Canadian Bioinformatics Workshops
www.bioinformatics.ca
Module 2
Finding over-represented pathways in gene lists

Quaid Morris
Pathway and Network Analysis of -omics Data
June 4-6, 2014

Random draws

... 7,834 draws later ...

Background population:
500 black genes,
4500 red genes

Learning Objectives of Module 2

• Be able to select the appropriate enrichment test for your data.
• Be able to determine the appropriate background gene list when running Fisher’s Exact Test (aka Hypergeometric test).
• Be able to determine when you need a multiple test correction.
• Be able to select whether to use a Bonferroni corrected P-value or a false discovery rate.
• Be able to explain, in plain language, how you calculate each correction.
• Understand how P-values can be computed by resampling.
Outline

• Introduction to enrichment analysis
• Hypergeometric Test, aka Fisher’s Exact Test
• GSEA enrichment analysis for ranked lists.
• Multiple test corrections:
  – Bonferroni correction
  – False Discovery Rate computation using Benjamini-Hochberg procedure

Types of enrichment analysis

• **Gene list** (e.g. expression change > 2-fold)
  – Answers the question: Are any gene sets surprisingly enriched (or depleted) in my gene list?
  – Statistical test: Fisher’s Exact Test (aka Hypergeometric test)

• **Ranked list** (e.g. by differential expression)
  – Answers the question: Are any gene set ranked surprisingly high or low in my ranked list of genes?
  – Statistical test: GSEA (+ others we won’t discuss)
Gene list enrichment analysis

- **Given:**
  1. Gene list: e.g. RRP6, MRD1, RRP7, RRP43, RRP42 (yeast)
  2. Gene sets or annotations: e.g. Gene ontology, transcription factor binding sites in promoter
- **Question:** Are any of the gene annotations **surprisingly enriched in the gene list?**
- **Details:**
  - Where do the gene lists come from?
  - How to assess “surprisingly” (statistics)
  - How to correct for repeating the tests
Two-class design for gene lists

Expression Matrix

Genes Ranked by Differential Statistic

Selection by Threshold

E.g.:
- Fold change
- Log (ratio)
- t-test
- Significance analysis of microarrays

Time-course design for gene lists

Expression Matrix

Gene Clusters

E.g.:
- K-means
- K-medoids
- SOM

Each cluster is a separate gene list

Module 2 bioinformatics.ca
Example gene list enrichment test

Microarray Experiment (gene expression table) → Gene list (e.g. UP-regulated) → Background (all genes on the array) → Gene-set Databases

Example gene list enrichment test

Microarray Experiment (gene expression table) → Gene list (e.g. UP-regulated) → Background (all genes on the array) → Gene-set → Gene-set Databases

Module 2 bioinformatics.ca
Enrichment Test

The output of an enrichment test is a P-value

The P-value assesses the probability that the overlap is at least as large as observed by random sampling the array genes.

Recipe for gene list enrichment test

- **Step 1:** Define your gene list and your background list,
- **Step 2:** Select your gene sets to test for enrichment,
- **Step 3:** Run enrichment tests and correct for multiple testing, if necessary,
- **Step 4:** Interpret your enrichments
- **Step 5:** Publish! ;)

Module 2

bioinformatics.ca
Why test enrichment in ranked lists?

• Possible problems with gene list test
  – No “natural” value for the threshold
  – Different results at different threshold settings
  – Possible loss of statistical power due to thresholding
    • No resolution between significant signals with different strengths
    • Weak signals neglected

Example ranked list enrichment test

<table>
<thead>
<tr>
<th>Gene-set</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spindle</td>
<td>0.0001</td>
</tr>
<tr>
<td>Apoptosis</td>
<td>0.025</td>
</tr>
</tbody>
</table>
Recipe for gene list enrichment test

- **Step 1:** Rank your gene list,
- **Step 2:** Select your gene sets to test for enrichment,
- **Step 3:** Run enrichment tests and correct for multiple testing, if necessary,
- **Step 4:** Interpret your enrichments
- **Step 5:** Publish! ;)

Outline of theory component

- Hypergeometric test for calculating enrichment P-values for gene lists
- GSEA for computing enrichment P-values for ranked lists
- Multiple test corrections:
  - Bonferroni
  - Benjamini-Hochberg FDR
The hypergeometric test
a.k.a., Fisher’s exact test

Gene list
- RRP6
- MRD1
- RRP7
- RRP43
- RRP42

Null hypothesis: List is a random sample from population
Alternative hypothesis: More black genes than expected

Background population:
500 black genes,
4500 red genes

---

The hypergeometric test
a.k.a., Fisher’s exact test

Gene list
- RRP6
- MRD1
- RRP7
- RRP43
- RRP42

Null distribution

P-value

Answer = 4.6 x 10^{-4}

Background population:
500 black genes,
4500 red genes

---
2x2 contingency table for Fisher’s Exact Test

<table>
<thead>
<tr>
<th>Gene list</th>
<th>In gene list</th>
<th>Not in gene list</th>
</tr>
</thead>
<tbody>
<tr>
<td>In gene set</td>
<td>4</td>
<td>496</td>
</tr>
<tr>
<td>Not in gene set</td>
<td>1</td>
<td>4499</td>
</tr>
</tbody>
</table>

Background population:
500 black genes,
4500 red genes

e.g.: http://www.graphpad.com/quickcalc/contingency1.cfm

Important details

- To test for under-enrichment of “black”, test for over-enrichment of “red”.
- Need to choose “background population” appropriately, e.g., if only portion of the total gene complement is queried (or available for annotation), only use that population as background.
- To test for enrichment of more than one independent types of annotation (red vs black and circle vs square), apply Fisher’s exact test separately for each type. ***More on this later***
Other enrichment tests

Gene list
- Fisher’s Exact Test, Binomial and Chi-squared.

Ranked list (semi-quantitative)
- GSEA, Wilcoxon ranksum, Mann-Whitney U, Kolmogorov-Smirnov

GSEA: Method

Steps
1. Calculate the ES score
2. Generate the ES distribution for the null hypothesis using permutations
   - see permutation settings
3. Calculate the empirical p-value

GSEA: Method

Where are the gene-set genes located in the ranked list?
Is there distribution random, or is there an enrichment in either end?

Every present gene (black vertical bar) gives a positive contribution,
every absent gene (no vertical bar) gives a negative contribution
**GSEA: Method**

**ES score calculation**

MAX (absolute value) running ES score $\Rightarrow$ Final ES Score

**Module 2**

**GSEA: Method**

**ES score calculation**

High ES score $\leftrightarrow$ High local enrichment

**Module 2**
GSEA: Method

Empirical p-value estimation (for every gene-set)

1. Generate null-hypothesis distribution from randomized data (see permutation settings)

![Distribution of ES from N permutations (e.g. 2000)]

GSEA: Method

Estimate empirical p-value by comparing observed ES score to null-hypothesis distribution from randomized data (for every gene-set)

![Distribution of ES from N permutations (e.g. 2000)]
GSEA: Method

Estimate empirical p-value by comparing observed ES score to null-hypothesis distribution from randomized data (for every gene-set)

More GSEA examples

Ranked gene list

Gene-set 1 (n=17) Enriched in treated

Gene-set 2 Not enriched

Gene-set 3 (n=22) Depleted in treated
Multiple test corrections

How to win the P-value lottery, part 1

Random draws

... 7,834 draws later ...

Expect a random draw with observed enrichment once every 1 / P-value draws

Background population:
500 black genes,
4500 red genes
How to win the P-value lottery, part 2
Keep the gene list the same, evaluate different annotations

![Observed draw and Different annotation](image)

Simple P-value correction: Bonferroni

If \( M = \# \) of annotations tested:

Corrected P-value = \( M \times \) original P-value

Corrected P-value is greater than or equal to the probability that one or more of the observed enrichments could be due to random draws. The jargon for this correction is “controlling for the Family-Wise Error Rate (FWER)”
Bonferroni correction caveats

- Bonferroni correction is very stringent and can “wash away” real enrichments leading to false negatives,
- Often one is willing to accept a less stringent condition, the “false discovery rate” (FDR), which leads to a gentler correction when there are real enrichments.

False discovery rate (FDR)

- FDR is the expected proportion of the observed enrichments due to random chance.
- Compare to Bonferroni correction which is a bound on the probability that any one of the observed enrichments could be due to random chance.
- Typically FDR corrections are calculated using the Benjamini-Hochberg procedure.
- FDR threshold is often called the “q-value”
### Benjamini-Hochberg example

<table>
<thead>
<tr>
<th>Rank</th>
<th>Category</th>
<th>P-value</th>
<th>Adjusted P-value</th>
<th>FDR / Q-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Transcriptional regulation</td>
<td>0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Transcription factor</td>
<td>0.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Initiation of transcription</td>
<td>0.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>Nuclear localization</td>
<td>0.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>51</td>
<td>RNAi activity</td>
<td>0.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>52</td>
<td>Cytoplasmic localization</td>
<td>0.06</td>
<td></td>
<td></td>
</tr>
<tr>
<td>53</td>
<td>Translation</td>
<td>0.07</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Sort P-values of all tests in decreasing order**

### Benjamini-Hochberg example

<table>
<thead>
<tr>
<th>Rank</th>
<th>Category</th>
<th>P-value</th>
<th>Adjusted P-value</th>
<th>FDR / Q-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Transcriptional regulation</td>
<td>0.001</td>
<td>0.001 x 53/1 = 0.053</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Transcription factor</td>
<td>0.01</td>
<td>0.01 x 53/2 = 0.27</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Initiation of transcription</td>
<td>0.02</td>
<td>0.02 x 53/3 = 0.35</td>
<td></td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>Nuclear localization</td>
<td>0.04</td>
<td>0.04 x 53/50 = 0.042</td>
<td></td>
</tr>
<tr>
<td>51</td>
<td>RNAi activity</td>
<td>0.05</td>
<td>0.05 x 53/51 = 0.052</td>
<td></td>
</tr>
<tr>
<td>52</td>
<td>Cytoplasmic localization</td>
<td>0.06</td>
<td>0.06 x 53/52 = 0.061</td>
<td></td>
</tr>
<tr>
<td>53</td>
<td>Translation</td>
<td>0.07</td>
<td>0.07 x 53/53 = 0.07</td>
<td></td>
</tr>
</tbody>
</table>

**Adjusted P-value = P-value \times \frac{\# \text{ of tests}}{\text{Rank}}**
**Benjamini-Hochberg example**

<table>
<thead>
<tr>
<th>Rank</th>
<th>Category</th>
<th>P-value</th>
<th>Adjusted P-value</th>
<th>FDR / Q-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Transcriptional regulation</td>
<td>0.001</td>
<td>0.001 x 53/1 = 0.053</td>
<td>0.042</td>
</tr>
<tr>
<td>2</td>
<td>Transcription factor</td>
<td>0.01</td>
<td>0.01 x 53/2 = 0.27</td>
<td>0.042</td>
</tr>
<tr>
<td>3</td>
<td>Initiation of transcription</td>
<td>0.02</td>
<td>0.02 x 53/3 = 0.35</td>
<td>0.042</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>50</td>
<td>Nuclear localization</td>
<td>0.04</td>
<td>0.04 x 53/50 = 0.042</td>
<td>0.042</td>
</tr>
<tr>
<td>51</td>
<td>RNAi activity</td>
<td>0.05</td>
<td>0.05 x 53/51 = 0.052</td>
<td>0.052</td>
</tr>
<tr>
<td>52</td>
<td>Cytoplasmic localization</td>
<td>0.06</td>
<td>0.06 x 53/52 = 0.061</td>
<td>0.061</td>
</tr>
<tr>
<td>53</td>
<td>Translation</td>
<td>0.07</td>
<td>0.07 x 53/53 = 0.07</td>
<td>0.07</td>
</tr>
</tbody>
</table>

**Q-value = minimum adjusted P-value at given rank or below**

---

**Benjamini-Hochberg example I**

<table>
<thead>
<tr>
<th>Rank</th>
<th>Category</th>
<th>P-value</th>
<th>Adjusted P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Transcriptional regulation</td>
<td>0.001</td>
<td>0.001 x 53/1 = 0.053</td>
</tr>
<tr>
<td>2</td>
<td>Transcription factor</td>
<td>0.002</td>
<td>0.002 x 53/2 = 0.053</td>
</tr>
<tr>
<td>3</td>
<td>Initiation of transcription</td>
<td>0.003</td>
<td>0.003 x 53/3 = 0.053</td>
</tr>
<tr>
<td>4</td>
<td>Nuclear localization</td>
<td>0.0031</td>
<td>0.0031 x 53/4 = 0.040</td>
</tr>
<tr>
<td>5</td>
<td>Chromatin modification</td>
<td>0.005</td>
<td>0.005 x 53/5 = 0.053</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>52</td>
<td>Cytoplasmic localization</td>
<td>0.97</td>
<td>0.985 x 53/52 = 1.004</td>
</tr>
<tr>
<td>53</td>
<td>Translation</td>
<td>0.99</td>
<td>0.99 x 53/53 = 0.99</td>
</tr>
</tbody>
</table>

**Adjusted P-value is “nominal” P-value times # of tests divided by the rank of the P-value in sorted list**
### Benjamini-Hochberg example II

<table>
<thead>
<tr>
<th>Rank</th>
<th>Category</th>
<th>P-value</th>
<th>Adjusted P-value</th>
<th>FDR / Q-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Transcriptional regulation</td>
<td>0.001</td>
<td>0.001 x 53/1 = 0.053</td>
<td>0.040</td>
</tr>
<tr>
<td>2</td>
<td>Transcription factor</td>
<td>0.002</td>
<td>0.002 x 53/2 = 0.053</td>
<td>0.040</td>
</tr>
<tr>
<td>3</td>
<td>Initiation of transcription</td>
<td>0.003</td>
<td>0.003 x 53/3 = 0.053</td>
<td>0.040</td>
</tr>
<tr>
<td>4</td>
<td>Nuclear localization</td>
<td>0.0031</td>
<td>0.0031 x 53/4 = 0.040</td>
<td>0.040</td>
</tr>
<tr>
<td>5</td>
<td>Chromatin modification</td>
<td>0.005</td>
<td>0.005 x 53/5 = 0.053</td>
<td>0.053</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>52</td>
<td>Cytoplasmic localization</td>
<td>0.97</td>
<td>0.985 x 53/52 = 1.004</td>
<td>0.99</td>
</tr>
<tr>
<td>53</td>
<td>Translation</td>
<td>0.99</td>
<td>0.99 x 53/53 = 0.99</td>
<td>0.99</td>
</tr>
</tbody>
</table>

**Q-value (or FDR) corresponding to a nominal P-value is the smallest adjusted P-value assigned to P-values with the same or larger ranks.**

### Benjamini-Hochberg example III

<table>
<thead>
<tr>
<th>Rank</th>
<th>Category</th>
<th>P-value</th>
<th>Adjusted P-value</th>
<th>FDR / Q-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Transcriptional regulation</td>
<td>0.001</td>
<td>0.001 x 53/1 = 0.053</td>
<td>0.040</td>
</tr>
<tr>
<td>2</td>
<td>Transcription factor</td>
<td>0.002</td>
<td>0.002 x 53/2 = 0.053</td>
<td>0.040</td>
</tr>
<tr>
<td>3</td>
<td>Initiation of transcription</td>
<td>0.003</td>
<td>0.003 x 53/3 = 0.053</td>
<td>0.040</td>
</tr>
<tr>
<td>4</td>
<td>Nuclear localization</td>
<td>0.0031</td>
<td>0.0031 x 53/4 = 0.040</td>
<td>0.040</td>
</tr>
<tr>
<td>5</td>
<td>Chromatin modification</td>
<td>0.005</td>
<td>0.005 x 53/5 = 0.053</td>
<td>0.053</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>52</td>
<td>Cytoplasmic localization</td>
<td>0.97</td>
<td>0.985 x 53/52 = 1.004</td>
<td>0.99</td>
</tr>
<tr>
<td>53</td>
<td>Translation</td>
<td>0.99</td>
<td>0.99 x 53/53 = 0.99</td>
<td>0.99</td>
</tr>
</tbody>
</table>

**P-value threshold for FDR < 0.05**

**P-value threshold is highest ranking P-value for which corresponding Q-value is below desired significance threshold**

**Red: non-significant**

**Green: significant at FDR < 0.05**
Reducing multiple test correction stringency

• The correction to the P-value threshold $\alpha$ depends on the # of tests that you do, so, no matter what, the more tests you do, the more sensitive the test needs to be
• Can control the stringency by reducing the number of tests: e.g. use GO slim; restrict testing to the appropriate GO annotations; or select only larger GO categories.

Summary

• Enrichment analysis:
  – Statistical tests
    • Gene list: Fisher’s Exact Test
    • Ranked list: GSEA, also see Wilcoxon ranksum, Mann-Whitney U-test, Kolmogorov-Smirnov test
  – Multiple test correction
    • Bonferroni: stringent, controls probability of at least one false positive*
    • FDR: more forgiving, controls expected proportion of false positives* -- typically uses Benjamini-Hochberg

* Type 1 error, aka probability that observed enrichment if no association
We are on a Coffee Break & Networking Session