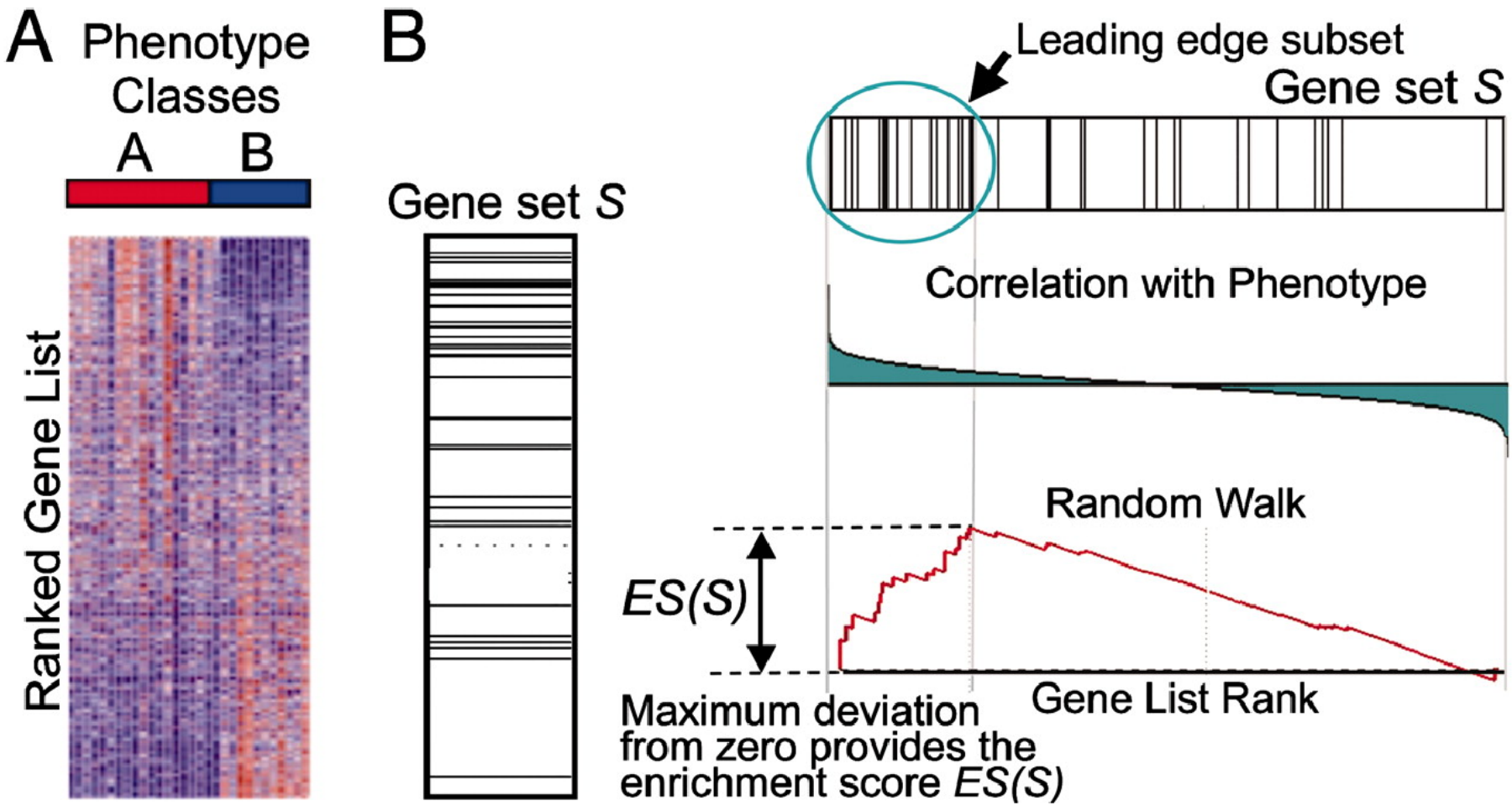
Bayesian variant-based gene set enrichment analysis using GWAS summary statistics

Xiang Zhu
University of Chicago

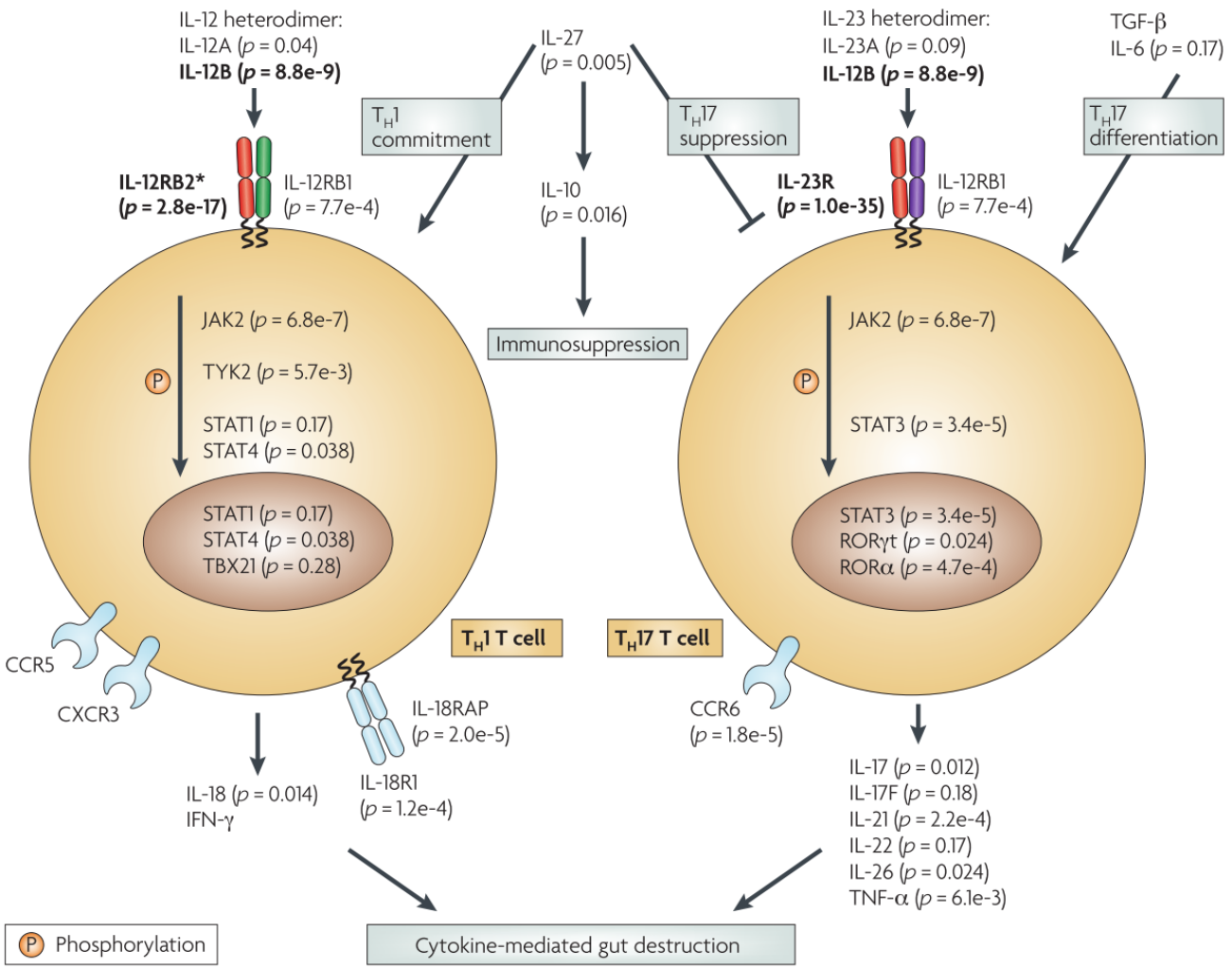
March 3, 2016

What is Gene Set Enrichment Analysis? (GSEA)



Aravind Subramanian et al. PNAS 2005;102:15545-15550

Similar ideas can be adopted in GWAS.



Wang et al. Nature Reviews Genetics 2010; 11; 843-854

GSEA illustrates the importance of combining information.

- **GWAS + Pathway** (this talk)
  Two reviews: Wang et al (2010); Mooney et al (2014)
- GWAS + Functional Annotation
  Pickrell (2014); Finucane et al (2015) ... ...
- GWAS + eQTL
  Nicolae et al (2010); He et al (2013) ... ...
- eQTL + Functional Annotation
  Veyrieras et al (2008); Wen et al (2015) ... ...
- ... ...

Most pathway approaches to GWAS do not tell us which genes are relevant.

Real examples

Name	Source	Database	# of Genes
Disease	Reactome	Pathway Commons	1206
Gene Expression	Reactome	NCBI BioSystems	1213
Homo sapiens	miRTarBase	Pathway Commons	11343

Two-stage process in Carbonetto and Stephens (2013):

1. Identify pathway enrichment
2. Prioritize variants within the enriched pathways

Software: BMApathway, `https://github.com/pcarbo/bmapathway`

Most pathway approaches to GWAS do not tell us which genes are relevant.

Real examples

Name	Source	Database	# of Genes
Disease	Reactome	Pathway Commons	1206
Gene Expression	Reactome	NCBI BioSystems	1213
Homo sapiens	miRTarBase	Pathway Commons	11343

Two-stage process in Carbonetto and Stephens (2013):

1. Identify pathway enrichment
2. Prioritize variants within the enriched pathways

Software: BMApathway, `https://github.com/pcarbo/bmapathway`

Carbonetto and Stephens (2013) propose an integrated approach.

- Likelihood
  - Continuous (e.g. height):

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

  - Binary (e.g. diabetes):

$$y_i|\boldsymbol{x}_i, \beta \sim \text{Bernoulli}(\eta(\boldsymbol{x}_i, \beta)), \ \text{logit}(\eta(\boldsymbol{x}_i, \beta)) = \beta_0 + \boldsymbol{x}_i^\intercal \beta$$

- Prior
  - effect size distribution:

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

  - probability of being “causal”:

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

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

BMApathway is complicated by the access to full GWAS data.

GWAS individual-level data can be hard to obtain.

GWAS summary statistics are widely available.

nature genetics

Asking for more

Because of the usefulness of genome-wide association study (GWAS) data for mapping regulatory variation in the human genome, the journal now asks authors to report the co-location of trait-associated variants with gene regulatory elements identified by epigenetic, functional and conservation criteria. We also ask that authors publish or database the genotype frequencies or association *P* values for all SNPs investigated, whether or not they reached genome-wide significance.

Can the integrated method be applied to GWAS summary data?

- The method in Carbonetto & Stephens (2013):

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

- A possible modification?

  $p(\text{Param}|\text{Summary Data}) \propto p(\text{Summary Data}|\text{Param}) \cdot p(\text{Param})$

Regression with Summary Statistics (RSS) provides a solution.

$$\widehat{\beta}|S,R,\beta \sim N(SRS^{-1}\beta, SRS)$$

- **single**-SNP data: $\widehat{\beta} := (\hat{\beta}_1, \ldots, \hat{\beta}_p)^\intercal$
- **multiple**-SNP parameter: $\beta := (\beta_1, \ldots, \beta_p)^\intercal$
- plug in $\{\widehat{S}, \widehat{R}\}$ for $\{S, R\}$:
  - $\widehat{S} := \text{diag}(\widehat{\boldsymbol{s}})$, $\widehat{\boldsymbol{s}} := (\hat{s}_1, \ldots, \hat{s}_p)^\intercal$, $\hat{s}_j \approx \text{se}(\hat{\beta}_j)$;
  - $\widehat{R}$: the estimated LD matrix [Wen & Stephens (2010)]

Regression with Summary Statistics (RSS) provides a solution.

- Likelihood:

$$\widehat{\beta}|S,R,\beta \sim N(SRS^{-1}\beta, SRS)$$

- Prior
  - effect size distribution:

$$\beta_j|S,\theta_0,\theta,h \sim (1-\pi_j)\delta_0 + \pi_j N(0,\sigma_B^2)$$

  - probability of being “causal”:

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

  - variance of causal effect:

$$\sigma_B^2 := h \cdot \left(\sum_{j=1}^{p} \pi_j n^{-1} s_j^{-2}\right)^{-1}$$

RSS retains the integrated characteristics of BMApathway.

- Test the pathway enrichment:

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

- Estimate the enrichment level:

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

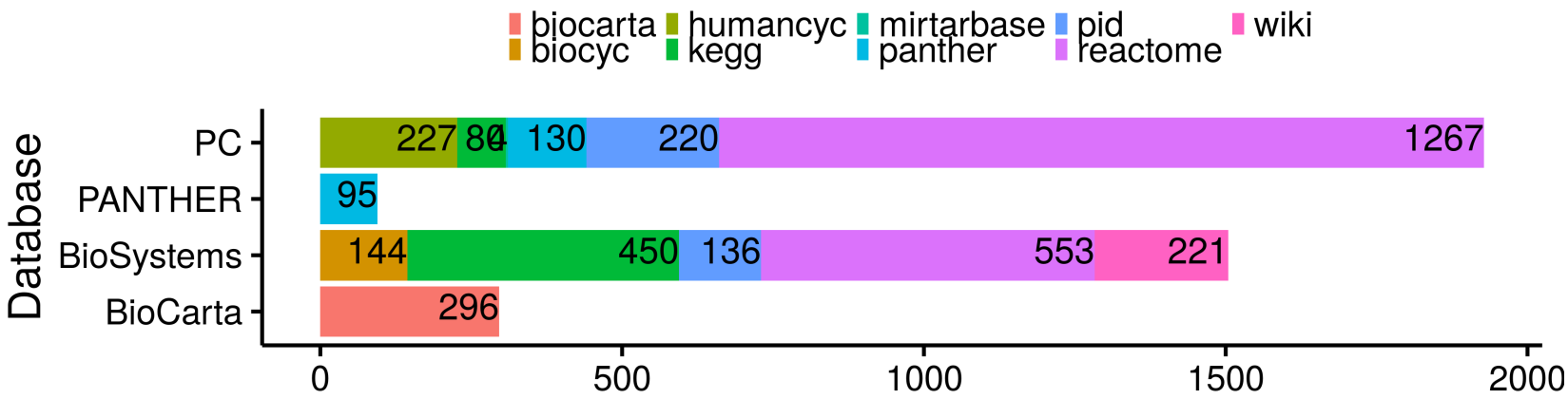
- Estimate the effect of SNP j given the enrichment:

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

where $\boldsymbol{a} := (a_1, \ldots, a_p)^\intercal$, $a_j := 1\{\text{SNP } j \text{ is in the pathway}\}$.

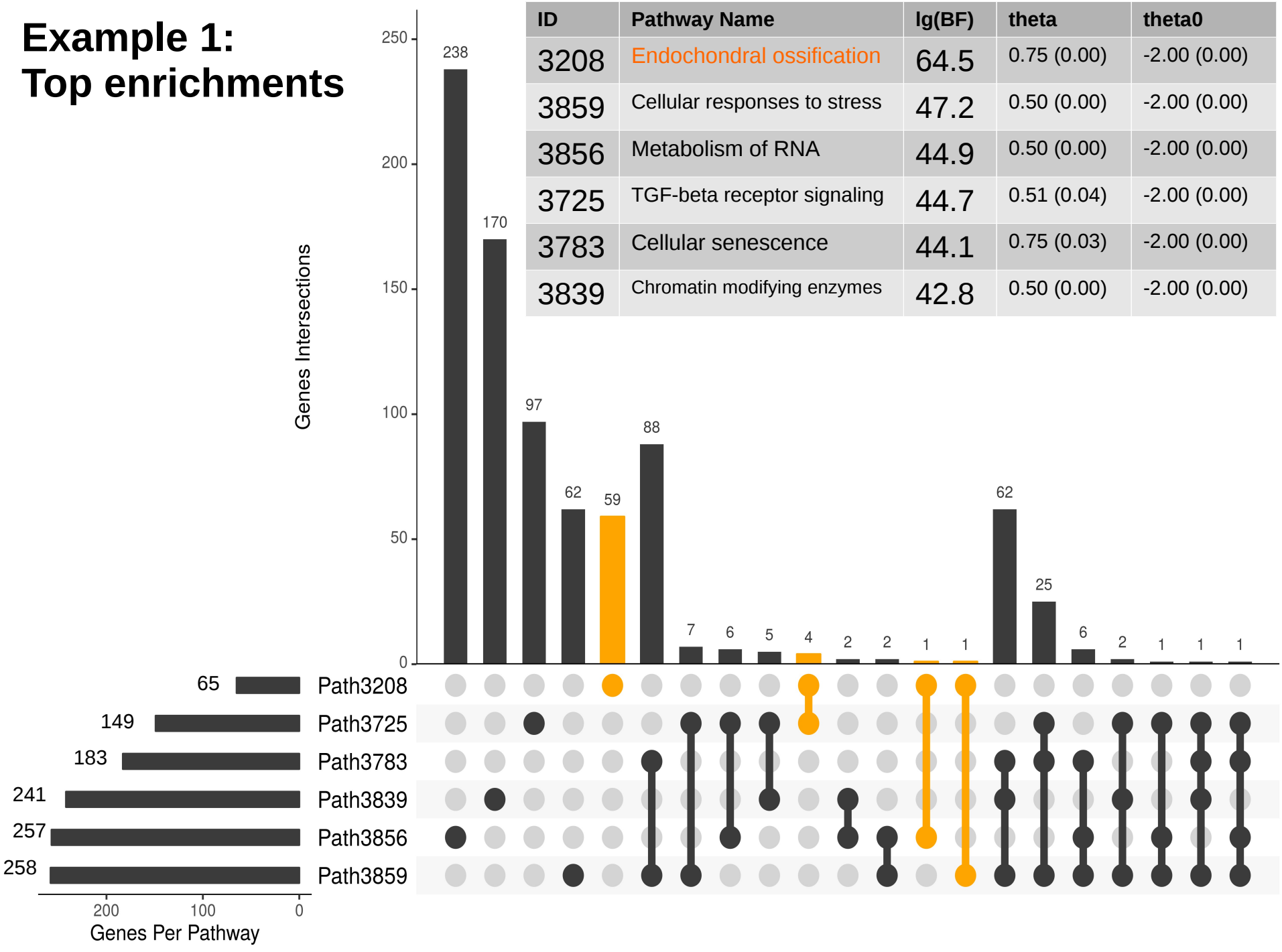
We apply RSS to a GWAS of adult human height.

- GWAS summary statistics (Wood et al, 2014)
  1,064,575 SNPs, **253,288** EUR individuals
- Reference genotypes (1000 Genomes Project, 2010)
  phased haplotypes from 379 EUR individuals
- **18,732** protein-coding genes on autosomes
- **3,823** curated biological pathways

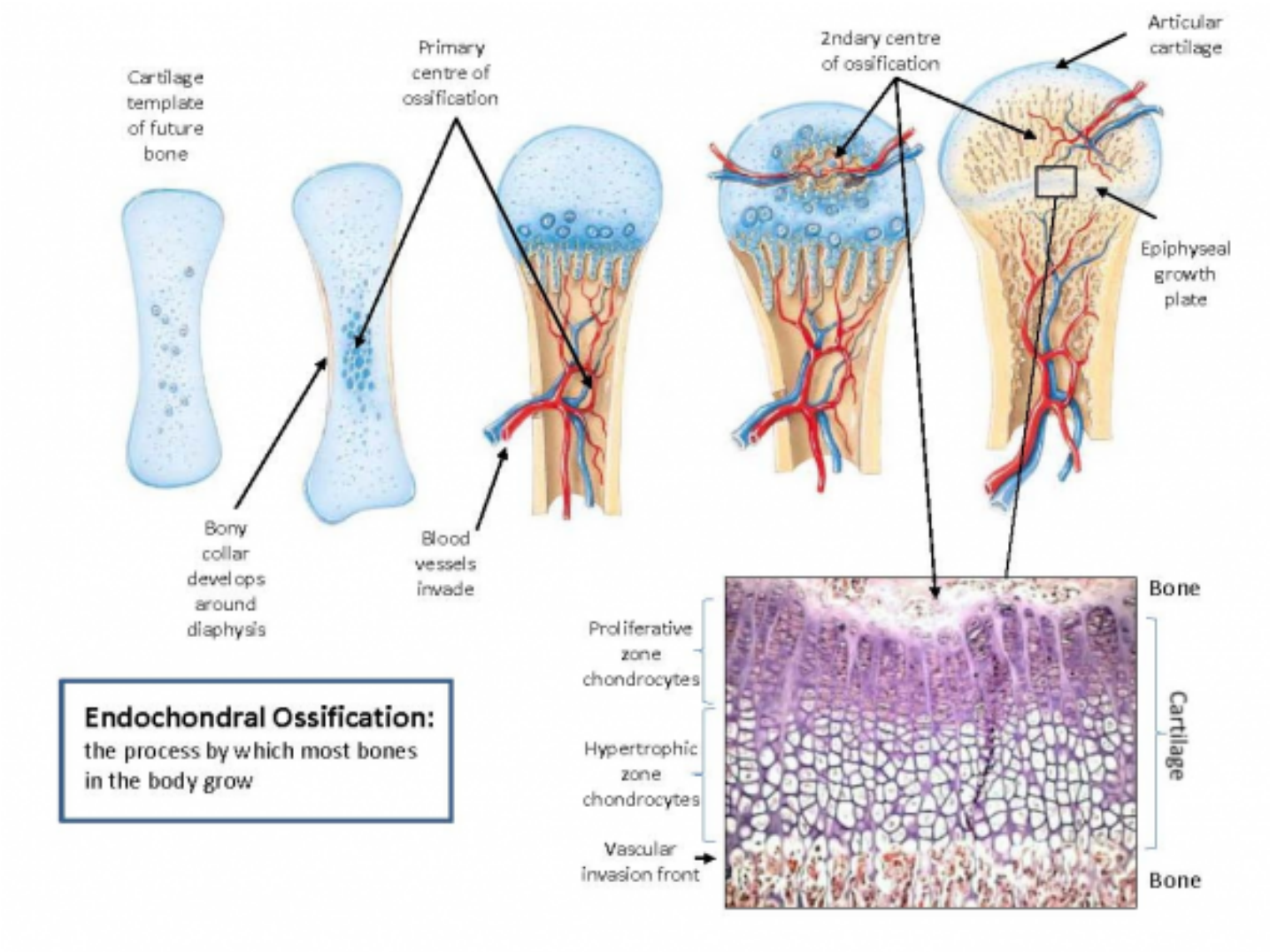


Example 1: Top enrichments

ID	Pathway Name	lg(BF)	theta	theta0
3208	Endochondral ossification	64.5	0.75 (0.00)	-2.00 (0.00)
3859	Cellular responses to stress	47.2	0.50 (0.00)	-2.00 (0.00)
3856	Metabolism of RNA	44.9	0.50 (0.00)	-2.00 (0.00)
3725	TGF-beta receptor signaling	44.7	0.51 (0.04)	-2.00 (0.00)
3783	Cellular senescence	44.1	0.75 (0.03)	-2.00 (0.00)
3839	Chromatin modifying enzymes	42.8	0.50 (0.00)	-2.00 (0.00)



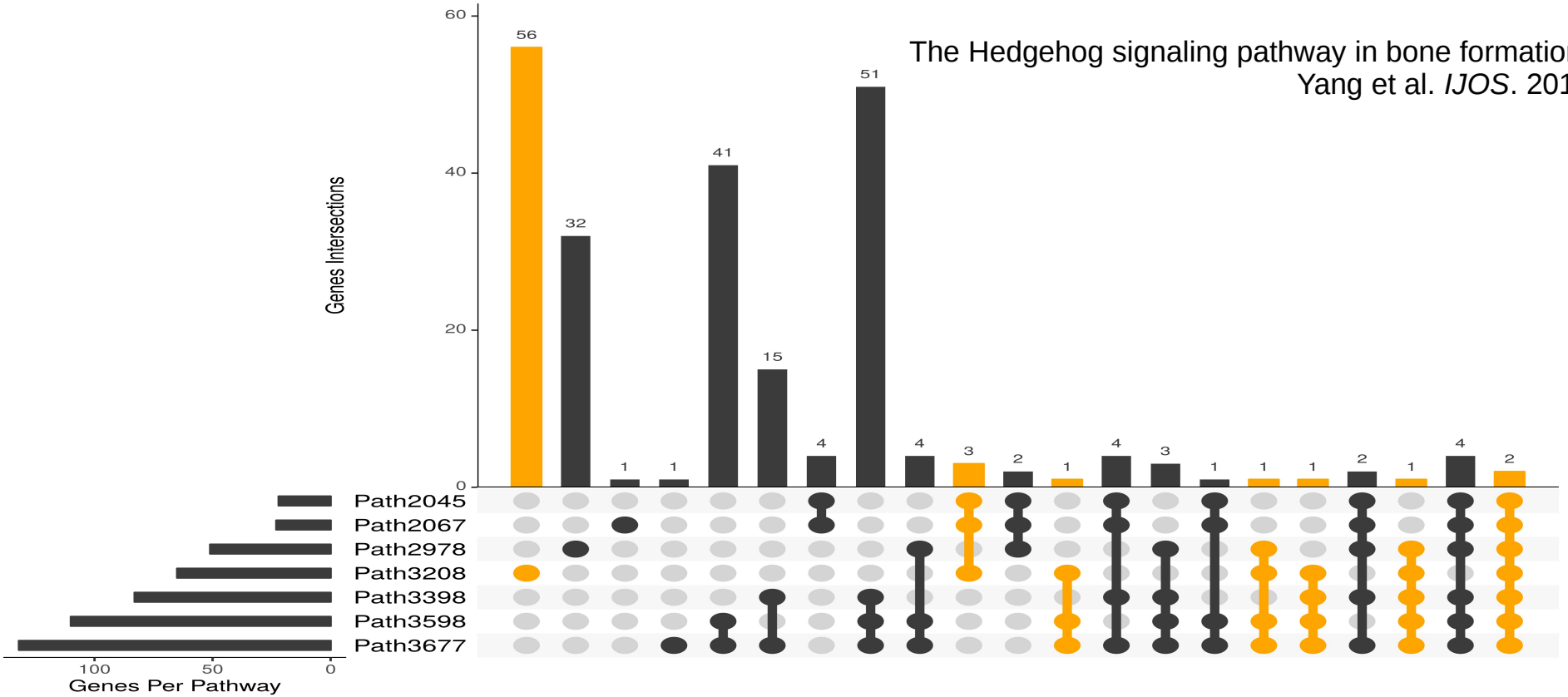
What is Endochondral Ossification?



Source: `http://www.wellcome-matrix.org/research_groups/ray-boot-handford.html`

Example 2: Hedgehog Signaling

ID	Pathway Name	Pathway Database	lg(BF)	theta	theta0
3208	Endochondral ossification	Wiki, BioSystems	64.5	0.75 (0.00)	-2.00 (0.00)
3398	Hedgehog 'on' state	Reactome, BioSystems	34.5	0.75 (0.01)	-2.00 (0.00)
3677	Signaling by Hedgehog	Reactome, BioSystems	32.3	0.50 (0.00)	-2.00 (0.00)
2978	Hedgehog signaling	KEGG, BioSystems	30.6	0.75 (0.00)	-2.00 (0.00)
2067	Signaling events mediated by the Hedgehog family	PID, PC	23.8	0.75 (0.03)	-2.00 (0.00)
2045	Signaling events mediated by the Hedgehog family	PID, BioSystems	23.7	0.75 (0.03)	-2.00 (0.00)
3598	Signaling by Hedgehog	Reactome, PC	22.3	0.50 (0.00)	-2.00 (0.00)

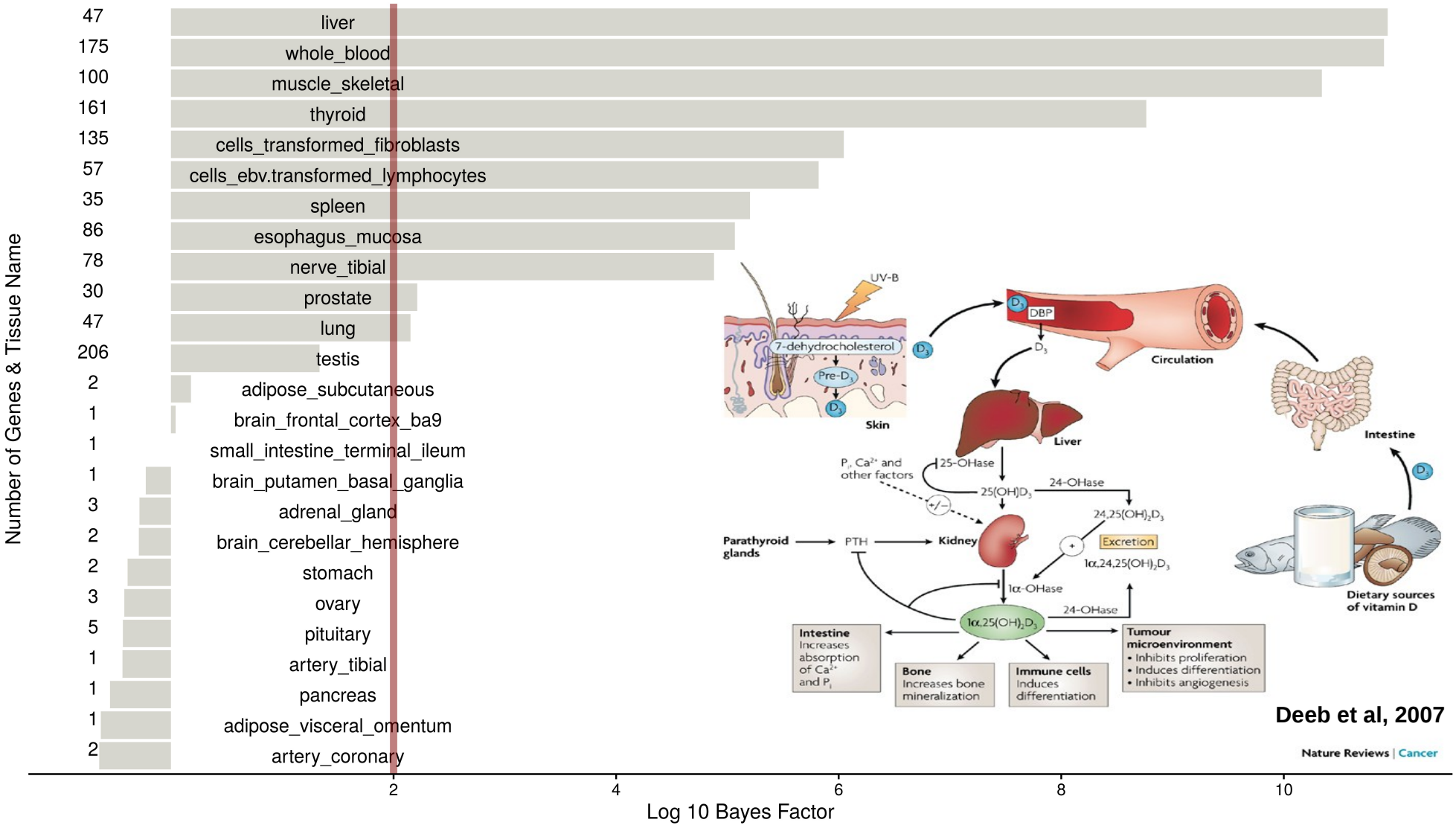


The Hedgehog signaling pathway in bone formation.
Yang et al. *IJOS*. 2015

Example 3:
GTEx eQTL Genes

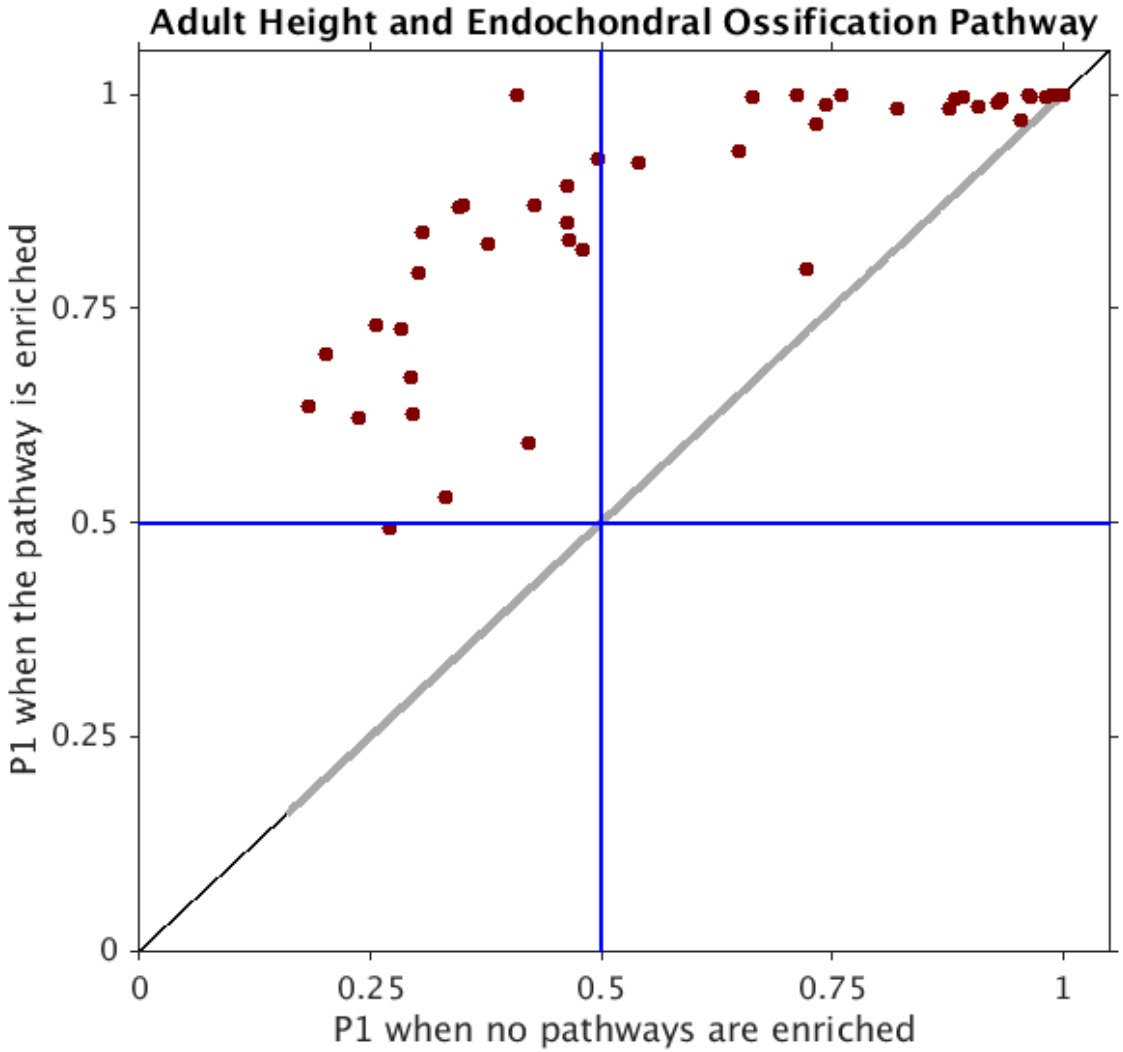
Liver:	log10(BF)=10.9,	theta=0.50(0.00), theta0=-2.00(0.00)
Whole Blood:	log10(BF)=10.9,	theta=0.25(0.00), theta0=-2.00(0.00)
Muscle Skeletal:	log10(BF)=10.3,	theta=0.50(0.02), theta0=-2.00(0.00)
Thyroid:	log10(BF)=8.8,	theta=0.25(0.00), theta0=-2.00(0.00)

Enrichment = Tissue + eQTL ?



GTEx tissue-specific eQTL gene lists courtesy of S. Urbut.

Additional associations can be informed by enriched pathways.



Acknowledgements

- those who help me conceive the study
  Matthew Stephens, Peter Carbonetto
- those who teach me genetics/biology
  Xin He, Nicholas Knoblauch, Siming Zhao, Sarah Urbut
- those who make their data public
  GIANT, 1000 Genomes, Pathway Commons, BioSystems, GTEx
- University of Chicago Research Computing Center

Appendix

- Appendix A: Variational Bayes algorithms
- Appendix B: Other applications of RSS
- Appendix C: Simulations

Appendix A: Variational Bayes algorithms

References:

- Bishop, C. *Pattern Recognition and Machine Learning*, Ch. 10
- Ormerod, J.T. & Wand, M.P. *Am. Stat.* 2010

Our posterior calculation exploits variational approximation.

- $\gamma := (\gamma_1, \ldots, \gamma_p)^\intercal$, where $\gamma_j = 1$ if SNP j is causal
- $D := \{\widehat{\beta}, S, R, \boldsymbol{a}\}$

Posterior distribution:

$$p(\beta, \gamma | D) = \int p(\beta, \gamma | D, \theta_0, \theta, h) p(\theta_0, \theta, h | D) \mathrm{d}\theta_0 \mathrm{d}\theta \mathrm{d}h$$

1. Estimate $p(\beta, \gamma | D, \theta_0, \theta, h)$
2. Estimate $p(\theta_0, \theta, h | D)$

Step 1: Estimate $p(\beta, \gamma | D, \theta_0, \theta, h)$

Aim: approximate $p(\beta, \gamma | D, \theta_0, \theta, h)$ by $q^\star(\beta, \gamma)$

- Decomposition of marginal likelihood:

$$\log p(D|\theta_0, \theta, h) = \underbrace{\mathrm{E}_q \log\left[\frac{q(\beta,\gamma)}{p(\beta,\gamma|D,\theta_0,\theta,h)}\right]}_{\text{Kullback-Leibler (KL) divergence}} + \underbrace{\mathrm{E}_q \log\left[\frac{p(D,\beta,\gamma|\theta_0,\theta,h)}{q(\beta,\gamma)}\right]}_{\text{Evidence lower bound (LB)}}$$

- Optimization over distributions:

$$q^\star = \arg\min_q \mathrm{KL}(q; \theta_0, \theta, h) = \arg\max_q \mathrm{LB}(q; \theta_0, \theta, h)$$

- Mean field approximation:

$$q(\beta, \gamma) = \prod_{j=1}^{p} q_j(\beta_j, \gamma_j)$$

Step 2: Estimate $p(\theta_0, \theta, h|D)$

- Put uniform (**grid**) prior on $\{\theta_0, \theta, h\}$:

$$p(\theta_0) \propto 1,\ \ p(\theta) \propto 1,\ \ p(h) \propto 1$$

- Approximate $p(\theta_0, \theta, h|D)$:

$$p(\theta_0, \theta, h|D) \propto \exp\{\mathrm{LB}(\theta_0, \theta, h)\}$$

- What about the exact calculation?

$$p(\theta_0, \theta, h|D) \propto p(D|\theta_0, \theta, h)$$

The estimate is discrete.

Appendix B: Other applications of RSS

Other features:

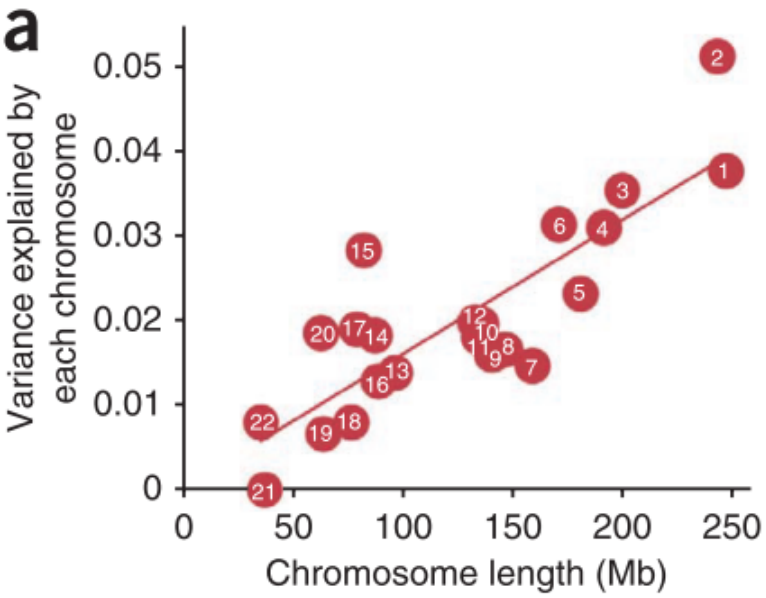
- Estimate SNP heritability (PVE) [ready to use]
- Detect genetic association [ready to use]
- Predict phenotype [in progress]

  `https://github.com/xiangzhu/dscr_blm`

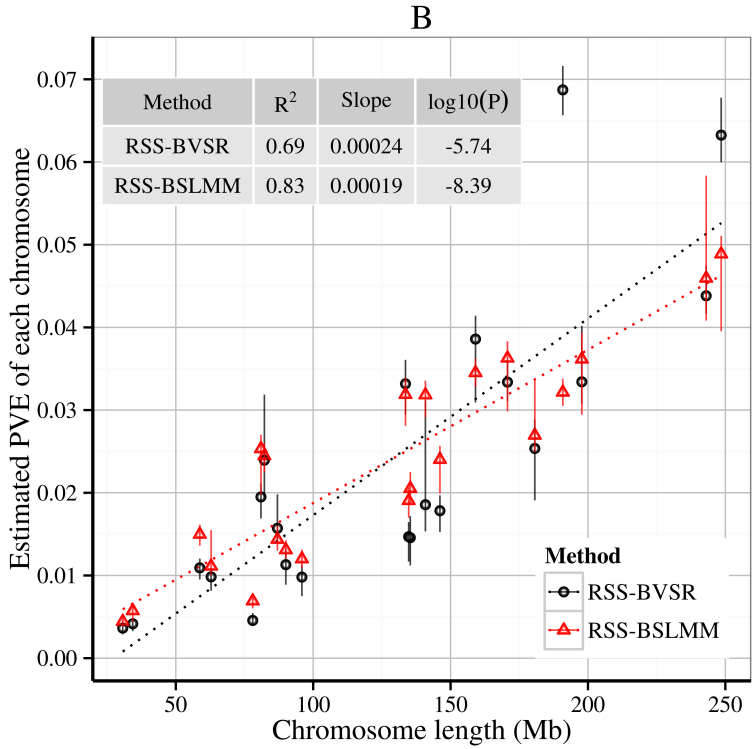
For more details:

- Manuscript: will be posted on bioRxiv soon
- Software: `https://github.com/stephenslab/rss`

RSS can estimate SNP heritability.



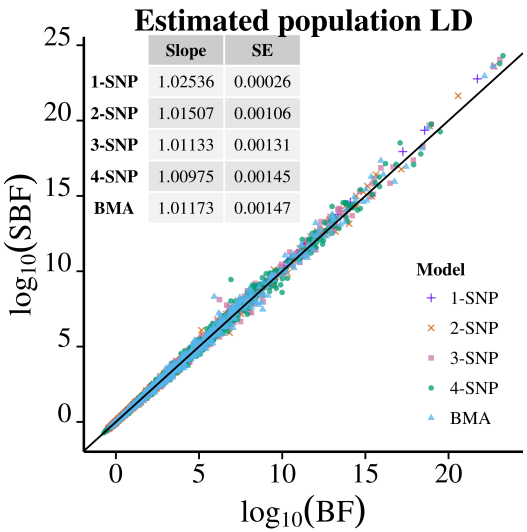
(a) Individual-level data + GCTA (Yang et al, 2011)



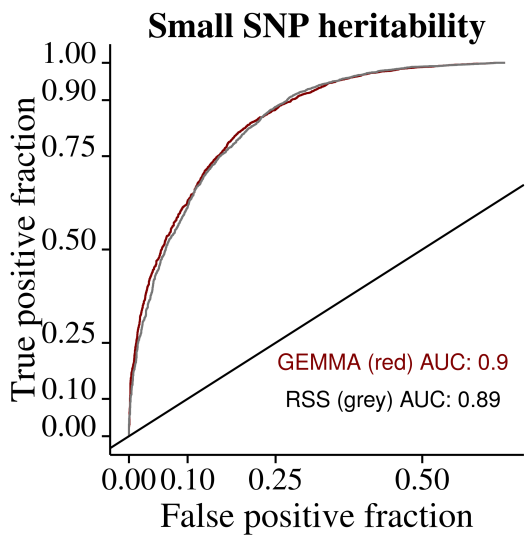
(b) Summary statistics + RSS

- RSS on summary data: 52.1%, [50.3%, 53.9%]
- GCTA on subsets of full data: 49.8% (4.4%)

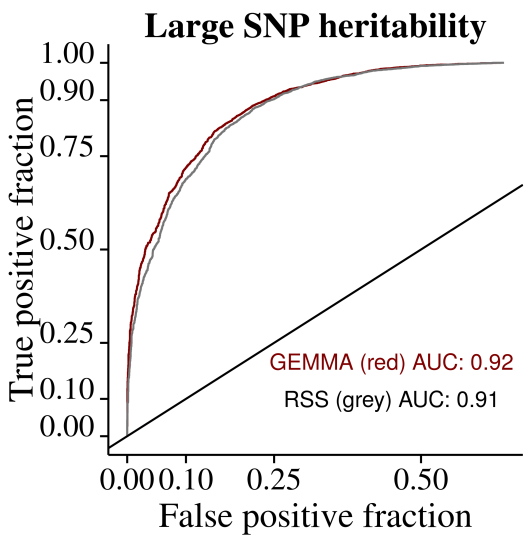
RSS can detect genetic association.



(a) BIMBAM vs RSS



(b) GEMMA vs RSS



(c) GEMMA vs RSS

- BIMBAM: multiple-SNP fine mapping of genes

  `http://www.haplotype.org/bimbam.html`

- GEMMA-BVSR: genome-wide multiple-SNP analysis

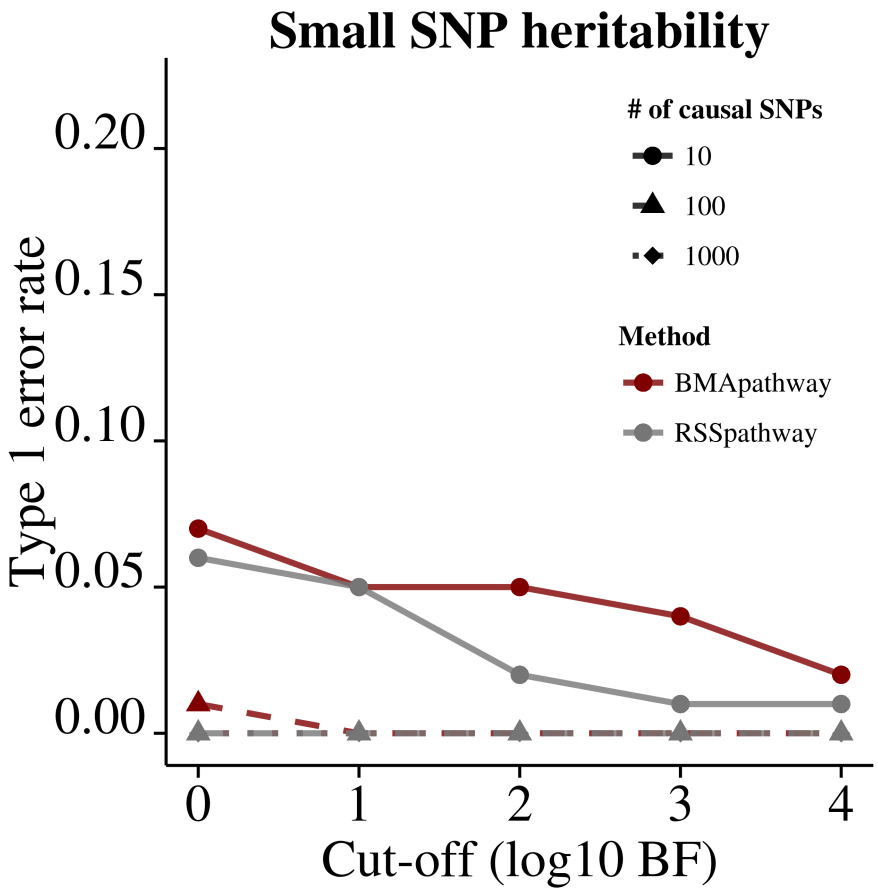
  `https://github.com/xiangzhou/GEMMA`

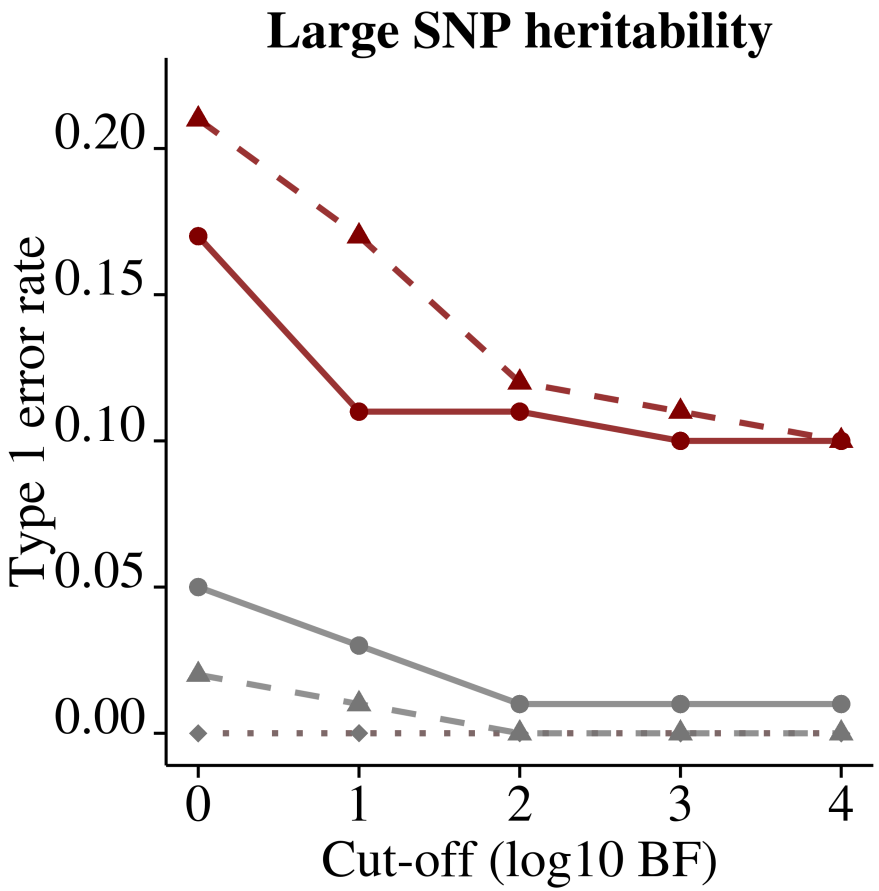
Appendix C: Simulations

We compare RSS with BMAPathway through simulations based on real genotype data (WTCCC, 2007).

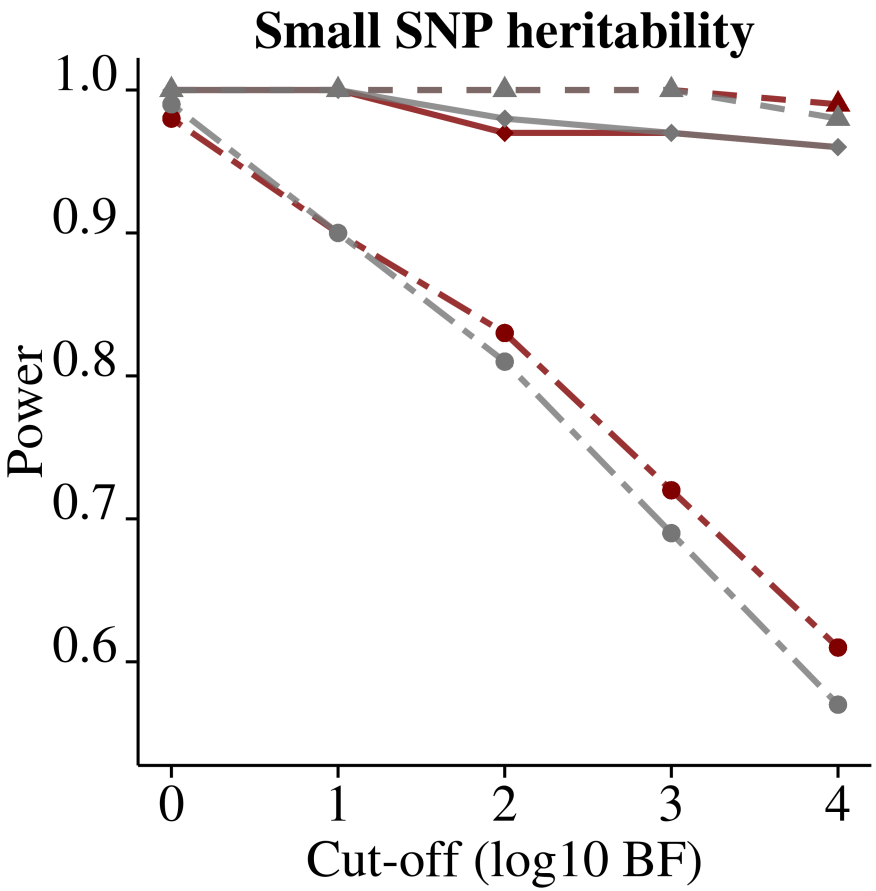
- Null dataset:
  each SNP is equally likely to be causal
- Enriched dataset:
  SNPs in the Signal Transduction pathway are more likely to be associated with the phenotypes

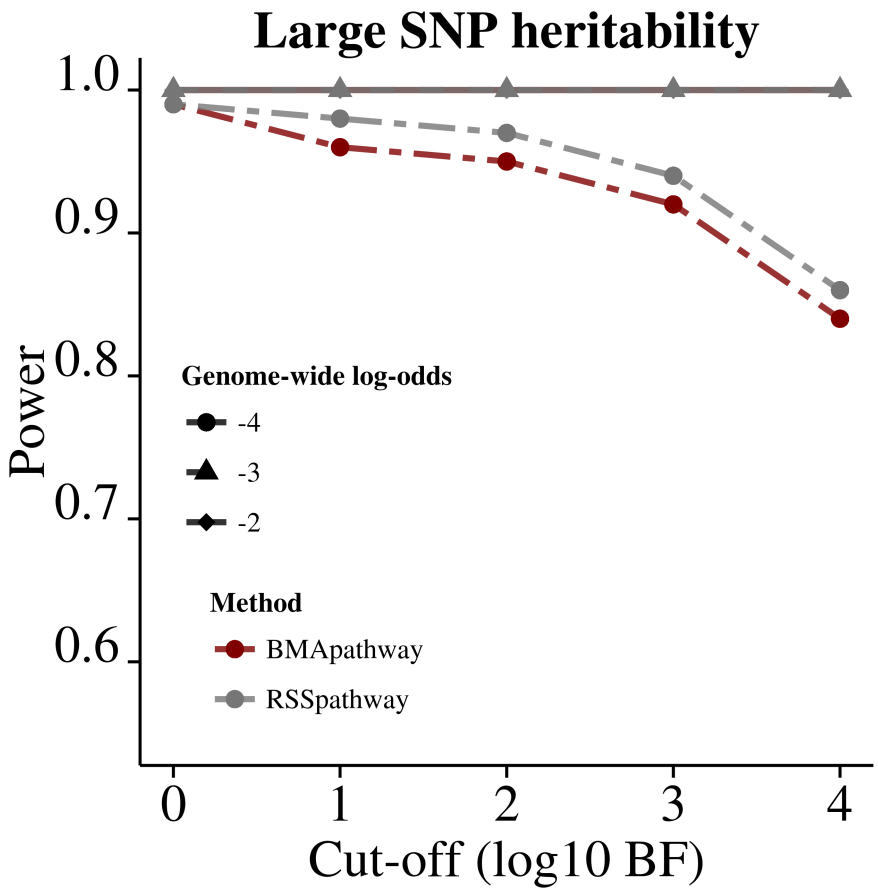
Type I Error


Small SNP heritability
# of causal SNPs
10
100
1000
Method
BMApathway
RSSpathway
Type 1 error rate
0.20
0.15
0.10
0.05
0.00
0
1
2
3
4
Cut-off (log10 BF)


Large SNP heritability
Type 1 error rate
0.20
0.15
0.10
0.05
0.00
0
1
2
3
4
Cut-off (log10 BF)

Power


Small SNP heritability
1.0
0.9
0.8
0.7
0.6
Power
0
1
2
3
4
Cut-off (log10 BF)


Large SNP heritability
1.0
0.9
0.8
0.7
0.6
Power
Genome-wide log-odds
-4
-3
-2
Method
BMApathway
RSSpathway
0
1
2
3
4
Cut-off (log10 BF)

Enrichment Estimation