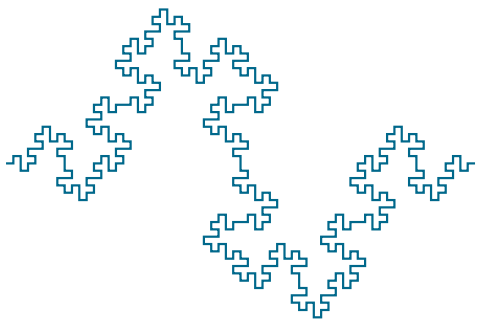
2021 MIC

Bulk RNA-seq: Pathway Analysis

Jichun Xie

Duke University



June 2021

Section 1

Overview

What is pathway analysis?

Many names for the same thing:

- Pathway analysis
- Gene set enrichment analysis
- Go-term analysis
- Gene list enrichment analysis

Single gene analysis

- Gene expression $X_1, X_2, \ldots, X_m$
- Phenotype expression Y
- Study the relationship between the genes and the phenotype.
- $$Y = \beta_{i0} + \beta_{i1} X_i + \epsilon$$

  or

  $$\mathsf{logit}\{P(Y = 1)\} = \beta_{i0} + \beta_{i1} X_i$$

  or other GLMs.
- For each gene, test the significance level

  $$\mathrm{H}_{0,i} : \beta_{i1} = 0.$$

Single gene analysis

- For each $H_{0,i}$, use Wald/score/likelihood ratio test to obtain test statistic and the corresponding P-value P_i.
- If P_i is large, then the chance that SNP/Gene i is associated with phenotype Y is small.
- If P_i is small, we think SNP/Gene i could be important.
- Thresholding P-values: Claim SNP/Gene i is significantly associated with the phenotype (Reject $H_{0,i}$) if $P_i < c$.

Single gene analysis

- How to decide the threshold?
- The threshold c depends on the desired type I error α and the number of genes m.
- Different type I error measures:
  - Family-wise error rate (FWER):

$$\mathsf{P}(\text{falsely reject any one gene})$$

  - False discovery rate (FDR):

$$\mathsf{E}\left(\frac{\text{number of the falsely rejected genes}}{\text{total number of the rejected genes}}\right)$$

Type I Error Rate

H_l: Gene set $\mathcal{S}_l$ is not associated with the phenotype, $l = 1, \ldots, m$.

	Claim significant	Claim non-significant	Total
True nulls	N_{00}	N_{01}	m_0
False nulls	N_{10}	N_{11}	m_1
Total	R	$m - R$	m

- FDR $= \mathsf{E}(N_{00}/(R \vee 1))$.
- FWER $= \mathsf{P}(N_{00} \geq 1)$.

Single gene analysis

- Typically, $\alpha = 0.05$.
- Assume all P-values are i.i.d Unif$(0, 1)$,

$$\text{FWER} = \alpha = (1 - c)^m.$$

- With $\alpha = 0.05$,

m	1	10	100	1000	10000
c	5E-2	5E-3	5E-4	5E-5	5E-6

Single gene analysis

- FWER is more conservative than FDR. This means, controlling FWER at level α will require $c(\alpha)$ to be smaller (than those for controlling FDR at level α).
- If the threshold c is smaller, fewer genes will be rejected (identified).
- Because m is very large (too many candidate genes), to control type I error (no matter which one is used) usually requires c to be very small. Thus, the power of the test will be very small.

Manhattan plot for single gene/SNP analysis

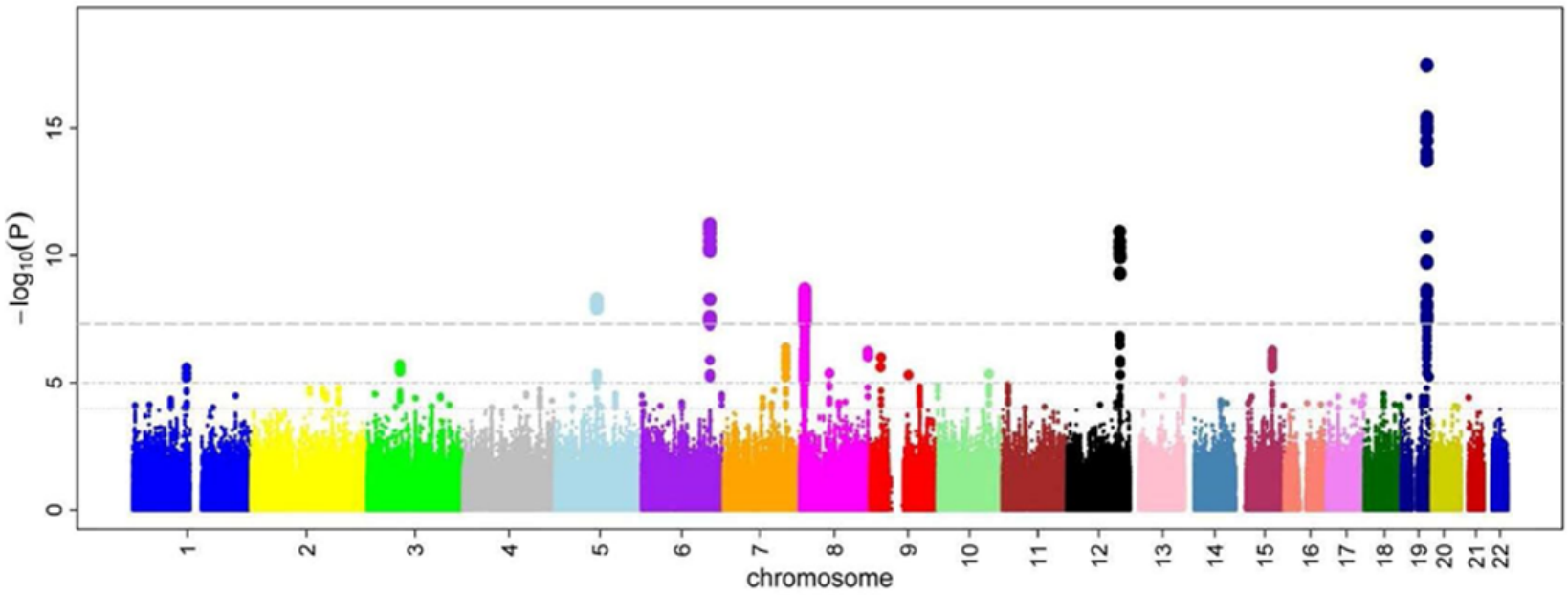


Figure: An example from Gibson (2010).

Pathway analysis

- An analysis to investigate the relationship between a disease phenotype and a set of genes on the basis of shared biological or functional properties.
- A set of genes:
  - Genes involved in a pathway
  - Genes corresponding to a Gene Ontology term
  - Genes mentioned in a paper to have certain similarities
- Are many genes in the pathway differentially expressed (up-regulated/down-regulated)?
- What is the probability of observing these changes just by chance?
- The trick is to reduce the number of candidate features.

Numer of genes >> number of gene sets

Why pathway analysis?

Single gene approach: List top 10-50 most-significant genes.

Pathway analysis: List the pathways whose genes have consistent trend to affect the phenotype.

Why pathway analysis?

Single gene approach: List top 10-50 most-significant genes.

- Assumption 1: Single gene work solely to largely increase the disease susceptibility

Pathway analysis: List the pathways whose genes have consistent trend to affect the phenotype.

- Assumption 1: Multiple Genes in the same pathway work together to confer disease susceptibility.

Why pathway analysis?

Single gene approach: List top 10-50 most-significant genes.

- Assumption 1: Single gene work solely to largely increase the disease susceptibility
- Assumption 2: The most associated gene is the best candidate for therapeutic intervention.

Pathway analysis: List the pathways whose genes have consistent trend to affect the phenotype.

- Assumption 1: Multiple Genes in the same pathway work together to confer disease susceptibility.
- Assumption 2: Targeting susceptibility pathways have clinical implications for finding additional drug targets.

Section 2

Statistical Issues

Two Types of Nulls

- Self-contained analysis: None of those genes in the gene set are associated with the phenotype.
- Competitive analysis: None of those genes in the gene set are associated with the phenotype.

Two Types of Nulls

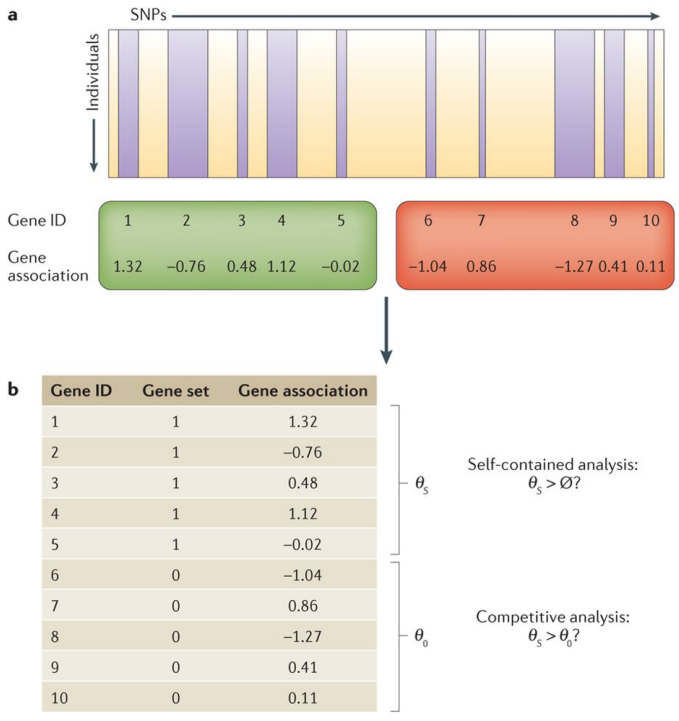
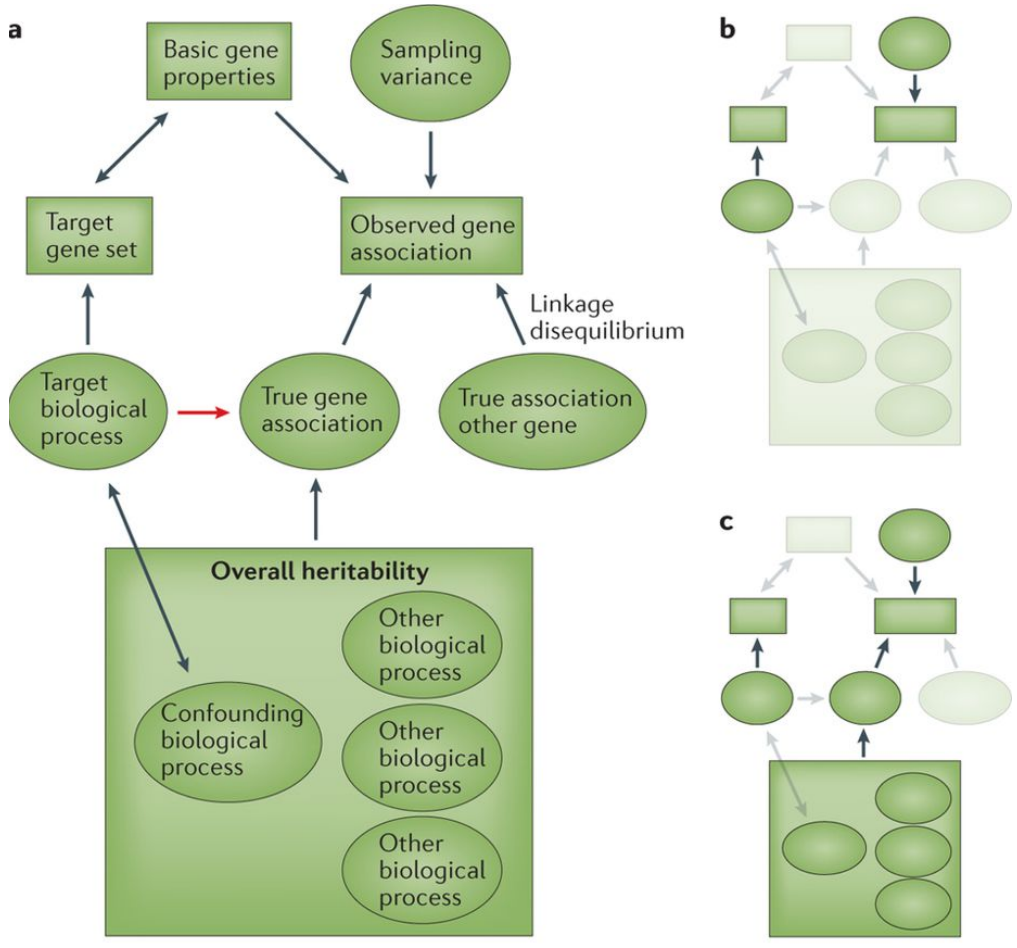


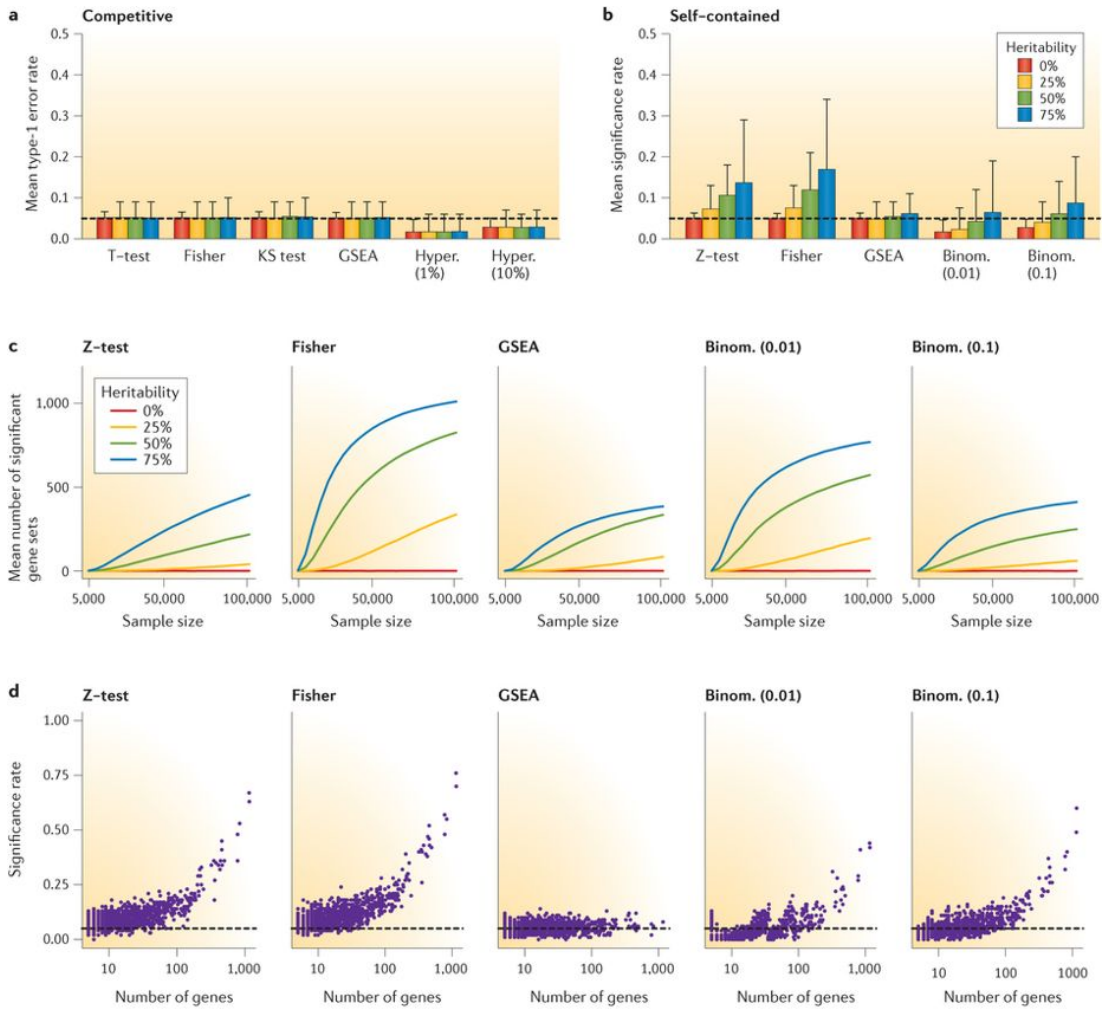
Figure: Schematic of the two-tier structures of GSA Leeuw et al. (2016).

Underlying Mechanism



Leeuw et al., 2016

Self-contained Tests Inflate Type I Error



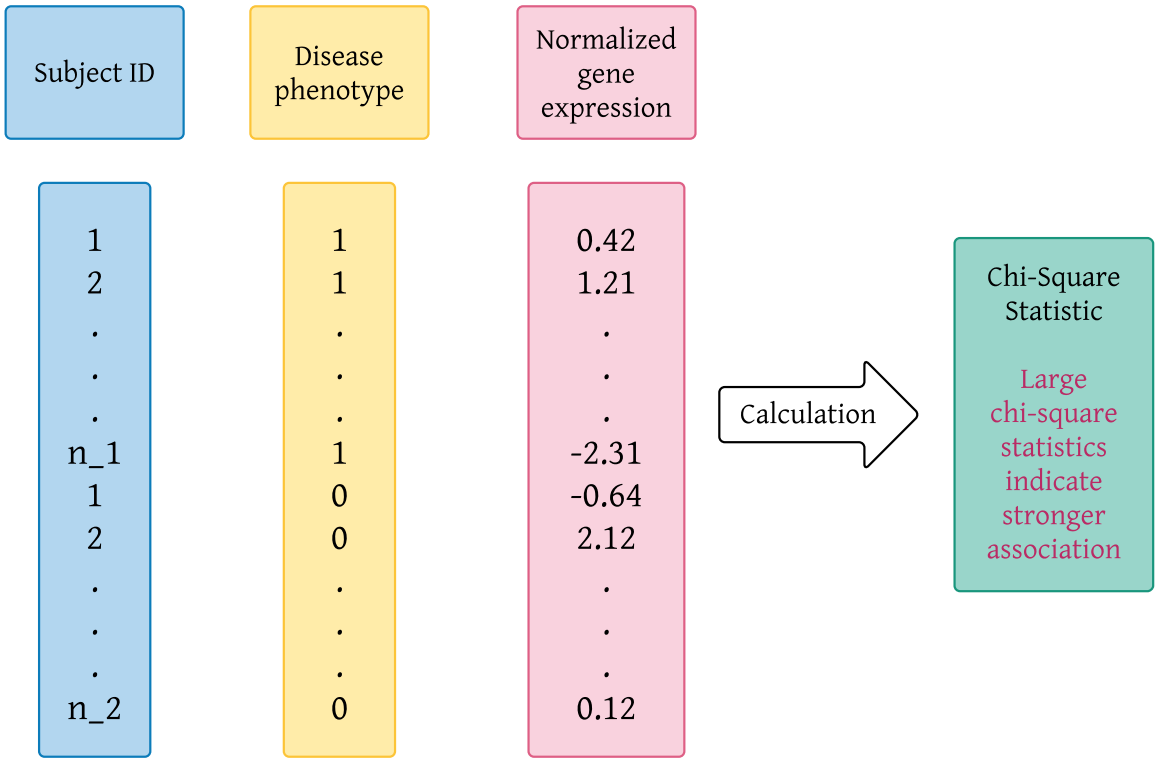
Section 3

Gene Set Enrichment Analysis (GSEA)

GSEA

- Gen-Gen: Kai Wang, Mingyao Li, and Maja Bucan (Dec. 2007). "Pathway-based approaches for analysis of genomewide association studies". In: *Am J Hum Genet* 81.6, pp. 1278–83. DOI: 10.1086/522374
- GSEA: Aravind Subramanian et al. (Oct. 2005). "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles". In: *Proc Natl Acad Sci U S A* 102.43, pp. 15545–50. DOI: 10.1073/pnas.0506580102

Normalized gene expression data



- Chi-square statistics cannot differentiate the over-expressed or under-expressed genes.
- Wald statistics can differentiate the over-expressed or under-expressed genes.

Summarize gene-phenotype association

- In total N genes.
- For gene j, get the test statistics r_j.
- Examples of r_j:
  - Score statistics
  - Wald statistics
  - Chi-square statistics

Enrichment Score

- A given gene set $\mathcal{S}$, $\mathsf{Card}(\mathcal{S}) = N_H$.
- For gene j, the larger the r_j is, the more associated gene j with the phenotype.
- Rank the association statistics from the largest to the smallest, denoted by

$$r_{(1)} \geq r_{(2)} \geq \ldots \geq r_{(N)}.$$

- Calculate a weighted Kolmogrov-Smirnov like running sum statistic

$$\text{ES}(\mathcal{S}) = \max_{1 \leq j \leq N} \left\{ \sum_{j^* \in \mathcal{S},\ j^* \leq j} \frac{|r_{(j^*)}|^p}{N_R} - \sum_{j^* \notin \mathcal{S},\ j^* \leq j} \frac{1}{N - N_H} \right\},$$

where $N_R = \sum_{j^* \in \mathcal{S}} |r_{(j^*)}|^p$.

Enrichment Score

Weighted Kolmogrov-Smirnov like running sum statistic

$$\text{ES}(\mathcal{S}) = \max_{1 \le j \le N} \left\{ \sum_{j^* \in \mathcal{S},\ j^* \le j} \frac{|r_{(j^*)}|^p}{N_R} - \sum_{j^* \notin \mathcal{S},\ j^* \le j} \frac{1}{N - N_H} \right\},$$

where $N_R = \sum_{j^* \in \mathcal{S}} |r_{(j^*)}|^p$.

- p is a parameter that gives higher weight to genes with extreme statistics.
- Common choice $p = 1$.
- $p = 0$ leads to regular KS statistic, usually not as powerful as $p = 1$.

Normalized Enrichment Score

- The enrichment score $\text{ES}(\mathcal{S})$ relies on the maximum statistic, so that a larger gene set $\mathcal{S}$ tends to produce larger $\text{ES}(\mathcal{S})$.
- Two-step normalization procedure:
  1. Permute the phenotype label of all samples
  2. During each permutation π, repeat the calculation of the enrichment score $\text{ES}(\mathcal{S}, \pi)$.
- Then

$$\text{NES}(\mathcal{S}) = \frac{\text{ES}(\mathcal{S}) - \mathsf{mean}\{\text{ES}(\mathcal{S}, \pi)\}}{\mathsf{sd}\{\text{ES}(\mathcal{S}, \pi)\}}$$

  - The NES adjusts for different sizes of genes.
  - THE NES preserves correlations between SNPs on the same gene.

Control FDR

- NES*: the normalized enrichment score in the observed data
- $$\widehat{\text{FDR}} = \frac{\%\text{ of all } (\mathcal{S}, \pi) \text{ with } \text{NES}(\mathcal{S}, \pi) \geq \text{NES}^*}{\%\text{ of observed } \mathcal{S} \text{ with } \text{NES}(\mathcal{S}) \geq \text{NES}^*}.$$
- Rationale
  - $\text{FDR} = \mathsf{E}\{N_{00}/(R \vee 1)\}$.
  - N_{00}/m: Estimated by % of all $(\mathcal{S}, \pi)$ with $\text{NES}(\mathcal{S}, \pi) \geq \text{NES}^*$.
  - R/m: Estimated by % of observed $\mathcal{S}$ with $\text{NES}(\mathcal{S}) \geq \text{NES}^*$.
- Larger NES* corresponds to smaller $\widehat{\text{FDR}}$.
- If $\widehat{\text{FDR}} \leq \alpha$, claim the corresponding gene set significant.

Control FWER

- NES*: the normalized enrichment score in the observed data
- $\widehat{\text{FWER}} = \%$ of all π with the highest $\text{NES}(\mathcal{S}, \pi) \geq \text{NES}^*$.
- Rationale:
  - $\text{FWER} = \mathsf{P}(N_{00} \geq 1) = \mathsf{E}\{I(N_{00} \geq 1)\}$.
  - Each permutation π can be viewed as a realization of the event. If the highest $\text{NES}(\mathcal{S}, \pi) \geq \text{NES}^*$, then there is a false rejection.
- Larger NES* corresponds to smaller $\widehat{\text{FWER}}$.
- If $\widehat{\text{FWER}} \leq \alpha$, claim the corresponding gene set significant.

Section 4

References

- Gibson, Greg (July 2010). "Hints of hidden heritability in GWAS". In: *Nat Genet* 42.7, pp. 558–60. DOI: 10.1038/ng0710-558.
- Leeuw, Christiaan A. de et al. (June 2016). "The statistical properties of gene-set analysis". In: *Nature Reviews Genetics* 17.6, pp. 353–364. ISSN: 1471-0064. DOI: 10.1038/nrg.2016.29.
- Subramanian, Aravind et al. (Oct. 2005). "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles". In: *Proc Natl Acad Sci U S A* 102.43, pp. 15545–50. DOI: 10.1073/pnas.0506580102.
- Wang, Kai, Mingyao Li, and Maja Bucan (Dec. 2007). "Pathway-based approaches for analysis of genomewide association studies". In: *Am J Hum Genet* 81.6, pp. 1278–83. DOI: 10.1086/522374.