Large-scale genome-wide enrichment analysis of 31 human phenotypes

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

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

THE UNIVERSITY OF CHICAGO

Examining associations between variables is a useful tool.

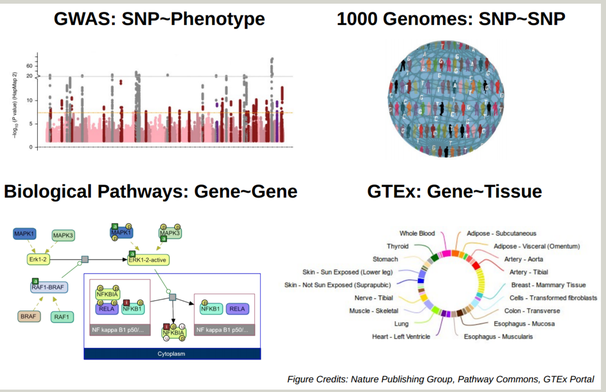


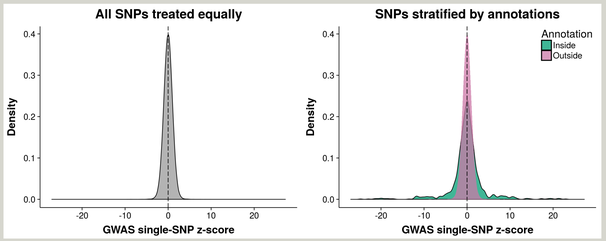
Figure Credits: Nature Publishing Group, Pathway Commons, GTEx Portal

Enrichment analysis combines multiple sources of association.

- SNP-Trait: genome-wide association study (GWAS)
- SNP-SNP: linkage disequilibrium (LD)
- Gene-Gene: biological pathways (e.g. Pathway Commons)
- Gene-Tissue: RNA-seq from different tissue samples (e.g. GTEx)

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)

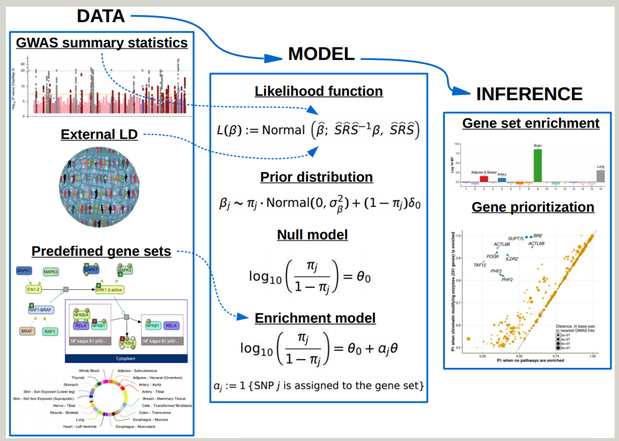


Recent reviews: de Leeuw *et al.* (2016); Pers (2016); Mooney *et al.* (2014); Wang *et al.* (2010).

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

- **Issue 1**
  If the gene set is truly enriched, we should relax significance threshold for "green" SNPs, but how much to relax?
- **Issue 2**
  The "inflated" pattern of green curve may be driven by correlation between SNPs, rather than enrichment of signal.

We develop a statistical method that systematically utilizes enrichment information.



Idea 1 ⇝ Issue 1: Learning enrichment from data

Model-based approach:

- Assume that SNP j is "causal" with probability π_j
- Represent π_j as a function of SNP j's annotation a_j

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

- Estimate enrichment parameter θ from data

Data-adaptive threshold:

- Enrichment data ⇝ large θ ⇝ large π_j ⇝ increased power
- Null data ⇝ $\theta \approx 0$ ⇝ unchanged π_j ⇝ maintained type 1 error

Reference: Carbonetto and Stephens (2013)

Idea 2 ⇝ Issue 2: Modeling linkage disequilibrium

Single-SNP summary data:

$$\hat{\beta}_j := (X_j^\intercal X_j)^{-1} X_j^\intercal y$$
$$\hat{\sigma}_j^2 := (nX_j^\intercal X_j)^{-1}(y - X_j\hat{\beta}_j)^\intercal(y - X_j\hat{\beta}_j)$$

- y: phenotype of n individuals
- X_j: genotype of n individuals at SNP j

Multiple-SNP likelihood function:

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

- multiple-SNP parameter: $\beta := (\beta_1, \ldots, \beta_p)^\intercal$
- single-SNP summary data: $\hat{\beta} := (\hat{\beta}_1, \ldots, \hat{\beta}_p)^\intercal$
- $\hat{S} := \text{diag}(\hat{s})$, $\hat{s} := (\hat{s}_1, \ldots, \hat{s}_p)^\intercal$, $\hat{s}_j^2 := \hat{\sigma}_j^2 + n^{-1}\hat{\beta}_j^2$
- $\hat{R}$: the shrinkage estimate of LD (Wen and Stephens, 2010)

Reference: Zhu and Stephens (2017)

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

This application is not small:

Total number of parameters in our analyses:

31 × (3,913+113) × 1.1 Million ≈ 137 Billion

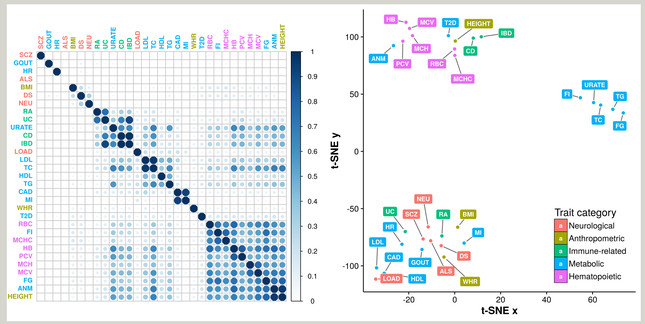
- **31** human phenotypes
- **3,913** biological pathways
- **113** tissue-based gene sets
- **1.1 million** common SNPs

One student can get this done:

- Publicly available summary data
- Variational Bayes algorithms
- Banded matrix approximation
- Parallel computing
- Hierarchical data format (HDF5)
- **High-performance computing at RCC**

We make our full analysis results publicly available.

- **Results:** `http://xiangzhu.github.io/rss-gsea/results`
- **Software:** `https://github.com/stephenslab/rss`



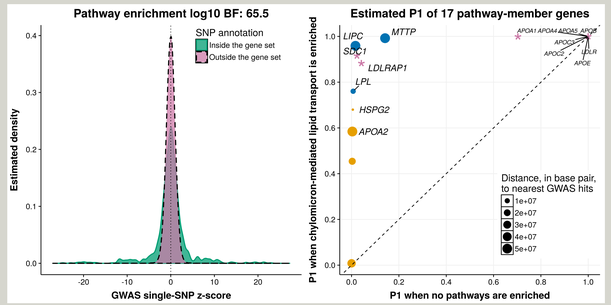
Acknowledgements

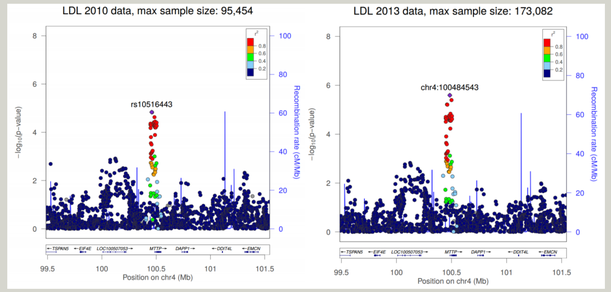
Discussions: Peter Carbonetto, Xin He
Data: Michael Turchin, Kushal Dey, Carl Anderson, John Perry, Ruth Loos, Marcel den Hoed, Simon Xi

NIH National Institutes of Health
GORDON AND BETTY MOORE FOUNDATION
THE UNIVERSITY OF CHICAGO RESEARCH COMPUTING CENTER

Our analyses yield new insights into complex human traits.

Example 1: Low-density lipoprotein & *MTTP* gene





Example 2: Alzheimer's disease & Liver