Gene Set Analysis

CMSC 702 Spring 2013
Hector Corrada Bravo
Computer Science
University of Maryland
Long lists of d.e. genes ≠ biological understanding

What happens next?
• Select some genes for validation?
• Do follow-up experiments on some genes?
• Publish a huge table with the results?
• Try to learn about all the genes on the list (read 100s of papers)?
• ....

Usually, some or all of the above will be done, and more.
Can we help further at this
Gene Set Enrichment

• Gene Set Enrichment Analysis is an approach to finding sets of biologically related genes that are enriched for differential expression.

• Gene set enrichment is a method of combining information across genes to make sense of gene lists.
Sets of genes

• There are usually many sets of genes that might be of interest in a given microarray experiment.

• Examples include:
  • genes in biological (e.g. biochemical, metabolic, and signalling) pathways,
  • genes associated with a particular location in the cell,
  • genes having a particular function or being involved in a particular process.
Sets of genes

• We could even include sets of genes for which all of the preceding are unknown, but we have reason believe could be of interest, typically from previous experiments.

• In thinking like this, it is important to remember that many genes (that is, their protein products) can have multiple functions, or be involved in many processes, etc.

• There are many databases (EcoCyc, KEGG, STRING, ..) of pathways and sets.

• We will focus on the most important related concept: the Gene Ontology.
The goal of the Gene Ontology ™ (GO) Consortium is to produce a controlled vocabulary that can be applied to all organisms even as knowledge of gene and protein roles in cells is accumulating and changing. GO provides three structured networks of defined terms to describe gene product attributes:

- **Molecular Function Ontology** (7304 terms as of April 5, 2004): the tasks performed by individual gene products; examples are carbohydrate binding and ATPase activity.

- **Biological Process Ontology** (8517 terms): broad biological goals, such as mitosis or purine metabolism, that are accomplished by ordered assemblies of molecular functions.

- **Cellular Component Ontology** (1394 terms): subcellular structures, locations, and macromolecular complexes; examples include nucleus, telomere, and origin recognition complex.
From the GO web site. The path back to each ontology from a gene.

We will call each term in a path a split.
Structure of a GO annotation

Annotated GO: GO:0006917

**Path:**

- GO:0003673 : Gene Ontology (46199)
  - ⊆ GO:0008150 : biological_process (30188)
    - ⊆ GO:0016265 : death (525)
      - ⊆ GO:0008219 : cell death (484)
      - ⊆ GO:0012501 : programmed cell death (447)
        - ⊆ GO:0006915 : apoptosis (419)
          - ⊆ GO:0006917 : induction of apoptosis (148)
          - ⊆ GO:0012502 : induction of programmed cell death (148)
            - ⊆ GO:0006917 : induction of apoptosis (148)
  - ⊆ GO:0005575 : cellular_component (22371)
  - ⊆ GO:0003674 : molecular_function (37018)

**Splits:**

- GO:0008150
- GO:0016265
- GO:0008219
- GO:0012501
- GO:0012502
- GO:0006915
- GO:0006917

Each gene can have several annotated GOs and each GO can have several splits. E.g. DNA topoisomerase II alpha has 8 GO annotations and 11 splits.
Annotation of genes to a node in the ontology

Each node is also connected to many other related nodes.
Pathways (KEGG)
Are sets of genes differentially expressed?

The sets we refer to here are all the outcomes of analyses. Later we discuss sets specified a priori.

Examples of sets. They could be the list of all genes whose differential expression (e.g. average $M$-value) exceeds a given threshold, typically a liberal one, which would not correspond to any real “significance”, e.g. 1.5-fold. They might be clusters.

What do we mean by a set being differentially expressed. Here it is a convenient shorthand for being unusual in relation to all the genes represented on the array, for example, by being functionally enriched, in the sense of having more genes of a given category than one would expect, by chance.
GO and microarray gene sets

**Hypothesis:** Functionally related, differentially expressed genes should accumulate in the corresponding GO-group.

**Problem:** to find a method which scores accumulation of differential gene expression in a node of the GO.

We describe the calculation from the program Gostat. For all the genes analysed, it determines the annotated GO terms and all splits. It then counts the # of appearances of each GO term for the genes in the set, as well as the # in the reference set, which is typically all genes on the array. Then a $2 \times 2$ table is formed, see over page, and a $p$-value calculated.
Is a GO term is specific for a set?

Contingency Table

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>51</td>
<td>416</td>
<td>467</td>
</tr>
<tr>
<td>125</td>
<td>8588</td>
<td>8713</td>
</tr>
<tr>
<td>173</td>
<td>9004</td>
<td>9177</td>
</tr>
</tbody>
</table>
Is a GO term is specific for a set?

Contingency Table

<table>
<thead>
<tr>
<th></th>
<th>51</th>
<th>416</th>
<th>467</th>
</tr>
</thead>
<tbody>
<tr>
<td>125</td>
<td>8588</td>
<td></td>
<td>8713</td>
</tr>
<tr>
<td>173</td>
<td>9004</td>
<td></td>
<td>9177</td>
</tr>
</tbody>
</table>
Is a GO term is specific for a set?

<table>
<thead>
<tr>
<th>Count Genes with GO Term in Set</th>
<th>Count Genes without GO Term in Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>51 416 467</td>
<td>125 8588 8713</td>
</tr>
<tr>
<td>173 9004 9177</td>
<td></td>
</tr>
</tbody>
</table>

Contingency Table

Wednesday, March 27, 13
Is a GO term is specific for a set?

Contingency Table

<table>
<thead>
<tr>
<th>count genes</th>
<th>count genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>with GO term in set</td>
<td>without GO term in set</td>
</tr>
<tr>
<td>51</td>
<td>125</td>
</tr>
<tr>
<td>416</td>
<td>8588</td>
</tr>
<tr>
<td>467</td>
<td>8713</td>
</tr>
<tr>
<td>173</td>
<td>9004</td>
</tr>
<tr>
<td>9177</td>
<td>9177</td>
</tr>
</tbody>
</table>

(count in set (e.g. differentially expressed genes))
Is a GO term specific for a set?

<table>
<thead>
<tr>
<th></th>
<th>Count genes with GO term in set</th>
<th>Count genes without GO term in set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count in set</td>
<td>51</td>
<td>125</td>
</tr>
<tr>
<td>Count in reference set</td>
<td>416</td>
<td>8588</td>
</tr>
<tr>
<td></td>
<td></td>
<td>467</td>
</tr>
<tr>
<td></td>
<td></td>
<td>173</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9004</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9177</td>
</tr>
</tbody>
</table>

Contingency Table

Count in reference set (e.g. all genes on array)
Is a GO term is specific for a set?

Contingency Table

<table>
<thead>
<tr>
<th>Count Genes with GO Term in Set</th>
<th>Count Genes without GO Term in Set</th>
<th>Count in Set (e.g. differentially expressed genes)</th>
<th>Count in Reference Set (e.g. all genes on array)</th>
</tr>
</thead>
<tbody>
<tr>
<td>51</td>
<td>125</td>
<td>173</td>
<td>416</td>
</tr>
<tr>
<td>416</td>
<td>8588</td>
<td>9004</td>
<td>467</td>
</tr>
<tr>
<td>467</td>
<td>8713</td>
<td>9177</td>
<td></td>
</tr>
</tbody>
</table>

Fisher's exact test or chi-square test
Is a GO term is specific for a set?

<table>
<thead>
<tr>
<th>Contingency Table</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>51 416 467</td>
<td>8x10^{-52}</td>
</tr>
<tr>
<td>125 8588 8713</td>
<td></td>
</tr>
<tr>
<td>173 9004 9177</td>
<td></td>
</tr>
</tbody>
</table>

- Count genes with GO term in set
- Count genes without GO term in set
- Count in set (e.g. differentially expressed genes)
- Count in reference set (e.g. all genes on array)

Fisher's exact test or chi-square test
The multiple testing problem

Naturally one doesn’t test a single GO term or split, but many, perhaps 1000s. As with testing of single genes, we need to deal with the multiple testing problem. Many of the solutions from there carry over: Bonferroni, Holm, step-down minP, FDR, and so on. But there are also special problems here, deriving from the nesting relationships between splits. In my view, these are not easily dealt with, and require more research.

Related questions. How can we compare the results of different lists being compared? And, rather than select a set of genes using a cut-off, can we make use the gene abundances or p-values for differential expression?
There are many tools

Here are a few.

GoStat
GenMAPP, and MAPPFinder
EASE (DAVID)
FunSpec
FatiGO

.....
Analyzing microarray data by functional gene sets defined *a priori*

Analysis at the level of single gene:

- Identifying differentially expressed genes becomes a challenge when the magnitude of differential expression is small.
- For some differences, many genes are involved.

Analysis at the level of functional group: why?

By incorporating biological knowledge, we can hope to detect modest but coordinate expression changes of sets of functionally related genes.
PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes

Mootha et al, Nature Genetics July 2003

Data: Affymetrix microarray data on 22,000 genes in skeletal muscle biopsy samples from 43 males, 17 with normal glucose tolerance (NGT), 8 with impaired glucose tolerance and 18 with Type 2 diabetes (DM2).

In their single gene analysis, a t-statistic was calculated for each gene. No significant difference found between NTG and DM2 after adjusting for multiple testing.

Their idea: test 149 a priori defined gene sets for association with disease phenotypes.


We prefer: Tian et al PNAS 2005 102 (38): 13544-13549
ES=enrichment score for each gene = scaled K-S dist

A set called OXPHOS got the largest ES score, with \( p=0.029 \) on 1,000 permutations.

From Mootha \textit{et al}
Gene Set Enrichment Analysis

- Large
- Value of Statistic
- Small

- Gene In A Relevant Set
- Gene Not In The Set
Gene Set Enrichment Analysis

Value of Statistic

Running Total

Statistic is Max Deviation From 0

Large

Small

Gene In A Relevant Set

Gene Not In The Set

Wednesday, March 27, 13
Permute The Labels and Recalculate

Leaves the relationship between genes unchanged.
Permuted GSEA Profile

Permuted Statistic is Max Deviation From 0

Value of Statistic

Running Total

Large

Small

Gene In A Relevant Set

Gene Not In The Set

Wednesday, March 27, 13
Calculating a P-value

\[ \text{P-value} = \frac{\# \left\{ S_{perm} \geq |S_{obs}| \right\}}{\text{# of Permutations}} \]
(A small difference for many genes)
• What are we really asking?
  – Hypothesis 1: The genes in a gene set show the same pattern of associations with the phenotype compared with the rest of the genes
  – Hypothesis 2: The gene set does not contain any gene associated with the phenotype
Hypothesis 1

Mootha *et al* did a two sample K-S test to compare genes in a specific gene set with genes not in that set.

Instead of doing this, why don’t we simply do a one sample test, comparing each gene set to the whole (population) directly?

Each gene set is small w.r.t. the entire set of genes, so all other genes $\approx$ all genes.
Hypothesis 1

Mootha et al. did a two sample K-S test to compare genes in a specific gene set with genes not in that set.

Statistic for gene-set k:

\[ T_k = \frac{1}{m_k} \sum_{i=1}^{G} I\{i \in k\} t_i \]
Hypothesis 1

Mootha et al did a two sample K-S test to compare genes in a specific gene set with genes not in that set.

Statistic for gene-set k:

\[ T_k = \frac{1}{m_k} \sum_{i=1}^{G} I\{i \in k\} t_i \]

What should the permutation be in this case then?
Hypothesis 1

Mootha et al did a two sample K-S test to compare genes in a specific gene set with genes not in that set.

Statistic for gene-set k:

\[ T_k = \frac{1}{m_k} \sum_{i=1}^{G} I\{i \in k\} t_i \]

What should the permutation be in this case then?

– *Permute the association statistics ti!* (not what Mootha et al. do)
• What are we really asking?
  – Hypothesis 1: The genes in a gene set show the same pattern of associations with the phenotype compared with the rest of the genes
  – Hypothesis 2: The gene set does not contain any gene associated with the phenotype
Hypothesis 2

Statistic for gene-set k:

\[ T_k = \frac{1}{m_k} \sum_{i=1}^{G} I\{i \in k\} t_i \]

What should the permutation be in this case then?
Hypothesis 2

Statistic for gene-set $k$:

$$T_k = \frac{1}{m_k} \sum_{i=1}^{G} I\{i \in k\} t_i$$

What should the permutation be in this case then?

– *Permute the phenotype labels!* (this *is* what Mootha et al. do)

– Take away message: it pays to be clear and direct when designing these procedures...
Geneset-based analysis

• Gene modules: Segal et al., [Nat. Genetics 2004]
• Functional maps: Troyanskova et al., [Genome Research 2009]
• Many more...