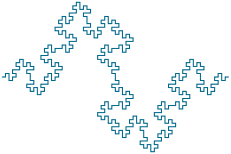
High-Throughput Sequencing Course

Gene-Set Analysis

Biostatistics and Bioinformatics



Summer 2018

Section 1

Introduction

What is Gene Set Analysis?

Many names for gene set analysis:

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

Single SNP/Gene Analysis

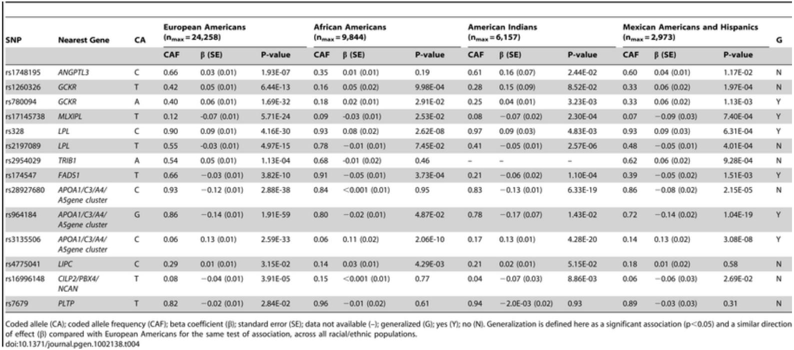
- SNP/Gene: $X_1, X_2, \ldots, X_p$
- Phenotype Y
- Study the relationship between X_i and Y
- $$Y = \beta_{i0} + \beta_{i1} X_i + Z_1$$

  or

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

  or other GLMs.
- Obtain the p-value P_i corresponding to the significance level of β_{i1}.
- Threshold p-values.

Typical Results of GWAS Analysis (Single SNP Approach)



SNP	Nearest Gene	CA	European Americans (n_{max} = 24,258)			African Americans (n_{max} = 9,844)			American Indians (n_{max} = 6,157)			Mexican Americans and Hispanics (n_{max} = 2,973)			G
			CAF	β (SE)	P-value	CAF	β (SE)	P-value	CAF	β (SE)	P-value	CAF	β (SE)	P-value	
rs1748195	*ANGPTL3*	C	0.66	0.03 (0.01)	1.93E-07	0.35	0.01 (0.01)	0.19	0.61	0.16 (0.07)	2.44E-02	0.60	0.04 (0.01)	1.17E-02	N
rs1260326	*GCKR*	T	0.42	0.05 (0.01)	6.44E-13	0.16	0.05 (0.02)	9.98E-04	0.28	0.15 (0.09)	8.52E-02	0.33	0.06 (0.02)	1.97E-04	N
rs780094	*GCKR*	A	0.40	0.06 (0.01)	1.69E-32	0.18	0.02 (0.01)	2.91E-02	0.25	0.04 (0.01)	3.23E-03	0.33	0.06 (0.02)	1.13E-03	Y
rs17145738	*MLXIPL*	T	0.12	-0.07 (0.01)	5.71E-24	0.09	-0.03 (0.01)	2.53E-02	0.08	−0.07 (0.02)	2.30E-04	0.07	−0.09 (0.03)	7.40E-04	Y
rs328	*LPL*	C	0.90	0.09 (0.01)	4.16E-30	0.93	0.08 (0.02)	2.62E-08	0.97	0.09 (0.03)	4.83E-03	0.93	0.09 (0.03)	6.31E-04	Y
rs2197089	*LPL*	T	0.55	-0.03 (0.01)	4.97E-15	0.78	−0.01 (0.01)	7.45E-02	0.41	−0.05 (0.01)	2.57E-06	0.48	−0.05 (0.01)	4.01E-04	N
rs2954029	*TRIB1*	A	0.54	0.05 (0.01)	1.13E-04	0.68	-0.01 (0.02)	0.46	–	–	–	0.62	0.06 (0.02)	9.28E-04	N
rs174547	*FADS1*	T	0.66	−0.03 (0.01)	3.82E-10	0.91	−0.05 (0.01)	3.73E-04	0.21	−0.06 (0.02)	1.10E-04	0.39	−0.05 (0.02)	1.51E-03	Y
rs28927680	*APOA1/C3/A4/ A5gene cluster*	C	0.93	−0.12 (0.01)	2.88E-38	0.84	<0.001 (0.01)	0.95	0.83	−0.13 (0.01)	6.33E-19	0.86	−0.08 (0.02)	2.15E-05	N
rs964184	*APOA1/C3/A4/ A5gene cluster*	G	0.86	−0.14 (0.01)	1.91E-59	0.80	−0.02 (0.01)	4.87E-02	0.78	−0.17 (0.07)	1.43E-02	0.72	−0.14 (0.02)	1.04E-19	Y
rs3135506	*APOA1/C3/A4/ A5gene cluster*	C	0.06	0.13 (0.01)	2.59E-33	0.06	0.11 (0.02)	2.06E-10	0.17	0.13 (0.01)	4.28E-20	0.14	0.13 (0.02)	3.08E-08	Y
rs4775041	*LIPC*	C	0.29	0.01 (0.01)	3.15E-02	0.14	0.03 (0.01)	4.29E-03	0.21	0.02 (0.01)	5.15E-02	0.18	0.01 (0.02)	0.58	N
rs16996148	*CILP2/PBX4/ NCAN*	T	0.08	−0.04 (0.01)	3.91E-05	0.15	<0.001 (0.01)	0.77	0.04	−0.07 (0.03)	8.86E-03	0.06	−0.06 (0.03)	2.69E-02	N
rs7679	*PLTP*	T	0.82	−0.02 (0.01)	2.84E-02	0.96	−0.01 (0.02)	0.61	0.94	−2.0E-03 (0.02)	0.93	0.89	−0.03 (0.03)	0.31	N

Coded allele (CA); coded allele frequency (CAF); beta coefficient (β); standard error (SE); data not available (–); generalized (G); yes (Y); no (N). Generalization is defined here as a significant association ($p<0.05$) and a similar direction of effect (β) compared with European Americans for the same test of association, across all racial/ethnic populations.

doi:10.1371/journal.pgen.1002138.t004

Figure: An example from Dumitrescu et al. (2011).

Typical Results of GWAS Analysis (Single SNP Approach)

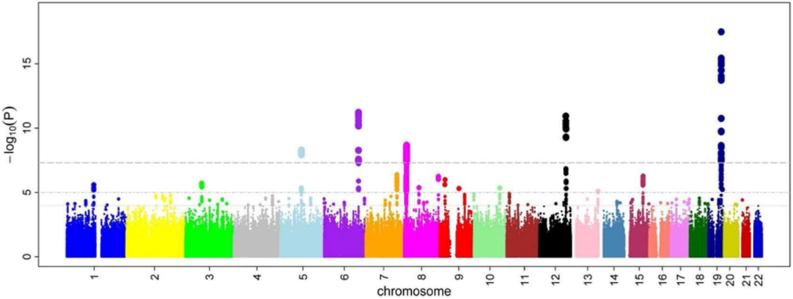


Figure: An example from Gibson (2010).

Gene Set Analysis (GSA)

- An analysis to investigate the relationship between a disease phenotype and a set of genes on the basis of shared biological or functional properties.
- Gene set: 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

Goal of GSA

Goal: give one number to measure the significance of a gene set as a whole.

- Are many genes in the pathway differentially expressed (up-regulated/down-regulated)?
- What is the probability of observing these changes just by chance?

Why GSA?

Single SNP approach: List top 20-50 most-significant SNPs and their neighboring genes.

GSA approach: List the pathways that have genes in the pathway have consistent trend to affect the phenotype.

Why GSA?

Single SNP approach: List top 20-50 most-significant SNPs and their neighboring genes.

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

GSA approach: List the pathways that have genes in the pathway have consistent trend to affect the phenotype.

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

Why GSA?

Single SNP approach: List top 20-50 most-significant SNPs and their neighboring 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.

GSA approach: List the pathways that have genes in the pathway 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.

Why GSA?

- Interpretation of genome-wide results
- Gene-sets are (typically) fewer than all the genes and have more descriptive names
- Difficult to manage a long list of significant genes
- Integrates external information into the analysis
- Less prone to false-positives on the gene-level
- Top genes might not be the interesting ones, several coordinated smaller changes
- Detect patterns that would be difficult to discern simply by manually going through, *e.g.*, the list of differentially expressed genes

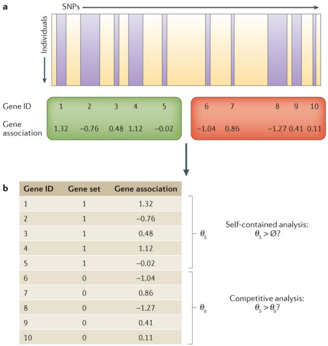
Section 2

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.

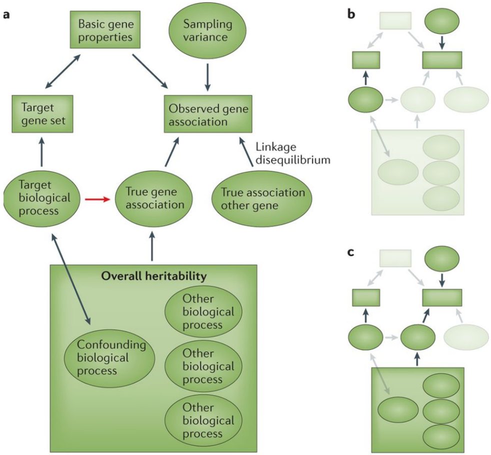
Two Types of Nulls



Gene ID	Gene set	Gene association
1	1	1.32
2	1	−0.76
3	1	0.48
4	1	1.12
5	1	−0.02
6	0	−1.04
7	0	0.86
8	0	−1.27
9	0	0.41
10	0	0.11

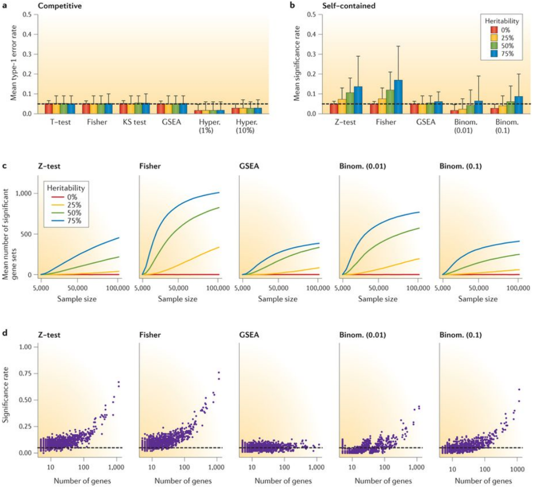
Figure: Schematic of the two-tier structures of GSA Leeuw et al. (2016).

Underlying Mechanism



Leeuw et al., 2016

Self-contained Tests Inflate Type I Error



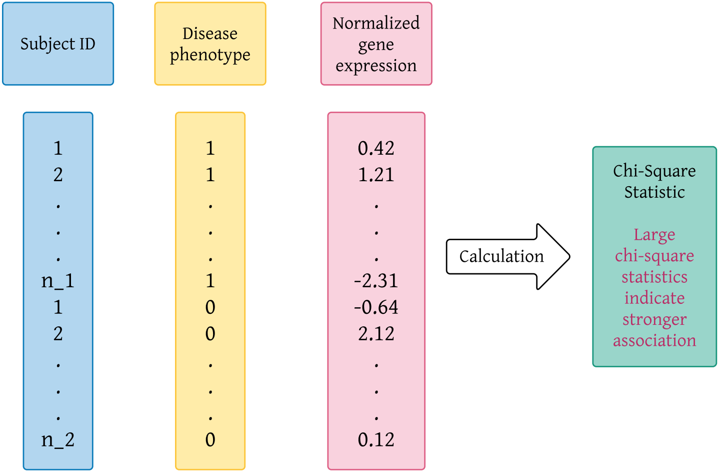
Section 3

Method: Gen-Gen/GSEA

Gen-Gen/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

Microarray Data

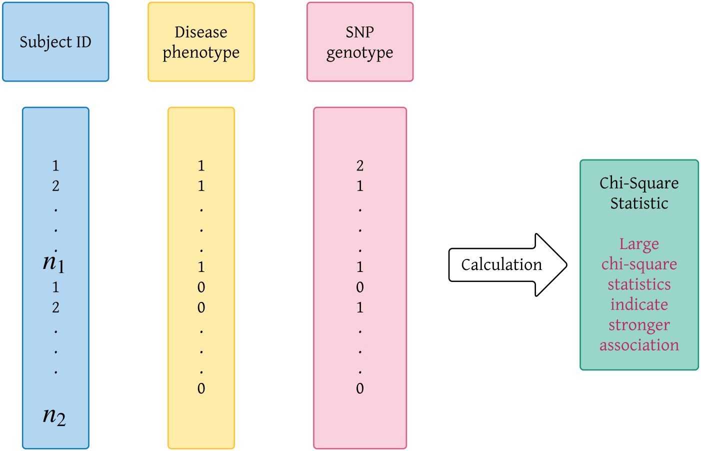


Subject ID	Disease phenotype	Normalized gene expression
1	1	0.42
2	1	1.21
.	.	.
.	.	.
.	.	.
n_1	1	-2.31
1	0	-0.64
2	0	2.12
.	.	.
.	.	.
.	.	.
n_2	0	0.12

Chi-Square Statistic

Large chi-square statistics indicate stronger association

Single Nucleotide Polymorphism Data



Subject ID	Disease phenotype	SNP genotype
1	1	2
2	1	1
.	.	.
.	.	.
.	.	.
n_1	1	1
1	0	0
2	0	1
.	.	.
.	.	.
.	.	.
n_2	0	0

Chi-Square Statistic

Large chi-square statistics indicate stronger association

Summarize SNP Associate on One Gene

- Map SNP V_i to gene j ($\mathcal{G}_j$) if the SNP is located within the gene or if the gene is the closest gene to the SNP.
- In total N genes.
- When one SNP is located within shared regions of two overlapping genes, the SNP is mapped to both genes.
- For each gene, assign the highest statistic value among all SNPs mapped to the gene as the statistic value of the gene, $r_j = \max_{v_i \in \mathcal{G}_j} t_i$.

Enrichment Score

- A given gene set $\mathcal{S}$, $\mathsf{Card}(\mathcal{S}) = N_H$.
- Calculate association chi-square statistics r_j, $j = 1, \ldots, N$.
- The larger the r_j is, the more associated gene Oj 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

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

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

$$\textsc{nes}(\mathcal{S}) = \frac{\textsc{es}(\mathcal{S}) - \mathsf{mean}\{\textsc{es}(\mathcal{S}, \pi)\}}{\mathsf{sd}\{\textsc{es}(\mathcal{S}, \pi)\}}$$

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

Type I Error Rate

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

	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

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

Control fdr

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

Control fwer

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

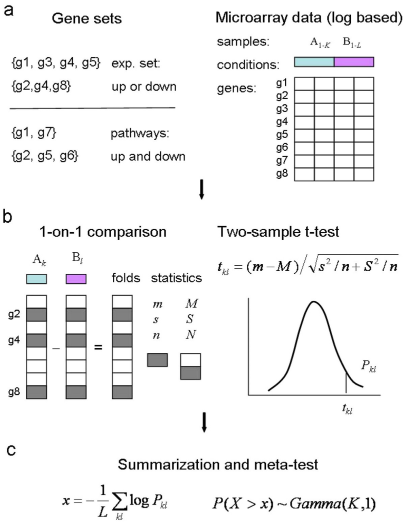
Section 4

Method: GAGE

GAGE

- **Luo2009GAGE**
- Gene expression data: RNA-Seq or Microarrary

GAGE Method Overview



$$t_{kl} = (m - M)\Big/\sqrt{s^2/n + S^2/n}$$

$$x = -\frac{1}{L}\sum_{kl} \log P_{kl} \qquad P(X > x) \sim Gamma(K,1)$$

Setting

- Gene: $i \in \{1, \ldots, N\}$
- Condition/Phenotype: $s \in 0, 1$
  - Paired (1-on-1): *e.g.*, one condition *vs.* another condition:
  - Unpaired (grp-on-grp): *e.g.*, one phenotype *vs.* another phenotype:
- Subject:
  - Paired: $k \in \{1, \ldots, K\}$
  - Unpaired: $k \in \{1, \ldots, K_1\}$ for cases and $k \in \{1, \ldots, K_0\}$ for controls.
- Gene expression:

$$G_{s,k,i} = \begin{cases} \text{Transcription level of gene } i & \text{Microarray} \\ \text{Read counts of gene } i/\text{Total counts} & \text{RNA-Seq} \end{cases}$$

$\log_2$ FOLD CHANGE

- Compare the gene expressions between two conditions or two phenotypes
  - Paired (1-on-1): $X_{k,i} = G_{1,k,i}/G_{0,k,i}$
  - Unpaired (grp-on-grp): $X_i = \bar{G}_{1,i}/\bar{G}_{0,i}$
  - Efficient but not recommended (1-on-grp): $X_{k,i} = G_{1,k,i}/\bar{G}_{0,i}$

Gene Set and T-statistic

- Gene set of interest $\mathcal{S}$
- mean fold change: $m = \mathsf{mean}_{i\in\mathcal{S}}(X_i)$ (gene set) *vs.* $M = \mathsf{mean}_{i\in\{1,\dots,N\}}(X_i)$ (all genes)
- standard deviation folde change: $s = \mathsf{sd}_{i\in\mathcal{S}}(X_i)$ (gene set) *vs.* $S = \mathsf{sd}_{i\in\{1,\dots,N\}}(X_i)$ (all genes)
- number of genes: n (gene set) *vs.* N (all genes)
- T-statistic:

$$T = (m - M)/\sqrt{s^2/n + S^2/n}$$

Remark:

- This is a two sample t-test between the interesting gene set containing n genes and a virtual random set of the same size derived from the background.
- Subscript k is left out for simplicity. We will discuss 1-on-1 setting (with subscript k) later.

P-Value

- Degree of freedom of T under the null

$$\mathrm{df} = (n-1)\frac{s^2 + S^2}{s^4 + S^4}.$$

- P-value:
    - Two sided: pathway set (genes may be het erogeneously regulated in either direction)
    - One sided: experimental set (genes are regulated in the same direction)
- Alternative choice of T: rank-based test (Wilcoxon Mann-Whitney test)

Summarizing P-Values

Recall that for 1-on-1 (paired) setting, the P-value for gene set $\mathcal{S}$ and subject k is $P_k(\mathcal{S})$.

$$X(\mathcal{S}) = \sum_k \log P_k(\mathcal{S}).$$

Under the null, $P_k(\mathcal{S})$ independently follows $\text{Unif}(0,1)$, and then $X(\mathcal{S})$ follows $\text{Gamma}(K,1)$.

Controlling FDR

If multiple gene sets are of interest, multiple testing methods are applied to control FDR.

- fdrtool: Korbinian Strimmer (July 2008). "A unified approach to false discovery rate estimation". In: *BMC Bioinformatics* 9, p. 303. DOI: 10.1186/1471-2105-9-303
- Benjamini and Hochberg (BH) procedure: Y. Benjamini and Y. Hochberg (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 57.1, pp. 289–300. ISSN: 00359246. URL: http://www.jstor.org/stable/2346101

Section 5

References

Benjamini, Y. and Y. Hochberg (1995). “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 57.1, pp. 289–300. ISSN: 00359246. URL: `http://www.jstor.org/stable/2346101`.

Dumitrescu, Logan et al. (June 2011). “Genetic determinants of lipid traits in diverse populations from the population architecture using genomics and epidemiology (PAGE) study”. In: *PLoS Genet* 7.6, e1002138. DOI: `10.1371/journal.pgen.1002138`.

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. URL: http://dx.doi.org/10.1038/nrg.2016.29.

Strimmer, Korbinian (July 2008). “A unified approach to false discovery rate estimation”. In: *BMC Bioinformatics* 9, p. 303. DOI: 10.1186/1471-2105-9-303.

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.