Pathway and Network Analysis of Genomics Data

June 2, 2011

Leloir, Buenos Aires

Gary Bader

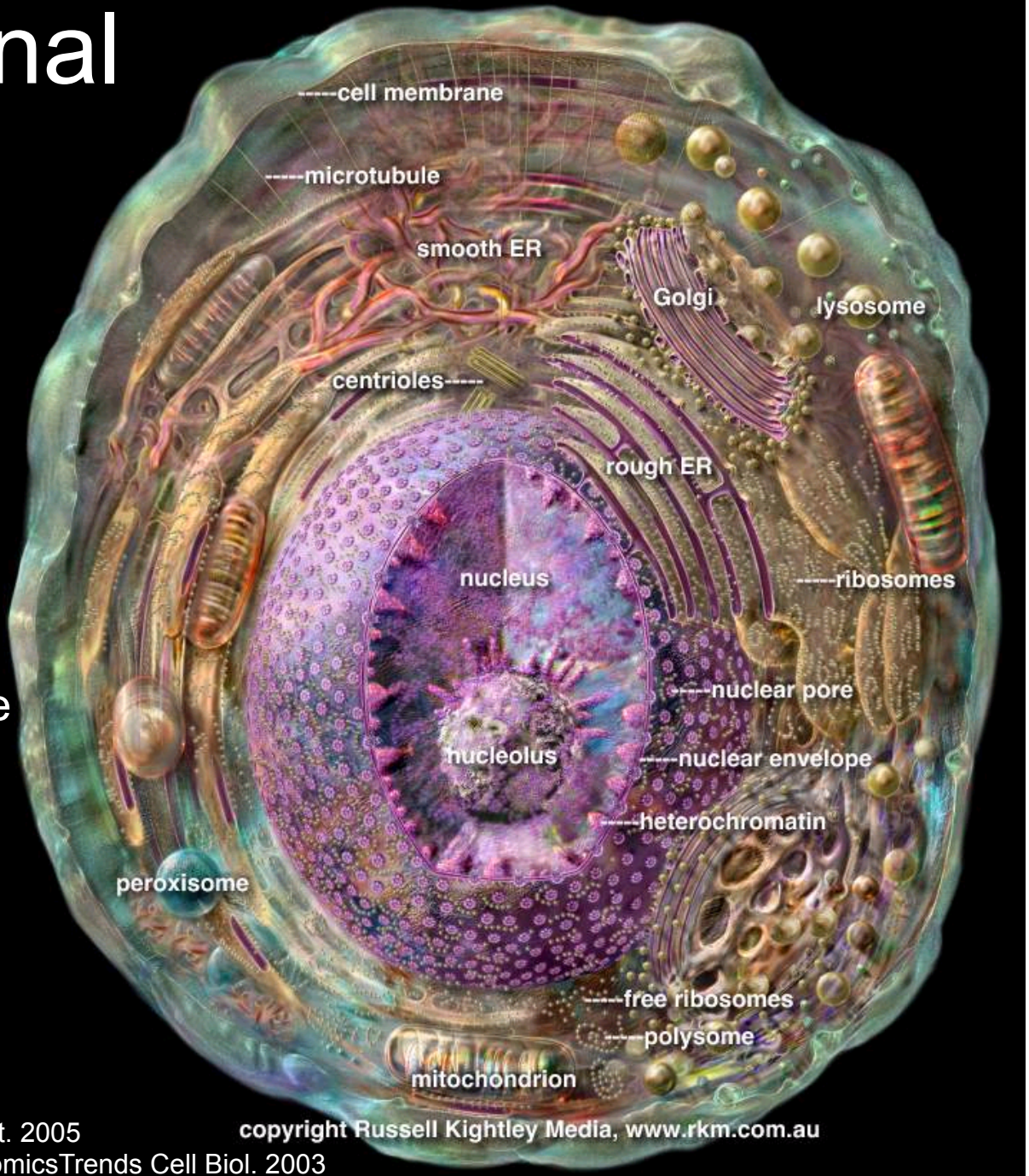
Donnelly Centre
for Cellular + Biomolecular Research

UNIVERSITY OF TORONTO

Computational Cell Map

Read map to understand
- Cell processes
- Gene function
- Disease effects
- Map evolution

Map the cell
- Predict map from genome
- Multiple perturbation mapping
- Active cell map
- Map visualization and analysis software



Cary MP et al. Pathway information… FEBS Lett. 2005
Bader GD et al. Functional genomics and proteomicsTrends Cell Biol. 2003

Outline

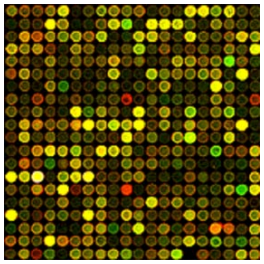
- Introduction, Gene lists and annotations, Pathway analysis using enrichment analysis
- Network visualization and analysis, Gene function prediction, Pathways
- Lunch
- Network Visualization and Analysis using Cytoscape (lab)
- Gene function prediction using GeneMANIA (lab)

Alex Pico
Allan Kuchinsky
Scooter Morris
Piet Molenaar
Gary Bader
Tero Aittokallio
Boris Steipe
Quaid Morris
Daniele Merico
Canadian Bioinformatics Workshops

Gene list introduction

Interpreting Gene Lists

- My cool new screen worked and produced 1000 hits! …Now what?
- Genome-Scale Analysis (Omics)
  - Genomics, Proteomics
- Tell me what's interesting about these genes



Ranking or clustering

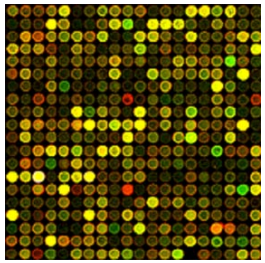
GNAQ
GNAS
DGKZ
GUCY1A3
PDE4B
PDE4D
ATP2A2
ATP2A3
NOS1
CNN1
GSTO1
NOS3
CNN2
MYLK2
CALD1
ACTA1
MYL2

?

GenMAPP.org

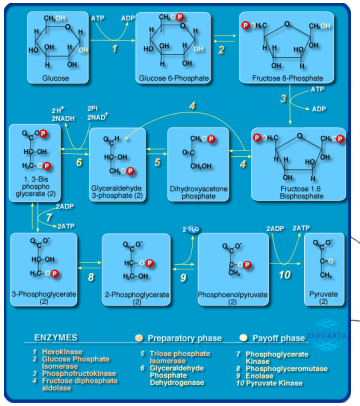
Interpreting Gene Lists

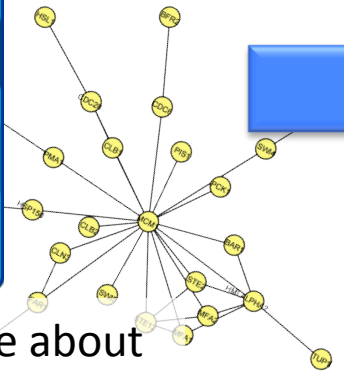
- My cool new screen worked and produced 1000 hits! …Now what?
- Genome-Scale Analysis (Omics)
  - Genomics, Proteomics
- Tell me what's interesting about these genes
  - Are they enriched in known pathways, complexes, functions



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!

Autism Spectrum Disorder (ASD)

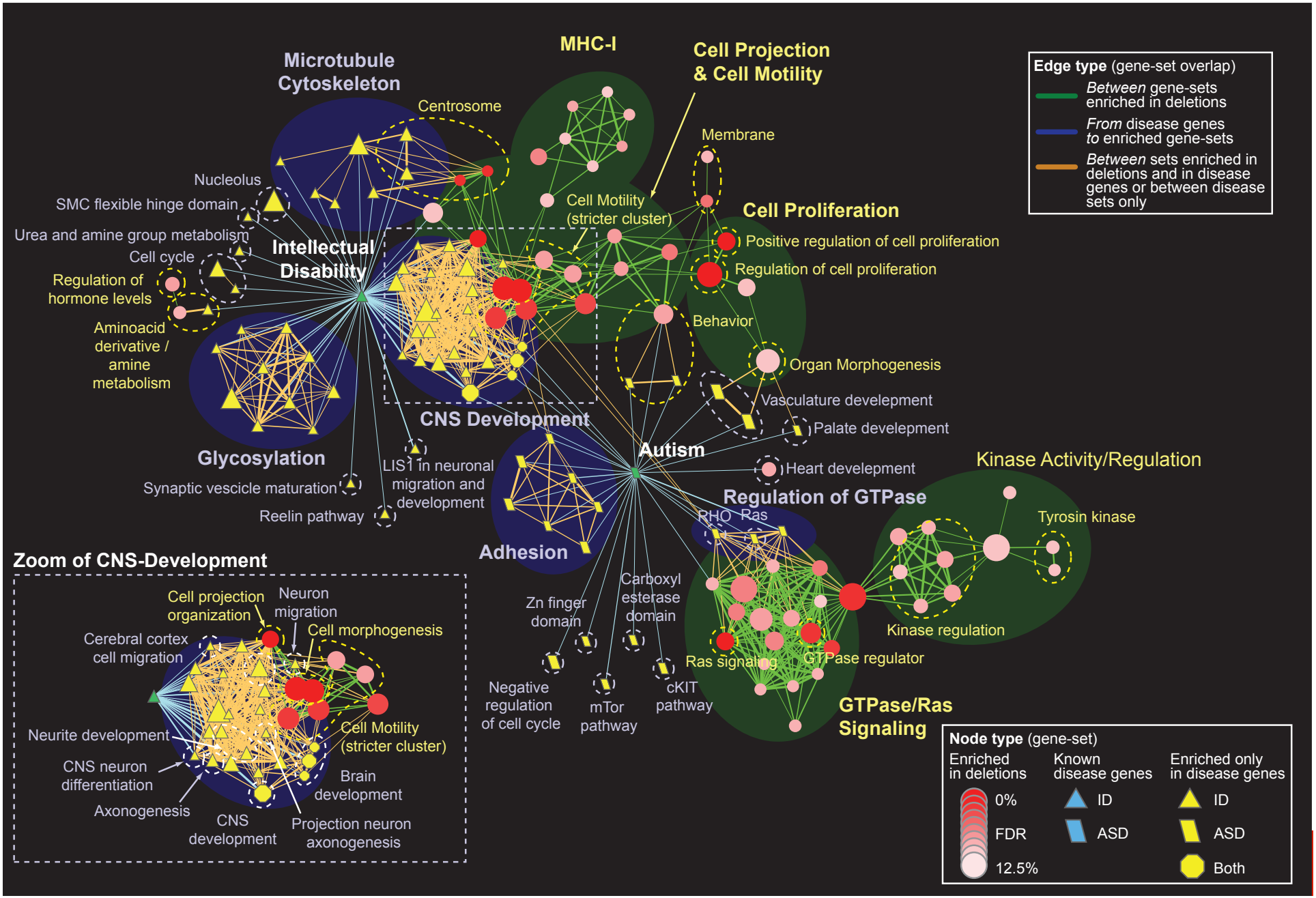
- Genetics
  - highly heritable
    - monozygotic twin concordance 60-90%
    - dizygotic twin concordance 0-10%
      (depending on the stringency of diagnosis)
  - known genetics:
    - 5-15% rare single-gene disorders and chromosomal re-arrangements
    - de-novo CNV previously reported in 5-10% of ASD cases
    - GWA (Genome-wide Association Studies) have been able to explain only a small amount of heritability

Pinto et al. Functional impact of global rare copy number variation in autism spectrum disorders. Nature. 2010 Jun 9.

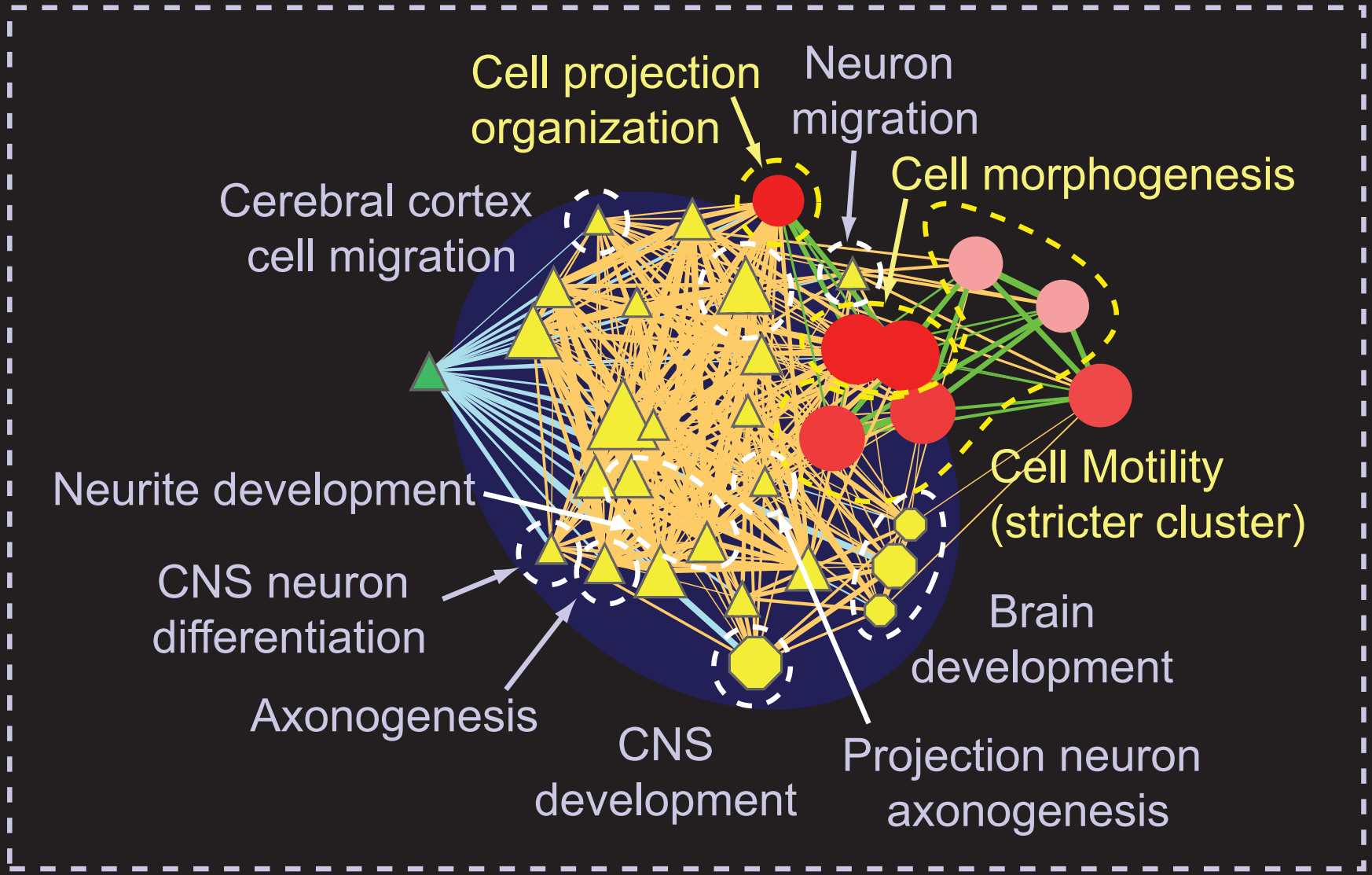
Rare copy number variants in ASD

- Rare Copy Number Variation screening (Del, Dup)
  - 889 Case and 1146 Ctrl (European Ancestry)
  - Illumina Infinium 1M-single SNP
  - high quality rare CNV (90% PCR validation)
    - identification by three algorithms required for detection
      - QuantiSNP, iPattern, PennCNV
    - frequency < 1%, length > 30 kb
- Results
  - average CNV size: 182.7 kb, median CNVs per individual: 2
  - > 5.7% ASD individuals carry at least one de-novo CNV
  - Top ~10 genes in CNVs associated to ASD

Pathways Enriched in Autism Spectrum Disorder



Zoom of CNS-Development



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)

Other examples?

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)

Biological Questions

- Step 1: What do you want to accomplish with your list (hopefully part of experiment design! ☺)
  - 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

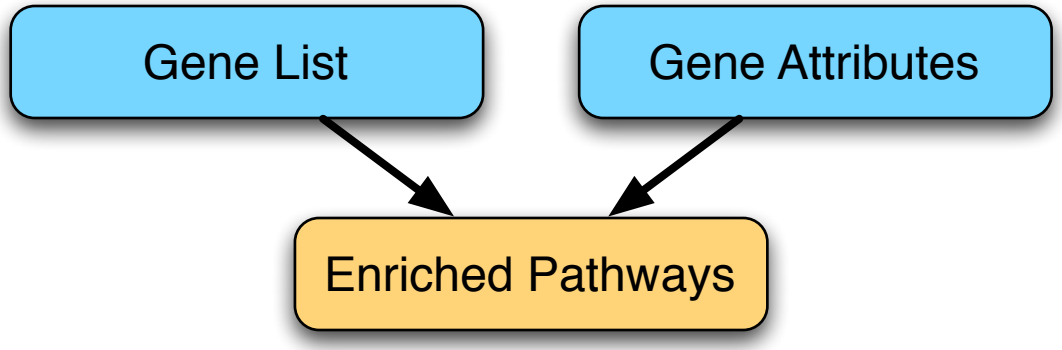
- Computational analysis methods we will cover
  - Pathway enrichment analysis: summarize and compare
  - Network visualization
  - Network analysis: predict gene function, find new pathway members, identify functional modules (new pathways)

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

Pathway Enrichment Analysis



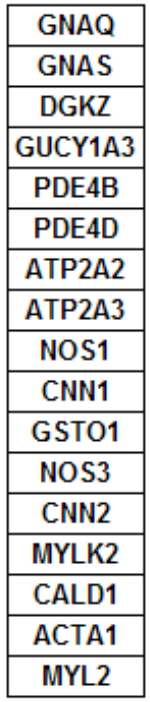
Annotation
Expression
Phenotypes

DAVID, GSEA

- 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.

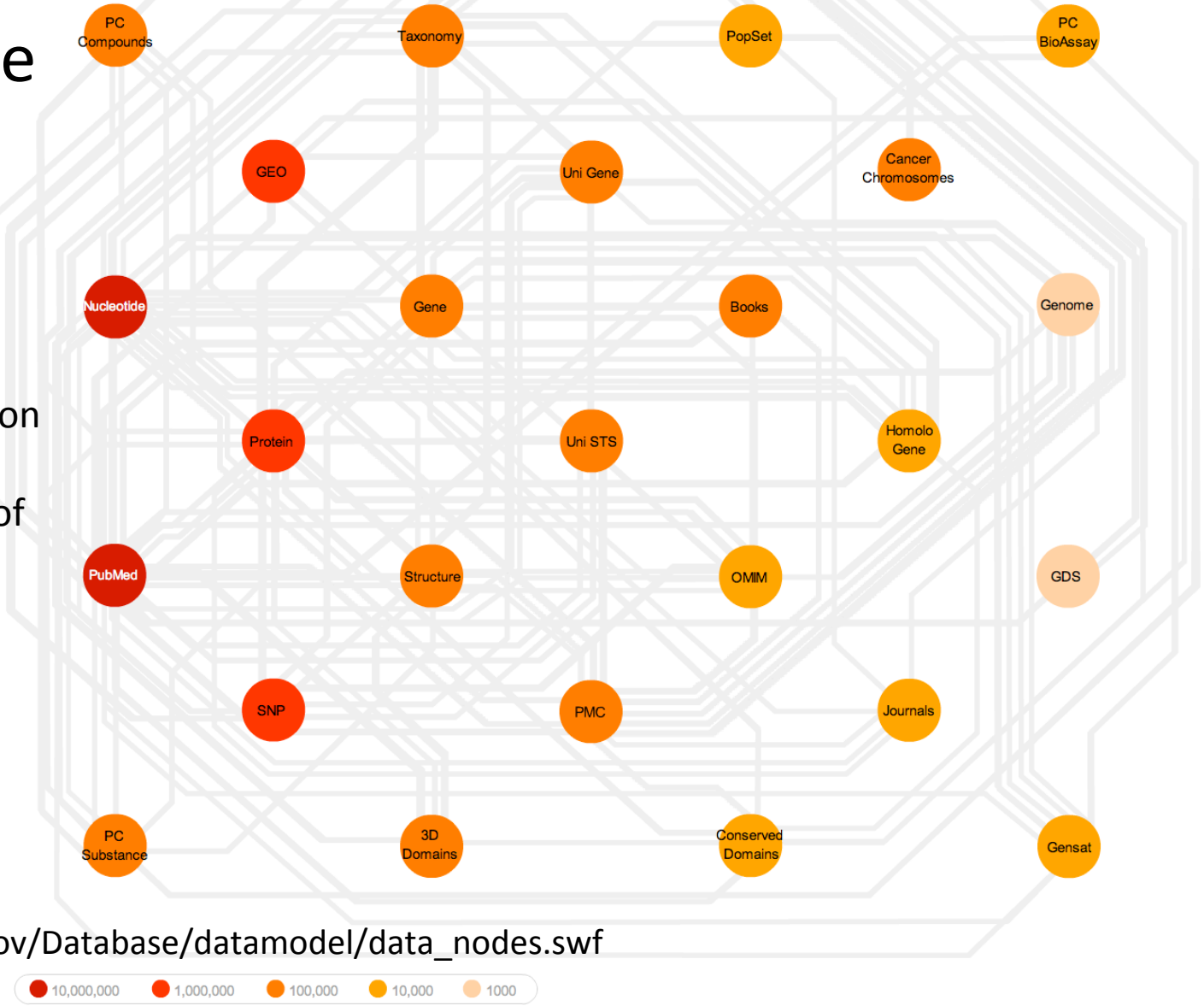


For your information

NCBI Database Links

NCBI:
U.S. National Center for Biotechnology Information

Part of National Library of Medicine (NLM)



http://www.ncbi.nlm.nih.gov/Database/datamodel/data_nodes.swf

For your information

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

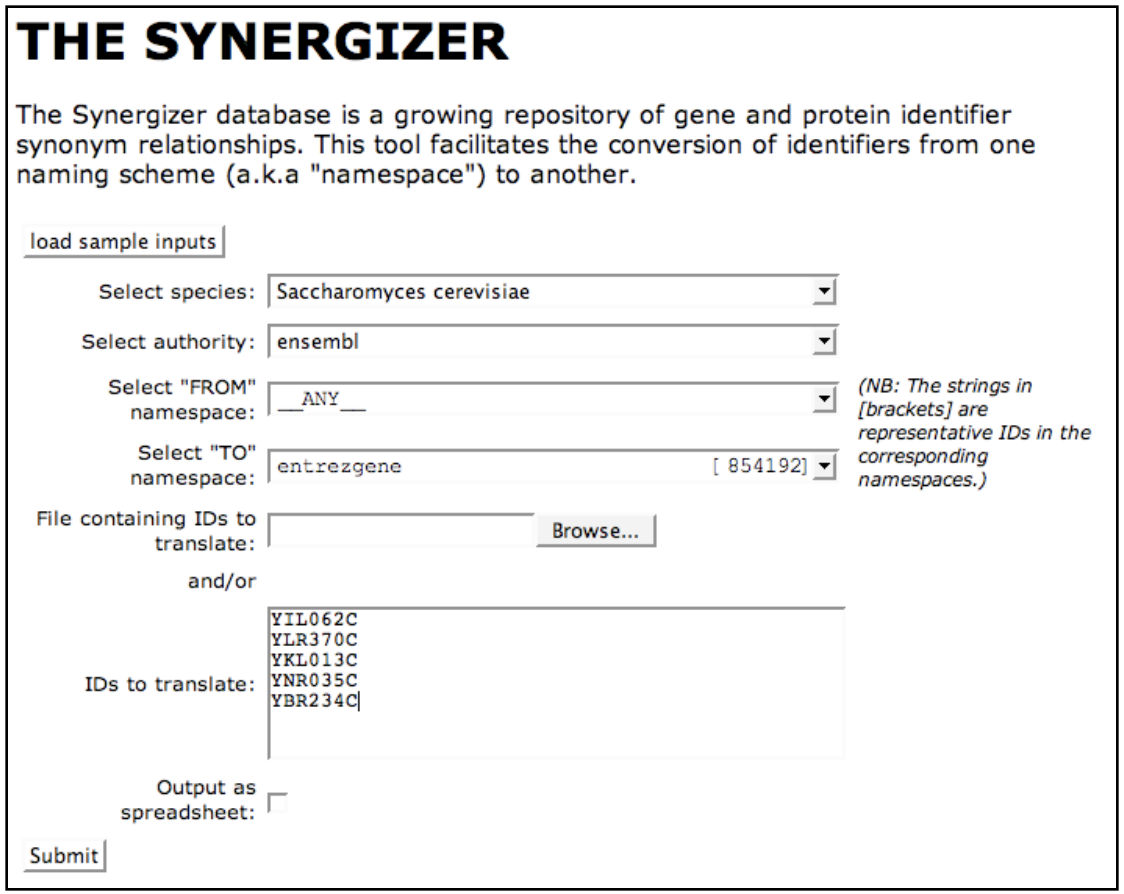
- So many IDs!
  - Software tools recognize only a handful
  - May need to map from your gene list IDs to standard IDs
- Four main uses
  - Searching for a favorite gene name
  - Link to related resources
  - Identifier translation
    - E.g. Proteins to genes, Affy ID to Entrez Gene
  - Merging data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

ID Challenges

- Avoid errors: map IDs correctly
- Gene name ambiguity – not a good ID
  - e.g. FLJ92943, LFS1, TRP53, p53
  - Better to use the standard gene symbol: TP53
- Excel error-introduction
  - OCT4 is changed to October-4
    - format cells as 'text' before pasting
- Problems reaching 100% coverage
  - E.g. due to version issues
  - Use multiple sources to increase coverage

Zeeberg BR et al. Mistaken identifiers: gene name errors can be introduced inadvertently when using Excel in bioinformatics BMC Bioinformatics. 2004 Jun 23;5:80

ID Mapping Services



THE SYNERGIZER

The Synergizer database is a growing repository of gene and protein identifier synonym relationships. This tool facilitates the conversion of identifiers from one naming scheme (a.k.a "namespace") to another.

load sample inputs

Select species: Saccharomyces cerevisiae

Select authority: ensembl

Select "FROM" namespace: __ANY__

Select "TO" namespace: entrezgene [854192]

(NB: The strings in [brackets] are representative IDs in the corresponding namespaces.)

File containing IDs to translate: Browse...

and/or

IDs to translate:
YIL062C
YLR370C
YKL013C
YNR035C
YBR234C

Output as spreadsheet:

Submit

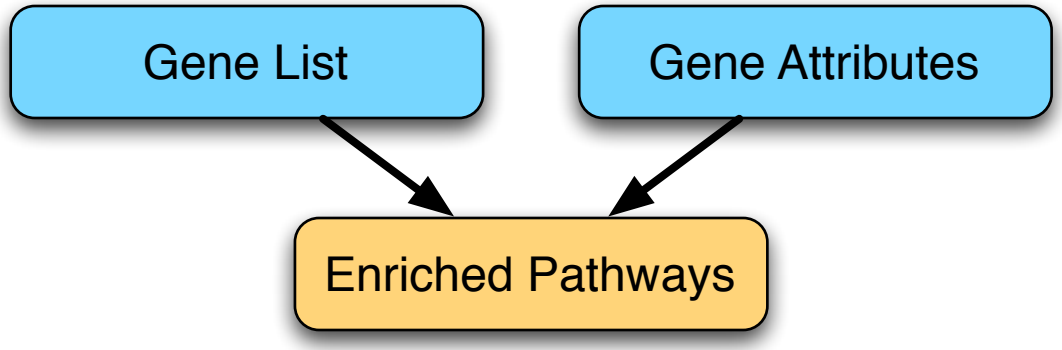
*	entrezgene
YIL062C	854748
YLR370C	851085
YKL013C	853856
YNR035C	855771
YBR234C	852536

- Synergizer
  - http://llama.med.harvard.edu/synergizer/translate/
- Ensembl BioMart
  - http://www.ensembl.org
- PICR (proteins only)
  - http://www.ebi.ac.uk/Tools/picr/

Recommendations

- For proteins and genes
  - (doesn't consider splice forms)
- Map everything to Entrez Gene IDs using a spreadsheet
- If 100% coverage desired, manually curate missing mappings
- Be careful of Excel auto conversions – especially when pasting large gene lists!
  - Remember to format cells as 'text' before pasting

Pathway Enrichment Analysis



Annotation
Expression
Phenotypes

DAVID, GSEA

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

Gene Attributes

- Available in databases
- 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

Gene Attributes

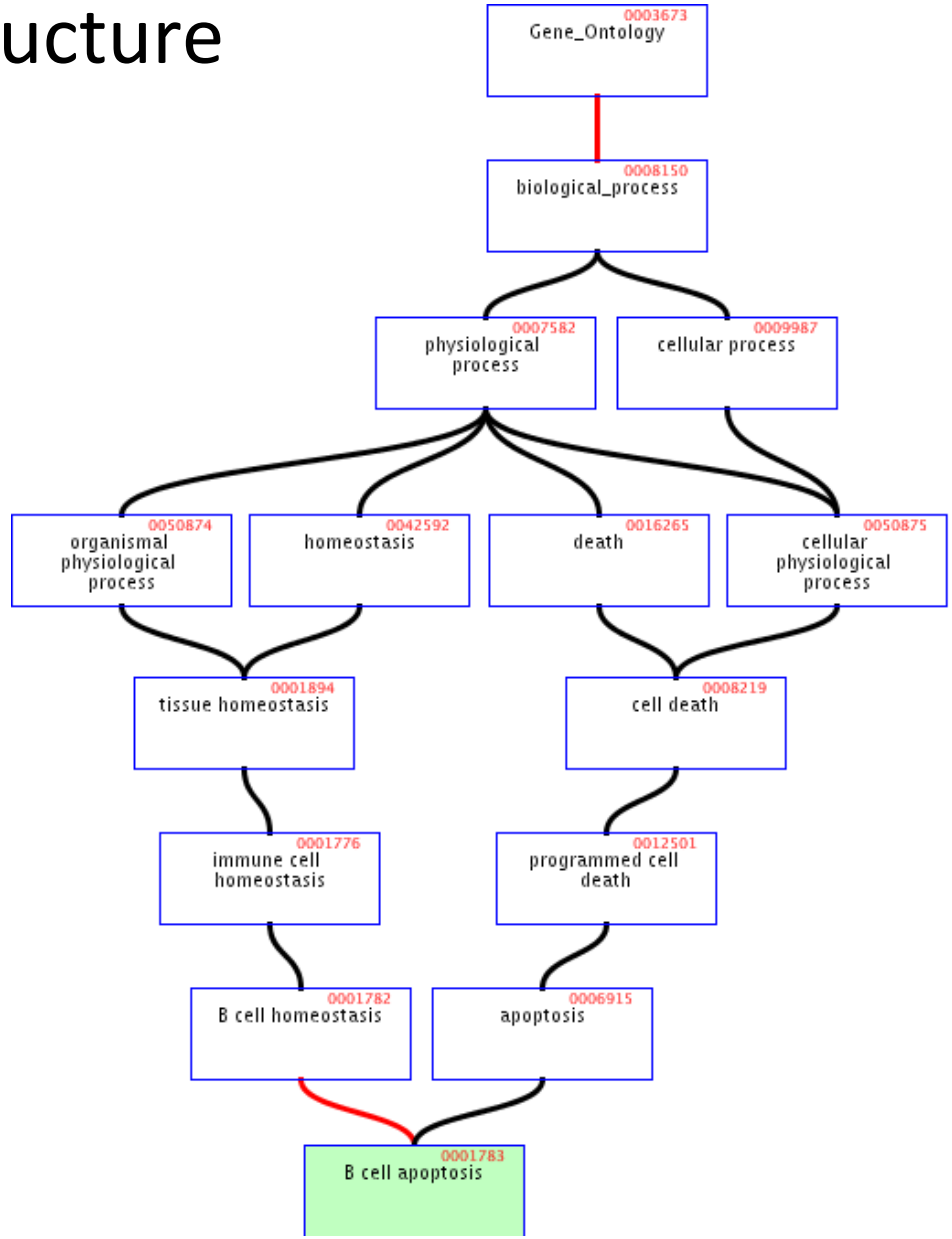
- Available in databases
- Function annotation
  - Biological process, molecular function, cell location
- Chromosome position
- Disease association
- DNA properties
  - TF binding sites, gene structure (intron/exon), SNPs
- Transcript properties
  - Splicing, 3’ UTR, microRNA binding sites
- Protein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

What is the Gene Ontology (GO)?

- Set of biological phrases (terms) which are applied to genes:
  - protein kinase (molecular function)
  - apoptosis (biological process)
  - membrane (cellular component)
- Dictionary: term definitions
- Ontology: A formal system for describing knowledge

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



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
  - 34065 terms, with definitions
    - 20703 biological_process
    - 2824 cellular_component
    - 9029 molecular_function
    - As of April 2011

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)

Annotation Sources

- Manual annotation
  - Curated by scientists
    - High quality
    - Small number (time-consuming to create)
  - Reviewed computational analysis
- Electronic annotation
  - Annotation derived without human validation
    - Computational predictions (accuracy varies)
    - Lower ‘quality’ than manual codes
- Key point: be aware of annotation origin

For your information

Evidence Types

- 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

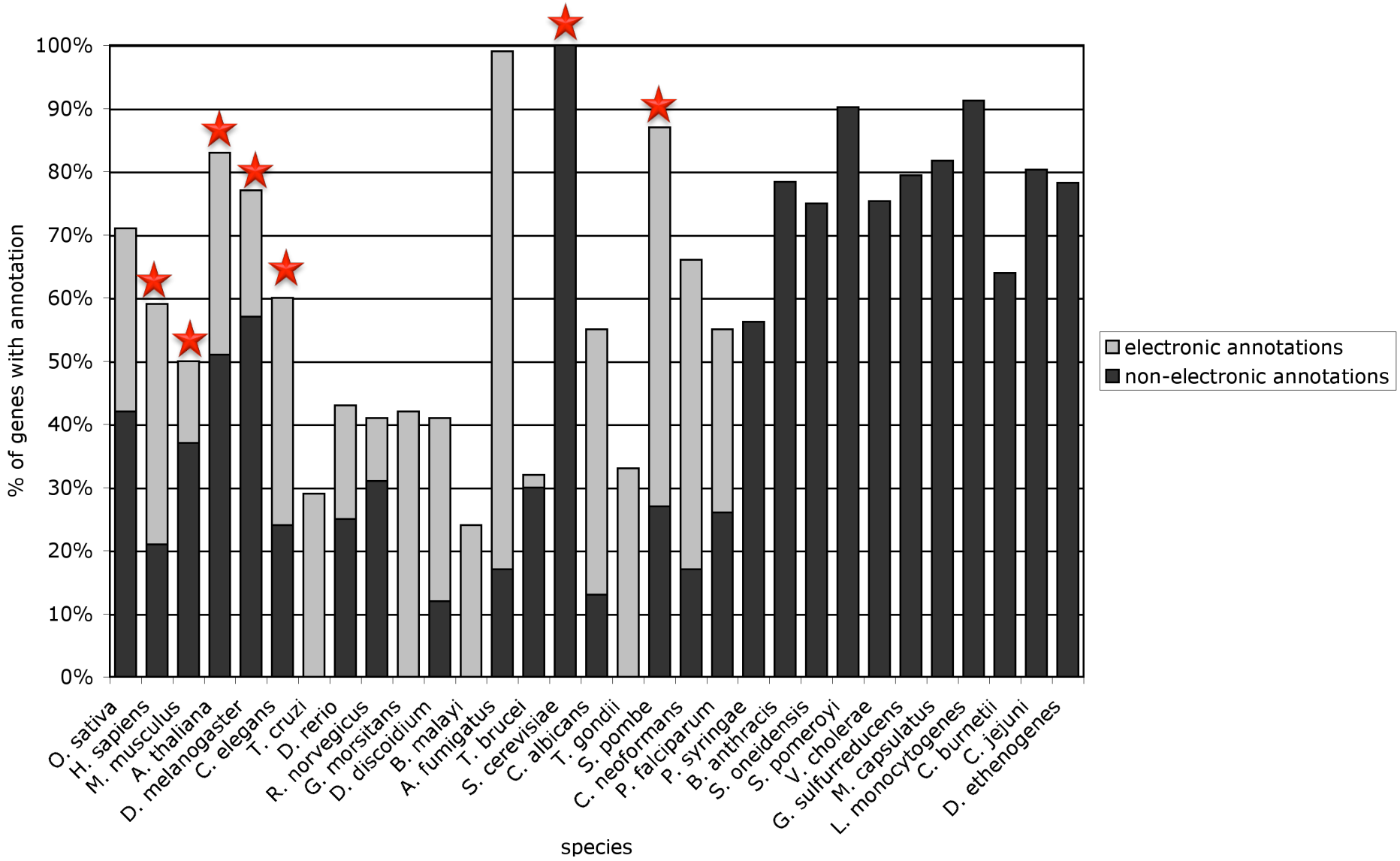
- Computational Analysis Evidence Codes
  - ISS: Inferred from Sequence or Structural Similarity
  - ISO: Inferred from Sequence Orthology
  - ISA: Inferred from Sequence Alignment
  - ISM: Inferred from Sequence Model
  - IGC: Inferred from Genomic Context
  - RCA: inferred from Reviewed Computational 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

- **IEA**: Inferred from electronic annotation

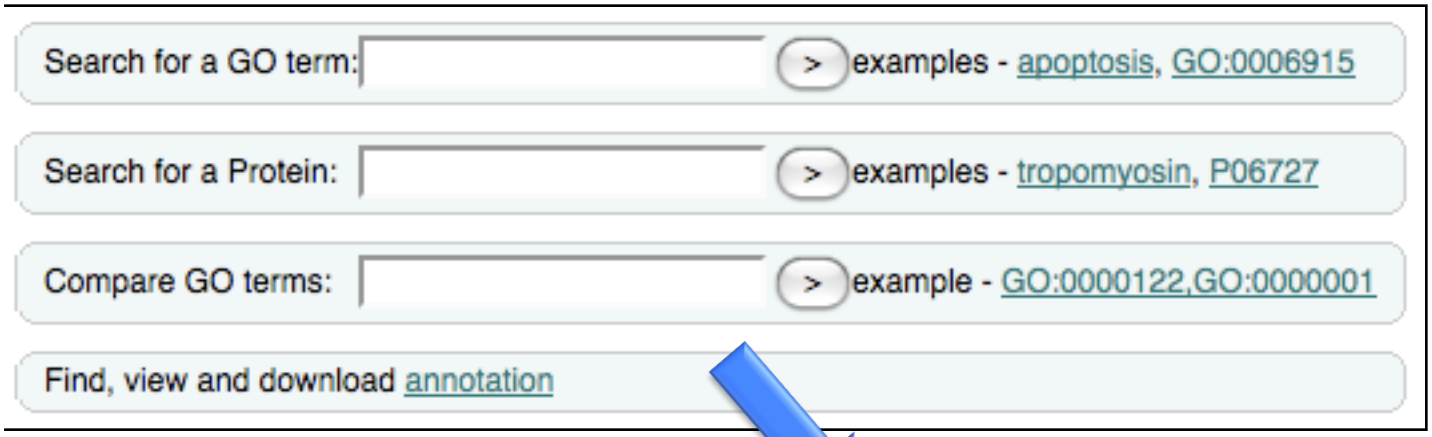
http://www.geneontology.org/GO.evidence.shtml

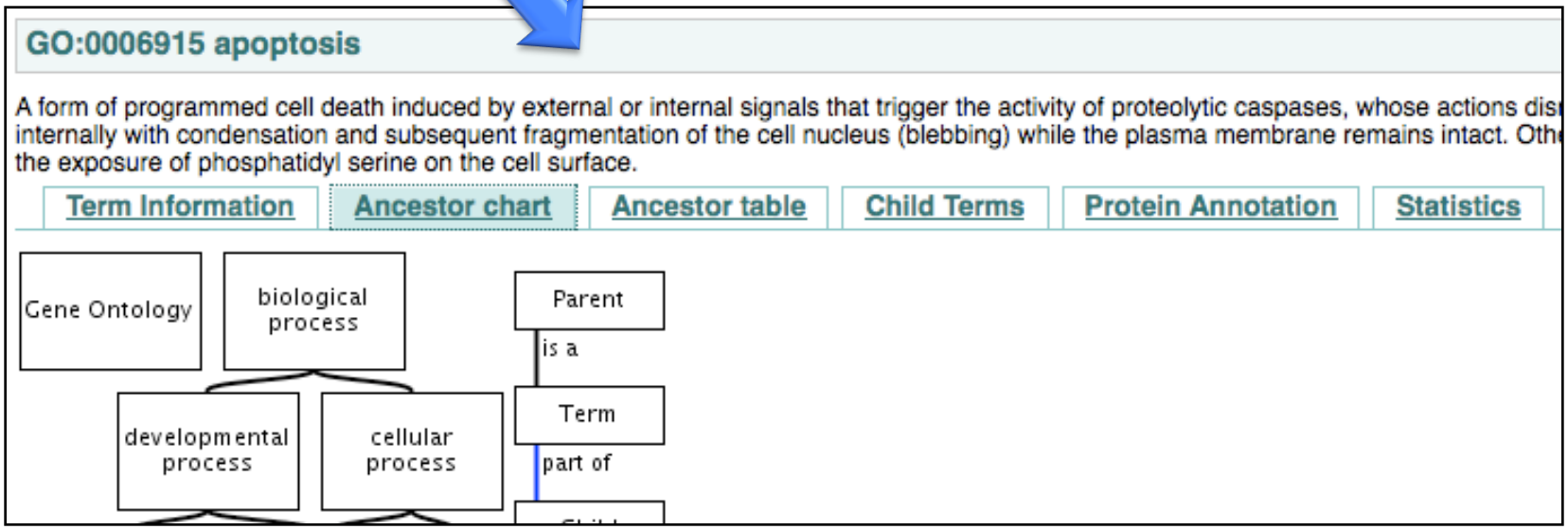
Variable Coverage



Lomax J. Get ready to GO! A biologist's guide to the Gene Ontology. Brief Bioinform. 2005 Sep;6(3):298-304.

Accessing GO: QuickGO





http://www.ebi.ac.uk/ego/

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

Ensembl BioMart

- Convenient access to gene list annotation

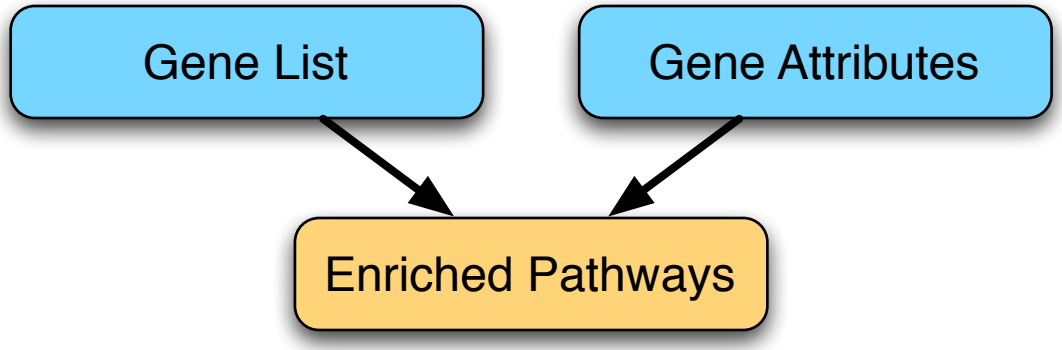
http://www.ensembl.org

Lab: Gene IDs, Attributes and Networks

- Objectives
  - Learn about gene identifiers, Synergizer and BioMart
- Use yeast demo gene list (module1YeastGenes.txt)
- Convert Gene IDs to Entrez Gene: Use Synergizer
- Get GO annotation + evidence codes
  - Use Ensembl BioMart
  - Summarize terms & evidence codes in a table
- Do it again with your own gene list
  - If compatible with covered tools, run the analysis. If not, instructors will recommend tools for you.

Pathway (gene set) enrichment analysis

Pathway Enrichment Analysis



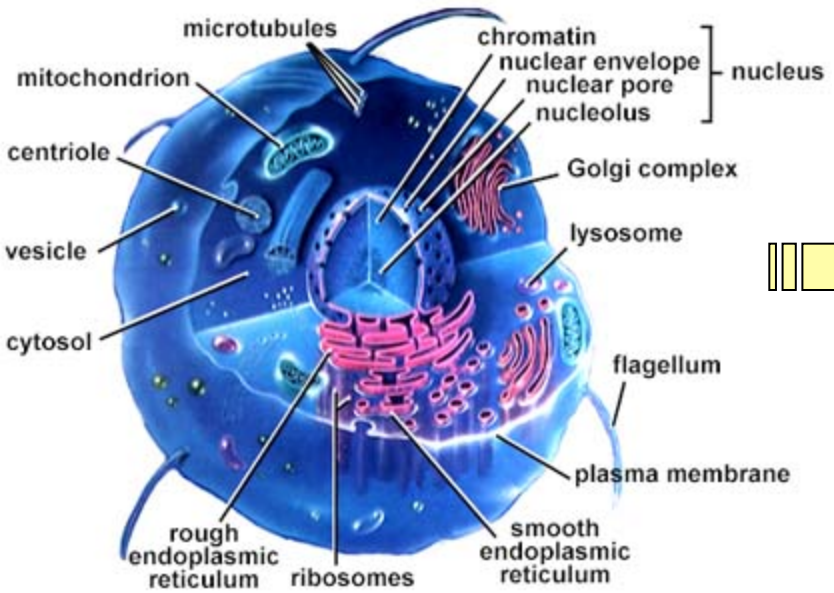
Annotation
Expression
Phenotypes

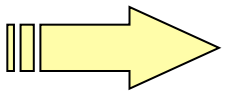
DAVID, GSEA

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

What is Gene Set Enrichment Analysis?

- Break down cellular function into gene sets
  - Every set of genes is associated to a specific cellular function, process, component or pathway





Nuclear Pore

Gene.AAA
Gene.ABA
Gene.ABC

Ribosome

Gene.RP1
Gene.RP2
Gene.RP3
Gene.RP4

Cell Cycle

Gene.CC1
Gene.CC2
Gene.CC3
Gene.CC4
Gene.CC5

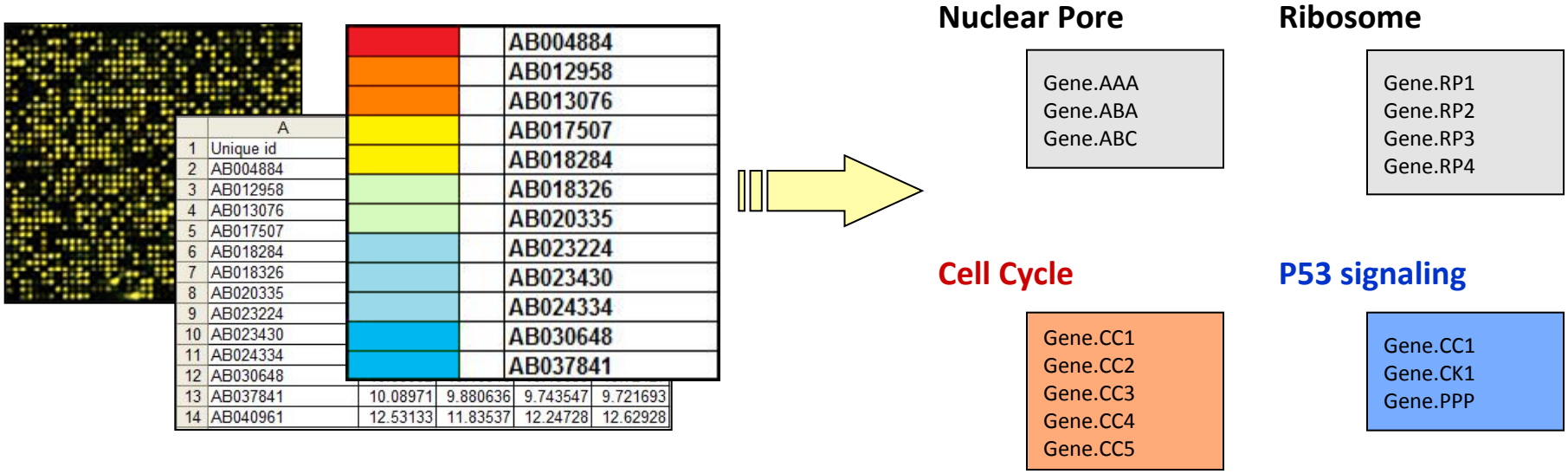
P53 signaling

Gene.CC1
Gene.CK1
Gene.PPP

Daniele Merico

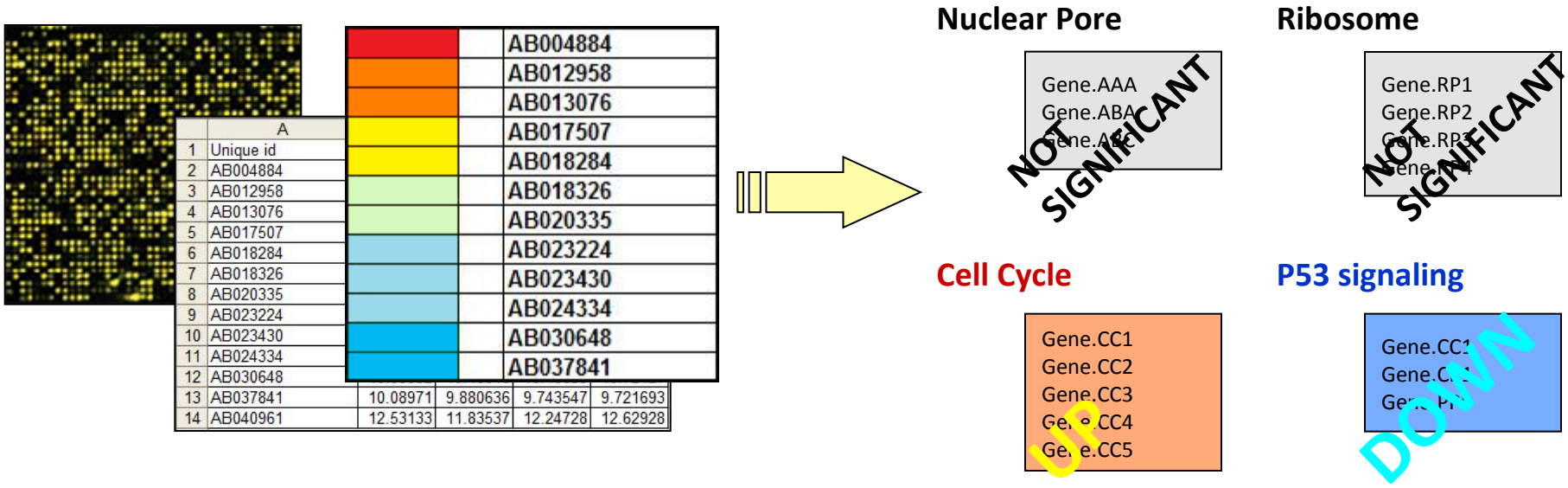
What is Gene Set Enrichment Analysis?

- Find known gene sets (e.g. pathways) enriched in a gene list (e.g. from gene expression)



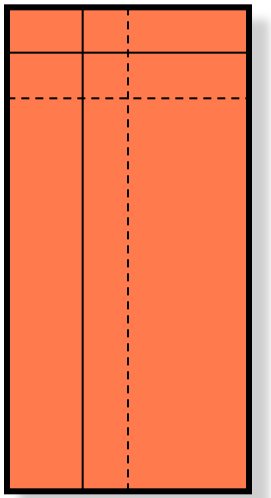
What is Gene Set Enrichment Analysis?

- Find known gene sets (e.g. pathways) enriched in a gene list (e.g. from gene expression)
  - Look for significant enrichment (more on how this works later)



Enrichment Test

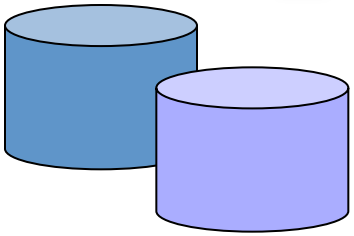
Microarray Experiment (gene expression table)



ENRICHMENT TEST

Enrichment Table

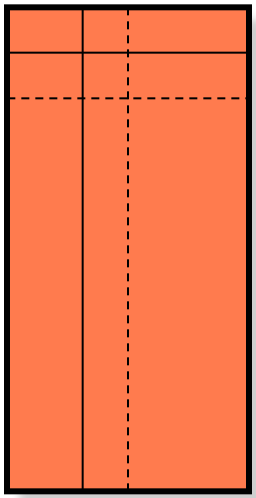
Spindle	**0.00001**
Apoptosis	**0.00025**



Gene-set Databases

Enrichment Test

Microarray Experiment (gene expression table)

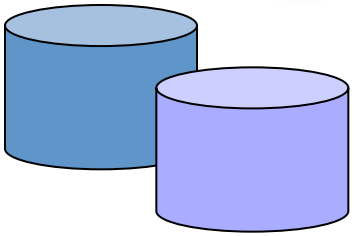


Experimental Data

ENRICHMENT TEST

Enrichment Table

Spindle	0.00001
Apoptosis	0.00025

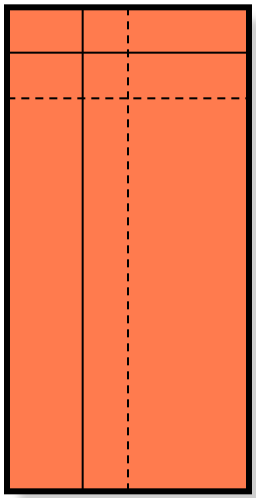


Gene-set Databases

A priori knowledge + existing experimental data

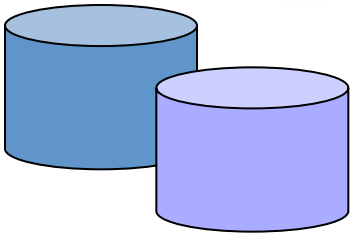
Enrichment Test

Microarray Experiment (gene expression table)



ENRICHMENT TEST

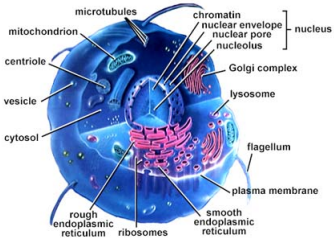
Gene-set Databases



Enrichment Table

Spindle	0.00001
Apoptosis	0.00025

Interpretation & Hypotheses



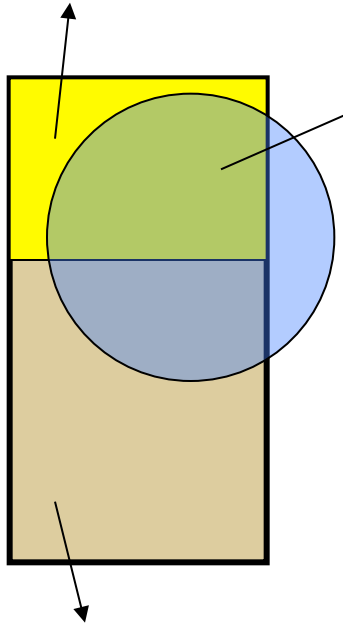
Enrichment Test

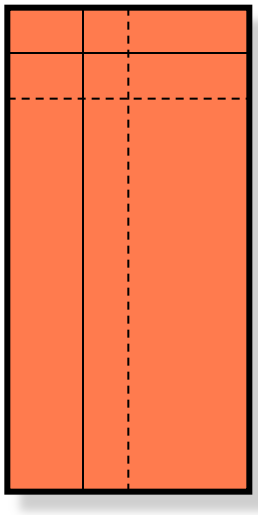
Significant genes (e.g UP)

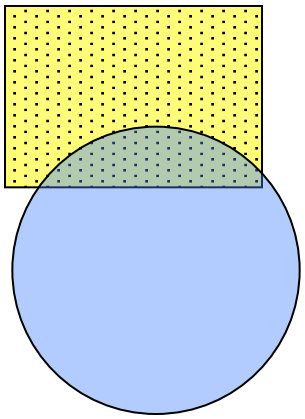
Overlap between gene list and gene-set

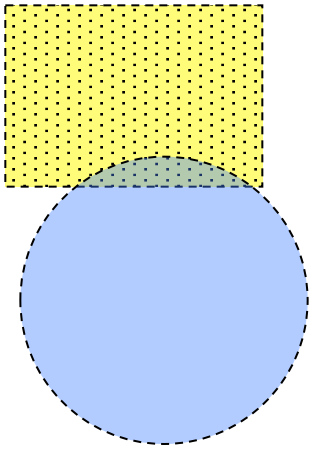
Is this overlap larger than expected by ***random sampling*** *of the array genes?*

Random samples of array genes









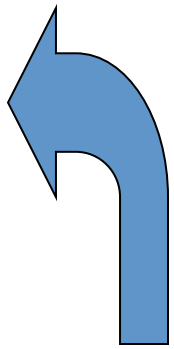
Background genes (array genes not significant)

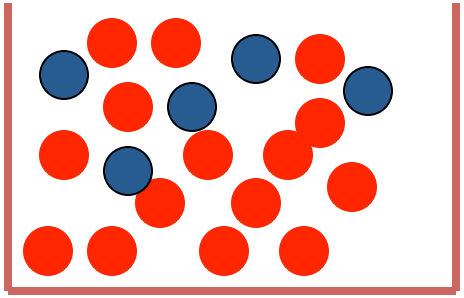
Fisher's exact test

a.k.a., the hypergeometric test

Gene list

- RRP6
- MRD1
- RRP7
- RRP43
- RRP42





Null hypothesis: List is a random sample from population
Alternative hypothesis: More black genes than expected

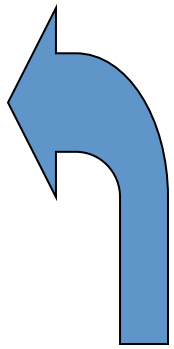
Background population:
500 black genes,
4500 red genes

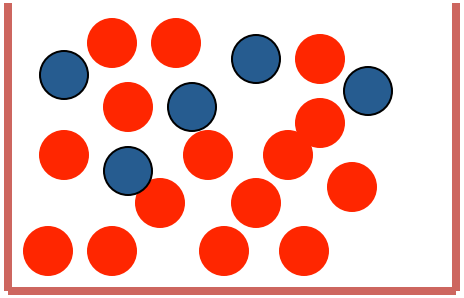
Fisher's exact test

a.k.a., the hypergeometric test

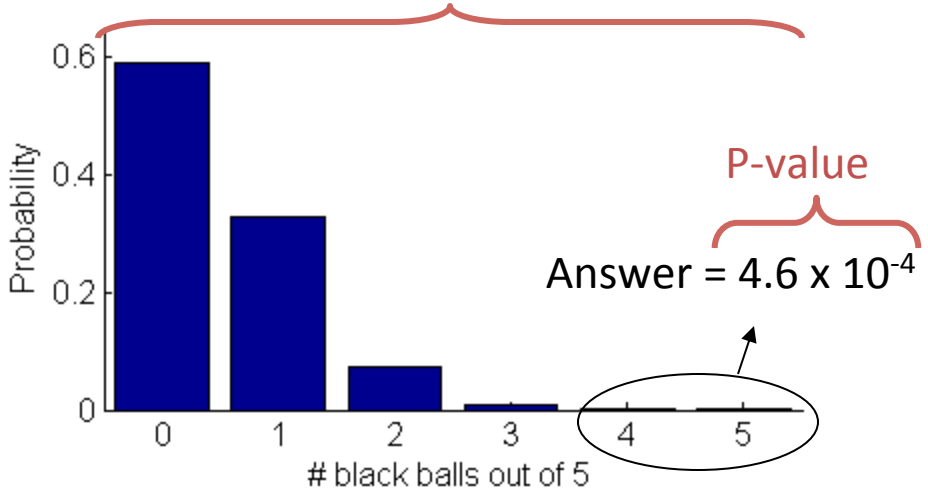
Gene list

- RRP6
- MRD1
- RRP7
- RRP43
- RRP42





Null distribution



P-value

Answer = 4.6×10^{-4}

Background population:

500 black genes,

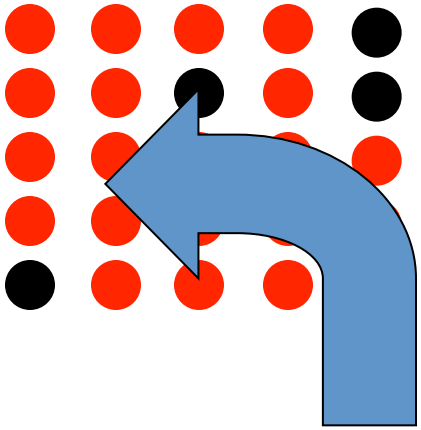
4500 red genes

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. ***More on this later***

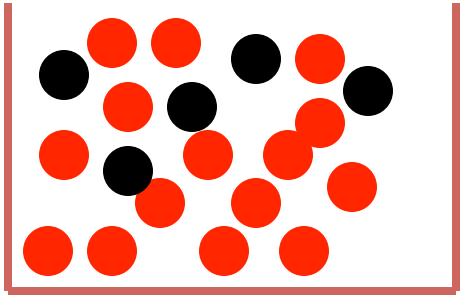
How to win the P-value lottery, part 1

Random draws



… 7,834 draws later …

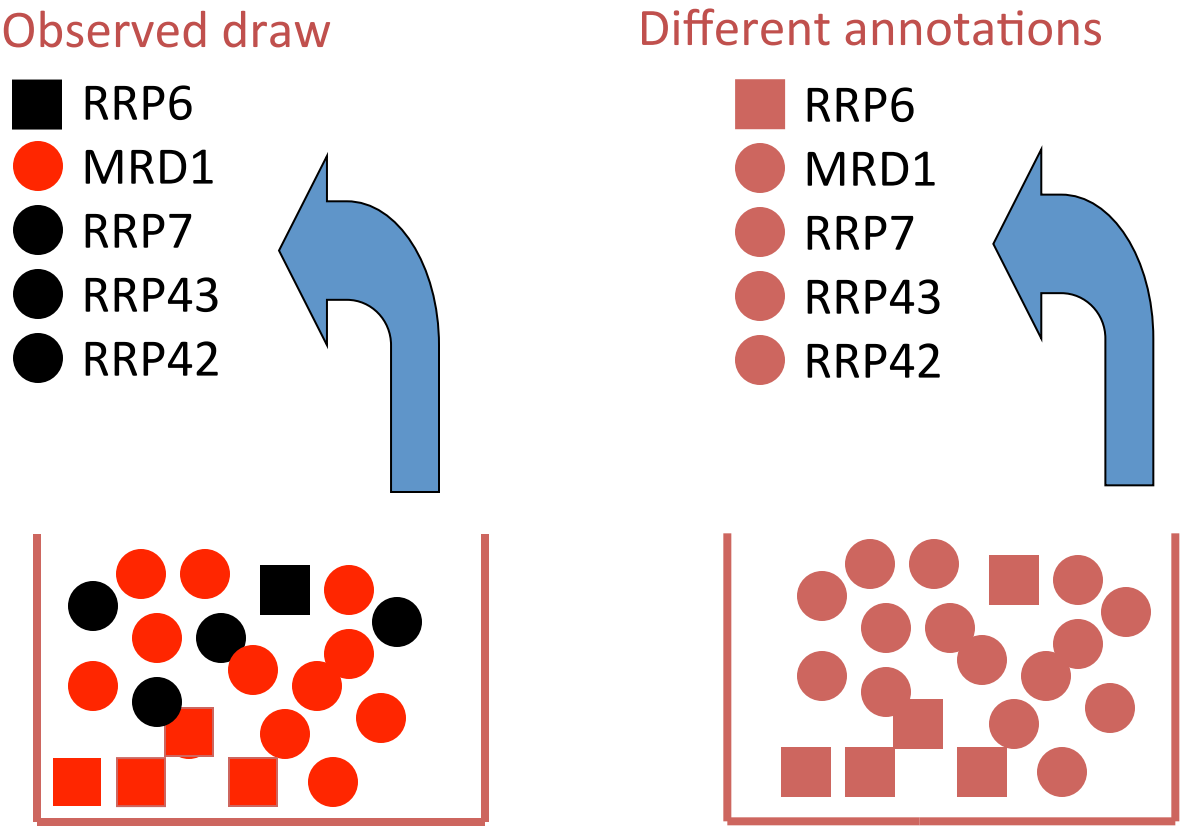
Expect a random draw with observed enrichment once every 1 / P-value draws



Background population:
500 black genes,
4500 red genes

How to win the P-value lottery, part 2

Keep the gene list the same, evaluate different annotations



Simple P-value correction: Bonferroni

If *M* = # of annotations tested:

Corrected P-value = *M* x original P-value

Corrected P-value is greater than or equal to the probability that any single one of the observed enrichments could be due to random draws. The jargon for this correction is **"controlling for the *Family-Wise Error Rate (FWER)*"**

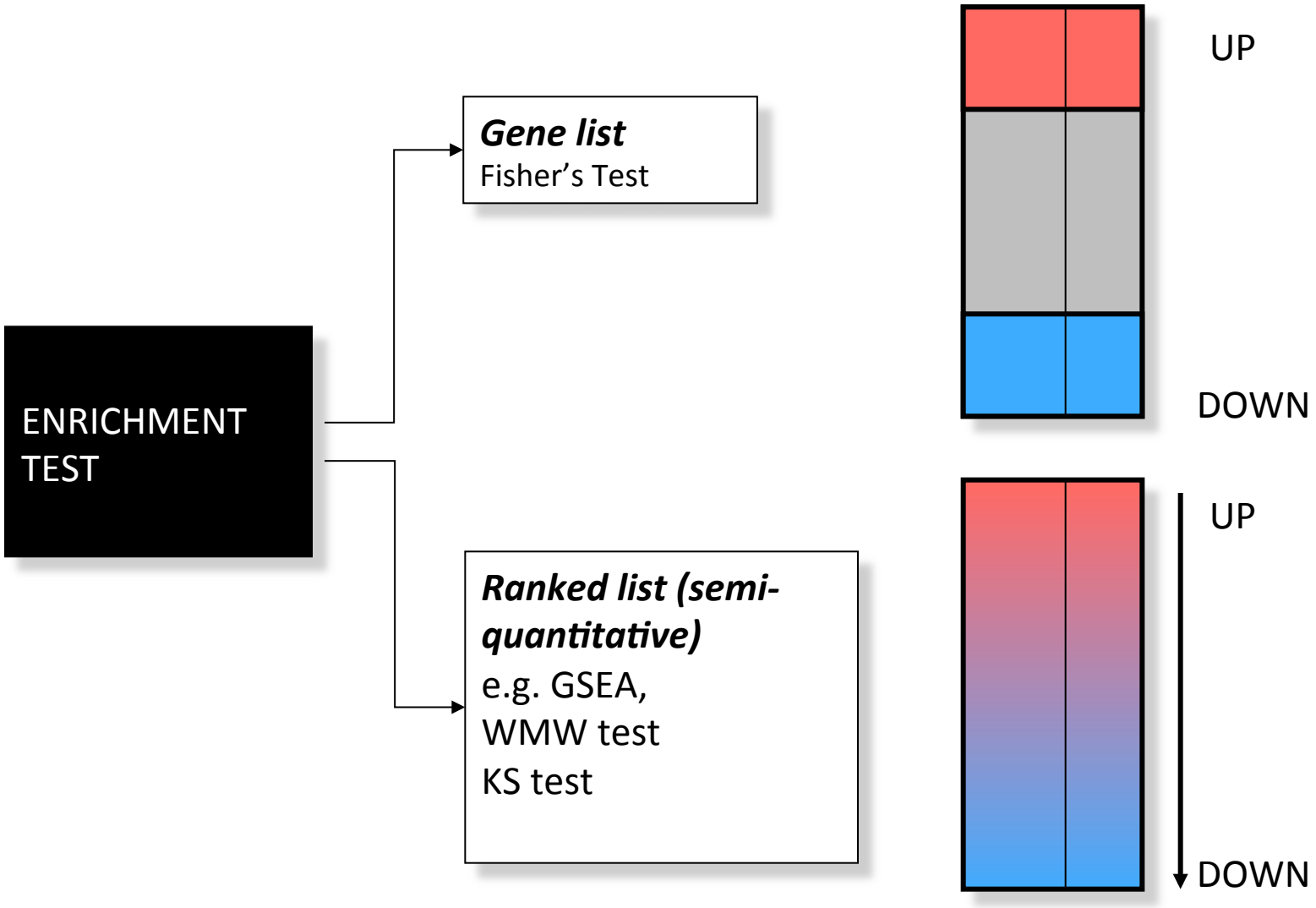
Bonferroni correction caveats

- Bonferroni correction is very stringent and can “wash away” real enrichments.
- Often users are willing to accept a less stringent condition, the “false discovery rate” (FDR), which leads to a gentler correction when there are real enrichments.

False discovery rate (FDR)

- FDR is *the expected **proportion** of the observed enrichments due to random chance.*
- Compare to Bonferroni correction which is a bound on *the probability that **any one** of the observed enrichments could be due to random chance.*
- Typically FDR corrections are calculated using the Benjamini-Hochberg procedure.
- FDR threshold is often called the “q-value”

Beyond Fisher's Exact Test



DAVID Bioinformatics Resources 6.7

National Institute of Allergy and Infectious Diseases (NIAID), NIH

Home | Start Analysis | Shortcut to DAVID Tools | Technical Center | Downloads & APIs | Term of Service | Why DAVID? | About Us

Shortcut to DAVID Tools

Functional Annotation

Gene-annotation enrichment analysis, functional annotation clustering , BioCarta & KEGG pathway mapping, gene-disease association, homologue match, ID translation, literature match and more

Gene Functional Classification

Provide a rapid means to reduce large lists of genes into functionally related groups of genes to help unravel the biological content captured by high throughput technologies. More

Gene ID Conversion

Convert list of gene ID/accessions to others of your choice with the most comprehensive gene ID mapping repository. The ambiguous accessions in the list can also be determined semi-automatically. More

Gene Name Batch Viewer

Display gene names for a given gene list; Search functionally related genes within your list or not in your list; Deep links to enriched detailed information. More

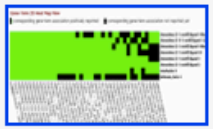
Recommending: A paper published in ***Nature Protocols*** describes step-by-step procedure to use DAVID!

Welcome to DAVID 6.7

2003 - 2010

The **D**atabase for **A**nnotation, **V**isualization and **I**ntegrated **D**iscovery (**DAVID**) v6.7 is an update to the sixth version of our original web-accessible programs. DAVID now provides a comprehensive set of functional annotation tools for investigators to understand biological meaning behind large list of genes. For any given gene list, DAVID tools are able to:

- Identify enriched biological themes, particularly GO terms
- Discover enriched functional-related gene groups
- Cluster redundant annotation terms
- Visualize genes on BioCarta & KEGG pathway maps
- Display related many-genes-to-many-terms on 2-D view.
- Search for other functionally related genes not in the list
- List interacting proteins
- Explore gene names in batch
- Link gene-disease associations
- Highlight protein functional domains and motifs
- Redirect to related literatures
- Convert gene identifiers from one type to another.
- And more


Screen Shot 1


Screen Shot 2

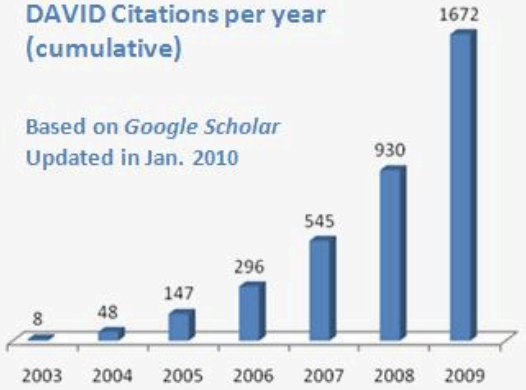

Screen Shot 3

Search

What's Important in DAVID?

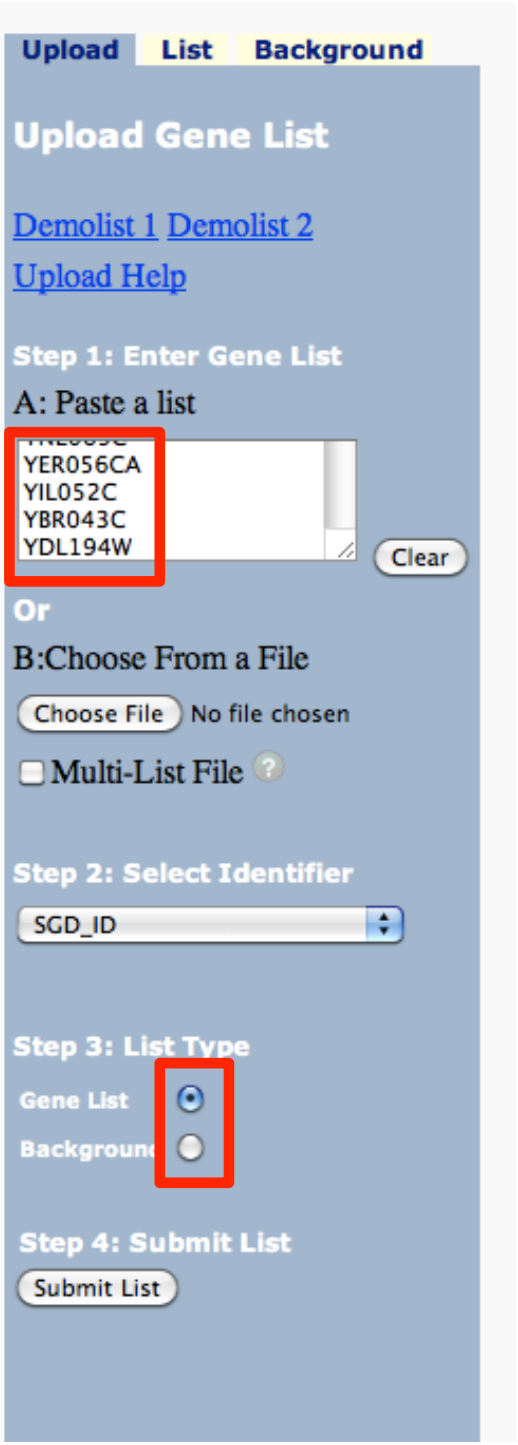
- Current (v 6.7) release note
- New requirement to cite DAVID
- IDs of Affy Exon and Gene arrays supported
- Novel Classification Algorithms
- Pre-built Affymetrix and Illumina backgrounds
- User's customized gene background
- Enhanced calculating speed

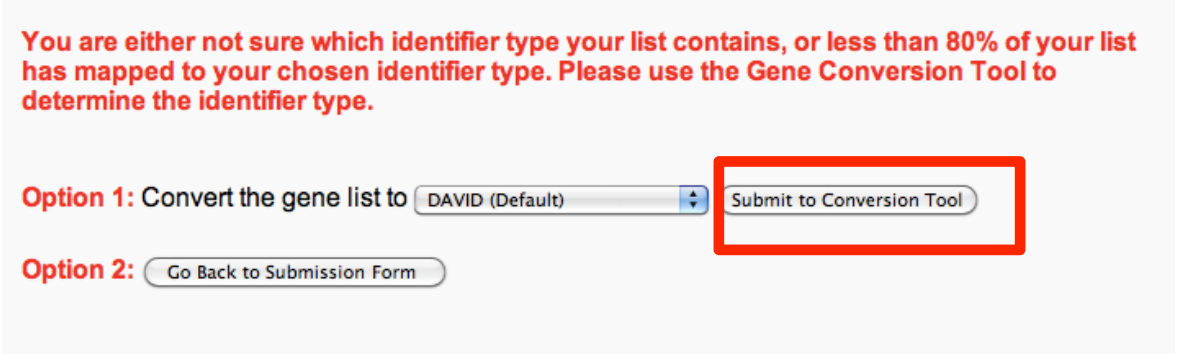
Statistics of DAVID



- Total: > 2,000 DAVID citations
- Daily Usage: ~1200 gene lists/sublists from ~400 unique researchers.
- Total Usage: ~800,000 gene lists/sublists from >5,000 research institutes world-wide

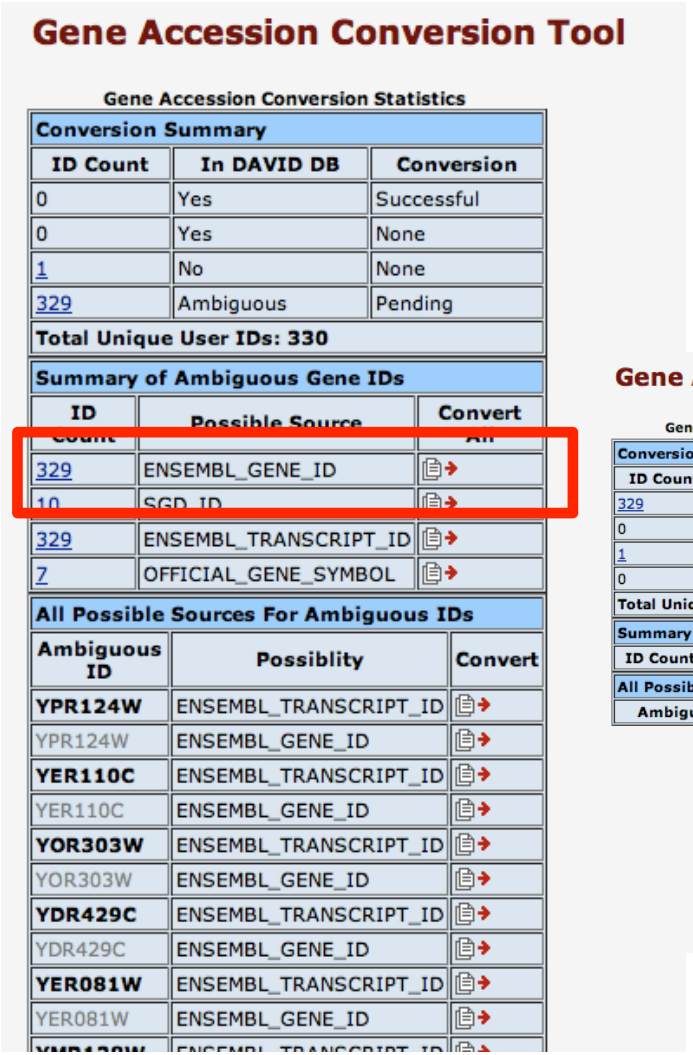
Upload your gene list





If DAVID doesn't recognize your genes, it can try to detect the correct identifiers to use

Gene ID mapping results



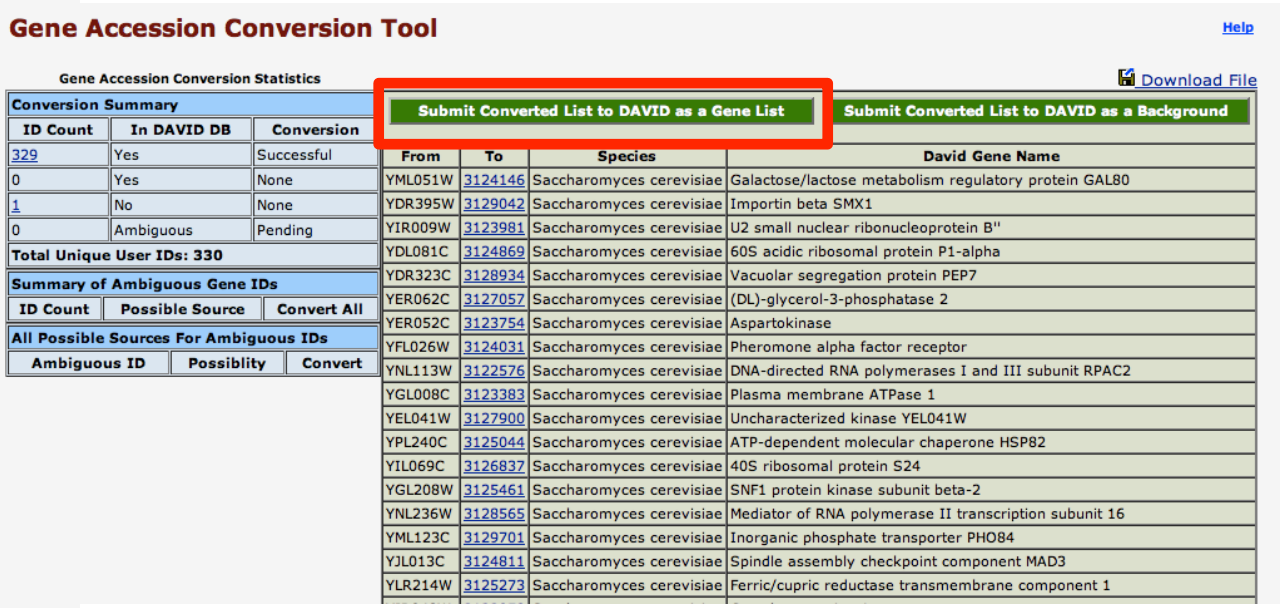
Gene Accession Conversion Tool

Gene Accession Conversion Statistics

Conversion Summary		
ID Count	**In DAVID DB**	**Conversion**
0	Yes	Successful
0	Yes	None
1	No	None
329	Ambiguous	Pending
Total Unique User IDs: 330		

Summary of Ambiguous Gene IDs		
ID Count	**Possible Source**	**Convert All**
329	ENSEMBL_GENE_ID	
10	SGD_ID	
329	ENSEMBL_TRANSCRIPT_ID	
7	OFFICIAL_GENE_SYMBOL	

All Possible Sources For Ambiguous IDs		
Ambiguous ID	**Possiblity**	**Convert**
YPR124W	ENSEMBL_TRANSCRIPT_ID	
YPR124W	ENSEMBL_GENE_ID	
YER110C	ENSEMBL_TRANSCRIPT_ID	
YER110C	ENSEMBL_GENE_ID	
YOR303W	ENSEMBL_TRANSCRIPT_ID	
YOR303W	ENSEMBL_GENE_ID	
YDR429C	ENSEMBL_TRANSCRIPT_ID	
YDR429C	ENSEMBL_GENE_ID	
YER081W	ENSEMBL_TRANSCRIPT_ID	
YER081W	ENSEMBL_GENE_ID	



Gene Accession Conversion Tool

Help

Gene Accession Conversion Statistics

Download File

Conversion Summary		
ID Count	**In DAVID DB**	**Conversion**
329	Yes	Successful
0	Yes	None
1	No	None
0	Ambiguous	Pending
Total Unique User IDs: 330		

Summary of Ambiguous Gene IDs		
ID Count	**Possible Source**	**Convert All**

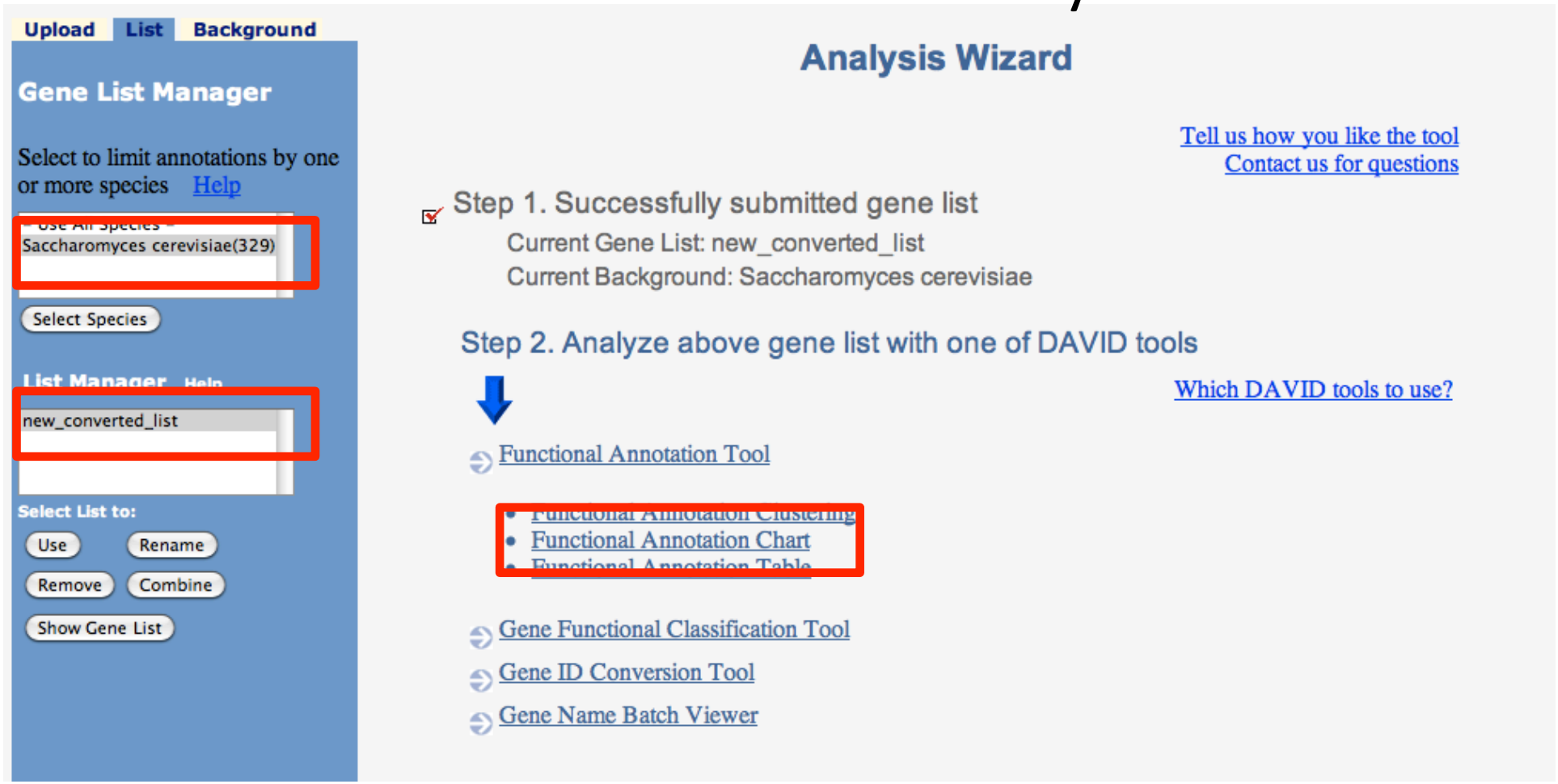
All Possible Sources For Ambiguous IDs		
Ambiguous ID	**Possiblity**	**Convert**

Submit Converted List to DAVID as a Gene List | **Submit Converted List to DAVID as a Background**

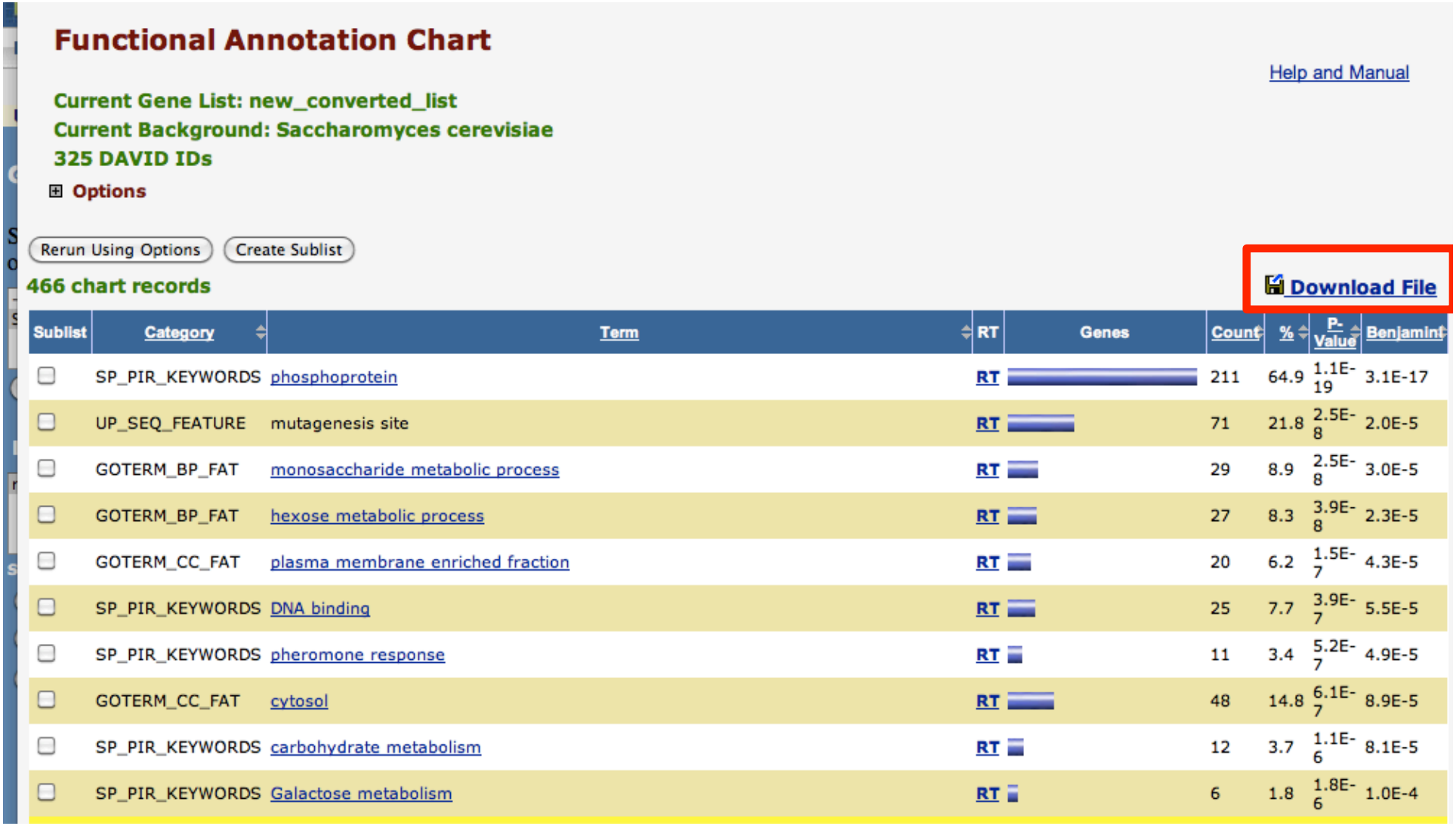
From	To	Species	David Gene Name
YML051W	3124146	Saccharomyces cerevisiae	Galactose/lactose metabolism regulatory protein GAL80
YDR395W	3129042	Saccharomyces cerevisiae	Importin beta SMX1
YIR009W	3123981	Saccharomyces cerevisiae	U2 small nuclear ribonucleoprotein B''
YDL081C	3124869	Saccharomyces cerevisiae	60S acidic ribosomal protein P1-alpha
YDR323C	3128934	Saccharomyces cerevisiae	Vacuolar segregation protein PEP7
YER062C	3127057	Saccharomyces cerevisiae	(DL)-glycerol-3-phosphatase 2
YER052C	3123754	Saccharomyces cerevisiae	Aspartokinase
YFL026W	3124031	Saccharomyces cerevisiae	Pheromone alpha factor receptor
YNL113W	3122576	Saccharomyces cerevisiae	DNA-directed RNA polymerases I and III subunit RPAC2
YGL008C	3123383	Saccharomyces cerevisiae	Plasma membrane ATPase 1
YEL041W	3127900	Saccharomyces cerevisiae	Uncharacterized kinase YEL041W
YPL240C	3125044	Saccharomyces cerevisiae	ATP-dependent molecular chaperone HSP82
YIL069C	3126837	Saccharomyces cerevisiae	40S ribosomal protein S24
YGL208W	3125461	Saccharomyces cerevisiae	SNF1 protein kinase subunit beta-2
YNL236W	3128565	Saccharomyces cerevisiae	Mediator of RNA polymerase II transcription subunit 16
YML123C	3129701	Saccharomyces cerevisiae	Inorganic phosphate transporter PHO84
YJL013C	3124811	Saccharomyces cerevisiae	Spindle assembly checkpoint component MAD3
YLR214W	3125273	Saccharomyces cerevisiae	Ferric/cupric reductase transmembrane component 1

Step 1

Run the enrichment analysis



Run the enrichment analysis



Functional Annotation Chart

Help and Manual

Current Gene List: new_converted_list
Current Background: Saccharomyces cerevisiae
325 DAVID IDs

⊞ Options

Rerun Using Options | Create Sublist

466 chart records

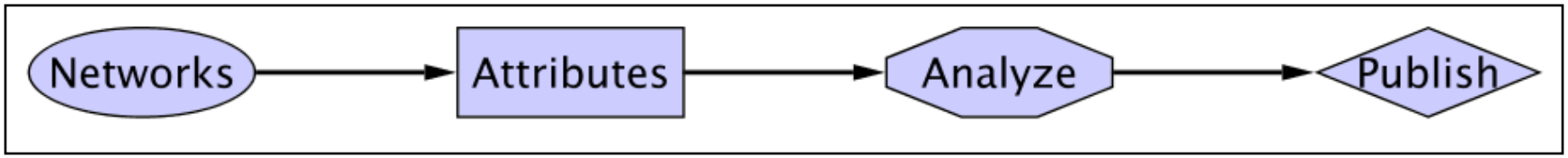
Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
☐	SP_PIR_KEYWORDS	phosphoprotein	RT		211	64.9	1.1E-19	3.1E-17
☐	UP_SEQ_FEATURE	mutagenesis site	RT		71	21.8	2.5E-8	2.0E-5
☐	GOTERM_BP_FAT	monosaccharide metabolic process	RT		29	8.9	2.5E-8	3.0E-5
☐	GOTERM_BP_FAT	hexose metabolic process	RT		27	8.3	3.9E-8	2.3E-5
☐	GOTERM_CC_FAT	plasma membrane enriched fraction	RT		20	6.2	1.5E-7	4.3E-5
☐	SP_PIR_KEYWORDS	DNA binding	RT		25	7.7	3.9E-7	5.5E-5
☐	SP_PIR_KEYWORDS	pheromone response	RT		11	3.4	5.2E-7	4.9E-5
☐	GOTERM_CC_FAT	cytosol	RT		48	14.8	6.1E-7	8.9E-5
☐	SP_PIR_KEYWORDS	carbohydrate metabolism	RT		12	3.7	1.1E-6	8.1E-5
☐	SP_PIR_KEYWORDS	Galactose metabolism	RT		6	1.8	1.8E-6	1.0E-4

Network visualization and analysis

Network Analysis Workflow

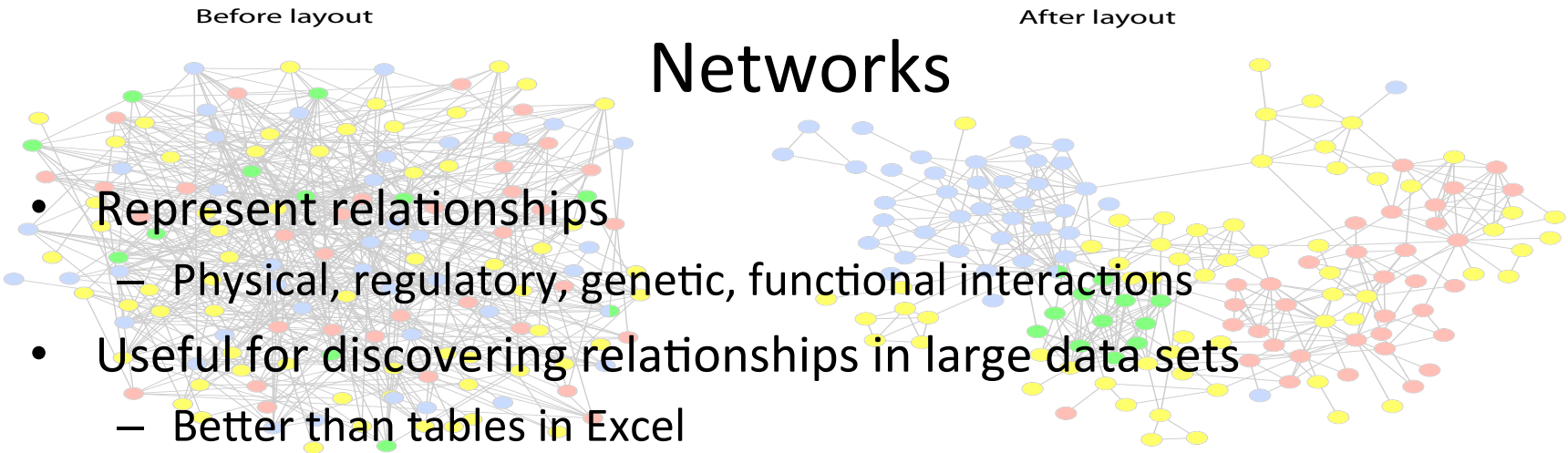
- Load Networks e.g. PPI data
  - Import network data into Cytoscape
- Load Attributes e.g. gene expression data
  - Get data about networks into Cytoscape
- Analyze and Visualize Networks
- Prepare for Publication
- A specific example of this workflow:
  - Cline, et al. "Integration of biological networks and gene expression data using Cytoscape", Nature Protocols, 2, 2366-2382 (2007).



Network Visualization and Analysis Outline

- Network introduction
- Network visualization
- Cytoscape software tool for network visualization and analysis
- Network analysis

Networks

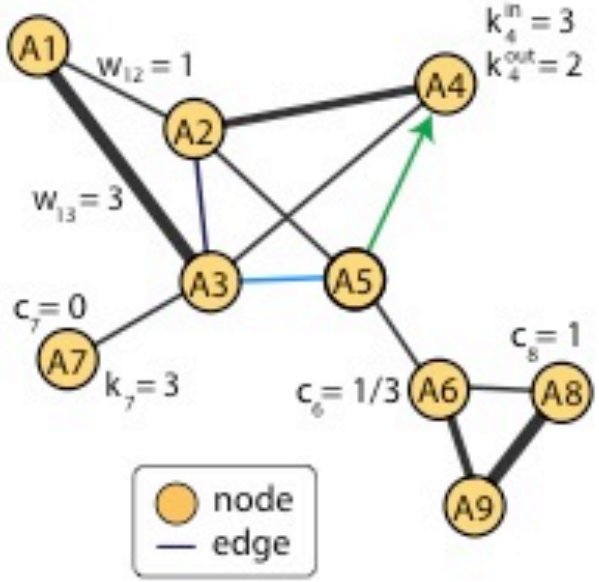


- Represent relationships
  - Physical, regulatory, genetic, functional interactions
- Useful for discovering relationships in large data sets
  - Better than tables in Excel
- Visualize multiple data types together
  - See interesting patterns

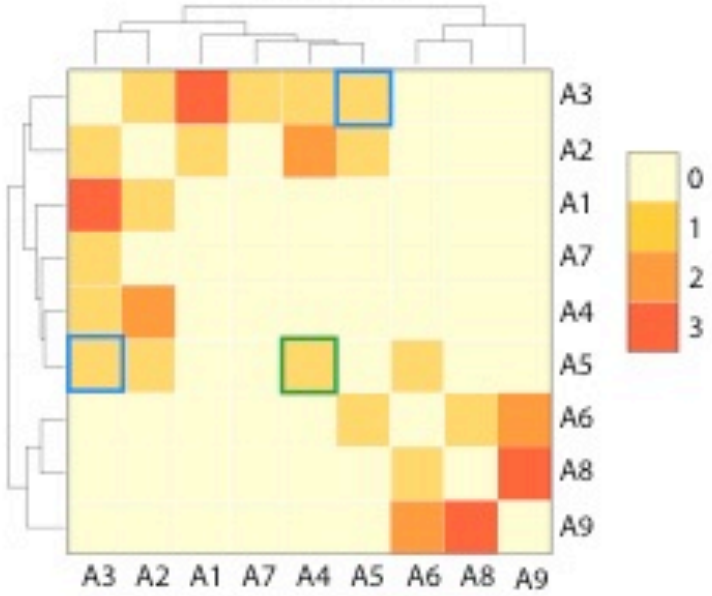
Network Representations

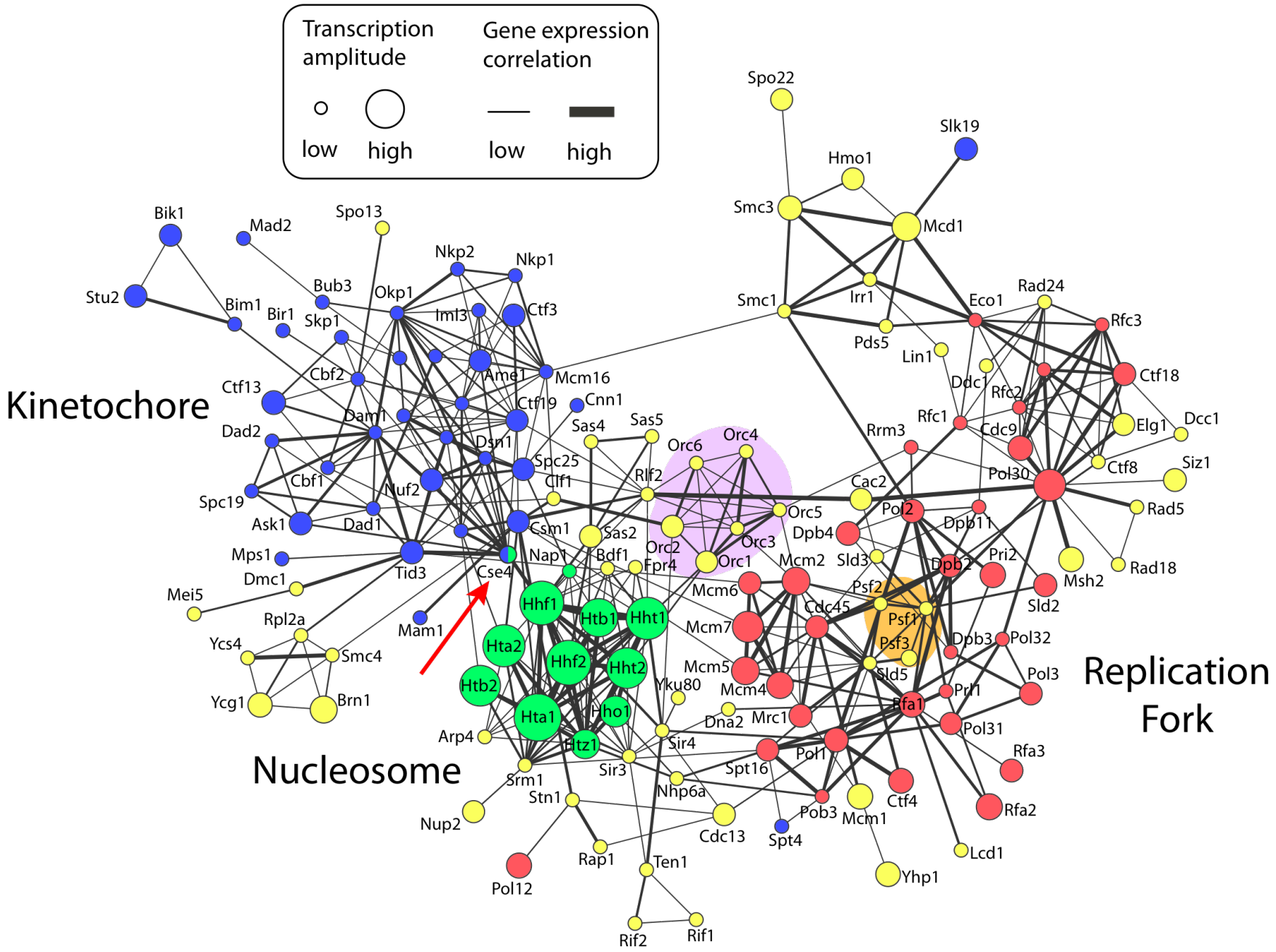
Relationships	Optional weight
A1 ↔ A2	1
A1 ↔ A3	3
A2 ↔ A3	1
A2 ↔ A4	2
A2 ↔ A5	1
A3 ↔ A4	1
A3 ↔ A5	1
A3 ↔ A7	1
A5 → A4	1
A5 ↔ A6	1
A6 ↔ A8	1
A6 ↔ A9	2
A8 ↔ A9	3

Network

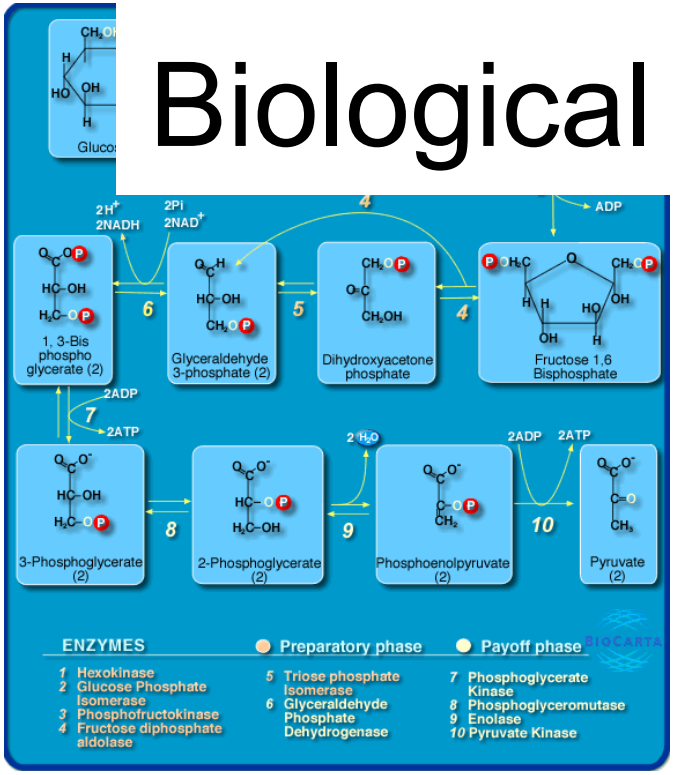


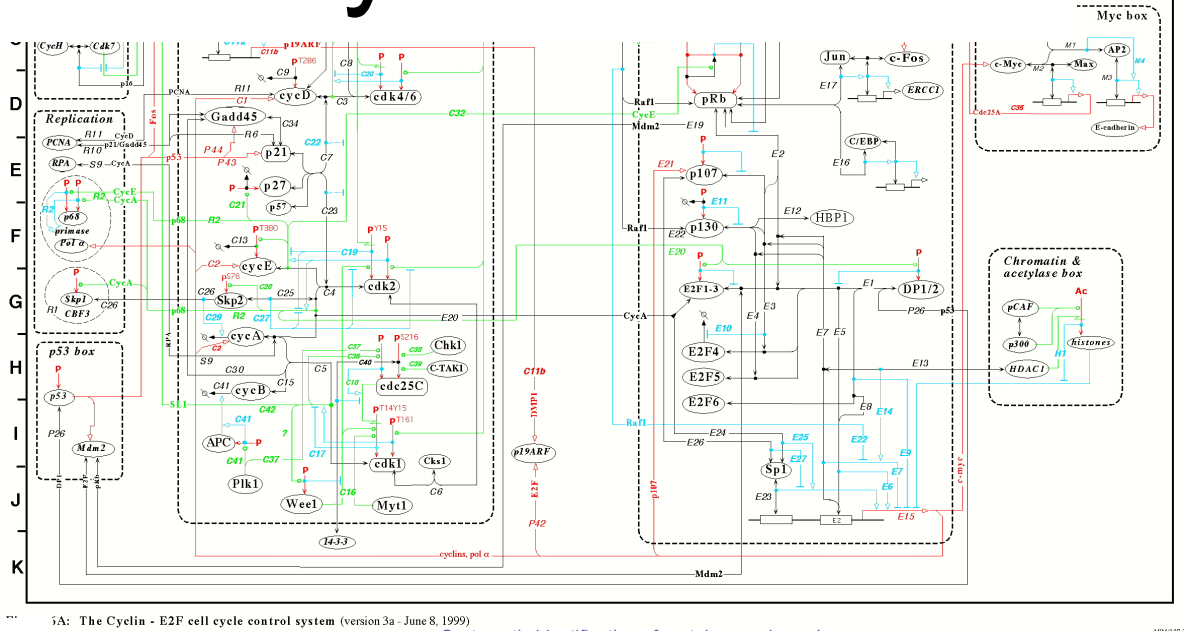
Heatmap



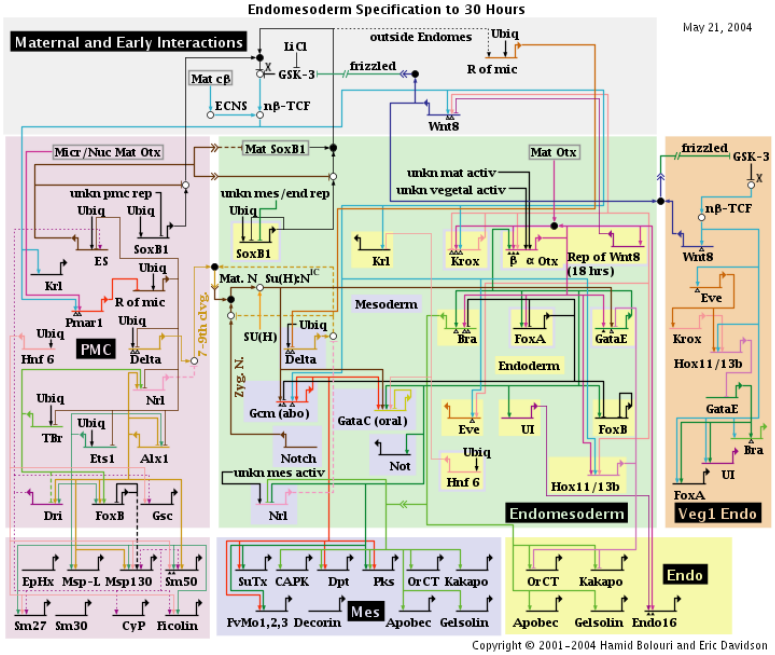

Transcription amplitude
low
high
Gene expression correlation
low
high
Kinetochore
Nucleosome
Replication Fork
Bik1
Stu2
Bim1
Mad2
Bir1
Spo13
Skp1
Bub3
Nkp2
Okp1
Nkp1
Iml3
Ctf3
Ame1
Mcm16
Ctf13
Cbf2
Ctf19
Cnn1
Dam1
Dad2
Dsn1
Sas4
Sas5
Spc25
Cbf1
Clf1
Rlf2
Spc19
Muf2
Ask1
Dad1
Csm1
Sas2
Mps1
Tid3
Nap1
Bdf1
Fpr4
Dmc1
Cse4
Mei5
Rpl2a
Mam1
Ycs4
Smc4
Ycg1
Brn1
Hhf1
Htb1
Hht1
Hta2
Hhf2
Hht2
Htb2
Hta1
Hho1
Htz1
Arp4
Srm1
Stn1
Nup2
Pol12
Rap1
Ten1
Rif2
Rif1
Cdc13
Nhp6a
Sir3
Sir4
Yku80
Dna2
Spo22
Slk19
Hmo1
Smc3
Mcd1
Smc1
Irr1
Pds5
Eco1
Rad24
Rfc3
Lin1
Ctf18
Ddc1
Rfc2
Rfc1
Cdc9
Elg1
Dcc1
Rrm3
Ctf8
Siz1
Pol30
Orc4
Orc6
Orc5
Orc3
Orc2
Orc1
Cac2
Pol2
Rad5
Dpb4
Dpb11
Sld3
Pri2
Dpb2
Msh2
Rad18
Mcm2
Mcm6
Mcm7
Mcm5
Mcm4
Cdc45
Psf2
Psf1
Psf3
Sld2
Dpb3
Pol32
Sld5
Pol3
Prl1
Rfa1
Pol31
Rfa3
Rfa2
Mrc1
Pol1
Spt16
Ctf4
Pob3
Mcm1
Spt4
Lcd1
Yhp1

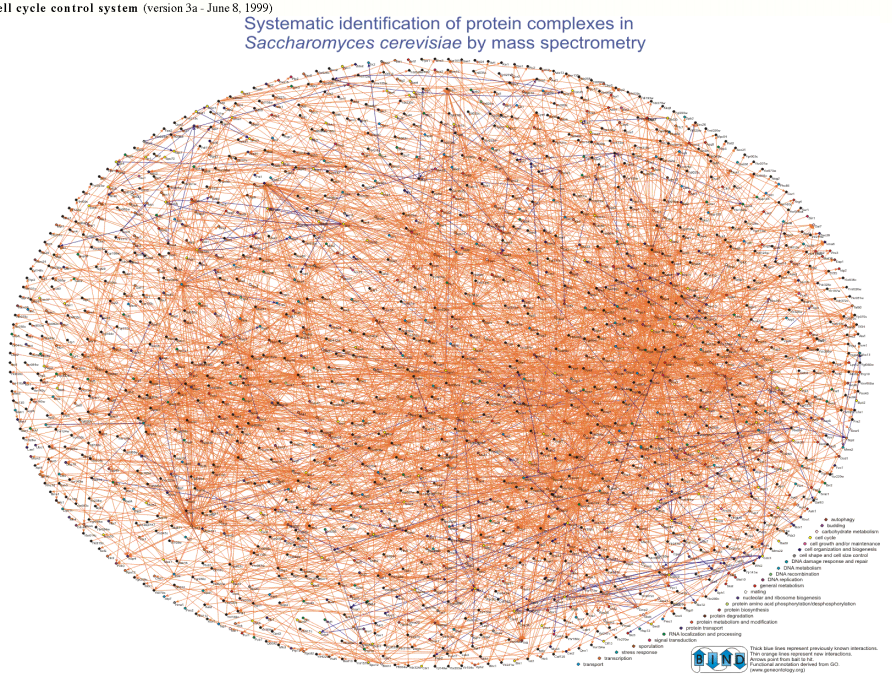
Biological Pathways/Networks?





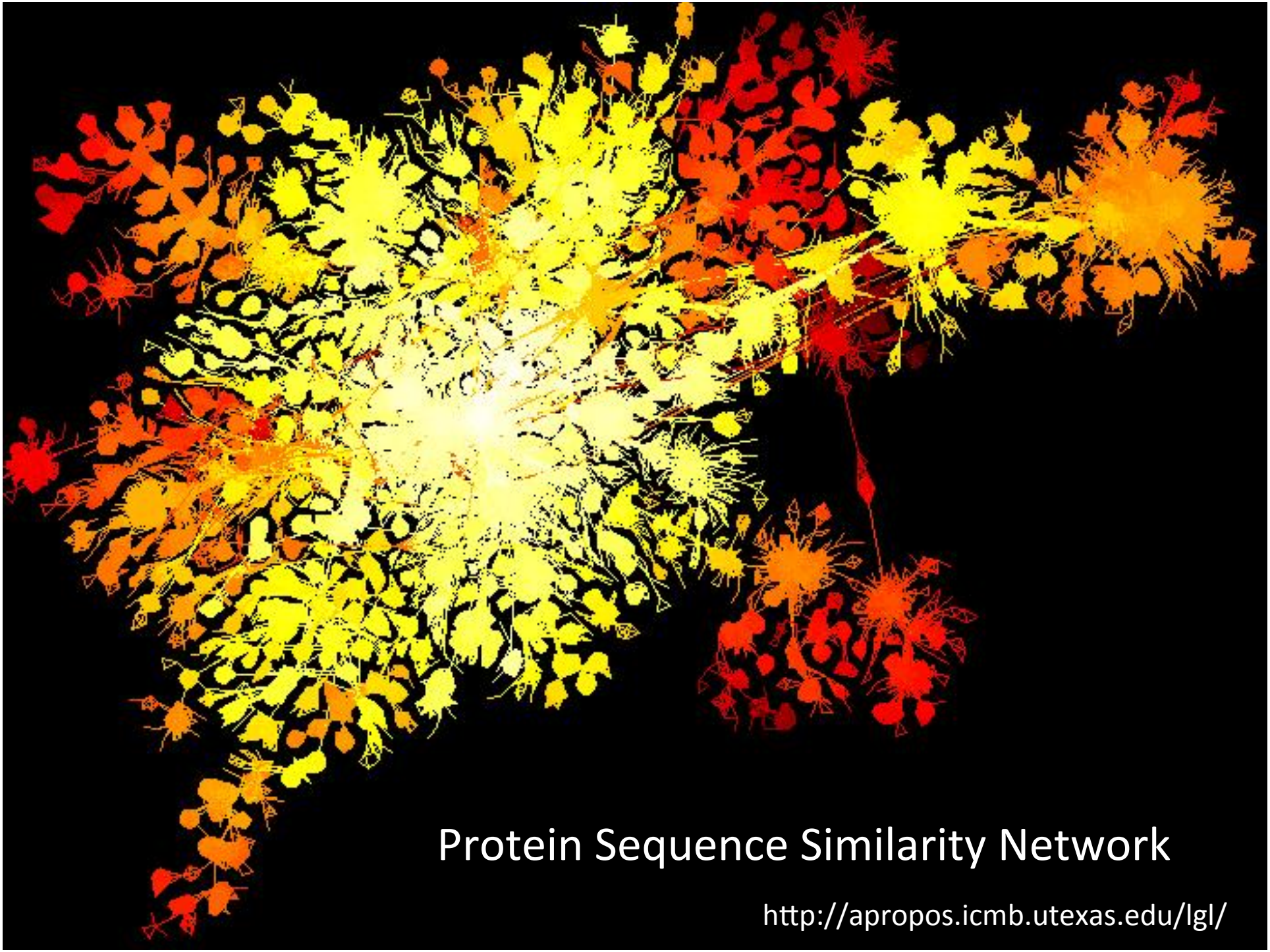
A: The Cyclin - E2F cell cycle control system (version 3a - June 8, 1999)





Mapping Biology to a Network

- A simple mapping
  - one compound/node, one interaction/edge
- A more realistic mapping
  - Cell localization, cell cycle, cell type, taxonomy
  - Only represent physiologically relevant interaction networks
- Edges can represent other relationships
- Critical: understand what nodes and edges mean

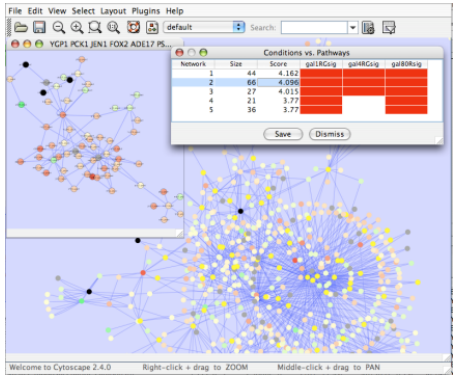

Protein Sequence Similarity Network
http://apropos.icmb.utexas.edu/lgl/

Six Degrees of Separation

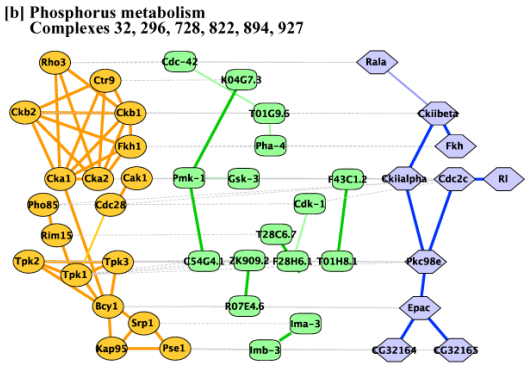
- Everyone in the world is connected
  by at most six links
- Which path should we take?
- Shortest path by breadth first search
  - If two nodes are connected, will find the shortest path betw…
- Are two proteins connected? If so, how?
- Biologically relevant?

http://www.time.com/time/techtime/200406/community.html

Applications of Network Biology

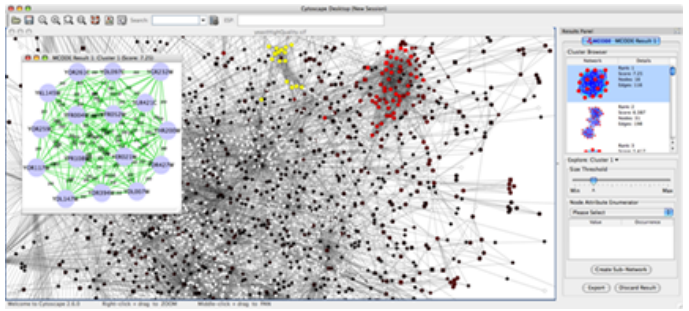


jActiveModules, UCSD

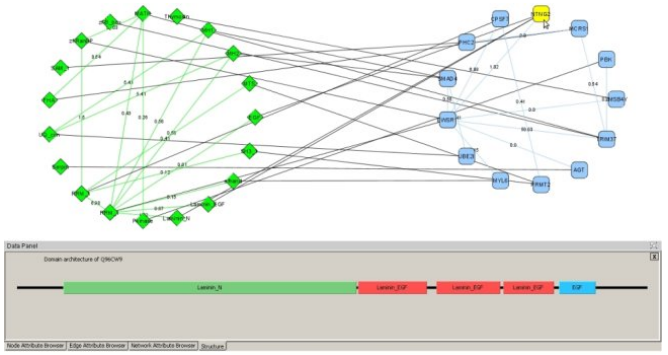


PathBlast, UCSD

- **Gene Function Prediction** – shows connections to sets of genes/proteins involved in same biological process
- **Detection of protein complexes/ other modular structures** – discover modularity & higher order organization (motifs, feedback loops)
- **Network evolution** – biological process(es) conservation across species
- **Prediction of new interactions and functional associations** – Statistically significant domain-domain correlations in protein interaction network to predict protein-protein or genetic interaction

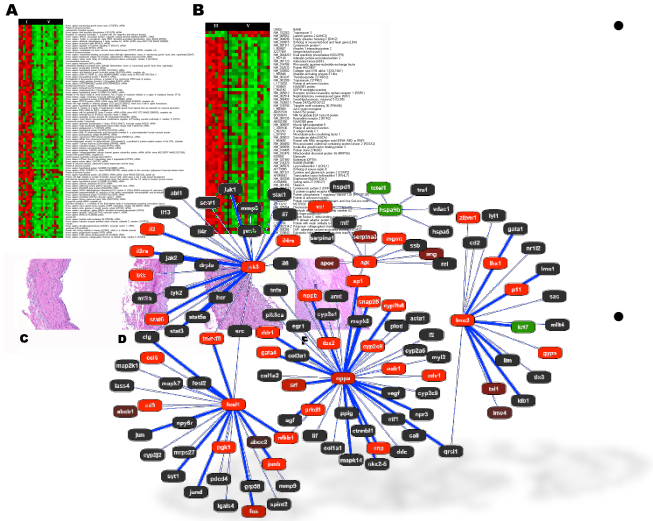


MCODE, University of Toronto

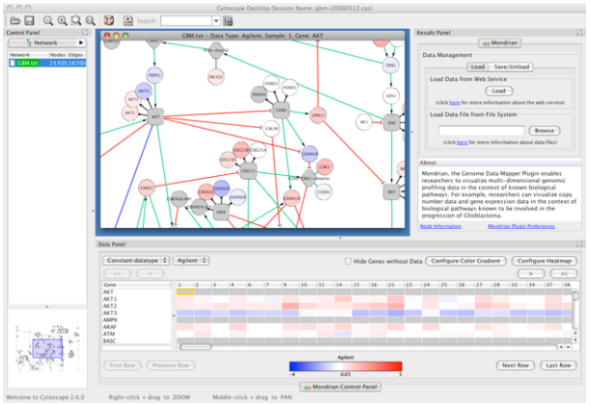


DomainGraph, Max Planck Institute

Applications of Network Informatics in Disease

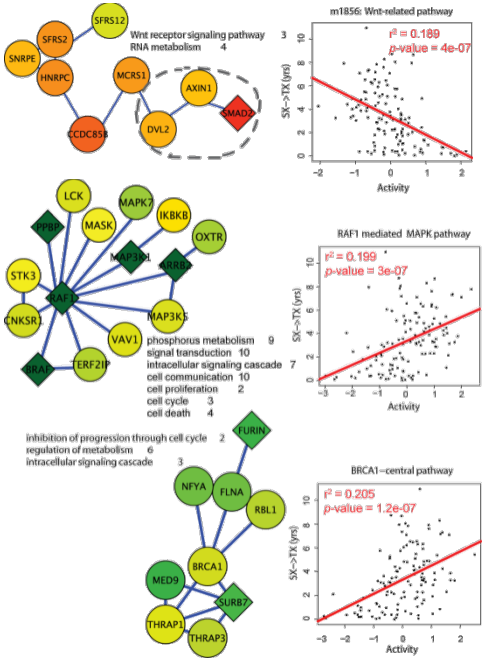


Agilent Literature Search



Mondrian, MSKCC

- **Identification of disease subnetworks** – identification of disease network subnetworks that are transcriptionally active in disease.
- **Subnetwork-based diagnosis** – source of biomarkers for disease classification, identify interconnected genes whose aggregate expression levels are predictive of disease state
- **Subnetwork-based gene association** – map common pathway mechanisms affected by collection of genotypes



PinnacleZ, UCSD

What Have We Learned?

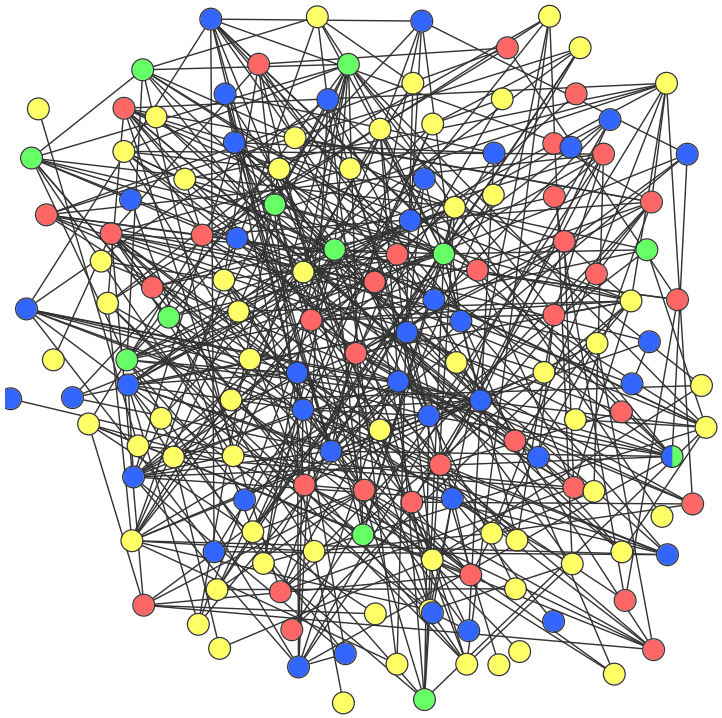
- Networks are useful for seeing relationships in large data sets
- Important to understand what the nodes and edges mean
- Important to define the biological question - know what you want to do with your gene list or network
- Many methods available for gene list and network analysis
  - Good to determine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

Network Visualization Outline

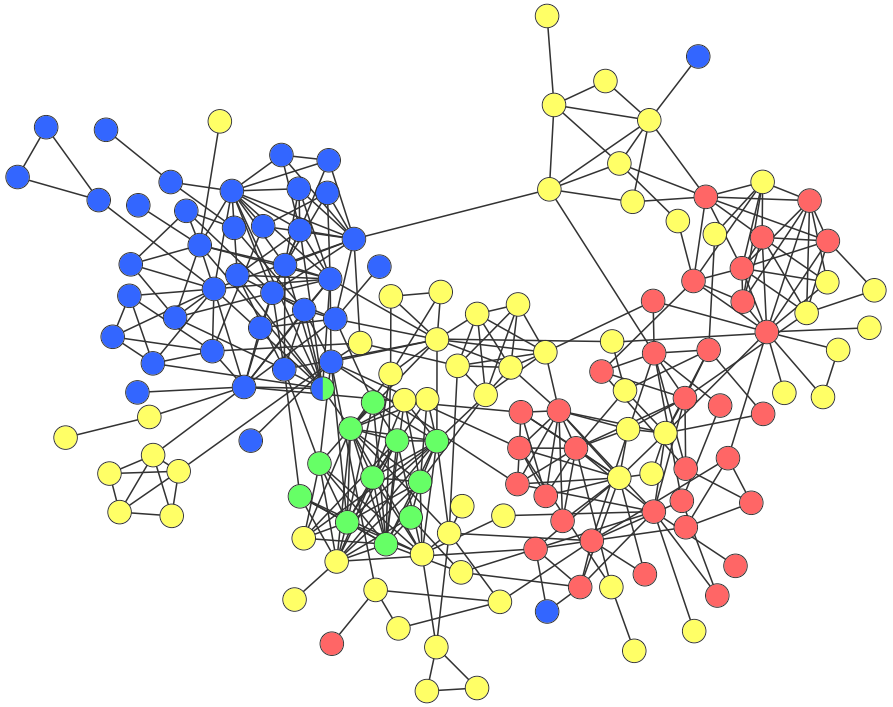
- Automatic network layout
- Visual features
- Visually interpreting a network

Automatic network layout

Before layout



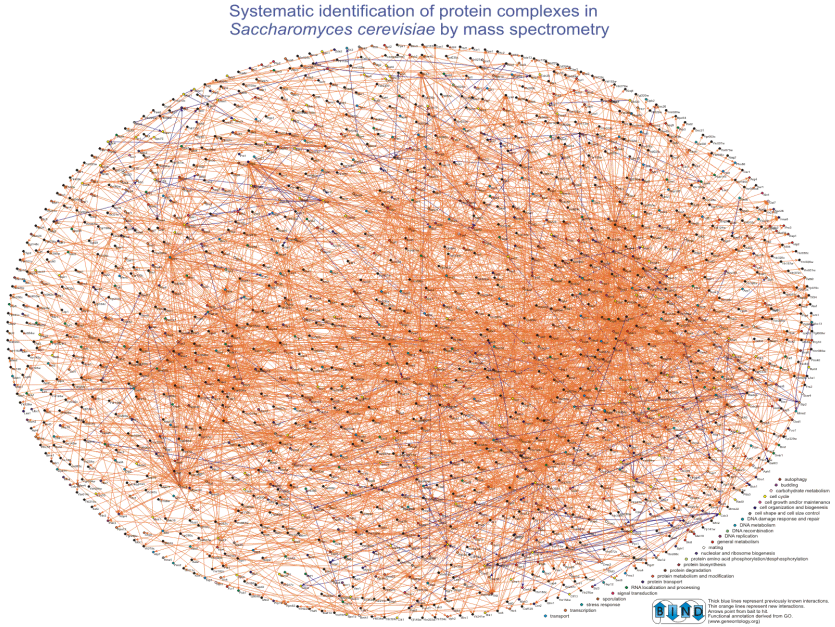
After layout



Automatic network layout

- Force-directed: nodes repel and edges pull
- Good for up to 500 nodes
  - Bigger networks give hairballs - Reduce number of edges
- Advice: try force directed first, or hierarchical for tree-like networks
- Tips for better looking networks
  - Manually adjust layout
  - Load network into a drawing program (e.g. Illustrator) and adjust labels

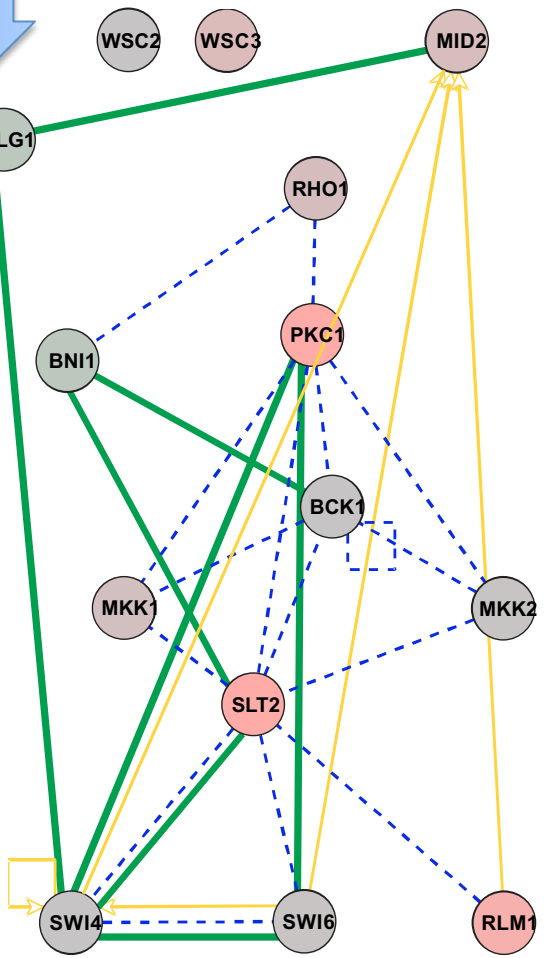
Overview



Zoom

Focus

PKC Cell Wall Integrity



Synthetic Lethal

Transcription Factor Regulation

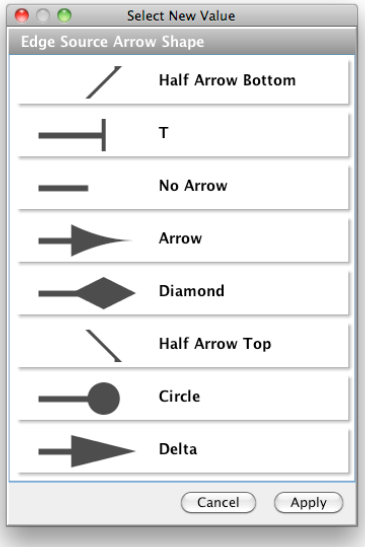
Protein-Protein Interaction

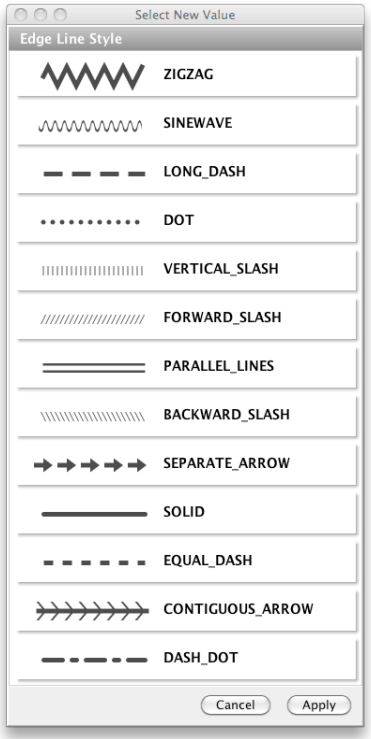
Up Regulated Gene Expression

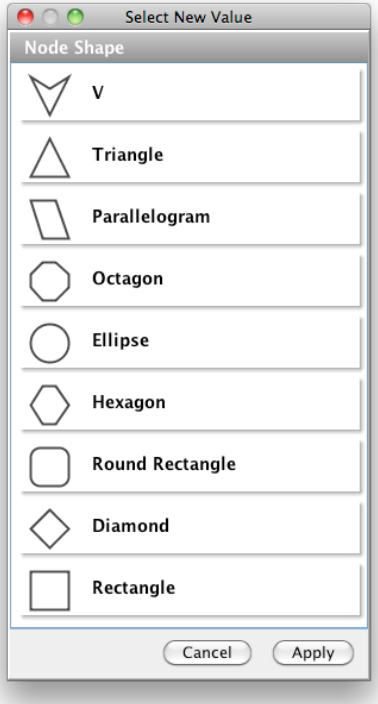
Down Regulated Gene Expression

Visual Features

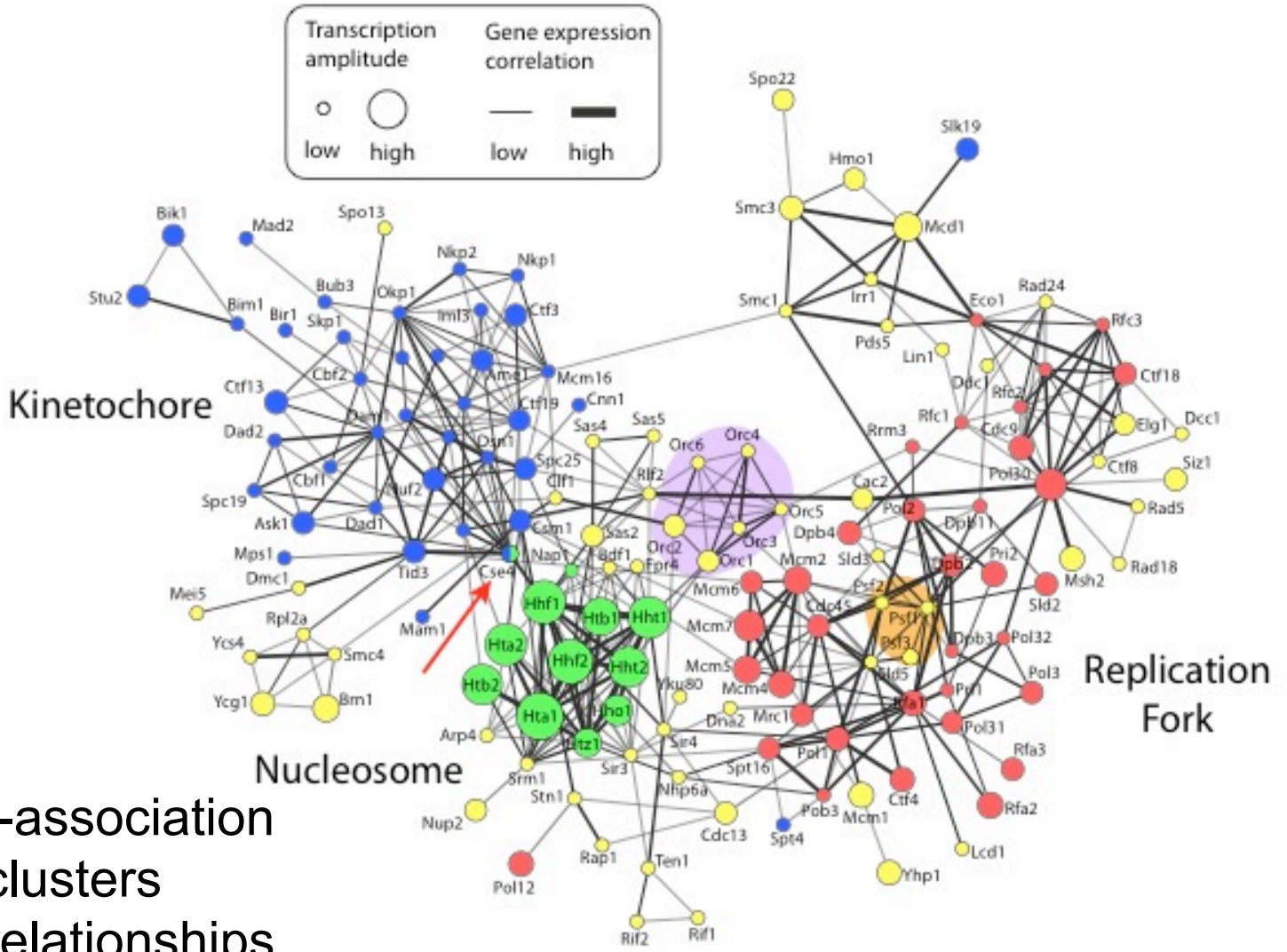
- Node and edge attributes
  - String, integer, float, Boolean, list
  - E.g. represent gene, interaction attributes
- Visual attributes
  - Node, edge visual properties
  - Colour, shape, size, borders, opacity...







Visually Interpreting a Network



Guilt-by-association
Dense clusters
Global relationships

What Have We Learned?

- Automatic layout is required to visualize networks
- Networks help you visualize interesting relationships in your data
- Avoid hairballs by focusing analysis
- Visual attributes enable multiple types of data to be shown at once – useful to see their relationships

Network Visualization and Analysis using Cytoscape

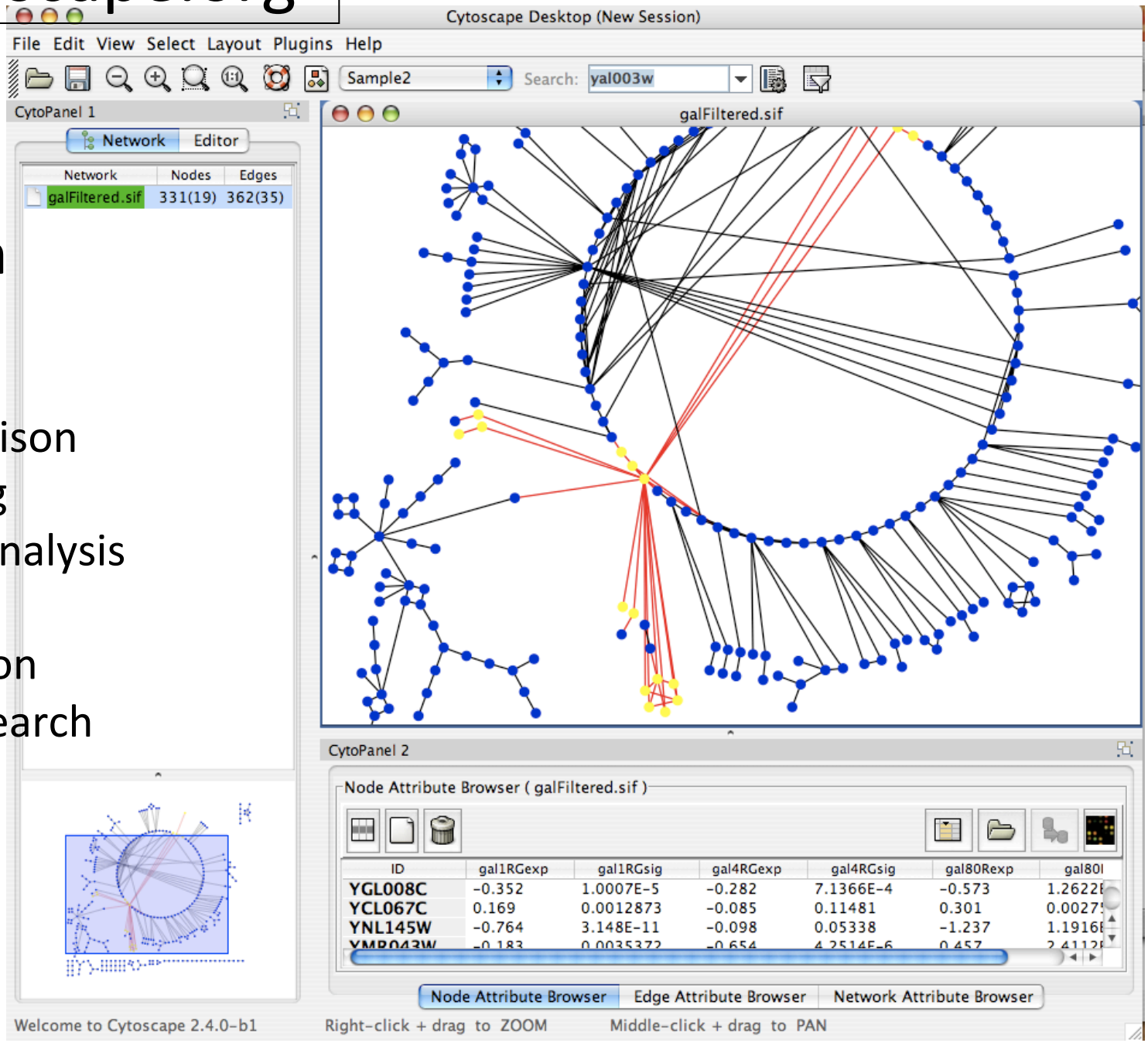
- Network visualization and analysis using Cytoscape software
- Cytoscape basics
- Cytoscape network analysis examples

http://cytoscape.org

Network visualization and analysis

Pathway comparison
Literature mining
Gene Ontology analysis
Active modules
Complex detection
Network motif search

UCSD, ISB, Agilent, MSKCC, Pasteur, UCSF, Unilever, UToronto, U Texas



Active Community

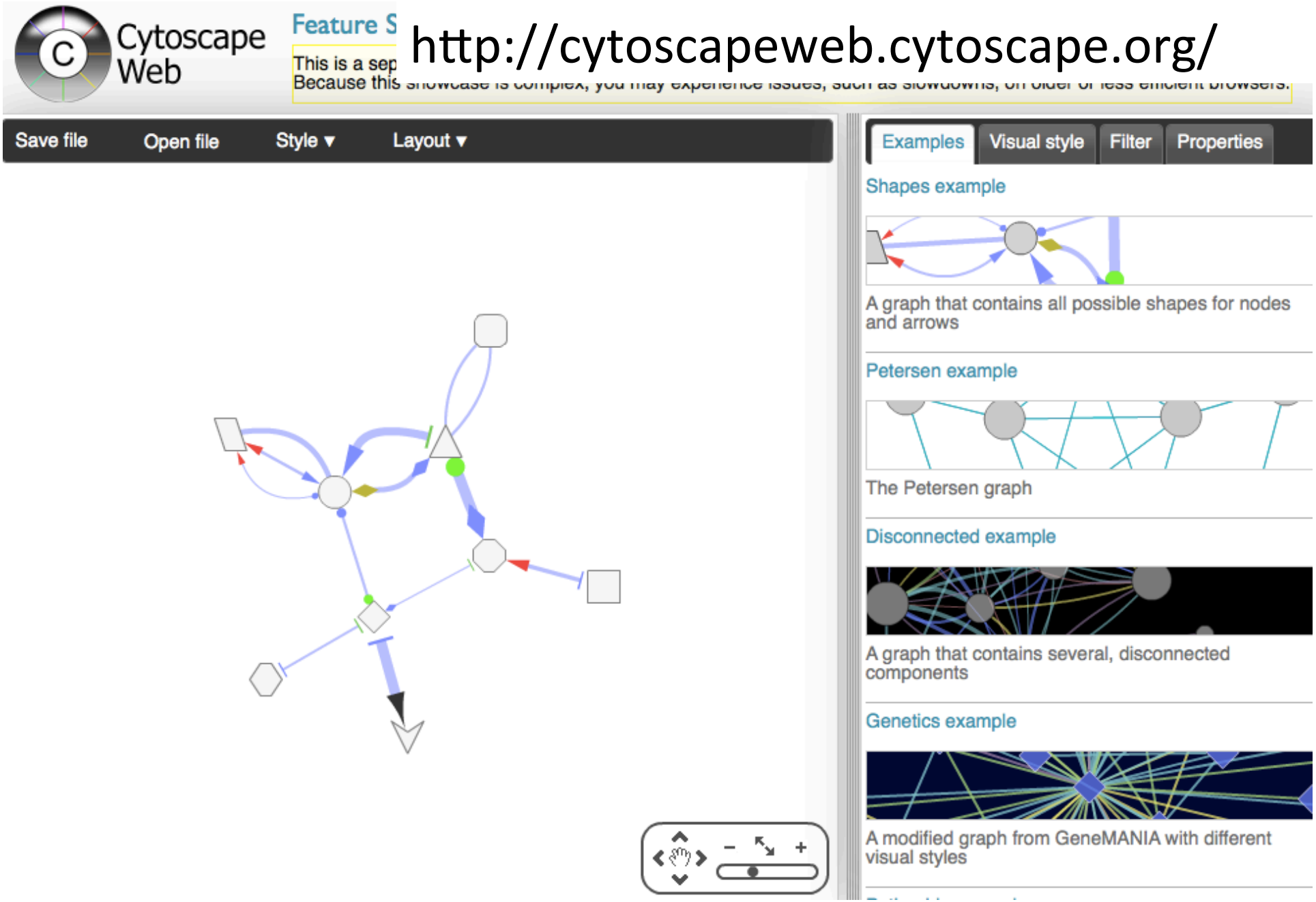
http://www.cytoscape.org

- Help
  - Tutorials, case studies
  - Mailing lists for discussion
  - Documentation, data sets
- Annual Conference: San Diego, May 18-21, 2011
- 10,000s users, 2500 downloads/month
- >100 Plugins 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


160
College Street

http://cytoscapeweb.cytoscape.org/

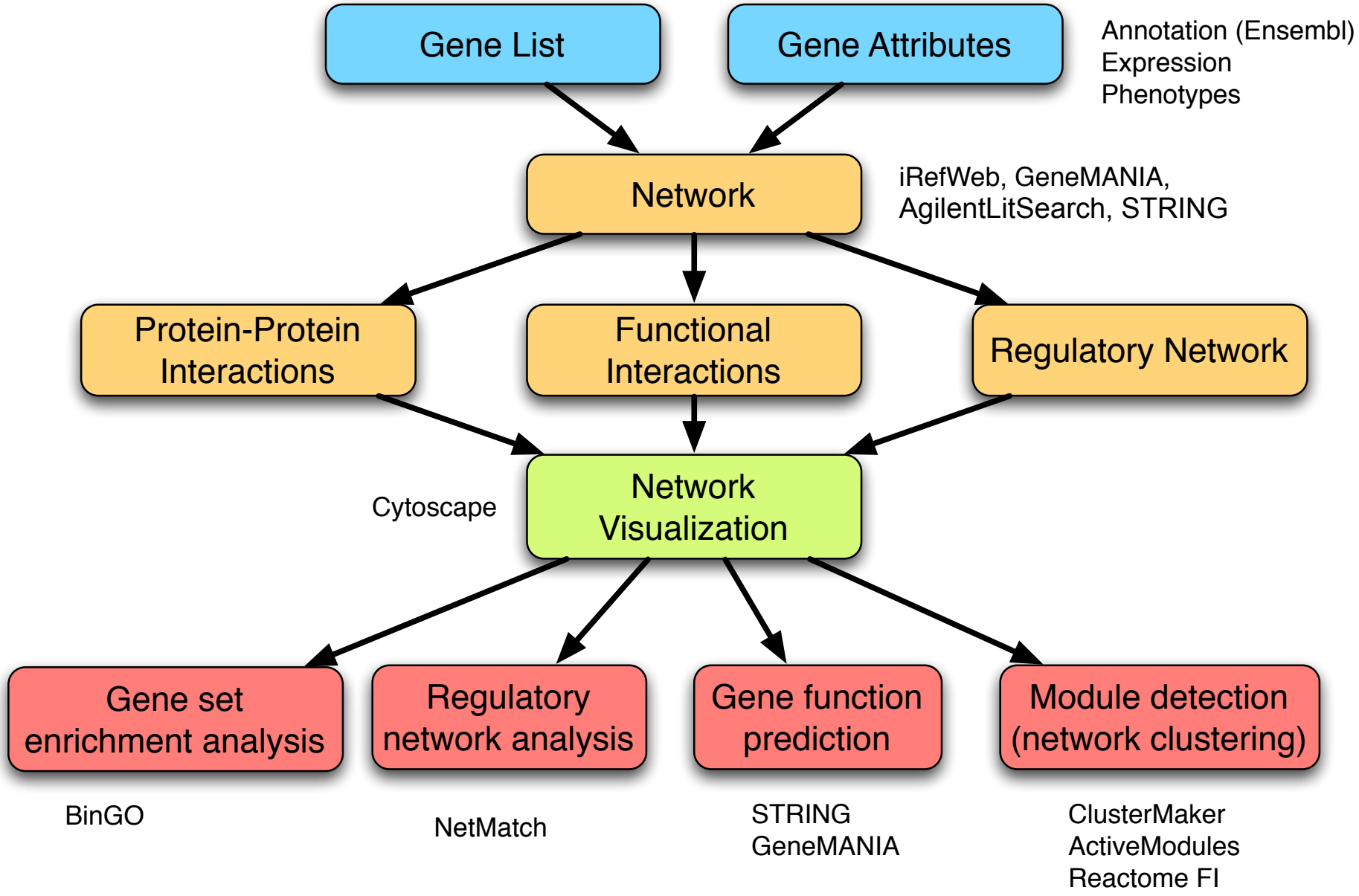


Cytoscape Demo

Version 2.8.0

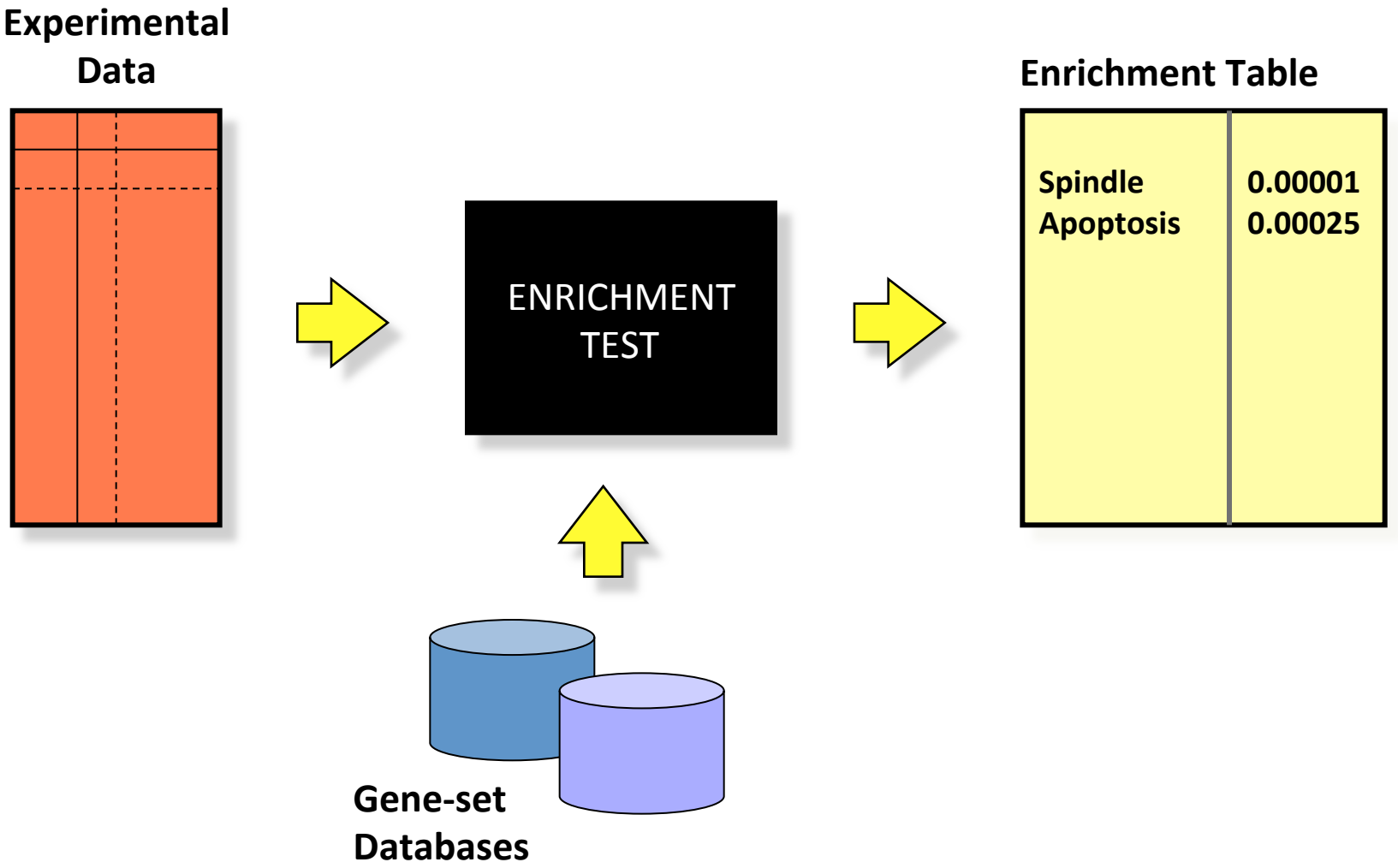
www.cytoscape.org

Gene List and Network Analysis Overview



Visualizing gene set enrichment analysis results

Enrichment Test: General Framework



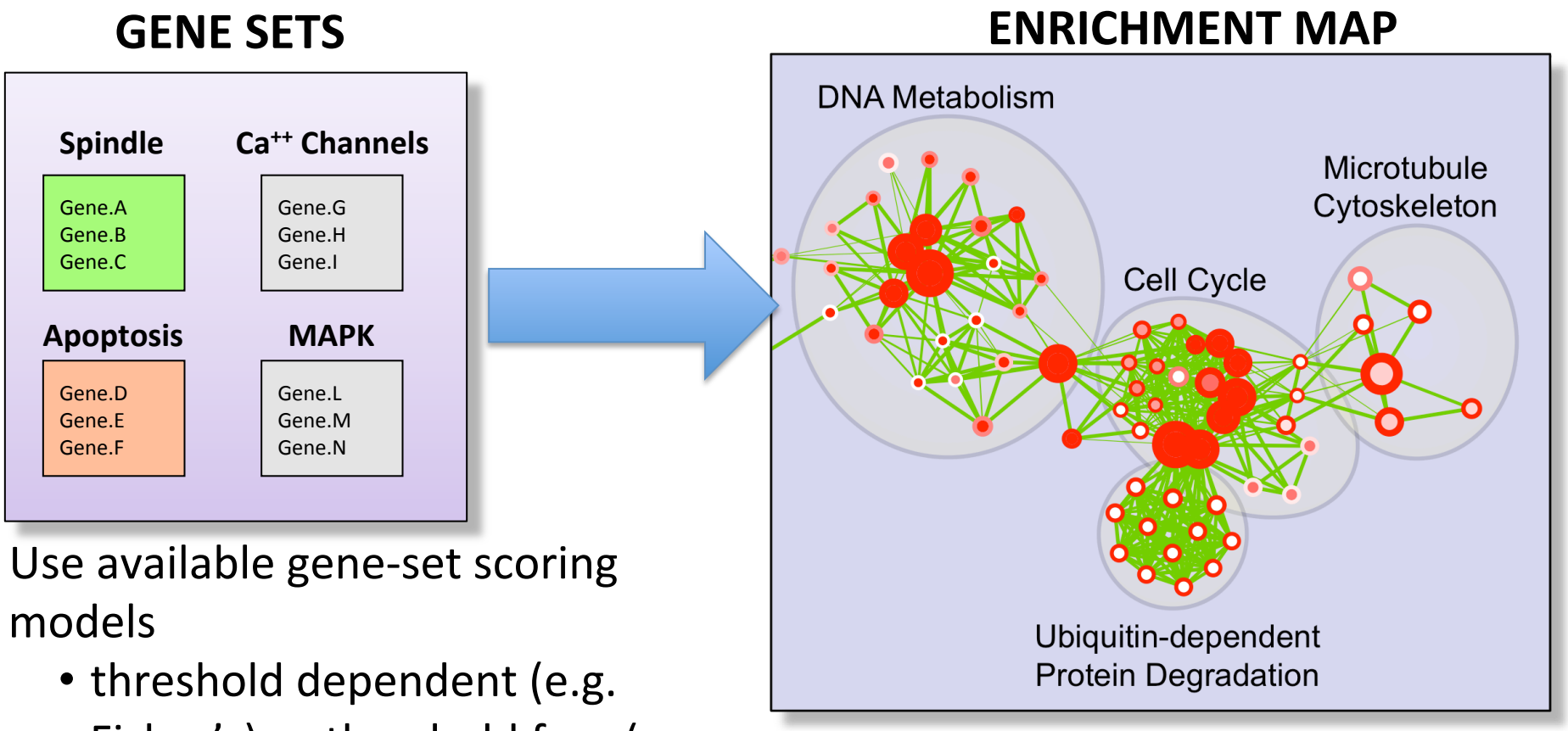
- Excellent idea used to interpret data in thousands of papers

GO.id	GO.name	p.value	cover	cover.rat	Deg.mdn	Deg.iqr
GO:0042330	taxis	2.18E-06	23	0.056930693	54.94499375	9.139238998
GO:0006935	**chemotaxis**	**2.18E-06**	**23**	**0.060209424**	**54.94499375**	**9.139238998**
GO:0002460	adaptive immune response based on somatic recombination	7.10E-05	25	0.111111111	57.32306955	16.97054864
GO:0002250	**adaptive immune response**	**7.10E-05**	**25**	**0.111111111**	**57.32306955**	**16.97054864**
GO:0002443	leukocyte mediated immunity	0.000419328	23	0.097046414	58.27890582	15.58333739
GO:0019724	**B cell mediated immunity**	**0.000683758**	**20**	**0.114285714**	**57.84161096**	**15.03496347**
GO:0030099	**myeloid cell differentiation**	**0.000691589**	**24**	**0.089219331**	**62.22171598**	**10.35284833**
GO:0002252	immune effector process	0.000775626	31	0.090116279	58.27890582	23.86214773
GO:0050764	regulation of phagocytosis	0.000792138	8	0.2	53.54786293	5.742849971
GO:0050766	positive regulation of phagocytosis	0.000792138	8	0.216216216	53.54786293	5.742849971
GO:0002449	lymphocyte mediated immunity	0.00087216	22	0.101851852	57.84161096	16.13171132
GO:0019838	**growth factor binding**	**0.000913285**	**15**	**0.068181818**	**83.0405088**	**10.58734852**
GO:0051258	protein polymerization	0.00108876	17	0.080952381	57.97543252	17.31639968
GO:0005789	**endoplasmic reticulum membrane**	**0.001178198**	**18**	**0.036072144**	**64.02284752**	**12.05209158**
GO:0016064	immunoglobulin mediated immune response	0.001444464	19	0.113095238	58.27890582	15.58333739
GO:0007507	**heart development**	**0.001991562**	**26**	**0.052313883**	**84.02538284**	**18.60761304**
GO:0009617	response to bacterium	0.002552999	10	0.027173913	52.75249873	23.23104637
GO:0030100	regulation of endocytosis	0.002658555	11	0.099099099	56.38041132	16.02486889
GO:0002526	acute inflammatory response	0.002660742	24	0.103004292	57.80098769	24.94311116
GO:0045807	positive regulation of endocytosis	0.002903401	9	0.147540984	54.94499375	6.769909171
GO:0002274	myeloid leukocyte activation	0.002969661	7	0.077777778	54.94499375	16.07042339
GO:0008652	**amino acid biosynthetic process**	**0.003502921**	**7**	**0.017241379**	**45.19797271**	**31.18248579**
GO:0050727	**regulation of inflammatory response**	**0.004999055**	**7**	**0.084337349**	**54.94499375**	**7.737346076**
GO:0002253	activation of immune response	0.00500146	23	0.116161616	60.29679989	18.41103376
GO:0002684	positive regulation of immune system process	0.006581245	27	0.111570248	60.29679989	22.05051447
GO:0050778	positive regulation of immune response	0.006581245	27	0.113924051	60.29679989	22.05051447
GO:0019882	antigen processing and presentation	0.007244488	7	0.029661017	54.94499375	16.58797889
GO:0002682	regulation of immune system process	0.007252134	29	0.099656357	61.05645008	22.65935206
GO:0050776	regulation of immune response	0.007252134	29	0.102112676	61.05645008	22.65935206
GO:0043086	negative regulation of enzyme activity	0.008017022	9	0.040723982	53.28031076	17.48904224
GO:0006909	phagocytosis	0.008106069	10	0.080645161	55.66270253	12.47536747
GO:0002573	myeloid leukocyte differentiation	0.008174948	10	0.092592593	62.86577216	9.401887596
GO:0006959	humoral immune response	0.008396095	16	0.044568245	55.05654091	18.94209565
GO:0046649	lymphocyte activation	0.009044401	29	0.059917355	61.92213317	21.03553355
GO:0030595	leukocyte chemotaxis	0.009707319	7	0.101449275	56.33116709	6.945510559
GO:0006469	negative regulation of protein kinase activity	0.010782155	7	0.046357616	52.22863516	12.58524145
GO:0051348	negative regulation of transferase activity	0.010782155	7	0.04516129	52.22863516	12.58524145
GO:0007179	**transforming growth factor beta receptor signaling pathw**	**0.012630825**	**13**	**0.071038251**	**83.49440788**	**12.63256309**
GO:0005520	**insulin-like growth factor binding**	**0.012950071**	**9**	**0.097826087**	**81.41963394**	**7.528247832**
GO:0042110	T cell activation	0.013410548	20	0.064516129	59.77891783	26.06174863
GO:0002455	humoral immune response mediated by circulating immunogl	0.016780163	10	0.125	54.70766244	14.2572143
GO:0005830	**cytosolic ribosome (sensu Eukaryota)**	**0.016907351**	**8**	**0.01843318**	**61.68933284**	**7.814673781**

- Excellent idea used to interpret data in thousands of papers
  - But! Major cognitive burden relating overlapping gene sets

GO.id	GO.name	p.value	cover	cover.rat	Deg.mdn	Deg.iqr
GO:0042330	taxis	2.18E-06	23	0.056930693	54.94499375	9.139238998
GO:0006935	**chemotaxis**	**2.18E-06**	**23**	**0.060209424**	**54.94499375**	**9.139238998**
GO:0002460	adaptive immune response based on somatic recombination	7.10E-05	25	0.111111111	57.32306955	16.97054864
GO:0002250	**adaptive immune response**	**7.10E-05**	**25**	**0.111111111**	**57.32306955**	**16.97054864**
GO:0002443	leukocyte mediated immunity	0.000419328	23	0.097046414	58.27890582	15.58333739
GO:0019724	**B cell mediated immunity**	**0.000683758**	**20**	**0.114285714**	**57.84161096**	**15.03496347**
GO:0030099	**myeloid cell differentiation**	**0.000691589**	**24**	**0.089219331**	**62.22171598**	**10.35284833**
GO:0002252	immune effector process	0.000775626	31	0.090116279	58.27890582	23.86214773
GO:0050764	regulation of phagocytosis	0.000792138	8	0.2	53.54786293	5.742849971
GO:0050766	positive regulation of phagocytosis	0.000792138	8	0.216216216	53.54786293	5.742849971
GO:0002449	lymphocyte mediated immunity	0.00087216	22	0.101851852	57.84161096	16.13171132
GO:0019838	**growth factor binding**	**0.000913285**	**15**	**0.068181818**	**83.0405088**	**10.58734852**
GO:0051258	protein polymerization	0.00108876	17	0.080952381	57.97543252	17.31639968
GO:0005789	**endoplasmic reticulum membrane**	**0.001178198**	**18**	**0.036072144**	**64.02284752**	**12.05209158**
GO:0016064	immunoglobulin mediated immune response	0.001444464	19	0.113095238	58.27890582	15.58333739
GO:0007507	**heart development**	**0.001991562**	**26**	**0.052313883**	**84.02538284**	**18.60761304**
GO:0009617	response to bacterium	0.002552999	10	0.027173913	52.75249873	23.23104637
GO:0030100	regulation of endocytosis	0.002658555	11	0.099099099	56.38041132	16.02486889
GO:0002526	acute inflammatory response	0.002660742	24	0.103004292	57.80098769	24.94311116
GO:0045807	positive regulation of endocytosis	0.002903401	9	0.147540984	54.94499375	6.769909171
GO:0002274	myeloid leukocyte activation	0.002969661	7	0.077777778	54.94499375	16.07042339
GO:0008652	**amino acid biosynthetic process**	**0.003502921**	**7**	**0.017241379**	**45.19797271**	**31.18248579**
GO:0050727	**regulation of inflammatory response**	**0.004999055**	**7**	**0.084337349**	**54.94499375**	**7.737346076**
GO:0002253	activation of immune response	0.00500146	23	0.116161616	60.29679989	18.41103376
GO:0002684	positive regulation of immune system process	0.006581245	27	0.111570248	60.29679989	22.05051447
GO:0050778	positive regulation of immune response	0.006581245	27	0.113924051	60.29679989	22.05051447
GO:0019882	antigen processing and presentation	0.007244488	7	0.029661017	54.94499375	16.58797889
GO:0002682	regulation of immune system process	0.007252134	29	0.099656357	61.05645008	22.65935206
GO:0050776	regulation of immune response	0.007252134	29	0.102112676	61.05645008	22.65935206
GO:0043086	negative regulation of enzyme activity	0.008017022	9	0.040723982	53.28031076	17.48904224
GO:0006909	phagocytosis	0.008106069	10	0.080645161	55.66270253	12.47536747
GO:0002573	myeloid leukocyte differentiation	0.008174948	10	0.092592593	62.86577216	9.401887596
GO:0006959	humoral immune response	0.008396095	16	0.044568245	55.05654091	18.94209565
GO:0046649	lymphocyte activation	0.009044401	29	0.059917355	61.92213317	21.03553355
GO:0030595	leukocyte chemotaxis	0.009707319	7	0.101449275	56.33116709	6.945510559
GO:0006469	negative regulation of protein kinase activity	0.010782155	7	0.046357616	52.22863516	12.58524145
GO:0051348	negative regulation of transferase activity	0.010782155	7	0.04516129	52.22863516	12.58524145
GO:0007179	**transforming growth factor beta receptor signaling pathw**	**0.012630825**	**13**	**0.071038251**	**83.49440788**	**12.63256309**
GO:0005520	**insulin-like growth factor binding**	**0.012950071**	**9**	**0.097826087**	**81.41963394**	**7.528247832**
GO:0042110	T cell activation	0.013410548	20	0.064516129	59.77891783	26.06174863
GO:0002455	humoral immune response mediated by circulating immunogl	0.016780163	10	0.125	54.70766244	14.2572143
GO:0005830	**cytosolic ribosome (sensu Eukaryota)**	**0.016907351**	**8**	**0.01843318**	**61.68933284**	**7.814673781**

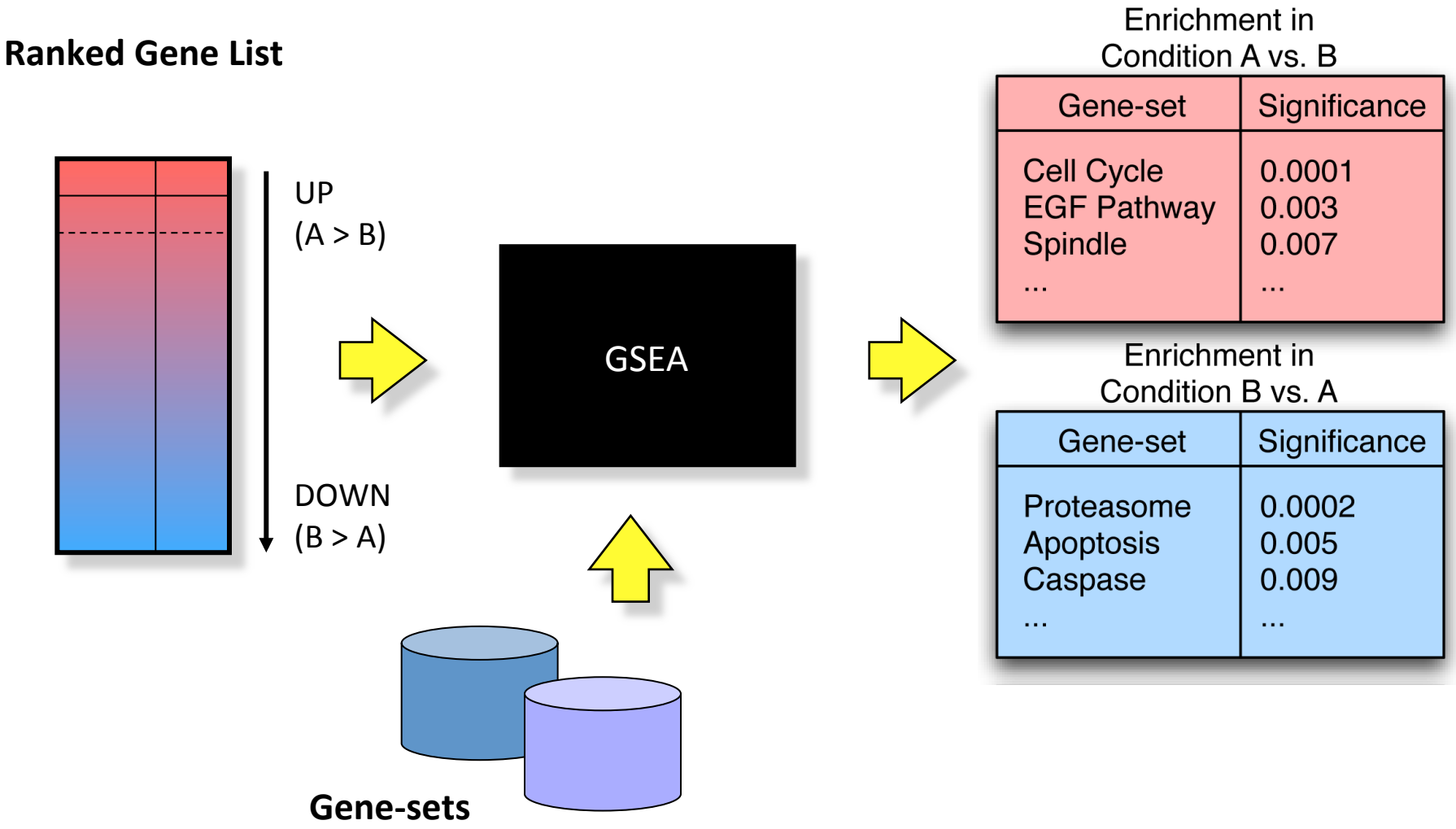
Enrichment Map



- Use available gene-set scoring models
  - threshold dependent (e.g. Fisher's) or threshold free (e.g. GSEA)
- Use the network framework to organize gene-sets exploiting their inter-dependencies

http://baderlab.org/Software/EnrichmentMap/

Gene Set Enrichment Analysis (GSEA)



Enrichment Map

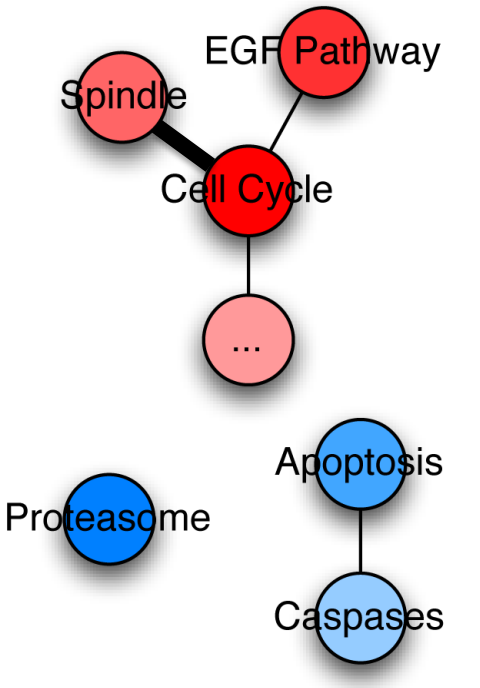
Enrichment in Condition A vs. B

Gene-set	Significance
Cell Cycle	0.0001
EGF Pathway	0.003
Spindle	0.007
...	...

Enrichment in Condition B vs. A

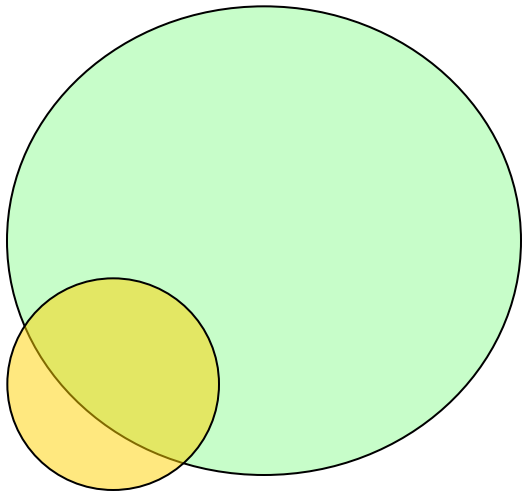
Gene-set	Significance
Proteasome	0.0002
Apoptosis	0.005
Caspase	0.009
...	...

GENE-SET LIST



ENRICHMENT MAP

Overlap



$$\frac{|A \cap B|}{\min(|A|,|B|)}$$

Enrichment Map: use case I
Single enrichment

Estrogen treatment of breast cancer cells

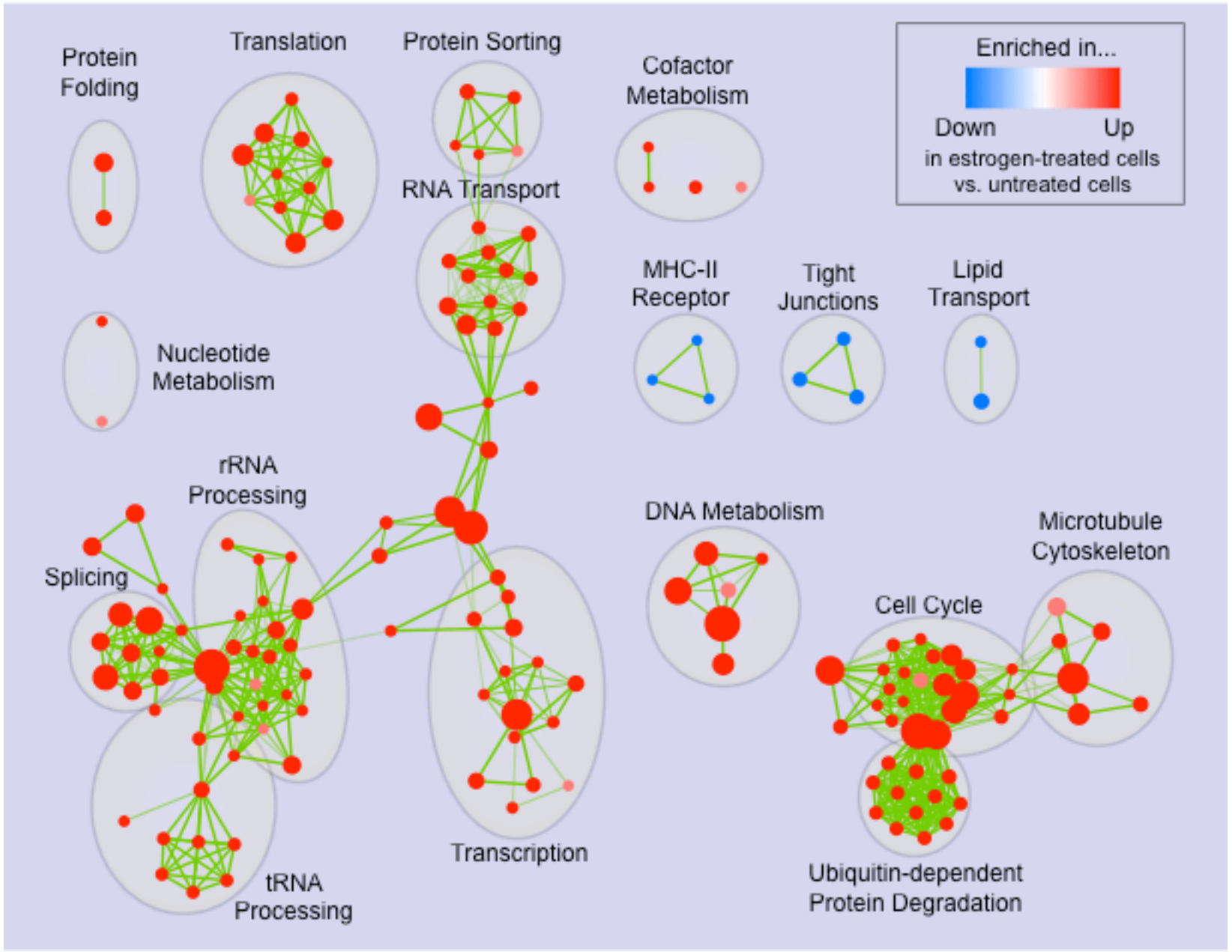
- Design:

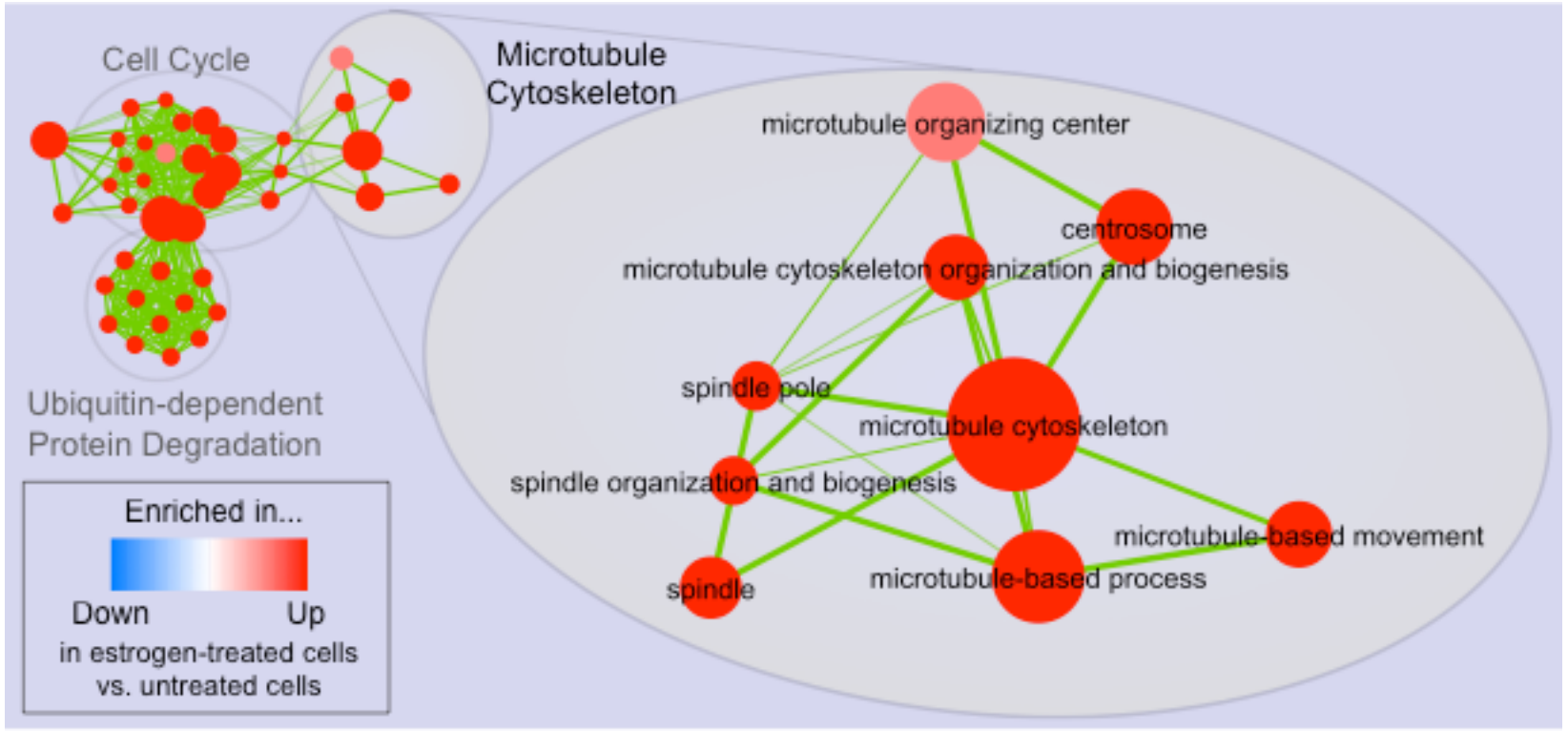
  2-time points, two-class

	12 hrs	**24 hrs**
Estrogen-treated	3	3
Untreated	3	3

- Gene set Database:

  Gene Ontology


Protein Folding
Translation
Protein Sorting
Cofactor Metabolism
Enriched in...
Down
Up
in estrogen-treated cells vs. untreated cells
RNA Transport
MHC-II Receptor
Tight Junctions
Lipid Transport
Nucleotide Metabolism
rRNA Processing
Splicing
DNA Metabolism
Microtubule Cytoskeleton
Cell Cycle
Transcription
tRNA Processing
Ubiquitin-dependent Protein Degradation


Cell Cycle
Ubiquitin-dependent
Protein Degradation
Microtubule
Cytoskeleton
microtubule organizing center
centrosome
microtubule cytoskeleton organization and biogenesis
spindle pole
microtubule cytoskeleton
spindle organization and biogenesis
microtubule-based movement
spindle
microtubule-based process
Enriched in...
Down
Up
in estrogen-treated cells
vs. untreated cells

Enrichment Map: use case II
Comparison of two enrichments

Estrogen treatment of breast cancer cells

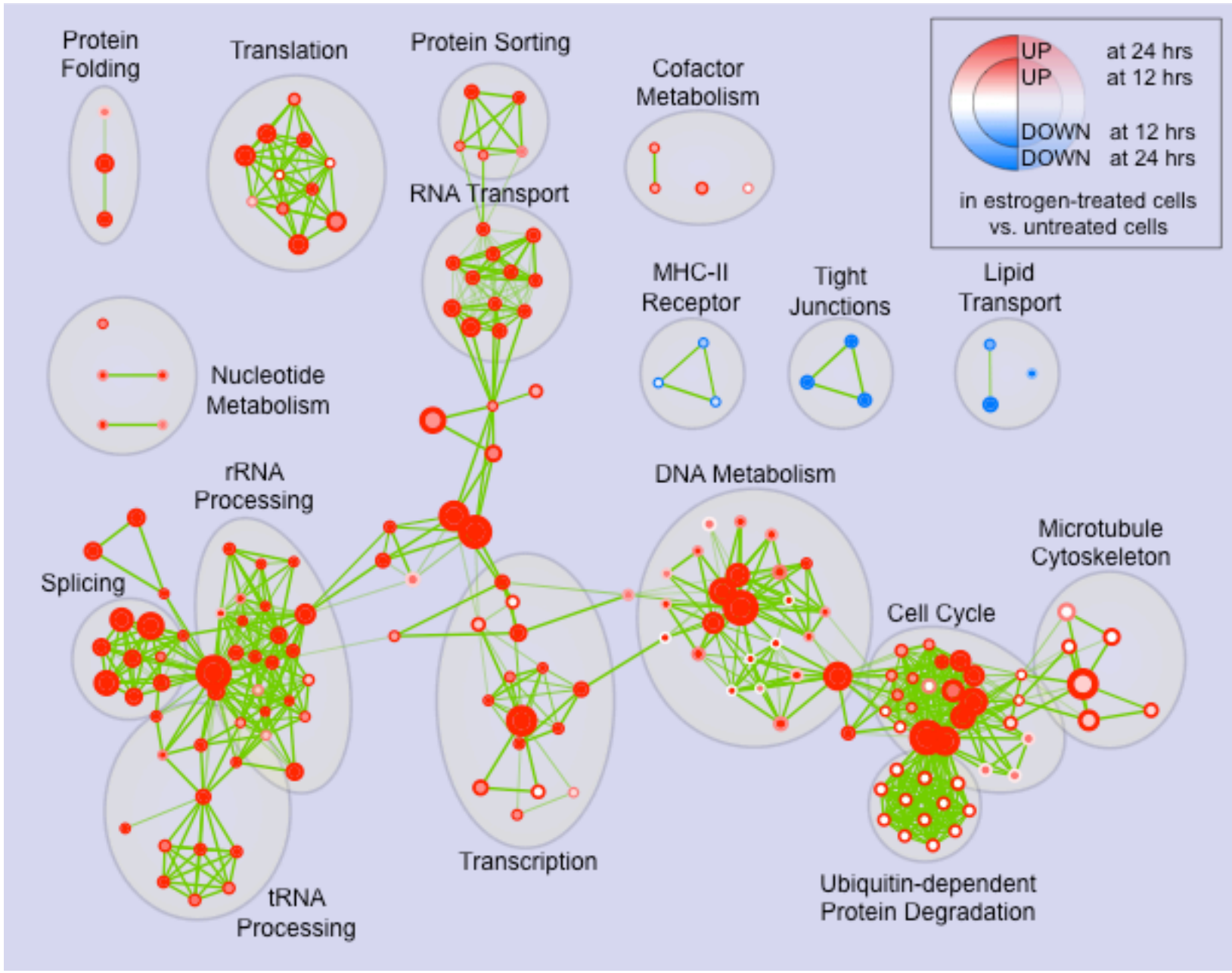
- Design:

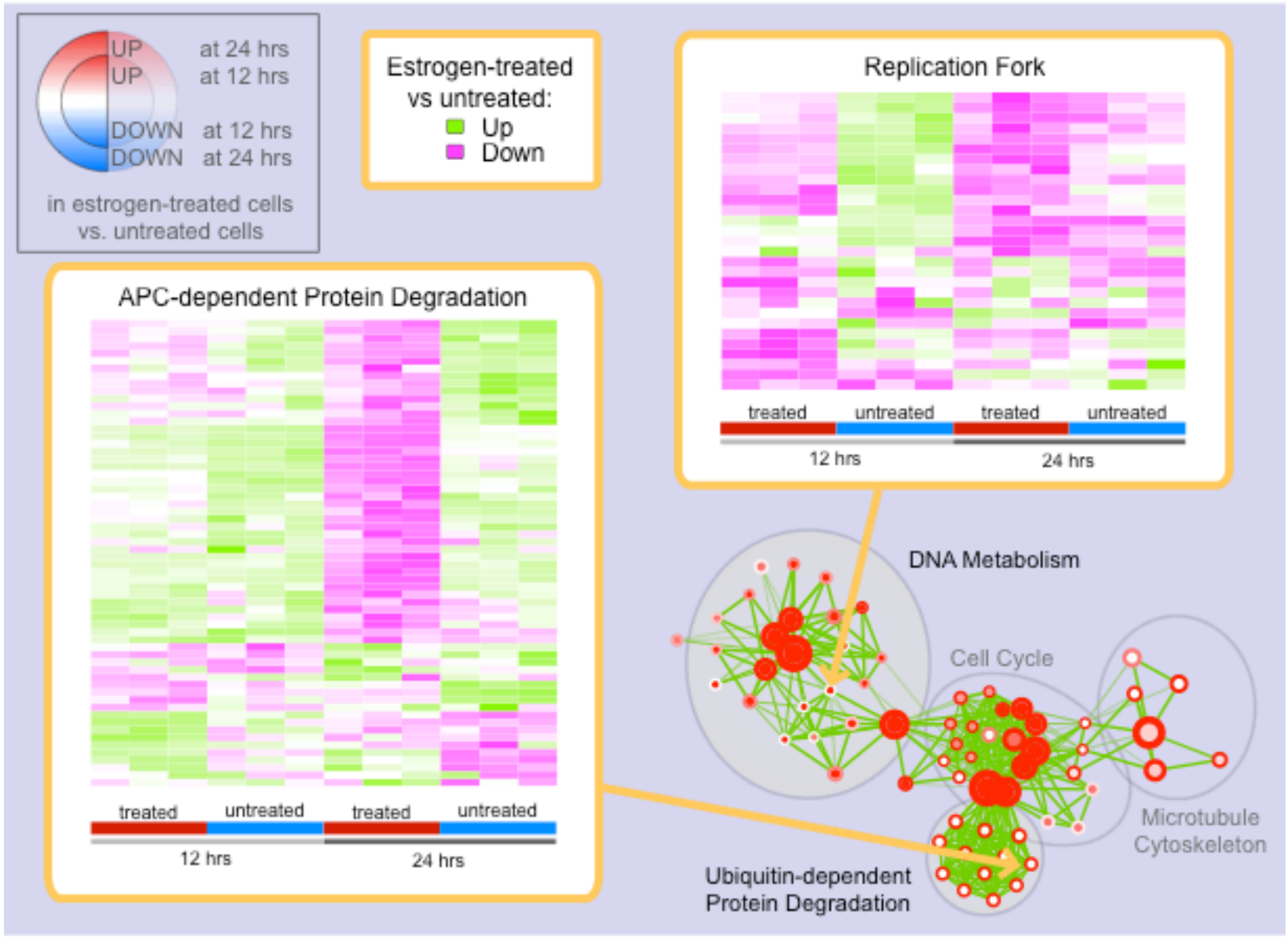
  2-time points, two-class

	12 hrs	**24 hrs**
Estrogen-treated	3	3
Untreated	3	3

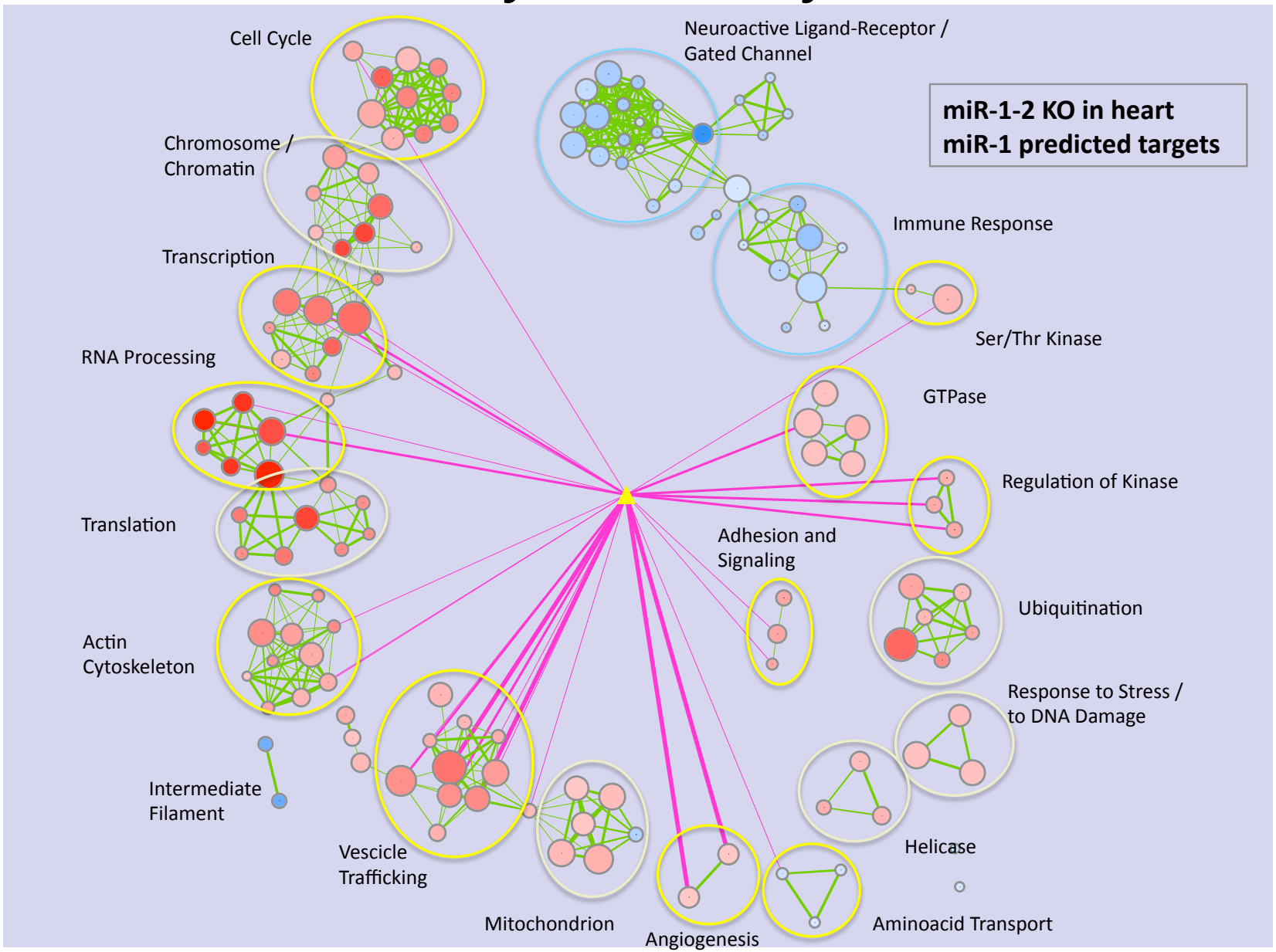
- Gene set Database:

  Gene Ontology

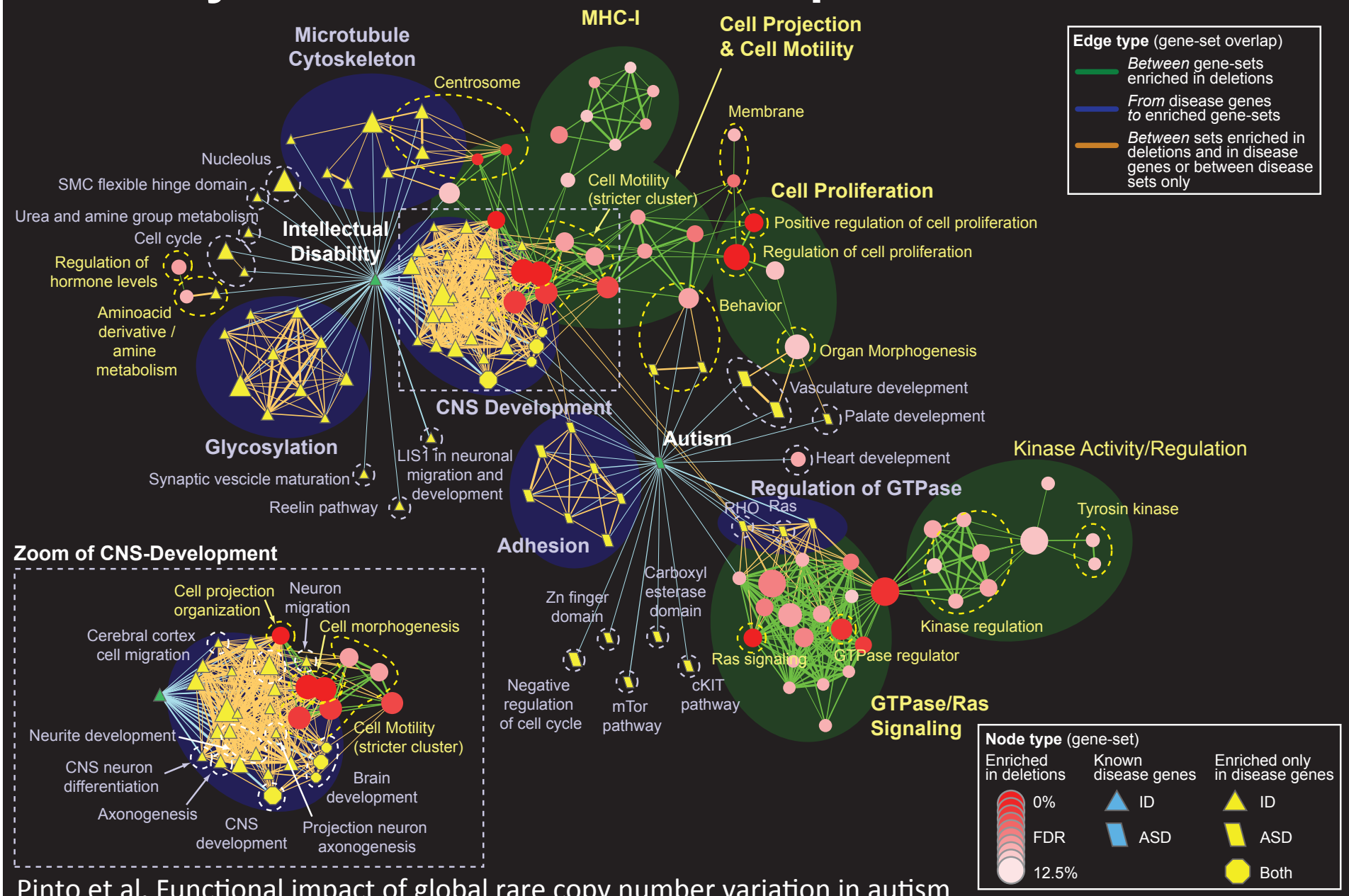

Protein Folding
Translation
Protein Sorting
Cofactor Metabolism
RNA Transport
UP at 24 hrs
UP at 12 hrs
DOWN at 12 hrs
DOWN at 24 hrs
in estrogen-treated cells vs. untreated cells
MHC-II Receptor
Tight Junctions
Lipid Transport
Nucleotide Metabolism
rRNA Processing
DNA Metabolism
Microtubule Cytoskeleton
Splicing
Cell Cycle
Transcription
Ubiquitin-dependent Protein Degradation
tRNA Processing


UP at 24 hrs
UP at 12 hrs
DOWN at 12 hrs
DOWN at 24 hrs
in estrogen-treated cells vs. untreated cells
Estrogen-treated vs untreated:
Up
Down
Replication Fork
treated
untreated
treated
untreated
12 hrs
24 hrs
APC-dependent Protein Degradation
treated
untreated
treated
untreated
12 hrs
24 hrs
DNA Metabolism
Cell Cycle
Microtubule Cytoskeleton
Ubiquitin-dependent Protein Degradation

Enrichment Map: use case III
Query Set Analysis

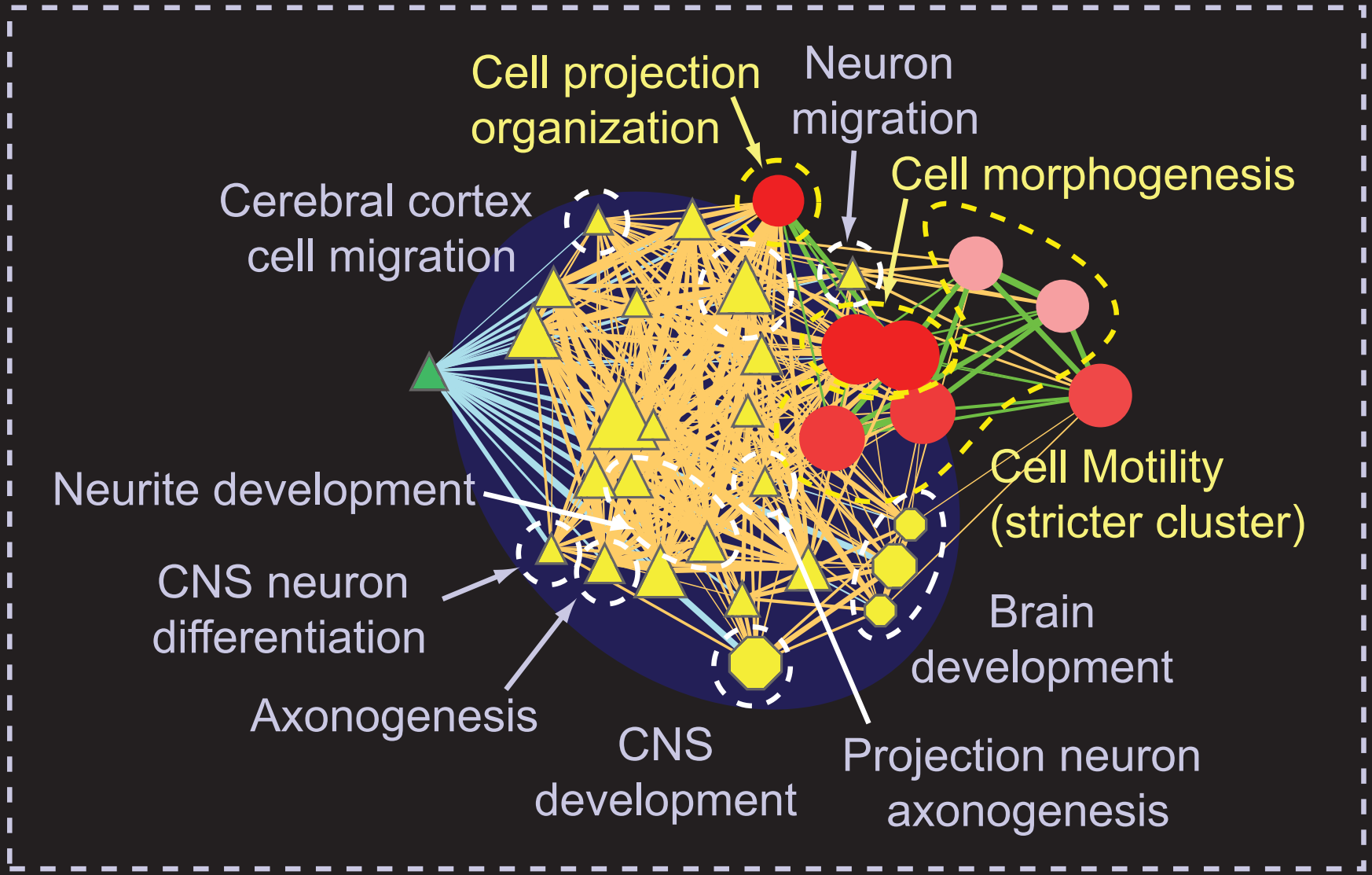


Pathways Enriched in Autism Spectrum Disorder



Pinto et al. Functional impact of global rare copy number variation in autism spectrum disorders. Nature. 2010 Jun 9.

Zoom of CNS-Development



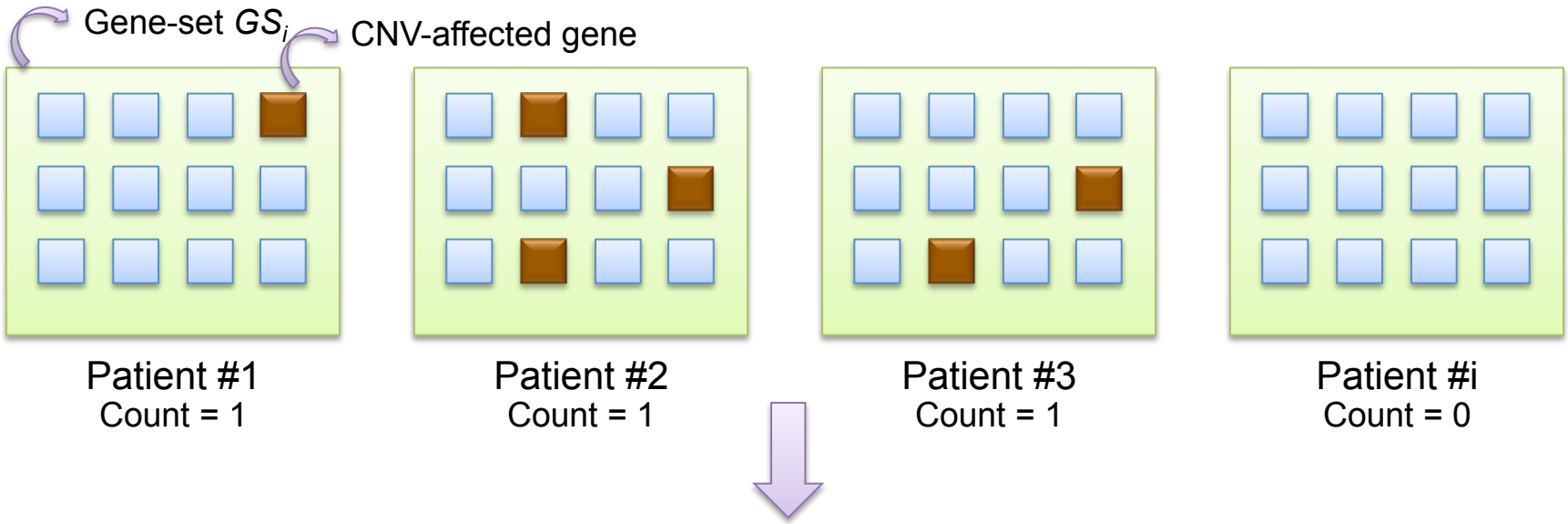
Gene-set sources

- **Gene Ontology**
  - Biological Process
  - Cellular Component
  - Molecular Function
- **Pathways**
  - KEGG
  - NCI
  - Reactome
- **PFAM domains**

- Number of gene-sets:
  - Unfiltered (all): 14,433
  - Filtered (5 << 700 genes): 6,129
  - Tested (counts > 0): 3,493

Pinto et al. Functional impact of global rare copy number variation in autism spectrum disorders. Nature. 2010 Jun 9.

Gene-set test



	Patient #1	Patient #2	Patient #3	…	Patient #i	…	Patient #n
GS_1	**1**	**1**	**1**	**…**	**0**	**…**	**0**
GS_2	0	0	1	…	1	…	0
GS_3	0	0	0	…	0	…	0

- If we have at least one CNV affecting at least one gene in a certain gene-set G_i, then we have a **perturbation potential** in that gene-set
- We count the presence / absence of such perturbation potential in patients

Gene-set test

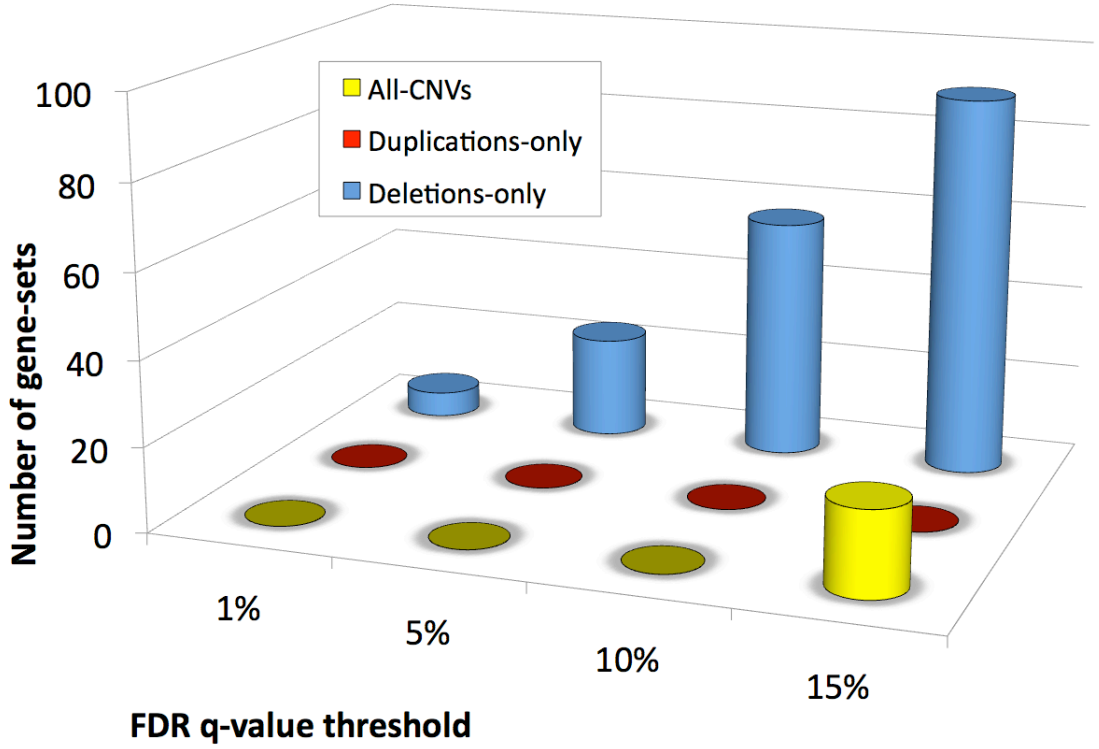
	Patient #1	Patient #2	Patient #3	…	Patient #i	…	Patient #n
GS_1	1	1	1	…	0	…	0
GS_2	0	0	1	…	1	…	0
GS_3	0	0	0	…	0	…	0

	Case	Control
GS_i	13	1
Not in GS_i	1146 - 13	889 - 1

Description:

- The significance of a gene-set is then assessed using the Fisher's Exact Test for association
- A *significant* gene-set is affected by a mutation potential *more frequently in cases than controls*
- The FDR is estimated by shuffling the columns in the 'Gene-set by patient' count table

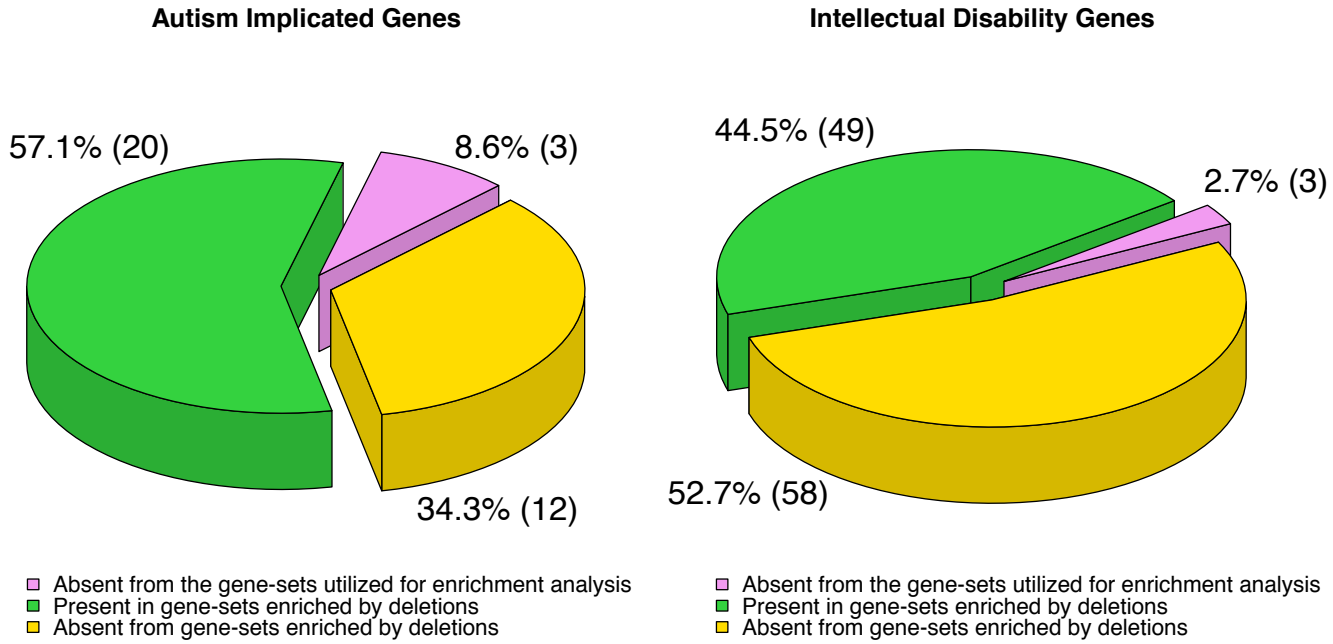
Enriched gene sets from deletions

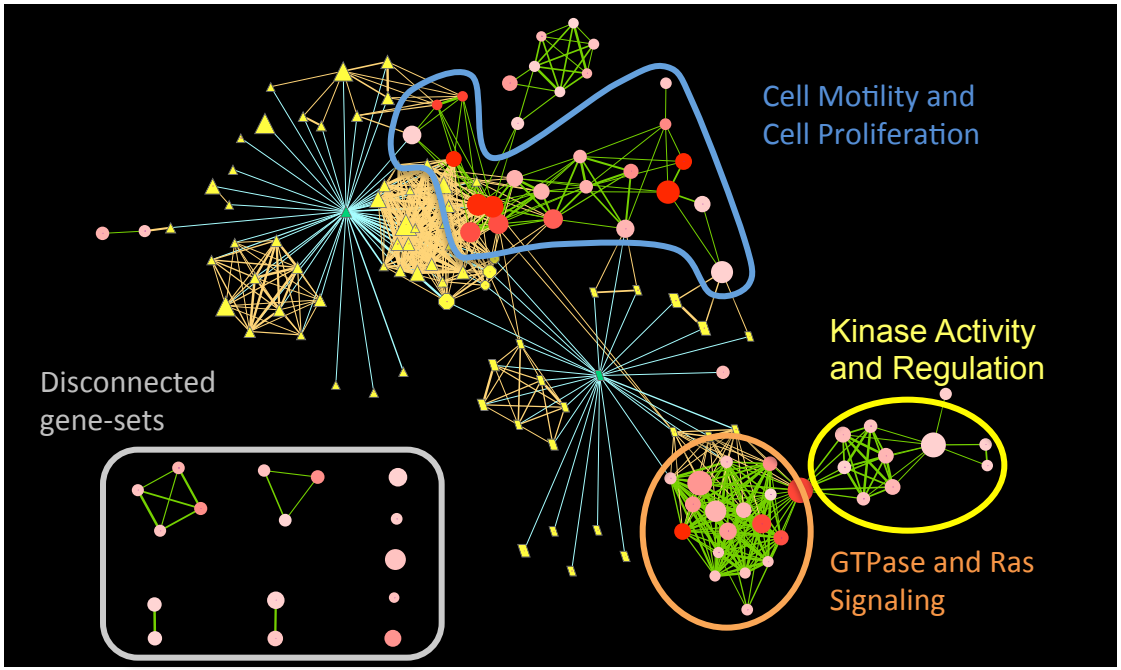


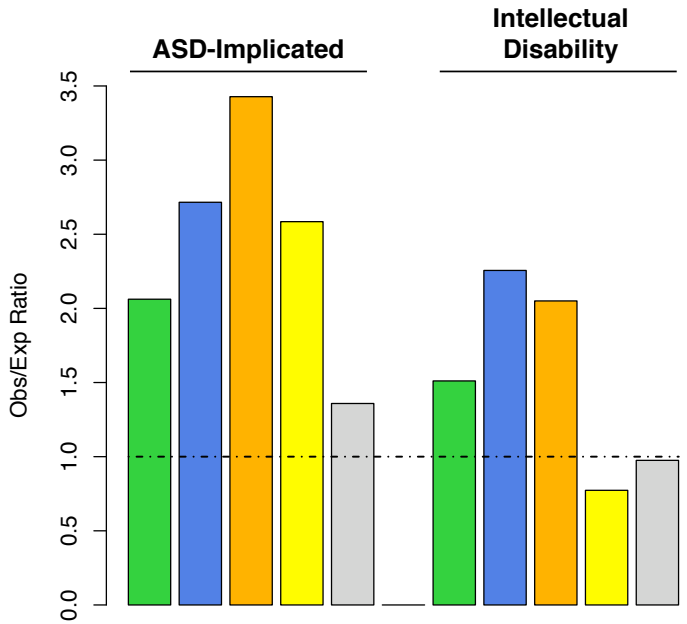
DEL: **7.6%**
gene-set x sample count cells > 1

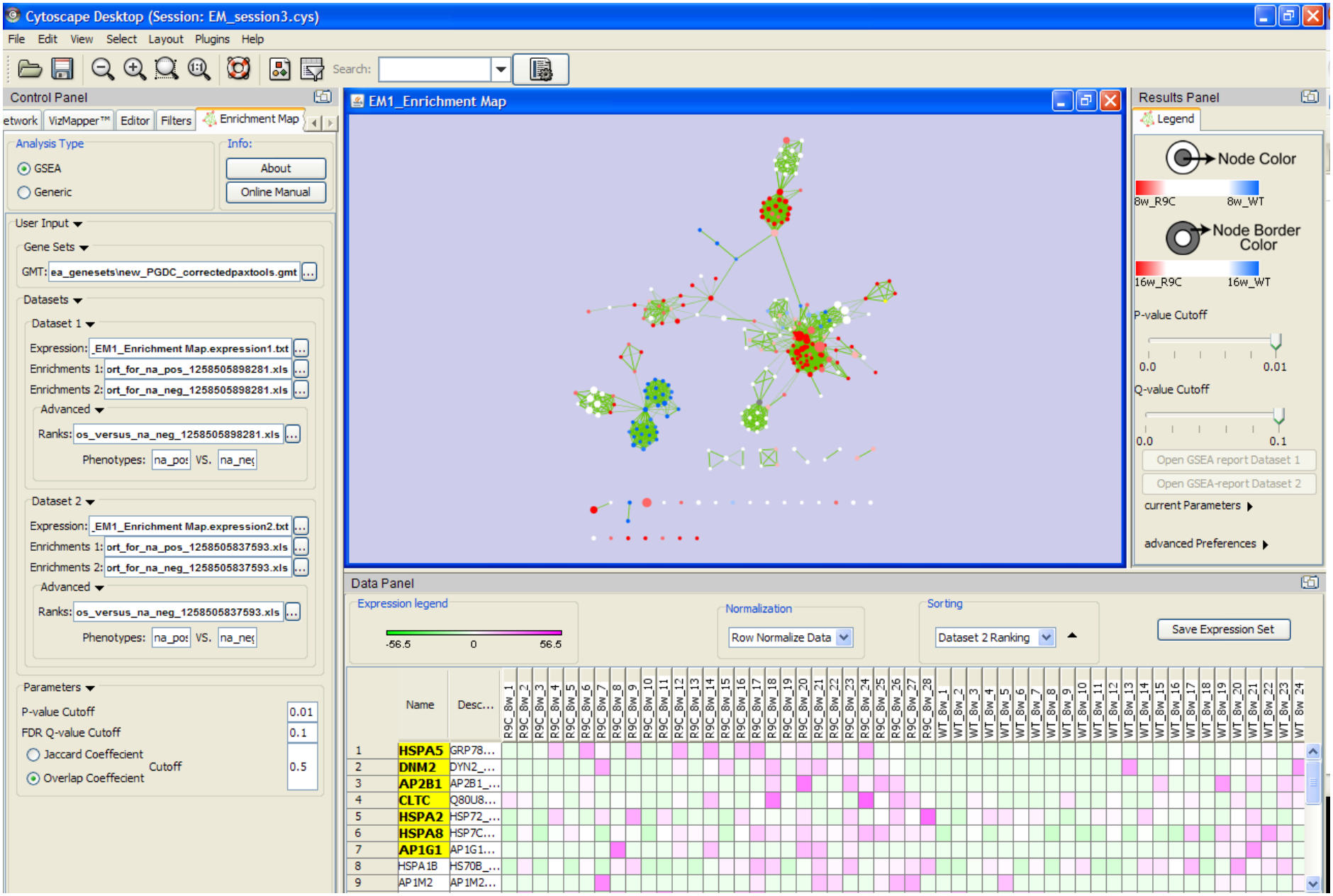
	CNV# Case	CNV# Control	CNV# / Sample# Case	CNV# / Sample# Control	Difference %
ALL	2382	3096	2.68	2.70	-0.21%
ALL (genes)	1451	1834	1.63	1.60	0.49%
DEL	1229	1527	1.38	1.33	0.92%
DEL (genes)	629	717	0.71	0.63	**3.07%**
DUP	1153	1569	1.30	1.37	-1.35%
DUP (genes)	822	1117	0.92	0.97	-1.32%

Case#	889
Ctrl#	1146


Autism Implicated Genes
57.1% (20)
8.6% (3)
34.3% (12)
Absent from the gene-sets utilized for enrichment analysis
Present in gene-sets enriched by deletions
Absent from gene-sets enriched by deletions
Intellectual Disability Genes
44.5% (49)
2.7% (3)
52.7% (58)
Absent from the gene-sets utilized for enrichment analysis
Present in gene-sets enriched by deletions
Absent from gene-sets enriched by deletions


Cell Motility and
Cell Proliferation
Kinase Activity
and Regulation
GTPase and Ras
Signaling
Disconnected
gene-sets

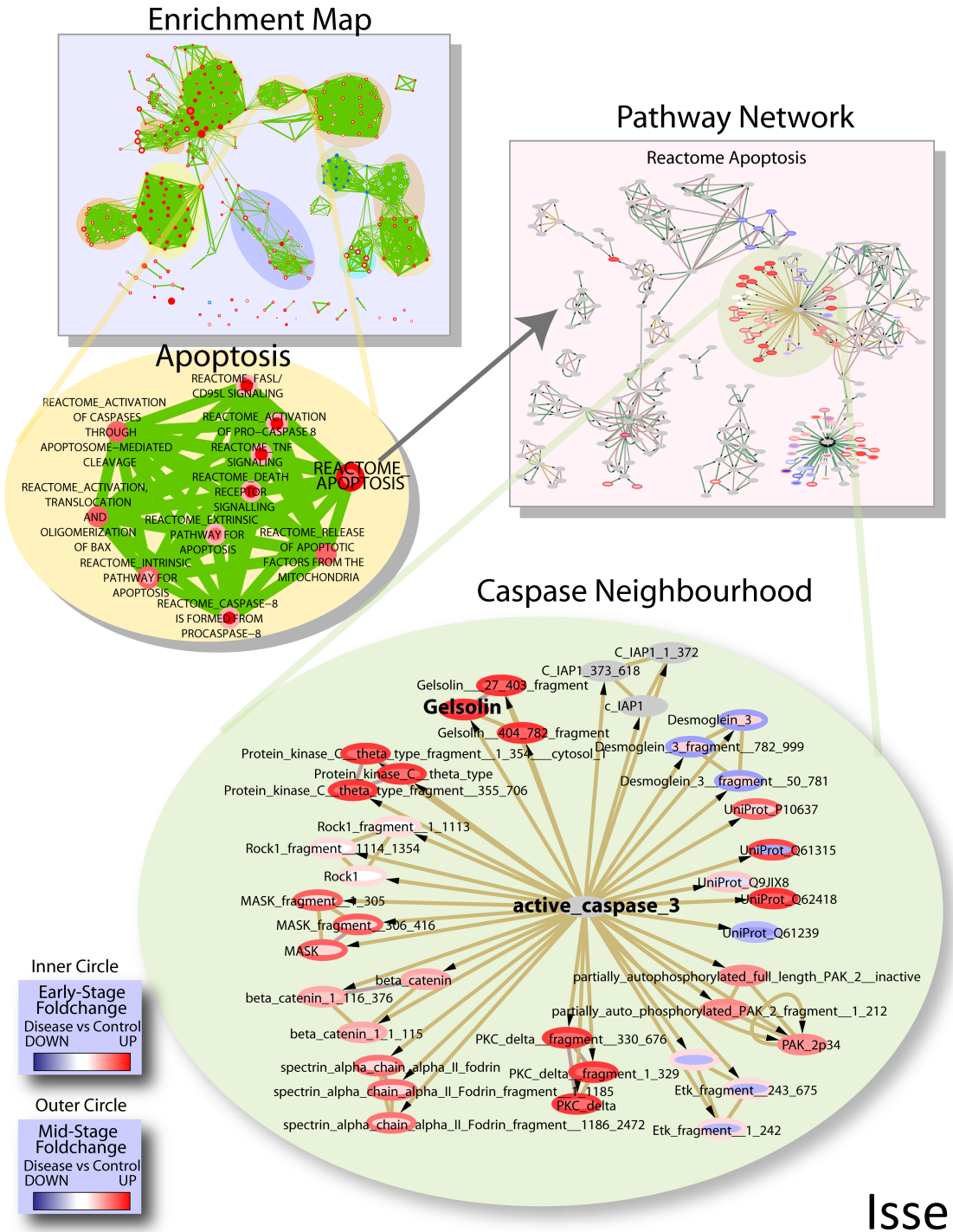

ASD-Implicated
Intellectual
Disability
Obs/Exp Ratio
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5



http://baderlab.org/Software/EnrichmentMap/

Enrichment Map Lab

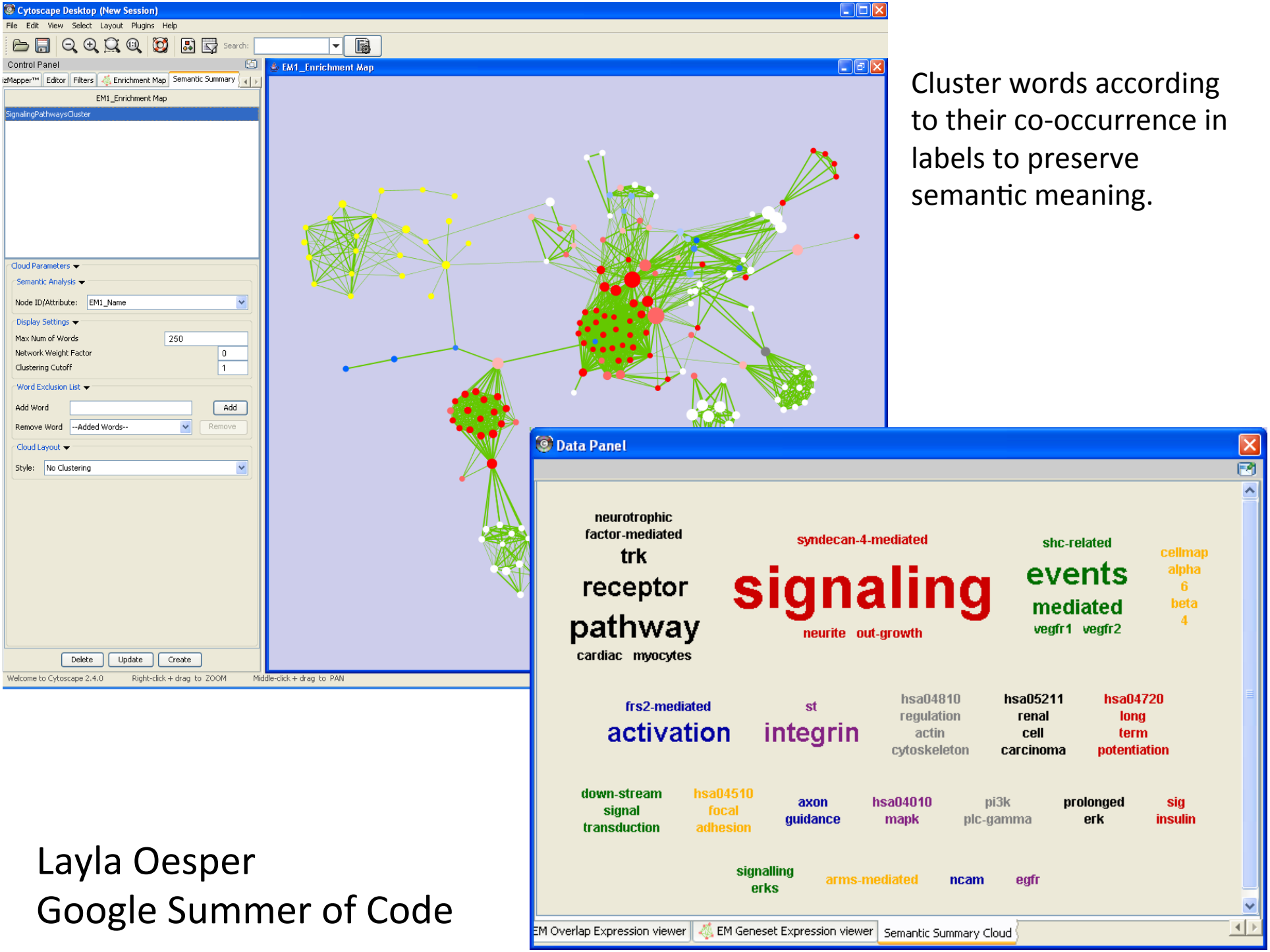
- Try out enrichment map – load the plugin from the plugin manager
- Load DAVID results – or - load the GSEA enrichment analysis file - EM_EstrogenMCF7_TestData.zip (unzip) available at
  - http://baderlab.org/Software/EnrichmentMap



Future Work

- Add network visualization support
- Pathway visualization and analysis tools

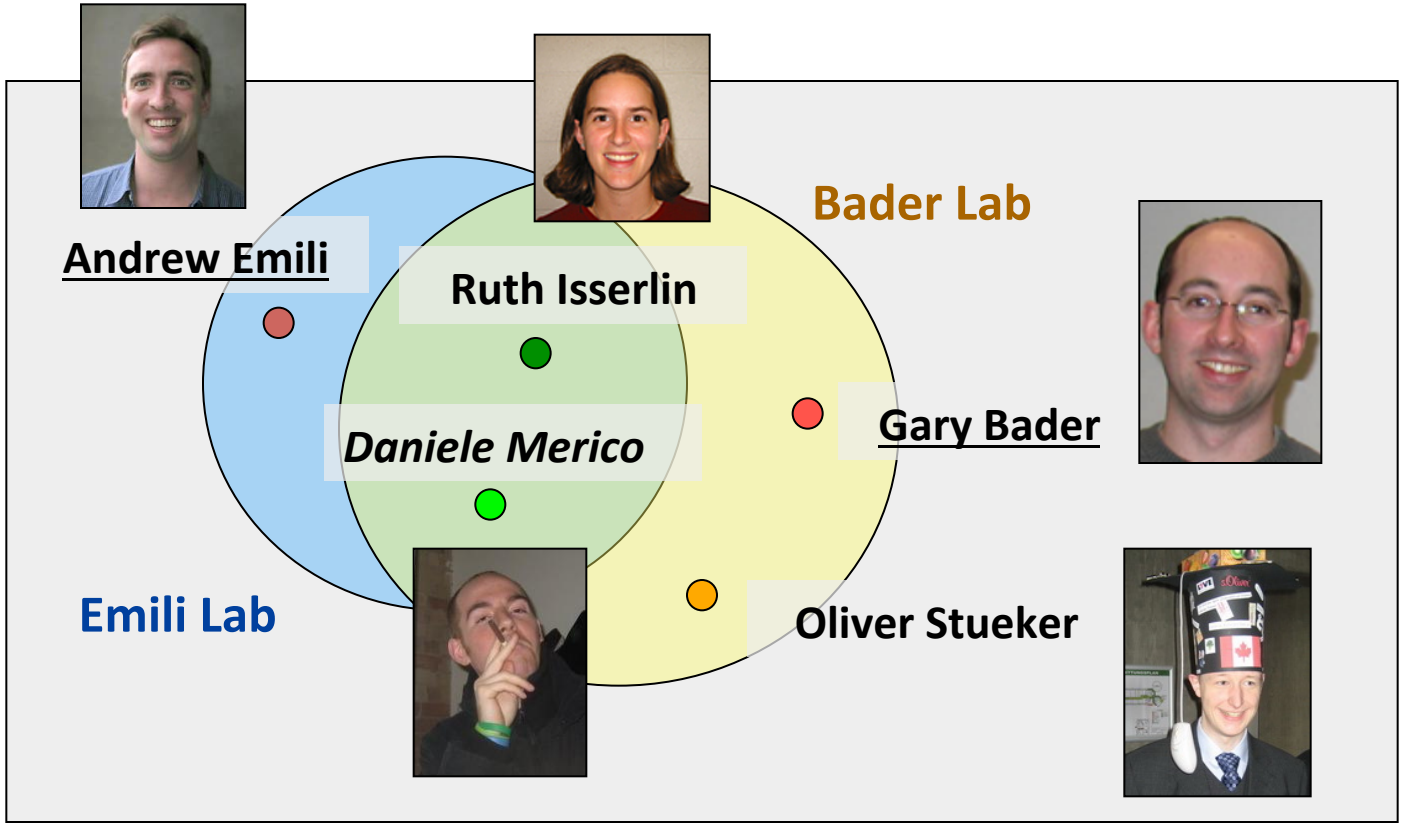
Isserlin et al. , Proteomics 2010



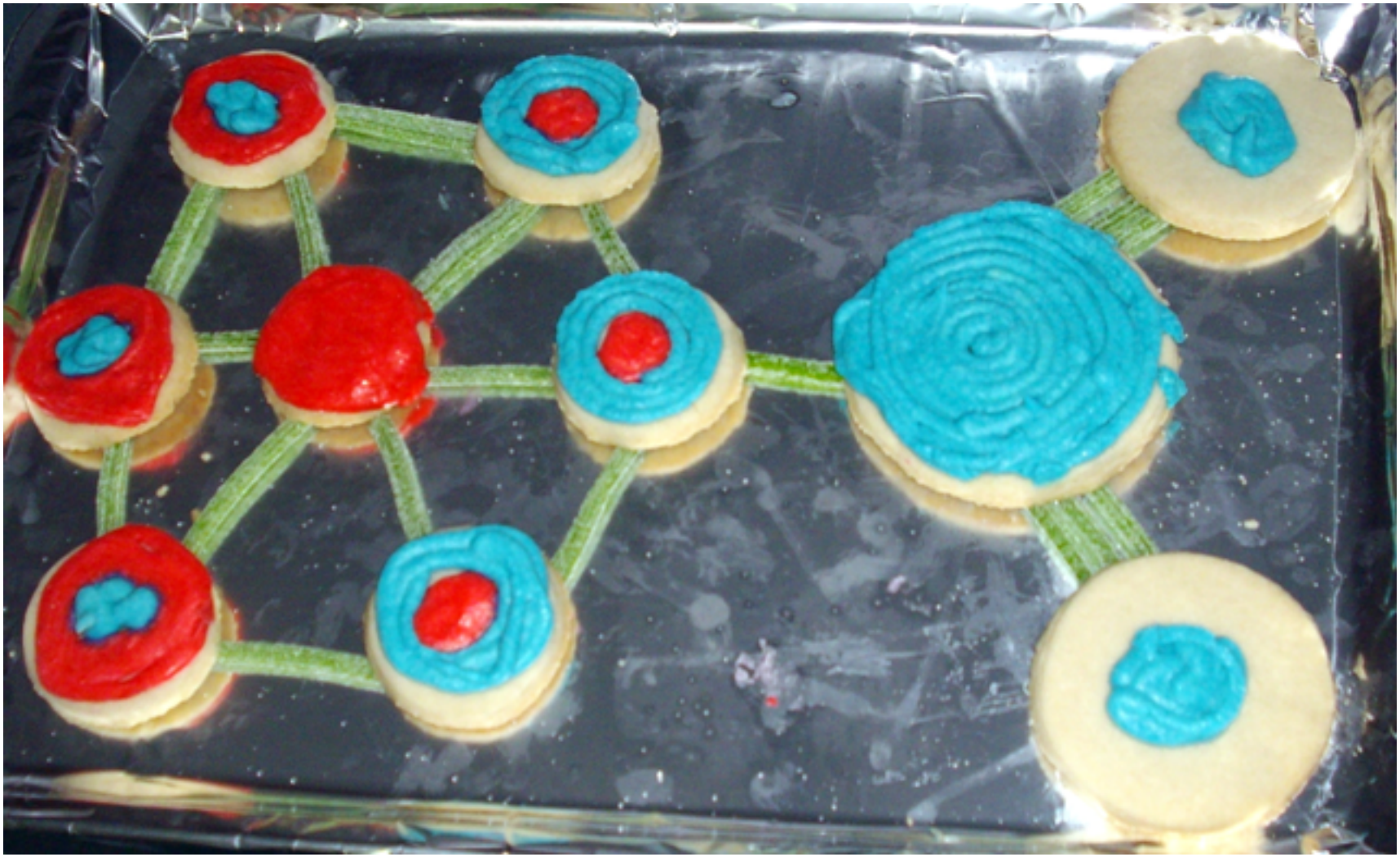
Cluster words according to their co-occurrence in labels to preserve semantic meaning.

Layla Oesper
Google Summer of Code

Enrichment Map
Acknowledgments



Donnelly Centre
(University of Toronto)



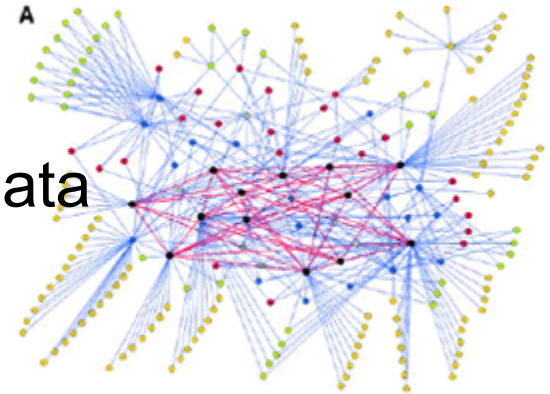
Gene function prediction

Outline

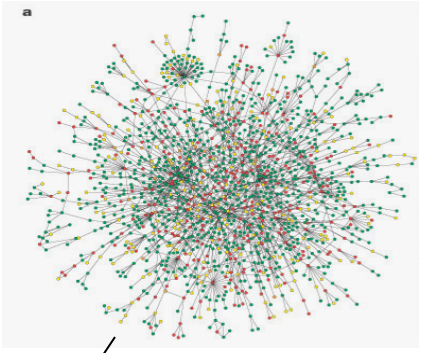
- Concepts in gene function prediction:
  - Guilt-by-association
  - Gene recommender systems
- Gene function prediction use cases
- Functional interaction networks
- Scoring interactions by guilt-by-association
- GeneMANIA & STRING
- GeneMANIA demo
- STRING demo

Using genome-wide data in the lab

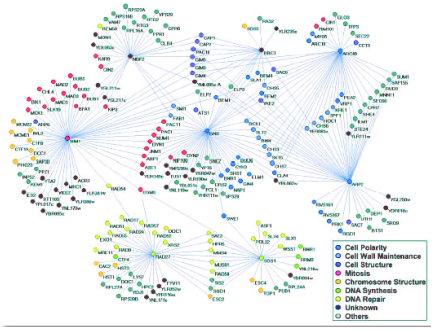
CHiP-chip regulation data



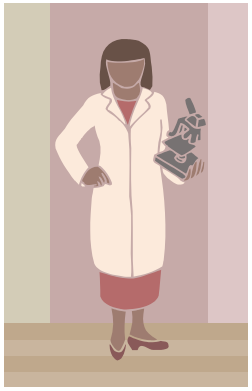
Protein-protein interaction data



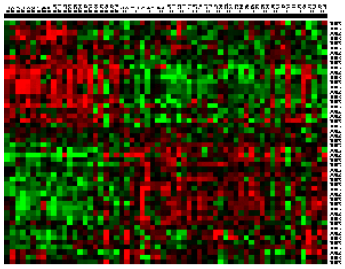
Genetic interaction data



?!?



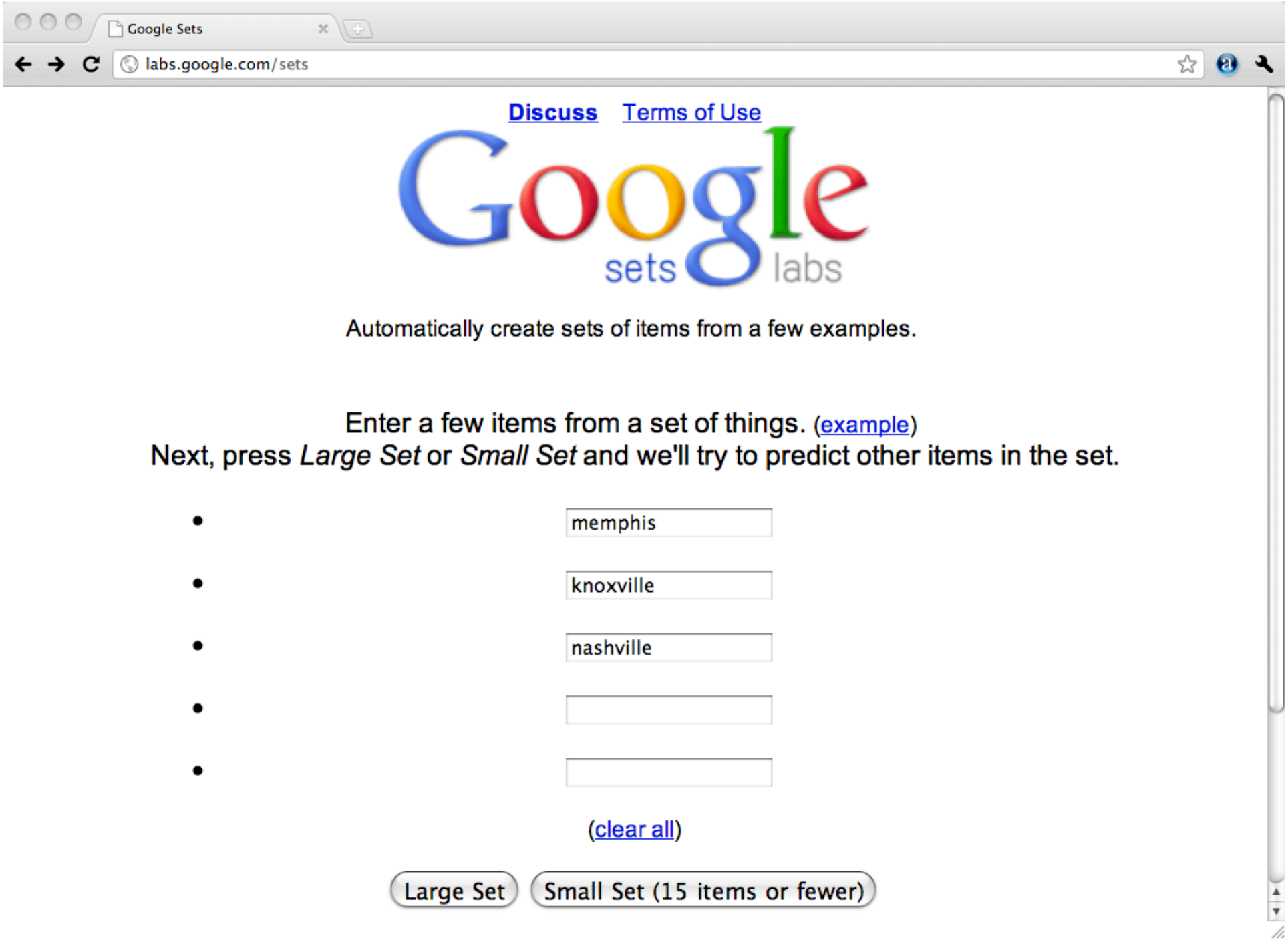
Microarray expression data

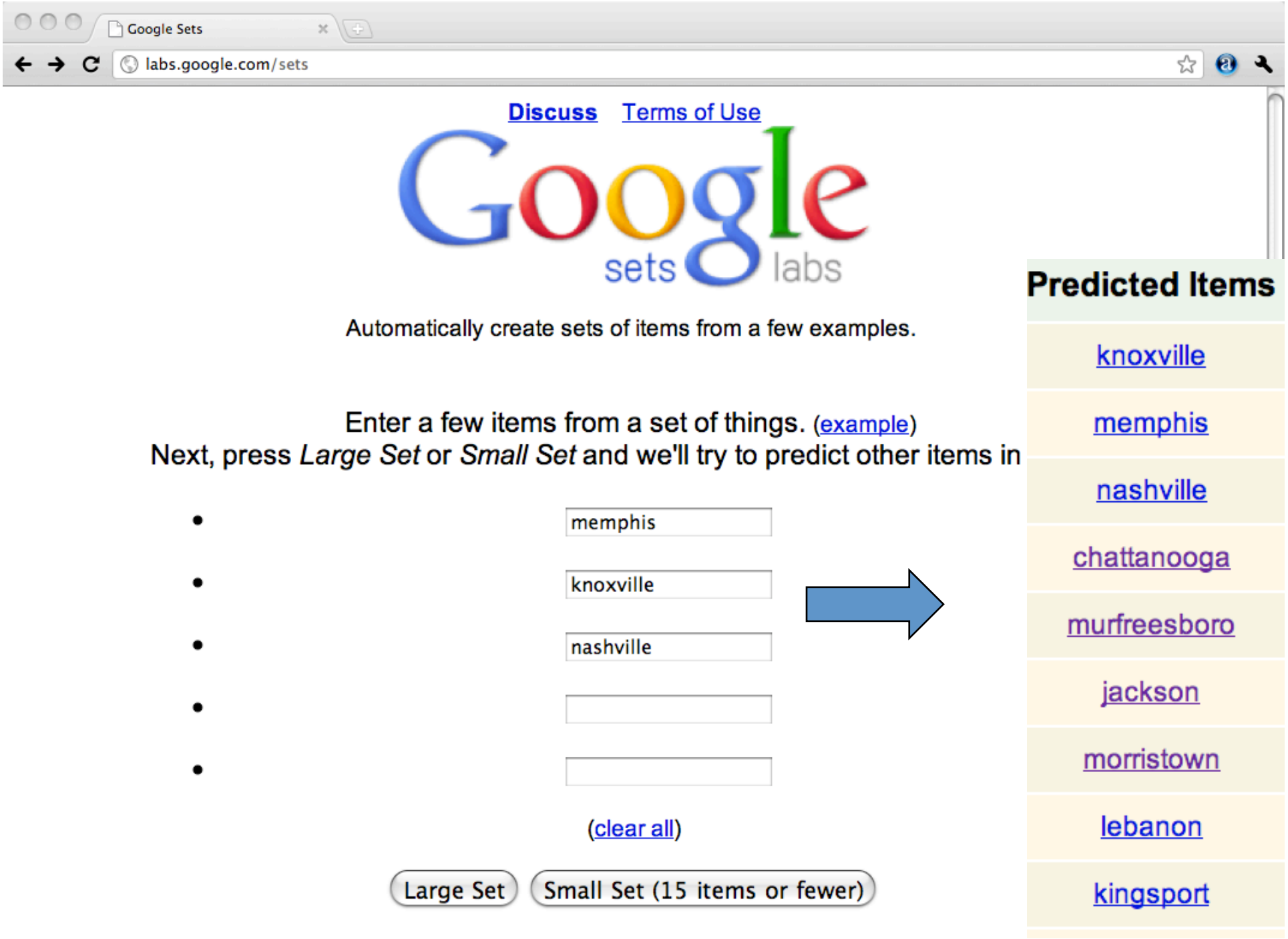


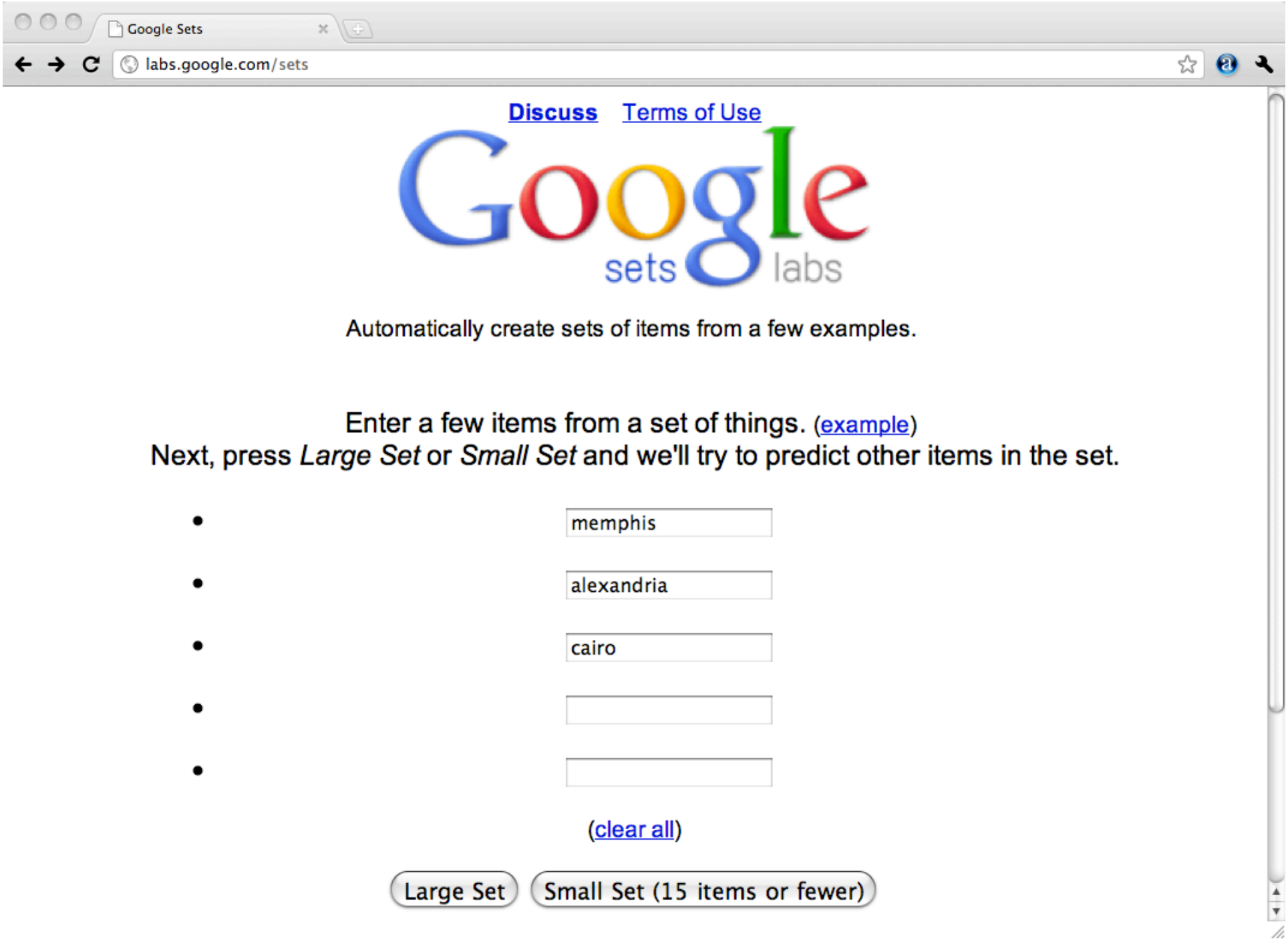
Genomics revolution, the bad news

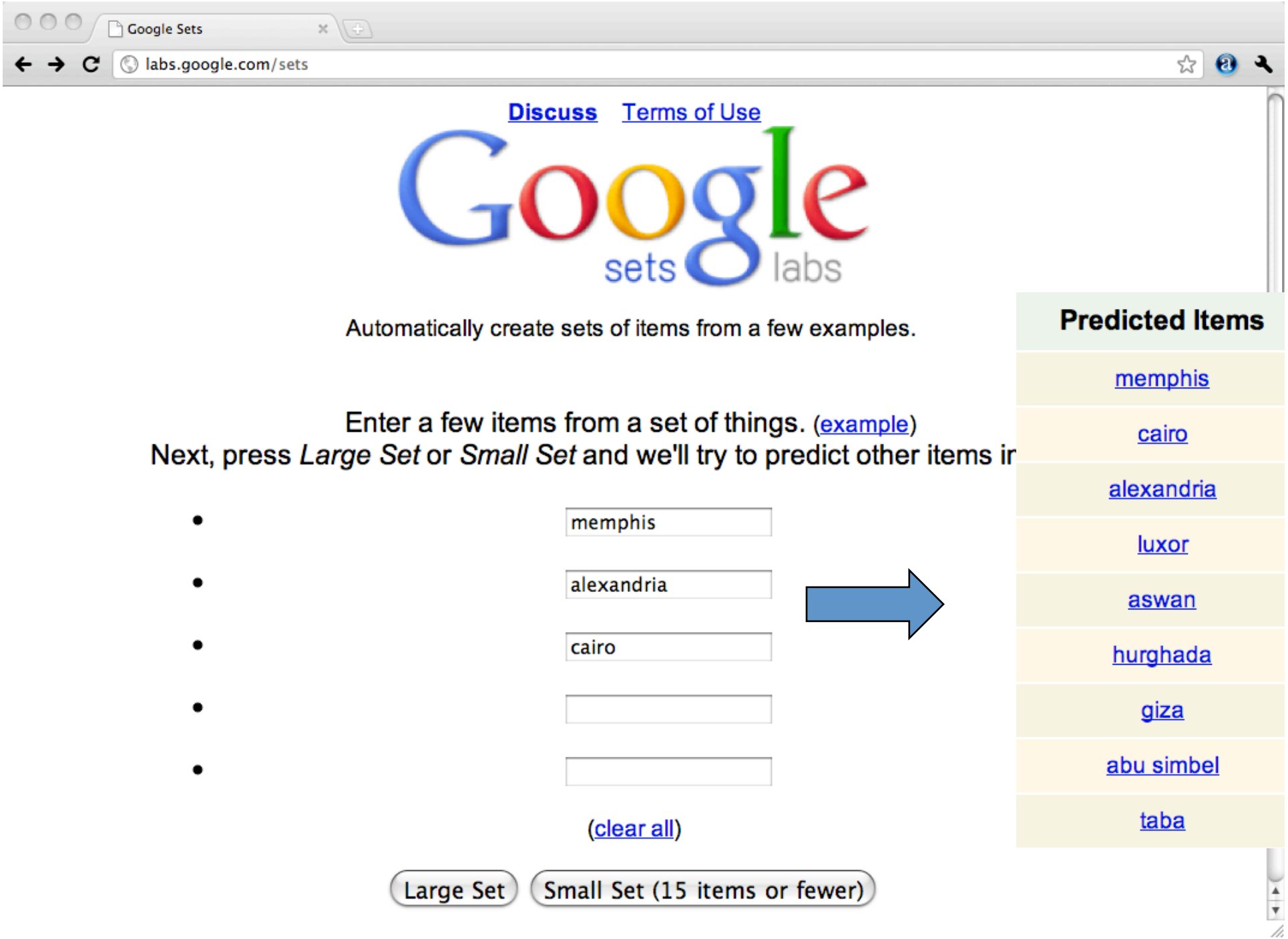
Genomics datasets are:

- noisy,
- redundant,
- incomplete,
- mysterious,
- massive


Google Sets
labs.google.com/sets
Discuss
Terms of Use
Google
sets labs
Automatically create sets of items from a few examples.
Enter a few items from a set of things. (example)
Next, press Large Set or Small Set and we'll try to predict other items in the set.
memphis
knoxville
nashville
(clear all)
Large Set
Small Set (15 items or fewer)


Google Sets
labs.google.com/sets
Discuss
Terms of Use
Google
sets labs
Automatically create sets of items from a few examples.
Enter a few items from a set of things. (example)
Next, press Large Set or Small Set and we'll try to predict other items in
memphis
knoxville
nashville
(clear all)
Large Set
Small Set (15 items or fewer)
Predicted Items
knoxville
memphis
nashville
chattanooga
murfreesboro
jackson
morristown
lebanon
kingsport


Google Sets
labs.google.com/sets
Discuss
Terms of Use
Google
sets labs
Automatically create sets of items from a few examples.
Enter a few items from a set of things. (example)
Next, press Large Set or Small Set and we'll try to predict other items in the set.
memphis
alexandria
cairo
(clear all)
Large Set
Small Set (15 items or fewer)


Google Sets
labs.google.com/sets
Discuss
Terms of Use
Google
sets labs
Automatically create sets of items from a few examples.
Enter a few items from a set of things. (example)
Next, press Large Set or Small Set and we'll try to predict other items ir
memphis
alexandria
cairo
(clear all)
Large Set
Small Set (15 items or fewer)
Predicted Items
memphis
cairo
alexandria
luxor
aswan
hurghada
giza
abu simbel
taba

Google can’t do biology

Automatically create sets of items from a few examples.

Enter a few items from a set of things. (example)
Next, press *Large Set* or *Small Set* and we'll try to predict other items in the set.

- CDC27
- APC11
- APC4
- CDC26
- DOC1

(clear all)

Large Set | Small Set (15 items or fewer)

Google can’t do biology

Automatically create sets of items from a few examples.

Next, pr

Predicted Items
cdc27
apc4

t.

Zero or only a few results? Try the following::

(clear all)

Large Set | Small Set (15 items or fewer)

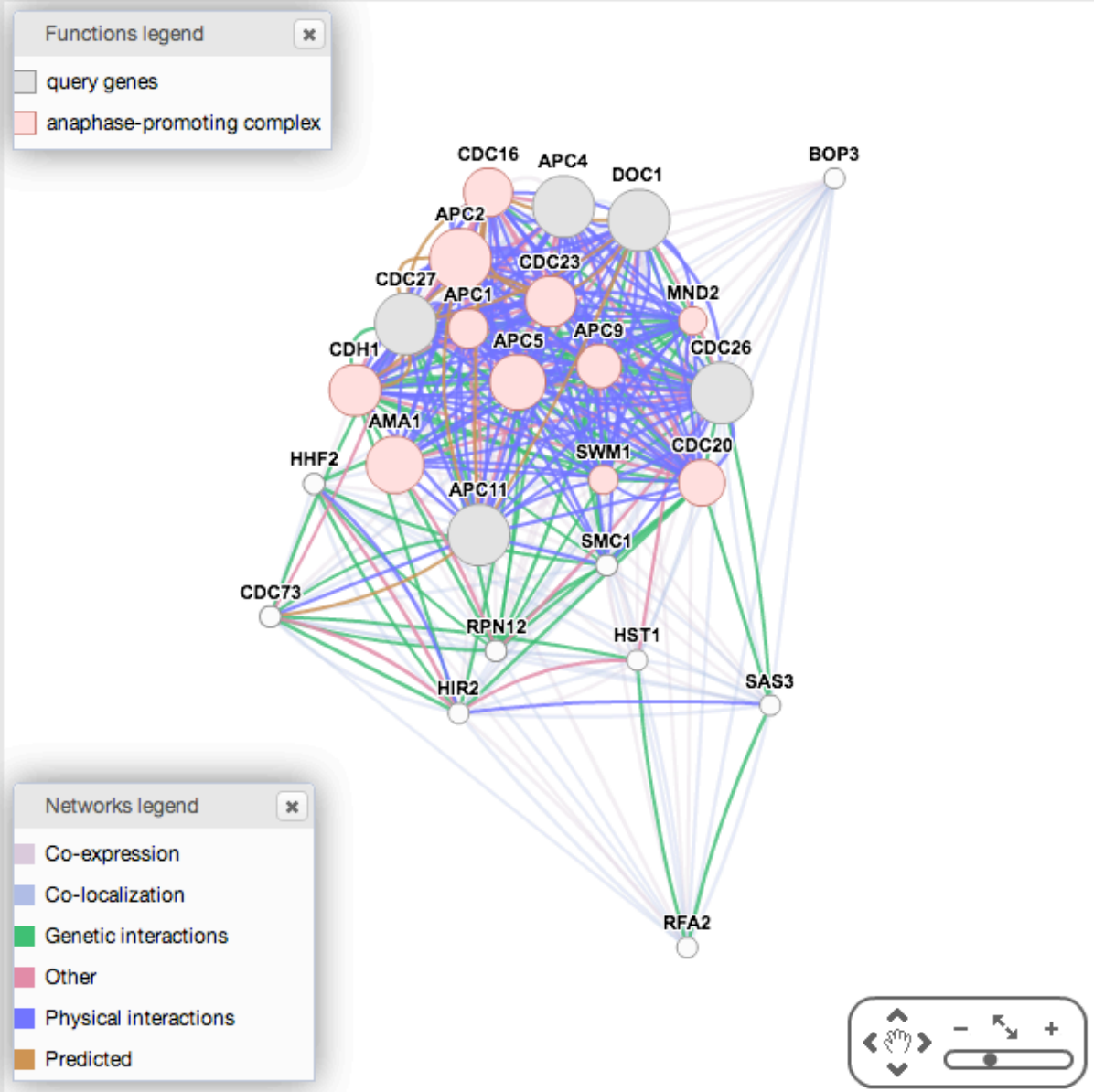
GeneMANIA

Find genes in S. cerevisiae (baker's yeast) related to cdc27; apc11; apc4; cdc26; doc1 Go

Showing 20 related genes

Show advanced options

File ▾ Actions ▾ Networks legend Functions legend



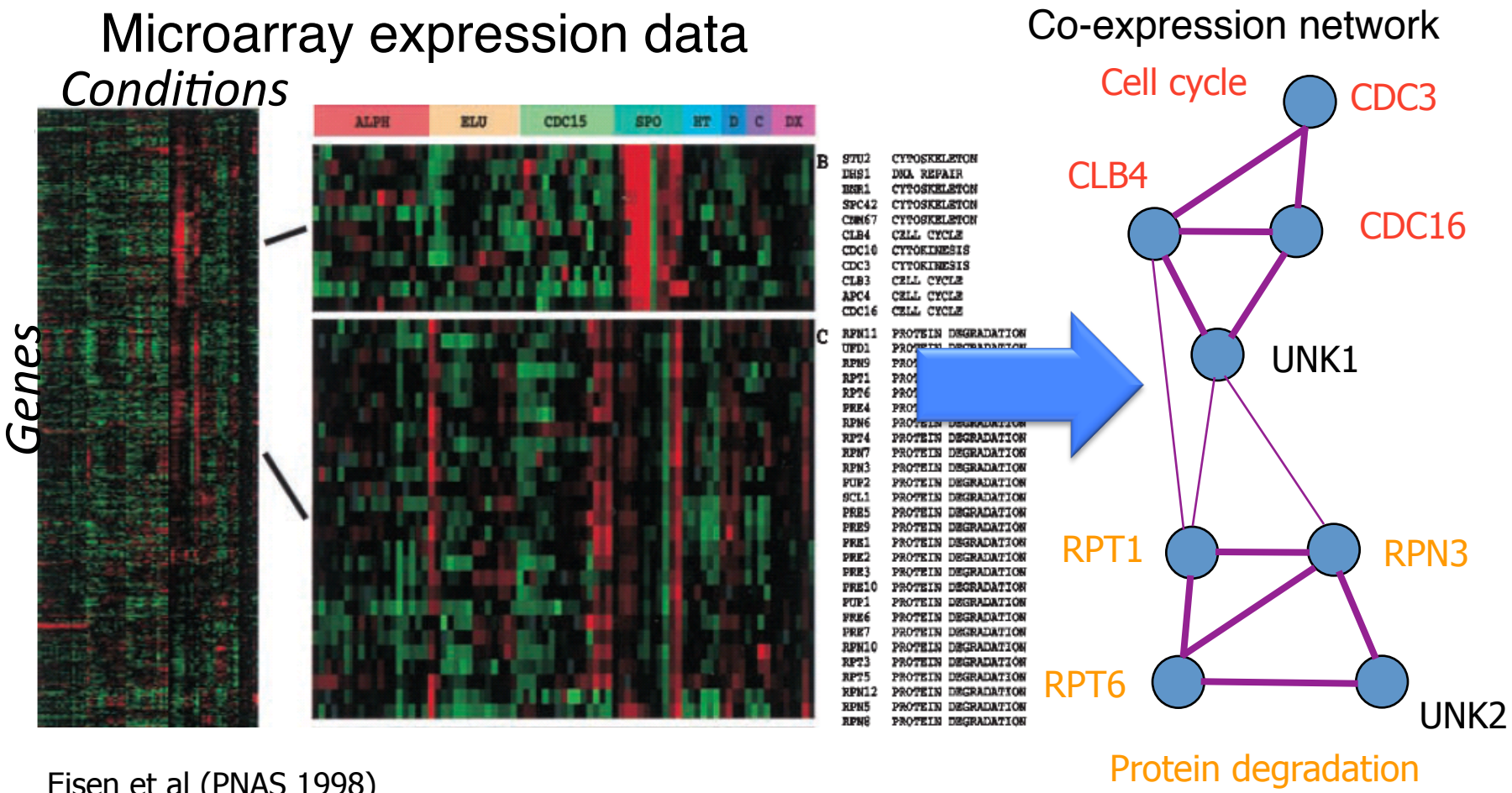
Networks Genes Functions Help

Sort by: name, score
Expand: all, none

DOC1	Processivity factor required for the ubiquitination activity of th	
APC4	Subunit of the Anaphase-Promoting Complex/Cyclosome (AI	
APC11	Catalytic core subunit of the Anaphase-Promoting Complex	
CDC26	Subunit of the Anaphase-Promoting Complex/Cyclosome (	
CDC27	Subunit of the Anaphase-Promoting Complex/Cyclosome (	
APC2	Subunit of the Anaphase-Promoting Complex/Cyclo	0.81
AMA1	Activator of meiotic anaphase promoting complex (A	0.75
APC5	Subunit of the Anaphase-Promoting Complex/Cyclo	0.73
CDH1	Cell-cycle regulated activator of the anaphase-promc	0.67
CDC23	Subunit of the Anaphase-Promoting Complex/Cyc	0.66
CDC16	Subunit of the anaphase-promoting complex/cyclc	0.64
CDC20	Cell-cycle regulated activator of anaphase-promotir	0.60
APC9	Subunit of the Anaphase-Promoting Complex/Cyclo	0.58
APC1	Largest subunit of the Anaphase-Promoting Comple	0.52
MND2	Subunit of the anaphase-promoting complex (APC);	0.37
SWM1	Subunit of the anaphase-promoting complex, whic	0.37
RPN12	Subunit of the 19S regulatory particle of the 26S pr	0.28
CDC73	Component of the Paf1p complex that binds to an	0.28
HHF2	Histone H4, core histone protein required for chroma	0.27
HIR2	Subunit of the HIR complex, a nucleosome assembly	0.27
HST1	NAD(+)-dependent histone deacetylase; essential sut	0.27
BOP3	Protein of unknown function, potential Cdc28p subs	0.27
RFA2	Subunit of heterotrimeric Replication Protein A (RPA),	0.27
SAS3	Histone acetyltransferase catalytic subunit of NuA3 c	0.27
SMC1	Subunit of the multiprotein cohesin complex, essen	0.27

Demo of GeneMANIA features

Guilt-by-association principle



Eisen et al (PNAS 1998)

Fraser AG, Marcotte EM - A probabilistic view of gene function - Nat Genet. 2004 Jun;36(6):559-64

Two types of functional prediction

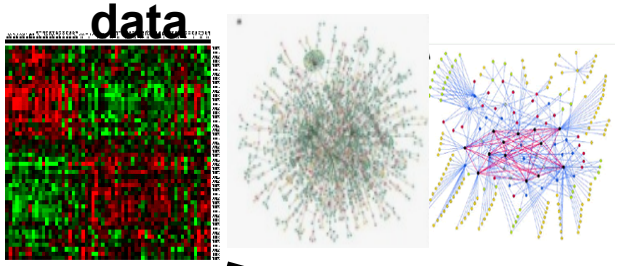
- **"Give me more genes like these"**,
  - e.g. find more genes in the Wnt signaling pathway, find more kinases, find more members of a protein complex

- **"What does my gene do?"**
  - Goal: determine a gene's function based on who it interacts with: "guilt-by-association".

“Give me more genes like these”

Input

Network and profile data



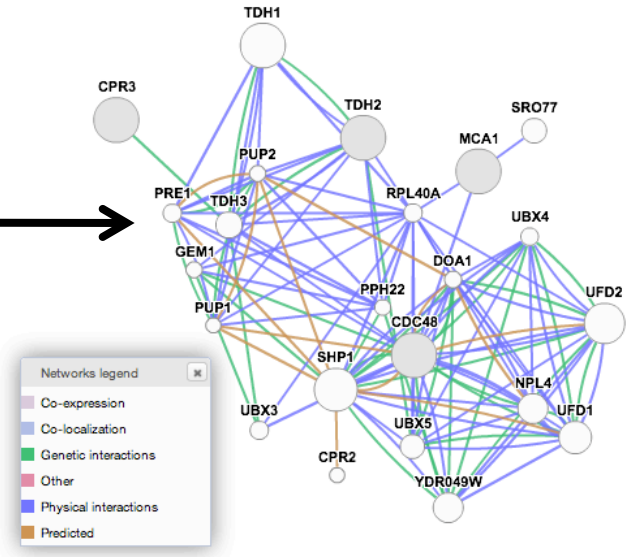
Query list

CDC48
CPR3
MCA1
TDH2

Gene recommender system

Output

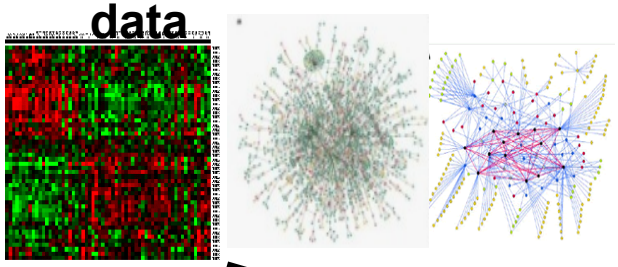
from GeneMANIA



“What does my gene do?”, Solution #1

Input

Network and profile data

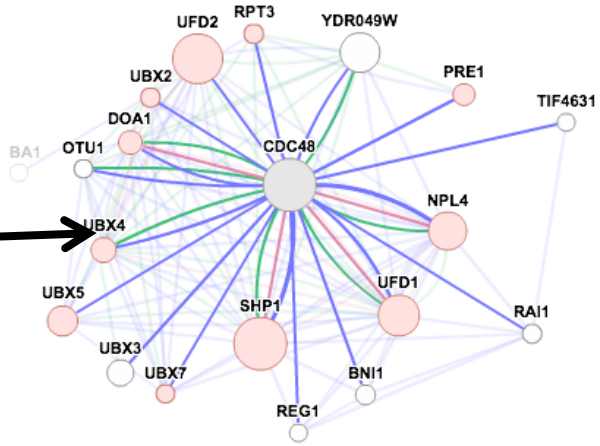


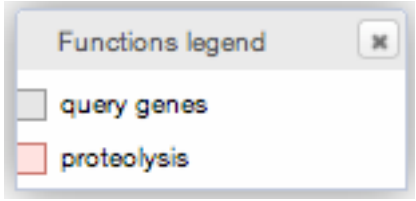
Query list

CDC48

Gene recommender system then enrichment analysis

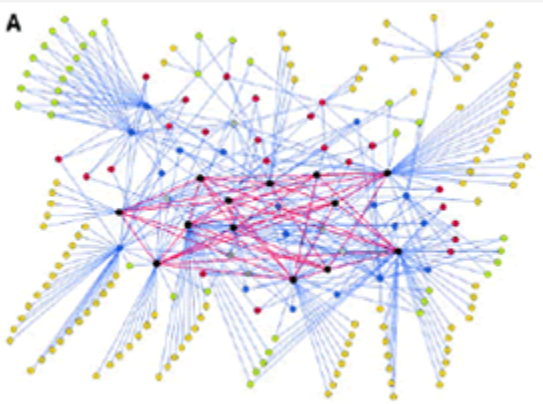
Output





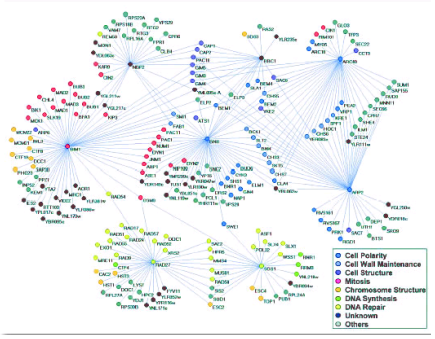
Composite functional interaction/linkage/ association networks

CHiP-chip regulation data

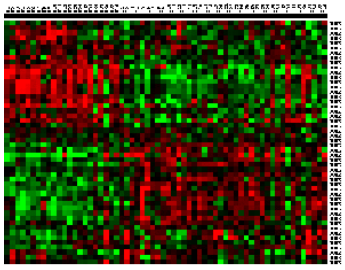


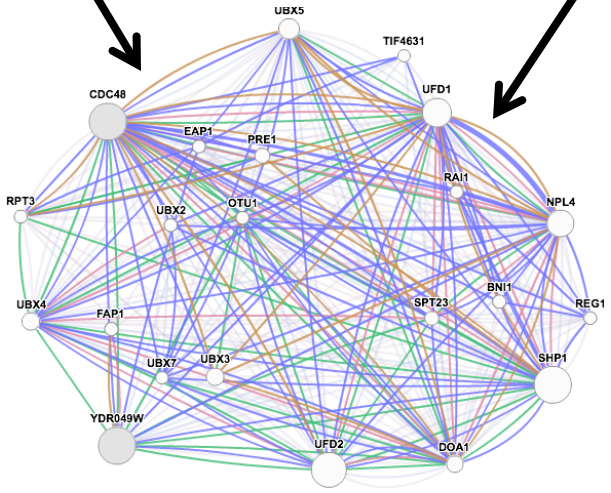
Protein-protein interaction data

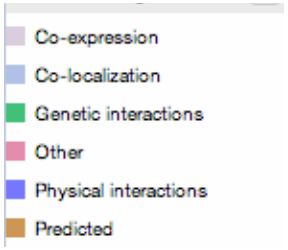
Genetic interaction data



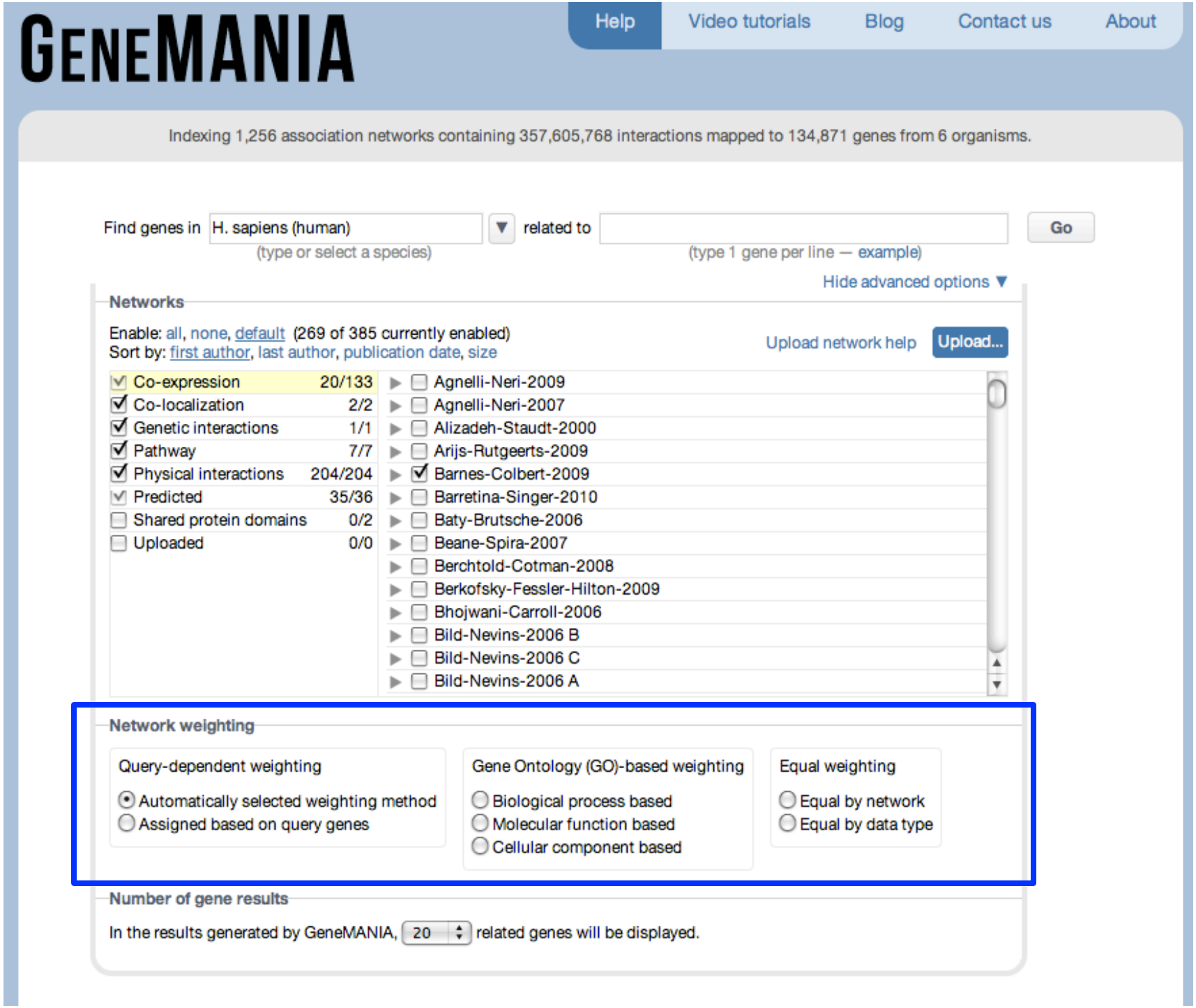
Microarray expression data





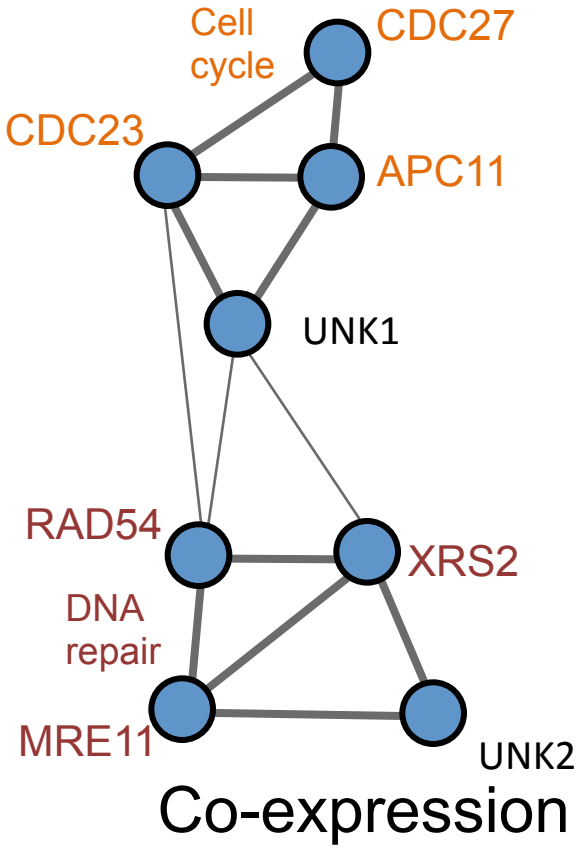


Composite functional association network


GeneMANIA
Help
Video tutorials
Blog
Contact us
About
Indexing 1,256 association networks containing 357,605,768 interactions mapped to 134,871 genes from 6 organisms.
Find genes in
H. sapiens (human)
(type or select a species)
related to
(type 1 gene per line — example)
Go
Hide advanced options
Networks
Enable: all, none, default (269 of 385 currently enabled)
Sort by: first author, last author, publication date, size
Upload network help
Upload...
Co-expression 20/133
Co-localization 2/2
Genetic interactions 1/1
Pathway 7/7
Physical interactions 204/204
Predicted 35/36
Shared protein domains 0/2
Uploaded 0/0
Agnelli-Neri-2009
Agnelli-Neri-2007
Alizadeh-Staudt-2000
Arijs-Rutgeerts-2009
Barnes-Colbert-2009
Barretina-Singer-2010
Baty-Brutsche-2006
Beane-Spira-2007
Berchtold-Cotman-2008
Berkofsky-Fessler-Hilton-2009
Bhojwani-Carroll-2006
Bild-Nevins-2006 B
Bild-Nevins-2006 C
Bild-Nevins-2006 A
Network weighting
Query-dependent weighting
Automatically selected weighting method
Assigned based on query genes
Gene Ontology (GO)-based weighting
Biological process based
Molecular function based
Cellular component based
Equal weighting
Equal by network
Equal by data type
Number of gene results
In the results generated by GeneMANIA, 20 related genes will be displayed.

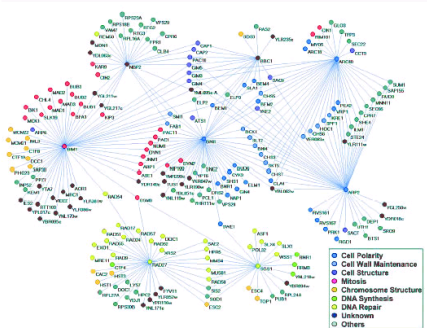
Query-independent composite networks

Pre-combine networks e.g. by simple addition or by pre-determined weights



Co-expression

+



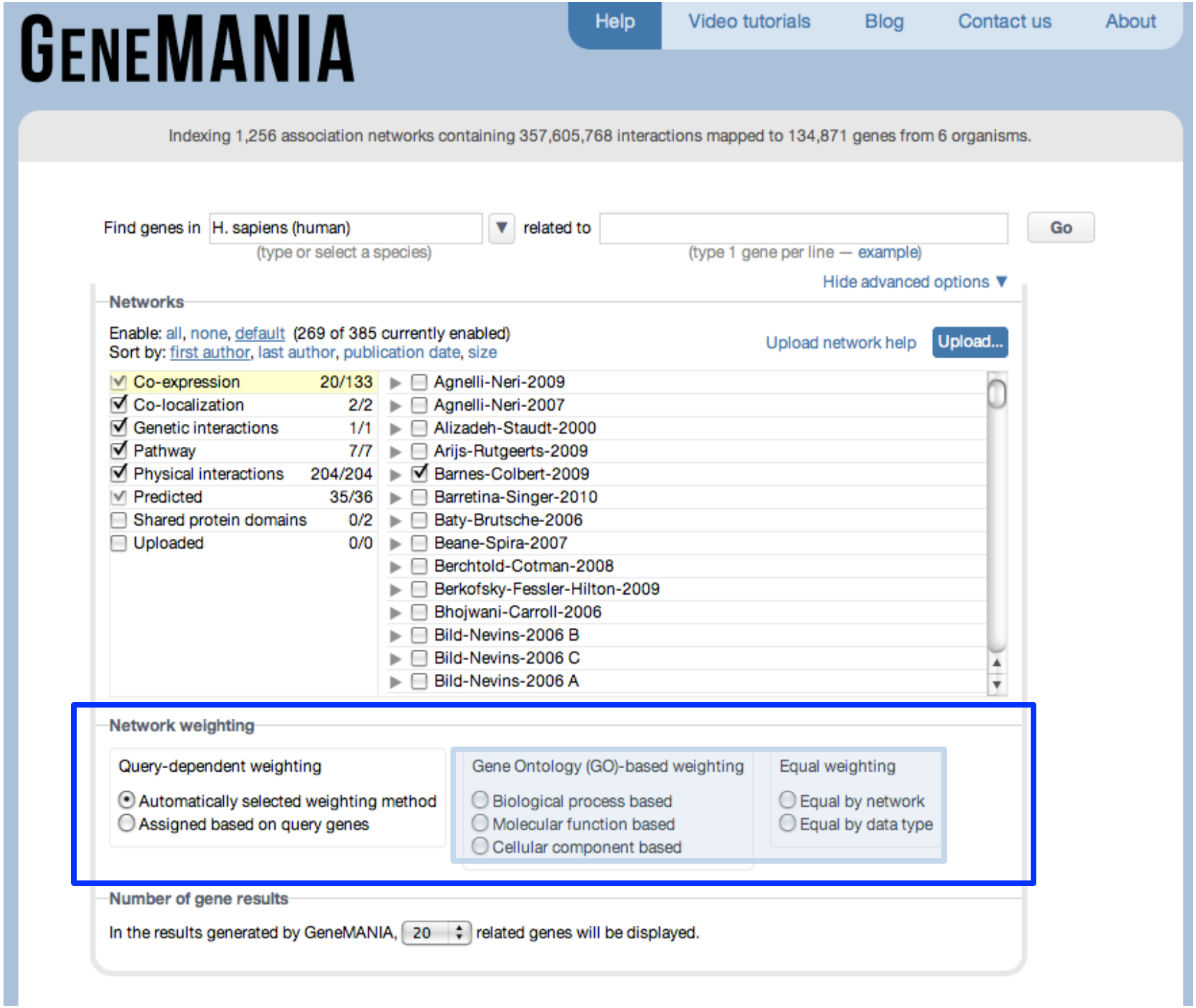
Genetic

Tong *et al.* 2001

+

Co-complexed

Jeong et al 2002


GeneMANIA
Help
Video tutorials
Blog
Contact us
About
Indexing 1,256 association networks containing 357,605,768 interactions mapped to 134,871 genes from 6 organisms.
Find genes in
H. sapiens (human)
(type or select a species)
related to
(type 1 gene per line — example)
Go
Hide advanced options
Networks
Enable: all, none, default (269 of 385 currently enabled)
Sort by: first author, last author, publication date, size
Upload network help
Upload...
Co-expression 20/133
Co-localization 2/2
Genetic interactions 1/1
Pathway 7/7
Physical interactions 204/204
Predicted 35/36
Shared protein domains 0/2
Uploaded 0/0
Agnelli-Neri-2009
Agnelli-Neri-2007
Alizadeh-Staudt-2000
Arijs-Rutgeerts-2009
Barnes-Colbert-2009
Barretina-Singer-2010
Baty-Brutsche-2006
Beane-Spira-2007
Berchtold-Cotman-2008
Berkofsky-Fessler-Hilton-2009
Bhojwani-Carroll-2006
Bild-Nevins-2006 B
Bild-Nevins-2006 C
Bild-Nevins-2006 A
Network weighting
Query-dependent weighting
Automatically selected weighting method
Assigned based on query genes
Gene Ontology (GO)-based weighting
Biological process based
Molecular function based
Cellular component based
Equal weighting
Equal by network
Equal by data type
Number of gene results
In the results generated by GeneMANIA, 20 related genes will be displayed.

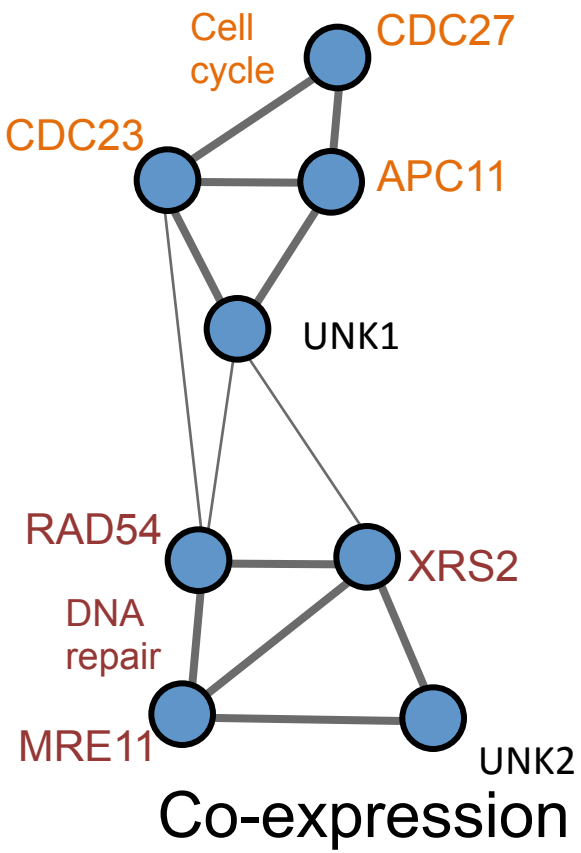
Composite networks: One size doesn't fit all

- Gene function could be a/the:
  - Biological process,
  - Biochemical/molecular function,
  - Subcellular/Cellular localization,
  - Regulatory targets,
  - Temporal expression pattern,
  - Phenotypic effect of deletion.

Some networks may be better for some types of gene function than others

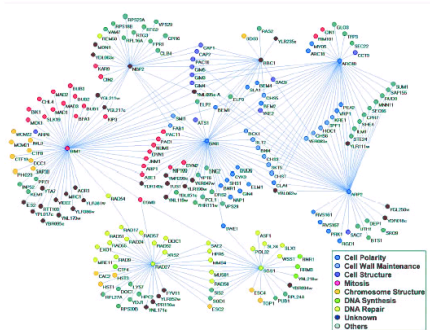
Solution: Query-specific weights

W_1 x ← weights → W_2 x W_3 x



Co-expression

\+



Genetic

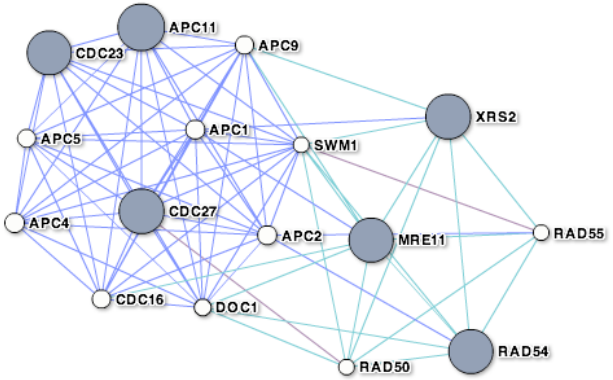
Tong *et al*, 2001

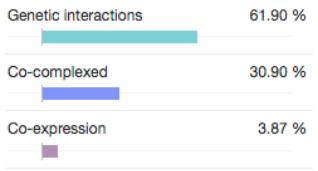
\+

Co-complexed

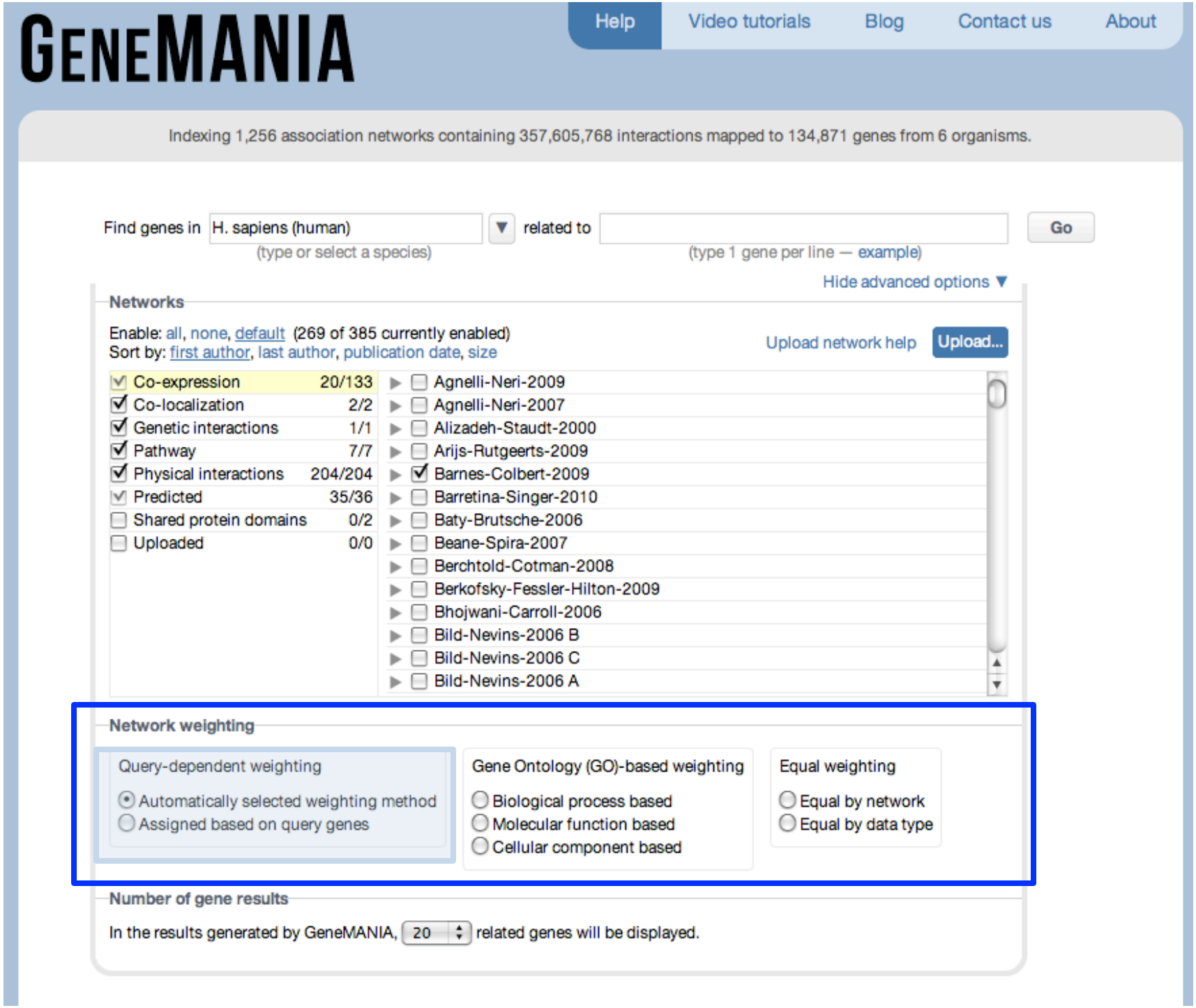
Jeong et al 2002

=





Pavlidis *et al*, 2002,
Lanckriet *et al,* 2004
Mostafavi et al, 2008


GeneMANIA
Help
Video tutorials
Blog
Contact us
About
Indexing 1,256 association networks containing 357,605,768 interactions mapped to 134,871 genes from 6 organisms.
Find genes in
H. sapiens (human)
(type or select a species)
related to
(type 1 gene per line — example)
Go
Hide advanced options
Networks
Enable: all, none, default (269 of 385 currently enabled)
Sort by: first author, last author, publication date, size
Upload network help
Upload...
Co-expression 20/133
Co-localization 2/2
Genetic interactions 1/1
Pathway 7/7
Physical interactions 204/204
Predicted 35/36
Shared protein domains 0/2
Uploaded 0/0
Agnelli-Neri-2009
Agnelli-Neri-2007
Alizadeh-Staudt-2000
Arijs-Rutgeerts-2009
Barnes-Colbert-2009
Barretina-Singer-2010
Baty-Brutsche-2006
Beane-Spira-2007
Berchtold-Cotman-2008
Berkofsky-Fessler-Hilton-2009
Bhojwani-Carroll-2006
Bild-Nevins-2006 B
Bild-Nevins-2006 C
Bild-Nevins-2006 A
Network weighting
Query-dependent weighting
Automatically selected weighting method
Assigned based on query genes
Gene Ontology (GO)-based weighting
Biological process based
Molecular function based
Cellular component based
Equal weighting
Equal by network
Equal by data type
Number of gene results
In the results generated by GeneMANIA, 20 related genes will be displayed.

Two rules for network weighting

Relevance

The network should be relevant to predicting the function of interest

- **Test:** Are the genes in the query list more often connected to one another than to other genes?

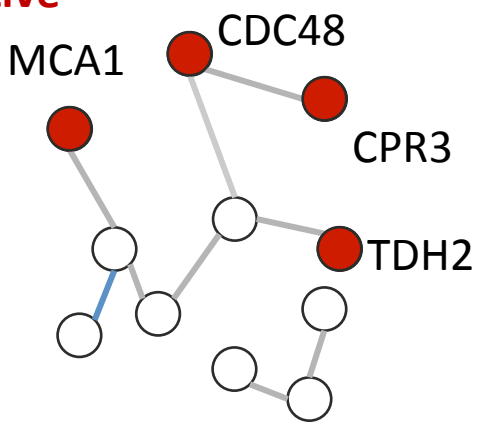
Redundancy

The network should not be redundant with other datasets – particularly a problem for co-expression

- **Test:** Do the two networks share many interactions
- *Caveat:* Shared interactions also provide more confidence that the interaction is real.

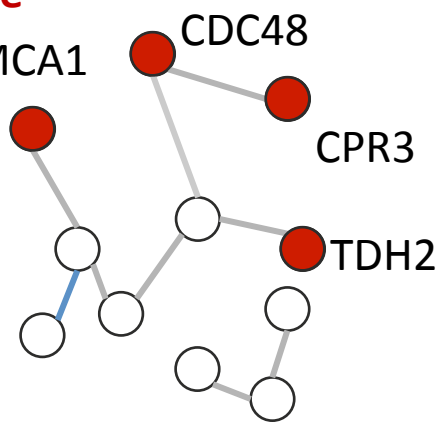
Scoring nodes by guilt-by-association

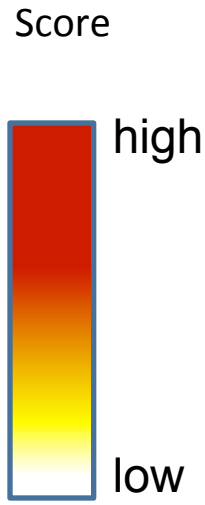
Query list: "positive examples"



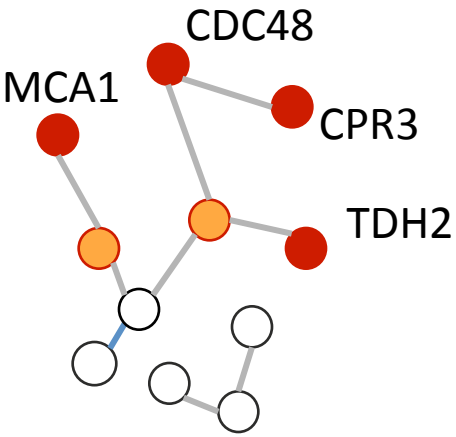
Scoring nodes by guilt-by-association

Query list: "positive examples"



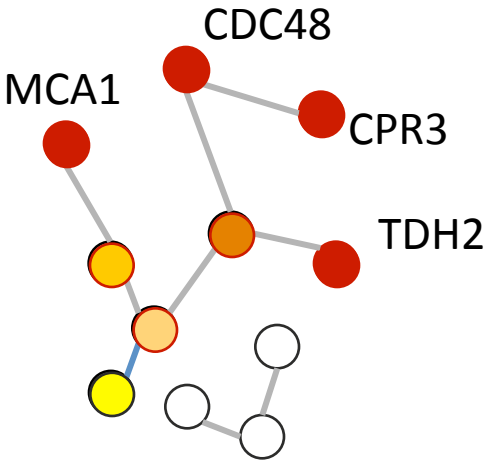


Direct neighborhood



Two main algorithms

Label propagation



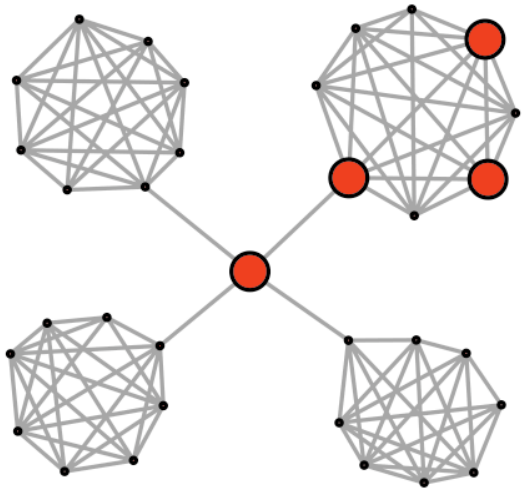
Node scoring algorithm details

- **Direct neighbour** node score depends on:
  - Strength of links to positive examples
  - # of positive neighbors

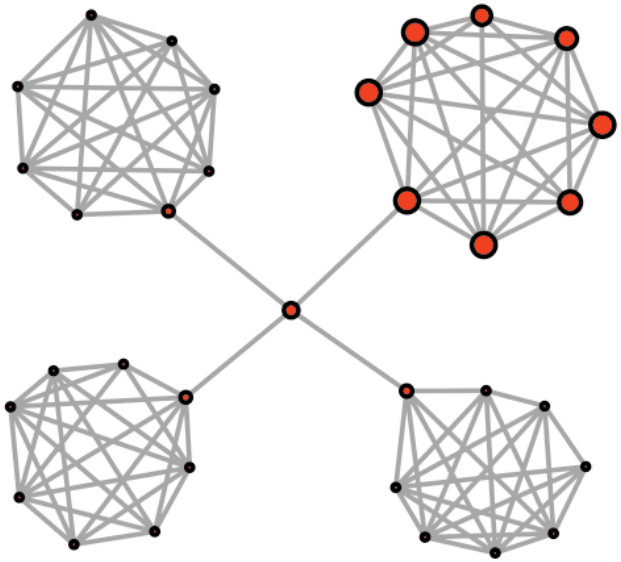
- **GeneMANIA Label propagation** node score depends on:
  - Strength of links and # of positive direct neighbors
  - # of shared neighbors with positive examples
  - “modular structure” of network

Label propagation example

Before



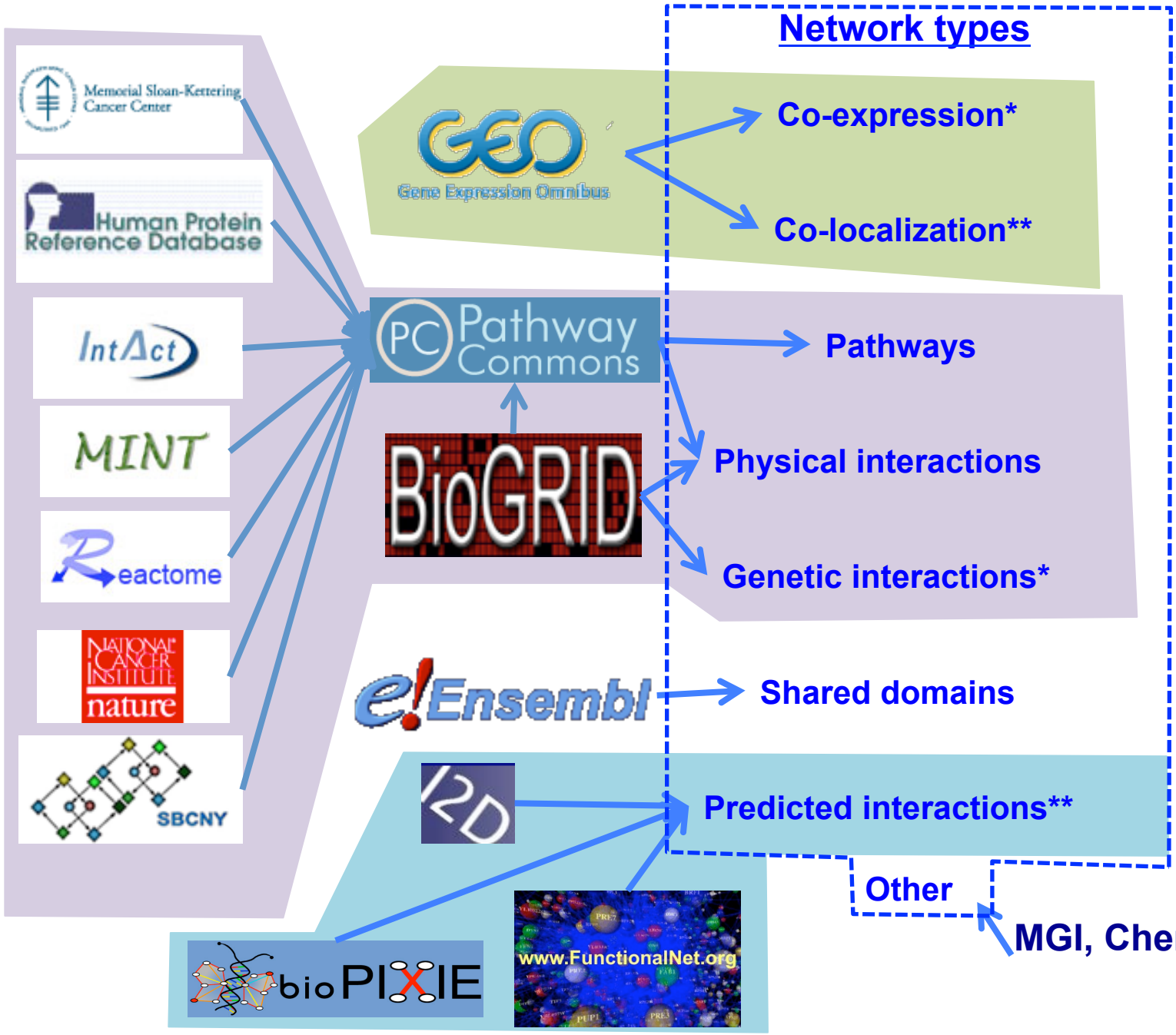
After



Three parts of GeneMANIA:

- A large, automatically updated collection of interactions networks.
- A query algorithm to find genes and networks that are functionally associated to your query gene list.
- An interactive, client-side network browser with extensive link-outs

GeneMANIA data sources



Legend
* minor curation
** major curation

-**Gene ID mappings** from Ensembl and Ensembl Plant

-**Network/gene descriptors** from Entrez-Gene and Pubmed

- **Gene annotations** from Gene Ontology, GOA, and model org. databases

Gene identifiers

- All unique identifiers within the selected organism: e.g.
  - Entrez-Gene ID
  - Gene symbol
  - Ensembl ID
  - Uniprot (primary)
  - also, some synonyms & organism-specific names
- We use Ensembl database for gene mappings (but we mirror it once / 3 months, so sometimes we are out of date)

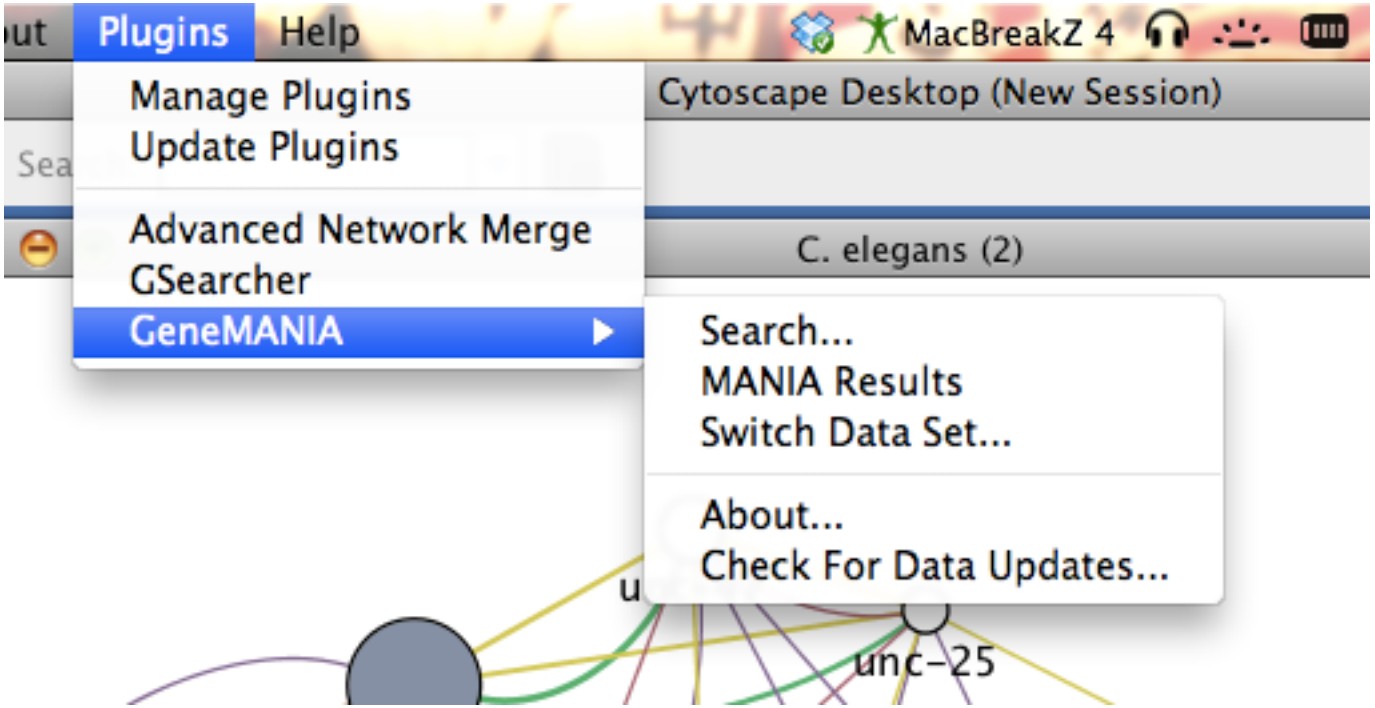
Current status

- Six organisms:
  - Human, Mouse, yeast, worm, fly, A Thaliana, [Rat coming soon]

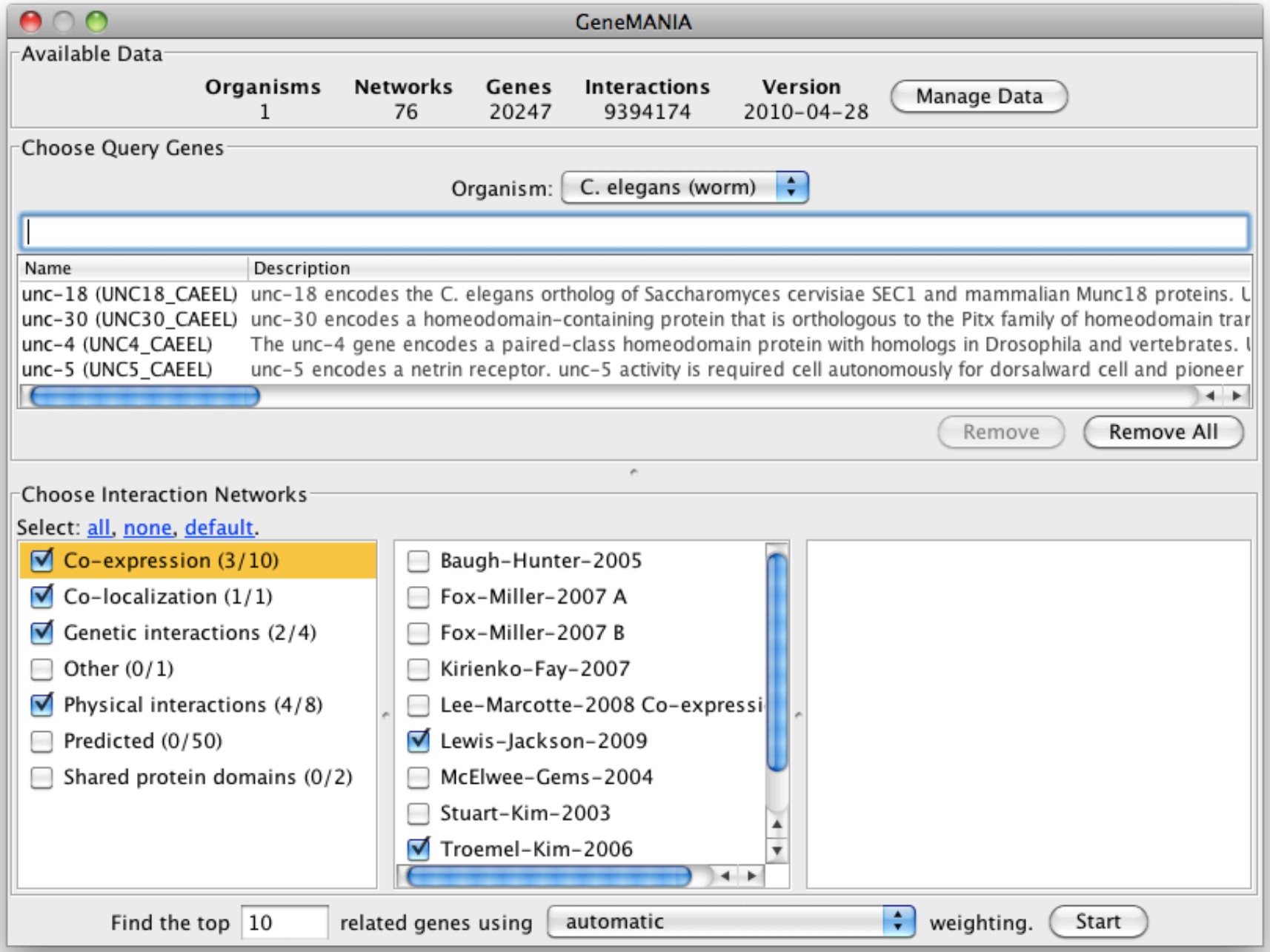
- ~1,250 networks (about 50% co-expression, 35% physical interaction)

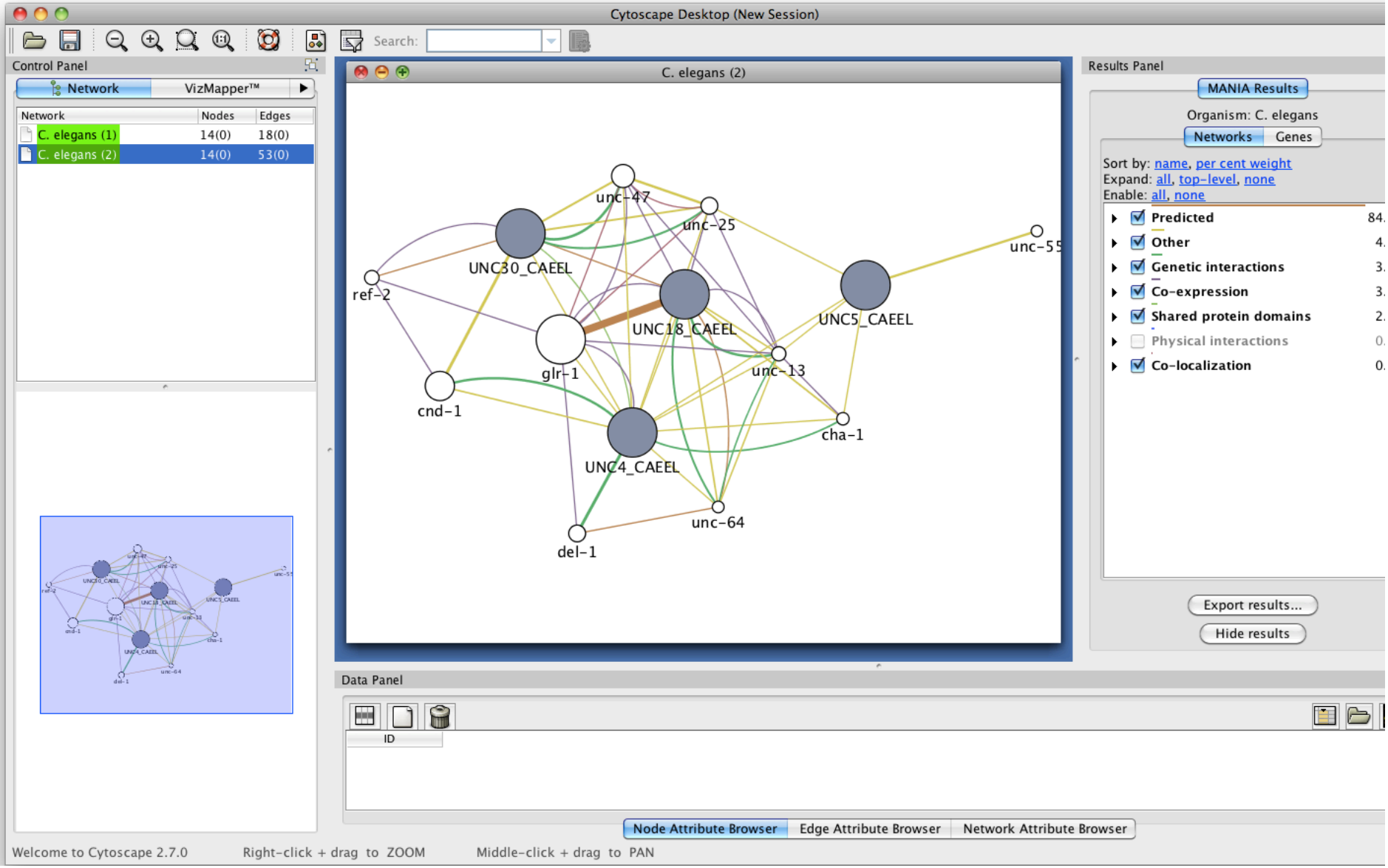
- Web network browser

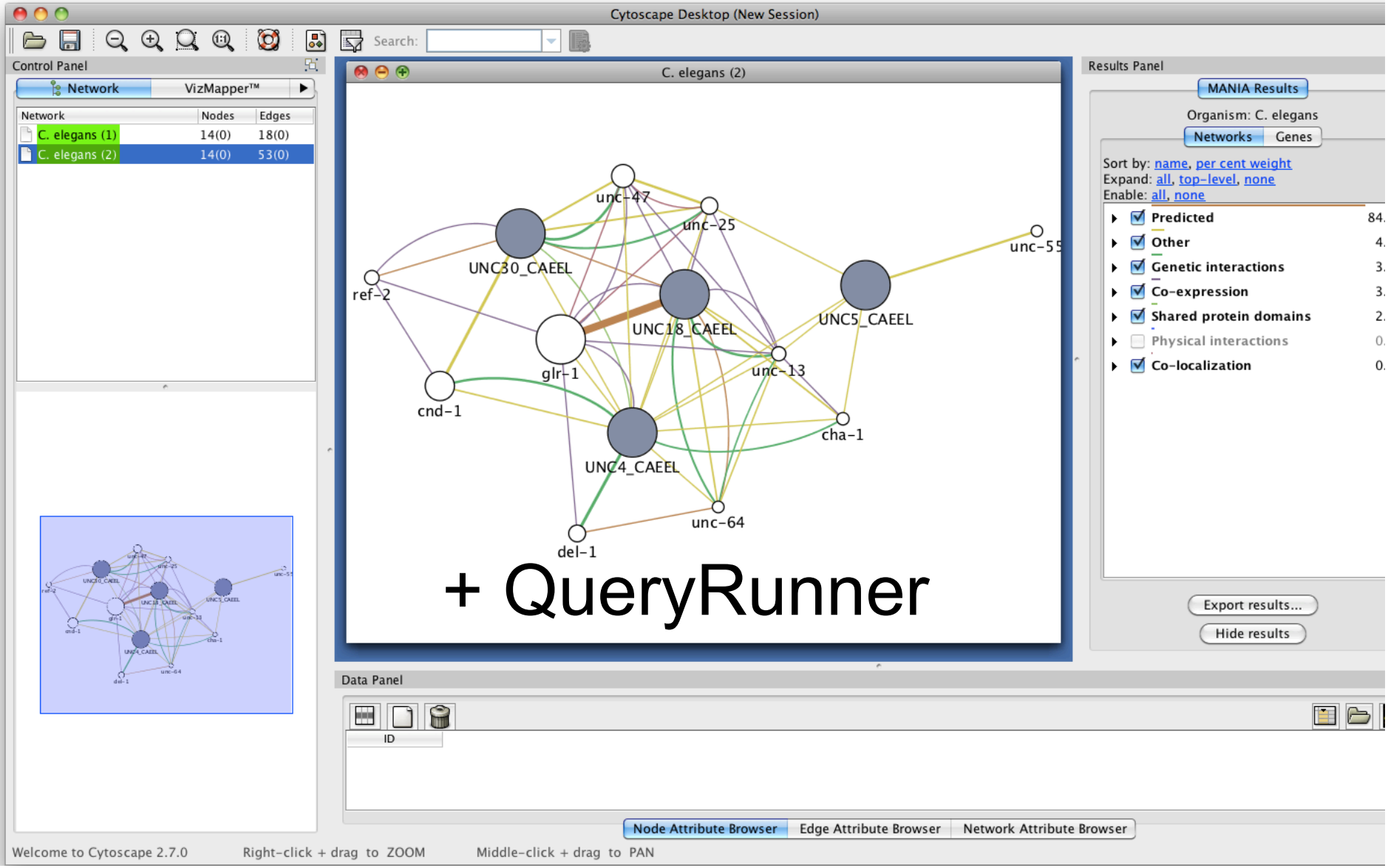
Cytoscape plugin



http://www.genemania.org/plugin/


GeneMANIA
Available Data
Organisms
1
Networks
76
Genes
20247
Interactions
9394174
Version
2010-04-28
Manage Data
Choose Query Genes
Organism:
C. elegans (worm)
Name
Description
unc-18 (UNC18_CAEEL)
unc-18 encodes the C. elegans ortholog of Saccharomyces cervisiae SEC1 and mammalian Munc18 proteins. U
unc-30 (UNC30_CAEEL)
unc-30 encodes a homeodomain-containing protein that is orthologous to the Pitx family of homeodomain trar
unc-4 (UNC4_CAEEL)
The unc-4 gene encodes a paired-class homeodomain protein with homologs in Drosophila and vertebrates. U
unc-5 (UNC5_CAEEL)
unc-5 encodes a netrin receptor. unc-5 activity is required cell autonomously for dorsalward cell and pioneer
Remove
Remove All
Choose Interaction Networks
Select: all, none, default.
Co-expression (3/10)
Co-localization (1/1)
Genetic interactions (2/4)
Other (0/1)
Physical interactions (4/8)
Predicted (0/50)
Shared protein domains (0/2)
Baugh-Hunter-2005
Fox-Miller-2007 A
Fox-Miller-2007 B
Kirienko-Fay-2007
Lee-Marcotte-2008 Co-expressi
Lewis-Jackson-2009
McElwee-Gems-2004
Stuart-Kim-2003
Troemel-Kim-2006
Find the top
10
related genes using
automatic
weighting.
Start


Cytoscape Desktop (New Session)
Search:
Control Panel
Network
VizMapper™
Network
Nodes
Edges
C. elegans (1)
14(0)
18(0)
C. elegans (2)
14(0)
53(0)
C. elegans (2)
unc-47
unc-25
UNC30_CAEEL
ref-2
UNC18_CAEEL
UNC5_CAEEL
unc-55
glr-1
unc-13
cnd-1
cha-1
UNC4_CAEEL
unc-64
del-1
Results Panel
MANIA Results
Organism: C. elegans
Networks
Genes
Sort by: name, per cent weight
Expand: all, top-level, none
Enable: all, none
Predicted
Other
Genetic interactions
Co-expression
Shared protein domains
Physical interactions
Co-localization
Export results...
Hide results
Data Panel
ID
Node Attribute Browser
Edge Attribute Browser
Network Attribute Browser
Welcome to Cytoscape 2.7.0
Right-click + drag to ZOOM
Middle-click + drag to PAN


Cytoscape Desktop (New Session)
Search:
Control Panel
Network
VizMapper™
Network
Nodes
Edges
C. elegans (1)
14(0)
18(0)
C. elegans (2)
14(0)
53(0)
C. elegans (2)
unc-47
unc-25
UNC30_CAEEL
ref-2
UNC18_CAEEL
UNC5_CAEEL
unc-55
glr-1
unc-13
cnd-1
cha-1
UNC4_CAEEL
unc-64
del-1
+ QueryRunner
Results Panel
MANIA Results
Organism: C. elegans
Networks
Genes
Sort by: name, per cent weight
Expand: all, top-level, none
Enable: all, none
Predicted
Other
Genetic interactions
Co-expression
Shared protein domains
Physical interactions
Co-localization
Export results...
Hide results
Data Panel
ID
Node Attribute Browser
Edge Attribute Browser
Network Attribute Browser
Welcome to Cytoscape 2.7.0
Right-click + drag to ZOOM
Middle-click + drag to PAN

GeneMANIA future directions

- Rat (1-3 weeks), next is probably E. Coli
- Non-coding genes (miRNAs!!!!)
- Regulatory networks (ChIP, RNA-protein, miRNA-mRNAs)
- More phenotypic information (OMIM, etc)
- Orthology mapping for inferring interologs

GeneMANIA URLs

Main site (stable but still fun):

http://www.genemania.org

Beta site (new and edgy but possibly unreliable):

http://beta.genemania.org

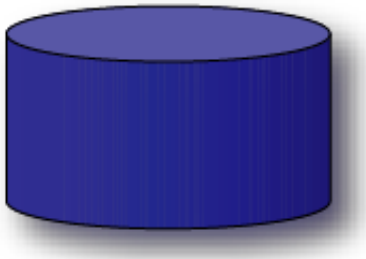
Pathways: representation and visualization

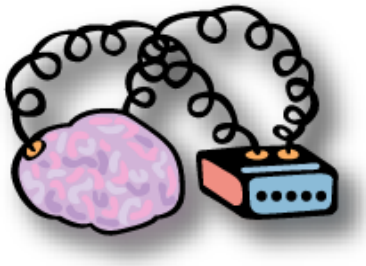
Pathways

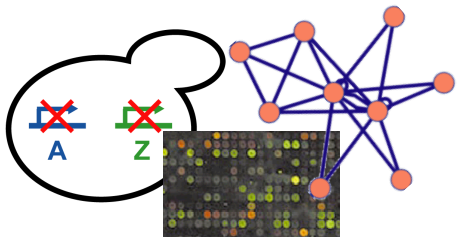
- A biological process
  - However, there is no *precise* biological definition of a pathway
- Start point is important or easily accessible stimulus (e.g. EGF hormone, drug)
- End point is a chosen readout (e.g. reporter gene expression, protein phosphorylation, production of a metabolite)

Pathway Information

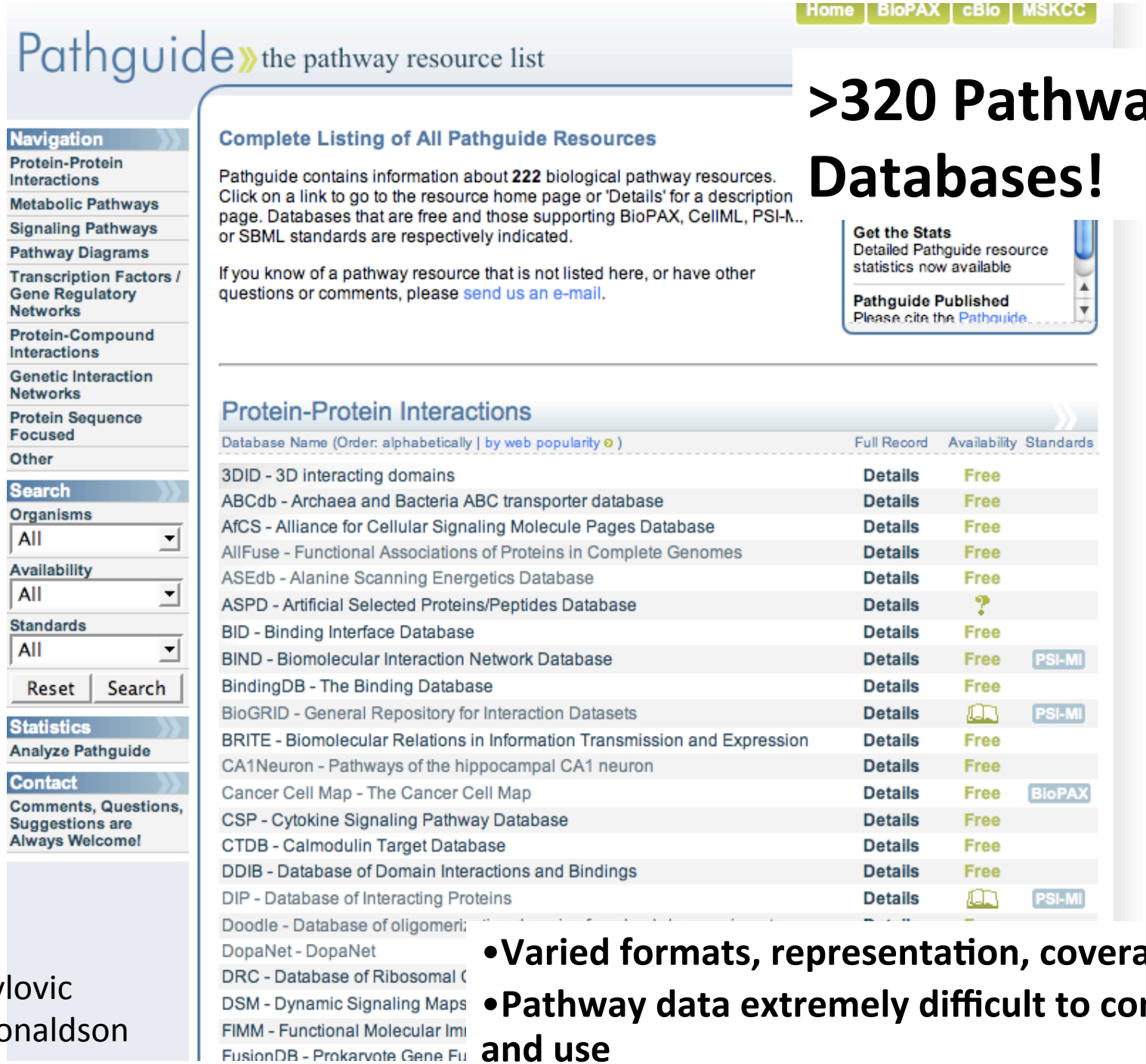
- Databases
  - Fully electronic
  - Easily computer readable
- Literature
  - Increasingly electronic
  - Human readable
- Biologist's brains
  - Richest data source
  - Limited bandwidth access
- Experiments
  - Basis for models







http://pathguide.org

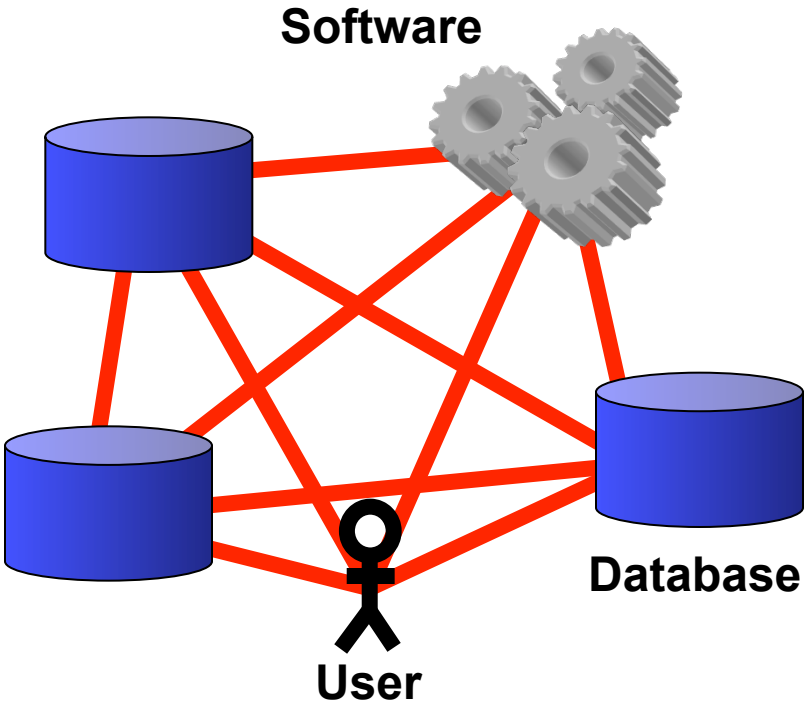


>320 Pathway Databases!

- Varied formats, representation, coverage
- Pathway data extremely difficult to combine and use

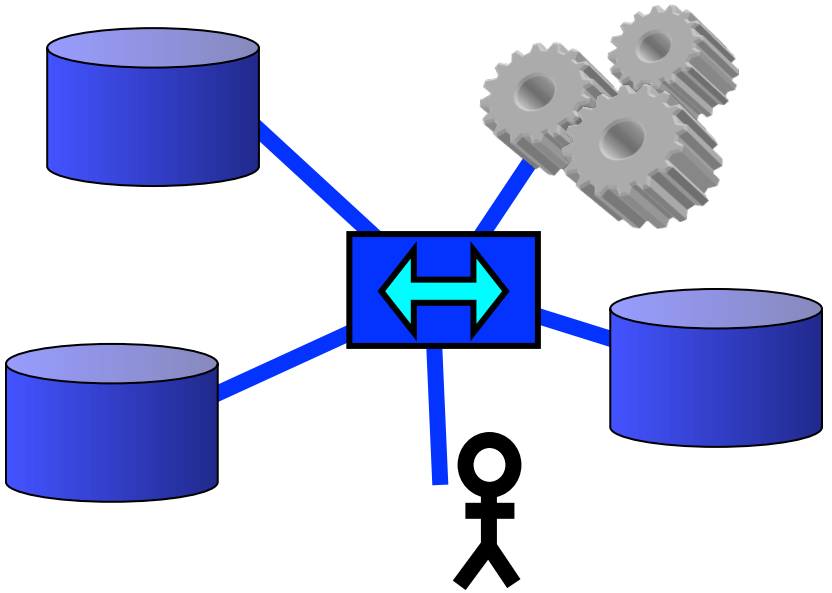
Vuk Pavlovic
Sylva Donaldson

Biological Pathway Exchange (BioPAX)



Before BioPAX

>100 DBs and tools
Tower of Babel



After BioPAX
Unifying language

Reduces work, promotes collaboration, increases accessibility

BioPAX Pathway Language

- Represent:
  - Metabolic pathways
  - Signaling pathways
  - Protein-protein, molecular interactions
  - Gene regulatory pathways
  - Genetic interactions
- Community effort: pathway databases distribute pathway information in standard format

www.biopax.org

BioPAX Supporting Groups

Many Participants

- Memorial Sloan-Kettering Cancer Center: E.Demir, M. Cary, C. Sander
- University of Toronto: G. Bader
- SRI Bioinformatics Research Group: P. Karp, S. Paley, J. Pick
- Bilkent University: U. Dogrusoz
- Université Libre de Bruxelles: C. Lemer
- CBRC Japan: K. Fukuda
- Dana Farber Cancer Institute: J. Zucker
- Millennium: J. Rees, A. Ruttenberg
- Cold Spring Harbor/EBI: G. Wu, M. Gillespie, P. D'Eustachio, I. Vastrik, L. Stein
- BioPathways Consortium: J. Luciano, E. Neumann, A. Regev, V. Schachter
- Argonne National Laboratory: N. Maltsev, E. Marland, M.Syed
- CST: Peter Hornbeck, David Merberg (Vertex)
- AstraZeneca: E. Pichler
- BIOBASE: E. Wingender, F. Schacherer
- NCI: M. Aladjem, C. Schaefer
- Università di Milano Bicocca, Pasteur, Rennes: A. Splendiani
- Vassar College: K. Dahlquist
- Columbia: A. Rzhetsky

Collaborating Organizations

- Proteomics Standards Initiative (PSI)
- Systems Biology Markup Language (SBML)
- CellML
- Chemical Markup Language (CML)

Databases

- BioCyc, WIT, KEGG, PharmGKB, aMAZE, INOH, Transpath, Reactome, PATIKA, eMIM, NCI PID, CellMap, NetPath

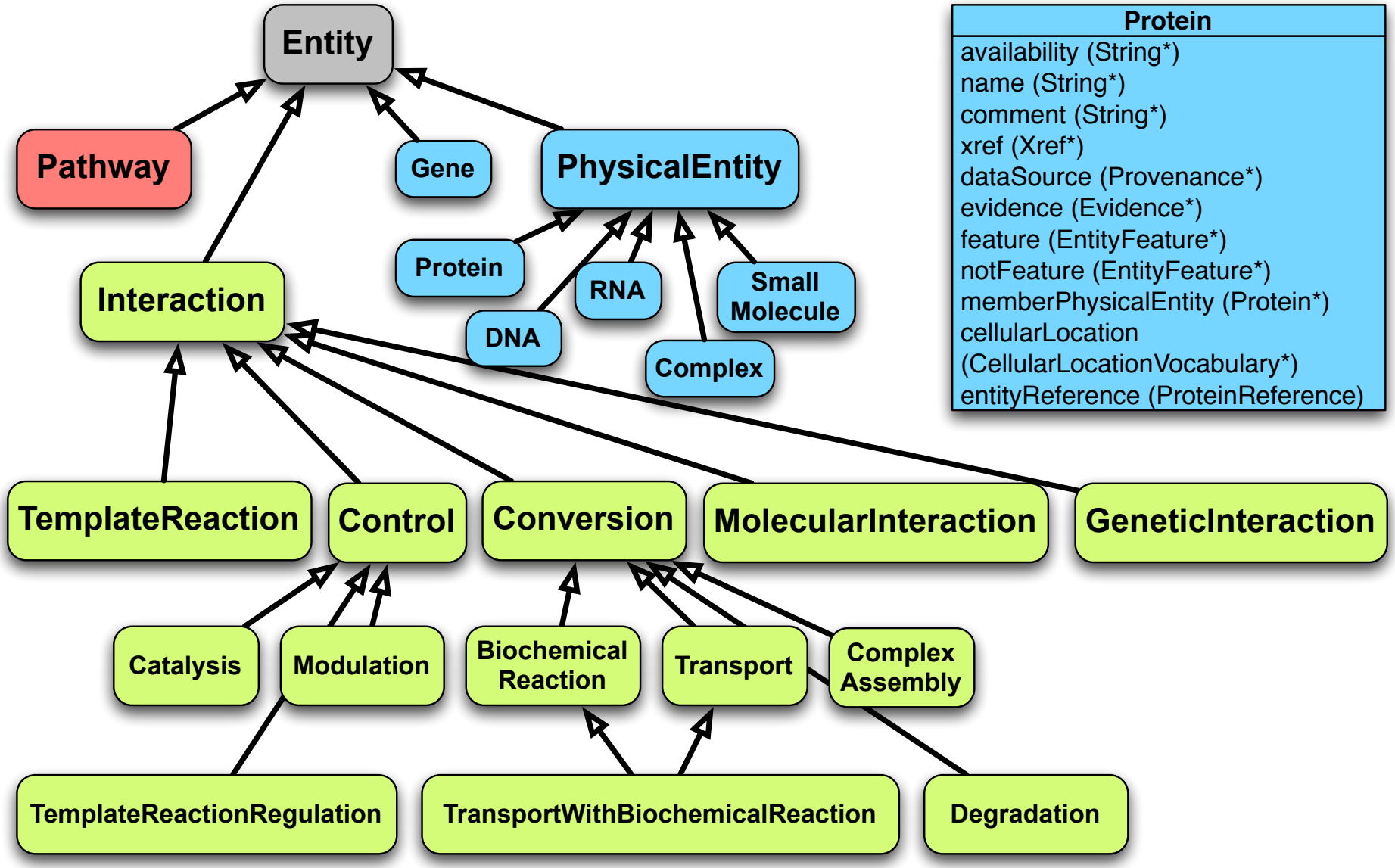
Wouldn't be possible without

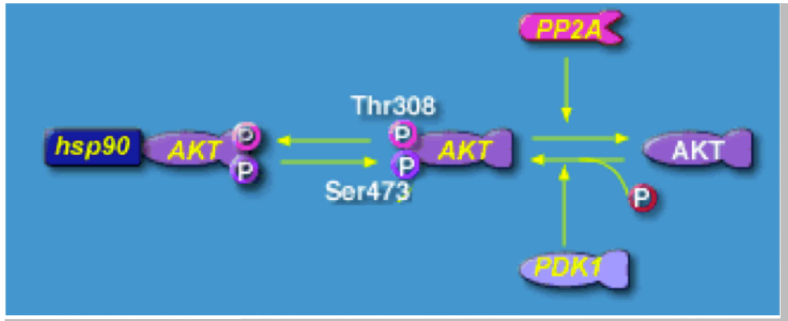
Gene Ontology

Protégé, U.Manchester, Stanford

Grants/Support

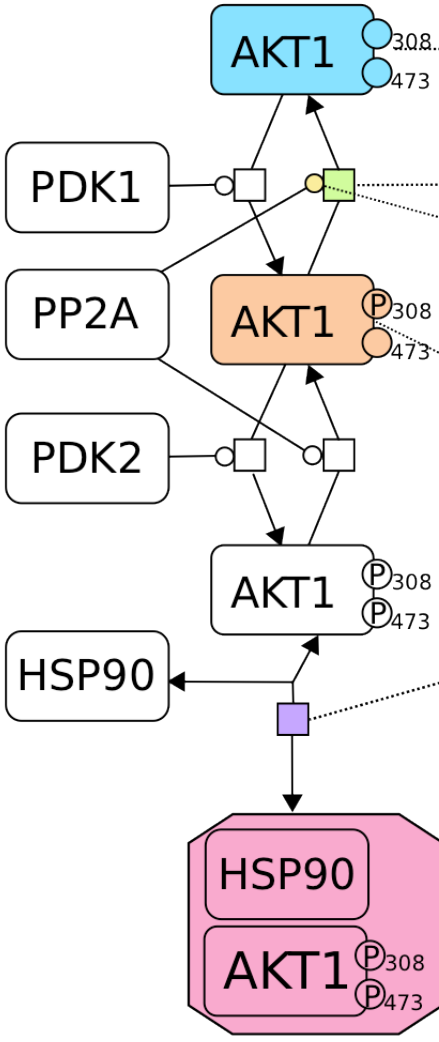
- Department of Energy (Workshop)
- caBIG


Entity
Pathway
Gene
PhysicalEntity
Protein
RNA
Small Molecule
DNA
Complex
Interaction
TemplateReaction
Control
Conversion
MolecularInteraction
GeneticInteraction
Catalysis
Modulation
Biochemical Reaction
Transport
Complex Assembly
TemplateReactionRegulation
TransportWithBiochemicalReaction
Degradation
Protein
availability (String*)
name (String*)
comment (String*)
xref (Xref*)
dataSource (Provenance*)
evidence (Evidence*)
feature (EntityFeature*)
notFeature (EntityFeature*)
memberPhysicalEntity (Protein*)
cellularLocation
(CellularLocationVocabulary*)
entityReference (ProteinReference)



rAKT1 is a *ProteinReference*
has *standard-name* "AKT1"
has *name* "PKB"
has *xref* **Uniprot-P31749**

p@308 is a *ModificationFeature*
has *featureLocation* **AKT1-308**
has *modificationType*
phosphorylation



AKT1.1 is a *Protein*
has *proteinReference* **rAKT1**
has *notFeature* **p@308**
has *notFeature* **p@473**

reaction1 is a *BiochemicalReaction*
has *left* **AKT1.2**
has *right* **AKT1.1**
is *left-to-right.*

catalysis1 is a *Catalysis*
has *controller* **PP2A.1**
has *controlled* **reaction1**
has *direction irr-left-to-right*

AKT1.2 is a *Protein*
has *proteinReference* **rAKT1**
has *feature* **p@308**
has *notFeature* **p@473**

assembly1 is a *ComplexAssembly*
has *left* **HSP90.1**
has *left* **AKT1.3**
has *right* **complex1**
is *reversible*

complex1 is a *Complex*
has *component* **AKT1.4**
has *component* **HSP90.2**

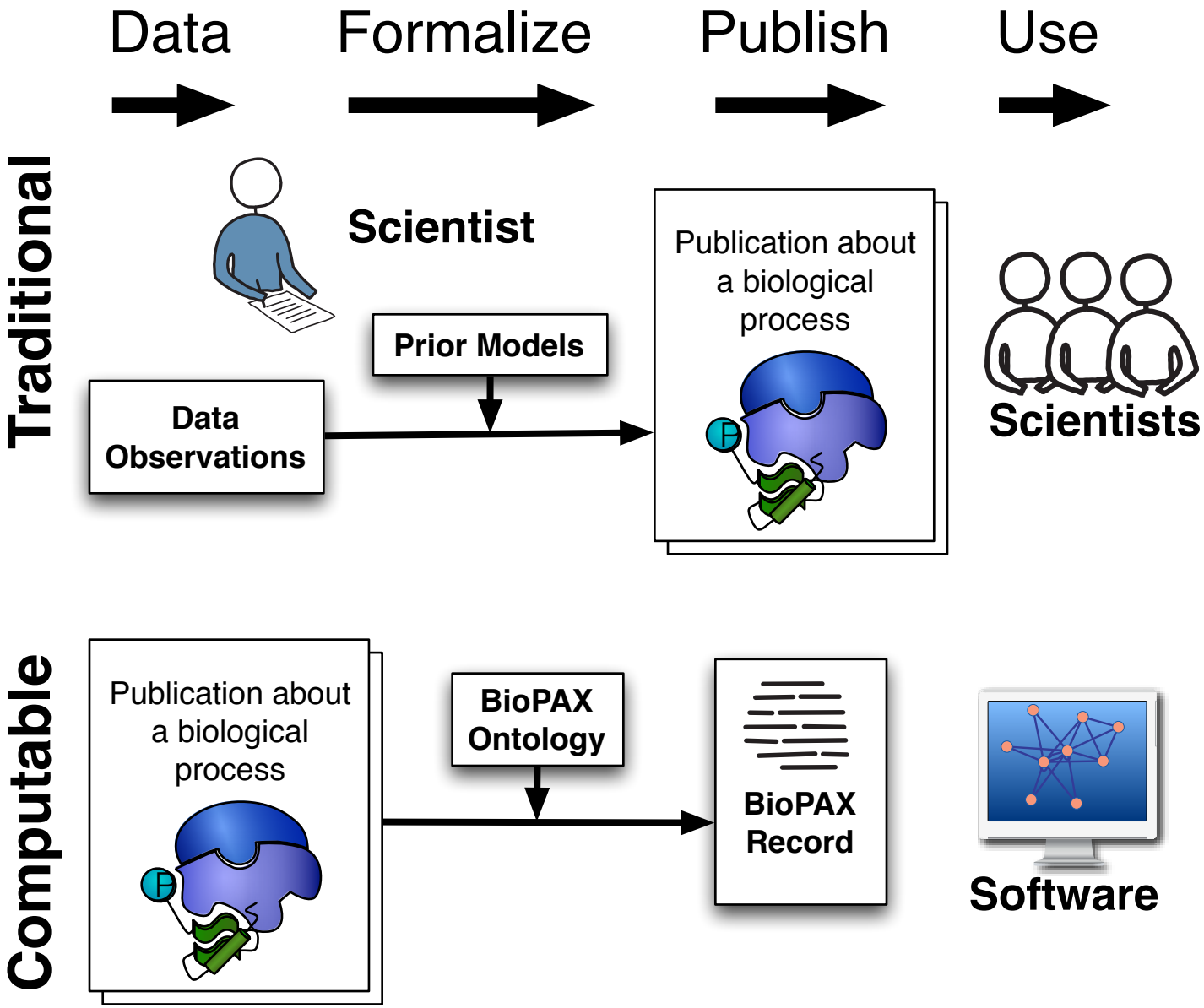
HSP90.2 is a *Protein*
has *proteinReference* **rHSP90**
is *boundTo* **AKT1.4**

AKT1.4 is a *Protein*
has *proteinReference* **rAKT1**
has *feature* **p@308**
has *feature* **p@473**
is *boundTo* **HSP90.2**

XML Snippet (OWL)

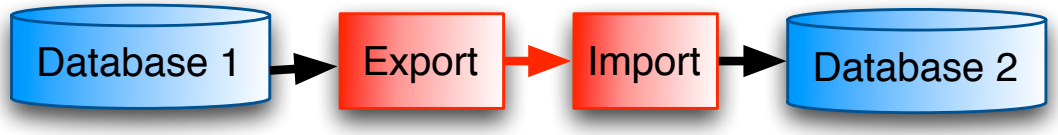
```
<bp:biochemicalReaction rdf:ID="biochemicalReaction37">
  <bp:DATA-SOURCE rdf:resource="#dataSource14"/>
  <bp:LEFT>
    <bp:physicalEntityParticipant rdf:ID="physicalEntityParticipant26">
      <bp:STOICHIOMETRIC-COEFFICIENT>1.0</bp:STOICHIOMETRIC-COEFFICIENT>
      <bp:PHYSICAL-ENTITY>
        <bp:smallMolecule rdf:ID="smallMolecule27">
          <bp:SHORT-NAME rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
          >a-D-glu-6-p</bp:SHORT-NAME>
          <bp:CHEMICAL-FORMULA rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
          >C6H13O9P</bp:CHEMICAL-FORMULA>
          <bp:SYNONYMS rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
          >&lt;FONT FACE="Symbol">a&lt;/FONT>-D-glucose-6-phoshate</bp:SYNONYMS>
          <bp:XREF>
            <bp:unificationXref rdf:ID="unificationXref30">
              <bp:ID rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
              >C00668</bp:ID>
              <bp:DB rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
              >KEGG</bp:DB>
            </bp:unificationXref>
          </bp:XREF>
          <bp:XREF rdf:resource="#unificationXref29"/>
          <bp:MOLECULAR-WEIGHT>260.14</bp:MOLECULAR-WEIGHT>
          <bp:AVAILABILITY rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
          >see http://www.amaze.ulb.ac.be/</bp:AVAILABILITY>
          <bp:SYNONYMS rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
          >glucose-6-P</bp:SYNONYMS>
          <bp:DATA-SOURCE rdf:resource="#dataSource14"/>
          <bp:SYNONYMS rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
          >alpha-D-glucose-6-p</bp:SYNONYMS>
          <bp:STRUCTURE>
            <bp:chemicalStructure rdf:ID="chemicalStructure28">
              <bp:STRUCTURE-FORMAT>SMILES</bp:STRUCTURE-FORMAT>
              <bp:STRUCTURE-DATA>C(OP(=O)(O)O)[CH]1([CH](O)[CH](O)[CH](O)[CH](O)O1)</bp:STRUCTURE-DATA>
            </bp:chemicalStructure>
          </bp:STRUCTURE>
          <bp:NAME>alpha-D-glucose 6-phosphate</bp:NAME>
          <bp:SYNONYMS rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
          >alpha-D-glucose-6-phosphate</bp:SYNONYMS>
          <bp:SYNONYMS rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
          > D-glucose-6-P</bp:SYNONYMS>
          <bp:DATA-SOURCE rdf:resource="#KB_439584_Individual_47"/>
        </bp:smallMolecule>
      </bp:PHYSICAL-ENTITY>
      <bp:CELLULAR-LOCATION rdf:resource="#openControlledVocabulary15"/>
    </bp:physicalEntityParticipant>
  </bp:LEFT>
  <bp:DELTA-G rdf:datatype="http://www.w3.org/2001/XMLSchema#double"
  >0.4</bp:DELTA-G>
  <bp:SYNONYMS rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >alpha-D-Glucose 6-phosphate &lt;=> beta-D-Fructose 6-phosphate </bp:SYNONYMS>
  <bp:RIGHT>
    <bp:physicalEntityParticipant rdf:ID="physicalEntityParticipant38">
      <bp:CELLULAR-LOCATION rdf:resource="#openControlledVocabulary15"/>
      <bp:PHYSICAL-ENTITY>
        <bp:smallMolecule rdf:ID="smallMolecule39">
```

Pathway Information Processing

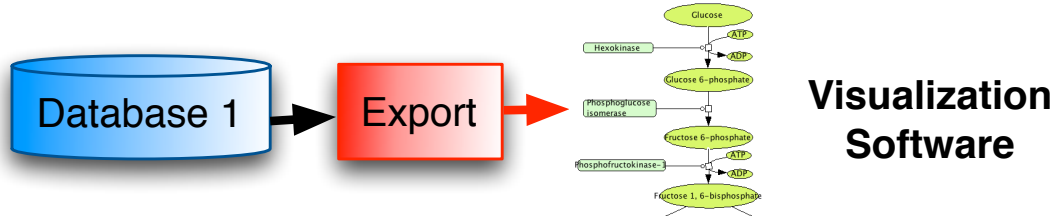


BioPAX uses

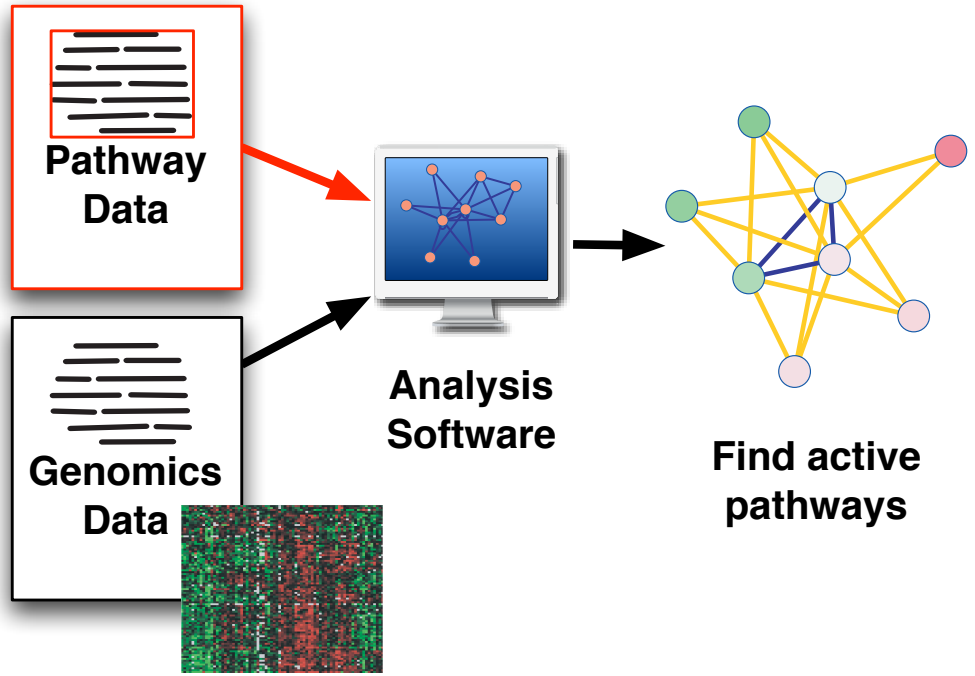
Data Exchange Between Database Groups



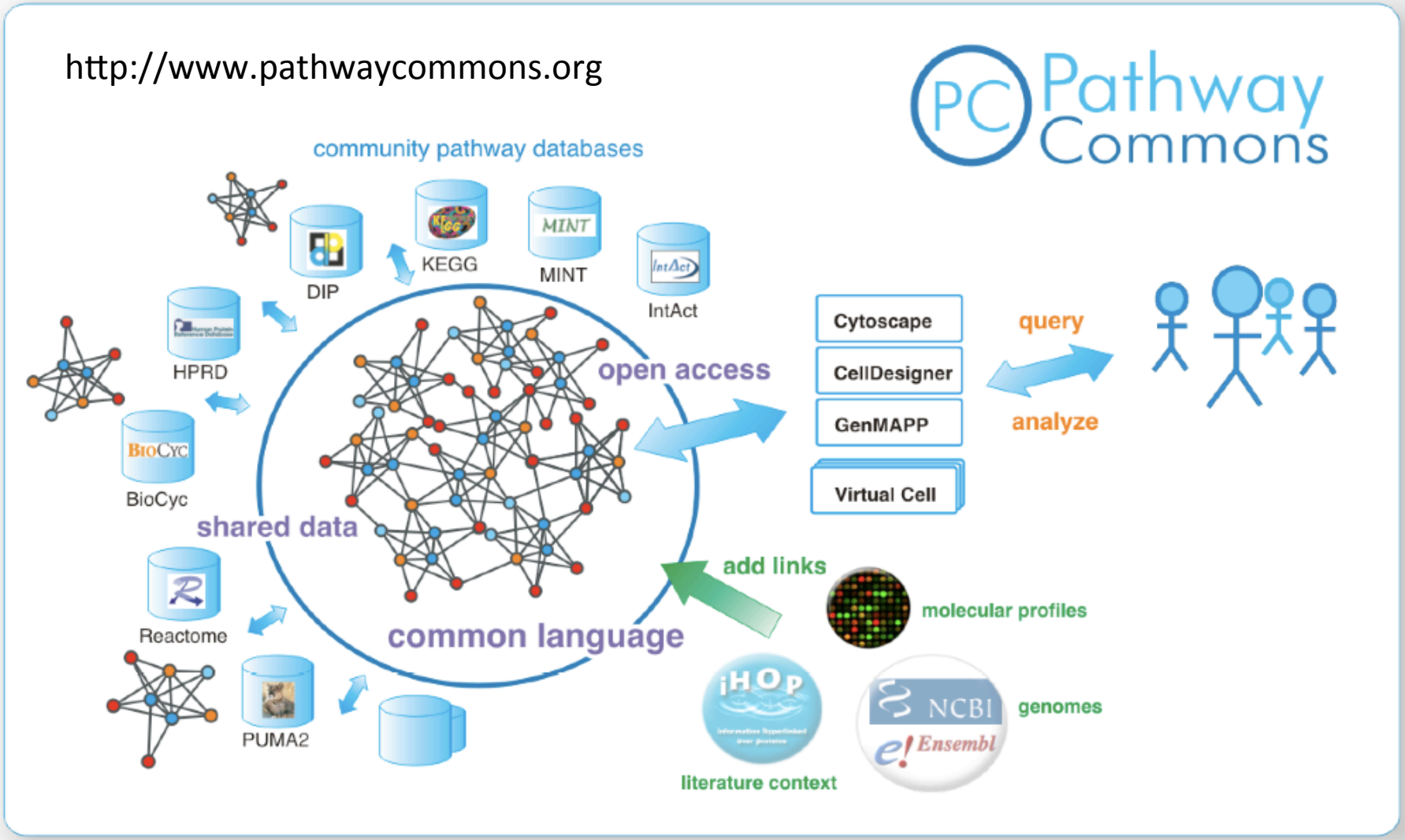
Pathway Visualization From Database



Pathway Analysis of Genomics Data



Aim: Convenient Access to Pathway Information



Facilitate creation and communication of pathway data
Aggregate pathway data in the public domain
Provide easy access for pathway analysis

Long term: Converge to integrated cell map

http://pathwaycommons.org

Pathway Commons

Search and visualize public biological pathway information. Single point of access. [more...]

Home | Filter | Data Sources | Download | FAQ | Web Service | About

Send us your feedback. Sign up for Pathway Commons announcements. RSS Feed

Search Pathway Commons:

Search

To get started, enter a gene name, gene identifier or pathway name in the text box above. For example: BRCA1, P38398 or mTOR.

To restrict your search to specific data sources or specific organisms, update your global filter settings.

What's New:

- NEW! July 2, 2009:
  - Batch Download of all Pathway Commons data in multiple file formats is now available.
  - Systems Biology Center New York - IMID data set (December 17, 2008 Version 27).
  - Latest Reactome data set (June 24, 2009 Version 29).
  - Latest HumanCyc data set (June 22, 2009 Version 13.1).
  - All yeast proteins are now annotated with UniProt functional annotation.
- March 1, 2009:
  - BioGRID data set (January 28, 2009 Version 2.0.49).
  - Latest Reactome data set (December 17, 2008 Version 27).
  - Latest HumanCyc data set (October 15, 2008 Version 12.5).
  - Neighborhood maps added to protein pages.
- July 24, 2008:
  - Latest Reactome data set (June 30, 2008 Version 25).
  - All human, mouse and rat proteins are now annotated with UniProt functional annotation.
  - Improved search support for gene symbols.
  - Stable links now available for linking out to protein pages.

Using Pathway Commons:

Biologists: Browse and search pathways across multiple valuable public pathway databases.

Computational biologists: Download an integrated set of pathways in BioPAX format for global analysis.

Software developers: Build software on top of Pathway Commons using our web service API. Download and install the cPath software to create a local mirror.

Current Data Sources:

Pathway Commons currently contains the following data sources (batch download):

BioGRID | Memorial Sloan-Kettering Cancer Center | Human Protein Reference Database
HumanCyc | SBCNY | IntAct
MINT | National Cancer Institute nature | Reactome

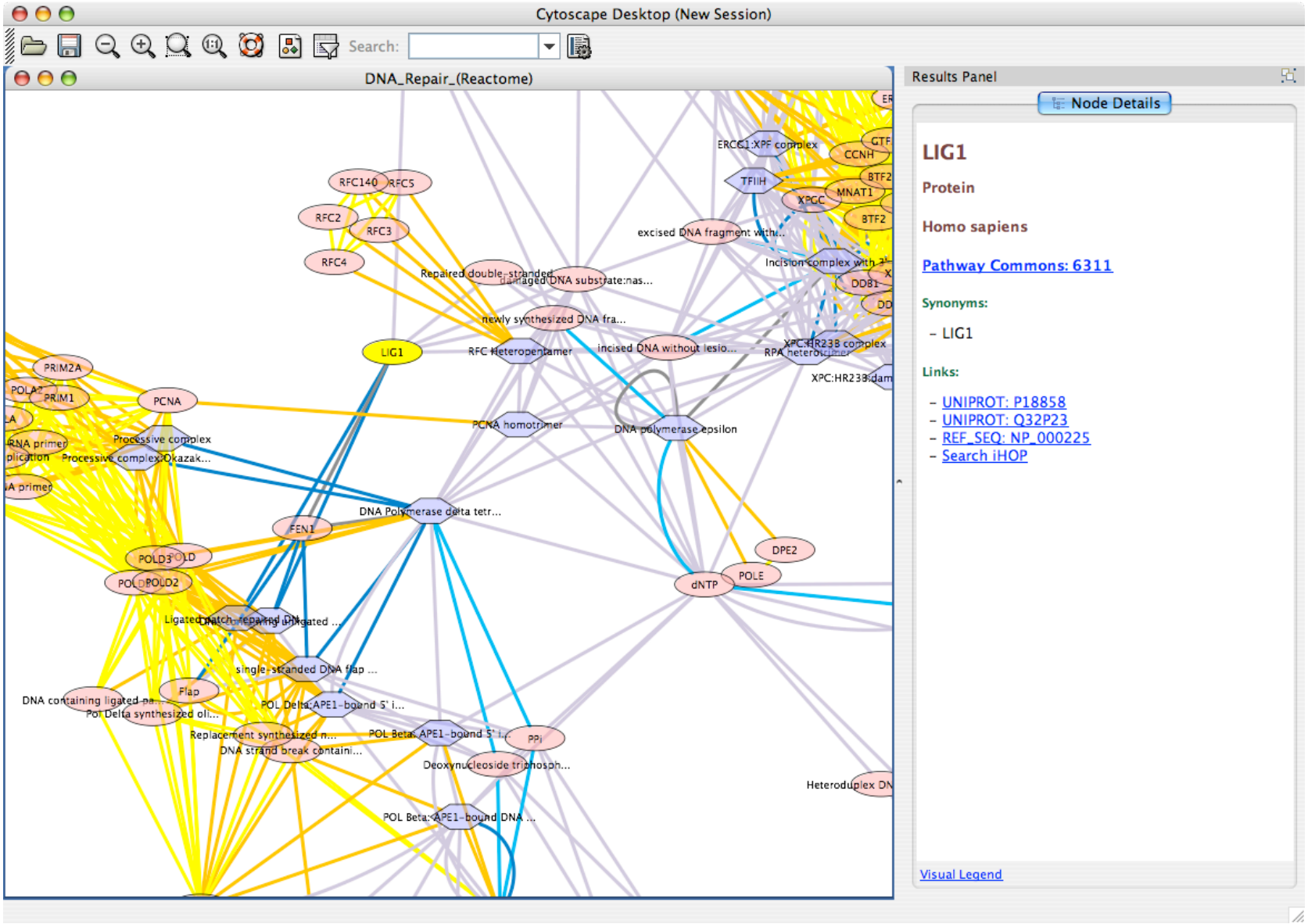
Pathway Commons Quick Stats:

Number of Pathways:	1,449
Number of Interactions:	421,395
Number of Physical Entities:	88,509
Number of Organisms:	441

Integration of additional data sources is planned in the near future. For a comprehensive directory of interaction and pathway databases, please refer to Pathguide.

http://pathwaycommons.org

Access From Cytoscape

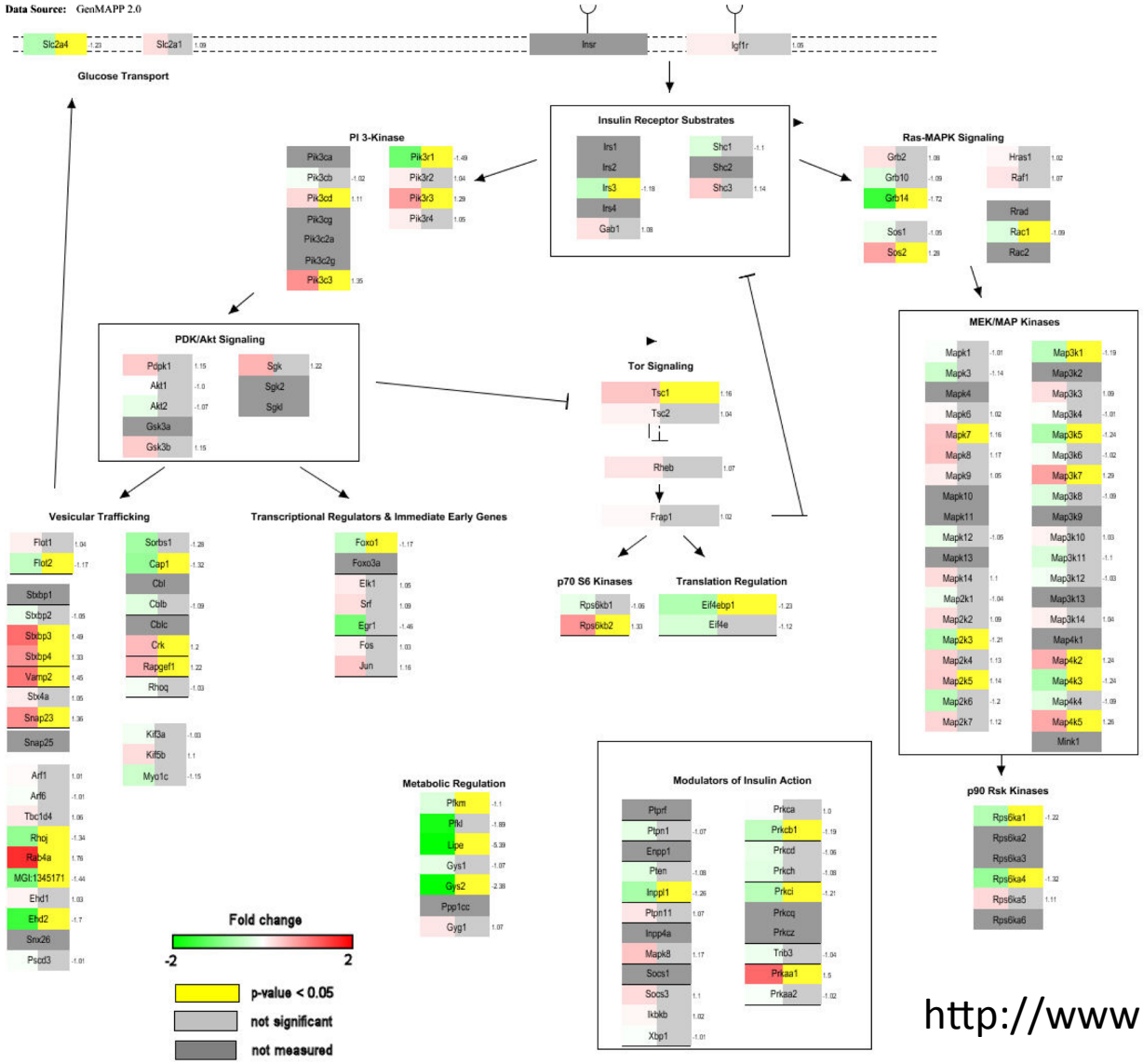


Download Service

http://www.pathwaycommons.org/pc-snapshot/

Name	Last modified	Size	Description
Parent Directory		-	
README.TXT	29-Jun-2009 12:28	4.1K	
biopax/	29-Jun-2009 12:27	-	
gene_sets/	29-Jun-2009 12:28	-	
gsea/	29-Jun-2009 12:28	-	
sif/	29-Jun-2009 12:28	-	
tab_delim_network/	29-Jun-2009 12:28	-	

Pathway Visualization: Pathvisio



http://www.pathvisio.org/

What Have We Learned?

- Where can you get pathway information?
  - Databases, literature, experts, experiments
- Many databases exist
  - Increasing convenience, but still difficult to combine and use

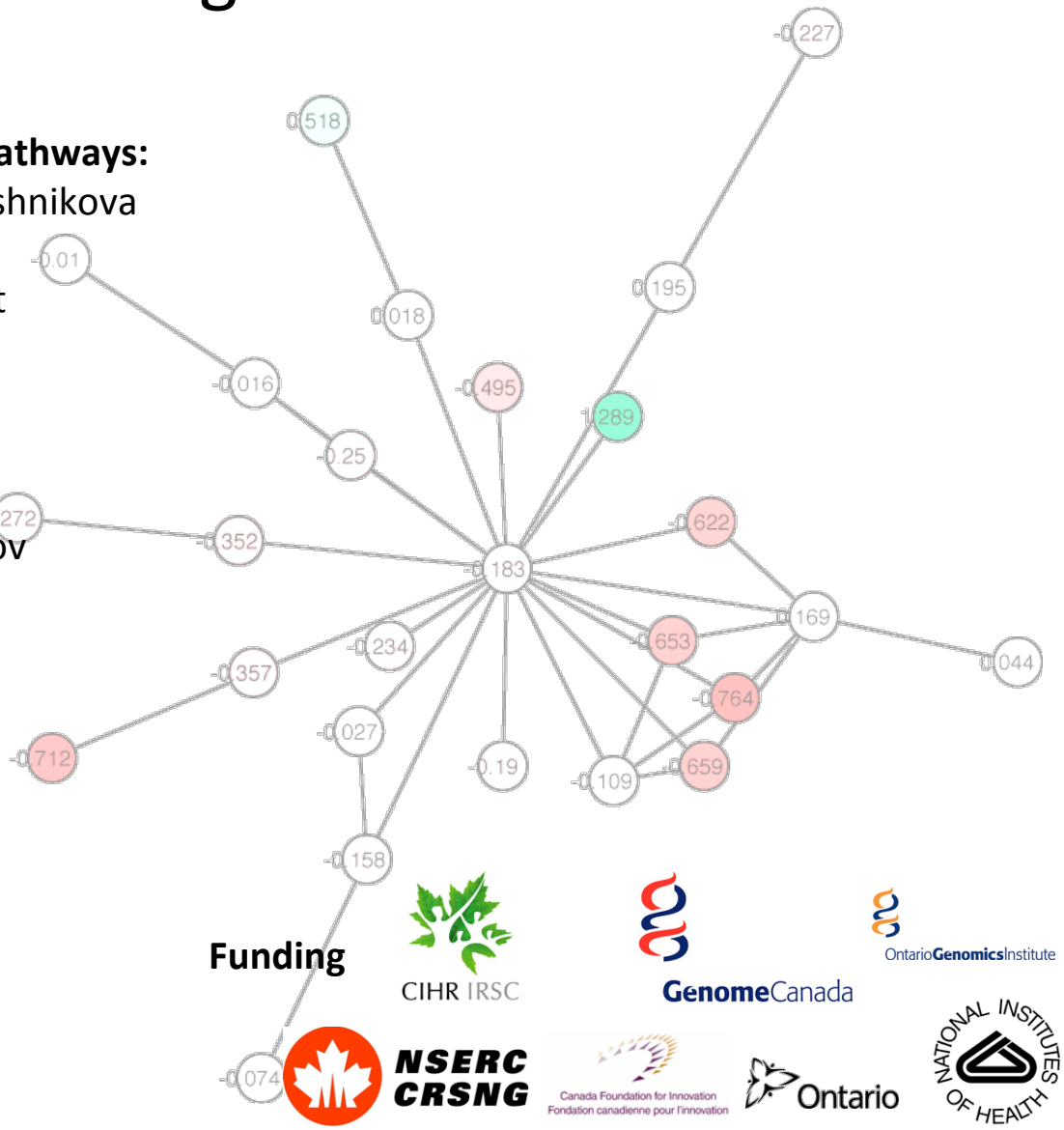
Acknowledgements

Bader Lab
Domain Interaction Team
Chris Tan
Shirley Hui
Shobhit Jain
Brian Law
Jüri Reimand

Former:
David Gfeller
Xiaojian Shao

Genetic Intx, Pathways:
Anastasia Baryshnikova
Iain Wallace
Magali Michaut
Ron Ammar
Daniele Merico
Ruth Isserlin
Vuk Pavlovic
Igor Rodchenkov

www.GeneMANIA.org
Quaid Morris (Donnelly)
Rashad Badrawi, Ovi Comes, Sylva Donaldson, Christian Lopes, Farzana Kazi, Jason Montojo, Harold Rodriguez, Khalid Zuberi



Funding
CIHR IRSC
GenomeCanada
OntarioGenomicsInstitute
NSERC CRSNG
Canada Foundation for Innovation
Fondation canadienne pour l'innovation
Ontario
National Institutes of Health

http://baderlab.org

Acknowledgements

Cytoscape
Trey Ideker (UCSD)
Mike Smoot, Kei Ono, Peng Liang Wang (Ryan Kelley, Nerius Landys, Chris Workman, Mark Anderson, Nada Amin, Owen Ozier, Jonathan Wang)

Lee Hood, Ilya Schmulevich (ISB)
Sarah Killcoyne (Iliana Avila-Campillo, Rowan Christmas, Andrew Markiel, Larissa Kamenkovich, Paul Shannon)

Benno Schwikowski (Pasteur)
Mathieu Michaud (Melissa Cline, Tero Aittokallio, Andrea Splendiani)

Chris Sander (MSKCC)
Ethan Cerami, Ben Gross (Robert Sheridan)

Annette Adler (Agilent) **Yeyejide Adeleye (Unilever)**
Allan Kuchinsky, Mike Creech (Aditya Vailaya) Noel Ruddock

Bruce Conklin (UCSF) **Scooter Morris (UCSF)** **David States (Texas)**
Alex Pico, Kristina Hanspers

Pathway Commons
Chris Sander
Ethan Cerami
Ben Gross
Emek Demir
Igor Rodchenkov
Nadia Anwar
Ozgun Babur

NIGMS GM070743-01

http://baderlab.org

Acknowledgments

Dalila Pinto[1], Alistair Pagnamenta[2], Lambertus Klei[3], Richard Anney[4], **Daniele Merico**[5], Regina Regan[6], Judith Conroy[6], Tiago Magalhaes[7], Catarina Correia[7], Brett S. Abrahams[8], Joana Almeida[9], Elena Bacchelli[10], **Gary D. Bader**[5], Anthony J. Bailey[11*], Gillian Baird[12], Agatino Battaglia[13*], Tom Berney[14], Nadia Bolshakova[4], Sven Bölte[15], Patrick F. Bolton[16], Thomas Bourgeron[17], Sean Brennan[4], Jessica Brian[18], Susan E. Bryson[19], Andrew R. Carson[1], Guillermo Casallo[1], Jillian Casey[6], Lynne Cochrane[4], Christina Corsello[20], Emily L. Crawford[21], Andrew Crossett[22], Geraldine Dawson[23,24*], Maretha de Jonge[25], Richard Delorme[26], Irene Drmic[18], Eftichia Duketis[15], Frederico Duque[9], Annette Estes[27], Penny Farrar[2], Bridget A. Fernandez[28], Tiziana Filippi[13], Eric Fombonne[29], Christine M. Freitag[15*], John Gilbert[30], Christopher Gillberg[31], Joseph T. Glessner[32], Jeremy Goldberg[33], Andrew Green[6], Jonathan Green[34], Stephen J. Guter[35], Hakon Hakonarson[32,36*], Elizabeth A. Heron[4], Matthew Hill[4], Richard Holt[2], Jennifer L. Howe[1], Gillian Hughes[4], Vanessa Hus[20], Roberta Igliozzi[13], Cecilia Kim[32], Sabine M. Klauck[37*], Alexander Kolevzon[38], Olena Korvatska[39], Vlad Kustanovich[40], Clara M. Lajonchere[40], Janine A. Lamb[41], Magdalena Laskawiec[11], Marion Leboyer[42], Ann Le Couteur[14], Bennett L. Leventhal[35], Anath C. Lionel[1], Xiao-Qing Liu[1], Catherine Lord[20], Linda Lotspeich[43], Sabata C. Lund[21], Elena Maestrini[10*], William Mahoney[44], Carine Mantoulan[45], Christian R. Marshall[1], Helen McConachie[14], Christopher J. McDougle[46], Jane McGrath[4], William M. McMahon[47*], Alison Merikangas[4], Ohsuke Migita[1], Nancy J. Minshew[48], Ghazala K. Mirza[2], Jeff Munson[25], Stanley F. Nelson[49*], Carolyn Noakes[18], Abdul Noor[50], Gudrun Nygren[31] , Guiomar Oliveira[9*], Katerina Papanikolaou[51], Jeremy R. Parr[11], Barbara Parrini[13], Tara Paton[1], Andrew Pickles[52], Marion Pilorge[53], Joseph Piven[54*], Chris P. Ponting[55], David J Posey[46], Annemarie Poustka[37X], Fritz Poustka[15], Aparna Prasad[1], Jiannis Ragoussis[2], Katy Renshaw[11], Jessica Rickaby[1], Wendy Roberts[18], Kathryn Roeder[22], Bernadette Roge[45], Michael L. Rutter[56], Laura J. Bierut[57], John P. Rice[57], SAGE Consortium, Jeff Salt[35], Katherine Sansom[1], Daisuke Sato[1], Ricardo Segurado[4], Lili Senman[18], Naisha Shah[6], Val C. Sheffield[58], Latha Soorya[38], Inês Sousa[2], Olaf Stein[59], Vera Stoppioni[60], Christina Strawbridge[33], Raffaella Tancredi[13], Katherine Tansey[4], Bhooma Thiruvahindrapduram[1], Ann P. Thompson[33], Susanne Thomson[21], Ana Tryfon[38], John Tsiantis[51], Herman Van Engeland[25], John B. Vincent[50], Fred Volkmar[61], Simon Wallace[11], Kai Wang[32], Zhouzhi Wang[1], Thomas H. Wassink[62*], Caleb Webber[55], Kirsty Wing[2], Kerstin Wittemeyer[45], Shawn Wood[3], Jing Wu[22], Brian L. Yaspan[21], Danielle Zurawiecki[38], Lonnie Zwaigenbaum[63], Joseph D. Buxbaum[38*], Rita M. Cantor[49*], Edwin H. Cook[35*], Hilary Coon[47], Michael L. Cuccaro[30], Bernie Devlin[3*], Sean Ennis[6*], Louise Gallagher[4*], Daniel H. Geschwind[8*], Michael Gill[4*], Jonathan L. Haines[64*], Joachim Hallmayer[43*], Judith Miller[47], Anthony P. Monaco[2*], John I. Nurnberger Jr.[46*], Andrew D. Paterson[1*], Margaret A. Pericak-Vance[30*], Gerard D. Schellenberg[65*], Peter Szatmari[33*], Astrid M. Vicente[7*], Veronica J. Vieland[59*], Ellen M. Wijsman[66*], **Stephen W. Scherer**[1*], James S. Sutcliffe[21*], Catalina Betancur[53*]

Sick Kids Hospital, Toronto

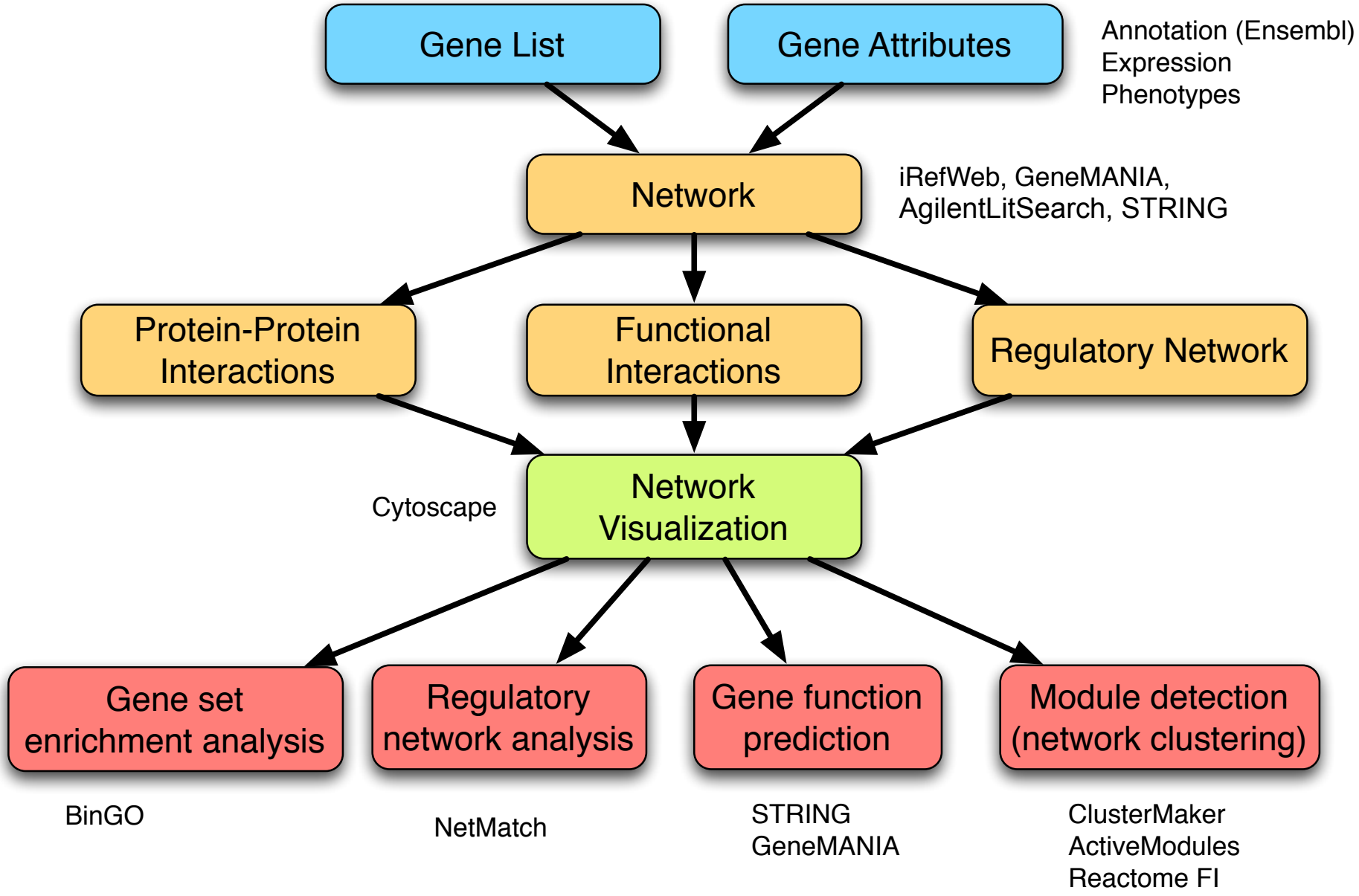
Donnelly

Lunch

12:30-14:00

Cytoscape Lab

Gene List and Network Analysis Overview



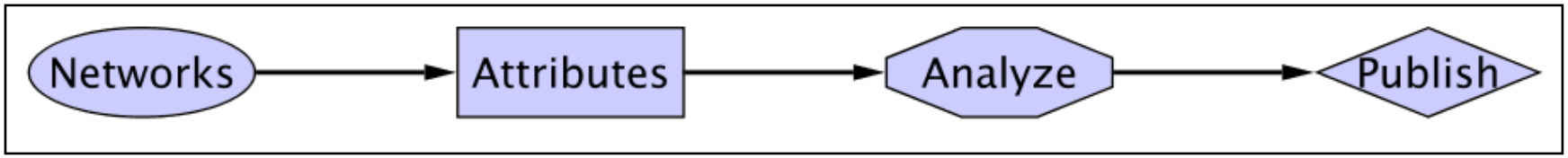
Cytoscape Workflow

Piet Molenaar

Cytoscape Workflow

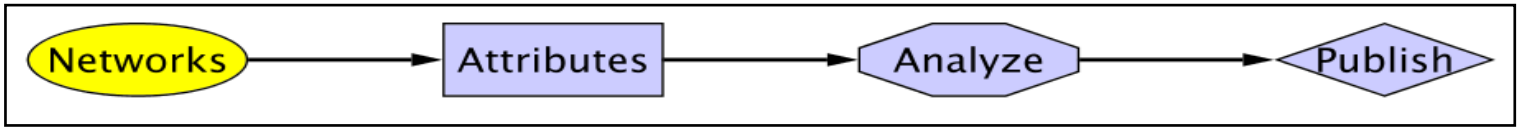
1. Load Networks (Import network data into Cytoscape)
2. Load Attributes (Get data about networks into Cytoscape)
3. Analyze and Visualize Networks
4. Prepare for Publication

- A specific example of this workflow:
  - Cline, et al. "Integration of biological networks and gene expression data using Cytoscape", Nature Protocols, 2, 2366-2382 (2007).



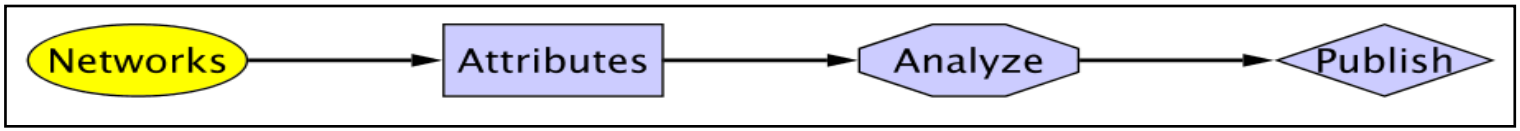
All kinds of network data…

- Physical interactions
  - Protein – Protein interactions
  - Protein – DNA interactions
  - Metabolic interactions
- Functional interactions
  - Co-expression relations
  - Genetic interactions
  - Knockout/siRNA – targets



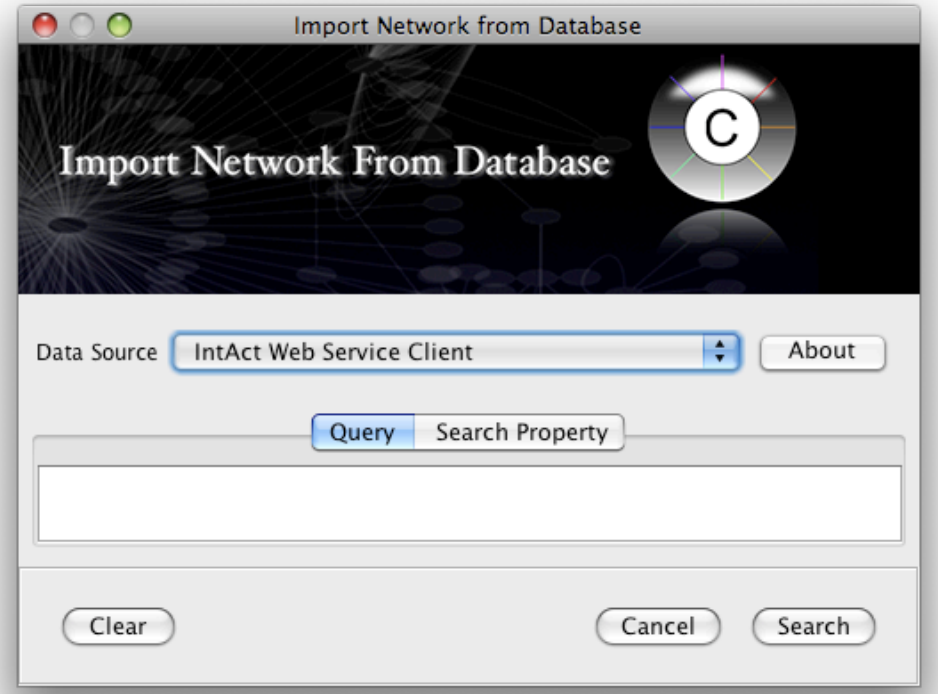
Pre-formatted Network Files

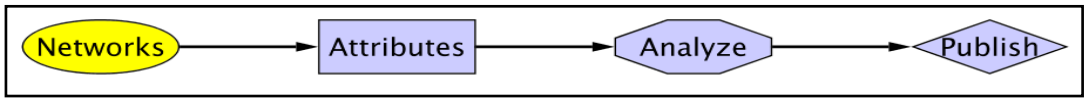
- Cytoscape supports many popular file formats:
  - SIF (Simple Interaction Format)
  - GML (Graph Markup Language)
  - XGMML (eXtensible Graph Markup and Modeling Language)
  - BioPAX (Biological Pathway Data)
  - PSI-MI 1 & 2.5 (Protein Standards Initiative)
  - SBML Level 2 (Systems Biology Markup Language)
  - KGML (KEGG Markup Language)
- Available for download from data sources (URLs, web-services, formatted table files)



Internet Databases

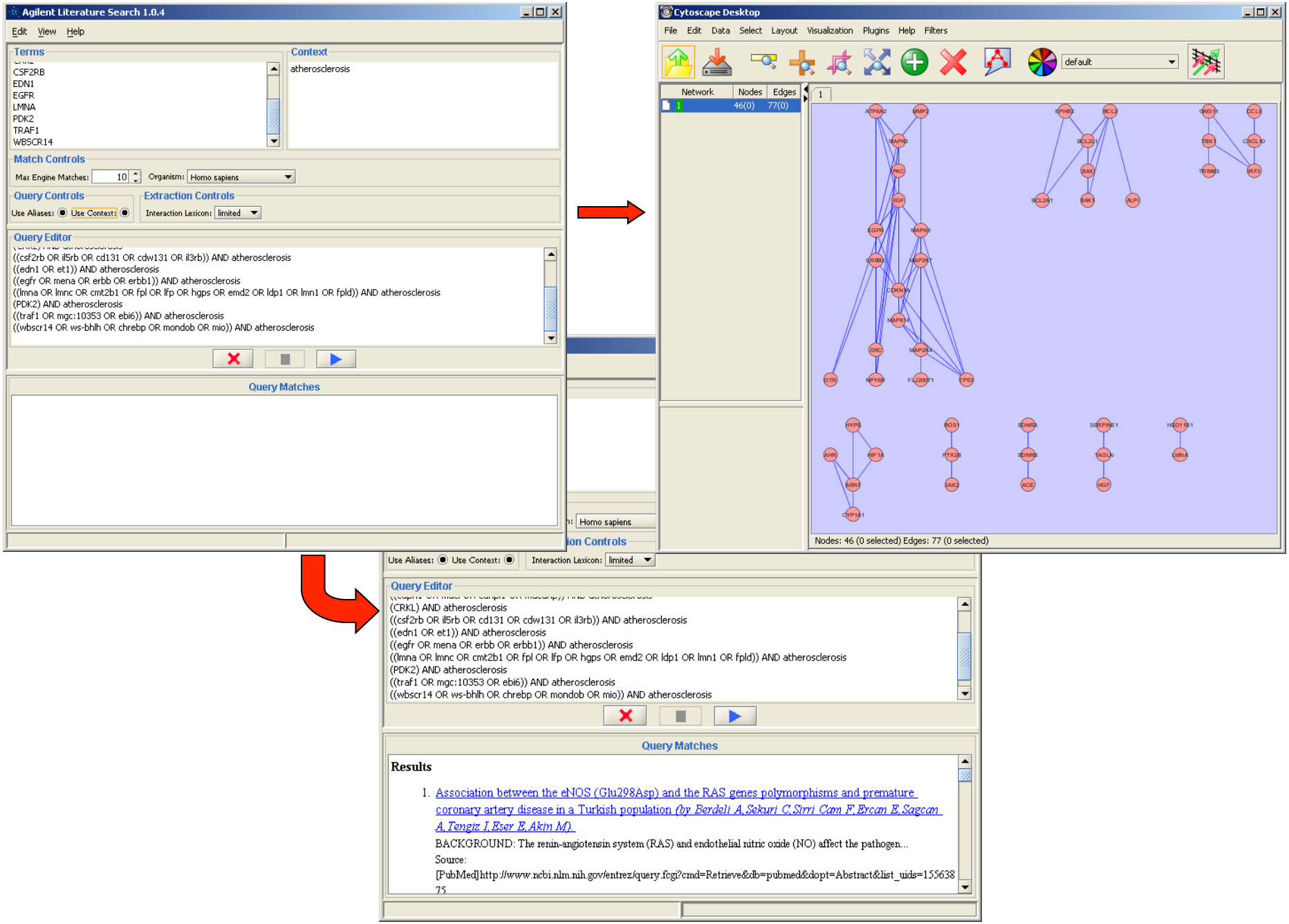
- Cytoscape version 2.6
  - web service clients: import networks directly from several trusted internet resources
    - IntAct (EMBL-EBI)
    - PathwayCommons (collection of data resources)
    - NCBI Entrez Gene
    - Many more will be included...



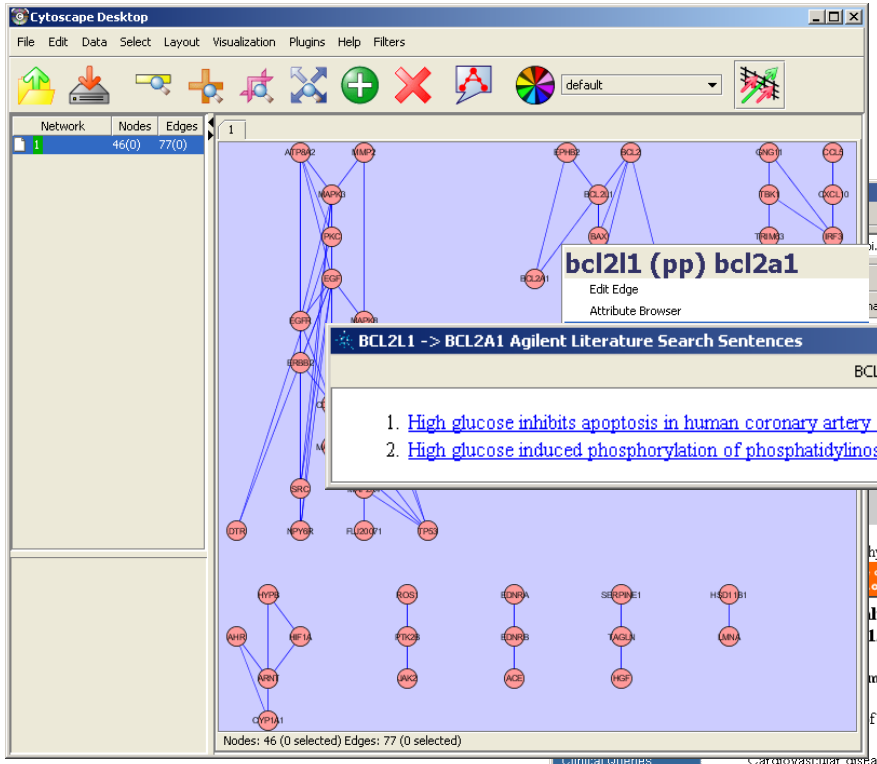


Text Mining

- Computationally extract gene relationships from text, usually PubMed abstracts
- Literature search tool, lots of network data
- BUT not perfect
  - Problems recognizing gene names
  - Natural language processing not perfect
- Agilent Literature Search Cytoscape plugin
- Others: E.g. iHOP
  - www.ihop-net.org/UniPub/iHOP/

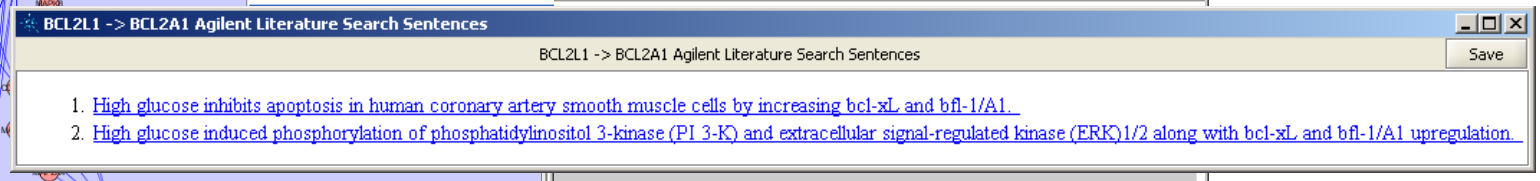

Agilent Literature Search 1.0.4
Edit View Help
Terms
CSF2RB
EDN1
EGFR
LMNA
PDK2
TRAF1
WBSCR14
Context
atherosclerosis
Match Controls
Max Engine Matches: 10
Organism: Homo sapiens
Query Controls
Use Aliases: Use Context:
Extraction Controls
Interaction Lexicon: limited
Query Editor
((csf2rb OR il5rb OR cd131 OR cdw131 OR il3rb)) AND atherosclerosis
((edn1 OR et1)) AND atherosclerosis
((egfr OR mena OR erbb OR erbb1)) AND atherosclerosis
((lmna OR lmnc OR cmt2b1 OR fpl OR lfp OR hgps OR emd2 OR ldp1 OR lmn1 OR fpld)) AND atherosclerosis
(PDK2) AND atherosclerosis
((traf1 OR mgc:10353 OR ebi6)) AND atherosclerosis
((wbscr14 OR ws-bhlh OR chrebp OR mondob OR mio)) AND atherosclerosis
Query Matches
Cytoscape Desktop
File Edit Data Select Layout Visualization Plugins Help Filters
default
Network Nodes Edges
1 46(0) 77(0)
Nodes: 46 (0 selected) Edges: 77 (0 selected)
Homo sapiens
Use Aliases: Use Context:
Interaction Lexicon: limited
Query Editor
(CRKL) AND atherosclerosis
((csf2rb OR il5rb OR cd131 OR cdw131 OR il3rb)) AND atherosclerosis
((edn1 OR et1)) AND atherosclerosis
((egfr OR mena OR erbb OR erbb1)) AND atherosclerosis
((lmna OR lmnc OR cmt2b1 OR fpl OR lfp OR hgps OR emd2 OR ldp1 OR lmn1 OR fpld)) AND atherosclerosis
(PDK2) AND atherosclerosis
((traf1 OR mgc:10353 OR ebi6)) AND atherosclerosis
((wbscr14 OR ws-bhlh OR chrebp OR mondob OR mio)) AND atherosclerosis
Query Matches
Results
1. Association between the eNOS (Glu298Asp) and the RAS genes polymorphisms and premature coronary artery disease in a Turkish population (by Berdeli A, Sekuri C, Sirri Can F, Ercan E, Sagcan A, Tengiz I, Eser E, Akin M)
BACKGROUND: The renin-angiotensin system (RAS) and endothelial nitric oxide (NO) affect the pathogen...
Source:
[PubMed]http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed&dopt=Abstract&list_uids=15563875

Cytoscape Network produced by Literature Search.

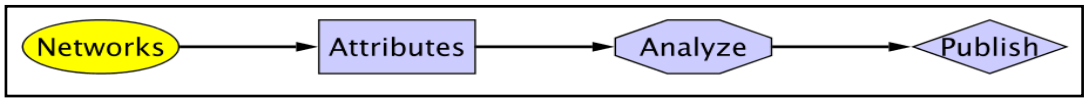


Abstract from the scientific literature

Sentences for an edge

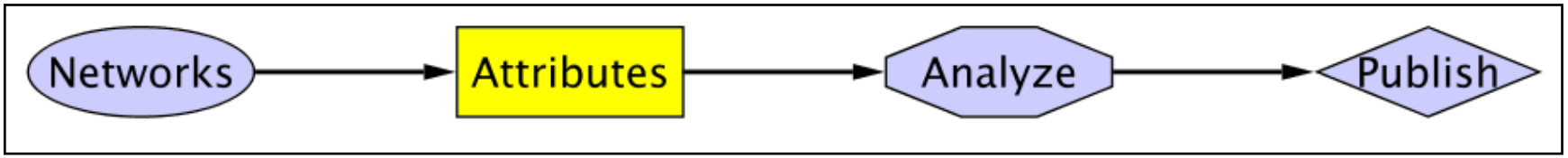


Demo Creating Network From Internet Database



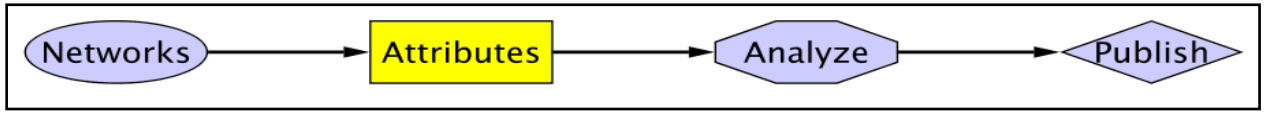
Cytoscape Workflow

1. Load Networks (Get network data into Cytoscape)
2. Load Attributes (Get data about networks into Cytoscape)
3. Analyze and Visualize Networks
4. Prepare for Publication



What are Attributes?

- Any data that describes or provides details about the nodes and edges in the network
  - Gene Expression Data
  - Mass Spectrometry Data
  - Protein Structure Information
  - Gene Ontology (GO) terms
  - Interaction Confidence Values, etc
- Cytoscape support multiple data types
  - Numbers (integer, float)
  - Text (string)
  - Logical (Boolean)
  - Lists…

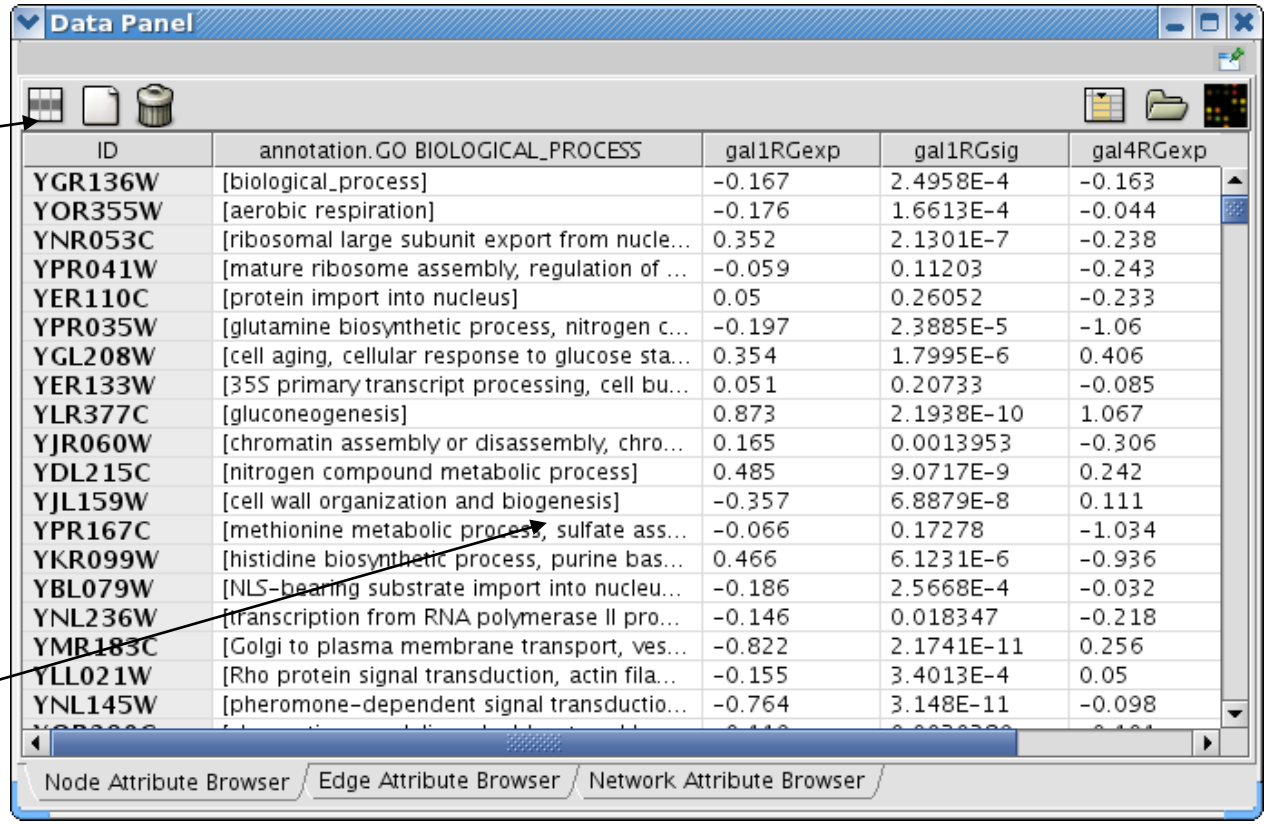


Attribute Management

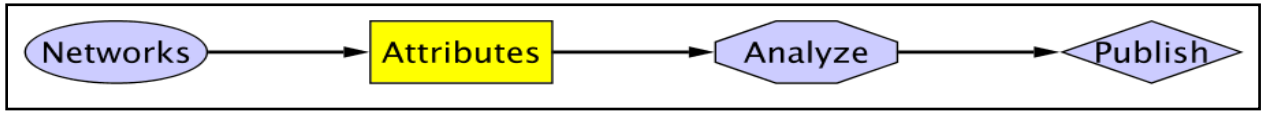
Select Attributes for Display

Node or Edge ID

Strings and floating type of attributes

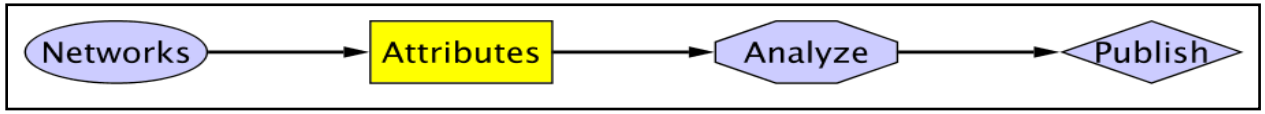


Specific Attribute Tabs



Load Attributes:
Import Attribute Files

- Map data about Networks onto Networks.
- Attributes can be loaded in many of the same ways as networks.
  - Import pre-formatted attribute files
  - Import formatted text or Excel files
  - Create attributes manually in attribute editor
  - Load attributes from web services
  - ID mapping though node attributes

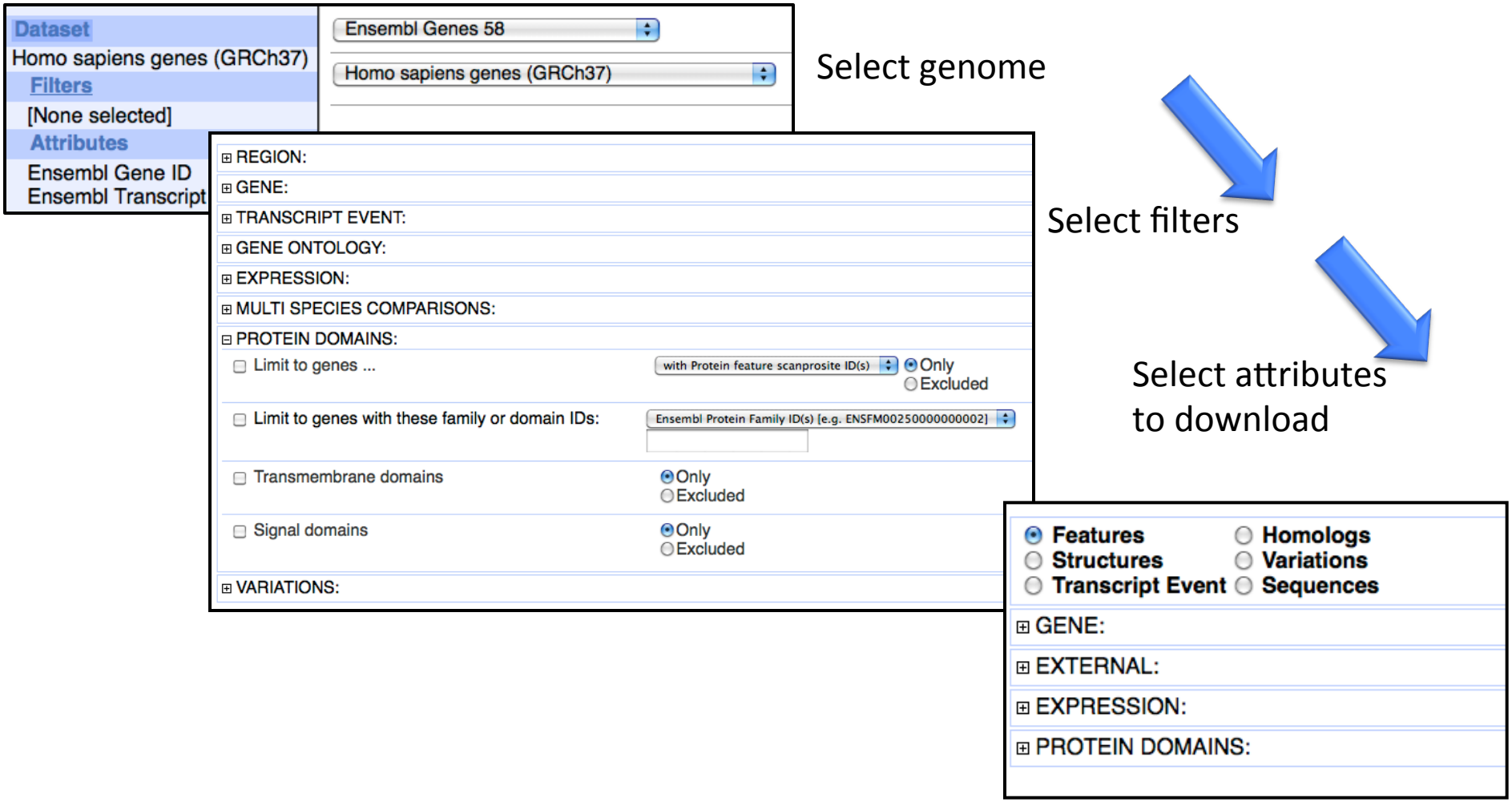


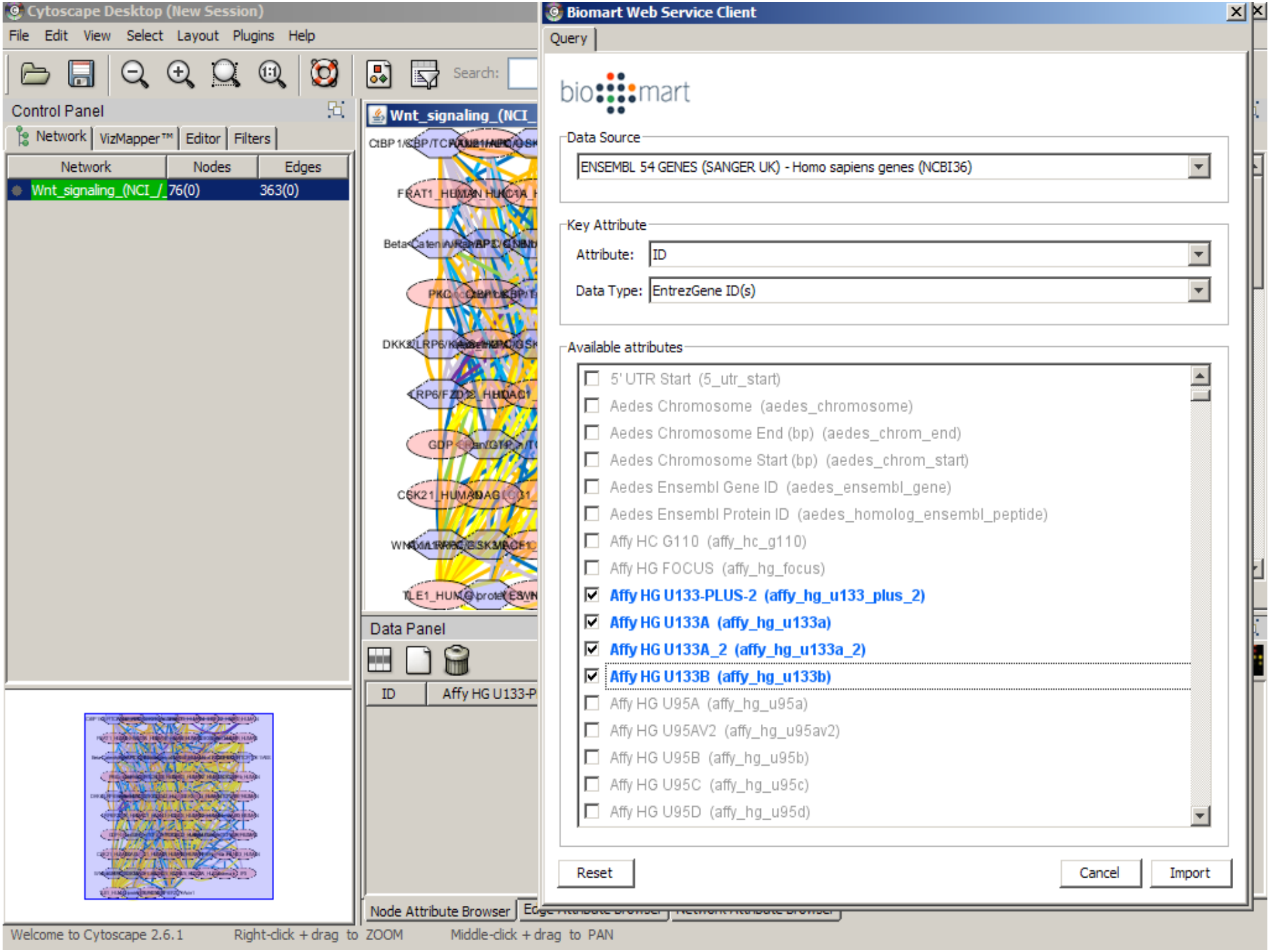
Public Sources of Gene Attributes

- Ensembl BioMart (eukaryotes)
  - 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: discuss during lab

Ensembl BioMart

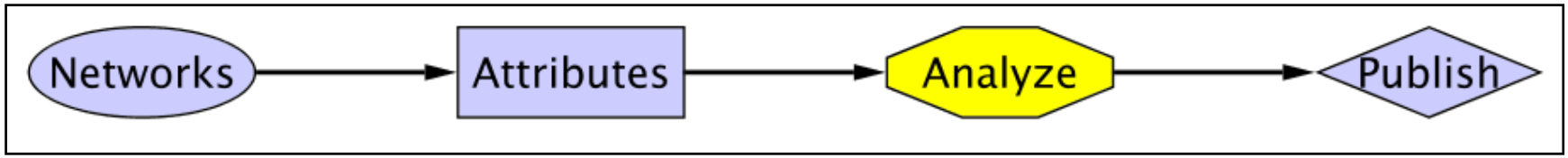
- Convenient access to gene list annotation




Cytoscape Desktop (New Session)
File Edit View Select Layout Plugins Help
Search:
Control Panel
Network VizMapper™ Editor Filters
Network Nodes Edges
Wnt_signaling_(NCI_/_ 76(0) 363(0)
Wnt_signaling_(NCI_
Data Panel
ID Affy HG U133-P
Node Attribute Browser
Welcome to Cytoscape 2.6.1
Right-click + drag to ZOOM
Middle-click + drag to PAN
Biomart Web Service Client
Query
biomart
Data Source
ENSEMBL 54 GENES (SANGER UK) - Homo sapiens genes (NCBI36)
Key Attribute
Attribute: ID
Data Type: EntrezGene ID(s)
Available attributes
5' UTR Start (5_utr_start)
Aedes Chromosome (aedes_chromosome)
Aedes Chromosome End (bp) (aedes_chrom_end)
Aedes Chromosome Start (bp) (aedes_chrom_start)
Aedes Ensembl Gene ID (aedes_ensembl_gene)
Aedes Ensembl Protein ID (aedes_homolog_ensembl_peptide)
Affy HC G110 (affy_hc_g110)
Affy HG FOCUS (affy_hg_focus)
Affy HG U133-PLUS-2 (affy_hg_u133_plus_2)
Affy HG U133A (affy_hg_u133a)
Affy HG U133A_2 (affy_hg_u133a_2)
Affy HG U133B (affy_hg_u133b)
Affy HG U95A (affy_hg_u95a)
Affy HG U95AV2 (affy_hg_u95av2)
Affy HG U95B (affy_hg_u95b)
Affy HG U95C (affy_hg_u95c)
Affy HG U95D (affy_hg_u95d)
Reset
Cancel
Import

Cytoscape Workflow

1. Load Networks (Get network data into Cytoscape)
2. Load Attributes (Get data about networks into Cytoscape)
3. Analyze and Visualize Networks
4. Prepare for Publication



Visual Data Integration

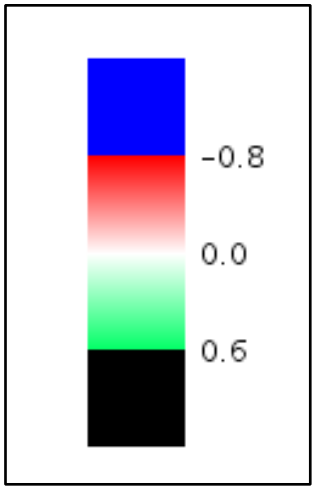
1. Network Data

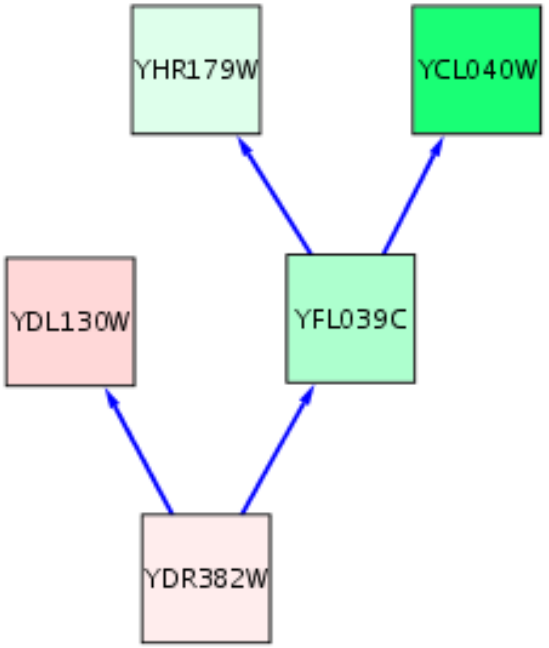
```
YDR382W pp YDL130W
YDR382W pp YFL039C
YFL039C pp YCL040W
YFL039C pp YHR179W
```

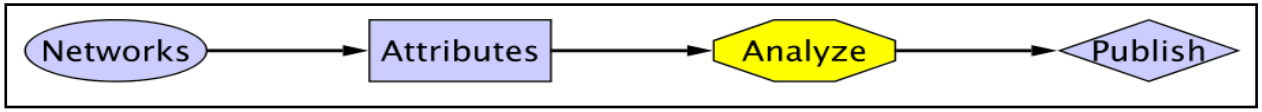
2. Attribute Data

```
ExpressionValue
YCL040W = 0.542
YDL130W = -0.123
YDR382W = -0.058
YFL039C = 0.192
YHR179W = 0.078
```

VizMapper

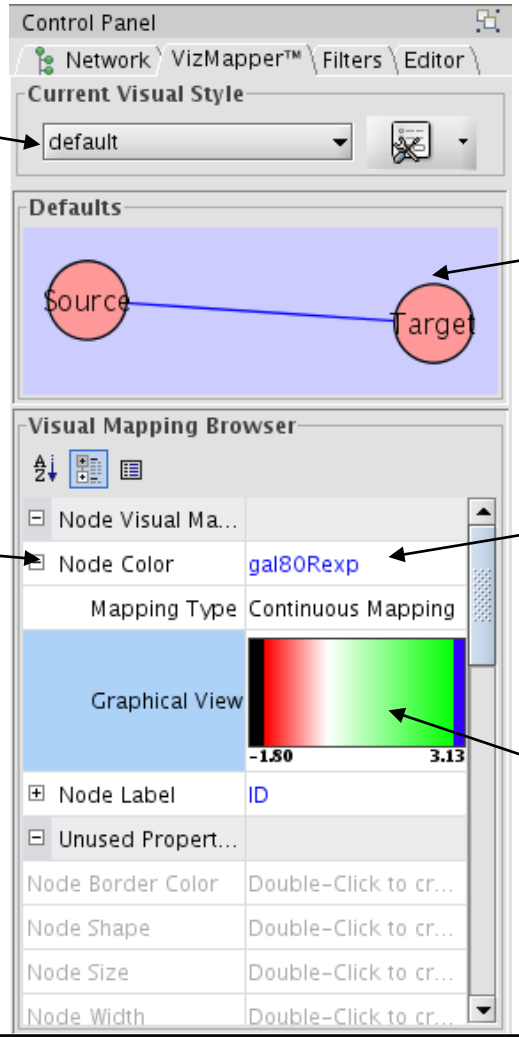






VizMapper

List of Visual Styles

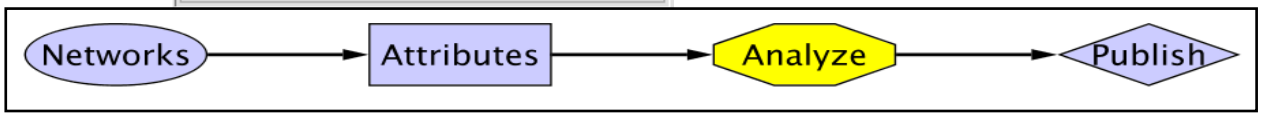


Default Visual Style Editor

List of Visual Attributes

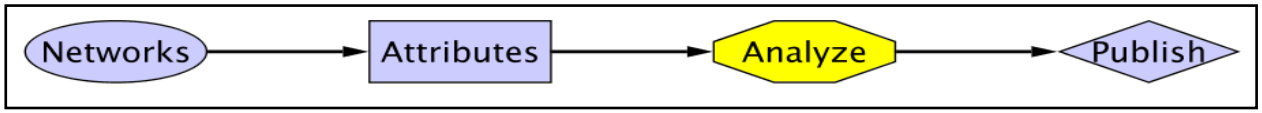
List of Data Attributes

Mapping definition

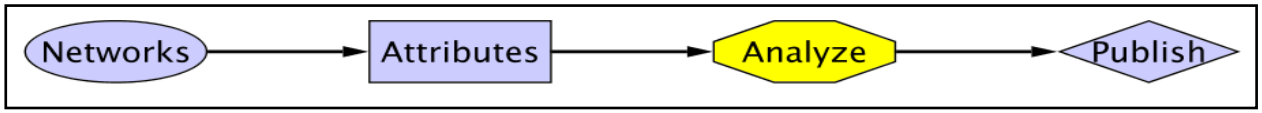


Types of mappings

- Continuous Data mapped to Continuous Visual Attributes (e.g. gene expression levels mapped to node color)
- Continuous Data mapped to Discrete Visual Attributes (e.g. p-value categories mapped to node shape)
- Discrete (categorical) Data to Discrete Visual Attributes (e.g. GO annotation mapped to node shape)
- Discrete Data mapped to Continuous Visual Attributes (e.g. multiple GO terms mapped to pie coloring)

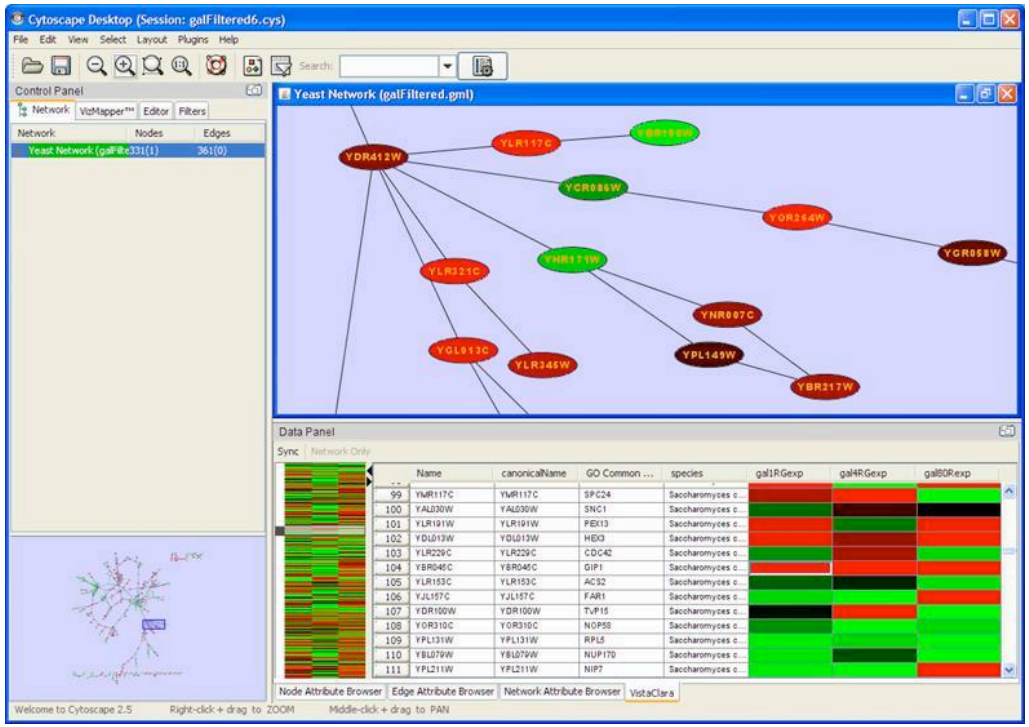


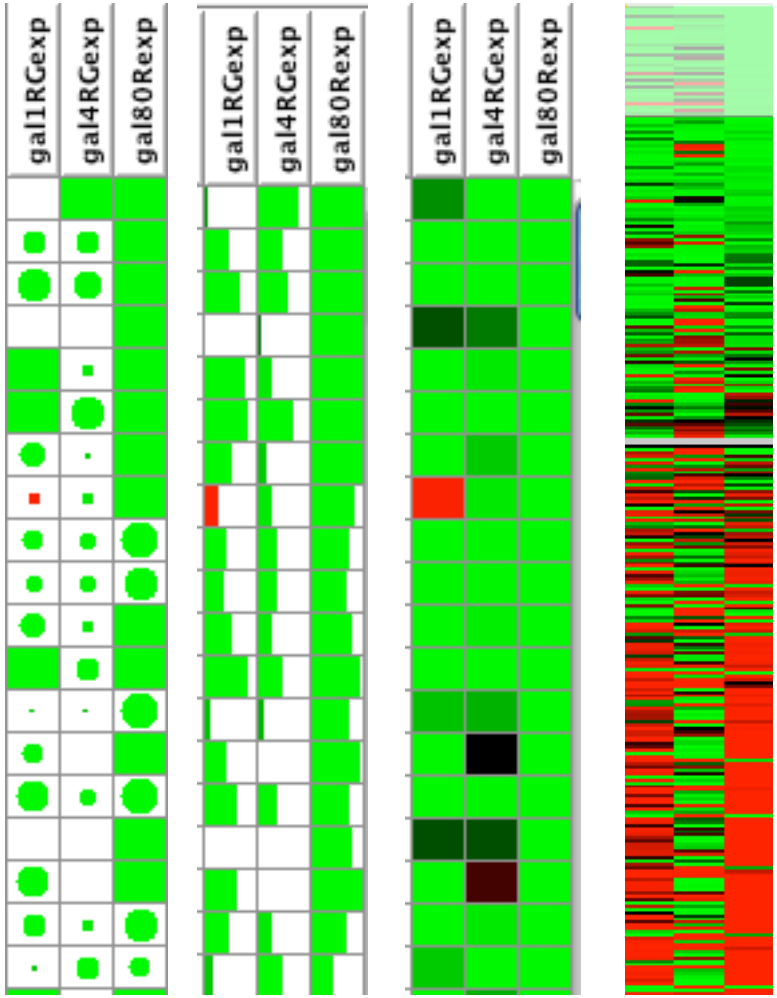
Demo Applying Vizmapping



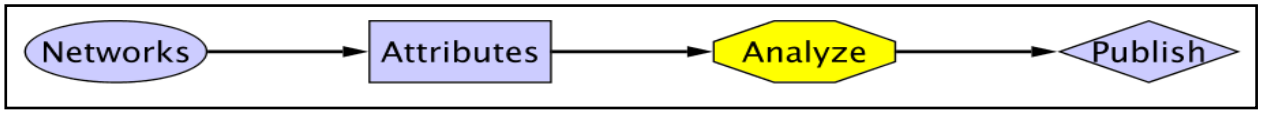
VistaClara

- Visualization for gene expression data
- Heat maps, sorting, animation

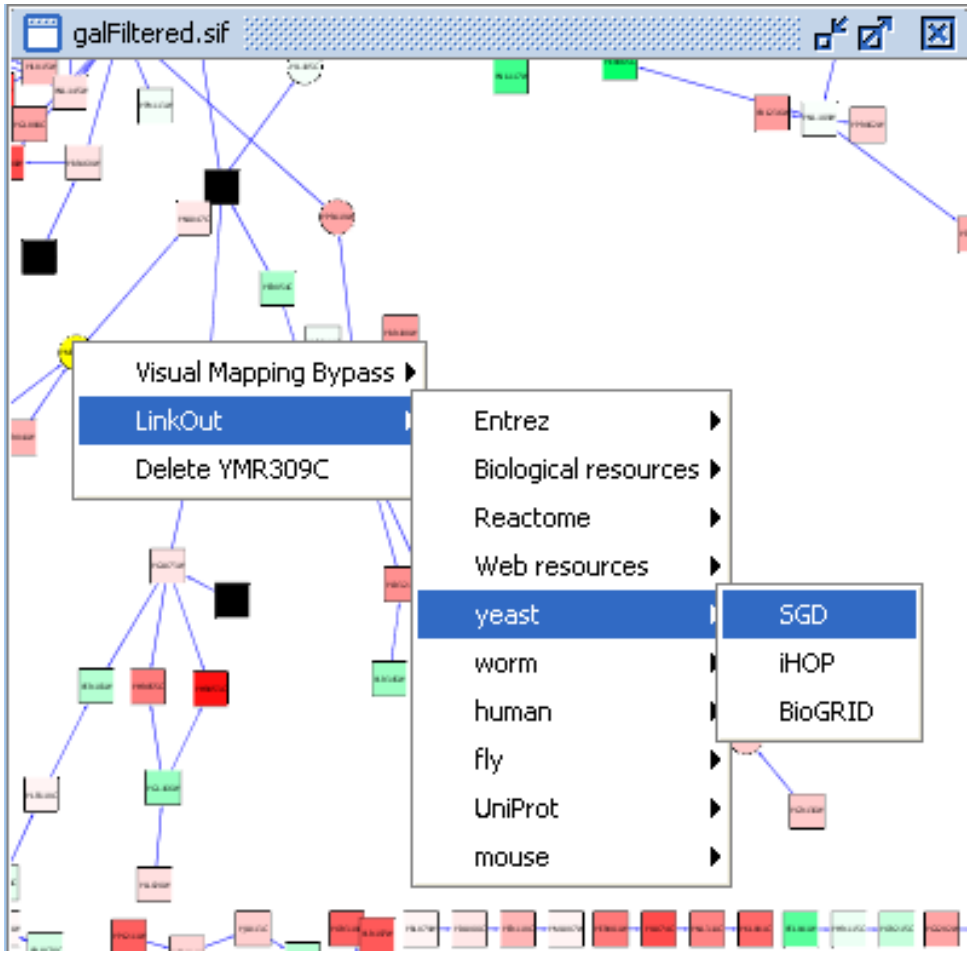




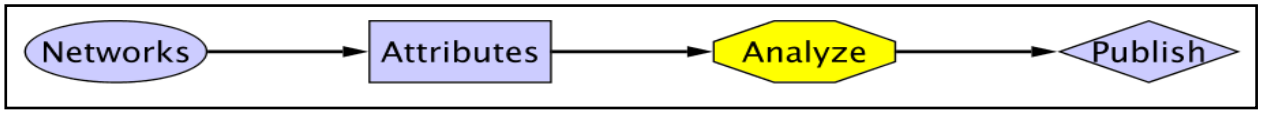
Demo network filtering and layout



Linkout

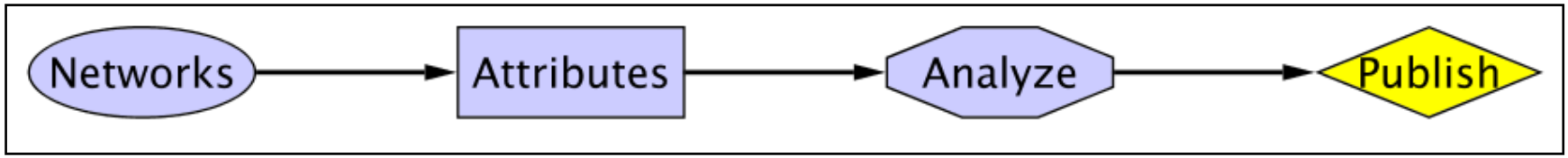


- Nodes and Edges act as hyperlinks to external databases.
- User-configurable URLs
- Collection of the biological results for the publication



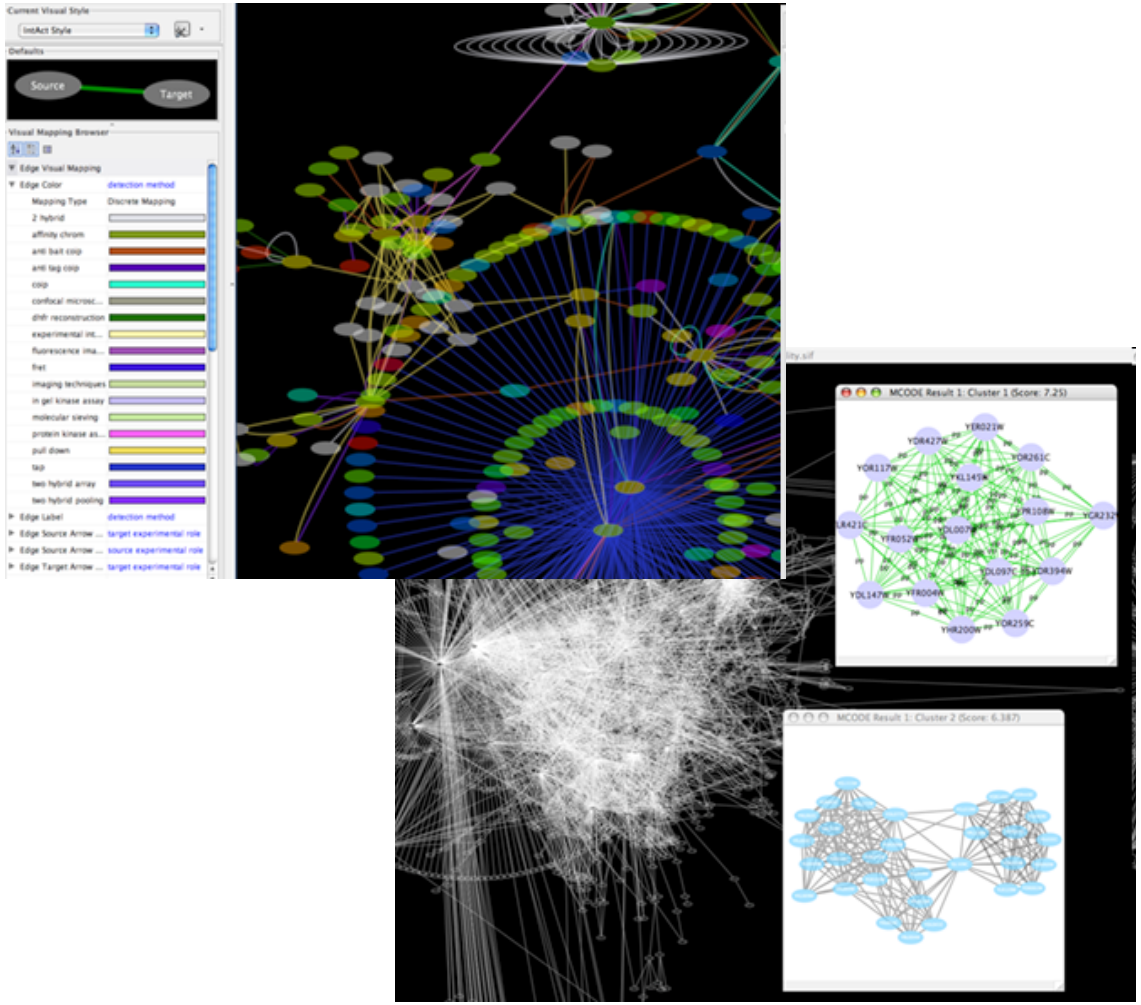
Cytoscape Workflow

1. Load Networks (Get network data into Cytoscape)
2. Load Attributes (Get data about networks into Cytoscape)
3. Analyze and Visualize Networks
4. Prepare for Publication



Publication quality figures

- Publication Quality Graphics in several formats
  - ➢ PDF, EPS, SVG, PNG, JPEG, and BMP
- Export Session to HTML for Web



Tips and Tricks

Gary Bader

Tips & Tricks

- "Root graph"
  - "There is one graph to rule them all…."
  - The networks in Cytoscape are all "views" on a single graph.
  - Changing the attribute for a node in one network *will* also change that attribute for a node with the same ID in all other loaded networks
  - There is no way to "copy" a node and keep the same ID
  - Make a copy of the session

Tips & Tricks

- Network views
  - When you open a large network, you will not get a view by default
  - To improve interactive performance, Cytoscape has the concept of "Levels of Detail"
    - Some visual attributes will only be apparent when you zoom in
    - The level of detail for various attributes can be changed in the preferences
    - To see what things will look like at full detail:
      - View→Show Graphics Details

Tips & Tricks

- Sessions
  - Sessions save pretty much everything:
    - Networks
    - Properties
    - Visual styles
    - Screen sizes
  - Saving a session on a large screen may require some resizing when opened on your laptop

Tips & Tricks

- Logging
  - By default, Cytoscape writes it's logs to the Error Dialog: Help→Error Dialog
  - Can change a preference to write it to the console
    - Edit→Preferences→Properties…
    - Set logger.console to true
    - Don't forget to save your preferences
    - Restart Cytoscape
  - (can also turn on debugging: cytoscape.debug, but I don't recommend it)

Tips & Tricks

- Memory
  - Cytoscape uses lots of it
  - Doesn't like to let go of it
  - An occasional restart when working with large networks is a good thing
  - Destroy views when you don't need them
  - Java doesn't give us a good way to get the memory right at start time
    - Cytoscape 2.7 does a much better job at "guessing" good default memory sizes than previous versions

Tips & Tricks

- .cytoscape directory
  - Your defaults and any plugins downloaded from the plugin manager will go here
  - Sometimes, if things get really messed up, deleting (or renaming) this directory can give you a "clean slate"
- Plugin manager
  - "Outdated" doesn't necessarily mean "won't work"
  - Plugin authors don't always update their plugins immediately after new releases
  - Click on "Show outdated plugins" to see the entire list of plugins.

Lab Time

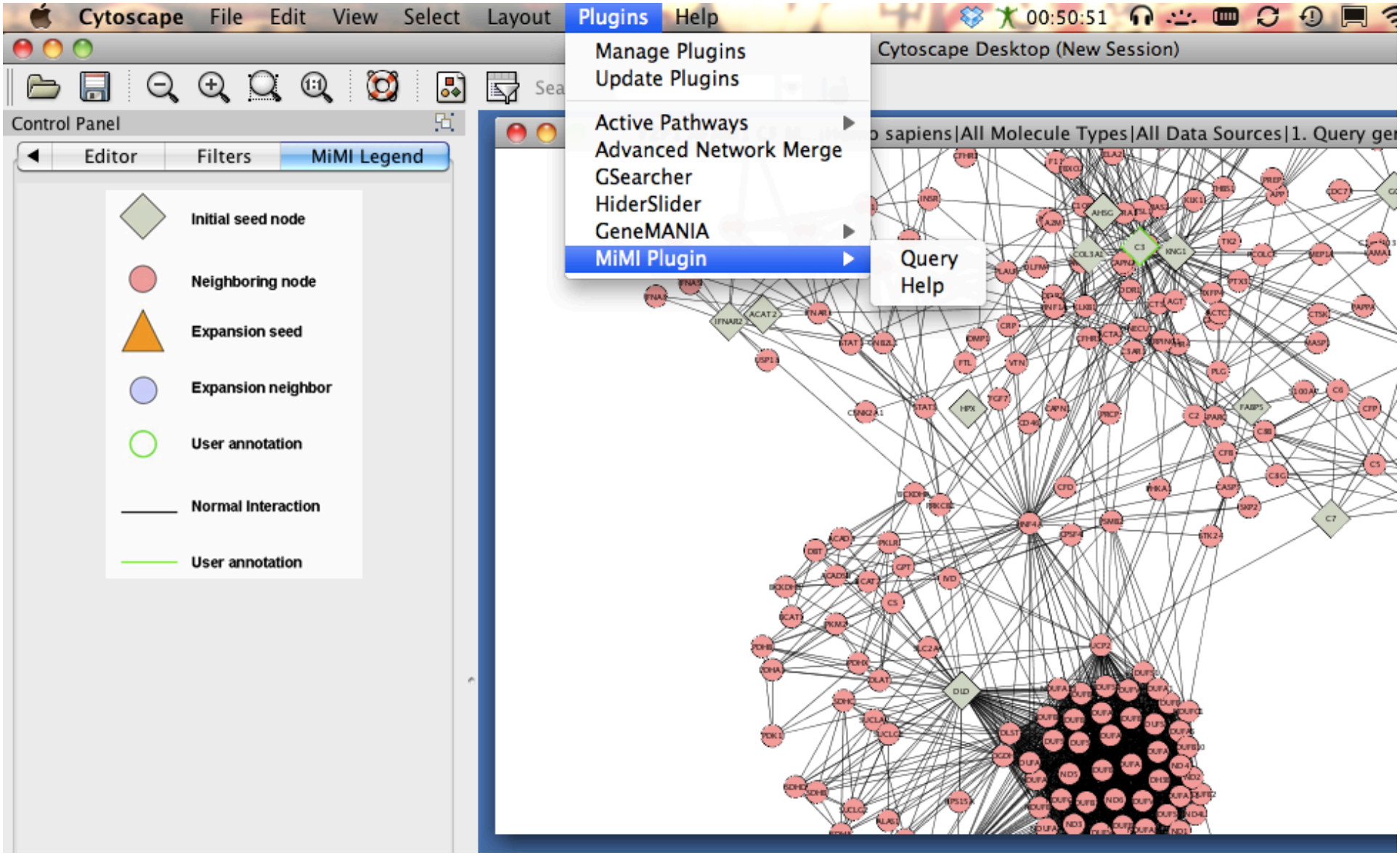
- Try out workflow
  - Agilent Literature Search, VistaClara
  - ID mapping services
  - Use your own data or sample data that comes with Cytoscape

- Resource:
  - http://opentutorials.rbvi.ucsf.edu/index.php/Tutorial:Introduction_to_Cytoscape
- Timing: 15:15-16:15

Gene List to Network Lab

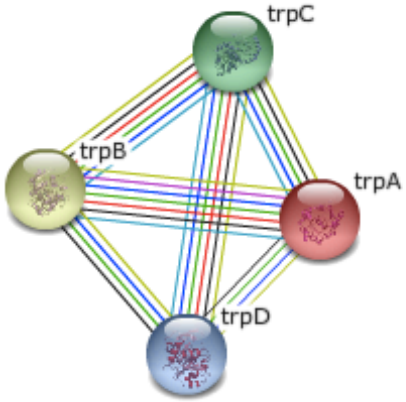
- Start with a gene list and find a network
  - MIMI – Protein-protein interactions (PPI)
  - STRING, GeneMANIA – Functional interactions
  - AgilentLitSearch – text mined interactions
  - BisoGeNet – another PPI source
- Gene function prediction with STRING and GeneMANIA

MiMI: Protein interactions



http://string.embl.de

STRING 8.3



This is the **evidence view**. Different line colors represent the types of evidence for the association.

confidence evidence actions interactive advanced more less save

(requires Flash player 10 or better)

Your Input:

- trpA Tryptophan synthase, alpha subunit; The alpha subunit is responsible for the aldol cleavage of indoleglycerol phosphate to indole and glyceraldehyde 3- phosphate (268 aa)
- trpB Tryptophan synthase beta subunit (Tryptophan synthase subunit B); The beta subunit is responsible for the synthesis of L- tryptophan from indole and L-serine (397 aa)
- trpC N-(5-phosphoribosyl)anthranilate isomerase and indole-3-glycerolphosphate synthetase; Bifunctional enzyme that catalyzes two sequential steps of tryptophan biosynthetic pathway. The first reaction is catalyzed by the isomerase, coded by the trpF domain; the second reaction is catalyzed by the synthase, coded by the trpC domain (452 aa)
- trpD fused glutamine amidotransferase (component II) of anthranilate synthase/anthranilate phosphoribosyl transferase (531 aa)

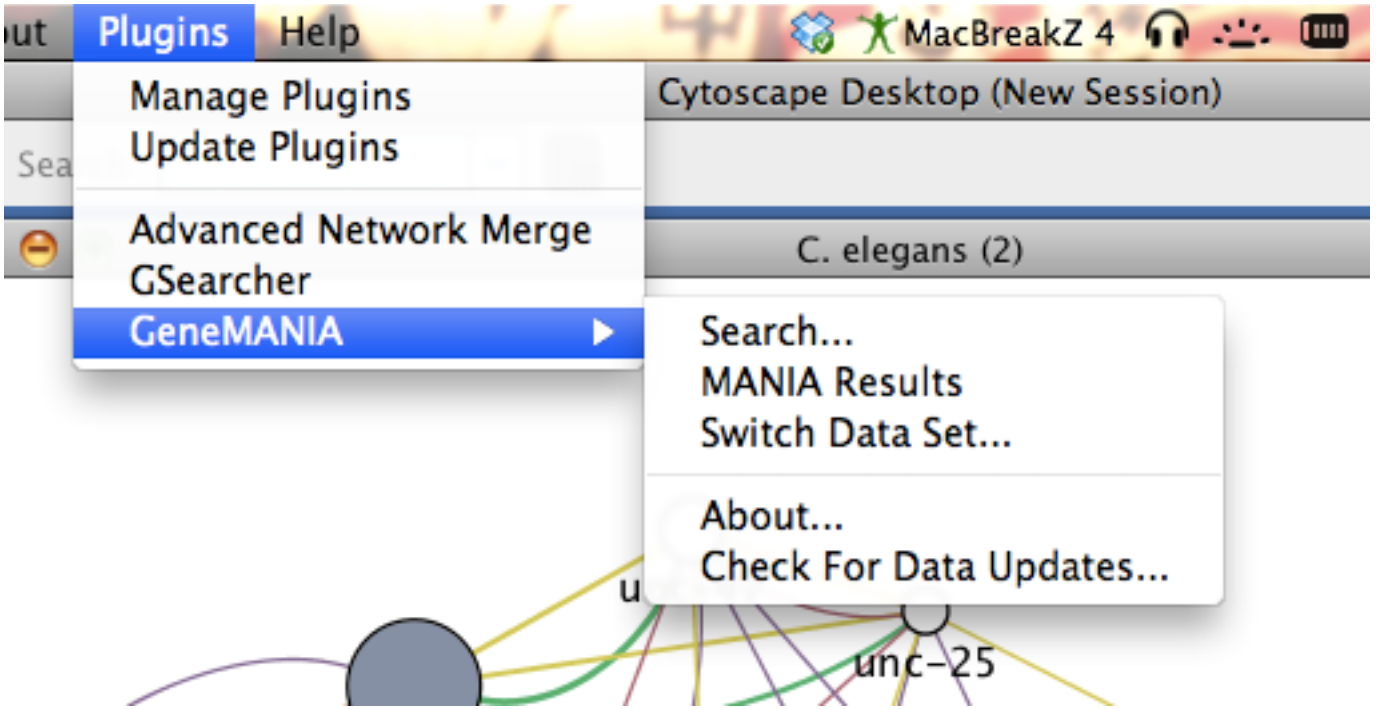
(Escherichia coli K12)

Predicted Functional Partners:

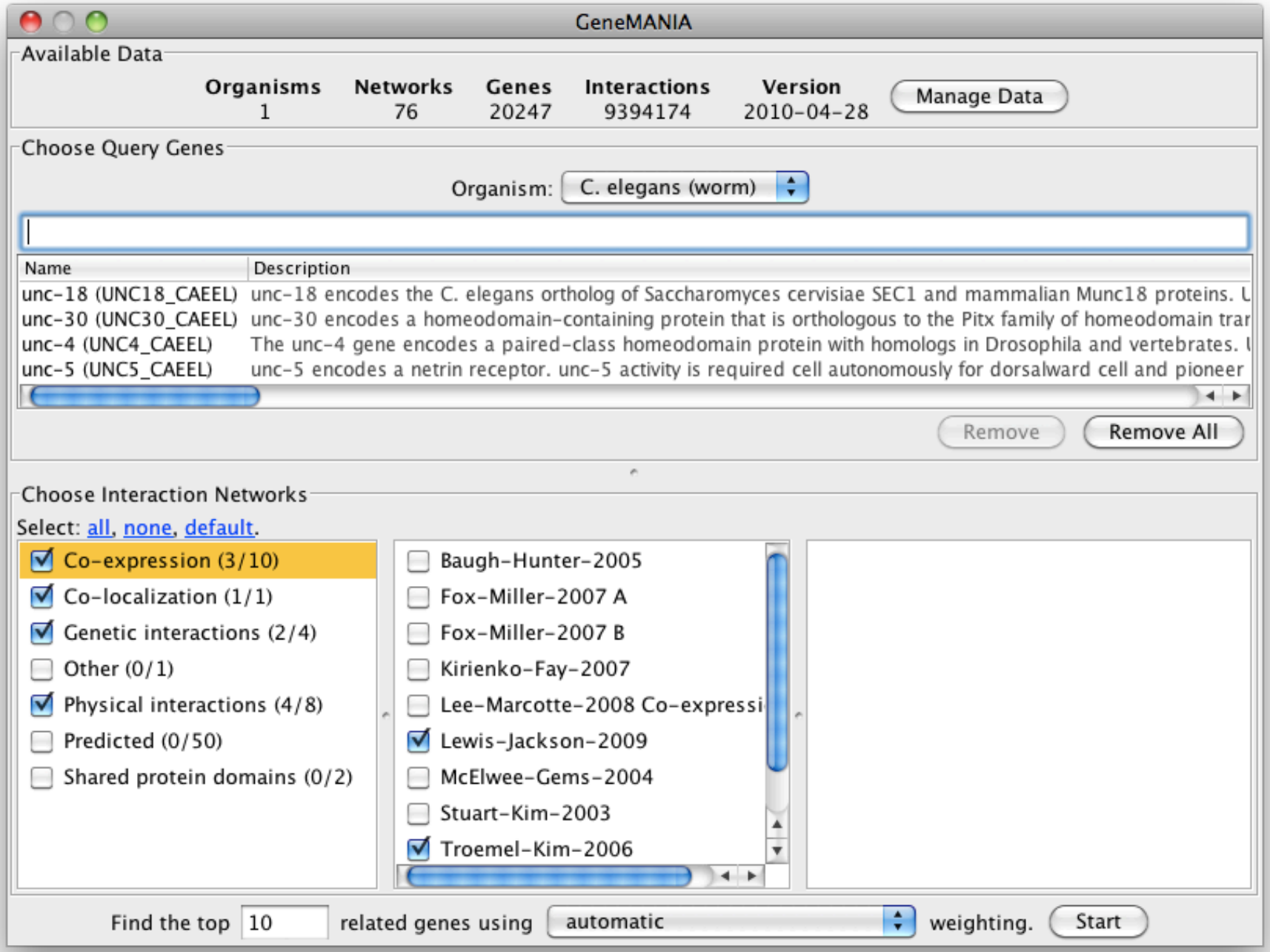
Neighborhood
Gene Fusion
Cooccurrence
Coexpression
Experiments
Databases
Textmining
[Homology]
Score

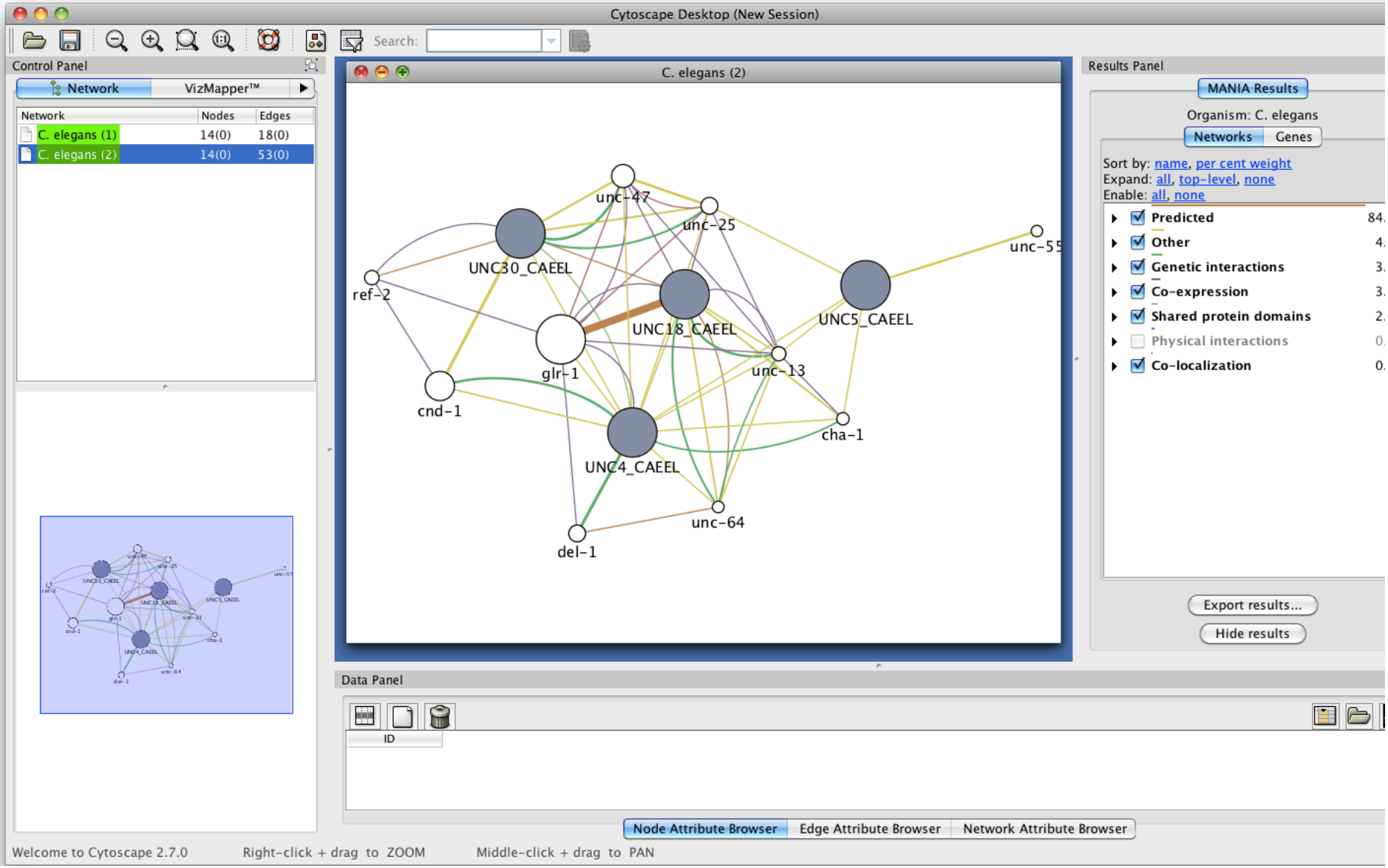
http://string.embl.de

tabdelimited.zSVr2AGatnE_.txt	Text Summary (TXT - simple tab delimited flatfile)
xml_summary.zSVr2AGatnE_.xml	XML Summary (PSI - Proteomics Standards Iniative)
network_medusa.zSVr2AGatnE_.dat	Graph Layout (Data for the 'Medusa' Network Viewer)
protein_sequences.zSVr2AGatnE_.fa	Network Proteins / Amino Acid Sequences (Multi-Sequence File; FASTA format)
proteins_desc.zSVr2AGatnE_.txt	Network proteins description (TXT - simple tab delimited flatfile)

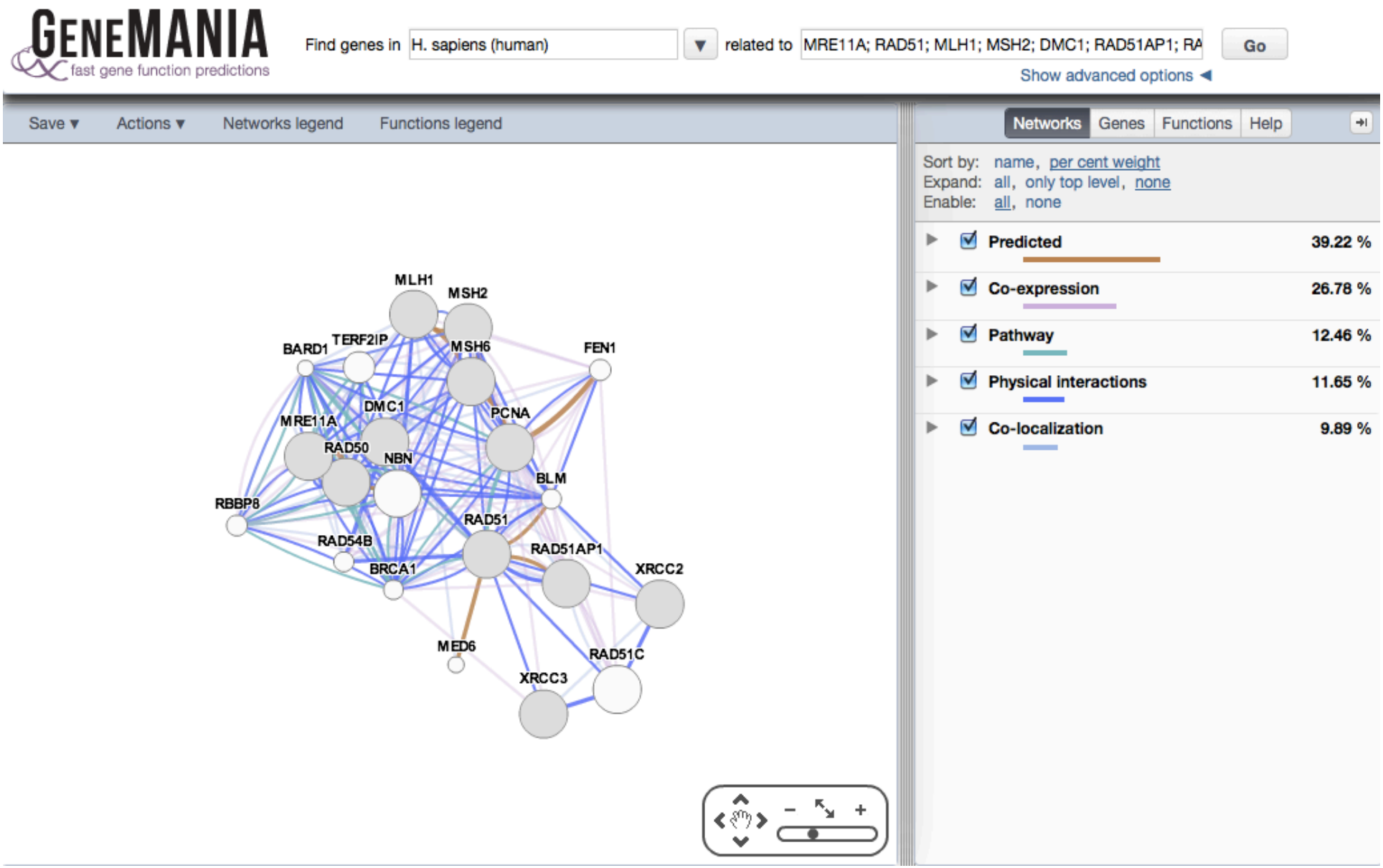


http://www.genemania.org/plugin/


GeneMANIA
Available Data
Organisms
1
Networks
76
Genes
20247
Interactions
9394174
Version
2010-04-28
Manage Data
Choose Query Genes
Organism:
C. elegans (worm)
Name
Description
unc-18 (UNC18_CAEEL)
unc-18 encodes the C. elegans ortholog of Saccharomyces cervisiae SEC1 and mammalian Munc18 proteins. U
unc-30 (UNC30_CAEEL)
unc-30 encodes a homeodomain-containing protein that is orthologous to the Pitx family of homeodomain trar
unc-4 (UNC4_CAEEL)
The unc-4 gene encodes a paired-class homeodomain protein with homologs in Drosophila and vertebrates. U
unc-5 (UNC5_CAEEL)
unc-5 encodes a netrin receptor. unc-5 activity is required cell autonomously for dorsalward cell and pioneer
Remove
Remove All
Choose Interaction Networks
Select: all, none, default.
Co-expression (3/10)
Co-localization (1/1)
Genetic interactions (2/4)
Other (0/1)
Physical interactions (4/8)
Predicted (0/50)
Shared protein domains (0/2)
Baugh-Hunter-2005
Fox-Miller-2007 A
Fox-Miller-2007 B
Kirienko-Fay-2007
Lee-Marcotte-2008 Co-expressi
Lewis-Jackson-2009
McElwee-Gems-2004
Stuart-Kim-2003
Troemel-Kim-2006
Find the top
10
related genes using
automatic
weighting.
Start


Cytoscape Desktop (New Session)
Search:
Control Panel
Network
VizMapper™
Network
Nodes
Edges
C. elegans (1)
14(0)
18(0)
C. elegans (2)
14(0)
53(0)
C. elegans (2)
unc-47
unc-25
UNC30_CAEEL
ref-2
UNC18_CAEEL
UNC5_CAEEL
unc-55
glr-1
unc-13
cnd-1
cha-1
UNC4_CAEEL
unc-64
del-1
Results Panel
MANIA Results
Organism: C. elegans
Networks
Genes
Sort by: name, per cent weight
Expand: all, top-level, none
Enable: all, none
Predicted
Other
Genetic interactions
Co-expression
Shared protein domains
Physical interactions
Co-localization
Export results...
Hide results
Data Panel
ID
Node Attribute Browser
Edge Attribute Browser
Network Attribute Browser
Welcome to Cytoscape 2.7.0
Right-click + drag to ZOOM
Middle-click + drag to PAN

http://www.genemania.org



Gene List to Network Lab

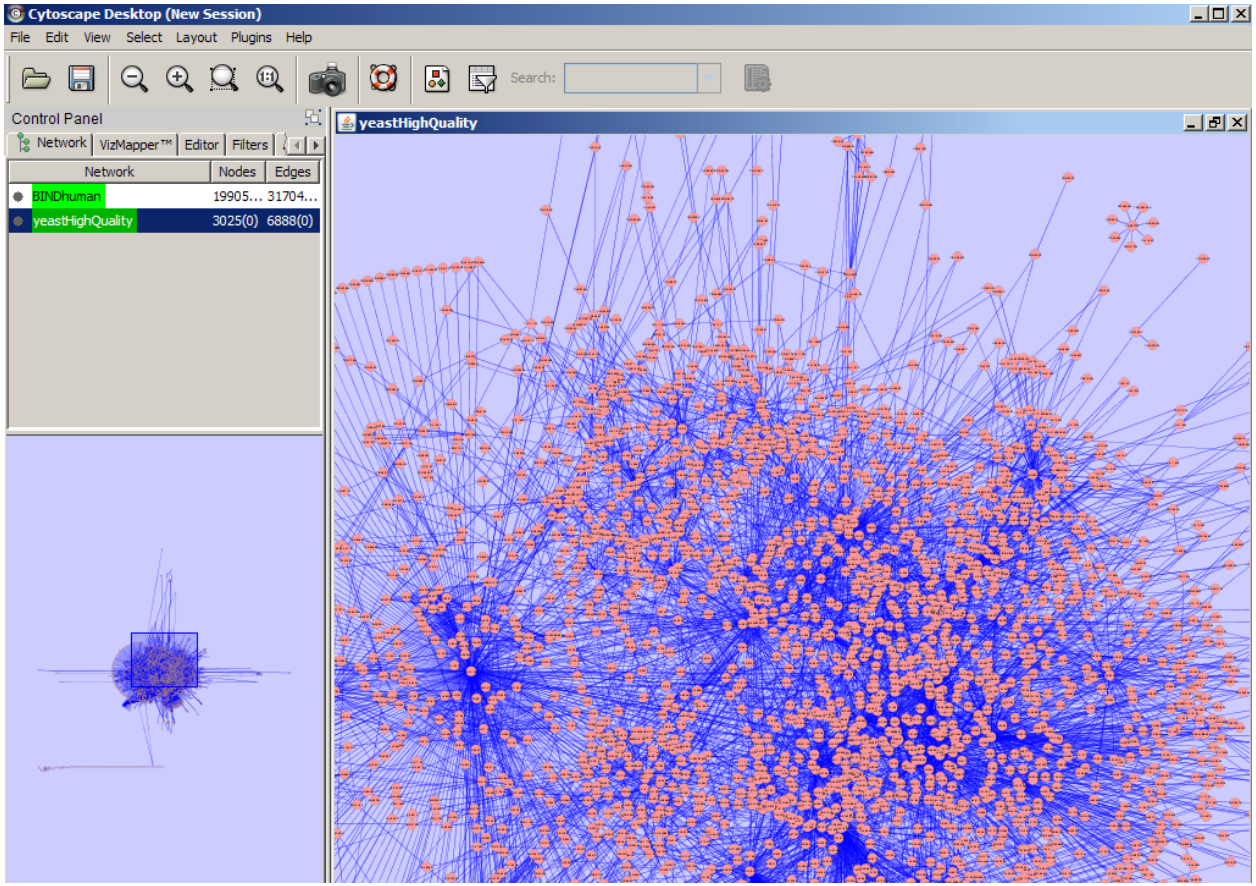
- Enter gene list into STRING and GeneMANIA websites
- Save results as text and load into Cytoscape

- Try MIMI plugin

- Gene function prediction: input a list of genes known to be in a given function. Ask STRING or GeneMANIA to find more genes like those (guilt by association)

PPI network analysis Lab

- Load protein-protein interaction network (e.g. YeastHighQuality: 7000 Yeast interactions among 3000 proteins)
- Visualize



PPI network analysis

- Load protein-protein interaction network (eg YeastHighQuality)
- Visualize
  - Large dataset; hairball
  - Layouts don't help

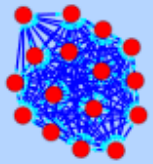
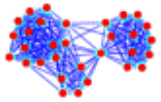
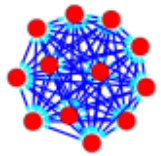
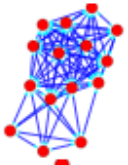
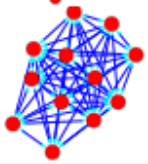
- Cluster - MCODE, ClusterMaker, ActiveModules

Analysis
Find Network Clusters - MCODE Plugin

- Network clusters are highly interconnected sub-networks that may be also partly overlapping
- Clusters in a protein-protein interaction network have been shown to represent protein complexes and parts of biological pathways
- Clusters in a protein similarity network represent protein families
- Network clustering is available through the MCODE Cytoscape plugin

MCODE plug-in demo

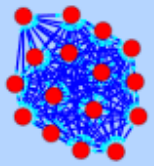
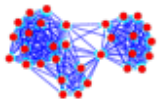
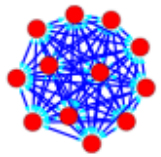
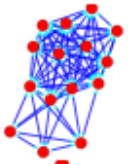
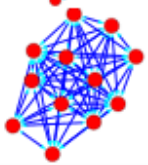
MCODE Results Summary

Rank	Score	Size	Names	Complex
1	7.25	16,116	YGR232W, YDL007W, YKL145W, YFR052W, YFR004W, YLR421C, YOR261C, YDL147W, YDR427W, YHR200W, YER021W, YOR117W, YDL097C, YOR259C, YPR108W, YDR394W	
2	6.387	31,198	YPL093W, YBL004W, YOR272W, YNL110C, YKL009W, YFL002C, YOL077C, YPL126W, YIL035C, YLR409C, YLR129W, YOR061W, YKR060W, YCR057C, YDR449C, YOR039W, YJL109C, YPL012W, YGR103W, YLR449W, YOR206W, YKL014C, YLL008W, YKL172W, YNL002C, YLR002C, YGL111W, YOL041C, YGL019W, YOR145C, YPR016C	
3	5.417	12,65	YGL011C, YOL038W, YPR103W, YMR314W, YBL041W, YOR362C, YER012W, YJL001W, YML092C, YGR253C, YER094C, YGR135W	
4	5	15,75	YPL043W, YMR290C, YER006W, YKR081C, YDR496C, YDL031W, YNL061W, YNL132W, YLR222C, YLR197W, YMR049C, YHR052W, YJL069C, YKL099C, YDL014W	
5	5	12,60	YPR187W, YPR010C, YPR110C, YNL248C, YOR341W, YNR003C, YKL144C, YOR207C, YPR190C, YNL113W, YOR116C, YBR154C	

Create a new child network. Save Done

Bader & Hogue, BMC Bioinformatics 2003 4(1):2

MCODE Results Summary

Rank	Score	Size	Names	Complex
1 **Proteasome 26S**	7.25	16,116	YGR232W, YDL007W, YKL145W, YFR052W, YFR004W, YLR421C, YOR261C, YDL147W, YDR427W, YHR200W, YER021W, YOR117W, YDL097C, YOR259C, YPR108W, YDR394W	
2 **Ribosome**	6.387	31,198	YPL093W, YBL004W, YOR272W, YNL110C, YKL009W, YFL002C, YOL077C, YPL126W, YIL035C, YLR409C, YLR129W, YOR061W, YKR060W, YCR057C, YDR449C, YOR039W, YJL109C, YPL012W, YGR103W, YLR449W, YOR206W, YKL014C, YLL008W, YKL172W, YNL002C, YLR002C, YGL111W, YOL041C, YGL019W, YOR145C, YPR016C	
3 **Proteasome 20S**	5.417	12,65	YGL011C, YOL038W, YPR103W, YMR314W, YBL041W, YOR362C, YER012W, YJL001W, YML092C, YGR253C, YER094C, YGR135W	
4 **RNA Splicing**	5	15,75	YPL043W, YMR290C, YER006W, YKR081C, YDR496C, YDL031W, YNL061W, YNL132W, YLR222C, YLR197W, YMR049C, YHR052W, YJL069C, YKL099C, YDL014W	
5 **RNA Pol core**	5	12,60	YPR187W, YPR010C, YPR110C, YNL248C, YOR341W, YNR003C, YKL144C, YOR207C, YPR190C, YNL113W, YOR116C, YBR154C	

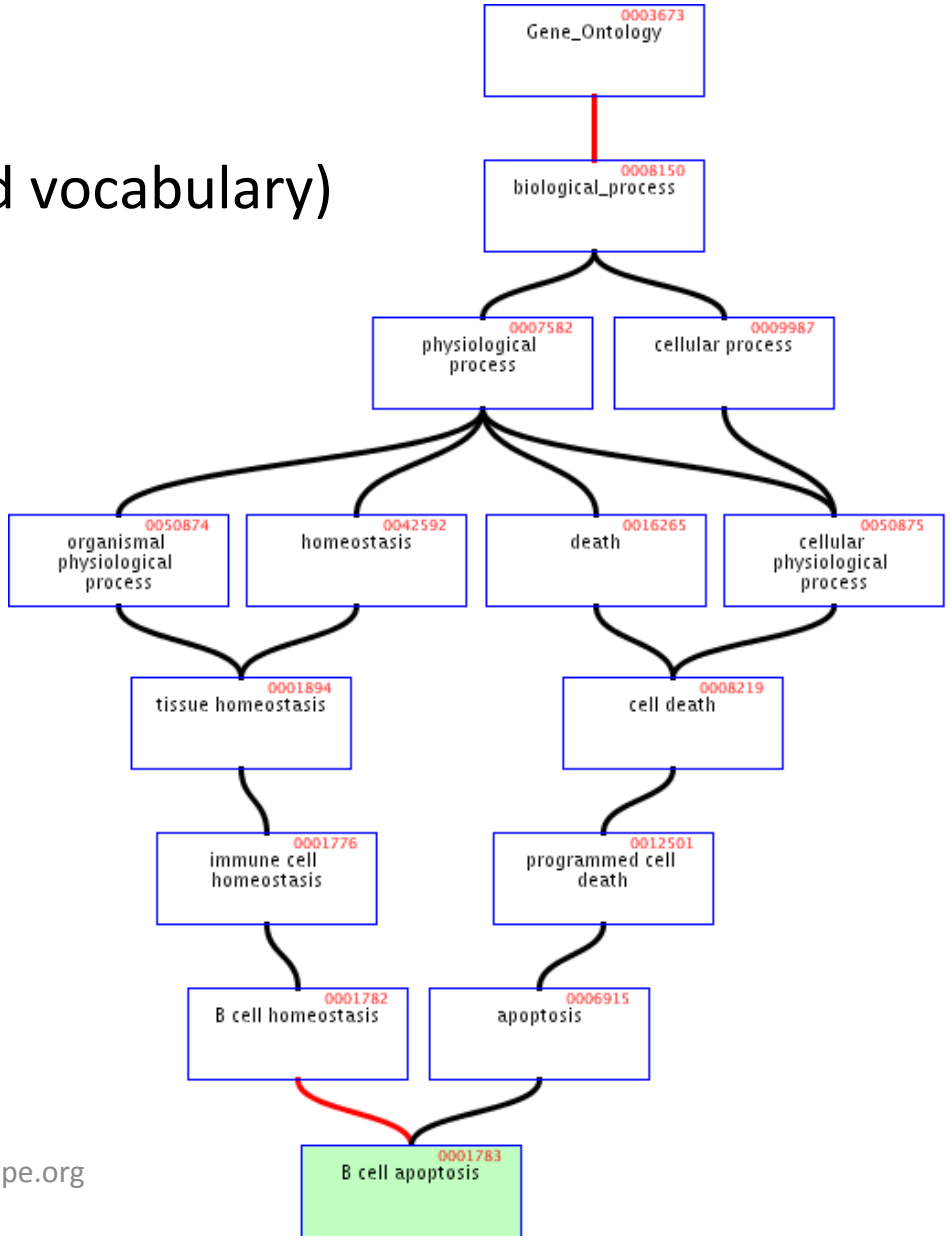
Create a new child network. Save Done

Gene Ontology analysis Lab

- Describes gene function

1. Agreed upon terms (controlled vocabulary)
   - Biological process
   - Cellular component
   - Molecular function
2. Genome annotation

www.geneontology.org



cytoscape.org

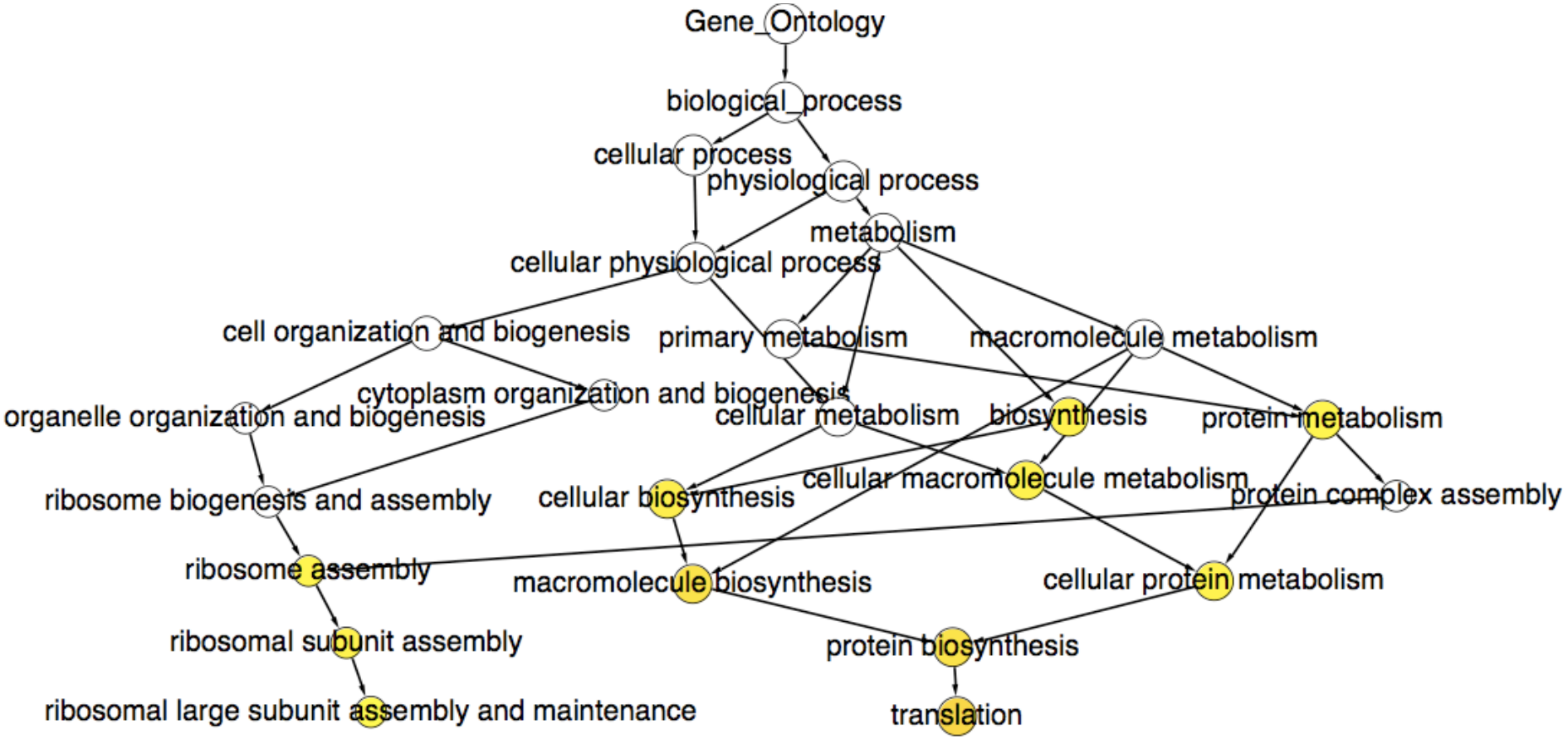
BinGO plugin

- Calculates over-representation of a subset of genes with respect to a background set in a specific GO category
- Input: subnetwork, or list
  - Background set by user
- Output: tree with nodes color reflecting overrepresentation; also as lists

- Caveats: Gene identifiers must match; low GO term coverage, GO bias, Background determining

BiNGO

Hypergeometric p-value
Multiple testing correction
(Benjamini-Hochberg FDR)



cytoscape.org

Maere, S., Heymans, K. and Kuiper, M
Bioinformatics 21, 3448-3449, 2005

Lab Time

- Try out MCODE with your dataset of choice
- Or use one of the other sets available
- If you find something interesting please share! ☺

- Try BINGO with subsets from e.g. MCODE

Analysis Lab
Find Active Subnetworks Lab

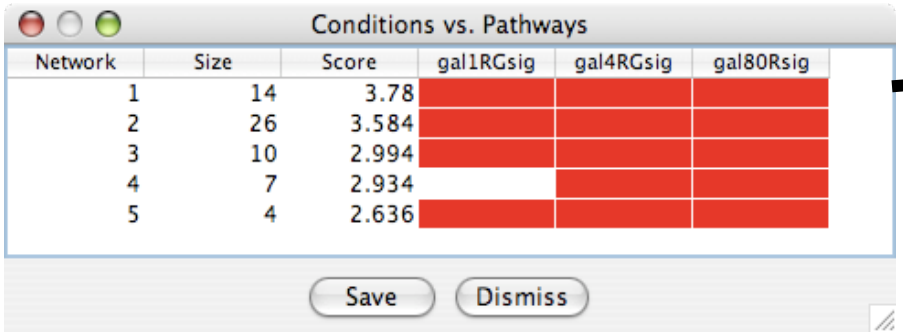
- Active modules are sub-networks that show differential expression over user-specified conditions or time-points
  - Microarray gene-expression attributes
  - Mass-spectrometry protein abundance
- Method
  - Calculate z-score/node, ZA score/subgraph, correct for random expression data sampling
  - Score over multiple experimental conditions
  - Simulated annealing-based search method is used to find the high scoring networks

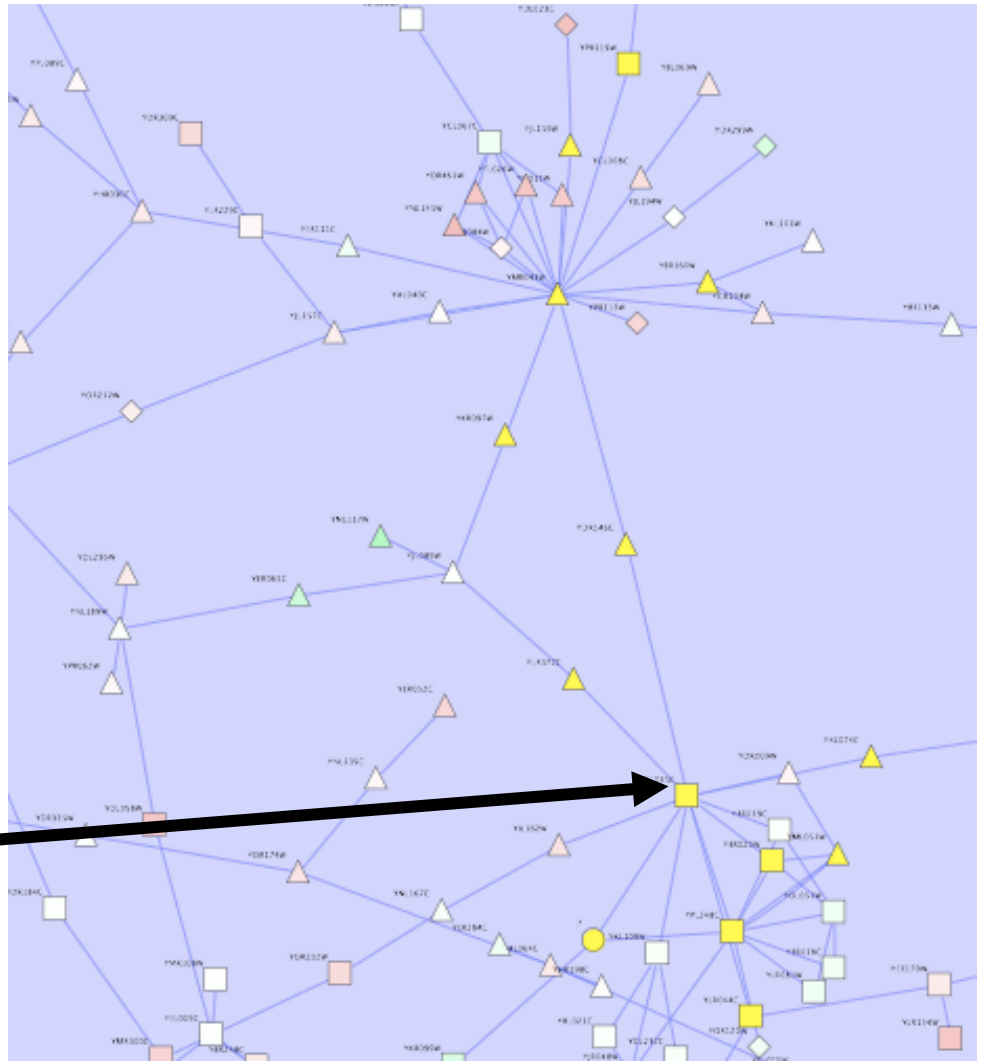
Analysis Lab
Find Active Subnetworks Lab

jActiveModules plug-in

Input: interaction network and p-values for gene expression values over several conditions

Output: significant sub-networks that show differential expression over one or several conditions





cytoscape.org

Ideker T et al. Science 2001; Bioinformatics 2002

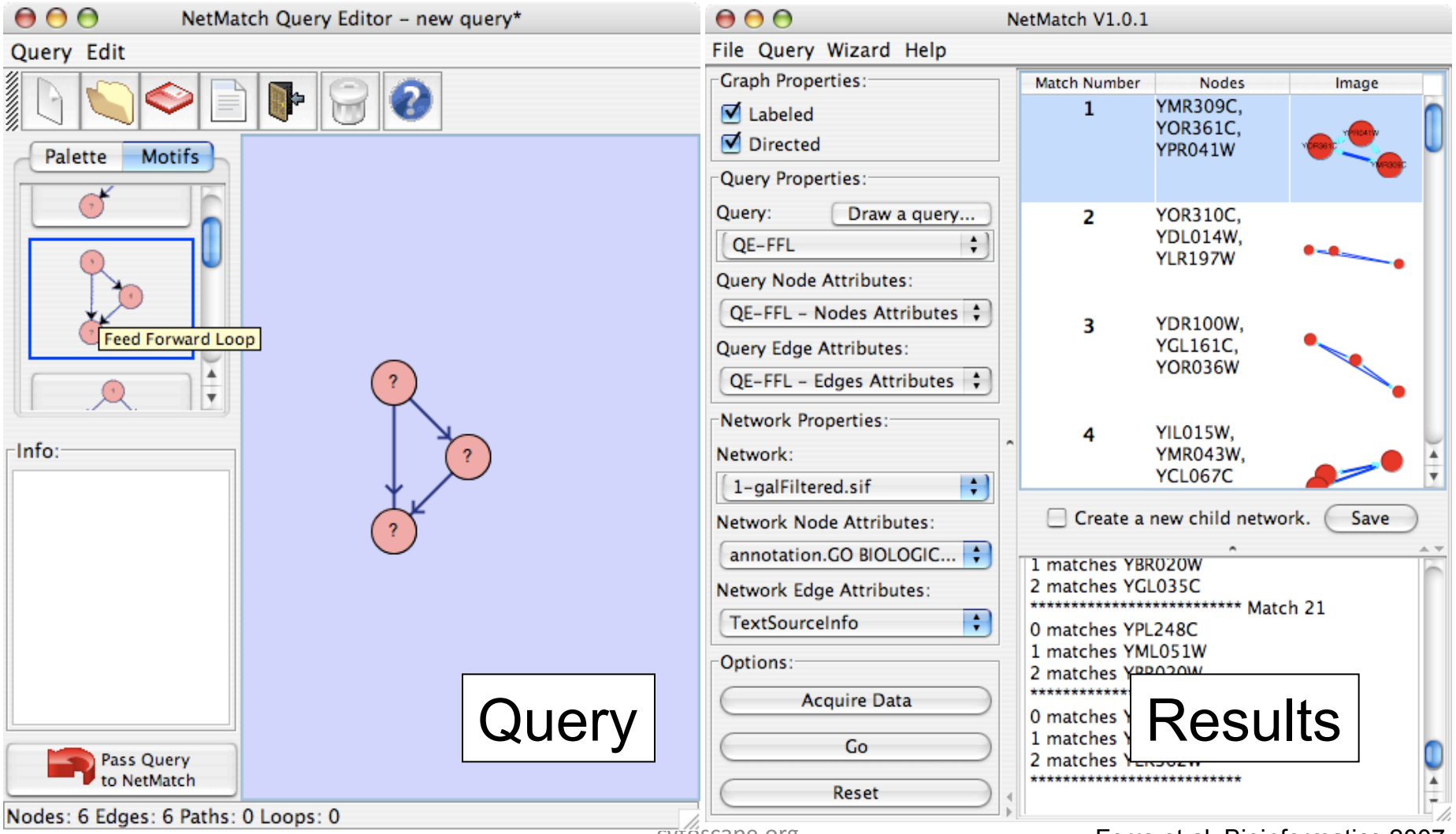
Lab Time

- Try out jActiveModules
- Use the gal expression dataset

Analysis Lab
Find Network Motifs - Netmatch plugin

- Network motif is a sub-network that occurs significantly more often than by chance alone
- Input: query and target networks, optional node/edge labels
- Output: topological query matches as subgraphs of target network
- Supports: subgraph matching, node/edge labels, label wildcards, approximate paths
- http://alpha.dmi.unict.it/~ctnyu/netmatch.html

Finding specific biological relevant TF-PPI sub-networks



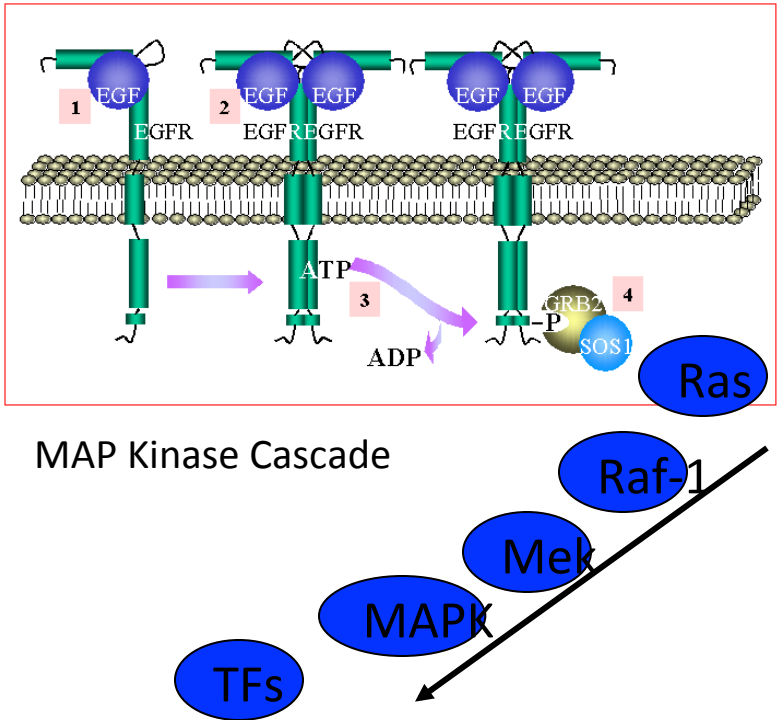
cytoscape.org

Ferro et al. Bioinformatics 2007

Find Signaling Pathways

- Potential signaling pathways from plasma membrane to nucleus via cytoplasm

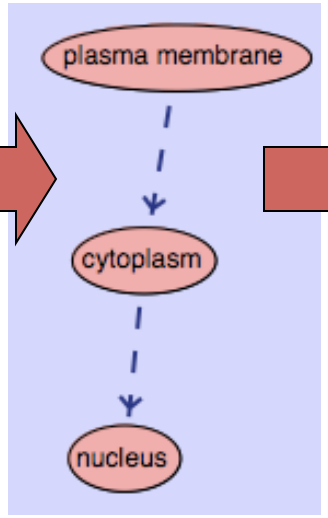
Signaling pathway example



MAP Kinase Cascade

Nucleus - Growth Control
Mitogenesis

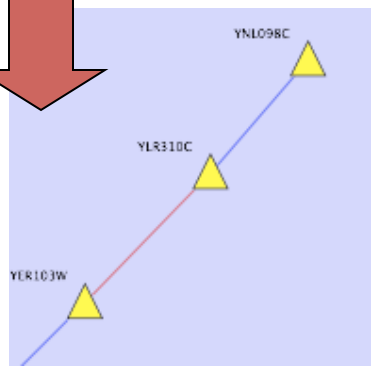
NetMatch query



Shortest path between subgraph matches

NetMatch Results

Match Number	Nodes	Image
	YGL008C	
4	YJL157C, YMR043W, YLR229C	
5	YJL157C, YAL040C, YLR229C	
6	YLR310C, YER103W, YNL098C	

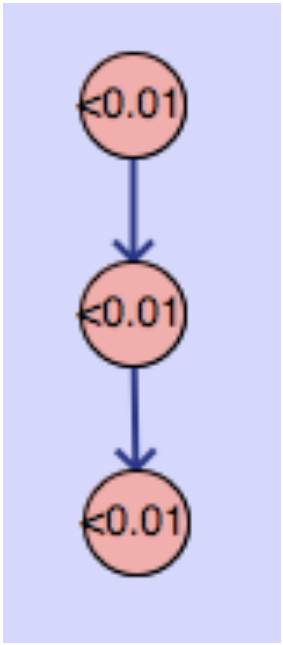


cytoscape.org

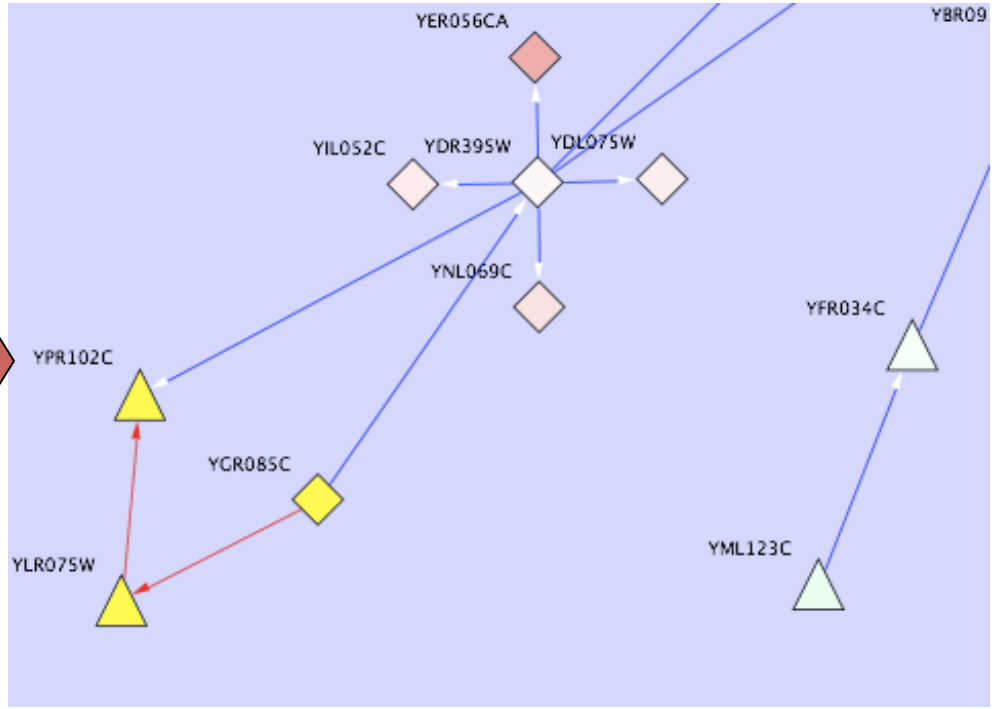
Find Expressed Motifs

Find specific subgraphs where certain nodes are significantly differentially expressed

NetMatch query



NetMatch Results



Protein	Differential Expression Significance
YLR075W	1.7255E-4
YGR085C	2.639E-4
YPR102C	3.7183E-4

Lab Time
Find motifs with Netmatch

- Use the provided dataset (not yet in sampleData)

GeneMANIA Lab