Extracting Biological Information from Gene Lists

Simon Andrews, Laura Biggins, Boo Virk

simon.andrews@babraham.ac.uk
laura.biggins@babraham.ac.uk
boo.virk@babraham.ac.uk

v2019-09
Biological material

Isolation of DNA, RNA or proteins

Sample for analysis

Sample processing

Analysis of processed sample: Data acquisition – sequencing, microarray analysis, mass spectrometry

Data analysis:
- identification of genes, transcripts or proteins

Public databases

Results Table

Containing hits – genes, transcripts or proteins

What does this mean???
Biological themes are not always obvious from gene lists

Relate the hits to existing knowledge
Descriptions aren’t always informative

<table>
<thead>
<tr>
<th>Gene</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gpr55</td>
<td>G protein-coupled receptor 55 [Source:MGI Symbol;Acc:MGI:2685064]</td>
</tr>
<tr>
<td>Ncl</td>
<td>nucleolin [Source:MGI Symbol;Acc:MGI:97286]</td>
</tr>
<tr>
<td>Aspm</td>
<td>asp (abnormal spindle)-like, microcephaly associated (Drosophila) [Source:MGI Symbol;Acc:MGI:1334448]</td>
</tr>
<tr>
<td>Tnfsf4</td>
<td>tumor necrosis factor (ligand) superfamily, member 4 [Source:MGI Symbol;Acc:MGI:104511]</td>
</tr>
<tr>
<td>Ephx1</td>
<td>epoxide hydrolase 1, microsomal [Source:MGI Symbol;Acc:MGI:95405]</td>
</tr>
<tr>
<td>Setx</td>
<td>senataxin [Source:MGI Symbol;Acc:MGI:2443480]</td>
</tr>
<tr>
<td>Angptl2</td>
<td>angiopoietin-like 2 [Source:MGI Symbol;Acc:MGI:1347002]</td>
</tr>
<tr>
<td>Ggta1</td>
<td>glycoprotein galactosyltransferase alpha 1, 3 [Source:MGI Symbol;Acc:MGI:95704]</td>
</tr>
<tr>
<td>Dab2ip</td>
<td>disabled homolog 2 (Drosophila) interacting protein [Source:MGI Symbol;Acc:MGI:1916851]</td>
</tr>
<tr>
<td>Neb</td>
<td>nebulin [Source:MGI Symbol;Acc:MGI:97292]</td>
</tr>
<tr>
<td>Ermn</td>
<td>ermin, ERM-like protein [Source:MGI Symbol;Acc:MGI:1925017]</td>
</tr>
<tr>
<td>Ckap5</td>
<td>cytoskeleton associated protein 5 [Source:MGI Symbol;Acc:MGI:1923036]</td>
</tr>
<tr>
<td>Prr5l</td>
<td>proline rich 5 like [Source:MGI Symbol;Acc:MGI:1919696]</td>
</tr>
<tr>
<td>Arhgap11a</td>
<td>Rho GTPase activating protein 11A [Source:MGI Symbol;Acc:MGI:2444300]</td>
</tr>
<tr>
<td>Bub1b</td>
<td>budding uninhibited by benzimidazoles 1 homolog, beta (S. cerevisiae) [Source:MGI Symbol;Acc:MGI:1333889]</td>
</tr>
<tr>
<td>Prnp</td>
<td>prion protein [Source:MGI Symbol;Acc:MGI:97769]</td>
</tr>
<tr>
<td>Fam102b</td>
<td>family with sequence similarity 102, member B [Source:MGI Symbol;Acc:MGI:3036259]</td>
</tr>
</tbody>
</table>
Reading up on individual genes can be slow and confusing.
Functional Analysis – Course Outline

• Gene Set Enrichment
  – Theory and data
  – Practical
  – Artefacts and Biases
  – Presenting Results

• Sequence analysis
  – Motif analysis theory
  – Motif analysis practical
Functional analysis relates current data to existing knowledge

Advantages:
• Biological insight
• Validation of experiment
• Generate new hypotheses

Limitations:
• You can only discover what is already known
  – Novel functionality will be missing
  – Existing annotations may be incorrect
  – Many species are poorly supported
Nothing is ever straight forward…

Best hit: “DNA Methylation” p<2e-10

- name: DNA methylation
- datasource: reactome
- organism: Human
- idtype: hgnc symbol
- Genes:
  - Methyltransferases: DNMT1 DNMT3A DNMT3B DNMT3L
  - Methyltransferase targeting protein: UHRF1
  - **Histones!!!** H2AFB1 H2AFJ H2AFV H2AFX H2AFZ H2BFS H3F3A H3F3B HIST1H2AB HIST1H2AC HIST1H2AD HIST1H2AE HIST1H2AJ HIST1H2BA HIST1H2BB HIST1H2BC HIST1H2BD HIST1H2BE HIST1H2BF HIST1H2BG HIST1H2BH HIST1H2BI HIST1H2BJ HIST1H2BK HIST1H2BL HIST1H2BM HIST1H2BN HIST1H2BO HIST1H3A HIST1H3B HIST1H3C HIST1H3D HIST1H3E HIST1H3F HIST1H3G HIST1H3H HIST1H3I HIST1H3J HIST1H4A HIST1H4B HIST1H4C HIST1H4D HIST1H4E HIST1H4F HIST1H4H HIST1H4I HIST1H4J HIST1H4K HIST1H4L HIST2H2AA3 HIST2H2AA4 HIST2H2AC HIST2H2BE HIST2H3A HIST2H3C HIST2H3D HIST2H4A HIST2H4B HIST3H2BB HIST4H4
Most functional analysis starts from gene lists

• Many considerations
  – Other start points
    • Genomic positions
    • Transcripts / Proteins
  – Gene nomenclature
  – Annotation sources / versions

• Types of list
  – Categorical (hit or not a hit)
  – Ordered
A functional gene set provides a group of genes with a common biological relationship.

**Germ-line stem cell division**

The self-renewing division of a germline stem cell to produce a daughter stem cell and a daughter germ cell, which will divide to form the gametes.

<table>
<thead>
<tr>
<th>Gene/product</th>
<th>Gene/product name</th>
<th>Organism</th>
<th>PANTHER family</th>
<th>Type</th>
<th>Source</th>
<th>Synonyms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hoxc4</td>
<td>homeobox C4</td>
<td>Mus musculus</td>
<td>family not named pthr24326</td>
<td>protein</td>
<td>MGI</td>
<td>Hox-3.5</td>
</tr>
<tr>
<td>Ing2</td>
<td>inhibitor of growth family, member 2</td>
<td>Mus musculus</td>
<td>inhibitor of growth protein pthr10333</td>
<td>protein</td>
<td>MGI</td>
<td>2810011M06Rik Ing11 P33ING2</td>
</tr>
<tr>
<td>Stra8</td>
<td>stimulated by retinoic acid gene 8</td>
<td>Mus musculus</td>
<td></td>
<td>protein</td>
<td>MGI</td>
<td></td>
</tr>
<tr>
<td>Zbtb16</td>
<td>zinc finger and BTB domain containing 16</td>
<td>Mus musculus</td>
<td>zinc finger protein pthr11389</td>
<td>protein</td>
<td>MGI</td>
<td>Green's luxoid PLZF Zfp145</td>
</tr>
<tr>
<td>Etv5</td>
<td>ets variant 5</td>
<td>Mus musculus</td>
<td>ets pthr11849</td>
<td>protein</td>
<td>MGI</td>
<td>1110005E01Rik 8430401F14Rikerm</td>
</tr>
</tbody>
</table>
Functional analysis relates your hits to a set of pre-defined functional groups

<table>
<thead>
<tr>
<th>A4galt</th>
<th>Flywch1</th>
<th>Mypop</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atl1</td>
<td>Gnpda2</td>
<td>Rnf6</td>
</tr>
<tr>
<td>Cdk19</td>
<td>Hoxc4</td>
<td>Serinc1</td>
</tr>
<tr>
<td>Cdon</td>
<td>Ing2</td>
<td>Stra8</td>
</tr>
<tr>
<td>Cecr2</td>
<td>Iigp1</td>
<td>Trp73</td>
</tr>
<tr>
<td>Etv5</td>
<td>Map3k9</td>
<td>Zbtb16</td>
</tr>
</tbody>
</table>
Functional analysis relates your hits to a set of pre-defined functional groups

A4galt  Flywch1  Mypop
Atl1      Gnpda2   Rnf6
Cdk19    Hoxc4    Serinc1
Cdon    Ing2    Stra8
Cecr2    ligp1    Trp73
Etv5    Map3k9    Zbtb16

Germ-line stem cell division

The self-renewing division of a germline stem cell to produce a daughter stem cell and a daughter germ cell, which will divide to form the gametes.
There are many sources of functional gene lists

- Human curated
  - Gene Ontology
  - Biological Pathways

- Domains / Patterns
  - Protein functional domains
  - Transcription factor regulated

- Experimental
  - Co-expressed genes
  - Interactions
  - Hits from other studies
Gene Ontology is a human curated functional database
GO has three domains and a hierarchical structure
Genes are placed into each domain as specifically as possible

Nanog homeobox [Source:HGNC Symbol;Acc:HGNC:20857]

- **Cellular Component**
  - GO:0005634 nucleus
  - GO:0005654 nucleoplasm
  - GO:0005730 nucleolus

- **Molecular Function**
  - GO:0003677 DNA binding
  - GO:0003700 transcription factor activity, sequence-specific DNA binding
  - GO:0003714 transcription corepressor activity
  - GO:0005515 protein binding
  - GO:0043565 sequence-specific DNA binding

- **Biological Process**
  - GO:0001714 endodermal cell fate specification
  - GO:0006351 transcription, DNA-templated
  - GO:0006355 regulation of transcription, DNA-templated
  - GO:0007275 multicellular organism development
  - GO:0008283 cell proliferation
  - GO:0019827 stem cell population maintenance
  - GO:0030154 cell differentiation
  - GO:0035019 somatic stem cell population maintenance
  - GO:0045595 regulation of cell differentiation
  - GO:0045944 positive regulation of transcription from RNA polymerase II promoter
  - GO:1903507 negative regulation of nucleic acid-templated transcription
Annotations come with evidence

• Experimental Evidence
  – 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)
Annotations come with evidence

• Computational Evidence
  – 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)
Annotations come with evidence

• Publications
  – Traceable Author Statement (TAS)
  – Non-traceable Author Statement (NAS)
• Curators
  – Inferred by Curator (IC)
  – No biological Data available (ND)
• Automated assignment
  – Inferred from Electronic Annotation (IEA)
Annotations come with evidence

- It looks like something which is annotated: 3.5e+06 counts
- Actual experimental evidence: 5e+05 counts
- Curator interpretation: 2e+05 counts
- Claimed in a paper: 1e+04 counts
- Mixture of sources: 5e+02 counts
- Annotated based on where it is in the genome: 1e+02 counts
Pathway databases trace metabolic pathways and their regulation
Interaction databases map out physical interactions between genes and their products.
Protein Domain databases map out functional subdomains within proteins

Pfam

InterPro

SMART
Co-expression databases group genes which are expressed together

GeneFriends

COXPRESSIONdb
Transcription Factor databases group genes by the motifs in their promoters
Some databases collate gene sets from many different sources
Testing for enriched gene sets
There are two basic ways to test for enrichment

• Categorical
  – Start from a list of hit genes
  – No ordering to hits
  – Compares proportions of hits

• Quantitative
  – Start with all genes
  – Associate a value with each gene
  – Look for functional sets with unusual distributions of values
Categorical Enrichment Analysis
Categorical tests for enrichment

13,101 genes on chip

3005 genes related to disease
3005/13,101 = 23.1%

Gene List

3005 genes related to disease
3005/13,101 = 23.1%

3005 genes related to disease
3005/13,101 = 23.1%

13,101 genes on chip

3005 genes related to disease
3005/13,101 = 23.1%

Gene List

Not related to disease

Related to disease
260/747 = 34.8%

Background

In disease annotated group
260
3005

Not in disease annotated group
487
10096

Not related to disease
## Fisher’s Exact test

<table>
<thead>
<tr>
<th>Gene List</th>
<th>Background</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>In disease annotated group</td>
<td>260</td>
<td>3005</td>
</tr>
<tr>
<td>E = 176.1</td>
<td>E = 3088.8</td>
<td></td>
</tr>
<tr>
<td>Not in disease annotated group</td>
<td>487</td>
<td>10096</td>
</tr>
<tr>
<td>E = 570.9</td>
<td>E = 10012.1</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>747</td>
<td>13101</td>
</tr>
</tbody>
</table>

\[
\frac{(260/487)}{(3005/10096)}
\]

```r
> counts <- matrix(c(260, 487, 3005, 10096), nrow = 2)
> fisher.test(counts)

Fisher's Exact Test for Count Data

data: counts
p-value = 9.769e-13
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval: 1.528462 2.10120
sample estimates:

    odds ratio
1.793564
```
Categorical tests are influenced by where you set the cutoff for “interesting” genes

- Function X
  - 3 hits out of 32 in ‘interesting’ list
  - Not significant (p=0.07)
Categorical tests are influenced by where you set the cutoff for “interesting” genes

- **Function X**
  - 3 hits out of 7 in ‘interesting’ list
  - Significant (p=0.02)
Ordered, but not quantitative lists allow sequential categorical analysis

• Function X
  – Length=1 \( p=0.60 \)
  – Length=2 \( p=0.80 \)
  – Length=3 \( p=0.30 \)
  – Length=4 \( p=0.35 \)
  – Length=5 \( p=0.40 \)
  – Length=6 \( p=0.45 \)
  – Length=7 \( p=0.05 \)
  – Length=8 \( p=0.08 \)
  – Length=9 \( p=0.10 \)
Quantitative Enrichment Analysis
Quantitative comparisons offer more power, if you have a suitable metric

• What quantitation can we use?
  – Differential p-value (normally $-10 \log(p)$)
  – Fold change
  – Absolute difference

• Measures often have odd distributions and biases
  – Z-scores
  – Ranks
What kind of changes do we expect in an interesting category?

Genes in that category all change, and by about the same amount?
What kind of changes do we expect in an interesting category?

Genes in that category all change in the same direction, but by different amounts?
What kind of changes do we expect in an interesting category?

Genes in that category all change in either direction, but by different amounts?
The expected change influences the statistical test used

• All changing by similar amounts
  – Student’s T-test

• All changing in the same direction but by different amounts
  – Kolmogorov Smirnov test
  – Wilcoxon Rank Sum test

• All changing but in either direction
  – Absolute Kolmogorov Smirnov test
Kolmogorov Smirnov

Annotated for function
Quantitated gene list (e.g. fold change)

- Looks for the biggest point of difference between the background and test lists
Multiple testing correction

- More annotations/functions being tested = more chance of increase in false-positives

**Bonferroni**
- Significance level (e.g. 0.05) /number of tests = new threshold
- Over correction if tests are correlated

**Benjamini-Hochberg**
- Rank the p-values
- Apply more stringent correction to the most significant, and least stringent to the least significant p-values
What do we get back from an enrichment test?

• A p-value
  – Remember that this reflects not only difference but also variance and power (number of observations)

• A difference value
  – Enrichment difference (odds ratio)
  – Mean quantitative difference
  – Remember large differences are easier to obtain with small numbers of observations
Tools for functional gene list analysis

• There are many different tools available, both free and commercial

• Popular tools include:
• Categorical Statistics
• Most popular system (mostly historic)
• Has been behind the latest annotation
  – Was recently updated though
• Lots of support for different IDs and Species
• Configurable gene sets
• Simple output presentation
• Categorical Statistics
• Biggest selection of gene sets
• Simple interface, but limited options
  – No species information
  – No background list option
• Simple interactive visualisation
• Novel scoring scheme to rank hits
• Categorical or Quantitative statistics
• Part of Gene Ontology Consortium
  – Annotations are up to date
• Simple enrichment analysis
• Functional lists and categorical break down
• Categorical or ranked analysis
• Mostly GO gene list support
• Interesting visualisation options
• Categorical or ordered statistics
• Lots of additional options
• Wide species support
• Interesting presentation
  – Doesn’t scale well to lots of hits
GSEA

- Quantitative enrichment
- Designed for expression datasets
- Local application
- Imports tab delimited expression data
- New version (v3) is open source – older versions are not
• Genes ranked based on correlation to annotation groups
• Genes from a gene set placed onto the ranked lists
• Look for sets where there is unusual grouping at the top or the bottom of the list
- Quantitative enrichment of sequencing datasets
- Local Java application
Gene List
Practical