpha_j, \mu_j, \sigma_j\}$ as the equations based on individual-level data [Equation 8-10 in Carbonetto and Stephens (2012)].

Proof. Following the notation in Carbonetto and Stephens (2012), we write $\sigma_\beta^2 = \sigma_a^2\sigma^2$, and write the coordinate updates of $\{\alpha_j, \mu_j, \sigma_j\}$ based on individual-level data [Equation 8-10 in Carbonetto and Stephens (2012)] as follows:

$$\sigma_j^2 = \frac{\sigma^2}{X_j^\top X_j + \sigma_a^{-2}}, \tag{E.38}$$

$$\mu_j = \frac{\sigma_j^2}{\sigma^2}\left(X_j^\top \mathbf{y} - \sum_{i \neq j} X_j^\top X_i \alpha_i \mu_i\right), \tag{E.39}$$

$$\frac{\alpha_j}{1-\alpha_j} = \frac{\pi_j}{1-\pi_j} \cdot \frac{\sigma_j}{\sigma_a \sigma} \cdot \exp\left\{\frac{\mu_j^2}{2\sigma_j^2}\right\}. \tag{E.40}$$

Based on the definition of $\widehat{\beta}$, S and $\widehat{R}^{\mathsf{sam}}$, we have:

$$X_j^\top \mathbf{y} = (X_j^\top X_j)\hat{\beta}_j, \;\; X_j^\top X_j = s_j^{-2}\hat{\sigma}_y^2, \;\; X_j^\top X_i = ||X_j|| \cdot ||X_i|| \cdot \widehat{R}_{ij}^{\mathsf{sam}}, \tag{E.41}$$

and then we can rewrite the updates above as follows:

$$\sigma_j^2 = \frac{\sigma^2}{s_j^{-2}\hat{\sigma}_y^2 + \sigma_a^{-2}}, \tag{E.42}$$

$$\mu_j = \frac{\sigma_j^2\hat{\sigma}_y^2}{\sigma^2}\left(\frac{\hat{\beta}_j}{s_j^2} - \sum_{i\neq j}\frac{\widehat{R}_{ij}^{\mathsf{sam}}\alpha_i\mu_i}{s_i s_j}\right), \tag{E.43}$$

$$\frac{\alpha_j}{1-\alpha_j} = \frac{\pi_j}{1-\pi_j}\cdot\frac{\sigma_j}{\sigma_a\sigma}\cdot\exp\left\{\frac{\mu_j^2}{2\sigma_j^2}\right\}. \tag{E.44}$$

Finally, if $\sigma^2 = \hat{\sigma}_y^2$ and $R_{ij} = \widehat{R}_{ij}^{\mathsf{sam}}$, we reproduce the coordinate descent updates (E.5), (E.6) and (E.7) that are based on summary-level data. □

Under the same conditions ($\sigma^2 = \hat{\sigma}_y^2$ and $R = \widehat{R}^{\mathsf{sam}}$), we can also show that the variational lower bound based on individual-level data and summary-level data are equivalent. The proof is similar to the previous one, so it is omitted here.

Proposition E.10.2. If $\sigma^2 = \hat{\sigma}_y^2$ and $R = \widehat{R}^{\mathsf{sam}}$, then the difference between the variational lower bound (E.10) based on summary-level data and individual-level data [Equation 13 in Carbonetto and Stephens (2012)] is a constant with respect to the variational parameters $\{\alpha_j, \mu_j, \sigma_j\}$.

E.11 Acknowledgements

We thank all the GWAS consortia for making their summary statistics publicly available. We also thank the GTEx consortium for making their RNA-sequencing data publicly available. A full list of acknowledgements appear as follows.

- **Genetic Investigation of ANthropometric Traits (GIANT) Consortium**
  Data on adult human height (Wood et al., 2014), body mass index (Locke et al., 2015) and body fat distribution (Shungin et al., 2015) have been contributed by GIANT investigators and have been downloaded from

http://portals.broadinstitute.org/collaboration/giant.

- **Psychiatric Genomics Consortium (PGC)**

  Data on schizophrenia (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014) have been contributed by PGC investigators and have been downloaded from http://www.med.unc.edu/pgc.

- **International Inflammatory Bowel Disease Genetics Consortium (IIBDGC)**

  Data on inflammatory bowel disease (Liu et al., 2015), including Crohn's disease and ulcerative colitis have been contributed by IIBDGC investigators and have been downloaded from https://www.ibdgenetics.org.

- **Coronary ARtery DIsease Genome wide Replication and Meta-analysis (CARDIoGRAM) plus The Coronary Artery Disease (C4D) Genetics (CARDIoGRAMplusC4D) Consortium**

  Data on coronary artery disease and myocardial infarction (Nikpay et al., 2015) have been contributed by CARDIoGRAMplusC4D investigators and have been downloaded from http://www.cardiogramplusc4d.org.

- **International Genomics of Alzheimer's Project (IGAP)**

  Data on Alzheimer's disease (Lambert et al., 2013) have been contributed by IGAP investigators and have been downloaded from http://web.pasteur-lille.fr/en/recherche/u744/igap/igap_download.php. We thank the International Genomics of Alzheimer's Project (IGAP) for providing summary results data for these analyses. The investigators within IGAP contributed to the design and implementation of IGAP and/or provided data but did not participate

in analysis or writing of this report. IGAP was made possible by the generous participation of the control subjects, the patients, and their families. The i-Select chips was funded by the French National Foundation on Alzheimer's disease and related disorders. EADI was supported by the LABEX (laboratory of excellence program investment for the future) DISTALZ grant, Inserm, Institut Pasteur de Lille, Université de Lille 2 and the Lille University Hospital. GERAD was supported by the Medical Research Council (Grant no. 503480), Alzheimer's Research UK (Grant no. 503176), the Wellcome Trust (Grant no. 082604/2/07/Z) and German Federal Ministry of Education and Research (BMBF): Competence Network Dementia (CND) grant no. 01GI0102, 01GI0711, 01GI0420. CHARGE was partly supported by the NIH/NIA grant R01 AG033193 and the NIA AG081220 and AGES contract N01-AG-12100, the NHLBI grant R01 HL105756, the Icelandic Heart Association, and the Erasmus Medical Center and Erasmus University. ADGC was supported by the NIH/NIA grants: U01 AG032984, U24 AG021886, U01 AG016976, and the Alzheimer's Association grant ADGC-10-196728.

- **Social Science Genetic Association Consortium (SSGAC)**

  Data on neuroticism and depressive symptoms (Okbay et al., 2016) have been contributed by SSGAC investigators and have been downloaded from `https://www.thessgac.org`. For financial support, the SSGAC thanks the U.S. National Science Foundation, the U.S. National Institutes of Health (National Institute on Aging, and the Office for Behavioral and Social Science Research), the Ragnar Söderberg Foundation, the Swedish Research Council, The Jan Wallander and Tom Hedelius Foundation, the European Research Council, and the Pershing Square Fund of the Foundations of Human Behavior.

- **GWAS summary statistics of rheumatoid arthritis (Okada et al., 2014)**

  Data on rheumatoid arthritis have been contributed by authors of Okada et al. (2014) and have been downloaded from `http://plaza.umin.ac.jp/yokada/datasource/software.htm`.

- **DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium**

  Data on type 2 diabetes (Morris et al., 2012) have been contributed by DIAGRAM investigators and have been downloaded from `http://www.diagram-consortium.org`.

- **Reproductive Genetics (ReproGen) Consortium**

  Data on age at natural menopause (Day et al., 2015) have been contributed by ReproGen investigators and have been downloaded from `http://www.reprogen.org`.

- **Global Urate Genetics Consortium (GUGC)**

  Data on serum urate concentrations and gout (Köttgen et al., 2013) have been contributed by GUGC investigators and have been downloaded from `http://metabolomics.helmholtz-muenchen.de/gugc`.

- **Global Lipids Genetics Consortium (GLGC)**

  Data on blood lipids (Teslovich et al., 2010), including levels of total cholesterol, low-density lipoprotein cholesterol, high-density lipoprotein cholesterol and triglycerides have been contributed by GLGC investigators and have been downloaded from `http://csg.sph.umich.edu//abecasis/public/lipids2010`.

- **Meta-Analyses of Glucose and Insulin-related traits Consortium (MAGIC)**

Data on glycaemic traits (Manning et al., 2012), including fasting glucose and insulin results accounting for body mass index, have been contributed by MAGIC investigators and have been downloaded from https://www.magicinvestigators.org.

- **Project MinE**

  Data on amyotrophic lateral sclerosis (van Rheenen et al., 2016) have been contributed by Project MinE and have been downloaded from http://databrowser.projectmine.com.

- **GWAS summary statistics of six red blood cell phenotypes (van der Harst et al., 2012)**

  Data on six red blood cell phenotypes have been contributed by authors of van der Harst et al. (2012) and have been downloaded from the European Genome-Phenome Archive (http://www.ebi.ac.uk/ega) under accession number EGAS00000000132.

- **Genotype-Tissue Expression (GTEx) Project**

  The GTEx Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from the GTEx Portal (https://gtexportal.org) on November 21, 2016.

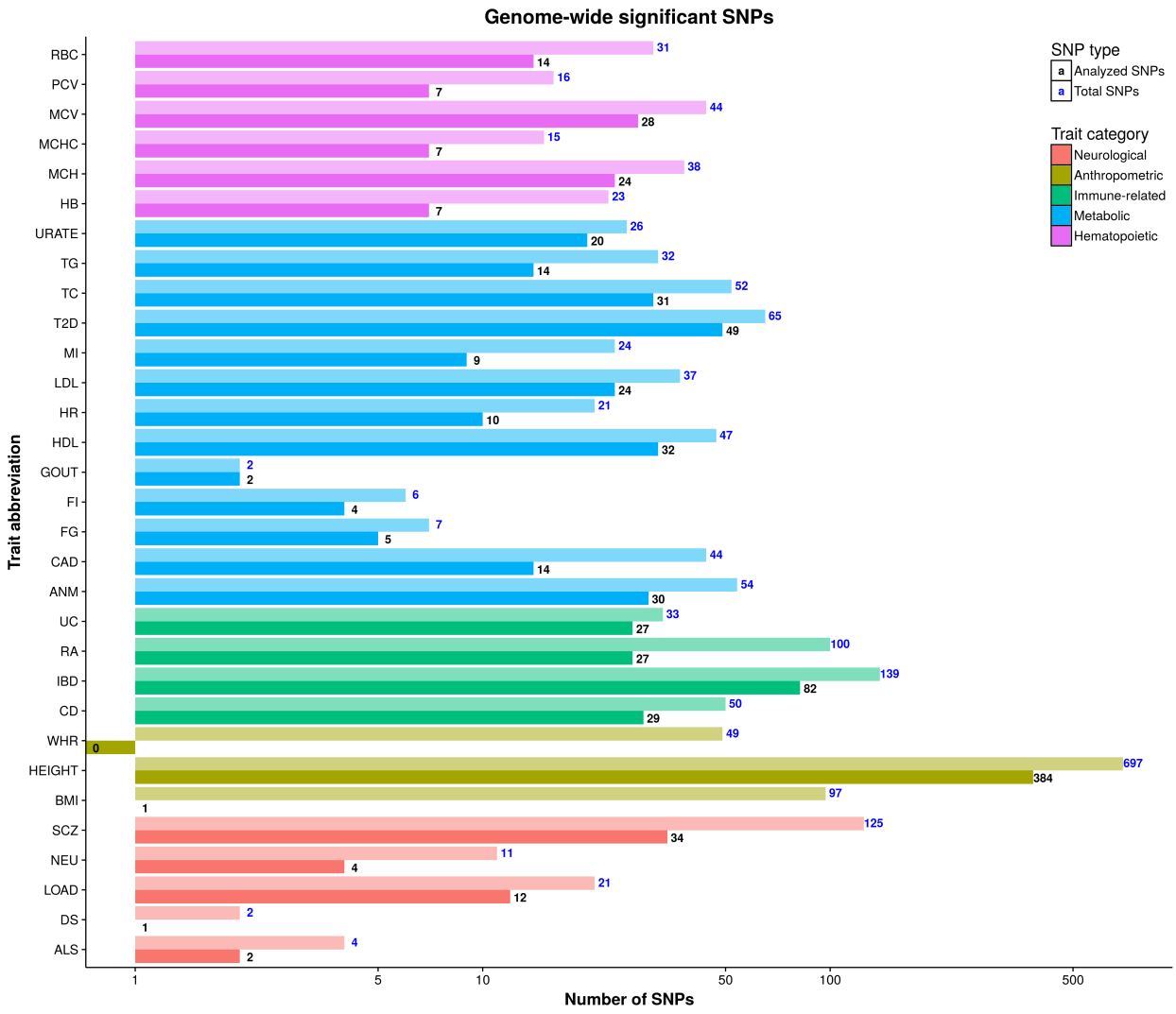
APPENDIX F

SUPPLEMENTARY FIGURES OF ZHU AND STEPHENS (2017B)

Supplementary Figure 1

Summary of genetic variants in GWAS of 31 human phenotypes. Panel **(a)** reports numbers of GWAS hits (i.e. loci or SNPs reaching genome-wide significance) reported in corresponding publications. Panel **(b)** reports numbers of all genetic variants available in corresponding GWAS. For both panels, "total SNPs" denote SNPs that were available in corresponding publications and/or summary data files (bar charts with higher transparency; blue numbers); "analyzed SNPs" denote SNPs analyzed in the present study (bar charts with lower transparency; black numbers).

(a) The "genome-wide significant SNPs" are sentinel SNPs reaching genome-wide significance ($p < 5 \times 10^{-8}$), which are directly retrieved from corresponding publications of GWAS. The horizontal axis uses a logarithmic scale (base 10). Note that the numbers of "analyzed" genome-wide significant SNPs for body mass index (BMI) (Locke et al., 2015) and waist-to-hip ratio (WHR) adjusted for BMI (Shungin et al., 2015) are extremely small. This is because summary data from both studies were combined results of GWAS arrays and custom arrays [Metabochip (Voight et al., 2012)], and we excluded SNPs on custom arrays from our analyses. Custom arrays harbored almost all GWAS hits for these two traits, so there are only zero and one GWAS hit left for WHR and BMI in our analyses.


Genome-wide significant SNPs
SNP type
a Analyzed SNPs
a Total SNPs
Trait category
Neurological
Anthropometric
Immune-related
Metabolic
Hematopoietic
Trait abbreviation
RBC 31 14
PCV 16 7
MCV 44 28
MCHC 15 7
MCH 38 24
HB 23 7
URATE 26 20
TG 32 14
TC 52 31
T2D 65 49
MI 24 9
LDL 37 24
HR 21 10
HDL 47 32
GOUT 2 2
FI 6 4
FG 7 5
CAD 44 14
ANM 54 30
UC 33 27
RA 100 27
IBD 139 82
CD 50 29
WHR 49 0
HEIGHT 697 384
BMI 97 1
SCZ 125 34
NEU 11 4
LOAD 21 12
DS 2 1
ALS 4 2
1
5
10
50
100
500
Number of SNPs

(b) Total numbers of SNPs available in corresponding GWAS.

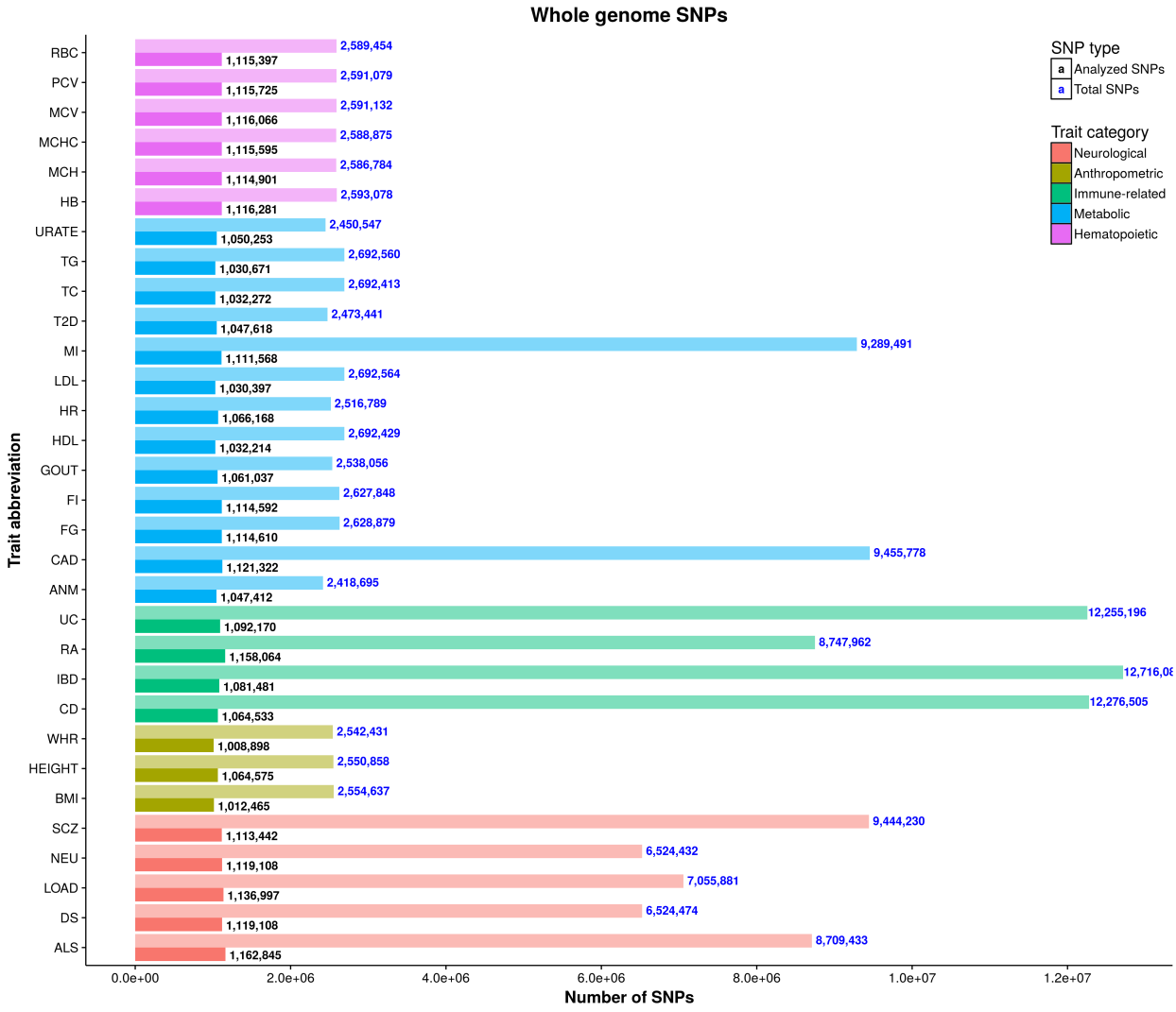


Figure F.1: Summary of genetic variants in GWAS of 31 human phenotypes.

Supplementary Figure 2

Inferred effect size distributions of 31 human phenotypes, assuming that no pathways are enriched. For each trait, its effect size distribution is summarized as the fraction of trait-associated SNPs and the standardized effect size of trait-associated SNPs. See Supplementary Note of Zhu and Stephens (2017b) for details of these two quantities. Each dot represents a trait, where the horizontal point range indicates the posterior mean and 95% credible interval (C.I.) of fraction of trait-associated SNPs, and the vertical point range indicates the posterior mean and 95% C.I. of standardized effect size. Both axes use a logarithmic scale (base 10).

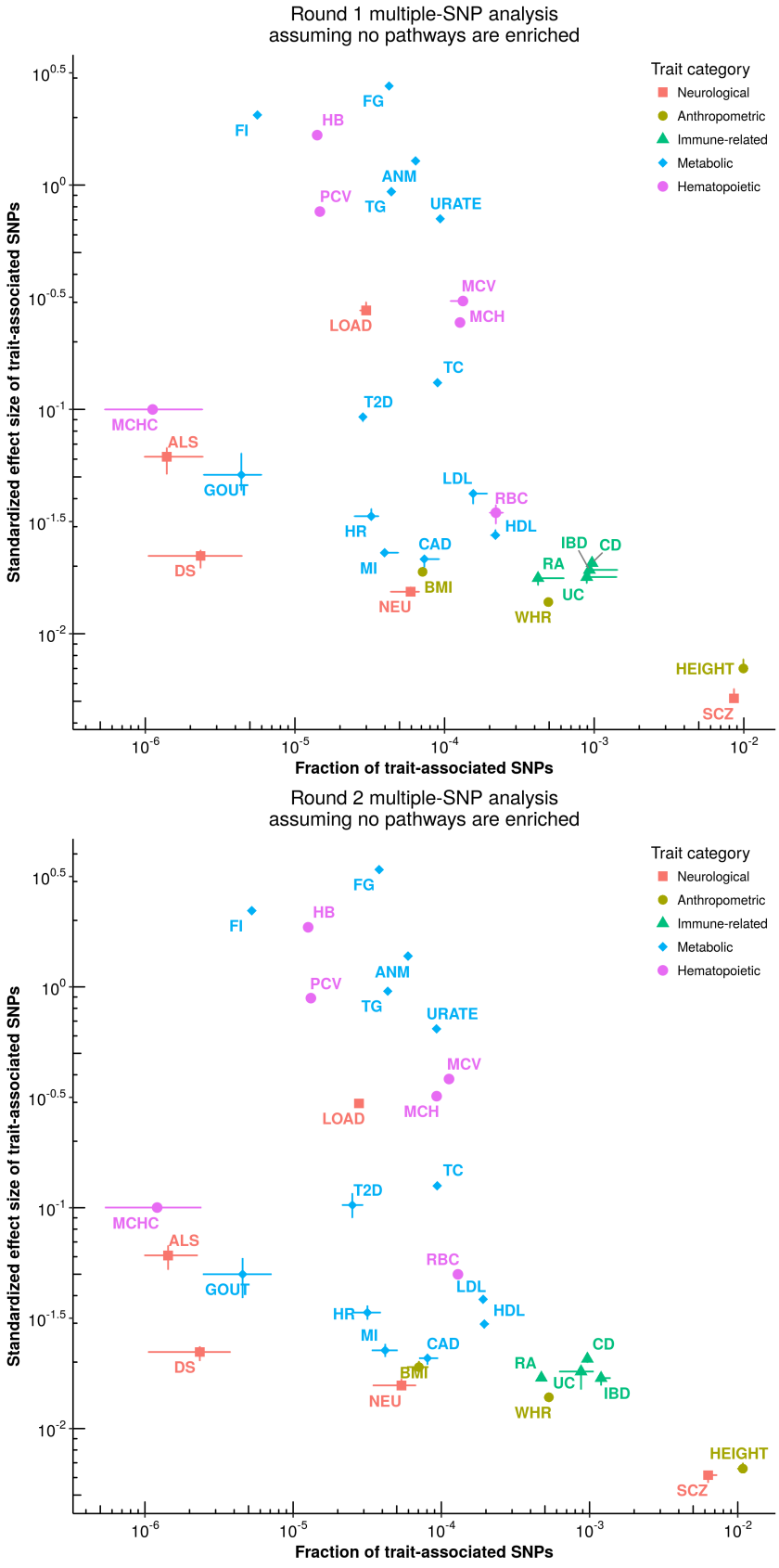


Figure F.2: Inferred effect size distributions of 31 human phenotypes.

Supplementary Figure 3

Ranking similarity between genome-wide multiple-SNP and single-SNP analyses, both assuming that no pathways are enriched. We first divide the entire genome into overlapped loci of 50 SNPs (with an overlap of 25 SNPs between neighboring loci). For each trait and each locus, we then compute i) the maximum single-SNP $|z|$-score; ii) the posterior probability that the locus contains at least one trait-associated SNP (P_1); and iii) the posterior expected number of trait-associated SNPs in the locus (ENS). Based on these three locus-level statistics, we obtain three ranked lists of loci for each trait, and then evaluate their similarity via i) Spearman's ρ statistic; ii) Kendall's τ statistic and iii) rank biased overlap (RBO) (Webber et al., 2010). The Spearman's ρ and Kendall's τ statistic are computed by `R` function `cor`. The RBO is computed by the function `rbo` in `R` package `gespeR` (Schmich, 2015).

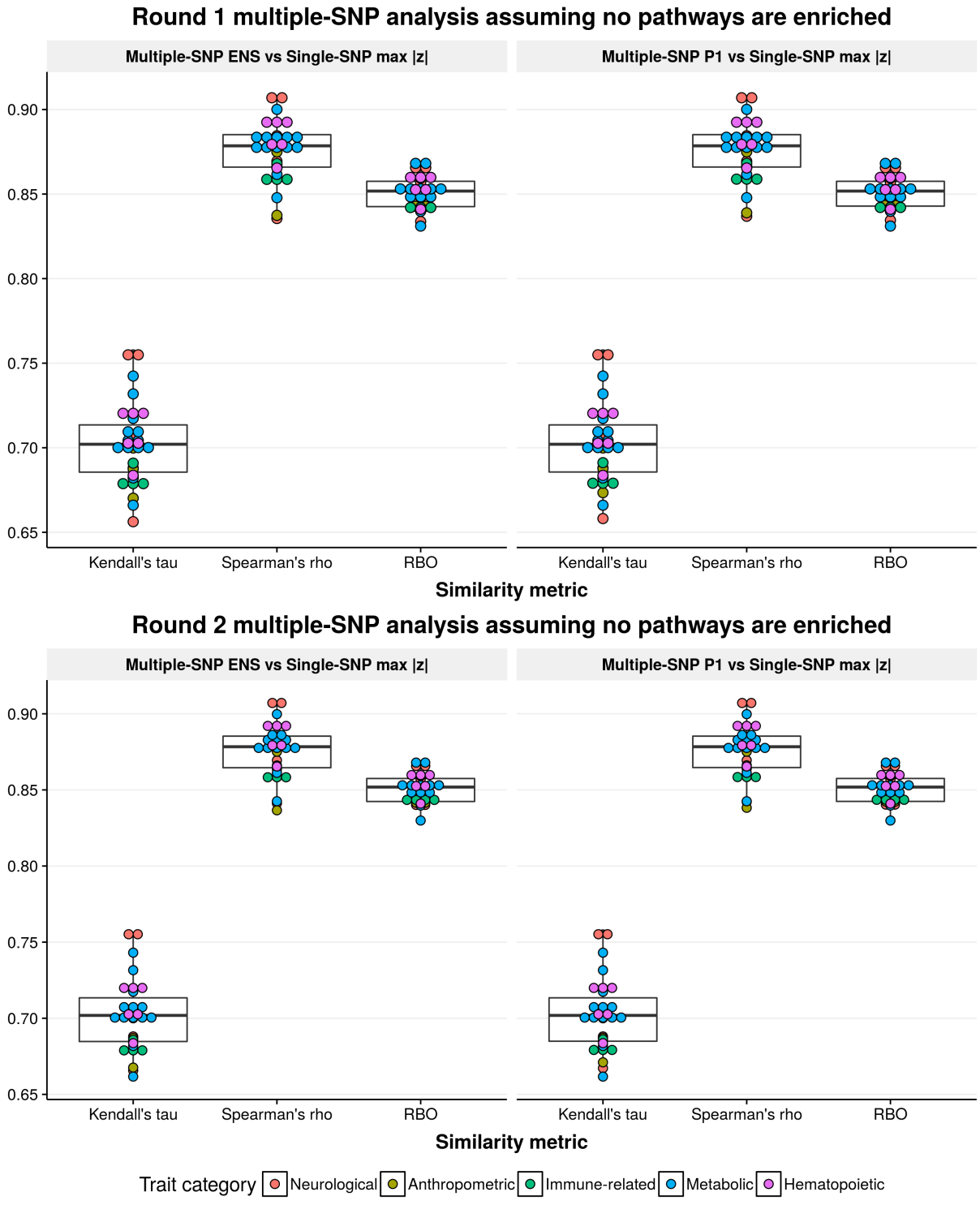
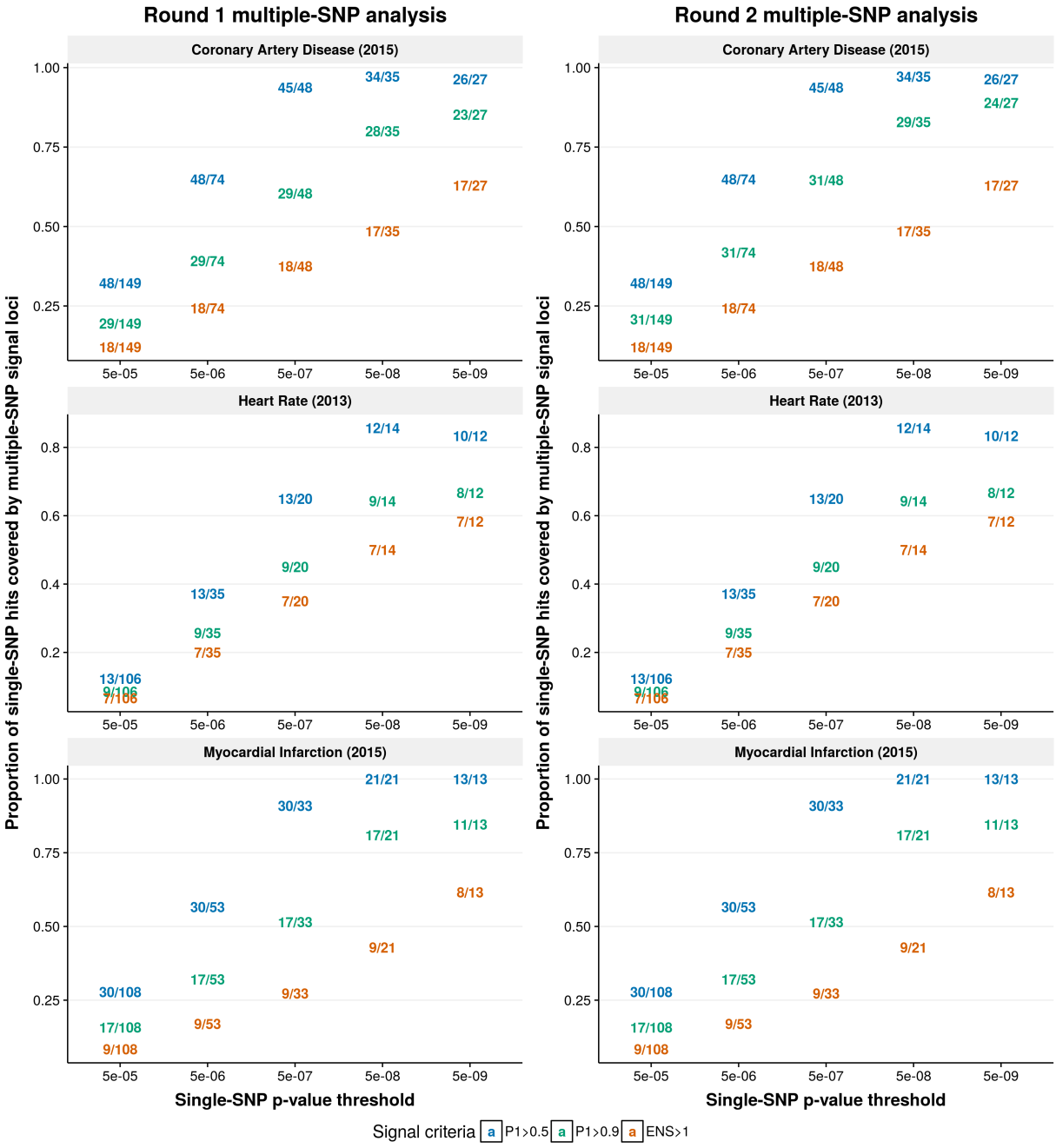


Figure F.3: Ranking similarity between genome-wide multiple-SNP and single-SNP analyses, both assuming that no pathways are enriched.

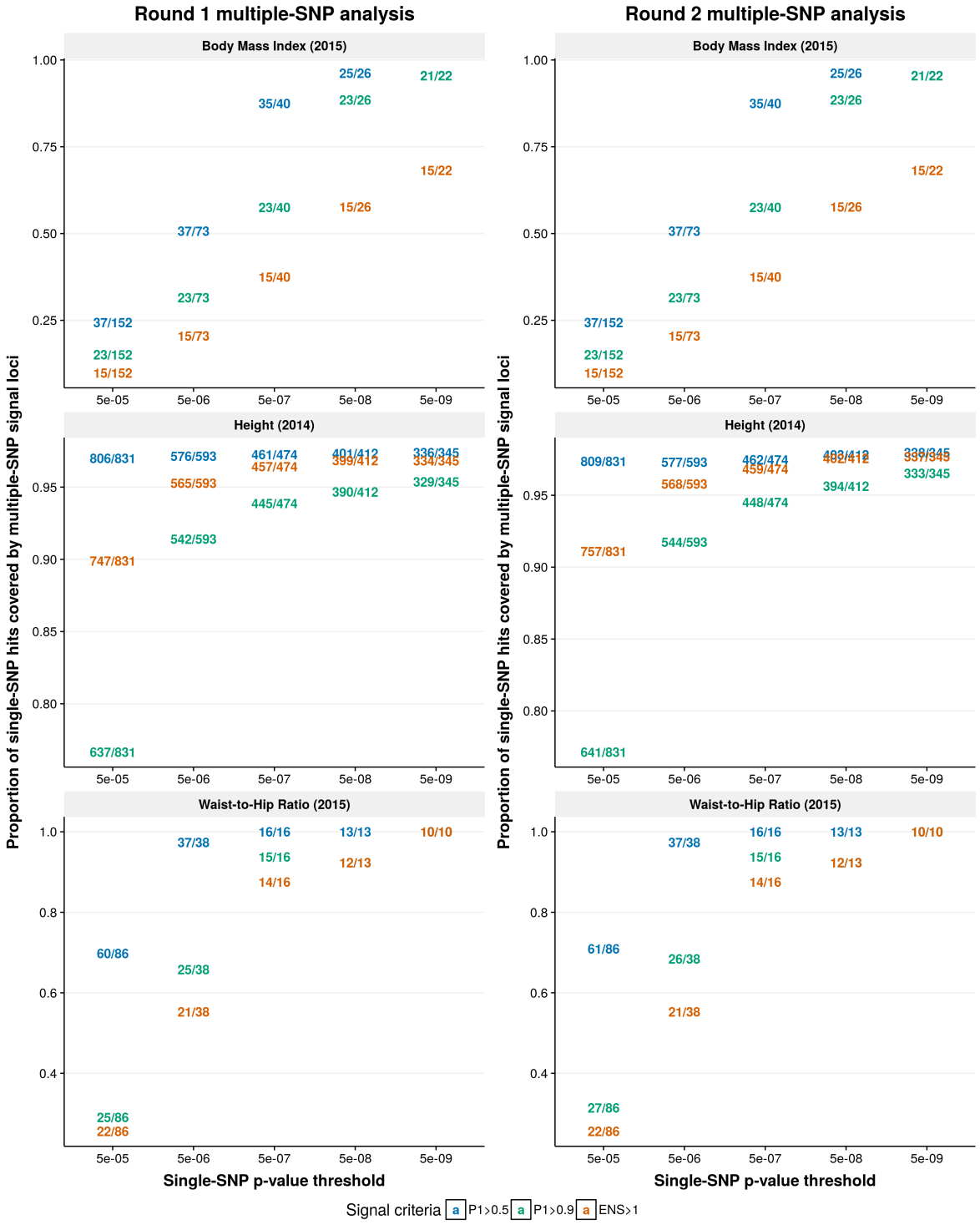
Supplementary Figure 4

Concordance between genome-wide single-SNP and multiple-SNP analyses of 31 phenotypes, both assuming that no pathways are enriched. "Single-SNP hits" are SNPs reaching significance (for a given p-value threshold) and separated by at least 1 Mb. "Multiple-SNP signal loci" are predefined genomic regions satisfying certain criteria (estimated $P_1 > 0.5$, $P_1 > 0.9$ or ENS > 1). For each phenotype, both single-SNP and multiple-SNP analyses are performed on the **same** summary-level data. For a given trait, the concordance between single-SNP and multiple-SNP analyses is measured by the proportion of single-SNP hits covered by multiple-SNP signal loci. See Supplementary Figure 3 of Zhu and Stephens (2017b) for the definition of locus.

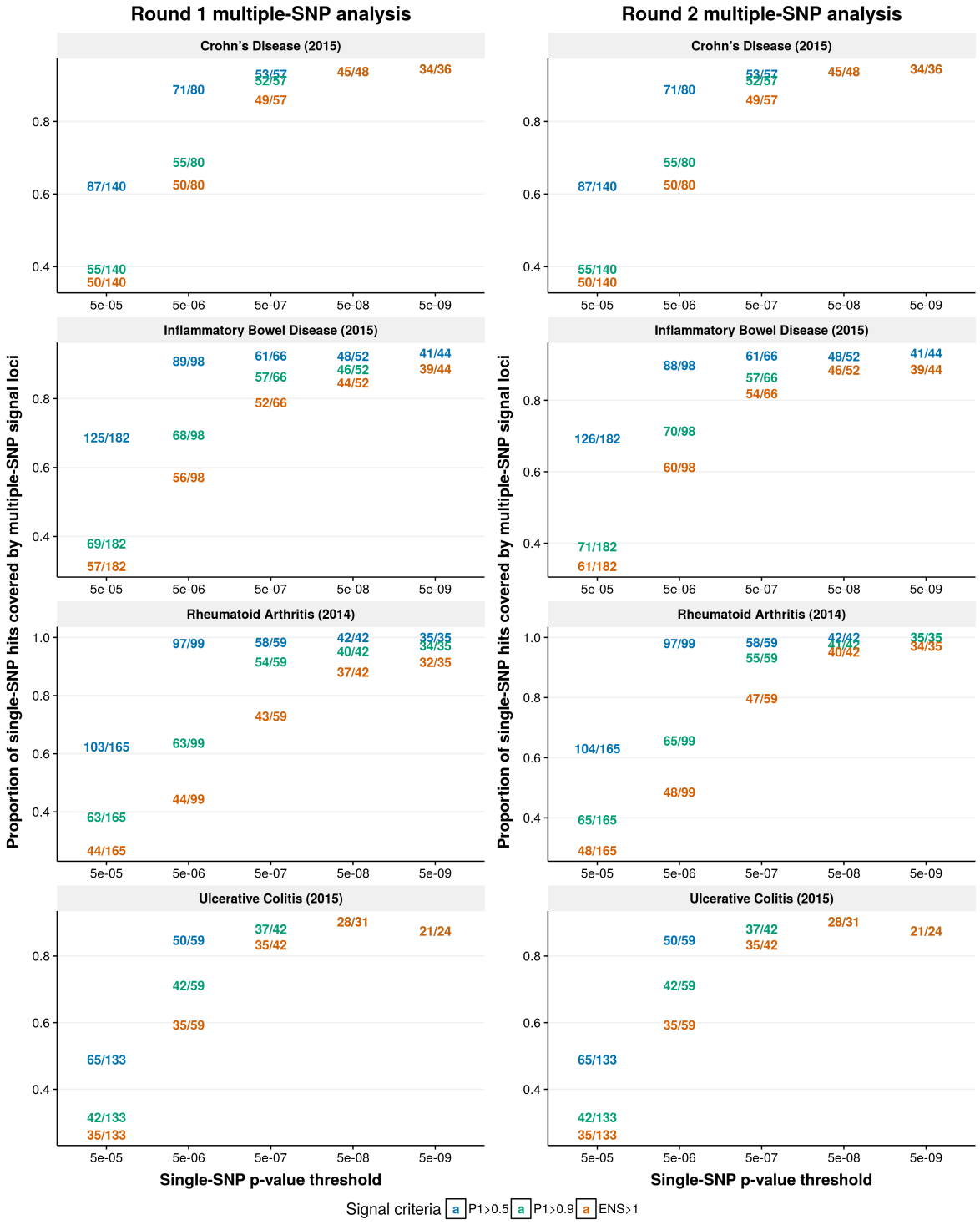
(a) Three heart-related traits: heart rate (Den Hoed et al., 2013), coronary artery disease and myocardial infarction (Nikpay et al., 2015).



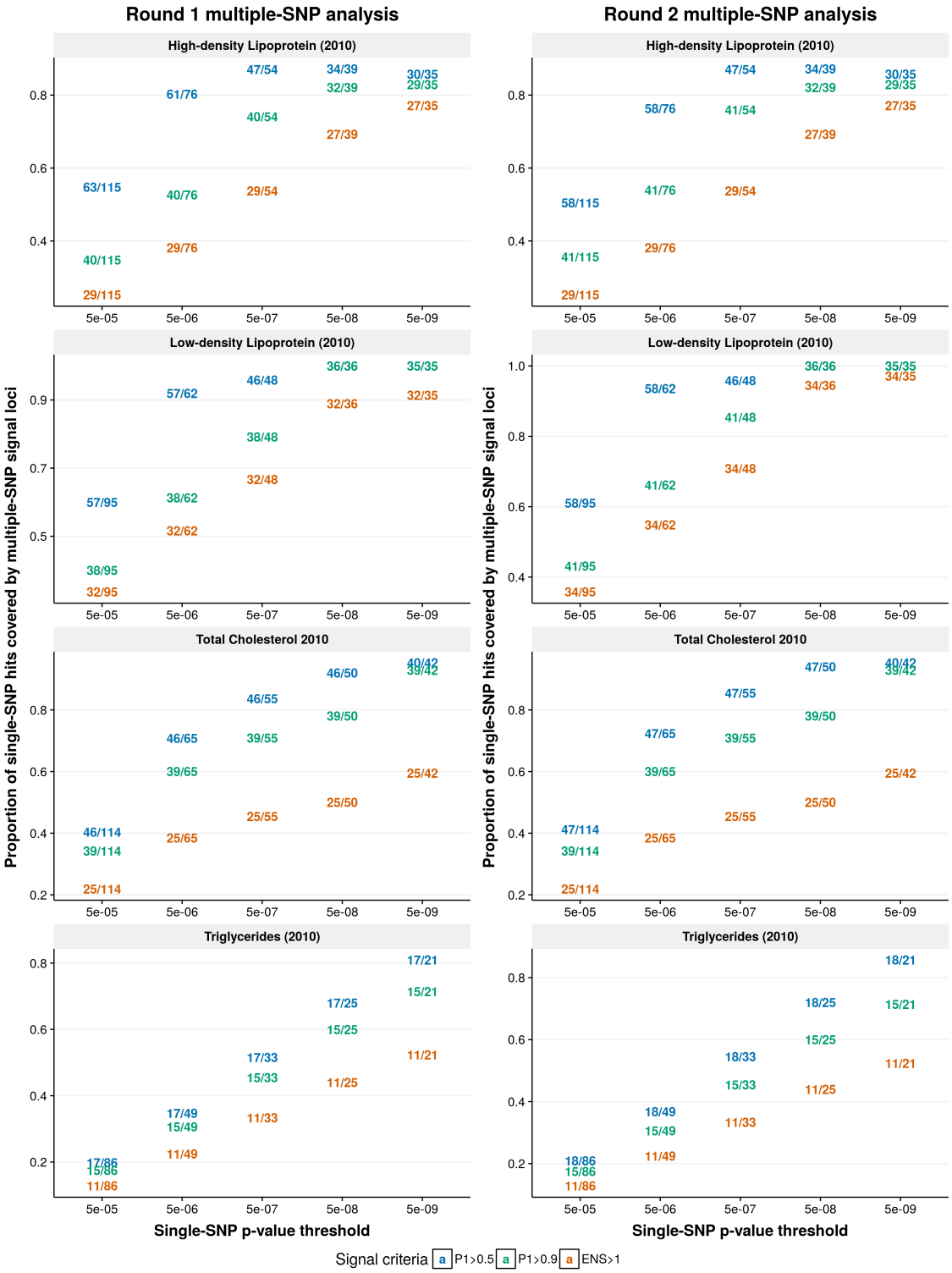
(b) Three anthropometric traits: adult height (Wood et al., 2014), body mass index (Locke et al., 2015) and waist-to-hip ratio after adjusting for body mass index (Shungin et al., 2015).



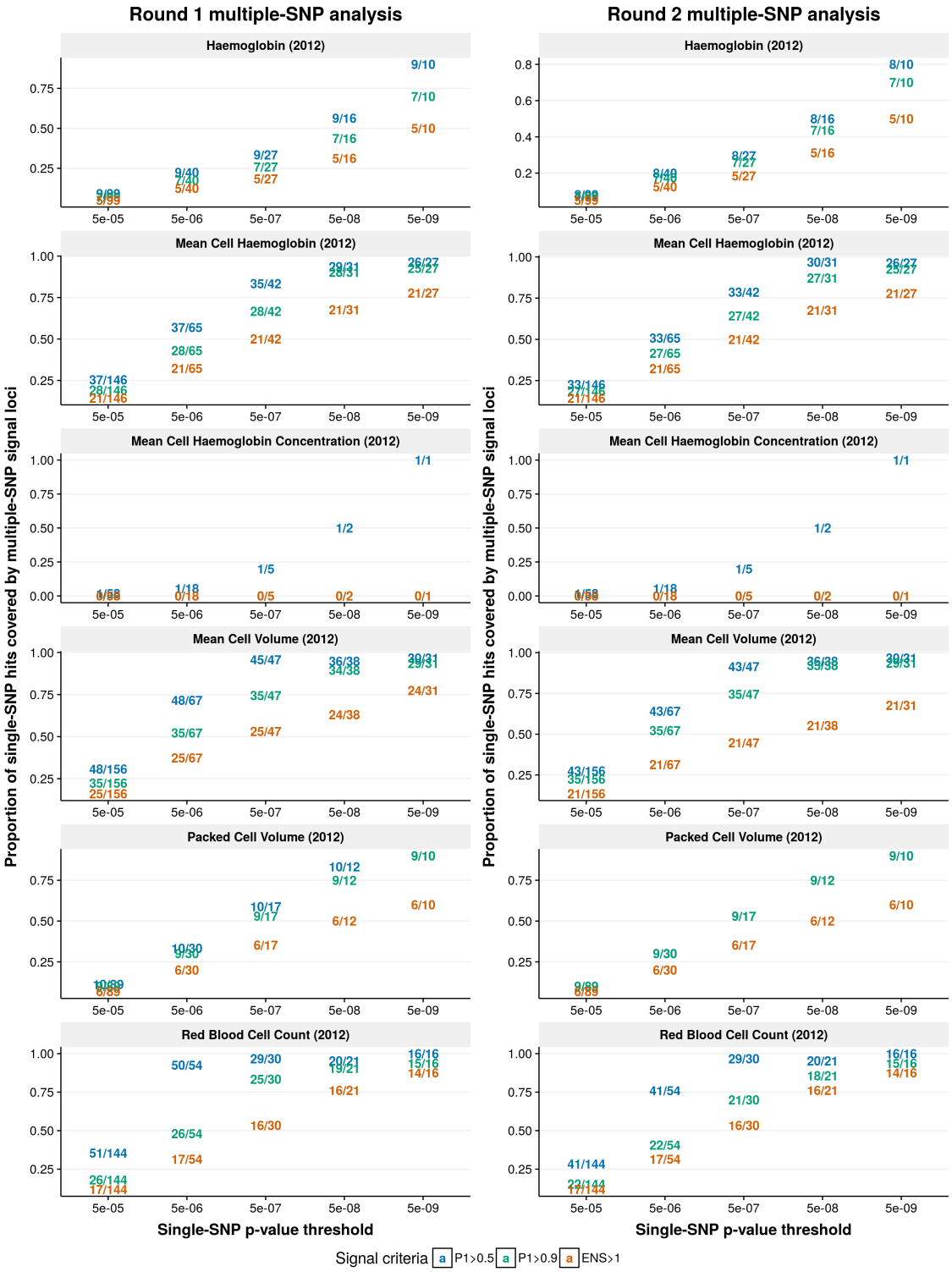
(c) Four immune-related traits: rheumatoid arthritis (Okada et al., 2014), inflammatory bowel disease, Crohn's disease and ulcerative colitis (Liu et al., 2015).



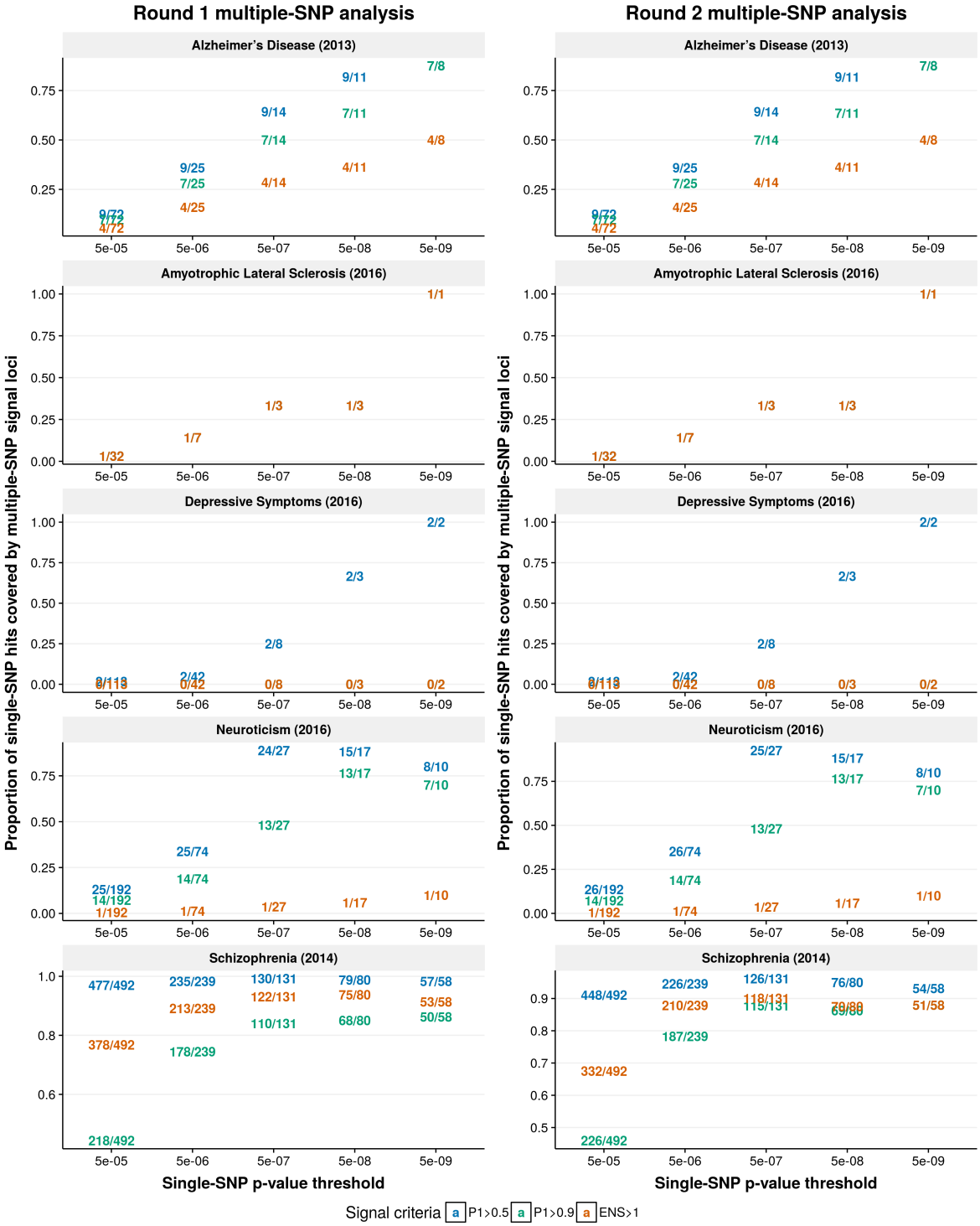
(d) Four blood lipid traits (Teslovich et al., 2010).



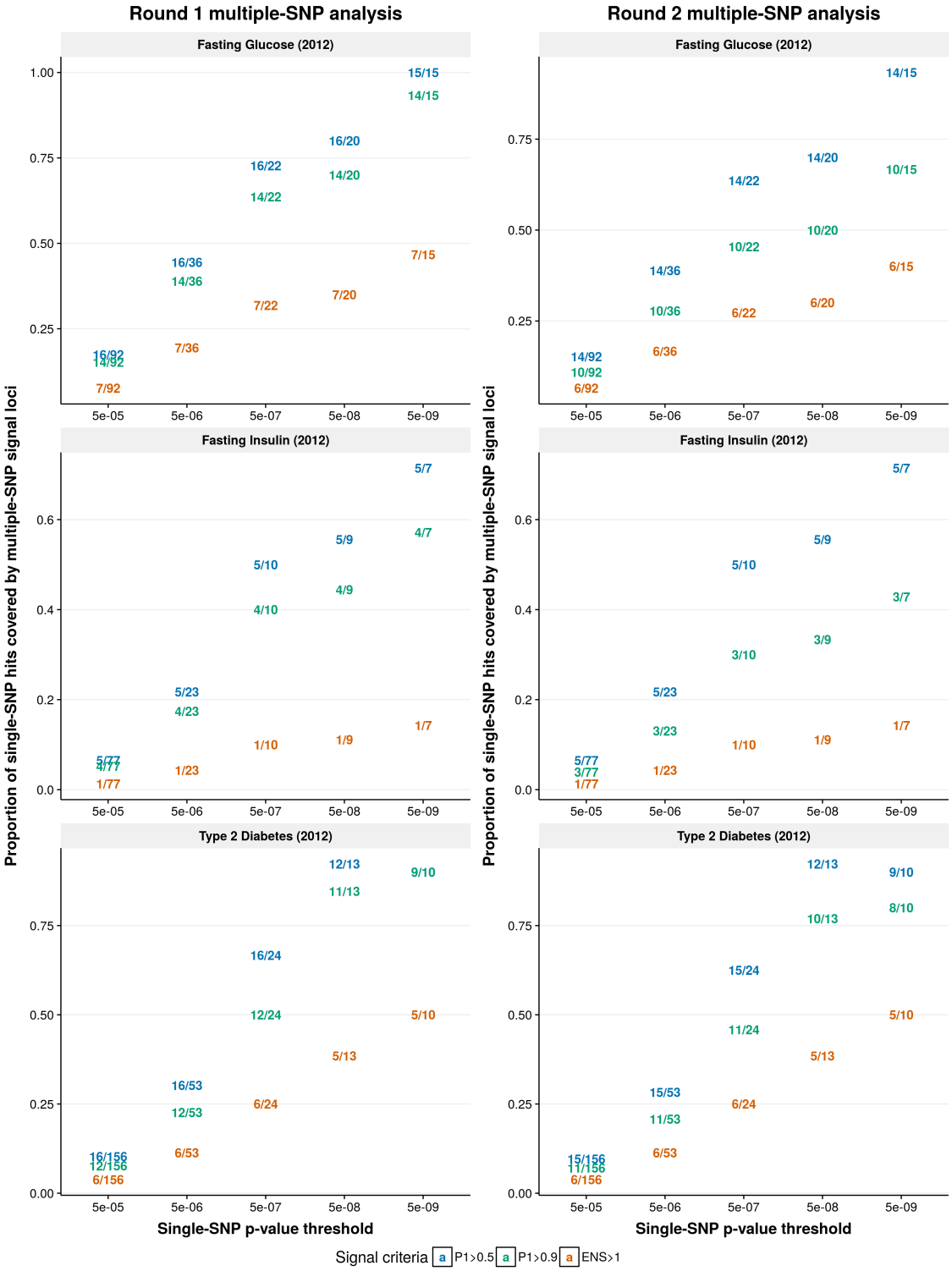
(e) Six red blood cell traits (van der Harst et al., 2012).



(f) Five neurological phenotypes: Alzheimer's disease (Lambert et al., 2013), schizophrenia (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014), Amyotrophic lateral sclerosis (van Rheenen et al., 2016), depressive symptoms and neuroticism (Okbay et al., 2016).



(g) Fasting glucose, fasting insulin (Manning et al., 2012) and type 2 diabetes (Morris et al., 2012).



(h) Serum urate, gout (Köttgen et al., 2013) and age at natural menopause (Day et al., 2015).

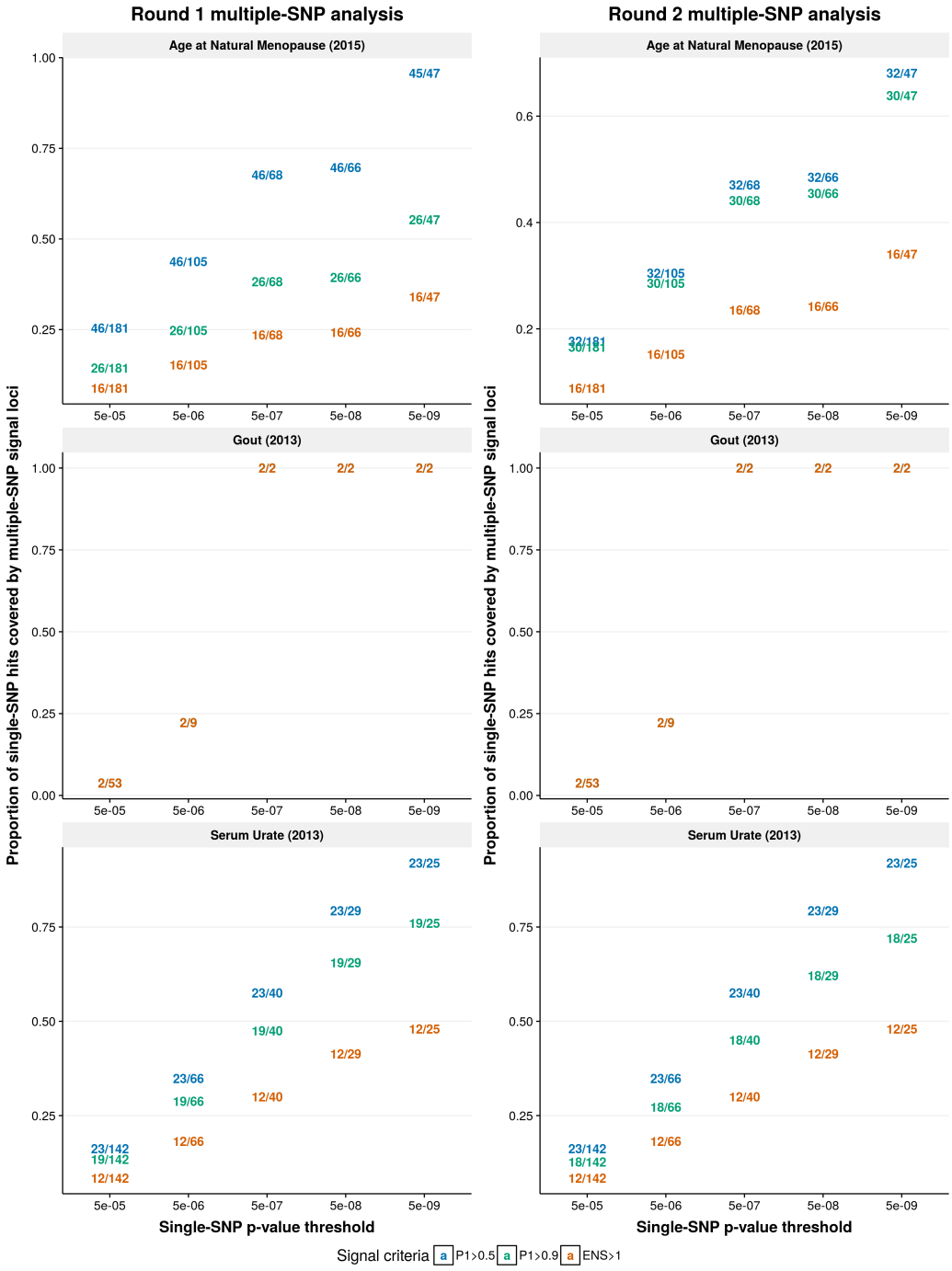


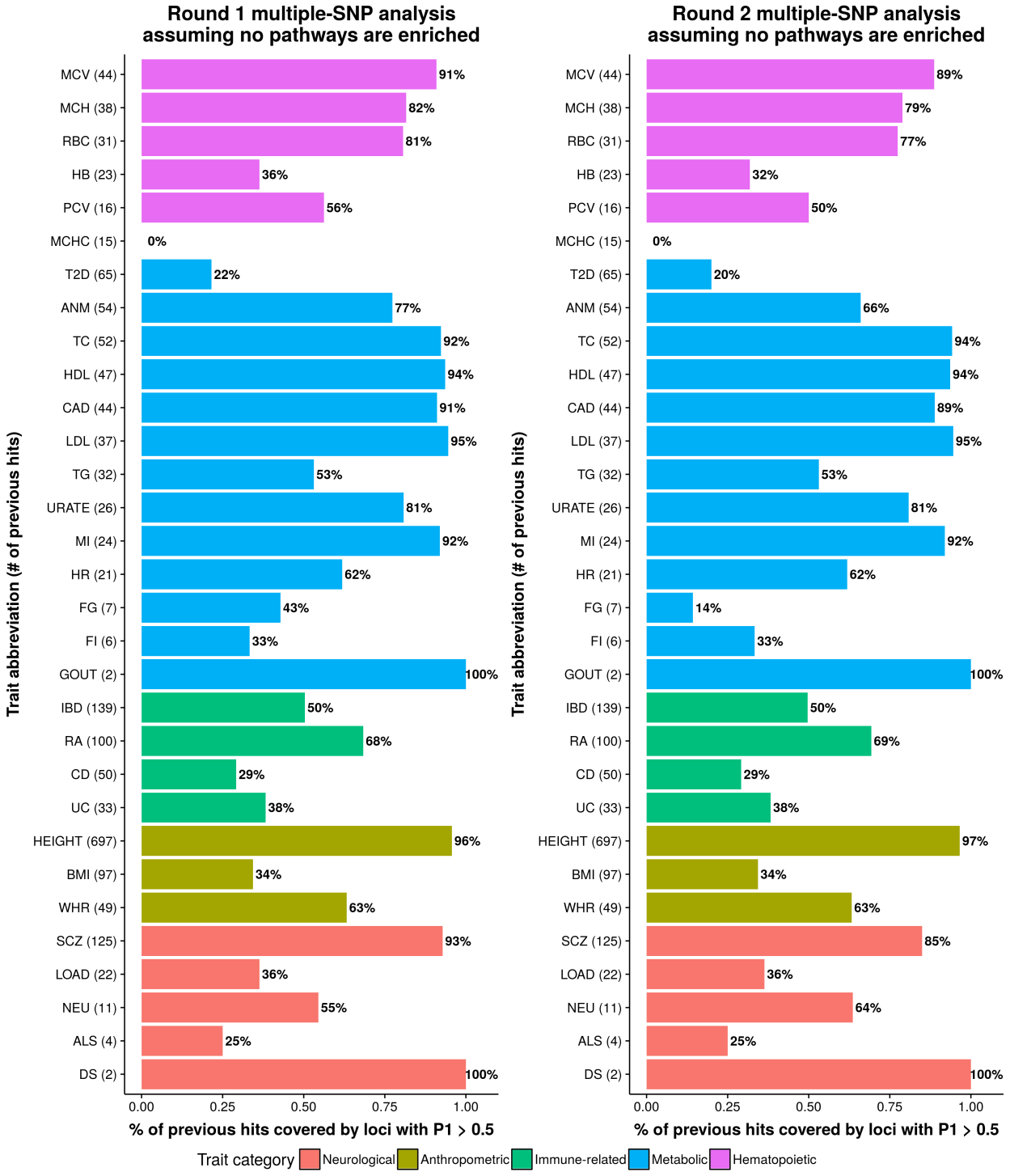
Figure F.4: Concordance between genome-wide single-SNP and multiple-SNP analyses of 31 phenotypes, both assuming that no pathways are enriched.

Supplementary Figure 5

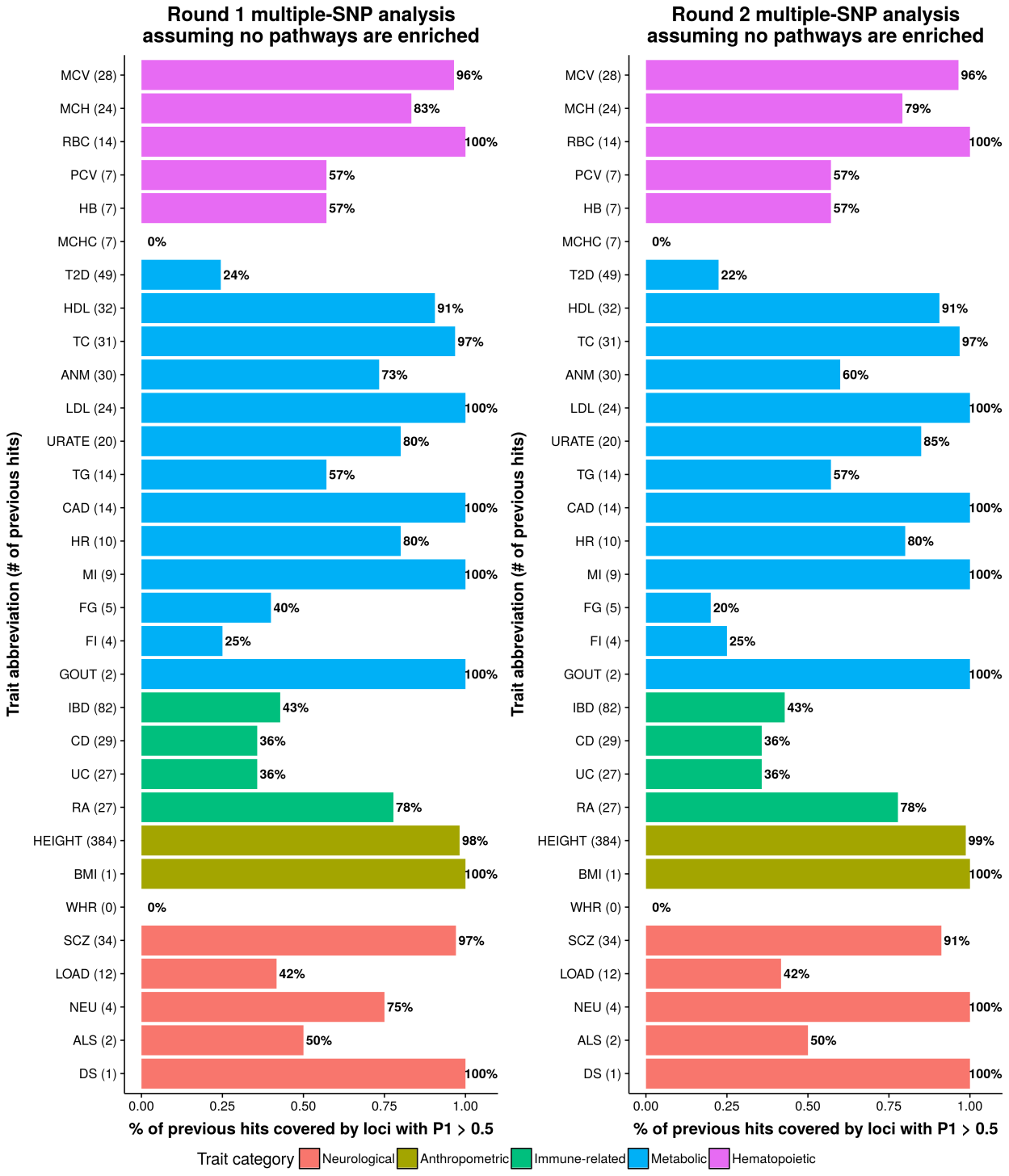
Proportion of previously-reported genome-wide significant variants that are detected by genome-wide multiple-SNP analyses, assuming that no pathways are enriched. For each phenotype, we manually extract the genome-wide significant variants (a.k.a. GWAS hits) from corresponding publications (e.g. Tables, Supplementary Files). [In contrast, we derive a list of GWAS hits from the summary statistics file for each trait in Supplementary Figure 4 of Zhu and Stephens (2017b).] We call a GWAS hit "detected by multiple-SNP analyses" if the variant is covered by a predefined locus satisfying certain multiple-SNP association criteria (estimated $P_1 > 0.5$, $P_1 > 0.9$ or ENS > 1). See Supplementary Figure 3 of Zhu and Stephens (2017b) for the definition of locus. For each panel, phenotypes are ordered first by trait category, then the number of reported GWAS hits.

It is important to note that some reported GWAS hits are not necessary (genome-wide) significant in the corresponding summary data file. For example, rs34856868 shows $p = 9.80 \times 10^{-9}$ for association with inflammatory bowel disease in Table 2 of Liu et al. (2015); however, the p value of the same SNP is 0.27 in the data file (`EUR.IBD.gwas.assoc.gz`). For this SNP, the result in Table 2 of Liu et al. (2015) was obtained from a combined analysis of data on both GWAS and custom arrays [Immunochip, Cortes and Brown (2011)], whereas the result in the summary data file was only based on GWAS data. Because of this type of potential discrepancy between results in summary data files and corresponding publications, the concordance rates shown in Supplementary Figure 5 of Zhu and Stephens (2017b) are lower than those shown in Supplementary Figure 4 of Zhu and Stephens (2017b) for several traits.

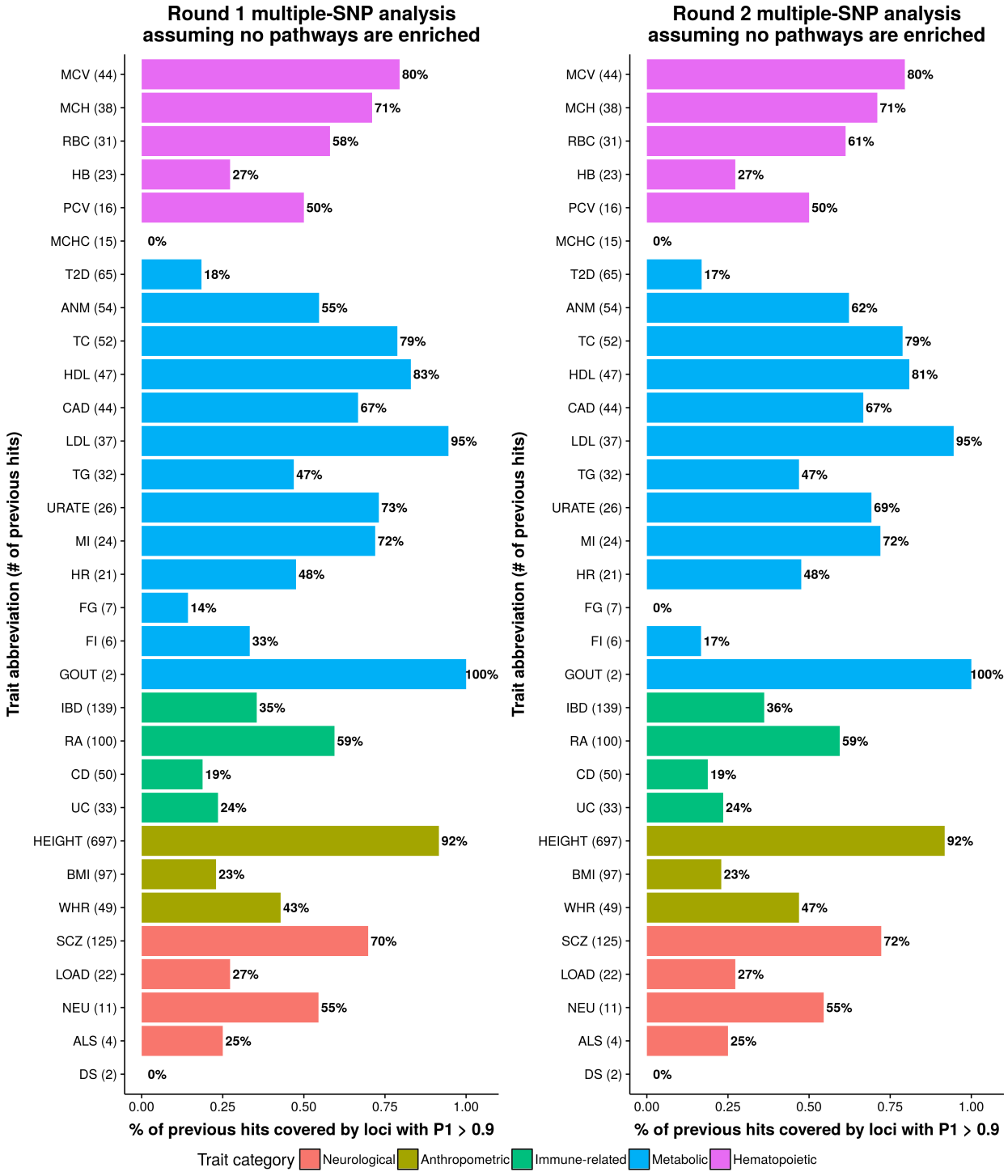
(a) Proportion of all previous GWAS hits that are covered by loci with $P_1 > 0.5$.



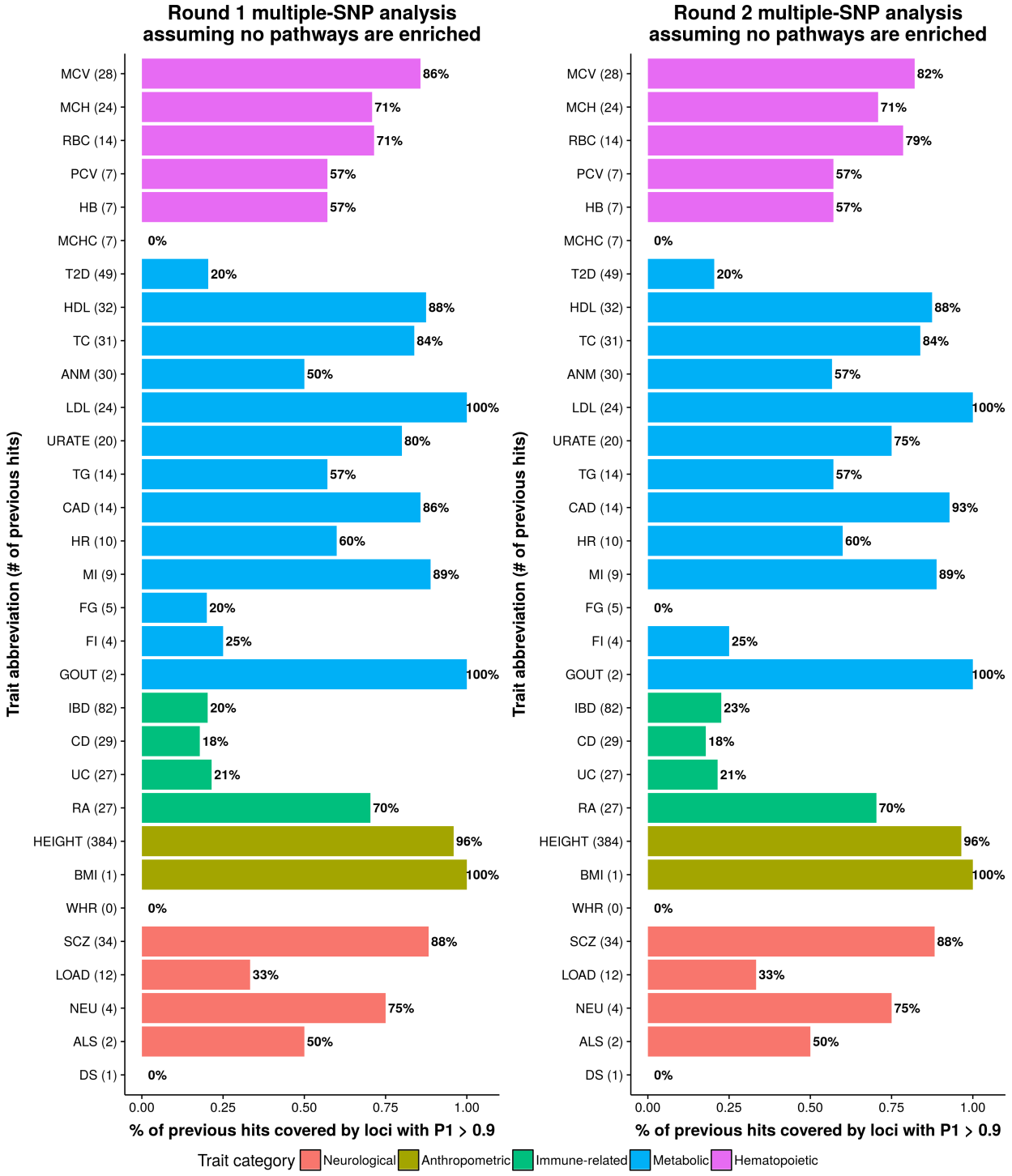
(b) Proportion of previous GWAS hits that are included in genome-wide multiple-SNP analyses and are covered by loci with $P_1 > 0.5$.



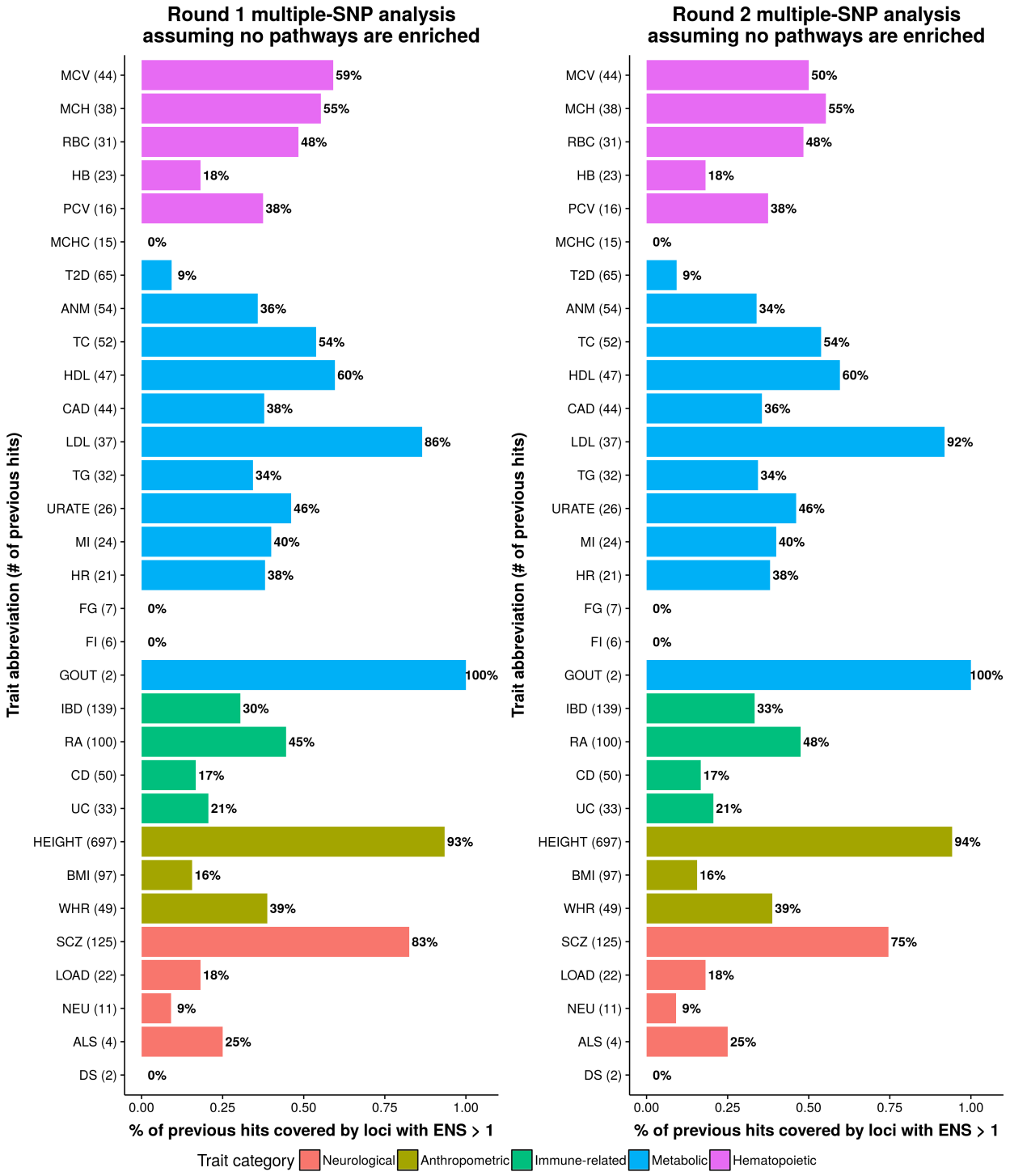
(c) Proportion of all previous GWAS hits that are covered by loci with $P_1 > 0.9$.



(d) Proportion of all previous GWAS hits that that are included in genome-wide multiple-SNP analyses and are covered by loci with $P_1 > 0.9$.



(e) Proportion of all previously-reported GWAS hits that are covered by loci with ENS > 1.



(f) Proportion of previously-reported GWAS hits that are included in genome-wide multiple-SNP analyses and are covered by loci with ENS > 1.

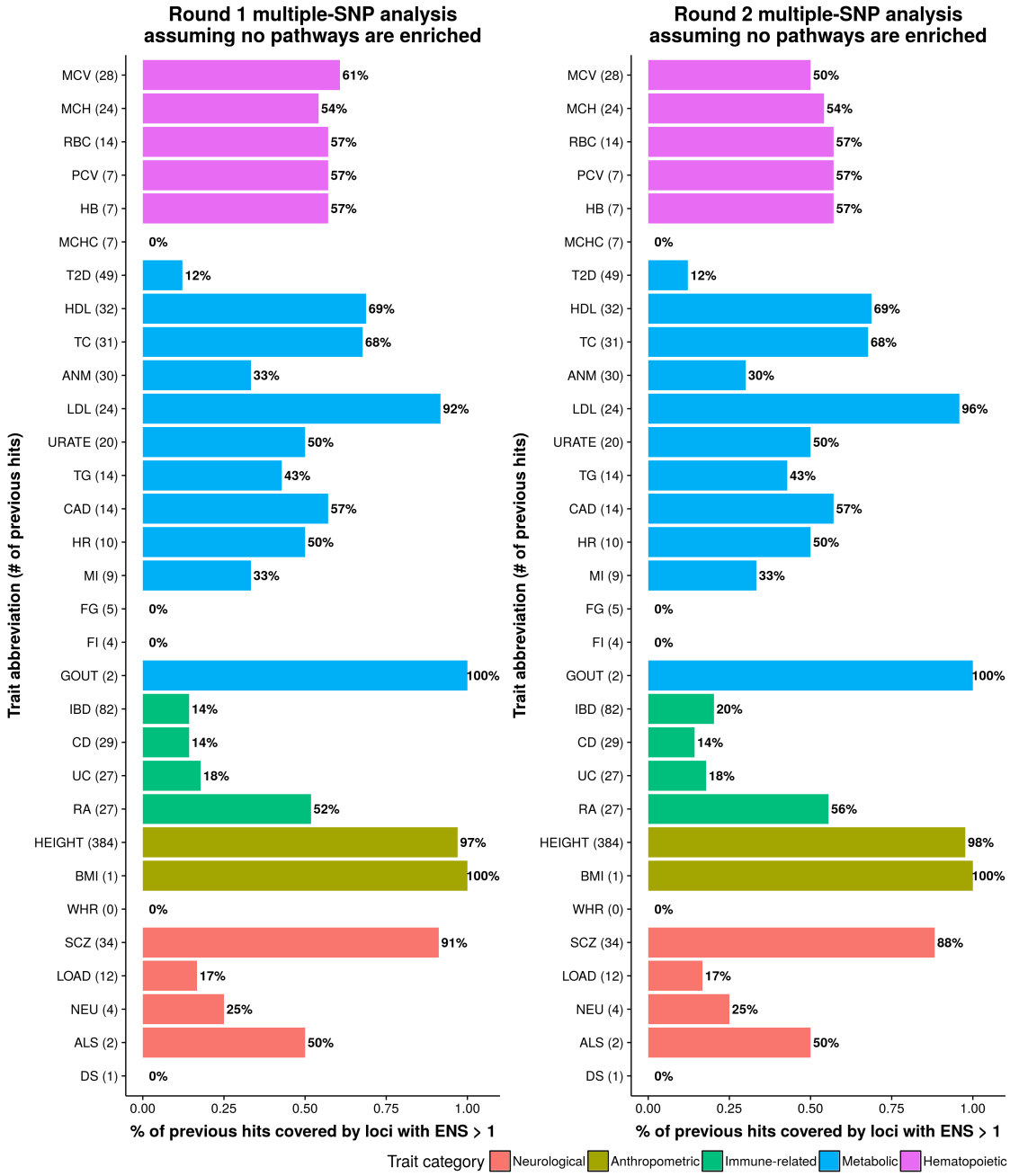


Figure F.5: Proportion of previously-reported genome-wide significant variants that are detected by genome-wide multiple-SNP analyses, assuming that no pathways are enriched.

Supplementary Figure 6

Compare the number of signals from genome-wide multiple-SNP and single-SNP analyses, both assuming that no pathways are enriched. Both axes use a logarithmic scale (base 10). Dashed lines are reference lines with intercept zero and slope one. Each point range along y-axis denotes the posterior mean and 95% credible interval of the posterior expected total number of trait-associated SNPs (ENS). See Supplementary Note of Zhu and Stephens (2017b) for the detail of computing this quantity. "Hits from single-SNP analysis" are the genome-wide significant genetic variants reported in corresponding publications [Supplementary Figure 5 of Zhu and Stephens (2017b)].

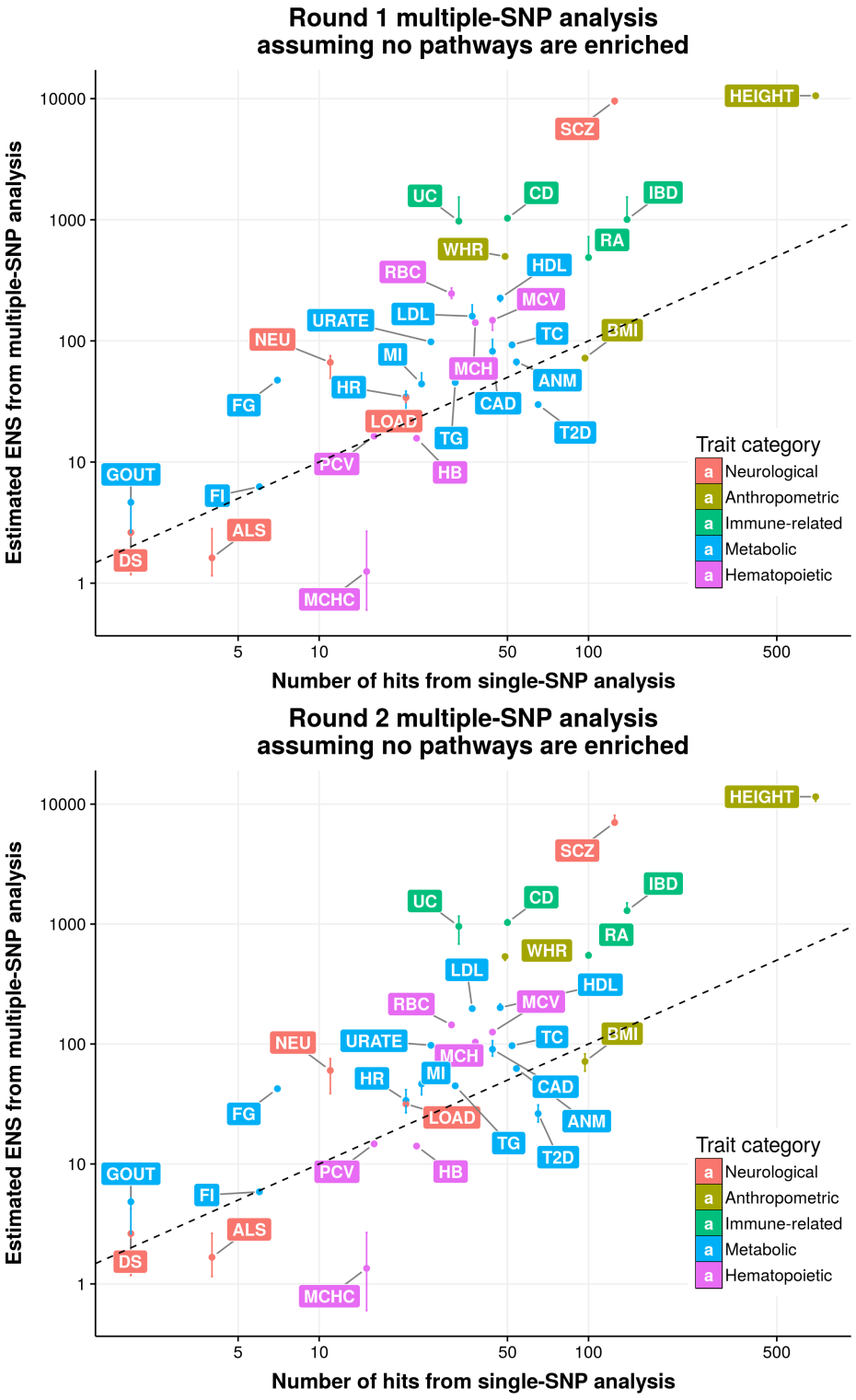
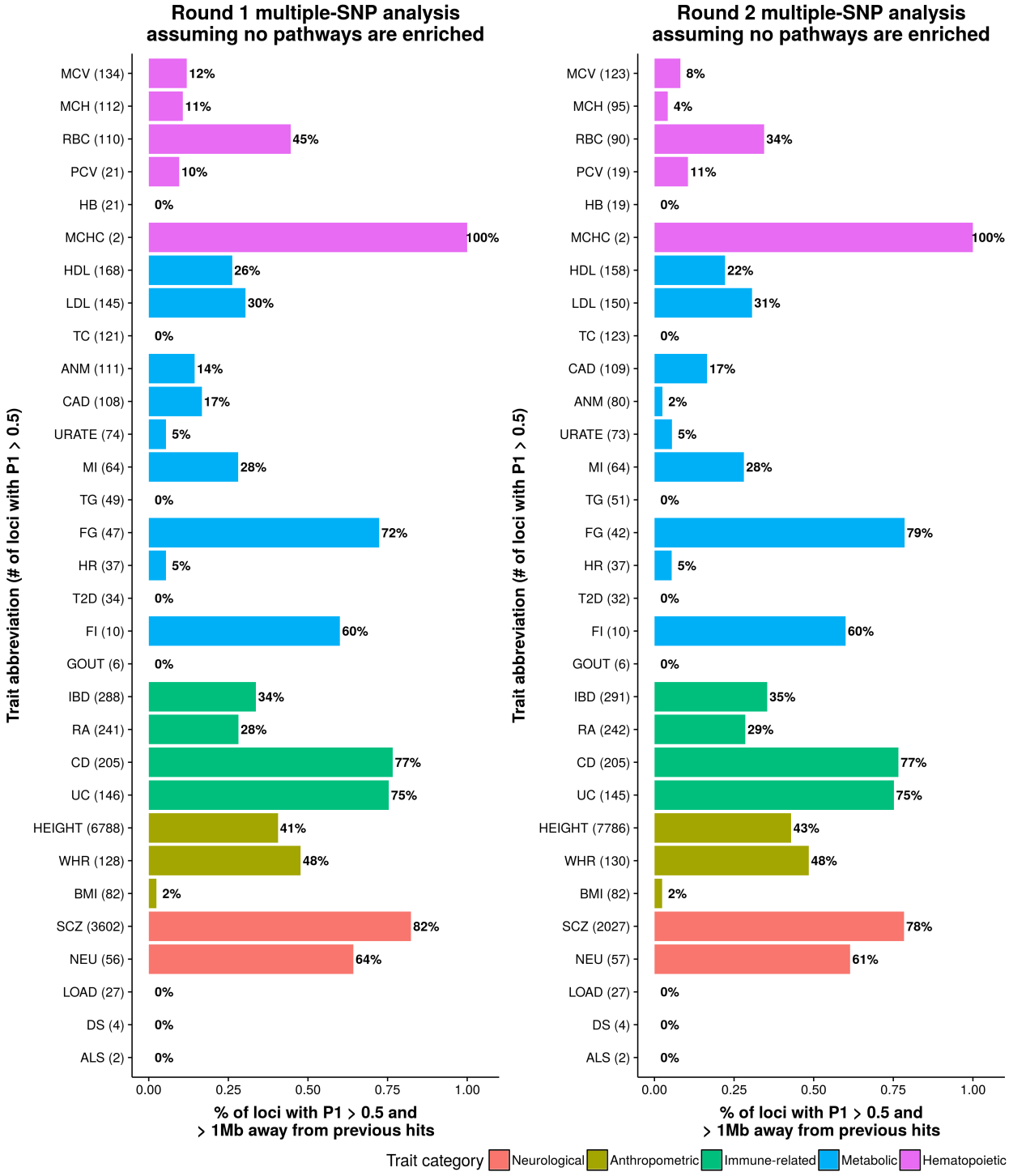


Figure F.6: Compare the number of signals from genome-wide multiple-SNP and single-SNP analyses, both assuming that no pathways are enriched.

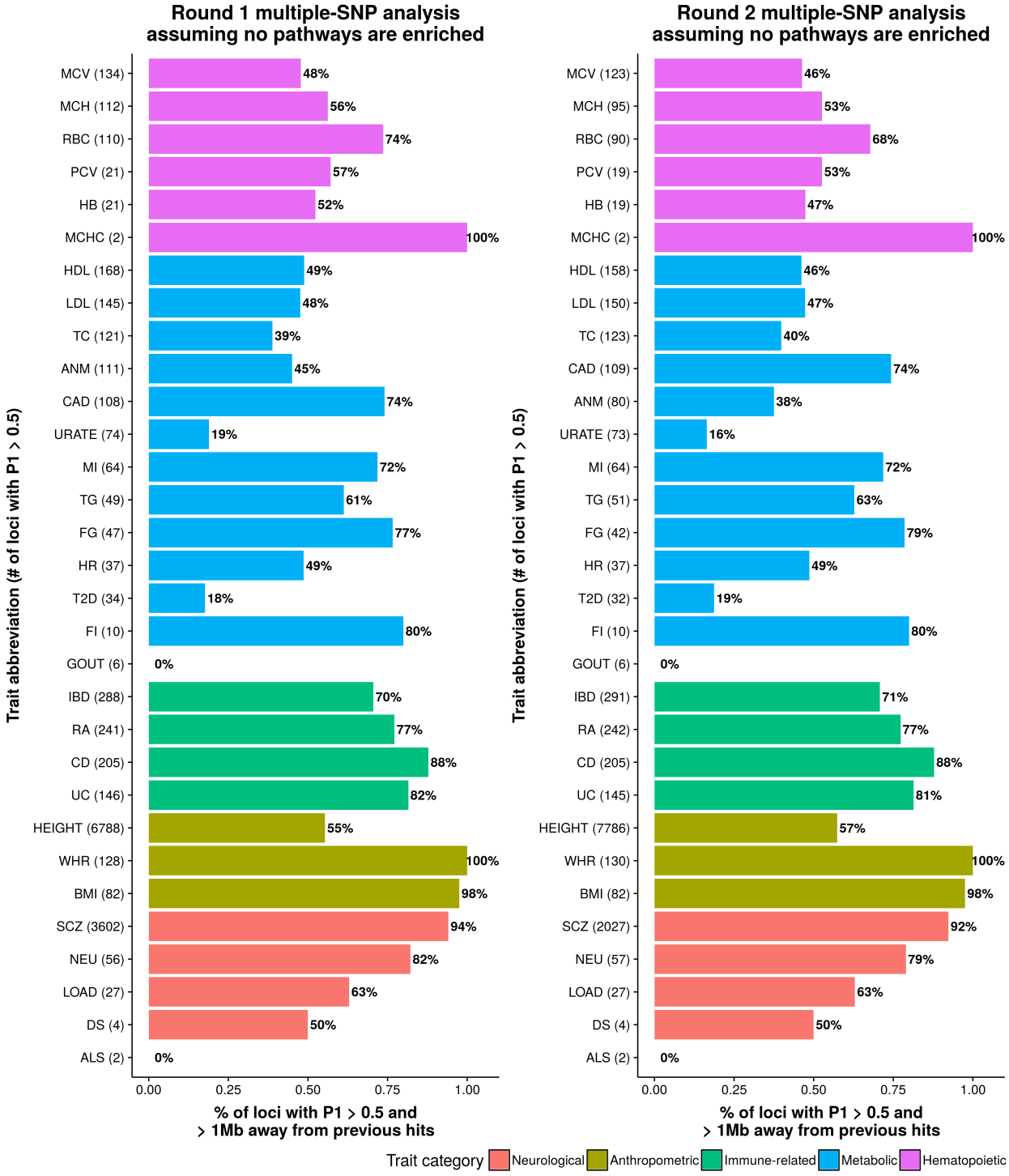
Supplementary Figure 7

Proportion of loci identified by genome-wide multiple-SNP analyses that are at least 1 Mb away from previously-reported GWAS hits, assuming that no pathways are enriched. We call a predefined locus "detected by multiple-SNP analyses" if the locus satisfies certain multiple-SNP association criteria (estimated $P_1 > 0.5$, $P_1 > 0.9$ or ENS > 1). See Supplementary Figure 3 of Zhu and Stephens (2017b) for the definition of locus. See Supplementary Figure 5 of Zhu and Stephens (2017b) for the definition of "previously-reported GWAS hits". For each panel, phenotypes are ordered first by category, then the number of loci identified from multiple-SNP analyses.

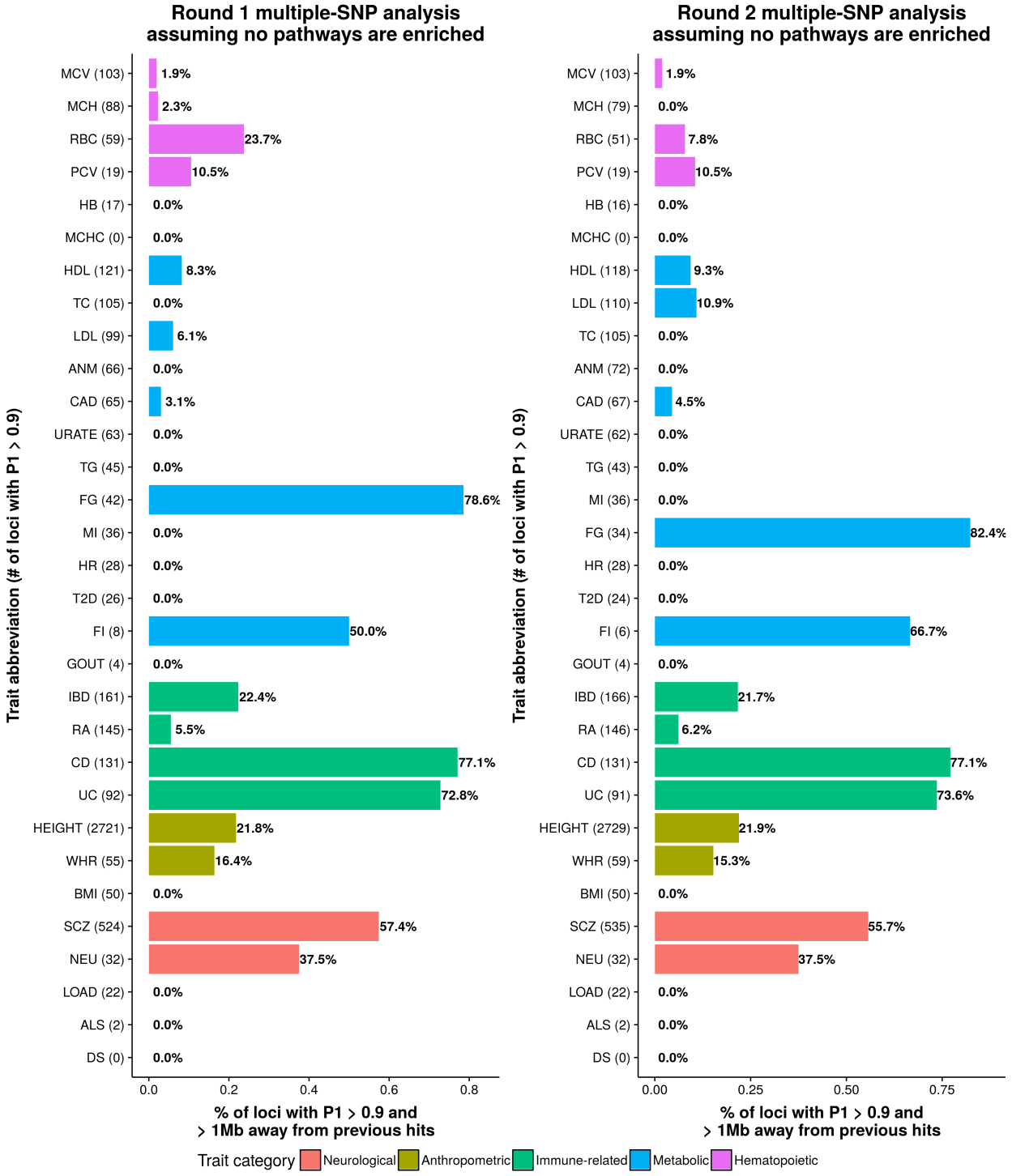
(a) Proportion of loci with $P_1 > 0.5$ that are at least 1 Mb away from all previously-reported GWAS hits.



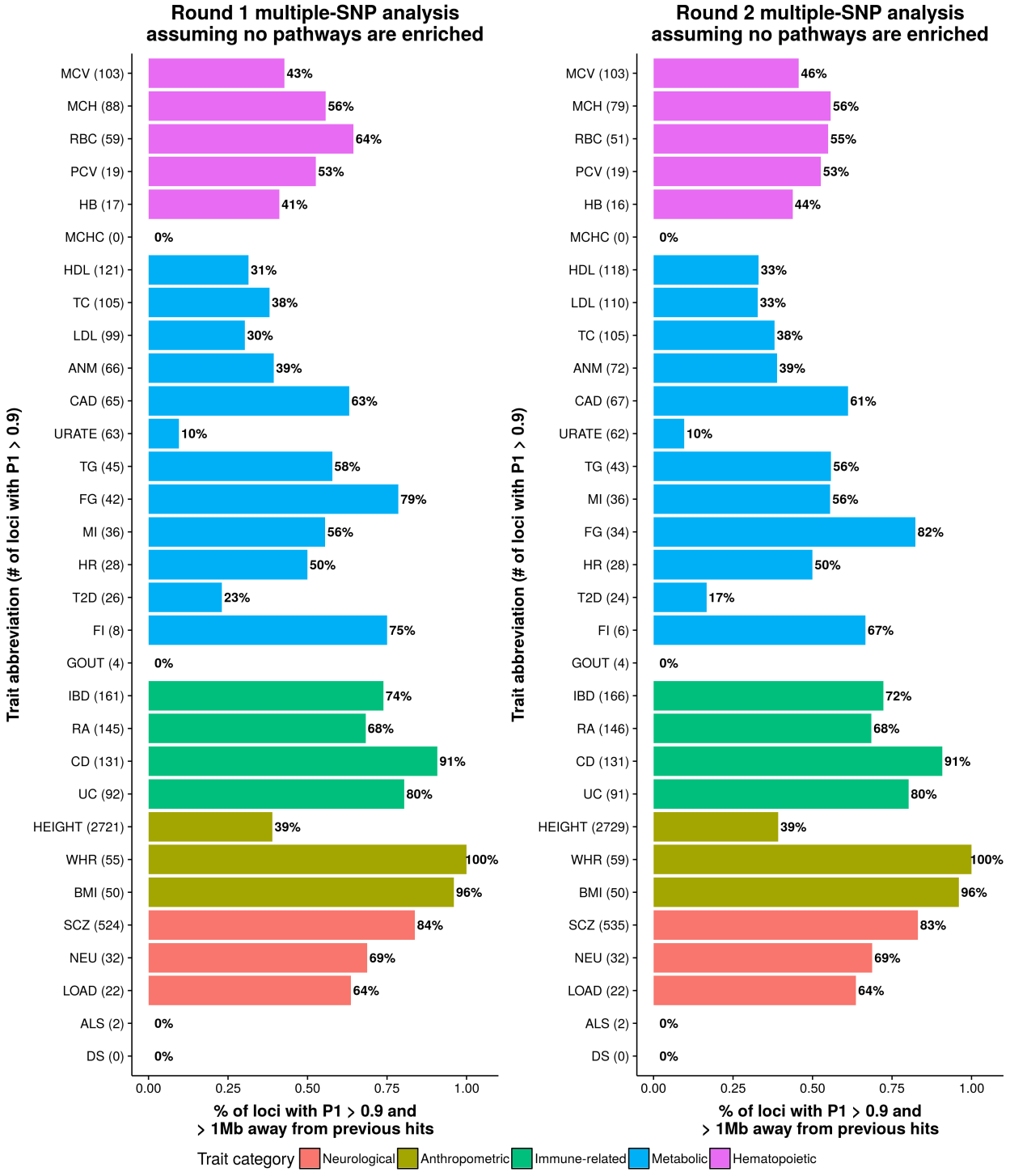
(b) Proportion of loci with $P_1 > 0.5$ that are at least 1 Mb away from previously-reported GWAS hits that are included in genome-wide multiple-SNP analyses.



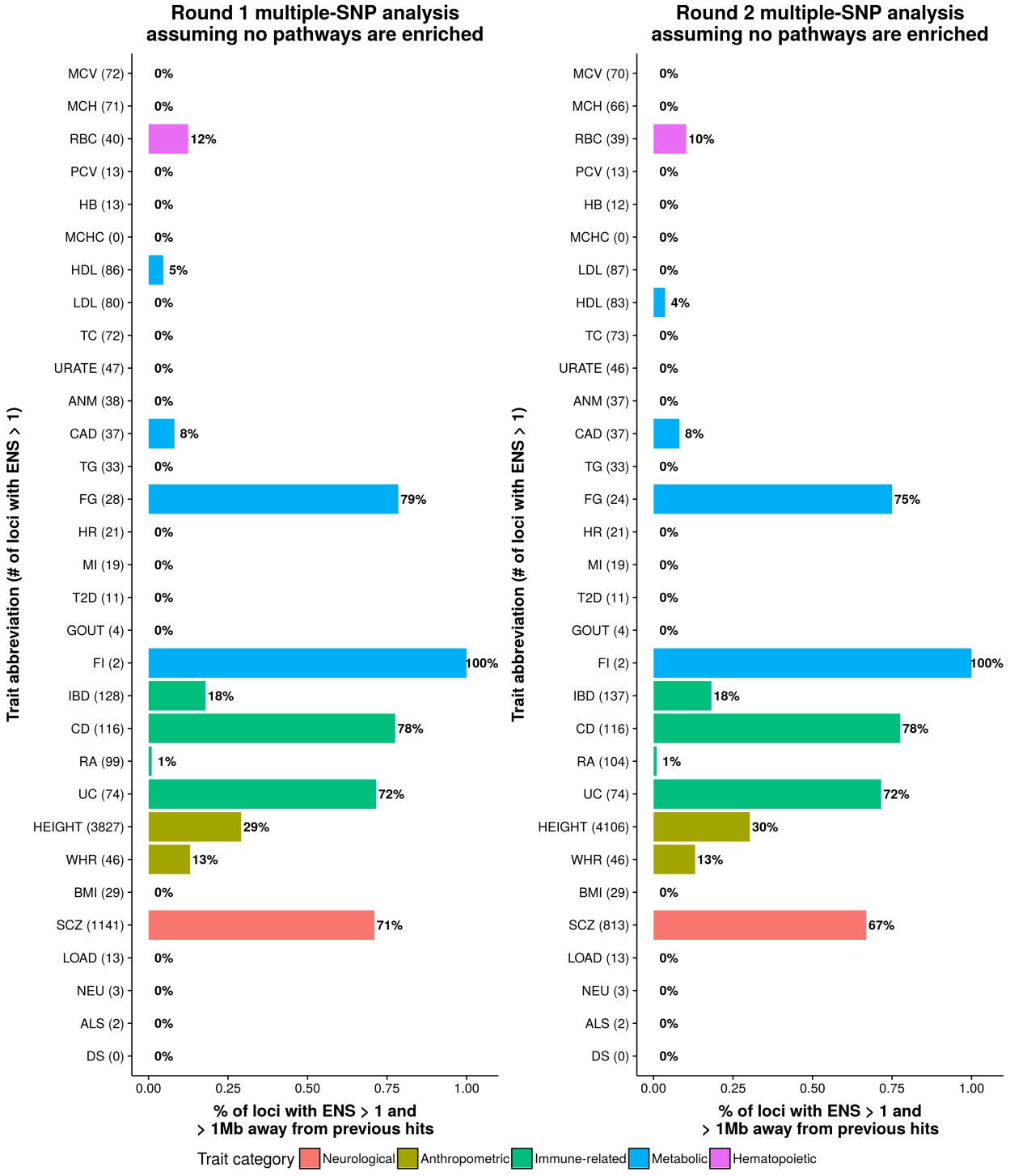
(c) Proportion of loci with $P_1 > 0.9$ that are at least 1 Mb away from all previously-reported GWAS hits.



(d) Proportion of loci with $P_1 > 0.9$ that are at least 1 Mb away from previously-reported GWAS hits that are included in genome-wide multiple-SNP analyses.



(e) Proportion of loci with ENS > 1 that are at least 1 Mb away from all previously-reported GWAS hits.



(f) Proportion of loci with ENS > 1 that are at least 1 Mb away from previously-reported GWAS hits that are included in genome-wide multiple-SNP analyses.

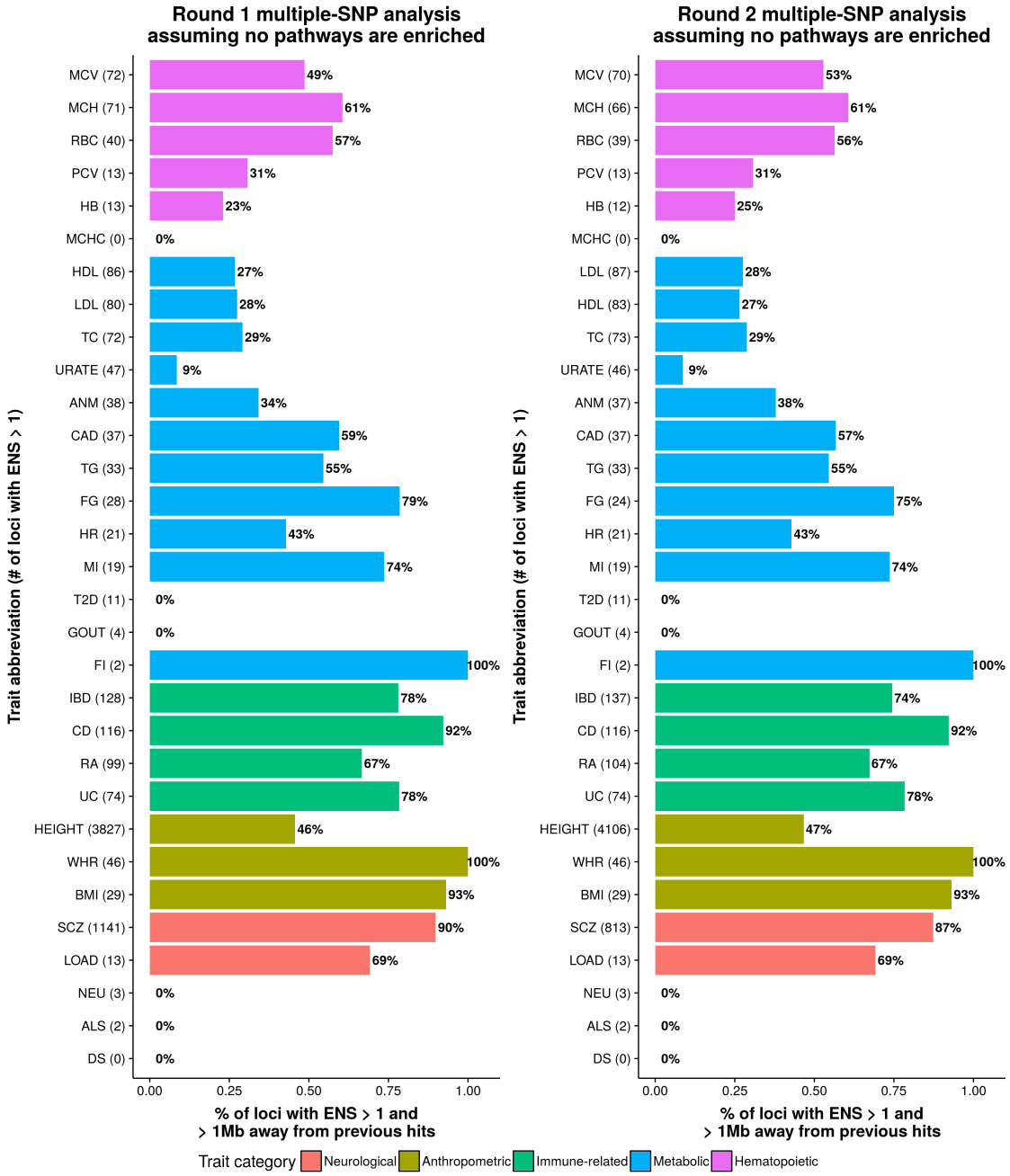


Figure F.7: Proportion of loci identified by genome-wide multiple-SNP analyses that are at least 1 Mb away from previously-reported GWAS hits, assuming that no pathways are enriched.

Supplementary Figure 8

Summary of biological pathways. Biological pathway definitions are retrieved from the Pathway Commons 2 [version 7, Cerami et al. (2011)], NCBI Biosystems (Geer et al., 2010), PANTHER [version 3.3, Mi and Thomas (2009)] and BioCarta [used in Carbonetto and Stephens (2013)]. The Pathway Commons 2 database includes gene sets derived from Reactome (Croft et al., 2014), Nature Pathway Interaction Database [PID, Schaefer et al. (2009)], HumanCyc (Romero et al., 2004), PANTHER, miRTarBase (Hsu et al., 2014) and Kyoto Encyclopedia of Genes and Genomes [KEGG, Wrzodek et al. (2013)] pathways. The NCBI BioSystem database contains pathways from KEGG, BioCyc (Caspi et al., 2014), PID, Reactome and WikiPathways (Pico et al., 2008). See Supporting Information Text S1 of Carbonetto and Stephens (2013) for the rationale for choosing these pathway databases (`https://doi.org/10.1371/journal.pgen.1003770.s015`). For each panel, the bar chart on the right side (labeled as "Total") shows the total number of pathways retrieved from the corresponding database, and the one on the left side (labeled as "Analyzed") shows the number of pathways used in our enrichment analyses.

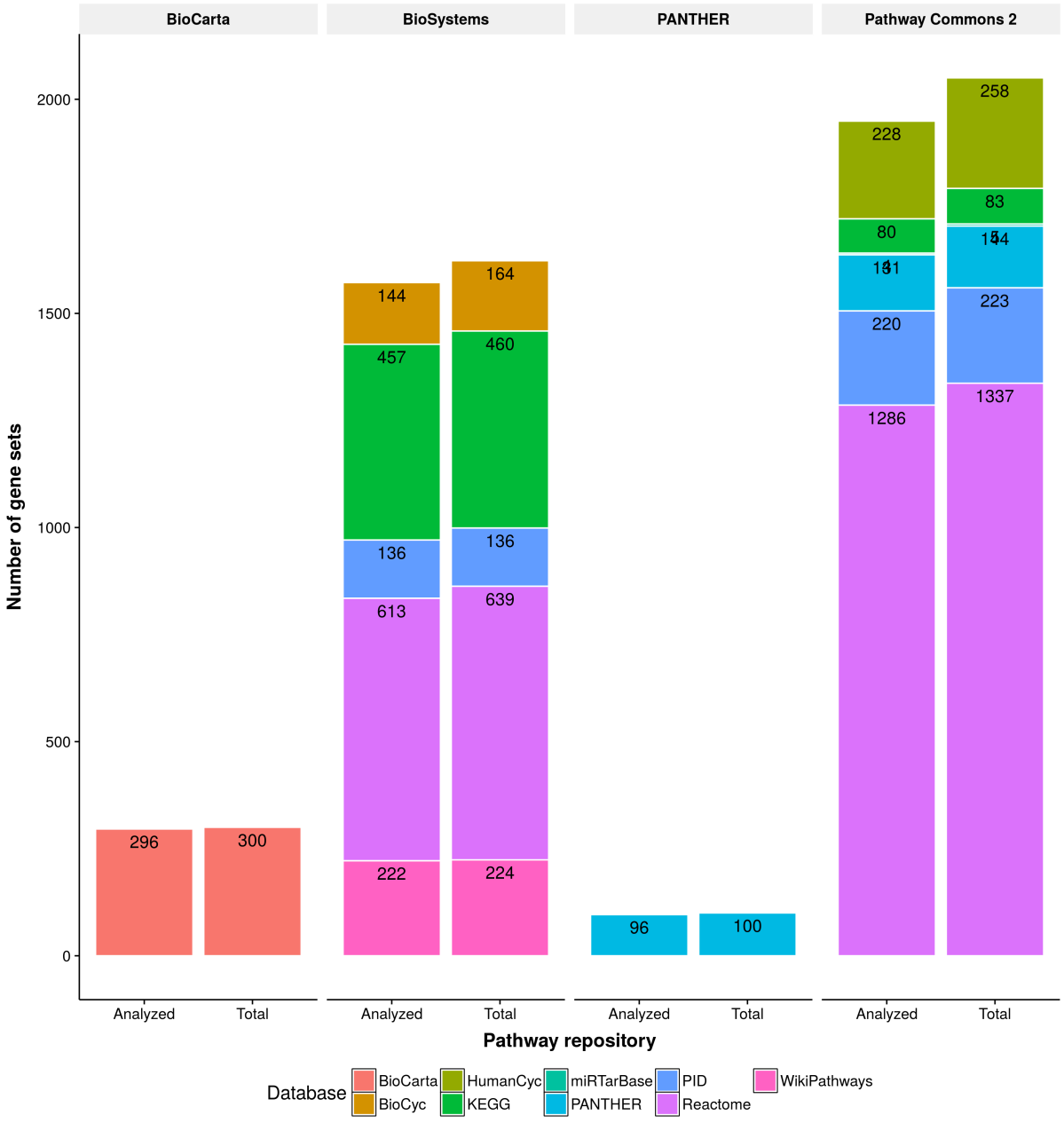
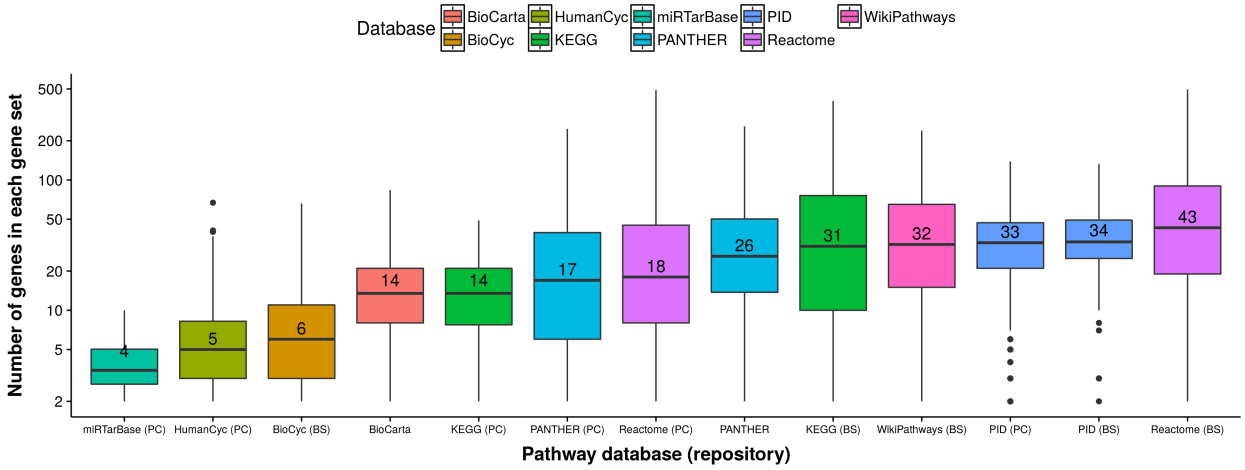


Figure F.8: Summary of biological pathways.

Supplementary Figure 9

Summary of genes. Genomic definitions for genes are derived from *Homo sapiens* reference genome GRCh37. In the present pathway analysis, we consider 18,313 autosomal protein-coding genes that are mapped to the reference sequence.

(a) Distributions of pathway sizes for every pathway database-repository combination. Combinations are ordered by median numbers of genes in pathways, which are displayed in each box plot. The vertical axis uses a logarithmic scale (base: 10). PC: Pathway Commons. BS: NCBI BioSystems.



(b) Manhattan plot of the number of pathway annotations for each gene. The highlighted genes (colored in green and labeled by their HGNC symbols) belong to more than 270 of 3,913 analyzed biological pathways. The Manhattan plot is produced by `R` package `qqman` (Turner, 2014).

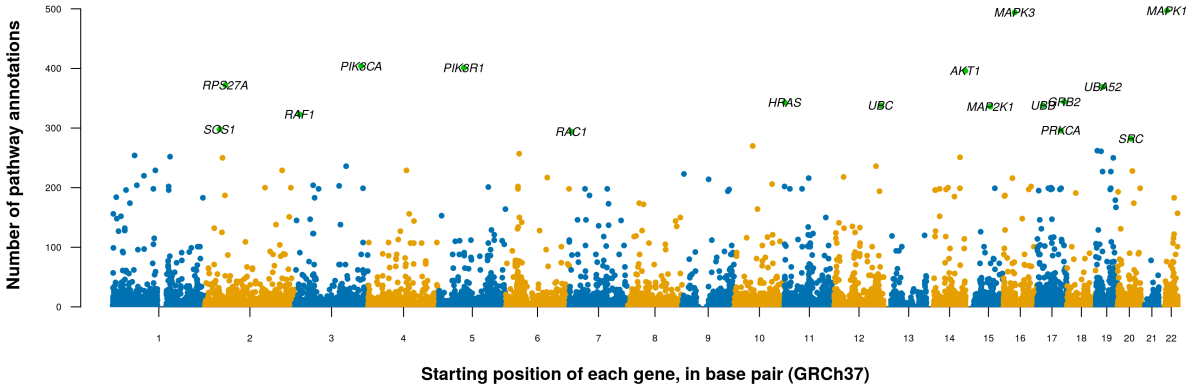


Figure F.9: Summary of genes.

Supplementary Figure 10

Sanity checks of top-ranked gene set enrichments for 31 phenotypes. To quickly evaluate whether the strong enrichments identified in our model-based analyses can possibly be true, we develop two sanity checks and apply them to the top enriched gene sets.

The first sanity check is an "eyeball test" that visualizes the distribution of GWAS single-SNP z-scores for a target trait, stratified by SNP-level annotations of a target gene set. Specifically, we plot two estimated density curves for each pair of trait and gene set:

- a **solid red curve** estimated from z-scores of SNPs within $\pm$ 100 kb of the transcribed region of a gene in the gene set ("inside gene set" SNPs);
- a **dashed black curve** estimated from z-scores of remaining SNPs ("outside gene set" SNPs).

For a typical pair of trait and gene set that is deemed to pass the "eyeball test", its dashed black curve is often more "spiky" at zero, and its solid red curve is more spread out. The density curves are produced by the function `geom_density` in `R` package `ggplot2` (default setting).

The second sanity check computes a likelihood ratio (LR) for the following two models:

- Null model (`a2`): "Inside gene set" SNPs have the same effect size distribution as "outside gene set" SNPs, which can be estimated by `a1$fitted_g` based on the (merged) whole genome data;
- Alternative model (`a3`): "Inside gene set" and "outside gene set" SNPs have different effect size distributions, which should be estimated separately.

For a strongly enriched gene set, its LR value tends to be very large, since the data should favor the alternative over the null hypothesis. The second check based on LR computation complements the first visual check in cases where the "eyeball test" results are not clearly

visible. The LR calculation is based on R package ashr (Stephens, 2017). Below are some R codes that illustrate the LR calculation.

```
suppressPackageStartupMessages(library(R.matlab))
suppressPackageStartupMessages(library(ashr))

sumstat.file <- "ldl2010_path1698_sumstat.mat"
sumstat <- R.matlab::readMat(sumstat.file)

# load GWAS summary statistics
betahat <- c(sumstat$betahat)
se <- c(sumstat$se)

# load SNP-level annotations
snps <- c(sumstat$snps)

# analyze summary data of the whole genome
a1 <- ashr::ash(betahat=betahat, sebetahat=se,
                mixcompdist="halfuniform", method="shrink")

# analyze summary data of SNPs that are "inside gene set"
# where the prior is estimated from data
a2 <- ashr::ash(betahat=betahat[snps], sebetahat=se[snps],
                mixcompdist="halfuniform", method="shrink")

# analyze summary data of SNPs that are "inside gene set"
# where the prior is fixed as the one estimated in a1
a3 <- ashr::ash(betahat=betahat[snps], sebetahat=se[snps],
```

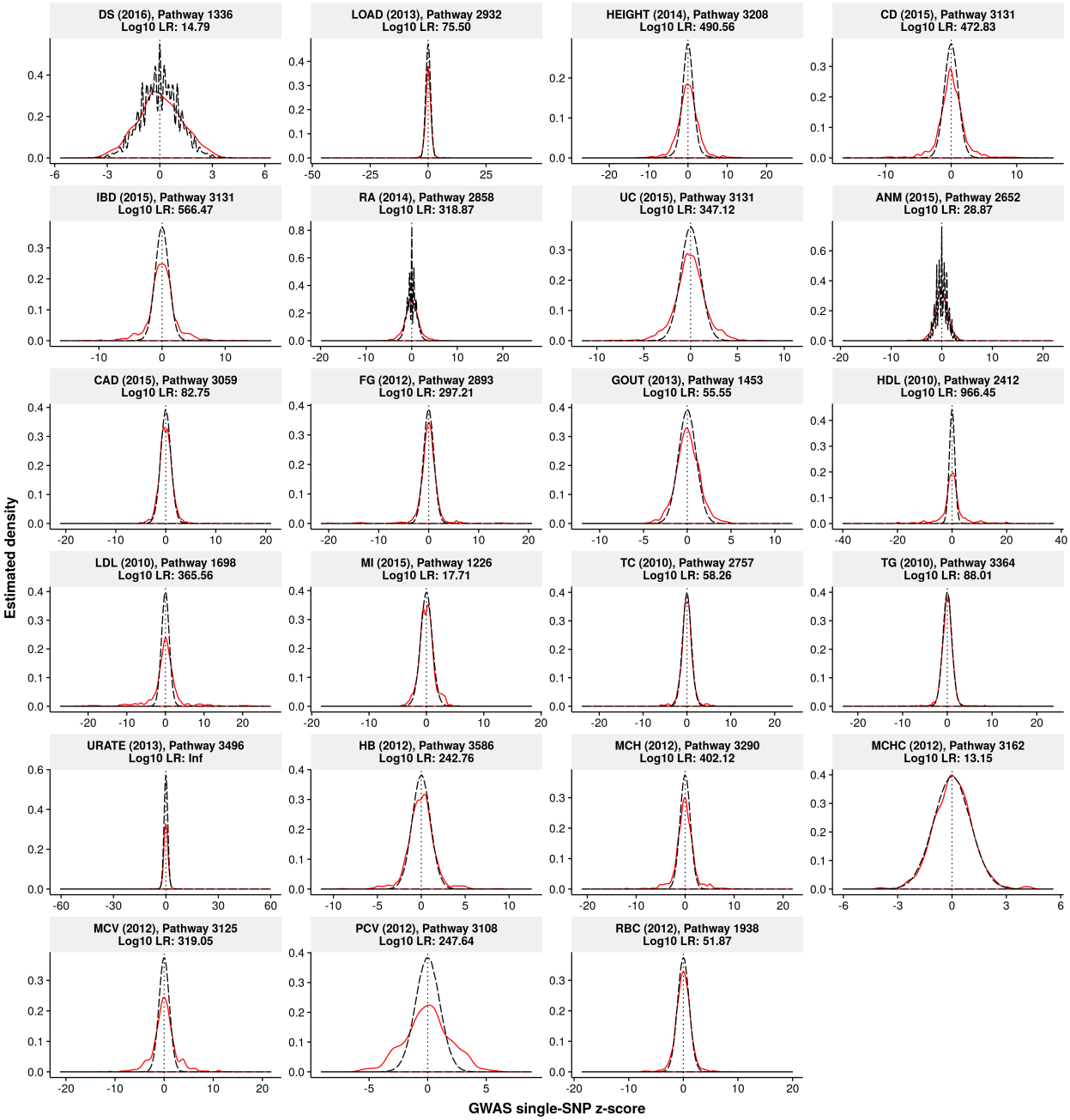
```
                mixcompdist="halfuniform", method="shrink",
                fixg=TRUE, g=a1$fitted_g)

# compute log10 likelihood ratio statistics
log10LR <- (a2$logLR - a3$logLR) / log(10)
```

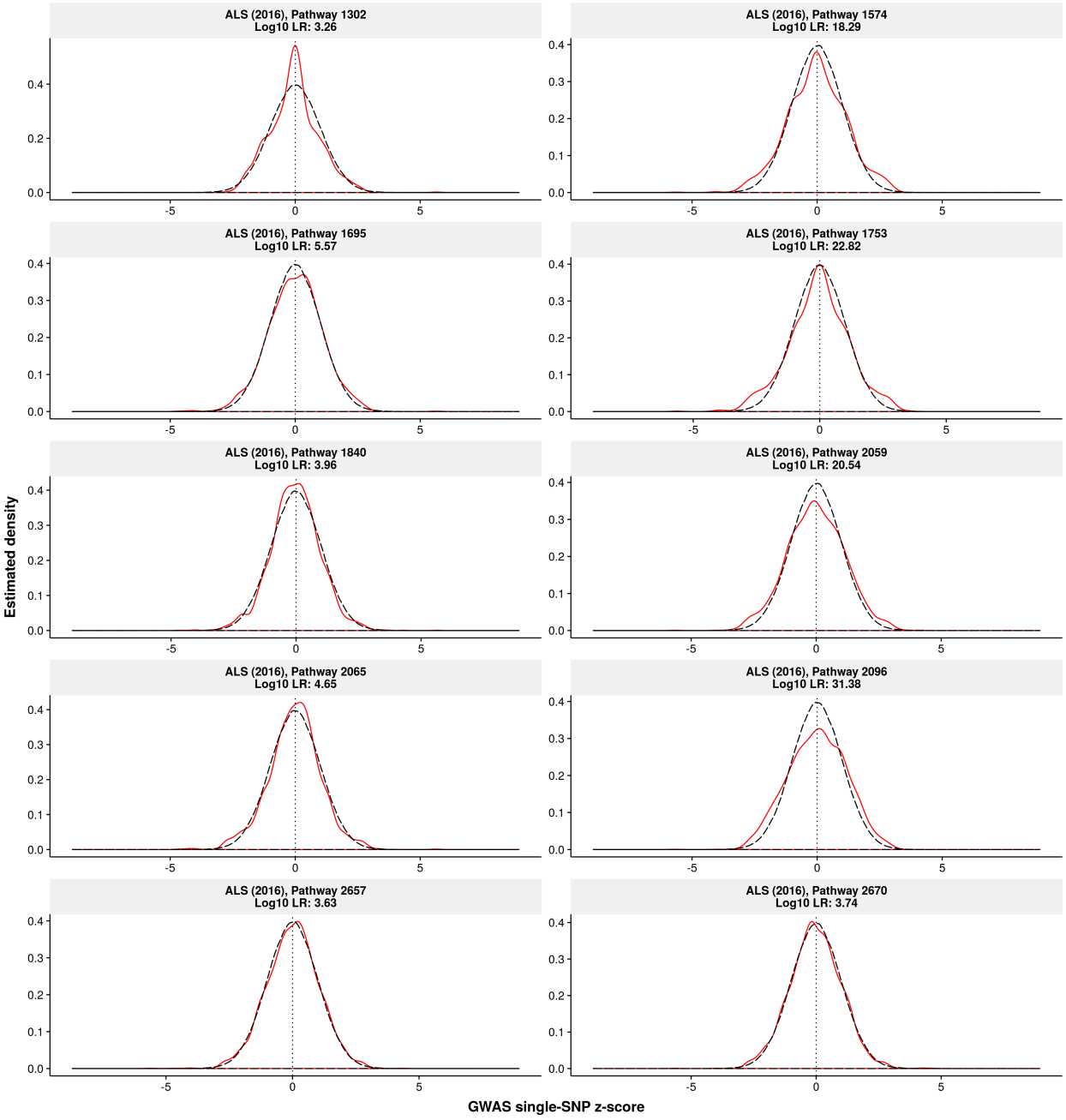
We first perform sanity checks on the trait-pathway pairs reported in Table 1 of Zhu and Stephens (2017b). We then perform sanity checks on the top 10 most enriched pathways with at least 10 member genes for each of the 31 traits. Finally we perform sanity checks on the trait-tissue pairs reported in Table 2 of Zhu and Stephens (2017b). Full information about these top enriched pathways and tissue-based gene sets is available at `http://xiangzhu.github.io/rss-gsea/results`.

For each pair of trait and gene set, the "eyeball test" result is presented as two density curves, and the log10 LR result is reported in the figure title. Suggested thresholds of LR are the same as thresholds of enrichment Bayes factors used in the main text: 1×10^8 for pathways and 1×10^3 for tissue-based gene sets.

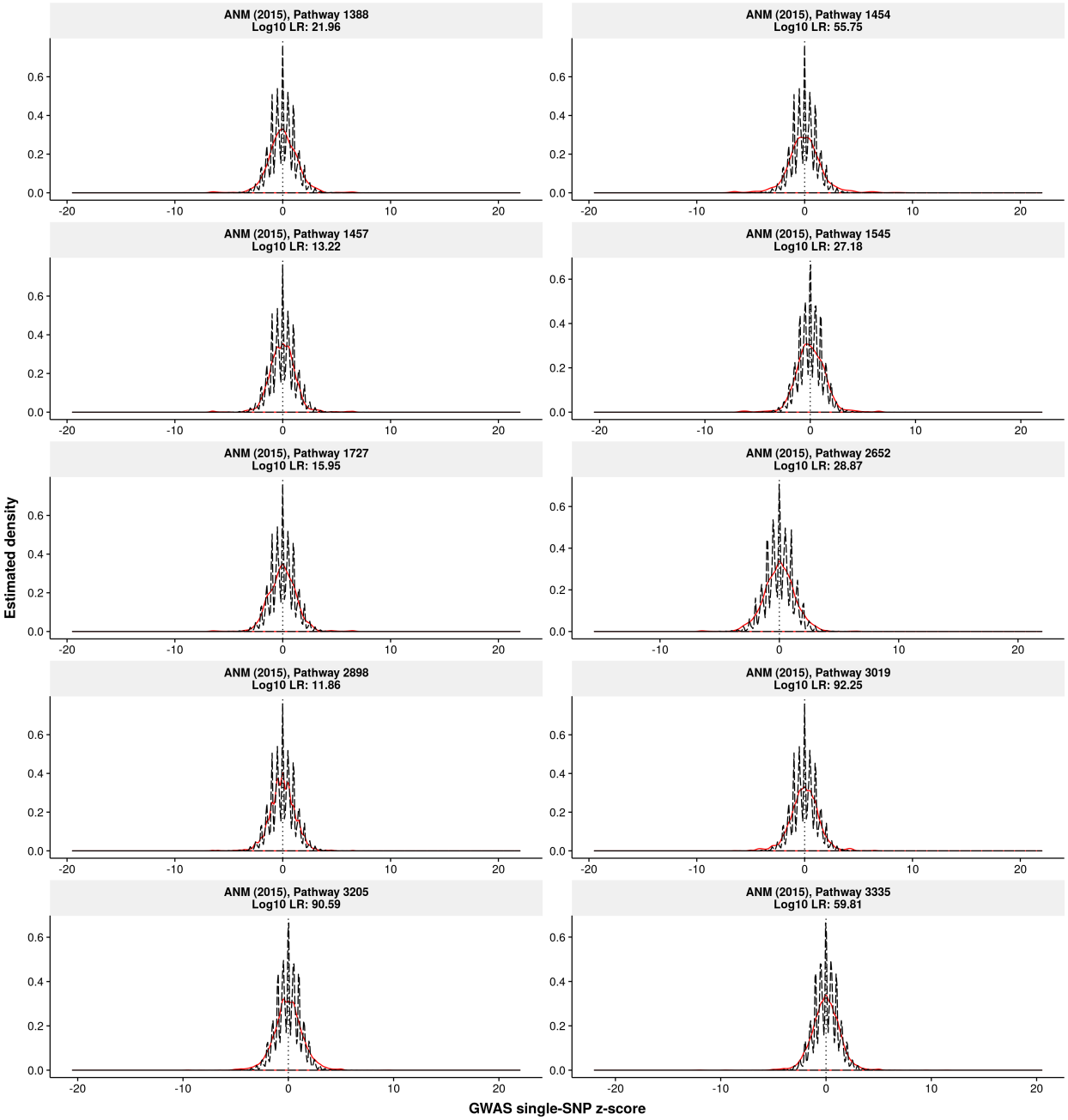
Pathway-trait pairs reported in Table 1 of Zhu and Stephens (2017b)



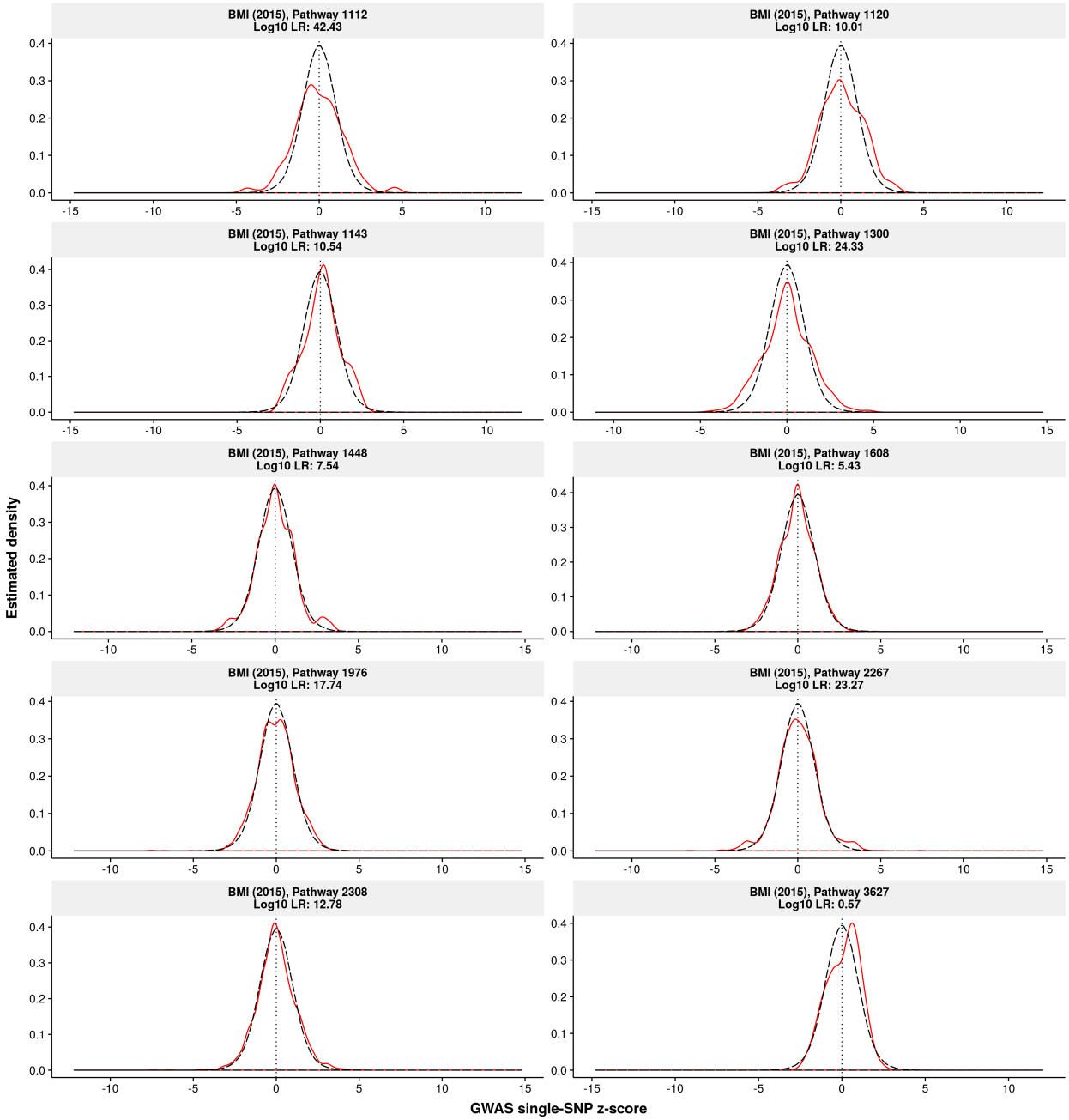
Amyotrophic lateral sclerosis (van Rheenen et al., 2016)



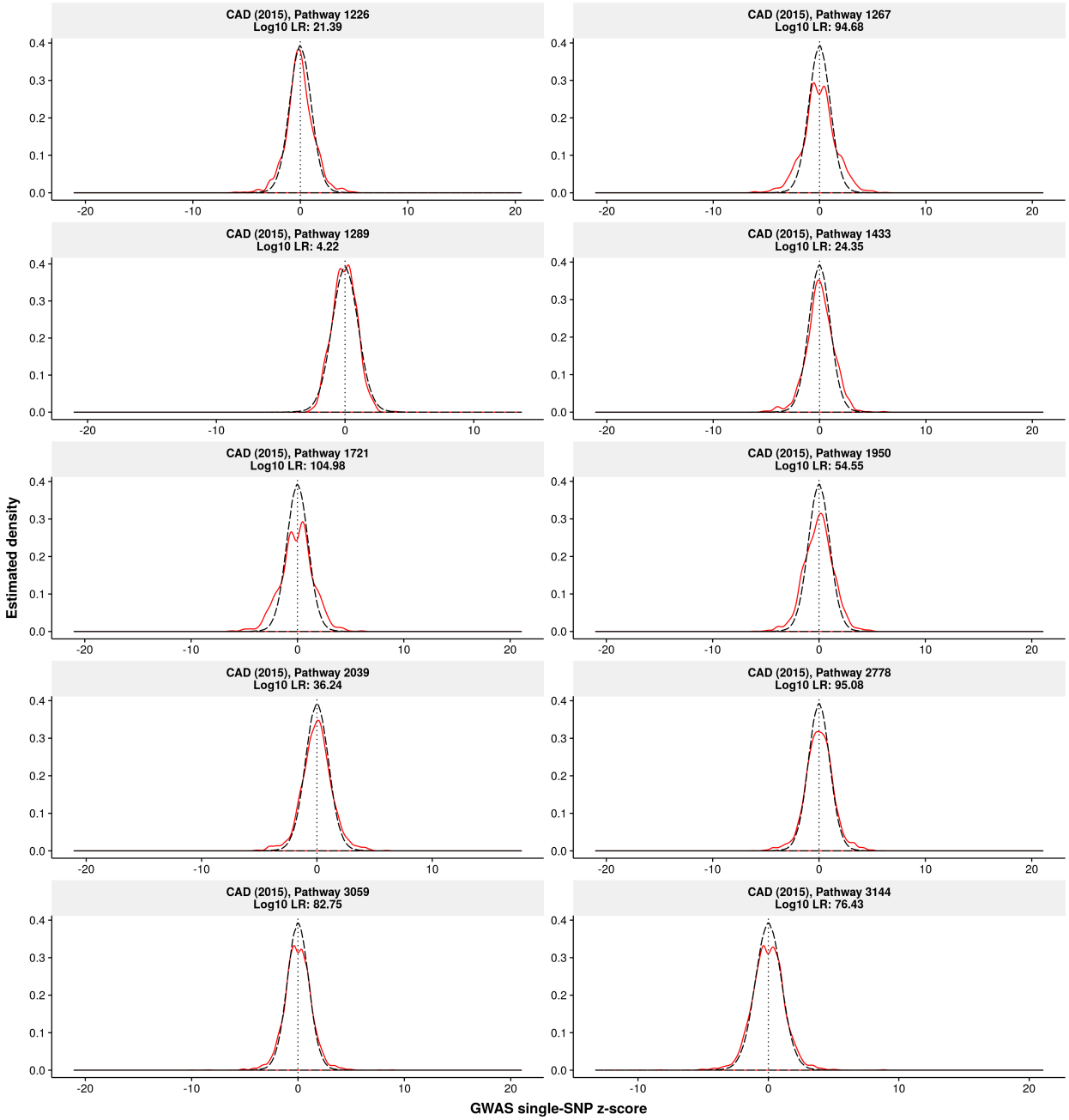
Age at natural menopause (Day et al., 2015)



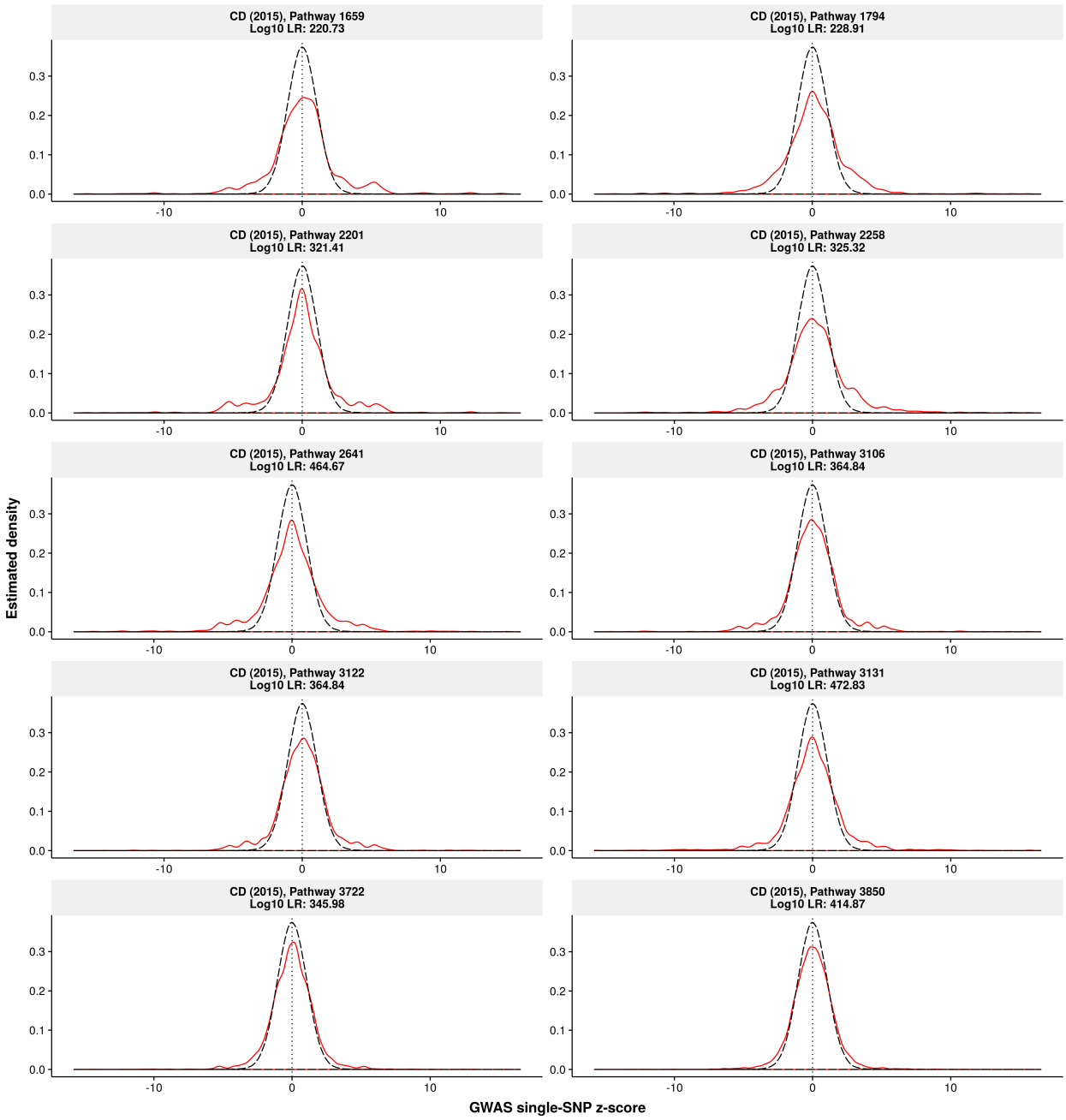
Body mass index (Locke et al., 2015)



Coronary artery disease (Nikpay et al., 2015)

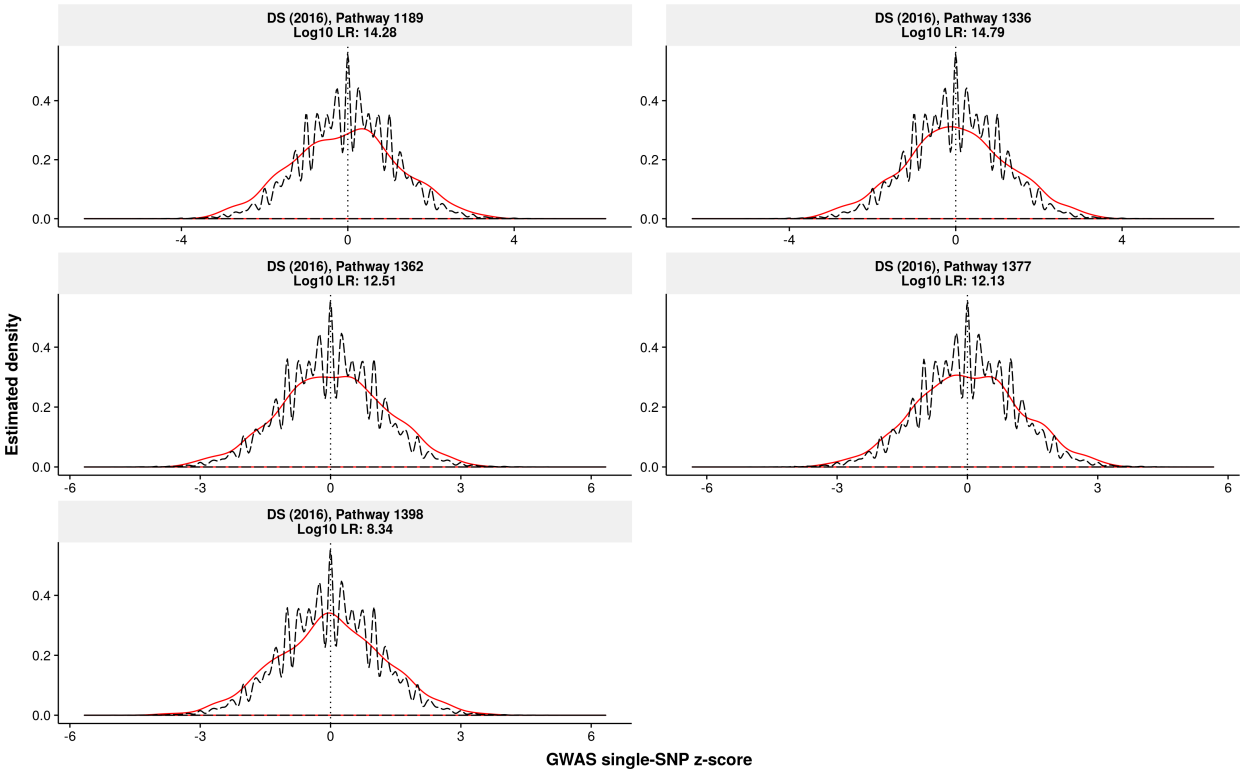


Crohn's disease (Liu et al., 2015)

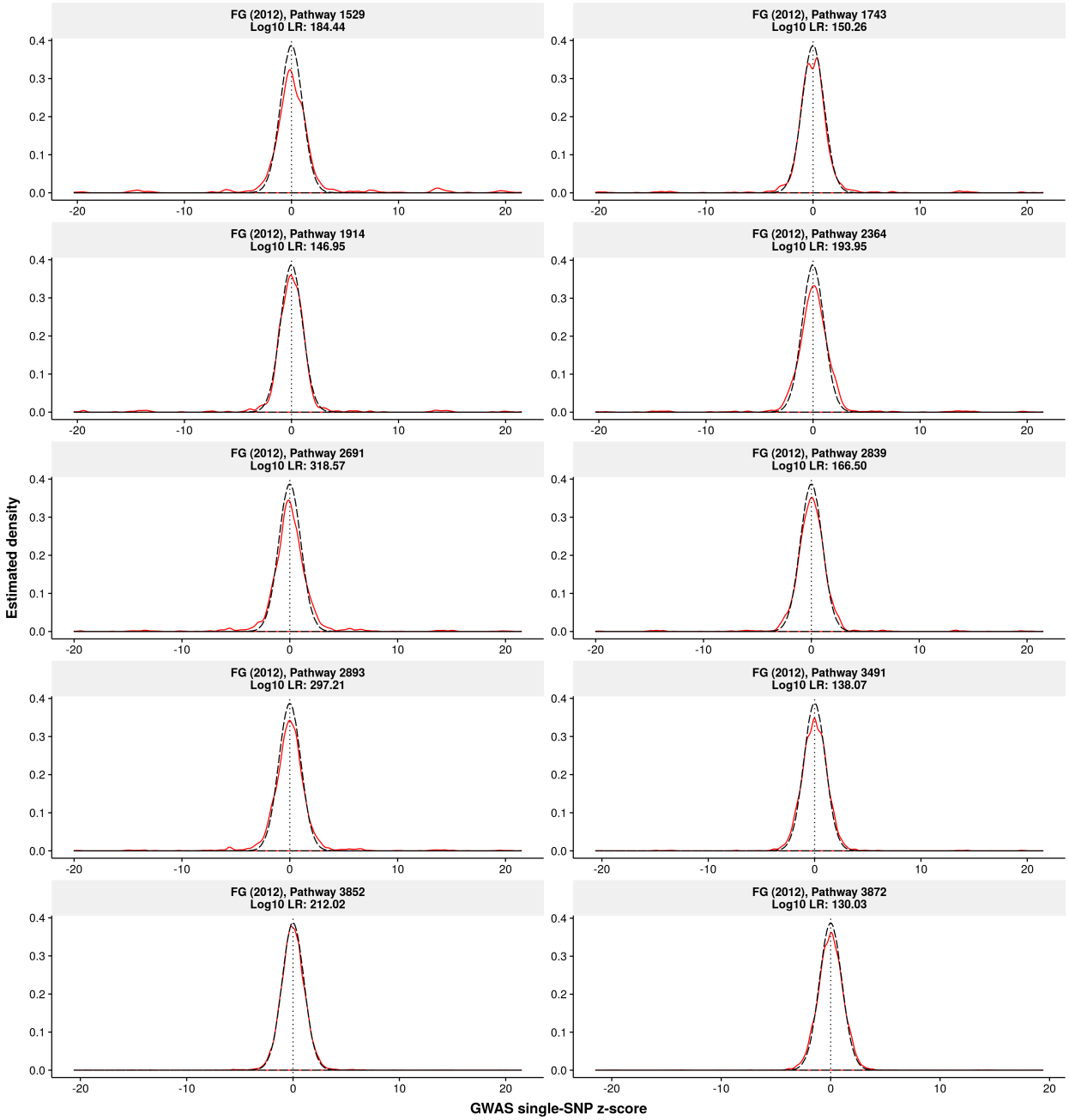


Depressive symptoms (Okbay et al., 2016)

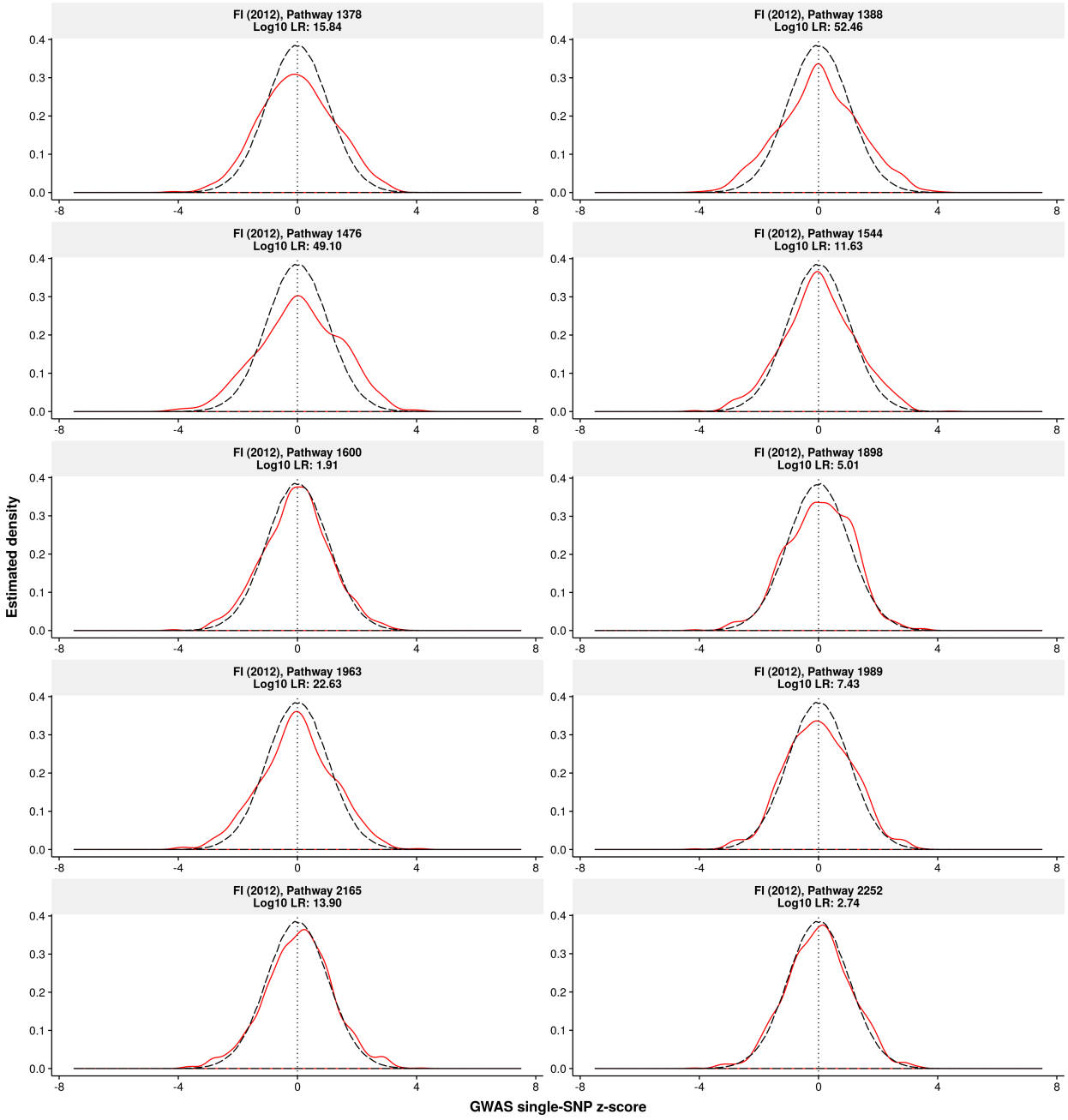
Note that there are only five pathways with at least 10 member genes in Round 2 enrichment analysis of depressive symptoms; see `http://xiangzhu.github.io/rss-gsea/results/ds2016-2.l`



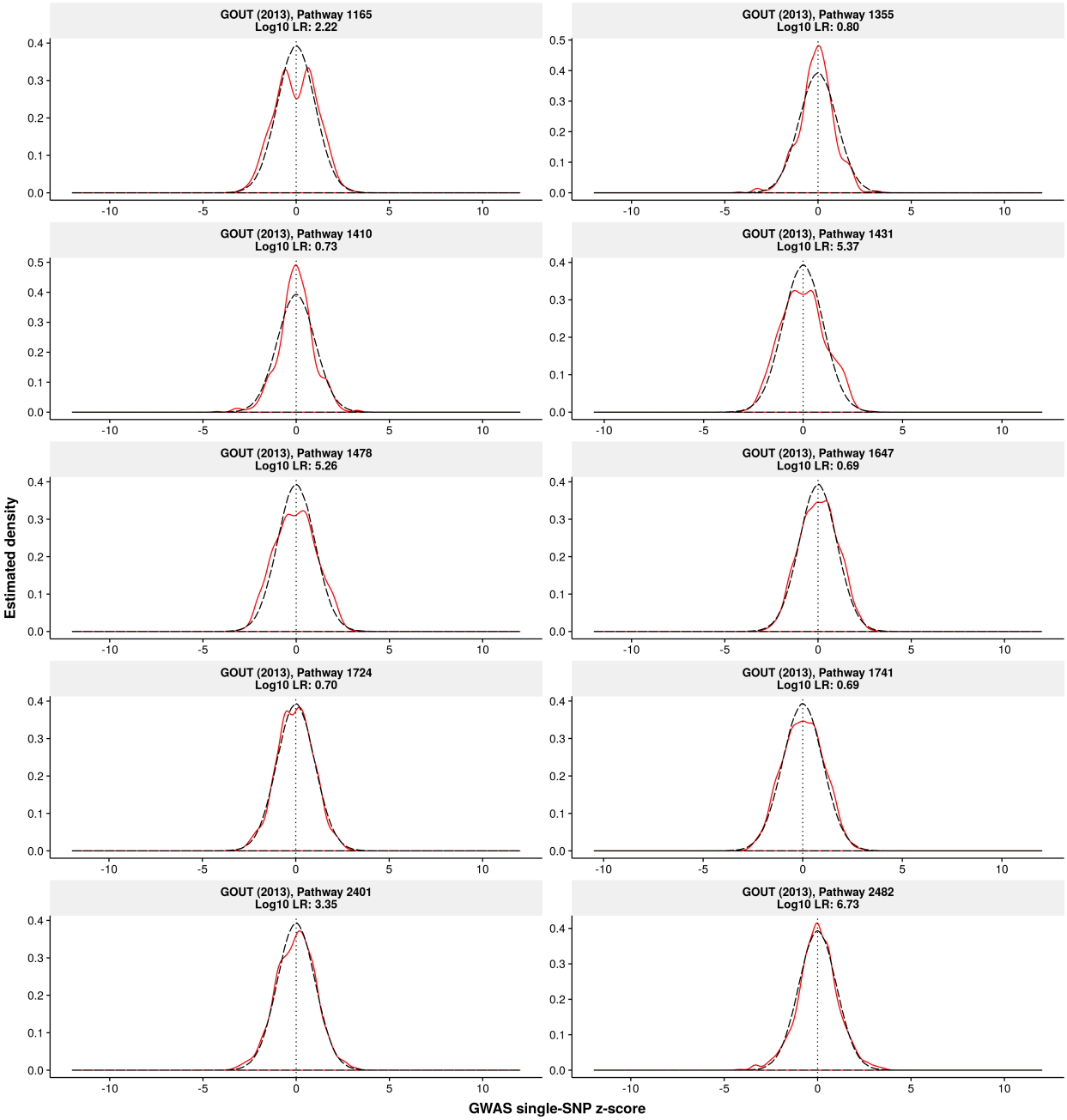
Fasting glucose levels (Manning et al., 2012)



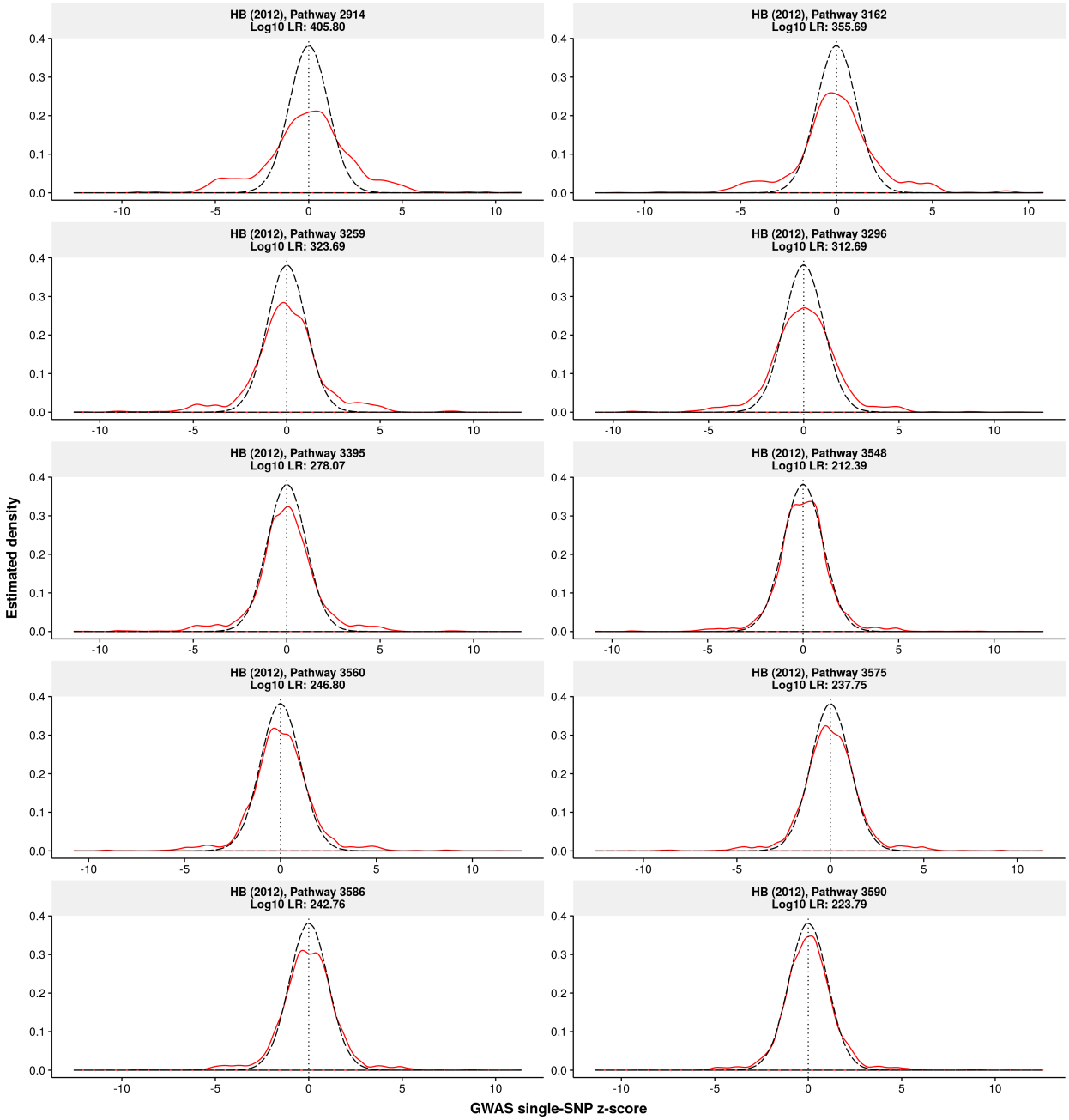
Fasting insulin levels (Manning et al., 2012)



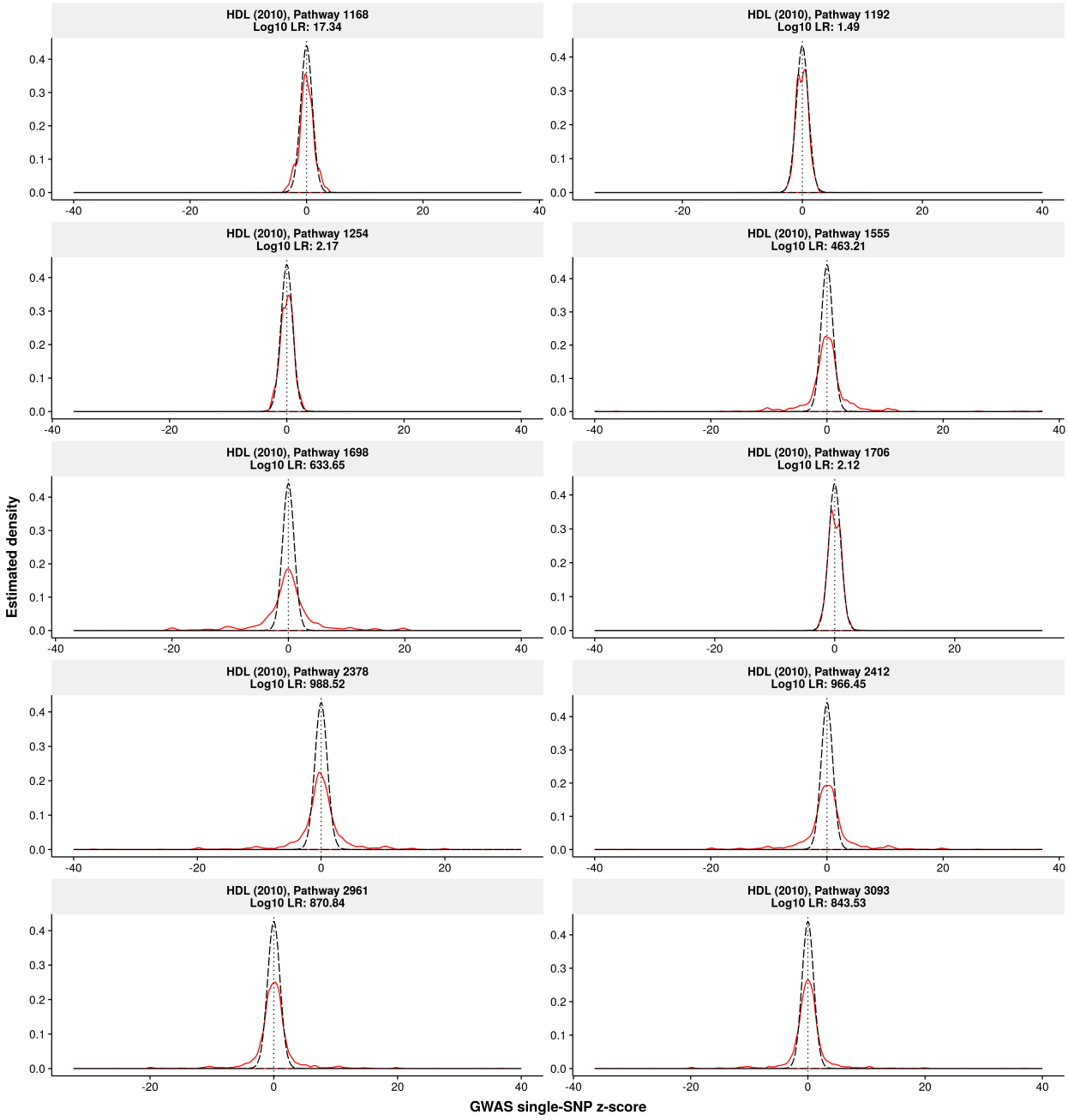
Gout (Köttgen et al., 2013)



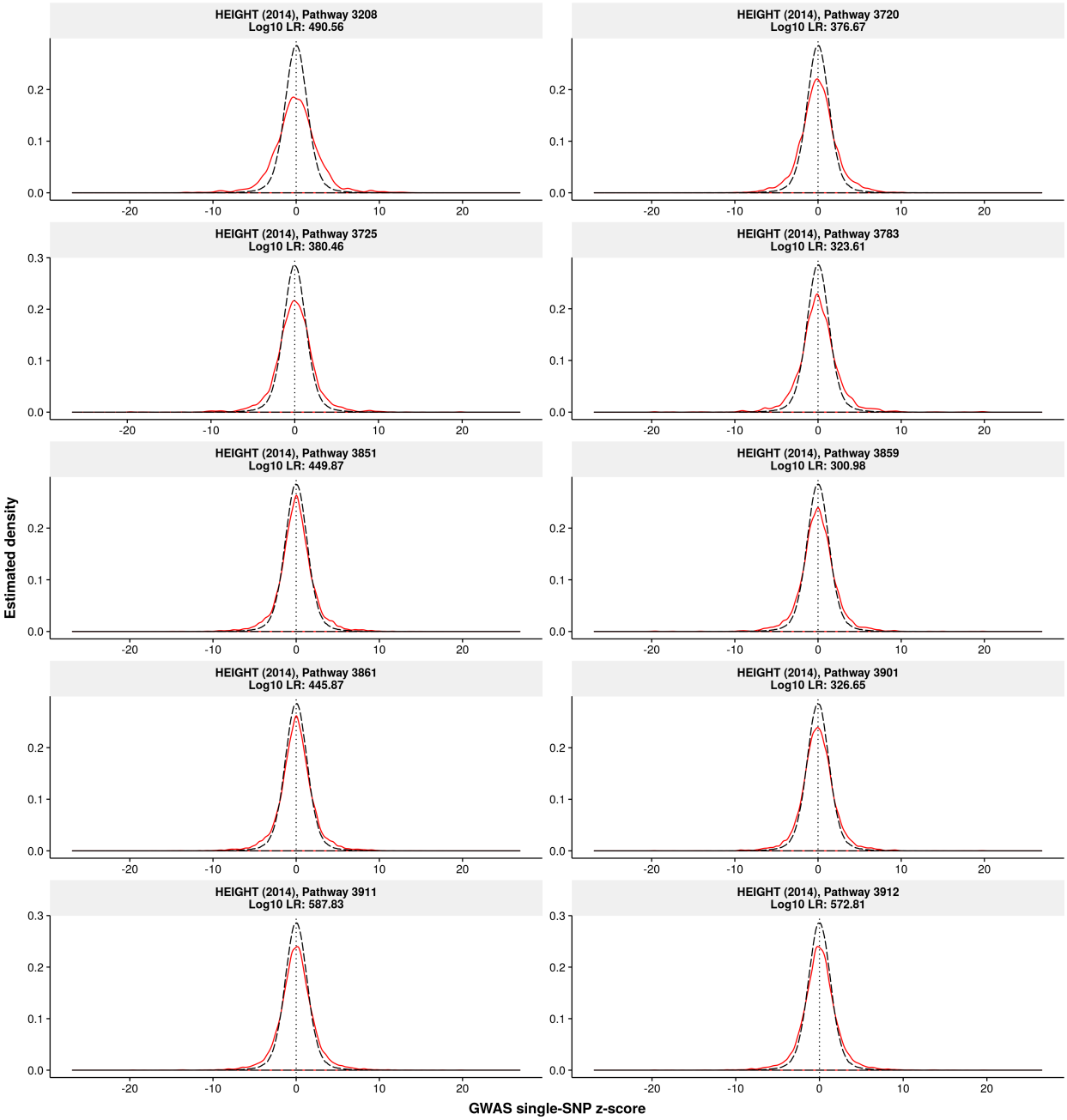
Haemoglobin (van der Harst et al., 2012)



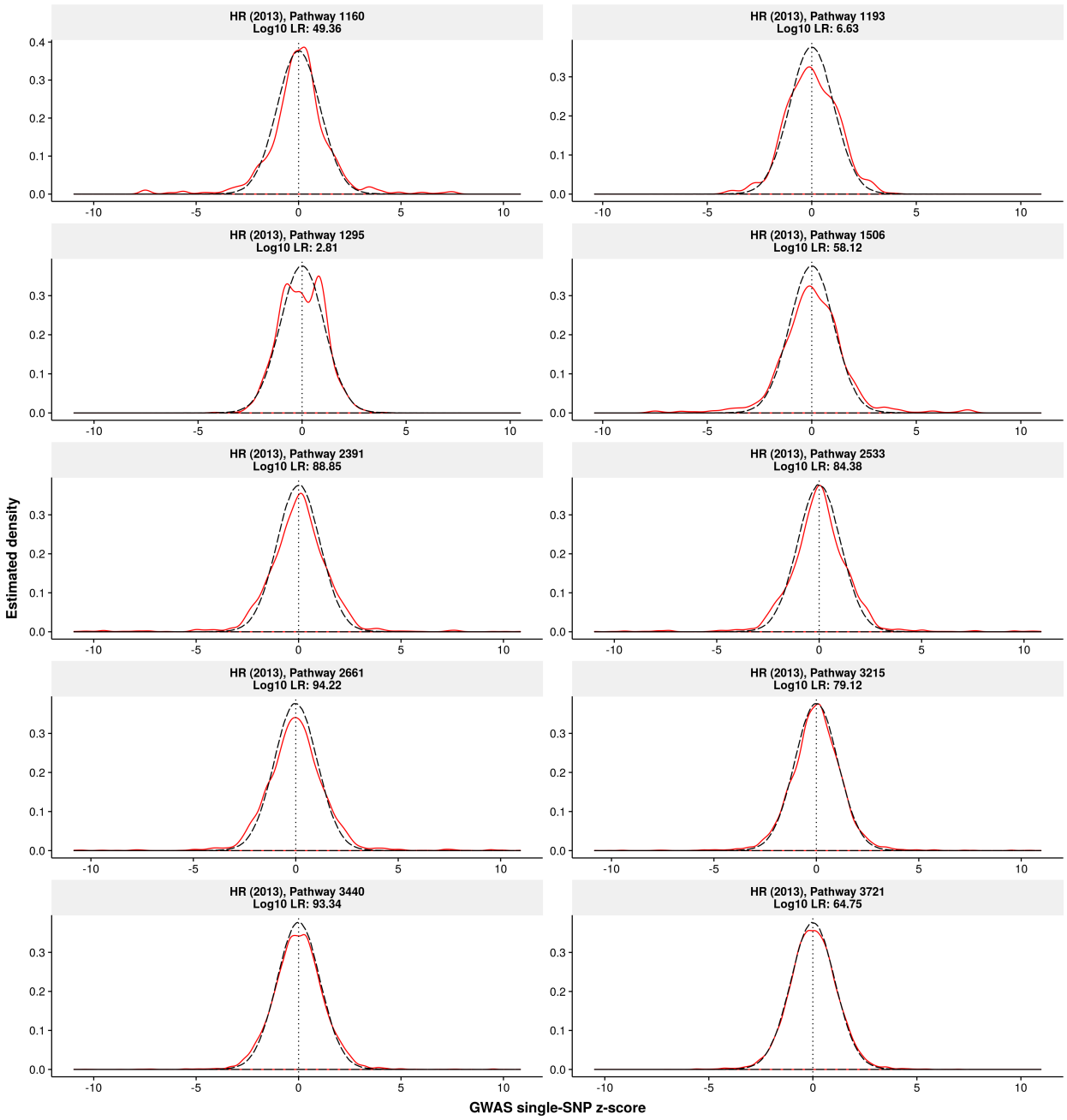
High-density lipoprotein (Teslovich et al., 2010)



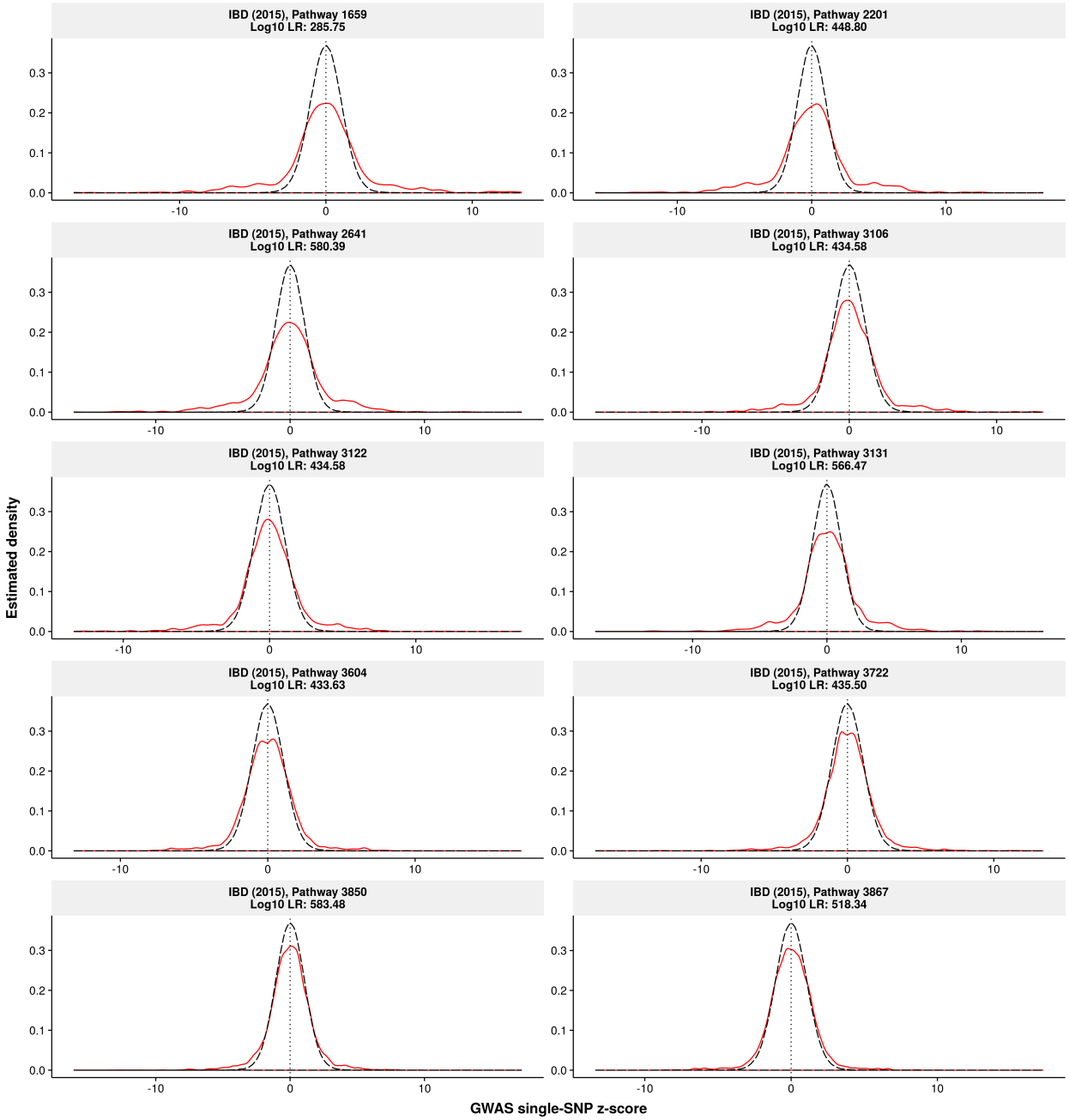
Adult height (Wood et al., 2014)



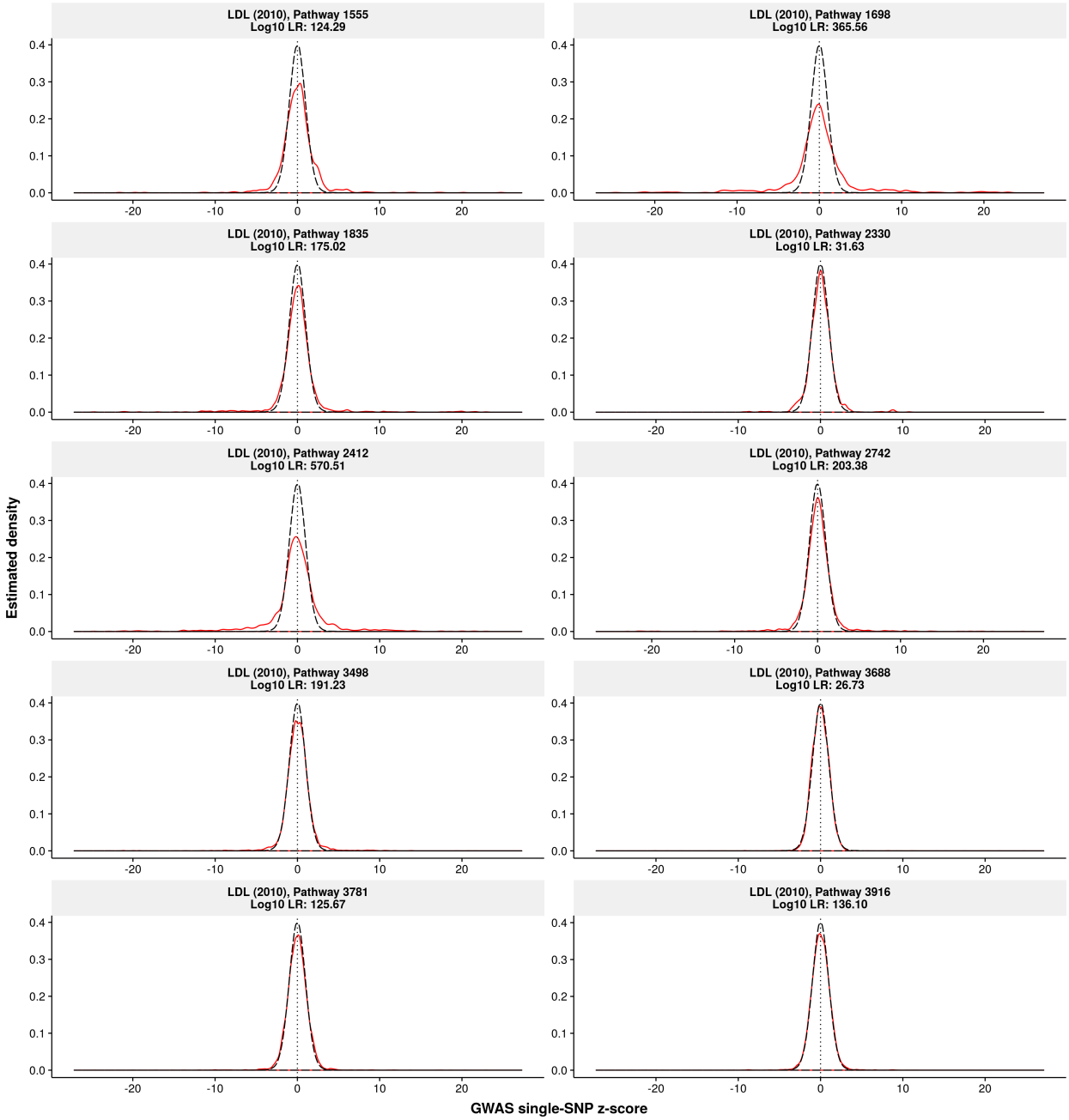
Heart rate (Den Hoed et al., 2013)



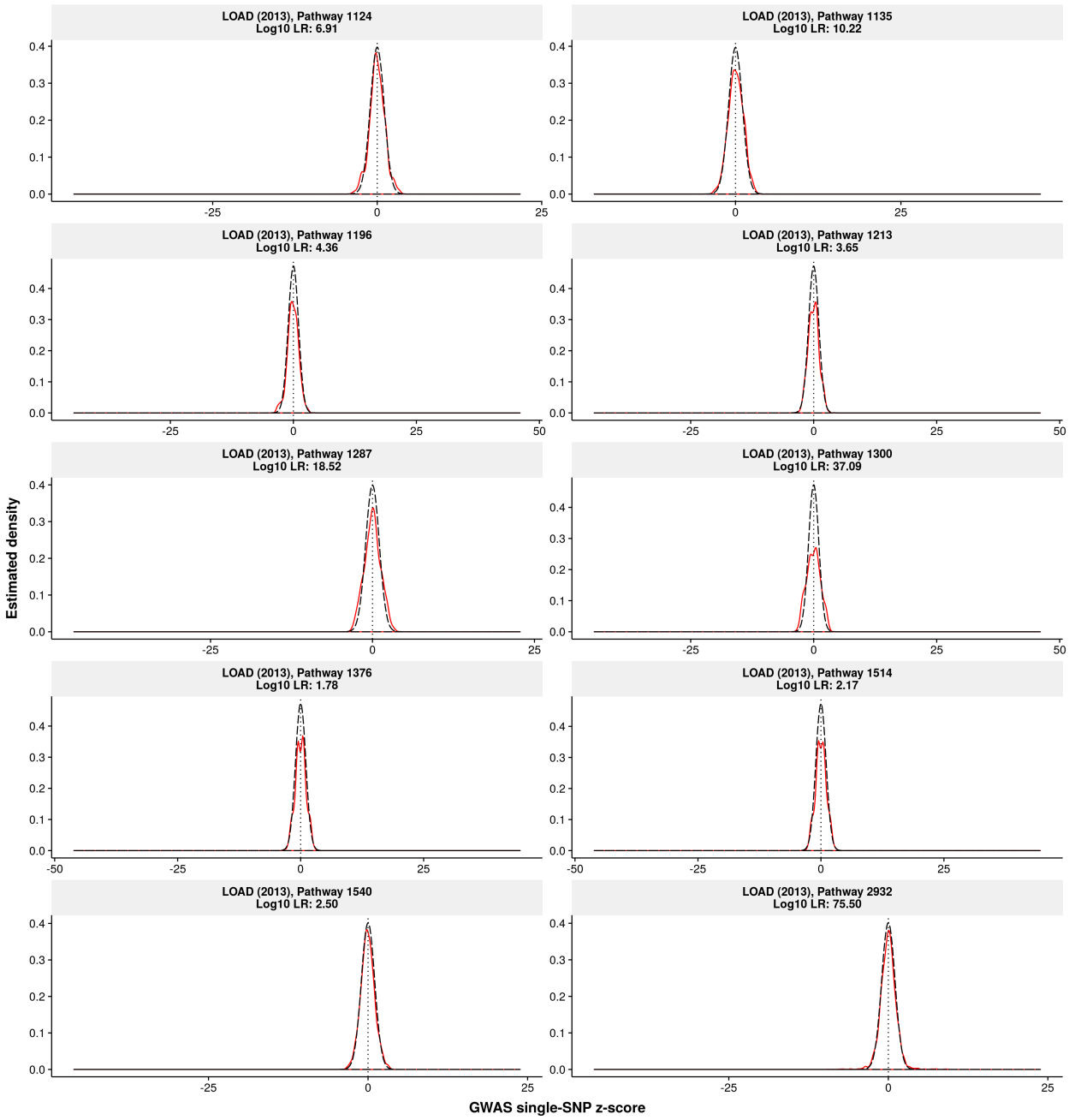
Inflammatory bowel disease (Liu et al., 2015)



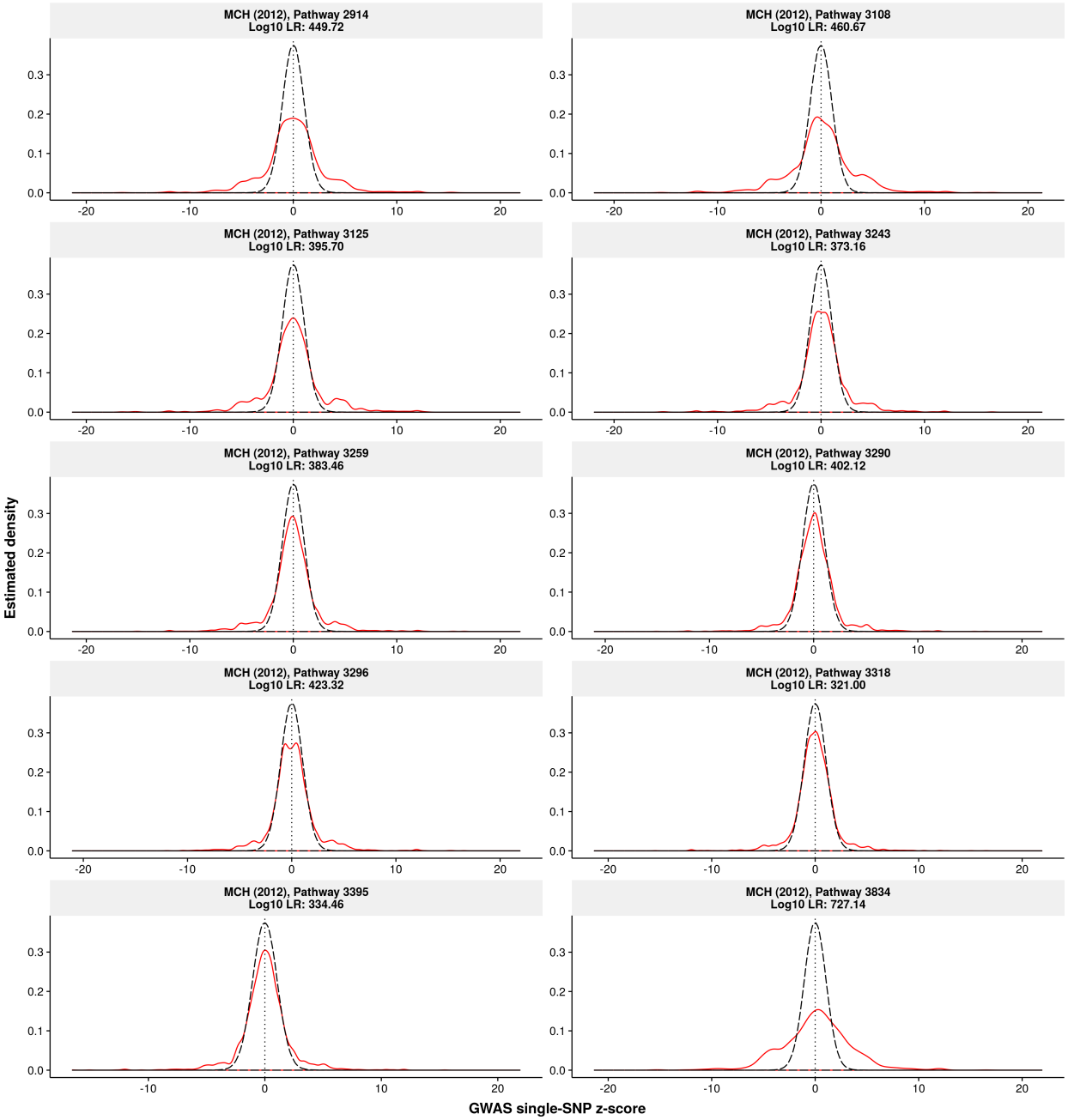
Low-density lipoprotein (Teslovich et al., 2010)



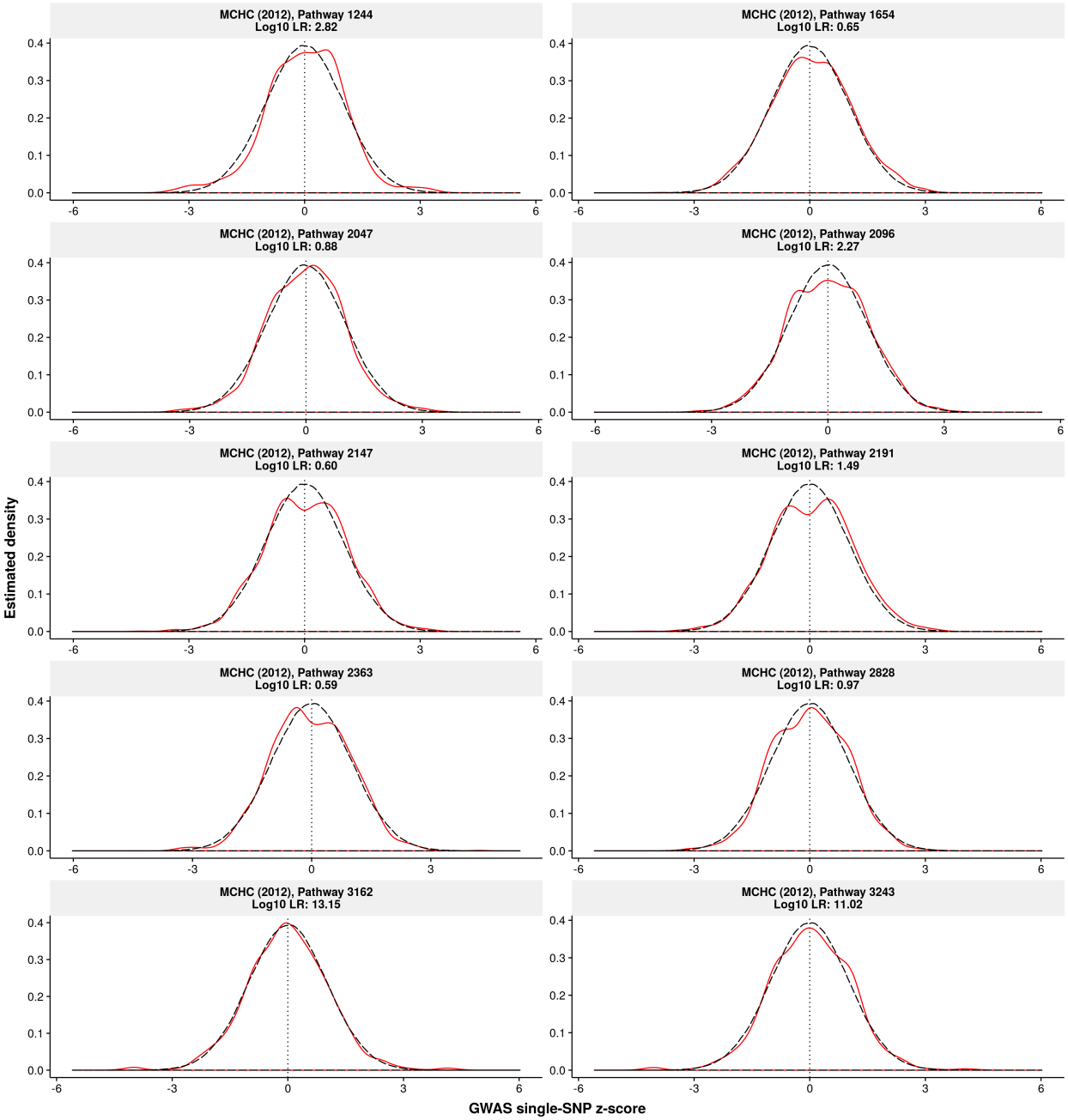
Alzheimer's disease (Lambert et al., 2013)



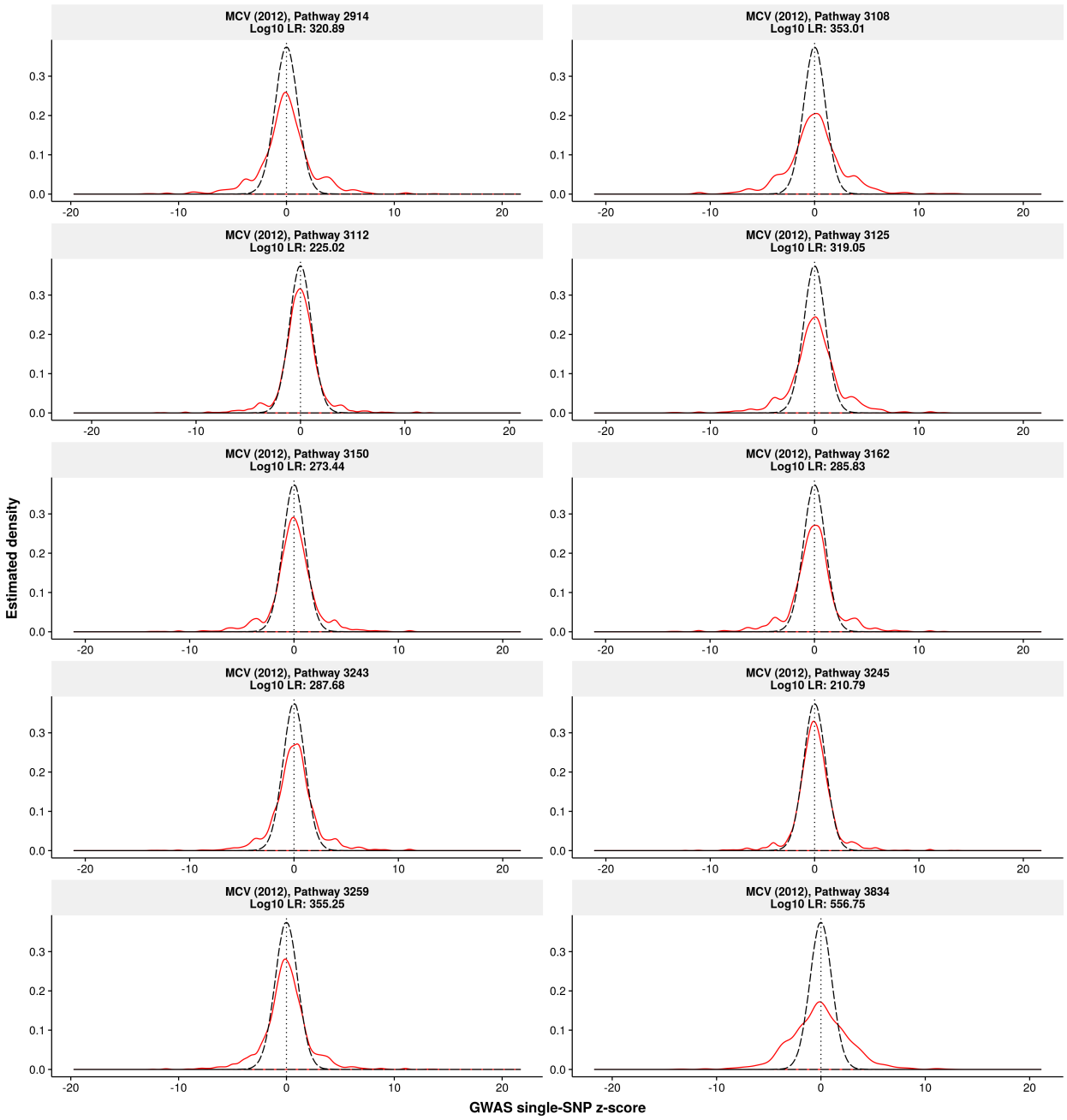
Mean cell haemoglobin (van der Harst et al., 2012)



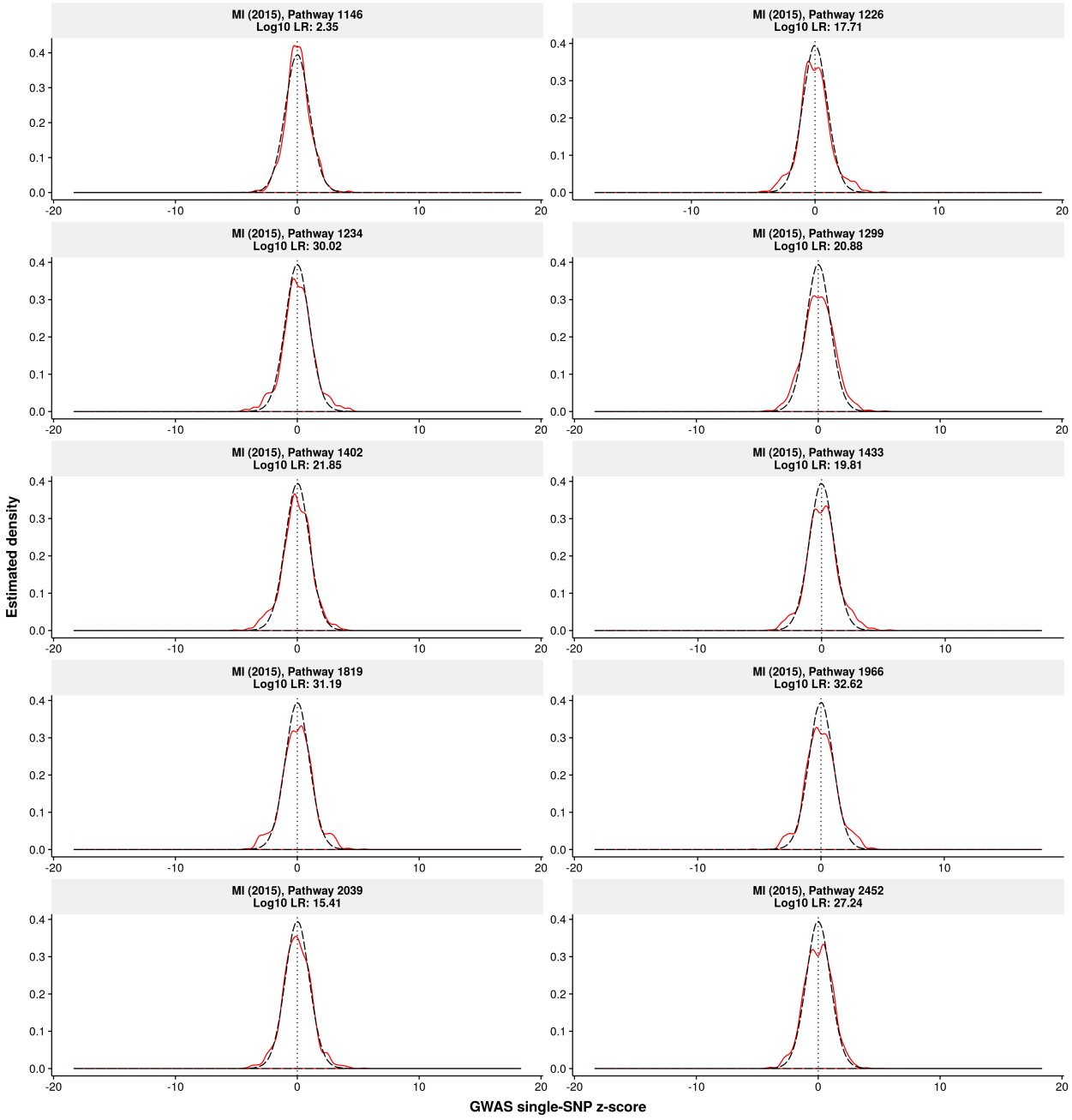
Mean cell haemoglobin concentration (van der Harst et al., 2012)



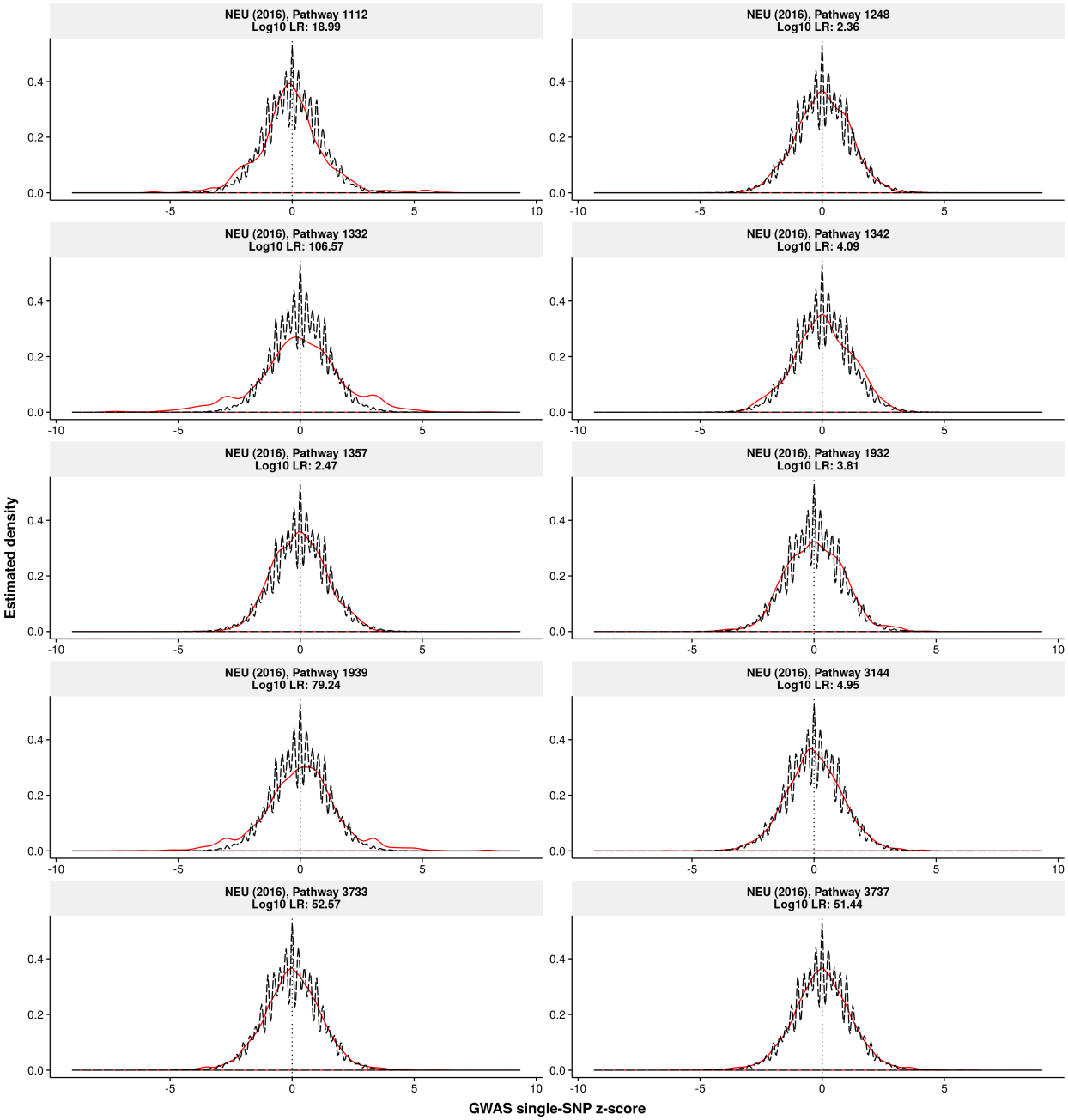
Mean cell volume (van der Harst et al., 2012)



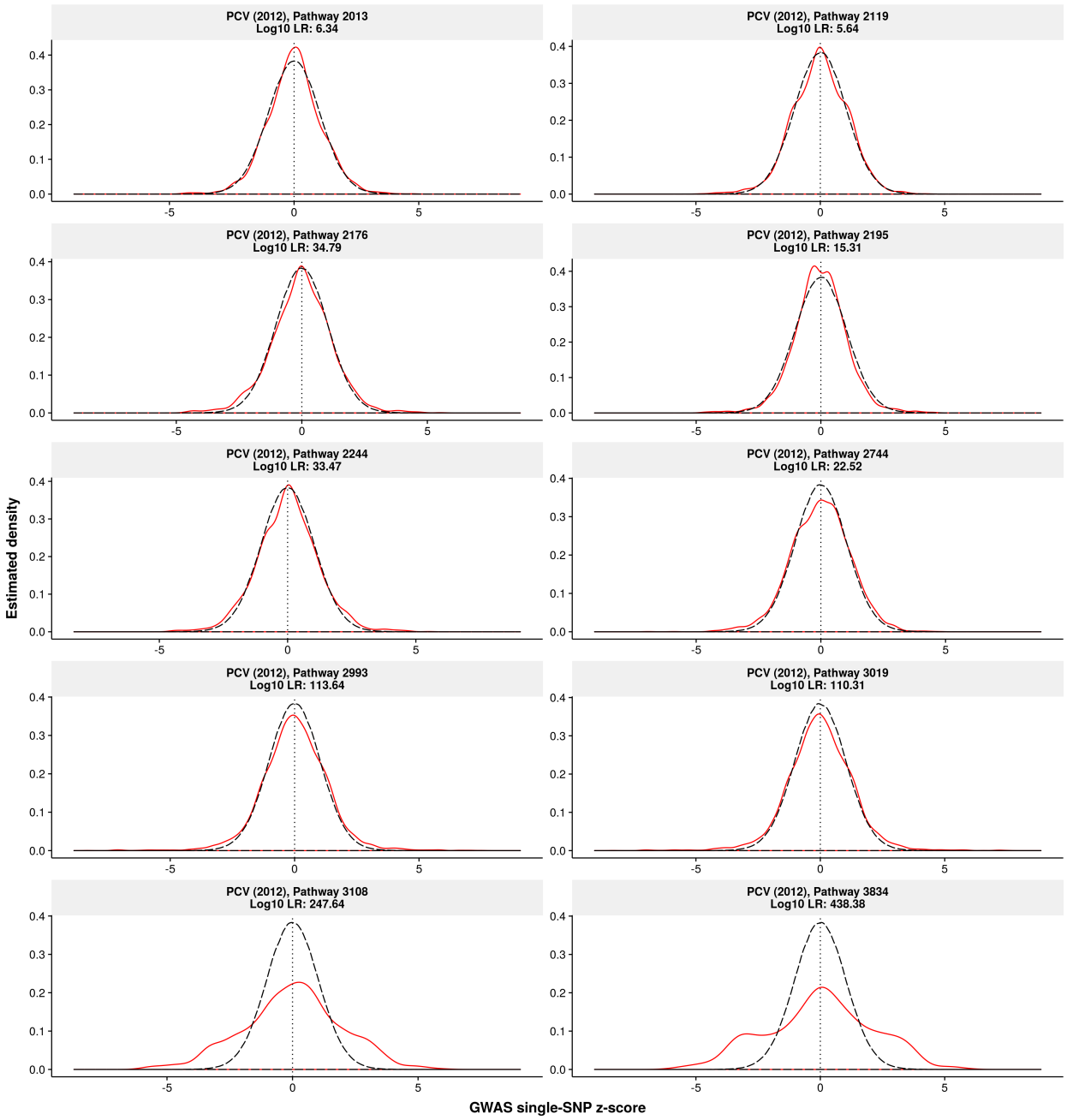
Myocardial infarction (Nikpay et al., 2015)



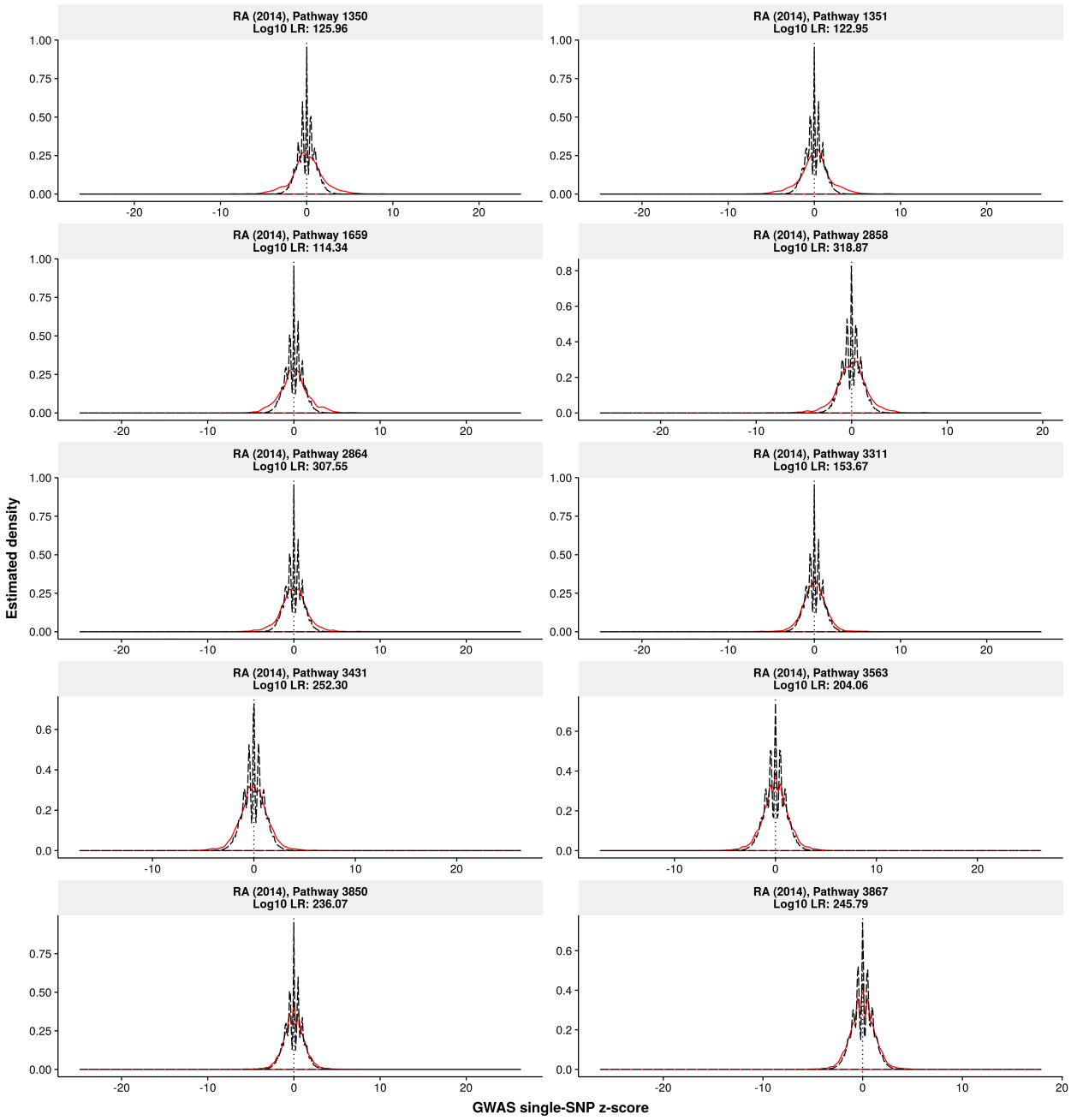
Neuroticism (Okbay et al., 2016)



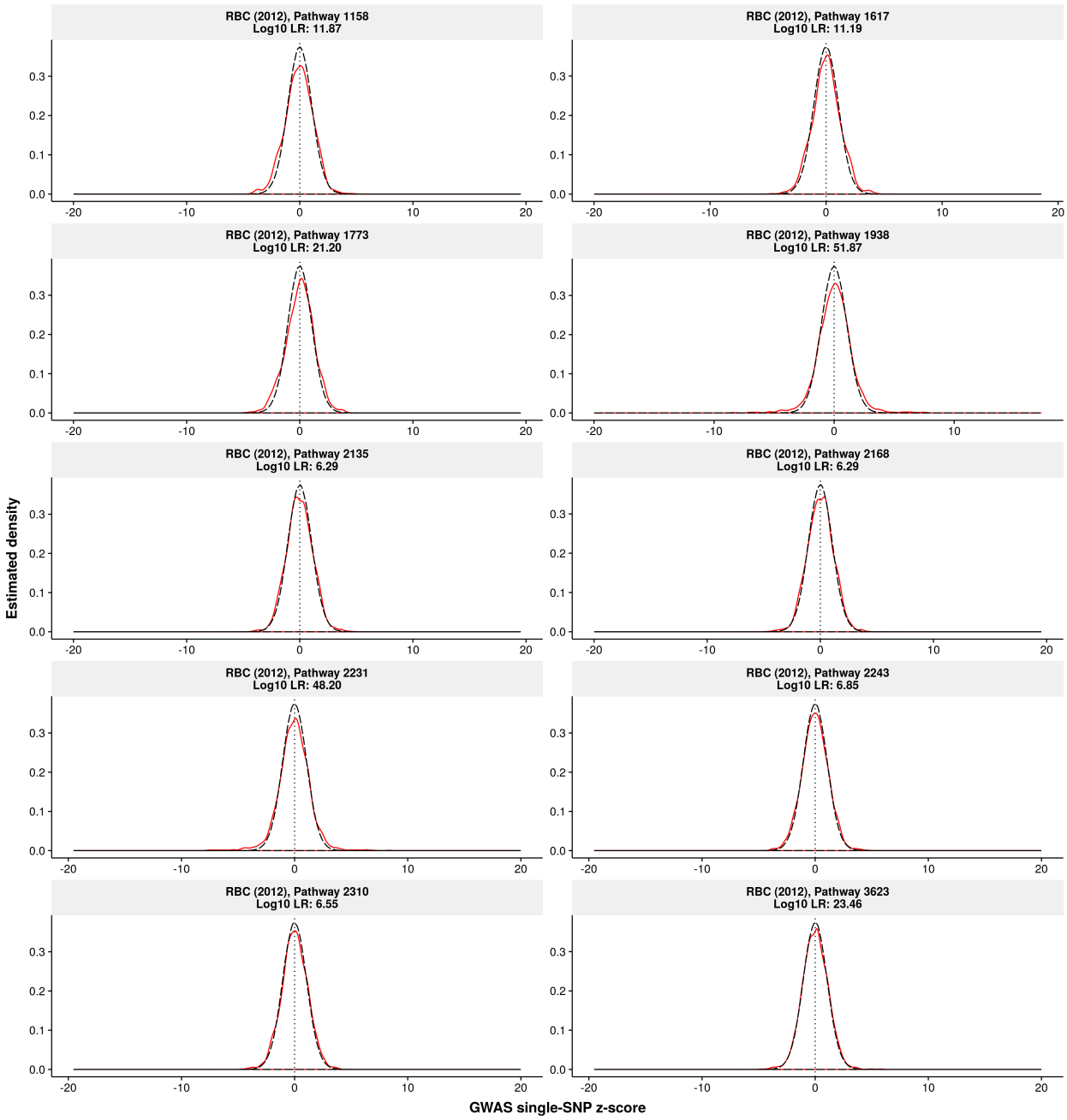
Packed cell volume (van der Harst et al., 2012)



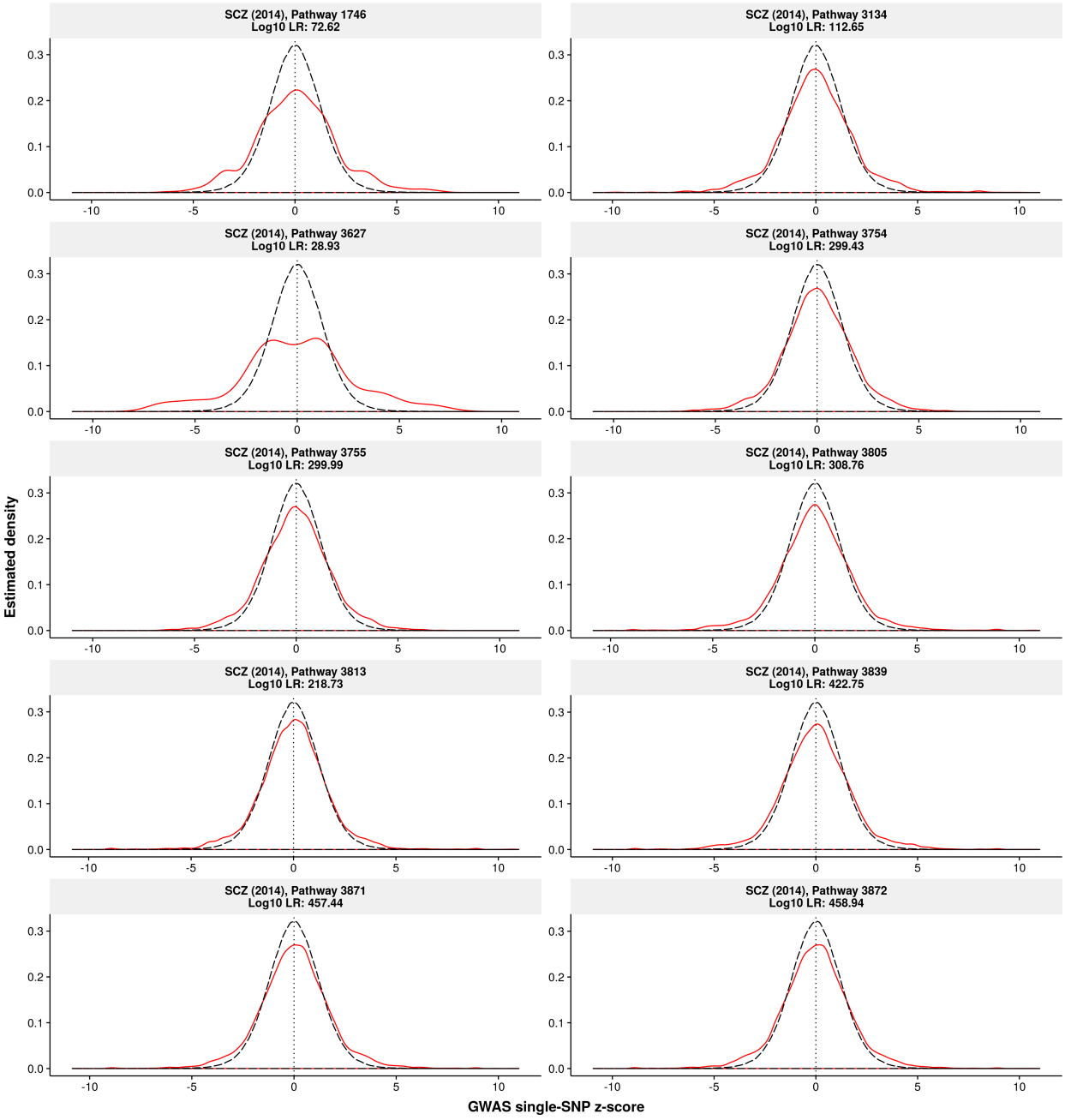
Rheumatoid arthritis (Okada et al., 2014)



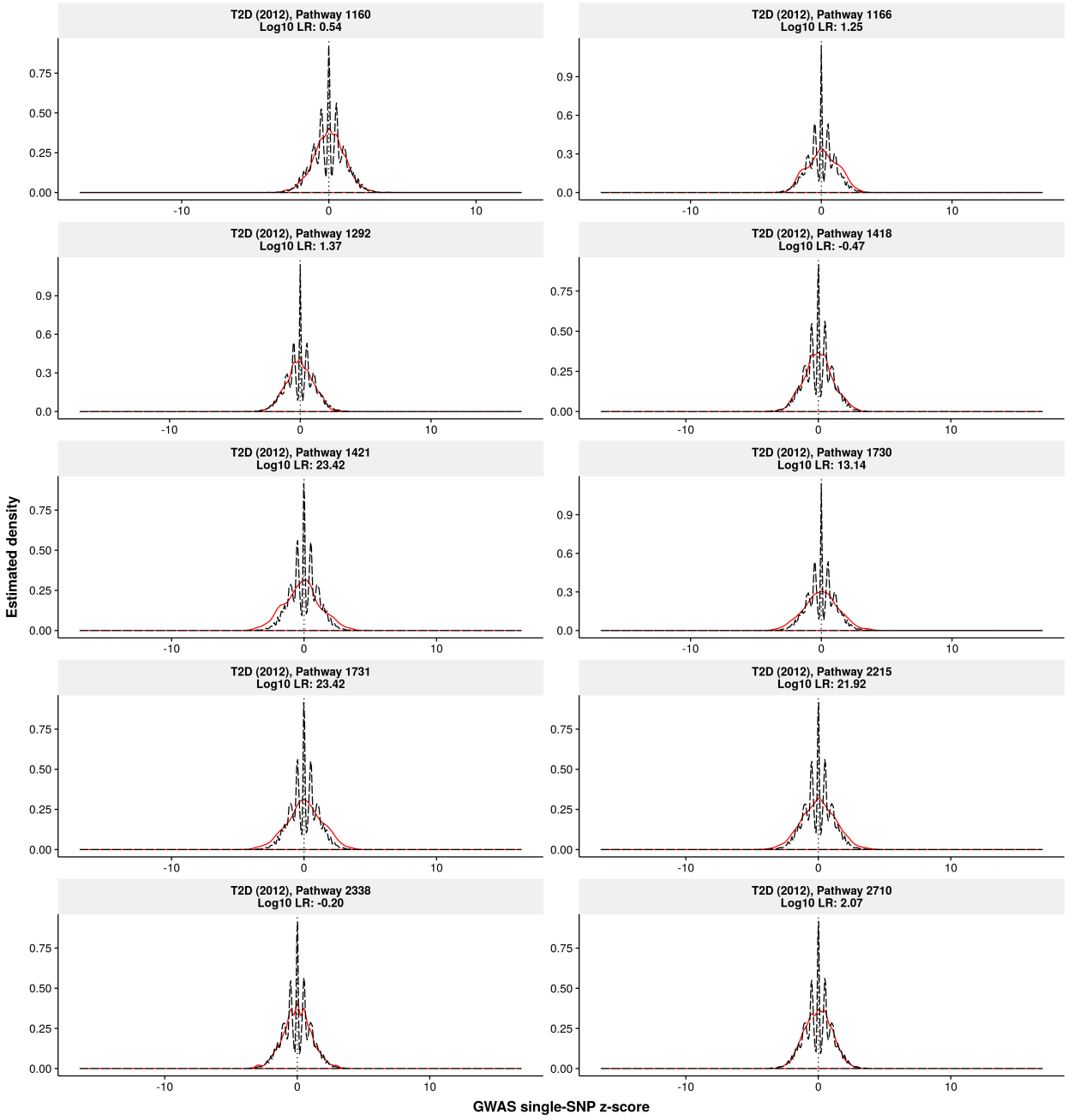
Red blood cell count (van der Harst et al., 2012)



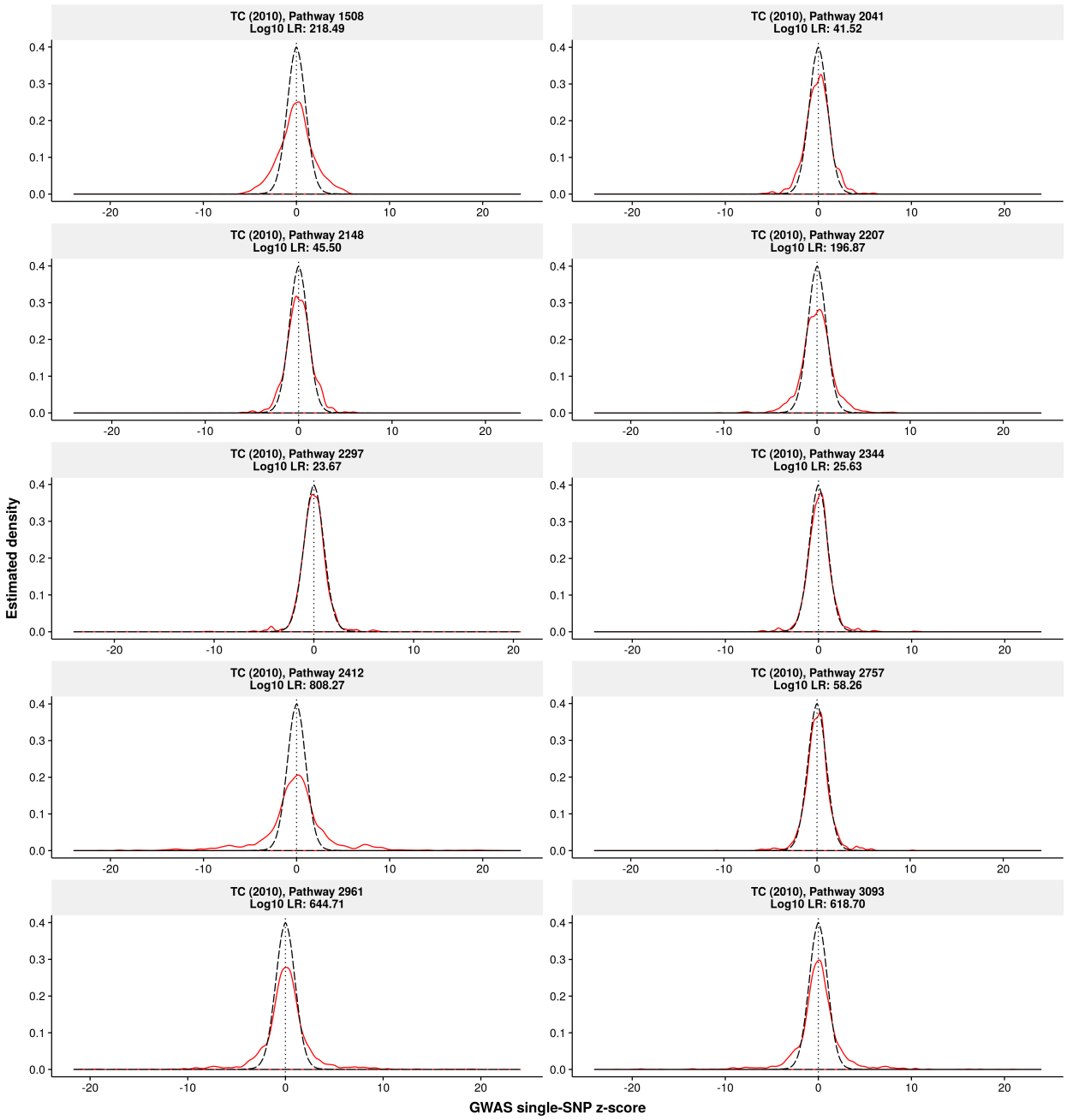
Schizophrenia (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014)



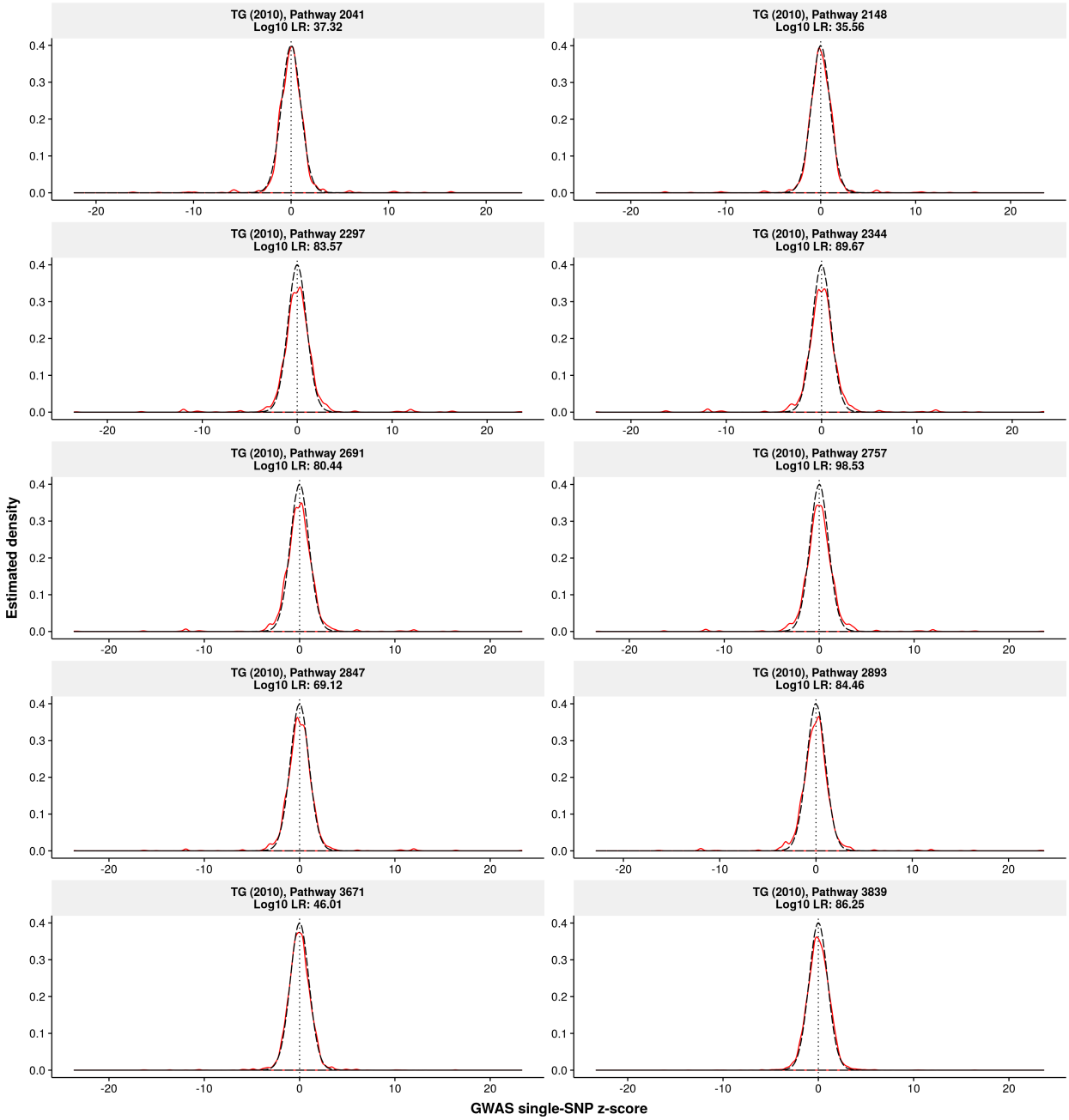
Type 2 diabetes (Morris et al., 2012)



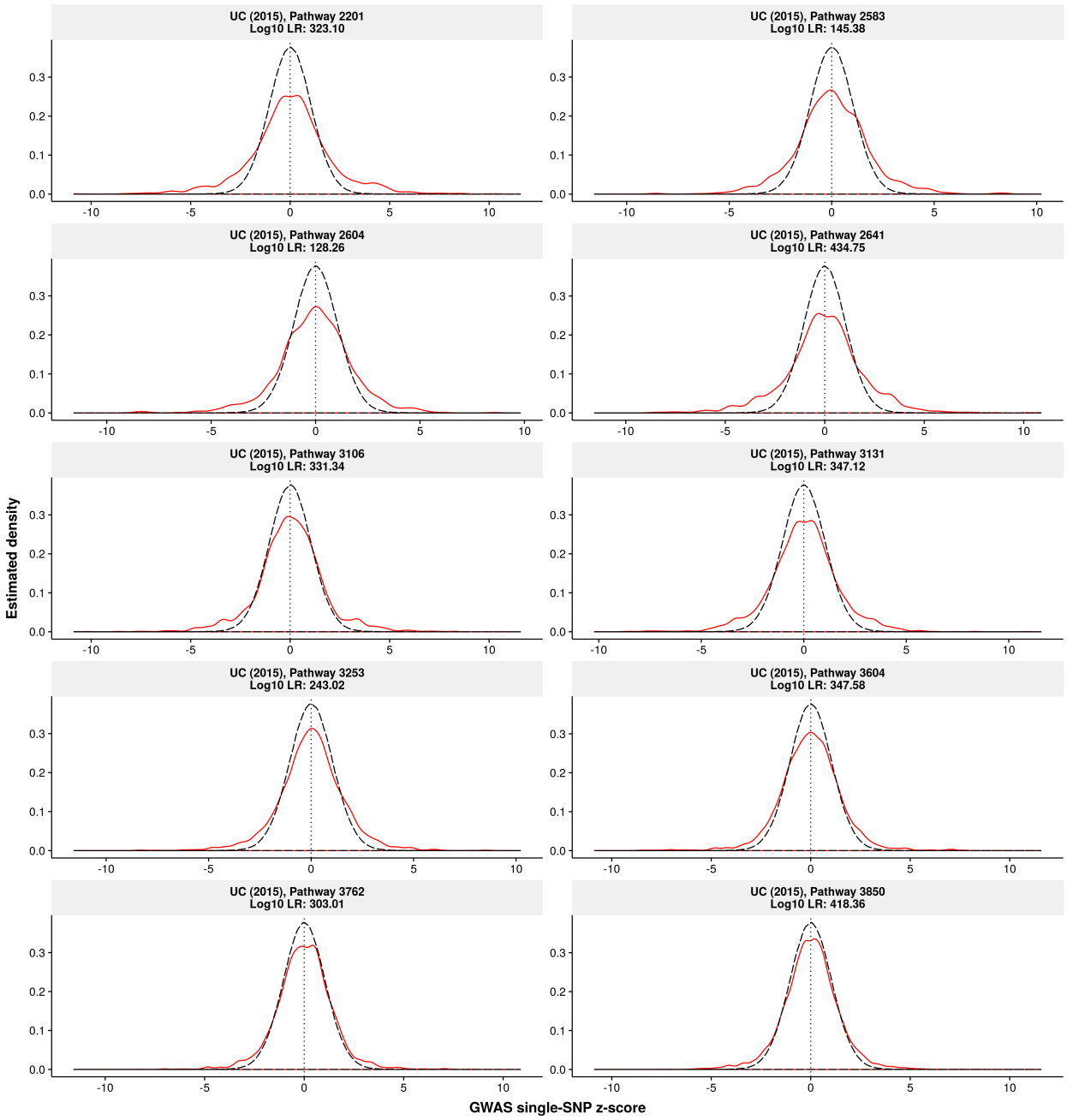
Total cholesterol (Teslovich et al., 2010)

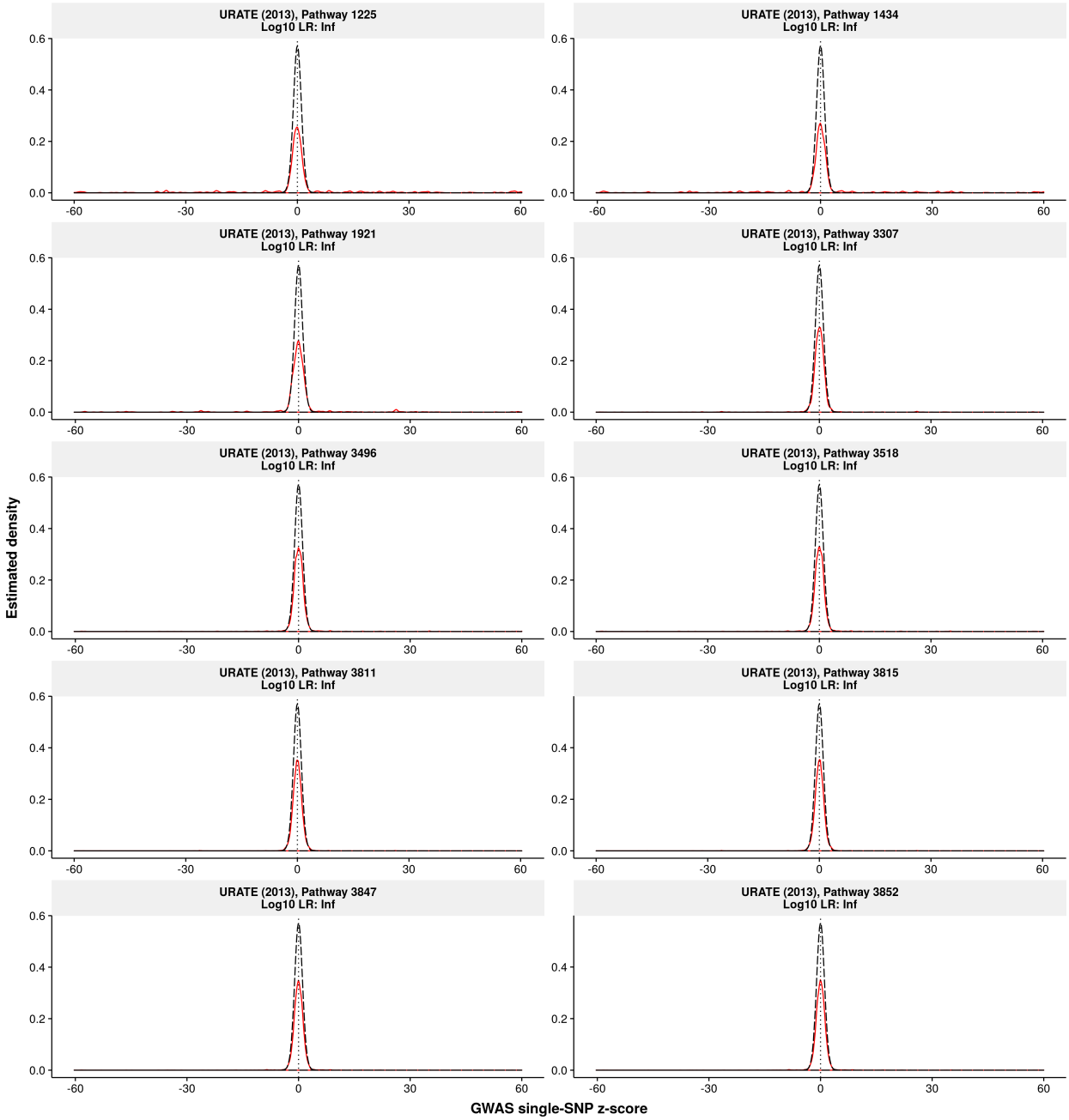


Triglycerides (Teslovich et al., 2010)

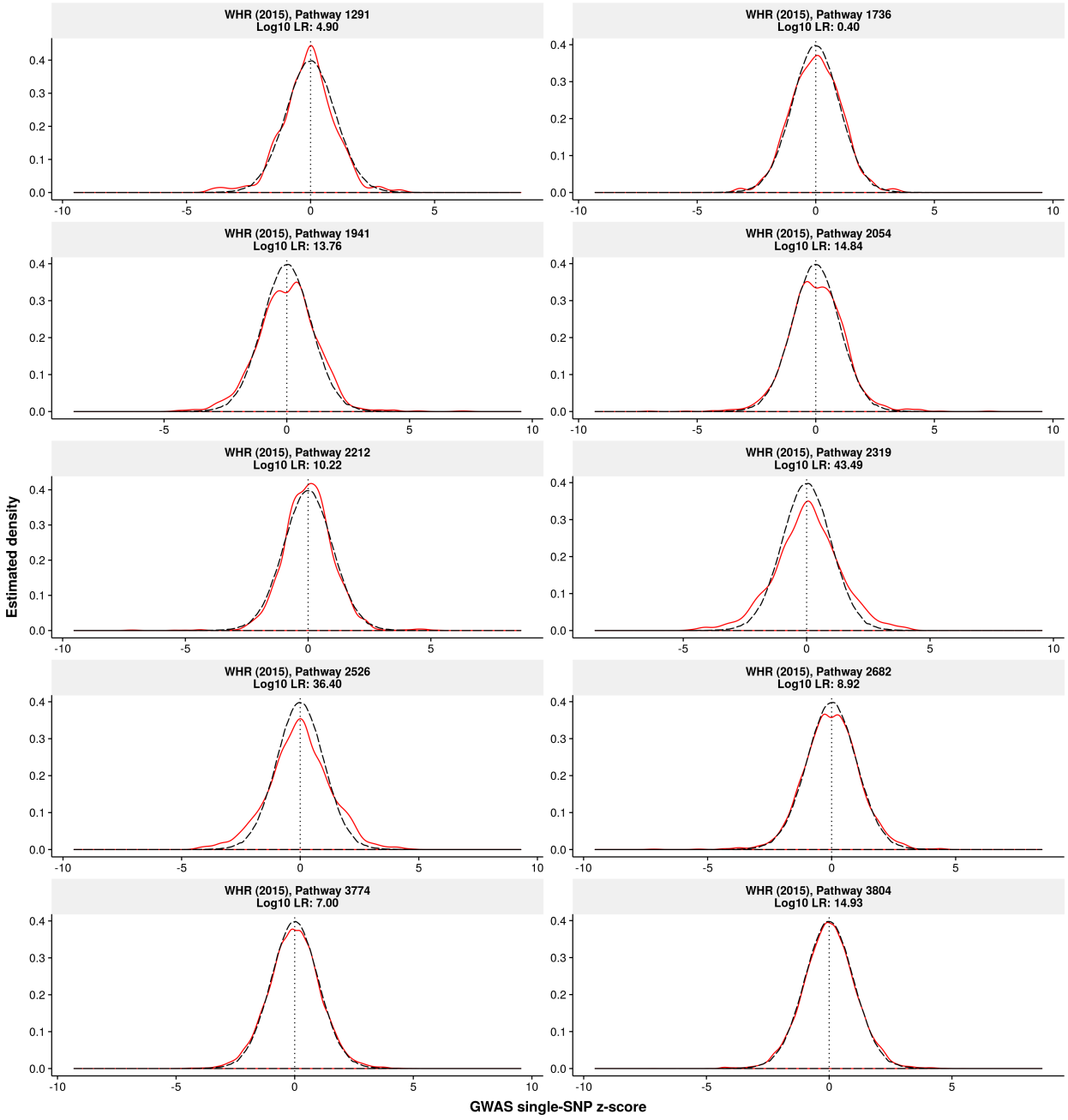


Ulcerative colitis (Liu et al., 2015)




URATE (2013), Pathway 1225
Log10 LR: Inf
URATE (2013), Pathway 1434
Log10 LR: Inf
URATE (2013), Pathway 1921
Log10 LR: Inf
URATE (2013), Pathway 3307
Log10 LR: Inf
URATE (2013), Pathway 3496
Log10 LR: Inf
URATE (2013), Pathway 3518
Log10 LR: Inf
URATE (2013), Pathway 3811
Log10 LR: Inf
URATE (2013), Pathway 3815
Log10 LR: Inf
URATE (2013), Pathway 3847
Log10 LR: Inf
URATE (2013), Pathway 3852
Log10 LR: Inf
Estimated density
0.0
0.2
0.4
0.6
-60
-30
0
30
60
GWAS single-SNP z-score

Waist-to-hip ratio adjusted for body mass index (Shungin et al., 2015)



Tissue-trait pairs reported in Table 2 of Zhu and Stephens (2017b)

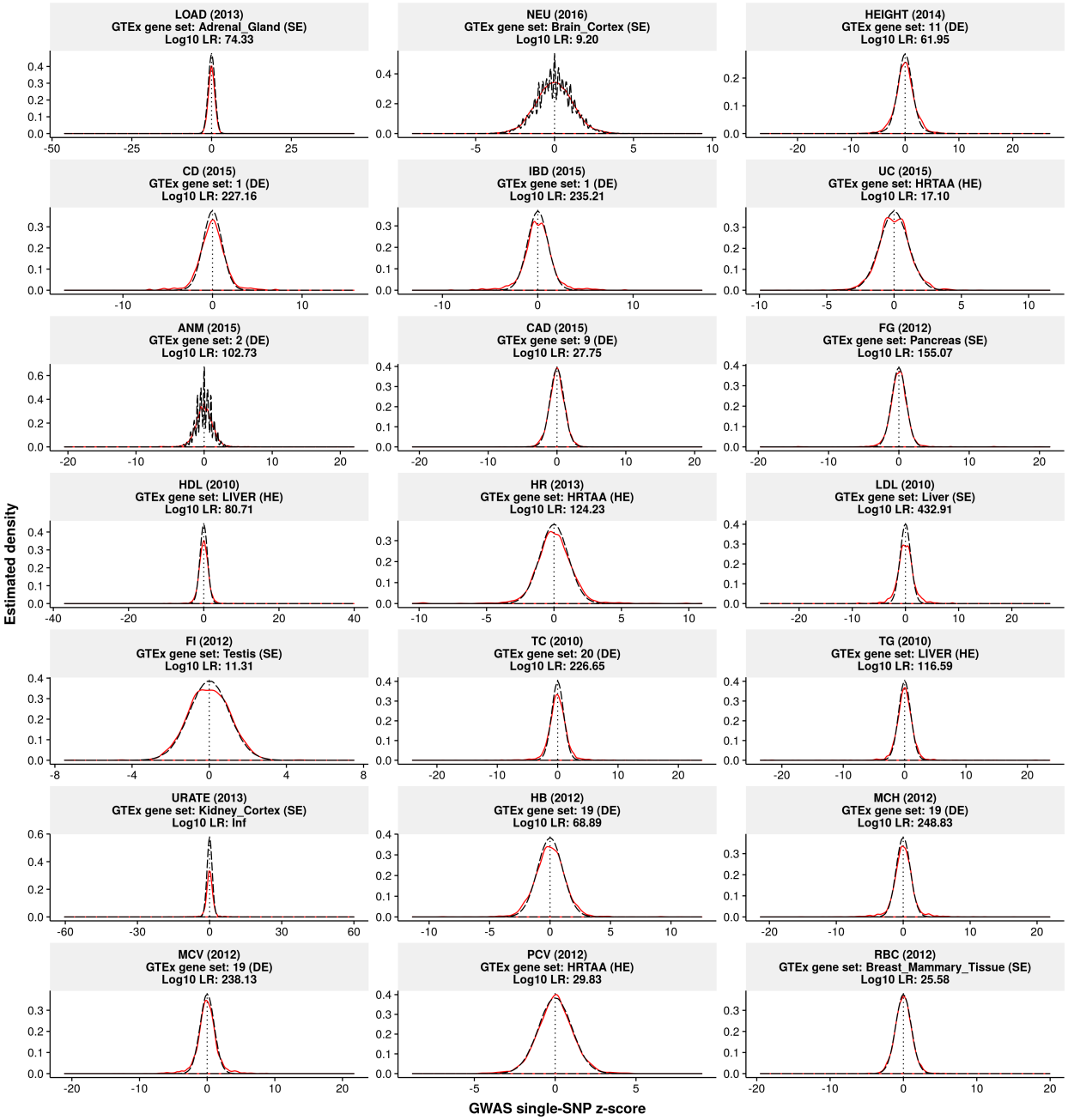


Figure F.10: Sanity checks of top-ranked gene set enrichments for 31 phenotypes.

Supplementary Figure 11

Bayes factors for enrichment of genetic associations near all genes in 31 phenotypes. For each phenotype, the enrichment hypothesis is that SNPs within ± 100 kb of the transcribed region of any protein-coding gene [18,313 in total in the present study; Supplementary Figure 9 of Zhu and Stephens (2017b)] are more likely to be trait-associated. The x-axis uses a normal scale inside the range [-1.5, 1.5], and a logarithmic scale (base 10) outside (-1.5, 1.5).

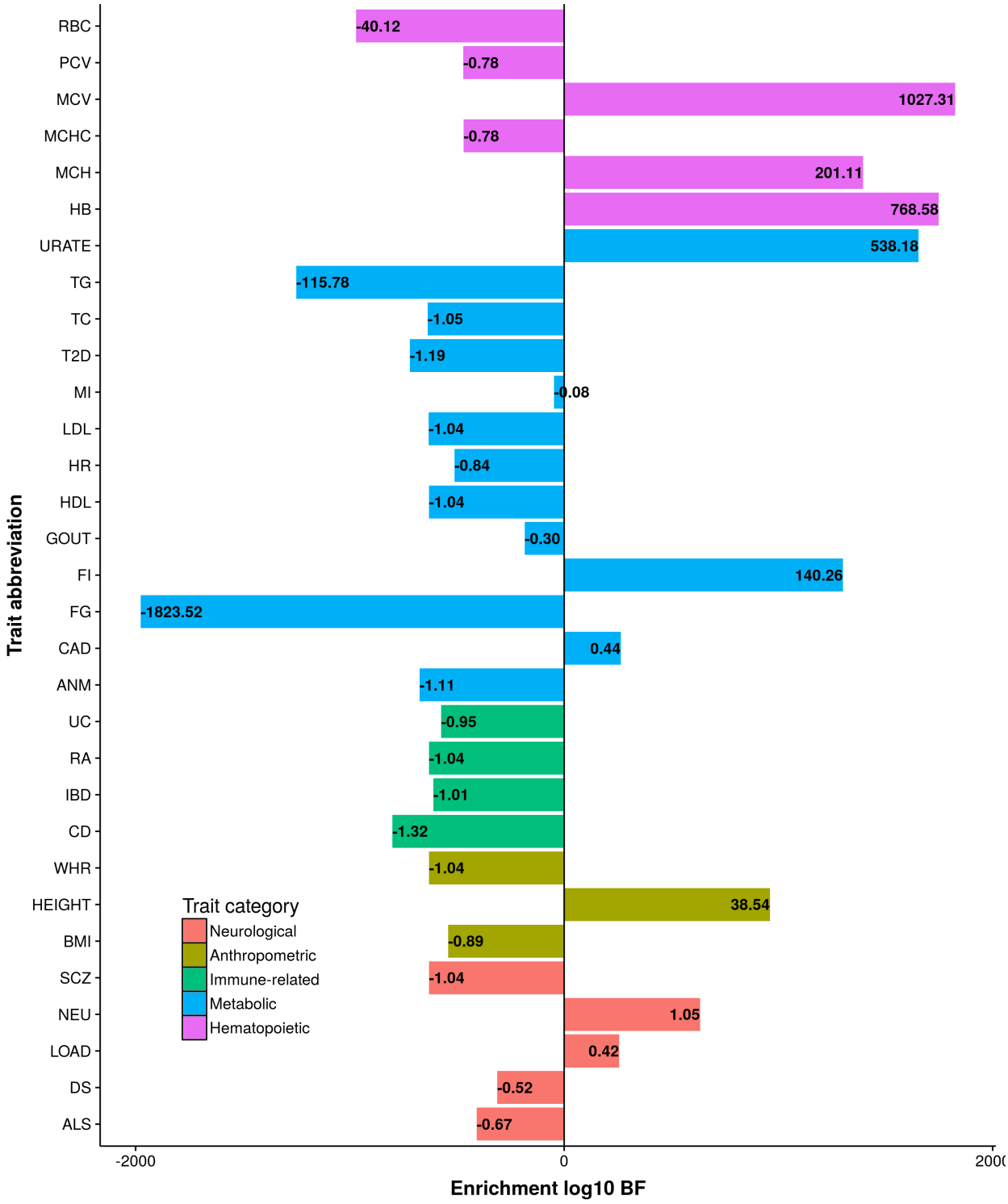


Figure F.11: Bayes factors for enrichment of genetic associations near all genes in 31 phenotypes.

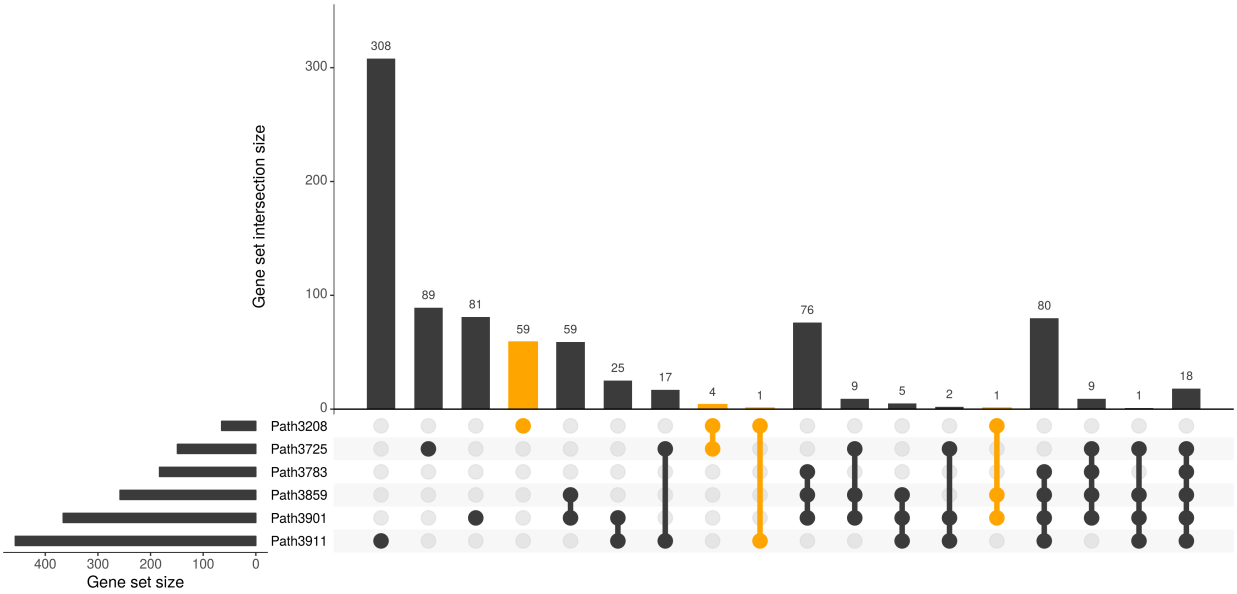
Supplementary Figure 12

Figure F.12: Distribution of Bayes factors for enrichment of 3,913 biological pathways in 31 phenotypes.

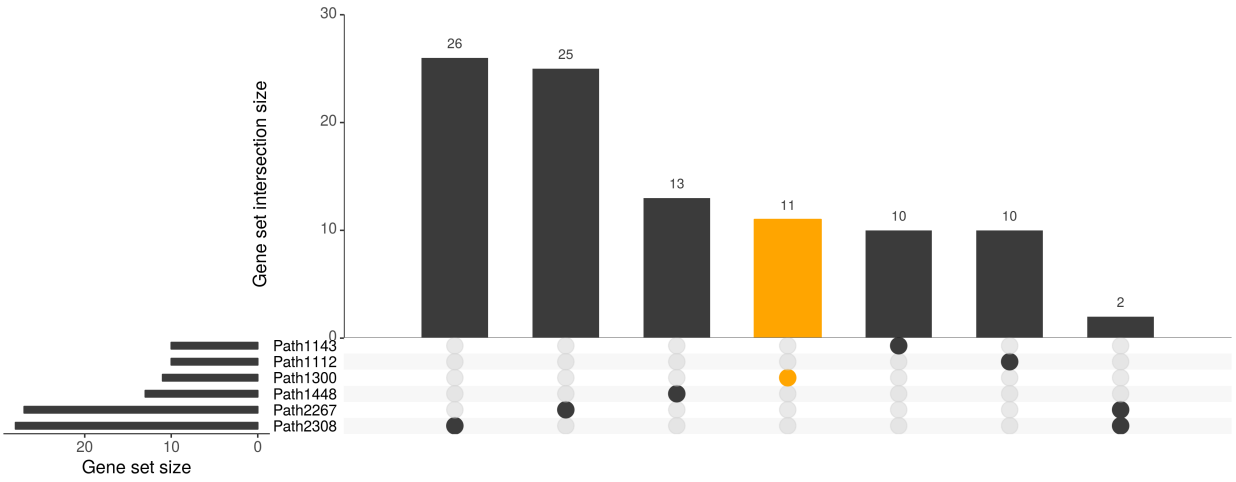
Supplementary Figure 13

Gene set overlap among top 6 most enriched pathways for each of 31 phenotypes. For each trait, yellow bars correspond to the most enriched pathway (i.e. pathways showing the largest enrichment BFs). When selecting the top enriched pathways, if multiple pathways from different databases have the same description/name, only the one with the largest BF is displayed here. Full information about the enriched pathways can be found at `http://xiangzhu.github.io/rss-gsea/results`. Intersections of top six pathways are visualized as a UpSet plot (Lex et al., 2014), using `R` package `UpSetR`.

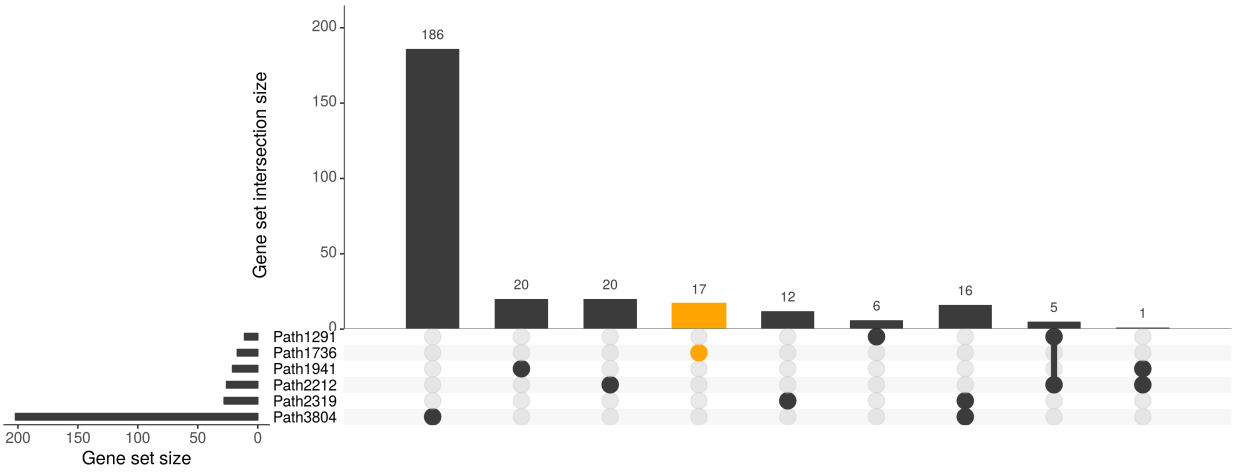
(a) Adult height (Wood et al., 2014).



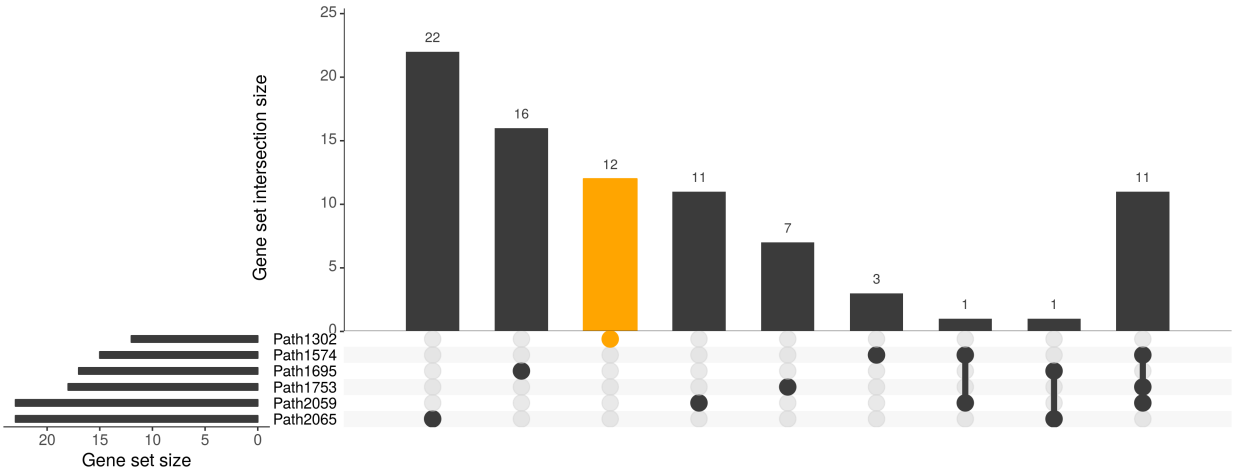
(b) Body mass index (Locke et al., 2015).



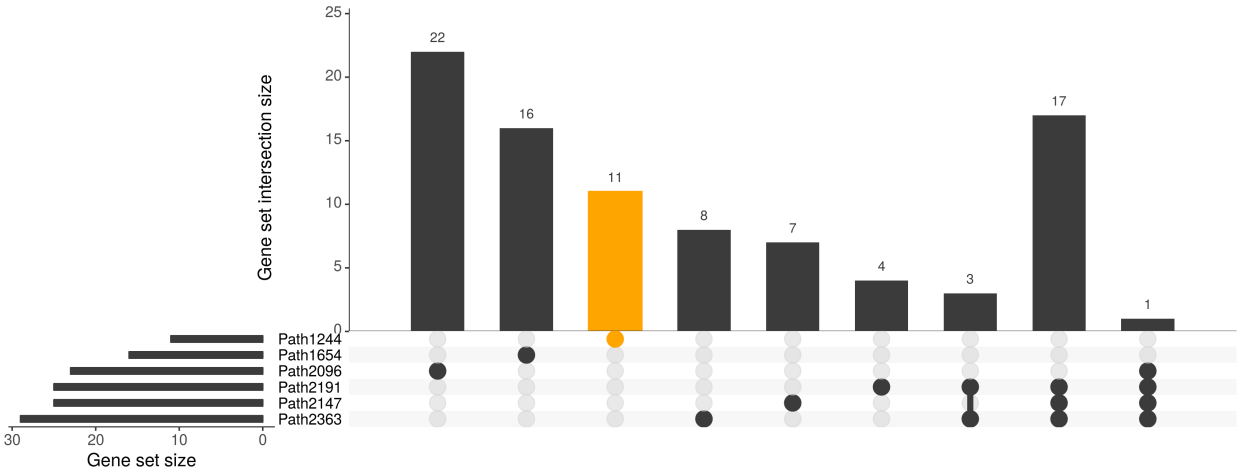
(c) Waist-to-hip ratio adjusted for body mass index (Shungin et al., 2015).

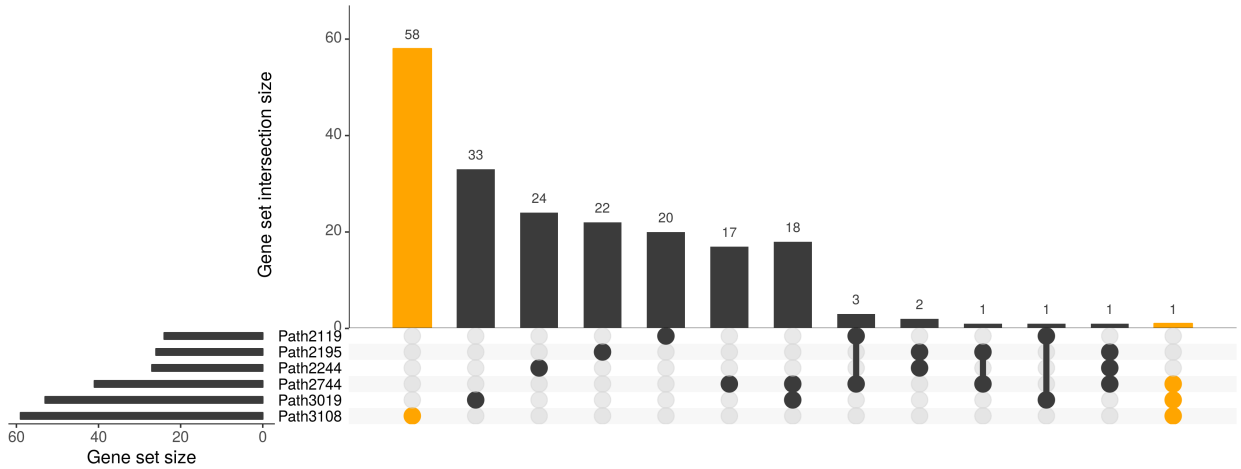


(d) Amyotrophic lateral sclerosis (van Rheenen et al., 2016).

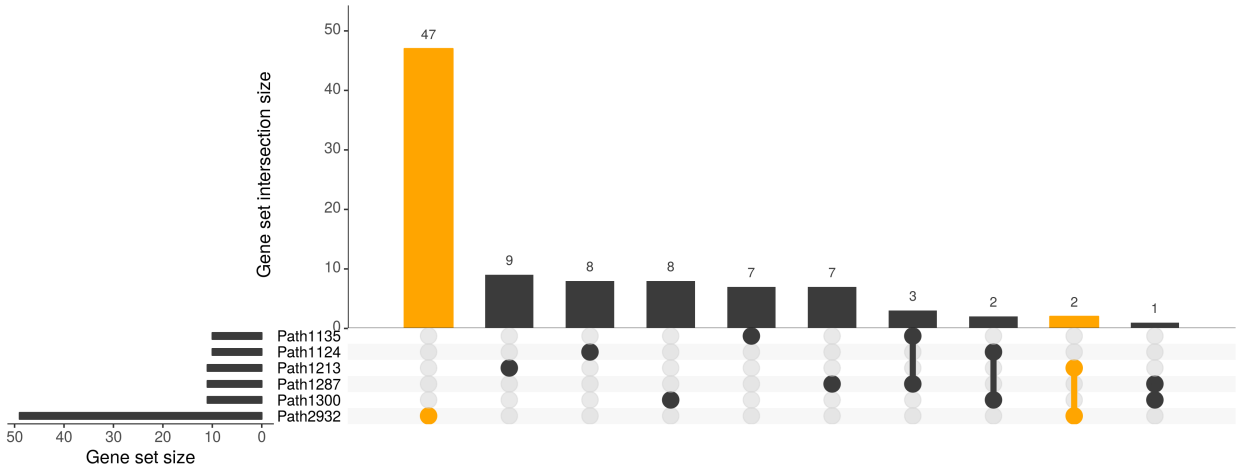


(e) Mean cell haemoglobin concentration (top) and packed cell volume (bottom) (van der Harst et al., 2012).

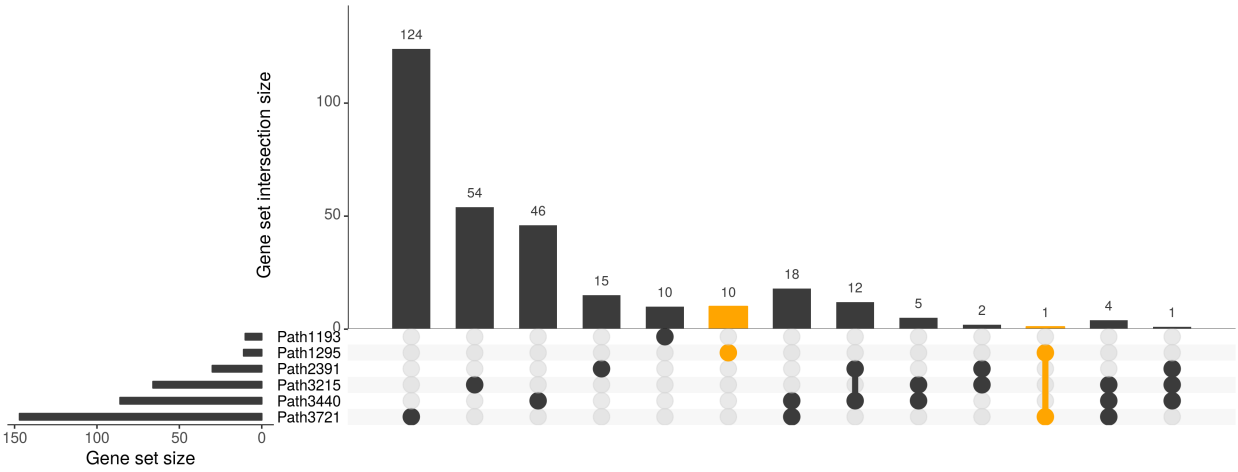




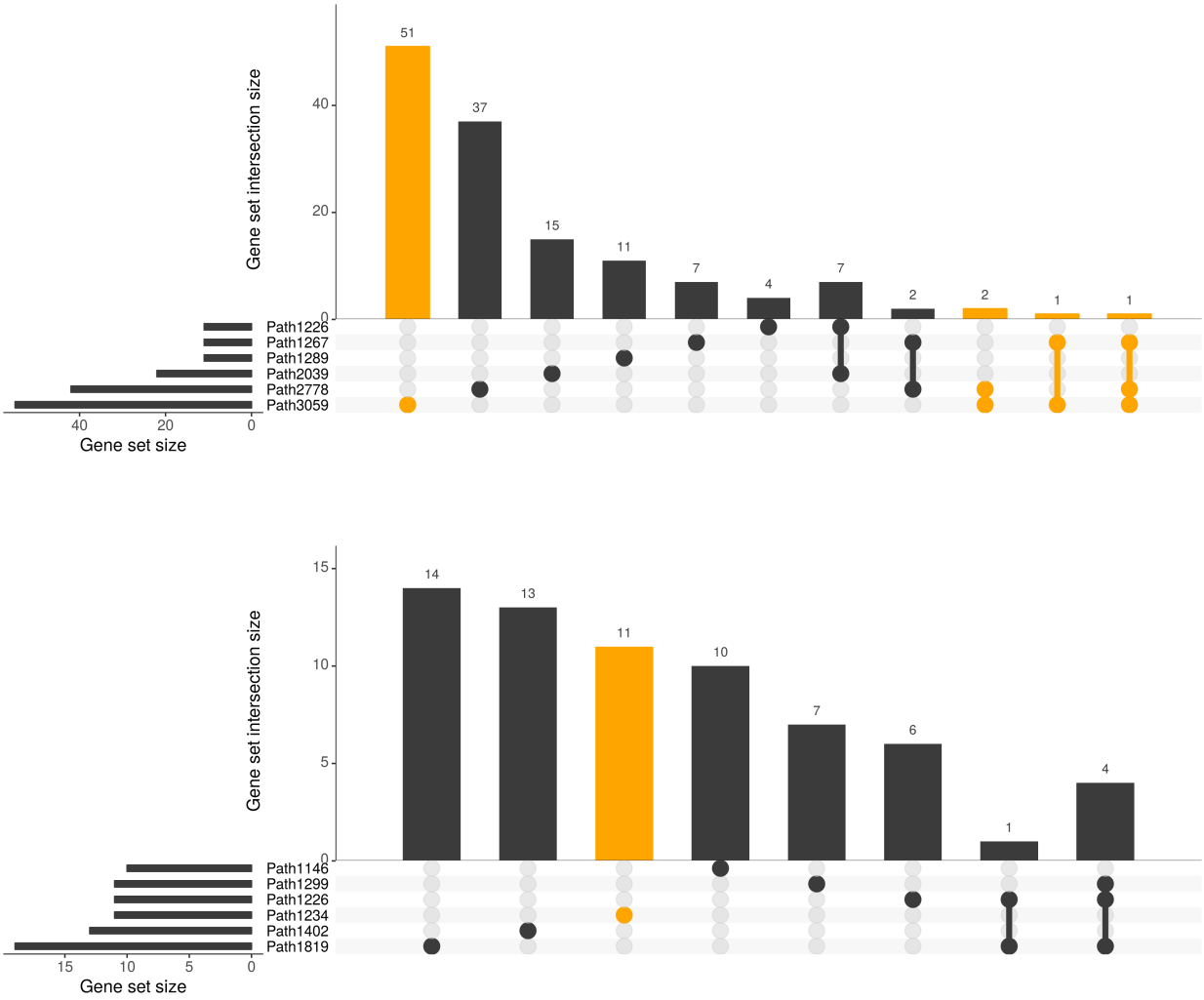
(f) Alzheimer's disease (Lambert et al., 2013).



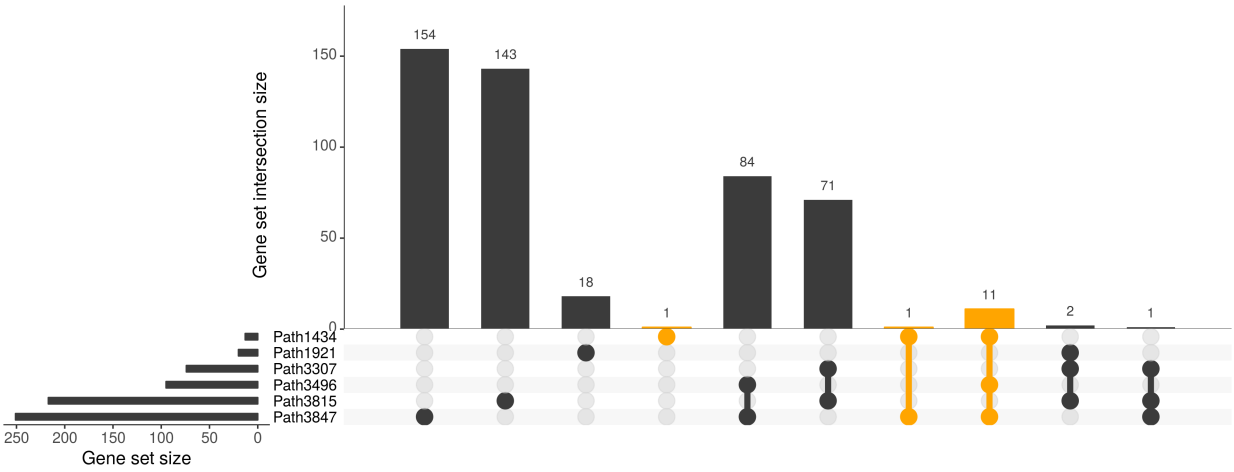
(g) Heart rate (Den Hoed et al., 2013).

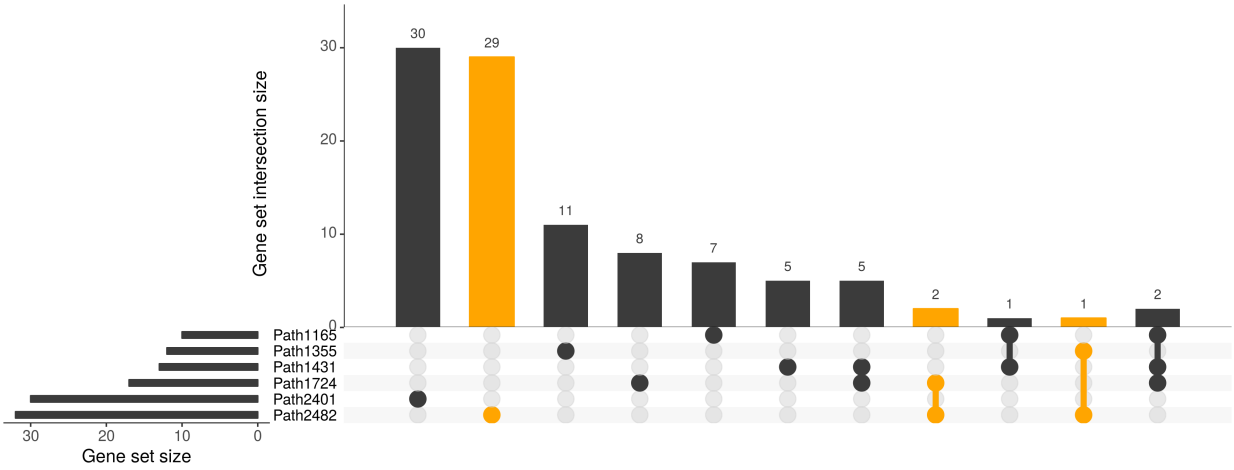


(h) Coronary artery disease (top) and myocardial infarction (bottom) (Nikpay et al., 2015).

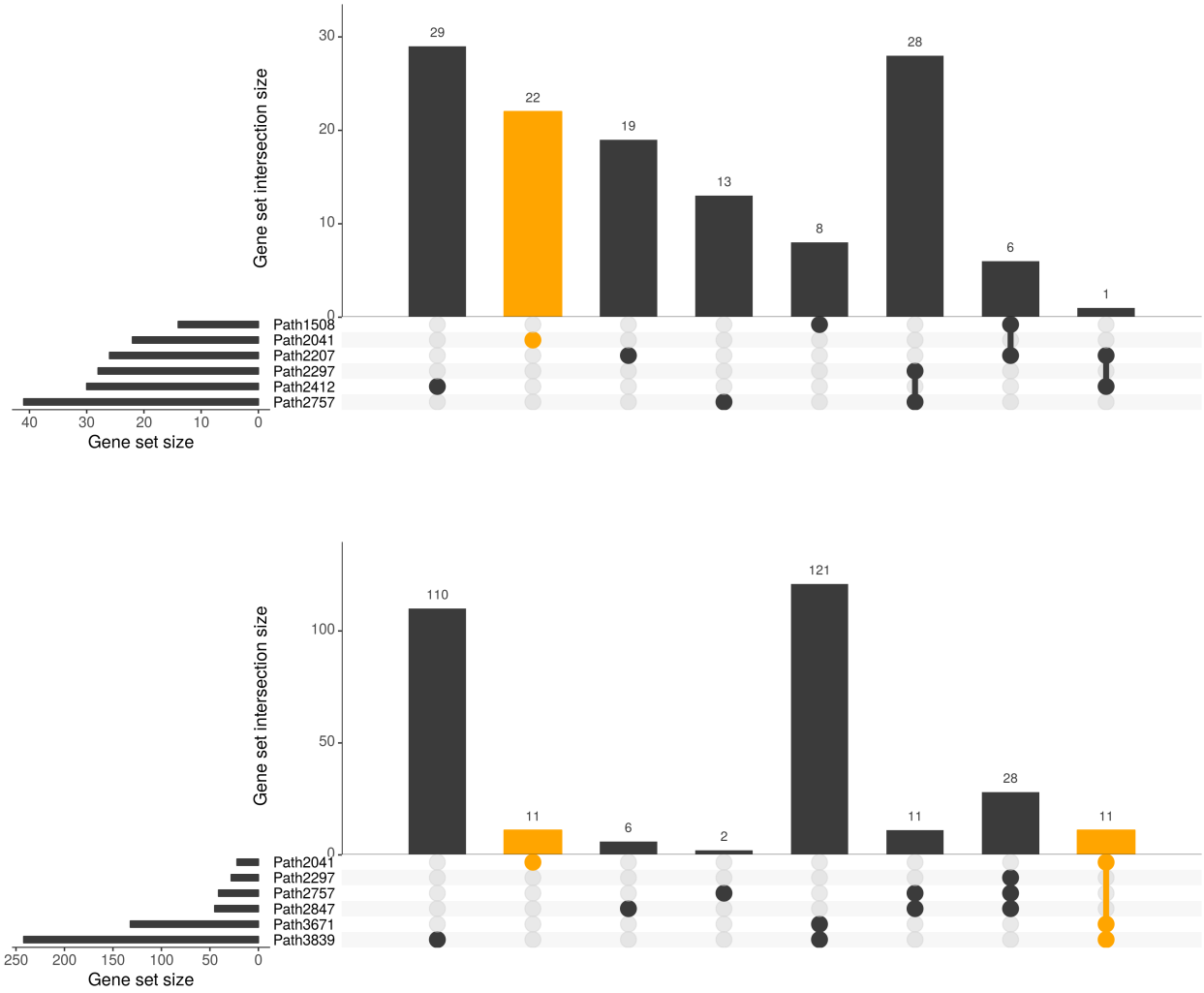


(i) Serum urate concentrations (top) and gout (bottom) (Köttgen et al., 2013).

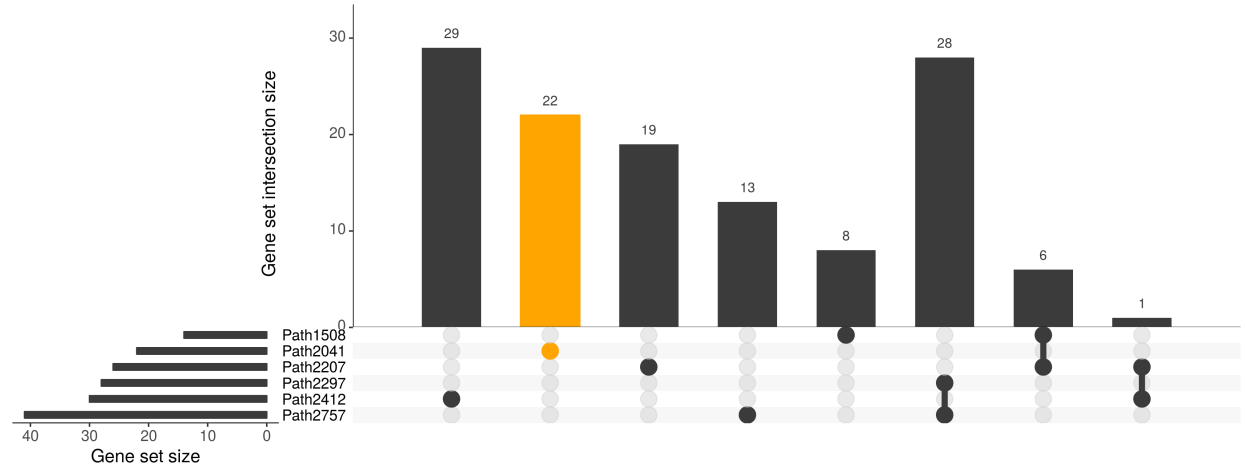


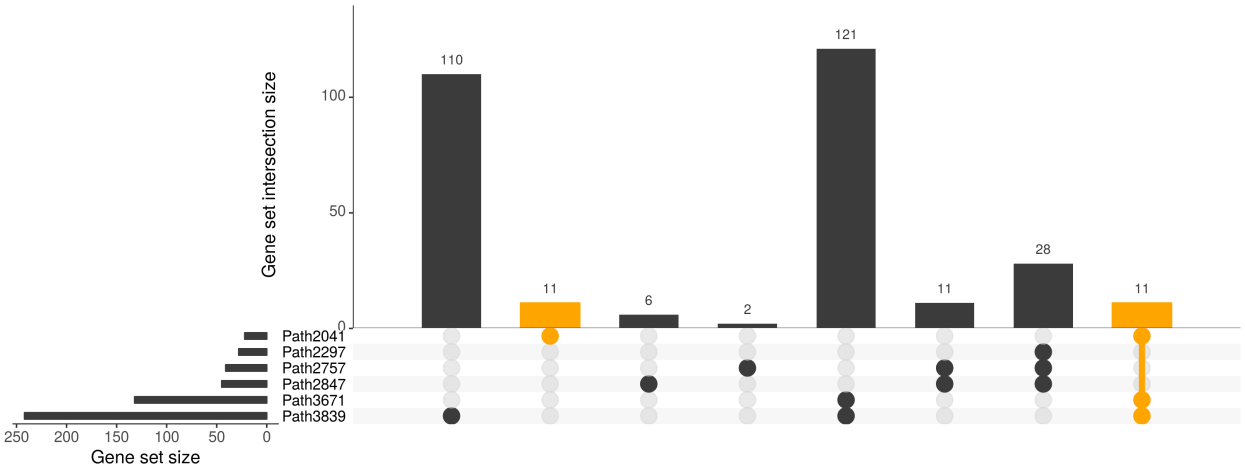


(j) Total cholesterol (top) and triglycerides (bottom) (Teslovich et al., 2010).

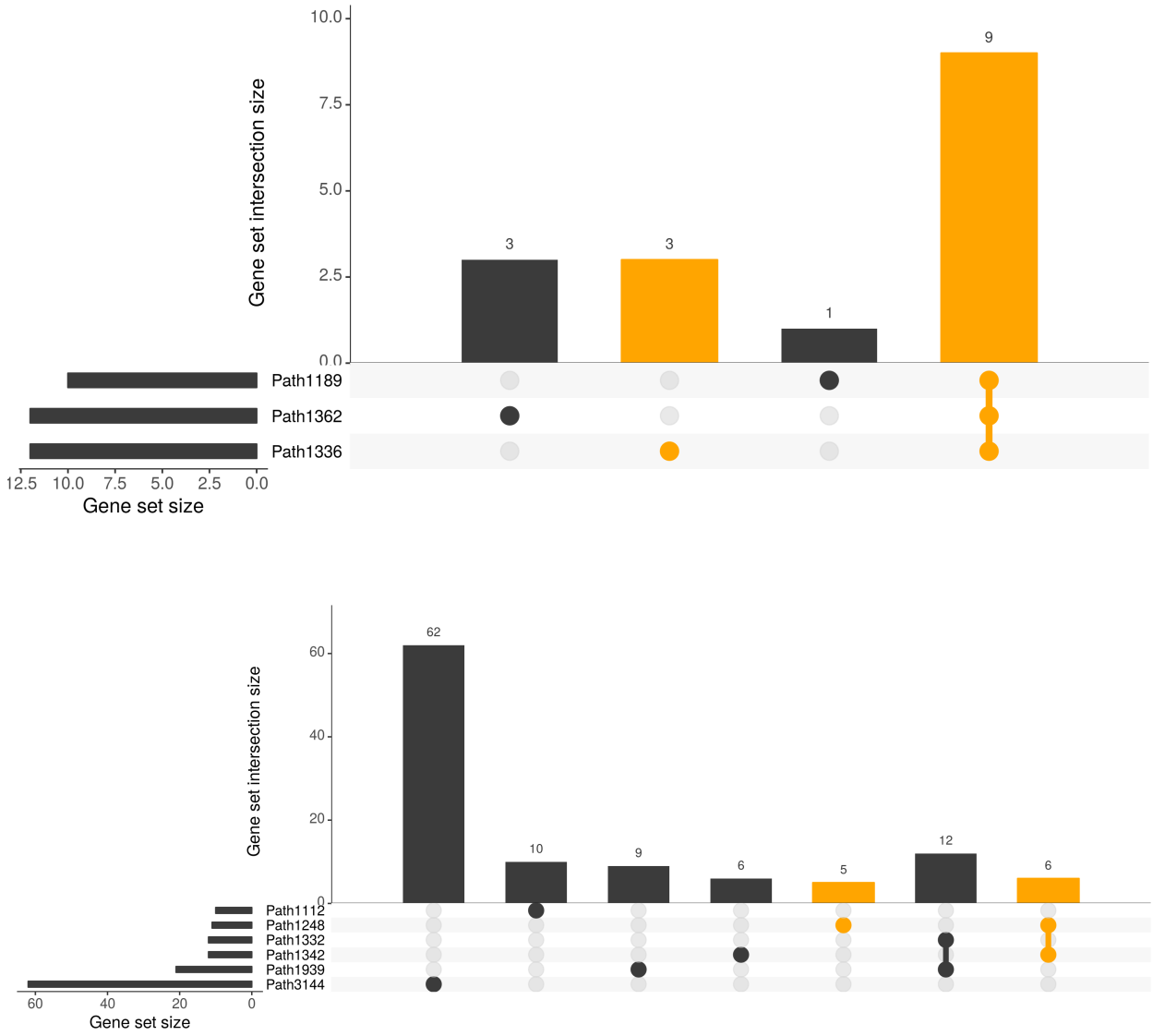


(k) High-density lipoprotein (top) and low-density lipoprotein (bottom) (Teslovich et al., 2010).

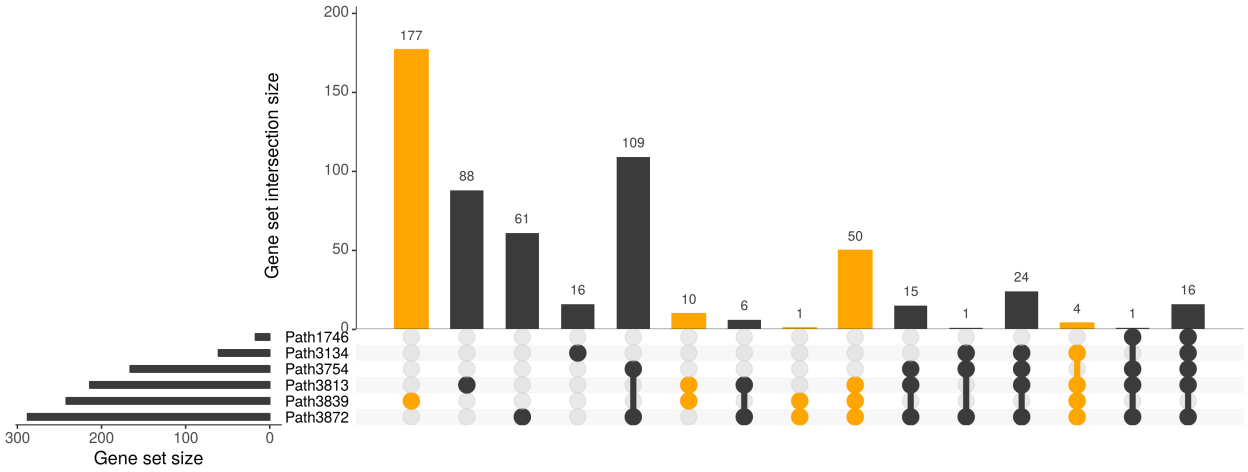




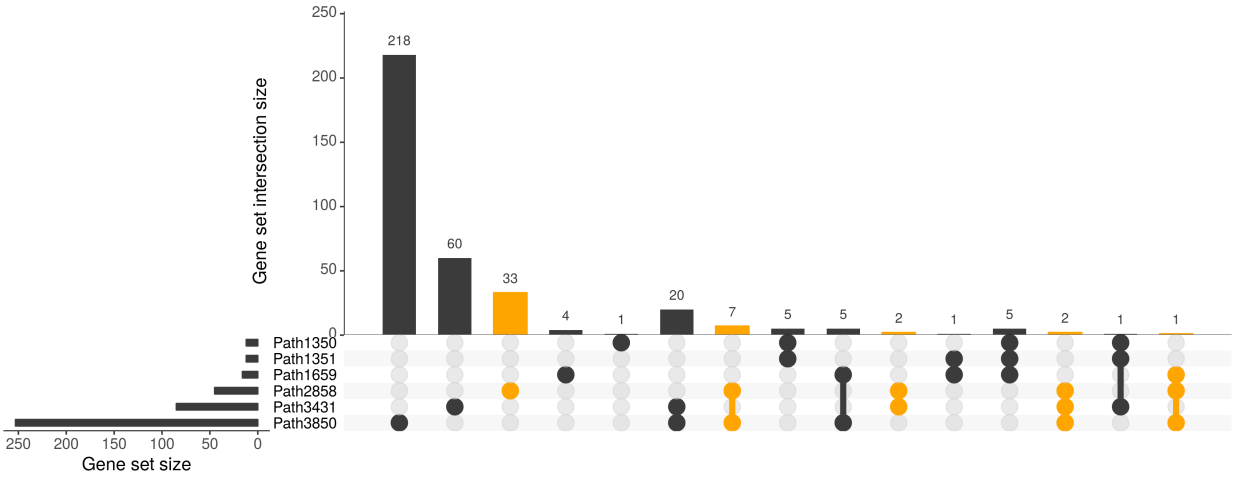
(l) Depressive symptoms (top) and neuroticism (bottom) (Okbay et al., 2016).



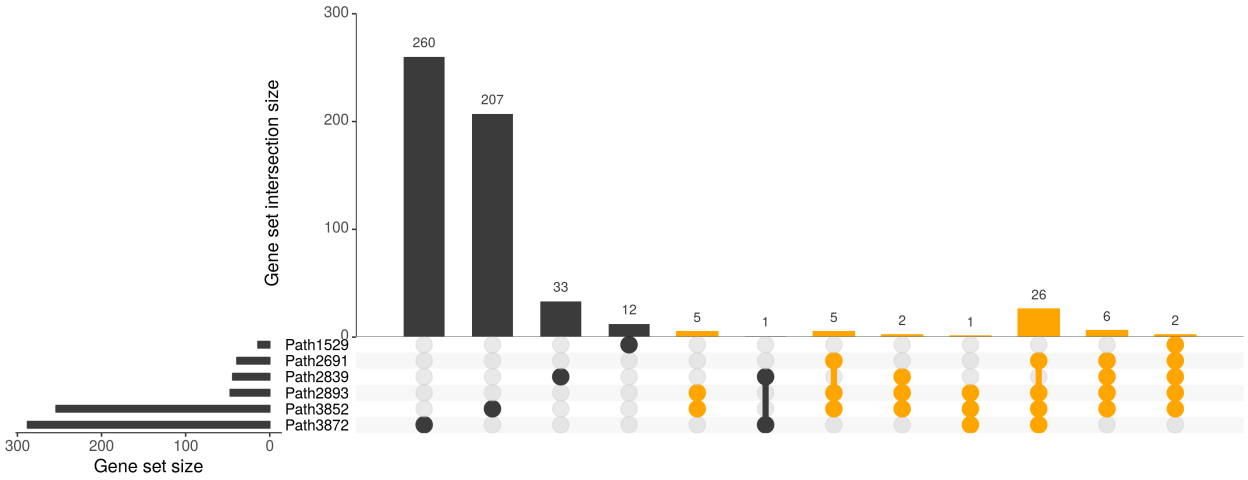
(m) Schizophrenia (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014).

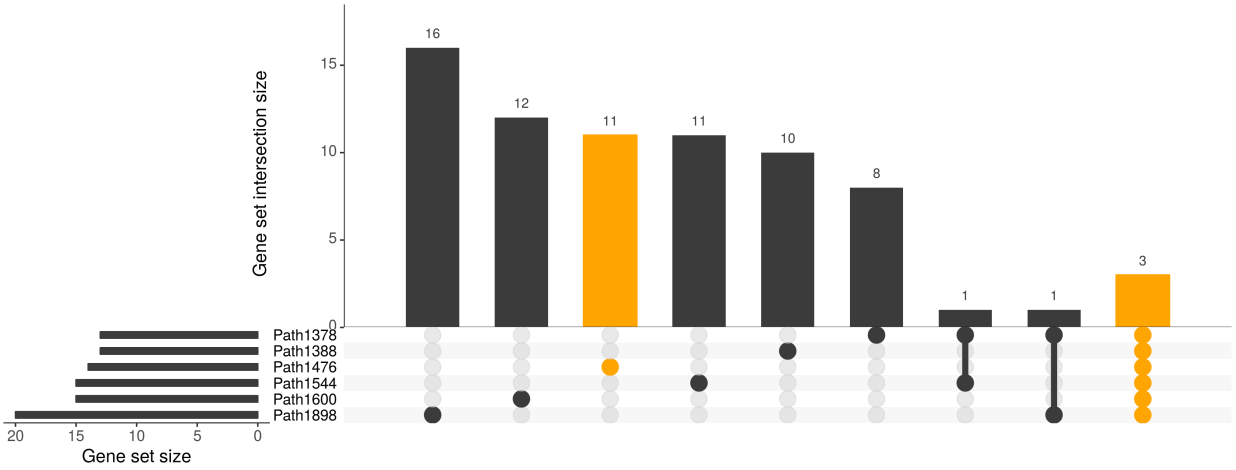


(n) Rheumatoid arthritis (Okada et al., 2014).

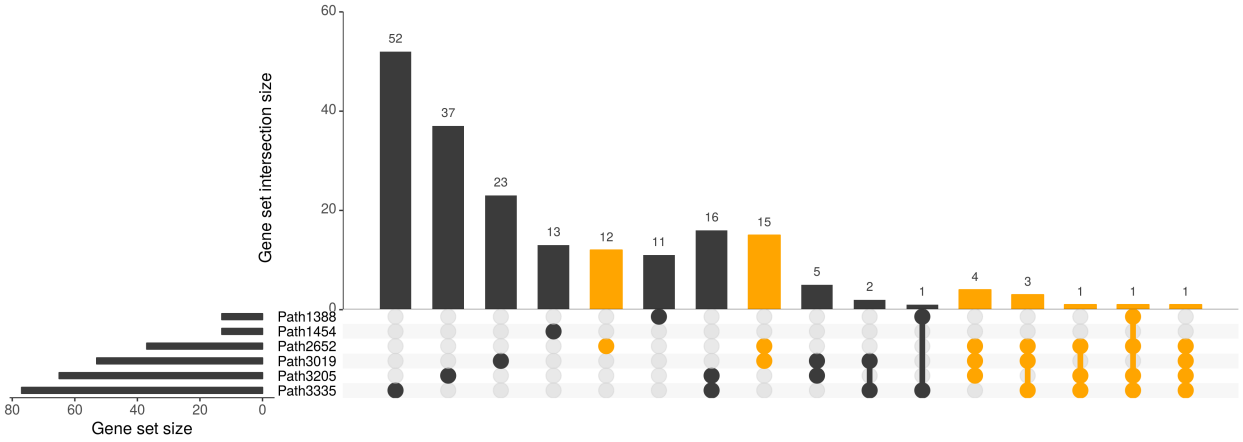


(o) Fasting glucose levels (top) and Fasting insulin levels (bottom) (Manning et al., 2012).

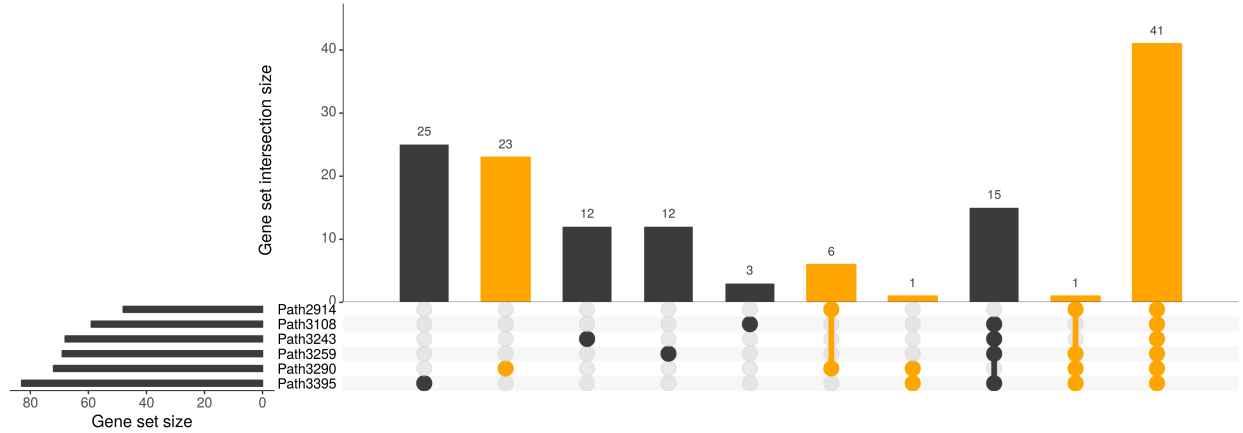


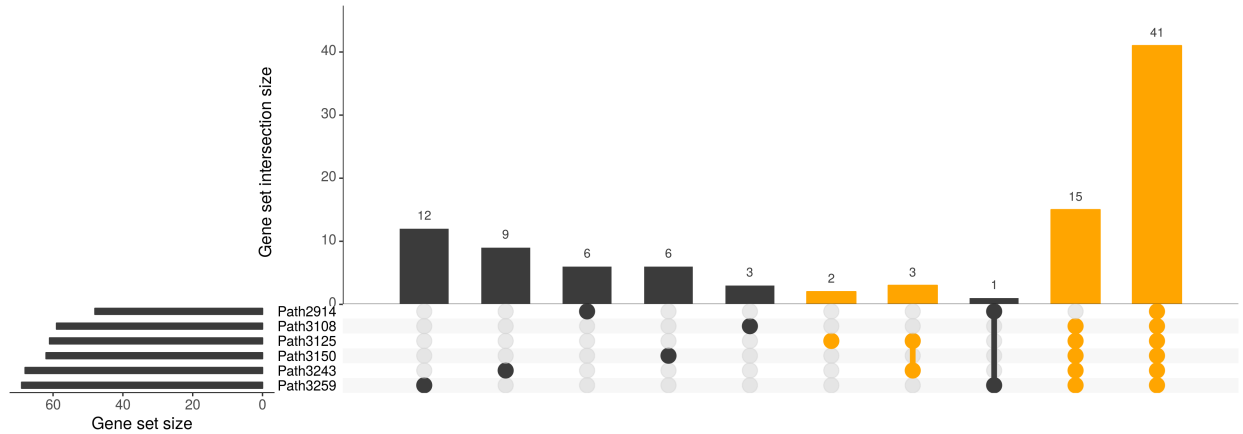


(p) Age at natural menopause (Day et al., 2015).

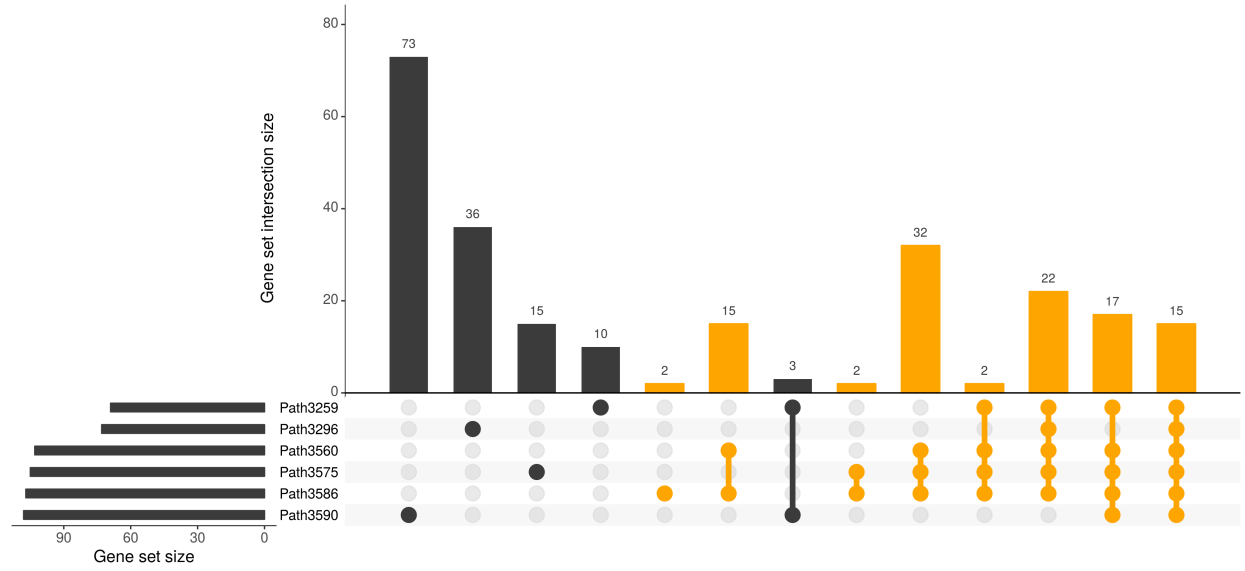


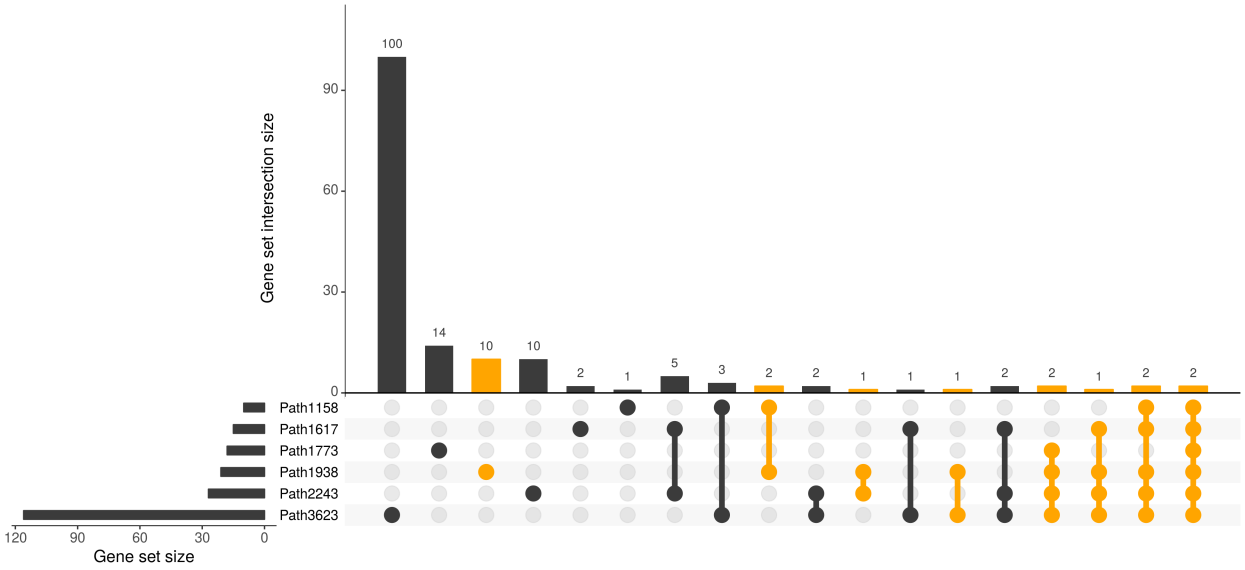
(q) Mean cell haemoglobin (top) and mean cell volume (bottom) (van der Harst et al., 2012).





(r) Haemoglobin (top) and red blood cell count (bottom) (van der Harst et al., 2012).





(s) Inflammatory bowel disease (top), Crohn's disease (middle) and ulcerative colitis (bottom) (Liu et al., 2015).

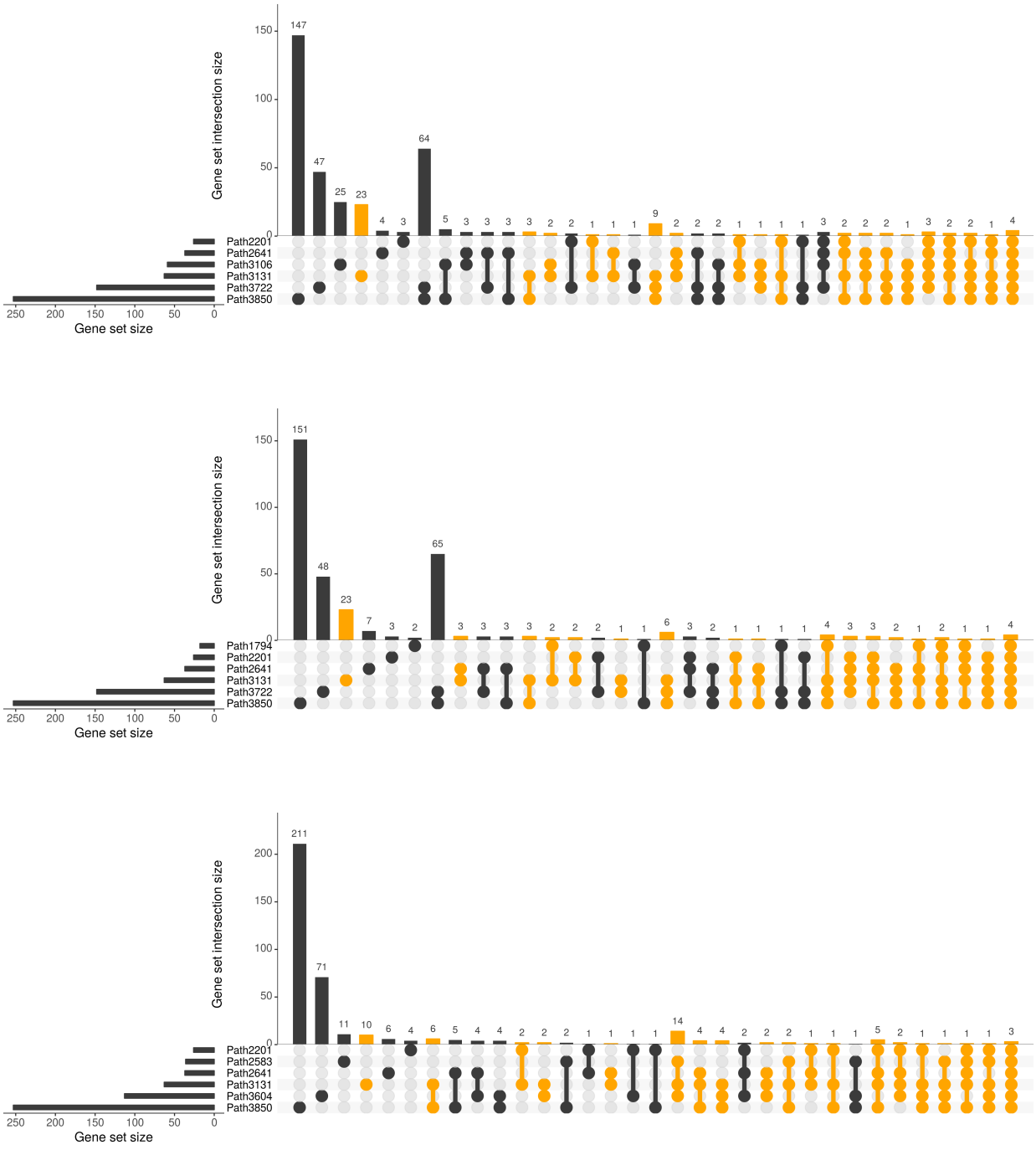


Figure F.13: Gene set overlap among top 6 most enriched pathways for each of 31 phenotypes.

Supplementary Figure 14

Compare the number of trait-associated loci detected under the baseline hypothesis with the number of trait-associated loci detected under the enrichment hypothesis. The baseline hypothesis assumes that no pathways are enriched. The enrichment hypothesis assumes that a candidate pathway is enriched. For each phenotype, each dot corresponds to one of the top 10 most enriched pathways [shown in Supplementary Figure 10 of Zhu and Stephens (2017b)]. A positive value in x-axis indicates that more trait-associated loci ($P_1 > 0.9$) are identified under the enrichment hypothesis than under the null hypothesis. See Supplementary Figure 3 of Zhu and Stephens (2017b) for the definition of locus. Note that some dots are overlapped due to their similar values in x-axis.

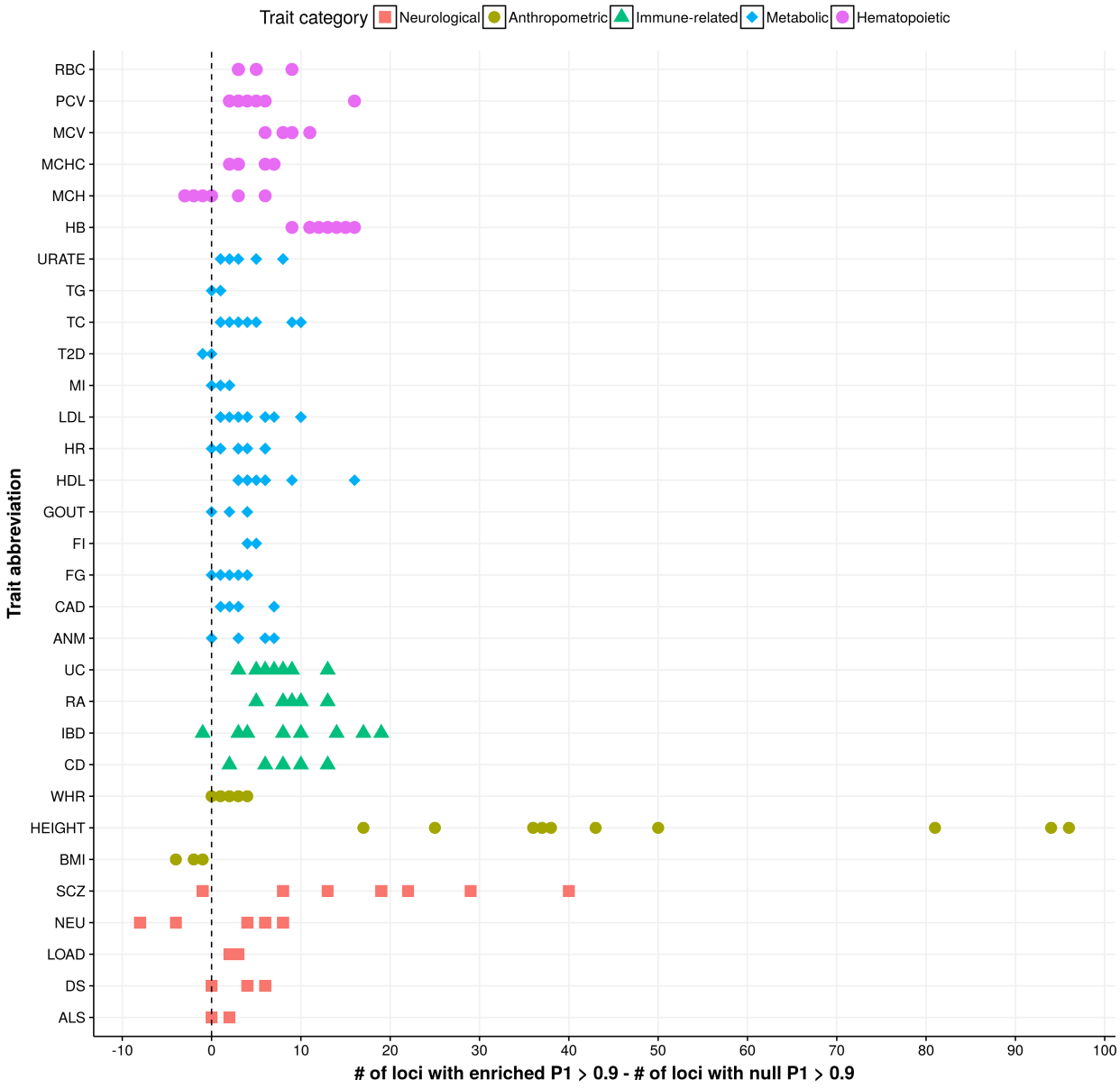
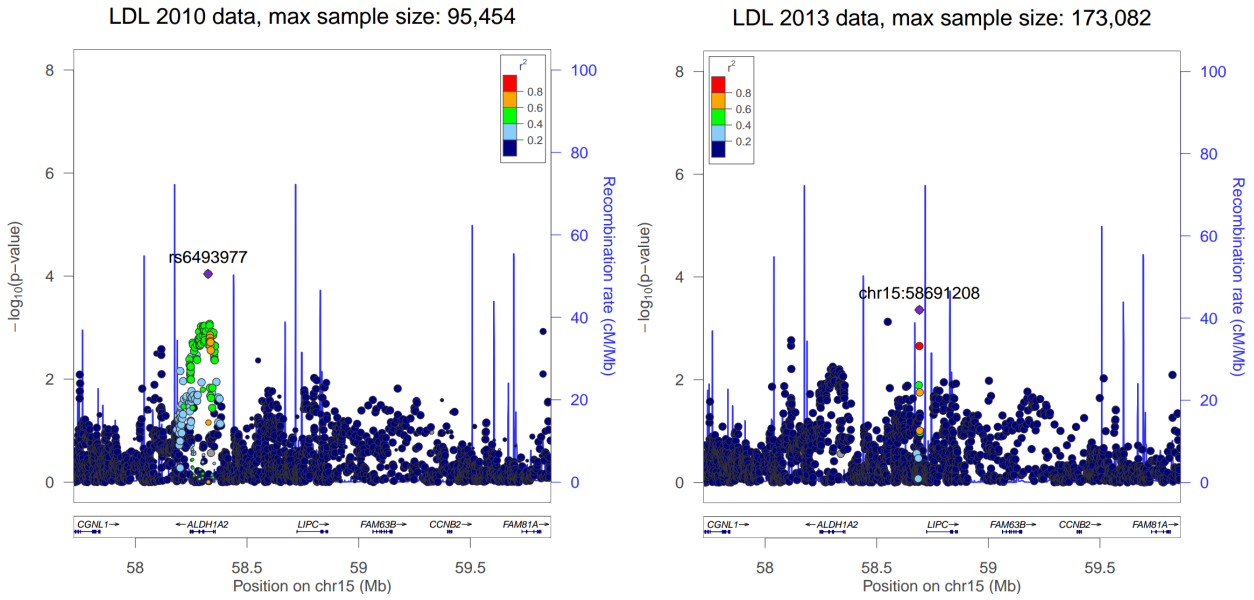


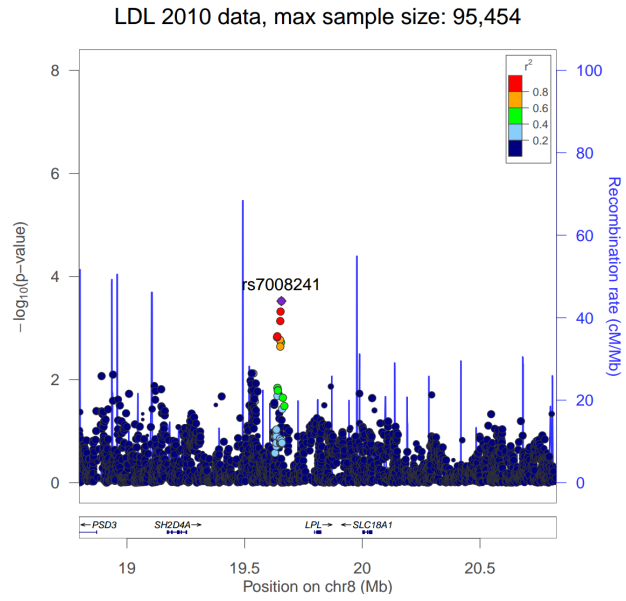
Figure F.14: Compare the number of trait-associated loci detected under the baseline hypothesis with the number of trait-associated loci detected under the enrichment hypothesis.

Supplementary Figure 15

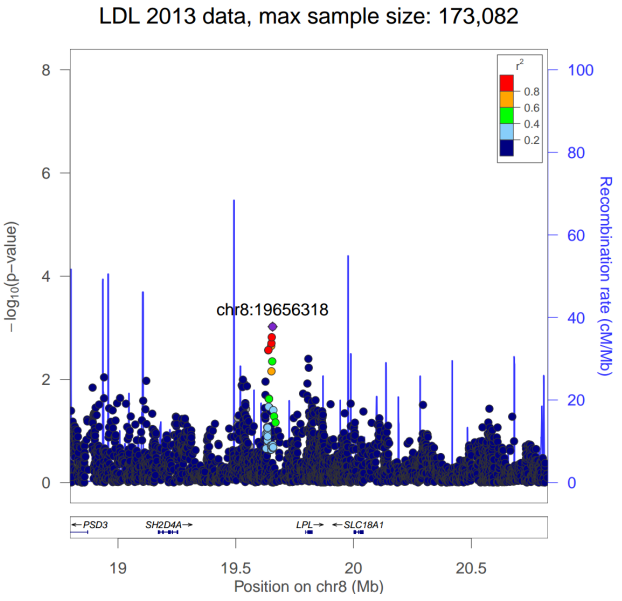


(a) *LIPC* (2010)

(b) *LIPC* (2013).



(c) *LPL* (2010).



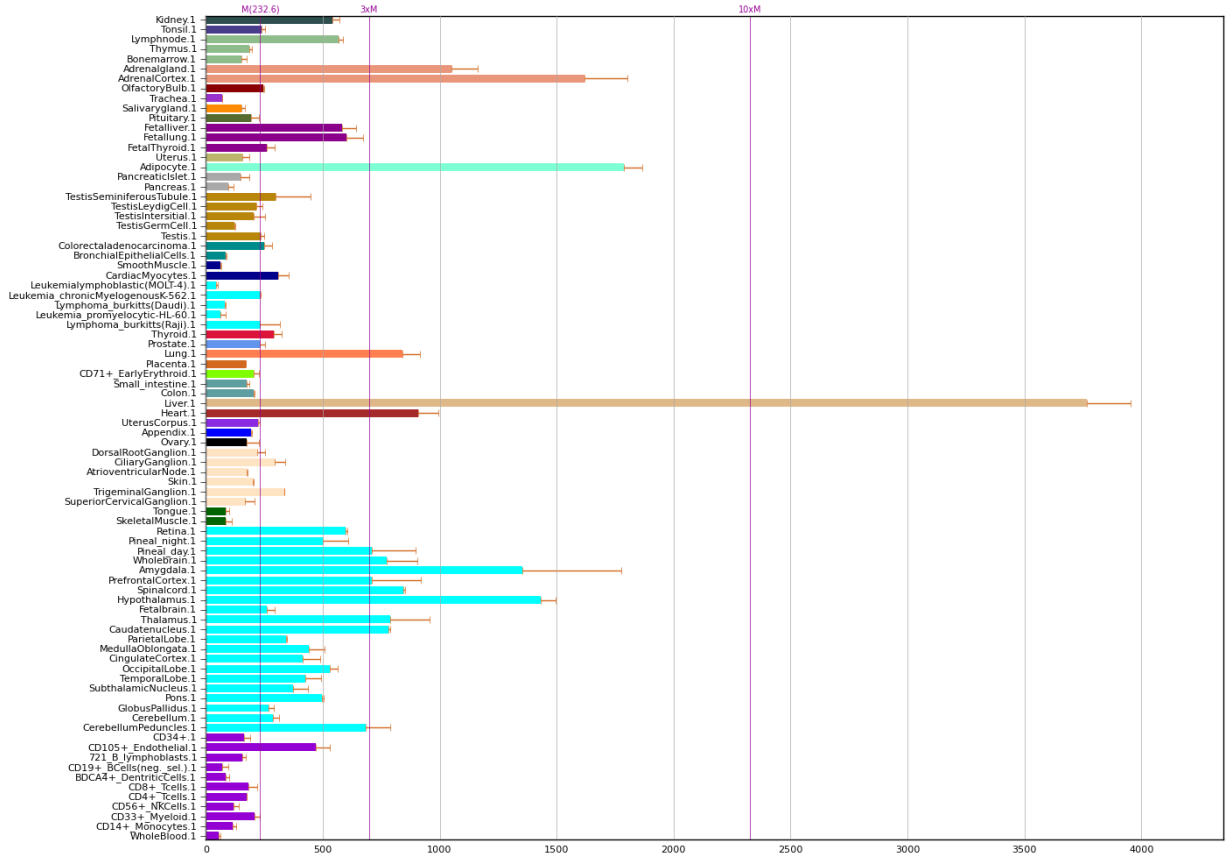
(d) *LPL* (2013).

Figure F.15: Regional association plots of genes *LIPC* and *LPL* based on single-SNP summary data of low-density lipoprotein cholesterol levels (Teslovich et al., 2010; Global Lipids Genetics Consortium, 2013)

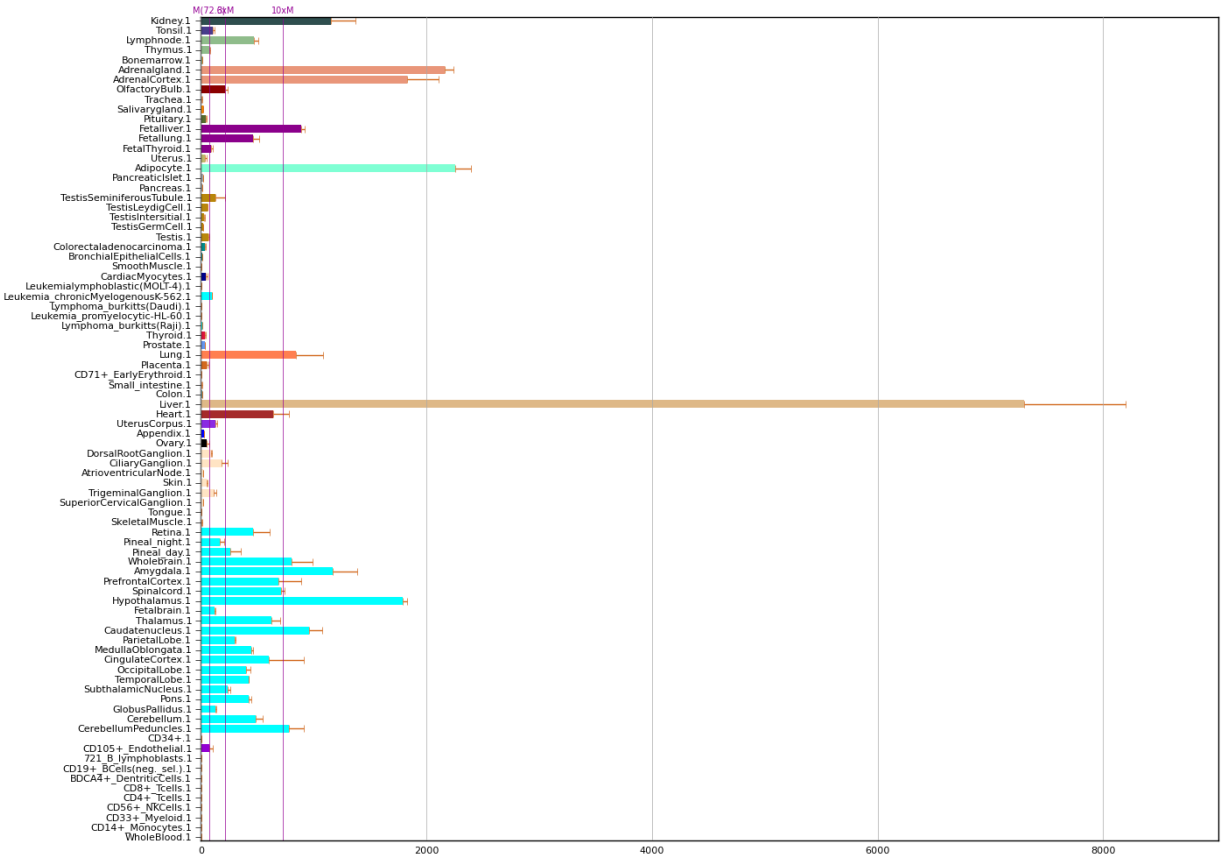
Supplementary Figure 16

Expression pattern of gene *APOE* across human tissues. Panels **(a)**-**(c)** are retrieved from `http://biogps.org/\#goto=genereport\&id=348`. Panels **(d)**-**(f)** are retrieved from `https://www.ncbi.nlm.nih.gov/gene/348`. Panel **(g)** is retrieved from `http://www.gtexportal.org/home/gene/APOE`.

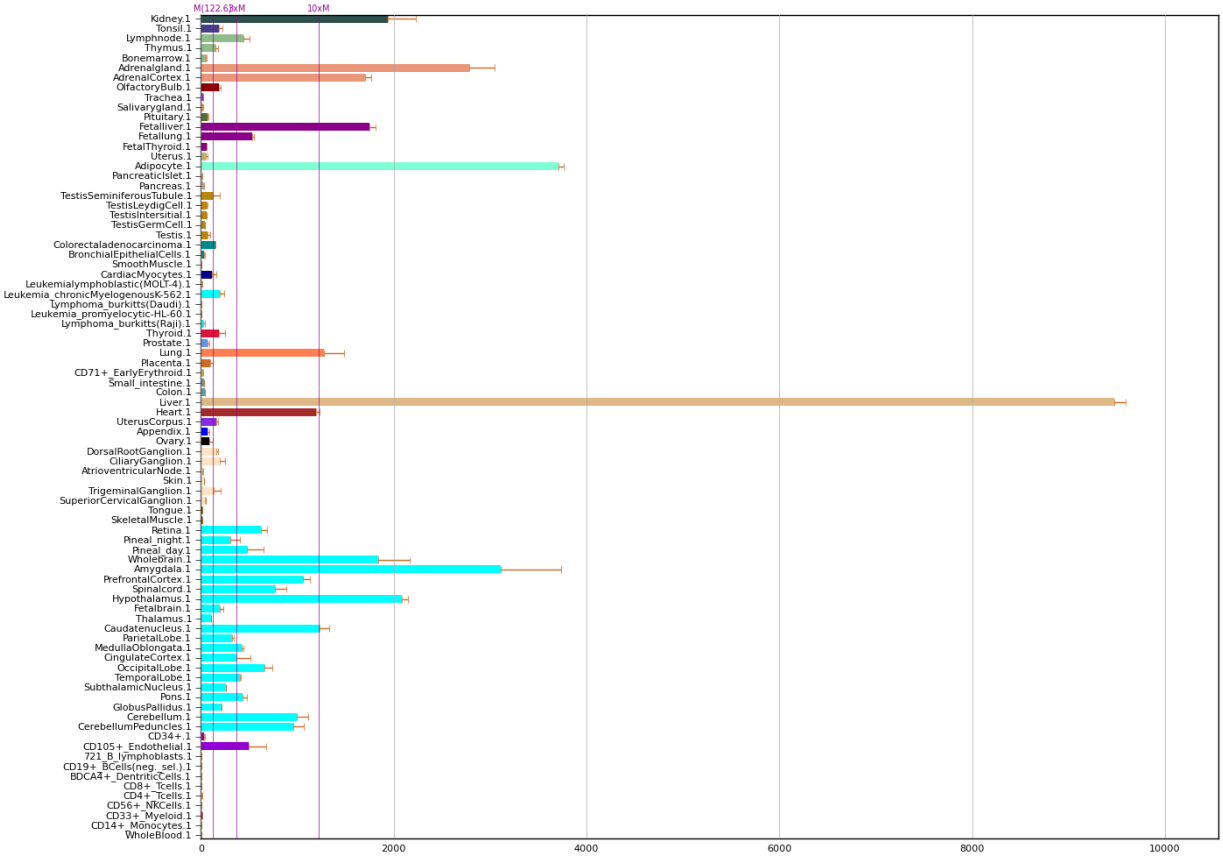
(a) Expression pattern of *APOE* across 76 human tissues, based on microarray data from GeneAtlas U133A, probeset 212884_x_at (Su et al., 2004).



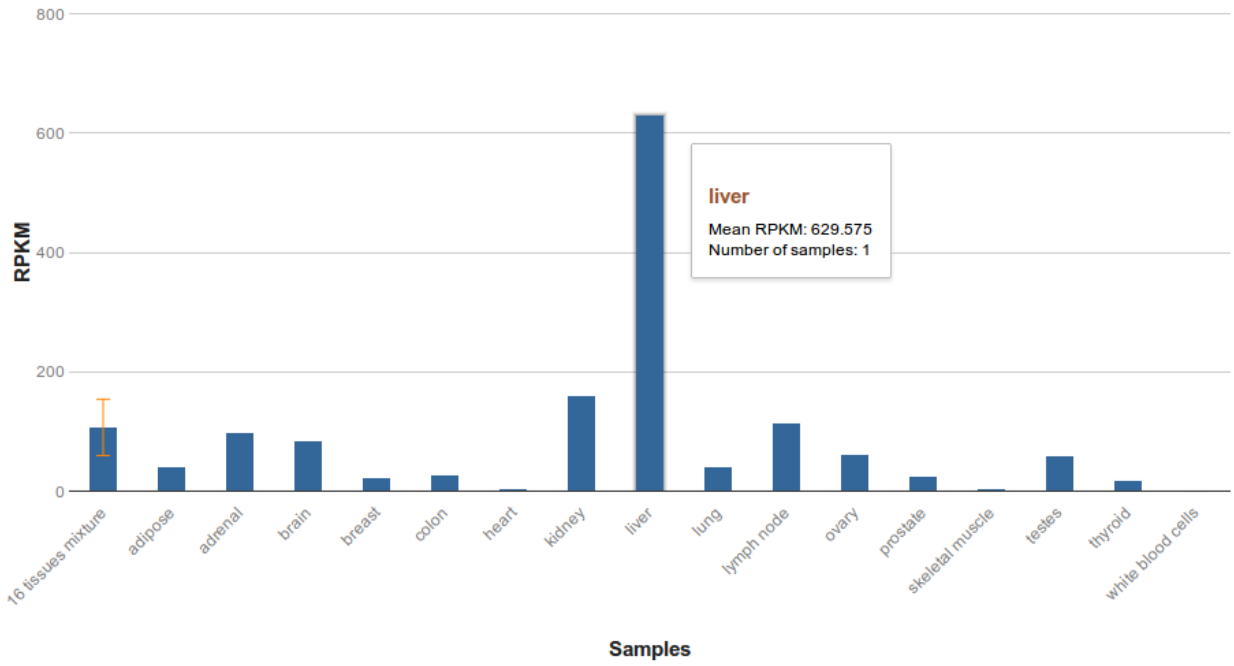
(b) Expression pattern of *APOE* across 76 human tissues, based on microarray data from GeneAtlas U133A, probeset 203382_s_at (Su et al., 2004).



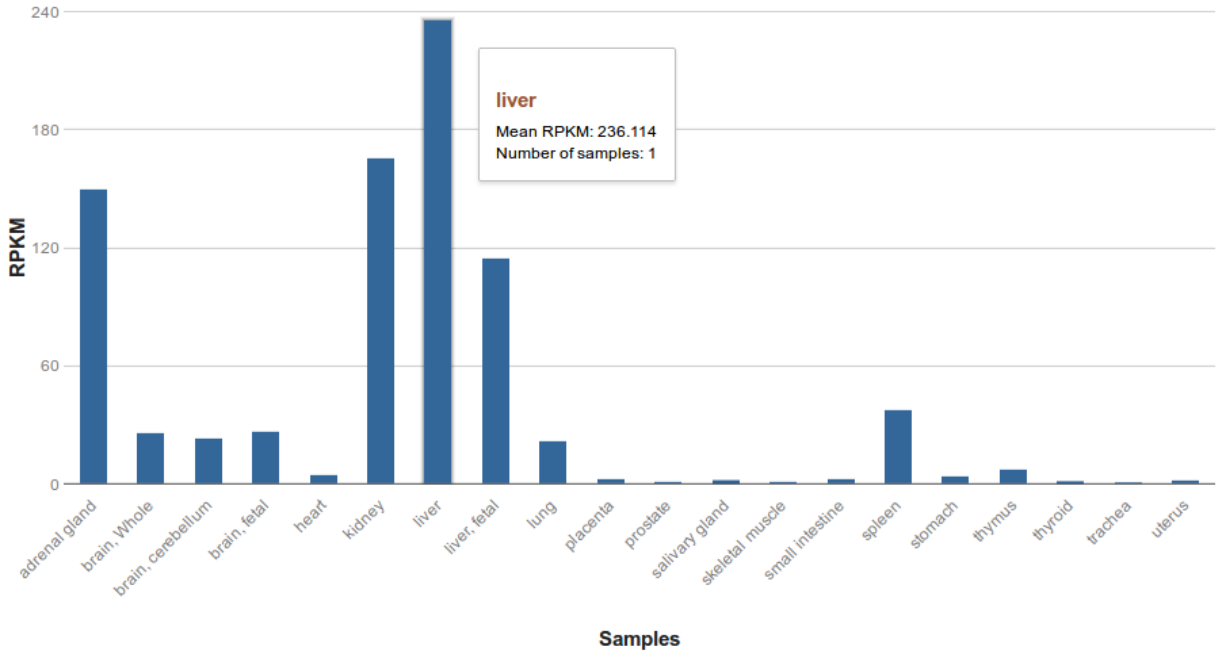
(c) Expression pattern of *APOE* across 76 human tissues, based on microarray data from GeneAtlas U133A, probeset 203381_s_at (Su et al., 2004).



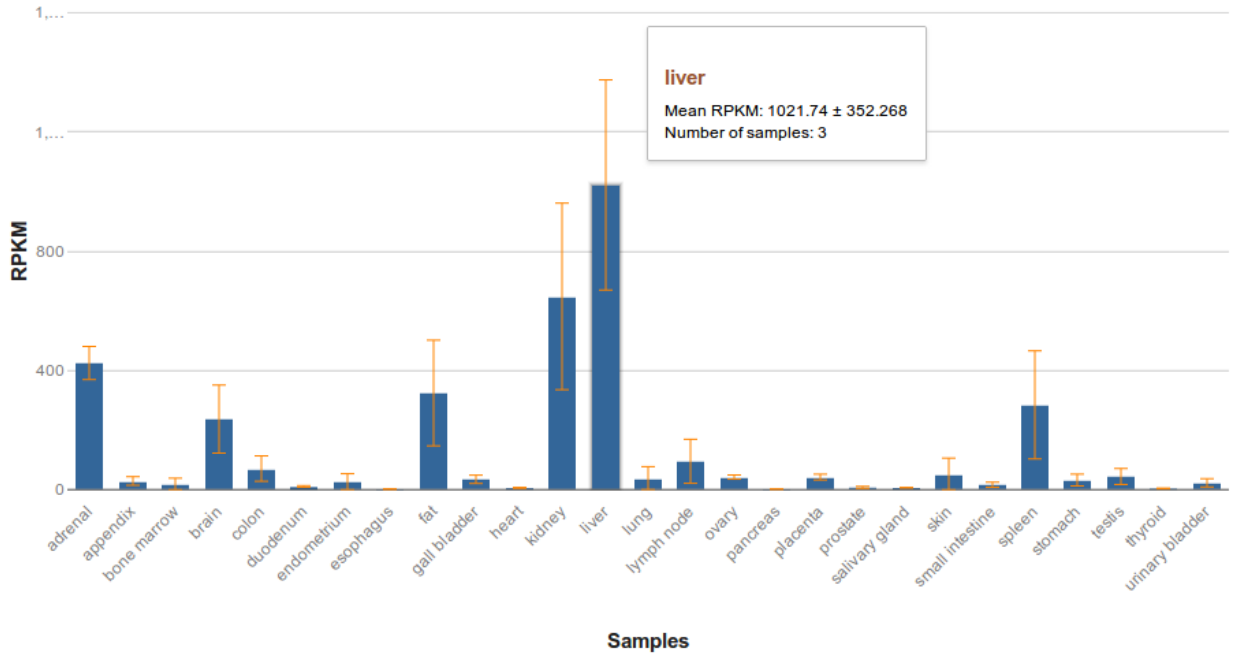
(d) Expression pattern of *APOE* across 16 human tissues, based on RNA-seq data from Illumina bodyMap2 transcriptome project (`https://www.ncbi.nlm.nih.gov/bioproject/PRJEB2445/`).



(e) Expression pattern of *APOE* across 20 human tissues based on RNA-seq data (Duff et al., 2015) (`https://www.ncbi.nlm.nih.gov/bioproject/PRJNA280600/`).



(f) Expression pattern of *APOE* across 27 human tissues, based on RNA-seq data from 95 human individuals (Fagerberg et al., 2014) (`https://www.ncbi.nlm.nih.gov/bioproject/PRJEB4337/`).



(g) Expression pattern of *APOE* across 53 human tissues, based on RNA-seq data from GTEx Analysis Release V6p (dbGaP Accession phs000424.v6.p1).

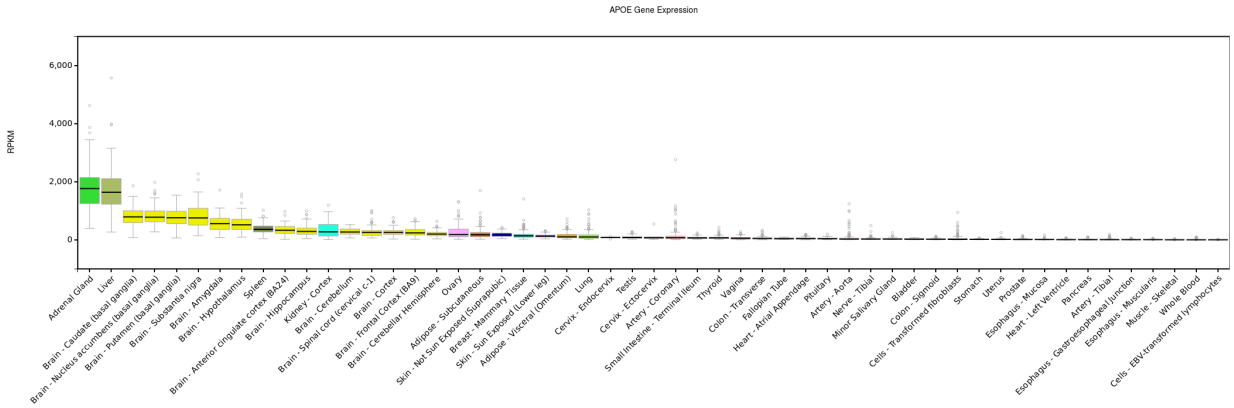
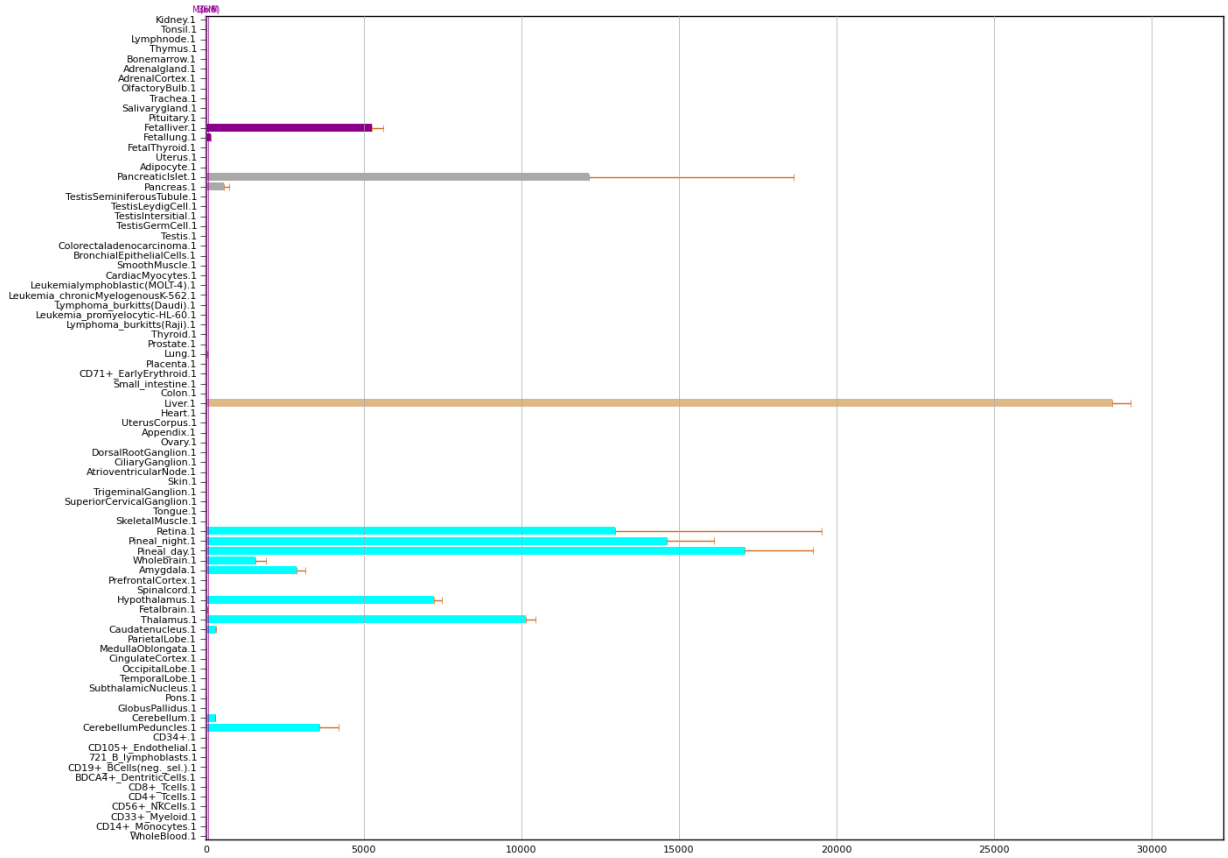


Figure F.16: Expression pattern of gene *APOE* across human tissues.

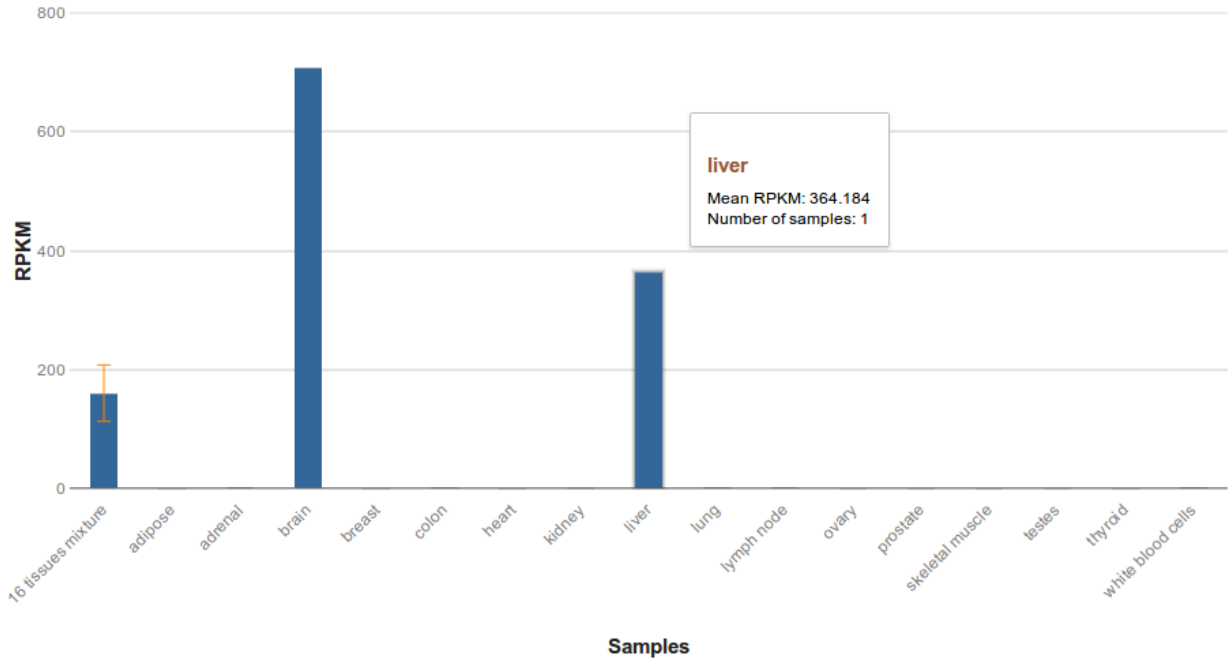
Supplementary Figure 17

Expression pattern of gene *TTR* across human tissues. Panels **(a)**-**(c)**) are retrieved from `http://biogps.org/\#goto=genereport\&id=7276`. Panels **(d)**)-**(f)** are retrieved from `https://www.ncbi.nlm.nih.gov/gene/7276`. Panel **(g)** is retrieved from `http://www.gtexportal.org/home/gene/TTR`.

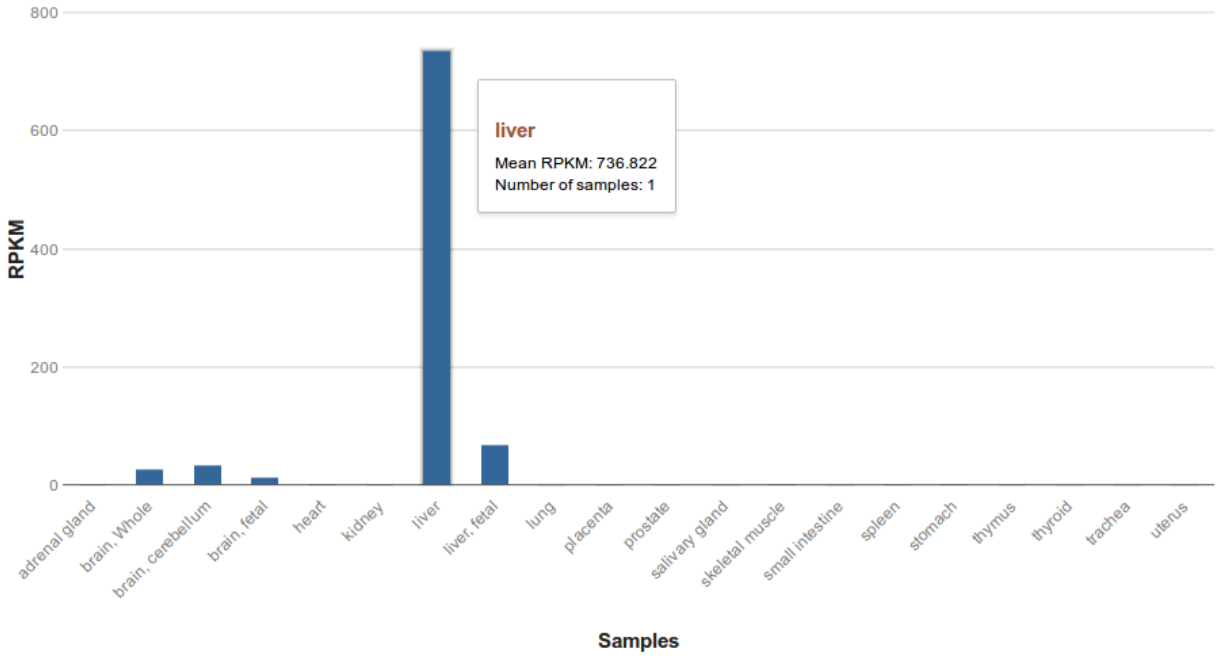
(a) Expression pattern of *TTR* across 76 human tissues, based on microarray data from GeneAtlas U133A, probeset 209660_at (Su et al., 2004).



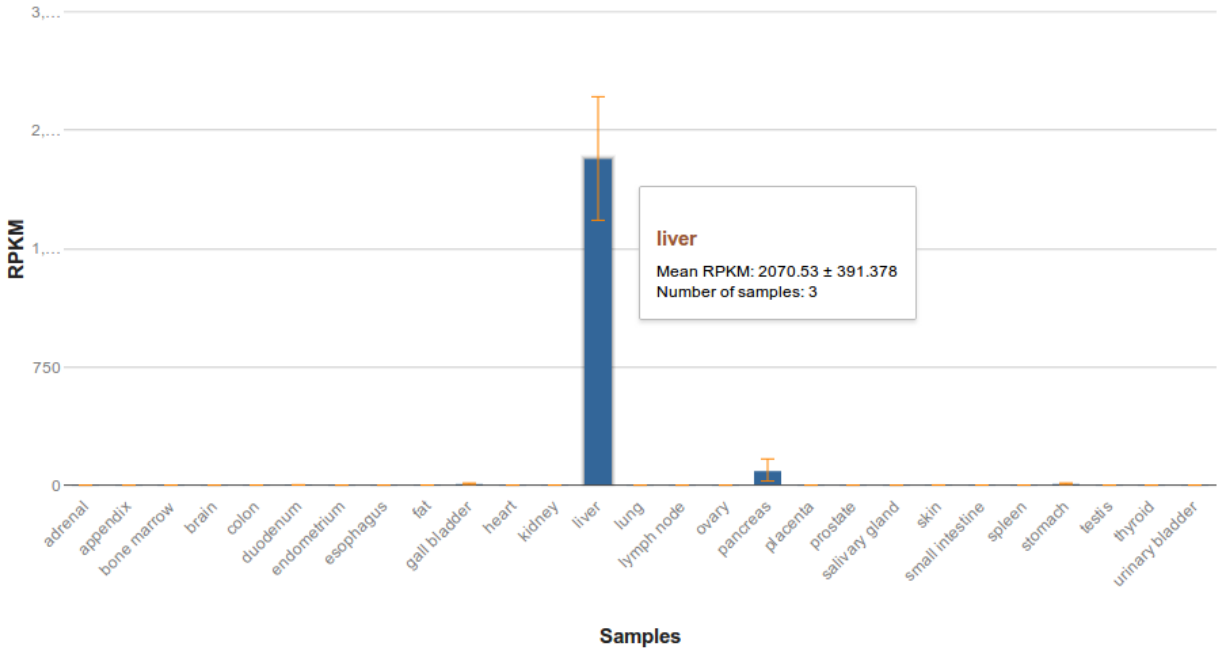
(b) Expression pattern of *TTR* across 16 human tissues, based on RNA-seq data from Illumina bodyMap2 transcriptome project (`https://www.ncbi.nlm.nih.gov/bioproject/PRJEB2445/`).



(c) Expression pattern of *TTR* across 20 human tissues based on RNA-seq data (Duff et al., 2015) (`https://www.ncbi.nlm.nih.gov/bioproject/PRJNA280600/`).



(d) Expression pattern of *TTR* across 27 human tissues, based on RNA-seq data from 95 human individuals (Fagerberg et al., 2014) (`https://www.ncbi.nlm.nih.gov/bioproject/PRJEB4337/`).



(e) Expression pattern of *TTR* across 53 human tissues, based on RNA-seq data from GTEx Analysis Release V6p (dbGaP Accession phs000424.v6.p1).

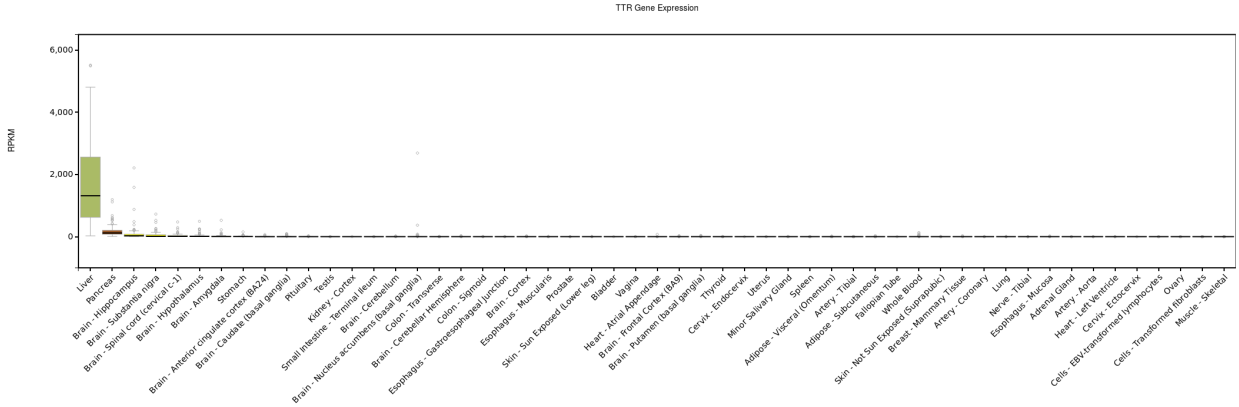


Figure F.17: Expression pattern of gene *TTR* across human tissues.

Supplementary Figure 18

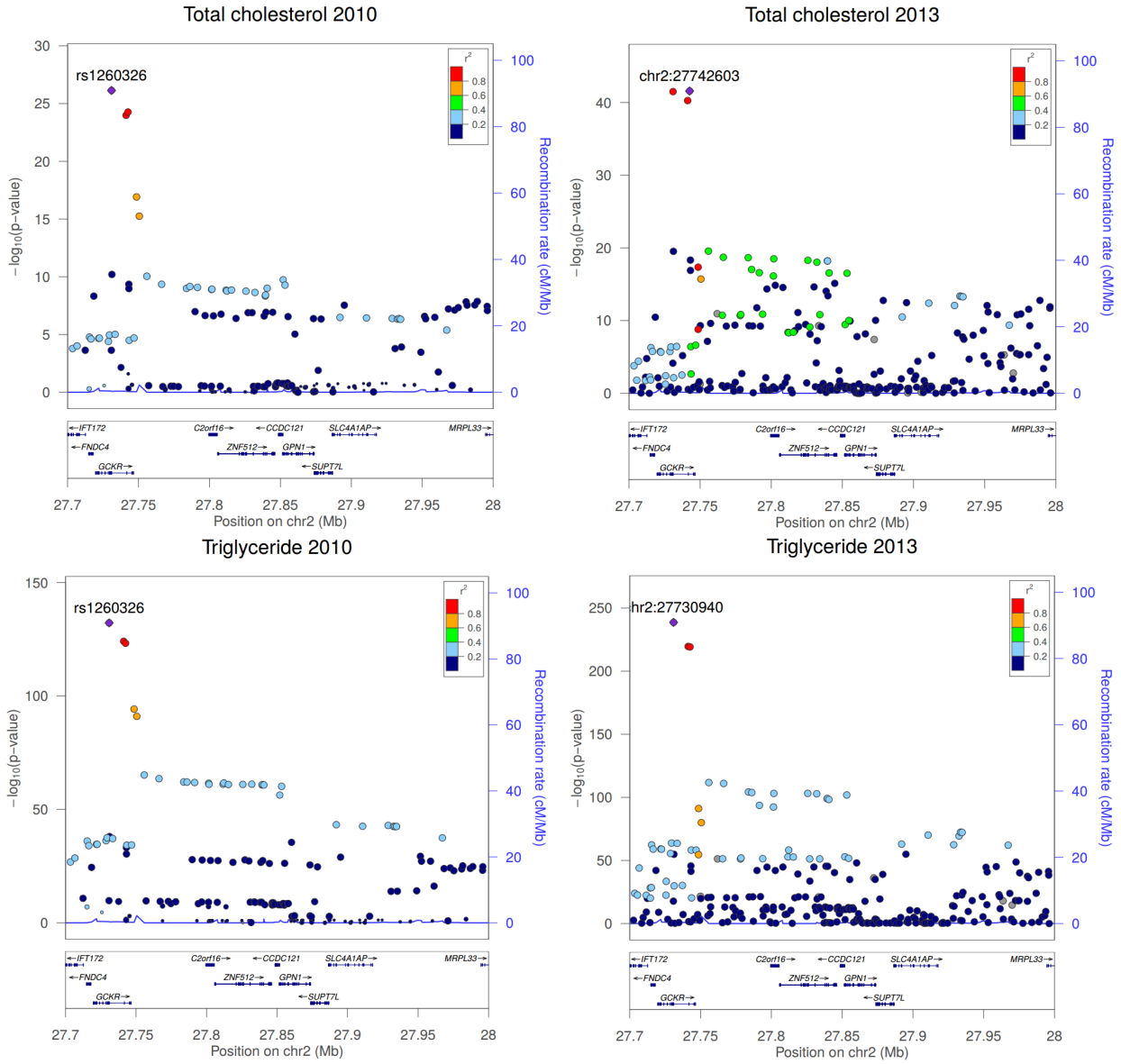


Figure F.18: Regional association plots of genes *C2orf16* and *GCKR* based on single-SNP summary data of total cholesterol and triglycerides levels (Teslovich et al., 2010; Global Lipids Genetics Consortium, 2013).

Supplementary Figure 19

Estimated proportion of trait-associated SNPs across 31 phenotypes. Each dot denotes a phenotype. The value of each dot along x-axis denotes the estimated proportion of trait-associated SNPs based on hyper-parameter estimates (θ_0). The point range along y-axis denotes the estimated proportion of trait-associated SNPs and 95% credible interval based on variational parameter estimates (α). See Supplementary Note of Zhu and Stephens (2017b) for details of computing these quantities. The dashed lines are reference lines with slope one and intercept zero. Both axes use a logarithmic scale (base 10). Note that the y-axis is the same as the x-axis of Supplementary Figure 2 of Zhu and Stephens (2017b).

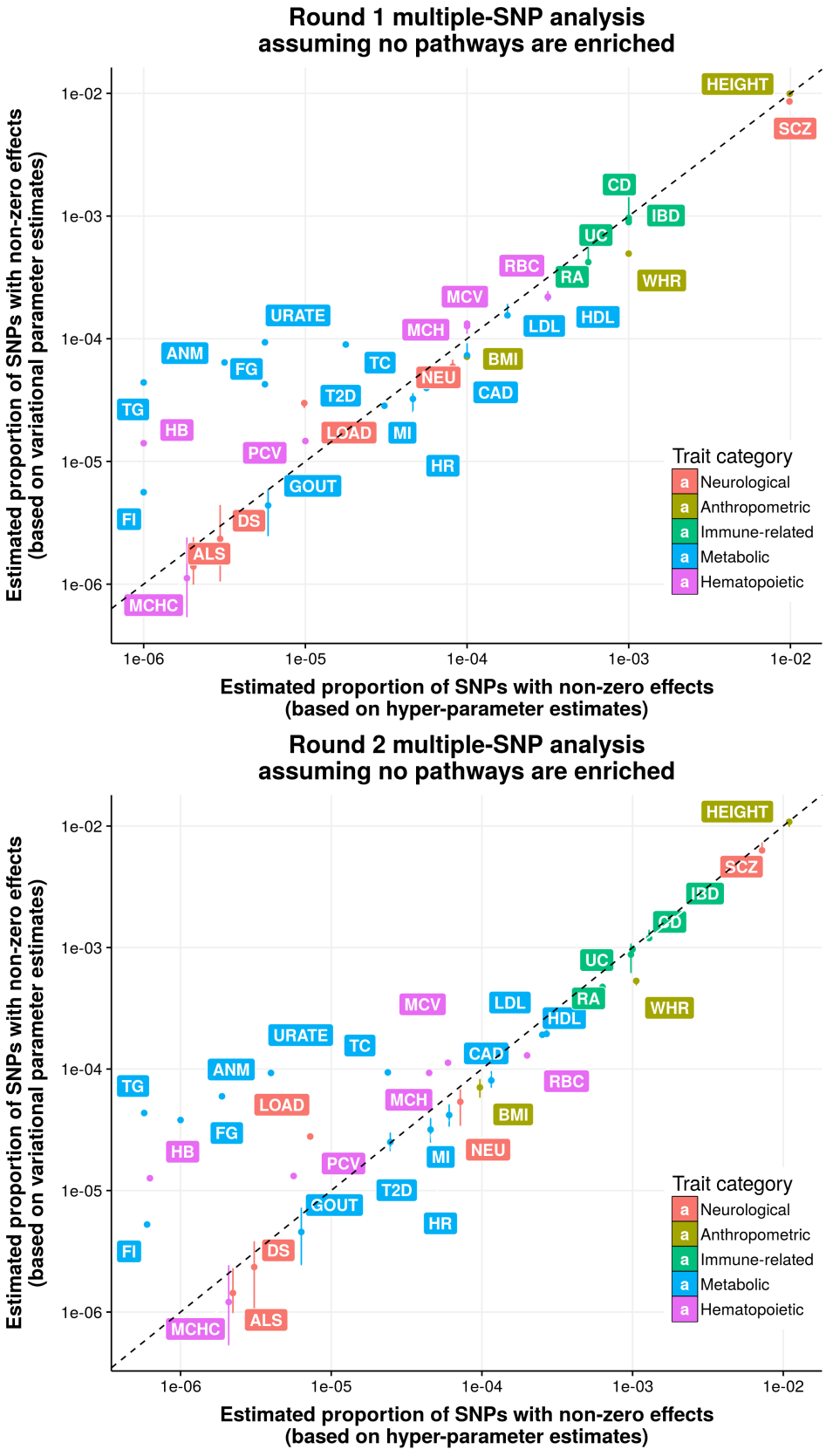
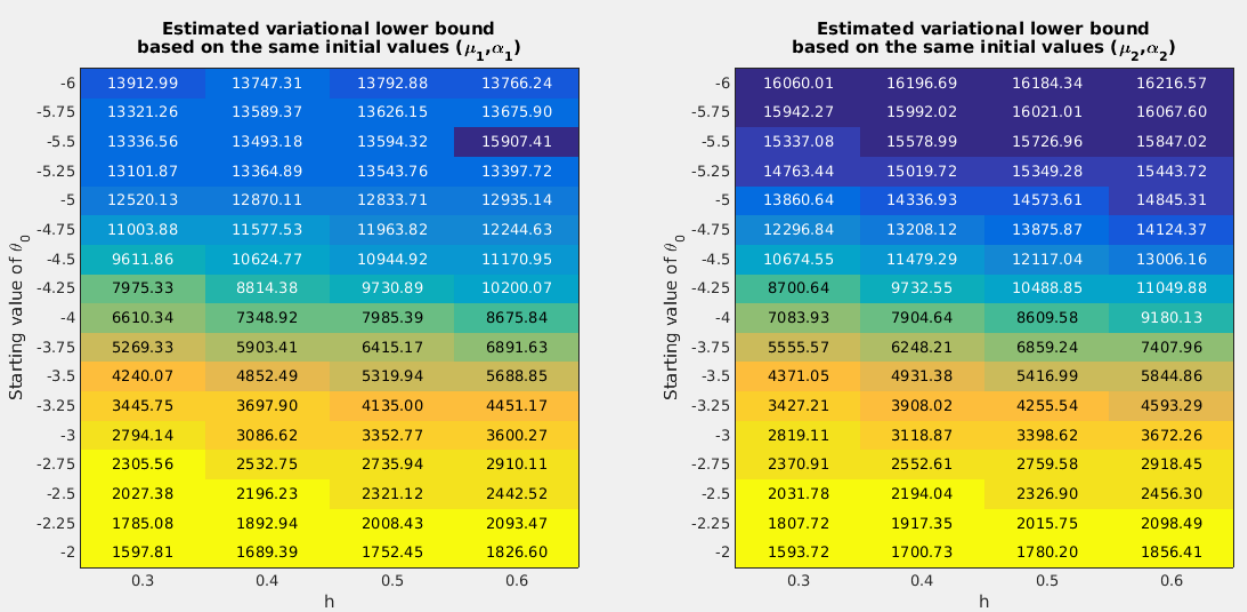


Figure F.19: Estimated proportion of trait-associated SNPs across 31 phenotypes.

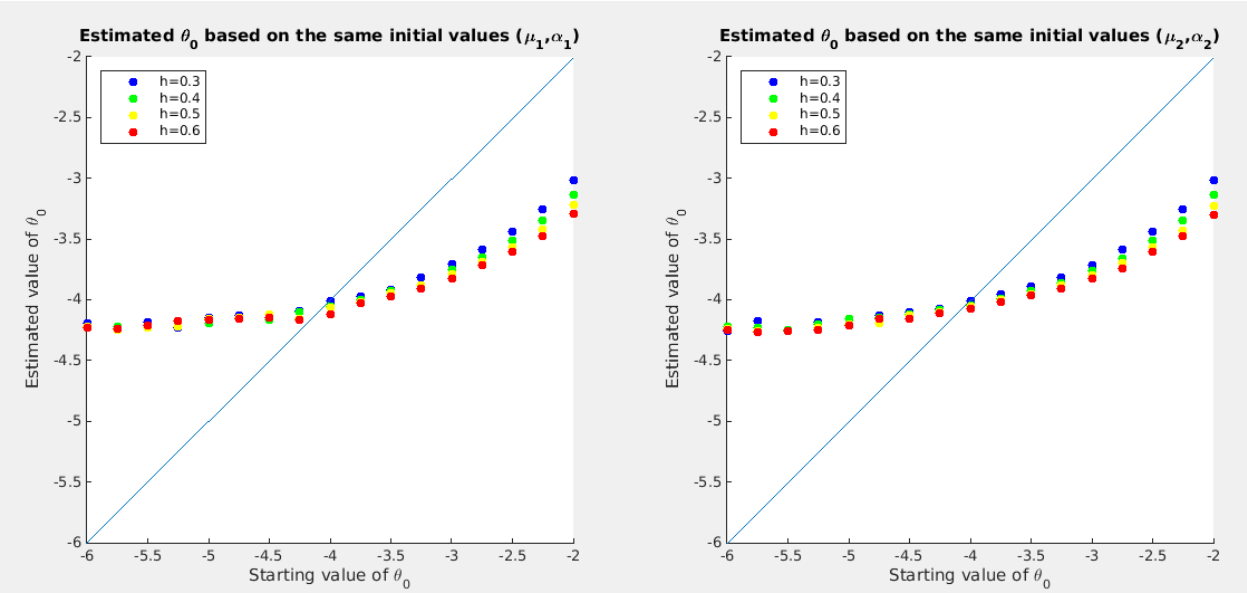
Supplementary Figure 20

A modified variational algorithm improves estimates of hyper-parameter and variational lower bound in the genome-wide multiple-SNP analysis of triglyceride summary data (Teslovich et al., 2010). The modified variational algorithm estimates both parameter β and hyper-parameter θ_0; in contrast, the default variational algorithm only estimates β with θ_0 fixed. See Supplementary Note of Zhu and Stephens (2017b) for details of the modified and default variational algorithms.

(a) Estimated variational lower bounds using the modified variational algorithm and two different random initializations.



(b) Estimated values of θ_0 versus starting values of θ_0, using the modified variational algorithm and two different random initializations. This result suggests that the modified algorithm can produce more consistent posterior estimate of θ_0 than the default algorithm for triglyceride summary data (Teslovich et al., 2010) [Supplementary Figure 19 of Zhu and Stephens (2017b)].



(c) Estimated variational lower bounds using the default variational algorithm, where the initial value is set as the optimal solution from the modified algorithm. This result suggests that the solution from the modified algorithm can serve as a better initialization for the default algorithm, compared with the random start strategy used in the default algorithm.

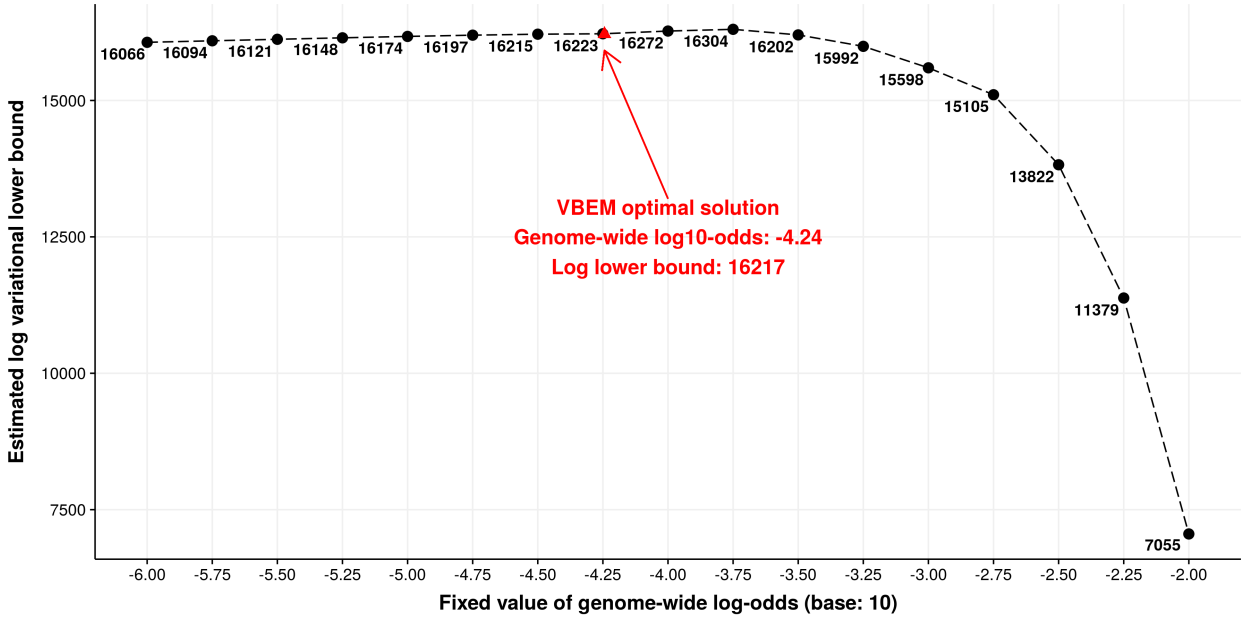
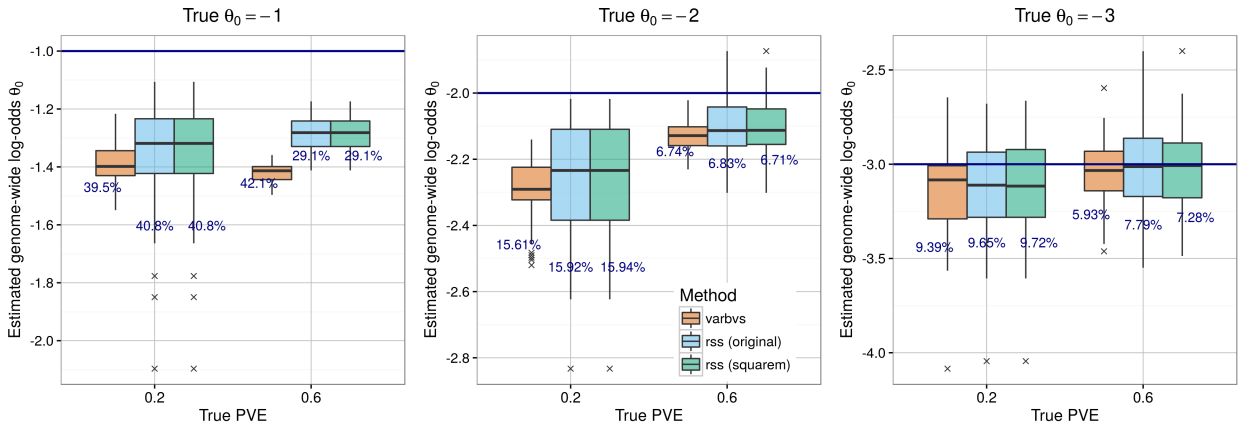


Figure F.20: A modified variational algorithm improves estimates of hyper-parameter and variational lower bound in the genome-wide multiple-SNP analysis of triglyceride summary data (Teslovich et al., 2010).

Supplementary Figure 21

Comparing analyses of individual-level data (Carbonetto and Stephens, 2012) with analyses of summary-level data under the baseline hypothesis. Here we use real genotypes of 12,758 SNPs on chromosome 16 from 1458 individuals in the UK Blood Service Control Group (Wellcome Trust Case Control Consortium , 2007) to simulate phenotype data under the baseline hypothesis. Specifically, we randomly select $12,758 \times (1+10^{-\theta_0})^{-1}$ "causal" SNPs with effect sizes coming from $\mathcal{N}(0,1)$, and then set the effect sizes of remaining SNPs as zero. We create datasets with true PVE $\in \{0.2, 0.6\}$ and true $\theta_0 \in \{1,2,3\}$ (100 independent replicates for each combination of true θ_0 and PVE values). We estimate LD matrix of 12,758 SNPs using genotypes of 1480 individuals from 1958 British Birth Cohort (Wellcome Trust Case Control Consortium , 2007). We use *Signal Transduction Pathway* (Biosystem, Reactome) to create SNP-pathway annotations when computing enrichment Bayes factor (BF).

(a) Estimates of genome-wide log10-odds (θ_0) from VARBVS (orange), RSS (blue) and RSS with SQUAREM (Varadhan and Roland, 2008) (green) in the baseline simulations. The accuracy of estimation is measured by the relative RMSE, which is defined as the root of mean squared error (RMSE) between the ratio of estimated over true θ_0 and 1. Relative RMSE for each method is reported (percentages under each box plot). The true values of θ_0 are shown as the solid horizontal lines. Each box plot summarizes results from 100 replicates.



(b) Type 1 error rates of VARBVS (orange), RSS (blue) and RSS with SQUAREM (green) in the baseline simulations. For each simulated dataset, a type 1 error is made if the enrichment BF is greater than the given cutoff for BF.

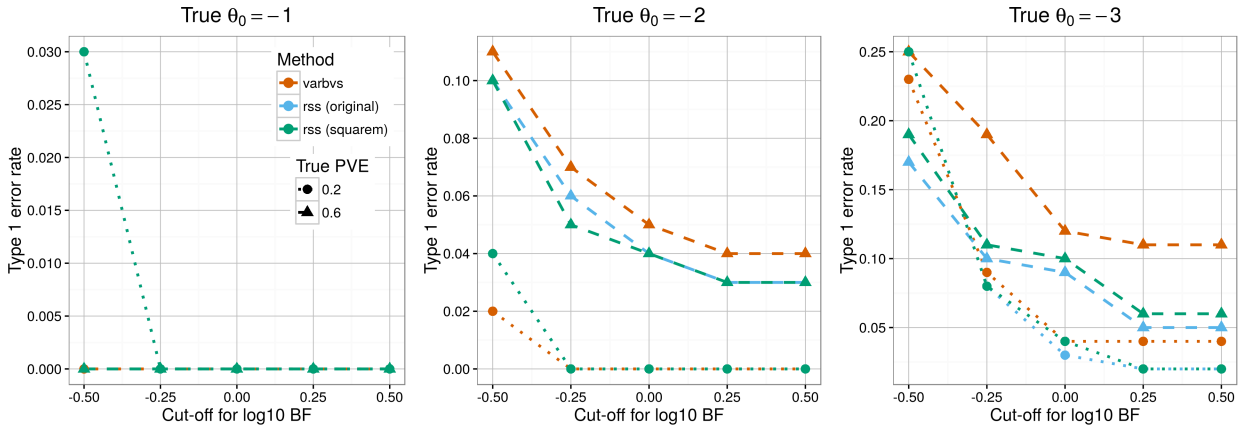
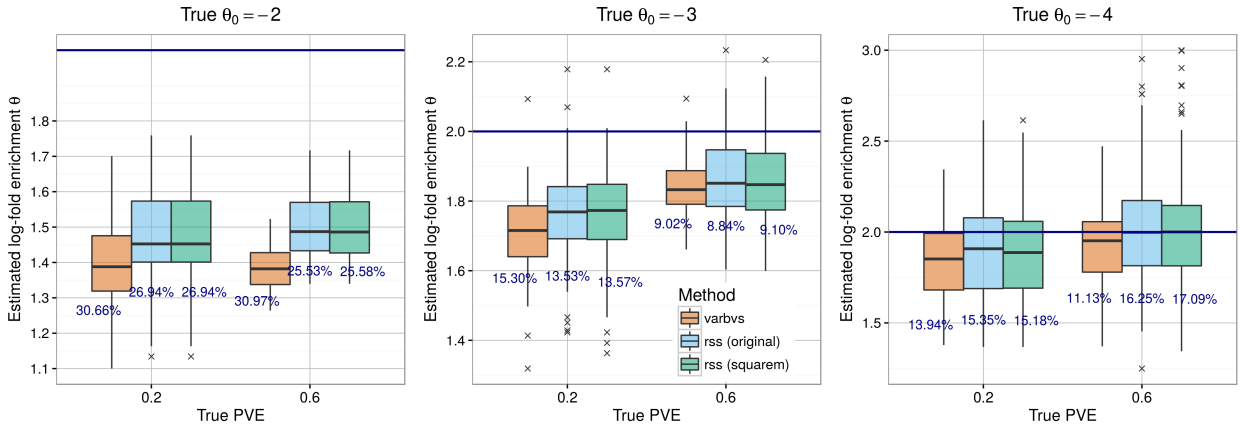


Figure F.21: Comparing analyses of individual-level data (Carbonetto and Stephens, 2012) with analyses of summary-level data under the baseline hypothesis.

Supplementary Figure 22

Comparing analyses of individual-level data (Carbonetto and Stephens, 2012) with analyses of summary-level data under the enrichment hypothesis. Here we use real genotypes of 12,758 SNPs on chromosome 16 from 1458 individuals in the UK Blood Service Control Group (Wellcome Trust Case Control Consortium , 2007) to simulate phenotype data under the enrichment hypothesis. Specifically, for each SNP j, we simulate its effect size from $\beta_j \sim (1-\pi_j)\delta_0 + \pi_j \mathcal{N}(0,1)$, where δ_0 denotes point mass at zero, $\pi_j = (1+10^{\theta_0+\theta a_j})^{-1}$, and $a_j = 1$ if SNP j is within $\pm$ 100 kb of transcribed region of a member gene in *Signal Transduction Pathway* (Biosystem, Reactome). We create datasets with true PVE $\in \{0.2, 0.6\}$, true $\theta_0 \in \{1,2,3\}$ and true $\theta = 2$ (100 independent replicates for each combination of true θ, θ_0 and PVE values).

(a) Estimates of log10-fold enrichment parameter (θ) from VARBVS (orange), RSS (blue) and RSS with SQUAREM (Varadhan and Roland, 2008) (green) in the enrichment simulations. Relative RMSE for each method is reported (percentages on top of box plots). The true values of θ are shown as the solid horizontal lines. Each box plot summarizes results from 100 replicates.



(b) Power of VARBVS (orange), RSS (blue) and RSS with SQUAREM (green) in the enrichment simulations. For each scenario, the power is computed as the fraction of datasets whose enrichment BFs are greater than the given cutoff for BF.

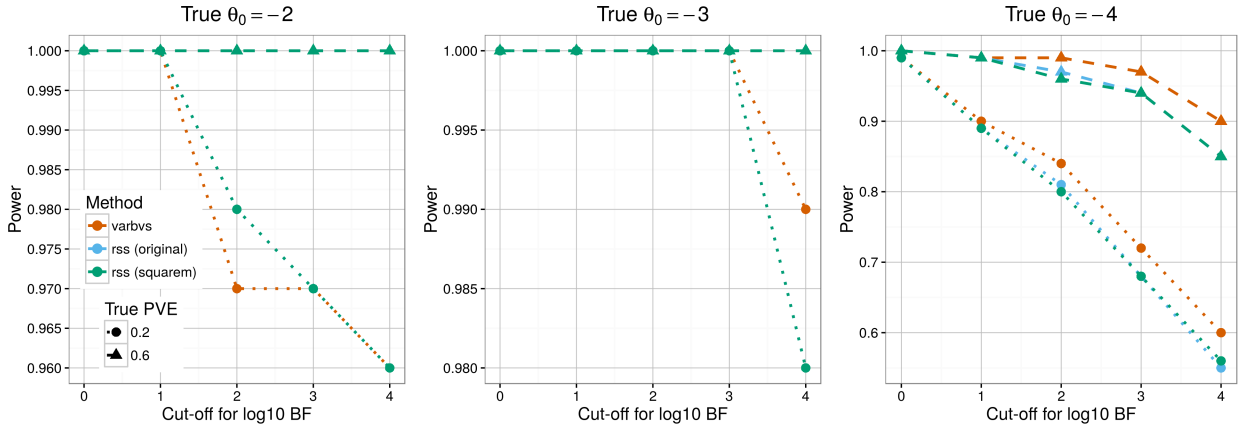


Figure F.22: Comparing analyses of individual-level data (Carbonetto and Stephens, 2012) with analyses of summary-level data under the enrichment hypothesis.

APPENDIX G

SUPPLEMENTARY TABLES OF ZHU AND STEPHENS (2017B)

Supplementary Table 1

Sample sizes and numbers of genetic variants in GWAS of 31 human phenotypes. For each trait, the number of "total" SNPs is the number of SNPs reported in the corresponding publication and/or summary statistics file, and the number of "analyzed" SNPs is the number of SNPs used in our analyses. Both columns are visualized in Supplementary Figure 1 of Zhu and Stephens (2017b).

Phenotype (abbreviation)	PMID	Number of SNPs		Sample size
		Total	Analyzed	(cases+controls)
Neurological phenotypes				
Amyotrophic lateral sclerosis (ALS)	27455348	8,709,433	1,162,845	12,577+23,475
Depressive symptoms (DS)	27089181	6,524,474	1,119,108	161,460
Alzheimer's disease (LOAD)	24162737	7,055,881	1,136,997	17,008+37,154
Neuroticism (NEU)	27089181	6,524,432	1,119,108	170,911
Schizophrenia (SCZ)	25056061	9,444,230	1,113,442	152,805
Anthropometric traits				
Body mass index (BMI)	25673413	2,554,637	1,012,465	234,069
Height (HEIGHT)	25282103	2,550,858	1,064,575	253,288
Waist-to-hip ratio (WHR)	25673412	2,542,431	1,008,898	142,762
Immune-related traits				
Crohn's disease (CD)	26192919	12,276,505	1,064,533	5,956+14,927
Inflammatory bowel disease (IBD)	26192919	12,716,083	1,081,481	12,882+21,770
Rheumatoid arthritis (RA)	24390342	8,747,962	1,158,064	14,361+43,923
Ulcerative colitis (UC)	26192919	12,255,196	1,092,170	6,968+20,464
Metabolic phenotypes				
Age at natural menopause (ANM)	26414677	2,418,695	1,047,412	69,360
Coronary artery disease (CAD)	26343387	9,455,778	1,121,322	60,801+123,504
Fasting glucose (FG)	22581228	2,628,879	1,114,610	58,074
Fasting insulin (FI)	22581228	2,627,848	1,114,592	51,750
Gout (GOUT)	23263486	2,538,056	1,061,037	2,115+67,259
High-density lipoprotein (HDL)	20686565	2,692,429	1,032,214	99,900
Heart rate (HR)	23583979	2,516,789	1,066,168	92,355
Low-density lipoprotein (LDL)	20686565	2,692,564	1,030,397	95,454
Myocardial infarction (MI)	26343387	9,289,491	1,111,568	42,561+123,504
Type 2 diabetes (T2D)	22885922	2,473,441	1,047,618	12,171+56,862
Total cholesterol (TC)	20686565	2,692,413	1,032,272	100,184
Triglycerides (TG)	20686565	2,692,560	1,030,671	96,598
Serum urate (URATE)	23263486	2,450,547	1,050,253	110,347
Hematopoietic traits				
Haemoglobin (HB)	23222517	2,593,078	1,116,281	61,155
Mean cell HB (MCH)	23222517	2,586,784	1,114,901	51,711
Mean cell HB concentration (MCHC)	23222517	2,588,875	1,115,595	56,475
Mean cell volume (MCV)	23222517	2,591,132	1,116,066	58,114
Packed cell volume (PCV)	23222517	2,591,079	1,115,725	53,089
Red blood cell count (RBC)	23222517	2,589,454	1,115,397	53,661

Table G.1: Sample sizes and numbers of genetic variants in GWAS of 31 human phenotypes.

Supplementary Table 2

Confounding adjustment in GWAS of 31 human phenotypes. Columns left to right: (1) phenotype and its abbreviation; (2) genomic control (GC) factor (Devlin and Roeder, 1999); (3) LD score (LDSC) regression intercept (Bulik-Sullivan et al., 2015b); (4) the number of top genotype-derived principal components (PCs) that were included as covariates in the single-SNP association testing (Price et al., 2006); (5) other covariates included in the single-SNP association testing. The genomic control factor λ_{GC} and the LD score regression intercept λ_{LDSC} are two measures of confounding biases such as population stratification. Values of $\lambda_{\mathrm{GC}} \approx 1$ or $\lambda_{\mathrm{LDSC}} \approx 1$ indicate little confounding effects, whereas $\lambda_{\mathrm{GC}} \geq 1$ or $\lambda_{\mathrm{LDSC}} \geq 1$ suggest possible existence of confounding biases. The "cohort" covariate denote all factors that are specific to study cohorts (e.g. genotyping array, study site).

Phenotype (abbreviation)	λ_{GC}	λ_{LDSC}	# of PCs	Other covariates
Neurological phenotypes				
Amyotrophic lateral sclerosis (ALS)	1.12	1.10	1-4	not shown
Depressive symptoms (DS)	1.17	1.01	4-15	sex, age, cohort
Alzheimer's disease (LOAD)	1.09	1.04	2-8	sex, age
Neuroticism (NEU)	1.32	1.00	4-15	sex, age, cohort
Schizophrenia (SCZ)	1.47	1.07	10	not shown
Anthropometric traits				
Body mass index (BMI)	1.08	1.02	not shown	sex, age, cohort
Height (HEIGHT)	1.94	1.05	not shown	sex, age, cohort
Waist-to-hip ratio (WHR)	1.01	0.93	not shown	sex, age, cohort, BMI
Immune-related traits				
Crohn's disease (CD)	1.13	1.03	10	not shown
Inflammatory bowel disease (IBD)	1.16	1.06	15	not shown
Rheumatoid arthritis (RA)	1.07	0.98	5-10	not shown
Ulcerative colitis (UC)	1.11	1.04	7	not shown
Metabolic phenotypes				
Age at natural menopause (ANM)	not shown	not shown	not shown	cohort
Coronary artery disease (CAD)	1.18	1.05	not shown	not shown
Fasting glucose (FG)	not shown	not shown	not shown	sex, age, cohort, BMI
Fasting insulin (FI)	1.07	1.02	not shown	sex, age, cohort, BMI
Gout (GOUT)	1.03	not shown	2-10	sex, age, cohort
High-density lipoprotein cholesterol (HDL)	1.14	1.01	not shown	sex, age, cohort
Heart rate (HR)	1.11	1.01	not shown	sex, age, cohort, BMI
Low-density lipoprotein cholesterol (LDL)	1.10	1.00	not shown	sex, age, cohort
Myocardial infarction (MI)	not shown	not shown	not shown	not shown
Type 2 diabetes (T2D)	1.10	1.03	not shown	cohort
Total cholesterol (TC)	1.11	1.01	not shown	sex, age, cohort
Triglycerides (TG)	1.12	1.00	not shown	sex, age, cohort
Serum urate (URATE)	1.12	1.01	2-10	sex, age, cohort
Hematopoietic traits				
Haemoglobin (HB)	1.10	not shown	2-10	sex, age, cohort
Mean cell HB (MCH)	1.13	not shown	2-10	sex, age, cohort
Mean cell HB concentration (MCHC)	1.08	not shown	2-10	sex, age, cohort
Mean cell volume (MCV)	1.14	not shown	2-10	sex, age, cohort
Packed cell volume (PCV)	1.10	not shown	2-10	sex, age, cohort
Red blood cell count (RBC)	1.14	not shown	2-10	sex, age, cohort

Table G.2: Confounding adjustment in GWAS of 31 human phenotypes.

Supplementary Table 3

	Round 1 analysis		Round 2 analysis	
Phenotype (abbreviation)	h	θ_0	h	θ_0
Neurological phenotypes				
Amyotrophic lateral sclerosis (ALS)	(0.3:0.1:0.6)	(-6:0.25:-3)	(0.3:0.1:0.6)	(-6:0.05:-5)
Depressive symptoms (DS)	(0.3:0.1:0.6)	(-6:0.25:-2)	(0.3:0.1:0.6)	(-6:0.05:-5)
Alzheimer's disease (LOAD)	(0.3:0.1:0.6)	(-5.25:0.25:-3.25)	0.3	(-5.25:0.025:-4.75)
Neuroticism (NEU)	(0.3:0.1:0.6)	(-4.5:0.25:-2)	0.3	(-4.5:0.025:-4)
Schizophrenia (SCZ)	(0.3:0.1:0.6)	(-4:0.25:-1)	0.3	(-2.25:0.025:-1.75)
Anthropometric traits				
Body mass index (BMI)	(0.3:0.1:0.6)	(-5:0.25:-1)	0.3	(-4.25:0.025:-3.75)
Height (HEIGHT)	(0.3:0.1:0.6)	(-4:0.25:-1)	(0.3:0.1:0.4)	(-2.25:0.025:-1.75)
Waist-to-hip ratio (WHR)	(0.3:0.1:0.6)	(-6:0.25:-3)	(0.3:0.1:0.6)	(-6:0.05:-5)
Immune-related traits				
Crohn's disease (CD)	(0.3:0.1:0.6)	(-4:0.25:-2)	0.3	(-3.25:0.025:-2.75)
Inflammatory bowel disease (IBD)	(0.3:0.1:0.6)	(-4:0.25:-2)	0.3	(-3.25:0.025:-2.75)
Rheumatoid arthritis (RA)	(0.3:0.1:0.6)	(-4.5:0.25:-2)	0.3	(-3.5:0.025:-3)
Ulcerative colitis (UC)	(0.3:0.1:0.6)	(-4:0.25:-2)	0.3	(-3.25:0.025:-2.75)
Metabolic phenotypes				
Age at natural menopause (ANM)	(0.3:0.1:0.6)	(-6:0.25:-2)	0.4	(-5.75:0.025:-5.25)
Coronary artery disease (CAD)	(0.3:0.1:0.6)	(-5:0.25:-2)	0.3	(-4.25:0.25:-3.75)
Fasting glucose (FG)	(0.3:0.1:0.6)	(-6:0.25:-3)	0.6	(-6:0.05:-5)
Fasting insulin (FI)	(0.3:0.1:0.6)	(-6:0.25:-3)	0.6	(-6.25:0.025:-5.75)
Gout (GOUT)	(0.3:0.1:0.6)	(-5.5:0.25:-2)	(0.3:0.1:0.6)	(-5.5:0.05:-4.5)
High-density lipoprotein (HDL)	(0.3:0.1:0.6)	(-5:0.25:-2)	0.3	(-3.75:0.025:-3.25)
Heart rate (HR)	(0.3:0.1:0.6)	(-5:0.25:-2)	(0.3:0.1:0.4)	(-4.5:0.025:-4)
Low-density lipoprotein (LDL)	(0.3:0.1:0.6)	(-5:0.25:-2)	0.3	(-4:0.025:-3.5)
Myocardial infarction (MI)	(0.3:0.1:0.6)	(-5:0.25:-2)	0.3	(-4.5:0.025:-4)
Type 2 diabetes (T2D)	(0.3:0.1:0.6)	(-5:0.25:-2)	(0.3:0.1:0.6)	(-4.75:0.025:-4.25)
Total cholesterol (TC)	(0.3:0.1:0.6)	(-5:0.25:-2)	0.6	(-5:0.025:-4.5)
Triglycerides (TG)	(0.3:0.1:0.6)	(-6:0.25:-3)	0.5	(-6.25:0.025:-5.75)
Serum urate (URATE)	(0.3:0.1:0.6)	(-5.5:0.25:-2)	0.5	(-5.5:0.025:-5)
Hematopoietic traits				
Haemoglobin (HB)	(0.3:0.1:0.6)	(-6:0.25:-3)	0.6	(-6.25:0.025:-5.75)
Mean cell HB (MCH)	(0.3:0.1:0.6)	(-5:0.25:-2)	0.6	(-4.75:0.05:-3.75)
Mean cell HB concentration (MCHC)	(0.3:0.1:0.6)	(-6:0.25:-3)	(0.3:0.1:0.6)	(-6:0.05:-5)
Mean cell volume (MCV)	(0.3:0.1:0.6)	(-5:0.25:-2)	0.6	(-4.25:0.025:-3.75)
Packed cell volume (PCV)	(0.3:0.1:0.6)	(-5:0.25:-2)	0.6	(-5.25:0.025:-4.75)
Red blood cell count (RBC)	(0.3:0.1:0.6)	(-5:0.25:-2)	(0.3:0.1:0.6)	(-3.75:0.025:-3.25)

Table G.3: Grids of hyper-parameters used in genome-wide multiple-SNP analyses of 31 human phenotypes, assuming no pathways are enriched.

Supplementary Table 4

	Round 1 analysis			Round 2 analysis		
Phenotype (abbreviation)	h	θ_0	θ	h	θ_0	θ
Neurological phenotypes						
Amyotrophic lateral sclerosis (ALS)	(0.3:0.1:0.6)	(-6:0.25:-5)	(0:0.6:6)	(0.3:0.1:0.6)	(-6:0.05:-5)	(0:0.1:4)
Depressive symptoms (DS)	(0.3:0.1:0.6)	(-6:0.25:-5)	(0:0.6:6)	(0.3:0.1:0.6)	(-6:0.05:-5)	(0:0.15:6)
Alzheimer's disease (LOAD)	0.6	-5	(0:0.025:5)	0.6	(-5.150:0.025:-5.075)	(0:0.01:4)
Neuroticism (NEU)	(0.3:0.1:0.4)	(-4.5:0.25:-4)	(0:0.1:4)	0.3	(-4.5:0.025:-4)	(0:0.037:3.7)
Schizophrenia (SCZ)	0.3	-2	(0:0.01:2)	0.3	(-2.2:0.025:-2.05)	(0:0.01:2)
Anthropometric traits						
Body mass index (BMI)	0.3	-4	(0:0.02:4)	0.3	(-4.2:0.025:-3.8)	(0:0.018:3.5)
Height (HEIGHT)	0.3	-2	(0:0.01:2)	0.3	(-2.075:0.025:-1.925)	(0:0.01:1)
Waist-to-hip ratio (WHR)	0.3	-3	(0:0.015:3)	0.3	(-3:0.025:-2.95)	(0:0.01:3)
Immune-related traits						
Crohn's disease (CD)	0.3	-3	(0:0.015:3)	0.3	-3	(0:0.01:2)
Inflammatory bowel disease (IBD)	0.3	-3	(0:0.015:3)	0.3	(-3:0.025:-2.8)	(0:0.02:2)
Rheumatoid arthritis (RA)	0.3	-3.25	(0:0.016:3.25)	0.3	(-3.25:0.025:-3.175)	(0:0.01:2)
Ulcerative colitis (UC)	0.3	-3	(0:0.015:3)	0.3	(-3.175:0.025:-2.775)	(0:0.025:2.5)
Metabolic phenotypes						
Age at natural menopause (ANM)	0.4	-5.5	(0:0.028:5.5)	0.4	(-5.75:0.025:-5.7)	(0:0.04:4.5)
Coronary artery disease (CAD)	0.3	-4	(0:0.02:4)	0.3	(-4.025:0.025:-3.775)	(0:0.035:3.5)
Fasting glucose (FG)	0.6	-5.25	(0:0.026:5.25)	0.6	(-6:0.05:-5.75)	(0.1:0.1:4.5)
Fasting insulin (FI)	0.6	-6	(0:0.03:6)	0.6	(-6.25:0.025:-6)	(0:0.019:3.8)
Gout (GOUT)	(0.3:0.1:0.6)	(-5.5:0.25:-4.75)	(0:0.46:5.5)	(0.3:0.1:0.6)	(-5.5:0.05:-4.6)	(0:0.1:5)
High-density lipoprotein (HDL)	0.3	-3.5	(0:0.018:3.5)	0.3	(-3.575:0.025:-3.5)	(0:0.01:3)
Heart rate (HR)	(0.3:0.1:0.4)	(-4.5:0.25:-4.25)	(0:(4.5/50):4.5)	(0.3:0.1:0.4)	(-4.5:0.025:-4.1)	(0:0.038:3.8)
Low-density lipoprotein (LDL)	0.3	-3.75	(0:0.019:3.75)	0.3	(-3.625:0.025:-3.55)	(0:0.01:3)
Myocardial infarction (MI)	0.3	(-4.5:0.25:-4)	(0:0.067:4.5)	0.3	(-4.475:0.025:-4)	(0:0.045:4.5)
Type 2 diabetes (T2D)	(0.4:0.1:0.6)	-4.5	(0:0.067:4.5)	(0.3:0.1:0.6)	(-4.75:0.025:-4.35)	(0:0.3:3)
Total cholesterol (TC)	0.6	-4.75	(0:0.024:4.75)	0.6	(-4.8:0.025:-4.55)	(0:0.02:4)
Triglycerides (TG)	0.5	-4	(0:0.03:4)	0.5	(-6.25:0.025:-6.1)	(0:0.02:5.2)
Serum urate (URATE)	0.5	-5.25	(0:0.026:5.25)	0.5	(-5.4:0.025:-5)	(0:0.1:4.7)
Hematopoietic traits						
Haemoglobin (HB)	0.6	-6	(0:0.03:6)	0.6	(-6.25:0.025:-5.9)	(0:0.04:4.4)
Mean cell HB (MCH)	0.6	-4	(0:0.02:4)	0.6	(-4.7:0.05:-4.35)	(0:0.015:3)
MCH concentration (MCHC)	(0.3:0.1:0.6)	(-6:0.25:-5.25)	(0:0.5:6)	(0.3:0.1:0.6)	(-6:0.05:-5)	(0:0.1:5)
Mean cell volume (MCV)	0.6	-4	(0:0.02:4)	0.6	(-4.225:0.025:-4.125)	(0:0.02:3)
Packed cell volume (PCV)	0.6	-5	(0:0.025:5)	0.6	(-5.25:0.025:-5.15)	(0:0.02:4.5)
Red blood cell count (RBC)	(0.3:0.1:0.6)	-3.5	(0:0.07:3.5)	(0.5:0.1:0.6)	(-3.7:0.025:-3.6)	(0:0.035:3.5)

Table G.4: Grids of hyper-parameters used in genome-wide multiple-SNP analyses of 31 human phenotypes, assuming a candidate pathway is enriched.

REFERENCES

1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, 2010.

1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

F. W. Albert and L. Kruglyak. The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics*, 16(4):197–212, 2015.

C. A. Anderson, G. Boucher, C. W. Lees, A. Franke, M. D'Amato, K. D. Taylor, J. C. Lee, P. Goyette, M. Imielinski, A. Latiano, et al. Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nature Genetics*, 43(3):246–252, 2011.

G. D. Anderson, S. D. Hauser, K. L. McGarity, M. E. Bremer, P. C. Isakson, and S. A. Gregory. 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 (11):2672, 1996.

R. I. Aqeilan, M. Q. Hassan, A. de Bruin, J. P. Hagan, S. Volinia, T. Palumbo, S. Hussain, S.-H. Lee, T. Gaur, G. S. Stein, et al. The *wwox* tumor suppressor is essential for postnatal survival and normal bone metabolism. *Journal of Biological Chemistry*, 283(31):21629–21639, 2008.

O. Aseem, B. T. Smith, M. A. Cooley, B. A. Wilkerson, K. M. Argraves, A. T. Remaley, and W. S. Argraves. Cubilin maintains blood levels of HDL and albumin. *Journal of the American Society of Nephrology*, 25(5):1028–1036, 2014.

N. E. Banovich, X. Lan, G. McVicker, B. Van de Geijn, J. F. Degner, J. D. Blischak, J. Roux, J. K. Pritchard, and Y. Gilad. Methylation QTLs are associated with coordinated changes

in transcription factor binding, histone modifications, and gene expression levels. *PLoS Genetics*, 10(9):e1004663, 2014.

R. F. Barber and E. J. Candès. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, 2015.

M. S. Beeri, M. Rapp, J. Silverman, J. Schmeidler, H. Grossman, J. Fallon, D. Purohit, D. Perl, A. Siddiqui, G. Lesser, et al. Coronary artery disease is associated with Alzheimer disease neuropathology in APOE4 carriers. *Neurology*, 66(9):1399–1404, 2006.

F. Begum, D. Ghosh, G. C. Tseng, and E. Feingold. Comprehensive literature review and statistical considerations for GWAS meta-analysis. *Nucleic Acids Research*, 40(9):3777, 2012.

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

D. D. Boos. A converse to Scheffe's Theorem. *The Annals of Statistics*, 13(1):423–427, 03 1985.

L. Bottolo and S. Richardson. Evolutionary stochastic search for Bayesian model exploration. *Bayesian Analysis*, 5(3):583–618, 2010.

L. Bottolo, M. Chadeau-Hyam, D. I. Hastie, T. Zeller, B. Liquet, P. Newcombe, L. Yengo, P. S. Wild, A. Schillert, A. Ziegler, et al. GUESS-ing polygenic associations with multiple phenotypes using a gpu-based evolutionary stochastic search algorithm. *PLoS Genetics*, 9(8):e1003657, 2013.

E. A. Boyle, Y. I. Li, and J. K. Pritchard. An expanded view of complex traits: From polygenic to omnigenic. *Cell*, 169(7):1177–1186, 2017.

D. Brzyski, C. B. Peterson, P. Sobczyk, E. J. Candes, M. Bogdan, and C. Sabatti. Controlling the rate of GWAS false discoveries. *Genetics*, 205(1):61–75, 2017.

P. Bühlmann, M. Kalisch, and L. Meier. High-dimensional statistics with a view toward applications in biology. *Annual Review of Statistics and Its Application*, 1:255–278, 2014.

B. Bulik-Sullivan, H. K. Finucane, V. Anttila, A. Gusev, F. R. Day, P.-R. Loh, L. Duncan, J. R. Perry, N. Patterson, E. B. Robinson, et al. An atlas of genetic correlations across human diseases and traits. *Nature Genetics*, 47(11):1236–1241, 2015a.

B. K. Bulik-Sullivan, P.-R. Loh, H. K. Finucane, S. Ripke, J. Yang, N. Patterson, M. J. Daly, A. L. Price, B. M. Neale, S. W. G. of the Psychiatric Genomics Consortium, et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, 47(3):291–295, 2015b.

J. N. Buxbaum, Z. Ye, N. Reixach, L. Friske, C. Levy, P. Das, T. Golde, E. Masliah, A. R. Roberts, and T. Bartfai. Transthyretin protects Alzheimer's mice from the behavioral and biochemical effects of Aβ toxicity. *Proceedings of the National Academy of Sciences*, 105 (7):2681–2686, 2008.

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

P. Carbonetto and M. Stephens. 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(10):e1003770, 2013.

C. M. Carvalho, N. G. Polson, and J. G. Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010.

G. Casella and C. P. Robert. Rao-blackwellisation of sampling schemes. *Biometrika*, 83(1): 81–94, 1996.

R. Caspi, T. Altman, R. Billington, K. Dreher, H. Foerster, C. A. Fulcher, T. A. Holland, I. M. Keseler, A. Kothari, A. Kubo, M. Krummenacker, M. Latendresse, L. A. Mueller, Q. Ong, S. Paley, P. Subhraveti, D. S. Weaver, D. Weerasinghe, P. Zhang, and P. D. Karp. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Research*, 42(D1):D459–D471, 2014.

E. G. Cerami, B. E. Gross, E. Demir, I. Rodchenkov, O. Babur, N. Anwar, N. Schultz, G. D. Bader, and C. Sander. Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Research*, 39(suppl 1):D685–D690, 2011.

S. R. Chamberlin. *The foundation and application of inferential estimation.* PhD thesis, University of Waterloo, 1989.

R. W. Cheloha, S. H. Gellman, J.-P. Vilardaga, and T. J. Gardella. PTH receptor-1 signalling – mechanistic insights and therapeutic prospects. *Nature Reviews Endocrinology*, 11(12): 712–724, 2015.

W. Chen and D. J. Schaid. PedBLIMP: extending linear predictors to impute genotypes in pedigrees. *Genetic Epidemiology*, 38(6):531–541, 2014.

W. Chen, B. R. Larrabee, I. G. Ovsyannikova, R. B. Kennedy, I. H. Haralambieva, G. A. Poland, and D. J. Schaid. Fine mapping causal variants with an approximate bayesian method using marginal test statistics. *Genetics*, 200(3):719–736, 2015.

A. Cortes and M. A. Brown. Promise and pitfalls of the Immunochip. *Arthritis Research & Therapy*, 13(1):101, 2011.

D. Cox. Unbiased estimating equations derived from statistics that are functions of a parameter. *Biometrika*, 80(4):905–909, 1993.

D. Croft, A. F. Mundo, R. Haw, M. Milacic, J. Weiser, G. Wu, M. Caudy, P. Garapati, M. Gillespie, M. R. Kamdar, B. Jassal, S. Jupe, L. Matthews, B. May, S. Palatnik, K. Rothfels, V. Shamovsky, H. Song, M. Williams, E. Birney, H. Hermjakob, L. Stein, and P. D'Eustachio. The Reactome pathway knowledgebase. *Nucleic Acids Research*, 42 (D1):D472–D477, 2014.

F. R. Day, K. S. Ruth, D. J. Thompson, K. L. Lunetta, N. Pervjakova, D. I. Chasman, L. Stolk, H. K. Finucane, P. Sulem, B. Bulik-Sullivan, et al. Large-scale genomic analyses link reproductive aging to hypothalamic signaling, breast cancer susceptibility and BRCA1-mediated DNA repair. *Nature Genetics*, 47(11):1294–1303, 2015.

R. A. Daynes and D. C. Jones. Emerging roles of PPARs in inflammation and immunity. *Nature Reviews Immunology*, 2(10):748–759, 2002.

C. A. de Leeuw, J. M. Mooij, T. Heskes, and D. Posthuma. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Computational Biology*, 11(4):e1004219, 2015.

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

G. de los Campos, D. Sorensen, and D. Gianola. Genomic heritability: what is it? *PLoS Genetics*, 11(5):e1005048, 2015.

S. Del Mare, K. C. Kurek, G. S. Stein, J. B. Lian, and R. I. Aqeilan. Role of the *wwox* tumor suppressor gene in bone homeostasis and the pathogenesis of osteosarcoma. *American Journal of Cancer Research*, 1(5):585, 2011.

M. Den Hoed, M. Eijgelsheim, T. Esko, B. J. Brundel, D. S. Peal, D. M. Evans, I. M. Nolte, A. V. Segrè, H. Holm, R. E. Handsaker, et al. Identification of heart rate-associated loci and their effects on cardiac conduction and rhythm disorders. *Nature Genetics*, 45(6): 621–631, 2013.

B. Devlin and K. Roeder. Genomic control for association studies. *Biometrics*, 55(4):997–1004, 1999.

K. K. Dey, C. J. Hsiao, and M. Stephens. Visualizing the structure of RNA-seq expression data using grade of membership models. *PLoS Genetics*, 13(3):e1006599, 2017.

G. Di Paolo and T.-W. Kim. Linking lipids to Alzheimer's disease: cholesterol and beyond. *Nature Reviews Neuroscience*, 12(5):284–296, 2011.

P. Donnelly. Progress and challenges in genome-wide association studies in humans. *Nature*, 456(7223):728–731, 2008.

H. Du, L. Li, D. Bennett, Y. Guo, I. Turnbull, L. Yang, F. Bragg, Z. Bian, Y. Chen, J. Chen, et al. Fresh fruit consumption in relation to incident diabetes and diabetic vascular complications: A 7-y prospective study of 0.5 million Chinese adults. *PLoS Medicine*, 14(4): e1002279, 2017.

S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid monte carlo. *Physics Letters B*, 195(2):216–222, 1987.

P. C. Dubois, G. Trynka, L. Franke, K. A. Hunt, J. Romanos, A. Curtotti, A. Zhernakova, G. A. Heap, R. Ádány, A. Aromaa, et al. Multiple common variants for celiac disease influencing immune gene expression. *Nature Genetics*, 42(4):295–302, 2010.

M. O. Duff, S. Olson, X. Wei, S. C. Garrett, A. Osman, M. Bolisetty, A. Plocik, S. E. Celniker, and B. R. Graveley. Genome-wide identification of zero nucleotide recursive splicing in Drosophila. *Nature*, 521(7552):376–379, 2015.

B. Efron. Bayes and likelihood calculations from confidence intervals. *Biometrika*, 80(1): 3–26, 1993.

B. Efron. R. A. Fisher in the 21st century (Invited paper presented at the 1996 R. A. Fisher Lecture). *Statistical Science*, 13(2):95–122, 05 1998.

G. B. Ehret, D. Lamparter, C. J. Hoggart, J. C. Whittaker, J. S. Beckmann, Z. Kutalik, Genetic Investigation of Anthropometric Traits Consortium, et al. A multi-SNP locus-association method reveals a substantial fraction of the missing heritability. *The American Journal of Human Genetics*, 91(5):863–871, 2012.

S. L. Elshaer and A. B. El-Remessy. Implication of the neurotrophin receptor p75NTR in vascular diseases: beyond the eye. *Expert Review of Ophthalmology*, 12(2):149–158, 2017.

M. Erbe, B. Hayes, L. Matukumalli, S. Goswami, P. Bowman, C. Reich, B. Mason, and M. Goddard. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *Journal of Dairy Science*, 95(7):4114–4129, 2012.

E. Evangelou and J. P. Ioannidis. Meta-analysis methods for genome-wide association studies and beyond. *Nature Reviews Genetics*, 14(6):379–389, 2013.

M. Evangelou, F. Dudbridge, and L. Wernisch. Two novel pathway analysis methods based on a hierarchical model. *Bioinformatics*, 30(5):690–697, 2014.

L. Fagerberg, B. M. Hallström, P. Oksvold, C. Kampf, D. Djureinovic, J. Odeberg, M. Habuka, S. Tahmasebpoor, A. Danielsson, K. Edlund, et al. Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Molecular & Cellular Proteomics*, 13(2):397–406, 2014.

K. K.-H. Farh, A. Marson, J. Zhu, M. Kleinewietfeld, W. J. Housley, S. Beik, N. Shoresh, H. Whitton, R. J. Ryan, A. A. Shishkin, et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, 518(7539):337–343, 2015.

H. K. Finucane, B. Bulik-Sullivan, A. Gusev, G. Trynka, Y. Reshef, P.-R. Loh, V. Anttila, H. Xu, C. Zang, K. Farh, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics*, 47(11):1228–1235, 2015.

R. A. Fisher. Inverse probability. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 26, pages 528–535. Cambridge University Press, 1930.

K. A. Frazer, D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L. Stuve, R. A. Gibbs, J. W. Belmont, A. Boudreau, P. Hardenbol, S. M. Leal, et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164):851–861, 2007.

E. R. Gamazon, H. E. Wheeler, K. P. Shah, S. V. Mozaffari, K. Aquino-Michaels, R. J. Carroll, A. E. Eyler, J. C. Denny, D. L. Nicolae, N. J. Cox, et al. A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics*, 47(9):1091–1098, 2015.

L. Y. Geer, A. Marchler-Bauer, R. C. Geer, L. Han, J. He, S. He, C. Liu, W. Shi, and S. H. Bryant. The NCBI BioSystems database. *Nucleic Acids Research*, 38(suppl 1):D492–D496, 2010.

E. I. George and R. E. McCulloch. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

E. I. George and R. E. McCulloch. Approaches for Bayesian variable selection. *Statistica Sinica*, pages 339–373, 1997.

C. Giambartolomei, D. Vukcevic, E. E. Schadt, L. Franke, A. D. Hingorani, C. Wallace, and V. Plagnol. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genetics*, 10(5):e1004383, 2014.

Global Lipids Genetics Consortium. Discovery and refinement of loci associated with lipids levels. *Nature Genetics*, 45(11):1274–1283, 2013.

P. Gopalan, W. Hao, D. Blei, and J. Storey. Scaling probabilistic models of genetic variation to millions of humans. *Nature Genetics*, 48(12):1587, 2016.

W. H. Greene. *Econometric Analysis, 7th Edition*. Pearson Education, 2012.

Y. Guan and S. Krone. Small-world mcmc and convergence to multi-modal distributions: From slow mixing to fast mixing. *The Annals of Applied Probability*, 17(1):284–304, 2007.

Y. Guan and M. Stephens. Practical issues in imputation-based association mapping. *PLoS Genetics*, 4(12):e1000279, 2008.

Y. Guan and M. Stephens. Bayesian variable selection regression for genome-wide association studies, and other large-scale problems. *The Annals of Applied Statistics*, 5(3): 1780–1815, 2011.

Y. Guan and K. Wang. Whole-genome multi-SNP-phenotype association analysis. In K.-A. Do, Z. S. Qin, and M. Vannucci, editors, *Advances in Statistical Bioinformatics*, pages 224–243. Cambridge University Press, 2013. ISBN 9781139226448.

A. Gusev, A. Ko, H. Shi, G. Bhatia, W. Chung, B. W. Penninx, R. Jansen, E. J. de Geus, D. I. Boomsma, F. A. Wright, et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics*, 48(3):245–252, 2016.

M. Gutierrez-Arcelus, T. Lappalainen, S. B. Montgomery, A. Buil, H. Ongen, A. Yurovsky, J. Bryois, T. Giger, L. Romano, A. Planchon, et al. Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *Elife*, 2:e00523, 2013.

S. F. Hansson, U. Andréasson, M. Wall, I. Skoog, N. Andreasen, A. Wallin, H. Zetterberg, and K. Blennow. Reduced levels of amyloid-β-binding proteins in cerebrospinal fluid from Alzheimer's disease patients. *Journal of Alzheimer's Disease*, 16(2):389–397, 2009.

J. W. Harmon. *The likelihood pivot: performing inference with confidence.* PhD thesis, University of Washington, 2015.

Y. Hasin, M. Seldin, and A. Lusis. Multi-omics approaches to disease. *Genome Biology*, 18 (1):83, 2017.

X. He, C. K. Fuller, Y. Song, Q. Meng, B. Zhang, X. Yang, and H. Li. Sherlock: detecting gene-disease associations by matching patterns of expression QTL and GWAS. *The American Journal of Human Genetics*, 92(5):667–680, 2013.

F. L. Heppner, R. M. Ransohoff, and B. Becher. Immune attack: the role of inflammation in Alzheimer disease. *Nature Reviews Neuroscience*, 16(6):358–372, 2015.

T. Hesterberg, N. H. Choi, L. Meier, C. Fraley, et al. Least angle and $\ell 1$ penalized regression: A review. *Statistics Surveys*, 2:61–93, 2008.

L. A. Hindorff, P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins, and T. A. Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, 106 (23):9362–9367, 2009.

P. Hoff and J. Wakefield. Bayesian sandwich posteriors for pseudo-true parameters: A discussion of "Bayesian inference with misspecified models" by Stephen Walker. *Journal of Statistical Planning and Inference*, 143(10):1638–1642, 2013. ISSN 0378-3758.

M. D. Hoffman, D. M. Blei, C. Wang, and J. W. Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

C. J. Hoggart, J. C. Whittaker, M. De Iorio, and D. J. Balding. Simultaneous analysis of all snps in genome-wide and re-sequencing association studies. *PLoS Genetics*, 4(7): e1000130, 2008.

D. Holland, Y. Wang, W. K. Thompson, A. Schork, C.-H. Chen, M.-T. Lo, A. Witoelar, T. Werge, M. O'Donovan, O. A. Andreassen, et al. Estimating effect sizes and expected replication probabilities from GWAS summary statistics. *Frontiers in Genetics*, 7, 2016.

F. Hormozdiari, E. Kostem, E. Y. Kang, B. Pasaniuc, and E. Eskin. Identifying causal variants at loci with multiple signals of association. *Genetics*, 198(2):497–508, 2014.

F. Hormozdiari, M. van de Bunt, A. V. Segre, X. Li, J. W. J. Joo, M. Bilow, J. H. Sul, S. Sankararaman, B. Pasaniuc, and E. Eskin. Colocalization of GWAS and eQTL signals detects target genes. *The American Journal of Human Genetics*, 99(6):1245–1260, 2016.

S.-D. Hsu, Y.-T. Tseng, S. Shrestha, Y.-L. Lin, A. Khaleel, C.-H. Chou, C.-F. Chu, H.-Y. Huang, C.-M. Lin, S.-Y. Ho, T.-Y. Jian, F.-M. Lin, T.-H. Chang, S.-L. Weng, K.-W. Liao, I.-E. Liao, C.-C. Liu, and H.-D. Huang. miRTarBase update 2014: an information resource for experimentally validated mirna-target interactions. *Nucleic Acids Research*, 42(D1):D78–D85, 2014.

P. J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 221–233, Berkeley, California, 1967. University of California Press.

T. Iioka, K. Furukawa, A. Yamaguchi, H. Shindo, S. Yamashita, and T. Tsukazaki. P300/cbp acts as a coactivator to cartilage homeoprotein-1 (cart1), paired-like homeoprotein, through acetylation of the conserved lysine residue adjacent to the homeodomain. *Journal of Bone and Mineral Research*, 18(8):1419–1429, 2003. ISSN 1523-4681.

International Consortium for Blood Pressure Genome-Wide Association Studies. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature*, 478(7367):103–109, 2011.

International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52–58, 2010.

H. Ishwaran and J. S. Rao. Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics*, 33(2):730–773, 04 2005.

A. E. Jaffe, P. Murakami, H. Lee, J. T. Leek, M. D. Fallin, A. P. Feinberg, and R. A. Irizarry. Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *International Journal of Epidemiology*, 41(1):200–209, 2012.

L. Janson, R. F. Barber, and E. Candès. EigenPrism: inference for high dimensional signal-to-noise ratios. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2016. ISSN 1467-9868. doi: 10.1111/rssb.12203. URL `http://dx.doi.org/10.1111/rssb.12203`.

V. E. Johnson and D. Rossell. On the use of non-local prior densities in Bayesian hypothesis tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(2): 143–170, 2010.

V. E. Johnson and D. Rossell. Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association*, 107(498):649–660, 2012.

L. Jostins, S. Ripke, R. K. Weersma, R. H. Duerr, D. P. McGovern, K. Y. Hui, J. C. Lee, L. P. Schumm, Y. Sharma, C. A. Anderson, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*, 491(7422):119–124, 2012.

J. D. Kalbfleisch and D. A. Sprott. Application of likelihood methods to models involving large numbers of parameters. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 175–208, 1970.

H. M. Kang, J. H. Sul, S. K. Service, N. A. Zaitlen, S.-y. Kong, N. B. Freimer, C. Sabatti, E. Eskin, et al. Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, 42(4):348–354, 2010.

M. A. Kennedy, G. C. Barrera, K. Nakamura, Á. Baldán, P. Tarr, M. C. Fishbein, J. Frank, O. L. Francone, and P. A. Edwards. ABCG1 has a critical role in mediating cholesterol efflux to HDL and preventing cellular lipid accumulation. *Cell Metabolism*, 1(2):121–131, 2005.

G. Kichaev, W.-Y. Yang, S. Lindstrom, F. Hormozdiari, E. Eskin, A. L. Price, P. Kraft, and B. Pasaniuc. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genetics*, 10(10):e1004722, 2014.

A. Kivitz, G. Eisen, and W. W. Zhao. Randomized placebo-controlled trial comparing efficacy and safety of valdecoxib with naproxen in patients with osteoarthritis. *Journal of Family Practice*, 51(6):530–537, 2002.

A. Köttgen, E. Albrecht, A. Teumer, V. Vitart, J. Krumsiek, C. Hundertmark, G. Pistis, D. Ruggiero, C. M. O'Seaghdha, T. Haller, et al. Genome-wide association analyses identify 18 new loci associated with serum urate concentrations. *Nature Genetics*, 45(2):145–154, 2013.

J. Kurkó, T. Besenyei, J. Laki, T. T. Glant, K. Mikecz, and Z. Szekanecz. Genetics of rheumatoid arthritis–a comprehensive review. *Clinical Reviews in Allergy & Immunology*, 45(2): 170–179, 2013.

I.-Y. Kwak and W. Pan. Adaptive gene-and pathway-trait association testing with GWAS summary statistics. *Bioinformatics*, 32(8):1178–1184, 2016.

J.-C. Lambert, C. A. Ibrahim-Verbaas, D. Harold, A. C. Naj, R. Sims, C. Bellenguez, G. Jun, A. L. DeStefano, J. C. Bis, G. W. Beecham, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature Genetics*, 45(12):1452–1458, 2013.

D. Lamparter, D. Marbach, R. Rueedi, Z. Kutalik, and S. Bergmann. Fast and rigorous computation of gene and pathway scores from SNP-based summary statistics. *PLoS Computational Biology*, 12(1):e1004714, 2016.

H. Lango Allen, K. Estrada, G. Lettre, S. I. Berndt, M. N. Weedon, F. Rivadeneira, C. J. Willer, A. U. Jackson, S. Vedantam, S. Raychaudhuri, et al. Hundreds of variants clustered

in genomic loci and biological pathways affect human height. *Nature*, 467(7317):832–838, 2010.

D. Lee, T. B. Bigdeli, B. P. Riley, A. H. Fanous, and S.-A. Bacanu. DIST: direct imputation of summary statistics for unmeasured snps. *Bioinformatics*, 29(22):2925–2927, 2013.

D. Lee, V. S. Williamson, T. B. Bigdeli, B. P. Riley, A. H. Fanous, V. I. Vladimirov, and S.-A. Bacanu. JEPEG: a summary statistics based tool for gene-level joint testing of functional variants. *Bioinformatics*, 31(8):1176–1182, 2015.

J. T. Leek and J. D. Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, 3(9):e161, 2007.

A. Lex, N. Gehlenborg, H. Strobelt, R. Vuillemot, and H. Pfister. UpSet: Visualization of intersecting sets. *IEEE Transactions on Visualization and Computer Graphics (InfoVis '14)*, 20(12):1983–1992, 2014. ISSN 1077-2626.

D. Li, R. Sakuma, N. A. Vakili, R. Mo, V. Puviindran, S. Deimling, X. Zhang, S. Hopyan, and C.-c. Hui. Formation of proximal and anterior limb skeleton requires early function of *irx3* and *irx5* and is negatively regulated by Shh signaling. *Developmental Cell*, 29(2): 233–240, 2014.

F. Li and N. R. Zhang. Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. *Journal of the American Statistical Association*, 105 (491):1202–1214, 2010.

Y. Li and M. Kellis. Joint bayesian inference of risk variants and tissue-specific epigenomic enrichments across multiple complex human diseases. *Nucleic Acids Research*, 44(18): e144, 2016.

Y. Li, C. Willer, S. Sanna, and G. Abecasis. Genotype imputation. *Annual Review of Genomics and Human Genetics*, 10:387–406, 2009.

Y. R. Li, J. van Setten, S. S. Verma, Y. Lu, M. V. Holmes, H. Gao, M. Lek, N. Nair, H. Chandrupatla, B. Chang, et al. Concept and design of a genome-wide association genotyping array tailored for transplantation-specific studies. *Genome Medicine*, 7(1):90, 2015.

Z. Li and M. J. Sillanpää. Estimation of quantitative trait locus effects with epistasis by variational Bayes algorithms. *Genetics*, 190(1):231–249, 2012.

F. Liang and W. H. Wong. Evolutionary Monte Carlo: Applications to c_p model sampling and change point problem. *Statistica Sinica*, pages 317–342, 2000.

F. Liang, R. Paulo, G. Molina, M. A. Clyde, and J. O. Berger. Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103(481), 2008.

W.-J. Liao, K.-C. Tsao, and R.-B. Yang. Electrostatics and N-glycan-mediated membrane tethering of *scube1* is critical for promoting bone morphogenetic protein signalling. *Biochemical Journal*, 473(5):661–672, 2016.

D. Lin. An efficient Monte Carlo approach to assessing statistical significance in genomic studies. *Bioinformatics*, 21(6):781–787, 2005.

C. Lippert, J. Listgarten, Y. Liu, C. M. Kadie, R. I. Davidson, and D. Heckerman. Fast linear mixed models for genome-wide association studies. *Nature Methods*, 8(10):833–835, 2011.

C.-C. Liu, T. Kanekiyo, H. Xu, and G. Bu. Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy. *Nature Reviews Neurology*, 9(2):106–118, 2013.

J. Z. Liu, A. F. Mcrae, D. R. Nyholt, S. E. Medland, N. R. Wray, K. M. Brown, N. K. Hayward, G. W. Montgomery, P. M. Visscher, N. G. Martin, et al. A versatile gene-based test for genome-wide association studies. *The American Journal of Human Genetics*, 87(1):139–145, 2010.

J. Z. Liu, S. van Sommeren, H. Huang, S. C. Ng, R. Alberts, A. Takahashi, S. Ripke, J. C. Lee, L. Jostins, T. Shah, et al. Association analyses identify 38 susceptibility loci for

inflammatory bowel disease and highlight shared genetic risk across populations. *Nature Genetics*, 47(9):979–986, 2015.

K. Liu and X.-L. Meng. There is individualized treatment. Why not individualized inference? *Annual Review of Statistics and Its Application*, 3:79–111, 2016.

A. E. Locke, B. Kahali, S. I. Berndt, A. E. Justice, T. H. Pers, F. R. Day, C. Powell, S. Vedantam, M. L. Buchkovich, J. Yang, et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518(7538):197–206, 2015.

B. A. Logsdon, G. E. Hoffman, and J. G. Mezey. A variational Bayes algorithm for fast and accurate multiple locus genome-wide association analysis. *BMC Bioinformatics*, 11(1):58, 2010.

B. A. Logsdon, J. Y. Dai, P. L. Auer, J. M. Johnsen, S. K. Ganesh, N. L. Smith, J. G. Wilson, R. P. Tracy, L. A. Lange, S. Jiao, et al. A variational Bayes discrete mixture test for rare variant association. *Genetic Epidemiology*, 38(1):21–30, 2014.

P.-R. Loh, G. Tucker, B. K. Bulik-Sullivan, B. J. Vilhjalmsson, H. K. Finucane, D. I. Chasman, P. M. Ridker, B. M. Neale, B. Berger, N. Patterson, et al. Efficient Bayesian mixed model analysis increases association power in large cohorts. *Nature Genetics*, 47(3):284–290, 2015.

F. Macian. NFAT proteins: key regulators of T-cell development and function. *Nature Reviews Immunology*, 5(6):472–484, 2005.

E. Mackie, Y. Ahmed, L. Tatarczuch, K.-S. Chen, and M. Mirams. Endochondral ossification: how cartilage is converted into bone in the developing skeleton. *The International Journal of Biochemistry & Cell Biology*, 40(1):46–62, 2008.

T. S. H. Mak, R. M. Porsch, S. W. Choi, X. Zhou, and P. C. Sham. Polygenic scores via

penalized regression on summary statistics. *Genetic Epidemiology*, 2017. ISSN 1098-2272. doi: 10.1002/gepi.22050.

H. Mallick and N. Yi. Bayesian methods for high dimensional linear models. *Journal of Biometrics & Biostatistics*, 1:005, 2013.

A. K. Manning, M.-F. Hivert, R. A. Scott, J. L. Grimsby, N. Bouatia-Naji, H. Chen, D. Rybin, C.-T. Liu, L. F. Bielak, I. Prokopenko, et al. A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nature Genetics*, 44(6):659–669, 2012.

T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, 2009.

J. Marchini and B. Howie. Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, 11(7):499–511, 2010.

J. Marchini, B. Howie, S. Myers, G. McVean, and P. Donnelly. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics*, 39(7): 906–913, 2007.

M. I. McCarthy, G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little, J. P. Ioannidis, and J. N. Hirschhorn. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, 9(5):356–369, 2008.

P. McCullagh and J. A. Nelder. *Generalized linear models (Second edition)*. London: Chapman & Hall, 1989.

F. M. McQueen, A. Chhana, and N. Dalbeth. Mechanisms of joint damage in gout: evidence from cellular and imaging studies. *Nature Reviews Rheumatology*, 8(3):173–181, 2012.

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

T. P. Minka. Expectation propagation for approximate Bayesian inference. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pages 362–369. Morgan Kaufmann Publishers Inc., 2001.

T. J. Mitchell and J. J. Beauchamp. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032, 1988.

O. A. Montesinos-López, A. Montesinos-López, J. Crossa, J. C. Montesinos-López, F. J. Luna-Vázquez, J. Salinas, J. Herrera-Morales, and R. Buenrostro-Mariscal. A variational Bayes genomic-enabled prediction model with genotype × environment interaction. *G3: Genes, Genomes, Genetics*, 2017. doi: 10.1534/g3.117.041202.

A. P. Morris, B. F. Voight, T. M. Teslovich, T. Ferreira, A. V. Segre, V. Steinthorsdottir, R. J. Strawbridge, H. Khan, H. Grallert, A. Mahajan, et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature Genetics*, 44(9):981–990, 2012.

G. Moser, S. H. Lee, B. J. Hayes, M. E. Goddard, N. R. Wray, and P. M. Visscher. Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. *PLoS Genetics*, 11(4):e1004969, 2015.

Nature Genetics. Asking for more. *Nature Genetics*, 44(7):733, 2012.

R. M. Neal. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2:113–162, 2011.

M. F. Neurath. Cytokines in inflammatory bowel disease. *Nature Reviews Immunology*, 14 (5):329–342, 2014.

P. J. Newcombe, D. V. Conti, and S. Richardson. JAM: a scalable Bayesian framework for joint analysis of marginal SNP effects. *Genetic Epidemiology*, 40(3):188–201, 2016.

A. C. Nica and E. T. Dermitzakis. Expression quantitative trait loci: present and future. *Philosophical Transactions of the Royal Society B*, 368(1620):20120362, 2013.

A. C. Nica, S. B. Montgomery, A. S. Dimas, B. E. Stranger, C. Beazley, I. Barroso, and E. T. Dermitzakis. Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genetics*, 6(4):e1000895, 2010.

S. J. Nicholls, E. M. Tuzcu, I. Sipahi, A. W. Grasso, P. Schoenhagen, T. Hu, K. Wolski, T. Crowe, M. Y. Desai, S. L. Hazen, et al. Statins, high-density lipoprotein cholesterol, and regression of coronary atherosclerosis. *Journal of the American Medical Association*, 297(5):499–508, 2007.

D. L. Nicolae, E. Gamazon, W. Zhang, S. Duan, M. E. Dolan, and N. J. Cox. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genetics*, 6(4):e1000888, 2010.

M. Nikpay, A. Goel, H.-H. Won, L. M. Hall, C. Willenborg, S. Kanoni, D. Saleheen, T. Kyriakou, C. P. Nelson, J. C. Hopewell, et al. A comprehensive 1000 genomes–based genome-wide association meta-analysis of coronary artery disease. *Nature Genetics*, 47(10):1121–1130, 2015.

A. Obri, M. P. Makinistoglu, H. Zhang, and G. Karsenty. HDAC4 integrates PTH and sympathetic signaling in osteoblasts. *The Journal of Cell Biology*, 205(6):771–780, 2014.

R. B. O'Hara and M. J. Sillanpää. A review of Bayesian variable selection methods: what, how and which. *Bayesian Analysis*, 4(1):85–117, 2009.

Y. Okada, D. Wu, G. Trynka, T. Raj, C. Terao, K. Ikari, Y. Kochi, K. Ohmura, A. Suzuki,

S. Yoshida, et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*, 506(7488):376–381, 2014.

A. Okbay, B. Baselmans, J. De Neve, P. Turley, M. Nivard, M. Fontana, S. Meddens, R. Linnér, C. Rietveld, J. Derringer, et al. Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nature Genetics*, 48(6):624–633, 2016.

L. Palla and F. Dudbridge. A fast method that uses polygenic scores to estimate the variance explained by genome-wide marker panels and the proportion of variants affecting a trait. *The American Journal of Human Genetics*, 97(2):250–259, 2015.

P. Papastamoulis, J. Hensman, P. Glaus, and M. Rattray. Improved variational Bayes inference for transcript expression estimation. *Statistical Applications in Genetics and Molecular Biology*, 13(2):203–216, 2014.

J.-H. Park, S. Wacholder, M. H. Gail, U. Peters, K. B. Jacobs, S. J. Chanock, and N. Chatterjee. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nature Genetics*, 42(7):570–575, 2010.

T. Park and G. Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

B. Pasaniuc and A. L. Price. Dissecting the genetics of complex traits using summary association statistics. *Nature Reviews Genetics*, 18:117–127, 2017.

B. Pasaniuc, N. Zaitlen, H. Shi, G. Bhatia, A. Gusev, J. Pickrell, J. Hirschhorn, D. P. Strachan, N. Patterson, and A. L. Price. Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics*, 30(20):2906–2914, 2014.

N. Patterson, A. L. Price, and D. Reich. Population structure and eigenanalysis. *PLoS Genetics*, 2(12):e190, 2006.

E. Peise, D. Fabregat-Traver, and P. Bientinesi. High performance solutions for big-data GWAS. *Parallel Computing*, 42:75–87, 2015.

T. Peltola, P. Marttinen, A. Jula, V. Salomaa, M. Perola, et al. Bayesian variable selection in searching for additive and dominant effects in genome-wide data. *PLoS ONE*, 7(1): e29115, 2012.

T. H. Pers, J. M. Karjalainen, Y. Chan, H.-J. Westra, A. R. Wood, J. Yang, J. C. Lui, S. Vedantam, S. Gustafsson, T. Esko, et al. Biological interpretation of genome-wide association studies using predicted gene functions. *Nature Communications*, 6, 2015.

J. K. Pickrell. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *The American Journal of Human Genetics*, 94(4):559–573, 2014.

J. K. Pickrell, T. Berisa, J. Z. Liu, L. Ségurel, J. Y. Tung, and D. A. Hinds. Detection and interpretation of shared genetic influences on 42 human traits. *Nature Genetics*, 48(7): 709–717, 2016.

A. R. Pico, T. Kelder, M. P. Van Iersel, K. Hanspers, B. R. Conklin, et al. Wikipathways: pathway editing for the people. *PLoS Biology*, 6(7):e184, 2008.

M. A. Portelli, M. Siedlinski, C. E. Stewart, D. S. Postma, M. A. Nieuwenhuis, J. M. Vonk, P. Nurnberg, J. Altmuller, M. F. Moffatt, A. J. Wardlaw, et al. Genome-wide protein QTL mapping identifies human plasma kallikrein as a post-translational regulator of serum uPAR levels. *The FASEB Journal*, 28(2):923–934, 2014.

A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909, 2006.

A. L. Price, N. A. Zaitlen, D. Reich, and N. Patterson. New approaches to population strat-

ification in genome-wide association studies. *Nature Reviews Genetics*, 11(7):459–463, 2010.

A. L. Price, C. C. Spencer, and P. Donnelly. Progress and promise in understanding the genetic basis of common diseases. In *Proceedings of the Royal Society B*, volume 282, page 20151684. The Royal Society, 2015.

J. K. Pritchard and M. Przeworski. Linkage disequilibrium in humans: models and data. *The American Journal of Human Genetics*, 69(1):1–14, 2001.

D. J. Rader and J. J. Kastelein. Lomitapide and mipomersen: Two first-in-class drugs for reducing low-density lipoprotein cholesterol in patients with homozygous familial hypercholesterolemia. *Circulation*, 129(9):1022–1032, 2014.

A. Raj, M. Stephens, and J. K. Pritchard. fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics*, 197(2):573–589, 2014.

S. Richardson, G. C. Tseng, and W. Sun. Statistical methods in integrative genomics. *Annual Review of Statistics and Its Application*, 3:181–209, 2016.

C. A. Rietveld, S. E. Medland, J. Derringer, J. Yang, T. Esko, N. W. Martin, H.-J. Westra, K. Shakhbazov, A. Abdellaoui, A. Agrawal, et al. GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science*, 340(6139):1467–1471, 2013.

D. Risso, J. Ngai, T. P. Speed, and S. Dudoit. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nature Biotechnology*, 32(9):896–902, 2014.

M. D. Ritchie, E. R. Holzinger, R. Li, S. A. Pendergrass, and D. Kim. Methods of integrating data to uncover genotype-phenotype interactions. *Nature Reviews Genetics*, 16(2):85–97, 2015.

P. Romero, J. Wagg, M. L. Green, D. Kaiser, M. Krummenacker, and P. D. Karp. Computational prediction of human metabolic pathways from the complete human genome. *Genome Biology*, 6(1):R2, 2004.

H. Rue, S. Martino, and N. Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392, 2009.

C. Sabatti. Multivariate linear models for GWAS. In K.-A. Do, Z. S. Qin, and M. Vannucci, editors, *Advances in Statistical Bioinformatics*, pages 188–207. Cambridge University Press, 2013. ISBN 9781139226448.

C. Sassi, P. G. Ridge, M. A. Nalls, R. Gibbs, J. Ding, M. K. Lupton, C. Troakes, K. Lunnon, S. Al-Sarraj, K. S. Brown, et al. Influence of coding variability in APP-Aβ metabolism genes in sporadic Alzheimer's Disease. *PLoS ONE*, 11(6):e0150079, 2016.

E. E. Schadt, J. Lamb, X. Yang, J. Zhu, S. Edwards, D. GuhaThakurta, S. K. Sieberts, S. Monks, M. Reitman, C. Zhang, et al. An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genetics*, 37(7):710–717, 2005.

C. F. Schaefer, K. Anthony, S. Krupa, J. Buchoff, M. Day, T. Hannay, and K. H. Buetow. PID: the Pathway Interaction Database. *Nucleic Acids Research*, 37(suppl 1):D674–D679, 2009.

Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, 511(7510):421–427, 2014.

F. Schmich. *gespeR: Gene-Specific Phenotype EstimatoR*, 2015. URL `http://www.cbg.ethz.ch/software/gespeR`. R package version 1.8.0.

H. Schunkert, I. R. König, S. Kathiresan, M. P. Reilly, T. L. Assimes, H. Holm, M. Preuss,

A. F. Stewart, M. Barbalic, C. Gieger, et al. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nature Genetics*, 43(4):333–338, 2011.

A. L. Schwarzman, L. Gregori, M. P. Vitek, S. Lyubski, W. J. Strittmatter, J. J. Enghilde, R. Bhasin, J. Silverman, K. H. Weisgraber, and P. K. Coyle. Transthyretin sequesters amyloid beta protein and prevents amyloid formation. *Proceedings of the National Academy of Sciences*, 91(18):8368–8372, 1994.

T. Schweder and N. L. Hjort. Confidence and likelihood. *Scandinavian Journal of Statistics*, 29(2):309–332, 2002. ISSN 1467-9469.

S. Seaman and B. Müller-Myhsok. Rapid simulation of p values for product methods and multiple-testing adjustment in association studies. *The American Journal of Human Genetics*, 76(3):399–408, 2005.

A. V. Segrè, L. Groop, V. K. Mootha, M. J. Daly, D. Altshuler, D. Consortium, M. Investigators, et al. Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *PLoS Genetics*, 6(8):e1001058, 2010.

B. Servin and M. Stephens. Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genetics*, 3(7):e114, 2007.

P. C. Sham and S. M. Purcell. Statistical power and significance testing in large-scale genetic studies. *Nature Reviews Genetics*, 15(5):335–346, 2014.

H. Shi, G. Kichaev, and B. Pasaniuc. Contrasting the genetic architecture of 30 complex traits from summary association data. *The American Journal of Human Genetics*, 99(1): 139–153, 2016a.

J. Shi, J.-H. Park, J. Duan, S. T. Berndt, W. Moy, K. Yu, L. Song, W. Wheeler, X. Hua, D. Silverman, et al. Winner's curse correction and variable thresholding improve performance

of polygenic risk modeling based on genome-wide association study summary-level data. *PLoS Genetics*, 12(12):e1006493, 2016b.

S.-Y. Shin, E. B. Fauman, A.-K. Petersen, J. Krumsiek, R. Santos, J. Huang, M. Arnold, I. Erte, V. Forgetta, T.-P. Yang, et al. An atlas of genetic influences on human blood metabolites. *Nature Genetics*, 46(6):543–550, 2014.

D. Shungin, T. W. Winkler, D. C. Croteau-Chonka, T. Ferreira, A. E. Locke, R. Mägi, R. J. Strawbridge, T. H. Pers, K. Fischer, A. E. Justice, et al. New genetic loci link adipose and insulin biology to body fat distribution. *Nature*, 518(7538):187–196, 2015.

J. Silence, F. Lupu, D. Collen, and H. Lijnen. Persistence of atherosclerotic plaque but reduced aneurysm formation in mice with stromelysin-1 (MMP-3) gene inactivation. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 21(9):1440–1445, 2001.

D. Sitara and A. O. Aliprantis. Transcriptional regulation of bone and joint remodeling by NFAT. *Immunological Reviews*, 233(1):286–300, 2010.

M. Slatkin. Linkage disequilibrium – understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, 9(6):477–485, 2008.

K. Slowikowski, X. Hu, and S. Raychaudhuri. SNPsea: an algorithm to identify cell types, tissues and pathways affected by risk loci. *Bioinformatics*, 30(17):2496–2497, 2014.

D. Smedley, S. Haider, S. Durinck, L. Pandini, P. Provero, J. Allen, O. Arnaiz, M. H. Awedh, R. Baldock, G. Barbiera, et al. The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Research*, 43(W1):W589–W598, 2015.

S. Smemo, J. J. Tena, K.-H. Kim, E. R. Gamazon, N. J. Sakabe, C. Gómez-Marín, I. Aneas, F. L. Credidio, D. R. Sobreira, N. F. Wasserman, et al. Obesity-associated variants within *FTO* form long-range functional connections with *IRX3*. *Nature*, 507(7492):371–375, 2014.

H.-C. So and P. C. Sham. Improving polygenic risk prediction from summary statistics by an empirical Bayes approach. *Scientific Reports*, 7:41262, 2017.

D. Sprott. Inferential estimation, likelihood, and linear pivotals. *Canadian Journal of Statistics*, 18(1):1–10, 1990.

A. L. Stark, R. J. Hause Jr, L. K. Gorsic, N. N. Antao, S. S. Wong, S. H. Chung, D. F. Gill, H. K. Im, J. L. Myers, K. P. White, et al. Protein quantitative trait loci identify novel candidates modulating cellular response to chemotherapy. *PLoS Genetics*, 10(4):e1004192, 2014.

O. Stegle, L. Parts, R. Durbin, and J. Winn. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Computational Biology*, 6(5):e1000770, 2010.

M. W. Steinberg, O. Turovskaya, R. B. Shaikh, G. Kim, D. F. McCole, K. Pfeffer, K. M. Murphy, C. F. Ware, and M. Kronenberg. A crucial role for HVEM and BTLA in preventing intestinal inflammation. *Journal of Experimental Medicine*, 205(6):1463–1476, 2008.

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

M. Stephens. False discovery rates: a new deal. *Biostatistics*, 18(2):275–294, 2017.

B. E. Stranger, E. A. Stahl, and T. Raj. Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics*, 187(2):367–383, 2011.

A. Strasser, P. J. Jost, and S. Nagata. The many roles of FAS receptor signaling in the immune system. *Immunity*, 30(2):180–192, 2009.

A. I. Su, T. Wiltshire, S. Batalov, H. Lapp, K. A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, et al. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America*, 101(16):6062–6067, 2004.

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

A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Medicine*, 12(3): e1001779, 2015.

K. Suhre and C. Gieger. Genetic variation in metabolic phenotypes: study designs and applications. *Nature Reviews Genetics*, 13(11):759–769, 2012.

W. Sun and Y. Hu. eQTL mapping using RNA-seq data. *Statistics in Biosciences*, 5(1): 198–219, 2013.

Y. Sun, N. R. Zhang, A. B. Owen, et al. Multiple hypothesis testing adjusted for latent variables, with an application to the AGEMAP gene expression data. *The Annals of Applied Statistics*, 6(4):1664–1688, 2012.

D. Susan-Resiga, R. Essalmani, J. Hamelin, M.-C. Asselin, S. Benjannet, A. Chamberland, R. Day, D. Szumska, D. Constam, S. Bhattacharya, et al. Furin is the major processing enzyme of the cardiac-specific growth factor bone morphogenetic protein 10. *Journal of Biological Chemistry*, 286(26):22785–22794, 2011.

T. J. Sweeting. On a converse to Scheffe's Theorem. *The Annals of Statistics*, 14(3):1252–1256, 09 1986.

L. Széles, D. Töröcsik, and L. Nagy. PPARγ in immunity and inflammation: cell types and diseases. *Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids*, 1771 (8):1014–1030, 2007.

S. Y. Tang, R.-P. Herber, S. P. Ho, and T. Alliston. Matrix metalloproteinase–13 is required for osteocytic perilacunar remodeling and maintains bone fracture resistance. *Journal of Bone and Mineral Research*, 27(9):1936–1950, 2012.

J. Taylor and R. J. Tibshirani. Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, 112(25):7629–7634, 2015.

M. W. Teng, E. P. Bowman, J. J. McElwee, M. J. Smyth, J.-L. Casanova, A. M. Cooper, and D. J. Cua. IL-12 and IL-23 cytokines: from discovery to targeted therapies for immune-mediated inflammatory diseases. *Nature Medicine*, 21(7):719–729, 2015.

T. M. Teslovich, K. Musunuru, A. V. Smith, A. C. Edmondson, I. M. Stylianou, M. Koseki, J. P. Pirruccello, S. Ripatti, D. I. Chasman, C. J. Willer, et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, 466(7307):707–713, 2010.

The GTEx Consortium. The genotype-tissue expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235):648–660, 2015.

W. K. Thompson, Y. Wang, A. J. Schork, A. Witoelar, V. Zuber, S. Xu, T. Werge, D. Holland, O. A. Andreassen, A. M. Dale, et al. An empirical Bayes mixture model for effect size distributions in genome-wide association studies. *PLoS Genetics*, 11(12):e1005717, 2015.

R. Tibshirani. Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):273–282, 2011.

Tobacco and Genetics Consortium. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nature Genetics*, 42(5):441–447, 2010.

G. Tucker, A. L. Price, and B. Berger. Improving the power of GWAS and avoiding confounding from population stratification with PC-Select. *Genetics*, 197(3):1045–1049, 2014.

S. D. Turner. qqman: an R package for visualizing GWAS results using QQ and Manhattan plots. *bioRxiv*, page 005165, 2014.

P. van der Harst, W. Zhang, I. M. Leach, A. Rendon, N. Verweij, J. Sehmi, D. S. Paul, U. Elling, H. Allayee, X. Li, et al. Seventy-five genetic loci influencing the human red blood cell. *Nature*, 492(7429):369–375, 2012.

A. W. van der Vaart. *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998. ISBN 0-521-49603-9.

W. van Rheenen, A. Shatunov, A. M. Dekker, R. L. McLaughlin, F. P. Diekstra, S. L. Pulit, R. A. van der Spek, U. Võsa, S. de Jong, M. R. Robinson, et al. Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. *Nature Genetics*, 48(9):1043–1048, 2016.

R. Varadhan and C. Roland. Simple and globally convergent methods for accelerating the convergence of any EM algorithm. *Scandinavian Journal of Statistics*, 35(2):335–353, 2008.

M. Varjosalo and J. Taipale. Hedgehog: functions and mechanisms. *Genes & Development*, 22(18):2454–2472, 2008.

R. B. Vega, K. Matsuda, J. Oh, A. C. Barbosa, X. Yang, E. Meadows, J. McAnally, C. Pomajzl, J. M. Shelton, J. A. Richardson, et al. Histone deacetylase 4 controls chondrocyte hypertrophy during skeletogenesis. *Cell*, 119(4):555–566, 2004.

L. Velayudhan, R. Killick, A. Hye, A. Kinsey, A. Güntert, S. Lynham, M. Ward, R. Leung, A. Lourdusamy, A. W. To, et al. Plasma transthyretin as a candidate marker for Alzheimer's disease. *Journal of Alzheimer's Disease*, 28(2):369–375, 2012.

B. Vilhjalmsson, J. Yang, H. K. Finucane, A. Gusev, S. Lindstrom, S. Ripke, G. Genovese, P.-R. Loh, G. Bhatia, R. Do, et al. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *The American Journal of Human Genetics*, 97(4):576–592, 2015.

P. M. Visscher, W. G. Hill, and N. R. Wray. Heritability in the genomics era–concepts and misconceptions. *Nature Reviews Genetics*, 9(4):255–266, 2008.

P. M. Visscher, M. A. Brown, M. I. McCarthy, and J. Yang. Five years of GWAS discovery. *The American Journal of Human Genetics*, 90(1):7–24, 2012.

B. F. Voight, H. M. Kang, J. Ding, C. D. Palmer, C. Sidore, P. S. Chines, N. P. Burtt, C. Fuchsberger, Y. Li, J. Erdmann, et al. The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genetics*, 8(8): e1002793, 2012.

J. Wakefield. Bayes factors for genome-wide association studies: comparison with p-values. *Genetic Epidemiology*, 33:79–86, 2009.

K. Wang, M. Li, and H. Hakonarson. Analysing biological pathways in genome-wide association studies. *Nature Reviews Genetics*, 11(12):843–854, 2010.

W. Webber, A. Moffat, and J. Zobel. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)*, 28(4):20, 2010.

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.

D. Welter, J. MacArthur, J. Morales, T. Burdett, P. Hall, H. Junkins, A. Klemm, P. Flicek, T. Manolio, L. Hindorff, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, 42(D1):D1001–D1006, 2014.

X. Wen and M. Stephens. Using linear predictors to impute allele frequencies from summary or pooled genotype data. *The Annals of Applied Statistics*, 4(3):1158–1182, 2010.

X. Wen, Y. Lee, F. Luca, and R. Pique-Regi. Efficient integrative multi-SNP association analysis via deterministic approximation of posteriors. *The American Journal of Human Genetics*, 98(6):1114–1129, 2016.

X. Wen, R. Pique-Regi, and F. Luca. Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. *PLoS Genetics*, 13(3):e1006646, 2017.

M. A. Wilson, E. S. Iversen, M. A. Clyde, S. C. Schmidler, and J. M. Schildkraut. Bayesian model search and multilevel inference for SNP association studies. *The Annals of Applied Statistics*, 4(3):1342, 2010.

T. W. Winkler, F. R. Day, D. C. Croteau-Chonka, A. R. Wood, A. E. Locke, R. Mägi, T. Ferreira, T. Fall, M. Graff, A. E. Justice, et al. Quality control and conduct of genome-wide association meta-analyses. *Nature Protocols*, 9(5):1192–1212, 2014.

A. R. Wood, T. Esko, J. Yang, S. Vedantam, T. H. Pers, S. Gustafsson, A. Y. Chu, K. Estrada, J. Luan, Z. Kutalik, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics*, 46(11):1173–1186, 2014.

C. Wrzodek, F. Büchel, M. Ruff, A. Dräger, and A. Zell. Precise generation of systems biology models from KEGG pathways. *BMC Systems Biology*, 7(1):15, 2013.

Q. Xiang, R. Bi, M. Xu, D.-F. Zhang, L. Tan, C. Zhang, Y. Fang, and Y.-G. Yao. Rare genetic variants of the transthyretin gene are associated with Alzheimer's disease in Han Chinese. *Molecular Neurobiology*, pages 1–9, 2016.

Z. Xu, Q. Duan, S. Yan, W. Chen, M. Li, E. Lange, and Y. Li. DISSCO: direct imputation of summary statistics allowing covariates. *Bioinformatics*, page btv168, 2015.

C. Yang, X. Wan, J. Liu, and M. Ng. Introduction to statistical methods for integrative data

analysis in genome-wide association studies. In *Big Data Analytics in Genomics*, pages 3–23. Springer, 2016.

J. Yang, B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, et al. Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, 42(7):565–569, 2010.

J. Yang, T. A. Manolio, L. R. Pasquale, E. Boerwinkle, N. Caporaso, J. M. Cunningham, M. de Andrade, B. Feenstra, E. Feingold, M. G. Hayes, et al. Genome partitioning of genetic variation for complex traits using common SNPs. *Nature Genetics*, 43(6):519–525, 2011.

J. Yang, T. Ferreira, A. P. Morris, S. E. Medland, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. N. Weedon, R. J. Loos, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature Genetics*, 44(4):369–375, 2012.

J. Yang, N. A. Zaitlen, M. E. Goddard, P. M. Visscher, and A. L. Price. Advantages and pitfalls in the application of mixed-model association methods. *Nature Genetics*, 46(2): 100–106, 2014.

E. Zeggini and J. P. Ioannidis. Meta-analysis in genome-wide association studies. *Pharmacogenomics*, 10(2):191–201, 2009.

A. Zellner. On assessing prior distributions and bayesian regression analysis with g prior distributions. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*. Elsevier, New York, 1986.

H. Zhang, W. Wheeler, P. L. Hyland, Y. Yang, J. Shi, N. Chatterjee, and K. Yu. 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(6): e1006122, 06 2016.

X. Zhou and M. Stephens. Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, 44(7):821–824, 2012.

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

X. Zhu and M. Stephens. Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. *The Annals of Applied Statistics*, To appear, `https://doi.org/10.1101/042457`, 2017a.

X. Zhu and M. Stephens. A large-scale genome-wide enrichment analysis identifies new trait-associated genes, pathways and tissues across 31 human phenotypes. Submitted, `https://doi.org/10.1101/160770`, 2017b.

X. Zhu, T. Feng, B. O. Tayo, J. Liang, J. H. Young, N. Franceschini, J. A. Smith, L. R. Yanek, Y. V. Sun, T. L. Edwards, et al. Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension. *The American Journal of Human Genetics*, 96(1):21–36, 2015.

Z. Zhu, F. Zhang, H. Hu, A. Bakshi, M. R. Robinson, J. E. Powell, G. W. Montgomery, M. E. Goddard, N. R. Wray, P. M. Visscher, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature Genetics*, 48(5):481–487, 2016.