Manuscript: March 4, 2016

BAYESIAN LARGE-SCALE MULTIPLE REGRESSION WITH SUMMARY STATISTICS FROM GENOME-WIDE ASSOCIATION STUDIES[*]

By Xiang Zhu and Matthew Stephens

University of Chicago

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 data into the priors they can increase power and yield new biological insights. However, these methods require access to individual genotypes and phenotypes, which are often not easily available. 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, sampling posteriors by Markov chain Monte Carlo. In a wide range of simulations RSS performs similarly to analyses using the individual data, both for estimating heritability and detecting associations. We apply RSS to a GWAS of human height that contains 253,288 individuals typed at 1.06 million SNPs, for which analyses of individual-level data are practically impossible. Estimates of heritability (52%) are consistent with, but more precise, than previous results using subsets of these data. We also identify many previously-unreported loci that show evidence for association with height in our analyses. Software implementing RSS is available at `https://github.com/stephenslab/rss`.

[*] This work is supported by a grant from the Gordon and Betty Moore Foundation (Grant GBMF # 4559) and NIH grant HG02585 to M.S.

Keywords and phrases: Summary statistics, Bayesian regression, Genome wide, Association study, Multiple-SNP analysis, Variable selection, Heritability, Explained variation, Markov chain Monte Carlo

1. Introduction.

Consider the multiple linear regression model:

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

where $\mathbf{y}$ is an $n \times 1$ (centered) vector, X is an $n \times p$ (column-centered) matrix, $\boldsymbol{\beta}$ is the $p \times 1$ vector of multiple regression coefficients, and $\boldsymbol{\epsilon}$ is the error term. Assuming the "individual-level" data $\{X, \mathbf{y}\}$ are available, many methods exist to infer $\boldsymbol{\beta}$. 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, $j \in \{1, \ldots, p\}$. We also assume that information on the correlation structure among $\{X_j\}$ is available. With this in hand, we address the question: *how do we infer* $\boldsymbol{\beta}$ *using* $\{\hat{\beta}_j, \hat{\sigma}_j^2\}$*?* Specifically, we derive a likelihood for $\boldsymbol{\beta}$ given $\{\hat{\beta}_j, \hat{\sigma}_j^2\}$, and combine it with suitable priors to perform Bayesian inference for $\boldsymbol{\beta}$.

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 (Donnelly, 2008; McCarthy et al., 2008). 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 $\boldsymbol{\beta}$ 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" analyses on a very large total sample size. Furthermore these summary data are often made freely available on the Internet (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)] is also available through public databases such as 1000 Genomes Project Consortium (2010). 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, a single analyst (X.Z.) performed multiple-SNP analyses of GWAS data on adult height (Wood et al., 2014) involving 253,288 individuals typed at $\sim$ 1.06 million SNPs, using modest computational resources. Doing this for the individual-level data appears impractical.

Multiple-SNP analyses of GWAS compliment standard single-SNP analyses in several ways. Multiple-SNP analyses are particularly helpful in fine-mapping causal loci, allowing for multiple causal variants in a region [e.g. Servin and Stephens (2007); Yang et al. (2012)]. In addition, they can increase power to identify associations [e.g. Hoggart et al. (2008); Guan and Stephens (2011)]; 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, Carbonetto and Stephens (2013)]. 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 [Peise, Fabregat-Traver and Bientinesi (2015); Loh et al. (2015)]. Our methods help with both these issues, removing the need for individual-level data, and reducing computation by exploiting the banded structure of the estimated LD matrix (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 detecting multiple-SNP associations (Yang et al., 2012) and allelic heterogeneity (Ehret et al., 2012), single-SNP analysis with correlated phenotypes (Stephens, 2013) and heterogeneous subgroups (Wen and Stephens, 2014), gene-level testing of functional variants (Lee et al., 2015), joint analysis of functional genomic data and GWAS (Pickrell, 2014; Finucane et al., 2015), imputation of allele frequencies (Wen and Stephens, 2010) and single-SNP association statistics (Lee et al., 2013), fine mapping of causal variants (Hormozdiari et al., 2014; Chen et al., 2015), correction of inflated test statistics (Bulik-Sullivan et al., 2015), estimation of SNP heritability (Palla and Dudbridge, 2015), and prediction of polygenic risk scores (Vilhjalmsson et al., 2015). 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 $\boldsymbol{\beta}$, 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 $\boldsymbol{\beta}$, and specifically to estimate SNP heritability and detect associations.

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 $\boldsymbol{\beta}$ in (1.1), and perform Bayesian inference for $\boldsymbol{\beta}$. These approaches are closely related, but working directly with $\boldsymbol{\beta}$ 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, $\boldsymbol{\beta}$ maintains a consistent interpretation across studies. And working with $\boldsymbol{\beta}$ allows us to exploit previous work developing prior distributions for $\boldsymbol{\beta}$ for multiple-SNP analysis [e.g. Guan and Stephens (2011); Zhou, Carbonetto and Stephens (2013)]. We also give a more rigorous statement and derivation of the likelihood being used, which provides insight into what approximations are being made and when they may be valid. Finally, this previous work focused only on small genomic regions, whereas here we analyze whole chromosomes.

2. Likelihood based on summary data.

We first introduce some notation. For any vector $\mathbf{v}$, $\mathrm{diag}(\mathbf{v})$ denotes the diagonal matrix with diagonal elements $\mathbf{v}$. Let $\widehat{\boldsymbol{\beta}} := (\hat{\beta}_1, \ldots, \hat{\beta}_p)^\intercal$, $\widehat{\mathbf{s}} := (\hat{s}_1, \ldots, \hat{s}_p)^\intercal$ and $\widehat{S} := \mathrm{diag}(\widehat{\mathbf{s}})$, where $\hat{s}_j^2 := \hat{\sigma}_j^2 + n^{-1}\hat{\beta}_j^2$ 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 4.1), and has the advantage that it produces a sparse, banded matrix $\widehat{R}$, which speeds computation for large genomic regions. 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 $\boldsymbol{\beta}$ is

$$L_{\mathsf{rss}}(\boldsymbol{\beta};\widehat{\boldsymbol{\beta}},\widehat{S},\widehat{R}) := \mathcal{N}(\widehat{\boldsymbol{\beta}};\widehat{S}\widehat{R}\widehat{S}^{-1}\boldsymbol{\beta},\widehat{S}\widehat{R}\widehat{S}). \tag{2.1}$$

We refer to (2.1) as the "Regression with Summary Statistics" (RSS) likelihood. We provide a formal derivation in Section 2.4, 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. It is also important to note the assumption, implicit in the definition (1.2, 1.3), that all summary statistics were computed from the same samples. We illustrate the importance of this in Section 5.

2.1. *Intuition.* The RSS likelihood (2.1) is obtained by first deriving an approximation for $p(\widehat{\boldsymbol{\beta}}|S,R,\boldsymbol{\beta})$, where S is the diagonal matrix with the jth diagonal entry $s_j \approx \mathrm{Var}^{1/2}(\hat{\beta}_j)$, of which $\hat{S}$ is an estimate (see Section 2.4 for details). Specifically, we have

$$\widehat{\boldsymbol{\beta}}|S,R,\boldsymbol{\beta} \dot{\sim} \mathcal{N}(SRS^{-1}\boldsymbol{\beta},SRS), \tag{2.2}$$

from which L_{rss} is derived by plugging in the estimates $\{\widehat{S},\widehat{R}\}$ for $\{S,R\}$.

The distribution (2.2) captures three key features of 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):

$$\mathrm{E}(\hat{\beta}_j|S,R,\boldsymbol{\beta}) = s_j \cdot \textstyle\sum_{i=1}^{p} r_{ij} s_i^{-1} \beta_i, \tag{2.3}$$

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:

$$\mathrm{Var}(\hat{\beta}_j|S,R,\boldsymbol{\beta}) = s_j^2 \approx \hat{s}_j^2 = (nX_j^{\intercal}X_j)^{-1}\mathbf{y}^{\intercal}\mathbf{y}. \tag{2.4}$$

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:

$$\text{Corr}(\hat{\beta}_j, \hat{\beta}_k | S, R, \boldsymbol{\beta}) = r_{jk}, \tag{2.5}$$

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 (2.3). 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 (2.5). 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 (Section 5).

2.2. *Connection with the full-data likelihood.* When individual-level data are available the multiple regression model is

$$\mathbf{y} | X, \boldsymbol{\beta}, \tau \sim \mathcal{N}(X\boldsymbol{\beta}, \tau^{-1} I). \tag{2.6}$$

If we further assume the residual variance τ^{-1} is *known*, model (2.6) specifies a likelihood for $\boldsymbol{\beta}$, which we denote $L_{\mathsf{mvn}}(\boldsymbol{\beta}; \mathbf{y}, X, \tau)$. The following Proposition gives conditions under which this full-data likelihood and RSS likelihood are equivalent.

Proposition 2.1. Let $\widehat{R}^{\mathsf{sam}}$ denote the sample LD matrix computed from the genotypes X of the study cohort, $\widehat{R}^{\mathsf{sam}} := D^{-1} X^\intercal X D^{-1}$ where $D := \text{diag}(\mathbf{d})$, $\mathbf{d} := (||X_1||, \ldots, ||X_p||)^\intercal$, $||X_j|| := (X_j^\intercal X_j)^{1/2}$.

If $n > p$, $\tau^{-1} = n^{-1} \mathbf{y}^\intercal \mathbf{y}$ and $\widehat{R} = \widehat{R}^{\mathsf{sam}}$ then

$$\log L_{\mathsf{rss}}(\boldsymbol{\beta}; \widehat{\boldsymbol{\beta}}, \widehat{S}, \widehat{R}) - \log L_{\mathsf{mvn}}(\boldsymbol{\beta}; \mathbf{y}, X, \tau) = C \tag{2.7}$$

where C is some constant that does not depend on $\boldsymbol{\beta}$.

When fine mapping a genomic region, it often holds that $n > p$, and also that $\tau^{-1} \approx n^{-1} \mathbf{y}^\intercal \mathbf{y}$ since SNPs in a region typically explain a very small proportion of phenotypic variation. (In contrast, these two conditions do not hold in genome-wide context.) 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 4.1 (Figure 1); see also Chen et al. (2015).

2.3. *Connection with other summary-data methods.* To connect RSS with previous work, we first assume that $\hat{s}_j = \hat{\sigma}_j$ for all SNP j, which is justified by the fact that $\{\hat{s}_j\}$ and $\{\hat{\sigma}_j\}$ differ negligibly in most GWAS due to the large sample size and small trait-SNP correlations (Table 1).

If $\widehat{R}$ is an identity matrix, then $\widehat{\boldsymbol{\beta}}|\boldsymbol{\beta}, \widehat{S} \sim \mathcal{N}(\boldsymbol{\beta}, \widehat{S}^2)$, which is the implied likelihood based on the standard confidence interval (Efron, 1993). Using this likelihood, Wakefield (2009) defines an approximate Bayes factor to link Bayesian with Frequentist inference, and Stephens (2016) proposes a novel Empirical Bayes method for large-scale hypothesis testing.

If we let $\mathbf{z}$ denote the vector of single-SNP z-scores, $\mathbf{z} := \widehat{S}^{-1}\widehat{\boldsymbol{\beta}}$, and plug $\{\widehat{S}, \widehat{R}\}$ into (2.2), then

$$\mathbf{z}|\widehat{S}, \widehat{R}, \boldsymbol{\beta} \sim \mathcal{N}(\widehat{R}\widehat{S}^{-1}\boldsymbol{\beta}, \widehat{R}). \tag{2.8}$$

This is analogous to the likelihood proposed in Hormozdiari et al. (2014), $\mathbf{z} \sim \mathcal{N}(\widehat{R}^{\mathsf{sam}}\boldsymbol{\nu}, \widehat{R}^{\mathsf{sam}})$, where they refer to $\boldsymbol{\nu}$ as the "non-centrality parameter". If further $\boldsymbol{\beta} = \mathbf{0}$, then $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \widehat{R})$, a result that has been used for multiple testing adjustment [e.g. Seaman and Müller-Myhsok (2005); Lin (2005)] and gene-based association detection [e.g. Liu et al. (2010)].

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 (2.8) yields $\mathrm{E}(z_j^2|\widehat{S}, \widehat{R}) = 1 + \sum_{i=1}^p r_{ij}^2 s_i^{-2} \mathrm{E}(\beta_j^2|\widehat{S}, \widehat{R})$. This is the key idea behind the LD score regression (Bulik-Sullivan et al., 2015); see Supplement for detailed derivation.

2.4. *Derivation.* We treat the (unobserved) genotypes of each individual, $\boldsymbol{x}_i$ (the ith row of X), as being independent and identically distributed draws from some population. 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 $\mathrm{Var}(\boldsymbol{\epsilon}) = \tau^{-1}I_p$. 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) correlations between the phenotype and each SNP:

$$\mathbf{c} := \sigma_y^{-1}\mathrm{diag}^{-1}(\boldsymbol{\sigma}_x)\boldsymbol{\mu}_{xy} \tag{2.9}$$

where $\boldsymbol{\mu}_{xy} := \mathrm{E}(\boldsymbol{x}_i y_i) = \Sigma_x\boldsymbol{\beta}$ and $\sigma_y^2 := \mathrm{Var}(y_i) = \tau^{-1} + \boldsymbol{\beta}^{\intercal}\Sigma_x\boldsymbol{\beta}$.

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.

PROPOSITION 2.2. Let $S := n^{-\frac{1}{2}}\sigma_y \mathrm{diag}^{-1}(\boldsymbol{\sigma}_x)$ and $\Sigma := nS(R+\Delta(\mathbf{c}))S$.

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - SRS^{-1}\boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma), \tag{2.10}$$

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

Proposition 2.2 suggests that the sampling distribution of $\widehat{\boldsymbol{\beta}}$ is close to $\mathcal{N}(SRS^{-1}\boldsymbol{\beta}, n^{-1}\Sigma)$ for large n. Without additional assumptions, this may be the best probability statement that can be used to infer $\boldsymbol{\beta}$. However, it is difficult to work with this asymptotic distribution, mainly because of the complicated form of $\Delta(\mathbf{c})$. However, we can justify ignoring this term in a typical GWAS by the fact that $\{c_j\}$ are typically small in GWAS (Table 1), and the following proposition:

PROPOSITION 2.3. For each $\boldsymbol{\beta} \in \mathbb{R}^p$,

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

These propositions justify the approximate asymptotic distribution of $\widehat{\boldsymbol{\beta}}$ given in 2.2, provided n is large and $\{c_j^2\}$ close to zero, yielding

$$L_{\mathsf{rss}}(\boldsymbol{\beta}; \widehat{\boldsymbol{\beta}}, S, R) := \mathcal{N}(\widehat{\boldsymbol{\beta}}; SRS^{-1}\boldsymbol{\beta}, SRS). \tag{2.11}$$

Finally, the RSS likelihood (2.1) is obtained by replacing $\{S, R\}$ with their estimates $\{\widehat{S}, \widehat{R}\}$. Replacing $\widehat{S}$ with S is motivated by the Weak Law of Large Numbers, that is, $\sqrt{n}(\widehat{S} - S) \xrightarrow{p} \mathbf{0}$. However, there remains obvious potential for errors in the estimates $\{\widehat{S}, \widehat{R}\}$ to impact inference, and we assess this impact empirically through simulations (Section 4) and real data analyses (Section 6).

3. Bayesian inference based on summary data.

Using the RSS likelihood, we perform Bayesian inference for the multiple regression coefficients.

3.1. *Prior specification.*

If $\{S, R\}$ were known, then one could perform Bayesian inference by specifying a prior on $\boldsymbol{\beta}$:

$$\underbrace{p(\boldsymbol{\beta}|\widehat{\boldsymbol{\beta}}, S, R)}_{\text{Posterior}} \propto \underbrace{p(\widehat{\boldsymbol{\beta}}|S, R, \boldsymbol{\beta})}_{\text{Likelihood}} \cdot \underbrace{p(\boldsymbol{\beta}|S, R)}_{\text{Prior}}. \tag{3.1}$$

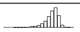
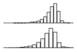
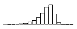
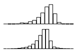
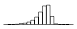
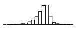
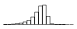
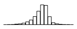
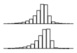
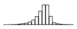
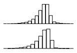
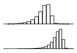
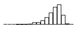
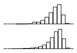
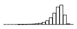
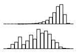
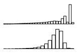
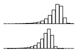
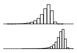
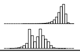
GWAS Phenotype	$\log_{10}(\hat{c}^2)$				$\log_{10}(n)$		
	Median	Mean	SD	Histogram	Median	Mean	SD
Height (GIANT, 2010)	−5.60	−5.77	0.93		5.26	5.26	0
Height (GIANT, 2014)	−5.41	−5.57	0.93		5.40	5.37	0.09
BMI (GIANT, 2015)	−5.65	−5.83	0.90		5.37	5.34	0.09
HDL (Global Lipids, 2010)	−5.25	−5.41	0.96		5.00	4.89	0.33
HDL (Global Lipids, 2013)	−5.25	−5.43	0.91		4.97	4.97	0.06
LDL (Global Lipids, 2010)	−5.23	−5.39	0.96		4.98	4.87	0.33
LDL (Global Lipids, 2013)	−5.24	−5.43	0.91		4.95	4.95	0.06
TC (Global Lipids, 2010)	−5.25	−5.41	0.96		5.00	4.89	0.33
TC (Global Lipids, 2013)	−5.26	−5.45	0.91		4.98	4.97	0.06
TG (Global Lipids, 2010)	−5.24	−5.39	0.96		4.98	4.87	0.33
TG (Global Lipids, 2013)	−5.24	−5.42	0.90		4.96	4.96	0.06
Cigarettes per day (TAG, 2010)	−5.19	−5.40	0.96		4.87	4.87	0
Smoking age of onset (TAG, 2010)	−5.18	−5.36	0.85		4.87	4.87	0
Ever smoked (TAG, 2010)	−5.17	−5.37	0.93		4.87	4.87	0
Former smoker (TAG, 2010)	−5.19	−5.39	0.94		4.87	4.87	0
Years of education (SSGAC, 2013)	−5.30	−5.36	0.66		5.10	5.10	0
College or not (SSGAC, 2013)	−1.23	−1.34	0.26		5.10	5.10	0
Schizophrenia (PGC, 2014)	−5.35	−5.55	0.95		5.18	5.18	0
Alzheimer (IGAP, 2013)	−5.04	−5.24	0.94		4.73	4.73	0
CAD (CARDIoGRAM, 2011)	−5.18	−5.39	0.97		4.91	4.88	0.08
T2D (DIAGRAM, 2012)	−5.49	−5.54	0.54		4.80	4.78	0.10

TABLE 1

Summary of sample squared correlation $\{\hat{c}_j^2\}$ and sample size n for several large-scale GWAS. The full names of phenotypes and references are provided in Supplementary Table 1. The medians, means, SDs and histograms are across SNPs. The sample correlation $\hat{c}_j$ between phenotype and SNP j is defined as $\hat{c}_j := (||\mathbf{y}|| \cdot ||X_j||)^{-1}(X_j^\intercal \mathbf{y})$. Note that $\hat{c}_j^2 = (n\hat{\sigma}_j^2 + \hat{\beta}_j^2)^{-1}\hat{\beta}_j^2 = (n\hat{s}_j^2)^{-1}\hat{\beta}_j^2$, and $\hat{c}_j \xrightarrow{p} c_j$.

To deal with unknown $\{S, R\}$ the RSS likelihood (2.1) approximates the likelihood in (3.1) by replacing $\{S, R\}$ with their estimates $\{\widehat{S}, \widehat{R}\}$. We take a similar approach to prior specification: we specify a prior $p(\boldsymbol{\beta}|S, R)$ and replace $\{S, R\}$ with $\{\widehat{S}, \widehat{R}\}$.

Our prior specification is based on the prior from Zhou, Carbonetto and Stephens (2013) which was designed for analysis of individual-level GWAS data. This prior assumes that $\boldsymbol{\beta}$ is independent of R *a priori*, with the prior on β_j being a mixture of two normal distributions

$$\beta_j \sim \pi\mathcal{N}(0, \sigma_B^2 + \sigma_P^2) + (1 - \pi)\mathcal{N}(0, \sigma_P^2). \tag{3.2}$$

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, Carbonetto and Stephens (2013) introduce two free parameters $h, \rho \in [0, 1]$ where h 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):

$$\log \pi \sim \mathcal{U}(\log(1/p), \log 1), \;\; h \sim \mathcal{U}(0,1), \;\; \rho \sim \mathcal{U}(0,1). \tag{3.3}$$

See Zhou, Carbonetto and Stephens (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 does not appear in our likelihood). Specifically we define

$$\sigma_B^2(S) := h\rho(\pi \textstyle\sum_{j=1}^p n^{-1} s_j^{-2})^{-1}, \;\; \sigma_P^2(S) := h(1-\rho)(\sum_{j=1}^p n^{-1} s_j^{-2})^{-1}, \tag{3.4}$$

where s_j is the jth diagonal entry of S. Because $ns_j^2 = \sigma_y^2 \sigma_{x,j}^{-2}$, definitions (3.4) 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 3.1. If $\boldsymbol{\beta}|S$ is distributed as (3.2), with (3.4), then

$$\mathrm{E}[V(X\boldsymbol{\beta})] = h \cdot \mathrm{E}[V(\mathbf{y})], \tag{3.5}$$

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

Because of its similarity with the prior from "Bayesian sparse linear mixed model" [BSLMM, Zhou, Carbonetto and Stephens (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.

3.2. *Posterior inference.* We use Markov chain Monte Carlo (MCMC) to sample from the posterior distribution of $\boldsymbol{\beta}$; see Supplement for details. Software implementing the methods is available at `https://github.com/stephenslab/rss`.

Compared with most existing summary-based methods, an important practical advantage of RSS is that multiple tasks can be performed simultaneously using the same posterior sample of $\boldsymbol{\beta}$. Here we focus on estimating PVE (SNP heritability) and detecting associations.

3.2.1. *Estimating PVE.* Given the full data $\{X, \mathbf{y}\}$ and the true value of $\{\boldsymbol{\beta}, \tau\}$ in model (1.1), Guan and Stephens (2011) define the PVE as

$$\text{PVE}(\boldsymbol{\beta}, \tau) := V(X\boldsymbol{\beta}) / (\tau^{-1} + V(X\boldsymbol{\beta})). \tag{3.6}$$

By this definition, PVE reflects the total proportion of sample phenotypic variation explained by available genotypes. Guan and Stephens (2011) then estimate PVE using the posterior sample of $\{\boldsymbol{\beta}, \tau\}$.

Because X is unknown here, we cannot compute PVE as defined above even if $\boldsymbol{\beta}$ 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:

$$\text{SPVE}(\boldsymbol{\beta}) := \sum_{i,j} \frac{\widehat{R}_{ij}\beta_i\beta_j}{\sqrt{(n\hat{\sigma}_i^2 + \hat{\beta}_i^2)(n\hat{\sigma}_j^2 + \hat{\beta}_j^2)}}. \tag{3.7}$$

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

$$\text{PVE} \approx \frac{V(X\boldsymbol{\beta})}{V(\mathbf{y})} = \sum_{i,j} \frac{X_i^\intercal X_j}{\mathbf{y}^\intercal \mathbf{y}}\beta_i\beta_j = \sum_{i,j} \frac{\widehat{R}_{ij}^{\mathsf{sam}}\beta_i\beta_j}{\sqrt{(n\hat{\sigma}_i^2 + \hat{\beta}_i^2)(n\hat{\sigma}_j^2 + \hat{\beta}_j^2)}}, \tag{3.8}$$

where $\widehat{R}^{\mathsf{sam}}$ is the (unknown) sample LD matrix of the study cohort, which we approximate in SPVE by $\widehat{R}$, and the last equality in (3.8) holds because of Equations (1.2) and (1.3). Simulations using both synthetic and real genotypes show that SPVE is a highly accurate approximation to PVE, given the true value of $\boldsymbol{\beta}$ (Supplementary Figure 1).

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. Unlike the original PVE (3.6), the definition of SPVE (3.7) is not bounded above by 1. Although we have not seen any estimates above 1 in our simulations or real data analyses, we expect this could occur if the posterior of $\boldsymbol{\beta}$ is poorly simulated and/or $\widehat{R}$ is severely misspecified.

3.2.2. *Detecting genome-wide associations.* 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 based on summary data

$$\text{SPIP}(j) = \Pr(\beta_j \neq 0 | \widehat{\boldsymbol{\beta}}, \widehat{S}, \widehat{R}). \tag{3.9}$$

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

4. Simulations.

We benchmark the RSS method through simulations, using real genotypes from Wellcome Trust Case Control Consortium (2007) (specifically, the $n = 1458$ individuals from the UK Blood Service Control Group) and simulated phenotypes. 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 are often larger than would be expected in a typical GWAS. This is, in some ways, not an ideal case for RSS, because the likelihood derivation assumes that effect sizes are small (Proposition 2.3). We use the simulations to i) investigate the effect of different choices for $\widehat{R}$; and ii) demonstrate that inferences from RSS agree well with both the simulation ground truth, and with results from methods based on the full data [BVSR and BSLMM implemented in the software package GEMMA (Zhou and Stephens, 2012)].

4.1. *Choice of LD matrix.*

The LD matrix $\widehat{R}$ 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)]. 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.

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*. We simulated 20 independent datasets, each with 10 causal SNPs and PVE=0.2. For each dataset, we ran RSS-BVSR with two strategies for computing $\widehat{R}$ from a reference panel (here, the 1480 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 (BVSR), and also with our RSS approach using the *cohort* LD matrix (RSS-C), which by Proposition 2.1 should produce results similar to the full data analysis. The results (Figure 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.

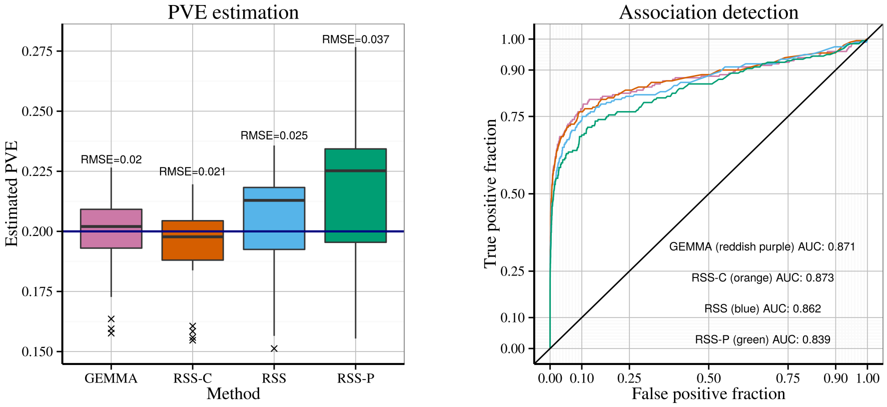


Fig 1: *Comparison of PVE estimation and association detection on three types of LD matrix* R*: cohort sample LD (RSS-C), shrinkage panel sample LD (RSS) and panel sample LD (RSS-P). Performance of estimating PVE is measured by the root of mean square error (RMSE), where a lower value indicates better performance. Performance of detecting associations is measured by the area under the curve (AUC), where a higher value indicates better performance.*

4.2. *Estimating PVE from summary data.* Here we use simulations to assess the performance of RSS for estimating PVE. Using the WTCCC genotypes from 12,758 SNPs on chromosome 16, we simulated phenotypes under two genetic architectures:

- Scenario 1.1 (sparse): randomly select 50 "causal" SNPs, with effects $\sim \mathcal{N}(0,1)$; effects of remaining SNPs are zero.
- Scenario 1.2 (polygenic): randomly select 50 "causal" SNPs, with effects $\sim \mathcal{N}(0,1)$; effects of remaining SNPs are $\sim \mathcal{N}(0,0.001^2)$.

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 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 2.3. (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.)

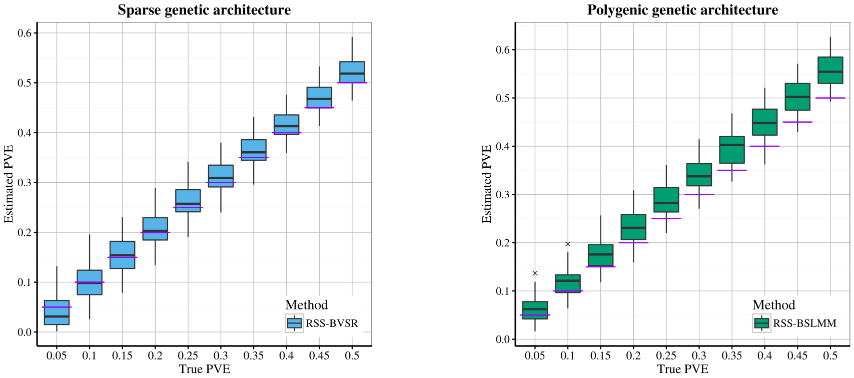


Fig 2: *Comparison of true PVE with estimated PVE (posterior median of SPVE) in Scenarios 1.1 (sparse) and 1.2 (polygenic). The purple lines indicate the true PVEs. Each box plot summarizes results from 50 replicates.*

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 effect 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 statistic methods (RSS-BVSR and RSS-BSLMM) with the corresponding full data methods (BVSR and 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, Carbonetto and Stephens, 2013); see also Supplement. Figure 3 summarizes the results. With modest true PVE (0.2), BVSR and RSS-BVSR perform better than other methods when the true model is very sparse (e.g. Scenario 2.1, $T = 10$), whereas BSLMM and RSS-BSLMM perform better when the true model is highly polygenic (e.g. Scenario 2.2, $T = 1000$).

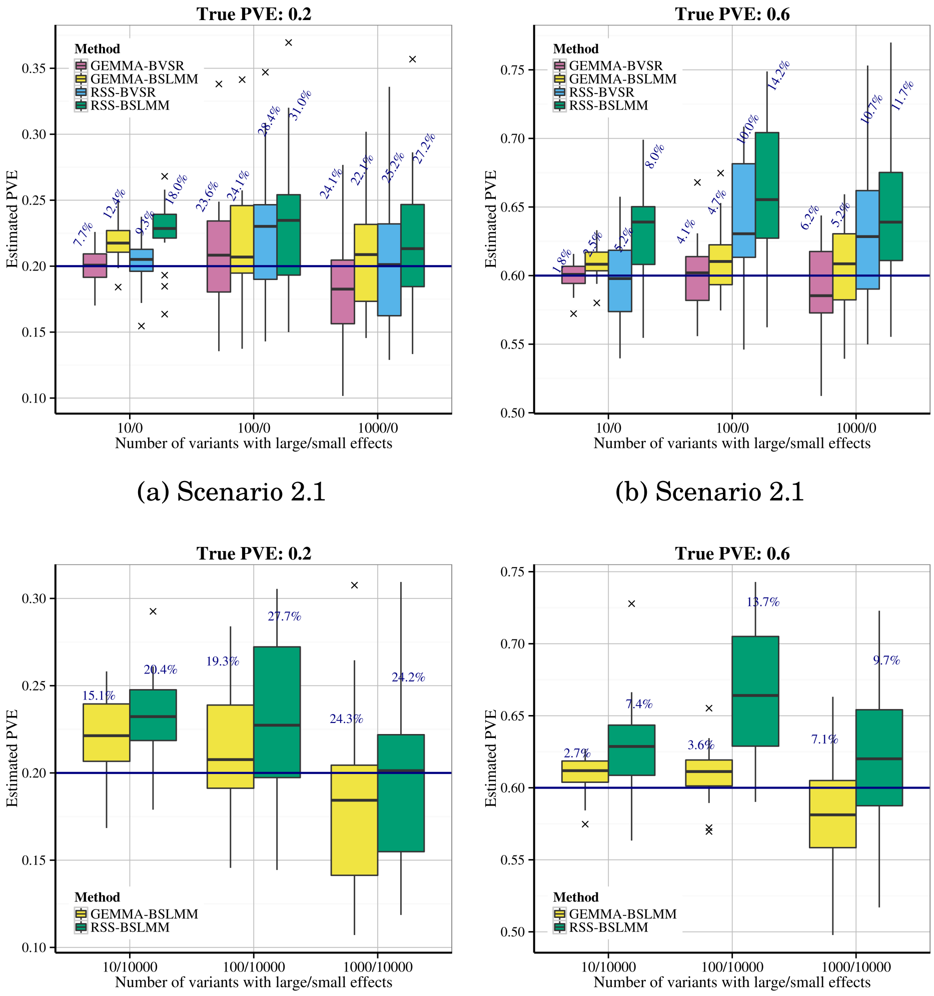
When the true PVE is large (0.6), the summary-based methods show an upward bias (Figure 3b and 3d), consistent with Figure 2. This bias is less severe when the true signals are more "diluted" (e.g. $T = 1000$), consistent with our speculation above that the bias is due to deviations from the "small effects" 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.

4.3. *Power to detect associations from summary data.* Previous studies using individual-level data have shown that multiple-SNP model can have higher power to detect SNP-phenotype associations than single-SNP analyses [e.g. Servin and Stephens (2007); Hoggart et al. (2008); Guan and Stephens (2011); 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 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 BVSR, we simulated data under Scenario 2.1 above. With Bayesian multiple-SNP 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 "Expected Number of included SNPs" (ENS), which summarizes the strength of association in that region. Results (Figure 4) 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). As above, the agreement of RSS-BVSR with BVSR is highest when PVE is diluted among many SNPs (e.g. $T = 1000$).

5. The importance of imputing to the same samples.

The derivation of the RSS likelihood assumes that the summary statistics are generated from the same individuals at each SNP. Specifically, the covariance matrix in (2.1) depends on this assumption. (In contrast, the mean in (2.1) holds even if different individuals are used at each SNP.) 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



(a) Scenario 2.1

(b) Scenario 2.1

(c) Scenario 2.2

(d) Scenario 2.2

Fig 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 in navy). The true PVEs are shown as navy horizontal line. Each box plot summarizes results from 20 replicates.*

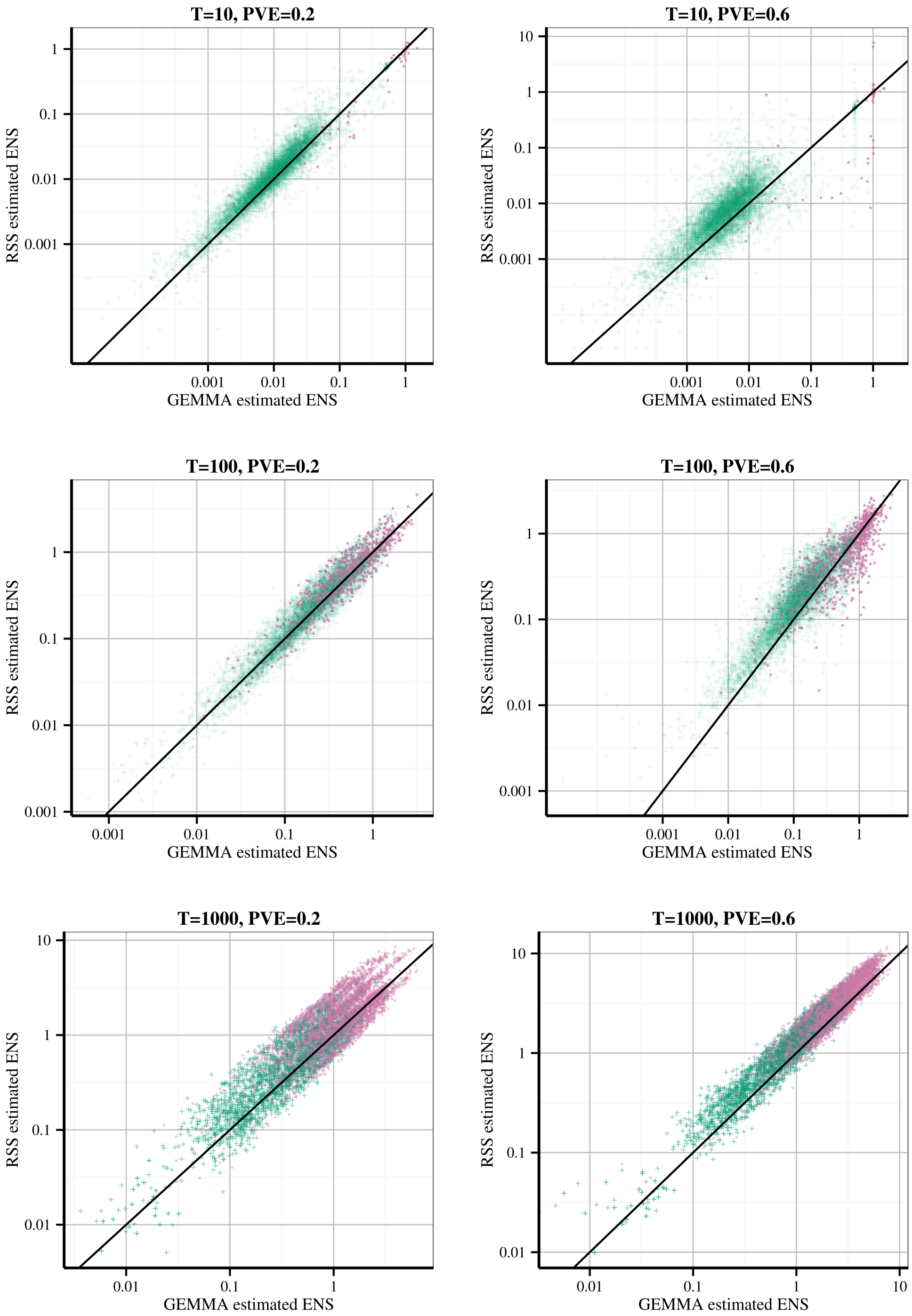


Fig 4: *Comparison of the 200-kb region posterior expected numbers of included SNPs (ENS) for GEMMA-BVSR (x-axis) and RSS-BVSR (y-axis), based on the simulation study of Scenario 2.1. Each point is a 200-kb genomic region, colored according to whether it contains at least one causal SNP (reddish purple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

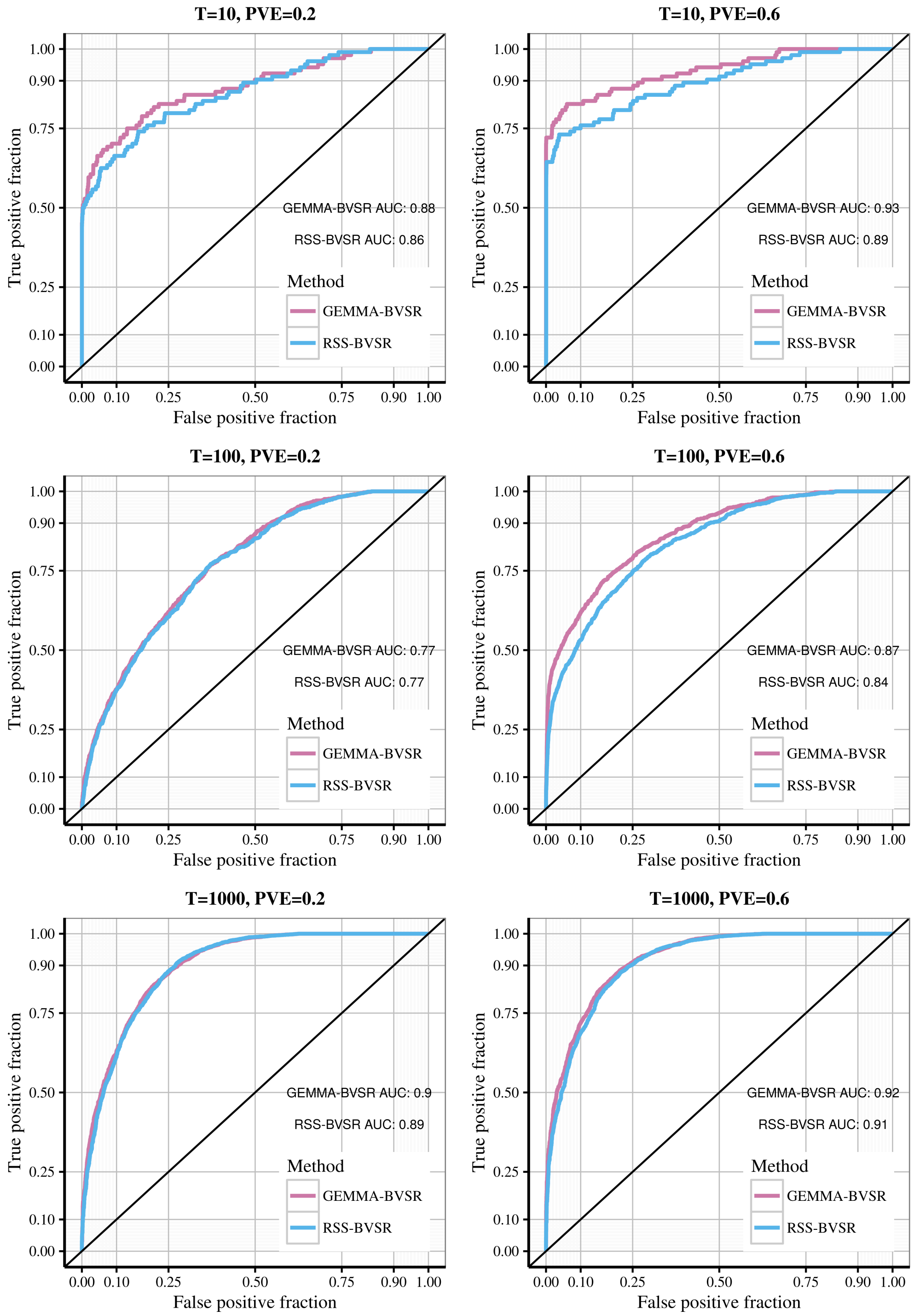


Fig 5: *Trade-off between true and false positives for GEMMA-BVSR (reddish purple) and RSS-BVSR (blue) in simulations of Scenario 2.1.*

β) will be 0, even if the SNPs are in complete LD.

In practice, we have found that problems can arise if RSS is applied to summary data that violate this assumption. Table 2 provides an illustrative example. Using summary statistics for high-density lipoprotein (HDL) cholesterol (Global Lipids Genetics Consortium, 2013), we computed the 1-SNP and 2-SNP Bayes factors (BFs) [as in Servin and Stephens (2007); see also Chen et al. (2015)] for 22 SNPs in the gene *ADH5*. Table 2 shows results for seven SNPs that are in complete LD with one another in the reference panel (1000 Genomes EUR $r^2 = 1$). One of these SNPs (rs7683704) has summary data based on approximately twice as many individuals as the other SNPs (187,000 versus 94,000 individuals). None of the SNPs shows evidence for marginal association with HDL (log10 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 all extremely large.

While the RSS likelihood – specifically, the covariance matrix in (2.1) – could in principle be modified to account for this issue, this is unattractive because it would require specification of sample overlaps for many pairs of SNPs. Instead, we suggest that genotype imputation [e.g. Servin and Stephens (2007); Marchini et al. (2007)] be used when generating GWAS summary data for public release, so that summary statistics are computed on the same individuals for each SNP.

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 2

Example of potential problems that can arise when RSS is applied to summary statistics computed from different samples. The table reports the sample sizes, single-SNP effect size estimates, SEs, and log 10 1-SNP BFs of seven SNPs that are in complete LD in the reference panel. The 2-SNP BFs reported are for rs7683704 with each of the other SNPs. These very large 2-SNP BFs appear unreasonable, likely due to the fact that the summary data were computed on different individuals.

6. Analysis of summary data on adult height. We applied RSS to summary statistics from a GWAS of human adult height, involving 253,288 individuals of European ancestry typed at $\sim$ 1.06 million SNPs

(Wood et al., 2014). Accessing the individual-level genotypes and phenotypes would be a considerable undertaking; in contrast the summary data are easily and freely available[1]. These data satisfy our requirement that summary statistics were computed on the same individuals [Section 1.1.2 of Supplementary Note in Wood et al. (2014)].

We filtered SNPs as in Bulik-Sullivan et al. (2015), and then removed SNPs that were absent in the genetic map of HapMap CEU 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). 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 2.6 GHz 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 28 hours, and the time of RSS-BSLMM ranged from 4 to 30 hours, both depending on the length of chromosomes.

We first inferred PVE (SNP heritability) from these summary data. Figure 6 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, 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. This is 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 co-

[1] `https://www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files`

horts (number of SNPs: 0.97-1.12 million; sample size: 1,145-5,668). The improved precision of the PVE estimates illustrates one benefit of being able to analyze summary data with a large sample size.

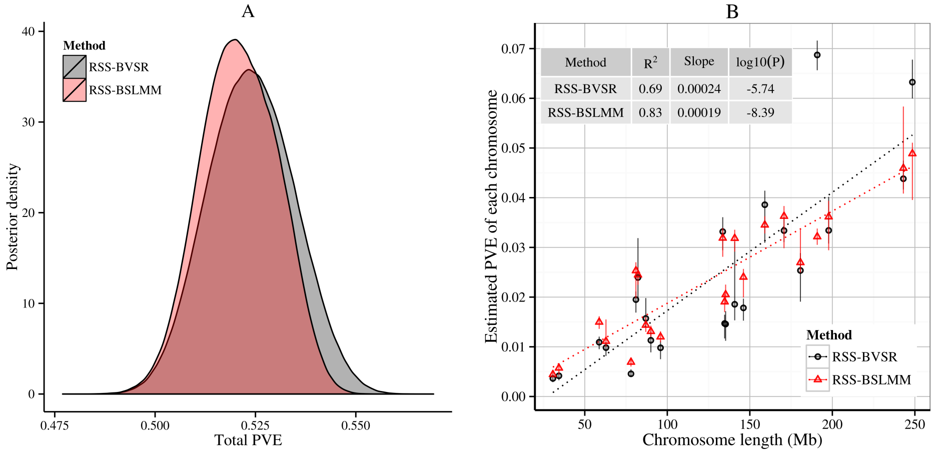


Fig 6: *Posterior inference of SNP heritability (PVE) for adult human height. Panel A: posterior distribution of the total PVE. Panel B: posterior median and 95% credible interval for PVE of each chromosome against the chromosome length, where the dotted lines 6 the fitted regression line.*

Next, we used RSS-BVSR 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. Among them, 384 SNPs were included in our filtered set of SNPs. Taking a region of ±40-kb around each of these SNPs, our analysis identified almost all of these regions (379/384) as showing strong signal for association (estimated ENS ≥ 1). However, 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 5194 regions with ENS ≥ 1, of which 2138 are putatively novel in that they are not near any of the previous 697 GWAS hits (distance > 1 Mb). Some of these 2138 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 (23 regions with estimated ENS > 3), and identified several loci harboring genes that seem plausibly related to height. These include the gene *WWOX*, which is a tumor suppressor linked to skeletal system morphogenesis (Del Mare et al., 2011; Aqeilan et al., 2008), and the gene *ALX1* (a.k.a. *CART1*), which is involved in bone development (Iioka et al., 2003).

7. Discussion.

We have presented a novel Bayesian method to infer 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, and thus provides a unified framework for various genome-wide analyses. We illustrate the applications of our framework on heritability estimation and association detection. Other potential applications include training phenotype prediction models, prioritizing causal variants and testing gene-level effects.

Our work highlights three conditions that should ideally hold for RSS to be applied. First, the marginal phenotype-genotype correlation of each covariate (SNP) must be small. In GWAS this holds, empirically, in a very wide range of studies (Table 1), but it may not hold in other contexts. Second, RSS depends on having an adequate estimate of the matrix R, which captures correlations among the covariates. In GWAS we often have available large suitable reference panels which help here. Third, our current implementation of RSS requires that the input summary data are computed on the same samples. Otherwise, misleading results can be obtained (Section 5). This last point suggests that, more generally, if the estimate $\widehat{R}$ is badly misspecified (e.g. computed from a reference panel that is not a good match to the study sample) then results of RSS could be problematic; however we have not studied this issue in detail.

We view the present work as the first stage of what could be done with RSS using GWAS summary statistics. One important extension is to integrate additional genomic information into the prior distributions. For example, Carbonetto and Stephens (2013) allow the prior probability of each SNP being included to depend on a covariate, such as biological pathway membership,

$$\beta_j | S \sim (1-\pi_j)\delta_0 + \pi_j \mathcal{N}(0, \sigma^2(S)), \;\; \text{logit}(\pi_j) = \theta_0 + \theta a_j, \tag{7.1}$$

where $a_j = 1$ if and only if SNP j is in the pathway. Unlike prior (3.2), prior (7.1) reflects that biologically related gene sets might preferentially harbor associated SNPs, essentially integrating the idea of gene set enrichment into GWAS (Wang, Li and Hakonarson, 2010). Second, some functional categories of the genome could contribute disproportionately to the heritability of complex traits (Gusev et al., 2014), which could be incorporated by letting the prior variance of the SNP effects depend on functional categorization, for example by

$$\beta_j | S \sim \mathcal{N}(0, \sigma_j^2(S)), \;\; \log(\sigma_j^2) = w_0 + \textstyle\sum_{g=1}^{G} w_g f_{j,g}, \tag{7.2}$$

where $f_{j,g} = 1$ when SNP j belongs to category g, w_0 captures the baseline (log) heritability and $\{w_g\}$ reflect the contribution of each category. This could provide a different way to partition heritability by functional annotation using GWAS summary statistics (Finucane et al., 2015).

Acknowledgements. We thank Xin He, Rina Foygel Barber, Yongtao Guan, Peter Carbonetto, Xiaoquan Wen and Xiang Zhou for discussions. We thank Raman Shah and John Zekos for technical support. This study makes use of data generated by the Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. We acknowledge the University of Chicago Research Computing Center for support.

References.

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

Bulik-Sullivan, B., Loh, P.-R., Finucane, H., Ripke, S., Yang, J., Psychiatric Genomics Consortium, Schizophrenia Working Group, Patterson, N., Daly, M. J., Price, A. L. and Neale, B. M. (2015). LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics* **47** 291–295.

Carbonetto, P. and Stephens, M. (2013). Integrated enrichment analysis of variants and pathways in genome-wide association studies indicates central role for IL-2 Signaling genes in Type 1 Diabetes, and Cytokine Signaling genes in Crohn's Disease. *PLoS Genetics* **9** e1003770.

Casella, G. and Robert, C. P. (1996). Rao-Blackwellisation of sampling schemes. *Biometrika* **83** 81–94.

Chen, W., Larrabee, B. R., Ovsyannikova, I. G., Kennedy, R. B., Haralambieva, I. H., Poland, G. A. and Schaid, D. J. (2015). Fine mapping causal variants with an approximate Bayesian method using marginal test statistics. *Genetics* **200** 719–736.

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

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

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

Del Mare, S., Kurek, K. C., Stein, G. S., Lian, J. B. and Aqeilan, R. I. (2011). Role of the WWOX tumor suppressor gene in bone homeostasis and the pathogenesis of osteosarcoma. *American Journal of Cancer Research* **1** 585.

Donnelly, P. (2008). Progress and challenges in genome-wide association studies in humans. *Nature* **456** 728–731.

Efron, B. (1993). Bayes and likelihood calculations from confidence intervals. *Biometrika* **80** 3–26.

Ehret, G. B., Lamparter, D., Hoggart, C. J., Whittaker, J. C., Beckmann, J. S., Kutalik, Z., of Anthropometric Traits Consortium, G. I. et al. (2012). A multi-SNP locus-association method reveals a substantial fraction of the missing heritability. *The American Journal of Human Genetics* **91** 863–871.

Finucane, H. K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V., Xu, H., Zang, C., Farh, K. et al. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics*.

Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., Belmont, J. W., Boudreau, A., Hardenbol, P., Leal, S. M. et al. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449** 851–861.

Nature Genetics (2012). Asking for more. *Nature Genetics* **44** 733.

Guan, Y. and Stephens, M. (2008). Practical issues in imputation-based association mapping. *PLoS Genetics* **4** e1000279.

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

Guan, Y. and Wang, K. (2013). Whole-genome multi-SNP-phenotype association analysis. In *Advances in Statistical Bioinformatics* (K.-A. Do, Z. S. Qin and M. Vannucci, eds.) 224–243. Cambridge University Press.

Gusev, A., Lee, S. H., Trynka, G., Finucane, H., Vilhjálmsson, B. J., Xu, H., Zang, C., Ripke, S., Bulik-Sullivan, B., Stahl, E., Kähler, A. K., Hultman, C. M., Purcell, S. M., McCarroll, S. A., Daly, M., Pasaniuc, B., Sullivan, P. F., Neale, B. M., Wray, N. R., Raychaudhuri, S. and Price, A. L. (2014). Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *The American Journal of Human Genetics* **95** 535–552.

Hoggart, C. J., Whittaker, J. C., De Iorio, M. and Balding, D. J. (2008). Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genetics* **4** e1000130.

Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B. and Eskin, E. (2014). Identifying causal variants at loci with multiple signals of association. *Genetics* **198** 497–508.

Iioka, T., Furukawa, K., Yamaguchi, A., Shindo, H., Yamashita, S. and Tsukazaki, T. (2003). 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** 1419–1429.

Lee, D., Bigdeli, T. B., Riley, B. P., Fanous, A. H. and Bacanu, S.-A. (2013). DIST: direct imputation of summary statistics for unmeasured SNPs. *Bioinformatics* **29** 2925–2927.

Lee, D., Williamson, V. S., Bigdeli, T. B., Riley, B. P., Fanous, A. H., Vladimirov, V. I. and Bacanu, S.-A. (2015). JEPEG: a summary statistics based tool for gene-level joint testing of functional variants. *Bioinformatics* **31** 1176–1182.

Lin, D. (2005). An efficient Monte Carlo approach to assessing statistical significance in genomic studies. *Bioinformatics* **21** 781–787.

Liu, J. Z., Mcrae, A. F., Nyholt, D. R., Medland, S. E., Wray, N. R., Brown, K. M., Hayward, N. K., Montgomery, G. W., Visscher, P. M., Martin, N. G. et al. (2010). A versatile gene-based test for genome-wide association studies. *The American Journal of Human Genetics* **87** 139–145.

Loh, P.-R., Tucker, G., Bulik-Sullivan, B. K., Vilhjalmsson, B. J., Finucane, H. K., Chasman, D. I., Ridker, P. M., Neale, B. M., Berger, B., Patterson, N. et al. (2015). Efficient Bayesian mixed model analysis increases association power in large cohorts. *Nature Genetics* **47** 284-290.

Marchini, J., Howie, B., Myers, S., McVean, G. and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics* **39** 906–913.

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. and Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics* **9** 356–369.

Moser, G., Lee, S. H., Hayes, B. J., Goddard, M. E., Wray, N. R. and Visscher, P. M. (2015). Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. *PLoS Genetics* **11** e1004969.

Palla, L. and Dudbridge, F. (2015). 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** 250–259.

Peise, E., Fabregat-Traver, D. and Bientinesi, P. (2015). High performance solutions for big-data GWAS. *Parallel Computing* **42** 75–87.

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

Pritchard, J. K. and Przeworski, M. (2001). Linkage disequilibrium in humans: models and data. *The American Journal of Human Genetics* **69** 1–14.

Sabatti, C. (2013). Multivariate linear models for GWAS. In *Advances in Statistical Bioinformatics* (K.-A. Do, Z. S. Qin and M. Vannucci, eds.) 188–207. Cambridge University Press.

Seaman, S. and Müller-Myhsok, B. (2005). Rapid simulation of P values for product methods and multiple-testing adjustment in association studies. *The American Journal of Human Genetics* **76** 399–408.

Servin, B. and Stephens, M. (2007). Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genetics* **3** e114.

Smedley, D., Haider, S., Durinck, S., Pandini, L., Provero, P., Allen, J., Arnaiz, O., Awedh, M. H., Baldock, R., Barbiera, G. et al. (2015). The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Research* **43** W589-W598.

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

Stephens, M. (2016). False discovery rates: A new deal. *bioRxiv*.

Vilhjalmsson, B., Yang, J., Finucane, H. K., Gusev, A., Lindstrom, S., Ripke, S., Genovese, G., Loh, P.-R., Bhatia, G., Do, R. et al. (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *The American Journal of Human Genetics* **97** 576–592.

Wakefield, J. (2009). Bayes factors for genome-wide association studies: comparison with P-values. *Genetic Epidemiology* **33** 79–86.

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

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

Wen, X. and Stephens, M. (2014). Bayesian methods for genetic association analysis with heterogeneous subgroups: From meta-analyses to gene–environment interactions. *The Annals of Applied Statistics* **8** 176–203.

Wood, A. R., Esko, T., Yang, J., Vedantam, S., Pers, T. H., Gustafsson, S., Chu, A. Y., Estrada, K., Luan, J., Kutalik, Z. et al. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics* **46** 1173–1186.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W. et al. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* **42** 565–569.

Yang, J., Manolio, T. A., Pasquale, L. R., Boerwinkle, E., Caporaso, N., Cunningham, J. M., de Andrade, M., Feenstra, B., Feingold, E., Hayes, M. G. et al. (2011). Genome partitioning of genetic variation for complex traits using common SNPs. *Nature Genetics* **43** 519–525.

Yang, J., Ferreira, T., Morris, A. P., Medland, S. E., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Weedon, M. N., Loos, R. J. et al. (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature Genetics* **44** 369–375.

Zhou, X., Carbonetto, P. and Stephens, M. (2013). Polygenic modeling with Bayesian sparse linear mixed models. *PLoS Genetics* **9** e1003264.

Zhou, X. and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics* **44** 821–824.

Xiang Zhu
Department of Statistics
University of Chicago
E-mail: xiangzhu@uchicago.edu

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