Bayesian large-scale regression with GWAS summary statistics

Xiang Zhu[1] and Matthew Stephens[1,2]

[1]Department of Statistics, [2]Department of Human Genetics

THE UNIVERSITY OF CHICAGO

How can summary statistics be used in multiple-SNP analysis?

- Recent work has revealed potential merits of multiple-SNP analysis.
- Existing methods are often complicated by access to full data.
- Summary statistics from single-SNP analysis are widely available.

A novel statistical problem

Consider the multiple linear regression,

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

where $\mathbf{y}$ is an $n \times 1$ vector, X is an $n \times p$ matrix, $\boldsymbol{\beta}$ is the $p \times 1$ regression coefficient, and $\boldsymbol{\epsilon}$ is the error term. In regression analysis, we observe the individual-level data $\{X, \mathbf{y}\}$ and use them to infer the parameter of interest $\boldsymbol{\beta}$. Here we assume that the full data $\{X, \mathbf{y}\}$ are not available, and only summary statistics of simple linear regression are provided:

$$\hat{\beta}_j := (X_j^\intercal X_j)^{-1} X_j^\intercal \mathbf{y}, \quad s_j^2 := (nX_j^\intercal X_j)^{-1}(\mathbf{y} - X_j\hat{\beta}_j)^\intercal(\mathbf{y} - X_j\hat{\beta}_j)$$

where X_j is the jth column of X, $j = 1, \ldots, p$.

How do we infer $\boldsymbol{\beta}$ using $\{\hat{\beta}_j, s_j\}$?

Examples of tools for multiple-SNP analysis

A growing number of GWAS summary statistics-based methods have recently been published.

- `GCTA-COJO` [1]: approximate the standard multiple linear regression
- `CAVIAR` [2]: model z-scores at a locus as multivariate normal
- `LDSC` [3]: regress genome-wide χ^2 statistics on "LD scores"

Shortcomings of existing methods

- Their connections with methods using full data are not clear.
- They cannot be easily applied to various multiple-SNP problems.

These concerns can be addressed if

$\boldsymbol{\beta}$ has an explicit likelihood based on summary-level data.

References

[1] J. Yang, *et al.*, *Nature Genetics* **44**, 369 (2012).

[2] F. Hormozdiari, E. Kostem, E. Y. Kang, B. Pasaniuc, E. Eskin, *Genetics* **198**, 497 (2014).

[3] B. Bulik-Sullivan, *et al.*, *Nature Genetics* **47**, 291 (2015).

[4] E. L. Lehmann, *Elements of large-sample theory* (Springer Science & Business Media, 1999).

[5] X. Wen, M. Stephens, *The Annals of Applied Statistics* **4**, 1158 (2010).

[6] P. Carbonetto, M. Stephens, *PLoS Genetics* **9**, e1003770 (2013).

[7] J. K. Pickrell, *The American Journal of Human Genetics* **94**, 559 (2014).

[8] G. Kichaev, *et al.*, *PLoS Genetics* **10**, e1004722 (2014).

[9] A. Gusev, *et al.*, *The American Journal of Human Genetics* **95**, 535 (2014).

[10] X. Zhu, M. Stephens, *Probabilistic Modeling in Genomics* (Cold Spring Harbor Laboratory, 2015).

[11] A. R. Wood, *et al.*, *Nature Genetics* **46**, 1173 (2014).

[12] 1000 Genomes Project Consortium, *Nature* **467**, 1061 (2010).

[13] Wellcome Trust Case Control Consortium , *Nature* **447**, 661 (2007).

[14] Y. Guan, M. Stephens, *The Annals of Applied Statistics* **5**, 1780 (2011).

[15] X. Zhou, M. Stephens, *Nature Genetics* **44**, 821 (2012).

[16] X. Zhou, P. Carbonetto, M. Stephens, *PLoS Genetics* **9**, e1003264 (2013).

[17] B. Servin, M. Stephens, *PLoS Genetics* **3**, e114 (2007).

[18] Y. Guan, M. Stephens, *PLoS Genetics* **4**, e1000279 (2008).

Acknowledgements

We thank Xin He, Yongtao Guan, Peter Carbonetto for helpful discussions. We thank Raman Shah and John Zekos for expert 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 of this work.

Regression with Summary Statistics (RSS) provides a solution.

Likelihood

We derive the following regression model for GWAS summary statistics:

$$\hat{\boldsymbol{\beta}} | S, R, \boldsymbol{\beta} \sim N(SRS^{-1}\boldsymbol{\beta}, SRS),$$

- $\hat{\boldsymbol{\beta}} := (\hat{\beta}_1, \ldots, \hat{\beta}_p)^\intercal$, where $\hat{\beta}_j$ is the single-SNP effect size estimate of SNP j;
- $S := \text{diag}(\mathbf{s})$, $\mathbf{s} := (s_1, \ldots, s_p)^\intercal$, where s_j is the standard error of $\hat{\beta}_j$;
- R is the population linkage disequilibrium (LD) matrix.

We term the model *Regression with Summary Statistics.*

Features of RSS model

- It produces an explicit likelihood of multiple-SNP effect $\boldsymbol{\beta}$.
- It is mathematically justified by asymptotic theory [4].
- It is computationally tractable for genome-wide analysis.
- It answers multiple questions within a single framework.

Dual role of population LD

- $\hat{\beta}_j$ includes the effects of all SNPs that SNP j tags.

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

- $\hat{\beta}_j$ and $\hat{\beta}_k$ are correlated if SNP j and k are in LD.

$$\text{Cov}(\hat{\beta}_j, \hat{\beta}_k | S, R, \boldsymbol{\beta}) = s_j s_k R_{jk}.$$

We estimate R using a shrinkage method based on population genetic principles [5].

Prior

Four types of prior on $\boldsymbol{\beta}$ are considered.

- Linear mixed model (LMM) prior:

$$\beta_j \sim N(0, \sigma_P^2)$$

- Bayesian variable selection regression (BVSR) prior:

$$\beta_j \sim \pi N(0, \sigma_B^2) + (1 - \pi)\delta_0$$

- Bayesian sparse linear mixed model (BSLMM) prior:

$$\beta_j \sim \pi N(0, \sigma_B^2 + \sigma_P^2) + (1 - \pi) N(0, \sigma_P^2)$$

- Adaptive shrinkage (ASH) prior:

$$\beta_j \sim \pi_1 N(0, \sigma_1^2) + \cdots + \pi_K N(0, \sigma_K^2)$$

They depict three genetic architectures.

infinitesimal (LMM), *sparse* (BVSR), *hybrid* (BSLMM & ASH)

Posterior

We provide efficient MCMC schemes to simulate posterior distributions of $\boldsymbol{\beta}$. Multiple tasks can be performed simultaneously using the same posterior samples.

Extension

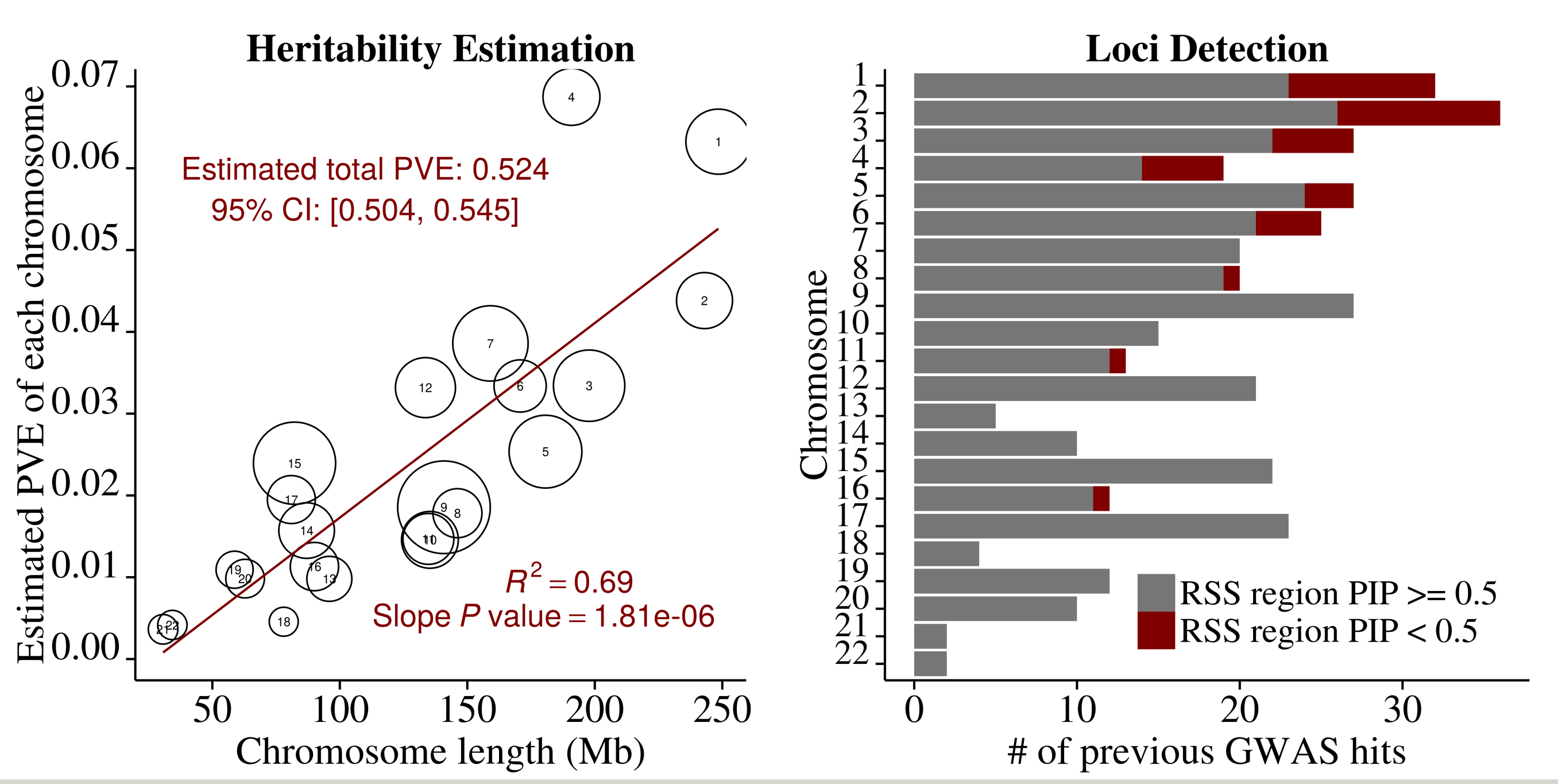
One important extension is to integrate additional genomic information with the RSS model [6, 7, 8, 9]. For example, together with the prior from [6],

$$\beta_j \sim (1 - \pi_j)\delta_0 + \pi_j N(0, \sigma_B^2), \quad \text{logit}(\pi_j) = \theta_0 + \theta \cdot \mathbf{1}\{\text{SNP } j \text{ is in the gene set}\}$$

RSS is able to infer gene set enrichment. Details will be presented at [10].

RSS on height GWAS supports a polygenic architecture of human stature.

We applied the RSS model on GWAS summary statistics of 1.06 million SNPs for adult human height from 253,288 individuals of European (EUR) ancestry [11]. The population LD matrix R was estimated from the 1000 Genomes [12] EUR samples.



Our heritability estimation (left) and loci detection (right) were comparable to results in [11], and supported a polygenic architecture hypothesis for human height.

Software

Software of fitting the RSS model is freely available from `https://github.com/stephenslab/rss`.

RSS yields results comparable to methods that require full data.

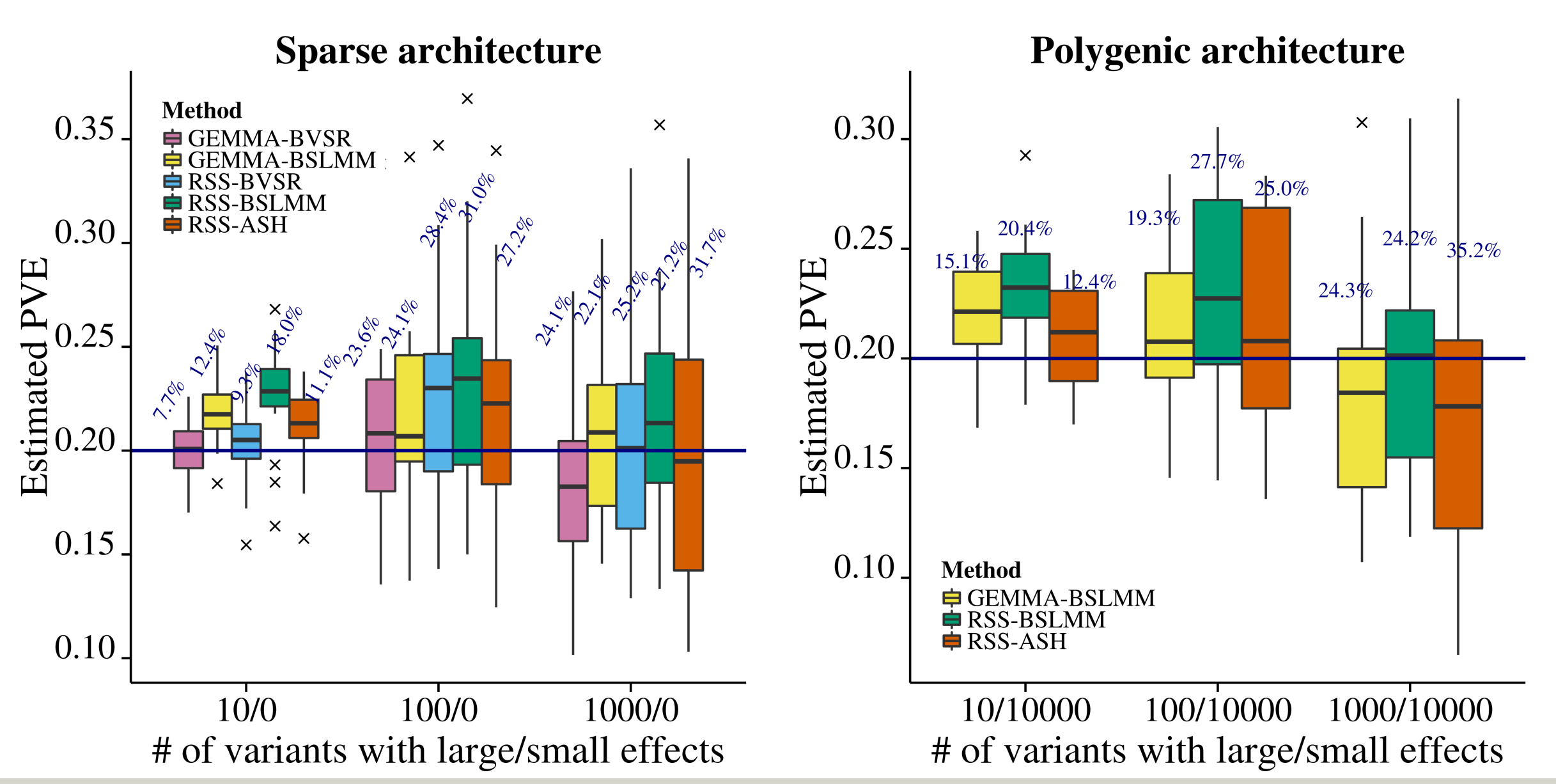
We compare RSS with individual-level data-based methods through simulations based on real genotype data [13].

Estimating SNP heritability

Phenotypic variation explained (PVE) by available genotypes:

$$\text{SPVE}(\boldsymbol{\beta}) := \sum_{i,j} \frac{R_{ij}\beta_i\beta_j}{\sqrt{(ns_i^2 + \hat{\beta}_i^2)(ns_j^2 + \hat{\beta}_j^2)}}$$

Full-data counterpart: `GEMMA-BVSR` and `GEMMA-BSLMM` [14, 15, 16]



Conclusion:

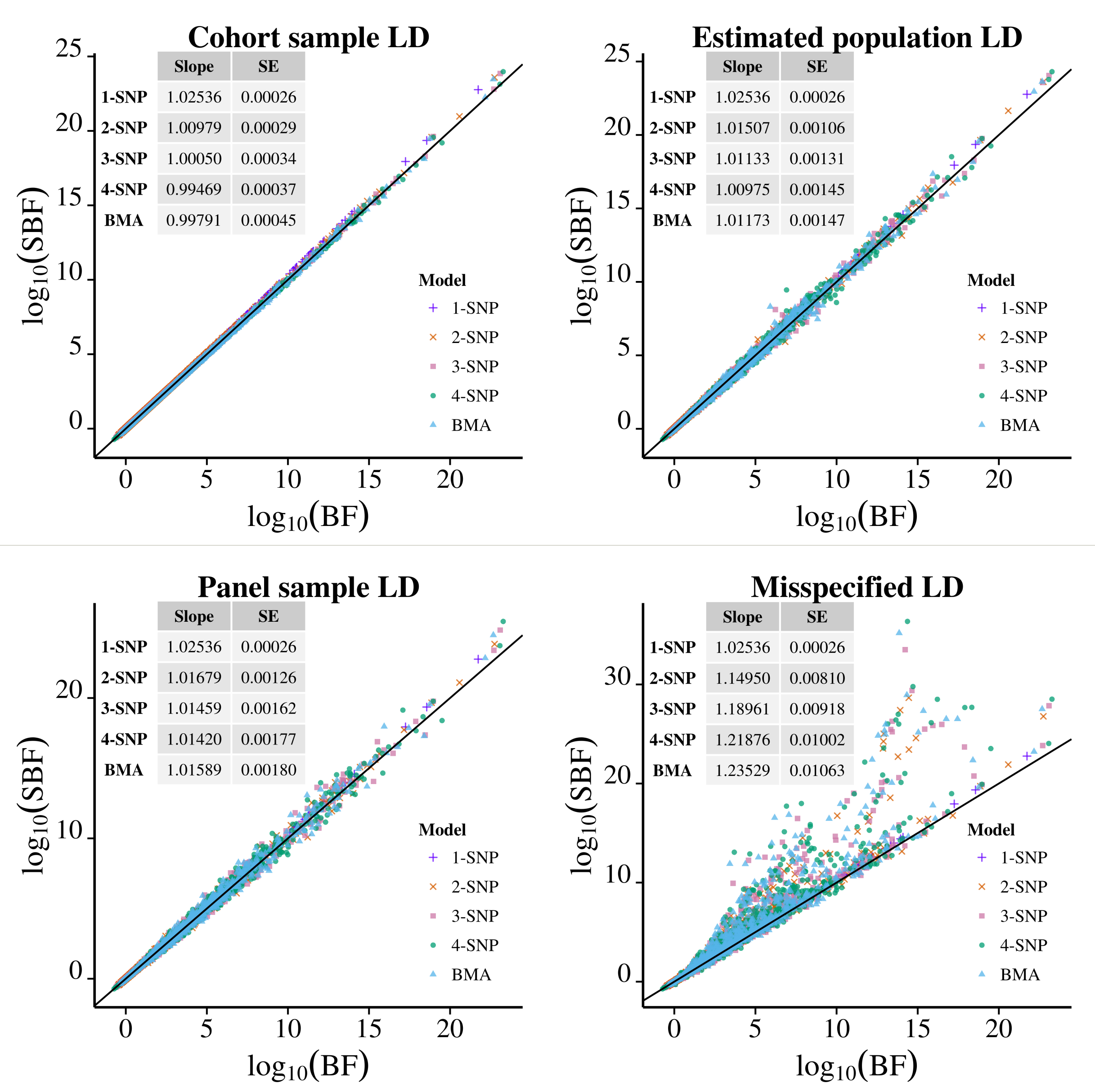
- PVE estimates using summary and individual-level data generally agree.
- Choice of prior is equally important for tools using summary and full data.

Testing SNP set association

Multiple-SNP Bayes factor (BF) of SNP set C under LMM prior:

$$\text{SBF}(C) = p(\hat{\boldsymbol{\beta}} | S, R, \sigma_P \neq 0) \;/\; p(\hat{\boldsymbol{\beta}} | S, R, \sigma_P = 0)$$

Full-data counterpart: `BIMBAM` [17, 18]



Cohort sample LD

	Slope	SE
1-SNP	1.02536	0.00026
2-SNP	1.00979	0.00029
3-SNP	1.00050	0.00034
4-SNP	0.99469	0.00037
BMA	0.99791	0.00045

Estimated population LD

	Slope	SE
1-SNP	1.02536	0.00026
2-SNP	1.01507	0.00106
3-SNP	1.01133	0.00131
4-SNP	1.00975	0.00145
BMA	1.01173	0.00147

Panel sample LD

	Slope	SE
1-SNP	1.02536	0.00026
2-SNP	1.01679	0.00126
3-SNP	1.01459	0.00162
4-SNP	1.01420	0.00177
BMA	1.01589	0.00180

Misspecified LD

	Slope	SE
1-SNP	1.02536	0.00026
2-SNP	1.14950	0.00810
3-SNP	1.18961	0.00918
4-SNP	1.21876	0.01002
BMA	1.23529	0.01063

Conclusion:

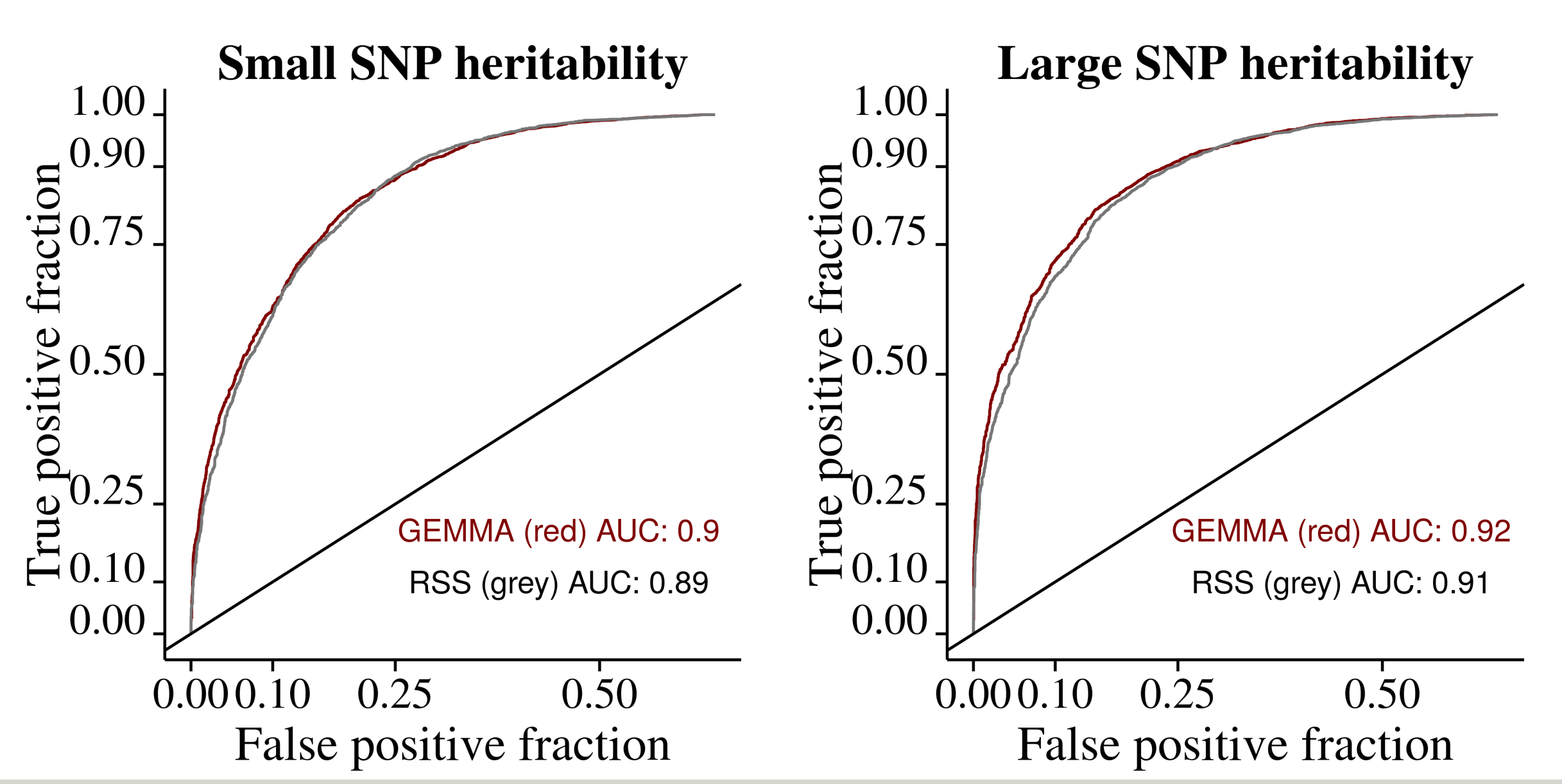
- SBF from summary data is an accurate approximation of BF from full data.
- Poorly specified LD can distort the summary-based method.

Detecting genome-wide association

Posterior inclusion probability (PIP) of SNP j under BVSR prior:

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

Full-data counterpart: `GEMMA-BVSR` [14, 15]



Conclusion:

- Association methods based on summary and full data have similar power.

More information at `xiangzhu@uchicago.edu`