Pathway Analysis

Min Kim
Bioinformatics Core Facility
2/28/2018
Outline

1. Background
2. Databases:
   - KEGG, Reactome, Biocarta, Gene Ontology, MSigDB, MetaCyc, SMPDB, IPA.
3. Statistical Methods: Overlap (ORA) vs GSEA
4. Web-based:
   - DAVID, Panther, Pathway Commons, ConsensusPathDB, MetaboAnalyst.
1. R Package – clusterProfiler.
• Genomic and proteomic studies allow us to identify the genes and proteins of interest.
• Understanding the functional role of these genes/proteins can be challenging.
• Existing knowledge base of annotated biological pathways and protein-protein interactions allow us to gain further insight.
• “Pathway Analysis” – is a broad term used to describe the methods used to query these knowledge-base resources using our ‘gene sets’ of interest.
**Input**: gene/protein (single, list, all).

**Database**: pathway database (predicted, evidence-based).

**Statistical methods**: ORA, FCS, PT.

**Result**: Significant pathways.
High-throughput profiling studies:

- Proteomics Analysis
- Transcriptomic analysis.

- List of genes, proteins, metabolites, or mutation.
Note on ID Conversion

Many ID formats exist for naming genes:

Ex: Entrez ID, Ensemble, GeneBank, HUGO Gene Symbol, ProbeID, PubChem, Uniprot, GI, RefSeq, etc.

www.biomart.org

www.david.ncifcrf.gov

https://biodbnet-abcc.ncifcrf.gov/db/db2db.php
Pathway Databases

• Curated collection of biological molecules (nucleic acids, proteins, metabolites, small molecules, drugs) grouped by biological process (functional, metabolic, signaling, regulatory, cellular component, disease).

• Popular - KEGG, Reactome, Gene Ontology, MSigDB

• Based on what is currently known.*
KEGG: Kyoto Encyclopedia of Genes and Genomes

http://www.kegg.jp

• Developed by Dr. Minoru Kanehisa at Kyoto University.
• Last updated Feb 1, 2018
• Besides pathways also contains annotation: disease related networks (KEGG Network), small molecules (KEGG Compound), glycans (KEGG Glycan), enzymes (KEGG Reaction).
• Also KEGG Cancer, KEGG Pathogen, KEGG Virus, KEGG Plant.
• Free Academic use.
• License required if you wish to use KEGG as part of service.
KEGG Mapping
Reactome

https://reactome.org

• Open-source, open access, manually curated and peer-reviewed pathway database.

• Developed by Lincoln Stein of OICR, Peter D’Eustachio of NYULMC, Henning Hermjakob of EMBL-EBI, and Guanming Wu of OHSU.

• Last updated December 18, 2017.
<table>
<thead>
<tr>
<th>SPECIES</th>
<th>PROTEINS</th>
<th>COMPLEXES</th>
<th>REACTIONS</th>
<th>PATHWAYS</th>
</tr>
</thead>
<tbody>
<tr>
<td>D. discoideum</td>
<td>2081</td>
<td>1957</td>
<td>1669</td>
<td>820</td>
</tr>
<tr>
<td>P. falciparum</td>
<td>790</td>
<td>744</td>
<td>590</td>
<td>450</td>
</tr>
<tr>
<td>S. pombe</td>
<td>1456</td>
<td>1440</td>
<td>1186</td>
<td>671</td>
</tr>
<tr>
<td>S. cerevisiae</td>
<td>1623</td>
<td>2098</td>
<td>1869</td>
<td>836</td>
</tr>
<tr>
<td>C. elegans</td>
<td>4992</td>
<td>3341</td>
<td>2820</td>
<td>1136</td>
</tr>
<tr>
<td>S. scrofa</td>
<td>12363</td>
<td>8293</td>
<td>7201</td>
<td>1584</td>
</tr>
<tr>
<td>B. taurus</td>
<td>9751</td>
<td>8336</td>
<td>7230</td>
<td>1585</td>
</tr>
<tr>
<td>C. familiaris</td>
<td>10962</td>
<td>8079</td>
<td>6912</td>
<td>1570</td>
</tr>
<tr>
<td>M. musculus</td>
<td>12316</td>
<td>9324</td>
<td>8109</td>
<td>1605</td>
</tr>
<tr>
<td>R. norvegicus</td>
<td>11571</td>
<td>8534</td>
<td>7398</td>
<td>1581</td>
</tr>
<tr>
<td>*H. sapiens</td>
<td>10732</td>
<td>11366</td>
<td>11426</td>
<td>2179</td>
</tr>
<tr>
<td>G. gallus</td>
<td>9036</td>
<td>7286</td>
<td>6215</td>
<td>1598</td>
</tr>
<tr>
<td>T. guttata</td>
<td>7299</td>
<td>6211</td>
<td>5187</td>
<td>1480</td>
</tr>
<tr>
<td>X. tropicalis</td>
<td>9149</td>
<td>7323</td>
<td>6250</td>
<td>1533</td>
</tr>
<tr>
<td>D. rerio</td>
<td>15582</td>
<td>7295</td>
<td>6196</td>
<td>1536</td>
</tr>
<tr>
<td>D. melanogaster</td>
<td>9925</td>
<td>4593</td>
<td>3899</td>
<td>1373</td>
</tr>
<tr>
<td>A. thaliana</td>
<td>6621</td>
<td>1935</td>
<td>1679</td>
<td>785</td>
</tr>
<tr>
<td>O. sativa</td>
<td>13455</td>
<td>1831</td>
<td>1587</td>
<td>777</td>
</tr>
<tr>
<td>M. tuberculosis</td>
<td>13</td>
<td>50</td>
<td>40</td>
<td>12</td>
</tr>
</tbody>
</table>
Reactome Demo
Gene Ontology

www.geneontology.org

• Formal ontologies to classify gene function in three aspects:
  1. Molecular function.
  2. Cellular component.
  3. Biological process.

• Funded by National Human Research Institute (NHGRI)
• Last updated September 26, 2017.
• Contains both experimental evidence based and computationally predicted annotations.
Levels
MSigDB: Molecular Signature Database

http://software.broadinstitute.org/gsea/

- Collection of annotated gene sets for use with GSEA software.
- Developed at Board Institute.
- Last updated October 2017.
Gene sets

H hallmark gene sets are coherently expressed signatures derived by aggregating many MSigDB gene sets to represent well-defined biological states or processes.

C1 positional gene sets for each human chromosome and cytogenetic band.

C2 curated gene sets from online pathway databases, publications in PubMed, and knowledge of domain experts.

C3 motif gene sets based on conserved cis-regulatory motifs from a comparative analysis of the human, mouse, rat, and dog genomes.

C4 computational gene sets defined by mining large collections of cancer-oriented microarray data.

C5 GO gene sets consist of genes annotated by the same GO terms.

C6 oncogenic gene sets defined directly from microarray gene expression data from cancer gene perturbations.

C7 immunologic gene sets defined directly from microarray gene expression data from immunologic studies.
IPA : Ingenuity Knowledge Base

- Proprietary database. Owned by QIAGEN.
- Requires license.
- Contains pathways not present in other public databases.
Statistical Methods

• Over-representation analysis (ORA).
  – Most commonly used.
  – Also known as functional enrichment analysis.
  – Statistical evaluation of fraction of genes in a pathway found among the set of input genes.
  – Hypergeometric test, chi-square test, binomial distribution...
Limitation of ORA

• Requires a subset of ‘most significant genes’ based on some arbitrary cutoff (fold-change ≥ 2 and/or p-value ≤ 0.05).
• Does not consider significance of change such as read counts, probe intensities, fold change. Highly expressed genes are considered equally important as low expressed genes.
• Assumes genes are independent.
• Assumes pathways are independent.

(Networks analysis)
Functional Class Scoring

• Hypothesis that although large changes in individual genes can have significant effects on pathways, weaker but coordinated changes in sets of functionally related genes can also have significant changes.

• GSEA – Gene set enrichment analysis.
GSEA – Gene Set Enrichment Analysis.

Step 1: Rank genes based on their correlation between their expression and phenotype label.

Step 2: Going down the ranked list, keep a running-sum score when gene encountered in pathway set.

Enrichment Score (ES).

Step 3: Calculate statistical significance of ES based on permutation test (phenotype or gene set).
GSEA Results

GSEA Report for Dataset brca_hd_tep_ranks

Enrichment in phenotype: na
- 2464 / 3795 gene sets are upregulated in phenotype na_pos
- 1250 gene sets are significant at FDR < 25%
- 578 gene sets are significantly enriched at nominal pvalue < 1%
- 926 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in excel format (tab delimited text)
- Guide to interpret results

Enrichment in phenotype: na
- 1331 / 3795 gene sets are upregulated in phenotype na_neg
- 679 gene sets are significantly enriched at FDR < 25%
- 394 gene sets are significantly enriched at nominal pvalue < 1%
- 550 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in excel format (tab delimited text)
- Guide to interpret results

Dataset details
- The dataset has 8191 features (genes)
- No probe set => gene symbol collapsing was requested, so all 8191 features were used
Running GSEA

<table>
<thead>
<tr>
<th>Column 1: Row identifiers. Typically probe set ids or clone ids. These must be UNIQUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Column 2: Row descriptions. Ignored by the program – can be dummy values (e.g. “na”)</td>
</tr>
<tr>
<td>Each column contains expression values from 1 sample. Missing values are allowed (leave empty).</td>
</tr>
</tbody>
</table>

If editing, in Excel, make sure to save your data as “tab delimited text”
GSEA Demo
Limitations

GSEA (like ORA) assumes pathway are independent.
GSEA assumes genes are independent.

Alternative : Network Analysis

Considers the internal connection between genes through graph representation allowing for more accurate representation of the interactions between genes and also between pathways.

– https://genemania.org/
– Cytoscape
Web-based Pathway Analysis

• DAVID: https://david.ncifcrf.gov/home.jsp
• Panther: http://www.pantherdb.org/
• ConsensusPathDB: http://cpdb.molgen.mpg.de/
• Reactome: https://reactome.org/PathwayBrowser/#TOOL=AT
• Metaboanalyst: http://www.metaboanalyst.ca/
clusterProfiler (R package)

- It is sometimes necessary to compare pathways across different test conditions.
- clusterProfiler allows for the comparison of pathway analysis results across multiple gene sets.
Four nanocourse workshops taught by NIH-NCBI trainers equipping you to utilize NCBI resources in your research

NCBI Resources For:
*Microbiology Research, Wed. 9AM - Noon
Resources for using bacterial sequence data, including human pathogens and microbiomes
*Genetic Disease Discovery & Clinical Support, 1-4PM
Clinically-focused resources, including genetic variation and testing in patient care
*Genetic Variation & Gene Expression, Thurs. 9AM - Noon
Utilizing genomic sequence, annotation & expression databases and command-line tools
*NIH Biosketches & Public Access Policy Compliance, 1-4PM
An array of My NCBI tools to assist researchers and grants staff with preparation of federal grant proposals and progress reports

**All sessions are stand-alone and have open attendance. Academic credit is available for participants who attend all sessions.
Location: E4.350
Time: 10am – 11am daily.
Contact: BICF@UTSouthwestern.edu
Phone: 214-645-1707
References

