From Genes to Biology:
The Gene Ontology / Pathway enrichment

Biol4230  Tues, April 10, 2018
Bill Pearson  wrp@virginia.edu  4-2818  Pinn 6-057

• Gene Ontology (GO)
  – "Ontology" – a directed acyclic graph (DAG)
  – molecular function, biological process, cellular component
  – evidence and evidence codes
  – positives and negatives, missing data

• Function/Pathway enrichment analysis
  – do sets (subsets) of differentially expressed genes reflect a pathway?
  – Over representation analysis (ORA)
  – functional class scoring – GSEA (gene set enrichment analysis)

To learn more:

2. www.geneontology.org
   conjecture with comparative functional genomic data from mammals. 
5. Thomas, P. D.  et al. On the Use of Gene Ontology Annotations to Assess Functional 
   Similarity among Orthologs and Paralogs: A Short Report. 
   approaches and outstanding challenges. 
   for interpreting genome-wide expression profiles. 
Gene Ontology/Enrichment Analysis

- I have a set of differentially expressed genes – what is happening to the cell?
- Gene Ontology (GO)
  - "Ontology" – a directed acyclic graph (DAG)
  - molecular function, biological process, cellular component
  - evidence and evidence codes
  - positives and negatives, missing data
  - One of many
- Function/Pathway enrichment analysis
  - do sets (subsets) of differentially expressed genes reflect a pathway?
  - Over representation analysis (ORA)
  - functional class scoring – GSEA (gene set enrichment analysis)
What is happening to the cell?

- Cellular functions are chemical
- Fundamental biochemical processes are lined chemical reactions: pathways
  - cell division: DNA replication, mitosis, segregation
  - metabolism: energy, amino-acids, detoxification
  - response to stimuli: signaling
- Some pathways are better understood than others

KEGG pathways (energy metabolism)
Differential gene expression on pathways

- Goal: to identify the (known) biological pathways that are activated during biological transitions
  - from stationary/resting to growth phase
  - from normal to cancerous
  - from pluripotent to lineage specific
  - in response to environmental stimuli
- Requirements:
  - list of genes turned on or off / up or down
  - (known) relationships between genes/proteins
    - Gene Ontology
    - shared pathways/processes KEGG/Reactome
- Measure of over-representation
  - hypergeometric (Fisher's Exact test)
  - permutation
  - GSEA

GO: The Gene Ontology (geneontology.org)

- Ontology relationships – Directed Acyclic Graph (DAG) of relationships
  - is-a
  - part-of / has-part
  - regulates / positively-regulates / negatively-regulates
- Hierarchical – three orthogonal hierarchies
  - molecular function
  - biological process
  - cellular location
  - (no sense of time, or developmental stage)
- Curated, with Evidence codes
  - experimental
  - similarity based (but curated)
  - IEA Inferred from Electronic Annotation (no human)
- Absence of activity/process annotation does NOT guarantee absence of activity/process
The Gene Ontology (GO)

**Figure 1** | simple trees versus directed acyclic graphs. Boxes represent nodes and arrows represent edges. a | an example of a simple tree, in which each child has only one parent and the edges are directed, that is, there is a source (parent) and a destination (child) for each edge. b | A directed acyclic graph (DAG), in which each child can have one or more parents. The node with multiple parents is coloured red and the additional edge is coloured grey. c | An example of a node, vesicle fusion, in the biological process ontology with multiple parentage. The dashed edges indicate that there are other nodes not shown between the nodes and the root node (biological process). A root is a node with no incoming edges, and at least one leaf (also called a sink). A leaf node is a node with no outgoing edges, that is, a terminal node with no children (vesicle fusion). Similar to a simple tree, a DAG has directed edges and does not have cycles, that is, no path starts and ends at the same node, and will always have at least one root node. The depth of a node is the length of the longest path from the root to that node, whereas the height is the length of the longest path from that node to a leaf. is_a and part_of are types of relationships that link the terms in the GO ontology. More information about the relationships between GO terms are found online (An Introduction to the Gene Ontology).

Gene Ontology Relationships:
is-a, part-of
(regulates/up-regulates/down-regulates)
Gene Ontology Relationships:
is-a, part-of, has-part, regulates

- **is-a**: identity (synonyms, reversible)
- **part-of**: sub-set, not reversible
- **has-part**: converse of part-of, not reversible
- **regulates/**up-regulates/down-regulates
- can be combined in logically consistent ways

Gene Ontology (GO) hierarchies for GSTM1_HUMAN

- **Molecular function (chemistry)**
  - glutathione binding, GST activity, enzyme binding, homodimerization, detoxification of nitrogen compound

- **Biological process (pathway, function)**
  - xenobiotic metabolic process, glutathione derivative biosynthetic process, small molecular metabolic process

- **Cellular location**
  - cytosol, cytoplasm
Gene Ontology (GO) hierarchies for GSTM1_HUMAN

<table>
<thead>
<tr>
<th>Gene product name</th>
<th>Gene product name</th>
<th>Qualifier</th>
<th>Direct annotation</th>
<th>Annotation extension</th>
<th>Source</th>
<th>Taxon</th>
<th>Evidence</th>
<th>Evidence with</th>
<th>Panther family</th>
<th>Isoform</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSTM1_HUMAN</td>
<td>Glutathione S-transferase 4a 1</td>
<td></td>
<td>Glutathione S-transferase family</td>
<td>P09488</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>fasta.bioch.virginia.edu/biol4559</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>fasting.bioch.virginia.edu/biol4230</td>
</tr>
</tbody>
</table>
Gene Ontology (GO) hierarchy for glutathione binding molecular function
(annotating the ontology, NOT a protein)
Gene Ontology (GO) hierarchy for xenobiotic metabolic process Biological Process

<table>
<thead>
<tr>
<th>Term Information</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accession</strong></td>
</tr>
<tr>
<td><strong>Name</strong></td>
</tr>
<tr>
<td><strong>Ontology</strong></td>
</tr>
<tr>
<td><strong>Synonyms</strong></td>
</tr>
<tr>
<td><strong>Definition</strong></td>
</tr>
<tr>
<td><strong>Comment</strong></td>
</tr>
<tr>
<td><strong>History</strong></td>
</tr>
<tr>
<td><strong>Subset</strong></td>
</tr>
<tr>
<td><strong>Community</strong></td>
</tr>
<tr>
<td><strong>Related</strong></td>
</tr>
</tbody>
</table>

Gene Ontology (GO) hierarchy for xenobiotic metabolic process Biological Process
Gene Ontology (GO) annotations have evidence codes

Gene Ontology entries have Evidence Codes
geneontology.org/page/guide-go-evidence-codes

Experimental:
• Inferred from Experiment (EXP)
• Inferred from Direct Assay (IDA)
• Inferred from Physical Interaction (IPI)
• Inferred from Mutant Phenotype (IMP)
• Inferred from Genetic Interaction (IGI)
• Inferred from Expression Pattern (IEP)

Literature based:
• Traceable author statement (TAS)

Computational (and someone looked at it)
• Inferred from Sequence or structural Similarity (ISS)
• Inferred from Sequence Orthology (ISO)
• Inferred from Sequence Alignment (ISA)
• Inferred from Sequence Model (ISM)
• Inferred from Genomic Context (IGC)
• Inferred from Biological aspect of Ancestor (IBA)
• Inferred from Biological aspect of Descendant (IBD)
• Inferred from Key Residues (IKR)
• Inferred from Rapid Divergence (IRD)
• Inferred from Reviewed Computational Analysis (RCA)

Computational (no human curation)
• Inferred from Electronic Annotation (IEA)

Evidence codes are not statements of the quality of the annotation. Within each evidence code classification, some methods produce annotations of higher confidence or greater specificity than other methods…. Thus evidence codes cannot be used as a measure of the quality of the annotation.
Gene Ontology entries have Evidence Codes

**Gene Ontology coverage, by evidence (2007)**

<table>
<thead>
<tr>
<th>Evidence code</th>
<th>Evidence code description</th>
<th>Source of evidence</th>
<th>Manually checked</th>
<th>Current number of annotations*</th>
</tr>
</thead>
<tbody>
<tr>
<td>IA</td>
<td>Inferred from direct essay</td>
<td>Experimental</td>
<td>Yes</td>
<td>71,010</td>
</tr>
<tr>
<td>IE</td>
<td>Inferred from expression pattern</td>
<td>Experimental</td>
<td>Yes</td>
<td>8,585</td>
</tr>
<tr>
<td>IG</td>
<td>Inferred from genetic interaction</td>
<td>Experimental</td>
<td>Yes</td>
<td>8,311</td>
</tr>
<tr>
<td>IM</td>
<td>Inferred from mutant phenotype</td>
<td>Experimental</td>
<td>Yes</td>
<td>93,549</td>
</tr>
<tr>
<td>IP</td>
<td>Inferred from physical interaction</td>
<td>Experimental</td>
<td>Yes</td>
<td>17,043</td>
</tr>
<tr>
<td>ISS</td>
<td>Inferred from sequence or structural similarity</td>
<td>Computational</td>
<td>Yes</td>
<td>196,843</td>
</tr>
<tr>
<td>RCA</td>
<td>Inferred from reviewed computational analysis</td>
<td>Computational</td>
<td>Yes</td>
<td>103,792</td>
</tr>
<tr>
<td>IGC</td>
<td>Inferred from genomic context</td>
<td>Computational</td>
<td>Yes</td>
<td>4</td>
</tr>
<tr>
<td>IEA</td>
<td>Inferred from electronic annotation</td>
<td>Computational</td>
<td>No</td>
<td>15,687,352</td>
</tr>
<tr>
<td>IC</td>
<td>Inferred by curator</td>
<td>Indirectly derived from experimental or computational evidence made by a curator</td>
<td>Yes</td>
<td>5,167</td>
</tr>
<tr>
<td>TAS</td>
<td>Traceable author statement</td>
<td>Indirectly derived from experimental or computational evidence made by the author of the published article</td>
<td>Yes</td>
<td>44,594</td>
</tr>
<tr>
<td>NAS</td>
<td>Non-traceable author statement</td>
<td>No 'source of evidence' statement given</td>
<td>Yes</td>
<td>25,656</td>
</tr>
<tr>
<td>ND</td>
<td>No biological data available</td>
<td>No information available</td>
<td>Yes</td>
<td>132,102</td>
</tr>
<tr>
<td>NR</td>
<td>Not recorded</td>
<td>Unknown</td>
<td>Yes</td>
<td>1,185</td>
</tr>
</tbody>
</table>

*October 2007 release

### Gene Ontology coverage, by organism (2007)

Table 1: Distribution of gene ontology (GO) annotations for species with more than 5,000 annotations

<table>
<thead>
<tr>
<th>Species (NCBI taxon ID)</th>
<th>Genes× with experimental annotations</th>
<th>Total annotated genes</th>
<th>Percentage of genes× with at least one experimental annotation</th>
<th>Total genes×</th>
<th>Percentage annotated</th>
<th>Percentage known in genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schizosaccharomyces pombe (4496)</td>
<td>4,482</td>
<td>4,930</td>
<td>90.0%</td>
<td>4,390</td>
<td>100.0%</td>
<td>90.0%</td>
</tr>
<tr>
<td>Saccharomyces cerevisiae (3932)</td>
<td>4,647</td>
<td>5,794</td>
<td>85.4%</td>
<td>5,794</td>
<td>100.0%</td>
<td>85.4%</td>
</tr>
<tr>
<td>Mus musculus (10090)</td>
<td>10,621</td>
<td>10,186</td>
<td>57.8%</td>
<td>27,329</td>
<td>100.0%</td>
<td>39.0%</td>
</tr>
<tr>
<td>Canis familiaris (62198)</td>
<td>4,654</td>
<td>5,154</td>
<td>35.9%</td>
<td>20,165</td>
<td>79.2%</td>
<td>23.3%</td>
</tr>
<tr>
<td>Human (9606)</td>
<td>4,390</td>
<td>10,151</td>
<td>29.2%</td>
<td>20,847</td>
<td>91.5%</td>
<td>29.2%</td>
</tr>
<tr>
<td>Arabidopsis thaliana (13702)</td>
<td>5,530</td>
<td>2,637</td>
<td>20.0%</td>
<td>27,029</td>
<td>98.5%</td>
<td>20.1%</td>
</tr>
<tr>
<td>Rat (10116)</td>
<td>3,566</td>
<td>1,245</td>
<td>20.7%</td>
<td>17,893</td>
<td>96.3%</td>
<td>19.3%</td>
</tr>
<tr>
<td>Firefly (11219)**</td>
<td>2,790</td>
<td>9,569</td>
<td>92.8%</td>
<td>14,181</td>
<td>100.0%</td>
<td>19.7%</td>
</tr>
<tr>
<td>Canis familiaris (6479)</td>
<td>896</td>
<td>3,916</td>
<td>21.4%</td>
<td>6,166</td>
<td>50.0%</td>
<td>3.6%</td>
</tr>
<tr>
<td>Pseudomonas aeruginosa (PA01 (20994))</td>
<td>491</td>
<td>2,596</td>
<td>19.6%</td>
<td>5,168</td>
<td>85.0%</td>
<td>8.8%</td>
</tr>
<tr>
<td>Sine modii (44698)</td>
<td>797</td>
<td>6,892</td>
<td>11.6%</td>
<td>15,825</td>
<td>50.3%</td>
<td>5.9%</td>
</tr>
<tr>
<td>Hygroscopicus levis (1681)</td>
<td>440</td>
<td>9,314</td>
<td>15.5%</td>
<td>6,134</td>
<td>42.3%</td>
<td>4.9%</td>
</tr>
<tr>
<td>Zebrafish (7951)</td>
<td>2,125</td>
<td>13,274</td>
<td>5.8%</td>
<td>21,312</td>
<td>59.7%</td>
<td>5.7%</td>
</tr>
<tr>
<td>Pseudomonas putida (5813)</td>
<td>188</td>
<td>3,443</td>
<td>5.8%</td>
<td>5,242</td>
<td>93.8%</td>
<td>5.4%</td>
</tr>
<tr>
<td>Rice (15652)</td>
<td>654</td>
<td>29,837</td>
<td>2.2%</td>
<td>41,308</td>
<td>71.5%</td>
<td>1.5%</td>
</tr>
<tr>
<td>Chicken (8033)</td>
<td>75</td>
<td>6,063</td>
<td>1.0%</td>
<td>30,717</td>
<td>100.0%</td>
<td>0.4%</td>
</tr>
<tr>
<td>Cow (9913)</td>
<td>96</td>
<td>8,536</td>
<td>1.1%</td>
<td>21,716</td>
<td>39.2%</td>
<td>0.4%</td>
</tr>
</tbody>
</table>

- Percentage annotated is determined by dividing the number of genes annotated by total genes.
- Percentage known in genome is determined by multiplying the percentage of experimentally derived annotations by the percentage of the genome annotated. This is an approximation of the extent of knowledge about the portion of the genome that encodes proteins in an organism with a complete genome sequence that is captured by annotation.


fasta.bioch.virginia.edu/biol4559

### Gene Ontology coverage, by organism (2015)

<table>
<thead>
<tr>
<th>Species</th>
<th>Source</th>
<th>Genes</th>
<th>Annots</th>
<th>non-IEA</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>P. falciparum</td>
<td>GeneDB</td>
<td>2373</td>
<td>6250</td>
<td>6250</td>
<td>3/10/2015</td>
</tr>
<tr>
<td>E. coli</td>
<td>PortEco</td>
<td>3770</td>
<td>45842</td>
<td>13302</td>
<td>6/26/2014</td>
</tr>
<tr>
<td>E. coli</td>
<td>FlyBase</td>
<td>14646</td>
<td>102825</td>
<td>90887</td>
<td>3/10/2015</td>
</tr>
<tr>
<td>B. taurus</td>
<td>GO/EBI</td>
<td>20466</td>
<td>163368</td>
<td>35893</td>
<td>3/31/2015</td>
</tr>
<tr>
<td>G. gallus</td>
<td>GO/EBI</td>
<td>12945</td>
<td>101588</td>
<td>15119</td>
<td>3/31/2015</td>
</tr>
<tr>
<td>B. taurus</td>
<td>GO/EBI</td>
<td>17349</td>
<td>141466</td>
<td>33661</td>
<td>3/31/2015</td>
</tr>
<tr>
<td>C. lupus</td>
<td>GO/EBI</td>
<td>16016</td>
<td>123620</td>
<td>19392</td>
<td>3/31/2015</td>
</tr>
<tr>
<td>M. musculus</td>
<td>MGI</td>
<td>24177</td>
<td>354820</td>
<td>255070</td>
<td>4/2/2015</td>
</tr>
<tr>
<td>H. sapiens</td>
<td>SGD</td>
<td>20565</td>
<td>416906</td>
<td>251499</td>
<td>4/2/2015</td>
</tr>
<tr>
<td>S. pombe</td>
<td>PomBase</td>
<td>5382</td>
<td>39112</td>
<td>34278</td>
<td>3/25/2015</td>
</tr>
<tr>
<td>S. cerevisiae</td>
<td>SGD</td>
<td>6379</td>
<td>94252</td>
<td>48762</td>
<td>4/4/2015</td>
</tr>
<tr>
<td>A. thaliana</td>
<td>TAIR</td>
<td>30469</td>
<td>230073</td>
<td>184681</td>
<td>3/31/2015</td>
</tr>
<tr>
<td>C. elegans</td>
<td>WormBase</td>
<td>20318</td>
<td>134916</td>
<td>67739</td>
<td>9/30/2014</td>
</tr>
<tr>
<td>D. rerio</td>
<td>ZFIN</td>
<td>19655</td>
<td>167449</td>
<td>48986</td>
<td>4/6/2015</td>
</tr>
<tr>
<td>UniProt, no IEA</td>
<td>GO/EBI</td>
<td>148533</td>
<td>756506</td>
<td>756506</td>
<td>-</td>
</tr>
<tr>
<td>UniProt</td>
<td>GO/EBI</td>
<td>29516189</td>
<td>201248286</td>
<td>2114923</td>
<td>-</td>
</tr>
</tbody>
</table>

geneontology.org/page/download-annotations

fasta.bioch.virginia.edu/biol4230
Using GO to test functional conservation:
A cautionary tale

NL Nehrt, WT Clark, P Radivojac, MW Hahn (2011) "Testing the ortholog conjecture with comparative functional genomic data from mammals" PLOS Comp. Biol. 7:e1002073

A common assumption in comparative genomics is that orthologous genes share greater functional similarity than do paralogous genes (the "ortholog conjecture"). Many methods used to computationally predict protein function are based on this assumption, even though it is largely untested. Here we present the first large-scale test of the ortholog conjecture using comparative functional genomic data from human and mouse. We use the experimentally derived functions of more than 8,900 genes, as well as an independent microarray dataset, to directly assess our ability to predict function using both orthologs and paralogs. Both datasets show that paralogs are often a much better predictor of function than are orthologs, even at lower sequence identities. Among paralogs, those found within the same species are consistently more functionally similar than those found in a different species. We also find that paralogous pairs residing on the same chromosome are more functionally similar than those on different chromosomes, perhaps due to higher levels of interlocus gene conversion between these pairs. In addition to offering implications for the computational prediction of protein function, our results shed light on the relationship between sequence divergence and functional divergence. We conclude that the most important factor in the evolution of function is not amino acid sequence, but rather the cellular context in which proteins act.

Nehrt et al, (2011)
Testing the ortholog conjecture...

100% identical sequences behave differently 35 – 60% of the time
But only in different organisms

Figure 1. The relationship between functional similarity and sequence identity for human-mouse orthologs (red) and all paralogs (blue). Standard error bars are shown. (A) Biological Process ontology, (B) Molecular Function ontology. doi:10.1371/journal.pcbi.1002386


fasta.bioch.virginia.edu/biol4230 27
A recent paper (Nehrt et al., PLoS Comput. Biol. 7:e1002073, 2011) has proposed a metric for the “functional similarity” between two genes that uses only the Gene Ontology (GO) annotations directly derived from published experimental results. Applying this metric, the authors concluded that paralogous genes within the mouse genome or the human genome are more functionally similar on average than orthologous genes between these genomes, an unexpected result with broad implications if true. We suggest, based on both theoretical and empirical considerations, that this proposed metric should not be interpreted as a functional similarity, and therefore cannot be used to support any conclusions about the “ortholog conjecture” (or, more properly, the “ortholog functional conservation hypothesis”). First, we reexamine the case studies presented by Nehrt et al. as examples of orthologs with divergent functions, and come to a very different conclusion: they actually exemplify how GO annotations for orthologous genes provide complementary information about conserved biological functions. We then show that there is a global ascertainment bias in the experiment-based GO annotations for human and mouse genes: particular types of experiments tend to be performed in different model organisms. We conclude that the reported statistical differences in annotations between pairs of orthologous genes do not reflect differences in biological function, but rather complementarity in experimental approaches. Our results underscore two general considerations for researchers proposing novel types of analysis based on the GO: 1) that GO annotations are often incomplete, potentially in a biased manner, and subject to an “open world assumption” (absence of an annotation does not imply absence of a function), and 2) that conclusions drawn from a novel, large-scale GO analysis should whenever possible be supported by careful, in-depth examination of examples, to help ensure the conclusions have a justifiable biological basis.

• MAP4K2 (Map kinase kinase kinase kinase)
  – 94% human mouse orthologous sequence identity; 5% orthologous functional similarity
  – functional similarity within mouse paralogs 69%
  – human proteins belong to intracellular protein kinase cascade and protein phosphorylation (kinase activity)
  – mouse Map4K2 only annotated as vesicle targeting
  – both protein are active in the same biological processes, but different processes annotated in different organisms

• Nuclear receptors
  – THRA/ThrA (thyroid hormone receptor) vs. estrogen receptors
  – again, paralogs annotated as more similar, because ligand-specific activities not consistent in human/mouse.

• Absence of activity/process annotation does not guarantee absence of activity/process.
On the Use of Gene Ontology Annotations to Assess Functional Similarity among Orthologs and Paralogs: A Short Report

- Testing the Ortholog Conjecture (Nerht, 2011) is wrong
- By focusing on the "highest quality" annotations (experiment based), Nerht discovered that similar experiments are done in the same organism (human, mouse), but the same experiment is often not done in two different organisms (why duplicate effort?)
- Absence of activity/process annotation does not guarantee absence of activity/process.
- Very few true negative annotations

GO: The Gene Ontology (geneontology.org)

- Ontology relationships – Directed Acyclic Graph (DAG) of relationships
  - is-a
  - part-of / has-part
  - regulates / positively-regulates / negatively-regulates
- Hierarchical – three orthogonal hierarchies
  - molecular function
  - biological process
  - cellular location
  - (no sense of time, or developmental stage)
- Curated, with Evidence codes
  - experimental
  - similarity based (but curated)
  - IEA Inferred from Electronic Annotation (no human)
- Absence of activity/process annotation does NOT guarantee absence of activity/process
Gene Ontology/Enrichment Analysis

- I have a set of differentially expressed genes – what is happening to the cell?
- Gene Ontology (GO)
  - "Ontology" – a directed acyclic graph (DAG)
  - molecular function, biological process, cellular component
  - evidence and evidence codes
  - positives and negatives, missing data
  - One of many
- Function/Pathway enrichment analysis
  - do sets (subsets) of differentially expressed genes reflect a pathway?
  - Over representation analysis (ORA)
  - functional class scoring – GSEA (gene set enrichment analysis)

From Genes to Pathways: enrichment analysis

Enrichment analysis

• Given a set of differentially expressed (up/down) genes
• And a set of Gene Ontology or Pathway relationships
• Can we use the differentially expressed genes to identify the biological process/pathway involved

GO/KEGG/PFAM enrichment

• are my 100's of candidates involved in similar process/pathways/functions?
• hypergeometric test for independence:

\[
P(X = k) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}
\]

\[
\frac{a!}{b!(a-b)!} = \frac{\binom{a}{b}}{\binom{a}{b}} = \frac{a!}{b!(a-b)!}
\]

What should 'N' be?
• Total number of genes?
• Number of genes expressed?
• Number of genes up? down?

<table>
<thead>
<tr>
<th></th>
<th>difference significant</th>
<th>insignificant difference</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>in group:</td>
<td>k</td>
<td>m-k</td>
<td>m</td>
</tr>
<tr>
<td>not in group:</td>
<td>n-k</td>
<td>N+k-n-m</td>
<td>N-m</td>
</tr>
<tr>
<td>total:</td>
<td>n</td>
<td>N-n</td>
<td>N</td>
</tr>
</tbody>
</table>
The significance of differences: Fisher's Exact Test

1. Around 1930, Muriel Bristol claimed, in a conversation with R. A. Fisher, that she could tell when milk was poured into tea, which was much preferable to tea being poured into milk.
2. Fisher choose to test this hypothesis by preparing 8 cups of tea, 4 tea first, 4 milk first, and asking Ms. Bristol to identify the 4 cups with tea first.
3. If she has no ability to identify milk first/tea first, then one expects her to be right 50% of the time (2 cups). But what if she was right for 3 of the 4 cups?

```r
> fisher.test(matrix(c(4,0,0,4),nrow=2),
+ alternative='greater')
Fisher's Exact Test for Count Data

data:  matrix(c(4, 0, 0, 4), nrow = 2)
p-value = 0.01427
alternative hypothesis: true odds ratio is not equal to 1
```

Enrichment: In group / Not in group


<table>
<thead>
<tr>
<th></th>
<th>significant</th>
<th>insignificant</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>in group:</td>
<td>k</td>
<td>m-k</td>
<td>m</td>
</tr>
<tr>
<td>not in group:</td>
<td>n-k</td>
<td>N+k-n-m</td>
<td>N-m</td>
</tr>
<tr>
<td>total:</td>
<td>n</td>
<td>N-n</td>
<td>N</td>
</tr>
</tbody>
</table>

What should 'N' be?
- Total number of genes?
- Number of genes expressed?
- Number of genes up? down?
Many levels of GO annotation:

Correcting for multiple inheritance
From Genes to Pathways: enrichment analysis

- over-representation analysis (ORA)
  - expected vs. observed #s of DEGs that share:
    - a GO term
    - a KEGG/Reactome/IPA pathway
    - TF/cis-regulatory promoter elements
    - miRNA targets in 3’ UTR
    - disease associations (GWAS, etc)
- hundreds of tools for this, differing by environment, statistics, database, visualization
- one favorite: GOrilla
  - [http://cbl-gorilla.cs.technion.ac.il/](http://cbl-gorilla.cs.technion.ac.il/)

Over Representation Analysis - Reproducibility

(A) “Stemness” genes. (B) ESC-enriched genes (C) NPC-enriched genes. (D) Overlap of “stemness” genes—two types of stem cell (ESC/NPC)-enriched genes

Issues with Over Representation Analysis (ORA)

1. arbitrary significance thresholds for inclusion
2. Differential Expression magnitude/directionality not considered
3. sensitive to choice of background “universe”
   – all genes, genes on chip, or genes with sufficient signal that could possibly be called DEG?
4. correlation between genes ignored
5. correlation/cross-talk between pathways

Functional Class Scoring (FCS) methods fix #1-3

FCS: Gene Set Enrichment Analysis (GSEA)

Given an a priori defined set of genes $S$ (e.g., genes encoding products in a metabolic pathway, located in the same cytogenetic band, or sharing the same GO category), the goal of GSEA is to determine whether the members of $S$ are randomly distributed throughout list $L$ or primarily found at the top or bottom.

- no P value/FDR threshold
- more sensitive than hypergeometric tests
- statistics calculated by permutation testing

FCS: Gene Set Enrichment Analysis (GSEA)

The distribution of three gene sets, from the C2 functional collection, in the list of genes in the male female lymphoblastoid cell line example ranked by their correlation with gender: S1, a set of chromosome X inactivation genes; S2, a pathway describing vitamin c import into neurons; S3, related to chemokine receptors expressed by T helper cells. Shown are plots of the running sum for the three gene sets: S1 is significantly enriched in females as expected, S2 is randomly distributed and scores poorly, and S3 is not enriched at the top of the list but is nonrandom, so it scores well. Arrows show the location of the maximum enrichment score and the point where the correlation (signal-to-noise ratio) crosses zero. The new method reduces the significance of sets like S3.

FCS: Gene Set Enrichment Analysis (GSEA)


from genes to pathways: enrichment analysis
Pathway Topology: PT vs ORA set enrichment vs. pathway impact

SPIA, DEAP, CePa, PathwayExpress

The X-axis shows the over-representation evidence, while the Y-axis shows the perturbation evidence. In the top-left plot, areas 2, 3 and 6 together will include pathways that meet the over-representation criterion (PNDE < α). Areas 1, 2 and 4 together will include pathways that meet the perturbation criterion (PPERT < α). Areas 1, 2, 3 and 5 will include the pathways that meet the combined SPIA criteria (PG < α). Note how SPIA results are different from a mere logical operation between the two criteria (OR would be areas 1, 2, 3, 4 and 6; AND would be area 2).

Pathway analysis results on the Colorectal cancer (top right), LaborC (bottom left) and Vessels (bottom right) datasets. Each pathway is represented by a point. Pathways above the oblique red line are significant at 5% after Bonferroni correction, while those above the oblique blue line are significant at 5% after FDR correction. The vertical and horizontal thresholds represent the same corrections for the two types of evidence considered individually. Note that for the Colorectal cancer dataset (top right), the colorectal cancer pathway (ID = 5210) is only significant according to the combined evidence but not so according to any individual evidence PNDE or PPERT.

SPIA – Signaling Pathway Impact Analysis

pNDE: from ORA
pPERT: from perturbation

SPIA – Signaling Pathway Impact Analysis

Figure 3: SPIA evidence plot for the colorectal cancer dataset. Each pathway is represented by one dot. The pathways at the right of the red curve are significant after Bonferroni correction of the global p-values, \( p_G \), obtained by combining the \( p_{PERT} \) and \( p_{NDE} \) using the normal inversion method. The pathways at the right of the blue curve line are significant after a FDR correction of the global p-values, \( p_G \).

The green dot shows the KEGG:05210 colon cancer pathway. This pathway is marginally significant (RDR < 0.05) with "normal inversion" combination of PERT and NDE, but not significant with Fisher's method.

pathway crosstalk yields false positives:

A rank pathway \( p(\text{fdr}) \)
1 Parkinson's disease 0.0007
2 Alzheimer's disease 0.0006
3 Huntington's disease 0.0005
4 Leukemia 0.0004
5 Diabetes 0.0003
6 Cell cycle 0.0011
7 Oocyte meiosis 0.0016
8 Oocyte meiosis contraction 0.0016
9 Toll-like receptor 0.0018
10 TNF signal pathway 0.0033
11 Chemokine signaling pathway 0.0054
12 Lysosome 0.0071
13 IRE1 signaling pathway 0.0095
14 Compl. and coag. cascade 0.0342
15 Cytokine-cytokine rec. inter. 0.0346
16 Urea cycle 0.0406
17 Progesterone-cyc. maturation 0.0530
18 Fc epsilon RI signaling pathway 0.0548
19 Leukocyte transendoth. migr. 0.0548

B rank pathway \( p(\text{fdr}) \)
1 Metabolism of alcohol 0.0010
2 MAPK signaling pathway 0.0115
3 Cell cycle 0.0145
4 TNF signaling pathway 0.0208
5 PI3K-Akt signaling path 0.0220
6 * Cytok-cytok. rec. int. 0.0343
7 Toll-like receptor signaling 0.0501
8 MAPK signaling pathway 0.1158
9 B-cell receptor signaling 0.1450
10 Lysosomes 0.1787
11 Nat. killer cell med. cyt. 0.1974
12 * Cell cycle 0.2204
13 Calcium signaling pathway 0.2209
14 Cell adhesion molecules 0.2588
15 NOD-like receptor signaling 0.2588
16 Vasc. smooth muscle contr. 0.4246
17 Dilated cardiomyopathy 0.4248
18 * Oocyte matosis 0.4320
19 Type I diabetes mellitus 0.4322
20 Wnt signaling pathway 0.4783

The results of the ORA analysis in the fat remodeling experiment for the comparison between days 3 and 0, before (A) and after (B) correction for crosstalk effects. All P-values are FDR corrected. The lines show the significance thresholds: (blue) 0.01, (yellow) 0.05. Pathways highlighted in red represent pathways not related to the phenomenon in analysis, while pathways highlighted in green are those for which we know, with reasonable confidence, are involved in the given phenomenon. The white background indicates pathways for which we do not have conclusive information on their involvement (or lack of it) with the phenomenon in analysis. (A) The top 20 pathways resulting from classical ORA before correction for crosstalk. The top four pathways are not related to fat remodeling. (B) The top 20 pathways after correction for crosstalk. Pathways ranked 1, 3, and 5 are modules that are functioning independently of the rest of their pathways in this particular condition. Starred pathways are pathways edited by removing such modules. Note the lack of any obvious false positive above the significance threshold(s).

Functional analysis: ORA, FC, PT

- Methods assume independence, but pathways and GO DAGs are anything but independent
  - statistics may be too generous (false positives)
  - statistics may be too strict (false negatives)
- What is the right control?
  - try different approaches?
  - compare to other published datasets?
  - do "positive control" on well understood pathways
- All methods need experimental confirmation
  - find a drug that blocks the pathway
  - ablate a gene (or genes) in the pathway

Function/Pathway Enrichment

- Function/Pathway enrichment analysis
  - do sets (subsets) of differentially expressed genes reflect a pathway?
- Over Representation Analyis (ORA)
  - Fisher exact test, hypergeometric
  - competitive vs. self-contained tests
- Functional Class Scoring (FTS)
  - GSEA : Gene Set Enrichment Analysis
- Pathway Topology (PT)
  - SPIA : Signaling Pathway Impact Analysis
- What are the right "controls"?
Gene Ontology/Enrichment Analysis

• Biological function in the cell: Pathways (chemistry)
• Gene Ontology (GO)
  – "Ontology" – a directed acyclic graph (DAG)
  – molecular function, biological process, cellular component
  – evidence and evidence codes
  – positives and negatives, missing data
  – One of many
• Function/Pathway enrichment analysis
  – do sets (subsets) of differentially expressed genes reflect a pathway?