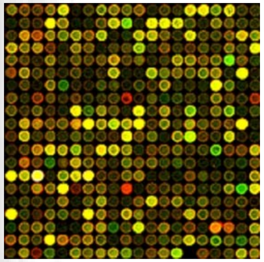
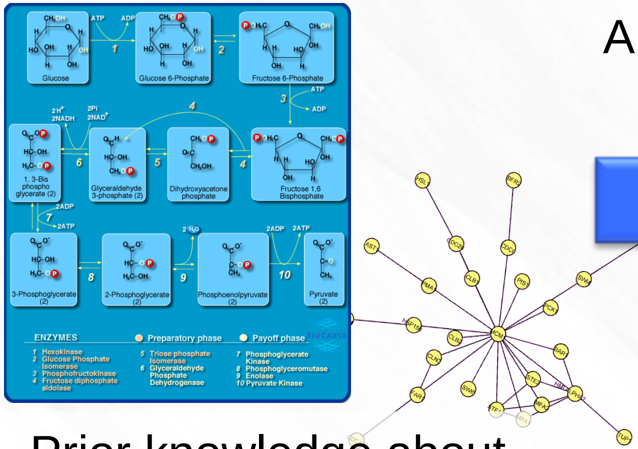
Interpreting Gene Lists

- The analysis produced 1000 hits-> Now what?
- Genome-Scale Analysis (Omics)
  - Genomics, Proteomics
- What’s interesting about these genes
  - Are they enriched in known pathways, complexes, functions



Ranking or clustering

GNAQ
GNAS
DGKZ
GUCY1A3
PDE4B
PDE4D
ATP2A2
ATP2A3
NOS1
CNN1
GSTO1
NOS3
CNN2
MYLK2
CALD1
ACTA1
MYL2



Prior knowledge about cellular processes

Analysis tools

Eureka! New heart disease gene!

Pathway and Network Analysis

- Any type of analysis that involves pathway or network information
- Most commonly applied to help interpret lists of genes
- Most popular type is pathway enrichment analysis, but many others are useful
- Helps gain mechanistic insight into ‘omics data

Correlation to Causation

- GWAS: find genetic markers correlated with disease – powerful approach, but:
  - genomics reduces statistical power (>multiple testing correction with >SNPs)
  - rare variants = more samples
- Associate pathways to increase power
  - Fewer pathways, organize many rare variants (damaging the system causes the disease)
- Use pathway knowledge to identify potential disease causes

Before Analysis

- ✓ Normalization
- ✓ Background adjustment
- ✓ Quality control (garbage in, garbage out)

- ✓ Use statistics that will increase signal and reduce noise specifically for your experiment
- ✓ Other analyses you may want to use to evaluate changes
- ✓ Make sure your gene IDs are compatible with software

Where Do Gene Lists Come From?

- Molecular profiling e.g. mRNA, protein
  - Identification → Gene list
  - Quantification → Gene list + values
  - Ranking, Clustering (biostatistics)
- Interactions: Protein interactions, microRNA targets, transcription factor binding sites (ChIP)
- Genetic screen e.g. of knock out library
- Association studies (Genome-wide)
  - Single nucleotide polymorphisms (SNPs)
  - Copy number variants (CNVs)

What Do Gene Lists Mean?

- Biological system: complex, pathway, physical interactors
- Similar gene function e.g. protein kinase
- Similar cell or tissue location
- Chromosomal location (linkage, CNVs)

Data

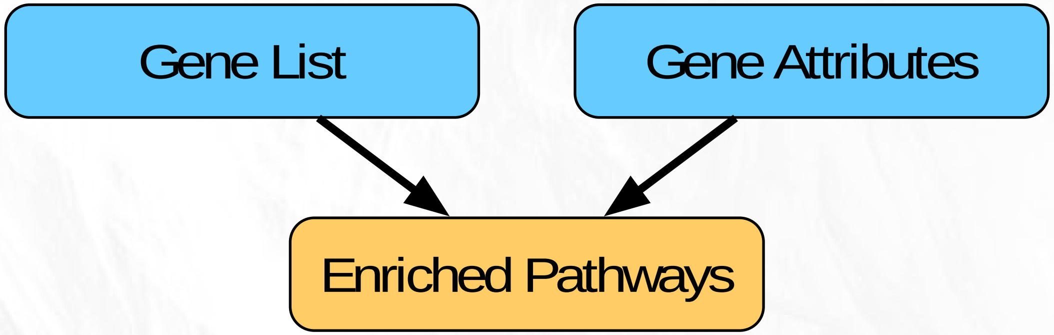
Biological Questions

- Step 1: What do you want to accomplish with your list
  - Summarize biological processes or other aspects of gene function
  - Perform differential analysis – what pathways are different between samples?
  - Find a controller for a process (TF, miRNA)
  - Find new pathways or new pathway members
  - Discover new gene function
  - Correlate with a disease or phenotype (candidate gene prioritization)

Biological Answers

- Computational analysis methods we will cover
  - Regulatory network analysis: find controllers
  - Pathway enrichment analysis: summarize and compare
  - Network analysis: predict gene function, find new pathway members, identify functional modules (new pathways)

Pathway Enrichment Analysis



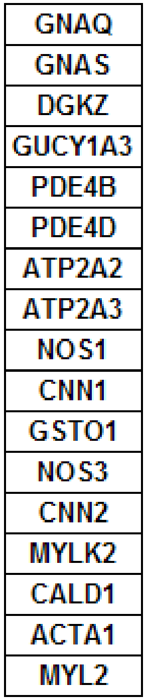
Annotation
Expression
Phenotypes

DAVID, GSEA, g:Profiler

- Gene identifiers
- Gene attributes/annotation
  - Gene Ontology
    - Ontology Structure
    - Annotation
  - BioMart + other sources

Gene and Protein Identifiers

- Identifiers (IDs) are ideally unique, stable names or numbers that help track database records
  - E.g. Social Insurance Number, Entrez Gene ID 41232
- Gene and protein information stored in many databases
  - → Genes have many IDs
- Records for: Gene, DNA, RNA, Protein
  - Important to recognize the correct record type
  - E.g. Entrez Gene records don’t store sequence. They link to DNA regions, RNA transcripts and proteins e.g. in RefSeq, which stores sequence.



GNAQ
GNAS
DGKZ
GUCY1A3
PDE4B
PDE4D
ATP2A2
ATP2A3
NOS1
CNN1
GSTO1
NOS3
CNN2
MYLK2
CALD1
ACTA1
MYL2

Common Identifiers

Gene

Ensembl ENSG00000139618
Entrez Gene 675
Unigene Hs.34012

RNA transcript

GenBank BC026160.1
RefSeq NM_000059
Ensembl ENST00000380152

Protein

Ensembl ENSP00000369497
RefSeq NP_000050.2
UniProt BRCA2_HUMAN or A1YBP1_HUMAN
IPI IPI00412408.1
EMBL AF309413
PDB 1MIU

Species-specific

HUGO HGNC BRCA2
MGI MGI:109337
RGD 2219
ZFIN ZDB-GENE-060510-3
FlyBase CG9097
WormBase WBGene00002299 or ZK1067.1
SGD S000002187 or YDL029W

Annotations

InterPro IPR015252
OMIM 600185
Pfam PF09104
Gene Ontology GO:0000724
SNPs rs28897757

Experimental Platform

Affymetrix 208368_3p_s_at
Agilent A_23_P99452
CodeLink GE60169
Illumina GI_4502450-S

Red = Recommended

Identifier Mapping

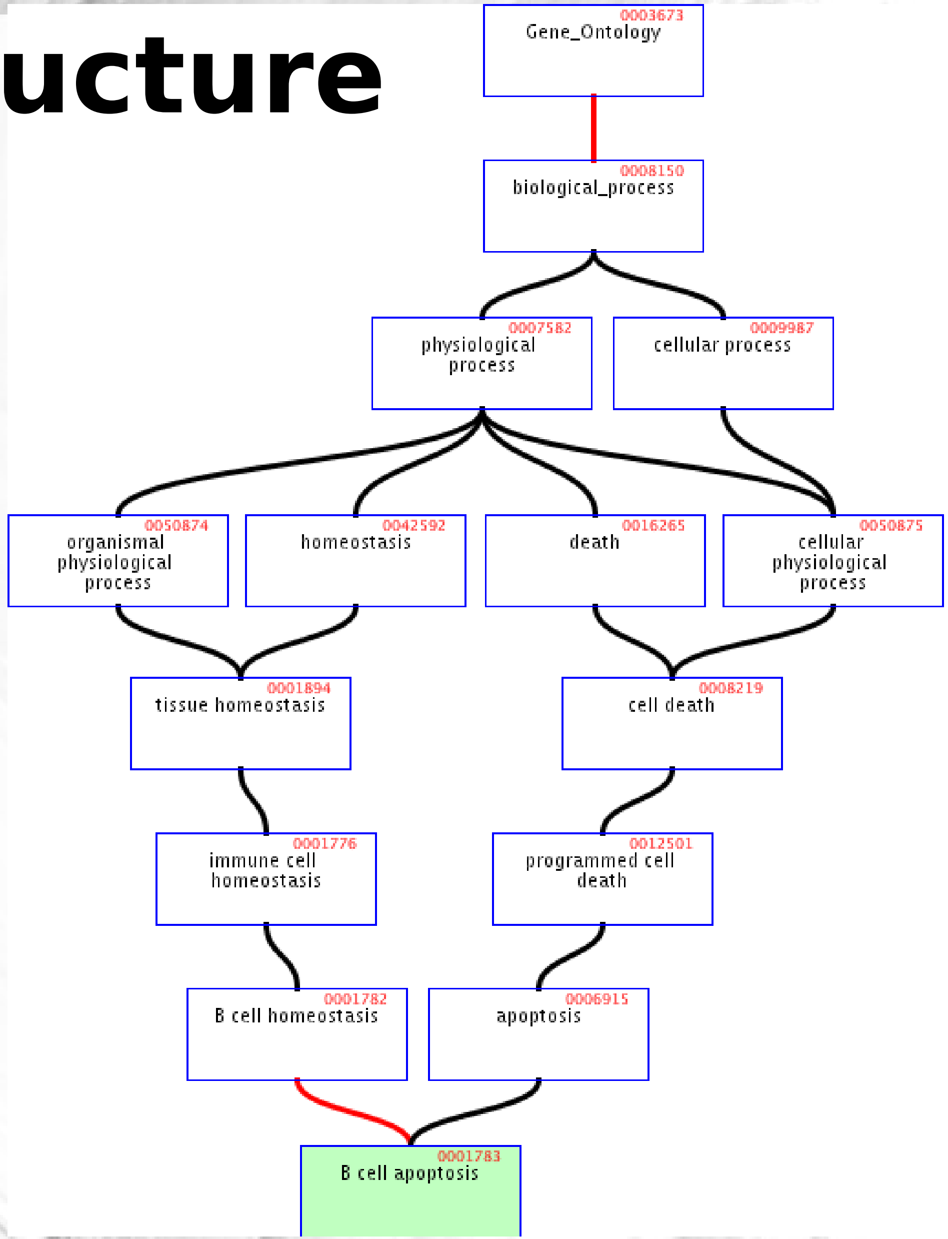
- So many IDs!
  - Software tools recognize only a handful
  - May need to map from your gene list IDs to standard IDs
- Four main uses
  - Searching for a favorite gene name
  - Link to related resources
  - Identifier translation
    - E.g. Proteins to genes, Affy ID to Entrez Gene
  - Merging data from different sources
    - Find equivalent records

What is the Gene Ontology (GO)?

- Set of biological phrases (terms) which are applied to genes:
  - protein kinase
  - apoptosis
  - membrane
- Dictionary: term definitions
- Ontology: A formal system for describing knowledge
- www.geneontology.org

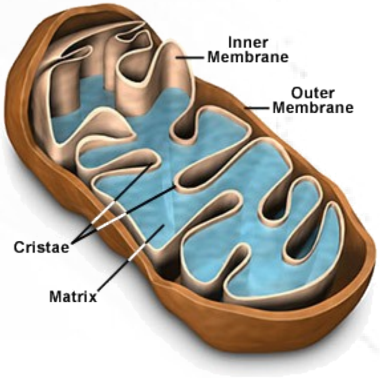
GO Structure

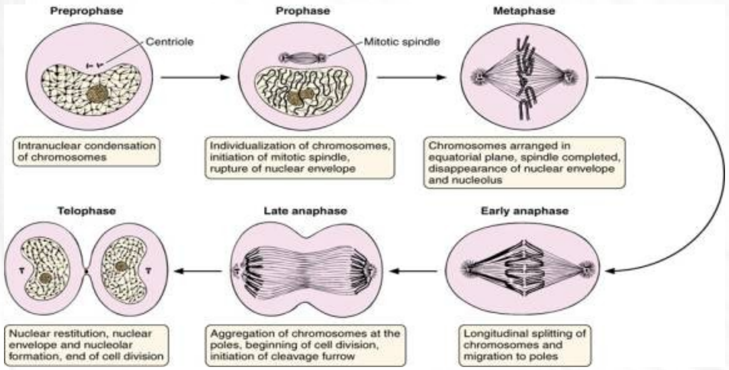
- Terms are related within a hierarchy
  - is-a
  - part-of
- Describes multiple levels of detail of gene function
- Terms can have more than one parent or child



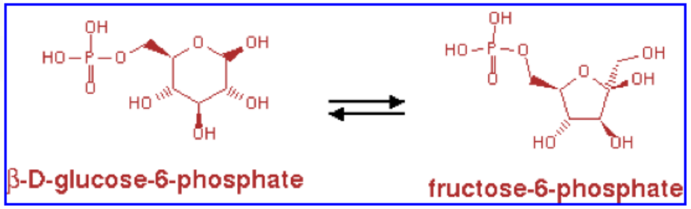
What GO Covers?

- GO terms divided into three aspects:
  - cellular component
  - molecular function
  - biological process




Preprophase
Centriole
Intranuclear condensation of chromosomes
Prophase
Mitotic spindle
Individualization of chromosomes, initiation of mitotic spindle, rupture of nuclear envelope
Metaphase
Chromosomes arranged in equatorial plane, spindle completed, disappearance of nuclear envelope and nucleolus
Early anaphase
Longitudinal splitting of chromosomes and migration to poles
Late anaphase
Aggregation of chromosomes at the poles, beginning of cell division, initiation of cleavage furrow
Telophase
Nuclear restitution, nuclear envelope and nucleolar formation, end of cell division

Cell division



glucose-6-phosphate isomerase activity

Part 1/2: Terms

- Where do GO terms come from?
  - GO terms are added by editors at EBI and gene annotation database groups
  - Terms added by request
  - Experts help with major development
  - 37104 terms, with definitions
    - 23074 biological_process
    - 2994 cellular_component
    - 9392 molecular_function
    - As of June 2012

Part 2/2: Annotations

- Genes are linked, or associated, with GO terms by trained curators at genome databases
  - Known as ‘gene associations’ or GO annotations
  - Multiple annotations per gene
- Some GO annotations created automatically (without human review)

Species Coverage

- All major eukaryotic model organism species and human
- Several bacterial and parasite species through TIGR and GeneDB at Sanger
- New species annotations in development
- Current list:
  - http://www.geneontology.org/GO.downloads.annotations.shtml

Gene Attributes

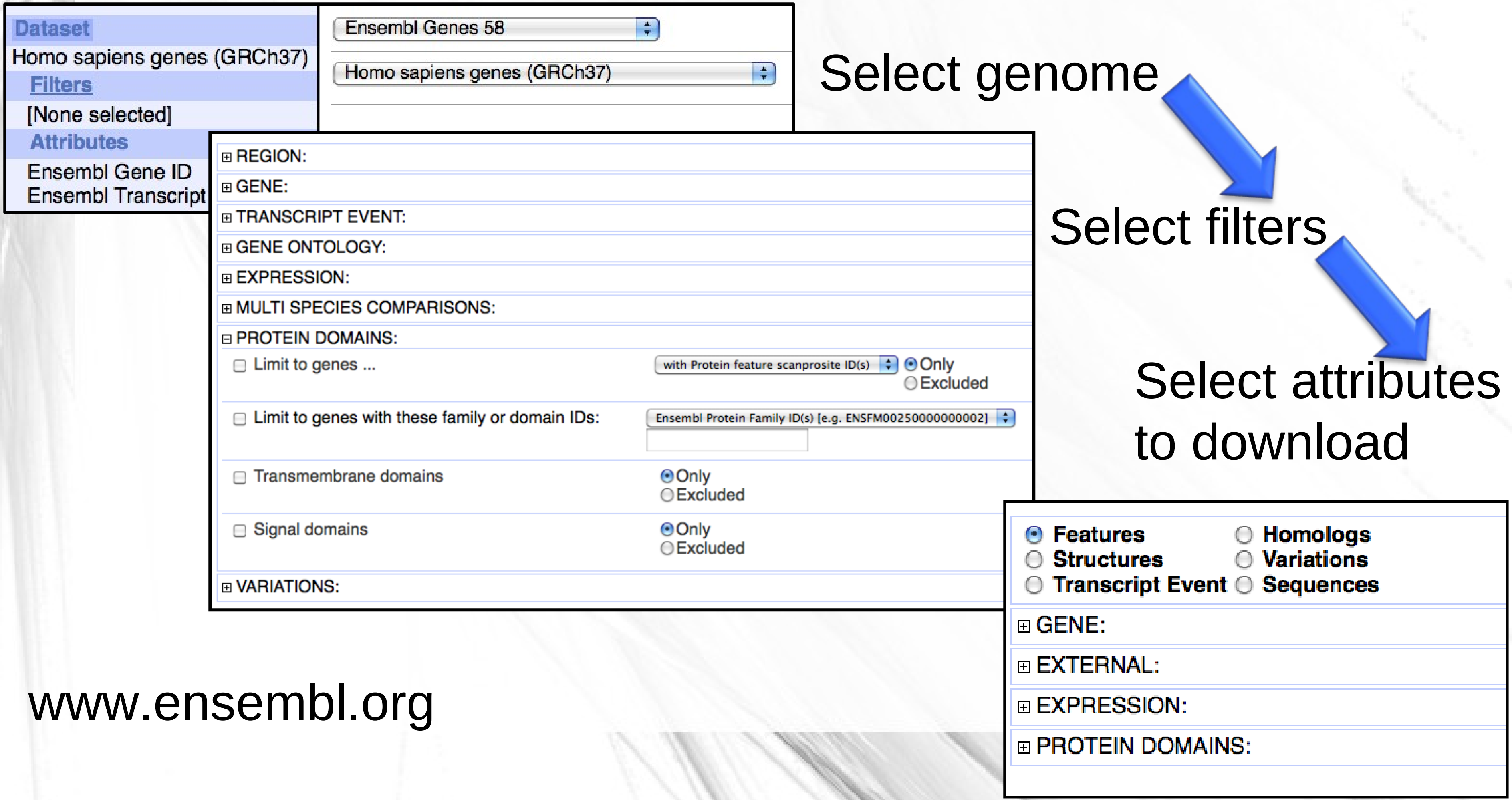
- Function annotation
  - Biological process, molecular function, cell location
- Chromosome position
- Disease association
- DNA properties
  - TF binding sites, gene structure (intron/exon), SNPs
- Transcript properties
  - Splicing, 3’ UTR, microRNA binding sites
- Protein properties
  - Domains, secondary and tertiary structure, PTM sites
- Interactions with other genes

Sources of Gene Attributes

- Ensembl BioMart (general)
  - http://www.ensembl.org
- Entrez Gene (general)
  - http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene
- Model organism databases
  - E.g. SGD: http://www.yeastgenome.org/
- Many others.....

Ensembl BioMart

- Convenient access to gene list annotation



Dataset
Homo sapiens genes (GRCh37)
Filters
[None selected]
Attributes
Ensembl Gene ID
Ensembl Transcript

Ensembl Genes 58
Homo sapiens genes (GRCh37)

Select genome

⊞ REGION:
⊞ GENE:
⊞ TRANSCRIPT EVENT:
⊞ GENE ONTOLOGY:
⊞ EXPRESSION:
⊞ MULTI SPECIES COMPARISONS:
⊟ PROTEIN DOMAINS:

□ Limit to genes ... — with Protein feature scanprosite ID(s) — ● Only ○ Excluded

□ Limit to genes with these family or domain IDs: — Ensembl Protein Family ID(s) [e.g. ENSFM00250000000002]

□ Transmembrane domains — ● Only ○ Excluded

□ Signal domains — ● Only ○ Excluded

⊞ VARIATIONS:

Select filters

Select attributes to download

● Features ○ Homologs
○ Structures ○ Variations
○ Transcript Event ○ Sequences

⊞ GENE:
⊞ EXTERNAL:
⊞ EXPRESSION:
⊞ PROTEIN DOMAINS:

www.ensembl.org

Enrichment analysis

- Introduction to enrichment analysis
- Hypergeometric Test, Fisher’s Exact Test
- GSEA enrichment analysis for ranked lists.
- Multiple test corrections:

– Bonferroni correction

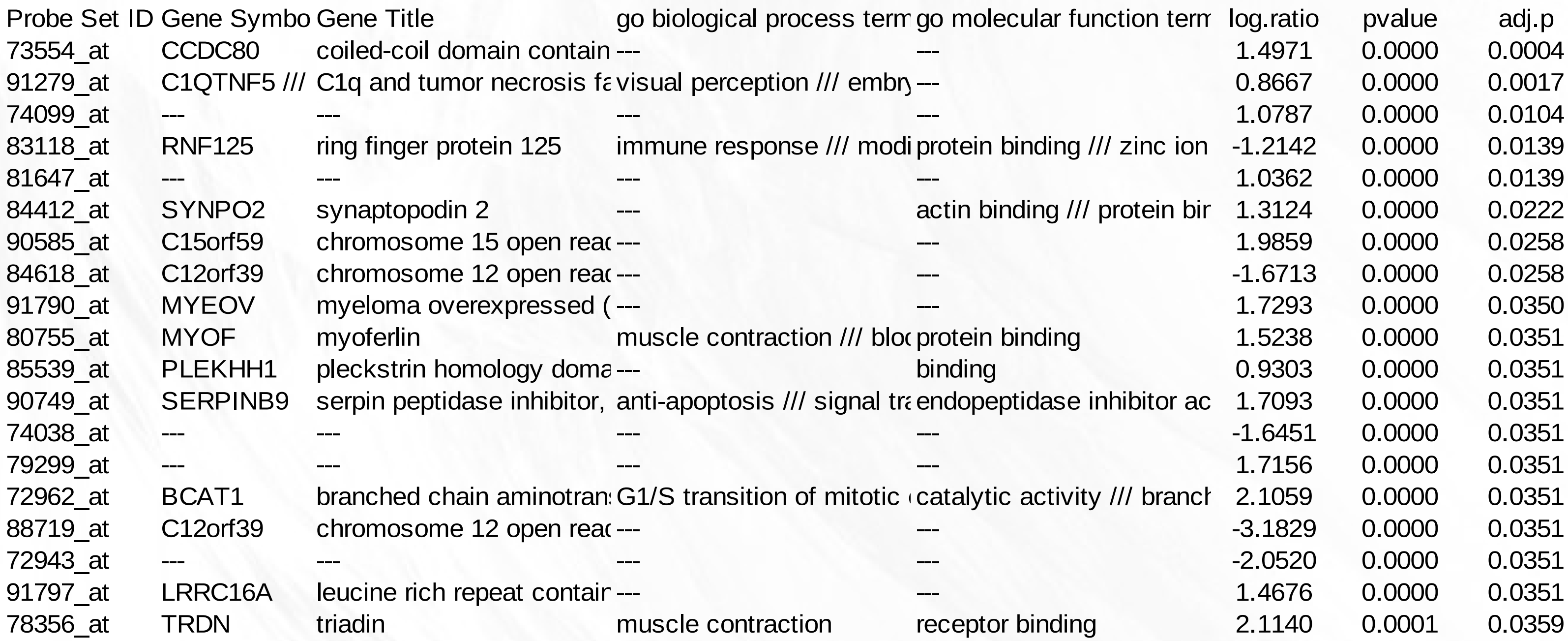
– False Discovery Rate computation using Benjamini-Hochberg procedure

The “result”

Probe Set ID	log.ratio	pvalue	adj.p
73554_at	1.4971	0.0000	0.0004
91279_at	0.8667	0.0000	0.0017
74099_at	1.0787	0.0000	0.0104
83118_at	-1.2142	0.0000	0.0139
81647_at	1.0362	0.0000	0.0139
84412_at	1.3124	0.0000	0.0222
90585_at	1.9859	0.0000	0.0258
84618_at	-1.6713	0.0000	0.0258
91790_at	1.7293	0.0000	0.0350
80755_at	1.5238	0.0000	0.0351
85539_at	0.9303	0.0000	0.0351
90749_at	1.7093	0.0000	0.0351
74038_at	-1.6451	0.0000	0.0351
79299_at	1.7156	0.0000	0.0351
72962_at	2.1059	0.0000	0.0351
88719_at	-3.1829	0.0000	0.0351
72943_at	-2.0520	0.0000	0.0351
91797_at	1.4676	0.0000	0.0351
78356_at	2.1140	0.0001	0.0359

What about the Biology???

Slightly more informative results



Probe Set ID	Gene Symbo	Gene Title	go biological process term	go molecular function term	log.ratio	pvalue	adj.p
73554_at	CCDC80	coiled-coil domain contain	---	---	1.4971	0.0000	0.0004
91279_at	C1QTNF5 ///	C1q and tumor necrosis fa	visual perception /// embry	---	0.8667	0.0000	0.0017
74099_at	---	---	---	---	1.0787	0.0000	0.0104
83118_at	RNF125	ring finger protein 125	immune response /// modi	protein binding /// zinc ion	-1.2142	0.0000	0.0139
81647_at	---	---	---	---	1.0362	0.0000	0.0139
84412_at	SYNPO2	synaptopodin 2	---	actin binding /// protein bir	1.3124	0.0000	0.0222
90585_at	C15orf59	chromosome 15 open reac	---	---	1.9859	0.0000	0.0258
84618_at	C12orf39	chromosome 12 open reac	---	---	-1.6713	0.0000	0.0258
91790_at	MYEOV	myeloma overexpressed (	---	---	1.7293	0.0000	0.0350
80755_at	MYOF	myoferlin	muscle contraction /// bloc	protein binding	1.5238	0.0000	0.0351
85539_at	PLEKHH1	pleckstrin homology doma	---	binding	0.9303	0.0000	0.0351
90749_at	SERPINB9	serpin peptidase inhibitor,	anti-apoptosis /// signal tra	endopeptidase inhibitor ac	1.7093	0.0000	0.0351
74038_at	---	---	---	---	-1.6451	0.0000	0.0351
79299_at	---	---	---	---	1.7156	0.0000	0.0351
72962_at	BCAT1	branched chain aminotrans	G1/S transition of mitotic c	catalytic activity /// branch	2.1059	0.0000	0.0351
88719_at	C12orf39	chromosome 12 open reac	---	---	-3.1829	0.0000	0.0351
72943_at	---	---	---	---	-2.0520	0.0000	0.0351
91797_at	LRRC16A	leucine rich repeat contair	---	---	1.4676	0.0000	0.0351
78356_at	TRDN	triadin	muscle contraction	receptor binding	2.1140	0.0001	0.0359

If we are lucky, some of the top genes mean something to us

But what if they don’t?

And how what are the results for other genes with similar biological functions

Apply some methods to incorporate biological knowledge into microarray analysis

The type of knowledge to deal with is rather simple: We know groups/sets of genes that for example

- Belong to the same pathway
- Have a similar function
- Are located on the same chromosome, etc…

We will assume these groupings to be given, i.e we will not discuss methods how to detect pathways, networks, gene clusters

What is a pathway?

- No clear definition
  - Wikipedia: “In biochemistry, **metabolic pathways** are series of chemical reactions occurring within a cell. In each pathway, a principal chemical is modified by chemical reactions.”
  - These pathways describe enzymes and metabolites
- But often the word “pathway” is also used to describe gene regulatory networks or protein interaction networks
- In all cases a pathway describes a biological function very specifically

What is a Gene Set?

- All genes involved in a pathway are an example of a Gene Set
- All genes corresponding to a Gene Ontology term are a Gene Set
- All genes mentioned in a paper of Smith et al might form a Gene Set

A Gene Set is a much more general and less specific concept than a pathway

What is Gene Set/Pathway analysis?

The aim is to give one number (score, p-value) to a Gene Set/Pathway to answer questions like:

- Are many genes in the pathway differentially expressed (up-regulated/downregulated)?
- Can we give a number (p-value) to the probability of observing these changes just by chance?

Pathway and Gene Set data resources

The Gene Ontology (GO) database

http://www.geneontology.org/

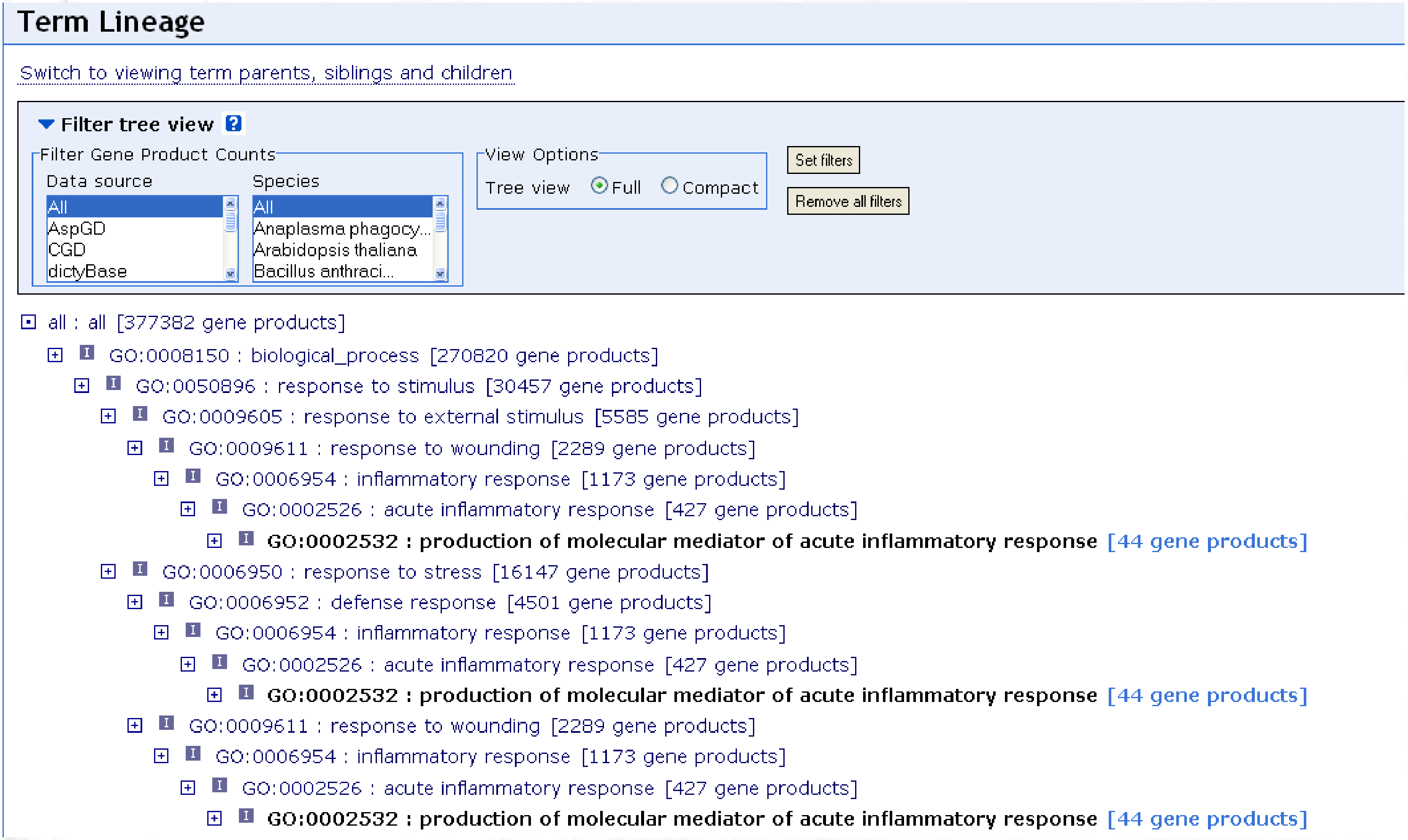
GO offers a relational/hierarchical database

- Parent nodes: more general terms
- Child nodes: more specific terms

At the end of the hierarchy there are genes/proteins

At the top there are 3 parent nodes: biological process, molecular function and cellular component

Example: we search the database for the term “inflammation”



Term Lineage

Switch to viewing term parents, siblings and children

Filter tree view

Filter Gene Product Counts

Data source	Species
All	All
AspGD	Anaplasma phagocy...
CGD	Arabidopsis thaliana
dictyBase	Bacillus anthraci...

View Options

Tree view ● Full ○ Compact

```
⊡ all : all [377382 gene products]
   ⊞ I GO:0008150 : biological_process [270820 gene products]
      ⊞ I GO:0050896 : response to stimulus [30457 gene products]
         ⊞ I GO:0009605 : response to external stimulus [5585 gene products]
            ⊞ I GO:0009611 : response to wounding [2289 gene products]
               ⊞ I GO:0006954 : inflammatory response [1173 gene products]
                  ⊞ I GO:0002526 : acute inflammatory response [427 gene products]
                     ⊞ I GO:0002532 : production of molecular mediator of acute inflammatory response [44 gene products]
         ⊞ I GO:0006950 : response to stress [16147 gene products]
            ⊞ I GO:0006952 : defense response [4501 gene products]
               ⊞ I GO:0006954 : inflammatory response [1173 gene products]
                  ⊞ I GO:0002526 : acute inflammatory response [427 gene products]
                     ⊞ I GO:0002532 : production of molecular mediator of acute inflammatory response [44 gene products]
            ⊞ I GO:0009611 : response to wounding [2289 gene products]
               ⊞ I GO:0006954 : inflammatory response [1173 gene products]
                  ⊞ I GO:0002526 : acute inflammatory response [427 gene products]
                     ⊞ I GO:0002532 : production of molecular mediator of acute inflammatory response [44 gene products]
```

The genes on our array that code for one of the 44 gene products would form the corresponding “inflammation” gene set

KEGG pathway database

KEGG = Kyoto Encyclopedia of Genes and Genomes

http://www.genome.jp/kegg/pathway.html

The pathway database gives far more detailed information than GO

- Relationships between genes and gene products

But: this detailed information is only available for selected organisms and processes

Types of enrichment analysis

- **Gene list** (e.g. expression change > 2-fold)
  - Answers the question: **Are any gene sets surprisingly enriched (or depleted) in my gene list?**
  - Statistical test: Fisher’s Exact Test (Hypergeometric test)
- **Ranked list** (e.g. by differential expression)
  - Answers the question: **Are any gene set ranked surprisingly high or low in my ranked list of genes?**
  - Statistical test: GSEA

Gene list enrichment analysis

- Given:

1. Gene list: e.g. RRP6, MRD1, RRP7, RRP43, RRP42 (yeast)
2. Gene sets or annotations: e.g. Gene ontology, transcription factor binding sites in promoter

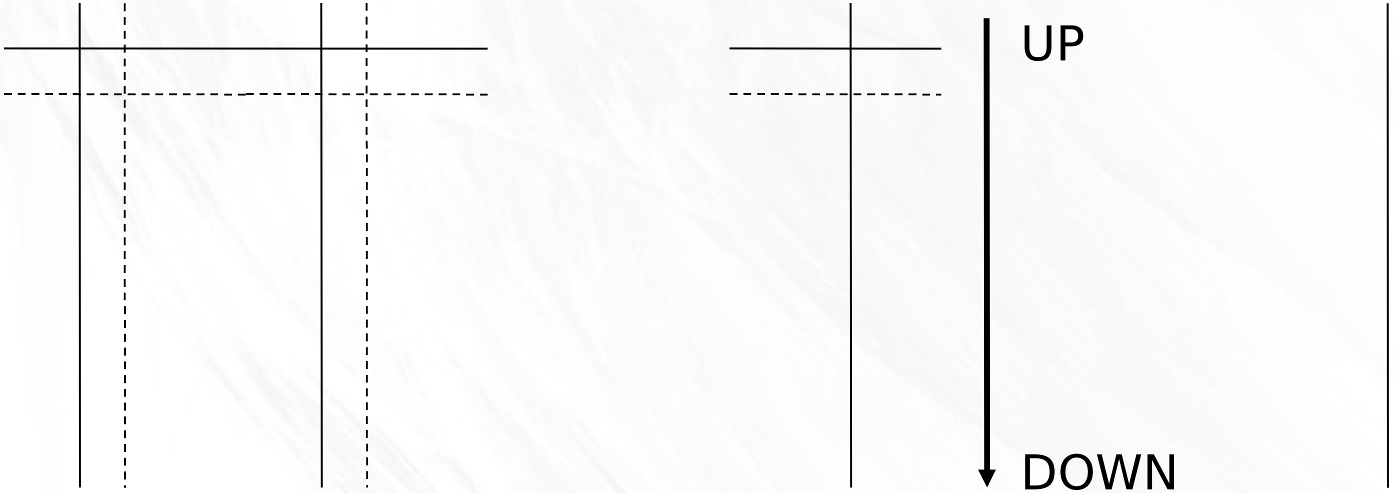
- Question: *Are any of the gene annotations <u>surprisingly</u> enriched in the gene list?*

Two-class design for gene lists

Expression Matrix

Genes Ranked by Differential Statistic

Selection by Threshold



UP

UP

DOWN

DOWN

Class-1 Class-2

E.g.:
- Fold change
- Log (ratio)
- t-test
- Significance analysis of microarrays

Recipe for gene list enrichment test

- **Step 1:** Rank your gene list,
- **Step 2:** Select your gene sets to test for enrichment,
- **Step 3:** Run enrichment tests and correct for multiple testing, if necessary,
- **Step 4:** Interpret your enrichments
- **Step 5:** Make conclusions

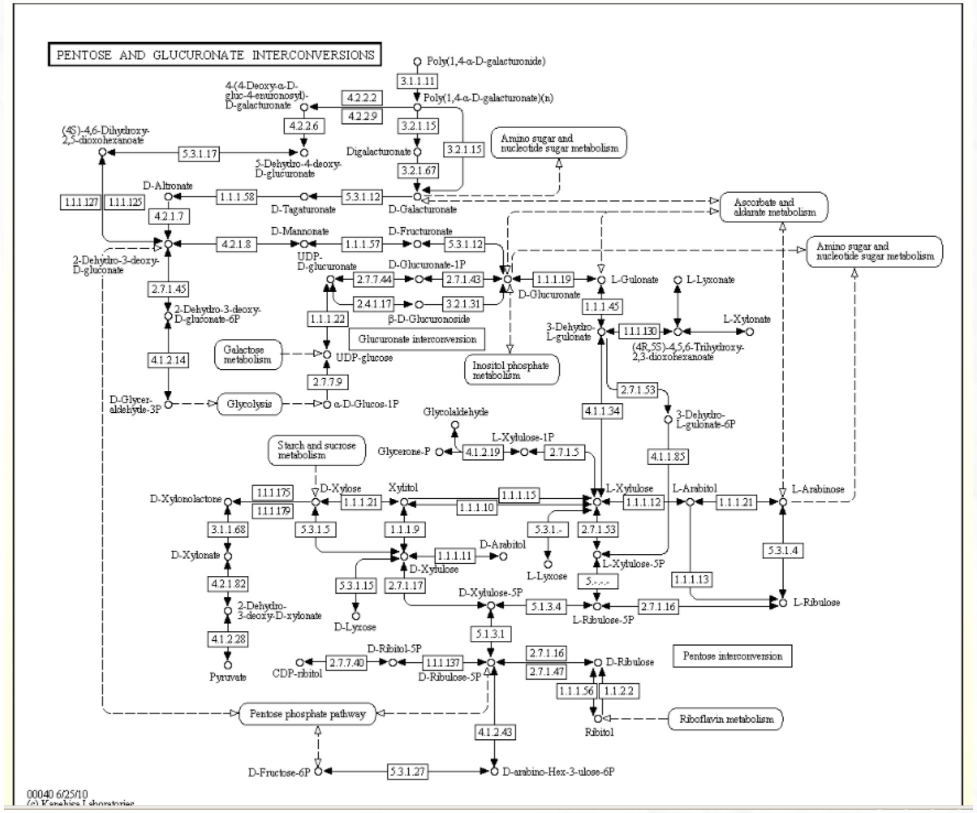
Theory component

- Hypergeometric test for calculating enrichment P-values for gene lists
- GSEA for computing enrichment P-values for ranked lists
- Multiple test corrections:
  - Bonferroni
  - Benjamini-Hochberg FDR

Important details

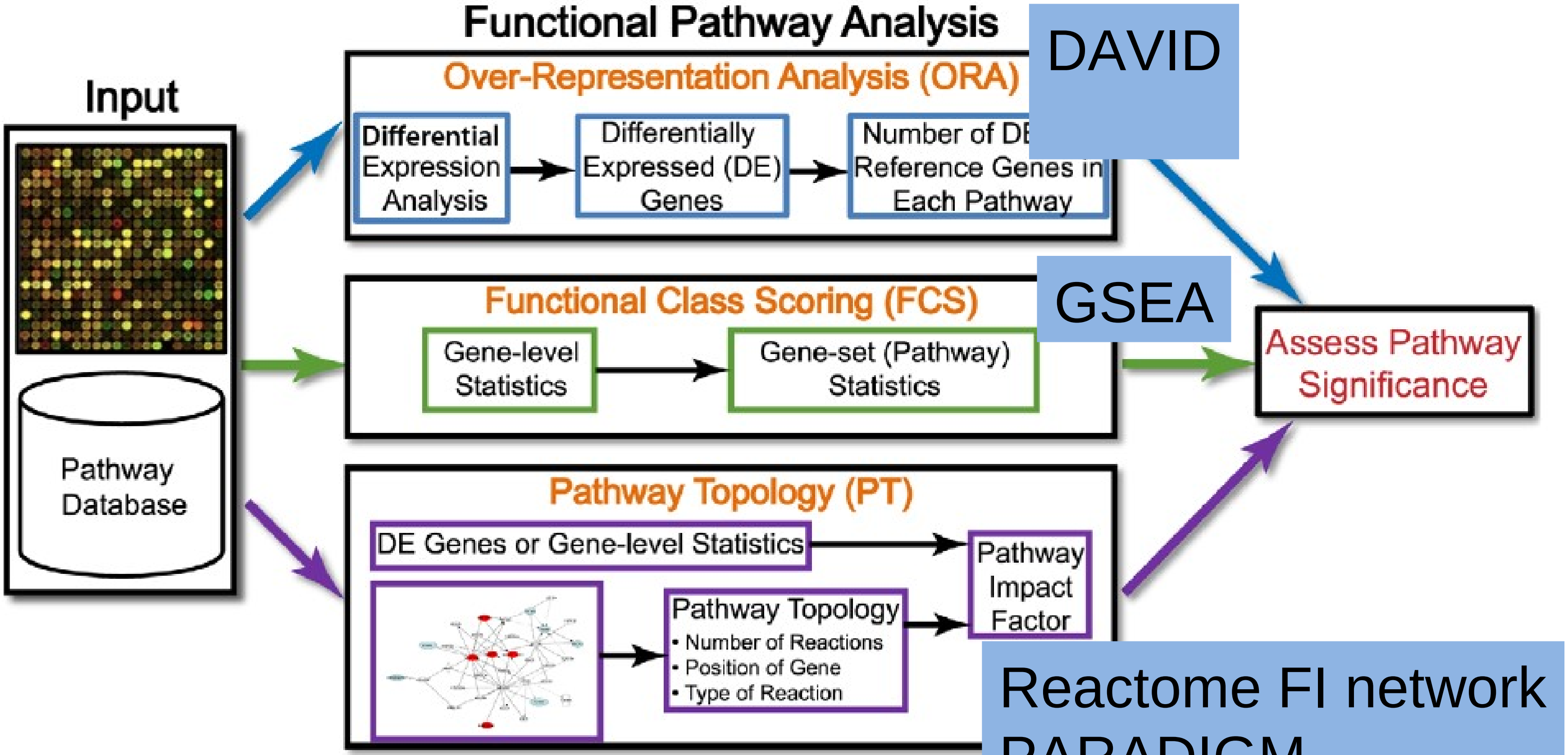
- To test for *under-enrichment* of “black”, test for *over-enrichment* of “red”.
- Need to choose “background population” appropriately, e.g., if only portion of the total gene complement is queried (or available for annotation), only use that population as background.
- To test for enrichment of more than one independent types of annotation (red vs black and circle vs square), apply Fisher’s exact test separately for each type.

Pathway and Network Analysis of –omics Data



00040 6/25/10
(c) Kanehisa Laboratories

Classes of Gene Set Analysis


Functional Pathway Analysis
Input
Pathway Database
Over-Representation Analysis (ORA)
Differential Expression Analysis
Differentially Expressed (DE) Genes
Number of DE Reference Genes in Each Pathway
Functional Class Scoring (FCS)
Gene-level Statistics
Gene-set (Pathway) Statistics
Pathway Topology (PT)
DE Genes or Gene-level Statistics
Pathway Topology
• Number of Reactions
• Position of Gene
• Type of Reaction
Pathway Impact Factor
Assess Pathway Significance

DAVID

GSEA

Reactome FI network
PARADIGM

Khatri *et al.* PLOS Comp Bio. 8:1 2012

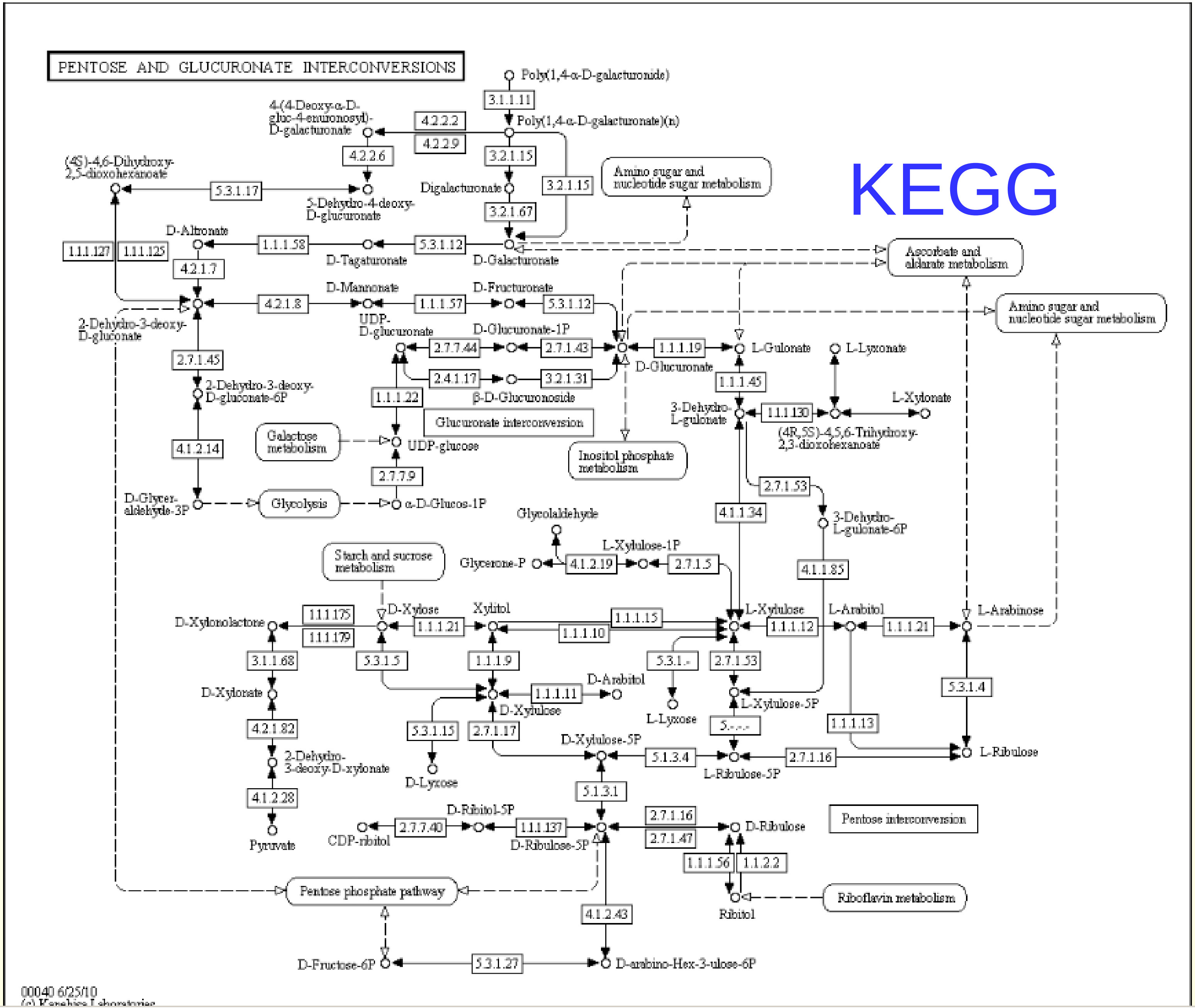
Limitations of Gene Set Enrichment Analysis

- Many possible gene sets – diseases, molecular function, biological process, cellular compartment, pathways...
- Gene sets are heavily overlapping; need to sort through lists of enriched gene sets!
- “Bags of genes” obscure regulatory relationships among them.

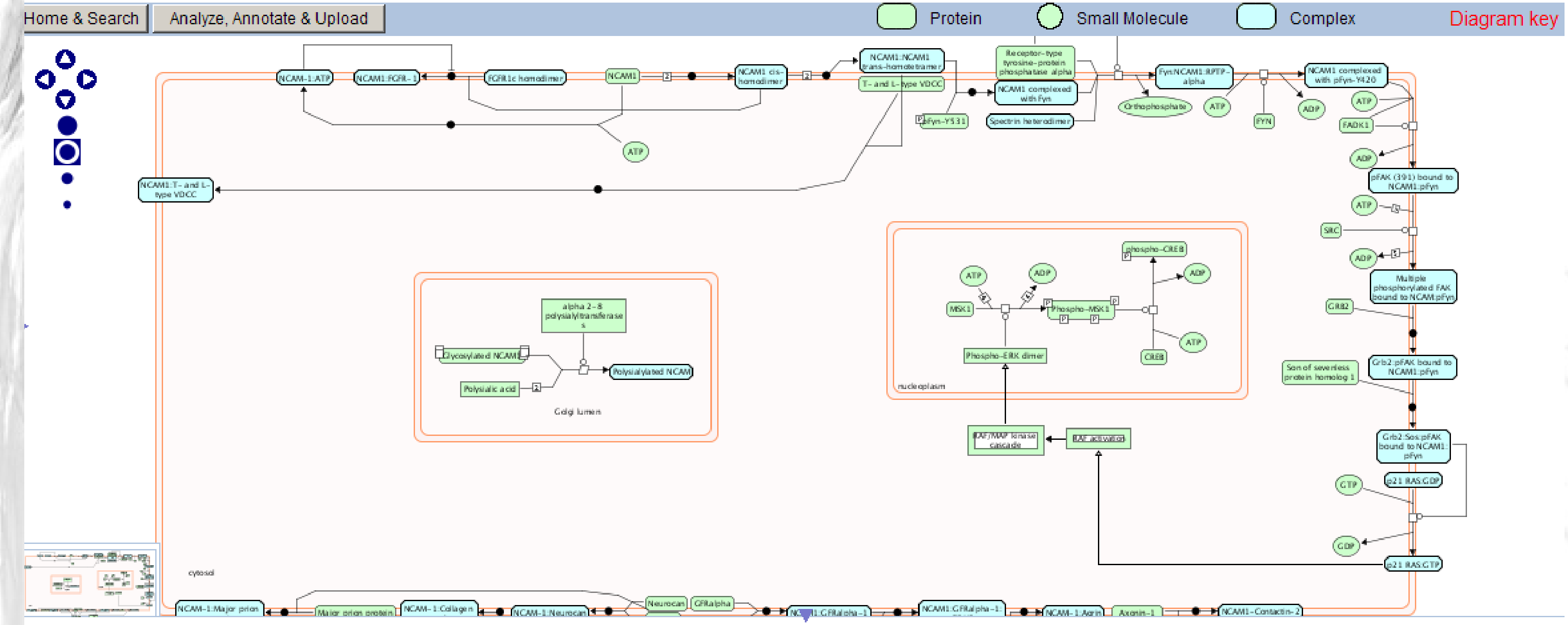
Pathway Databases

- Advantages:
  - Usually curated.
  - Biochemical view of biological processes.
  - Cause and effect captured.
  - Human-interpretable visualizations.
- Disadvantages:
  - Sparse coverage of genome.
  - Different databases disagree on boundaries of pathways.

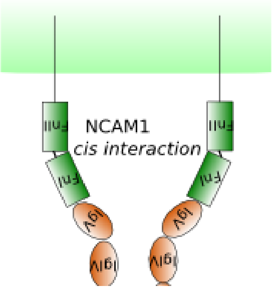
KEGG


PENTOSE AND GLUCURONATE INTERCONVERSIONS
Poly(1,4-α-D-galacturonide)
3.1.1.11
Poly(1,4-α-D-galacturonate)(n)
4-(4-Deoxy-α-D-gluc-4-enuronosyl)-D-galacturonate
4.2.2.2
4.2.2.9
4.2.2.6
3.2.1.15
(4S)-4,6-Dihydroxy-2,5-dioxohexanoate
5.3.1.17
5-Dehydro-4-deoxy-D-glucuronate
Digalacturonate
3.2.1.15
3.2.1.67
Amino sugar and nucleotide sugar metabolism
D-Altronate
1.1.1.58
5.3.1.12
1.1.1.127
1.1.1.125
4.2.1.7
D-Tagaturonate
D-Galacturonate
Ascorbate and aldarate metabolism
D-Mannonate
D-Fructuronate
4.2.1.8
1.1.1.57
5.3.1.12
Amino sugar and nucleotide sugar metabolism
2-Dehydro-3-deoxy-D-gluconate
UDP-D-glucuronate
D-Glucuronate-1P
2.7.7.44
2.7.1.43
1.1.1.19
L-Gulonate
L-Lyxonate
D-Glucuronate
2.7.1.45
2.4.1.17
3.2.1.31
1.1.1.45
2-Dehydro-3-deoxy-D-gluconate-6P
1.1.1.22
β-D-Glucuronoside
3-Dehydro-L-gulonate
1.1.1.130
L-Xylonate
Glucuronate interconversion
(4R,5S)-4,5,6-Trihydroxy-2,3-dioxohexanoate
Galactose metabolism
UDP-glucose
4.1.2.14
Inositol phosphate metabolism
2.7.7.9
2.7.1.53
D-Glyceraldehyde-3P
Glycolysis
α-D-Glucose-1P
Glycolaldehyde
4.1.1.34
3-Dehydro-L-gulonate-6P
L-Xylulose-1P
Starch and sucrose metabolism
Glycerone-P
4.1.2.19
2.7.1.5
4.1.1.85
D-Xylose
Xylitol
1.1.1.15
L-Xylulose
L-Arabitol
L-Arabinose
D-Xylonolactone
1.1.1.175
1.1.1.21
1.1.1.10
1.1.1.12
1.1.1.21
1.1.1.179
3.1.1.68
5.3.1.5
1.1.1.9
5.3.1.-
2.7.1.53
D-Arabitol
5.3.1.4
D-Xylonate
1.1.1.11
D-Xylulose
L-Xylulose-5P
L-Lyxose
4.2.1.82
5.3.1.15
2.7.1.17
5.-.-.-
1.1.1.13
D-Xylulose-5P
2-Dehydro-3-deoxy-D-xylonate
5.1.3.4
2.7.1.16
L-Ribulose
D-Lyxose
L-Ribulose-5P
5.1.3.1
4.1.2.28
D-Ribitol-5P
2.7.1.16
Pentose interconversion
2.7.7.40
1.1.1.137
D-Ribulose
Pyruvate
CDP-ribitol
D-Ribulose-5P
2.7.1.47
1.1.1.56
1.1.2.2
Pentose phosphate pathway
Riboflavin metabolism
Ribitol
4.1.2.43
D-Fructose-6P
5.3.1.27
D-arabino-Hex-3-ulose-6P
00040 6/25/10
(c) Kanehisa Laboratories

Reactome



NCAM1 mediated intracellular signal transduction is represented in the figure below. The Ig domains in NCAM1 are represented in orange ovals and Fn domains in green squares. The tyrosine residues susceptible to phosphorylation are represented in red circles and their positions are numbered. Phosphorylation is represented by red arrows and dephosphorylation by yellow. Ig, Immunoglobulin domain; Fn, Fibronectin domain; Fyn, Proto-oncogene tyrosine-protein kinase Fyn; FAK, focal adhesion kinase; RPTPalpha, Receptor-type tyrosine-protein phosphatase; Grb2, Growth factor receptor-bound protein 2; SOS, Son of sevenless homolog; Raf, RAF proto-oncogene serine/threonine-protein kinase; MEK, MAPK and ERK kinase; ERK, Extracellular signal-regulated kinase; MSK1, Mitogen and stress activated protein kinase 1; CREB, Cyclic AMP-responsive element-binding protein; CRE, cAMP response elements. [Ditlevsen *et al* 2008]

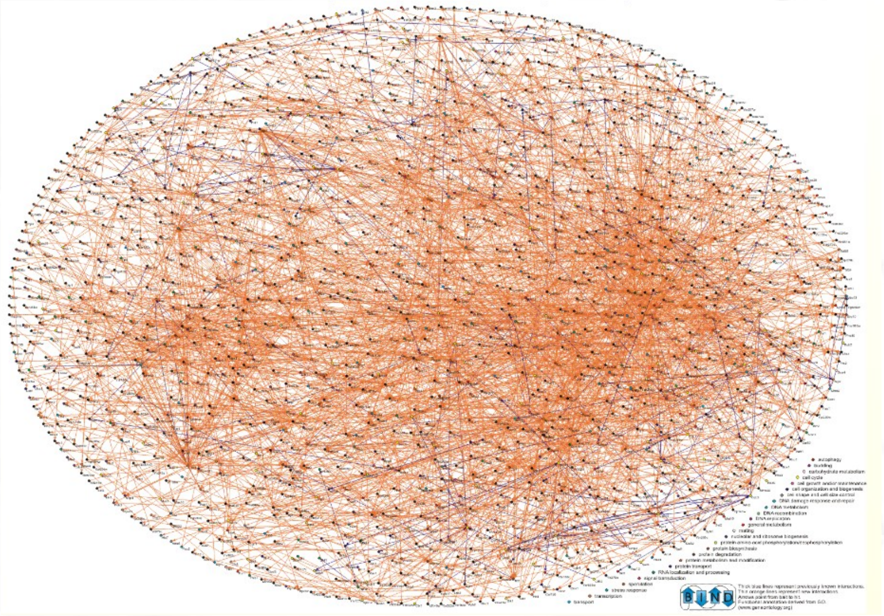


Pathway Colorization

- Main feature offered by all pathway databases.
- Upload a gene list
- Database calculates an enrichment score on each pathway and displays ranked list.
- Browse into pathways of interest; download colorized pictures.

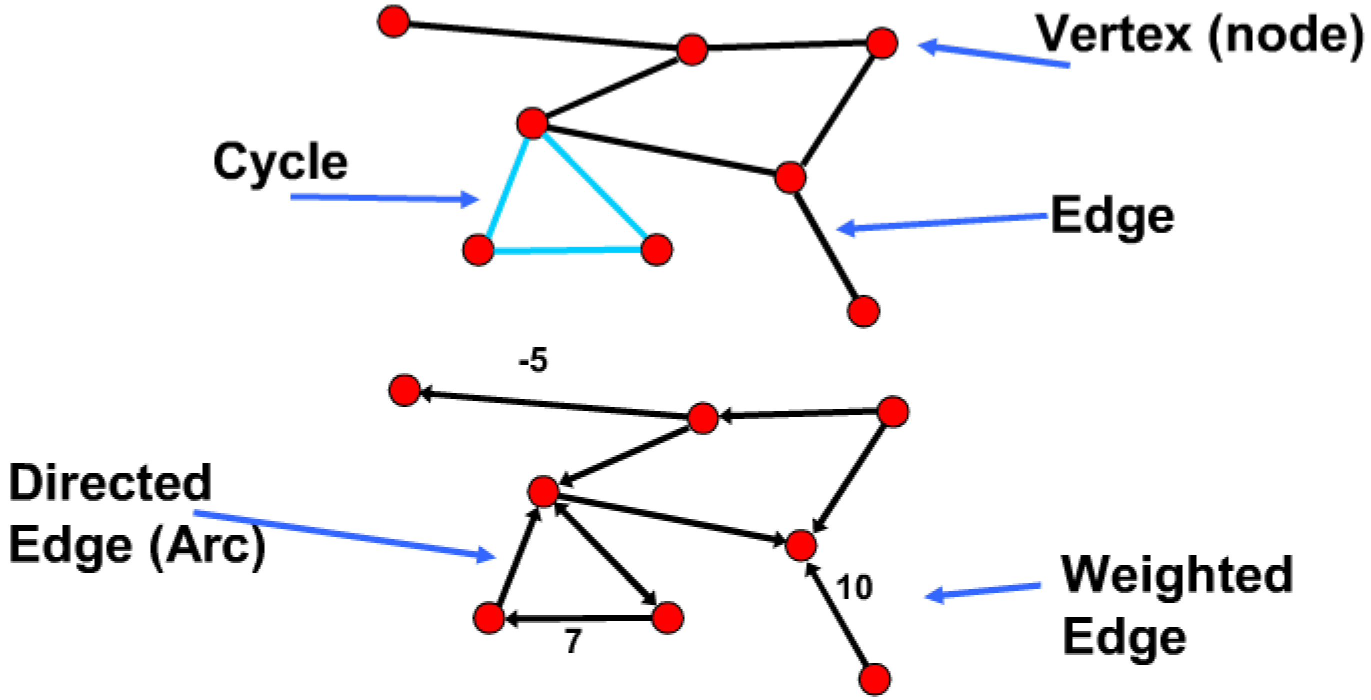
Networks

- Pathways capture only the “well understood” portion of biology.
- Networks cover less well understood relationships:
  - – Genetic interactions
  - – Physical interaction
  - – Coexpression
  - – GO term sharing
  - – Adjacency in pathways


autophagy
budding
carbohydrate metabolism
cell cycle
cell growth and/or maintenance
cell organization and biogenesis
cell shape and cell size control
DNA damage response and repair
DNA metabolism
DNA recombination
DNA replication
general metabolism
mating
nucleolar and ribosome biogenesis
protein amino acid phosphorylation/dephosphorylation
protein biosynthesis
protein degradation
protein metabolism and modification
protein transport
RNA localization and processing
signal transduction
sporulation
stress response
transcription
transport
BIND
Thick blue lines represent previously known interactions.
Thin orange lines represent new interactions.
Arrows point from bait to hit.
Functional annotation derived from GO.
(www.geneontology.org)

Networks

- E.g. Protein-protein interaction networks
- Useful if we don't know pathways
  - Could discover new pathways



Vertex (node)

Cycle

Edge

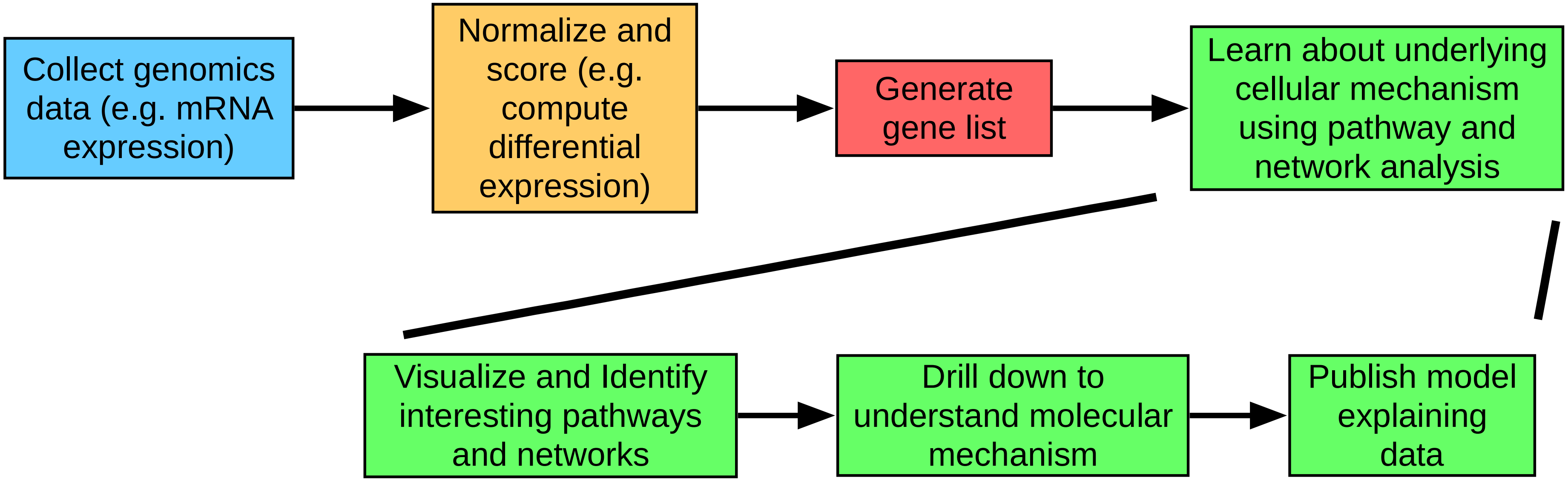
Directed Edge (Arc)

Weighted Edge

Mapping Biology to a Network

- A simple mapping: Protein-protein interactions
  - one protein/node, one interaction/edge
- Edges can represent other relationships
  - Physical e.g. protein-protein interaction
  - Regulatory e.g. kinase activates target
  - Genetic e.g. epistasis
  - Similarity e.g. protein sequence similarity
- Critical: understand the mapping for network analysis

Network Analysis Workflow



Collect genomics data (e.g. mRNA expression)

Normalize and score (e.g. compute differential expression)

Generate gene list

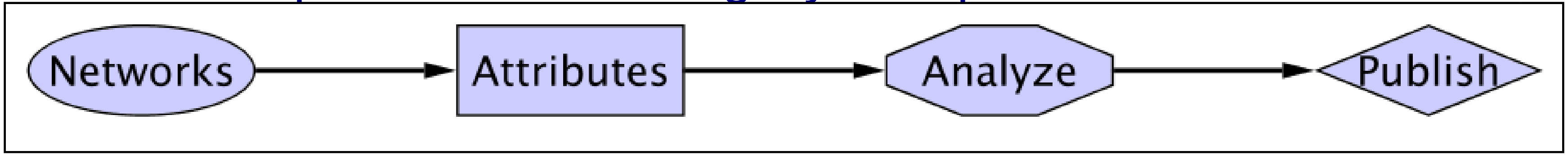
Learn about underlying cellular mechanism using pathway and network analysis

Visualize and Identify interesting pathways and networks

Drill down to understand molecular mechanism

Publish model explaining data

- A specific example of this workflow:
  - Cline, et al. “Integration of biological networks and gene expression data using Cytoscape”, Nature Protocols, 2,



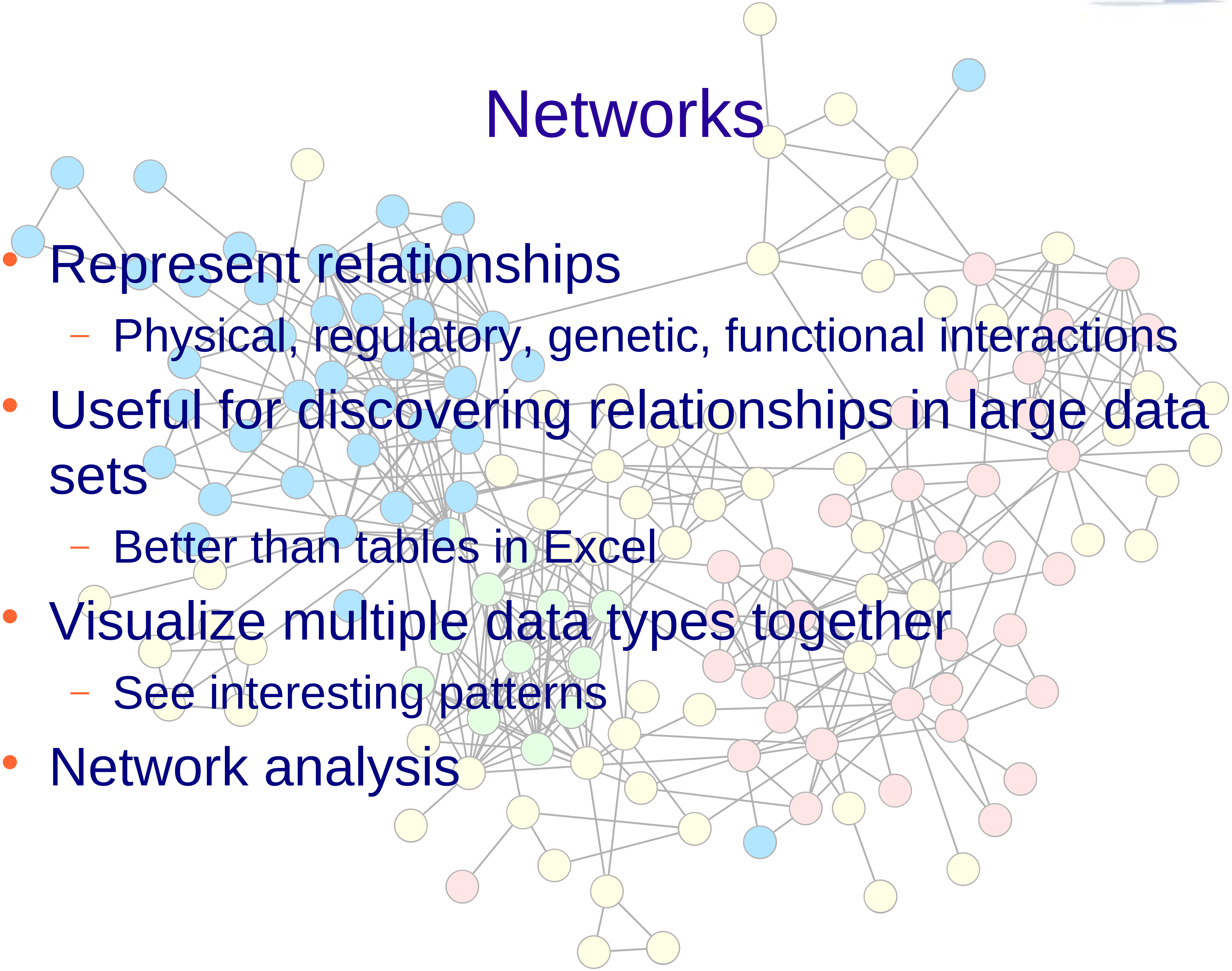
Networks

Attributes

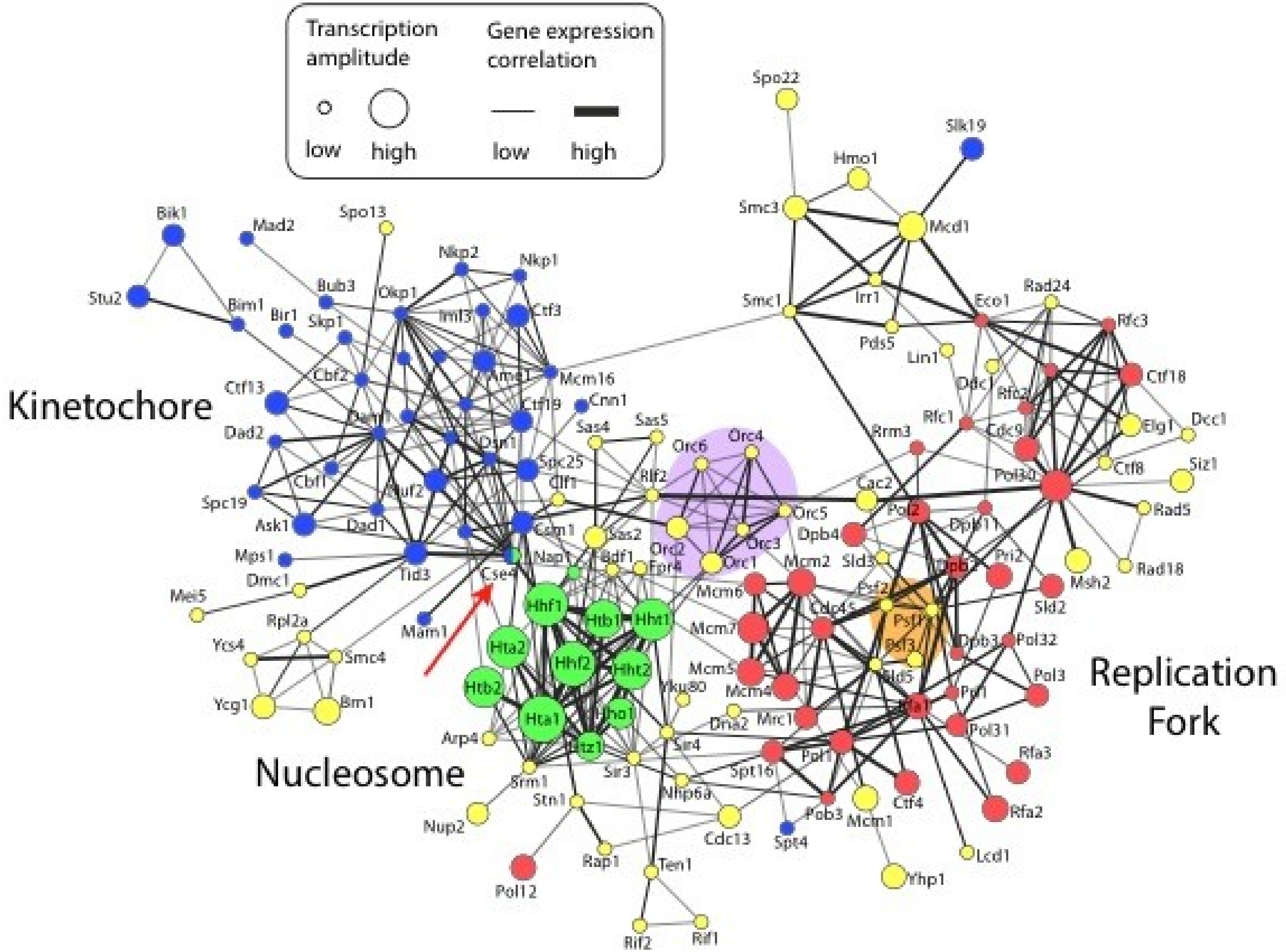
Analyze

Publish

Networks



- Represent relationships
  - Physical, regulatory, genetic, functional interactions
- Useful for discovering relationships in large data sets
  - Better than tables in Excel
- Visualize multiple data types together
  - See interesting patterns
- Network analysis

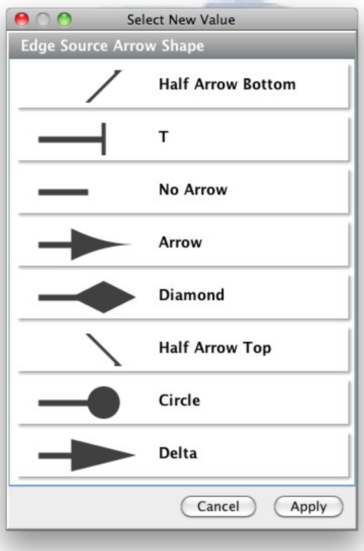

Transcription amplitude
low
high
Gene expression correlation
low
high
Kinetochore
Nucleosome
Replication Fork
Spo22
Slk19
Hmo1
Smc3
Mcd1
Smc1
Irr1
Pds5
Lin1
Eco1
Rad24
Rfc3
Ctf18
Odc1
Rfc2
Rfc1
Cdc9
Elg1
Dcc1
Rrm3
Ctf8
Siz1
Pol30
Rad5
Rad18
Msh2
Bik1
Mad2
Spo13
Nkp2
Nkp1
Stu2
Bim1
Bub3
Okp1
Iml3
Ctf3
Bir1
Skp1
Cbf2
Ame1
Mcm16
Cnn1
Ctf13
Ctf19
Dad2
Dam1
Dsn1
Spc25
Cbf1
Nuf2
Ctf1
Spc19
Ask1
Dad1
Cnm1
Mps1
Dmc1
Tid3
Cse4
Nap1
Mei5
Mam1
Rpl2a
Ycs4
Smc4
Ycg1
Brn1
Sas4
Sas5
Orc6
Orc4
Rif2
Orc5
Sas2
Orc2
Orc3
Orc1
Bdf1
Fpr4
Cac2
Pol2
Dpb4
Dpb11
Sld3
Dpb2
Pri2
Sld2
Mcm2
Mcm6
Psf2
Psf1
Psf3
Cdc45
Mcm7
Dpb3
Pol32
Mcm5
Mcm4
Sld5
Pol3
Pri1
Pfa1
Hhf1
Htb1
Hht1
Hta2
Hhf2
Hht2
Htb2
Hta1
Hho1
Htz1
Arp4
Srm1
Yku80
Sir4
Dna2
Mrc1
Pol1
Pol31
Sir3
Spt16
Nhp6a
Stn1
Nup2
Rap1
Cdc13
Spt4
Pob3
Mcm1
Ctf4
Rfa3
Rfa2
Lcd1
Yhp1
Pol12
Ten1
Rif2
Rif1

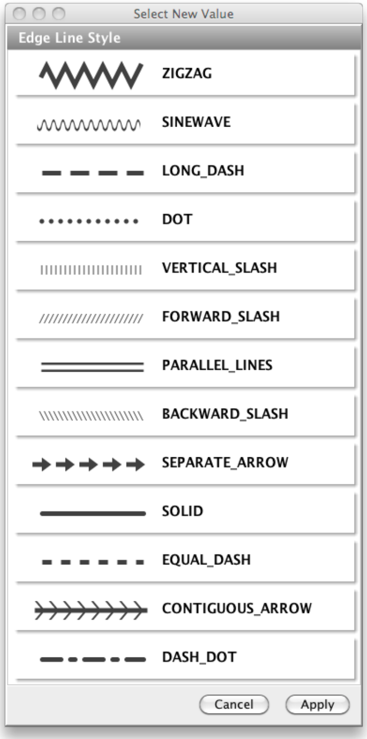
Summary

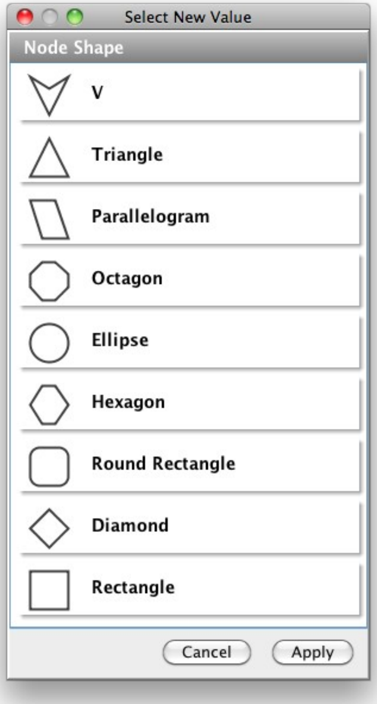
- Networks are useful for seeing relationships in large data sets
- Important to understand what the nodes and edges mean
- Important to define the biological question - know what you want to do with your gene list or network
- Many methods available for gene list and network analysis
  - Good to determine your question and search for a solution
  - Or get to know many methods and see how they can be applied to your data

Visual Features

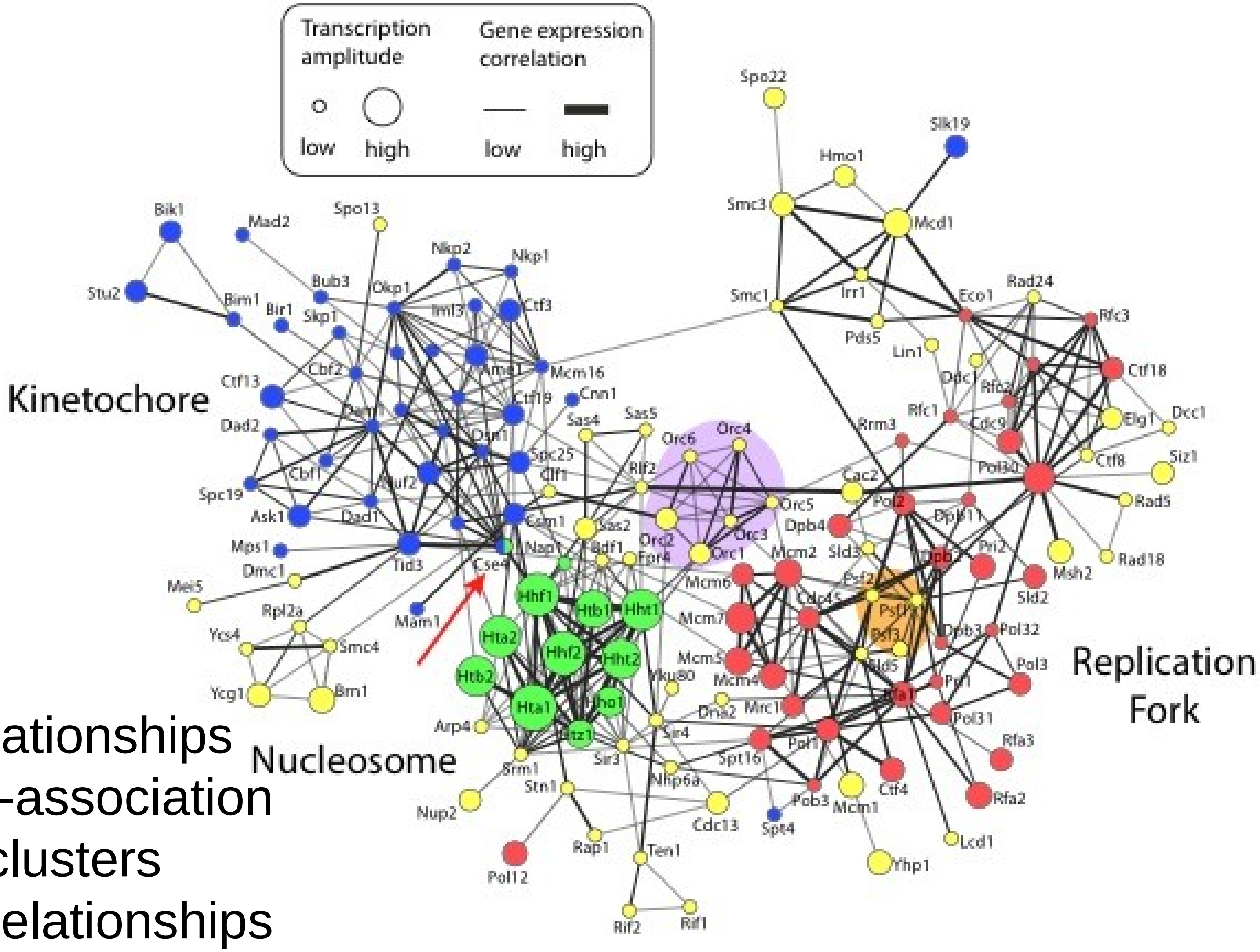
- Node and edge attributes
  - Text (string), integer, float, Boolean, list
  - E.g. represent gene, interaction attributes
- Visual attributes
  - Node, edge visual properties
  - Colour, shape, size, borders, opacity...





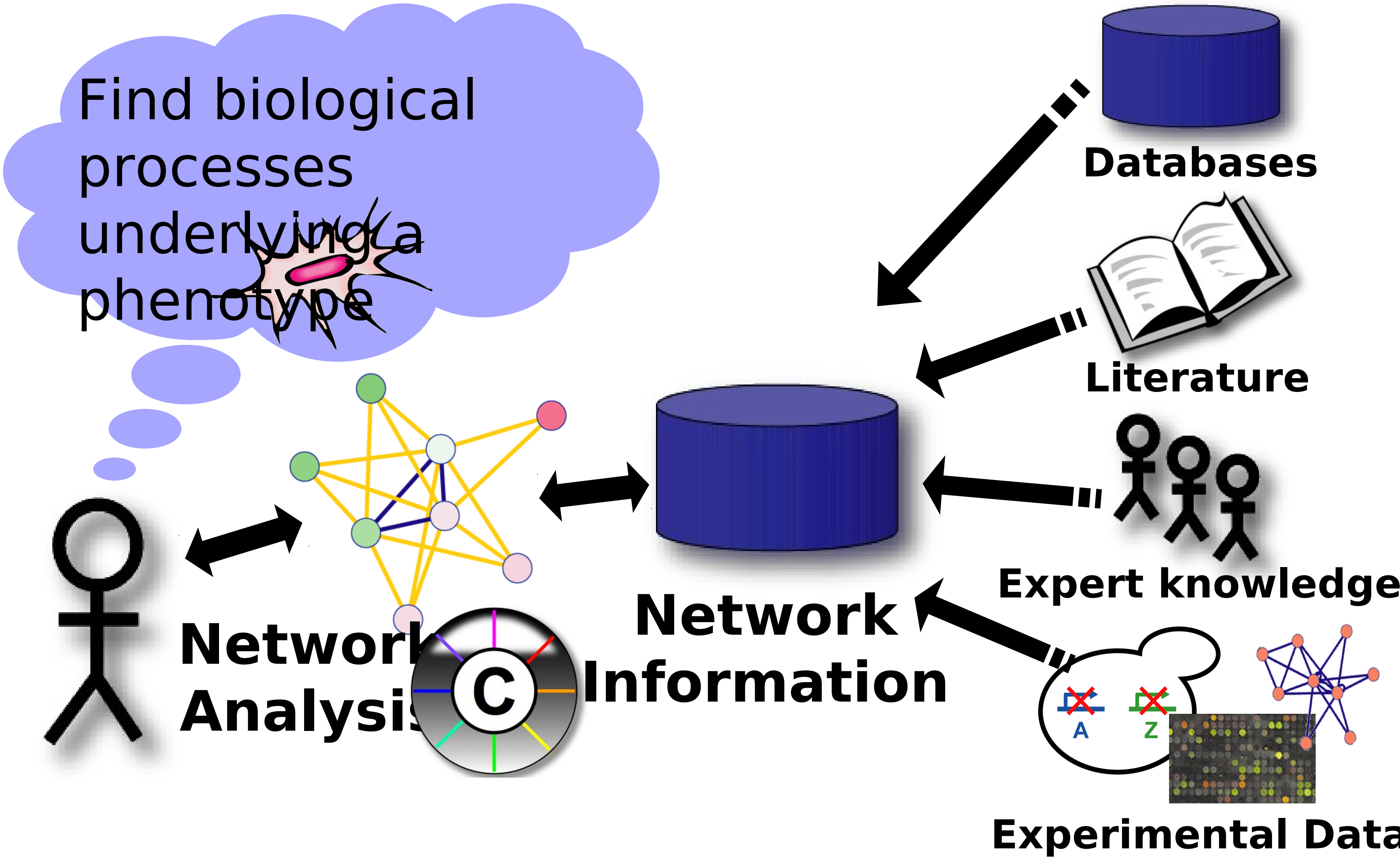


Visually Interpreting a Network



Data relationships
Guilt-by-association
Dense clusters
Global relationships

Network Analysis using Cytoscape



Find biological processes underlying a phenotype

Databases

Literature

Expert knowledge

Experimental Data

Network Information

Network Analysis

Active Community

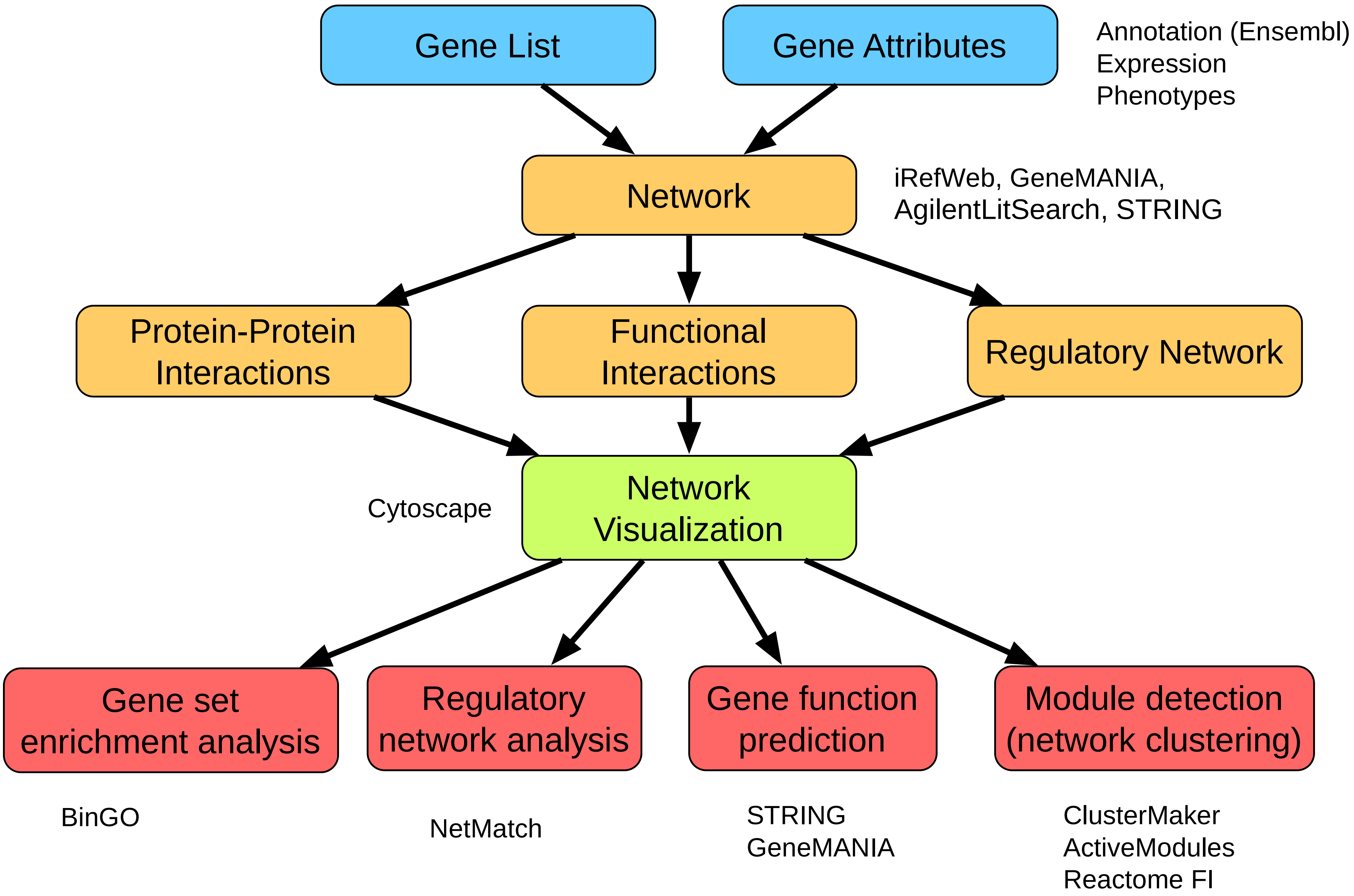
http://www.cytoscape.org

- 10,000s users, >5000 downloads/month
- Help
  - Documentation, data sets
  - Mailing lists
  - **http://tutorials.cytoscape.org**
- Annual Conference: TBD, North America 2014
- >200 Apps Extend Functionality
  - Build your own, requires programming

Cline MS et al. Integration of biological networks and gene expression data using Cytoscape Nat Protoc. 2007;2(10):2366-82

- Cytoscape is a useful, free software tool for network visualization and analysis
- Provides basic network manipulation features
- Apps are available to extend the functionality

Gene List and Network Analysis Overview



Gene List

Gene Attributes

Annotation (Ensembl)
Expression
Phenotypes

Network

iRefWeb, GeneMANIA,
AgilentLitSearch, STRING

Protein-Protein Interactions

Functional Interactions

Regulatory Network

Cytoscape

Network Visualization

Gene set enrichment analysis

Regulatory network analysis

Gene function prediction

Module detection (network clustering)

BinGO

NetMatch

STRING
GeneMANIA

ClusterMaker
ActiveModules
Reactome FI