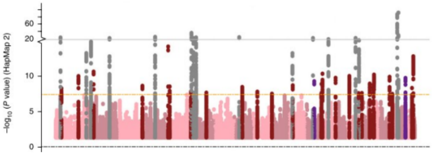
Large-scale genome-wide enrichment analyses of 31 human complex traits

Xiang Zhu
University of Chicago

JSM 2017, August 3

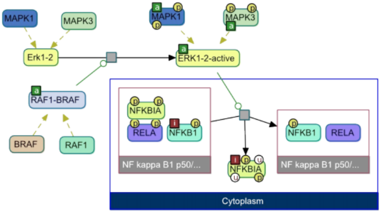
Examining associations between variables is a useful tool in genetics.

GWAS: SNP~Phenotype

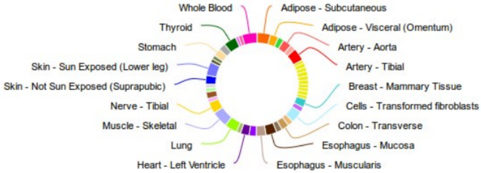


1000 Genomes: SNP~SNP

Biological Pathways: Gene~Gene



GTEx: Gene~Tissue



Enrichment analysis combines multiple sources of association.

- **SNP-Trait:** GWAS summary statistics
- **SNP-SNP:** linkage disequilibrium (LD)
- **Gene-Gene:** biological pathways
- **Gene-Tissue:** RNA-seq across tissues

Let’s keep this talk "jargon-free".

1. GWAS summary statistics
2. gene set enrichment analysis

What are GWAS summary statistics?

- **Data:** phenotype Y and genotype X
- **Size:** n (**>10K**) individuals and p (**>1M**) variants (SNPs)

$$Y := \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad X := \begin{bmatrix} x_{11} & \ldots & x_{1j} & \ldots & x_{1p} \\ x_{21} & \ldots & x_{2j} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ x_{n1} & \ldots & x_{nj} & \ldots & x_{np} \end{bmatrix}$$

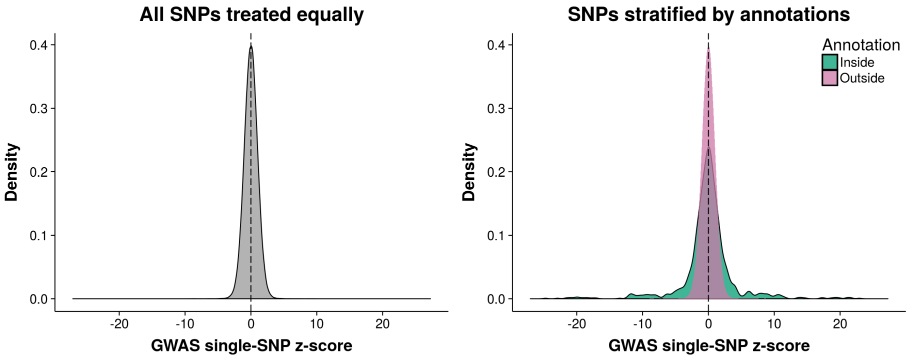
- **Model:** single-SNP association analysis

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \sim \begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{bmatrix} \quad \rightsquigarrow \quad \begin{cases} \hat{\beta}_j : & \text{marginal effect estimate} \\ \hat{\sigma}_j : & \text{standard error of } \hat{\beta}_j \end{cases}$$

- **Availability:** $\{\hat{\beta}_j,\ \hat{\sigma}_j\}\ >>\ \{Y,\ X\}$ (Nat. Genet. Editorial, 2012)

What is enrichment analysis?

- **Phenotype:** low-density lipoprotein (Teslovich *et al.*, 2010)
- **Pathway:** *chylomicron-mediated lipid transport* (17 genes)
- **Annotation:** Is the SNP “near” a pathway gene? (yes or no)



Recent reviews: de Leeuw *et al.* (2016); Pers (2016); Mooney *et al.* (2014); Wang *et al.* (2010).

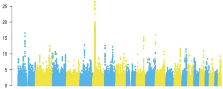
The “enrichment” idea is simple, but there are (at least) two technical issues.

1. If the gene set is truly enriched, we should relax significance threshold for “green” SNPs, but how much to relax?
   - **Data-driven** threshold ⟵ Function (Pathway, Phenotype)
   - Maintained type 1 error + Improved power
2. The “inflated” green curve may be driven by correlation between SNPs (LD), rather than enrichment of signal.
   - SNP 1 has a large genetic effect on a trait.
   - SNPs 2-100 have zero effect, but all are in **high LD** with SNP 1.
   - Thus, SNPs 1-100 all show very large **single-SNP** z-scores.

We develop a method that systematically utilizes enrichment information.

Public Data

GWAS summary statistics

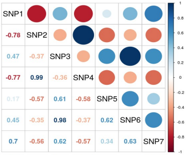


$\widehat{\beta} := (\hat{\beta}_1, \ldots, \hat{\beta}_p)^{\intercal}, \ \widehat{S} := \text{diag}\{(\hat{s}_1, \ldots, \hat{s}_p)^{\intercal}\}$

$\hat{\beta}_j$: marginal effect estimate of SNP j

$\hat{s}_j$: standard error of $\hat{\beta}_j$

External LD estimates



SNP1						
-0.78	SNP2					
0.47	-0.37	SNP3				
-0.77	0.99	-0.36	SNP4			
0.17	-0.57	0.61	-0.58	SNP5		
0.45	-0.35	0.98	-0.37	0.62	SNP6	
0.7	-0.56	0.62	-0.57	0.34	0.63	SNP7

$\widehat{R} := [\hat{r}_{ij}]_{p\times p}, \ \ \hat{r}_{ij}$: LD between SNP i and j

Predefined gene sets

Gene	Chr	Start	End
BRAF	7	140419127	140624564
MAPK1	22	22108789	22221970
MAPK3	16	30125426	30134827
NFKB1	4	103422486	103538459

$a_j := 1\{\text{SNP } j \text{ is "near" a gene in the set}\}$

Bayesian Model

Likelihood function

$$L(\beta) := \text{Normal}\left(\widehat{\beta};\ \widehat{S}\widehat{R}\widehat{S}^{-1}\beta,\ \widehat{S}\widehat{R}\widehat{S}\right)$$

Prior distribution

$$\beta_j \sim \pi_j \cdot \text{Normal}(0, \sigma_\beta^2) + (1-\pi_j)\delta_0$$

Baseline model (*M0*)

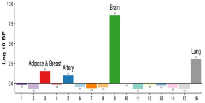
$$\log_{10}\left(\frac{\pi_j}{1-\pi_j}\right) = \theta_0$$

Enrichment model (*M1*)

$$\log_{10}\left(\frac{\pi_j}{1-\pi_j}\right) = \theta_0 + a_j\theta$$

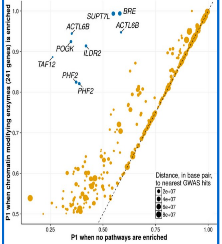
Inference

Gene set enrichment



$$\text{BF(gene set)} = \frac{\Pr(\text{Data} \mid M_1)}{\Pr(\text{Data} \mid M_0)}$$

Gene prioritization



$p(\beta \mid \text{Data},\ M_1)$ vs $p(\beta \mid \text{Data},\ M_0)$

Address Issue 1: Learning enrichment from data

Model-based approach:

- Assume that SNP j is trait-associated with probability π_j
- Represent π_j as a function of SNP-level annotation a_j

$$\log_{10}\left(\frac{\pi_j}{1-\pi_j}\right) := \theta_0 + a_j\theta$$

Data-adaptive threshold:

- Estimate enrichment parameter θ from data
- “Enriched” data $\rightsquigarrow$ $\theta > 0$ $\rightsquigarrow$ larger π_j for $a_j = 1$ $\rightsquigarrow$ increased power
- “Null” data $\rightsquigarrow$ $\theta \approx 0$ $\rightsquigarrow$ unchanged π_j $\rightsquigarrow$ maintained type 1 error

References: Veyrieras *et al.* (2008); Carbonetto & Stephens (2013); Pickrell (2014); Kichaev *et al.* (2014); Chen *et al.* (2016); Li & Kellis (2016); Wen *et al.* (2017).

Address Issue 2: Modeling linkage disequilibrium

Genome-wide multiple-SNP likelihood function:

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

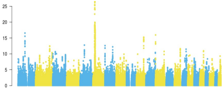
- multiple-SNP parameter: $\beta := (\beta_1, \ldots, \beta_p)'$
- single-SNP summary data: $\widehat{\beta} := (\hat{\beta}_1, \ldots, \hat{\beta}_p)'$
- $\widehat{S} := \mathrm{diag}(\widehat{s})$, $\widehat{s} := (\hat{s}_1, \ldots, \hat{s}_p)'$, $\hat{s}_j^2 := \hat{\sigma}_j^2 + n^{-1}\hat{\beta}_j^2 \simeq \hat{\sigma}_j^2$
- $\widehat{R}$: shrinkage estimate of LD (Wen & Stephens, 2010)
- **“Big data”** genetics: jointly analyze 1.1 million SNPs

Reference: Zhu and Stephens (2017 To appear). `http://dx.doi.org/10.1101/042457`.

We develop a method that systematically utilizes enrichment information.

Public Data

GWAS summary statistics

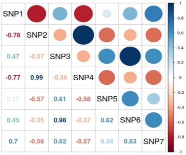


$\widehat{\beta} := (\hat{\beta}_1, \dots, \hat{\beta}_p)^\intercal$, $\widehat{S} := \text{diag}\{(\hat{s}_1, \dots, \hat{s}_p)^\intercal\}$

$\hat{\beta}_j$: marginal effect estimate of SNP j

$\hat{s}_j$: standard error of $\hat{\beta}_j$

External LD estimates



SNP1	●	●	●	●	●	●
-0.78	SNP2	●	●	●	●	●
0.47	-0.37	SNP3	●	●	●	●
-0.77	0.99	-0.36	SNP4	●	●	●
0.17	-0.57	0.61	-0.58	SNP5	●	●
0.45	-0.35	0.98	-0.37	0.62	SNP6	●
0.7	-0.56	0.62	-0.57	0.34	0.63	SNP7

$\widehat{R} := [\hat{r}_{ij}]_{p \times p}$, $\hat{r}_{ij}$: LD between SNP i and j

Predefined gene sets

Gene	Chr	Start	End
BRAF	7	140419127	140624564
MAPK1	22	22108789	22221970
MAPK3	16	30125426	30134827
NFKB1	4	103422486	103538459

$a_j := 1\,\{\text{SNP } j \text{ is “near” a gene in the set}\}$

Bayesian Model

Likelihood function

$$L(\beta) := \text{Normal}\left(\widehat{\beta};\ \widehat{S}\widehat{R}\widehat{S}^{-1}\beta,\ \widehat{S}\widehat{R}\widehat{S}\right)$$

Prior distribution

$$\beta_j \sim \pi_j \cdot \text{Normal}(0, \sigma_\beta^2) + (1 - \pi_j)\delta_0$$

Baseline model (*M0*)

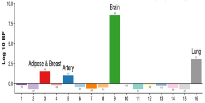
$$\log_{10}\left(\frac{\pi_j}{1 - \pi_j}\right) = \theta_0$$

Enrichment model (*M1*)

$$\log_{10}\left(\frac{\pi_j}{1 - \pi_j}\right) = \theta_0 + a_j\theta$$

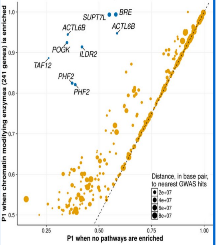
Inference

Gene set enrichment



$$\text{BF(gene set)} = \frac{\Pr(\text{Data} \mid M_1)}{\Pr(\text{Data} \mid M_0)}$$

Gene prioritization



$p(\beta \mid \text{Data},\ M_1)$ vs $p(\beta \mid \text{Data},\ M_0)$

We apply the method to analyze 31 complex traits and 4,026 gene sets.

This application is not small:

# of Parameters = 31 × (3,913+113) × 1.1 Million ≈ 137 Billion

- **31** human phenotypes
- **3,913** biological pathways
- **113** tissue-based gene sets
- **1.1 million** common SNPs

One student can get this done, aided by:

- Publicly available input data (GWAS & LD & gene sets)
- Variational inference; Squared extrapolation method (SQUAREM)
- Parallel computing; Hierarchical data format (HDF5)

Our full results are publicly available.

- **Results**

  `https://xiangzhu.github.io/rss-gsea/results`

- **Software**

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

- **Demonstration**

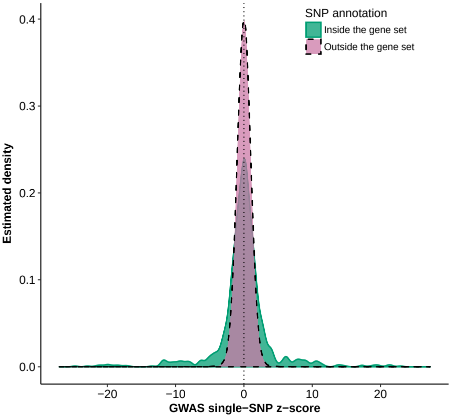
  `https://stephenslab.github.io/rss/Example-5`

- **R package (in progress)**

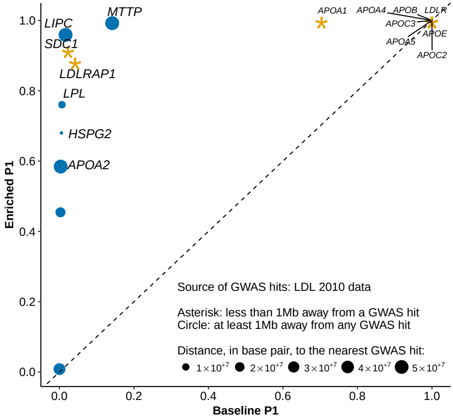
  `https://github.com/stephenslab/rssr` (N. Knoblauch)

Example: Low-density lipoprotein & *MTTP*

Chylomicron−mediated lipid transport pathway enrichment log10 Bayes factor: 65.5



Estimated P1 for 17 member genes of chylomicron−mediated lipid transport pathway



Source of GWAS hits: LDL 2010 data

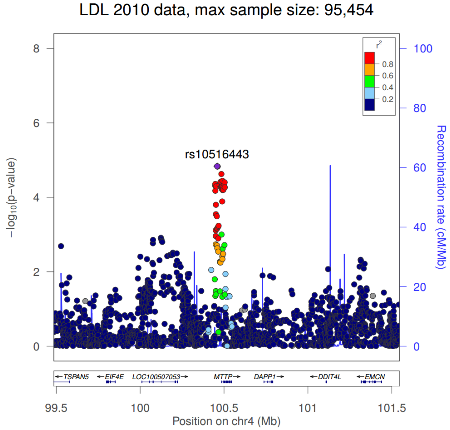
Asterisk: less than 1Mb away from a GWAS hit
Circle: at least 1Mb away from any GWAS hit

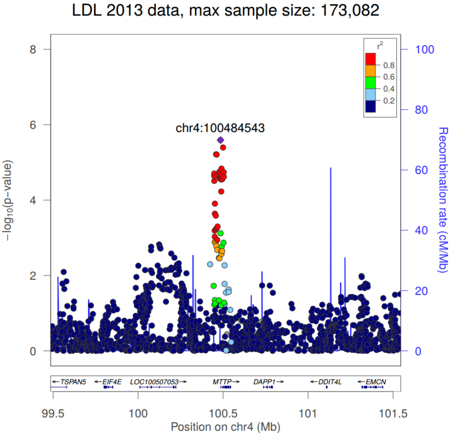
Distance, in base pair, to the nearest GWAS hit:

- $P_1 := 1 - \text{Prob}(\beta_j = 0, \forall \text{ SNP } j \in \text{locus} \mid \text{Data})$
- ***MTTP***: baseline $P_1 = \mathbf{0.14}$ vs enriched $P_1 = \mathbf{0.99}$

Example: Low-density lipoprotein & *MTTP*

- Total sample size: **95,454** (2010) $\longrightarrow$ **173,082** (2013)
- GWAS p-value: $\mathbf{1.5} \times \mathbf{10^{-5}}$ (2010) $\longrightarrow$ $\mathbf{2.6} \times \mathbf{10^{-6}}$ (2013) “<” 5×10^{-8}

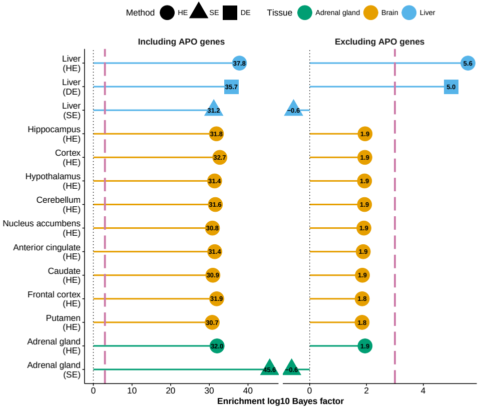

LDL 2010 data, max sample size: 95,454


LDL 2013 data, max sample size: 173,082

References: Global Lipids Genetics Consortium (2013); Teslovich *et al.* (2010).

Example: Alzheimer’s disease & Liver

- Tissue-based gene sets
  - HE: highly expressed
  - SE: selectively expressed
  - DE: distinctively expressed
- **GTEx RNA-seq data**
- Top-enriched tissues
  - Adrenal gland
  - Brain
  - **Liver** (even w/o APO)
- Non-APO, liver gene: ***TTR***
  - baseline $P_1 = \mathbf{0.64}$
  - enriched $P_1 = \mathbf{1.00}$



References: Xi *et al.* (2017+); Dey *et al.* (2017); The GTEx Consortium (2015); Lambert *et al.* (2013).

What’s next?

We develop a new enrichment analysis method that:

- uses **publicly available** data as input
- **efficiently** assesses thousands of gene sets
- **automatically** identifies trait-associated genes

Future work:

- one annotation at a time → jointly analyze **many** annotations
- gene-based annotations → **“finer-scale”** annotations
- “spike-and-slab prior” → more **flexible** effect size distributions

Acknowledgements

- Joint work with **Matthew Stephens**
- **GWAS Consortia**

  Teslovich *et al*. (2010); Manning *et al*. (2012); Morris *et al*. (2012); van der Harst *et al*. (2012); Köttgen *et al*. (2013); Lambert *et al*. (2013); Okada *et al*. (2014); Ripke *et al*. (2014); Wood *et al*. (2014); Day *et al*. (2015); Liu *et al*. (2015); Locke *et al*. (2015); Nikpay *et al*. (2015); Shungin *et al*. (2015); Okbay *et al*. (2016); van Rheenen *et al*. (2016)
- **GTEx Consortium**

  `https://gtexportal.org`

GORDON AND BETTY MOORE FOUNDATION

NIH National Institutes of Health

THE UNIVERSITY OF CHICAGO RESEARCH COMPUTING CENTER

Thank you!

- Preprint: `https://doi.org/10.1101/160770`
- Software: `https://github.com/stephenslab/rss`
- Contact: `xiangzhu[at]uchicago[dot]edu`