High-Throughput Sequencing Course
Gene-Set Analysis

Biostatistics and Bioinformatics

Summer 2019
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
$$\text{logit}\{P(Y = 1)\} = \beta_{i0} + \beta_{i1}X_i$$

or other GLMs.

- Obtain the $p$-value $P_i$ corresponding to the significance level of $\beta_{i1}$.
- Threshold $p$-values.
## Typical Results of GWAS Analysis (Single SNP Approach)

<table>
<thead>
<tr>
<th>SNP</th>
<th>Nearest Gene</th>
<th>CA (CAF)</th>
<th>European Americans (n&lt;sub&gt;max&lt;/sub&gt; = 24,258)</th>
<th>African Americans (n&lt;sub&gt;max&lt;/sub&gt; = 9,844)</th>
<th>American Indians (n&lt;sub&gt;max&lt;/sub&gt; = 6,157)</th>
<th>Mexican Americans and Hispanics (n&lt;sub&gt;max&lt;/sub&gt; = 2,973)</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs1748195</td>
<td>ANGPT13</td>
<td>C 0.66</td>
<td>0.03 (0.01)</td>
<td>1.93E-07</td>
<td>0.35 (0.01)</td>
<td>0.19 (0.07)</td>
</tr>
<tr>
<td>rs1260326</td>
<td>GCKR</td>
<td>T 0.42</td>
<td>0.05 (0.01)</td>
<td>6.44E-13</td>
<td>0.16 (0.02)</td>
<td>0.28 (0.09)</td>
</tr>
<tr>
<td>rs780094</td>
<td>GCKR</td>
<td>A 0.40</td>
<td>0.06 (0.01)</td>
<td>1.69E-12</td>
<td>0.18 (0.01)</td>
<td>0.25 (0.01)</td>
</tr>
<tr>
<td>rs17145738</td>
<td>MLXIP1</td>
<td>T 0.12</td>
<td>-0.07 (0.01)</td>
<td>5.71E-14</td>
<td>0.09 (0.01)</td>
<td>-0.07 (0.02)</td>
</tr>
<tr>
<td>rs328</td>
<td>LPL</td>
<td>C 0.90</td>
<td>0.09 (0.01)</td>
<td>4.16E-10</td>
<td>0.93 (0.02)</td>
<td>2.62E-08</td>
</tr>
<tr>
<td>rs2197089</td>
<td>LPL</td>
<td>T 0.35</td>
<td>-0.03 (0.01)</td>
<td>4.97E-15</td>
<td>0.78 (0.01)</td>
<td>-0.05 (0.01)</td>
</tr>
<tr>
<td>rs2954029</td>
<td>TRIB1</td>
<td>A 0.34</td>
<td>0.05 (0.01)</td>
<td>1.13E-04</td>
<td>0.68 (0.02)</td>
<td>-0.01 (0.01)</td>
</tr>
<tr>
<td>rs174547</td>
<td>FADS1</td>
<td>T 0.66</td>
<td>-0.03 (0.01)</td>
<td>3.62E-10</td>
<td>0.91 (0.01)</td>
<td>-0.05 (0.02)</td>
</tr>
<tr>
<td>rs28927680</td>
<td>APOA1/C3/A4/</td>
<td>C 0.93</td>
<td>-0.12 (0.01)</td>
<td>2.88E-18</td>
<td>0.84 (0.01)</td>
<td>-0.01 (0.01)</td>
</tr>
<tr>
<td>rs964184</td>
<td>APOA1/C3/A4/</td>
<td>G 0.86</td>
<td>-0.14 (0.01)</td>
<td>1.91E-09</td>
<td>0.80 (0.01)</td>
<td>-0.02 (0.01)</td>
</tr>
<tr>
<td>rs3135506</td>
<td>APOA1/C3/A4/</td>
<td>C 0.66</td>
<td>0.13 (0.01)</td>
<td>2.59E-13</td>
<td>0.06 (0.02)</td>
<td>0.11 (0.02)</td>
</tr>
<tr>
<td>rs4775041</td>
<td>LPC</td>
<td>C 0.29</td>
<td>0.01 (0.01)</td>
<td>3.15E-02</td>
<td>0.14 (0.01)</td>
<td>0.03 (0.01)</td>
</tr>
<tr>
<td>rs16996148</td>
<td>CILP2/PBX4/</td>
<td>T 0.08</td>
<td>-0.04 (0.01)</td>
<td>3.91E-05</td>
<td>0.15 (0.01)</td>
<td>-0.00 (0.01)</td>
</tr>
<tr>
<td>rs7679</td>
<td>PLTP</td>
<td>T 0.82</td>
<td>-0.02 (0.01)</td>
<td>2.84E-02</td>
<td>0.96 (0.02)</td>
<td>-0.01 (0.02)</td>
</tr>
</tbody>
</table>

Coded allele (CA); coded allele frequency (CAF); beta coefficient (); 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 () 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

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

**Microarray Data**

<table>
<thead>
<tr>
<th>Subject ID</th>
<th>Disease phenotype</th>
<th>Normalized gene expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.42</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1.21</td>
</tr>
<tr>
<td>n_1</td>
<td>1</td>
<td>-2.31</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>-0.64</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>2.12</td>
</tr>
</tbody>
</table>

Chi-Square Statistic

Large chi-square statistics indicate stronger association
**Single Nucleotide Polymorphism Data**

<table>
<thead>
<tr>
<th>Subject ID</th>
<th>Disease phenotype</th>
<th>SNP genotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>(n_1)</td>
<td>(n_2)</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>(\ldots)</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
</tr>
</tbody>
</table>

Chi-Square Statistic

Large chi-square statistics indicate stronger association
**Summarize SNP Association on One Gene**

- Map SNP $V_i$ to gene $j$ ($G_j$) 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, $r_j = \max_{v_i \in G_j} t_i$. 
**Enrichment Score**

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

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

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

- \( p \) 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 $ES(\mathcal{S})$ relies on the maximum statistic, so that a larger gene set $\mathcal{S}$ tends to produce larger $ES(\mathcal{S})$.

- Two-step normalization procedure:
  1. Permute the phenotype label of all samples
  2. During each permutation $\pi$, repeat the calculation of the enrichment score $ES(\mathcal{S}, \pi)$.

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

- The $NES$ adjusts for different sizes of genes.
- THE $NES$ preserves correlations between SNPs on the same gene.
**Type I Error Rate**

\( H_l : \) Gene set \( S_l \) is not associated with the phenotype, 
\[ l = 1, \ldots, m. \]

<table>
<thead>
<tr>
<th></th>
<th>Claim significant</th>
<th>Claim non-significant</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>True nulls</td>
<td>( N_{00} )</td>
<td>( N_{01} )</td>
<td>( m_0 )</td>
</tr>
<tr>
<td>False nulls</td>
<td>( N_{10} )</td>
<td>( N_{11} )</td>
<td>( m_1 )</td>
</tr>
<tr>
<td>Total</td>
<td>( R )</td>
<td>( m - R )</td>
<td>( m )</td>
</tr>
</tbody>
</table>

\[ \text{FDR} = \mathbb{E}(N_{00}/(R \lor 1)). \]
\[ \text{FWER} = \mathbb{P}(N_{00} \geq 1). \]
CONTROL FDR

▶ NES*: the normalized enrichment score in the observed data

\[
\widehat{FDR} = \frac{\% \text{ of all } (S, \pi) \text{ with } NES(S, \pi) \geq NES^*}{\% \text{ of observed } S \text{ with } NES(S) \geq NES^*}.
\]

▶ Rationale

▶ FDR = E{N_{00}/(R \lor 1)}.

▶ N_{00}/m: Estimated by % of all (S, \pi) with NES(S, \pi) \geq NES^*.

▶ R/m: Estimated by % of observed S with NES(S) \geq NES^*.

▶ Larger NES^* corresponds to smaller \(\widehat{FDR}\).

▶ If \(\widehat{FDR} \leq \alpha\), claim the corresponding gene set significant.
CONTROL FWER

- $\text{NES}^*$: the normalized enrichment score in the observed data
- $\hat{\text{FWER}} = \%$ of all $\pi$ with the highest $\text{NES}(\mathcal{S}, \pi) \geq \text{NES}^*$.
- Rationale:
  - $\text{FWER} = P(N_{00} \geq 1) = \mathbb{E}\{I(N_{00} \geq 1)\}$.
  - Each permutation $\pi$ 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 $\text{NES}^*$ corresponds to smaller $\hat{\text{FWER}}$.
- If $\hat{\text{FWER}} \leq \alpha$, claim the corresponding gene set significant.
Section 4

Method: GAGE
GAGE


- Gene expression data: RNA-Seq or Microarray
### GAGE Method Overview

**Gene sets**
- \{g_1, g_3, g_4, g_5\} exp. set
- \{g_2, g_4, g_8\} up or down
- \{g_1, g_7\} pathways
- \{g_2, g_5, g_6\} up and down

**Microarray data (log based)**
- samples: \(A_{1:t}, B_{1:t}\)
- conditions: 
  - green
  - pink
- genes:

**1-on-1 comparison**
- \(A_k, B_l\)
- folds
- statistics

**Two-sample t-test**
\[
t_{kl} = \frac{(m-M)/s^2}{n + S^2}/n
\]

**Summarization and meta-test**
\[
x = -\frac{1}{L} \sum_{kl} \log P_{kl}
\]
\(P(X > x) \sim Gamma(K,1)\)
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:**
  \[
  G_{s,k,i} = \begin{cases} 
  \text{Transcription level of gene } i & \text{Microarray} \\
  \text{Read counts of gene } i / \text{Total counts} & \text{RNA-Seq}
  \end{cases}
  \]
\[ \log_2 \text{ FOLD CHANGE} \]

- Compare the gene expressions between two conditions or two phenotypes
  - Paired (1-on-1): \[ X_{k,i} = \frac{G_{1,k,i}}{G_{0,k,i}} \]
  - Unpaired (grp-on-grp): \[ X_i = \frac{\bar{G}_{1,i}}{\bar{G}_{0,i}} \]
  - Efficient but not recommended (1-on-grp): \[ X_{k,i} = \frac{G_{1,k,i}}{\bar{G}_{0,i}} \]
Gene Set and T-statistic

- Gene set of interest $S$
- mean fold change: $m = \text{mean}_{i \in S}(X_i)$ (gene set) vs. $M = \text{mean}_{i \in \{1,...,N\}}(X_i)$ (all genes)
- standard deviation fold change: $s = \text{sd}_{i \in S}(X_i)$ (gene set) vs. $S = \text{sd}_{i \in \{1,...,N\}}(X_i)$ (all genes)
- number of genes: $n$ (gene set) vs. $N$ (all genes)
- T-statistic:
  \[
  T = \frac{(m - M)}{\sqrt{s^2/n + S^2/n}}
  \]

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

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

- $P$-value:
  - Two sided: pathway set (genes may be heterogeneously 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 $S$ and subject $k$ is $P_k(S)$.

\[ X(S) = \sum_k \log P_k(S). \]

Under the null, $P_k(S)$ independently follows Unif(0, 1), and then $X(S)$ follows Gamma($K, 1$).
CONTROLLING fdr

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


Section 5

References

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.


