Bayesian variant-based pathway enrichment analysis using GWAS summary statistics

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

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

THE UNIVERSITY OF CHICAGO

Can pathway enrichment and variant prioritization be integrated?

Carbonetto and Stephens [1] proposed a single framework that integrated *testing pathway enrichment*, *estimating enrichment level* and *prioritizing genetic variants* in the enriched pathways. The software, `BMApathway`, is available at `https://github.com/pcarbo/bmapathway`.

A novel integrated approach

The GWAS full data are genotypes $X := (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\mathsf{T}$ and phenotypes $\mathbf{y} := (y_1, \ldots, y_n)^\mathsf{T}$ from n unrelated individuals.

- For continuous traits, linear regression is used:

$$y_i|\mathbf{x}_i, \boldsymbol{\beta} \sim N(\beta_0 + \mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}, \tau^{-1}).$$

- For binary traits, logistic regression is used:

$$y_i|\mathbf{x}_i, \boldsymbol{\beta} \sim \text{Bernoulli}(\eta(\mathbf{x}_i, \boldsymbol{\beta})),\ \text{logit}(\eta(\mathbf{x}_i, \boldsymbol{\beta})) = \beta_0 + \mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}.$$

The multiple-SNP effect of p SNPs, $\boldsymbol{\beta} := (\beta_1, \ldots, \beta_p)^\mathsf{T}$, has prior

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

The *enrichment* of associations withing a pathway is modeled as

$$\text{logit}_{10}(\pi_j) = \theta_0 + a_j\theta,$$

where $a_j := \mathbf{1}\{\text{SNP } j \text{ is in the pathway}\}$, θ is the *log-fold enrichment* and θ_0 is the *genome-wide log-odds*.

Notations: $\mathbf{a} := (a_1, \ldots, a_p)^\mathsf{T}$ and $\boldsymbol{\gamma} := (\gamma_1, \ldots, \gamma_p)^\mathsf{T}$, $\gamma_j = \mathbf{1}\{\beta_j \neq 0\}$.

Can the integrated method be applied to GWAS summary data?

- Application of `BMApathway` is complicated by access to full data.
- Summary statistics from single-SNP analysis are widely available.
- The enrichment prior is useful even if full data are not provided.

A similar integrated analysis is possible if

we keep the prior and use a likelihood that only relies on summary data.

References

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

[2] X. Zhu, M. Stephens, *Presented at the 65th Annual Meeting of The American Society of Human Genetics, Baltimore, MD* (2015).

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

[4] D. Croft, *et al.*, *Nucleic Acids Research* **42**, D472 (2014).

[5] L. Y. Geer, *et al.*, *Nucleic Acids Research* **38**, D492 (2010).

[6] E. G. Cerami, *et al.*, *Nucleic Acids Research* **39**, D685 (2011).

[7] C. F. Schaefer, *et al.*, *Nucleic Acids Research* **37**, D674 (2009).

[8] H. Mi, A. Muruganujan, P. D. Thomas, *Nucleic Acids Research* **41**, D377 (2013).

[9] C. Wrzodek, F. Büchel, M. Ruff, A. Dräger, A. Zell, *BMC systems biology* **7**, 15 (2013).

[10] A. R. Pico, *et al.*, *PLoS Biology* **6**, e184 (2008).

[11] L. S. Chen, *et al.*, *The American Journal of Human Genetics* **86**, 860 (2010).

[12] H. Gui, M. Li, P. C. Sham, S. S. Cherny, *BMC research notes* **4**, 386 (2011).

[13] M.-X. Li, J. S. Kwan, P. C. Sham, *The American Journal of Human Genetics* **91**, 478 (2012).

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

[15] 1000 Genomes Project Consortium, *Nature* **467**, 1061 (2010).

Regression with Summary Statistics (RSS) provides a solution.

We propose the following regression model for GWAS summary statistics:

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

- $\widehat{\boldsymbol{\beta}} := (\hat{\beta}_1, \ldots, \hat{\beta}_p)^\mathsf{T}$, where $\hat{\beta}_j$ is the single-SNP effect estimate of SNP j;
- $S := \text{diag}(\mathbf{s})$, $\mathbf{s} := (s_1, \ldots, s_p)^\mathsf{T}$, 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* [2].

RSS uses an algorithm based on variational approximation.

Our posterior computation exploits the fact that

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

The strength of pathway enrichment is measured by

$$\text{SBF}(\mathbf{a}) := p(\widehat{\boldsymbol{\beta}}|S, R, \mathbf{a}, \theta > 0) \ / \ p(\widehat{\boldsymbol{\beta}}|S, R, \mathbf{a}, \theta = 0).$$

The log-fold enrichment θ is estimated from

$$p(\theta|\mathbf{D}), \quad \mathbf{D} := \{\widehat{\boldsymbol{\beta}}, S, R, \mathbf{a}\}.$$

The association signal of SNP j given the enrichment is summarized as

$$\text{SPIP}(j) := p(\gamma_j = 1|\widehat{\boldsymbol{\beta}}, S, R, \mathbf{a}).$$

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

Decomposition of marginal likelihood:

$$\log p(\mathbf{D}|\theta_0, \theta) = \underbrace{\mathsf{E}_q \log\left[\frac{q(\boldsymbol{\beta}, \boldsymbol{\gamma})}{p(\boldsymbol{\beta}, \boldsymbol{\gamma}|\mathbf{D}, \theta_0, \theta)}\right]}_{\text{Kullback-Leibler (KL) divergence}} + \underbrace{\mathsf{E}_q \log\left[\frac{p(\mathbf{D}, \boldsymbol{\beta}, \boldsymbol{\gamma}|\theta_0, \theta)}{q(\boldsymbol{\beta}, \boldsymbol{\gamma})}\right]}_{\text{Evidence lower bound (LB)}}$$

Optimization problem:

$$q^\star = \arg\min_q \text{KL}(q; \theta_0, \theta) = \arg\max_q \text{LB}(q; \theta_0, \theta)$$

The *only* assumption being made: *mean field approximation*

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

Optimal solution $q_j^\star$ for each each q_j:

$$q_j^\star(\beta_j, \gamma_j) = [\alpha_j N(\beta_j; \mu_j, \sigma_j^2)]^{\gamma_j}[(1 - \alpha_j)\delta_0(\beta_j)]^{1-\gamma_j}$$

Iterative scheme for obtaining $q_j^\star$:

$$\sigma_j^2 = (s_j^{-2} + \sigma_B^{-2})^{-1}$$
$$\mu_j = \sigma_j^2(s_j^{-2}\hat{\beta}_j - s_j^{-1}\textstyle\sum_{i \neq j} s_i^{-1}R_{ij}\alpha_i\mu_i)$$
$$\frac{\alpha_j}{1 - \alpha_j} = \frac{\pi_j}{1 - \pi_j} \cdot \frac{\sigma_j}{\sigma_B} \cdot \exp\left\{\frac{\mu_j^2}{2\sigma_j^2}\right\}$$

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

- Current approach is to use points of $\{\theta_0, \theta\}$ in a regular grid and set

$$p(\theta_0, \theta|\mathbf{D}) \propto \exp\{\text{LB}(q^\star; \theta_0, \theta)\}.$$

- We are investigating more efficient approaches.

Parallel implementation

$R_{ij} = 0$ *if SNP* i *and* j *are on different chromosomes.*

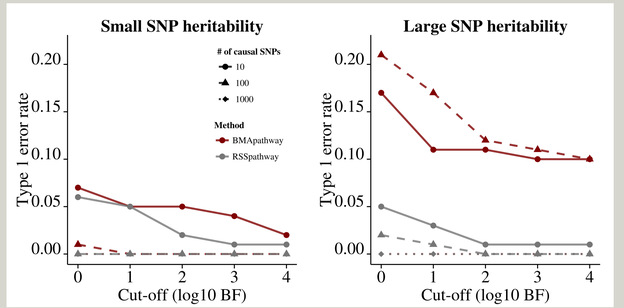
- Iterative update of $q_j^\star$ only requires data from SNP i that $R_{ij} \neq 0$.
- $\text{LB}(q^\star; \theta_0, \theta) = \sum_{c=1}^{22} \text{LB}_c$, LB_c only uses data from Chromosome c.

RSS yields results comparable to a method that requires genotype data.

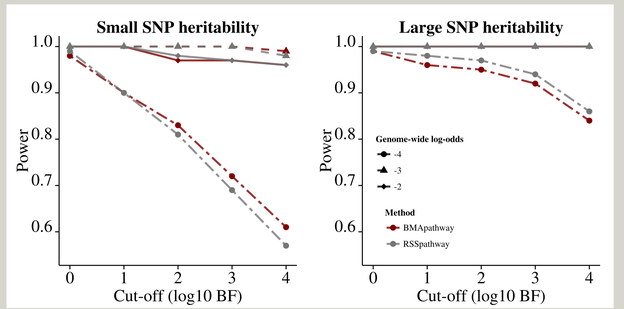
We compare RSS with a full data-based method, `BMApathway` [1], through simulations based on real genotype data [3].

- Null dataset assumes that each SNP is equally likely to be causal.
- Enriched dataset assumes that SNPs in the pathway are more likely to be associated with the phenotypes. The pathway used in simulations is signal transduction (Reactome [4]) retrieved from BioSystems [5].

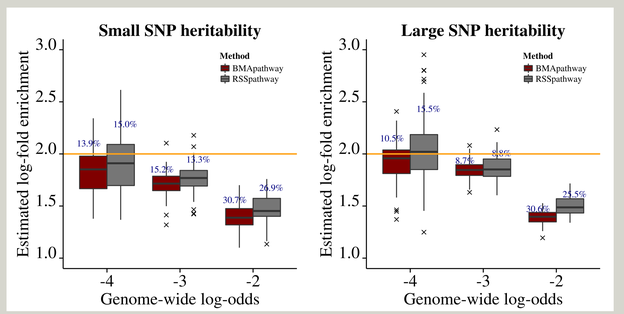
Type 1 error



Power



Enrichment estimation



Acknowledgements

We thank Peter Carbonetto and Xin He 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.

RSS provides a way to gain biological insights into complex human traits.

Pathways are retrieved from Pathway Commons (PC) [6], Biosystems (BS) [5] and BioCarta (BC), which include gene sets derived from Reactome [4], PID [7], PANTHER [8], KEGG [9] and WikiPathways (Wiki) [10].

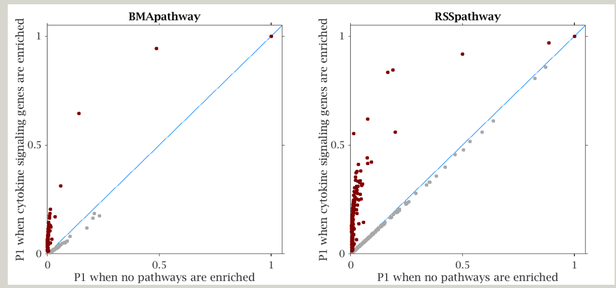
Crohn's disease

We applied RSS on 3160 curated pathways and GWAS summary statistics of 435,615 SNPs for Crohn's disease from 4,686 individuals (1,748 cases and 2,938 controls) in the British population [3].

The top five candidate pathways for enrichment of Crohn's disease detected by `RSSpathway` and `BMApathway` are the same, and their enrichments are also significant using other methods.

Pathway	Source	Database	$\log_{10}$(BF)		$-\log_{10}(P)$		
			RSSpathway	BMApathway [1]	GRASS [11]	Gates-Simes [12]	HYST [13]
IL12-mediated signaling events	PID	PC	9.42	4.00	11.06	10.80	15.65
Cytokine signaling in immune system	Reactome	BS	8.87	5.97	14.91	10.32	15.65
IL23-mediated signaling events	PID	PC	8.72	4.13	12.29	11.02	Inf
Immune system	Reactome	BS	6.24	3.30	10.72	9.97	8.39
Immune system	Reactome	PC	5.75	3.07	10.04	10.13	7.49

The following figure shows P_1, posterior probability that locus contains disease risk variants, with and without enrichment of cytokine signaling.



Adult height

We applied RSS on 3700 curated pathways and GWAS summary statistics of 1.06 million SNPs for adult human height from 253,288 individuals of European (EUR) ancestry [14]. The population LD matrix R was estimated from the 1000 Genomes [15] EUR samples.

The top-ranked pathways are listed below, and most of them are linked to skeletal development and homeostasis, regulation of human stature, cell migration, cancer and etiology of bone diseases.

Pathway	Source	Database	# of genes
Hedgehog signaling pathway	Wiki	BS	22
Hedgehog signaling pathway	KEGG	BS	51
Basal cell carcinoma	KEGG	BS	55
Hedgehog 'on' state	Reactome	BS	84
RAC1 signaling pathway	PID	PC	54
Rho cell motility signaling pathway	BC	BC	32
Signaling by Hedgehog	Reactome	BS	135
Regulatory role of PI3K subunit p85	BC	BC	16
How does salmonella hijack a cell	BC	BC	13
Y branching of actin filaments	BC	BC	20
Pathogenic *Escherichia coli* infection	KEGG	BS	55
Pathogenic *Escherichia coli* infection	Wiki	BS	56
How progesterone initiates oocyte membrane	BC	BC	33
Rho GTPases activate WASPs and WAVEs	Reactome	BS	35
Cytoskeletal regulation by Rho GTPase	PANTHER	PC	70
Signaling events mediated by the Hedgehog family	PID	BS	22
Signaling events mediated by the Hedgehog family	PID	PC	23
EPHB-mediated forward signaling	Reactome	PC	41
Ligand-receptor interactions	Reactome	BS	8

Software

Software of using RSS for integrated enrichment analysis, `RSSpathway`, will be available from `https://github.com/stephenslab/rss`.

More information at xiangzhu@uchicago.edu