On the statistics of gene set enrichment

Michael A. Newton

University of Wisconsin

UW Madison, March 2008
A common situation in data analysis

- multiple tissue samples
  e.g., tumors from patients
A common situation in data analysis

- multiple tissue samples
  e.g., tumors from patients
- molecular data
  microarray expression levels
A common situation in data analysis

- multiple tissue samples
  e.g., tumors from patients
- molecular data
  microarray expression levels
- phenotype
  treatment/control
  virus status
A common situation in data analysis

- multiple tissue samples
e.g., tumors from patients
- molecular data
  microarray expression levels
- phenotype
  treatment/control
  virus status
- other biological data
  exogenous
  collections of gene sets
Gene perspective

- expression levels in multiple tissue samples
- other properties of the gene (sets)
Gene perspective

- expression levels in multiple tissue samples
- other properties of the gene (sets)

General problem

- integrate data
- assess enrichment
- develop leads for followup
Analysis step 1: merge gene expression and phenotype

- genes \{1, 2, \ldots, G\}
Analysis step 1: merge gene expression and phenotype

- genes \( \{1, 2, \ldots, G\} \)
- gene-level statistics

\[ s = (s_1, s_2, \ldots, s_G) \]

- log fold change between treatment control
- correlation with virus status
- collapsed across tissue samples
Analysis step 1: merge gene expression and phenotype

- genes \( \{1, 2, \ldots, G\} \)
- gene-level statistics
  \[
  s = (s_1, s_2, \ldots, s_G)
  \]
  - log fold change between treatment control
  - correlation with virus status
  - collapsed across tissue samples
- large \( s_g \) \( \longleftrightarrow \) \( g \) is *interesting*
Analysis step 1: merge gene expression and phenotype

- genes \{1, 2, \ldots, G\}
- gene-level statistics

\[ s = (s_1, s_2, \ldots, s_G) \]

- log fold change between treatment control
- correlation with virus status
- collapsed across tissue samples

- large \( s_g \) \( \iff \) \( g \) is interesting
- statistical issues
Analysis step 1: merge gene expression and phenotype

- genes \( \{1, 2, \ldots, G\} \)
- gene-level statistics

\[ s = (s_1, s_2, \ldots, s_G) \]

- log fold change between treatment control
- correlation with virus status
- collapsed across tissue samples

- large \( s_g \) \( \iff \) \( g \) is interesting

- statistical issues
  - choice of \( s_g \)
Analysis step 1: merge gene expression and phenotype

- genes \{1, 2, \ldots, G\}
- gene-level statistics
  \[ s = (s_1, s_2, \ldots, s_G) \]
  - log fold change between treatment control
  - correlation with virus status
  - collapsed across tissue samples
- large \(s_g \leftarrow g\) is interesting
- statistical issues
  - choice of \(s_g\)
  - gene listing; FDR control; etc
Analysis step 2: merge s and gene sets

- gene set $c \subset \{1, 2, \ldots, G\}$ of size $m$
Analysis step 2: merge $s$ and gene sets

- gene set $c \subset \{1, 2, \ldots, G\}$ of size $m$
- $c = \{\text{genes having specific biological property}\}$
Analysis step 2: merge $s$ and gene sets

- gene set $c \subset \{1, 2, \ldots, G\}$ of size $m$
- $c = \{ \text{genes having specific biological property} \}$
- collections $\{c\}$
  - Gene Ontology (GO), Kyoto Encyclopedia (KEGG)
Analysis step 2: merge \(s\) and gene sets

- gene set \(c \subseteq \{1, 2, \ldots, G\}\) of size \(m\)
- \(c = \{\text{genes having specific biological property}\}\)
- collections \(\{c\}\)
  - Gene Ontology (GO), Kyoto Encyclopedia (KEGG)
- set-level statistic \(u(c, s)\)
  - enrichment of \(c\) for interesting genes
Analysis step 2: merge s and gene sets

- gene set \( c \subset \{1, 2, \ldots, G\} \) of size \( m \)
- \( c = \{ \text{genes having specific biological property} \} \)
- collections \( \{c\} \)
  - Gene Ontology (GO), Kyoto Encyclopedia (KEGG)
- set-level statistic \( u(c, s) \)
  - enrichment of \( c \) for interesting genes
- reporting
  - find the most interesting sets
Statistical issues

- choice of $u(s, c)$
  - selection
  - averaging
  - first, second-order
  - other
Statistical issues

- choice of $u(s, c)$
  - selection
  - averaging
  - first, second-order
  - other

- calibration
  - $u(s, C)$ competitive, random set
  - $u(S, c)$ self-contained
Statistical issues

- choice of $u(s, c)$
  
  selection
  
  averaging
  
  first, second-order
  
  other

- calibration

  $u(s, C)$ competitive, random set

  $u(S, c)$ self-contained

- random-set calibration

  global universe

  local universe
Gene ontology: www.geneontology.org

- networks of defined terms describing gene product attributes

- term: antigen presentation; endogenous antigen
  - id: GO:0019883
  - definition: the process by which antigen-presenting cells express self antigen on their surface in a form recognizable by lymphocytes
  - m = 48 probe sets on Affymetrix hgu133plus2 microarray

- lineage:
  - GO:0019882: antigen presentation (98)
  - GO:0006955: immune response (1494)
  - GO:0007582: organismal physiological process (21069)

- Newton Enrichment
Gene ontology: www.geneontology.org

- networks of **defined terms** describing gene product attributes
- example set c
  - term: antigen presentation; endogenous antigen
Gene ontology: www.geneontology.org

- networks of **defined terms** describing gene product attributes
- example set c
  - term: antigen presentation; endogenous antigen
  - id: GO:0019883
Gene ontology: www.geneontology.org

- networks of **defined terms** describing gene product attributes
- example set c
  - term: antigen presentation; endogenous antigen
  - id: GO:0019883
  - definition: *the process by which antigen-presenting cells express self antigen on their surface in a form recognizable by lymphocytes*
Gene ontology: www.geneontology.org

- networks of **defined terms** describing gene product attributes
- example set c
  - term: antigen presentation; endogenous antigen
  - id: GO:0019883
  - definition: *the process by which antigen-presenting cells express self antigen on their surface in a form recognizable by lymphocytes*
  - \( m = 48 \) probe sets on Affymetrix hgu133plus2 microarray
Gene ontology: www.geneontology.org

- networks of defined terms describing gene product attributes
- example set
  - term: antigen presentation; endogenous antigen
  - id: GO:0019883
  - definition: the process by which antigen-presenting cells express self antigen on their surface in a form recognizable by lymphocytes
  - $m = 48$ probe sets on Affymetrix hgu133plus2 microarray
  - lineage:
    - GO:0019882: antigen presentation (98)
    - GO:0006955: immune response (1494)
    - GO:0007582: organismal physiological process (21069)
Gene ontology: www.geneontology.org

- networks of **defined terms** describing gene product attributes
- example set: c
  - term: antigen presentation; endogenous antigen
  - id: GO:0019883
  - definition: *the process by which antigen-presenting cells express self antigen on their surface in a form recognizable by lymphocytes*
  - $m = 48$ probe sets on Affymetrix hgu133plus2 microarray
  - lineage:
    - GO:0019882: antigen presentation (98)
    - GO:0006955: immune response (1494)
    - GO:0007582: organismal physiological process (21069)
  - $\exists$ many gene sets
More on GO

[Diagram]

- GO level
- # genes per GO term
- GO terms per gene
- max GO level

Newton Enrichment
Case study: Nasopharyngeal carcinoma (NPC)

Scatterplot of one host/one virus: rank transformed

$cor = -0.62$
Many host genes are negatively correlated with EBNA1
Permutation analysis indicates significant association

Minimum correlation

fraction negatively correlated

Frequency

0.3 0.4 0.5 0.6 0.7

Newton Enrichment
Connecting to gene sets: enrichment

gene scores: \( s_g = -\text{atanh} \left( \text{Spearman} \ r_g \right) \)

selection: count extreme-scoring, interesting genes

\[ u_{\text{sel}}(s, c) = \frac{1}{m} \sum_{g \in c} 1[s_g > k] \]

averaging: combine evidence from all genes

\[ u_{\text{ave}}(s, c) = \frac{1}{m} \sum_{g \in c} s_g \]
Selection: identify interesting genes

Gene list targets 5% FDR using q-value method and $\text{atanh}(r) \sim \text{Gaussian}$
Selection: cross classify genes

<table>
<thead>
<tr>
<th>Select</th>
<th>yes</th>
<th>no</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0019883</td>
<td>8</td>
<td>40</td>
</tr>
<tr>
<td>other</td>
<td>566</td>
<td>54061</td>
</tr>
</tbody>
</table>

\[ \text{sel}(s, c) = \frac{8}{48} \]

\[ \text{sel}(s, \mathcal{C}) \sim \frac{1}{m} \text{Hypergeometric} \]

Fisher's \( p = 3.7 \times 10^{-8} \)

Common practice: NetAffx; GOHyperG; Fatigo; Gostat
Selection: cross classify genes

<table>
<thead>
<tr>
<th></th>
<th>Select</th>
<th>rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>GO:0019883</td>
<td>8</td>
<td>40</td>
</tr>
<tr>
<td>other</td>
<td>566</td>
<td>54061</td>
</tr>
<tr>
<td></td>
<td>574</td>
<td>54101</td>
</tr>
</tbody>
</table>

\[ u_{sel}(s, c) = \frac{8}{48} \]
Selection: cross classify genes

<table>
<thead>
<tr>
<th>Select</th>
<th>rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>GO:0019883 8 40</td>
<td>$m = 48$ 0.17</td>
</tr>
<tr>
<td>other 566 54061</td>
<td>54627 0.01</td>
</tr>
<tr>
<td>574 54101</td>
<td>54675</td>
</tr>
</tbody>
</table>

$$u_{sel}(s, c) = \frac{8}{48}$$

$$u_{sel}(s, C) \sim \frac{1}{m}\text{Hypergeometric}$$
Selection: cross classify genes

<table>
<thead>
<tr>
<th>Select</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>GO:0019883</td>
<td>8</td>
</tr>
<tr>
<td>other</td>
<td>566</td>
</tr>
<tr>
<td></td>
<td>574</td>
</tr>
</tbody>
</table>

\[
u_{\text{sel}}(s, c) = \frac{8}{48}
\]

\[
u_{\text{sel}}(s, C) \sim \left(\frac{1}{m}\right)\text{Hypergeometric}
\]

Fisher’s \( p = 3.7 \times 10^{-8} \)
Selection: cross classify genes

<table>
<thead>
<tr>
<th>Select</th>
<th>Rate</th>
<th>m</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>no</td>
<td>8</td>
<td>40</td>
</tr>
<tr>
<td>GO:0019883</td>
<td>566</td>
<td>54061</td>
<td>54627</td>
</tr>
<tr>
<td>other</td>
<td></td>
<td>574</td>
<td>54101</td>
</tr>
</tbody>
</table>

\[ u_{sel}(s, c) = \frac{8}{48} \]

\[ u_{sel}(s, C) \sim \left( \frac{1}{m} \right) \text{Hypergeometric} \]

Fisher’s \( p = 3.7 \times 10^{-8} \)

common practice: NetAffx; GOHyperG; Fatigo; Gostat
Equivalently

| gene selection | 1 | 2 | 3 | \cdots | \cdots | G = 54675 |
|----------------|---|---|---|\cdots | \cdots | 0           |
| set c          | 1 | 1 | 1 | \cdots | 1 | 0 | 0 | \cdots | 0 | 574 |
| on both        | 0 | 1 | 0 | \cdots | 1 | 1 | 0 | \cdots | 1 | 48 |
|                | 1 | 1 |   |         |   |   | 1 |         |   | 8/48 |
Equivalently

<table>
<thead>
<tr>
<th>gene</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>...</th>
<th>(G = 54675)</th>
</tr>
</thead>
<tbody>
<tr>
<td>selection</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>set (C)</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>on both</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

\(m = 48\)

\(u(s, C) = 8/48\)

Hypergeometric is obtained by shuffling a row

<table>
<thead>
<tr>
<th>gene</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>...</th>
<th>(G = 54675)</th>
</tr>
</thead>
<tbody>
<tr>
<td>selection</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>set (C)</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>on both</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

\(m = 48\)

\(u(s, C)\)
Equivalently

<table>
<thead>
<tr>
<th>gene</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>...</th>
<th>$G = 54675$</th>
</tr>
</thead>
<tbody>
<tr>
<td>selection</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>set $c$</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>on both</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Hypergeometric is obtained by shuffling a row

<table>
<thead>
<tr>
<th>gene</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>...</th>
<th>$G = 54675$</th>
</tr>
</thead>
<tbody>
<tr>
<td>selection</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>set $C$</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>on both</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$C$ is a **random set**. Uniform among $\binom{G}{m}$ size $m$ subsets of genome.
Beyond hypergeometric

\[
\begin{array}{cccccc}
gene & 1 & 2 & 3 & \cdots & G = 54675 \\
gene score & s_1 & s_2 & s_3 & \cdots & s_G \\
C & 0 & 1 & 0 & \cdots & 1 & 0 & 0 & \cdots & 0 \\
add & s_2 & s_g & & & \\
\end{array}
\]

\[m = 48\]

\[u(s, C)\]

where \[u(s, C) = \frac{1}{m} \sum_{g \in C} s_g.\]
Beyond hypergeometric

<table>
<thead>
<tr>
<th>gene score</th>
<th>s₁</th>
<th>s₂</th>
<th>s₃</th>
<th>⋯</th>
<th>s₇</th>
</tr>
</thead>
<tbody>
<tr>
<td>gene</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>⋯</td>
<td>G = 54675</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>⋯</td>
<td>1</td>
</tr>
<tr>
<td>add</td>
<td>s₂</td>
<td>s₇</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

where \( u(s, C) = \frac{1}{m} \sum_{g \in C} s_g \).

\[
\mu = E\{u(s, C)\} = \frac{\sum_{g=1}^{G} s_g}{G}
\]

\[
\sigma^2 = \text{var}\{u(s, C)\} = \left(\frac{1}{m} - \frac{1}{G}\right) \left(\frac{\sum_{g=1}^{G} (s_g - \bar{s})^2}{G - 1}\right)
\]

Standardized set score:
\[
z(s, c) = \frac{u(s, C) - \mu}{\sigma}
\]
Beyond hypergeometric

<table>
<thead>
<tr>
<th>gene</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>...</th>
<th>G = 54675</th>
</tr>
</thead>
<tbody>
<tr>
<td>gene score</td>
<td>s₁</td>
<td>s₂</td>
<td>s₃</td>
<td>...</td>
<td>sₖ</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>add</td>
<td>s₂</td>
<td>sₖ</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

where \( u(s, C) = \frac{1}{m} \sum_{g \in C} s_g \).

\[
\mu = E\{u(s, C)\} = \frac{\sum_{g=1}^{G} s_g}{G}
\]

\[
\sigma^2 = var\{u(s, C)\} = \left(\frac{1}{m} - \frac{1}{G}\right) \left(\frac{\sum_{g=1}^{G} (s_g - \bar{s})^2}{G - 1}\right)
\]

Standardized set score:

\[
z(s, c) = [u(s, c) - \mu]/\sigma
\]
2761 GO sets with $m \geq 10$
Sets with extreme $Z$

Selection: The number of *interesting* genes is extreme

Averaging: The average *interestingness* of genes is extreme
Look inside GO:0019883

GO class “antigen presentation, endogenous antigen”
(all 42 probesets)

31 tumors ranked by EBNA1 expression

low

HLA

A B C E F G

HFE

TAP2

highest

gene expression level

lowest
## Bio followup

<table>
<thead>
<tr>
<th>RT PCR</th>
<th>EBV transfection experiment</th>
</tr>
</thead>
</table>

**Graph A**

- **x-axis**: tumors grouped by EBNA1 RNA content
- **y-axis**: molecules

- Red line: HLA-A & HLA-F
- Blue line: EBNA1

**Graph 2**

- **x-axis**: normalized cell counts (log scale)
- **y-axis**: 0 to 100

- Red line: EBV+
- Blue line: EBV-

**Legend**

- 293 cells
Theoretical comparison: selection versus averaging

Set scores:

\[ u_{\text{ave}}(s, c) = \frac{1}{m} \sum_{g \in c} s_g \]

averaging

\[ u_{\text{sel}}(s, c) = \frac{1}{m} \sum_{g \in C} 1[s_g > k] \]

selection
Location model: $s_g = \delta l_g + \epsilon_g$

- $l_g$: latent Bernoulli indicator that $g$ is interesting
Location model: \( s_g = \delta I_g + \epsilon_g \)

- \( I_g \): latent Bernoulli indicator that \( g \) is interesting
- \( \delta > 0 \): expected score of truly interesting genes
Location model: \( s_g = \delta l_g + \epsilon_g \)

- \( l_g \): latent Bernoulli indicator that \( g \) is interesting
- \( \delta > 0 \): expected score of truly interesting genes
- \( \epsilon_g \): standard normal error
Location model: $s_g = \delta l_g + \epsilon_g$

- $l_g$: latent Bernoulli indicator that $g$ is interesting
- $\delta > 0$: expected score of truly interesting genes
- $\epsilon_g$: standard normal error
- $\tilde{\alpha}$: FDR of selected list
Location model: $s_g = \delta l_g + \epsilon_g$

- $l_g$: latent Bernoulli indicator that $g$ is interesting
- $\delta > 0$: expected score of truly interesting genes
- $\epsilon_g$: standard normal error
- $\tilde{\alpha}$: FDR of selected list
- $\pi$: proportion of interesting genes in system
Location model: $s_g = \delta l_g + \epsilon_g$

- $l_g$: latent Bernoulli indicator that $g$ is interesting
- $\delta > 0$: expected score of truly interesting genes
- $\epsilon_g$: standard normal error
- $\tilde{\alpha}$: FDR of selected list
- $\pi$: proportion of interesting genes in system
- $\pi_c$: proportion of interesting genes in $c$
Location model: $s_g = \delta l_g + \epsilon_g$

- $l_g$: latent Bernoulli indicator that $g$ is interesting
- $\delta > 0$: expected score of truly interesting genes
- $\epsilon_g$: standard normal error
- $\tilde{\alpha}$: FDR of selected list
- $\pi$: proportion of interesting genes in system
- $\pi_c$: proportion of interesting genes in $c$
- enrichment: $\pi_c - \pi$
Location model: \( s_g = \delta l_g + \epsilon_g \)

- \( l_g \): latent Bernoulli indicator that \( g \) is interesting
- \( \delta > 0 \): expected score of truly interesting genes
- \( \epsilon_g \): standard normal error
- \( \tilde{\alpha} \): FDR of selected list
- \( \pi \): proportion of interesting genes in system
- \( \pi_c \): proportion of interesting genes in \( c \)
- enrichment: \( \pi_c - \pi \)
- \( H_0: \pi_c = \pi \)
Testing enrichment: averaging

Test: Reject null for large $\bar{X}_{ave} = \frac{1}{m} \sum_{g \in c} s_g$
Testing enrichment: averaging

- Test: Reject null for large $$\bar{X}_{\text{ave}} = \frac{1}{m} \sum_{g \in c} s_g$$

- Power: $$1 - \Phi(\tau_{\text{ave}})$$ for standard normal cdf $$\Phi$$
  where $$\tau_{\text{ave}} = z_\alpha - \sqrt{m} \left( \pi_c - \pi \right)$$

  - enrichment
  - effect
Power: $\pi = .2$, $m = 20$, $\alpha = .05$
Testing enrichment: selection

Test: Reject null for large

$$\bar{X}_{sel} = \frac{1}{m} \sum_{g \in c} 1[s_g > k]$$
Testing enrichment: selection

- **Test:** Reject null for large
  \[ \bar{X}_{sel} = \frac{1}{m} \sum_{g \in c} 1[s_g > k] \]

- **Power:** \(1 - \Phi(\tau_{sel})\)

  where \( \tau_{sel} = z_{\alpha} \frac{\sigma(\pi)}{\sigma(\pi_c)} - \sqrt{m} \left( \pi_c - \pi \right) \left( \frac{\mu_1 - \mu_0}{\sigma(\pi_c)} \right) \)

  - \(z_{\alpha}\): Normalized enrichment
  - \(\sigma(\pi)/\sigma(\pi_c)\): f(effect)
Power: $\pi = .2, \ m = 20, \ \alpha = \tilde{\alpha} = .05$
Power Comparison

A. Power of selection

B. Power of averaging

C. Selection better

D. Averaging better
Put $0 < \kappa < 1$ where

\[ \kappa = \frac{\tilde{\alpha}}{1 - \tilde{\alpha}} \frac{\pi}{1 - \pi} \]

- Selection is more powerful than averaging if $m$ is sufficiently large and

\[ 2\Phi^{-1} \left( \frac{1}{1 + \kappa} \right) < \delta < \frac{1}{\sqrt{\kappa}} - \sqrt{\kappa} \]

- Averaging is more powerful than selection if $m$ is sufficiently large and

\[ 0 < \delta < \delta^*(\pi, \pi_c) \]
Other approaches: GSEA/SAFE

▶ SAFE: Virteniva et al. 2001; Barry et al. 2005
GSEA: Mootha et al. 2003; Subramanian et al. 2005
▶ Retain quantitative gene-level scores
▶ Combine to form category-level score
▶ Calibrate by label permutation
▶ Problems:
  ▶ structural
  ▶ computational
  ▶ inferential
SAFE and random sets on GO:0019883
SAFE and random sets: p-values, 2761 categories
Gene set enrichment analysis

Order gene level scores $s(1) \leq s(2) \leq s(G)$

$$a_g = \frac{\sum_{i=1}^{g} |s(i)|^p 1[i \in c]}{\sum_{i \in c} |s_g|^p} - \frac{\sum_{i=1}^{g} (1 - 1[i \in c])}{G - m}$$

and then construct enrichment score

$$ES = \max_g |a_g|$$
GSEA

Newton

Enrichment
Results show GSEA has no power in NPC data.

\[ p_v(GO:0019882) = 0.022 \]

\[ FDR = 0.7! \]

\[ \min(FDR) = 0.28 \]
Efron and Tibshirani’s maxmean

Let

$$s^+ = \frac{\sum_{g \in c} s_g 1[s_g > 0]}{\sum_{g \in c} 1[s_g > 0]}$$

and

$$s^- = \frac{\sum_{g \in c} s_g 1[s_g \leq 0]}{\sum_{g \in c} 1[s_g \leq 0]}$$

and use

$$u(s, c) = \max(s^+, -s^-)$$
Decorrelating enrichment scores: local vs global universe

- Sets \{c\} in GO have extensive overlap [DAG]

Correlation can be problematic [inheritance problem]

Solution: in standardizing \(u(s, c)\), restrict to random subsets of parent \(c'\)

Fact:

\[
\text{corr}\{z(s, \text{local}, C), z(s, \text{local}, C')\} = 0
\]
Decorrelating enrichment scores: local vs global universe

- Sets \( \{c\} \) in GO have extensive overlap [DAG]
- If \( C_1 \) and \( C_2 \) are two random sets of fixed sizes \( m_1 \) and \( m_2 \) and overlap \( m_{1,2} \), then

\[
\text{corr} \{ z(s, C_1), z(s, C_2) \} \approx \frac{m_{1,2}}{\sqrt{m_1 m_2}}
\]

Correlation can be problematic [inheritance problem]

Solution: in standardizing \( u(s, c) \), restrict to random subsets of parent \( c' \)

Fact:

\[
\text{corr} \{ z(s, C), z(s, C') \} = 0
\]
Decorrelating enrichment scores: local vs global universe

- Sets \{c\} in GO have extensive overlap [DAG]
- If \(C_1\) and \(C_2\) are two random sets of fixed sizes \(m_1\) and \(m_2\) and overlap \(m_{1,2}\), then
  \[
  \text{corr} \{z(s, C_1), z(s, C_2)\} \approx \frac{m_{1,2}}{\sqrt{m_1 m_2}}
  \]
- Correlation can be problematic [inheritance problem]
Decorrelating enrichment scores: local vs global universe

- Sets \{c\} in GO have extensive overlap [DAG]
- If \(C_1\) and \(C_2\) are two random sets of fixed sizes \(m_1\) and \(m_2\) and overlap \(m_{1,2}\), then

\[
corr\{z(s, C_1), z(s, C_2)\} \approx \frac{m_{1,2}}{\sqrt{m_1 m_2}}
\]

- Correlation can be problematic [inheritance problem]
- Solution: in standardizing \(u(s, c)\), restrict to random subsets of parent \(c'\)
Decorrelating enrichment scores: local vs global universe

- Sets \( \{c\} \) in GO have extensive overlap [DAG]
- If \( C_1 \) and \( C_2 \) are two random sets of fixed sizes \( m_1 \) and \( m_2 \) and overlap \( m_{1,2} \), then
  \[
  \text{corr} \left\{ z(s, C_1), z(s, C_2) \right\} \approx \frac{m_{1,2}}{\sqrt{m_1 m_2}}
  \]
- Correlation can be problematic [inheritance problem]
- Solution: in standardizing \( u(s, c) \), restrict to random subsets of parent \( c' \)
- Fact: \( \text{corr} \left\{ z(s_{\text{local}}, C), z(s_{\text{local}}, C') \right\} = 0 \)
Decorrelation by local calibration GO [HPV example]
Multiple categories: the balance of power

\[ E(Z_{\text{ave}}) = \sqrt{m} \left( \pi_C - \pi \right) \delta \]

enrichment effect
Second-order enrichment [HPV example]
Second-order enrichment [simulation]

- Variance
- Mean absolute
- Max−mean
- GSEA
The hypergeometric assessment of enrichment is one in a class of conditional tests.
Summary

- The hypergeometric assessment of enrichment is one in a class of conditional tests.
- Selection and averaging measure different aspects of the enrichment signal.
The hypergeometric assessment of enrichment is one in a class of conditional tests.

Selection and averaging measure different aspects of the enrichment signal.

Local scoring decorrelates enrichment scores in GO.
Summary

- The hypergeometric assessment of enrichment is one in a class of conditional tests.
- Selection and averaging measure different aspects of the enrichment signal.
- Local scoring decorrelates enrichment scores in GO.
- Sampling theory meets genomics.
The hypergeometric assessment of enrichment is one in a class of conditional tests.

Selection and averaging measure different aspects of the enrichment signal.

Local scoring decorrelates enrichment scores in GO.

Sampling theory meets genomics.

Comparisons … statistics/calibrations.
Summary

- The hypergeometric assessment of enrichment is one in a class of conditional tests.
- Selection and averaging measure different aspects of the enrichment signal.
- Local scoring decorrelates enrichment scores in GO.
- Sampling theory meets genomics.
- Comparisons ... statistics/calibrations.
- R package *allez*.
Enrichment collaboration

<table>
<thead>
<tr>
<th>NPC</th>
<th>Johan den Boon, Srikumar Sengupta, and Paul Ahlquist</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mixtures</td>
<td>Fernando Quintana</td>
</tr>
<tr>
<td>R</td>
<td>Deepayan Sarkar</td>
</tr>
<tr>
<td>$$</td>
<td>NCI</td>
</tr>
</tbody>
</table>
On threshold $k$ to get FDR $\tilde{\alpha}$

$$\tilde{\alpha} = P(I_g = 0| s_g > k) = \frac{\mu_0(1 - \pi)}{\mu_0(1 - \pi) + \mu_1 \pi}$$

So with $h(x) = [1 - \Phi(x - \delta)] / [1 - \Phi(x)]$, we have $k = h^{-1}(\kappa)$ where $\kappa = \pi \tilde{\alpha} / [(1 - \pi)(1 - \tilde{\alpha})]$. 
Genes versus probe sets

Probe sets per gene

Ideal: collapse probe scores to genes
compute $Z^*$

Simple: adjust $Z^* = Z_{\text{sqrt}\{m^*/m\}}$

Ideal $Z_{\text{q}}$ vs naive $Z_{\text{q}}$

Ideal $Z_{\text{q}}$ vs adjusted $Z_{\text{q}}$