Analyzing Gene Set Enrichment

BaRC Hot Topics – June 20, 2016

Yanmei Huang
Bioinformatics and Research Computing
Whitehead Institute
http://barc.wi.mit.edu/hot_topics/
Purpose of Gene Set Enrichment Analysis

A list of genes you wish to study

A known gene set

Gene Set Enrichment Analysis

Are genes in the known gene set overrepresented in your list?
Sources of Gene List to Study

For example:

- Genes differentially expressed in two conditions
  - RNA-seq, microarray

- Genes behave similarly in a set of conditions
  - a clade from a clustering result

- Genes bound by a particular TF or RBP, etc
  - ChIP-seq, RIP-seq, CLIP-seq

- Any list of genes you might be interested in
Sources of Known Gene Sets

For example:

- A set of genes sharing a particular GO term annotation
- A set of genes involved in a particular pathway
- A set of genes known to be regulated by a particular TF, miRNA, etc
- A set of genes defined by a previous experiment
- Any set of genes you might be interested in
Two Different Strategies

Sort/Rank

Strategy 1
Look only in a list of genes meeting certain cutoff, are genes in the known gene get overrepresented?

use **Hypergeometric Test**

Strategy 2
Look across all rank, are genes in the known gene get distributed in a non-random manner?

- Gene set 3
- Gene set 2
- Gene set 1
- Ranked list

use **Kolmogorov-Smirnov (K-S) Test**
Hypergeometric Test

Are black balls overrepresented in your pick?

Bowl: 7 black + 8 white balls
Cup: randomly take 6 balls from the bowl

Ask: what is the probability of getting 4 or more black balls in the cup?
Hypergeometric Test

All possible combinations of picking 6 balls from 15 balls:

\[ C(15, 6) = \frac{15!}{9! \times 6!} = \frac{15\times14\times13\times12\times11\times10}{6\times5\times4\times3\times2\times1} = 5005 \]

<table>
<thead>
<tr>
<th># of black</th>
<th>number of possible combinations</th>
<th>probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>( C(7, 0) \times C(8, 6) = \frac{7!}{7!\times0!} \times \frac{8!}{2!\times6!} = 28 )</td>
<td>( \frac{28}{5005} = 0.006 )</td>
</tr>
<tr>
<td>1</td>
<td>( C(7, 1) \times C(8, 5) = \frac{7!}{6!\times1!} \times \frac{8!}{3!\times5!} = 392 )</td>
<td>( \frac{392}{5005} = 0.078 )</td>
</tr>
<tr>
<td>2</td>
<td>( C(7, 2) \times C(8, 4) = \frac{7!}{5!\times2!} \times \frac{8!}{4!\times4!} = 1470 )</td>
<td>( \frac{1470}{5005} = 0.294 )</td>
</tr>
<tr>
<td>3</td>
<td>( C(7, 3) \times C(8, 3) = \frac{7!}{4!\times3!} \times \frac{8!}{5!\times3!} = 1960 )</td>
<td>( \frac{1960}{5005} = 0.392 )</td>
</tr>
<tr>
<td>4</td>
<td>( C(7, 4) \times C(8, 2) = \frac{7!}{3!\times4!} \times \frac{8!}{6!\times2!} = 980 )</td>
<td>( \frac{980}{5005} = 0.196 )</td>
</tr>
<tr>
<td>5</td>
<td>( C(7, 5) \times C(8, 1) = \frac{7!}{2!\times5!} \times \frac{8!}{7!\times1!} = 168 )</td>
<td>( \frac{168}{5005} = 0.034 )</td>
</tr>
<tr>
<td>6</td>
<td>( C(7, 6) \times C(8, 0) = \frac{7!}{1!\times6!} \times \frac{8!}{8!\times0!} = 7 )</td>
<td>( \frac{7}{5005} = 0.001 )</td>
</tr>
</tbody>
</table>

Probability of getting 4 or more black balls:
\[ 0.196 + 0.034 + 0.001 = 0.231 \]

Probability of getting 5 or more black balls:
\[ 0.034 + 0.001 = 0.035 \]
Hypergeometric Test

Are black balls overrepresented in your pick?

```
> sum(dhyper(4:6, 7, 8, 6))
[1] 0.2307692

> phyper(3, 7, 8, 6, lower.tail = FALSE)
[1] 0.2307692

> fisher.test(matrix(c(4, 3, 2, 6), nrow=2), alternative="greater")$p.value
[1] 0.2307692
```
**Hypergeometric Test**

Are genes in the known gene set overrepresented in your list?

**Example:** Are immune response genes overrepresented in your list of 100 genes

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>immune response</th>
<th>Not immune response</th>
</tr>
</thead>
<tbody>
<tr>
<td>background</td>
<td>30,000</td>
<td>600</td>
<td>29400</td>
</tr>
<tr>
<td>Your list</td>
<td>100</td>
<td>10</td>
<td>90</td>
</tr>
</tbody>
</table>

```r
> sum(dhyper(10:100, 600, 29400, 100))
[1] 3.276244e-05
```

```r
> phyper(9, 600, 29400, 100, lower.tail = FALSE)
[1] 3.276244e-05
```

```r
> fisher.test(matrix(c(10, 590, 90, 29310), nrow=2), alternative="greater")$p.value
[1] 3.276244e-05
```
### Common Outputs from Programs Using Fisher’s Exact Test

**Number of genes in your list: 564**

<table>
<thead>
<tr>
<th>Term</th>
<th>Count</th>
<th>%</th>
<th>List Total</th>
<th>Pop Hits</th>
<th>Pop Total</th>
<th>Fold Enrichment</th>
<th>PValue</th>
<th>Bonferroni</th>
<th>Benjamini</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>extracellular matrix</td>
<td>45</td>
<td>7.98</td>
<td>554</td>
<td>241</td>
<td>19235</td>
<td>6.48</td>
<td>4.93E-23</td>
<td>2.10E-20</td>
<td>2.10E-20</td>
<td>6.94E-20</td>
</tr>
<tr>
<td>GO:0005578~proteinaceous extracellular matrix</td>
<td>49</td>
<td>8.69</td>
<td>500</td>
<td>320</td>
<td>15908</td>
<td>4.87</td>
<td>8.92E-20</td>
<td>3.16E-17</td>
<td>3.16E-17</td>
<td>1.22E-16</td>
</tr>
<tr>
<td>GO:0031012~extracellular matrix</td>
<td>49</td>
<td>8.69</td>
<td>500</td>
<td>345</td>
<td>15908</td>
<td>4.52</td>
<td>2.10E-18</td>
<td>7.44E-16</td>
<td>3.72E-16</td>
<td>2.88E-15</td>
</tr>
<tr>
<td>GO:0001501~skeletal system development</td>
<td>45</td>
<td>7.98</td>
<td>459</td>
<td>319</td>
<td>14116</td>
<td>4.34</td>
<td>3.52E-16</td>
<td>8.60E-13</td>
<td>8.60E-13</td>
<td>5.88E-13</td>
</tr>
<tr>
<td>cell cycle</td>
<td>51</td>
<td>9.04</td>
<td>554</td>
<td>461</td>
<td>19235</td>
<td>3.84</td>
<td>5.86E-16</td>
<td>2.36E-13</td>
<td>1.18E-13</td>
<td>7.77E-13</td>
</tr>
<tr>
<td>signal</td>
<td>169</td>
<td>30</td>
<td>554</td>
<td>3250</td>
<td>19235</td>
<td>1.81</td>
<td>1.48E-15</td>
<td>6.15E-13</td>
<td>2.05E-13</td>
<td>2.03E-12</td>
</tr>
<tr>
<td>GO:0044421~extracellular region part</td>
<td>79</td>
<td>14</td>
<td>500</td>
<td>960</td>
<td>15908</td>
<td>2.62</td>
<td>5.63E-15</td>
<td>2.00E-12</td>
<td>6.68E-13</td>
<td>7.76E-12</td>
</tr>
<tr>
<td>hsa04512:ECM-receptor interaction</td>
<td>23</td>
<td>4.08</td>
<td>180</td>
<td>84</td>
<td>5085</td>
<td>7.74</td>
<td>4.14E-14</td>
<td>5.09E-12</td>
<td>5.09E-12</td>
<td>4.78E-11</td>
</tr>
<tr>
<td>GO:0005581~collagen</td>
<td>16</td>
<td>2.84</td>
<td>500</td>
<td>35</td>
<td>15908</td>
<td>14.54</td>
<td>4.15E-14</td>
<td>1.47E-11</td>
<td>3.67E-12</td>
<td>5.69E-11</td>
</tr>
<tr>
<td>GO:0007049~cell cycle</td>
<td>68</td>
<td>12.1</td>
<td>459</td>
<td>776</td>
<td>14116</td>
<td>2.69</td>
<td>1.36E-13</td>
<td>3.50E-10</td>
<td>1.75E-10</td>
<td>2.40E-10</td>
</tr>
<tr>
<td>collagen</td>
<td>22</td>
<td>3.9</td>
<td>554</td>
<td>95</td>
<td>19235</td>
<td>8.04</td>
<td>2.48E-13</td>
<td>1.06E-10</td>
<td>2.64E-11</td>
<td>3.49E-10</td>
</tr>
<tr>
<td>GO:0000279~M phase</td>
<td>41</td>
<td>7.27</td>
<td>459</td>
<td>329</td>
<td>14116</td>
<td>3.83</td>
<td>4.94E-13</td>
<td>1.27E-09</td>
<td>4.25E-10</td>
<td>8.74E-10</td>
</tr>
<tr>
<td>GO:0022403~cell cycle phase</td>
<td>46</td>
<td>8.16</td>
<td>459</td>
<td>414</td>
<td>14116</td>
<td>3.42</td>
<td>9.02E-13</td>
<td>2.33E-09</td>
<td>5.82E-10</td>
<td>1.60E-09</td>
</tr>
<tr>
<td>triple helix</td>
<td>14</td>
<td>2.48</td>
<td>554</td>
<td>31</td>
<td>19235</td>
<td>15.68</td>
<td>1.02E-12</td>
<td>4.36E-10</td>
<td>8.72E-11</td>
<td>1.44E-09</td>
</tr>
</tbody>
</table>
Fisher’s Exact Test - Technical Considerations

- Number of genes in your list annotated
  - stop if too few are annotated

- Size of the known gene set
  - ignore gene sets of very small size

- Size of the background
  - try to define the background more precisely

- Correction for multiple testing
  - try to reduce redundancy

- Fold enrichment versus p value
  - “Fold enrichment” not always comparable
Kolmogorov-Smirnov (K-S) Test

K-S Test: Test for equality of distribution

Null hypothesis: the two samples are taken from the same distribution

\[ F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{[-\infty, x]}(X_i) \]

\[ D_{n,n'} = \sup_x |F_{1,n}(x) - F_{2,n'}(x)| \]

The null hypothesis is rejected at level \( \alpha \) if

\[ D_{n,n'} > c(\alpha) \sqrt{\frac{n + n'}{nn'}} \]

<table>
<thead>
<tr>
<th>( \alpha )</th>
<th>0.10</th>
<th>0.05</th>
<th>0.025</th>
<th>0.01</th>
<th>0.005</th>
<th>0.001</th>
</tr>
</thead>
<tbody>
<tr>
<td>( c(\alpha) )</td>
<td>1.22</td>
<td>1.36</td>
<td>1.48</td>
<td>1.63</td>
<td>1.73</td>
<td>1.95</td>
</tr>
</tbody>
</table>
If you rank genes in your list in a certain way, do genes in the known gene set tend to appear in the top ranked area?

$$P_{hit}(S, i) = \sum_{g_j \in S} \frac{|r_j|^p}{N_R}, \quad \text{where } N_R = \sum_{g_j \in S} |r_j|^p$$

$$P_{miss}(S, i) = \sum_{g_j \in S} \frac{1}{(N - N_H)}$$

**Enrichment Score**  \( ES = P_{hit} - P_{miss} \)

**Nominal P value**  Create 1000 permutation of dataset, for each permutation calculate an ES to form a null distribution, \( ES_{\text{NULL}} \), compare ES with \( ES_{\text{NULL}} \)

**Normalized Enrichment Score**  \( NES = \frac{\text{actual ES}}{\text{mean(ESs against all permutations of the dataset)}} \)

**Multiple hypothesis testing correction**  Calculate false discovery rate (FDR) corresponding to each of the NESs in the 1000 permutation
Broad’s GSEA Tool – Technical Considerations

- Ranking of the gene list affects ES
  - try pre-rank your list in different ways
  - let GSEA rank your list, know the options

- Gene sets included in the analysis affect NES
  - reduce redundancy
  - test only relevant gene sets

- Modes of permutations affects statistics
  - permutation by gene get
  - permutation by phenotype
Purpose of Gene Set Enrichment Analysis

A list of genes you wish to study

A known gene set

Gene Set Enrichment Analysis

Are genes in the known gene set overrepresented in your list?
Popular Public Gene Set Databases

**Gene Ontology (GO) annotation (many organisms)** - Defines concepts/classes used to describe gene function, and relationships between these concepts. Classifies gene functions by three aspects:

1. Molecular Function (MF)
2. Cellular Component (CC)
3. Biological Process (BP)

**KEGG Pathways (many organisms)** - A collection of manually drawn pathway maps representing our knowledge on the molecular interaction and reaction networks for several categories:

1. Metabolism
2. Genetic Information Processing
3. Environmental Information Processing
4. Cellular Processes
5. Organismal Systems
6. Human Diseases
7. Drug Development (drug structure relationship, not relevant here)

**MSigDB (Human only)** - A collection of annotated gene sets in 8 major categories:

- **H** - hallmark gene sets
- **C1** - positional gene sets
- **C2** - curated gene sets
- **C3** - motif gene sets
- **C4** - computational gene sets
- **C5** - GO gene sets
- **C6** - oncogenic signatures
- **C7** - immunologic signatures
Gene Ontology (GO) Annotation

Gene Ontology (GO): Hierarchical relationship based on “is a” or “is part of”

Example of different levels of GO annotation:

Level 1: physiological processes
Level 2: response to external stimulus
Level 3: response to biotic stimulus
Level 4: defense response
Level 5: immune response

More General include more genes
More Specific include less genes
KEGG: Kyoto Encyclopedia of Genes and Genomes

KEGG is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies. See Release notes (June 1, 2016) for new and updated features.

Announcement: KEGG RPAIR to be discontinued

Main entry point to the KEGG web service
- KEGG2
- KEGG Table of Contents
- Update notes

Data-oriented entry points
- KEGG PATHWAY KEGG pathway maps [Pathway list]
- KEGG BRITE BRITE functional hierarchies [Brite list]
- KEGG MODULE KEGG modules [Module list | Statistics]
- KEGG ORTHOLOGY Ortholog groups [KO system | Annotation]
- KEGG GENOME Genomes [KEGG organisms]
- KEGG GENES Genes and proteins [Release history]
- KEGG COMPOUND Small molecules [Compound classification]
- KEGG GLYCAN Glycans [Monosaccharide codes]
- KEGG REACTION Biochemical reactions [Reaction modules]
- KEGG ENZYME Enzymes [EC number to sequence links]
- KEGG DISEASE Human diseases [Cancer | Pathogen]
- KEGG DRUG Drugs [ATC drug classification]
- KEGG MEDICUS Health information resource [Drug labels search]

Organism-specific entry points
- KEGG Organisms
  - Enter org code(s)
  - Go
  - hsa
  - hsa eco
KEGG PATHWAY Database

Wiring diagrams of molecular interactions, reactions, and relations

Pathway Maps

**KEGG PATHWAY** is a collection of manually drawn *pathway maps* representing our knowledge on the molecular interaction and reaction networks for:

1. **Metabolism**
   - Global/overview
   - Carbohydrate
   - Energy
   - Lipid
   - Nucleotide
   - Amino acid
   - Other amino
   - Glycan
   - Cofactor/vitamin
   - Terpenoid/PK
   - Other secondary metabolite
   - Xenobiotics
   - Chemical structure
2. **Genetic Information Processing**
3. **Environmental Information Processing**
4. **Cellular Processes**
5. **Organismal Systems**
6. **Human Diseases**

and also on the structure relationships (KEGG drug structure maps) in:

7. **Drug Development**
Molecular Signatures Database (MSigDB)

**H** hallmark gene sets are coherently expressed signatures derived by aggregating many MSigDB gene sets to represent well-defined biological states or processes.

**C1** positional gene sets for each human chromosome and cytogenetic band.

**C2** curated gene sets from online pathway databases, publications in PubMed, and knowledge of domain experts.

**C3** motif gene sets based on conserved cis-regulatory motifs from a comparative analysis of the human, mouse, rat, and dog genomes.

**C4** computational gene sets defined by mining large collections of cancer-oriented microarray data.

**C5** GO gene sets consist of genes annotated by the same GO terms.

**C6** oncogenic signatures defined directly from microarray gene expression data from cancer gene perturbations.

**C7** immunologic signatures defined directly from microarray gene expression data from immunologic studies.

- Human only
- Some gene sets manually curated by scientists
- Developed for convenient use within the GSEA program
- Gene sets available for download
<table>
<thead>
<tr>
<th>Tool</th>
<th>Species</th>
<th>Allow user defined gene set?</th>
<th>Has its own curated gene sets?</th>
<th>Gene Set Updates</th>
<th>Access</th>
<th>Graphic output?</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAVID</td>
<td>Various</td>
<td>No</td>
<td>No</td>
<td>occasional</td>
<td>Free</td>
<td>No</td>
</tr>
<tr>
<td>GSEA</td>
<td>Human</td>
<td>Yes</td>
<td>Yes</td>
<td>occasional</td>
<td>Free</td>
<td>Yes</td>
</tr>
<tr>
<td>BiNGO</td>
<td>Various</td>
<td>Yes</td>
<td>No</td>
<td>occasional</td>
<td>Free</td>
<td>Yes</td>
</tr>
<tr>
<td>GeneGO</td>
<td>Human</td>
<td>No</td>
<td>Yes</td>
<td>regular</td>
<td>WI license</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Mouse</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IPA</td>
<td>Human</td>
<td>?</td>
<td>Yes</td>
<td>regular</td>
<td>Need license</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Graphic Representation of Results

BiNGO

IPA

GeneGO
Practicalities Using the Tools

➤ Choose a tool that
  – includes your species
  – includes your genes’ identifiers
  – has up-to-date gene set annotation
    ~80% of papers in 2015 used outdated annotation (Wadi et al., 2016, BioRxiv)
  – allows user-defined gene sets
  – allows user-defined background

➤ Try a few tools and compare results

➤ Consider graphic representation of results
The problem:
Is one set of genes overrepresented in another?

The statistical methods:
Fisher’s exact test
K-S test

The practicalities:
The gene set databases to use
The programs to use
Suggested Readings

DEMO of GSEA

Step 1. Launching the GSEA software
   go to http://software.broadinstitute.org/gsea
   Click on “Download”, type in a registered email address
   Select desired memory and click “Launch”. A .jnlp file will be downloaded
   open the downloaded .jnlp file

Step 2. Load data
   Click on “Browser for files” and load the following three files:
   “TCGA_tumor_normal_pair.txt”
   “tumor_vs_normal.cls”
   “tumor_normal_DE.rnk”

Step 3. Run the main GESA analysis with the “TCGA_tumor_normal_pair.txt” file

Step 4. Run the GSEApreranked analysis with the “tumor_normal_DE.rnk” file

Step 5. Run Leading edge analysis
Demo of DAVID

Step 1. Find the tool
Go to DAVID website (https://david.ncifcrf.gov/)
Click on “Shortcut to DAVID Tools” -> Functional Annotation

Step 2. Obtain the list of gene to analyze
open the “tumor_normal_DE.txt” file in excel,
filter the genes such that padj < 0.01 and log2FC >= 1

Step 3. Upload the lists of interest
copy the list of genes passing the filter into “Paste a list” box
Select Identifier
Indicate list type
click “Submit List”

Step 4. Run the Analysis
select known gene sets to analyze
click on one of the buttons (clustering, chart or table) to display result