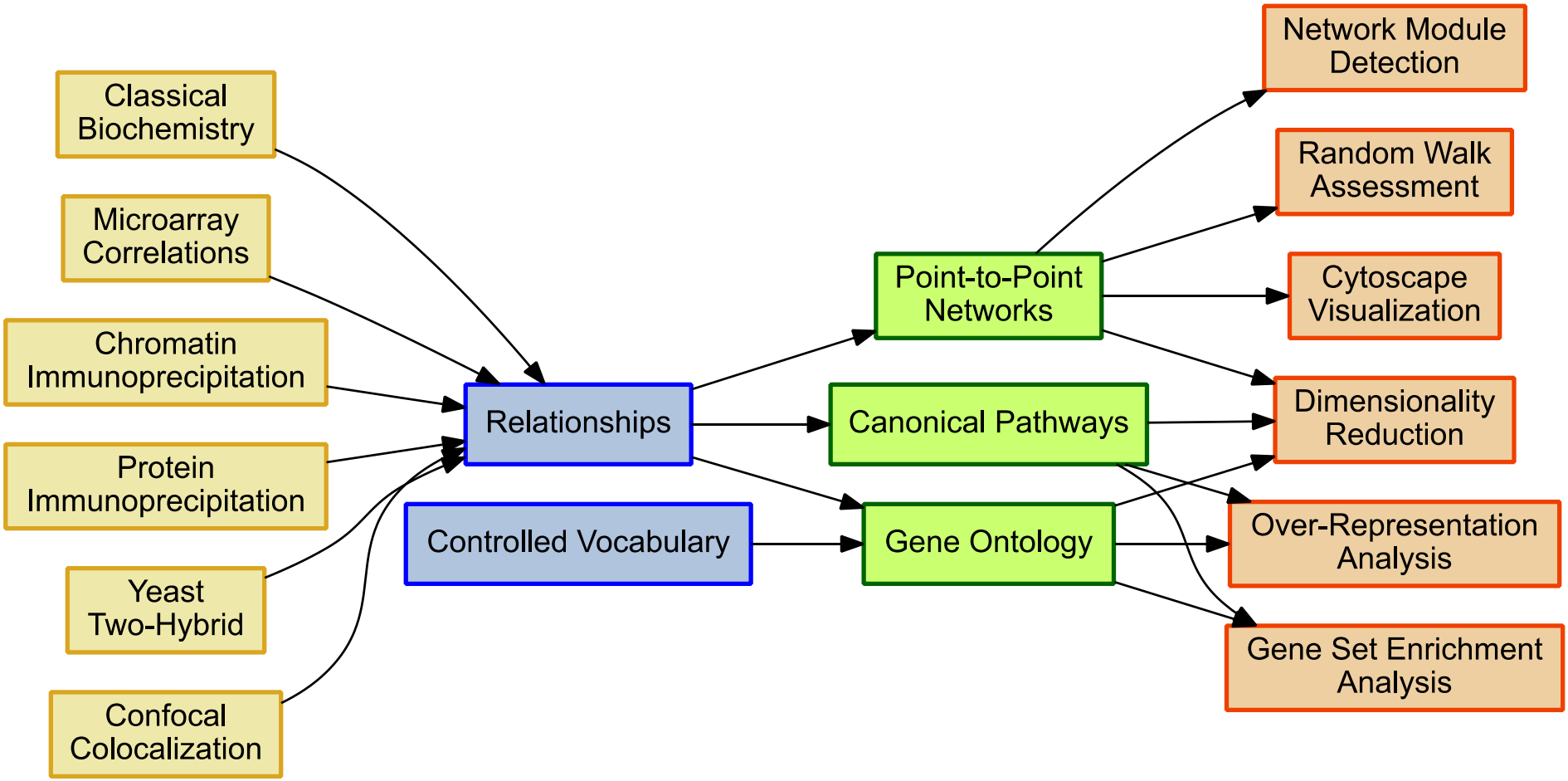
Biological Pathways and Networks

DAVID L. TABB, PH.D.

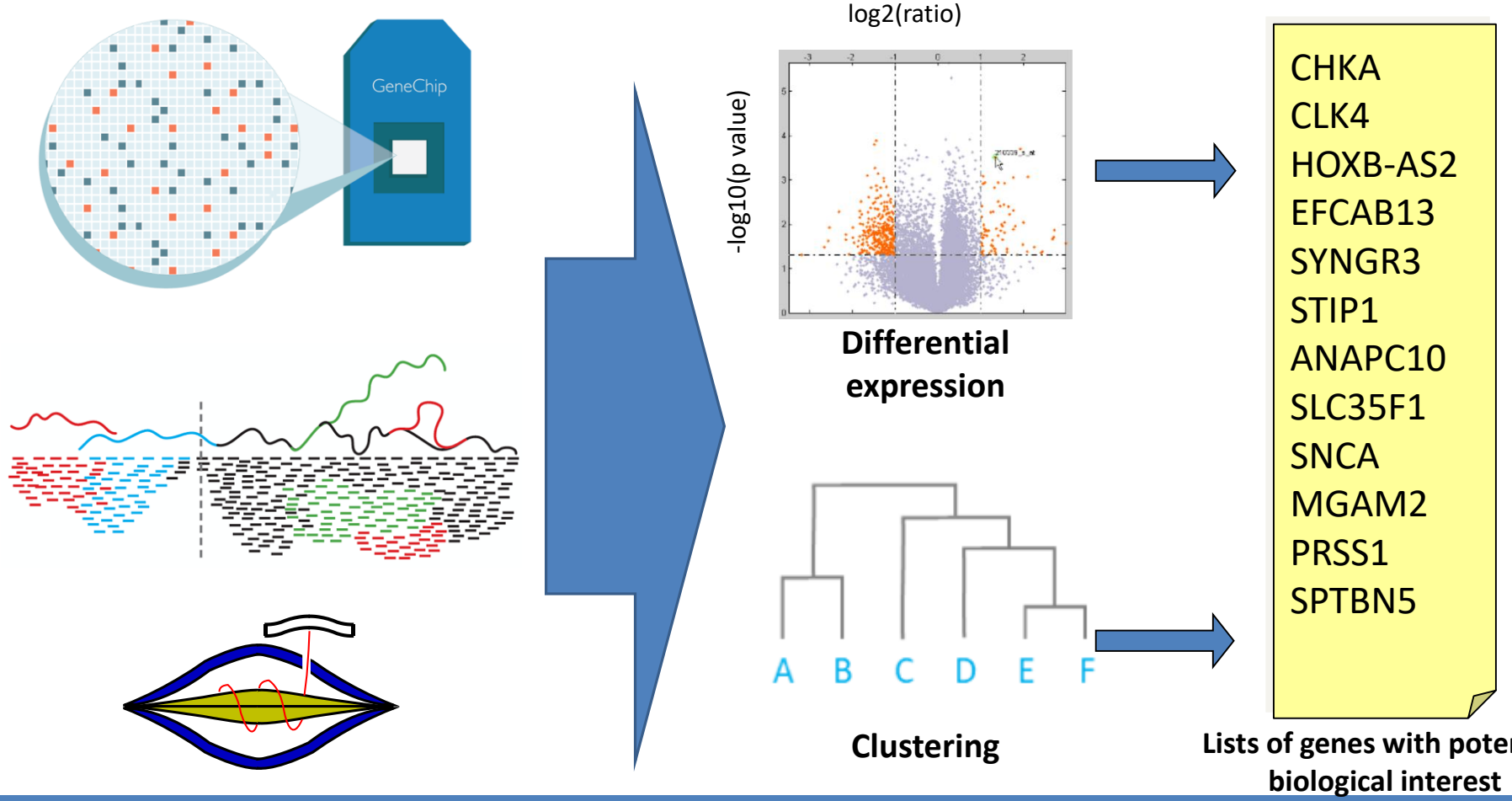
With many valuable slide contributions from
Bing Zhang, Baylor College of Medicine

Overview

- Organizing genes by “gene sets”
  - Pathways
  - Gene Ontology
  - Network modules
- Enrichment analysis methods
  - Over-representation analysis: WebGestalt
  - Gene Set enrichment analysis: GSEA


Classical Biochemistry
Microarray Correlations
Chromatin Immunoprecipitation
Protein Immunoprecipitation
Yeast Two-Hybrid
Confocal Colocalization
Relationships
Controlled Vocabulary
Point-to-Point Networks
Canonical Pathways
Gene Ontology
Network Module Detection
Random Walk Assessment
Cytoscape Visualization
Dimensionality Reduction
Over-Representation Analysis
Gene Set Enrichment Analysis

Omics studies generate lists of interesting genes



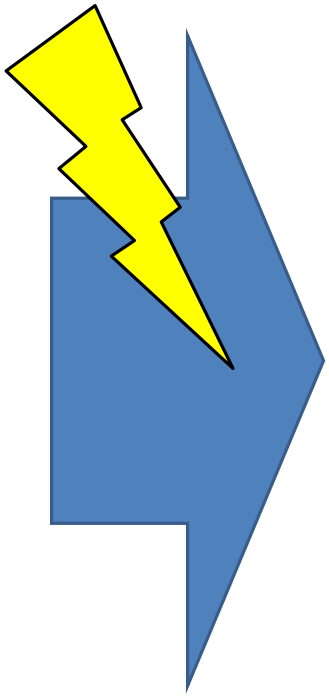
https://ischemicstrokelab.com/

Iyer and Chinnaiyan, *Nat. Biotech*. (2011) 29: 599-600

UNIVERSITEIT STELLENBOSCH UNIVERSITY

Reorganizing to pathways changes our perspective

- CASP
  - MAP Kinase
  - Apoptosis
- Ras
  - MAP Kinase
  - cAMP Signaling
- MEKK1
  - MAP Kinase
  - Apoptosis
- MLCP
  - cAMP Signaling



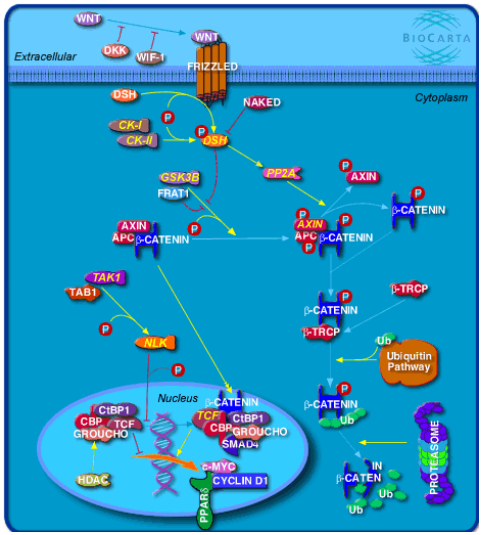
- MAP Kinase
  - CASP
  - Ras
  - MEKK1
- Apoptosis
  - CASP
  - MEKK1
- cAMP Signaling
  - Ras
  - MLCP

UNIVERSITEIT STELLENBOSCH UNIVERSITY

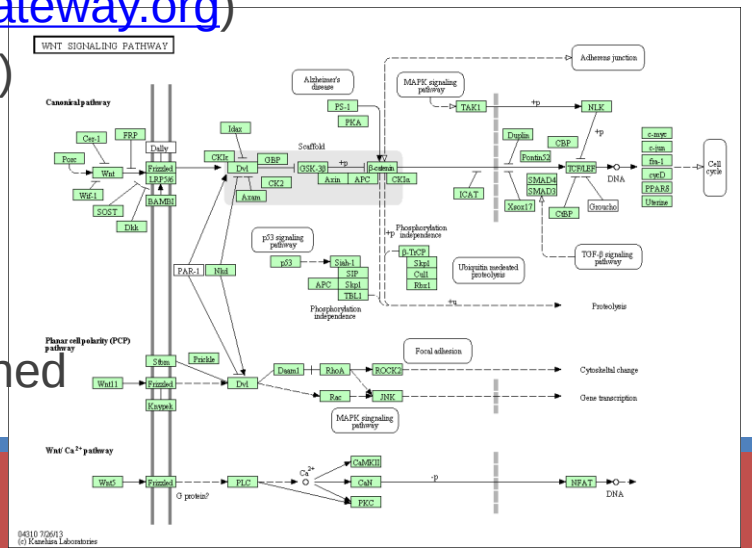
Advantages of pathway analysis

- Better interpretation
  - From interesting genes to interesting biological themes
- Improved robustness
  - Guards against noise in the data
- Improved sensitivity
  - Detecting minor but concordant changes in a pathway

Pathway databases



- Databases
  - BioCarta (http://www.biocarta.com/genes/index.asp)
  - KEGG (http://www.genome.jp/kegg/pathway.html)
  - MetaCyc (http://metacyc.org)
  - Pathway commons (http://www.pathwaycommons.org)
  - Reactome (http://www.reactome.org)
  - STKE (http://stke.sciencemag.org/cm)
  - Signaling Gateway (http://www.signaling-gateway.org)
  - Wikipathways (http://www.wikipathways.org)
- Limitation
  - Limited coverage
  - Inconsistency among different databases
  - Relationship between pathways is not defined



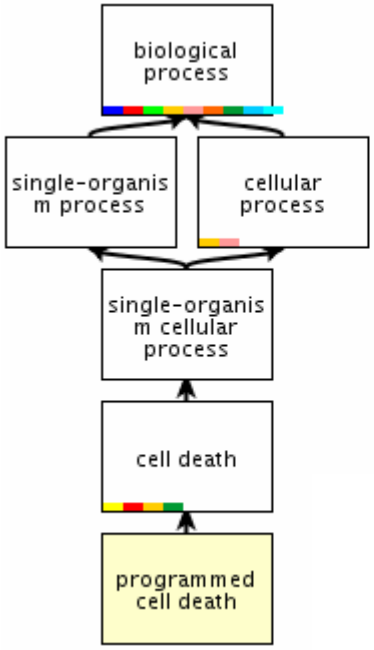
Gene Ontology

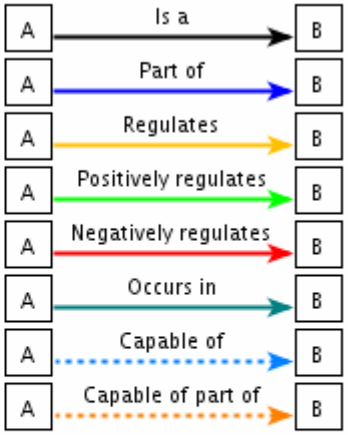
- Structured, precisely defined, *controlled vocabulary* for describing the roles of genes and gene products
- Three organizing principles: molecular function, biological process, and cellular component
  - Dopamine receptor D2, the product of human gene DRD2
    - *molecular function*: dopamine receptor activity
    - *biological process*: synaptic transmission
    - *cellular component*: plasma membrane
- Terms in GO are linked by several types of *relationships:*
  "IS A" *"PART OF"* *"HAS PART"* *"REGULATES"*

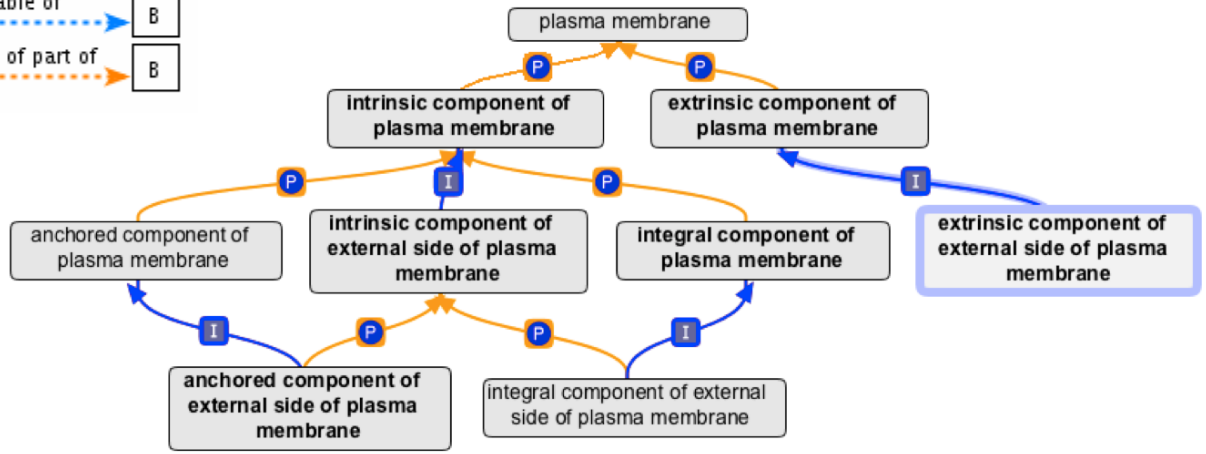
Ontology: a theory about the nature of being or the kinds of things that have existence

Biological ontologies generally feature controlled vocabularies and defined relationships

Relationship examples







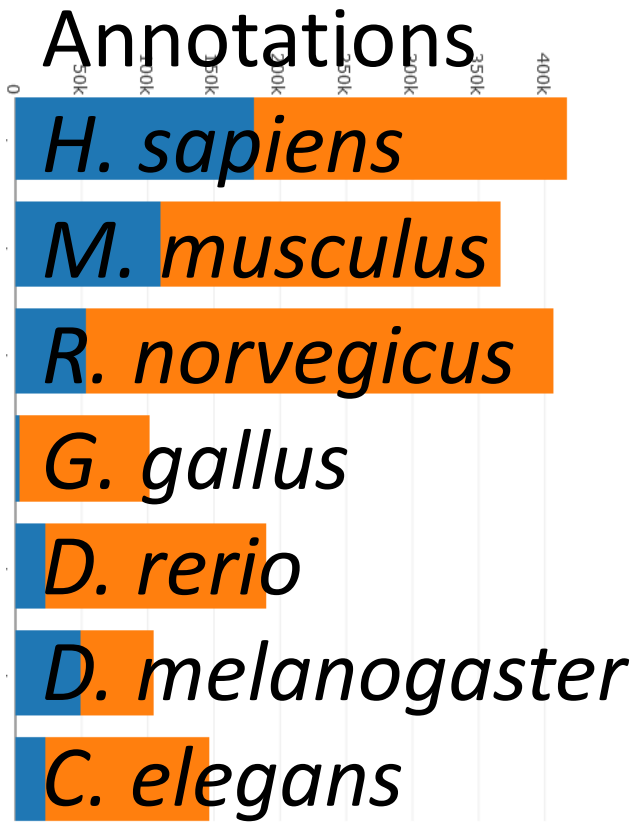
http://www.geneontology.org/

Annotation using GO terms

- Two types of GO annotations
  - Electronic annotation
  - Manual annotation
- All annotations must:
  - be attributed to a source
  - indicate what evidence was found to support the GO term-gene/protein association
- Types of evidence codes
  - Experimental codes - IDA, IMP, IGI, IPI, IEP
  - Computational codes - ISS, IEA, RCA, IGC
  - Author statement - TAS, NAS
  - Other codes - IC, ND

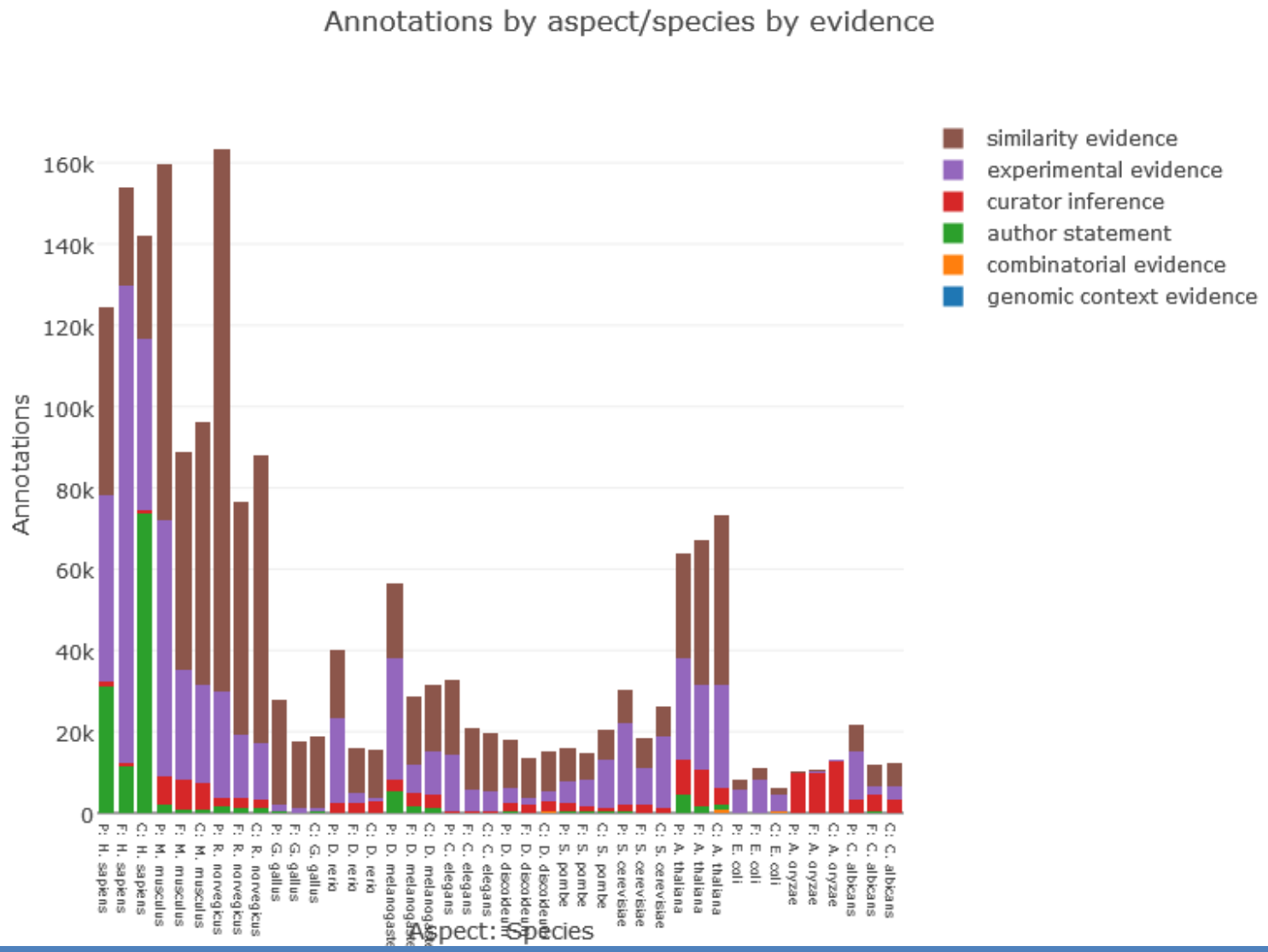
Handy interfaces to GO

- AmiGO 2 (GO service)
- Bioconductor (CRAN)
- QuickGO (EBI service)



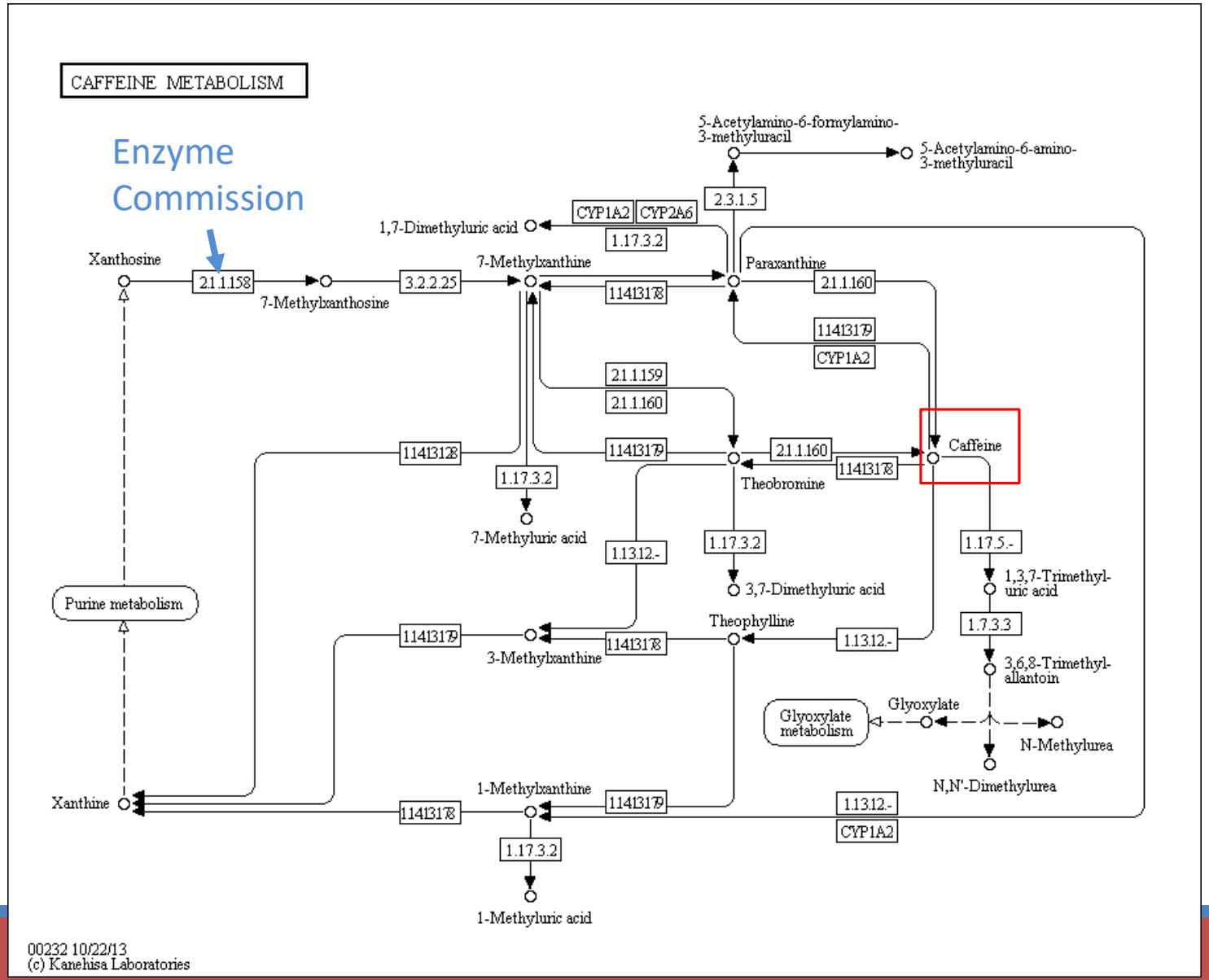
JA Blake. *PLoS Computat. Bio*. (2013) 9: e1003343

How do we add to GO?



http://amigo.geneontology.org/amigo/base_statistics

KEGG molecular networks

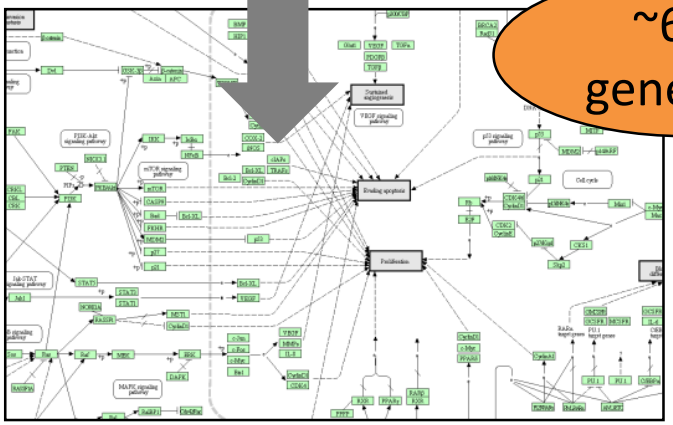


WebGestalt: enrichment by pathway

92546_r_at
92545_f_at
96055_at
102105_f_at
102700_at
……

~200 ID types

~60K gene sets



Statistical analysis

8 organisms
Human, Mouse, Rat, Dog, Fruitfly, Worm, Zebrafish, Yeast

Microarray Probe IDs
- Affymetrix
- Agilent
- Codelink
- Illumina

Genetic Variation IDs
- dbSNP

Gene IDs
- Gene Symbol
- GenBank
- Ensembl Gene
- RefSeq Gene
- UniGene
- Entrez Gene
- SGD
- MGI
- Flybase ID
- Wormbase ID
- ZFIN

Protein IDs
- UniProt
- IPI
- RefSeq Peptide
- Ensembl Peptide

196 ID types with mapping to Entrez Gene ID

WebGestalt

59,278 functional categories with genes identified by Entrez Gene IDs

Gene Ontology
- Biological Process
- Molecular Function
- Cellular Component

Pathway
- KEGG
- Pathway Commons
- WikiPathways

Network module
- Transcription factor targets
- microRNA targets
- Protein interaction modules

Disease and Drug
- Disease association genes
- Drug association genes

Chromosomal location
- Cytogenetic bands

Zhang et.al. *Nucleic Acids Res*. 33:W741, 2005
Wang et al. *Nucleic Acids Res*. 41:W77, 2013

Over-representation analysis

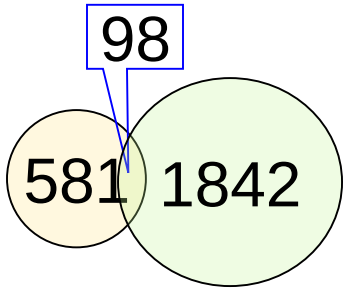
Hoxa5
Hoxa11
Ltbp3
Sox4
Foxc1
Edn1
Ror2
Gnag
Smad3
Wdr5
Trp63
Sox9
Pax1
Acd
Rai1
Pitx1
……

Development (1842 genes)

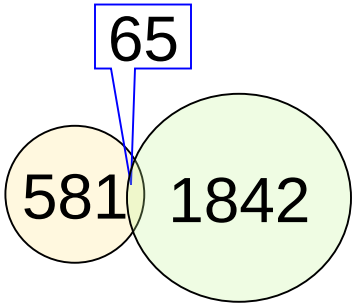
compare

Sash1
Cd24a
Agt
Psrc1
Ctla2b
Angptl4
Depdc7
Sorbs1
Macrod1
Enpp2
Tmem176a
……

Differentially expressed genes (581 genes)



Observe



Expect

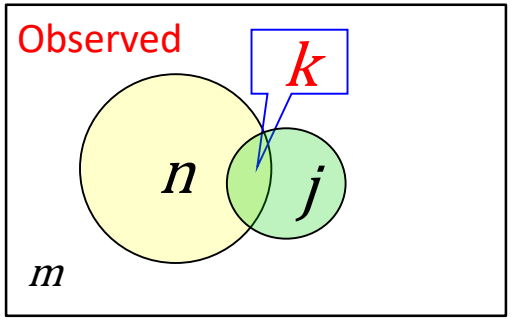
How often would one see even more genes than this at random?

Contingency Table

	Significant genes	Non-significant genes	Total
genes in the group	*k*	*j-k*	*j*
Other genes	*n-k*	*m-n-j+k*	*m-j*
Total	*n*	*m-n*	*m*

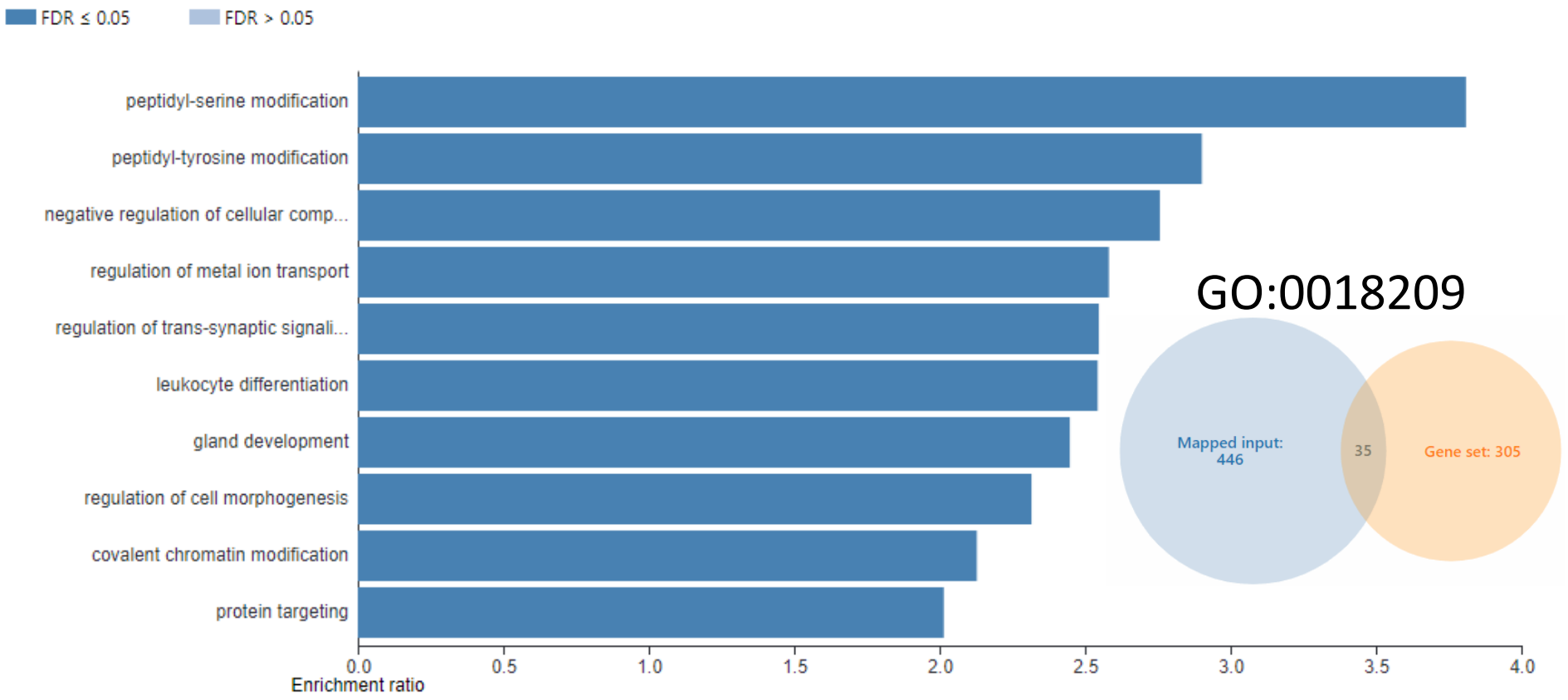
Hypergeometric test: given a total of *m* genes where *j* genes are in the functional group, if we pick *n* genes randomly, what is the probability of having *k* or more genes from the group?

$$p = \sum_{i=k}^{\min(n,j)} \frac{\binom{m-j}{n-i}\binom{j}{i}}{\binom{m}{n}}$$

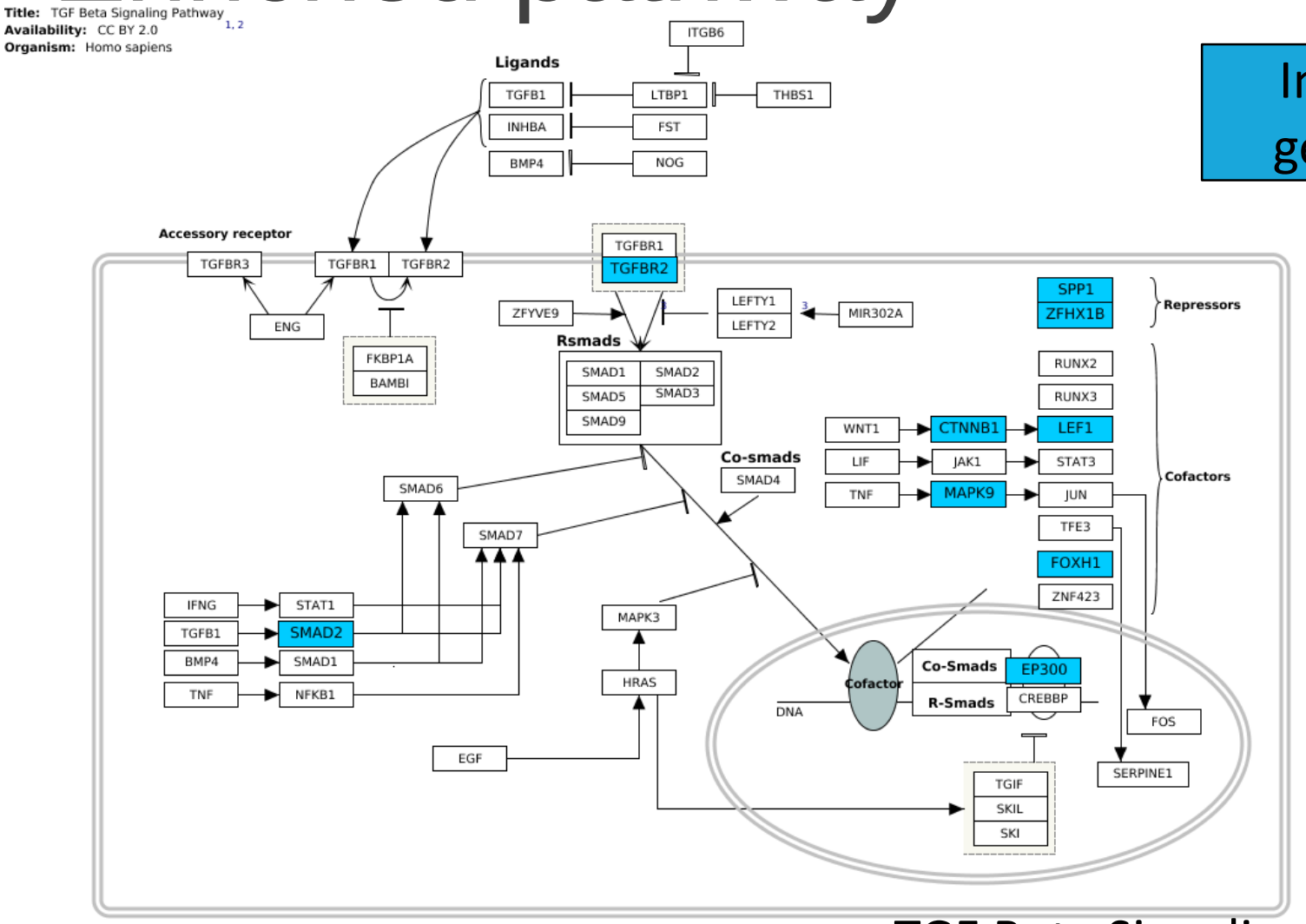


Zhang et.al. *Nucleic Acids Res*. 33:W741, 2005

Enriched GO terms



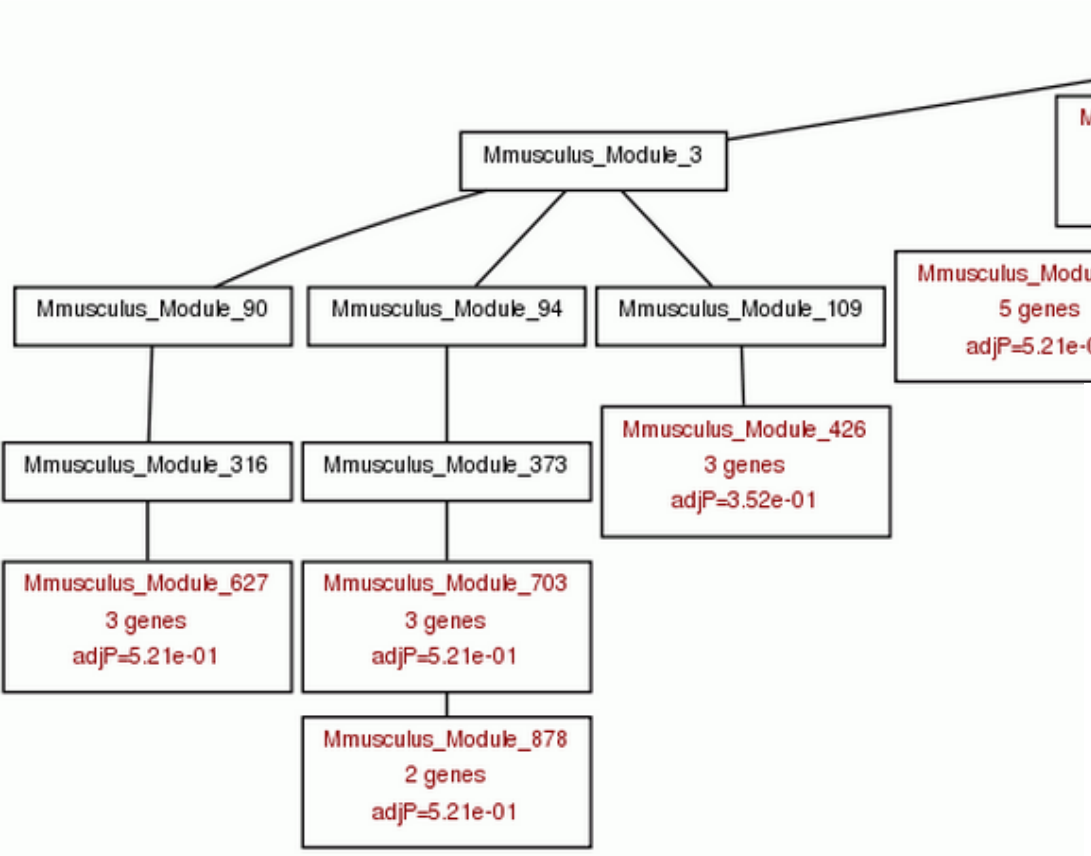
Enriched pathway

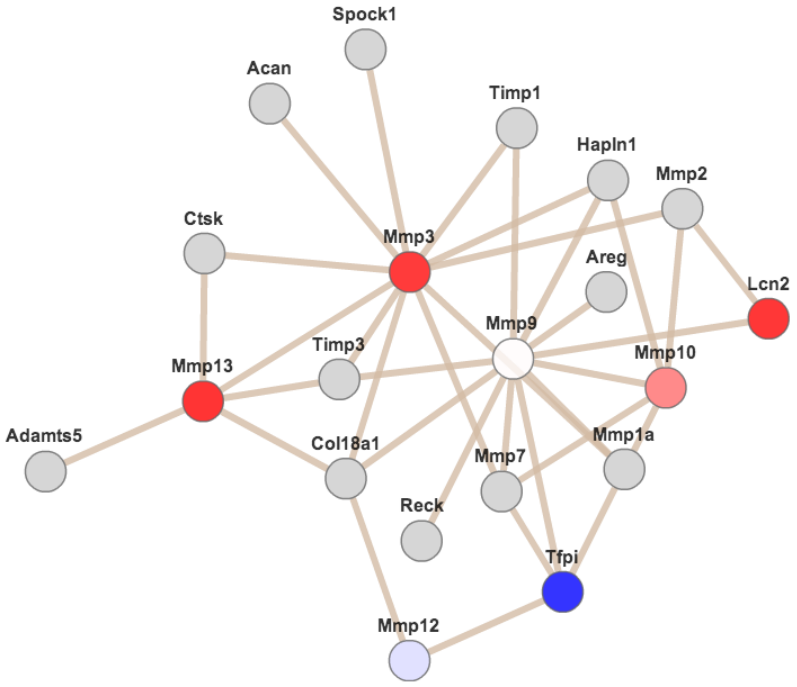


Input genes

TGF Beta Signaling Pathway

Enriched network modules



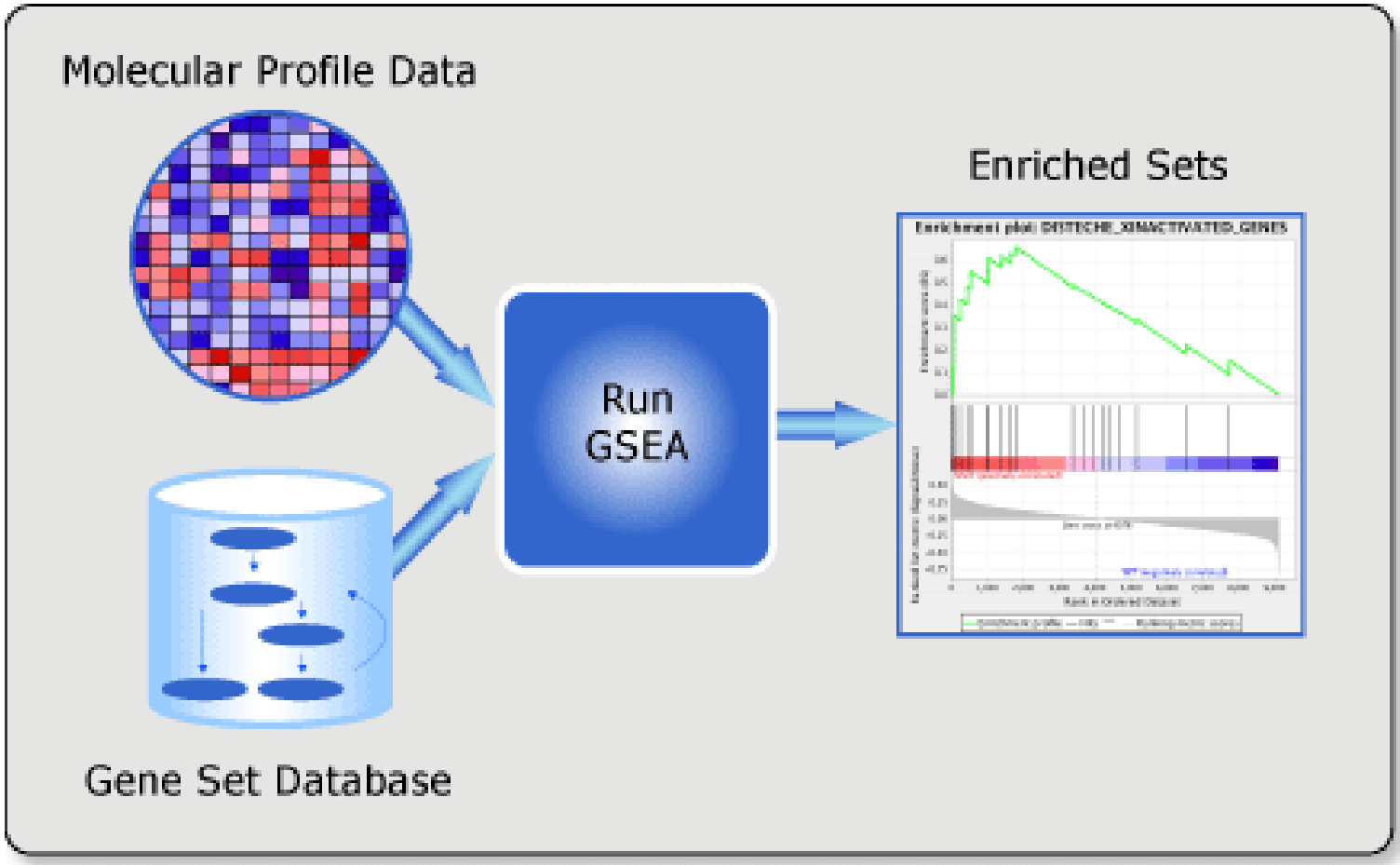


UNIVERSITEIT STELLENBOSCH UNIVERSITY

Over-representation analysis: limitations

- Thresholding can be quite arbitrary.
- Ignoring the order of genes in the significant gene list throws away magnitude of p-value.
- Treating pathway genes as a set ignores that some genes are central to a pathway while others are less affiliated.

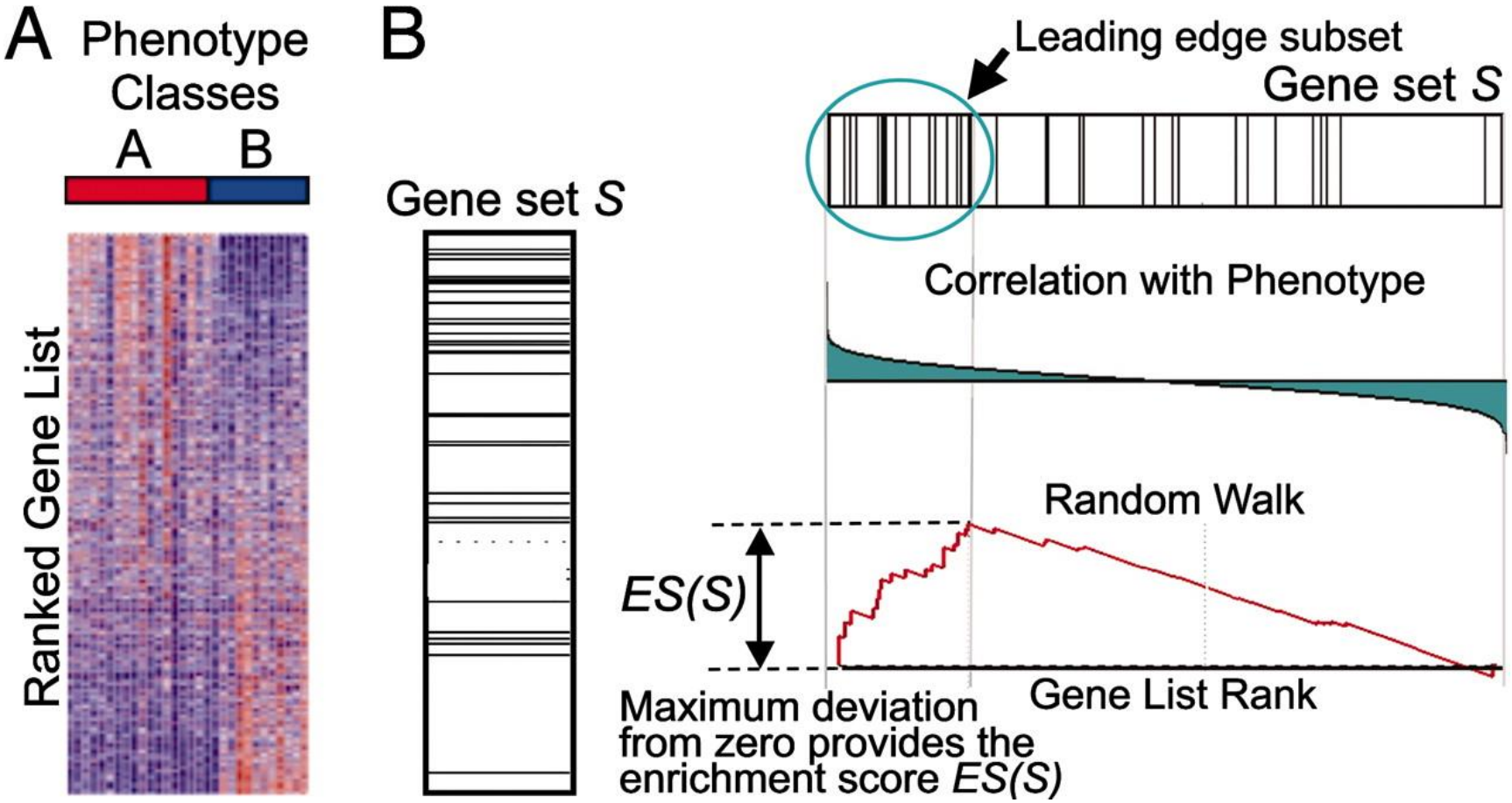
Gene Set Enrichment Analysis



www.broad.mit.edu/gsea

UNIVERSITEIT STELLENBOSCH UNIVERSITY

Gene Set Enrichment Analysis: method



Subramanian et al. *PNAS* (2005) 102:15545

Pathway-based analysis

- Organizing genes by
  - Pathways
  - Gene Ontology
- Enrichment analysis methods
  - Over-representation analysis
  - Gene Set enrichment analysis
- Major limitation: Existing knowledge of gene functions is far from complete

Intermission

Graph Theory Definitions

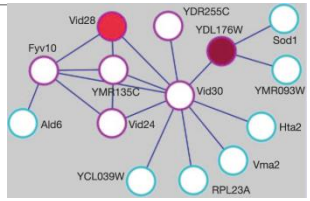
- Node: a vertex, generally representing an object or concept, particularly genes or proteins
- Edge: a relationship between a pair of nodes. May be directional (in *digraph*) or undirected
- Degree: the number of edges for a node
- Connected component: a set of interconnected nodes that have no edges to nodes outside the set

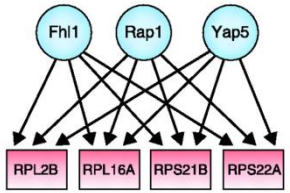
Advanced Definitions

- Clique: a set of nodes for which every possible connection is present
- Module: sets of nodes that are more strongly connected among the set than outside it.
- Path length: how many edges must be traversed to get from node A to node B?
- Hub: a node of high degree that is *between* many pairs of other nodes.

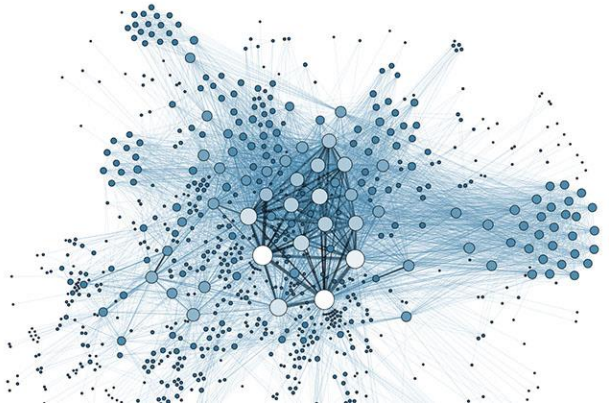
Biological networks

Networks		Nodes	Edges
Physical interaction networks	Protein-protein interaction network	Proteins	Physical interaction, undirected
	Signaling network	Proteins	Modification, directed
	Gene regulatory network	TFs/miRNAs Target genes	Physical interaction, directed
	Metabolic network	Metabolites	Metabolic reaction, directed
Functional association networks	Co-expression network	Genes/proteins	Co-expression, undirected
	Genetic network	Genes	Genetic interaction, undirected

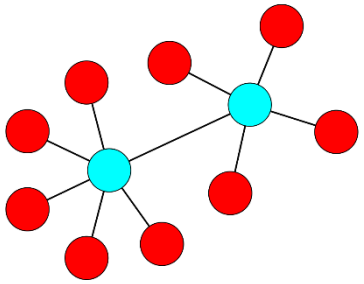




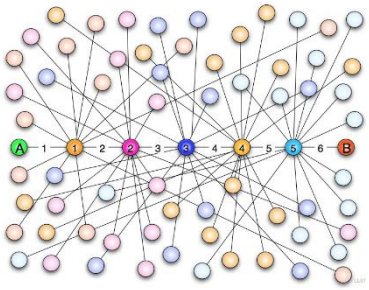
Properties of complex networks



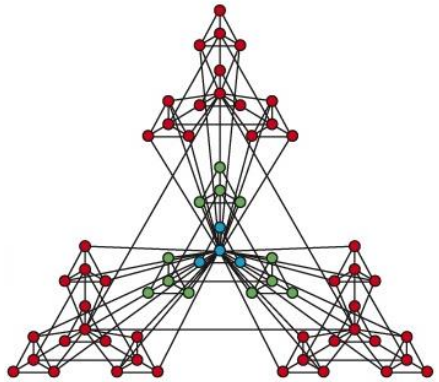
Human protein-protein interaction network
9,198 proteins and 36,707 interactions



Scale-free
(hubs)



Small world
(6° separation)



Hierarchical modular

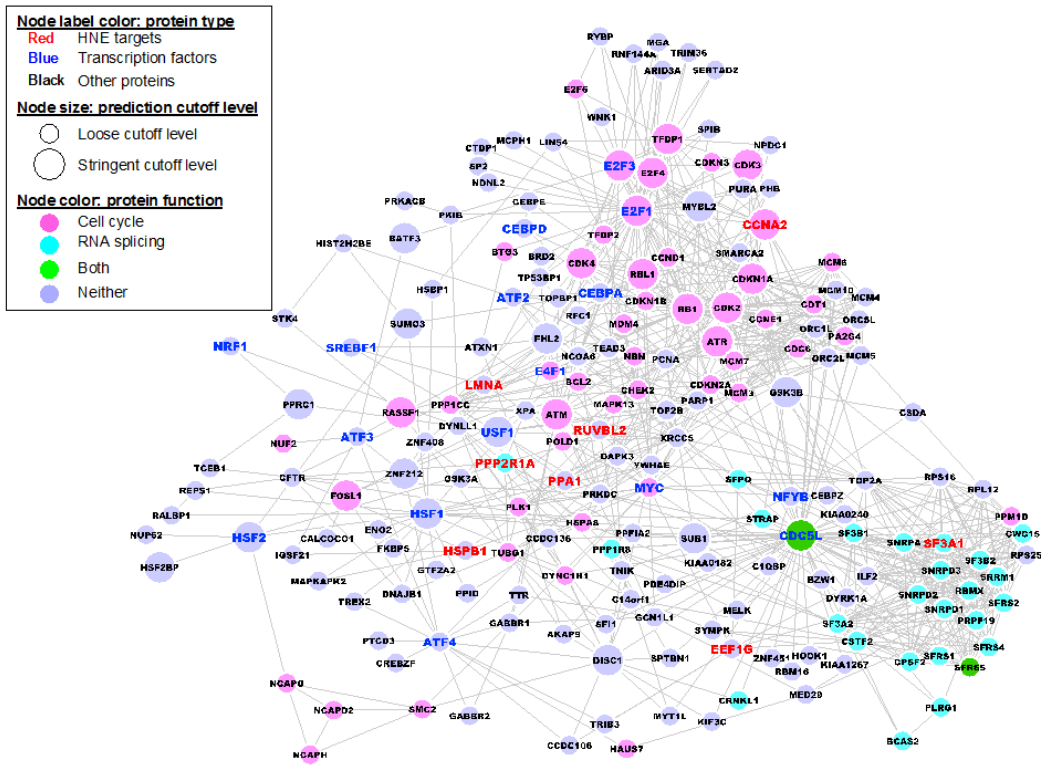
Barabási. *Nature Reviews Genet*. (2004) 5:101-113

Network visualization

ASSORTED TOOLS

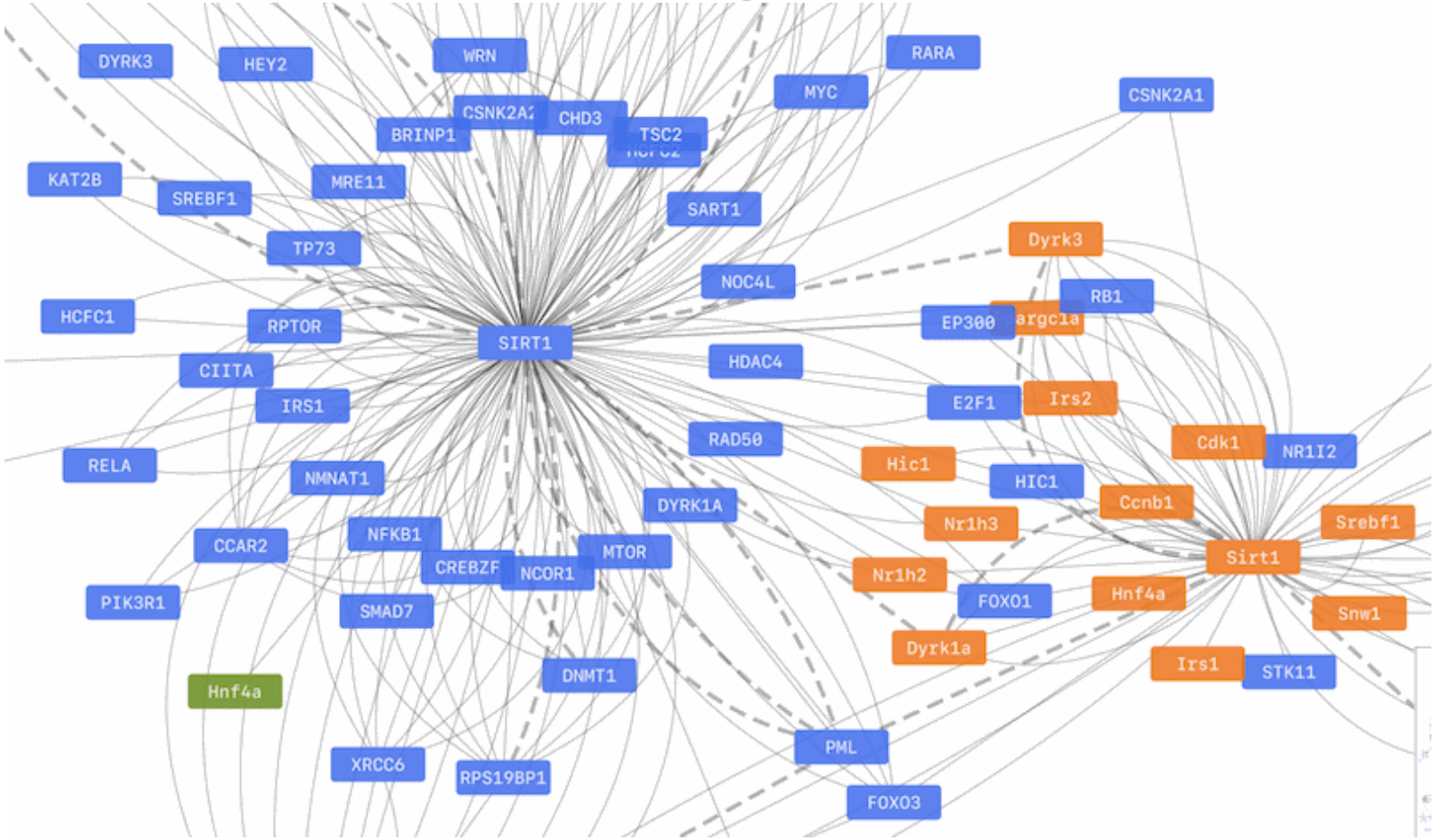
- GraphViz
- VizANT
- Medusa3
- Ondex
- Pajek
- BioLayout Express^{3D}

CYTOSCAPE



Gehlenborg et al. *Nature Methods* (2010) 7:S56

Cytoscape integrates and visualizes through networks.

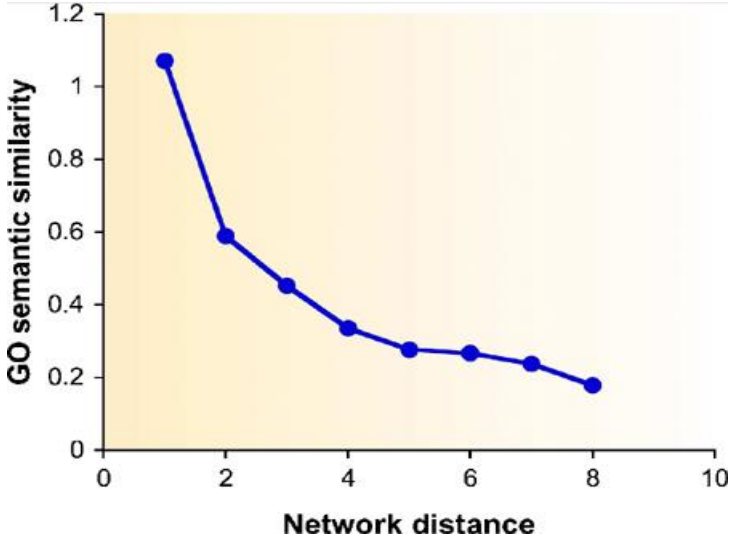


P Shannon et al. *Genome Research* (2003) 13: 2498-2504

ME Smoot et al. *Bioinformatics* (2011) 27: 431-432

Network distance *vs* functional similarity

- Proteins that lie closer to one another in a protein interaction network are more likely to have similar function and involve in similar biological process.
- Network-based gene function prediction
- Network-based disease gene prediction

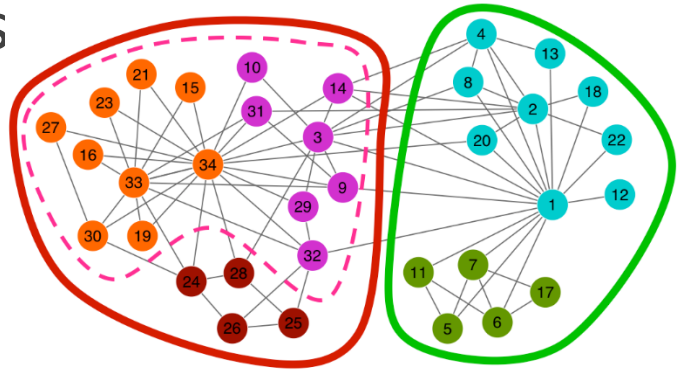


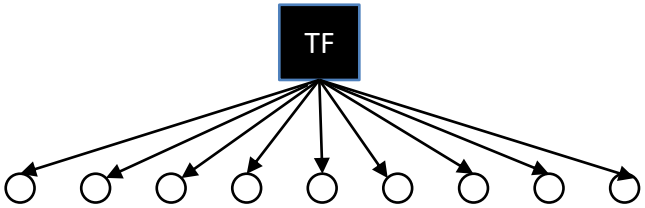
Sharan et al. *Mol Syst Biol*. (2007) 3:88

UNIVERSITEIT STELLENBOSCH UNIVERSITY

Organizing genes based on network modules

- Protein-protein interaction modules
- Transcriptional regulatory modules
  - Transcription factor targets
  - miRNA targets
- Network module-based analysis





https://www.bioconductor.org/packages/release/bioc/html/NetSAM.html

NetGestalt: network module-based interaction

Network modules

Gene names

T-test –log(p-values)

Microarray measurements

Shi et al. *Nature Methods* (2013) 10: 597-598.

Takeaway Messages

- Building your assessment on pathways or networks rather than genes or proteins may have two key effects:
  - Biological interpretability should be far greater.
  - You will incorporate more data in each statistical test.
- Biological pathways are built on a categorical basis, while biological networks borrow from graph theory for analysis.
- Gene Set Enrichment Analysis and Over-Representation Analysis are two of the most common statistical tests for pathway and network data.