AFRICAN CENTERS OF EXCELLENCE IN BIOINFORMATICS

KAMPALA, UGANDA

Enrichment, Pathways and Network Analysis

Mariam Quiñones, PhD

1

Today's Instructor

Dr. Mariam Quiñones, Ph.D. in Molecular Microbiology

Ongoing Computational Biology projects:
- 16S microbiome
- Ion Torrent Zika virus genome assembly

- Bioinformatics and Computational Biosciences Branch (BCBB), NIAID
- National Institutes of Health, Bethesda, MD USA.
- Contact our team via email:
  - Email: ace@icermali.org
  - Listserv: ACE-GLOBAL-ANNOUNCE@list.nih.gov
  - Instructor: mariam.quinones@nih.gov

2

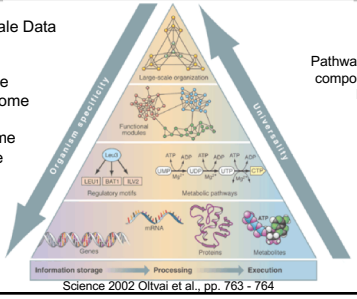
Overview

- Introduction to pathways, enrichment methods and networks
  - Pathways, ORA, GSEA
  - Theory on Gene Ontology
  - Networks analysis
- Useful databases/tools for pathways and networks analyses
  - Overview of web network/pathway tools such as KEGG, Reactome, String, DAVID
  - Explore String networks
- Practical
  - Run a Gene Set Enrichment using GSEA
  - Network visualization using Cytoscape

Biologists have abandoned the reductionist approach to adopt a systems biology approach to handle genome scale data

Genome Scale Data

- Genome
- Epigenome
- Transcriptome
- Proteome
- Metabolome
- Signalome



Pathways analysis is one component of a systems biology approach.

Science 2002 Oltvai et al., pp. 763 - 764

Common methods for knowledge-base gene and pathways analyses

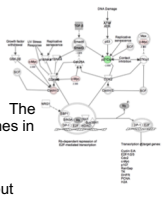
Enrichment: A gene set is tested for a significant association with a trait compared to an association of any other gene set (often the genes on a chip)

1- Over-representation analysis (ORA)
2- Functional Class Scoring (FCS)
3- Pathway Topology (PT)

Khatri P, PLOS 2012 http://www.ploscompbiol.org/article/info%3Adoi/%2F10.1371%2Fjournal.pcbi.1002375
Holmans P, Advances in Genetics 2010 http://www.sciencedirect.com/science/article/pii/B9780123808622000072#
Zhang Q, BMC Bioinformatics 2016 https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1333-x

1- Over-representation analysis (ORA)

- Before starting a pathway analysis, the researcher typically chooses genes that are differentially expressed in a given condition; these have an expression value and p-value
- For each pathway, input genes in that pathway are counted. The process is repeated for a background gene set (e.g. the genes in the microarray chip)
- Pathways are tested for over-representation in the list of input genes. If the proportion of genes in the pathway appearing on the list is significantly higher than the corresponding proportion of genes not in the pathway, the pathway is said to be overrepresented.



1- Over-representation analysis (ORA) - *cont*

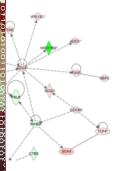
- Tools using ORA: DAVID, MetaCore, GeneGo, IPA
- Statistical tests: Fisher's exact test, the hypergeometric distribution, or a chi-square analysis
- Limitations: Gene IDs are the only requirement, gene expression levels are not used to provide weight. An arbitrary threshold cutoff is required to create the input gene list.

2- Functional Class Scoring (FCS) or Gene-set enrichment

- It expects that small coordinated changes in related genes within the same pathway can cause significant effects
- The FCS uses all genes. It computes differential expression of genes, aggregates all gene level statistics into pathway level statistic, and then determines the statistical significance of the pathway-level statistic.
- Limitations: Some pathways cross and overlap, therefore a pathway may appear affected due to overlapping genes.

Tools: GSEA, sigPathway (BioConductor)

3- Pathway Topology or Network based



- This method uses FCS but also use pathway topology to compute gene-level statistics.
- It takes into account the number of reactions needed to connect two genes in a pathway.
- It could incorporate biological factors such as gene expression, types of interactions and positions of genes in a pathway.

Why pathways analysis?

Pathways analyses
It is the analyses of sets of genes that are related to each other biologically. It helps interpret the data in the context of diseases, biological processes, pathways and networks.

- High-throughput experimental technologies often identify hundreds of genes but do not necessarily produce biological findings.
- Genes do not work alone but in a large network of interactions
- An associated pathway is likely to implicate function better than a hit in a single gene

General limitations of Pathways Analysis methods

- Incomplete annotation of genes and isoforms
- Missing cell type specific information
- Inability to understand pathway dynamics and how one pathway affects another one

Pathway Databases and tools

Some commercial tools:

- Manually curated database:
  - Ingenuity Pathways http://ingenuity.com/
  - GeneGo Metacore http://www.genego.com/
- Text processing:
  - Pathway Studio https://www.elsevier.com/solutions/pathway-studio-biological-research

3

Web and open source pathways databases and tools

- Comprehensive Databases for Pathways:
  - KEGG http://www.genome.jp/kegg/pathway.html
  - WikiPathways http://www.wikipathways.org
  - Reactome http://www.reactome.org/
  - Pathway Commons http://www.pathwaycommons.org/pc2/
  - MSigDB- http://www.broadinstitute.org/gsea/msigdb/index.jsp
- Selected Tools for enrichment and pathways analysis:
  - DAVID Bioinformatics http://david.abcc.ncifcrf.gov:8080/
  - g:Profiler http://biit.cs.ut.ee/gprofiler/
  - Gene Ontology enrichment http://geneontology.org/page/go-enrichment-analysis
  - ConsensusPathDB http://cpdb.molgen.mpg.de
  - Enrichr http://amp.pharm.mssm.edu/Enrichr/

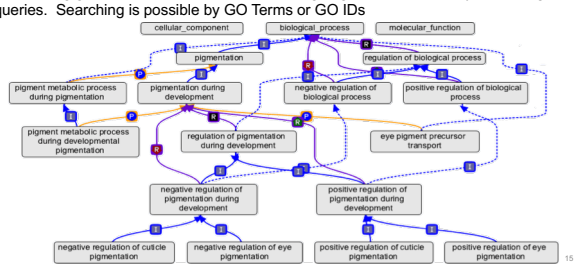
4

Gene Ontology (geneontology.org)

- Bioinformatics initiative with the aim of standardizing the representation of gene and gene product attributes across species and databases.
- GO describes how gene products behave in a cellular context. It consists of three controlled vocabularies:
  - Molecular function (what does the gene product do?)
  - Biological process (why does it perform the activity?)
  - Cellular component (where does it act?)
- It is very general and is available for a large number of organisms (>40 species).

•In the Gene Ontology, each term is linked by the relationships “is a”, “part of” or “regulates”.

•Annotating genes with GO terms allows sharing of gene information by facilitating queries. Searching is possible by GO Terms or GO IDs



15

6

Gene Ontology

1- It infers classification from experimental data and electronic annotation
2- It labels the terms with an evidence code to record the type of evidence used to make the annotation

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

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

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

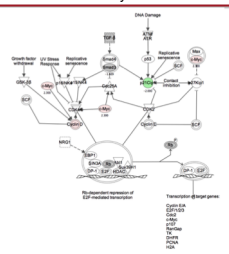
Automatically-assigned Evidence Codes
- IEA: Inferred from Electronic Annotation

Obsolete Evidence Codes
- NR: Not Recorded

7

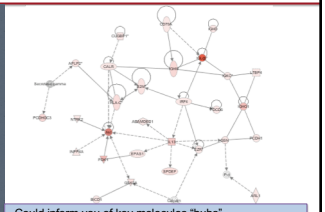
In addition to Gene Ontology and Pathways, let's explore Networks

Pathway



- It has directionality
- Curated in the literature

Network



- Could inform you of key molecules "hubs"
- Could provide information on possible link between two affected pathways not present in your dataset

- Triangular relationships
- De Novo and unbiased

Networks

Example of network types:
- Social (Facebook)
- Professional (LinkedIn)
- Biological (PPI)

Purpose of networks
- Study interactions
- Infer function based on association

Edge

Node

Software to create and analyze networks

Cytoscape — http://cytoscape.org
Best at biological visualization

Gephi makes graphs handy — https://gephi.org
Best at general visualization

Tulip — http://tulip.labri.fr/
Easy to use

20

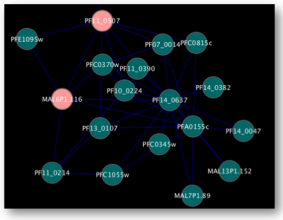
Networks use data from Protein-protein or Protein-DNA interaction databases

1. BioGRID – (The Biological General Repository for Interaction Datasets) – a curated database http://www.thebiogrid.org/
2. IntAct – EMBL molecular interaction database http://www.ebi.ac.uk/intact/
3. Pathways Commons http://www.pathwaycommons.org/pc/ - Integrates data from BioGRID, HumanCyc, IntAct, Reactome, MINT, NCI, SBCNY, HPRD and CancerCell Map. Important: Of these sources, IntAct and MINT contain data for some parasites.
4. STRING http://string.embl.de/ and GeneMania http://genemania.org
   1. Combines known and predicted protein-protein interactions
   2. Predictions are derived from Genomic context, high throughput experiments, co-expression and previous knowledge reported in PubMed.

21

Cytoscape www.cytoscape.org

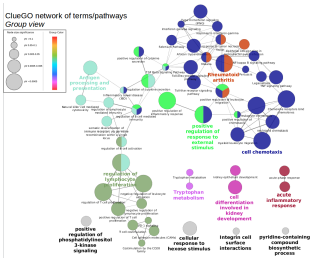
- Open source visualization tool for networks
- Framework; functionality is expanded with plugins
- It allows users to:
  - Modify networks to ease of visualization
  - Load custom datafiles or files from databases such as Pathways Commons
  - Explore large networks



32

Cytoscape has a library of plugins to perform enrichment and network analysis

Example: ClueGo plugin



33

Other plugins to analyze networks include MCODE

Identifies densely connected areas of protein interaction networks

MCODE (Molecular Complex Detection) v1.2 (Jan 2007) A Cytoscape PlugIn Version 1.2 by Vuk Pavlovic (Bader Lab, University of Toronto)

Bader GD, Hogue CW An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics. 2003 Jan 13;4(1):2

Next: Practical Session

- Exercise on enrichment, exploring interaction databases, visualizing and analyzing networks