Bayesian variant-based pathway enrichment analysis using GWAS summary statistics
Xiang Zhu\textsuperscript{1} and Matthew Stephens\textsuperscript{1,2}
\textsuperscript{1}Department of Statistics, \textsuperscript{2}Department of Human Genetics

Can pathway enrichment and variant prioritization be integrated?

Carbone et al. [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/stephenslab/rsspathway.

A novel integrated approach

The GWAS full data are genotypes $X \equiv (x_{ij}, y_{ij})$ and phenotypes $y = (y_1, \ldots, y_n)$ from $n$ individuals.

For continuous traits, linear regression is used:

$$ y_i \sim \mathcal{N}(x_i \beta, \sigma^2) $$

For binary traits, logistic regression is used:

$$ y_i \sim \text{Bernoulli}(x_i \beta) $$

The multiple-SNP effect of $p$ SNPs, $\beta = (\beta_1, \ldots, \beta_p)$, has prior

$$ \beta \sim (\beta_1, \ldots, \beta_p) \sim \mathcal{N}(0, \Sigma_p) $$

The enrichment of associations within a pathway is modeled as

$$ \log_{10}(\text{OR})^* = \beta_j $$

where $j = 1, \ldots, p$ is the log-fold enrichment and $\beta_j$ is the genome-wide log-odds.

Notations: $a = (a_1, \ldots, a_p)$ and $\gamma = (\gamma_1, \ldots, \gamma_p)$, $\gamma_j = 1$ if $\beta_j \neq 0$.

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.

Regression with Summary Statistics (RSS) provides a solution.

We propose the following regression model for GWAS summary statistics:

$$ \beta | \text{R}, \beta \sim \mathcal{N}(\text{RSS} \times \beta, \text{RSS}) $$

- $\beta = (\beta_1, \ldots, \beta_p)$, where $\beta_j$ is the single-SNP effect estimate of SNP $j$.

- $S = \text{diag}(s)$, $s = (s_1, \ldots, s_p)$, where $s_j$ is the standard error of $\hat{\beta}_j$.

- $\text{R}$ 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

$$ p(\beta | \text{R}, \beta \sim \mathcal{N}(\text{RSS} \times \beta, \text{RSS}) $$

The strength of pathway enrichment is measured by

$$ \text{SBF}(\gamma) = \frac{\sum_i \beta_i \gamma_i}{\sum_i \beta_i^2} $$

The log-fold enrichment level is estimated from

$$ \log(\text{OR})^* = \beta_j, \text{R} \sim \mathcal{N}(\text{RSS} \times \beta, \text{RSS}) $$

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

$$ \text{SBF}(\gamma) = \frac{\sum_i \beta_i \gamma_i}{\sum_i \beta_i^2} $$

Estimate $\beta(\gamma; \text{R}, \beta \sim \mathcal{N}(\text{RSS} \times \beta, \text{RSS})$ given $\gamma(\theta)$

Decomposition of marginal likelihood:

$$ \log p(\text{D} | \theta, \beta \sim \mathcal{N}(\text{RSS} \times \beta, \text{RSS}) $$

The iterative scheme for obtaining $\beta(\gamma; \theta, \beta(\gamma; \theta))$ is

$$ \beta_j = (\frac{1}{\gamma_j} + \frac{1}{\gamma_j})^{-1} - \frac{\sum_j \beta_i \gamma_i}{\sum_j \beta^2} $$

We are investigating more efficient approaches.

Parallel implementation

$R_j = 0$ if SNP $j$ and $s_j$ are on different chromosomes.

• Iterative update of $\gamma^*$ only requires data from SNP $j$ that $R_j \neq 0$.

• $\text{SBF}(\gamma^*; \beta(\gamma^*; \beta))$ only uses data from Chromosome $c$.

References


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 (Reactions [4]) retrieved from BioSystems [5].

• Type 1 error

Power

Estimation

Enrichment

Adult height

We applied RSS on 3700 curated pathways and GWAS summary statistics of 1.06 million SNPs for adult human height from 253,281 individuals of European (EUR) ancestry [14]. The population LD matrix was estimated from the 1000 Genomes [15] EUR samples.

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

Pathway

Source

Database

# of genes

Hedgehog signaling pathway

Wiki

RSDB

50

Hedgehog signaling pathway

KEGG

RSDB

51

Basal cell carcinoma

KEGG

RSDB

55

Hedgehog signaling

Reactome

RSDB

84

RAC signaling pathway

PID

RSDB

54

Rho cell motility signaling pathway

BC

RSDB

52

Signaling by Hedgehog

Reactome

BC

125

Regulatory role of PKC subunit zeta

BC

16

Y branching of actin filaments

BC

10

Pathogenic Euchromatic cell cycle

KEGG

RSDB

55

Pathogenic Euchromatic cell cycle

KEGG

RSDB

56

How progranulin initiates osteocyte membrane

BC

33

RhoC-GEF activity

RAPTOR

PANTHER

90

Cytokine-regulated by Rho-GEFs

PANTHER

70

Signaling events mediated by the Hedgehog family

PID

RSDB

22

Signaling events mediated by the Hedgehog family

PID

RSDB

23

EPIN-mediated forward signaling

PID

RSDB

41

Ligand-receptor interactions

Reactome

RSDB

8

Software

Software of using RSS for integrated enrichment analysis, RSSpathway, is available from https://github.com/stephenslab/rsspathway.