Gene List
Enrichment Analysis

George Bell, Ph.D.
BaRC Hot Topics
March 16, 2010

Outline

• Why do enrichment analysis?
• Main types
• Selecting or ranking genes
• Annotation sources
• Statistics
• Remaining issues
• Presenting findings
• Recommended tools

Why do enrichment analysis?

• Most array, sequencing, and screens produce
  – A measurement for most or all genes
  – List(s) of “interesting” genes
• Most cellular processes involve sets of genes.
• Can we compare the above two datasets?
• Is the overlap different than expected?
• Does this tell us something about cellular mechanisms?

Why not just link genes to physiology?

• Too many genes to examine in detail.
• Are we biased?
• How do we know that what we’re seeing is surprising?

Genome = 20,000 genes
Our list = 100 genes
Schmooase activity = 1000 genes
Intersection = 10 genes
Main types of enrichment analysis

- **List-based: inputs are**
  - A subset of all genes chosen by some relevant method
  - A list of annotations, each linked to genes
- **Rank-based: inputs are**
  - A set of all genes ranked by some metric (ratio, fold change, etc.)
  - A list of annotations, each linked to genes
- **List-based with relationships: inputs are**
  - A subset of all genes
  - A list of annotations, each linked to genes, organized in some relationship (e.g., a hierarchy)

Getting your list

- **Goal:** Identify a list of genes (or probes) that appear to be working together in some way.
- **What identifiers to use?**
  - **Most common method:** Get a list of differentially expressed genes
    - P-value or fold change?
    - Threshold?
  - **Alternatives:**
    - Define a cluster
    - Sort data and/or apply a model to rank genes
- **Recommendations:**
  - Try lists of varying length
  - Try to maximize signal / noise (What produces the smallest p-values for enrichment?)

Annotation sources

- **Gene Ontology (most popular)**
  - biological process, molecular function, cellular component
  - Terms may have >1 “parent” (more general term)
  - GO Slim: includes only general categories
- **KEGG; REACTOME pathways**
- **Genes sharing a motif of regulated by the same protein/miRNA**
- **Genes found on the same chromosome**
- **Also ... see Broad’s Molecular Signatures Database (MSigDB)**
- [any grouping that is biologically sensible]

Statistics to test for enrichment

```
Genome = 20,000 genes

schmooase activity = 1000 genes
100 genes with 10% overlap
1000 genes
1000 genes
1000 genes

stromphase activity = 20 genes
1000 genes
1000 genes
1000 genes

Intersection = 10 genes
Intersection = 2 genes
```

- Our list = 100 genes
- Our list = 1000 genes
- Intersection = 10 genes
- Intersection = 2 genes
Tests for enrichment

- Fisher’s exact
- Hypergeometric
- Binomial
- Chi-squared
- Z
- Kolmogorov-Smirnov
- Permutation

Statistics to test for enrichment

- What is the chance of observing enrichment at least this extreme due to chance?
- Different tests produce very different ranges of p-values
- All look for over-enrichment; some look for under-enrichment
- Recommendation: Use p-values as a tool to rank genes but don’t take them literally
- Most methods correct for multiple testing (e.g., with FDR), which is necessary

Other statistical issues

- Goal: Identifying theme(s) of maximal biological significance
  – but this is not perfectly correlated with statistical significance
- What is your background gene set?
  – All genes that could appear in your list
- What about sparse annotation groups?
- Some annotation terms may be subsets of other terms.

Practicalities

- Choose a tool that
  – Includes your species
  – Includes your gene / probe identifiers
  – Has up-to-date annotation
  – Lets you define your background (if possible)
- Get recommendations from the usual sources.
- Try at least a few tools.
Presenting results

- Generally ignore enriched categories which
  - Contain very few genes
  - Show high overlap with other categories
- When in doubt, select more general category.
- Simplify complex results.
- Graphical or text summary?
- Plan to share your gene lists when you publish.

Enrichment tools


Some recommended tools

- DAVID
- GSEA
- BIOBASE (Whitehead has license)
- BiNGO (uses Cytoscape)
- GoMiner: http://discover.nci.nih.gov/gominer
- GOstat: http://gostat.wehi.edu.au

DAVID

- Database for Annotation, Visualization and Integrated Discovery (NIAID)
- List-based
  - http://david.abcc.ncifcrf.gov/
- Lots of identifiers; lots of species
- Allows background definition
- Statistic is a modified Fisher exact test
Welcome to the new, temporary home of DAVID2008. We have extended the retirement of this version until 30/12/2016. Please complete any analysis using this version by this date as it will no longer be available. Thanks for using and supporting DAVID.

Functional Annotation Chart

Input:
- pre-ranked gene list

GSEA
- Gene Set Enrichment Analysis
- Rank-based
- http://www.broadinstitute.org/gsea/
- As a Java Web Start or desktop application
- Linked to MSigDB (annotated gene lists)
- Also permits custom annotation

BiNGO
- BiNGO: A Biological Network Gene Ontology tool
- Works with Cytoscape network visualization tool
- Also permits custom annotation

Shows relationship between annotation categories
BIOBASE

- BIOBASE Knowledge Library
- Use Internet Explorer
- Go to “Gene Set Analysis”

References

- Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. (PMID: 19033363)  \textit{Review}

- Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. (PMID: 19131956)  \textit{DAVID}

- Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. (PMID: 16199517)  \textit{GSEA}