Integrated enrichment analysis of genetic variants and biological pathways 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

Integrated analysis of GWAS is often limited by the access to full data.

Bayesian hierachical modelling has been used for joint analysis of genetic variants and biological pathways in GWAS recently [1, 2].

These methods, however, are often complicated by the limited access to individual-level data.

Bayesian hierarchical framework in [1]

Likelihood requires the individual-level data.

- Quantitative traits (e.g. height):

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

- Binary traits (e.g. schizophrenia):

$$y_i|\mathbf{x}_i, \beta \sim \mathrm{B}(1, \eta(\mathbf{x}_i, \beta)),\ \mathrm{logit}(\eta(\mathbf{x}_i, \beta)) = \beta_0 + \mathbf{x}_i^\intercal\beta$$

Prior specification does not depend on data.

- Effect size distribution:

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

- Probability of being "causal":

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

where $a_j := 1\{\text{SNP } j \text{ is in the pathway}\}$

RSS (Regression with Summary Statistics) offers a solution [3].

Unlike the individual-level data, GWAS summary statistics are often publicly available.

A possible solution is to develop similar methods for summary-level data.

Revisit the **Bayes' Theorem**:

$$p(\beta|\text{Individual Data}) \propto p(\text{Individual Data}|\beta)\cdot p(\beta)$$

$$\underbrace{p(\beta|\text{Summary Data})}_{\text{Posterior}} \propto \underbrace{p(\text{Summary Data}|\beta)}_{\text{Likelihood}}\cdot\underbrace{p(\beta)}_{\text{Prior}}$$

The only missing piece is $p(\text{Summary Data}|\beta)$.

RSS: a likelihood based on summary data

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

- multiple-SNP parameter: $\beta := (\beta_1, \ldots, \beta_p)'$
- single-SNP summary data: $\widehat{\beta} := (\hat{\beta}_1, \ldots, \hat{\beta}_p)'$
- $\widehat{S} := \mathrm{diag}(\widehat{s})$, $\widehat{s} := (\hat{s}_1, \ldots, \hat{s}_p)'$, $\hat{s}_j \approx \mathrm{se}(\hat{\beta}_j)$;
- $\widehat{R}$: the shrinkage estimate of LD matrix [4]

Joint analysis of variants and pathways

- Test the significance of enrichment:

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

- Estimate the level of enrichment:

$$p(\theta|\widehat{\beta}, \widehat{S}, \widehat{R}, a)$$

- Estimate the effect of SNP j under enrichment:

$$p(\beta_j|\widehat{\beta}, \widehat{S}, \widehat{R}, a)$$

The posterior distributions are approximated by **Variational Bayes** methods.

RSS is applied to 31 human phenotypes.

The input of RSS is usually **publicly available**.

Data	Summary
Phenotype	**31** complex traits/diseases
Reference panel	1000 Genomes Phase 3 [5]
Genetic variant	**∼1.1M** HapMap3 SNPs
Gene	**∼19K** protein-coding genes
Pathway	**3913** curated pathways

Software

`https://github.com/stephenslab/rss`.

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

Top-ranked candidate pathways for enrichment of genetic associations

RSS not only **tests** the significance of enrichment (BF), but also **estimates** the level of enrichment (θ) simultaneously.

Phenotype	Pathway	Source	Size	BF	Estimated θ_0	Estimated θ
Height [6]	Endochondral ossification	WikiPathways (BS)	65	7.7×10^{68}	-2.05, [-2.08, -2.05]	0.80, [0.74, 0.86]
Schizophrenia [7]	Chromatin modifying enzymes	Reactome (BS)	241	1.0×10^{3}	-2.13, [-2.15, -2.08]	1.33, [0.79, 1.76]
Myocardial infarction [8]	Thyroid hormone metabolism II	BioCyc (BS)	4	1.3×10^{209}	-4.05, [-4.05, -4.03]	4.50, [4.46, 4.50]
Low-density lipoprotein [9]	Chylomicron-mediated lipid transport	Reactome (PC)	17	3.4×10^{65}	-3.60, [-3.60, -3.58]	2.38, [2.36, 2.43]
Inflammatory bowel disease [10]	IL23-mediated signaling events	PID (PC)	37	3.1×10^{23}	-2.98, [-3.00, -2.90]	1.38, [1.20, 1.50]
Neuroticism [11]	Digestion of dietary carbohydrate	Reactome (PC)	9	7.3×10^{168}	-4.48, [-4.50, -4.45]	2.77, [2.66, 2.89]

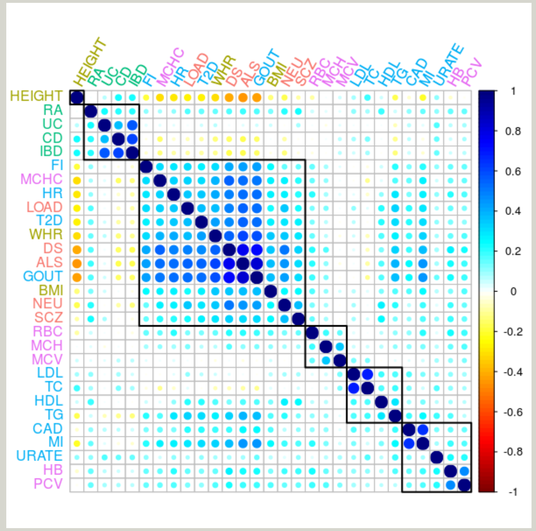
Abbreviations used in table: BS=NCBI BioSystems [12], PC=Pathway Commons 2 [13].

The complete analysis results are publicly available online:

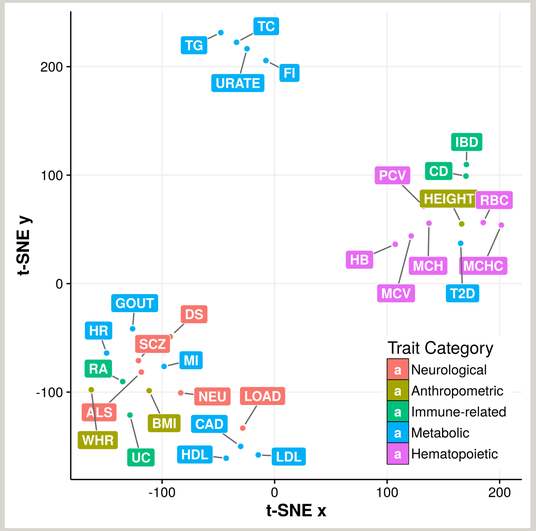
`http://xiangzhu.github.io/rss-gsea/_book/`

Sharing and specificity of enrichment profiles across phenotypes

For each trait, the enrichment profile consists of 3913 gene-set (log 10) BFs.



(a) Kendall's τ correlation matrix



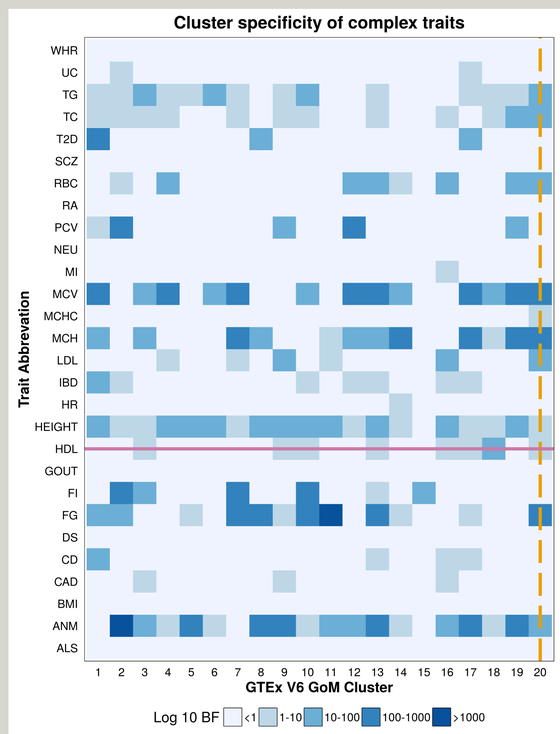
(b) t-Distributed stochastic neighbor embedding

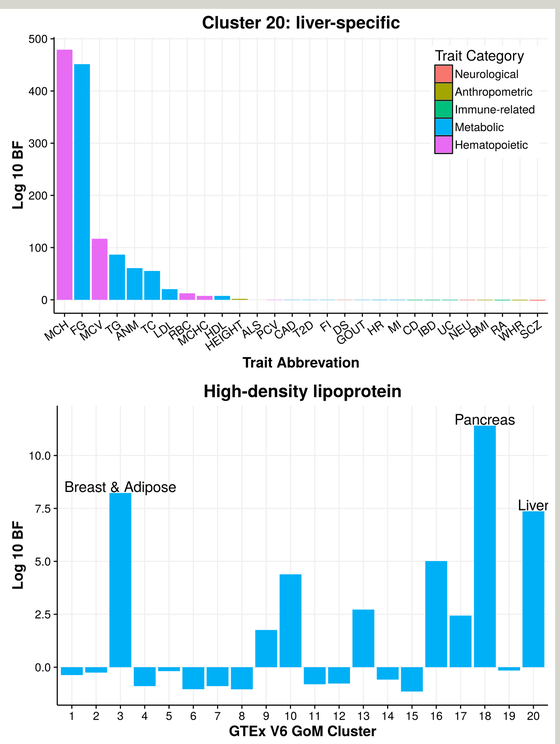
Integration of GWAS summary data and high-throughput molecular data

Two-step analysis:

Recipe	Example
High-dimensional molecular data	GTEx V6 RNA-seq gene expression [14]
↓ Data-driven methods	CountClust [15] ↓
Molecularly-derived gene sets	Clusters of distinctively expressed genes
↓ RSS	RSS ↓
Enrichment analysis of GWAS	Enrichment analysis of GWAS

- **Top right:**
  Given a cluster, examine the enrichment patterns across different phenotypes.
- **Bottom right:**
  Given a phenotype, examine the enrichment patterns across different clusters.





RSS reveals putatively novel loci not previously implicated by GWAS.

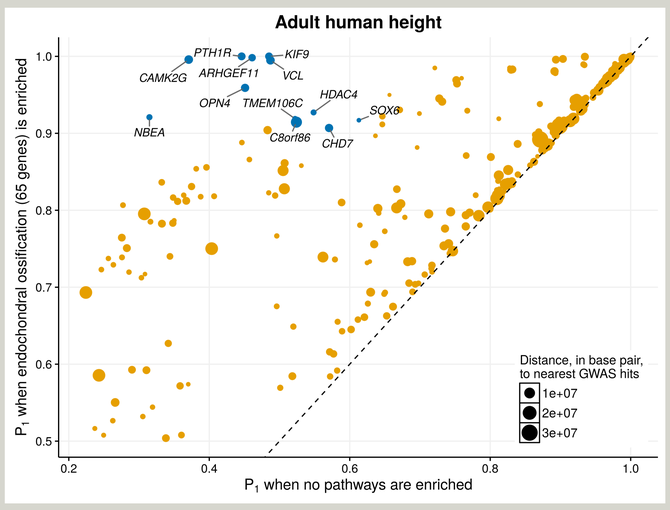
The genetic association between loci and phenotype is measured by P_1, the posterior probability that at least one SNP in the loci is associated with the phenotype.

Additional associations are detected when:

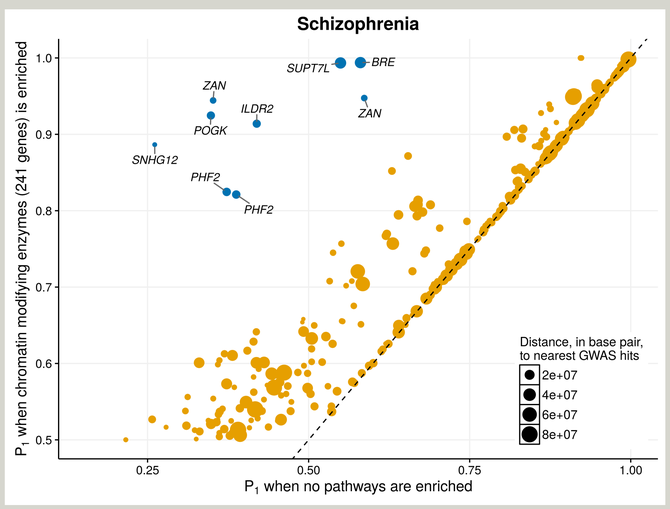
- $P_1(\cdot|\theta > 0)$ is much larger than $P_1(\cdot|\theta = 0)$;
- and, $P_1(\cdot|\theta > 0)$ is close to 1.

$P_1(\cdot|\theta > 0)$ and $P_1(\cdot|\theta = 0)$ are **automatically** obtained from the output of RSS.

Quantitative trait: adult human height [6]



Binary trait: schizophrenia [7]



References

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

[2] M. Evangelou, F. Dudbridge, L. Wernisch, *Bioinformatics* **30**, 690 (2014).

[3] X. Zhu, M. Stephens, *bioRxiv* (2016).

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

[5] 1000 Genomes Project Consortium, *Nature* **526**, 68 (2015).

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

[7] Schizophrenia Working Group of the Psychiatric Genomics Consortium, *Nature* **511**, 421 (2014).

[8] M. Nikpay, *et al.*, *Nature Genetics* **47**, 1121 (2015).

[9] T. M. Teslovich, *et al.*, *Nature* **466**, 707 (2010).

[10] J. Z. Liu, *et al.*, *Nature Genetics* **47**, 979 (2015).

[11] A. Okbay, *et al.*, *Nature Genetics* **48**, 624 (2016).

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

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

[14] The GTEx Consortium, *Science* **348**, 648 (2015).

[15] K. K. Dey, C. J. Hsiao, M. Stephens, *bioRxiv* (2016).

Acknowledgments

We thank Peter Carbonetto and Xin He for helpful discussions. We thank the GWAS consortia that made summary statistics publicly available.

This work was completed in part with resources provided by the University of Chicago Research Computing Center.

More information at `xiangzhu@uchicago.edu`