Large-scale genome-wide enrichment analysis of 31 human phenotypes
Xiang Zhu¹ and Matthew Stephens¹,²
¹Department of Statistics, ²Department of Human Genetics

Exercising associations between variables is a useful tool.

We develop a statistical method that systematically utilizes enrichment information.

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.

Our analyses yield new insights into complex human traits.

Acknowledgements

Example 1:
Low-density lipoprotein & MTTP gene

Example 2:
Alzheimer’s disease & Liver

Examining associations between variables is a useful tool.

We develop a statistical method that systematically utilizes enrichment information.

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.

Our analyses yield new insights into complex human traits.

Acknowledgements

Example 1:
Low-density lipoprotein & MTTP gene

Example 2:
Alzheimer’s disease & Liver

There are a few things that we noticed.

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)

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.

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)

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

Idea 1 → Issue 1:
Learning enrichment from data

Model-based approach:
- Assume that SNP j is “causal” with probability πj
- Represent πj as a function of SNP j’s annotation aj

log10(πj/(1−πj)) := θ0 + ajθ

Data-adaptive threshold:
- Enrichment data → large θ → large πj → increased power
- Null data → θ ≈ 0 → unchanged πj → maintained type 1 error

Reference: Carbonetto and Stephens (2013)

Idea 2 → Issue 2:
Modeling linkage disequilibrium

Single-SNP summary data:
- βj := (X⊺WX)−1X⊺Wy
- σj2 := (X⊺WX)−1y⊺X(1−X⊺X)Xy

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

Multiple-SNP likelihood function:

LRSS(β, S, R) := Normal(β, S−1R)−1β, SRS−1β, SRS

- multiple-SNP parameter: β := (β1, ..., βp)T
- single-SNP summary data: β := (β1, ..., βp)T
- S := diag(Sj), Sj := (σ1j, ..., σpj, σ2j−1, ..., σpj−1, σpj)
- R := the shrinkage estimate of LD (Wen and Stephens, 2010)