bioinformatics.ca

Canadian Bioinformatics Workshops

www.bioinformatics.ca

creative commons

Attribution-Share Alike 2.5 Canada

You are free:

to Share — to copy, distribute and transmit the work

to Remix — to adapt the work

Free Cultural APPROVED FOR Works

Under the following conditions:

Attribution. You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).

Share Alike. If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar licence to this one.

- For any reuse or distribution, you must make clear to others the licence terms of this work.
- Any of the above conditions can be waived if you get permission from the copyright holder.
- The author's moral rights are retained in this licence.

Module 7 part 2
FROM GENE LISTS TO PATHWAYS

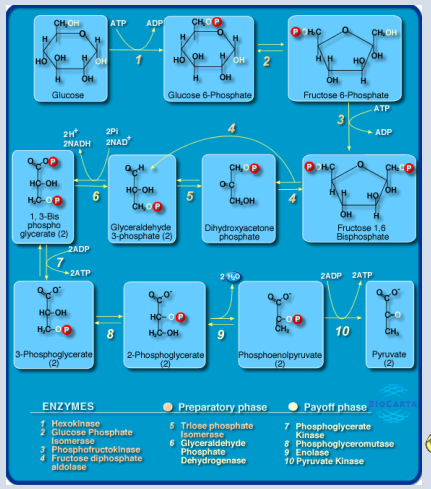
Veronique Voisin

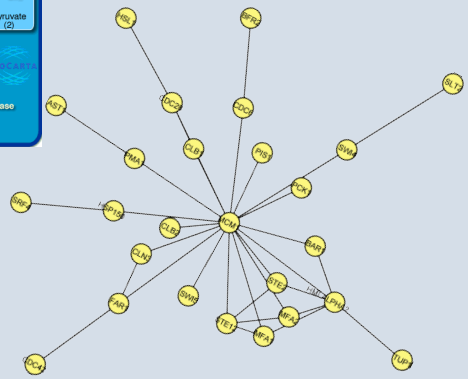
Bioinformatics for Cancer Genomics

May 27-31, 2013

bioinformatics.ca

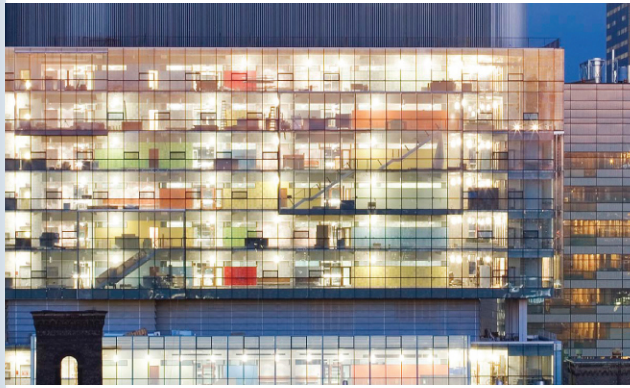
GNAQ
GNAS
DGKZ
GUCY1A3
PDE4B
PDE4D
ATP2A2
ATP2A3
NOS1
CNN1
GSTO1
NOS3
CNN2
MYLK2
CALD1
ACTA1
MYL2





Donnelly Centre for Cellular + Biomolecular Research

UNIVERSITY OF TORONTO



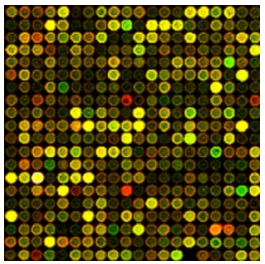
http://baderlab.org

Learning Objectives of Module

- To understand the basic concepts of pathway and network analysis.
- Be able to recognize different gene identifiers and gene attributes.
- To understand how simple enrichment analysis tools work.
- To introduce network visualization using Cytoscape.

Interpreting Gene Lists

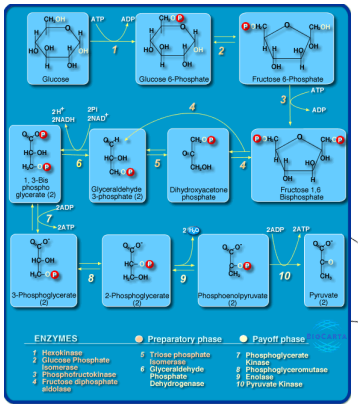
- My cool new screen worked and produced 1000 hits! …Now what?
- Genome-Scale Analysis (Omics)
  - Genomics, Proteomics
- Tell me what's interesting about these genes
  - Are they enriched in known pathways, complexes, functions

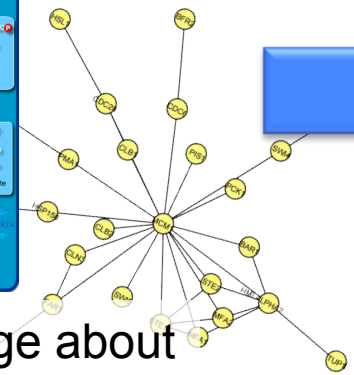


Ranking or clustering

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

Pathway database



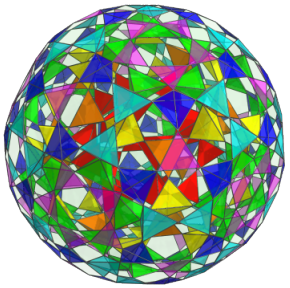


Prior knowledge about cellular processes

Analysis tools

Eureka! New heart disease gene!

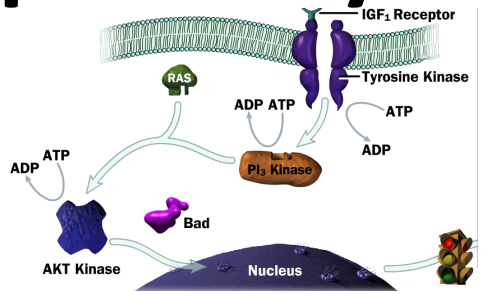
Are these genes enriched in known pathways?



http://eusebeia.dyndns.org/4d/rect120cell

a cell

?

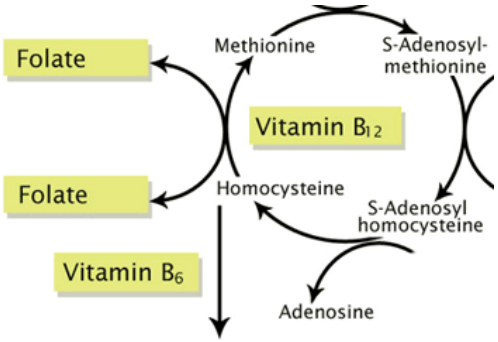


http://outreach.mcb.harvard.edu/animations/biochem.swf

signaling pathway

active or inactive ?

?



http://proventigen.com/bvitamins

metabolic pathway

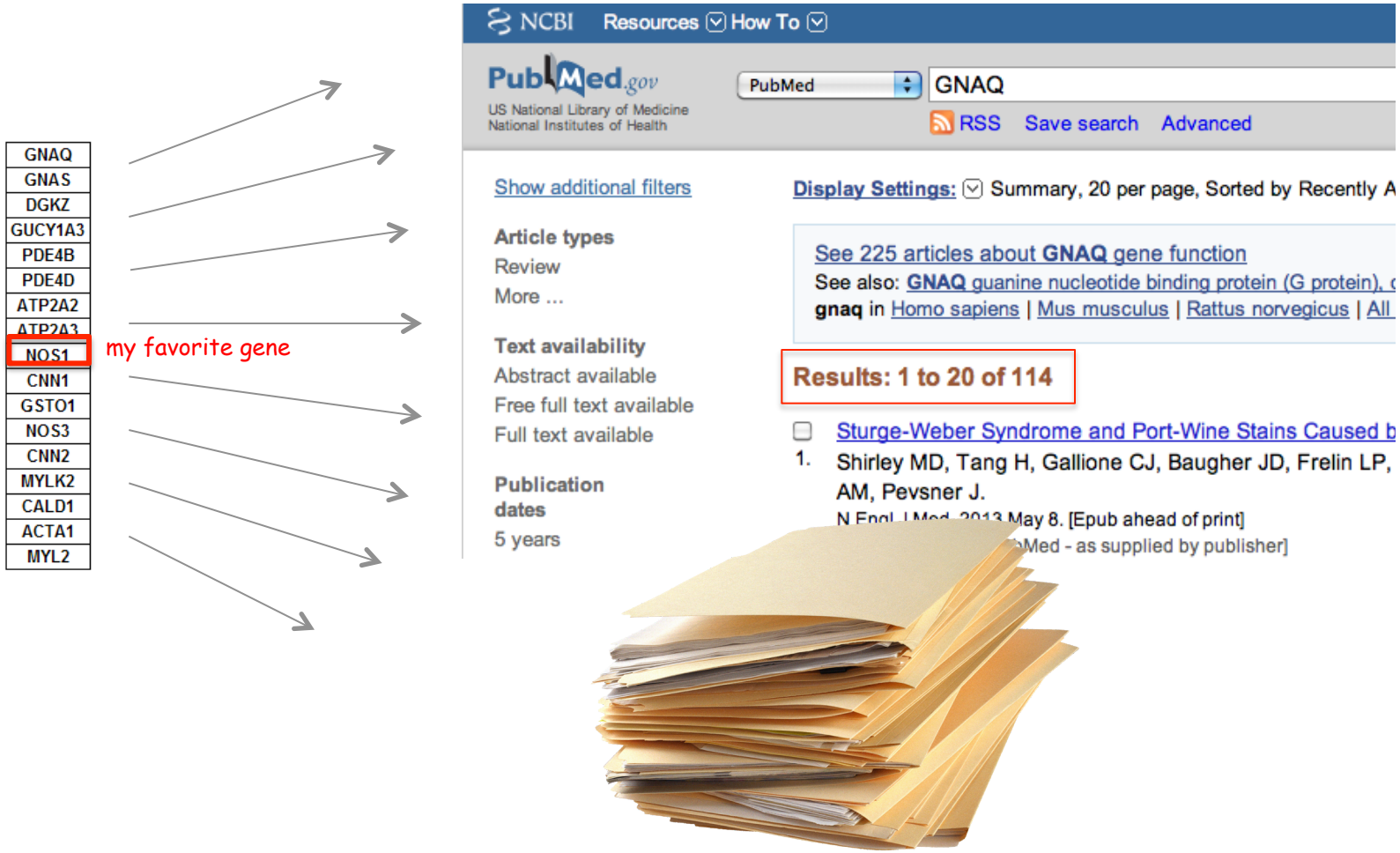
active or inactive ?

<u>gene-set definition:</u>
contains all genes in a defined pathway

but also disease or drug –pathway association

Pathway and network analysis

- Save time compared to traditional approach



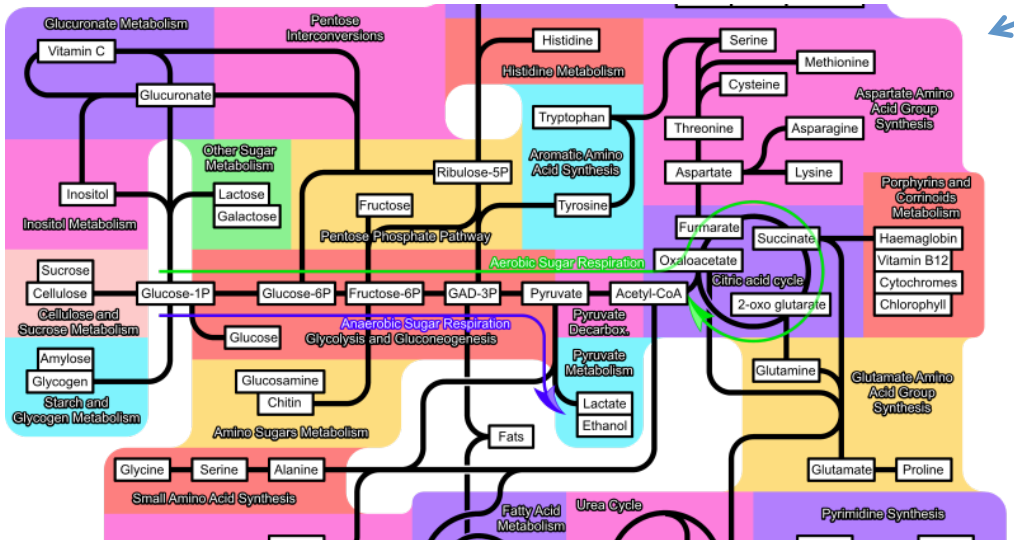
Pathway and Network analysis

- Intuitive way of analyzing results

CELL MAP

GENE LIST

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

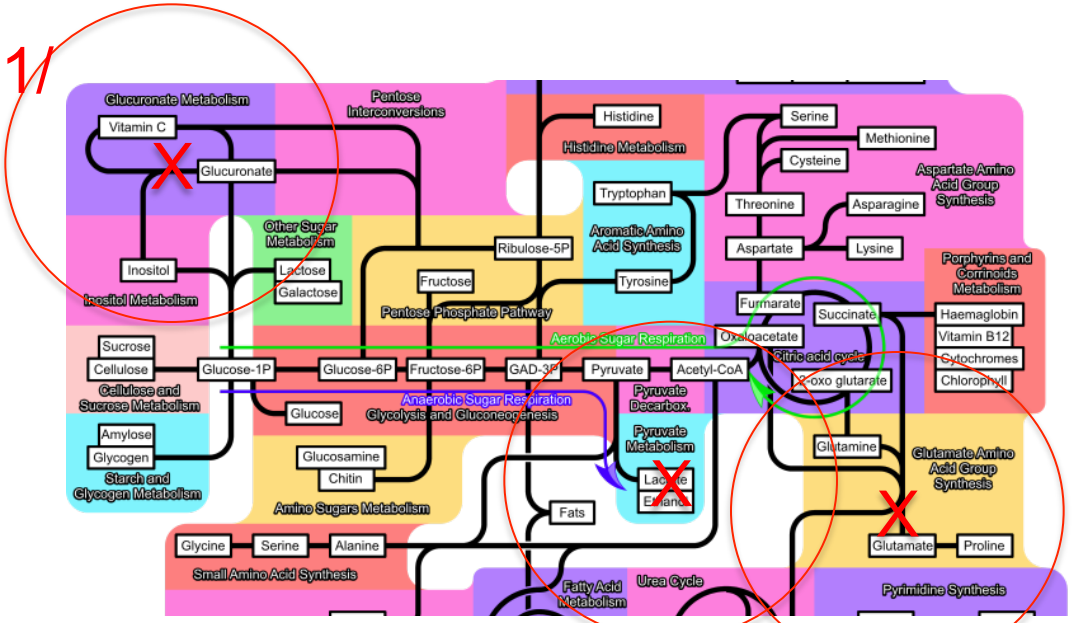


PATHWAYS

Interpreting Gene Lists

- Overlap at the pathway level

Mutation 1/
patient 1



Mutation 2
patient 2

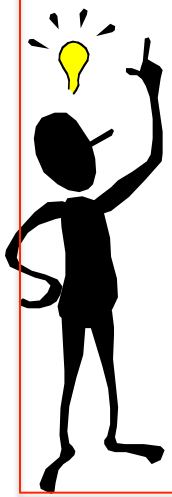
Mutation 3/
patient 3

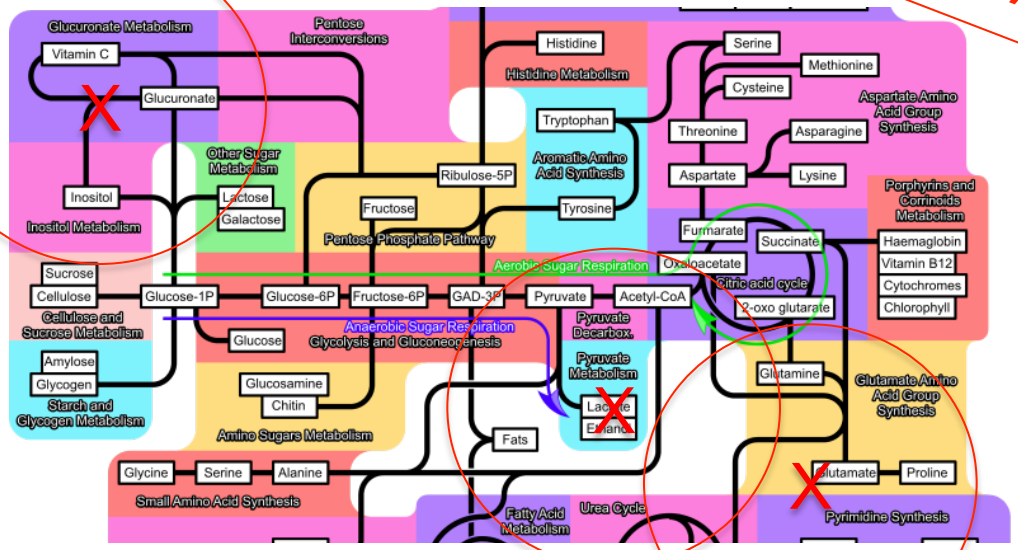
Interpreting Gene Lists

- Overlap at the pathway level

PATHWAY X IS TARGETED IN ALL PATIENTS!

Mutation 1/ patient 1



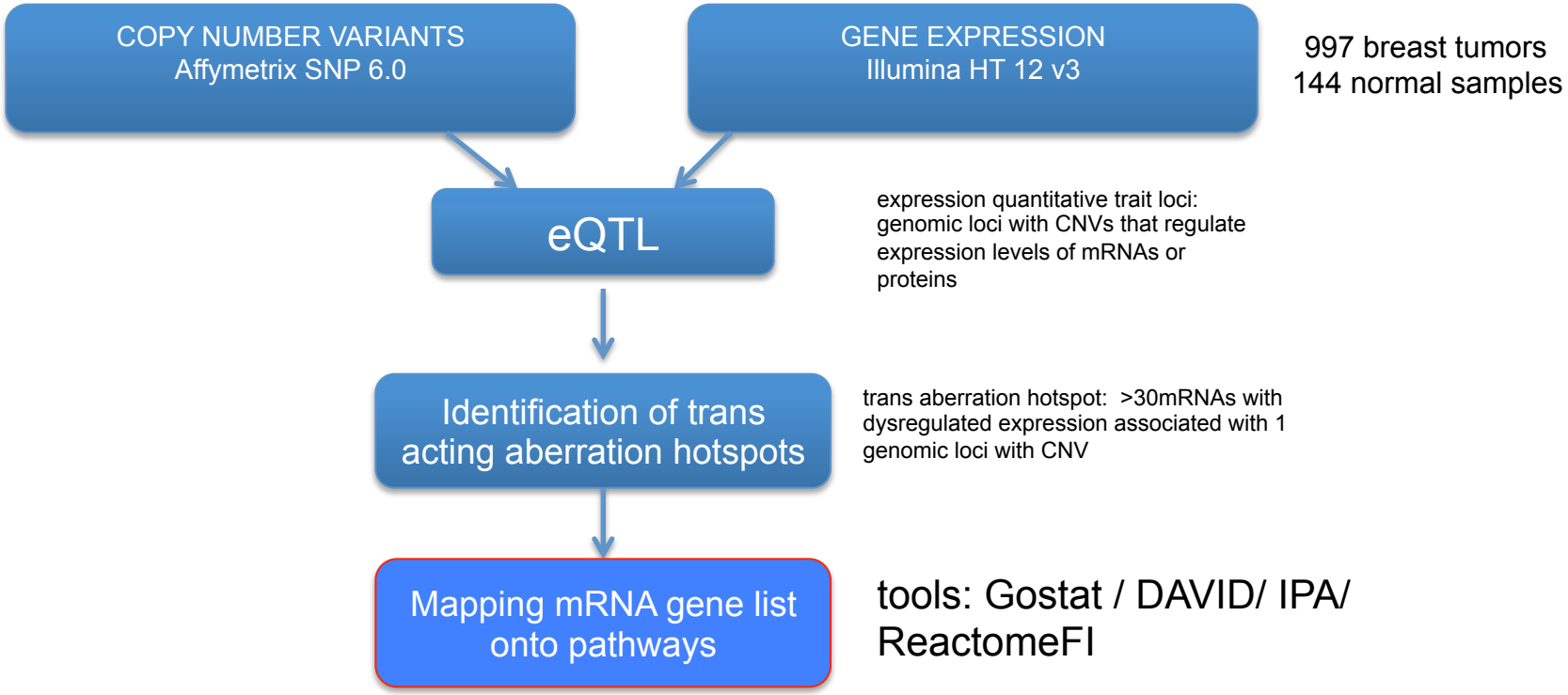


Mutation 2 patient 2

Mutation 3/ patient 3

Interpreting Gene Lists

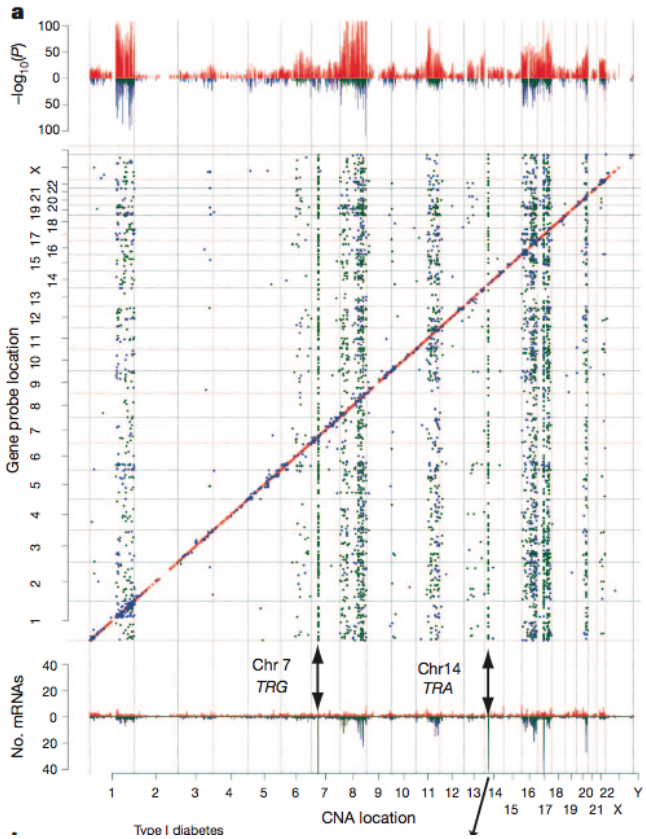
- Example: genomic and transcriptomic architecture of breast cancer



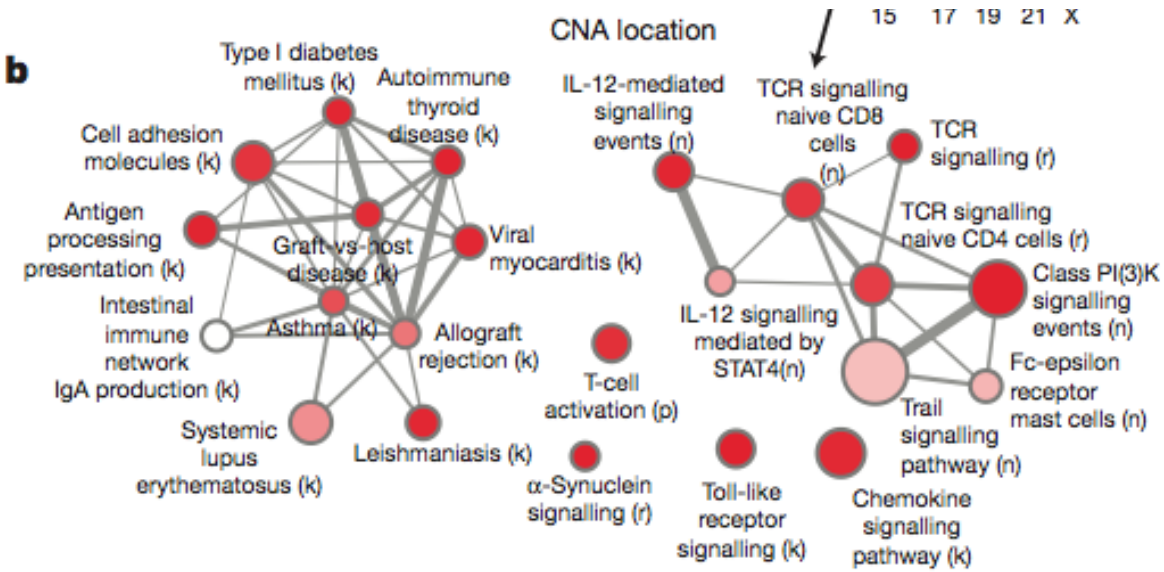
Nature. 2012 Apr 18;486(7403):346-52.**The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups.**

Interpreting Gene Lists

- Example: genomic and transcriptomic architecture of breast cancer



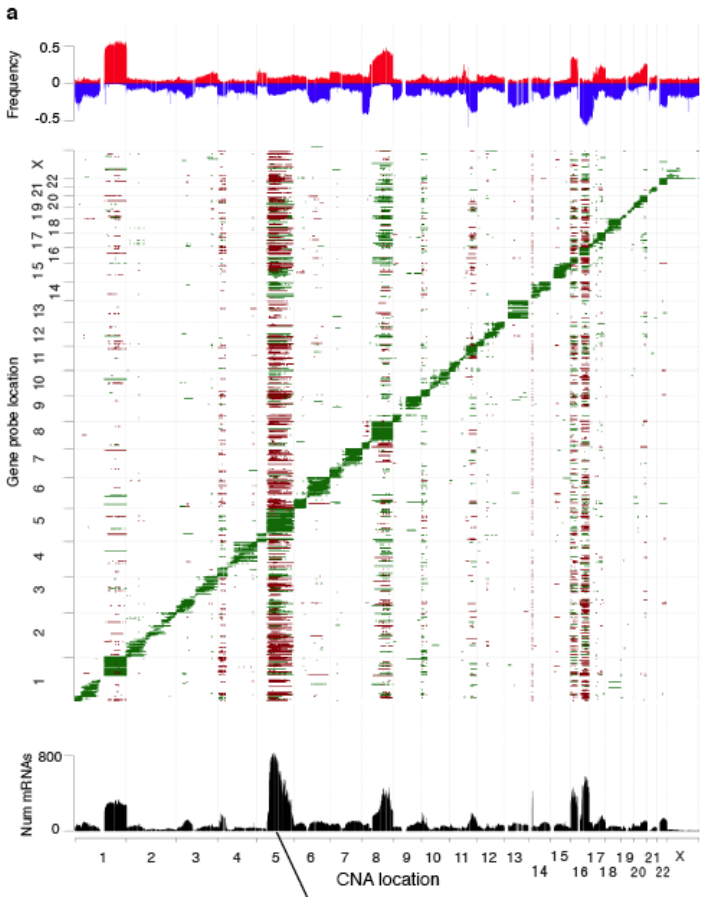
- TCR deletion-mediated adaptive immune response in CNA devoid subgroup

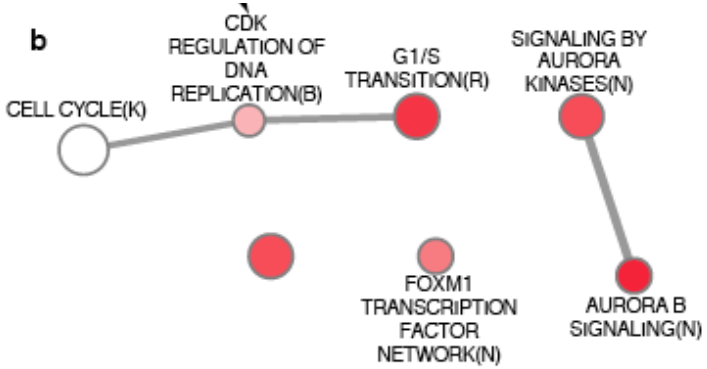


Nature. 2012 Apr 18;486(7403):346-52.**The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups.**

Interpreting Gene Lists

- Basal specific chromosome 5 deletion associated mitotic network

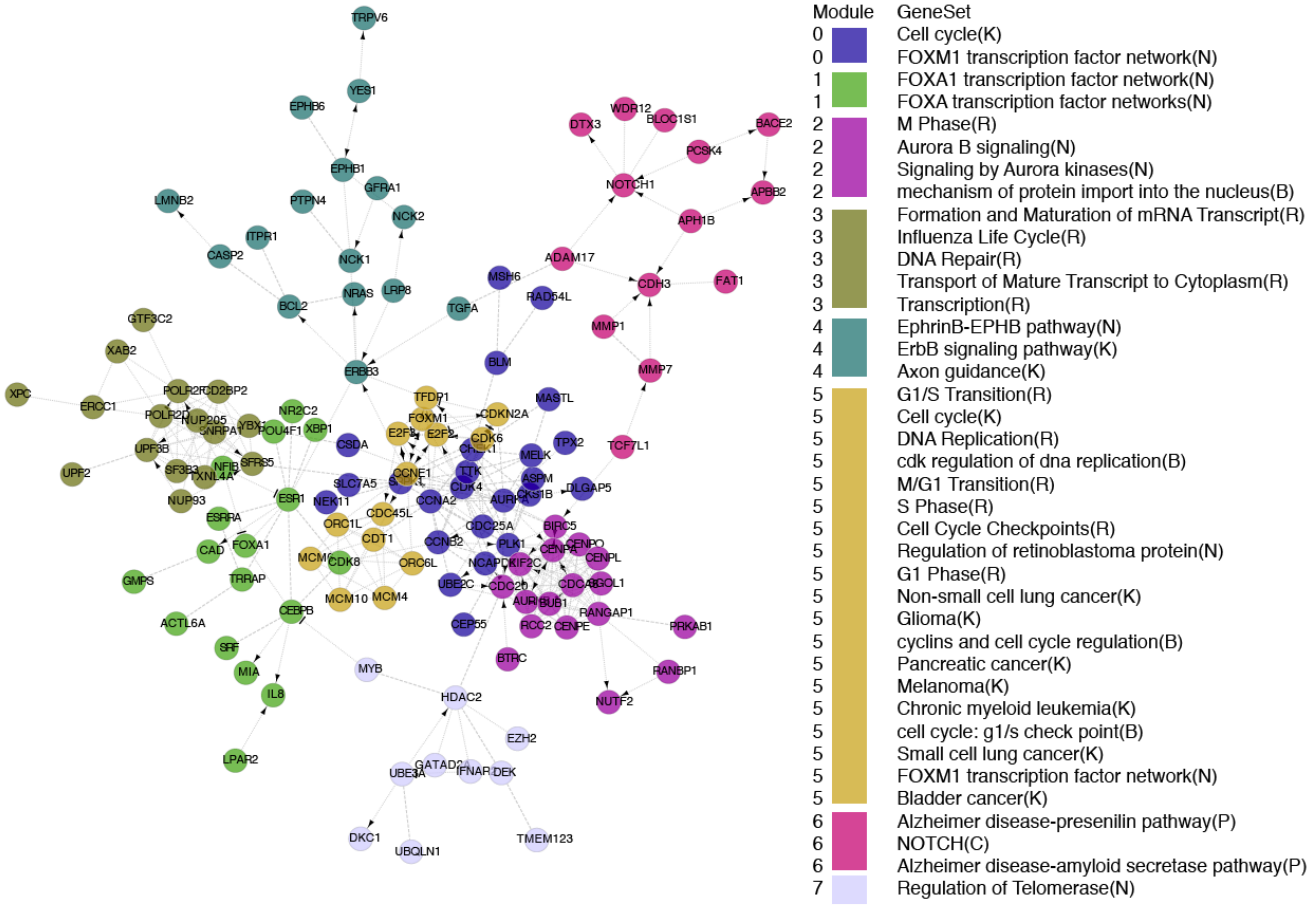




Nature. 2012 Apr 18;486(7403):346-52.**The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups.**

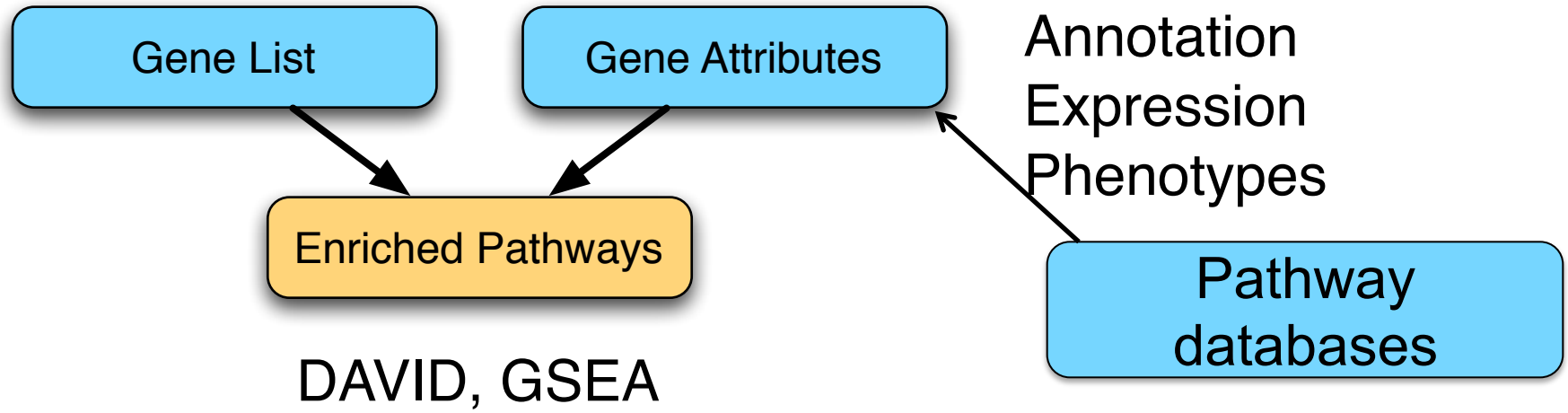
Interpreting Gene Lists

- Basal specific chromosome 5 deletion associated mitotic network



Nature. 2012 Apr 18;486(7403):346-52.**The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups.**

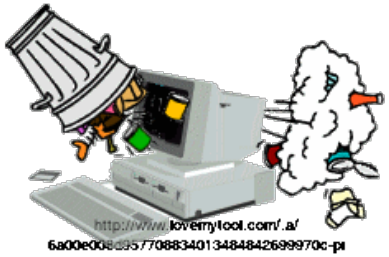
Pathway Enrichment Analysis



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

Before you start a pathway and network analysis

- ✓ Use statistics that will increase signal and reduce noise specifically for your experiment
  - ✓ Normalization
  - ✓ Background adjustment
  - ✓ Quality control



(garbage in, garbage out)

- ✓ Gene list size
- ✓ Make sure your gene IDs are compatible with software

Where Do Gene Lists Come From?

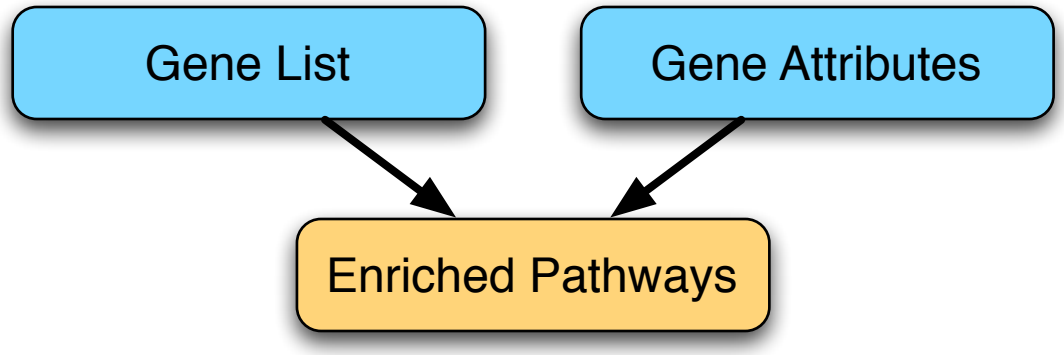
- Molecular profiling e.g. mRNA (arrays/ RNAseq), protein
- Interactions: Protein interactions, microRNA targets, transcription factor binding sites (ChIP)
- Genetic screen e.g. of knock out library
- Association studies (Genome-wide)
  - Single nucleotide polymorphisms (SNPs)
  - Copy number variants (CNVs)

Other examples?

Biological Questions?

- 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)

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.

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

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 (official gene symbols)
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

tool

Common Identifiers

FOR HUMAN GENES

GeneCards®
The Human Gene Compendium
מכון ויצמן למדע
WEIZMANN INSTITUTE OF SCIENCE
with
XENOMIX a division of LifeMap SCIENCES
Free for academic non-profit institutions. ALL other users need a commercial license from LifeMap Sciences, Inc.

Home | GeneCards Guide | Suite | Terms and Conditions | About Us | User Feedback | Mirror sites

Set Analyses: GeneALaCart GeneDecks
keyword(s) Search Advanced Search

BRCA2 Gene
protein-coding GIFtS: 68
GCID: GC13P032889

breast cancer 2, early onset
(Previous names: **Fanconi anemia, complementation group D1**)
(Previous symbols: **FANCD1, FACD, FANCD**)

Explore 119 diseases affiliated with BRCA2 via MalaCards our new Human Malady Compendium

Antibodies / cDNA/ RNAi Proteins & Enzymes Assays & Kits/ Pathways
SABiosciences Gene Network TFBS PCR Arrays Primers: ChIP / RT²
Biological research products for BRCA2
ORIGENE Proteins Antibodies Assays/ Genes/ shRNA/ Primers
GenScript Proteins Antibodies Assays / Cell Lines / Clones

Jump to Section...

Aliases
for BRCA2 gene

(According to [1]HGNC, [2]Entrez Gene, [3]UniProtKB/Swiss-Prot, [4]UniProtKB/TrEMBL, [5]OMIM, [6]GeneLoc, [7]Ensembl, [8]DME, [9]miRBase, and/or [10]tRNAdb)
About This Section

Aliases

Breast Cancer 2, Early Onset[1 2]	GLM3[2 5]
FANCD1[1 2 3 5]	PNCA2[2 5]
FACD[1 2 3]	Fanconi Anemia, Complementation Group D1[1]
BRCC2[1 2]	FANCB[2]
FAD[1 2]	BRCA1/BRCA2-Containing Complex, Subunit 2[2]
FAD1[1 2]	Breast And Ovarian Cancer Susceptibility Gene, Early Onset[2]
FANCD[1 2]	Breast Cancer 2 Tumor Suppressor[2]
Fanconi Anemia Group D1 Protein[2 3]	Breast Cancer Type 2 Susceptibility Protein[2]
BROVCA2[2 5]	Truncated Breast And Ovarian Cancer Susceptibility Protein 2[2]

External Ids: HGNC: 1101[1] Entrez Gene: 675[2] Ensembl: ENSG00000139618[7] OMIM: 600185[5] UniProtKB: P51587[3]

Export aliases for BRCA2 gene to outside databases

Previous GC identifers: GC13P030875 GC13P026876 GC13P031826 GC13P030687 GC13P031787 GC13P013701

Jump to Section...

Summaries
for BRCA2 gene

(According to Entrez Gene, Tocris Bioscience, Wikipedia's Gene Wiki, PharmGKB,

Entrez Gene summary for BRCA2:
Inherited mutations in BRCA1 and this gene, BRCA2, confer increased lifetime risk of developing breast or ovarian cancer. Both BRCA1 and BRCA2 are involved in maintenance of genome stability, specifically the homologous recombination pathway for double-strand DNA repair. The BRCA2 protein contains several copies of a 70 aa motif called the BRC motif, and these motifs mediate binding to the RAD51 recombinase which functions in DNA repair. BRCA2 is considered a tumor suppressor gene, as tumors with BRCA2 mutations generally exhibit loss of heterozygosity (LOH) of the wild-type allele. (provided by RefSeq, Dec 2008)

UniProtKB/Swiss-Prot: BRCA2_HUMAN, P51587

http://www.genecards.org/cgi-bin/carddisp.pl?gene=BRCA2

Entrez Gene ID

NCBI Resources How To

Gene | Gene | 6234[uid]

Save search Limits Advanced

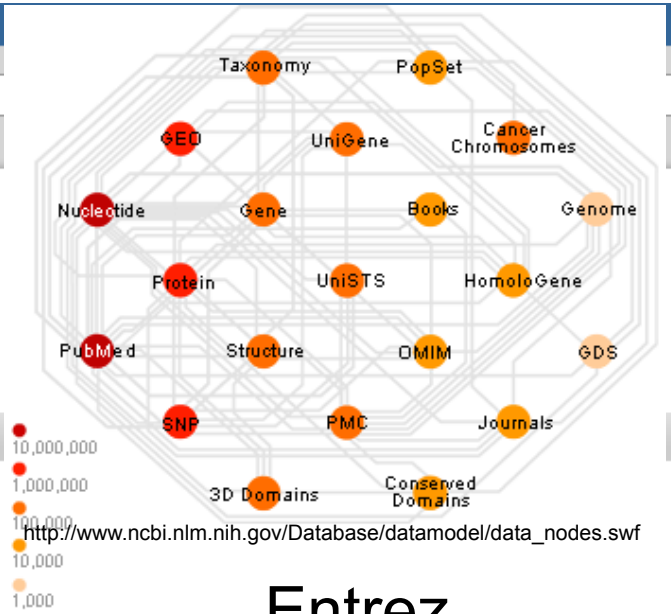
Display Settings: Full Report

RPS28 ribosomal protein S28 [*Homo sapiens* (human)]

Gene ID: 6234, updated on 5-May-2013

Summary

Official Symbol	RPS28 provided by HGNC
Official Full Name	ribosomal protein S28 provided by HGNC
Primary source	HGNC:10418
See related	HPRD:04731; MIM:603685
Gene type	protein coding
RefSeq status	REVIEWED
Organism	Homo sapiens
Lineage	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominid
Also known as	S28
Summary	Ribosomes, the organelles that catalyze protein synthesis, consist of a small 40S subunit and a large 60S subunit. Together these subunits are com RNA species and approximately 80 structurally distinct proteins. This gene encodes a ribosomal protein that is a component of the 40S subunit. The belongs to the S28E family of ribosomal proteins. It is located in the cytoplasm. As is typical for genes encoding ribosomal proteins, there are multip processed pseudogenes of this gene dispersed through the genome. [provided by RefSeq, Jul 2008]



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

Entrez

RefSeq (NCBI Reference Sequences)

These reference sequences exist independently of genome builds. Explain

mRNA and Protein(s)

NM_001031.4 → NP_001022.1 40S ribosomal protein S28

Status: REVIEWED

Source sequence(s)	BC021239, BC070217
Consensus CDS	CCDS45953.1
UniProtKB/TrEMBL	B2R4R9
UniProtKB/Swiss-Prot	P62857

Conserved Domains (1) summary

cd04457 Location:10 – 55 Blast Score: 208	S1_S28E; S1_S28E: S28E, S1-like RNA-binding domain. S1-like RNA-binding domains are fo associated proteins. S28E protein is a component of the 30S ribosomal subunit. S28E is highly and eukaryotes. S28E may ...

RefSeqs of Annotated Genomes: Homo sapiens Annotation Release 104

The following sections contain reference sequences that belong to a specific genome build. Explain

Reference GRCh37.p10 Primary Assembly

Genomic

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

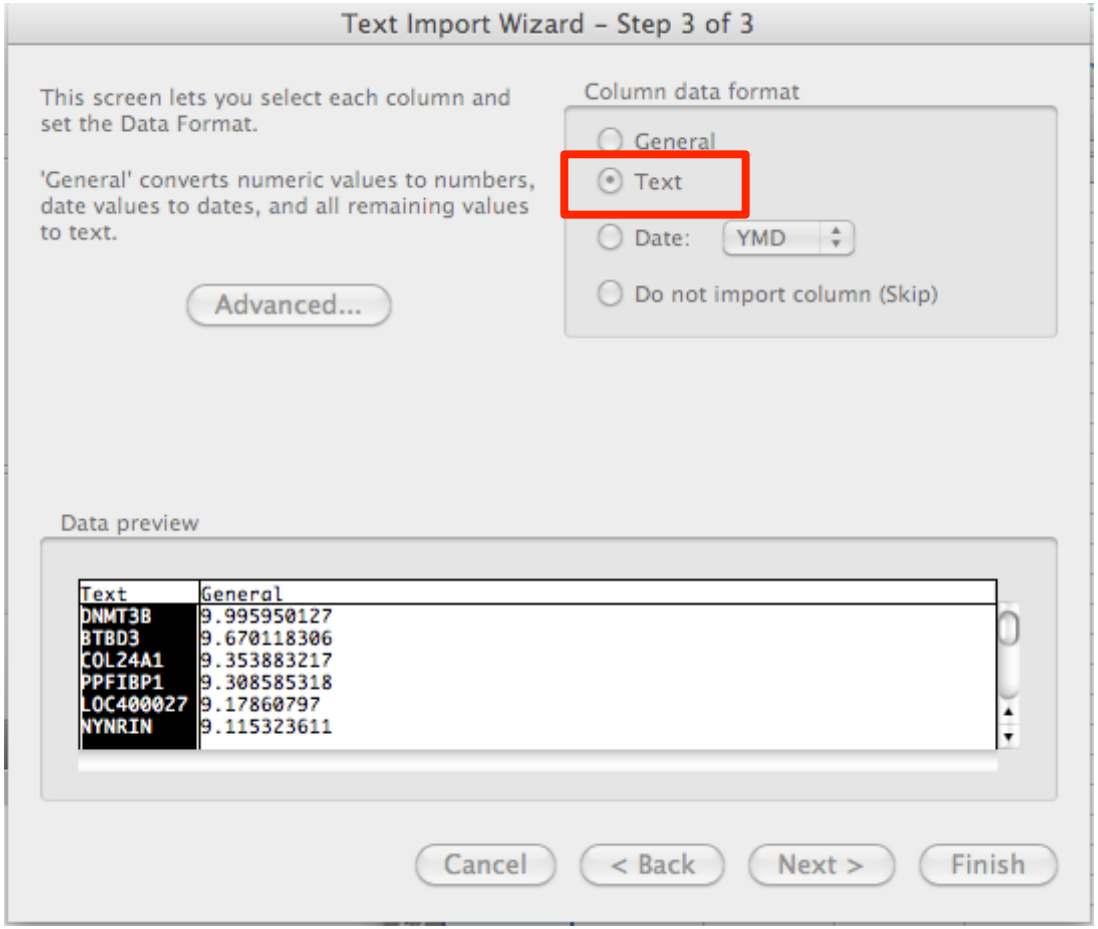
- Avoid errors: map IDs correctly
- Gene name ambiguity – not a good ID
  - e.g. FLJ92943, LFS1, TRP53, p53
  - Better to use the standard gene symbol: TP53
- Excel error-introduction
  - OCT4 is changed to October-4
- Problems reaching 100% coverage
  - E.g. due to version issues
  - Use multiple sources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

- Excel auto-format

01-Mar	NM_017923.
02-Mar	NM_001005
03-Mar	NM_178450.
04-Mar	NM_020814.
05-Mar	NM_017824.
06-Mar	NM_005885.
07-Mar	NM_022826.
08-Mar	NM_145021.
09-Mar	NM_138396.
10-Mar	NM_001100
11-Mar	NM_001102
01-Sep	XM_944593.
02-Sep	NM_001008
03-Sep	NM_019106.
04-Sep	NM_080415.
05-Sep	NM_002688.
06-Sep	NM_145799.
07-Sep	NM_001788.
08-Sep	NM_001098
09-Sep	NM_006640.
10-Sep	NM_144710.
11-Sep	NM_018243.
12-Sep	NM_144605.
13-Sep	NR_024271.
14-Sep	NM_207366.
15-Sep	NM_004261.
01-Dec	NM_017418.



ID Challenges

- Gene name

Letters to Nature

Nature **426**, 100 (6 November 2003) | doi:10.1038/nature02141

Retraction: Hes1 is a target of microRNA-23 during retinoic-acid-induced neuronal differentiation of NT2 cells

Hiroaki Kawasaki & Kazunari Taira

Nature **423**, 838–842 (2003).

In this Article, the messenger RNA that is identified to be a target of microRNA-23 (miR-23) is from the gene termed human 'homolog of ES1' (HES1), accession number Y07572, and not from the gene encoding the transcriptional repressor 'Hairy enhancer of split' HES1 (accession number NM_00524) as stated in our paper. We incorrectly identified the gene because of the confusing nomenclature. The function of HES1 Y07572 is unknown but the encoded protein shares homology with a protein involved in isoprenoid biosynthesis. Our experiments in NT2 cells had revealed that the protein levels of the repressor Hes1 were diminished by miR-23. Although we have unpublished data that suggest the possibility that miR-23 might also interact with Hes1 repressor mRNA, the explanation for the finding that the level of repressor Hes1 protein decreases in response to miR-23 remains undefined with respect to mechanism and specificity. Given the interpretational difficulties resulting from our error, we respectfully retract the present paper. Further studies aimed at clarifying the physiological role of miR-23 will be submitted to a peer-reviewed journal subject to the outcome of our ongoing research.

will be used during lab

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/

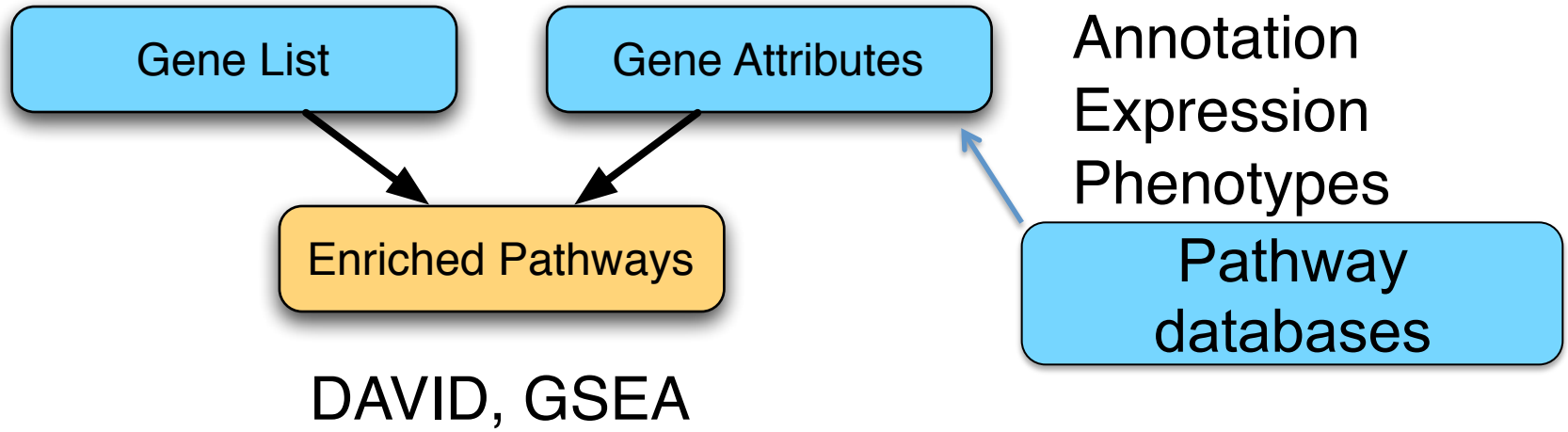
Recommendations

- 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

What Have We Learned?

- Genes and their products and attributes have many identifiers (IDs)
- Genomics often requires conversion of IDs from one type to another
- ID mapping services are available
- Use standard, commonly used IDs to reduce ID mapping challenges

Pathway Enrichment Analysis



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

What Are 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

DATABASES containing annotated pathways (function annotation)

Gene Ontology

MSigDB-c2

REACTOME

KEGG

NCI

BIOCARTA

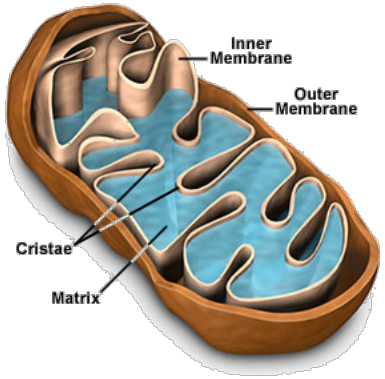
HumanCyc

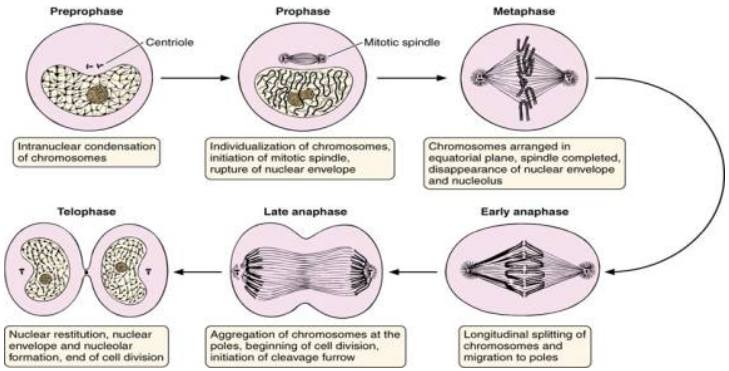
What is the Gene Ontology (GO)?

- Largest database
- 41,007 gene products (proteins) annotated for human
- Updated every 3 months
- Organism independent /many model organisms (Homo Sapiens, Mus musculus, Danio Rerio…)
- GO resources are freely available to anyone without restriction

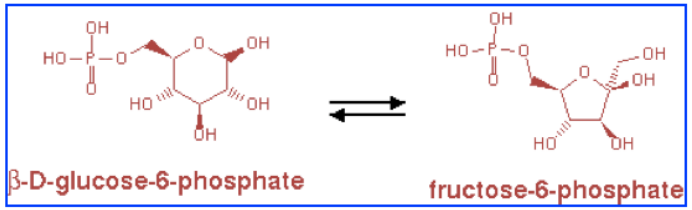
What GO Covers?

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





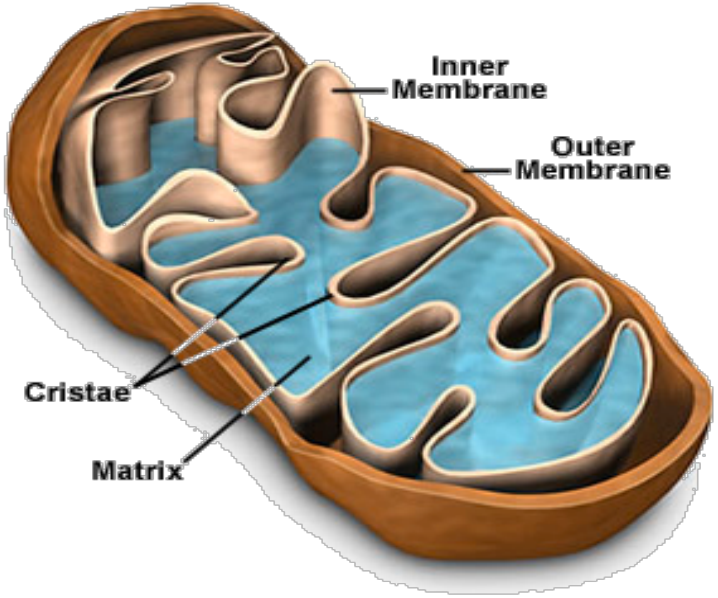
Cell division



glucose-6-phosphate isomerase activity

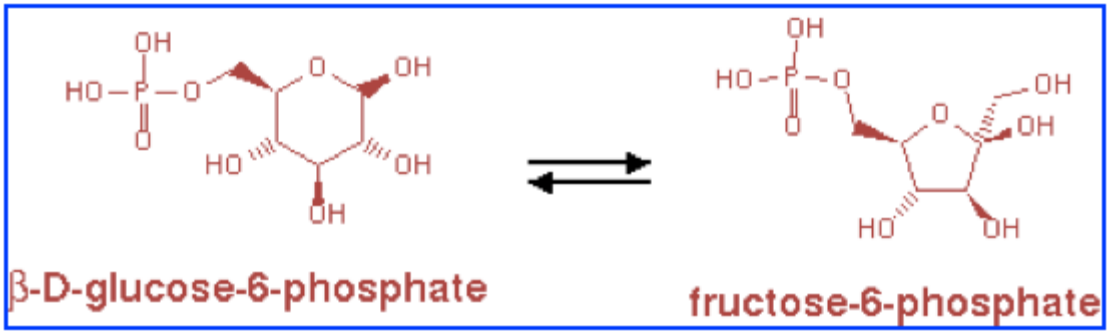
What GO Covers?

- **cellular component**: the parts of a cell or its extracellular environment;



What GO Covers?

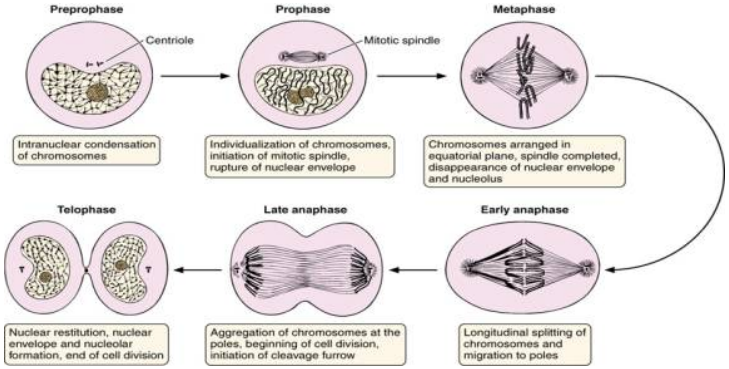
- **molecular function**: the elemental activities of a gene product at the molecular level, such as binding or catalysis;



glucose-6-phosphate isomerase activity

What GO Covers?

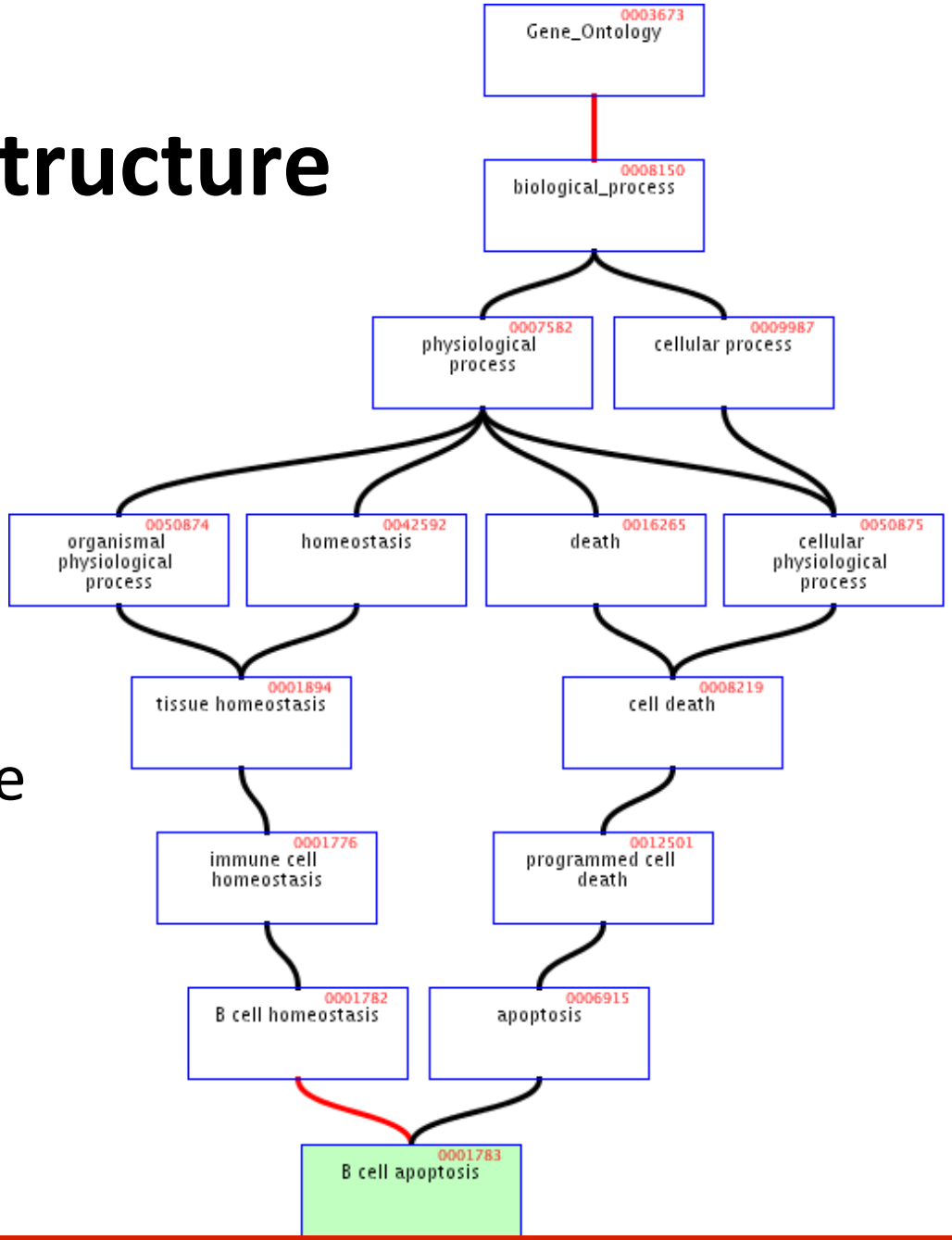
- **biological process**: operations or sets of molecular events with a defined beginning and end, pertinent to the functioning of integrated living units: cells, tissues, organs, and organisms.



Cell division

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



How genes are linked, or associated, with GO terms by trained curators

In this study, we report the isolation and molecular characterization of the B. napus **PERK1** cDNA, that is predicted to encode a novel receptor-like kinase. We have shown that like other plant RLKs, the kinase domain of PERK1 has serine/threonine kinase activity. In addition, the location of a PERK1-GTP fusion protein to the plasma membrane supports the prediction that PERK1 is an integral membrane protein… these kinases have been implicated in early stages of wound response…

Molecular function

Molecular component

Biological process

Manual
Or
Electronic
curation

GO Evidence Types

information about how the annotation was created

alpha-1,6-mannosyltransferase activity

ALG12	GO:0000009	IC	Dol-P-Man:Man(7)GlcNAc(2)-PP-Dol alpha-1,6-mannosyltransferase
TNP1	GO:0000012	IDA	Spermatid nuclear transition protein 1
PSMD4	GO:0000209	TAS	26S proteasome non-ATPase regulatory subunit 4
COX6B1	GO:0005743	TAS	Cytochrome c oxidase subunit 6B1
MEF2C	GO:0000002	IEA	Uncharacterized protein

IC: inferred by curator

IDA: Inferred from direct assay

TAS: Traceable Author Statement

IEA: Inferred by electronic annotations

Guide to GO Evidence Codes: http://www.geneontology.org/GO.evidence.shtml

Note: Evidence codes cannot be used as a measure of the quality of the annotation.

will be used during lab

Accessing GO: QuickGO

Search for a GO term: > examples - apoptosis, GO:0006915

Search for a Protein: > examples - tropomyosin, P06727

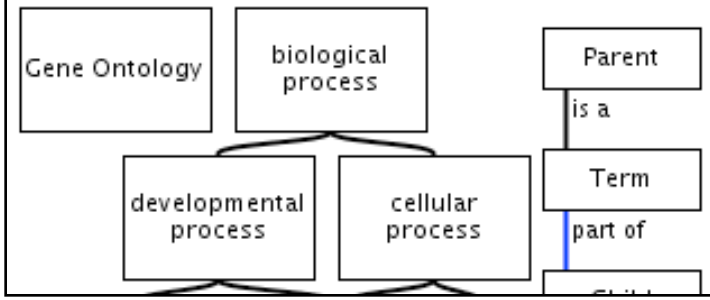
Compare GO terms: > example - GO:0000122,GO:0000001

Find, view and download annotation

GO:0006915 apoptosis

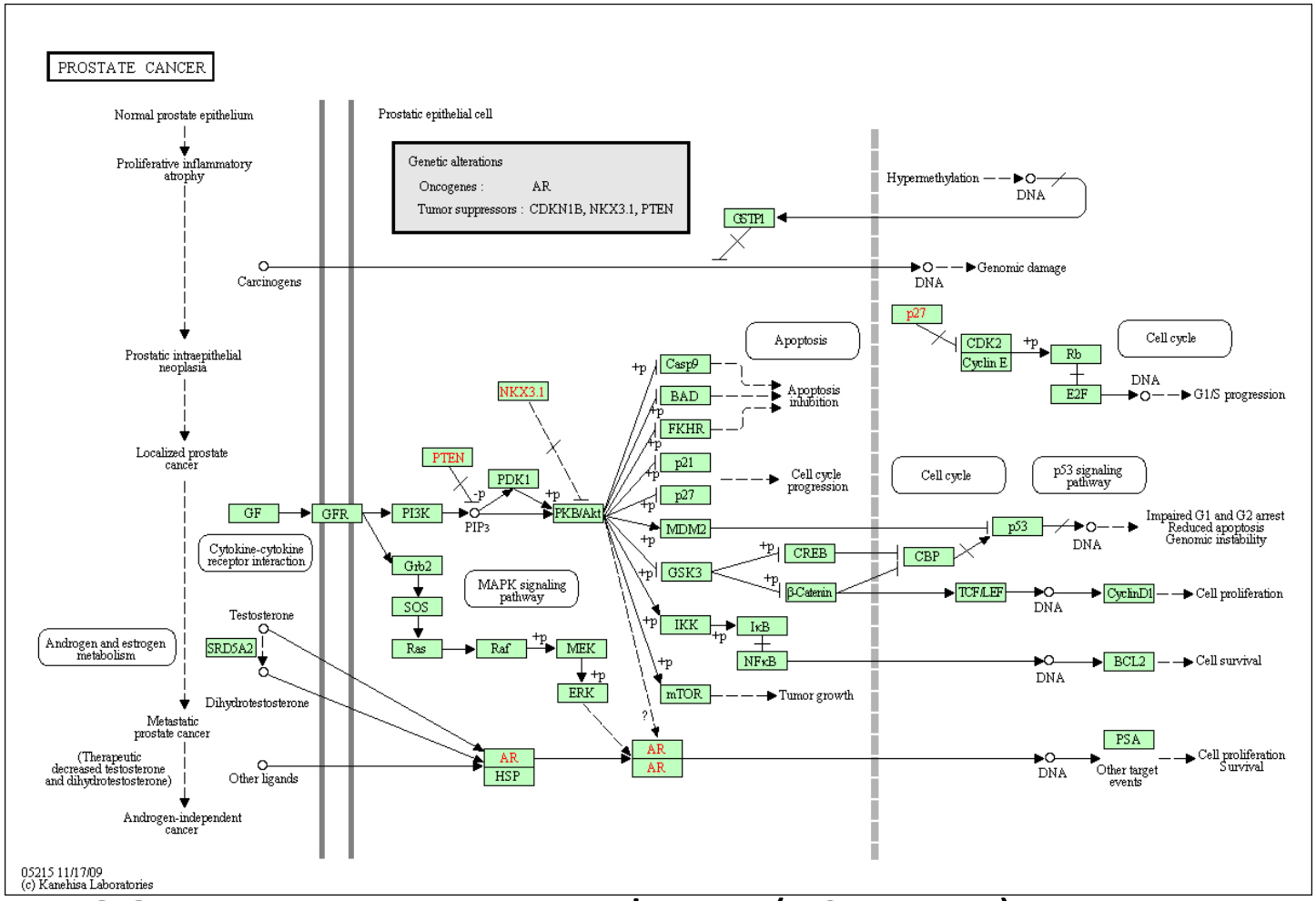
A form of programmed cell death induced by external or internal signals that trigger the activity of proteolytic caspases, whose actions dis internally with condensation and subsequent fragmentation of the cell nucleus (blebbing) while the plasma membrane remains intact. Oth the exposure of phosphatidyl serine on the cell surface.

Term Information | Ancestor chart | Ancestor table | Child Terms | Protein Annotation | Statistics



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

KEGG

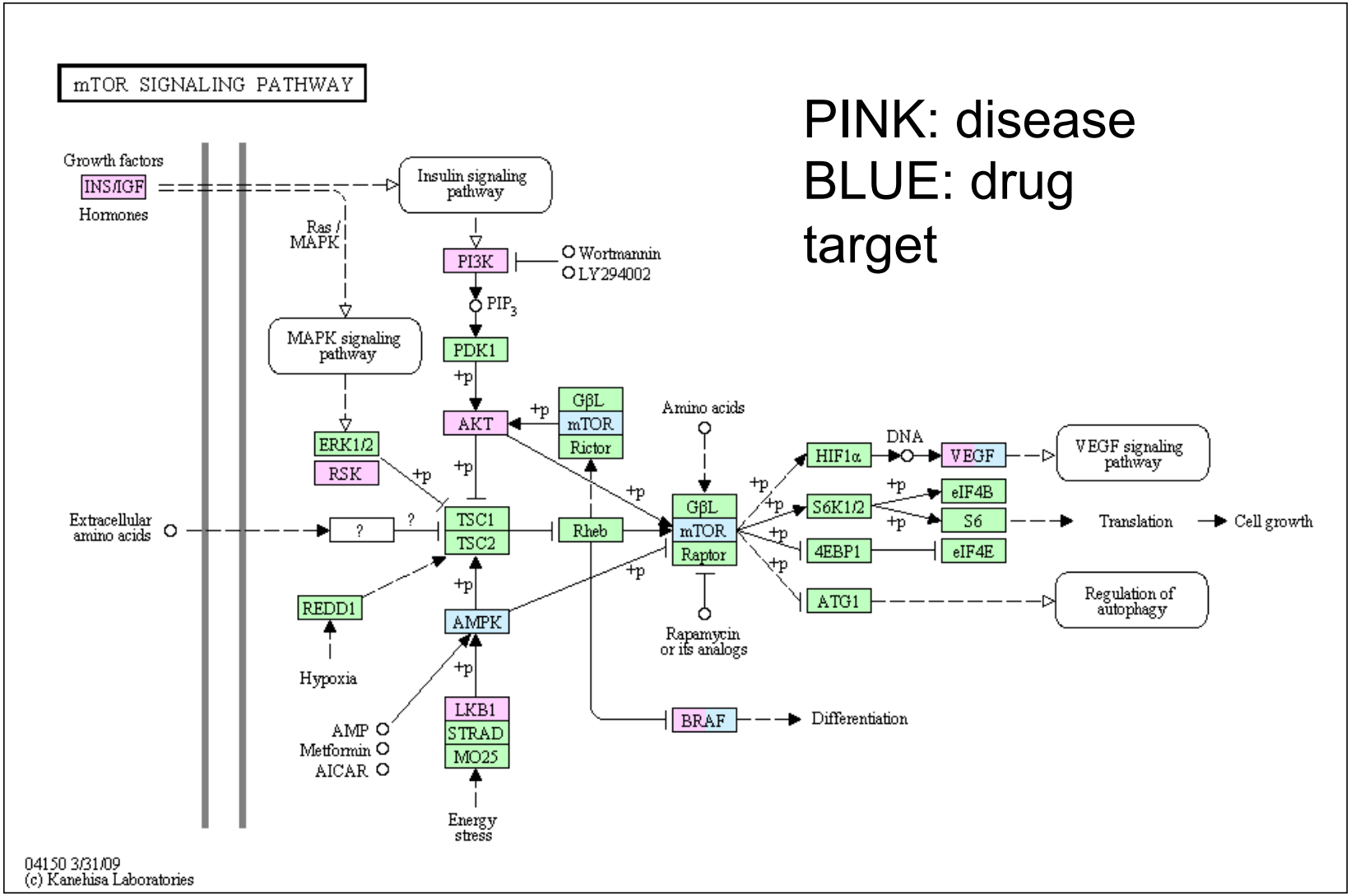


KEGG prostate cancer pathway (42 genes)

KEGG

- Pathway maps for metabolism and other cellular processes, as well as human diseases; manually created from published materials
- 5.2M genes; 1024 species; 100K pathways.
- Most pathways are projected across species.
- Features:
  - pathway/gene lookup;
  - colorize pathways with gene lists.
- Free for academic use; need license to download
- Current statistics: http://www.genome.jp/kegg/docs/statistics.html

KEGG: disease and drug annotations



Biocarta

PATHWAYS ▸ Activation of Src by Protein-tyrosine phosphatase alpha

Submitted by: Guru:

COMMENT ON THIS PATHWAY | DESCRIPTION | CONTRIBUTORS | SAVE THIS LINK | SUBMIT | LEGEND

PRODUCT INDEX

PRODUCT SEARCH

SEARCH

Contains Exact

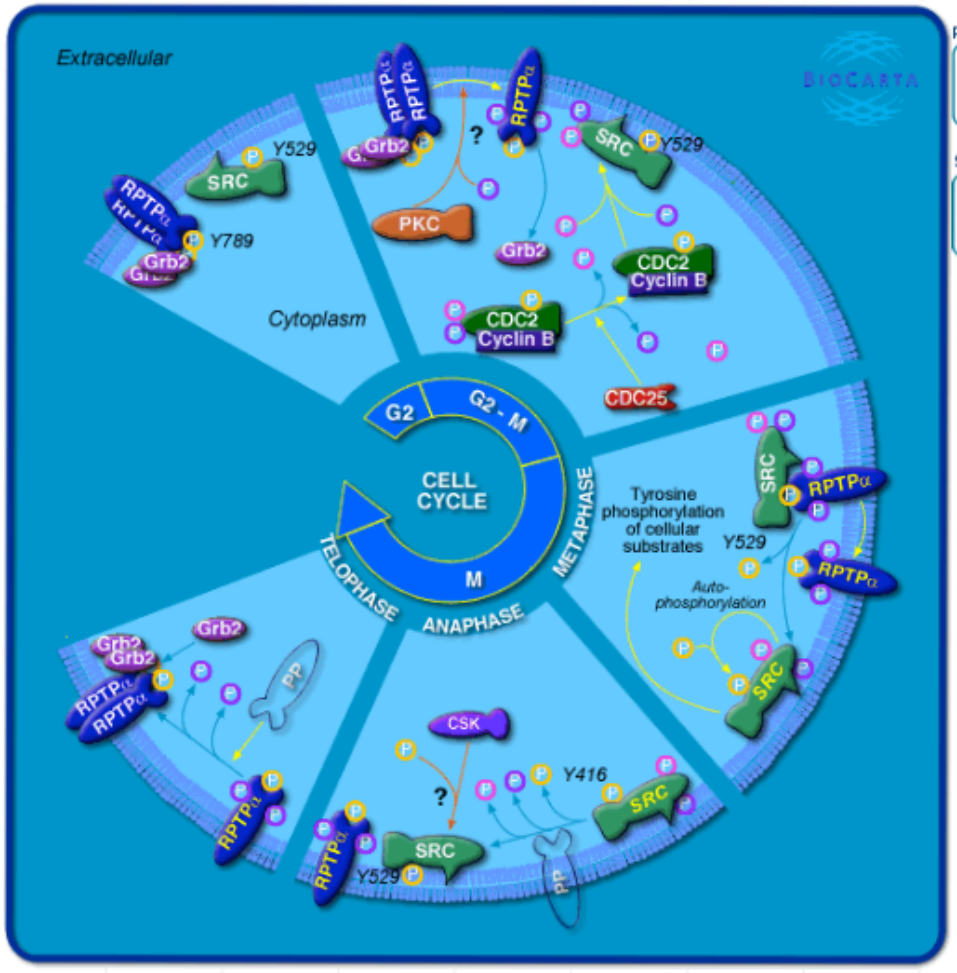
Advanced Search

PRODUCT HIGHLIGHT

ON OFF

PROTEIN LIST

REQUEST A CATALOG



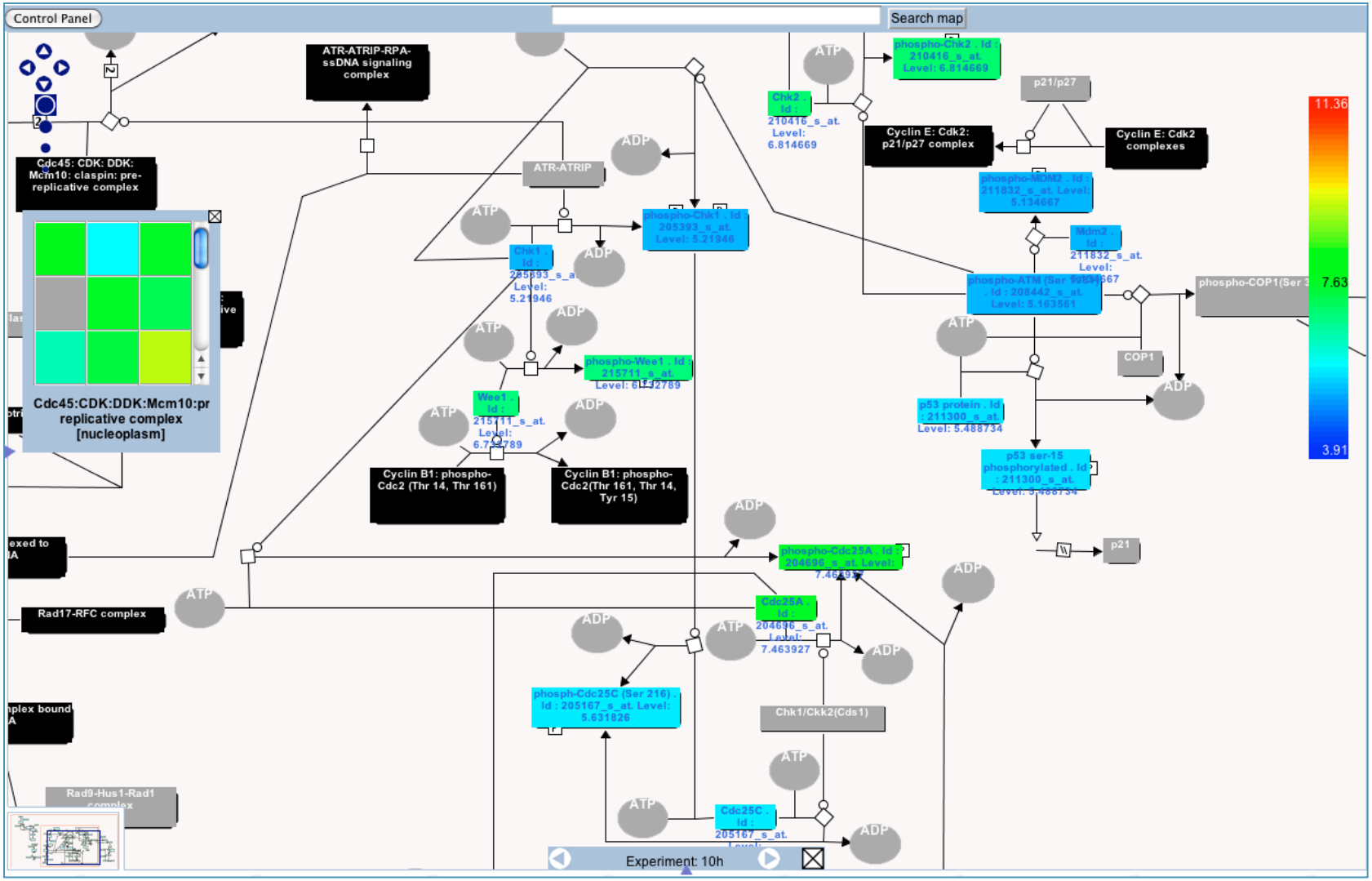
This Pathway:

Other Species:

Biocarta

- Beautiful hand-drawn pathway diagrams;
- company which develops reagents and assays for biopharmaceutical research: community-annotation service mostly used by drug companies to overlay drug ads;
- No underlying database; can't automate for gene list interpretation;
- pathways from diverse fields like apoptosis, cell cycle, cell signalling, development, immunology, neuroscience, adhesion, and metabolism.

Reactome


Control Panel
Search map
ATR-ATRIP-RPA-ssDNA signaling complex
Cdc45: CDK: DDK: Mcm10: claspin: pre-replicative complex
ATR-ATRIP
Cyclin E: Cdk2: p21/p27 complex
Cyclin E: Cdk2 complexes
p21/p27
phospho-MDM2 . Id . 211832_s_at. Level: 5.134667
Mdm2 . Id : 211832_s_at. Level:
phospho-ATM (Ser 2367) . Id : 208442_s_at. Level: 5.163561
phospho-COP1(Ser 3
COP1
p53 protein . Id : 211300_s_at. Level: 5.488734
p53 ser-15 phosphorylated . Id : 211300_s_at. Level: 5.488734
p21
Chk2 . Id : 210416_s_at. Level: 6.814669
phospho-Chk1 . Id . 205393_s_at. Level: 5.21946
Chk1 . Id : 205393_s_at. Level: 5.21946
phospho-Wee1 . Id . 215711_s_at. Level: 6.732789
Wee1 . Id : 215711_s_at. Level: 6.732789
Cyclin B1: phospho-Cdc2 (Thr 14, Thr 161)
Cyclin B1: phospho-Cdc2(Thr 161, Thr 14, Tyr 15)
phospho-Cdc25A . Id : 204696_s_at. Level: 7.463927
Cdc25A . Id : 204696_s_at. Level: 7.463927
Rad17-RFC complex
phosph-Cdc25C (Ser 216) . Id : 205167_s_at. Level: 5.631826
Chk1/Ckk2(Cds1)
Cdc25C . Id : 205167_s_at.
Rad9-Hus1-Rad1 complex
Cdc45:CDK:DDK:Mcm10:pr replicative complex [nucleoplasm]
ATP
ADP
11.36
7.63
3.91
Experiment: 10h

Reactome

- Hand-curated pathways in human.
- Rigorous curation standards – every reaction traceable to primary literature.
- Automatically-projected pathways to non-human species.
- 22 species; 1112 human pathways; 5078 proteins.
- Features:
  - Google-map style reaction diagrams with overlays;
  - Find pathways containing your gene list;
  - Calculate gene overrepresentation in pathways;
  - Find corresponding pathways in other species.
- Open access.

Ingenuity

INGENUITY® SYSTEMS

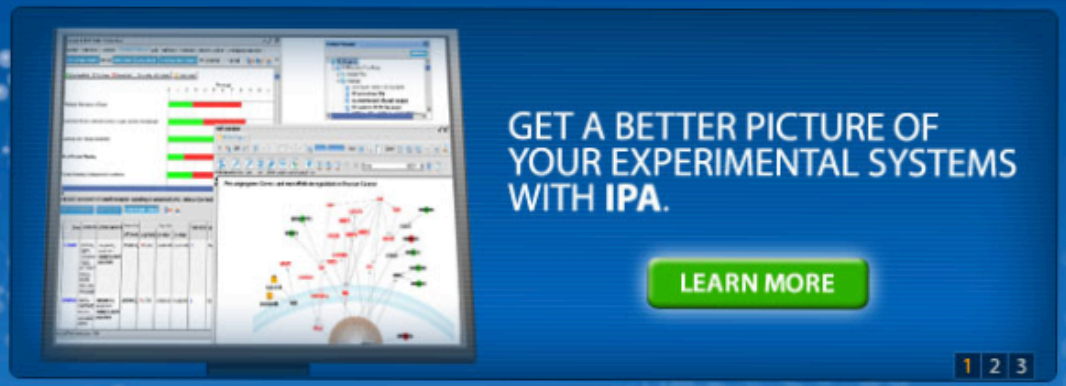
Customer Support | Log into IPA *(Login help)* | Add Me to My Institution's License | FREE TRIAL

Home | Products & Services | Library | News & Events | Partners | About Us | Careers

Discover The Biology

Ingenuity Systems is the leading provider of information solutions and custom services to the life science community.

- IPA
- Ingenuity Answers
- Other products and services



GET A BETTER PICTURE OF YOUR EXPERIMENTAL SYSTEMS WITH **IPA**.

LEARN MORE

1 2 3

NEWS

RSS/Blog | Facebook | Twitter

Ingenuity Announces 2010 Regional Training Series in 11 Cities Worldwide

Ingenuity CTO invited to participate in NIH-initiated workshop on biosecurity

PRESS RELEASES

SCIENCE SPOTLIGHT

"The goal of our study was to provide knowledge about intracellular signaling events in glioma cancer stem cells in response to perturbations by hypoxia, inhibition of STAT3 phosphorylation and IL-6 stimulation. Glioma cancer stem cells (gCSC) are refractory to traditional therapies and new insights are needed to understand their underlying biology. We used IPA to compliment our analysis of multiple comparisons between gCSC treatments. We will continue to employ IPA tools in analysis of global phosphoproteomic data sets."

- Charles Conrad, M.D.
Professor, Department of Neuro-Oncology
The University of Texas M. D. Anderson Cancer Center

See how IPA was used in this publication:

EVENTS

TRAINING

Ingenuity offers free weekly training webinars for IPA.

» View Schedule

EVENTS

Regional Training: Indianapolis, IN
August 3-4

Adapt 2010
September 13-16

Ingenuity

- Popular $$$$ commercial application.
- Very polished user interface.
- Combination of curation, integration and machine learning, but algorithms unpublished.
- Content stats unavailable.
- Features:
  - Identify pathways containing list of genes;
  - Extract and build custom pathways/networks;
  - Integration with pharmaceutical information
- Subscription required.

Pathway Commons

Pathway Commons

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

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

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

Search Pathway Commons:

Find Pathways | Find Molecules

Search

For example, if you enter: BRCA1, you will **get back the list of pathways** containing the keyword "BRCA1", and the list of pathways that contain the BRCA1 gene.

Current filter settings: All Organisms, All Data Sources. Set filters.

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

- Browse and search Pathways from multiple databases in a uniform format.
- 564 species; 1,623 pathways / All data is freely available.

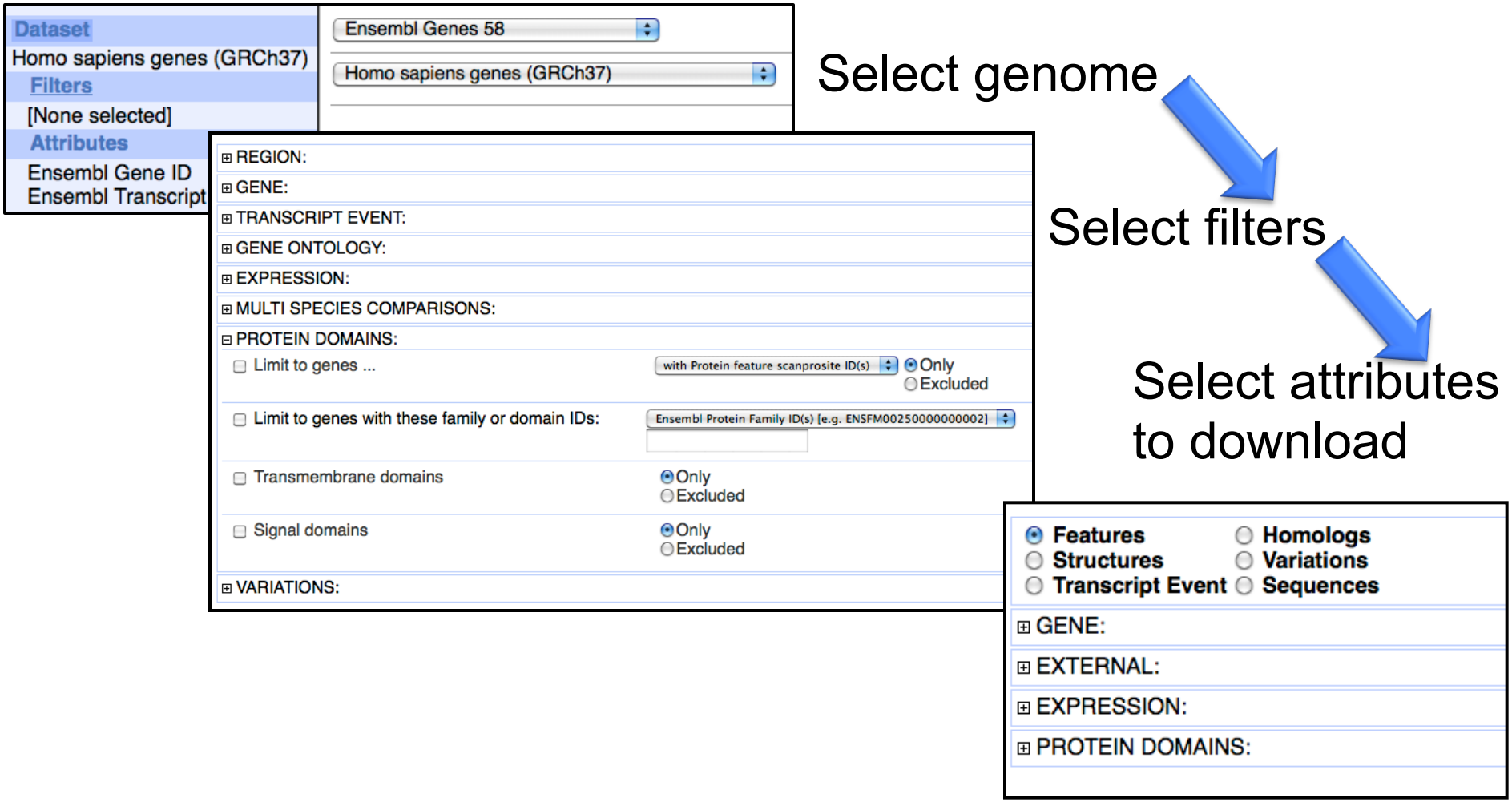
Gene Attributes

- 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

will be used during lab

Ensembl BioMart

- Convenient access to gene list annotation



What Have We Learned?

- Gene attributes define functions, characteristics of a gene
- Many gene attributes in databases
  - Gene Ontology (GO) provides gene function annotation
    - GO is a classification system and dictionary for biological concepts
  - KEGG
  - Reactome
- Many gene attributes available from Ensembl and Entrez Gene

for your information

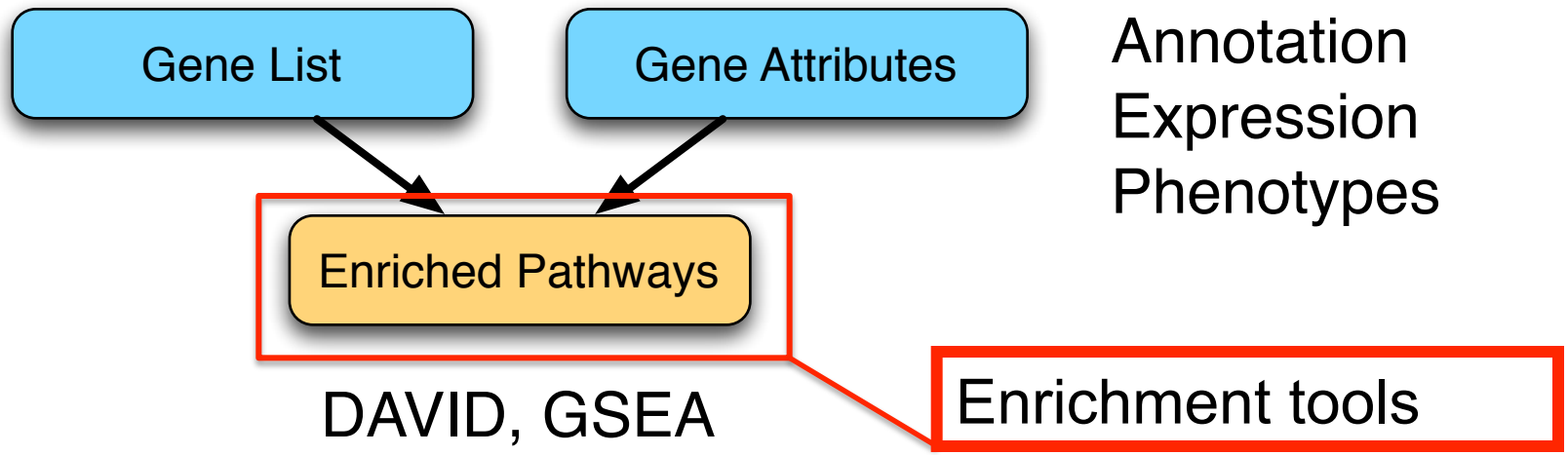
URLs

- GO-http://www.geneontology.org
- KEGG – www.genome.jp/kegg
- Biocarta – www.biocarta.com
- WikiPathways – www.wikipathways.org
- Reactome – www.reactome.org
- NCI/PID – pid.nci.nih.gov
- Ingenuity – www.ingenuity.com
- Pathway Commons – www.pathwaycommons.org/pc/

Sources of Gene Attributes

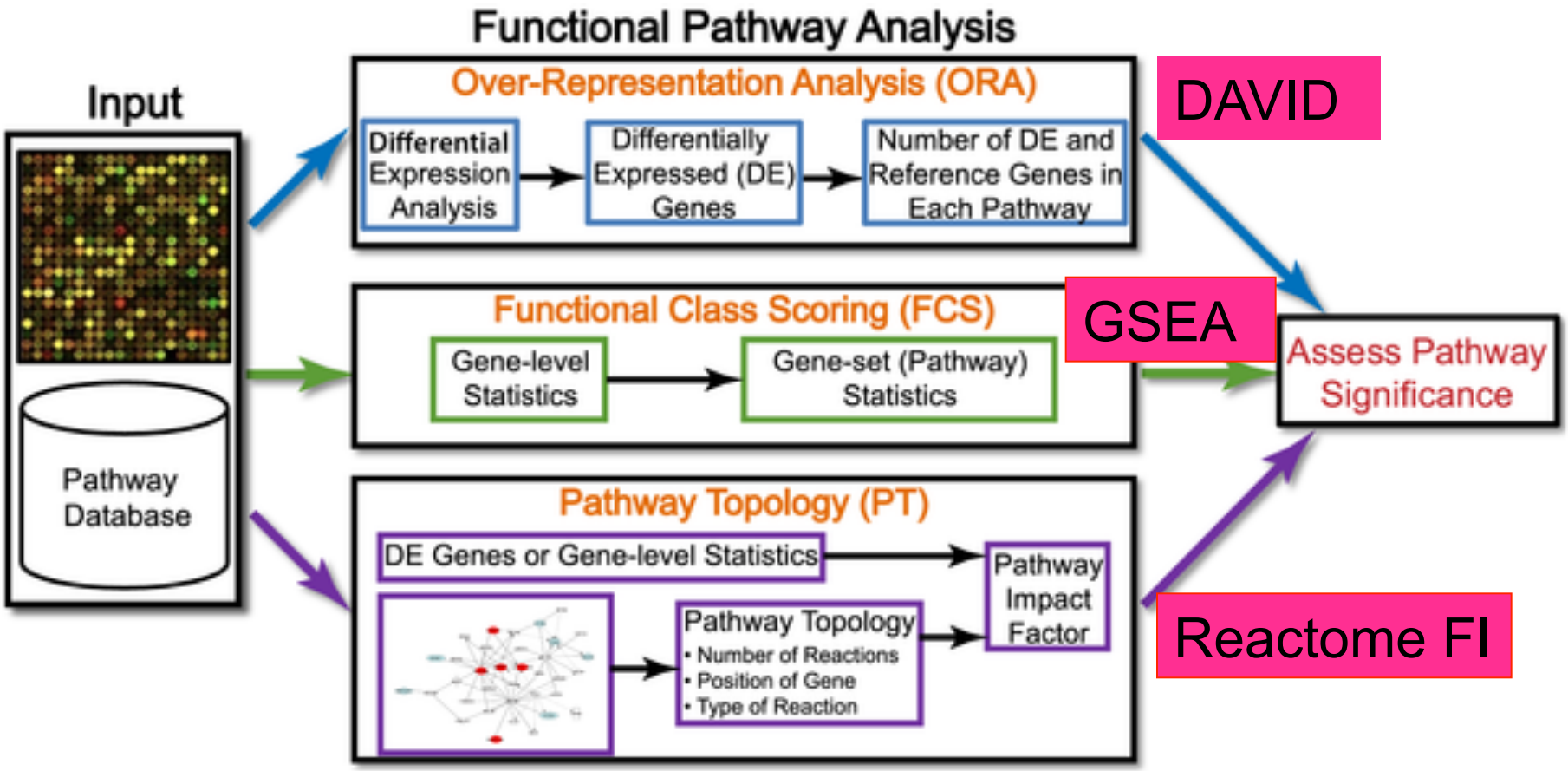
- Ensembl BioMart (general)
  - http://www.ensembl.org
- Entrez Gene (general)
  - http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene
- Model organism databases
  - E.g. SGD: http://www.yeastgenome.org/

Pathway Enrichment Analysis



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

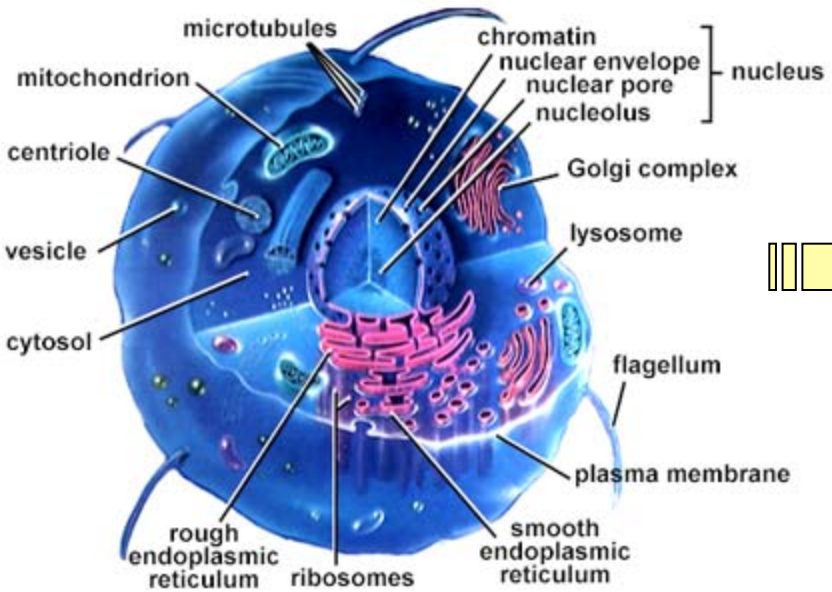
Overview of existing pathway analysis methods



Khatri P, Sirota M, Butte AJ (2012) Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. PLoS Comput Biol 8(2): e1002375. doi:10.1371/journal.pcbi.1002375

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

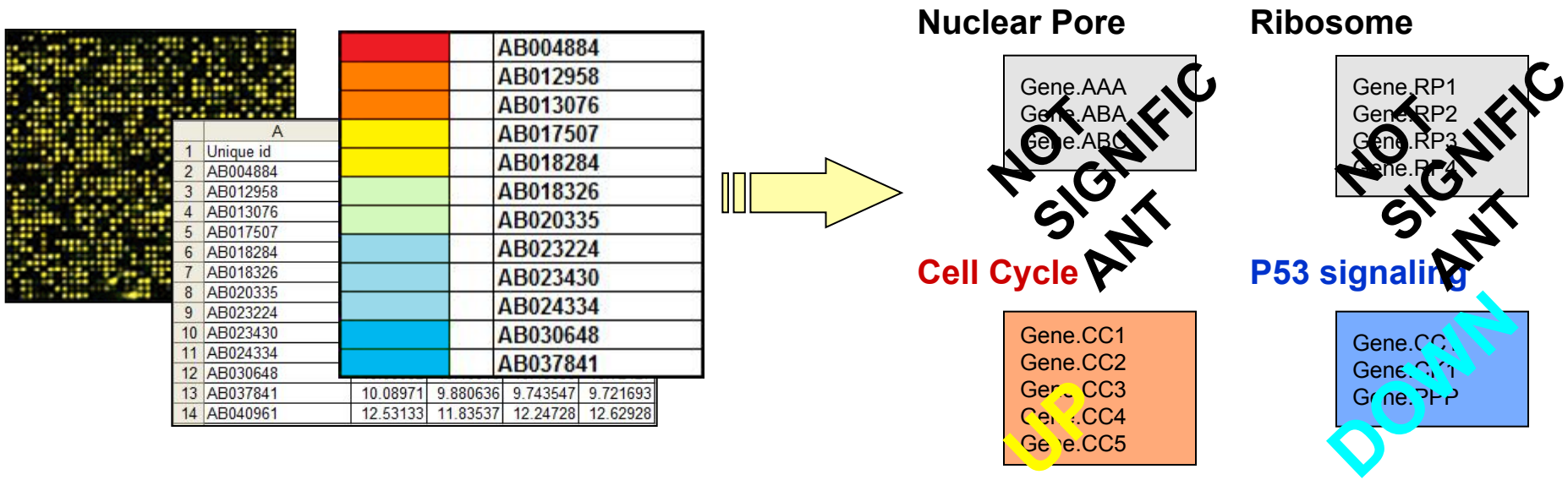
P53 signaling

Gene.CC1
Gene.CK1
Gene.PPP

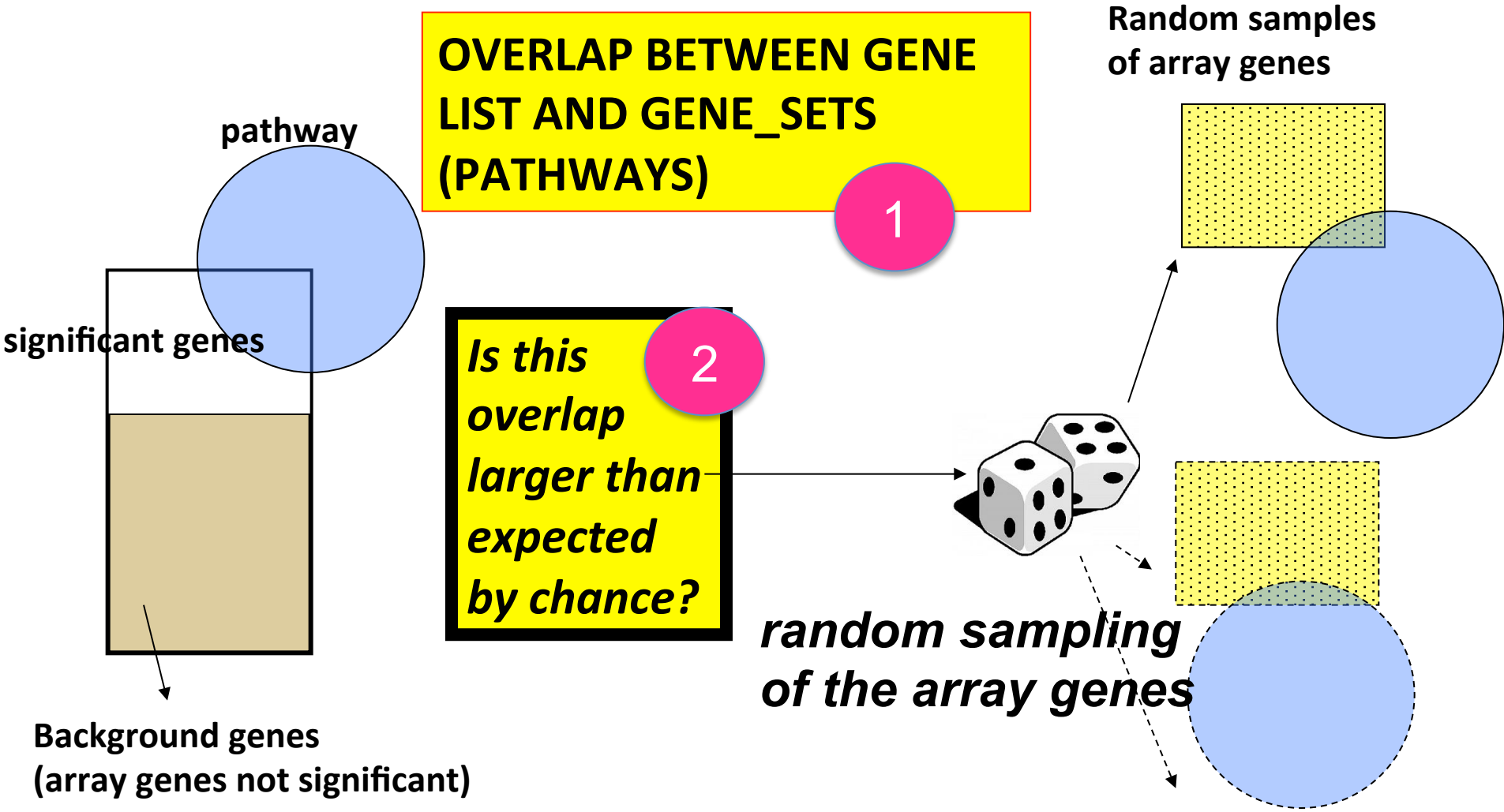
Daniele Merico

What is Gene Set Enrichment Analysis?

- Find known gene sets (e.g. pathways) enriched in a gene list
  - Look for significant overlap between gene list and pathways



How do simple enrichment tests work?

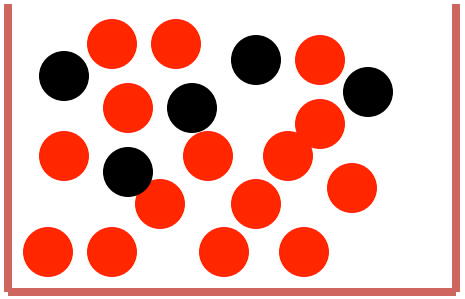


Fisher's exact test

Gene list

- RRP6
- MRD1
- RRP7
- RRP43
- RRP42

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



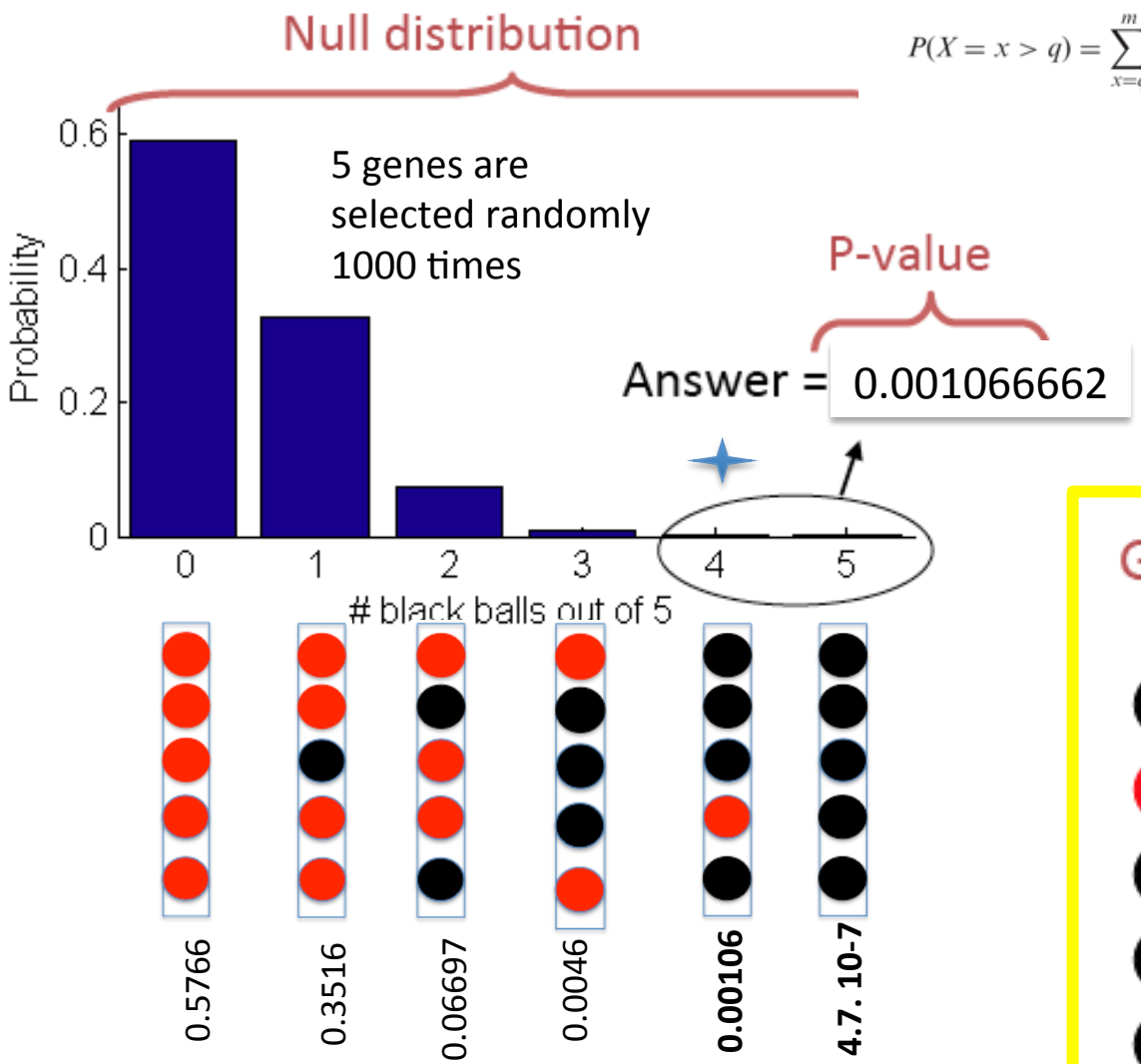
Background population:
500 black genes,
4500 red genes

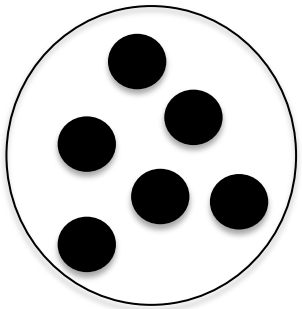
Fisher's exact test

gene universe

all genes in the genome or in array: 5 red, 45 black

$$P(X = x > q) = \sum_{x=q}^{m} \frac{\binom{m}{x}\binom{t-m}{k-x}}{\binom{t}{k}}.$$





1 gene-set (apoptosis)

Gene list

- RRP6
- MRD1
- RRP7
- RRP43
- RRP42

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

Different steps of enrichment analysis

1. The overlap is tested with each gene-set present in the pathway database (>3,000 gene-sets ?)
2. The gene-sets are ranked by the enrichment p-value to find out the most significant gene-sets (you want the lowest p-values)
3. The enrichment p-values need to be corrected for multiple hypothesis testing (FDR, Benjamini-Hochberg for example)

False discovery rate (FDR)

- FDR corrects for multiple hypothesis testing
- FDR is *the expected* ***proportion*** *of the observed enrichments due to random chance.*
- Typically FDR corrections are calculated using the Benjamini-Hochberg procedure.
- FDR threshold is often called the "q-value"

What Have We Learned?

Typical output of an enrichment analysis is:

Pathway name	Number of overlapping genes	Number of genes in pathway	P-value	Adjusted p-value
…	….	…	…	….

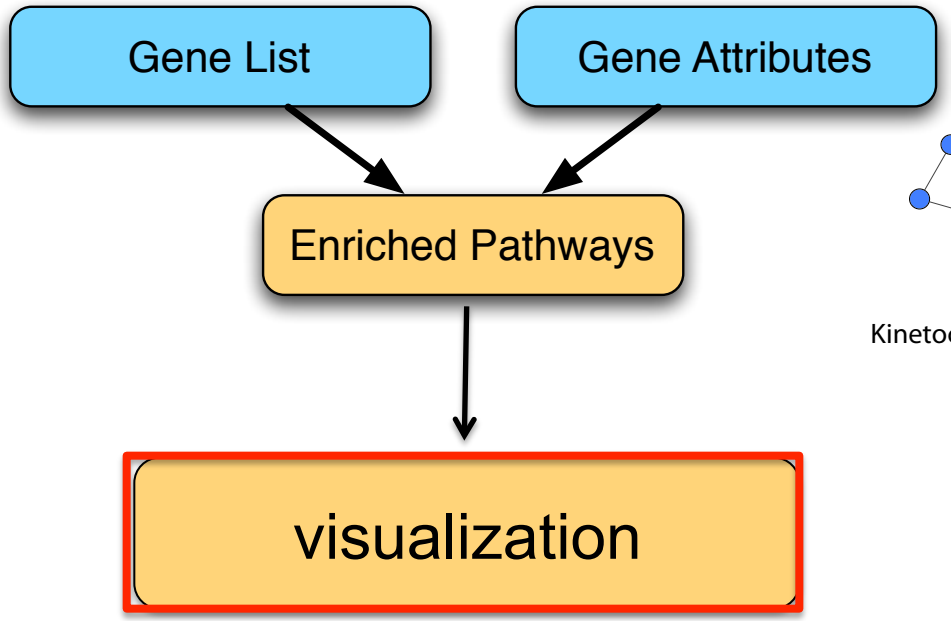
Typical output

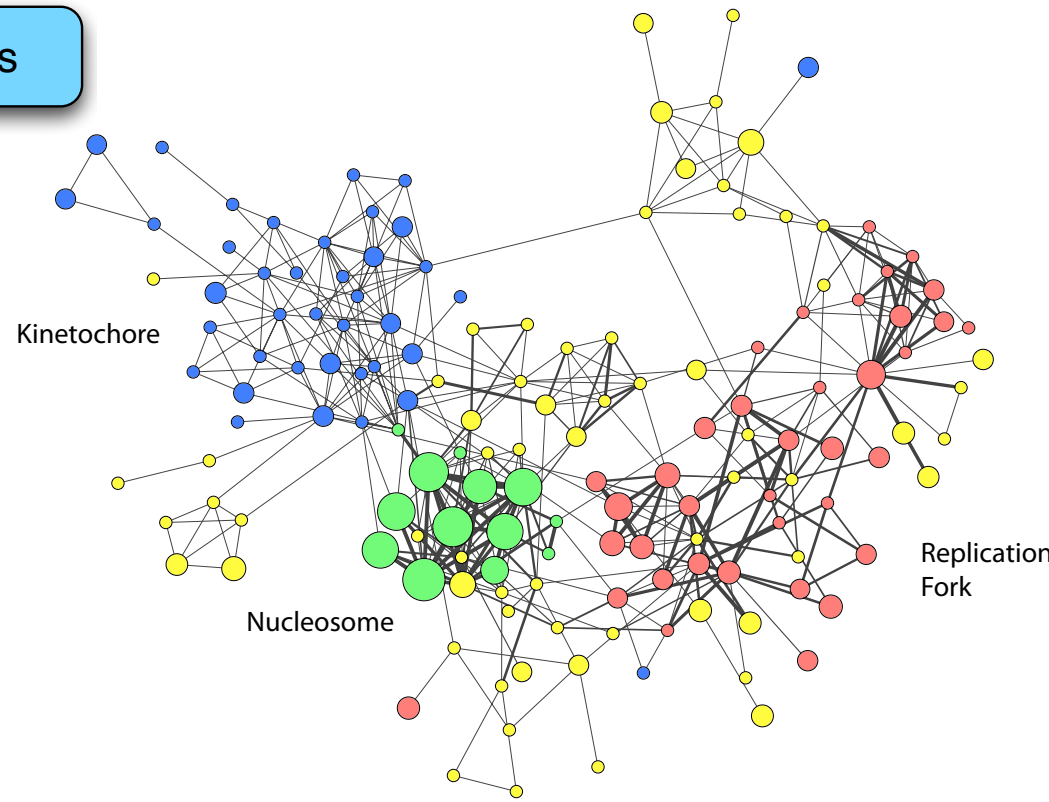
RNA HELICASE ACTIVITY%GO%GO:0003724	28	1.77	0.0041	0.0464386
MRNA SURVEILLANCE PATHWAY%KEGG%HSA03015	82	1.77	0	0.0466167
UBIQUITIN-DEPENDENT DEGRADATION OF CYCLIN D1%REACTOME%REACT_4.1	50	1.77	0.0021	0.0486015
BIOCARTA_CD40_PATHWAY%MSIGDB_C2%BIOCARTA_CD40_PATHWAY	15	1.77	0.0048	0.0483781
IGF1 PATHWAY%PATHWAY INTERACTION DATABASE NCI-NATURE CURATED DATA%IGF1 PATHWAY	29	1.76	0.003	0.0489742
UBIQUITIN-DEPENDENT PROTEIN CATABOLIC PROCESS%GO%GO:0006511	204	1.76	0	0.0488442
PHAGOSOME%KEGG%HSA04145	147	1.76	0	0.0486164
PROTEASOME COMPLEX%GO%GO:0000502	29	1.76	0.007	0.0490215
ANTIGEN PRESENTATION: FOLDING, ASSEMBLY AND PEPTIDE LOADING OF CLASS I MHC%REACTOME%REACT_7	24	1.76	0.0041	0.0505599
ABORTIVE ELONGATION OF HIV-1 TRANSCRIPT IN THE ABSENCE OF TAT%REACTOME%REACT_6261.3	23	1.75	0	0.0529242
DNA DAMAGE RESPONSE, SIGNAL TRANSDUCTION BY P53 CLASS MEDIATOR RESULTING IN CELL CYCLE ARREST%	67	1.75	0	0.052886
REGULATION OF MACROPHAGE ACTIVATION%GO%GO:0043030	11	1.75	0.003	0.0534709
PROTEIN FOLDING%REACTOME%REACT_16952.2	52	1.75	0.002	0.0537717
ENDOPLASMIC RETICULUM UNFOLDED PROTEIN RESPONSE%GO%GO:0030968	73	1.75	0	0.0546052
PROTEIN EXPORT%KEGG%HSA03060	24	1.75	9.75E-04	0.0548699
TRANSCRIPTION INITIATION FROM RNA POLYMERASE II PROMOTER%GO%GO:0006367	64	1.75	0.001	0.0545783
S PHASE%REACTOME%REACT_899.4	110	1.75	0	0.0546003
PROTEASOMAL PROTEIN CATABOLIC PROCESS%GO%GO:0010498	163	1.75	0	0.0550066
ATP-DEPENDENT RNA HELICASE ACTIVITY%GO%GO:0008026	20	1.74	0.0059	0.0556722
ACID-AMINO ACID LIGASE ACTIVITY%GO%GO:0016881	217	1.74	0	0.0560217
GO%GO:0072474	67	1.74	0.002	0.0565978
GO%GO:0035966	107	1.74	0	0.0562957
GO%GO:0072413	67	1.74	9.81E-04	0.05761
BIOCARTA_IL4_PATHWAY%MSIGDB_C2%BIOCARTA_IL4_PATHWAY	11	1.74	0.0082	0.0581508
ASSOCIATION OF TRIC CCT WITH TARGET PROTEINS DURING BIOSYNTHESIS%REACTOME%REACT_16907.2	28	1.74	0.0039	0.0581298
UBIQUITIN-DEPENDENT DEGRADATION OF CYCLIN D%REACTOME%REACT_938.4	50	1.74	0.0029	0.057876
MODIFICATION-DEPENDENT PROTEIN CATABOLIC PROCESS%GO%GO:0019941	207	1.74	0	0.0576579
TRANSLATION INITIATION COMPLEX FORMATION%REACTOME%REACT_1979.1	55	1.74	0.0021	0.0575181
GO%GO:0001906	13	1.74	0.0117	0.0572877
G1 S TRANSITION%REACTOME%REACT_1783.2	107	1.74	0	0.0572618
GO%GO:0034620	73	1.73	0.0021	0.0576606
SIGNALING BY NOTCH%REACTOME%REACT_299.2	19	1.73	0.0069	0.0578565
RESPONSE TO UNFOLDED PROTEIN%GO%GO:0006986	102	1.73	0	0.0583864
SIGNAL TRANSDUCTION INVOLVED IN G1 S TRANSITION CHECKPOINT%GO%GO:0072404	68	1.73	0.002	0.0582213
GO%GO:0072431	67	1.73	0	0.058551
BIOCARTA_PROTEASOME_PATHWAY%MSIGDB_C2%BIOCARTA_PROTEASOME_PATHWAY	19	1.73	0.0099	0.0586655
HOST INTERACTIONS OF HIV FACTORS%REACTOME%REACT_6288.4	117	1.73	0	0.0586888
AUTOPHAGIC VACUOLE ASSEMBLY%GO%GO:0000045	13	1.73	0.0122	0.0588271
CYCLIN A:CDK2-ASSOCIATED EVENTS AT S PHASE ENTRY%REACTOME%REACT_9029.2	66	1.73	0	0.0610099

TABULAR FORMAT

→ NETWORK VISUALIZATION

Network Visualization





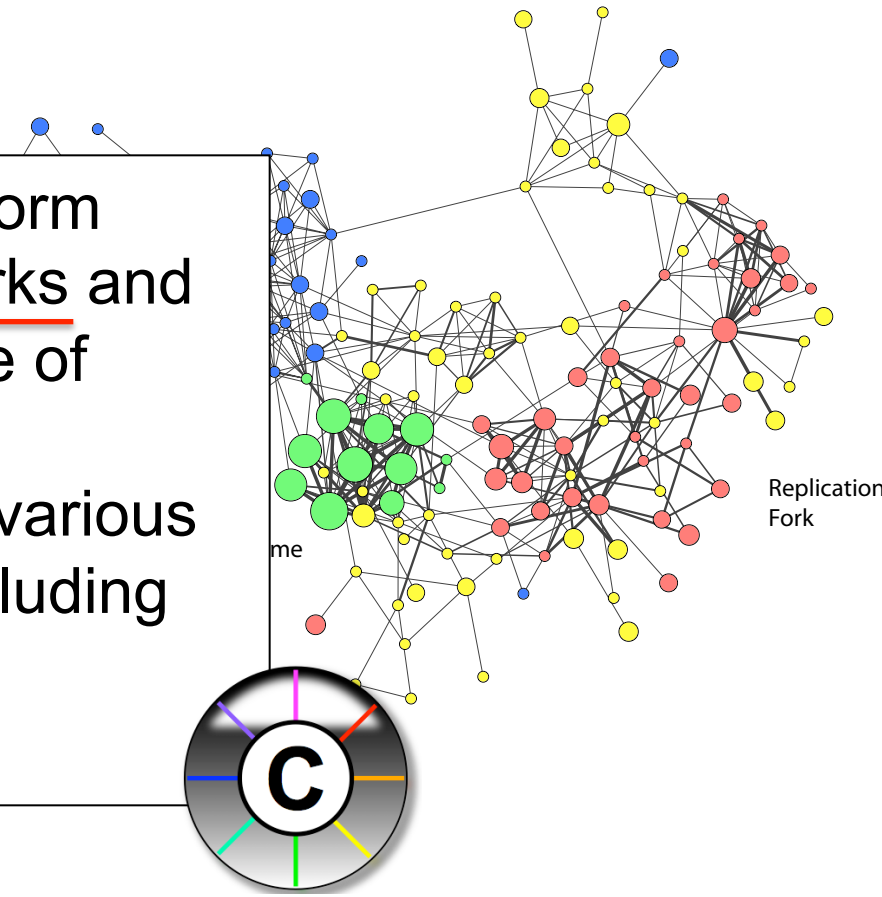
Cytoscape

www.cytoscape.org

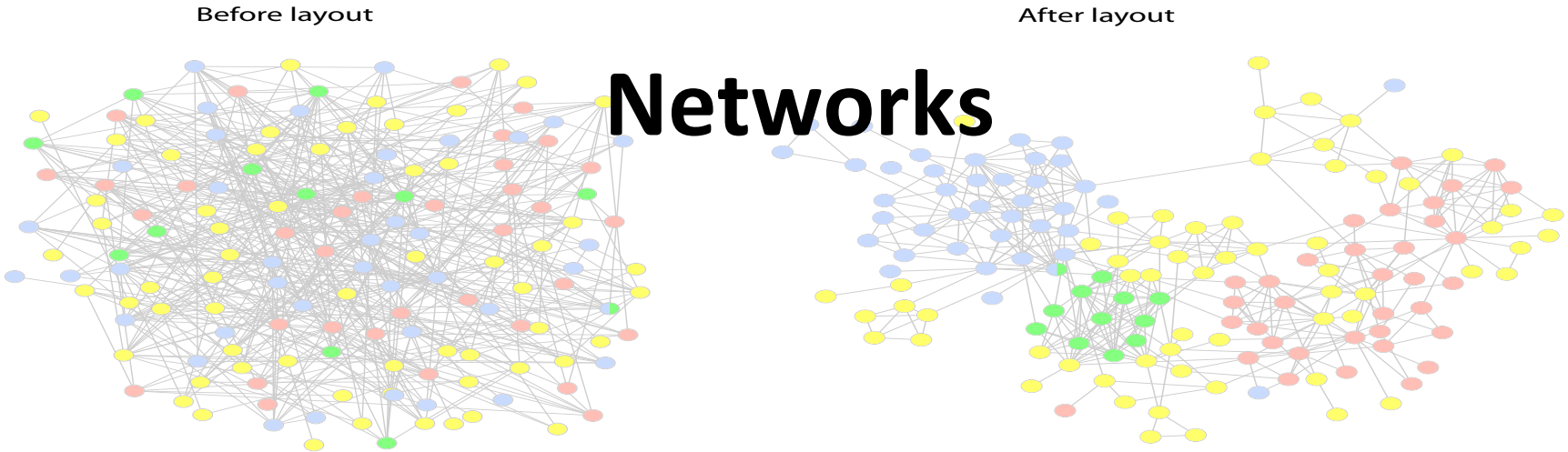
Network Visualization

Cytoscape is

- an open source software platform
- for visualizing complex networks and integrating these with any type of attribute data.
- a lot of apps are available for various kinds of problem domains, including bioinformatics, social network analysis, and semantic web.



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 basics: 1/2
Nodes and Edges

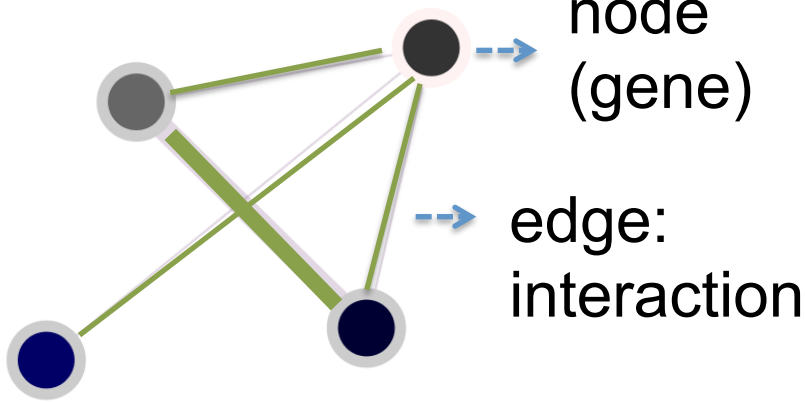
A simple mapping
- one compound/node, one interaction/edge

A more realistic mapping
- Cell localization, cell cycle, cell type, taxonomy
- Only represent physiologically relevant interaction networks

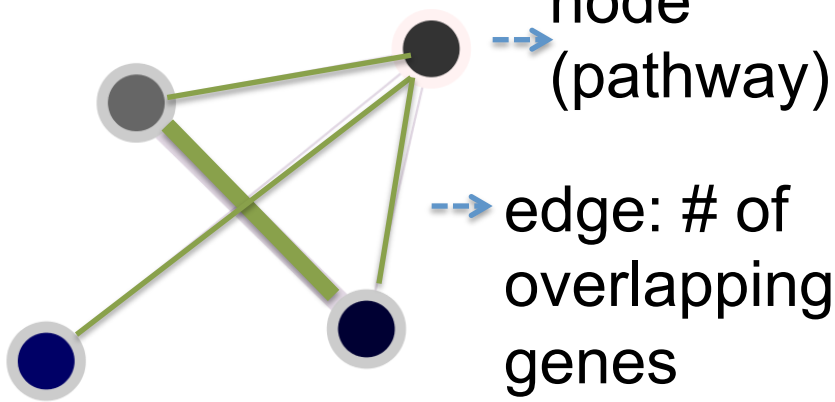
Edges can represent other relationships

Critical: understand what nodes and edges mean

gene-gene network

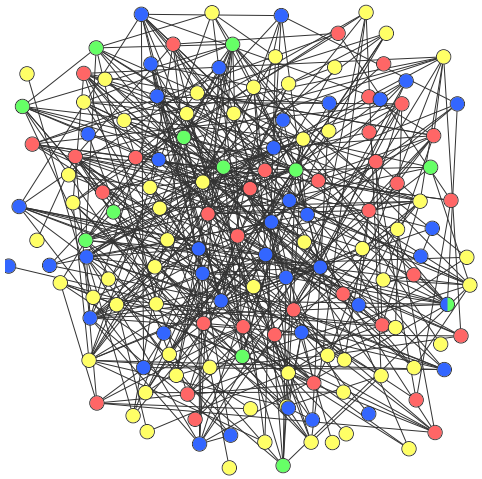


pathway network

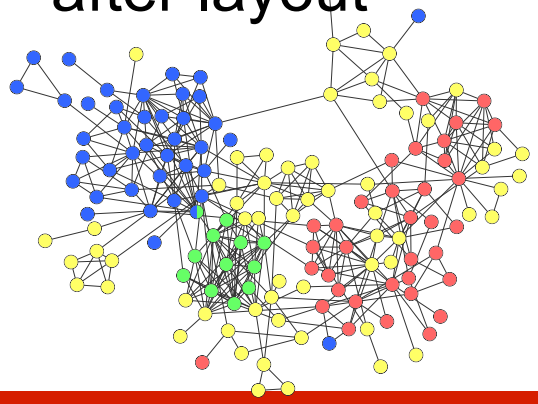


Network basics 2/2: Automatic network layout

before layout



after layout



- Force-directed layout: 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

will be used during lab

Introduction to Cytoscape (2.8.3)

save your session

Control panel

Results panel

Data panel

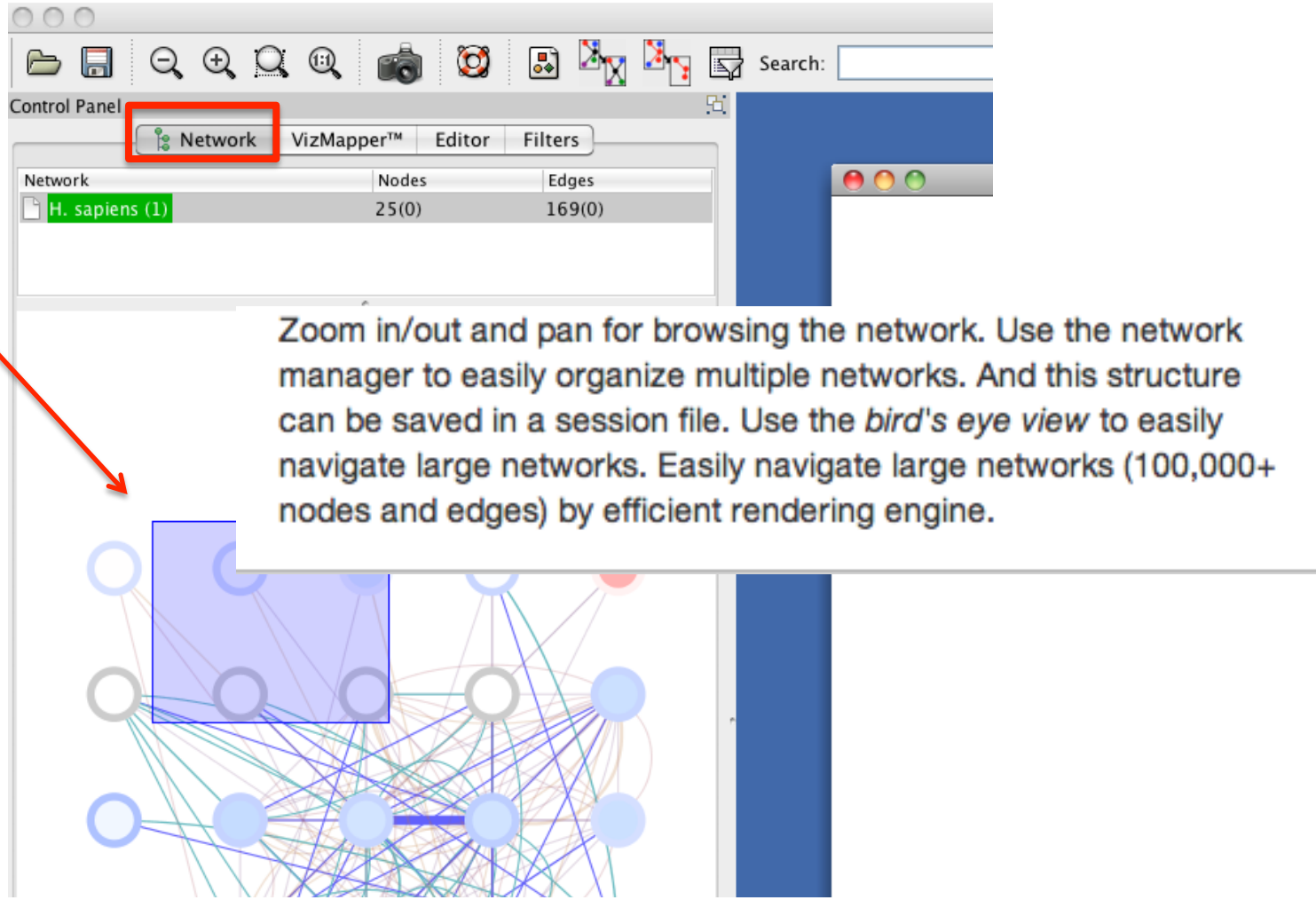
Session File

You can save your work by a single click. All of settings, data files, and visualizations are packed in a session file. It is called *Cytoscape Session (.cys) file*. Cytoscape Session file includes networks, attributes (for node/edge/network), Desktop states (selected/hidden nodes and edges, window sizes), Properties, some plugin states, and Visual Styles.

will be used during lab

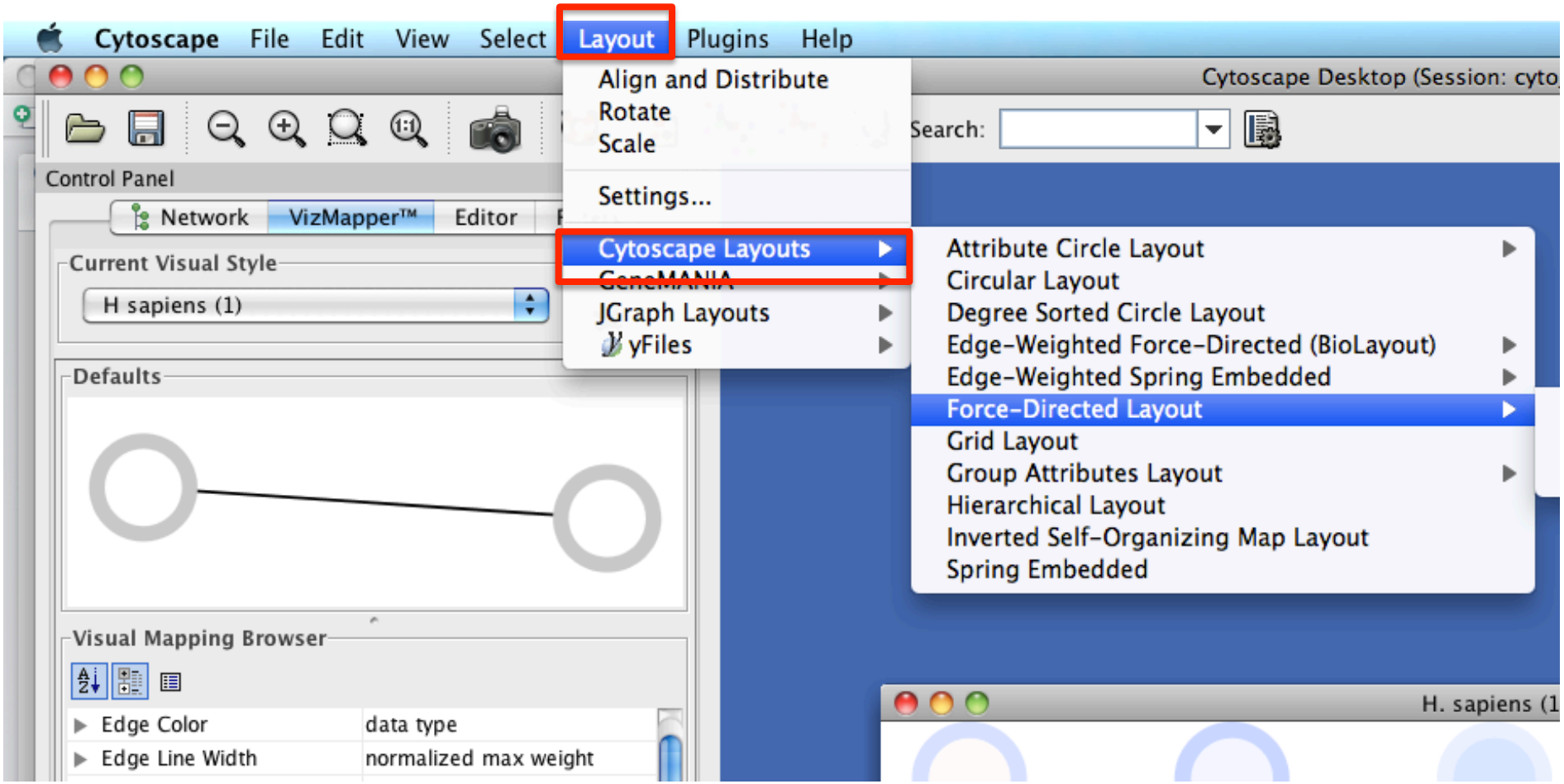
Navigate through the network (2.8.3)

move the blue square to naviage through the network



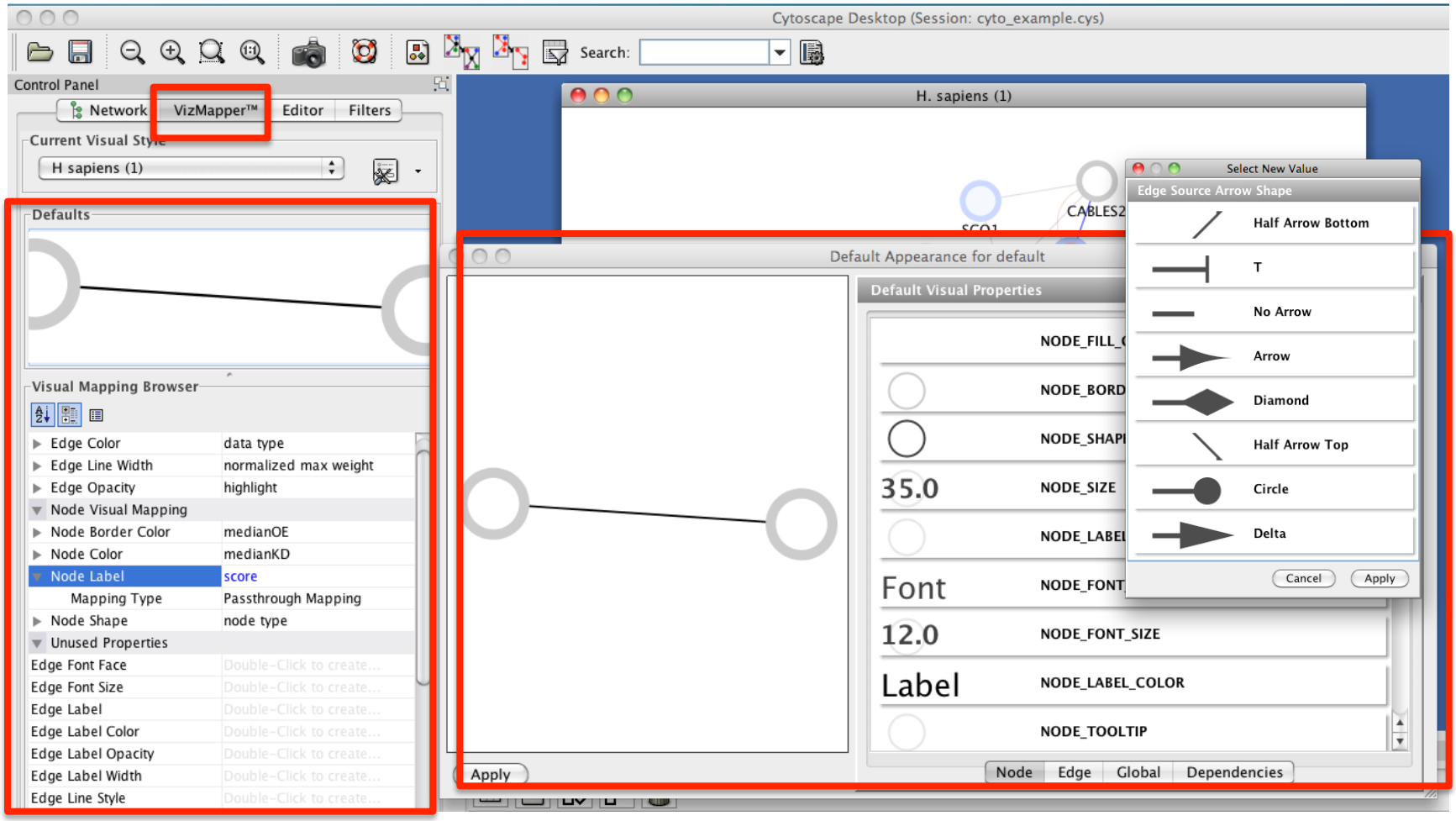
will be used during lab

Cytoscape layout (2.8.3)



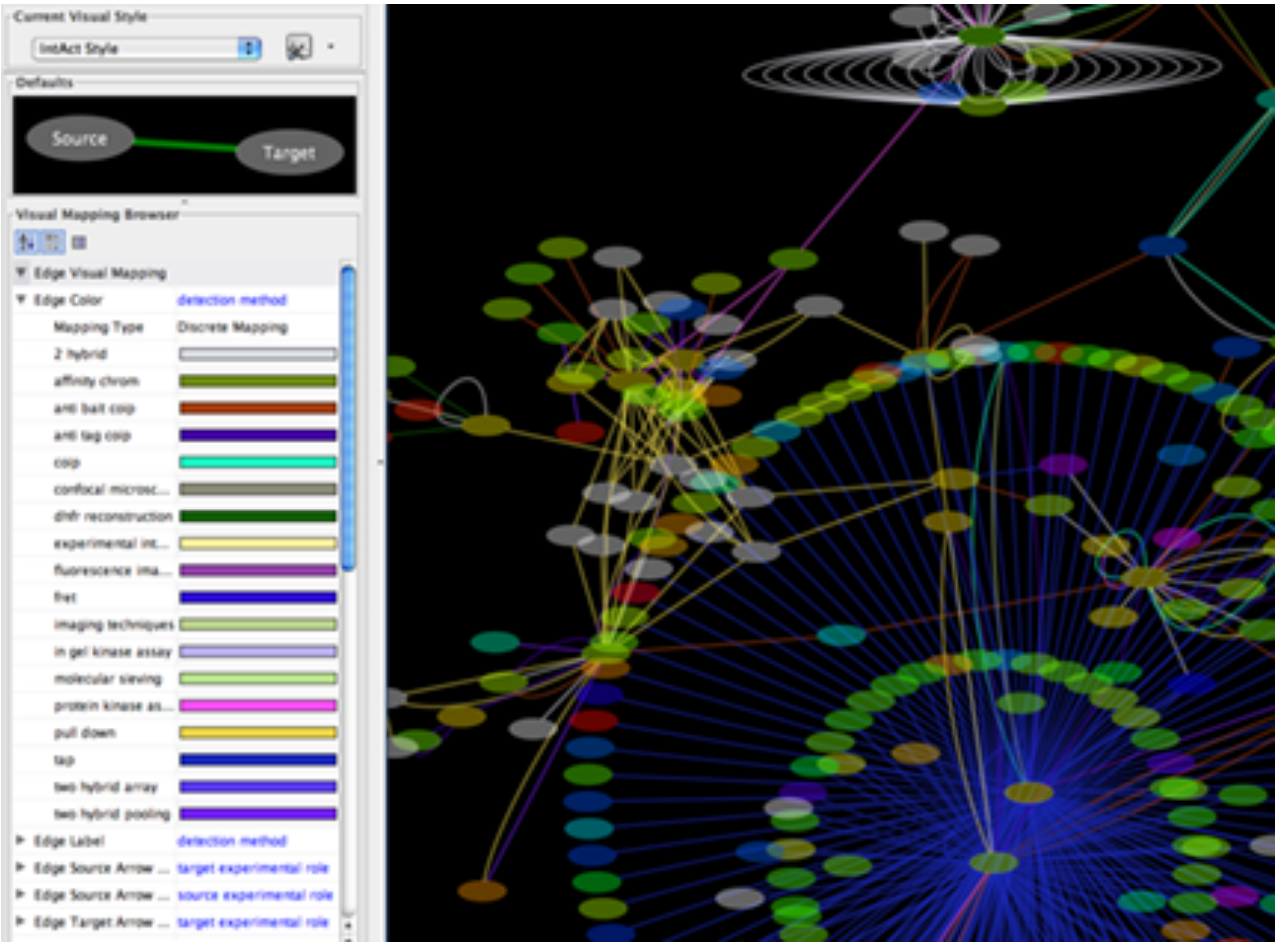
will be used during lab

Visual Features (2.8.3)



will be used during lab

Visual Features: customize network data



What Have We Learned?

- Networks are useful for seeing relationships in large data sets
- Important to understand what the nodes and edges mean
- Automatic layout is required to visualize networks
- Visual attributes enable multiple types of data to be shown at once – useful to see their relationships

Example of Cytoscape plugins

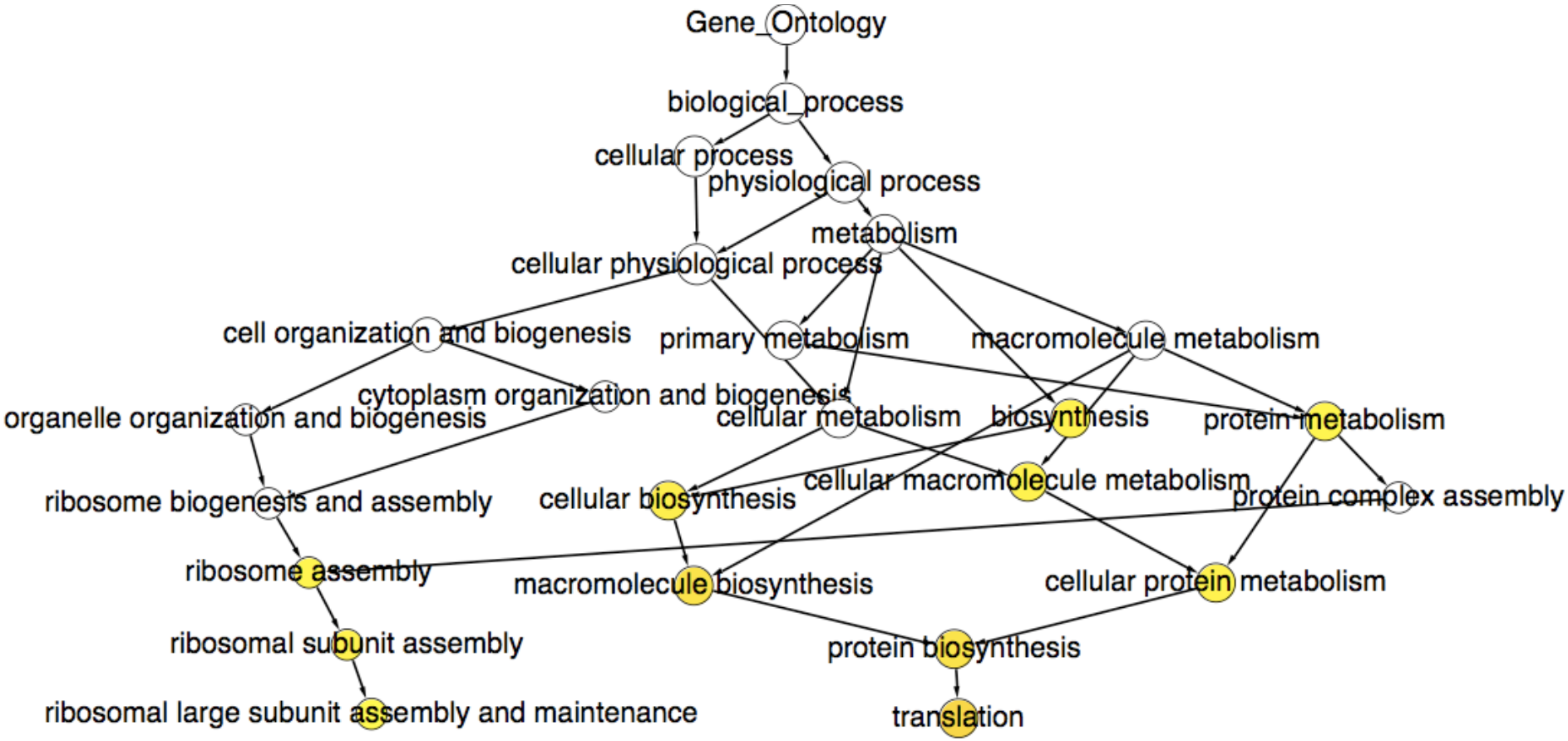
BiNGO plugin

- 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)



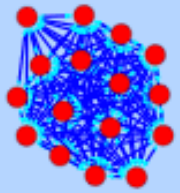
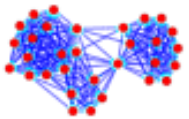
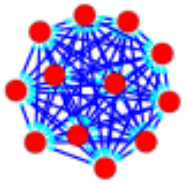
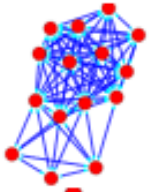
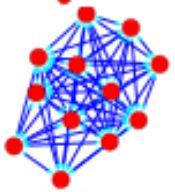
Maere, S., Heymans, K. and Kuiper, M
Bioinformatics 21, 3448-3449, 2005

Network Clustering

- Clusters in a protein-protein interaction network have been shown to represent protein complexes and parts of pathways
- Clusters in a protein similarity network represent protein families
- Network clustering is available through the ClusterMaker Cytoscape plugin

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

Cytoscape 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

Cytoscape 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

Cytoscape 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

Cytoscape 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)

Cytoscape 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
    - Since version 2.7, Cytoscape does a much better job at "guessing" good default memory sizes than previous versions

Cytoscape 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

for your information

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

Slides from

Gary Bader

Quaid Morris

Lincoln Stein

Veronique Voisin

We are on a Coffee Break & Networking Session