Learning Objectives of Module

• What identifiers can be used for genes?
• What gene annotations are available for genes?
• What is a gene-set enrichment test and how does it work?
• Why do I need multiple test correction for gene-set enrichment and how does it work?
• How to visualize gene-set enrichment results using Enrichment Map
• General principles of network visualization in Cytoscape
Introduction

Interpreting Gene Lists

• My cool new screen worked and produced 1000 hits! ...Now what?
• Genome-Scale Analysis (Omics)
  – Genomics, Proteomics
• Tell me what’s interesting about these genes
  – Are they enriched in known pathways, complexes, functions
Pathway and network analysis

- Save time compared to traditional approach
Gene Identification

• Getting gene IDs right is important
  – Identify the right entity
  – Stable and traceable

• Issues to keep in mind:
  – What is the output of the experiment?
  – Are the annotations used to analyze the experimental data in a compatible ID system?
  – Is the statistical test appropriate? (most used tests assume a random uniform distribution over genes)
ID Challenges

• Avoid errors: map IDs correctly
  – Beware of 1-to-many mappings

• Gene name ambiguity – not a good ID
  – e.g. FLJ92943, LFS1, TRP53, p53
  – Better to use the standard gene symbol: TP53

• Excel error-introduction
  – OCT4 is changed to October-4 (paste as text)

• Problems reaching 100% coverage
  – E.g. due to version issues
  – Use multiple sources to increase coverage

Zeeberg BR et al. Mistaken identifiers: gene name errors can be introduced inadvertently when using Excel in bioinformatics BMC Bioinformatics. 2004 Jun 23;5:80

ID Mapping Services

• g:Convert
  • http://bit.cs.ut.ee/gprofiler/gconvert.cgi

• Ensembl Biomart
  • http://www.ensembl.org
Gene Annotations and Gene-sets

Module 8

From Cell Biology to Gene-sets

Where can I get these gene-sets?
How were the gene-sets compiled?
How are they structured?
Pathways and other gene function attributes

• Available in databases

• Pathways
  – Gene Ontology biological process, pathway databases e.g. Reactome

• Other annotations
  – Gene Ontology molecular function, cell location
  – Chromosome position
  – Disease association
  – DNA properties
    • TF binding sites, gene structure (intron/exon), SNPs
  – Transcript properties
    • Splicing, 3’ UTR, microRNA binding sites
  – Protein properties
    • Domains, secondary and tertiary structure, PTM sites
  – Interactions with other genes
What is the Gene Ontology (GO)?

- Set of biological phrases (terms) applied to genes:
  - protein kinase
  - apoptosis
  - membrane
- Dictionary: term definitions
- Ontology: A formal system for describing knowledge
- www.geneontology.org

GO Structure

- Terms are related within a hierarchy
  - is-a
  - part-of
- Describes multiple levels of detail of gene function
- Terms can have more than one parent or child
What does GO cover?

• GO terms divided into three aspects:
  – cellular component
  – molecular function
  – biological process

Cell division

glucose-6-phosphate isomerase activity

Part 1/2: Terms

• Where do GO terms come from?
  – GO terms are added by editors at EBI and gene annotation database groups
  – Terms added by request
  – Experts help with major development

<table>
<thead>
<tr>
<th></th>
<th>Jun 2012</th>
<th>Apr 2015</th>
<th>increase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biological process</td>
<td>23,074</td>
<td>28,158</td>
<td>22%</td>
</tr>
<tr>
<td>Molecular function</td>
<td>9,392</td>
<td>10,835</td>
<td>15%</td>
</tr>
<tr>
<td>Cellular component</td>
<td>2,994</td>
<td>3,903</td>
<td>30%</td>
</tr>
<tr>
<td>total</td>
<td>37,104</td>
<td>42,896</td>
<td>16%</td>
</tr>
</tbody>
</table>
Part 2/2: Annotations

• Genes are linked, or associated, with GO terms by trained curators at genome databases
  – Known as ‘gene associations’ or GO annotations
  – Multiple annotations per gene
• Some GO annotations created automatically (without human review)

Hierarchical annotation

• Genes annotated to specific term in GO automatically added to all parents of that term

AURKB
Annotation Sources

• Manual annotation
  – Curated by scientists
    • High quality
    • Small number (time-consuming to create)
  – Reviewed computational analysis

• Electronic annotation
  – Annotation derived without human validation
    • Computational predictions (accuracy varies)
    • Lower ‘quality’ than manual codes

• Key point: be aware of annotation origin

Evidence Types

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

• Author Statement Evidence Codes
  • TAS: Traceable Author Statement
  • NAS: Non-traceable Author Statement

• Curator Statement Evidence Codes
  • IC: Inferred by Curator
  • ND: No biological data available

• Computational Analysis Evidence Codes
  • ISS: Inferred from Sequence or Structural Similarity
  • ISO: Inferred from Sequence Orthology
  • ISA: Inferred from Sequence Alignment
  • ISM: Inferred from Sequence Model
  • IGC: Inferred from Genomic Context
  • RCA: Inferred from Reviewed Computational Analysis

• IEA: Inferred from electronic annotation

Evidence codes in g:Profiler

- Input gene list

<table>
<thead>
<tr>
<th>Evidence codes</th>
<th>Gene annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO and gene annotations are evolving rapidly</td>
<td>Wadi et al. in review; bioRxiv [<a href="http://dx.doi.org/10.1101/049288">http://dx.doi.org/10.1101/049288</a>]</td>
</tr>
</tbody>
</table>
Many GO tools are out of date

Out-of-date tools (2010) miss up to 80% of enriched gene sets

Wadi et al. in review; bioRxiv http://dx.doi.org/10.1101/049288
Pathways

What are Pathways?

• Depict mechanistic details of metabolic, signaling and other biological processes

• Advantages:
  – Curated, accurate
  – Cause and effect captured.
  – Human-interpretable visualizations

• Disadvantages:
  – More sparse coverage of genome than functional sets
  – More complex models are required to score pathways
  – Static model of dynamic systems
Gene-set Enrichment Tests

Activity Profiles / Somatic Mutations

Prior Knowledge about genes

Activity Maps

Gene sets
Gene networks
Pathways

Scoring models
Search algorithms
Informatics
The output of an enrichment test is a *P-value*. The P-value assesses the probability that, by random sampling the “detectable” genes, the overlap is at least as large as observed.

*Most used statistical model: Fisher’s Exact Test*
**Fisher’s Exact Test**

### 2 x 2 Contingency Table

<table>
<thead>
<tr>
<th></th>
<th>Exp_positive=yes</th>
<th>Exp_positive=no</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene-Set=yes</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>Gene-Set=no</td>
<td>c</td>
<td>d</td>
</tr>
</tbody>
</table>

Probability of one table to occur by random sampling:

**Hypergeometric distribution** formula:

\[
p = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!}
\]

Test **p-value**: sum of random sampling probabilities for tables as extreme or more extreme than the real table

http://en.wikipedia.org/wiki/Fisher's_exact_test

**Importance of the Background**

- Inappropriate modeling of the background will lead to incorrectly biased results
  - E.g.: kinase phosphorylation assay: only kinases can be detected
- Depending on the experiment, the background may be easy or difficult to define
Multiple test corrections

How to win the P-value lottery, part 1

Background population: 500 black genes, 4500 red genes

Random draws

… 7,834 draws later …

Expect a random draw with observed enrichment once every $1 / P$-value draws
How to win the P-value lottery, part 2
Keep the gene list the same, evaluate different annotations

Observed draw
- RRP6
- MRD1
- RRP7
- RRP43
- RRP42

Different annotation
- RRP6
- MRD1
- RRP7
- RRP43
- RRP42

Black vs red nodes
Square vs round nodes

Simple P-value correction: Bonferroni

If $M =$ number of annotations tested:

Corrected P-value = $M \times$ original P-value

Corrected P-value is greater than or equal to the probability that at least one (or more) of the observed enrichments is 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 (e.g. 5%).
- 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 I

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

Sort P-values of all tests in decreasing order

Benjamini-Hochberg example II

<table>
<thead>
<tr>
<th>Rank</th>
<th>Category</th>
<th>(Nominal) 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 number of tests divided by the rank of the P-value in sorted list
### Benjamini-Hochberg example III

<table>
<thead>
<tr>
<th>Rank</th>
<th>Category</th>
<th>(Nominal) 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</td>
<td>0.053</td>
</tr>
<tr>
<td>2</td>
<td>Transcription factor</td>
<td>0.002</td>
<td>0.002 x 53/2</td>
<td>0.053</td>
</tr>
<tr>
<td>3</td>
<td>Initiation of transcription</td>
<td>0.003</td>
<td>0.003 x 53/3</td>
<td>0.053</td>
</tr>
<tr>
<td>4</td>
<td>Nuclear localization</td>
<td>0.0031</td>
<td>0.0031 x 53/4</td>
<td>0.040</td>
</tr>
<tr>
<td>5</td>
<td>Chromatin modification</td>
<td>0.005</td>
<td>0.005 x 53/5</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</td>
<td>1.004</td>
</tr>
<tr>
<td>53</td>
<td>Translation</td>
<td>0.99</td>
<td>0.99 x 53/53</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 higher ranks.

---

### Benjamini-Hochberg example III

<table>
<thead>
<tr>
<th>Rank</th>
<th>Category</th>
<th>(Nominal) 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</td>
<td>0.053</td>
</tr>
<tr>
<td>2</td>
<td>Transcription factor</td>
<td>0.002</td>
<td>0.002 x 53/2</td>
<td>0.053</td>
</tr>
<tr>
<td>3</td>
<td>Initiation of transcription</td>
<td>0.003</td>
<td>0.003 x 53/3</td>
<td>0.053</td>
</tr>
<tr>
<td>4</td>
<td>Nuclear localization</td>
<td>0.0031</td>
<td>0.0031 x 53/4</td>
<td>0.040</td>
</tr>
<tr>
<td>5</td>
<td>Chromatin modification</td>
<td>0.005</td>
<td>0.005 x 53/5</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</td>
<td>1.004</td>
</tr>
<tr>
<td>53</td>
<td>Translation</td>
<td>0.99</td>
<td>0.99 x 53/53</td>
<td>0.99</td>
</tr>
</tbody>
</table>

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

Red: non-significant
Green: significant at FDR < 0.05

**Q-value (or FDR) corresponding to a nominal P-value**

is the smallest adjusted P-value assigned to P-values with the same or higher ranks.
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. restrict testing to the appropriate GO annotations; or filter gene sets by size.

Enrichment Results Visualization:
Enrichment Map
### Gene-set Enrichment Redundancy Problem

- Many redundant gene-sets
  - Gene Ontology has a very large number of gene-sets, often with slight differences
  - Different pathway databases have different yet overlapping definitions of pathways
  - Globally, it is useful to grasp the overlap relations between enriched gene-sets

--> we need a visualization framework going beyond the enrichment table
Summarising pathway analysis with Enrichment Map

Nodes – gene sets reflecting pathways, processes
Edges – sets share many common genes
Network clustering groups processes/pathways as themes

Ependymoma subtyping

- A form of pediatric and adult cancer in brain and CNS
- Pathology is primary means of classification and grading, limited clinical utility
- New methylation biomarkers developed for classification
- Reveals 9 subtypes of ependymoma
- Characterised by distinct molecular alterations, gene expression, clinical properties
- What are common and subtype-specific pathways?

Molecular classification of ependymal tumors across all CNS compartments, histopathological grades, and age groups - KW Pajtler et al Cancer Cell 2015
Pathways activated in EPN subtypes

Cytoscape Network Visualization
Network Representations

How to visually interpret biological data using networks
Merico D, Gfeller D, Bader GD
Nature Biotechnology 2009 Oct 27

Cytoscape

- Cytoscape is a freely available, open-source, Java-based application
- Very popular in the community, provides key functionalities, extended by plugins (now called “apps” to be cool)
Key Ideas in Network Visualization

• Layout
  – Automatic layout algorithms are necessary to arrange a network in a way that suggests the existence of patterns to the human eye

• Node and Edge visual attributes
  – Can be used to map a number of information items relating to gene / proteins and their interactions / similarity

Layout: Before and After
Visual Attributes