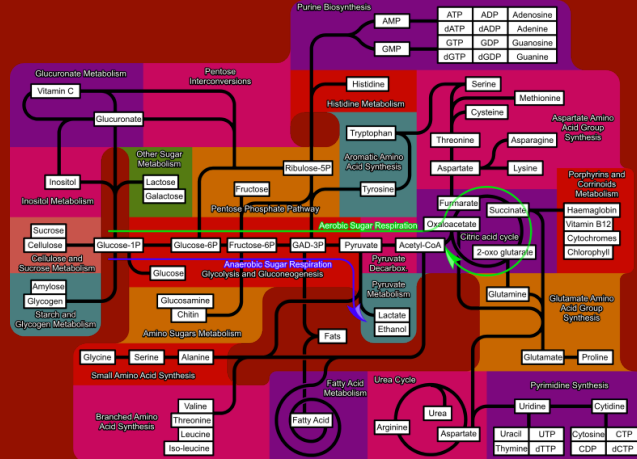
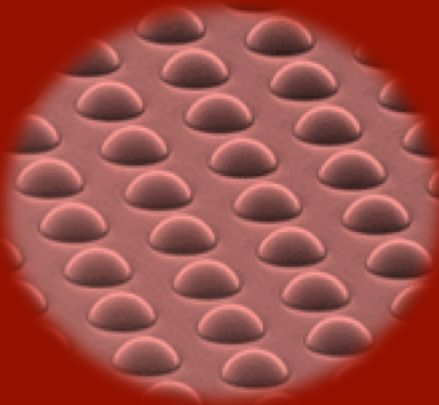


http://en.wikipedia.org/wiki/File:Metabolism_790px.png

Pathway analysis of genomics data part 1

Veronique Voisin

Bader Lab

Bioinformatics core, OICR Cancer Stem Cell

Toronto Western Hospital, Dec 5, 2011





COURSE OUTLINE

PART 1 (lecture)

- 1.1 What is Pathway and Network analysis ?
- 1.2 Brief description of a few databases
- 1.3 How to manipulate gene identifiers
- 1.4 Presentation of two enrichment analysis techniques: *GSEA* and *DAVID*



COURSE OUTLINE

PART 2 (hands-on training)

2.1 use The Synergizer tool to convert gene-lists

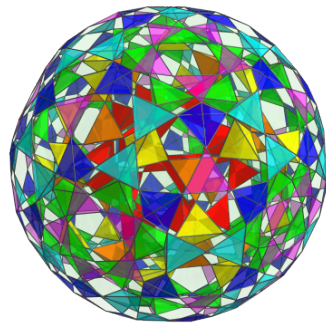
2.2 DAVID

2.3 GSEA

2.4 create Enrichment Maps

2.5 open cytoscape and navigate through one map

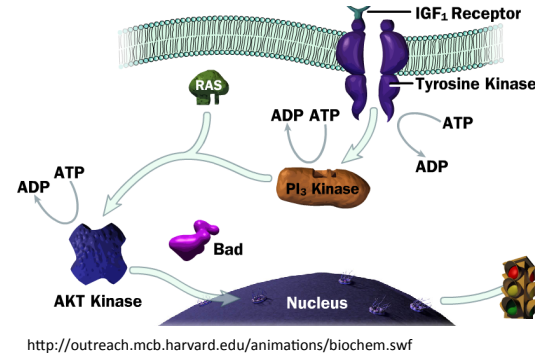
1.1 PATHWAY ANALYSIS



a cell

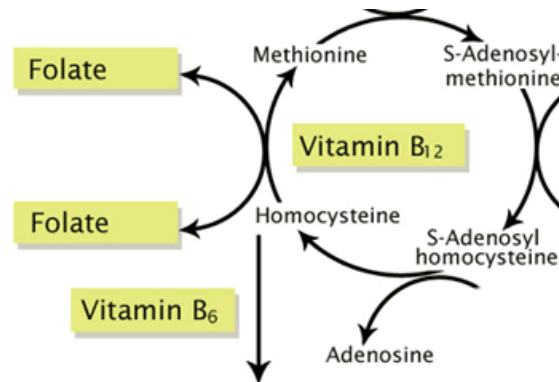
?

?



signaling pathway

active
or
inactive
?



metabolic pathway

active
or
inactive
?

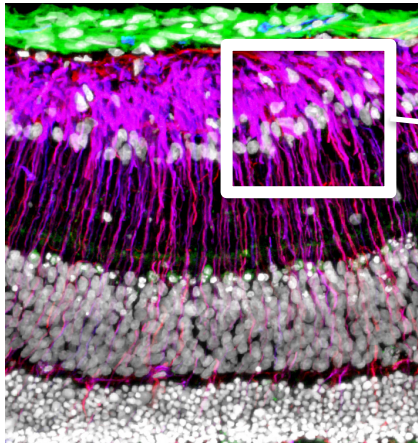
Pathway analysis is an alternative to traditional analysis

- gene by gene basis
- requires literature searching
- time-consuming



To perform pathway and network analysis, you first need a gene list.

Gene expression data (arrays, RNA seq)

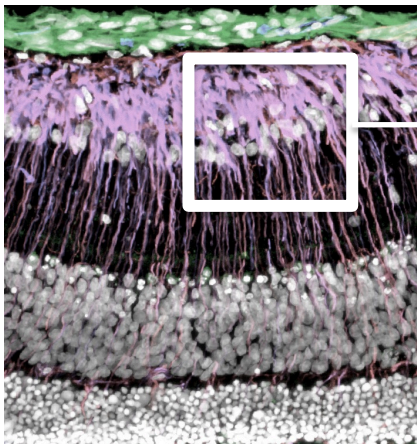


RNA



<http://www.neuroscience.cam.ac.uk/directory/profile.php?thomasvjohanson>

Wild-type



RNA



mutant

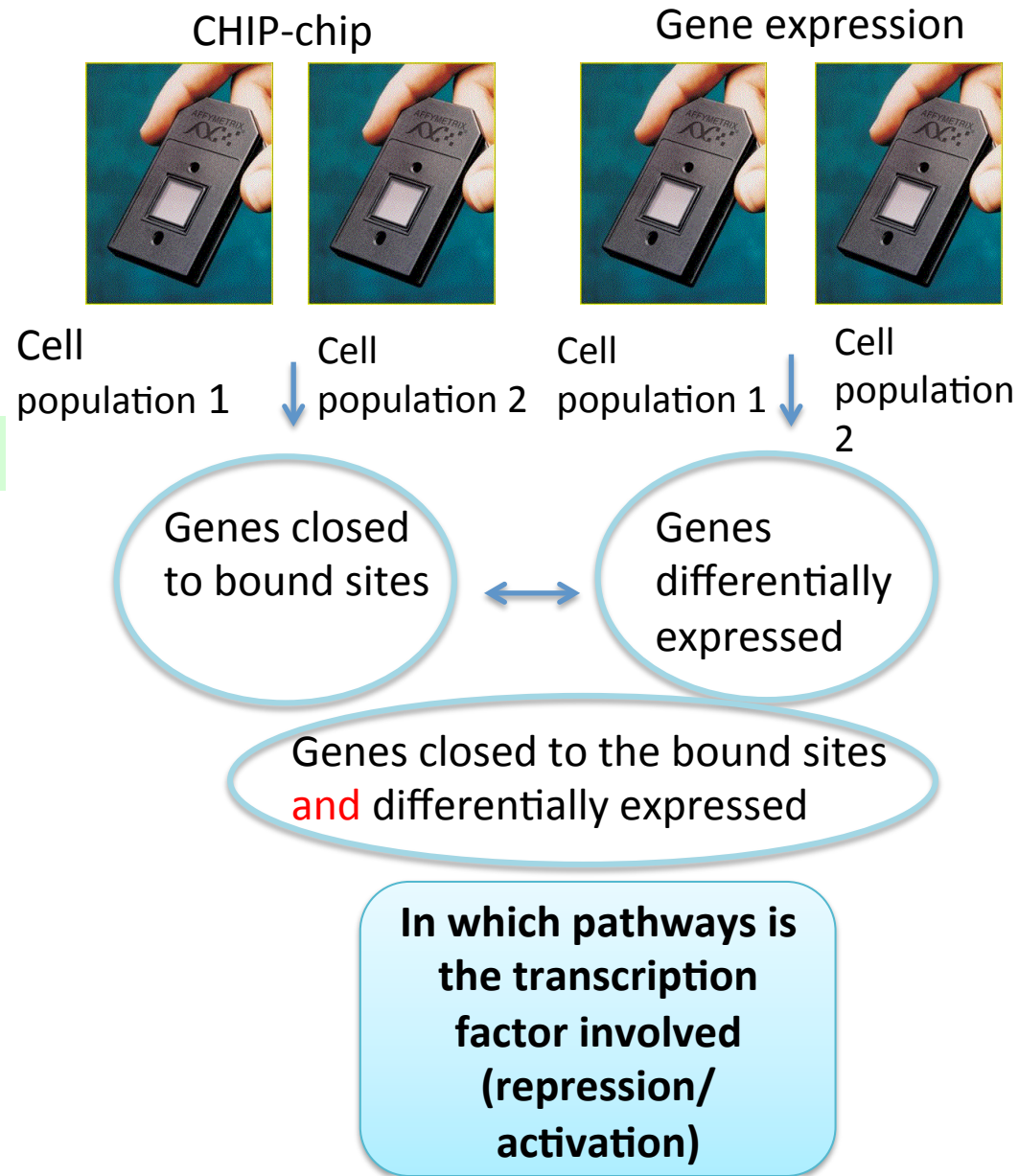
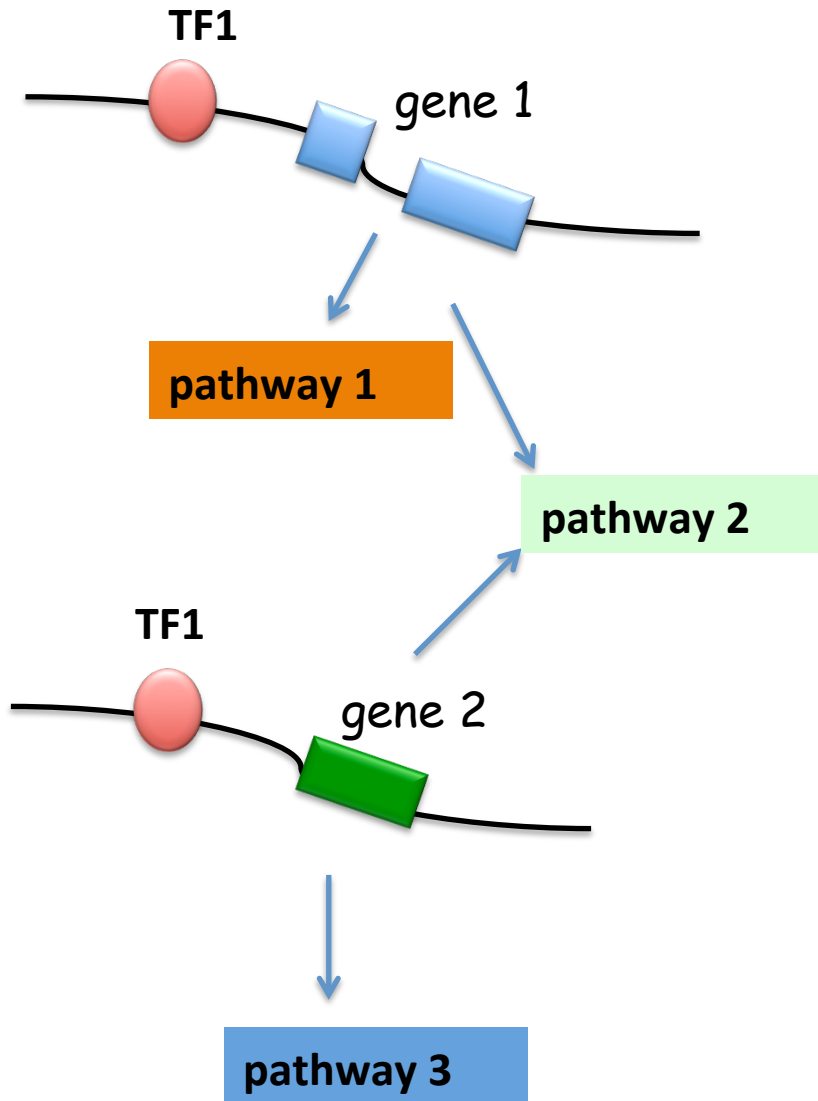
List of genes that are differentially expressed between the 2 conditions

4IN_1661266	3119	HLA-DQB1	10.03470934	8.922700486	-1.112008851	-12.448491
4IN_1770338	4071	TM4SF1	8.120788593	9.988782534	1.867993941	12.358159
4IN_1738742	5327	PLAT	6.970287144	8.037981147	1.067694003	12.222051
4IN_1688580	820	CAMP	8.21099204	10.19708797	1.986095928	12.036168
4IN_2157099	8900	CCNA1	8.26682593	10.12988686	1.86306926	11.984898
4IN_1659645	10394	PRG3	10.5015173	13.61803116	3.116513854	11.971558
4IN_2134974	23682	RAB38	8.046506351	8.987338794	0.940832443	1.902203
4IN_1725193	3485	IGFBP	10.4871578	11.4001185	1.552854081	1.455347
4IN_1692223	3485	IGFBP	10.4871578	11.4001185	1.552854081	1.455347
4IN_1784272	3485	IGFBP	10.4871578	11.4001185	1.552854081	1.455347
4IN_1777519	3485	IGFBP	10.4871578	11.4001185	1.552854081	1.455347
4IN_1668092	3485	IGFBP	10.4871578	11.4001185	1.552854081	1.455347
4IN_1808405	3485	IGFBP	10.4871578	11.4001185	1.552854081	1.455347
4IN_3214389	1E+08	LOC100133583	12.17455888	11.16680984	-1.007749043	0.789857
4IN_1814015	81753	TLR10	7.80909033	6.952626127	-0.856464203	0.739456
4IN_1715991	81753	TLR10	7.80909033	6.952626127	-0.856464203	0.739456
4IN_3249667	81753	TLR10	7.80909033	6.952626127	-0.856464203	0.739456
4IN_1719599	1306	COL15A1	6.40439199	7.736579374	1.332187384	0.168232
4IN_1790692	909	CD1A	12.01374084	10.73998081	-1.273760031	-10.041818
4IN_1723520	255231	MCOLN2	10.22567089	9.137018542	-1.08865235	-9.9938267
4IN_1744968	7881	KCNAB1	7.050821049	8.153666284	1.102845234	9.9406443
4IN_1837428	NA		9.720692155	10.37376104	1.153068881	9.9287522
4IN_1714335						9.906329
4IN_3200140						9.8611335
4IN_1732269						9.8422557
4IN_1784706						9.6937495
4IN_1656310						9.6886052
4IN_2324202						9.6861555
4IN_1805410						9.6816224
4IN_1737314						9.6585706
4IN_1704870						9.6513091
4IN_1687306						9.6223065
4IN_2342579						9.508609
4IN_2376204						9.4240632
4IN_1719905						9.4117547
4IN_1749131						9.3568685
4IN_1738725						9.3186401
4IN_1700024						9.298592
4IN_1723004						9.2680378
4IN_1677920						9.2076905
4IN_1672097						9.1697852
4IN_1729487						9.086912
4IN_1663793	116449	MIST	7.832837381	7.084649469	-0.748187913	-9.0802038
4IN_1787526	84281	MGC13057	7.836453975	8.690609657	0.854155682	9.0308169
4IN_2148668	1102	RCBTB2	10.92310685	10.16047995	-0.7626269	-9.0157949
4IN_1749070	3115	HLA-DPB1	10.23399637	9.174405097	-1.059591269	-9.011587
4IN_2102670	2624	GATA2	8.477088879	9.451979962	0.974891083	8.9792179

How many pathways are dysregulated between the 2 conditions?

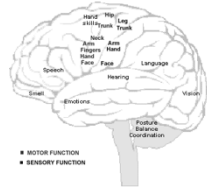
Enrichment analysis works with any gene lists.

CHIP chip + gene expression

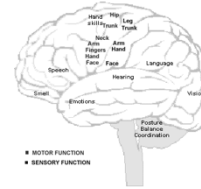


CNA CNV SNPs

DNA obtained
from blood or
buccal swabs



Patient
(autism)



reference



SNPs arrays,
Deep sequencing,
arrayCGH,...

Regions of the genome are altered when comparing the patient with the reference



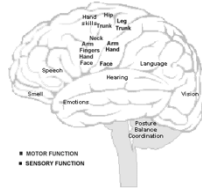
list of genes that
are within the
altered regions

**Do these genes
belong to same
functional
pathways?**

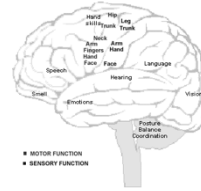


Pathway and network analysis

CNA CNV SNPs



patient (brain cancer)



patient (normal tissue)

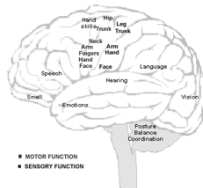
SNPs arrays,

List1: Non silent somatic mutations within genes

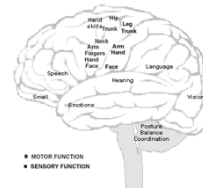
Pathway and network analysis

List 2: Genes differentially expressed
in cancer versus normal

gene expression



patient (brain cancer)

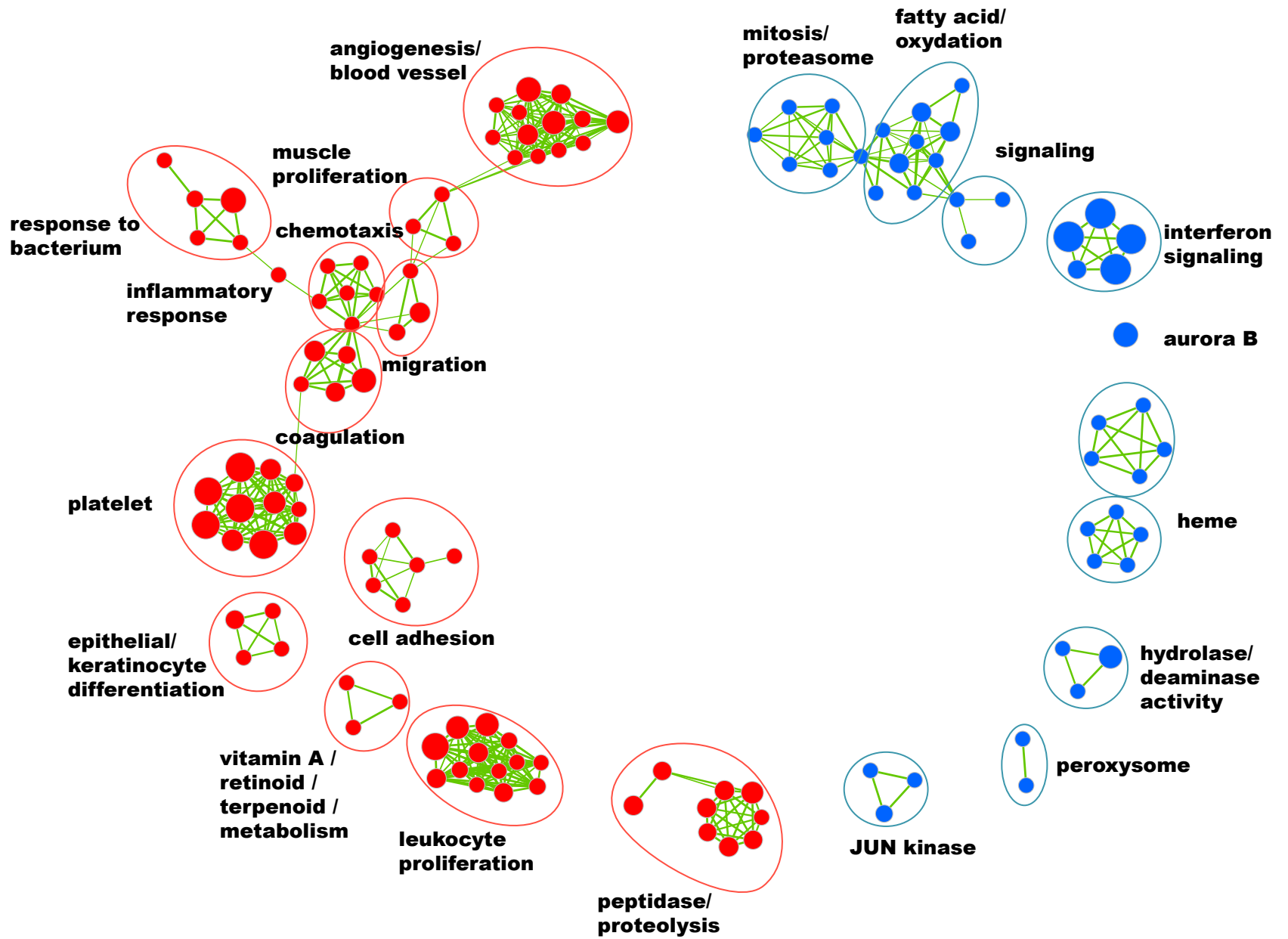


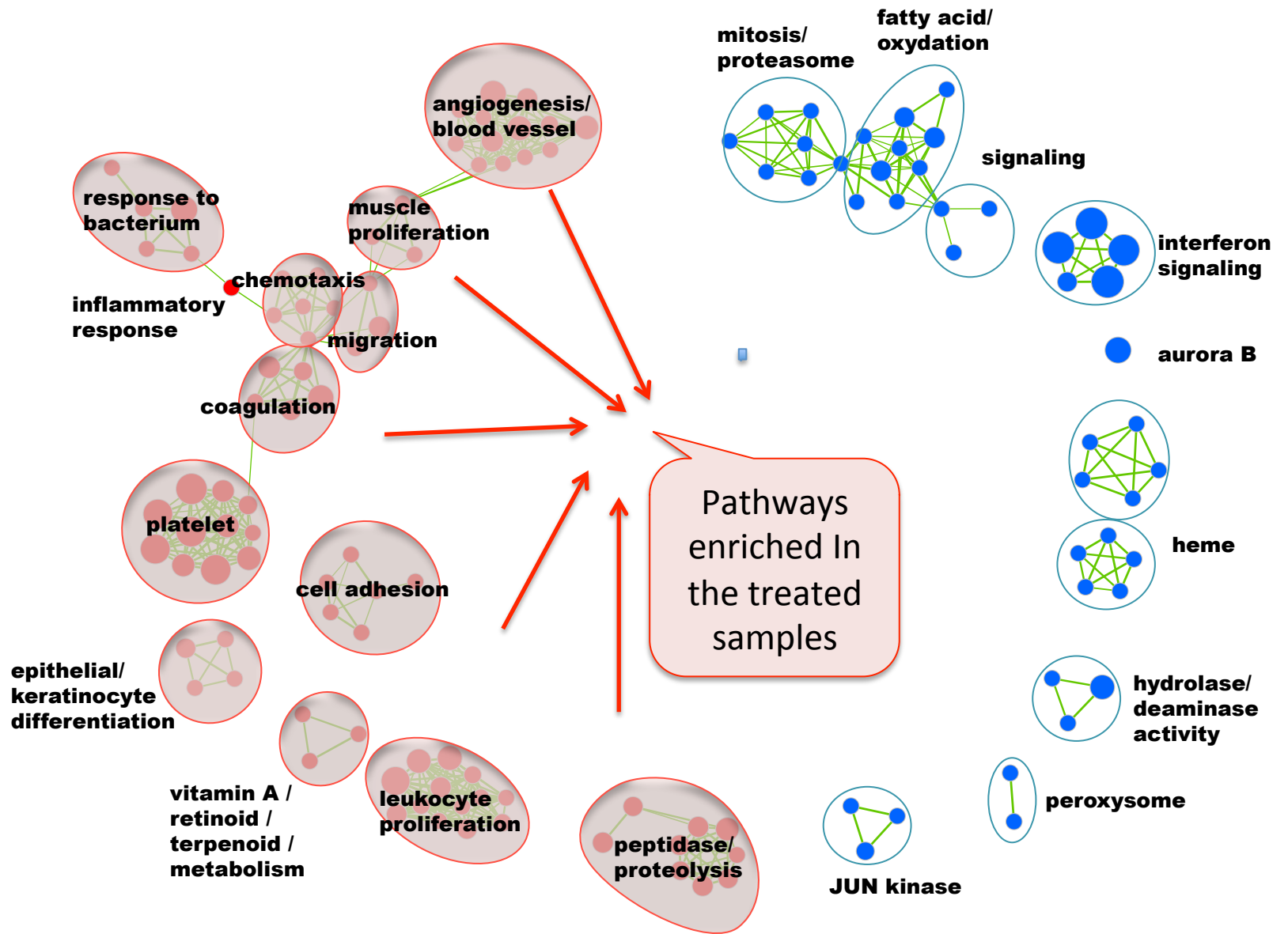
patient (normal tissue)

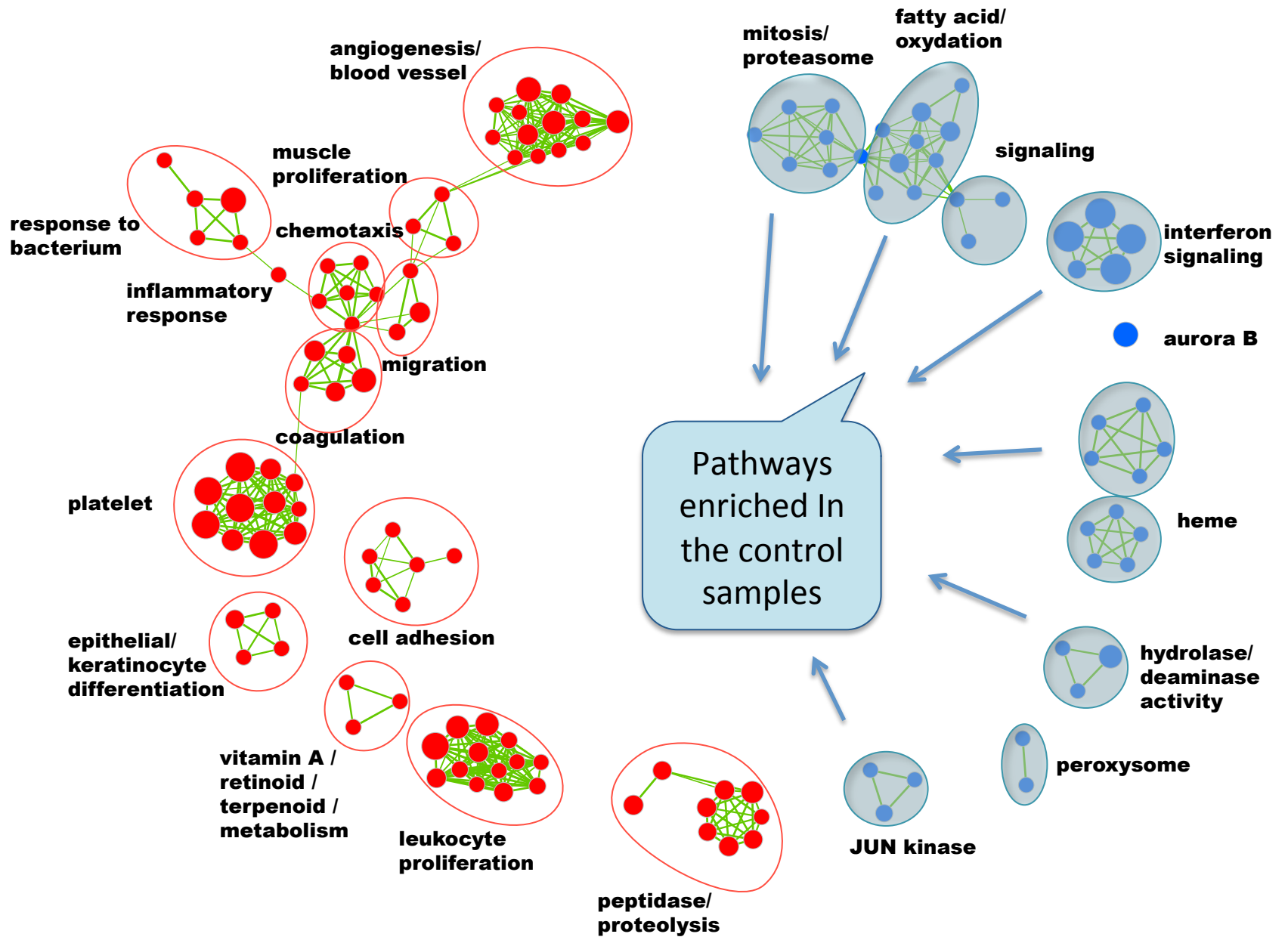
**Can we correlate at
the pathway level
the initiation
events (somatic
non silent
mutations) and the
resulting
phenotype (gene
expression)**

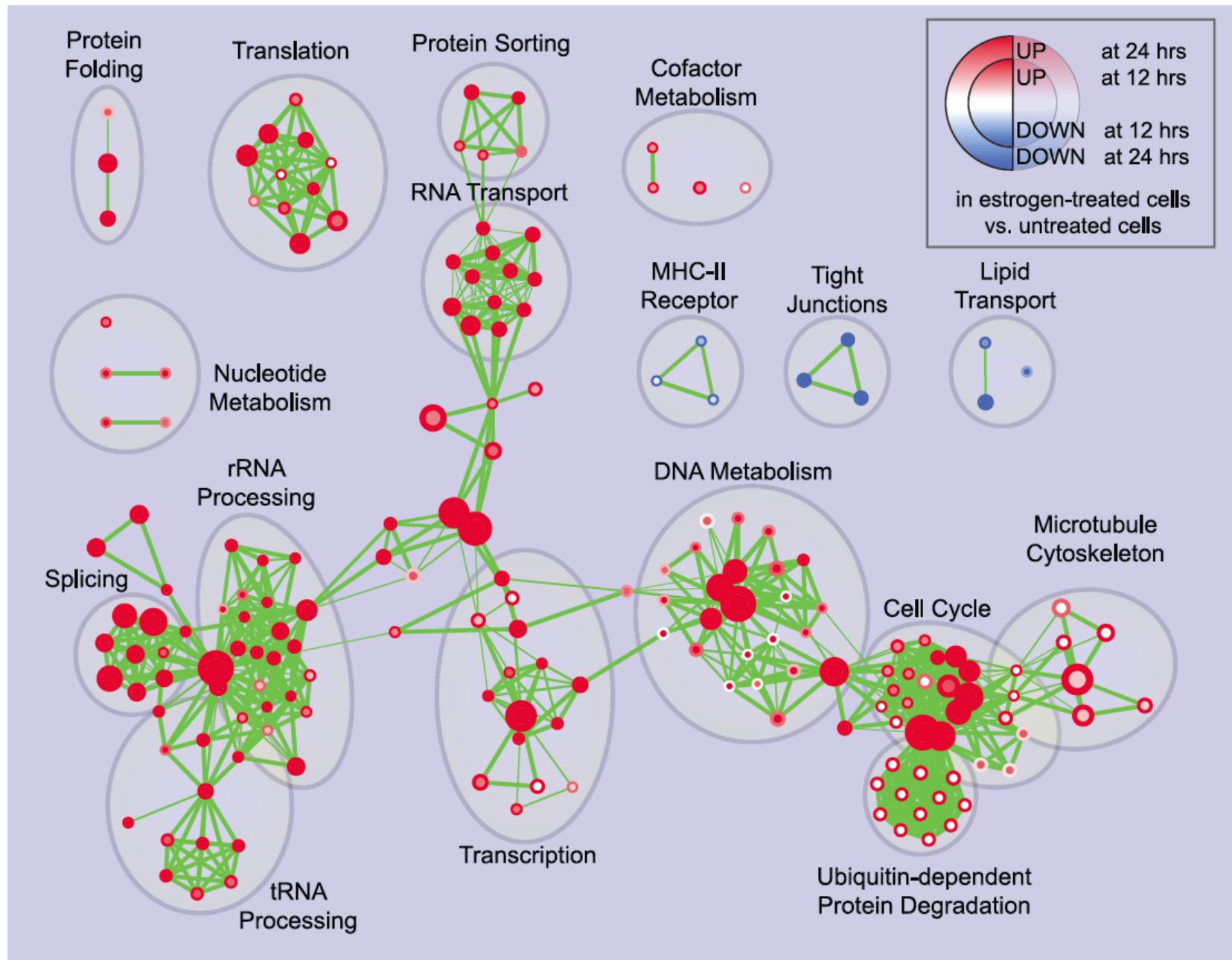
etc....

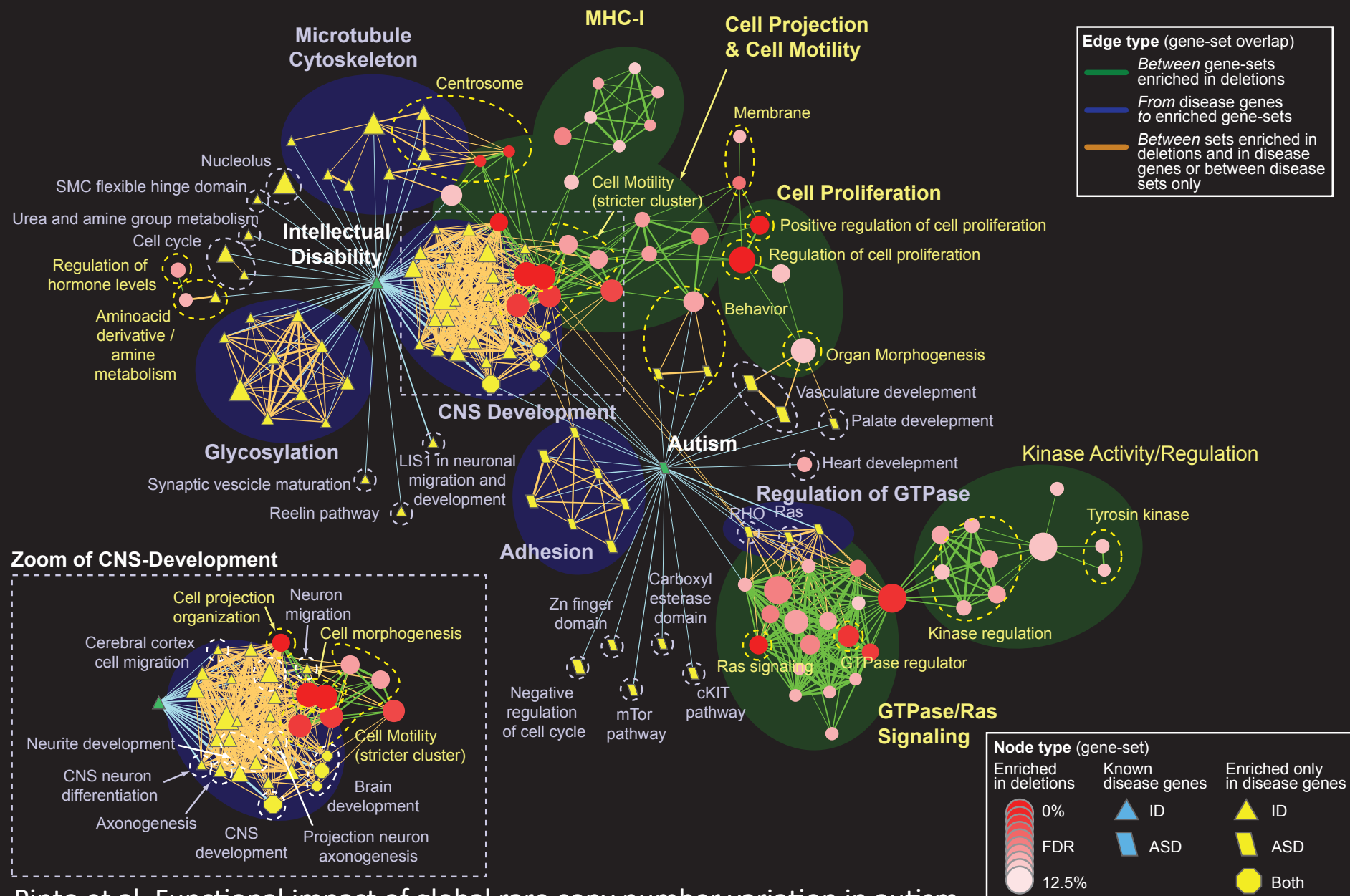
Some examples



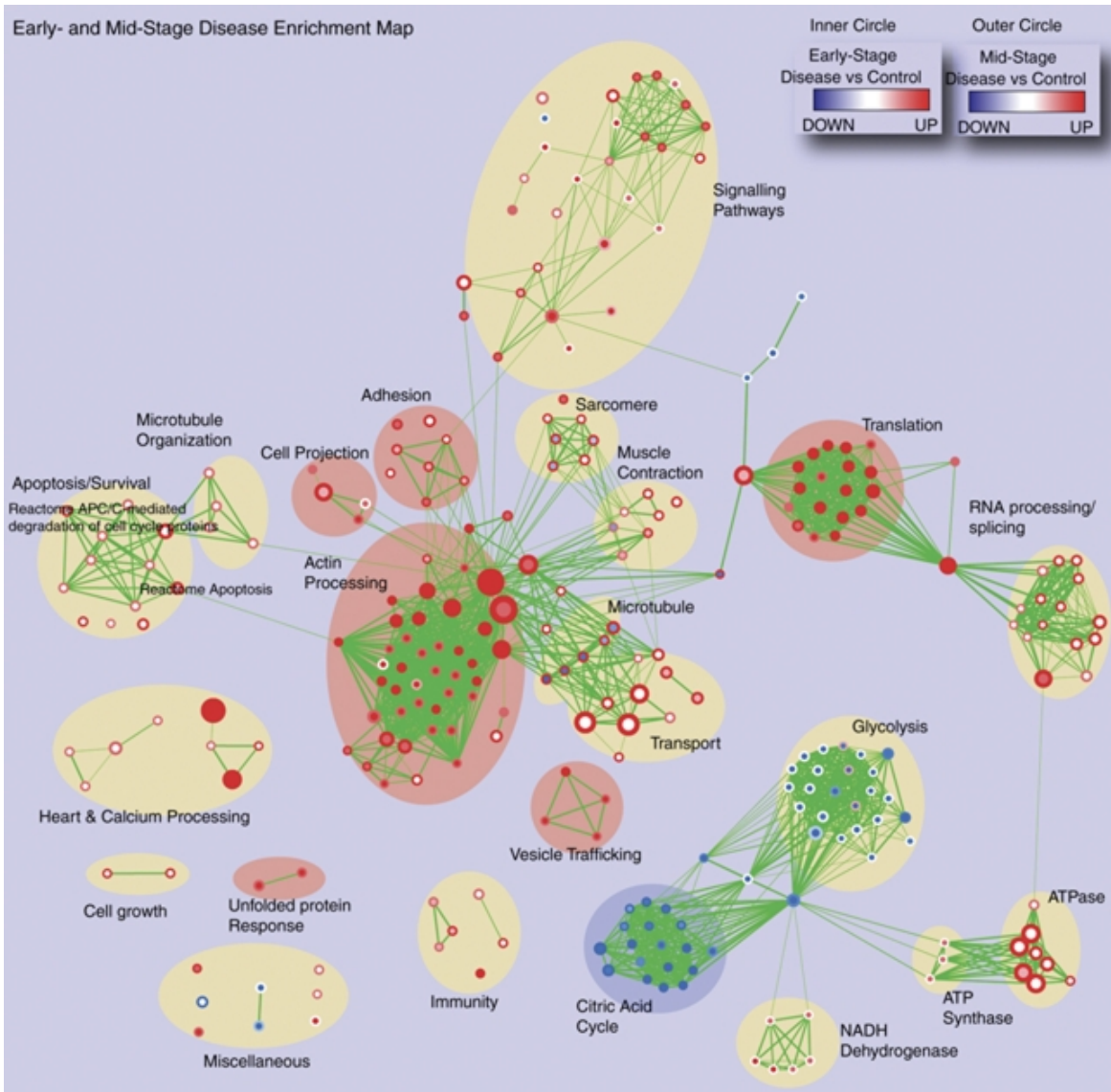






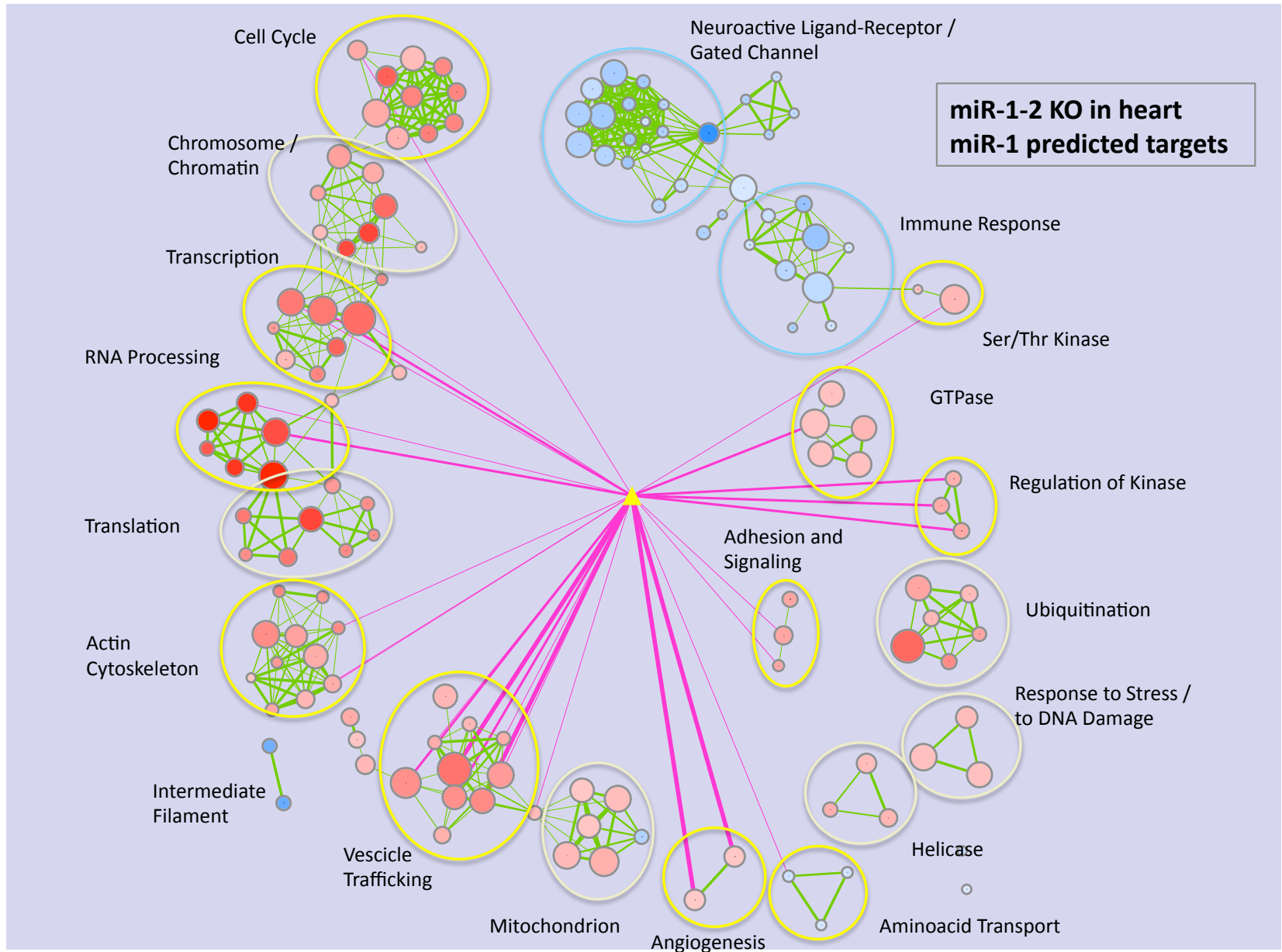


Pinto et al. Functional impact of global rare copy number variation in autism spectrum disorders. Nature. 2010 Jun 9.



understanding of the progressive alterations associated with functional decline in dilated cardiomyopathy (in mice)

microRNA knock-out



Comparison of the gene list with
databases containing annotated
pathways

DATABASES containing annotated pathways

Gene Ontology



MSigDB-c2



Netpath

NetPath

KEGG



REACTOME



NCI



BIOCARTA



HumanCyc



Gene Ontology (GO)



- <http://www.geneontology.org>
- Largest database
- 41.007 gene products (proteins) annotated for human
- Updated every 3 months
- Organism independent ?
- Covers many model organisms (Homo Sapiens, Mus musculus, Danio Rerio...)

- 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](#)...

- Some GO annotations created automatically

- 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...

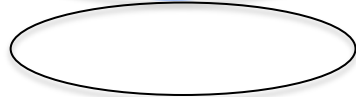
- Some GO annotations created automatically



Molecular function



Molecular component



Biological process

GO Evidence Codes: information about how the annotation was created

alpha-1,6-mannosyltransferase activity

ALG12	GO:0000009
TNP1	GO:0000012
PSMD4	GO:0000209
COX6B1	GO:0005743
MEF2C	GO:0000002

IC
IDA
TAS
TAS
IEA

Dol-P-Man:Man(7)GlcNAc(2)-PP-Dol alpha-1,6-mannosyltransferase
Spermatid nuclear transition protein 1
26S proteasome non-ATPase regulatory subunit 4
Cytochrome c oxidase subunit 6B1
Uncharacterized protein

IC: inferred by curator

IDA: Inferred from direct assay

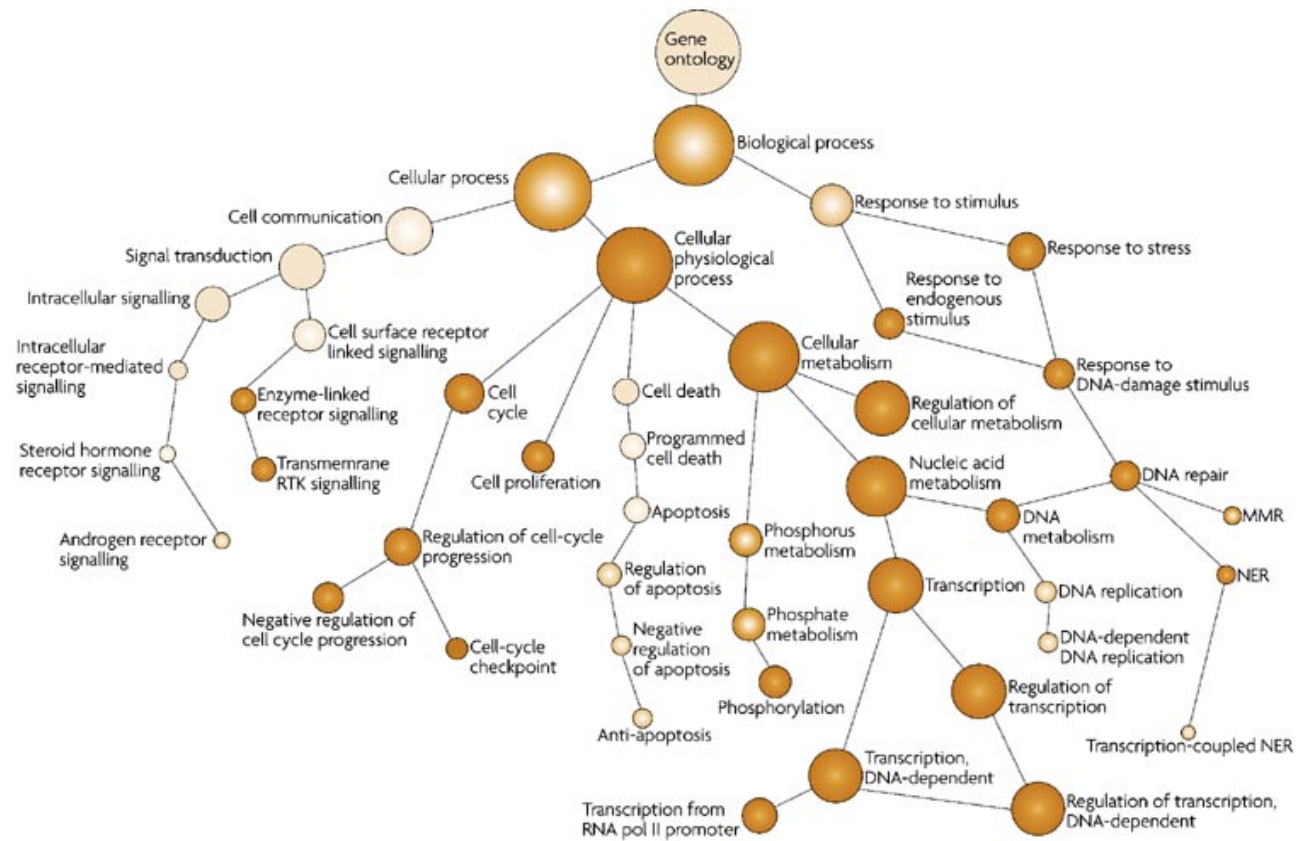
TAS: Traceable Author Statement

IEA: Inferred by electronic annotations

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

GO covers 3 domains:

- **cellular component:** the parts of a cell or its extracellular environment;
- **molecular function:** the elemental activities of a gene product at the molecular level, such as binding or catalysis;
- **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.



Nature Reviews | Cancer

Hu *et al.* *Nature Reviews Cancer* 7, 23–34 (January 2007) | doi:10.1038 / nrc2036



KEGG: Kyoto Encyclopedia of Genes and Genomes



- Pathway maps for metabolism and other cellular processes, as well as human diseases; manually created from published materials
- 413 pathways
- Homo sapiens and Mus musculus model organism (Mus musculus inferred from Homo sapiens).

Example of 2 categories of pathway:

Cellular Process

Cell Growth and Death

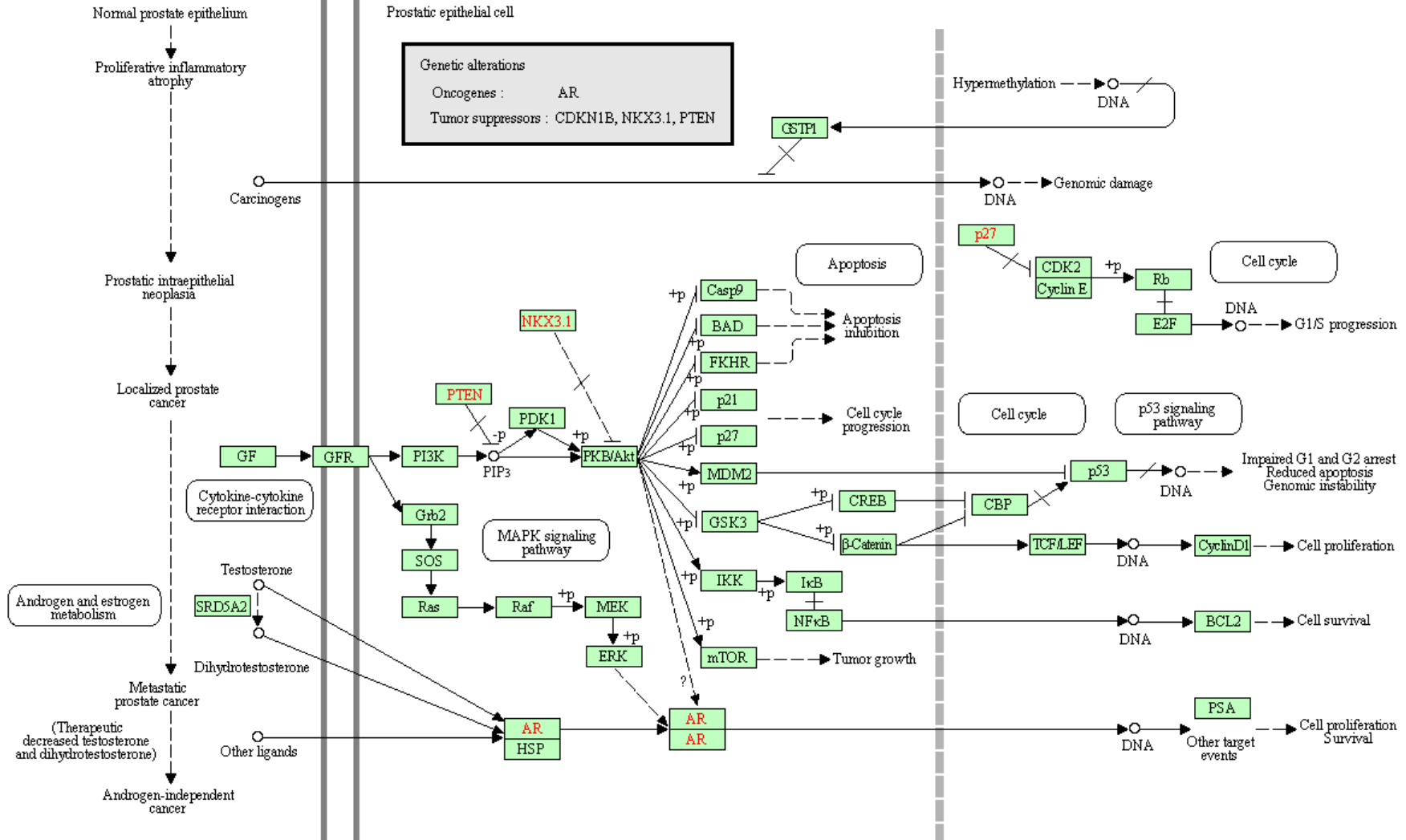
P53 signaling pathway

Metabolism

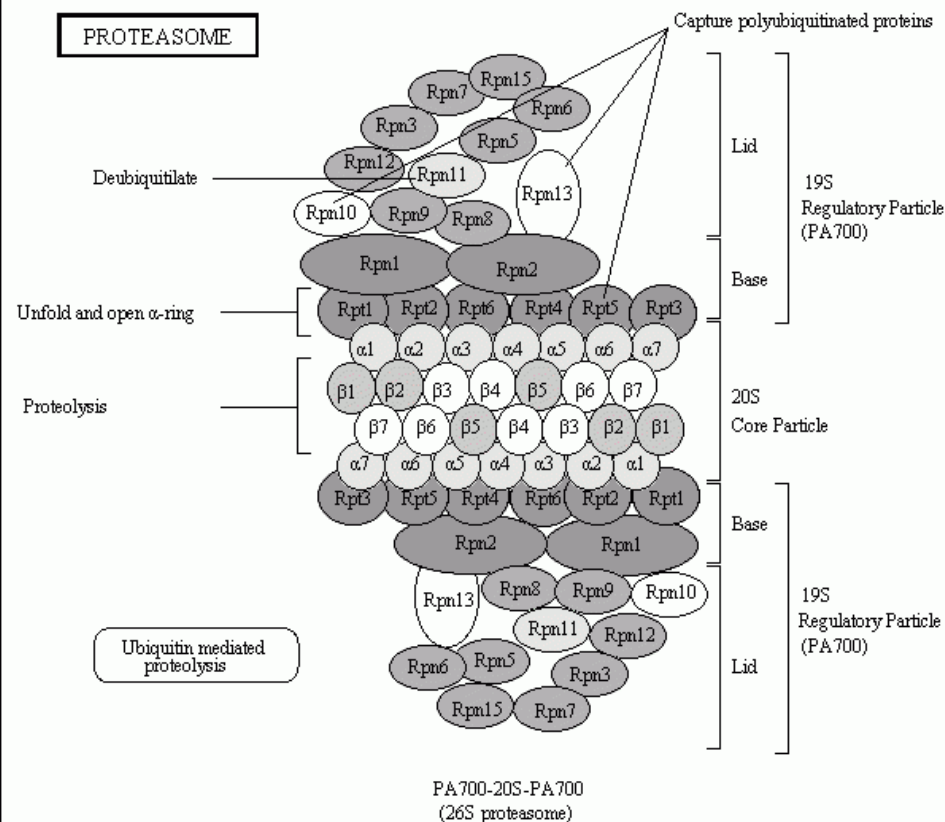
Metabolism of cofactors and vitamins

Retinol metabolism

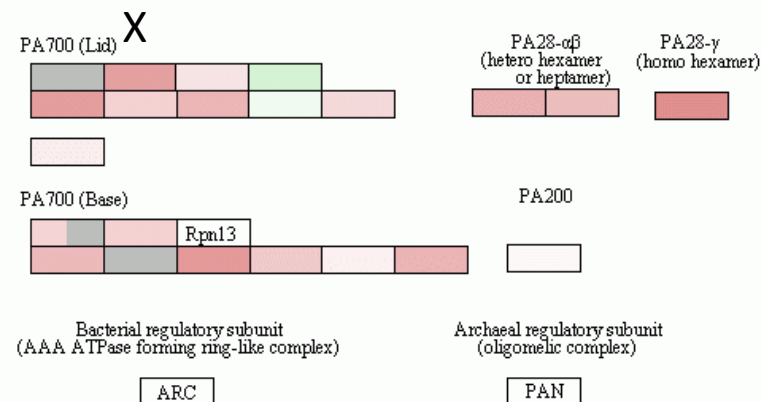
KEGG prostate cancer pathway (42 genes)



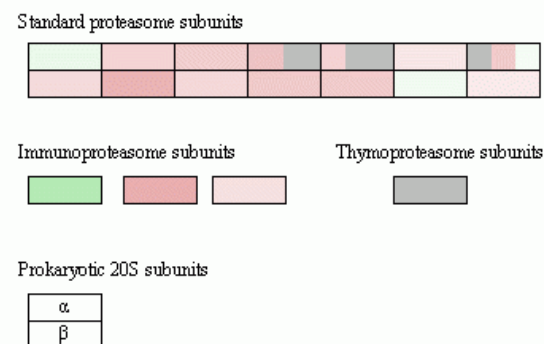
PROTEASOME



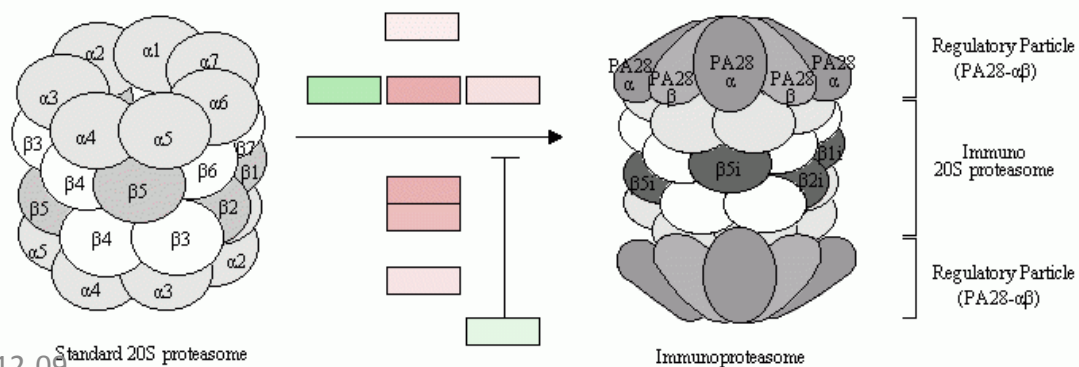
Regulatory Particles



Core Particles (20S proteasome)



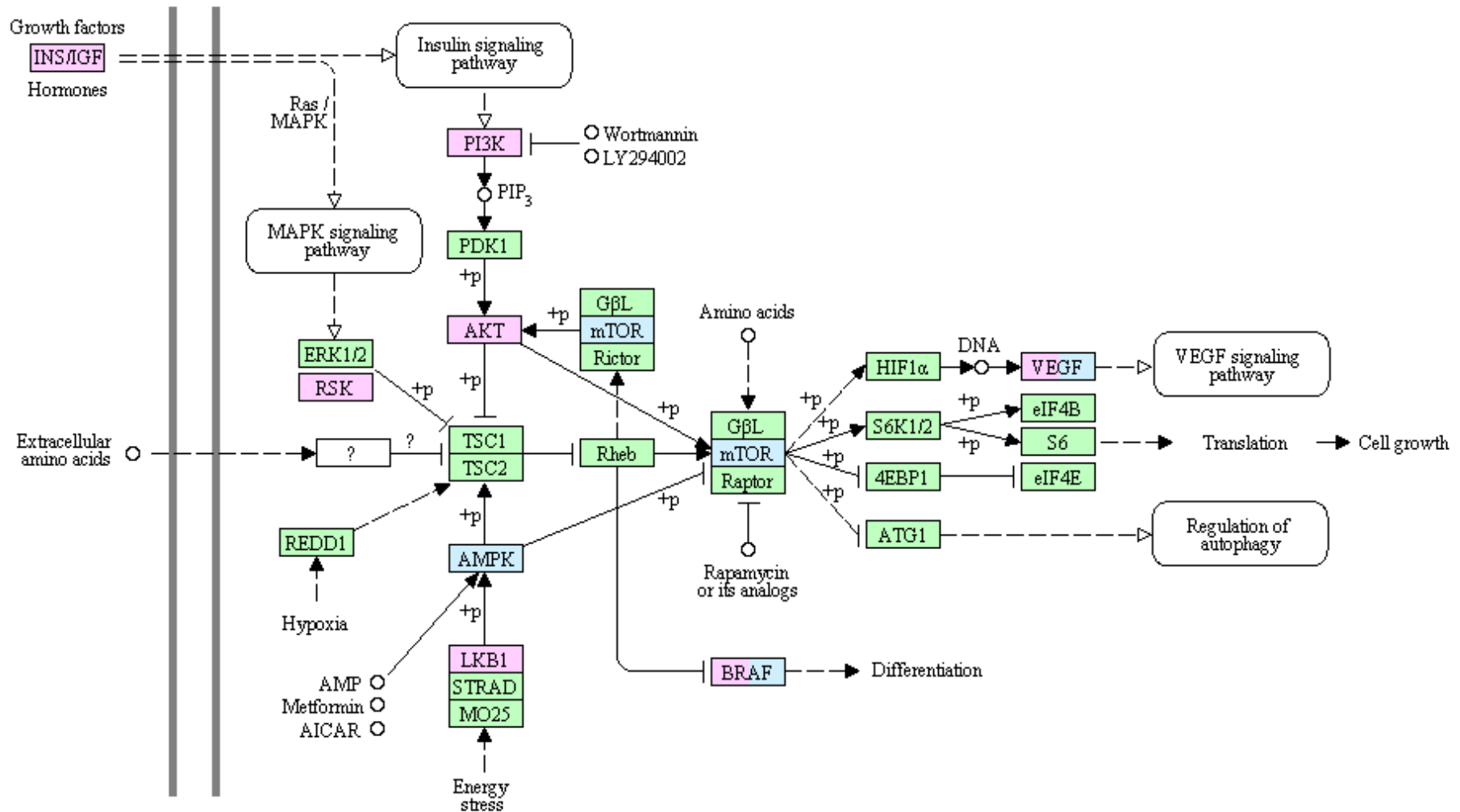
Formation of immunoproteasomes



11-12-09

mTOR SIGNALING PATHWAY

PINK: disease
BLUE: drug target

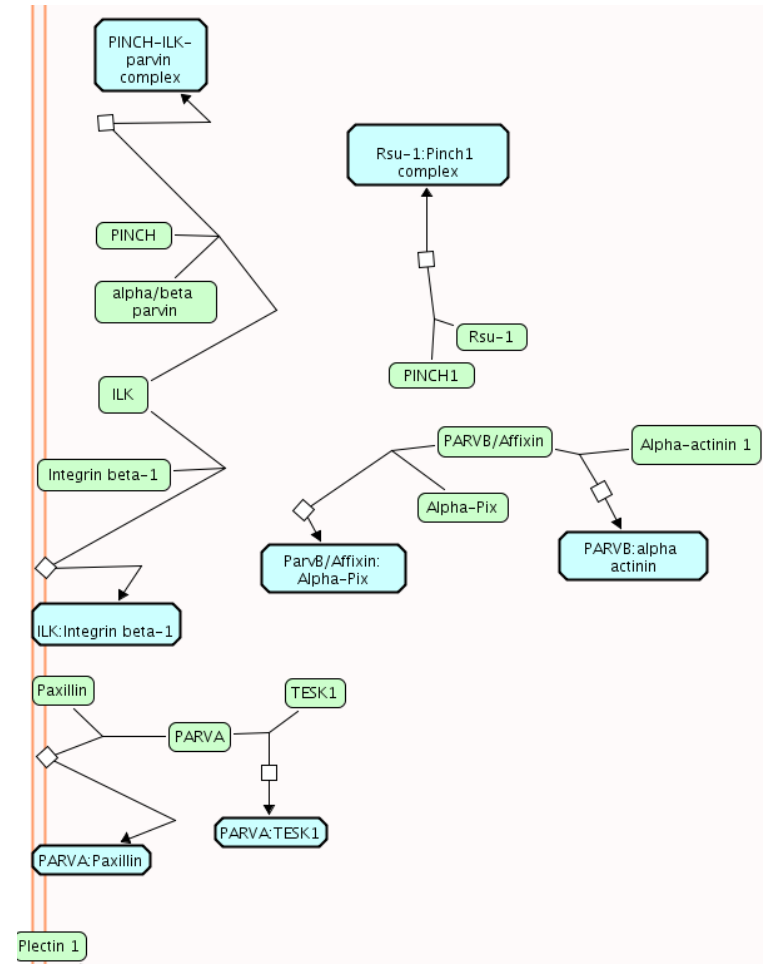




Reactome



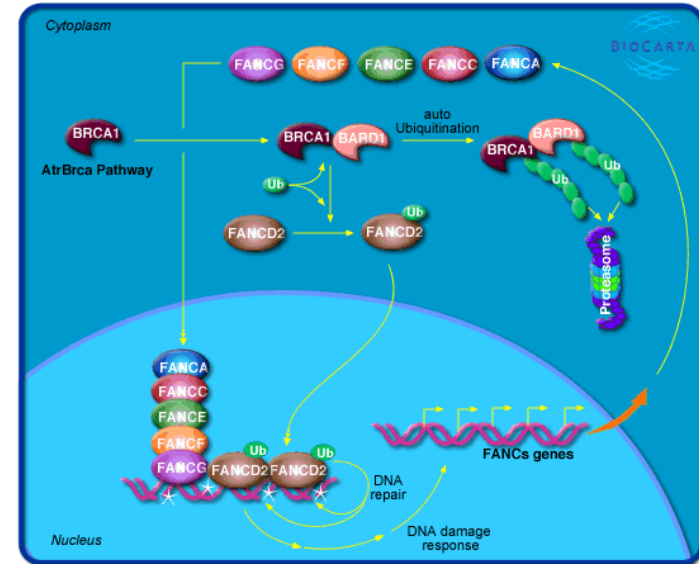
- open-source
- curated pathway database
- encompassing many areas of human biology.
- Information is authored by expert biological researcher
- 5.000 distinct human proteins
- model organism: human, other species inferred from human data
- Cytoscape plugin available



- curated resource of signal transduction pathways in humans.
- At this time, 10 immune and 10 cancer signaling pathways are available.
- diagram available

Cancer Signaling Pathways	Immune Signaling Pathways
▪ EGFR1	▪ B Cell Receptor
▪ TGF beta Receptor	▪ T Cell Receptor
▪ TNF alpha/NF-kB	▪ IL-1
▪ Alpha6 Beta4 Integrin	▪ IL-2
▪ ID	▪ IL-3
▪ Hedgehog	▪ IL-4
▪ Notch	▪ IL-5
▪ Wnt	▪ IL-6
▪ AR	▪ IL-7
▪ Kit Receptor	▪ IL-9
▪ TSH	▪ RANKL
▪ Leptin	▪ TSLP
	▪ FSH

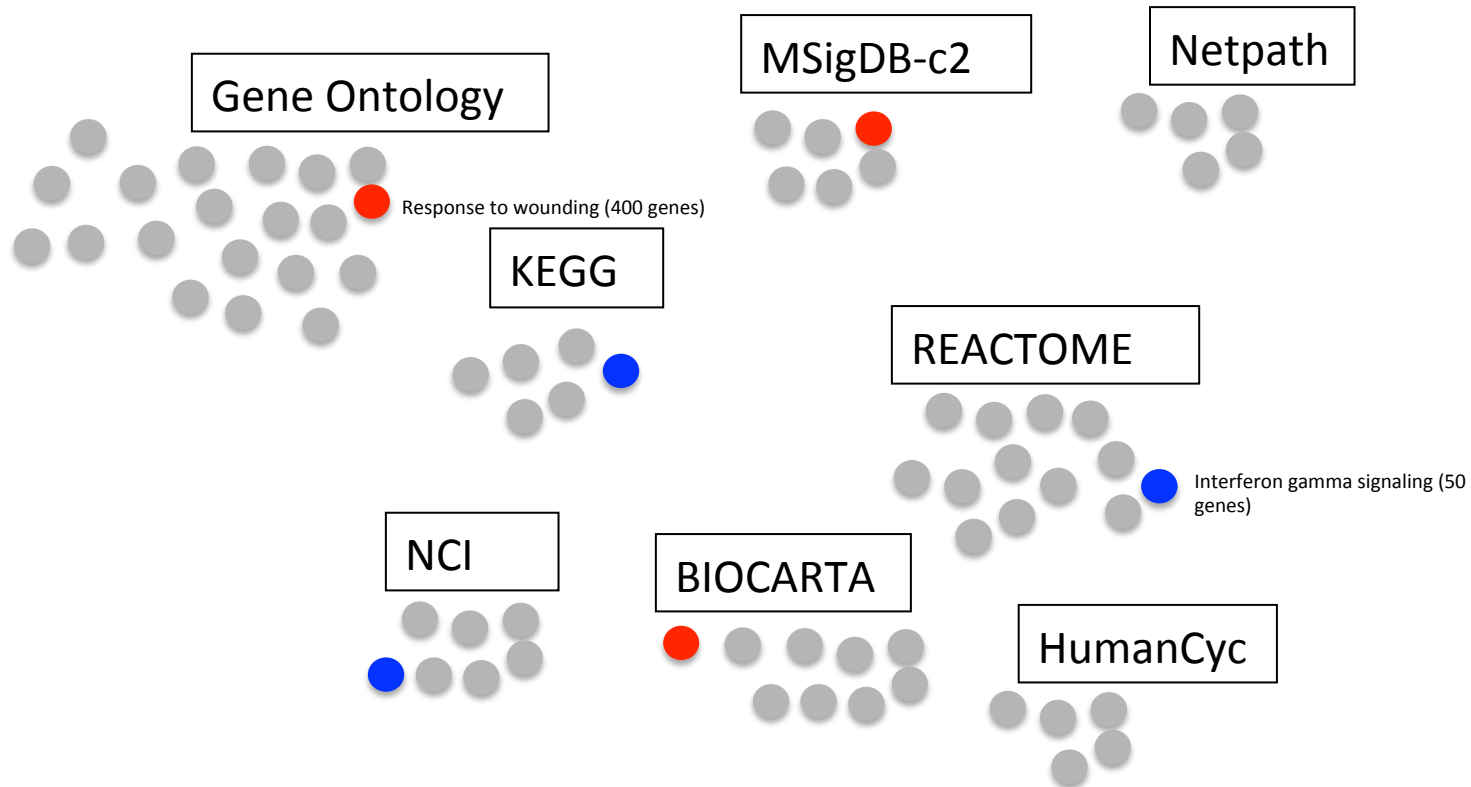
- company which develops reagents and assays for biopharmaceutical research
- expert-curated interactive graphic models of many pathways from diverse fields like apoptosis, cell cycle, cell signalling, development, immunology, neuroscience, adhesion, and metabolism.
- organisms: Homo sapiens and Mus musculus



Description: BRCA1 is a breast and ovarian cancer tumor suppressor protein that associates with BARD1 to form a RING/RING heterodimer. The BRCA1/BARD1 RING complex functions as an ubiquitin (Ub) ligase with activity substantially greater than individual BRCA1 or BARD1 subunits. The BRCA1 tumor suppressor forms a heterodimer with the BARD1 protein, and the resulting complex functions as an E3 ubiquitin ligase that catalyzes the synthesis of polyubiquitin chains. Ubch5c and Ubch7 also interact with the BRCA1/BARD1 complex with similar affinity (not shown on this figure). Although the in vivo substrate(s) is not yet known, BRCA1 has been observed to undergo autoubiquitination and is capable of monoubiquitinating histones 2A and 2AX in vitro

Gene-sets

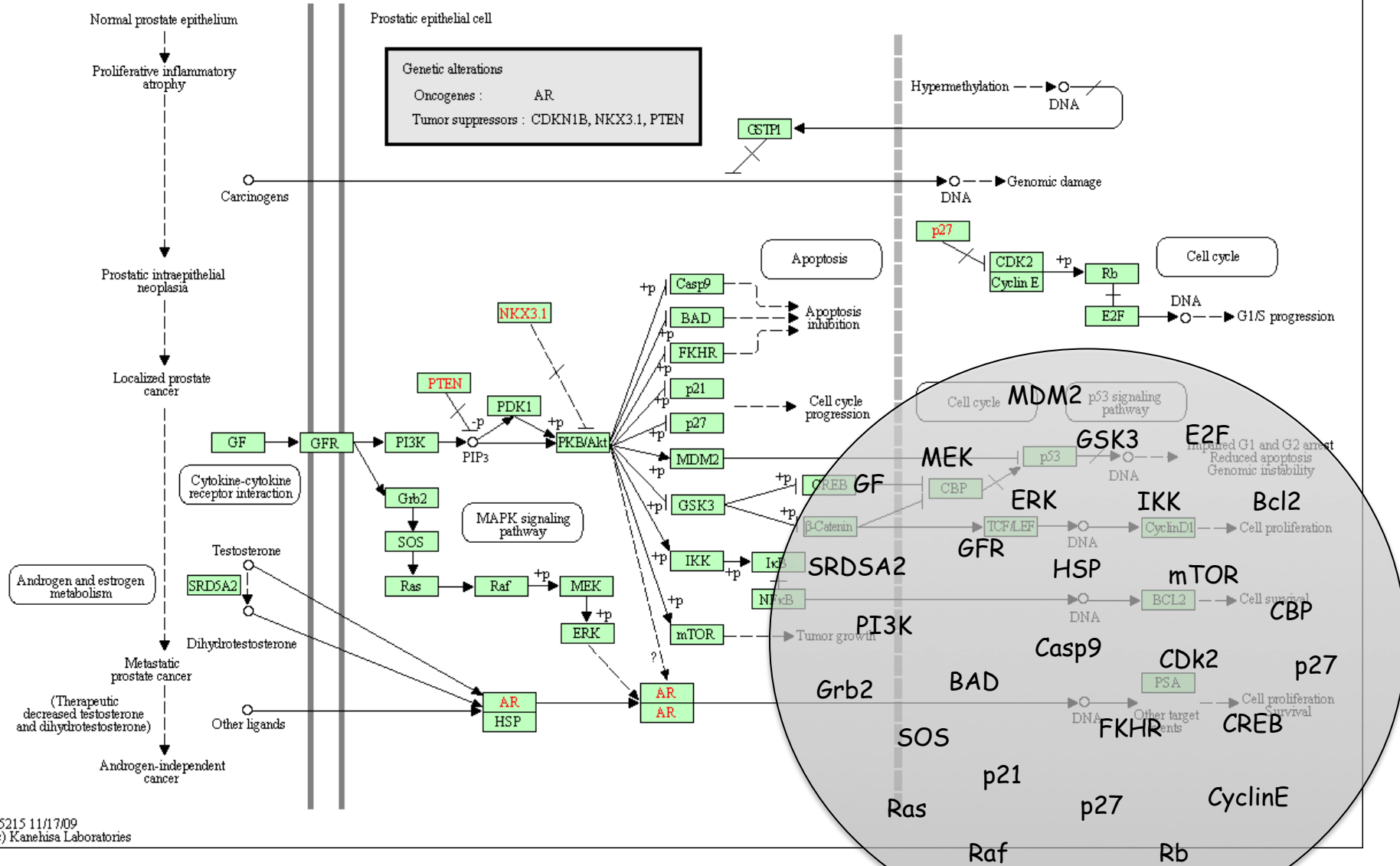
DATABASES containing annotated pathways



● a gene-set : group of genes with “similar function or annotation”

PROSTATE CANCER

KEGG prostate cancer pathway (42 genes)



KEGG prostate cancer gene-set
(42 genes)

gene-set identifier

gene-set name

KEGG|hsa05215

Prostate cancer

AKT1	AKT2	AKT3	AR	ARAF	ATF4	BAD	BCL2	BRAF
CASP9	CCND1	CCNE1	CCNE2	CDK2	CDKN1A	CDKN1B	CHUK	CREB1
CREB3	CREB3L1	CREB3L2	CREB3L3	CREB3L4	CREB5	CREBBP	CTNNB1	E2F1
E2F2	E2F3	EGF	EGFR	EP300	ERBB2	FGFR1	FGFR2	FOXO1
GRB2	GSK3B	GSTP1	HRAS	HSP90AA1		HSP90AB1	HSP90B1	IGF1
IGF1R	IKBKB	IKBKGIN5	INSRR	KLK3	KRAS	LEF1	MAP2K1	
MAP2K2	MAPK1	MAPK3	MDM2	MTOR	NFKB1	NFKBIA	NKX3-1	NRAS
PDGFA	PDGFB	PDGFC	PDGFD	PDGFRA	PDGFRB	PDPK1	PIK3CA	PIK3CB
PIK3CD	PIK3CG	PIK3R1	PIK3R2	PIK3R3	PIK3R5	PTEN	RAF1	RB1
RELA	SOS1	SOS2	SRD5A2	TCF7	TCF7L1	TCF7L2	TGFA	TP53

GO GO:0001972	retinoic acid binding			UGT1A7
UGT1A1	UGT1A4	CYP26B1	CYP26A1	
UGT1A6	CYP26C1	UGT2B4	UGT1A3	
UGT1A8	UGT1A9	UGT1A10	NR2F2	
UGT2B17	UGT2B7	UGT2B15	RARA	

REACTOME REACT_13552.1	Integrin cell surface interactions				ITGA1	ITGB5	BCAR1
LAMA5	ITGA8	LAMA2	ITGA6	ITGA5	ITGA7	PTPN1	COL1A2
ITGA2	RAP1B	RAP1A	ITGA9	BSG	ITGB1	SYK	ITGAV
COL4A4	ITGA2B	SPP1 FGB	FN1	COL4A1	COL4A2	COL4A3	CDH1
COL4A5	PDPK1	COL2A1	COL1A1	SHC1	THBS1	CSK	PTK2 VWF FGA
FGG	ITGB3	TLN1	APBB1IP	JAM3	AKT1	IBSP	ITGAL TNC
ICAM2	ICAM3	LAMC3	JAM2	LAMC1	C17ORF72	LAMB1	ICAM1
F11R	ITGA3	FBN1	RASGRP1	C17ORF57	ITGA4	ITGB8	LAMA1
ITGAX	AMICA1	ITGB6	ITGAM	ITGB2	LAMB2	VCAM1	ITGA11
SOS1	ITGAE	ITGA10	ITGB4				

How do enrichment analysis work?
(general concept)

Enrichment tests analyze the overlaps between our gene list and the genes in each gene-set that are contained in our pathway database.



gene list

AKT2	AR	CASP9	CCND1	CDK2	CREB3L4	CTNNB1	E2F3	EGFR	IGF1
IKBKB	KLK3	MAPK1	PIK3CD	PIK3CG	PIK3R2	PIK3R5	TP53		



Enriched in the Kegg
prostate cancer gene-
set?

enriched more by chance?

KEGG hsa05215		Prostate cancer							
AKT1	AKT2	AKT3	AR	ARAF	ATF4	BAD	BCL2	BRAF	
CASP9	CCND1	CCNE1	CCNE2	CDK2	CDKN1A	CDKN1B	CHUK	CREB1	
CREB3	CREB3L1	CREB3L2	CREB3L3	CREB3L4	CREB5	CREBBP	CTNNB1	E2F1	
E2F2	E2F3	EGF	EGFR	EP300	ERBB2	FGFR1	FGFR2	FOXO1	
GRB2	GSK3B	GSTP1	HRAS	HSP90AA1		HSP90AB1	HSP90B1	IGF1	
IGF1R	IKBKB	IKBKG	INSR	KLK3	KRAS	LEF1	MAP2K1		
MAP2K2	MAPK1	MAPK3	MDM2	MTOR	NFKB1	NFKBIA	NKX3-1	NRAS	
PDGFA	PDGFB	PDGFC	PDGFD	PDGFRA	PDGFRB	PDPK1	PIK3CA	PIK3CB	
PIK3CD	PIK3CG	PIK3R1	PIK3R2	PIK3R3	PIK3R5	PTEN	RAF1	RB1	
RELA	SOS1	SOS2	SRD5A2	TCF7	TCF7L1	TCF7L2	TGFA	TP53	

The output of an enrichment test is a p-value that estimates if the association between a gene-set and our gene list is random or not =

whether it exists a larger number of overlapping genes between our gene list and a gene-set than expected by chance.

Fisher's exact test

a.k.a., the hypergeometric test

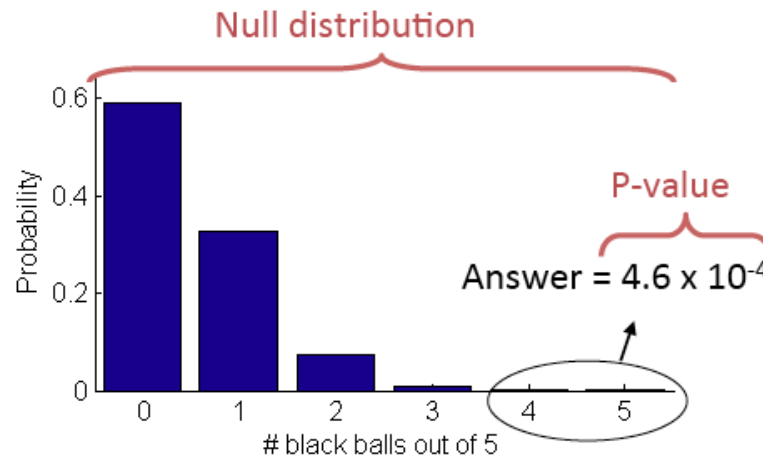
(or modified
Fisher's exact
test)

Gene list

- RRP6
- MRD1
- RRP7
- RRP43
- RRP42



Genes from all gene-sets

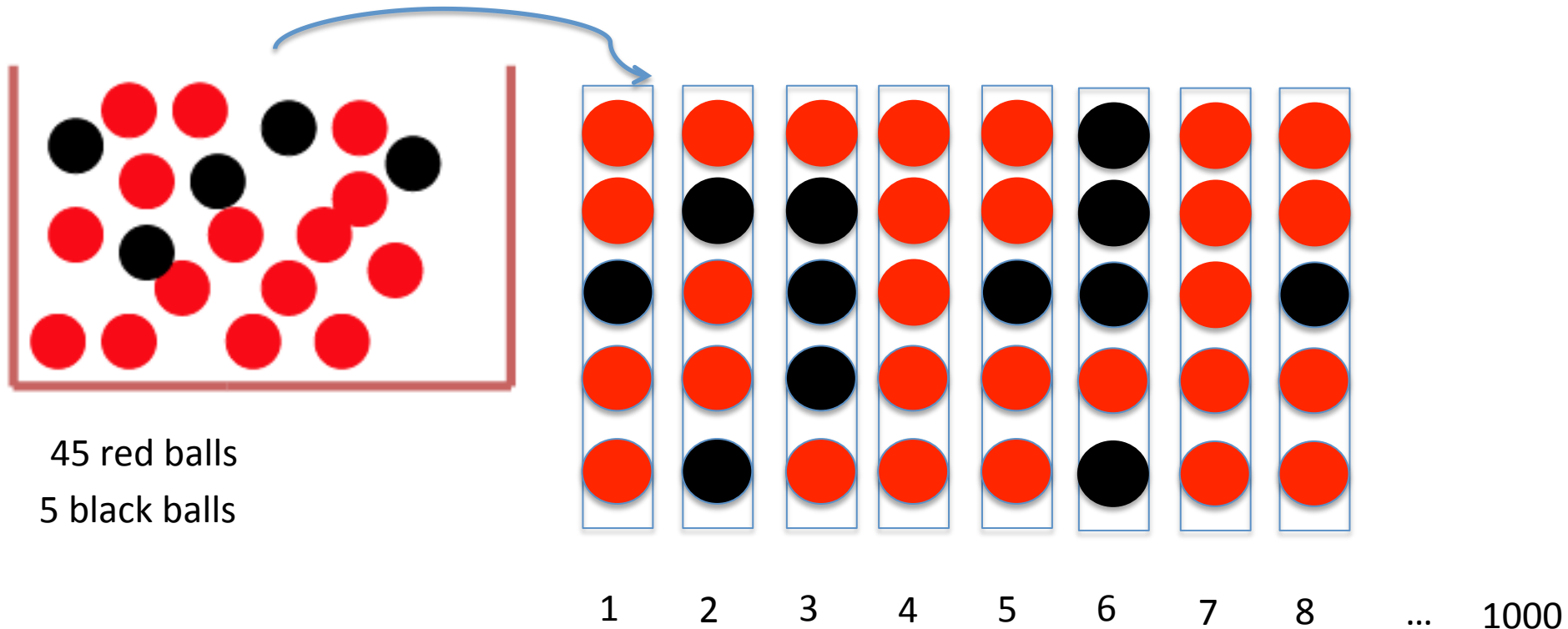


Background population:
500 black genes,
4500 red genes

Fisher's exact test

The null distribution (hypergeometric distribution)

5 balls are selected randomly 1000 times



What is the probability to select 4 black balls and one red?

Fisher's exact test

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

k= # of black balls selected (4)
q= total # of balls selected (5)
m= total # of black ball (5)
t= total # of balls (50)

Cut-off: my case = 0.000106

P-value: is the sum of p-value less than or equal to cut-off

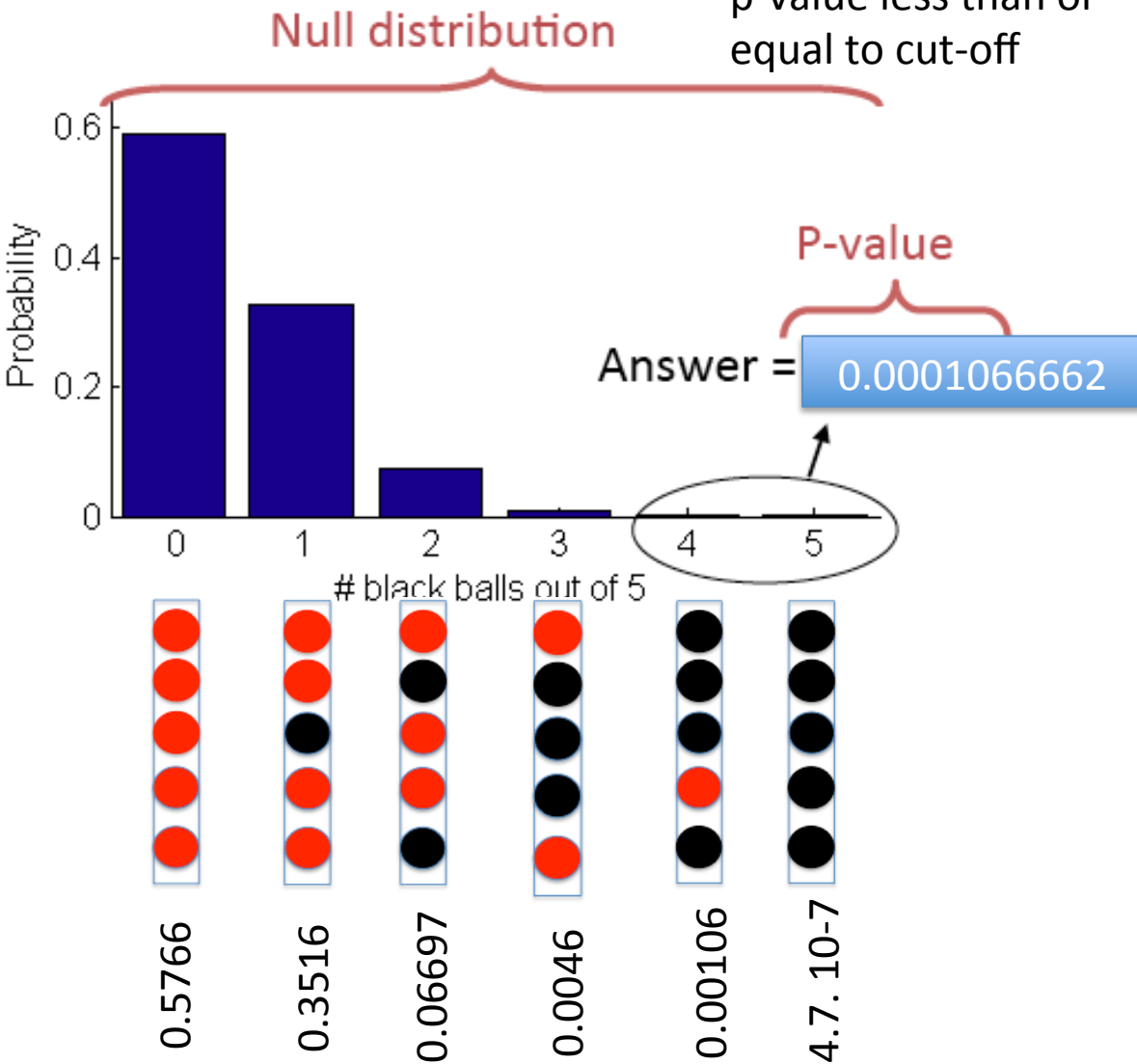


50 balls



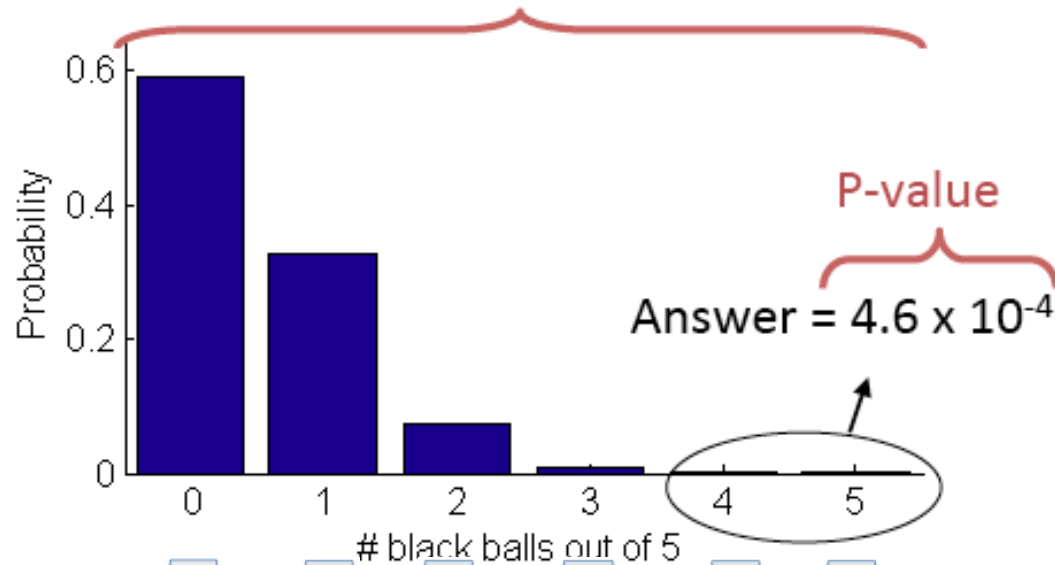
H₀: the selection of 4 black balls and 1 red ball is random

1 gene-set (prostate cancer)

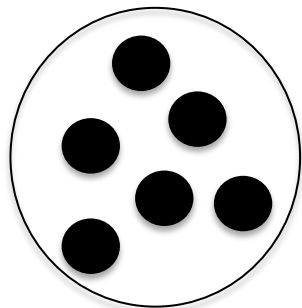


Fisher's exact test

The null distribution Null distribution



all genes in the
gene-set and in
our microarray



1 gene-set (apoptosis)

= gene list enriched in the apoptosis
gene-set)

Gene list

- RRP6
- MRD1
- RRP7
- RRP43
- RRP42

Different steps of the enrichment analysis:

1. The overlap of our gene list is going to be tested with each gene-set present in the pathway database (>3.000 gene-sets)
2. The gene-sets are going to be ranked by the enrichment p-value to find out the most significant gene-sets
3. The enrichment p-values need to be corrected for multiple hypothesis testing (FDR, Benjamini-Hochberg for example)

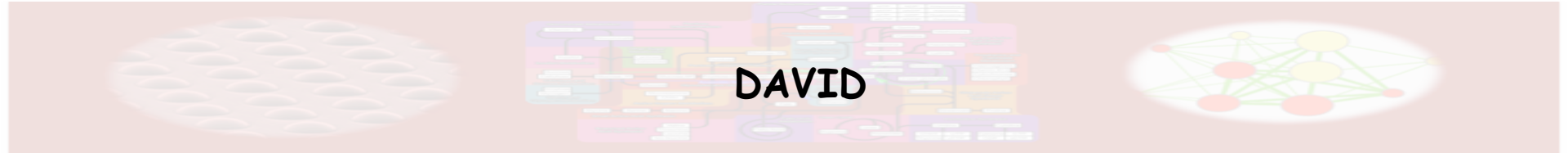


DAVID

(the database for annotation, visualization and integrated discovery)

<http://david.abcc.ncifcrf.gov/>





DAVID

- Free, On-line
- Gene-sets from diverse databases
- Use a modified Fisher's exact test to compute the enrichment p-values (called the Ease score)
- The user needs to enter a list of selected genes:

Typically 100 – 2.000 genes

The list should ideally contain the major important genes (=the most differentially expressed and the most reliable)

A larger gene list can have higher statistical power than a smaller one and the sensitivity is increased toward more specific terms.

- Different gene identifiers can be used as DAVID input (Entrez Gene ID, official gene names, Affy or Illumina probe IDs,...)

DAVID

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

Modified Fisher's exact test

k= # of black balls selected (4)

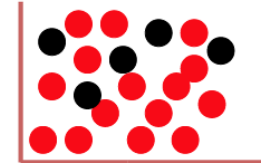
q= total # of balls selected (5)

m= total # of black ball (5)

t= total # of balls (50)

Gene list

- RRP6
- MRD1
- RRP7
- RRP43
- RRP42



2. EASE Score, a modified Fisher Exact P-Value

When members of two independent groups can fall into one of two mutually exclusive categories, Fisher Exact test is used to determine whether the proportions of those falling into each category differs by group. In DAVID annotation system, Fisher Exact is adopted to measure the gene-enrichment in annotation terms.

A Hypothetical Example:

In human genome background (30,000 gene total), 40 genes are involved in p53 signalling pathway. A given gene list has found that 3 out of 300 belong to p53 signalling pathway. Then we ask the question if 3/300 is more than random chance comparing to the human background of 40/30000.

A 2x2 contingency table is built on above numbers:

	User Genes	Genome
In Pathway	3-1	40
Not In Pathway	297	29960

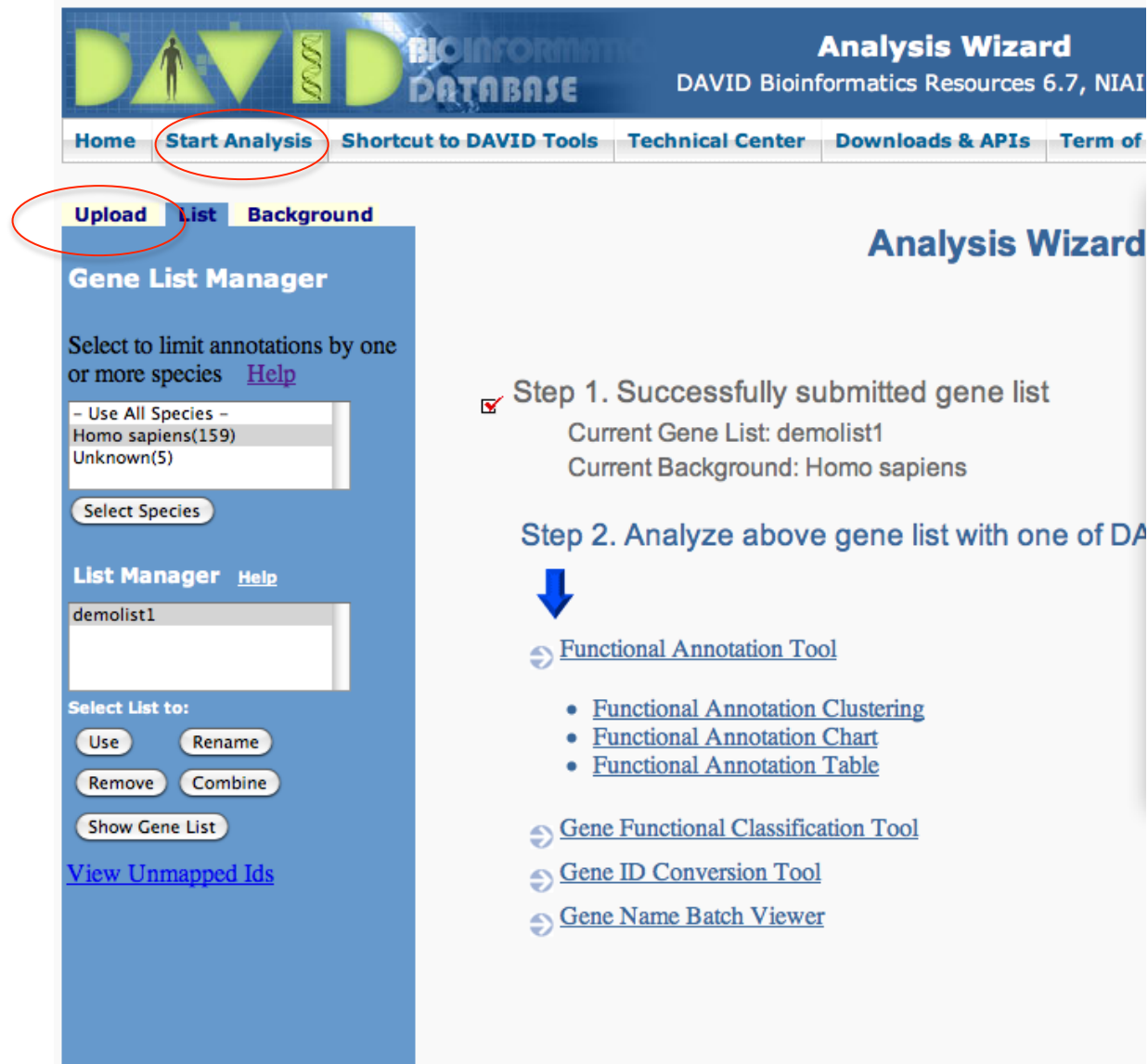
k-1: modified test (to be more conservative)

Fisher Exact P-Value = 0.008 (using 3 instead of 3-1). Since P-Value <= 0.01, this user gene list is specifically associated (enriched) in p53 signalling pathway than random chance

However, EASE Score is more conservative to examine the situation. EASE Score = 0.06 (using 3-1 instead of 3). Since P-Value > 0.01, this user gene list is specifically associated (enriched) in p53 signalling pathway no mmore than random chance

http://david.abcc.ncifcrf.gov/content.jsp?file=functional_annotation.html#fisher

1. Start the analysis



DAVID BIOINFORMATICS DATABASE Analysis Wizard
DAVID Bioinformatics Resources 6.7, NIAI

Home **Start Analysis** Shortcut to DAVID Tools Technical Center Downloads & APIs Term of

Upload List Background

Gene List Manager

Select to limit annotations by one or more species [Help](#)

– Use All Species –
Homo sapiens(159)
Unknown(5)

Select Species

List Manager [Help](#)

demolist1

Select List to:

Use Rename
Remove Combine
Show Gene List

[View Unmapped Ids](#)

Analysis Wizard

☒ Step 1. Successfully submitted gene list
Current Gene List: demolist1
Current Background: Homo sapiens

Step 2. Analyze above gene list with one of DA

↓

↻ [Functional Annotation Tool](#)

- [Functional Annotation Clustering](#)
- [Functional Annotation Chart](#)
- [Functional Annotation Table](#)

↻ [Gene Functional Classification Tool](#)

↻ [Gene ID Conversion Tool](#)

↻ [Gene Name Batch Viewer](#)

2. Upload Tab

The screenshot shows the 'Upload' tab of the DAVID bioinformatics tool. The interface is divided into four steps:

- Step 1: Enter Gene List**
 - A: Paste a list**: A text area containing gene identifiers (1007_at, 1053_at, 117_at, 121_at). An arrow points to this area with the number 1.
 - B: Choose From a File**: A file selection interface with a 'Browse...' button.
- Step 2: Select Identifier**: A dropdown menu currently set to 'AFFY_ID'. An arrow points to this dropdown with the number 2.
- Step 3: List Type**: Two radio buttons are present: 'Gene List' (selected) and 'Background'. An arrow points to the 'Gene List' radio button with the number 3.
- Step 4: Submit List**: A 'Submit List' button. An arrow points to this button with the number 4.

Annotations on the right side of the image provide further details:

- 1: Copy/paste gene list into the box; One gene per row without header row.
- 2: Tell DAVID corresponding gene identifier type of above genes
- 3: Gene List: for users' analysis, e.g. selected interesting gene from an array experiment. Background: for enrichment calculation only, e.g. entire genes in an array
- 4: Submit list to DAVID

3. List Tab

The screenshot shows the 'Gene List Manager' interface. At the top, there are three tabs: 'Upload', 'List', and 'Background'. The 'List' tab is selected and circled in red. Below the tabs, the title 'Gene List Manager' is displayed. A text prompt says 'Select to limit annotations by one or more species' followed by a blue 'Help' link. Below this is a dropdown menu with three options: '- Use All Species -', 'Homo sapiens(391)', and 'Synthetic construct(3)'. The 'Homo sapiens(391)' option is highlighted in blue. Below the dropdown is a 'Select' button. Further down, the 'List Manager' section has a blue 'Help' link. Below that is a list box containing 'demolist2' and 'demolist1', with 'demolist2' highlighted in blue. Below the list box is the text 'Select List to:'. Underneath this text are four buttons: 'Use', 'Rename', 'Remove', and 'Combine'. The 'Use' button is highlighted in blue. At the bottom is a large button labeled 'Show Gene List' with 'new!' in red text next to it. Five red callout lines with numbered circles point to specific elements: 1 points to 'demolist2', 2 points to the 'Use' button, 3 points to 'Homo sapiens(391)', 4 points to the 'Select' button, and 5 points to the 'List' tab.

Upload **List** Background

Gene List Manager

Select to limit annotations by one or more species [Help](#)

- Use All Species -
Homo sapiens(391)
Synthetic construct(3)

Select

List Manager [Help](#)

demolist2
demolist1

Select List to:

Use Rename
Remove Combine

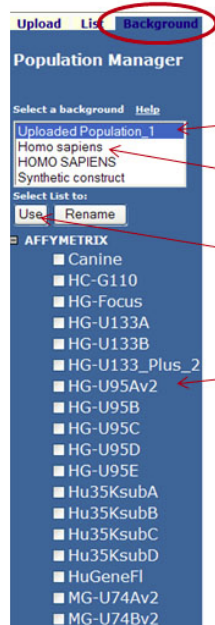
Show Gene List^{new!}

- ③ Species info. for the gene list selected in the box below
- ④ Always click "Select" button to switch species
- ① Highlight the gene list to be analyzed
- ② Always click "Use" button to switch gene lists

DAVID

Need to add a background correction to calculate the enrichment scores

4. Background Tab



User uploaded background

Whole genome background (DAVID default)

Switch background selected in above box

Pre-built array backgrounds



= background (all genes in the pathway database and on your array)

	Whole genome	Your custom chip
Gene-set size	200	20
Total number of genes in your database	23.000	10.000

Background Tab allows you to manage different background for enrichment an

- DAVID has an automatic procedure to 'guess' the background.
- Most of the studies are genome-wide or close to genome-wide studies.

5 .Select the databases

The screenshot displays the 'Population Manager' web application. The 'Background' tab is selected, showing a list of backgrounds with 'Homo sapiens' chosen. Below this, there are sections for 'Affymetrix 3' IVT Backgrounds' and 'Affymetrix Exon Backgrounds', each with a list of options and a 'Use' button. To the right, the 'Annotation Summary Results' section is highlighted with a red box. It shows the 'Current Gene List' as 'demolist2' and the 'Current Background' as 'Homo sapiens'. A list of annotation categories follows, with counts for each: Disease (1 selected), Functional_Categories (3 selected), Gene_Ontology (3 selected), General Annotations (0 selected), Literature (0 selected), Main_Accessions (0 selected), Pathways (3 selected), Protein_Domains (3 selected), Protein_Interactions (0 selected), and Tissue_Expression (0 selected). A red note states: '***Red annotation categories denote DAVID defined defaults***'. At the bottom, the 'Combined View for Selected Annotation' section offers three options: 'Functional Annotation Clustering', 'Functional Annotation Chart', and 'Functional Annotation Table'.

Population Manager

Select a background [Help](#)

Homo sapiens

Select List to:

Use Rename

Affymetrix 3' IVT Backgrounds [Help](#)

- ☐ Affymetrix Mu19KsubA
- ☐ Affymetrix Mu19KsubB
- ☐ Affymetrix Mu19KsubC
- ☐ Arabidopsis ATH1-121501 Genome Array

Affymetrix Exon Backgrounds [Help](#)

- ☐ HuGene-1_0-st-v1
- ☐ MoEx-1_0-st-v1
- ☐ MoGene-1_0-st-v1
- ☐ RaEx-1_0-st-v1
- ☐ RaGene-1_0-st-v1

Annotation Summary Results

Current Gene List: demolist2

Current Background: Homo sapiens

- ☒ **Disease** (1 selected)
- ☒ **Functional_Categories** (3 selected)
- ☒ **Gene_Ontology** (3 selected)
- ☐ **General Annotations** (0 selected)
- ☐ **Literature** (0 selected)
- ☐ **Main_Accessions** (0 selected)
- ☒ **Pathways** (3 selected)
- ☒ **Protein_Domains** (3 selected)
- ☒ **Protein_Interactions** (0 selected)
- ☐ **Tissue_Expression** (0 selected)

Red annotation categories denote DAVID defined defaults

Combined View for Selected Annotation

Functional Annotation Clustering

Functional Annotation Chart

Functional Annotation Table

Pathway database

in red: default

Upload List Background

Gene List Manager

Select to limit annotations by one or more species

Use All Species
Homo sapiens(15)
Unknown(5)

List Manager Help

demolist1

Select List to:

List Name Remove Compare

Show Gene List

View Unmapped Ids

Annotation Summary Results

[Help and Tool Manual](#)

Current Gene List: demolist1 155 DAVID IDs

Current Background: Homo sapiens Check Defaults

Disease (1 selected)

GENETIC_ASSOCIATION_DB_DISEASE	38.7%	60	Chart
GENETIC_ASSOCIATION_DB_DISEASE_CLASS	38.7%	60	Chart
OMIM_DISEASE	28.4%	44	Chart

Functional Categories (3 selected)

COG_ONTOLOGY	9.7%	15	Chart
PIR_SEQ_FEATURE	28.4%	44	Chart
SP_COMMENT_TYPE	94.2%	146	Chart
SP_PIR_KEYWORDS	95.5%	148	Chart
UP_SEQ_FEATURE	94.2%	146	Chart

Gene Ontology (3 selected)

GOTERM_BP_1	87.7%	136	Chart
GOTERM_BP_2	87.1%	135	Chart
GOTERM_BP_3	83.2%	129	Chart
GOTERM_BP_4	83.2%	129	Chart
GOTERM_BP_5	78.1%	121	Chart
GOTERM_BP_ALL	89.0%	138	Chart
GOTERM_BP_FAT	87.7%	136	Chart
GOTERM_CC_1	86.5%	134	Chart
GOTERM_CC_2	83.9%	130	Chart
GOTERM_CC_3	83.9%	130	Chart
GOTERM_CC_4	75.5%	117	Chart
GOTERM_CC_5	71.0%	110	Chart
GOTERM_CC_ALL	86.5%	134	Chart
GOTERM_CC_FAT	80.0%	124	Chart
GOTERM_MF_1	85.2%	132	Chart
GOTERM_MF_2	84.5%	131	Chart
GOTERM_MF_3	75.5%	117	Chart
GOTERM_MF_4	71.0%	110	Chart
GOTERM_MF_5	61.9%	96	Chart
GOTERM_MF_ALL	85.2%	132	Chart
GOTERM_MF_FAT	75.5%	117	Chart
PANTHER_BP_ALL	77.4%	120	Chart
PANTHER_MF_ALL	77.4%	120	Chart

General Annotations (0 selected)

CHROMOSOME	99.4%	154	Chart
CYTOBAND	98.7%	153	Chart
ENTREZ_GENE_SUMMARY	69.0%	107	Chart
HOMOLOGOUS_GENE	94.8%	147	Chart
OFFICIAL_GENE_SYMBOL	99.4%	154	Chart
PIR_SUMMARY	58.7%	91	Chart
SP_COMMENT	93.5%	145	Chart

Literature (0 selected)

GENERIF_SUMMARY	81.9%	127	Chart
HIV_INTERACTION_PUBMED_ID	11.0%	17	Chart
PUBMED_ID	98.7%	153	Chart

Main Accessions (0 selected)

ENSEMBL_GENE_ID	98.1%	152	Chart
ENTREZ_GENE_ID	99.4%	154	Chart

Pathways (3 selected)

BBID	4.5%	7	Chart
BIOCARTA	14.8%	23	Chart
EC_NUMBER	24.5%	38	Chart
KEGG_PATHWAY	38.7%	60	Chart
PANTHER_PATHWAY	24.5%	38	Chart
REACTOME_PATHWAY	20.6%	32	Chart

Literature (0 selected)			
GENERIF_SUMMARY	81.9%	127	Chart
HIV_INTERACTION_PUBMED_ID	11.0%	17	Chart
PUBMED_ID	98.7%	153	Chart

Main Accessions (0 selected)

ENSEMBL_GENE_ID	98.1%	152	Chart
ENTREZ_GENE_ID	99.4%	154	Chart

Pathways (3 selected)

BBID	4.5%	7	Chart
BIOCARTA	14.8%	23	Chart
EC_NUMBER	24.5%	38	Chart
KEGG_PATHWAY	38.7%	60	Chart
PANTHER_PATHWAY	24.5%	38	Chart
REACTOME_PATHWAY	20.6%	32	Chart

Protein Domains (3 selected)

BLOCKS	69.0%	107	Chart
COG_NAME	9.7%	15	Chart
INTERPRO	92.9%	144	Chart
PANTHER_FAMILY	89.7%	139	Chart
PANTHER_SUBFAMILY	76.8%	119	Chart
PFAM	91.6%	142	Chart
PIR_SUPERFAMILY	54.2%	84	Chart
PRINTS	41.3%	64	Chart
PRODOM	17.4%	27	Chart
PROFILE	54.8%	85	Chart
PROSITE	77.4%	120	Chart
SCOP_CLASS	20.6%	32	Chart
SCOP_FAMILY	20.6%	32	Chart
SCOP_FOLD	20.6%	32	Chart
SCOP_SUPERFAMILY	20.6%	32	Chart
SMART	60.6%	94	Chart
SSF	28.4%	44	Chart
TIGRFAMS	9.7%	15	Chart

Protein Interactions (0 selected)

BIND	38.1%	59	Chart
DIP	9.7%	15	Chart
HIV_INTERACTION	11.0%	17	Chart
HIV_INTERACTION_CATEGORY	11.0%	17	Chart
MINT	36.8%	57	Chart
NCICB_CAPATHWAY_INTERACTION	20.0%	31	Chart
REACTOME_INTERACTION	14.2%	22	Chart
UCSC_TFBS	96.1%	149	Chart

Tissue Expression (0 selected)

CGAP_EST_QUARTILE	78.7%	122	Chart
CGAP_SAGE_QUARTILE	76.1%	118	Chart
GNF_U133A_QUARTILE	95.5%	148	Chart
PIR_TISSUE_SPECIFICITY	40.0%	62	Chart
UNIGENE_EST_QUARTILE	84.5%	131	Chart
UP_TISSUE	95.5%	148	Chart

Red annotation categories denote DAVID defined defaults

UploadListBackground

Population Manager

Select a background [Help](#)

Homo sapiens

Select List to:

UseRename

Affymetrix 3' IVT Backgrounds [Help](#)

☐ Affymetrix Mu19KsubA
☐ Affymetrix Mu19KsubB
☐ Affymetrix Mu19KsubC
☐ Arabidopsis ATH1-121501 Genome Array

Affymetrix Exon Backgrounds [Help](#)

☐ HuGene-1_0-st-v1
☐ MoEx-1_0-st-v1
☐ MoGene-1_0-st-v1
☐ RaEx-1_0-st-v1
☐ RaGene-1_0-st-v1

Annotation Summary Results

Current Gene List: demolist2
Current Background: Homo sapiens

- ☒ **Disease** (1 selected)
- ☒ **Functional_Categories** (3 selected)
- ☒ **Gene_Ontology** (3 selected)
- ☐ **General Annotations** (0 selected)
- ☐ **Literature** (0 selected)
- ☐ **Main_Accessions** (0 selected)
- ☒ **Pathways** (3 selected)
- ☒ **Protein_Domains** (3 selected)
- ☐ **Protein_Interactions** (0 selected)
- ☐ **Tissue_Expression** (0 selected)

Red annotation categories denote DAVID defined defaults

Combined View for Selected Annotation

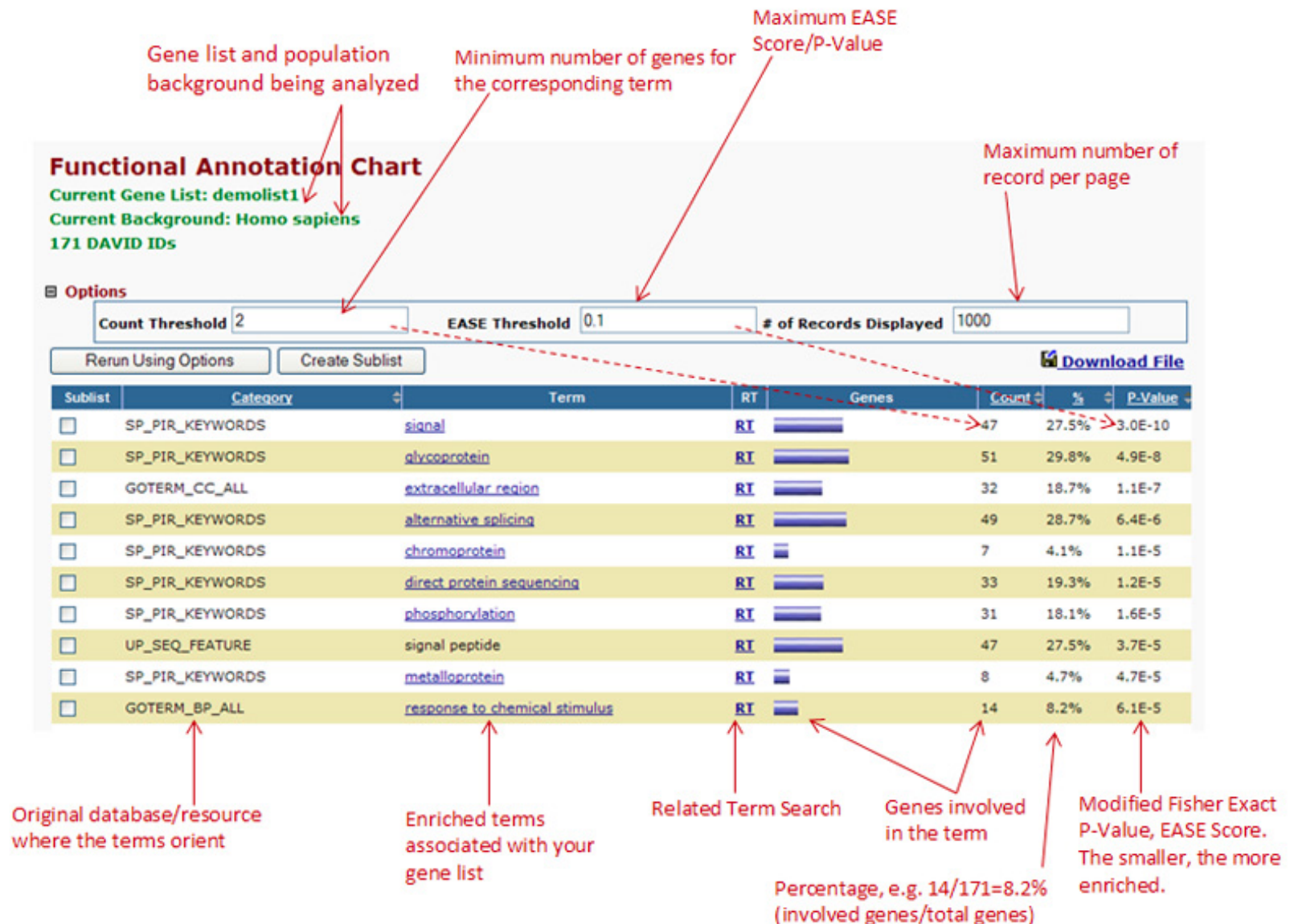
Functional Annotation Clustering

Functional Annotation Chart

Functional Annotation Table

6.Enrichment analysis

DAVID output (enrichment analysis)



DAVID output (fuzzy clustering)

Gene list being analyzed

Clustering options and stringency

The overall enrichment score for the group based on the EASE scores of each term members. The higher, the more enriched.

Every term in the annotation cluster

Related Term Search

Genes involved in individual term

ALL genes involved in this annotation cluster

A group of terms having similar biological meaning due to sharing similar gene members

Functional Annotation Clustering

Current Gene List: demolist1
171 DAVID IDs

Options Classification Stringency High

Rerun using options Create Sublist

Download File

Annotation Cluster 1		Enrichment Score: 3.69		
☐	SP_PIR_KEYWORDS	chromoprotein	RT	7 1.1E-5
☐	SP_PIR_KEYWORDS	metalloprotein	RT	8 4.7E-5
☐	SP_PIR_KEYWORDS	iron	RT	9 2.1E-4
☐	GOTERM_MF_ALL	iron ion binding	RT	10 2.5E-4
☐	SP_PIR_KEYWORDS	heme	RT	7 3.5E-4
☐	GOTERM_MF_ALL	tetrapyrrole binding	RT	6 1.3E-3
☐	GOTERM_MF_ALL	heme binding	RT	6 1.3E-3
Annotation Cluster 2		Enrichment Score: 3.52		
☐	SP_PIR_KEYWORDS	antibiotic	RT	5 2.2E-4
☐	SP_PIR_KEYWORDS	antimicrobial	RT	5 2.4E-4
☐	GOTERM_BP_ALL	defense response to bacteria	RT	6 5.4E-4
Annotation Cluster 3		Enrichment Score: 2.66		
☐	UP_SEQ_FEATURE	domain:Ig-like C2-type 1	RT	8 5.4E-4
☐	UP_SEQ_FEATURE	domain:Ig-like C2-type 2	RT	8 5.4E-4
☐	INTERPRO_NAME	Immunoglobulin	RT	6 3.6E-2
Annotation Cluster 4		Enrichment Score: 2.63		

EASE Score, the modified Fisher Exact P-Value. They are identical to that in the Chart Report. The smaller, the more enriched.



GSEA

(Gene Set Enrichment Analysis)





GSEA

- No cut-off (all gene list)
- Possible to detect situations where all genes in a predefined set change in a small but coordinated way
- Also possible to detect cases where the effect is due to large changes in a relatively few genes
- In microarray experiments where no single gene shows statistically significant differentially expressed genes, GSEA has identified significantly expressed set of genes
- Predictions of the method have been validated in independent laboratory experiments
- Easy-to-use software package, open-source
- Need a ranked list

The GSEA Algorithm

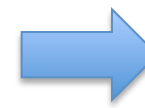
1. generate ranked gene list (Fold Change, t-Test, log-ratio)
2. for each Gene Set:
 - calculate running sum:
 - walk down ranked gene list
 - increase sum if the gene is in the current gene set
 - decrease sum if not
 - Enrichment Score (ES) is the max. deviation from zero
 - score is normalized for Gene Set size (NES)
3. Estimation of Significance
 - a) either Phenotype permutation
 - b) or Geneset permutation
 - correction for multiple testing (FDR)

The GSEA algorithm

Rank the genes using a differential expression value

UP

Ranked
list



> 40 000
probes



DOWN

CDKN1A
ZMAT3
PHLDA3
FAS
TNFSF4
SESN1
SIGLEC14
NCF1C
PT
TP53BP1
ATF3
D7
XPC
ADRB2
METTL5A
DNAH1
TP53BP1
TNFRSF10B
GADD45A
PPM1D
FBXO32
GDF15
IER5
NC1
CL2045
TRIM21
FBXO32
PLK2
CDN2
ZNF79
POLR1
FAS
TNFSF4
ANKRA2
PLNAB2
SIRPB1
BLOC1S2
ASC3
MDM2
TNFSF4
ACOT8
RPS27L
RGS12
BTID
RNF19B
ASC3
CDKN1A
APORC3H
E2F7
LOC525216
GADD45A
ZMAT3
TAP1
72
LOC789567
TNFRSF10B
CEACAM1
MIR1204
ISCU
TRIM18
REX1
SRA1
PRKAGIN
ISCU
WRBP
CL10f4
BRMS1
LOC1001290
ITB
PLNAB2
FOXK
ANKRA2
PCNA
AEN
CEACAM1
APORC3G
CCB7
SLC7A5
LOC348840
HL2
CYFIP2
KIBP
PRKAB1
ISG20L1
CYFIP2
RG1
DPYSL4
TNFRSF1
SLC7A5
PRK7
SERTAD1
FAM125B
NCF1B
BAX
ORAL3
APORC3G
TNFRSF2
ZNF501
NCDAS
P2Y1C
LRDD

EntrezGene ID	Probe_id	GENE	DESCRIPTION	mean expression value non-treated	mean expression value treated	log FC (treated vs non treated)	t (treated versus non treated)	p-Value (treated versus non treated)	adj.p-Value (treated versus non treated)
12544	ILMN_1784602	CDKN1A	cyclin-dependent kinase inhibitor 1A (p21, Cip1) (CDKN1A)	8.906	12.154	3.248	34.681	8.54E-16	1.79E-11
1005687	ILMN_1654262	ZMAT3	matrin type 3 (ZMAT3), transcript variant 2, mRNA.	10.444	12.190	1.815	20.733	1.70E-12	1.41E-08
85465	ILMN_1659106	PHLDA3	pleckstrin homology-like domain, family A, member 3 (PHLDA3), transcript variant 1, mRNA.	6.264	8.473	2.213	20.497	2.01E-12	1.41E-08
466855	ILMN_2319077	FAS	Fas (TNF receptor superfamily, member 6) (FAS), transcript variant 1, mRNA.	7.558	9.604	1.949	16.849	3.44E-11	1.57E-07
7458	ILMN_2089875	TNFSF4	tumor necrosis factor (ligand) superfamily, member 4 (TNFSF4), transcript variant 1, mRNA.	8.256	10.325	1.850	16.749	3.75E-11	1.57E-07
1235	ILMN_1800626	SESN1	sestrin 1 (SESN1), mRNA.	9.467	11.333	1.707	16.508	4.61E-11	1.61E-07
100589	ILMN_3243061	SIGLEC14	sialic acid binding Ig-like lectin 14 (SIGLEC14), mRNA.	9.658	11.620	1.885	16.209	5.99E-11	1.80E-07
256688	ILMN_2112988	NCF1C	neutrophil cytosolic factor 1C pseudogene (NCF1C), cDNA: FLJ21679 fis, clone COL09221	7.772	9.445	1.967	15.897	7.91E-11	2.07E-07
12544	ILMN_1907834	SEPT2	septin 2 (SEPT2), transcript variant 3, mRNA.	7.233	8.705	1.783	15.234	1.45E-10	3.38E-07
1005687	ILMN_1747665	SEPT2	septin 2 (SEPT2), transcript variant 3, mRNA.	6.415	6.496	-0.003	-0.025	9.81E-01	9.96E-01
85465	ILMN_1658422	LOC653242	similar to fumarylacetoacetate hydrolase domain containing 1 (FAM104A), transcript variant 1, mRNA.	6.736	7.080	-0.003	-0.025	9.81E-01	9.96E-01
466855	ILMN_1668228	LOC136143	similar to ribosomal protein L18 (LOC136143), mRNA.	11.174	11.328	-0.003	-0.025	9.80E-01	9.96E-01
7458	ILMN_1671905	C10orf78	chromosome 10 open reading frame 78 (C10orf78), transcript variant 1, mRNA.	9.504	9.596	-0.003	-0.025	9.80E-01	9.96E-01
1235	ILMN_1741564	DCTN4	dynactin 4 (p62) (DCTN4), mRNA.	8.272	8.180	-0.002	-0.025	9.80E-01	9.96E-01
110589	ILMN_3224555	LOC729397	hypothetical LOC729397 (LOC729397), mRNA.	7.050	6.551	-0.004	-0.025	9.80E-01	9.96E-01
256688	ILMN_1687316	NCDN	neurochondrin (NCDN), transcript variant 3, mRNA.	6.525	6.614	-0.003	-0.026	9.80E-01	9.96E-01
12544	ILMN_1708632	ZNF771	zinc finger protein 771 (ZNF771), mRNA.	6.998	6.965	-0.002	-0.026	9.80E-01	9.96E-01
5555	ILMN_1847363	LOC731835	Homo sapiens hypothetical protein LOC731835 (LOC731835), transcript variant 1, mRNA.	6.422	6.036	-0.003	-0.026	9.79E-01	9.96E-01
85465	ILMN_2400292	MAPK9	mitogen-activated protein kinase 9 (MAPK9), transcript variant 1, mRNA.	8.762	8.561	-0.003	-0.027	9.79E-01	9.96E-01
466855	ILMN_2141453	RPL18A	ribosomal protein L18a (RPL18A), mRNA.	15.124	15.133	-0.002	-0.027	9.79E-01	9.96E-01
7458	ILMN_1725018	LOC654220	similar to hypothetical protein FLJ33915, transcript variant 1, mRNA.	6.748	6.650	-0.002	-0.027	9.79E-01	9.96E-01
1235	ILMN_1680618	MYC	v-myc myelocytomatosis viral oncogene homolog (a) (MYC), transcript variant 1, mRNA.	12.491	11.974	-0.540	-5.601	4.96E-05	6.47E-03
100589	ILMN_1712803	CCNB1	cyclin B1 (CCNB1), mRNA.	9.688	8.718	-0.771	-5.625	4.75E-05	6.27E-03
14562	ILMN_3244929	LOC10013316	similar to breakpoint cluster region (LOC100133163), transcript variant 1, mRNA.	9.851	9.451	-0.507	-5.661	4.44E-05	6.01E-03
12544	ILMN_1815184	ASPM	asp (abnormal spindle) homolog, microcephaly associated (ASPM), transcript variant 1, mRNA.	10.494	9.725	-0.671	-5.784	3.54E-05	4.95E-03
1005687	ILMN_1651433	DCCK	deoxycytidine kinase (DCCK), mRNA.	10.316	9.932	-0.533	-5.930	2.71E-05	3.94E-03
85465	ILMN_1802951	CDC41	cell division cycle associated 1 (CDC41), transcript variant 1, mRNA.	7.670	6.816	-0.695	-6.279	1.44E-05	2.35E-03
466855	ILMN_1788251	SNN	stannin (SNN), mRNA.	7.747	6.851	-0.658	-6.323	1.34E-05	2.19E-03
7458	ILMN_1671933	CLCC1	chloride channel CLIC-like 1 (CLCC1), transcript variant 1, mRNA.	9.163	8.248	-0.724	-6.417	1.13E-05	1.93E-03
1235	ILMN_1783621	CMPK2	cytidine monophosphate (UMP-CMP) kinase 2, mitochondrial (CMPK2), transcript variant 1, mRNA.	8.833	8.327	-0.706	-6.447	1.07E-05	1.88E-03
1458	ILMN_3239771	DLGAP5	discs, large (Drosophila) homolog-associated protein 5 (DLGAP5), transcript variant 1, mRNA.	10.288	9.078	-1.075	-6.540	9.13E-06	1.71E-03
64477	ILMN_1658847	MGC61598	similar to ankyrin-repeat protein Nrarp (MGC61598), transcript variant 1, mRNA.	9.894	9.275	-0.570	-6.682	7.15E-06	1.42E-03
12544	ILMN_1666545	GCNT1	glucosaminyl (N-acetyl) transferase 1, core 2 (beta-1) (GCNT1), transcript variant 1, mRNA.	9.078	8.143	-0.879	-6.713	6.77E-06	1.37E-03
9854777	ILMN_1749829	DLGAP5	discs, large (Drosophila) homolog-associated protein 5 (DLGAP5), transcript variant 1, mRNA.	10.167	9.184	-1.006	-7.167	3.15E-06	7.70E-04

Down-regulated
in treated

Enrichment score

Enrichment score from Mootha et al.
2003:

N: gene list length (20000)

G: # of genes in the gene-set (50)

if the gene i is in the gene-set:

$$X_i = \sqrt{(N-G) / G}$$

$$X_i = \sqrt{(20000-50) / 50}$$

$$X_i = 20$$

if the gene i is in not in the gene-set:

$$X_i = -\sqrt{G / (N-G)}$$

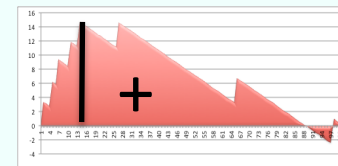
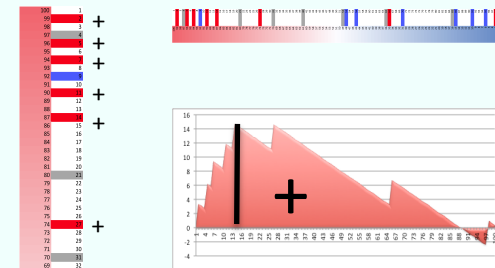
$$X_i = -\sqrt{50 / (20000-50)}$$

$$X_i = -0.05$$

The running sum of X is calculated
until we reach the last gene in the
ranked list.

UP in
treated

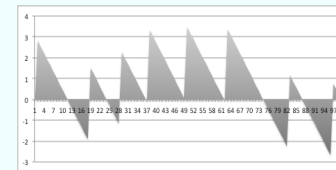
Ranked
gene list



Enrichment
Score

Gene-set 1 (17 genes)

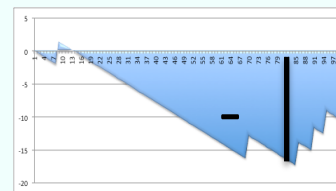
Enriched
in
"treated"



Enrichment
Score

Gene-set 2

"not
enriched"



Enrichment
Score

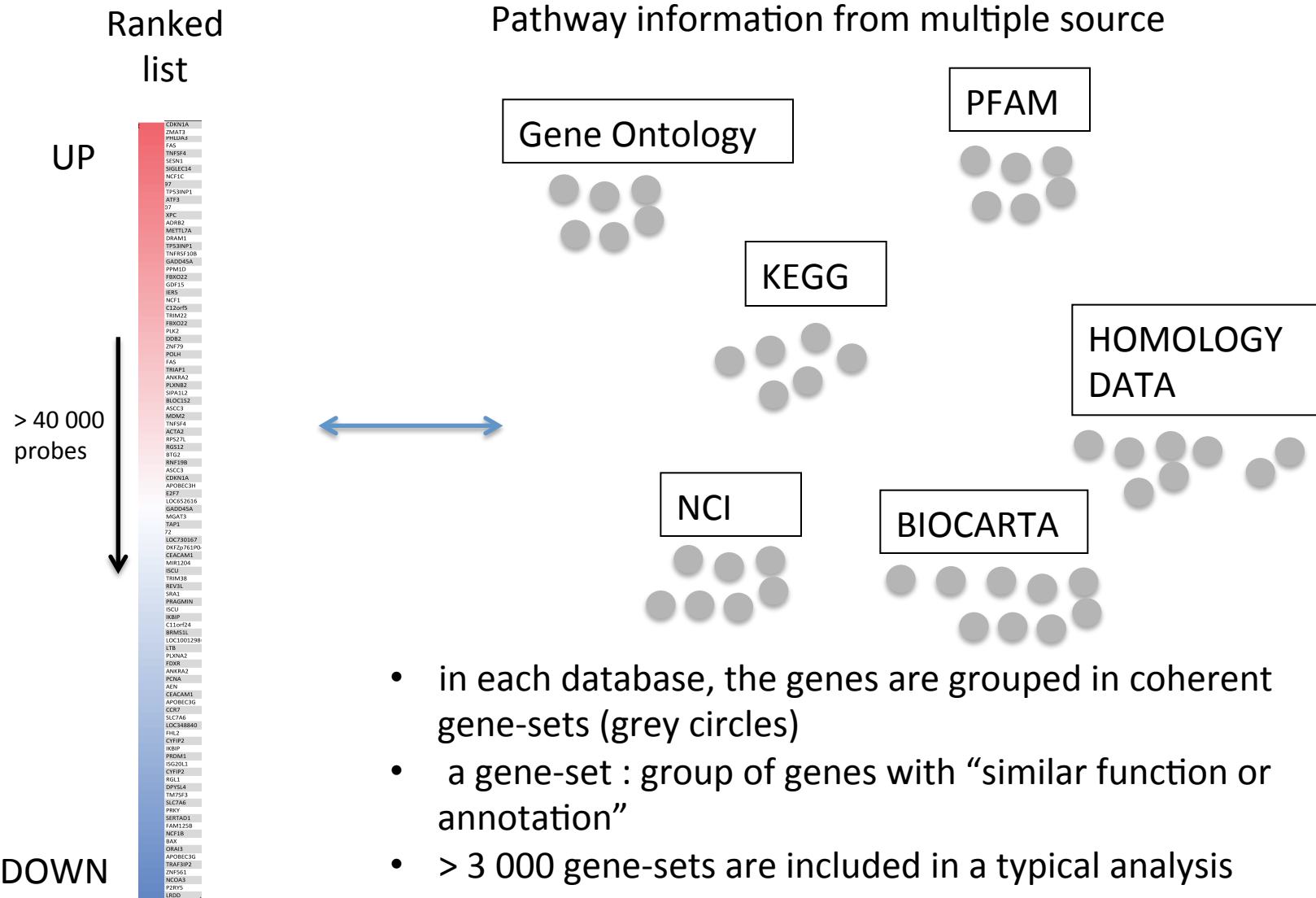
Gene-set 3 (22 genes)

Enriched in
"non
treated"



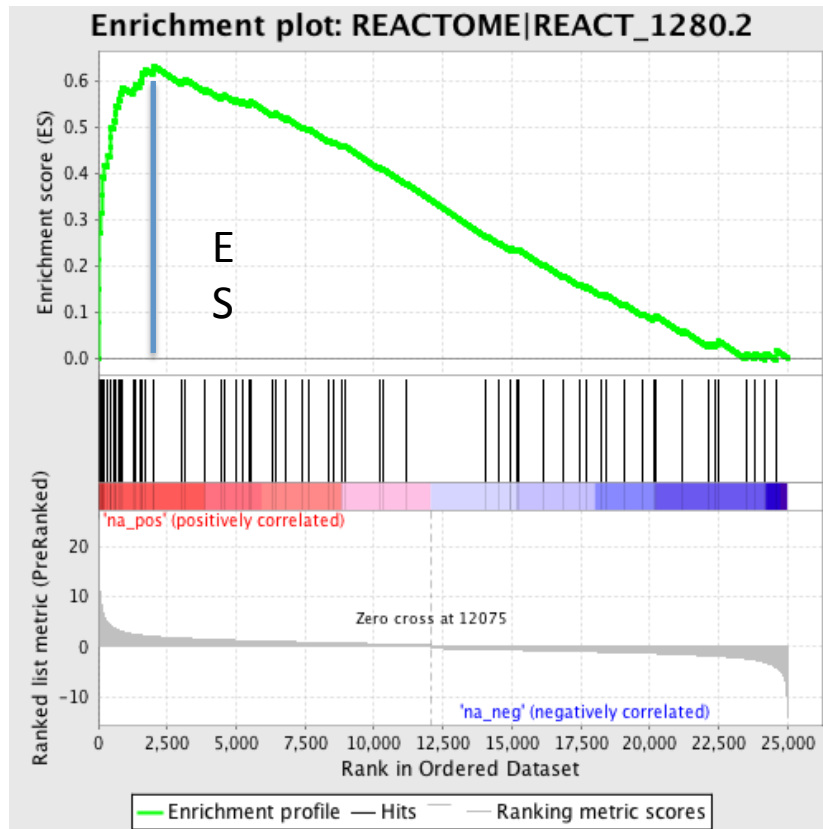
DOWN in
treated

Calculate the enrichment score for each gene-set

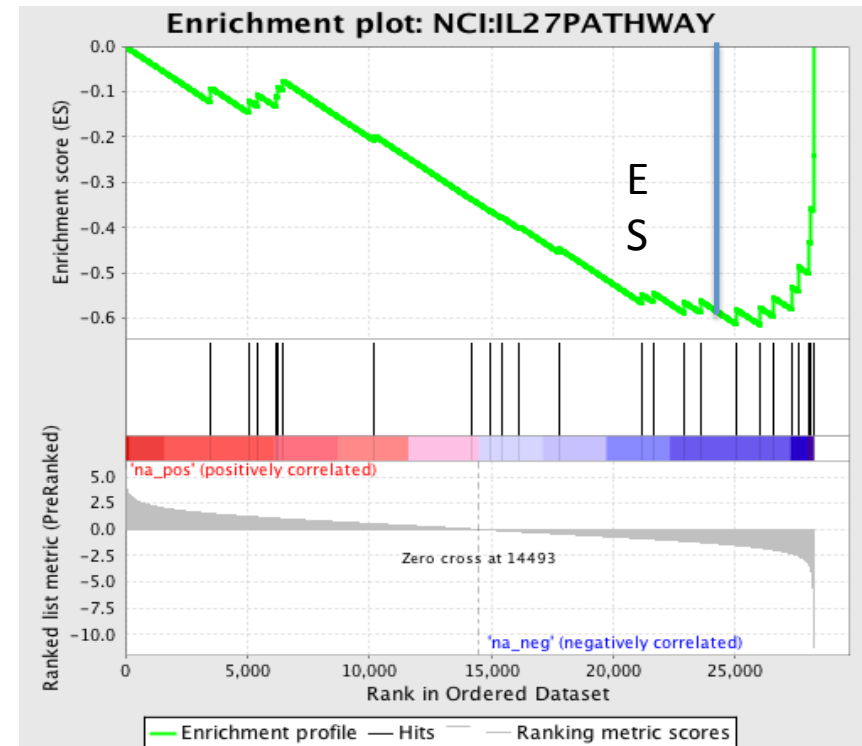


- in each database, the genes are grouped in coherent gene-sets (grey circles)
- a gene-set : group of genes with “similar function or annotation”
- > 3 000 gene-sets are included in a typical analysis

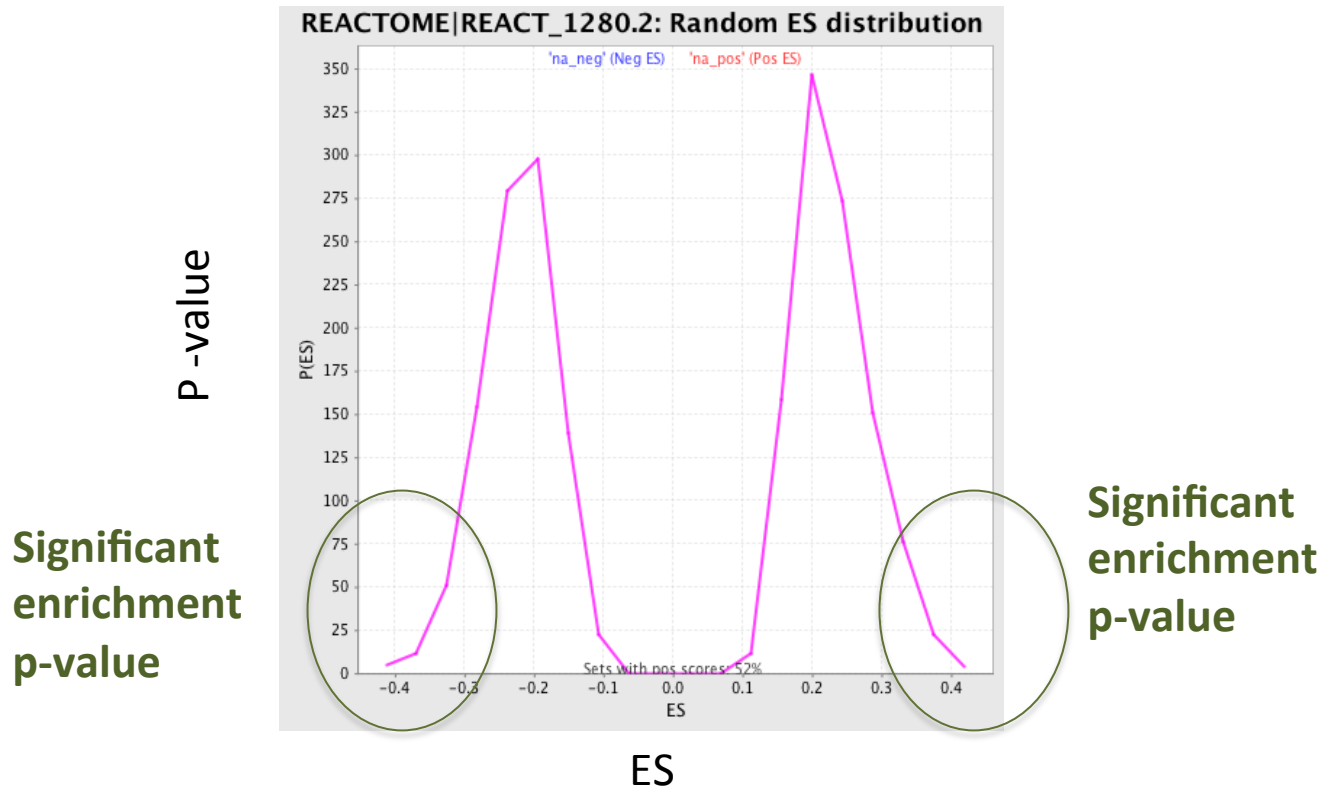
positive enriched geneset



negative enriched geneset



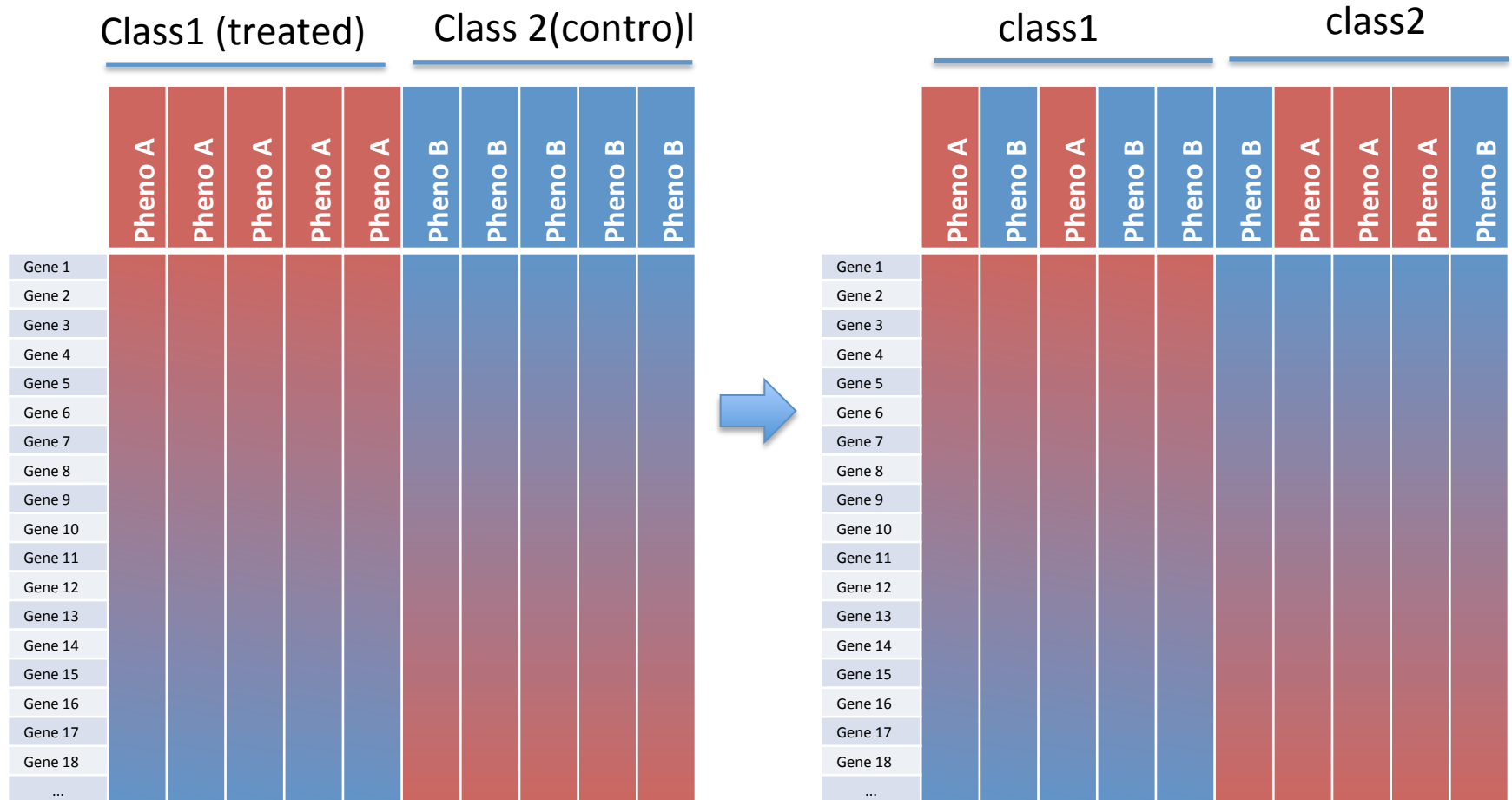
1. Null distribution



2. Correct p-value for multiple hypothesis testing (FDR)

Phenotype Permutation

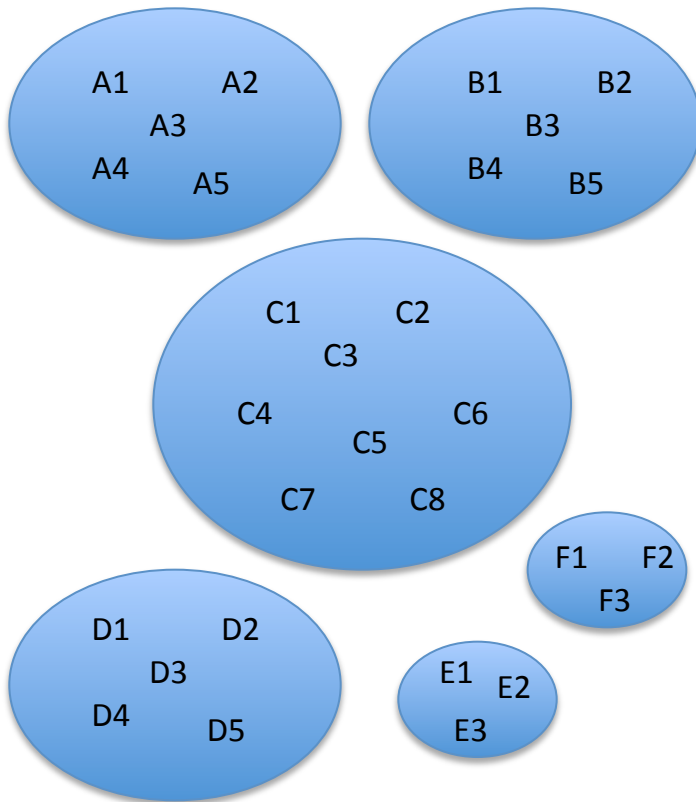
shuffling phenotype labels (1000-2000 times)



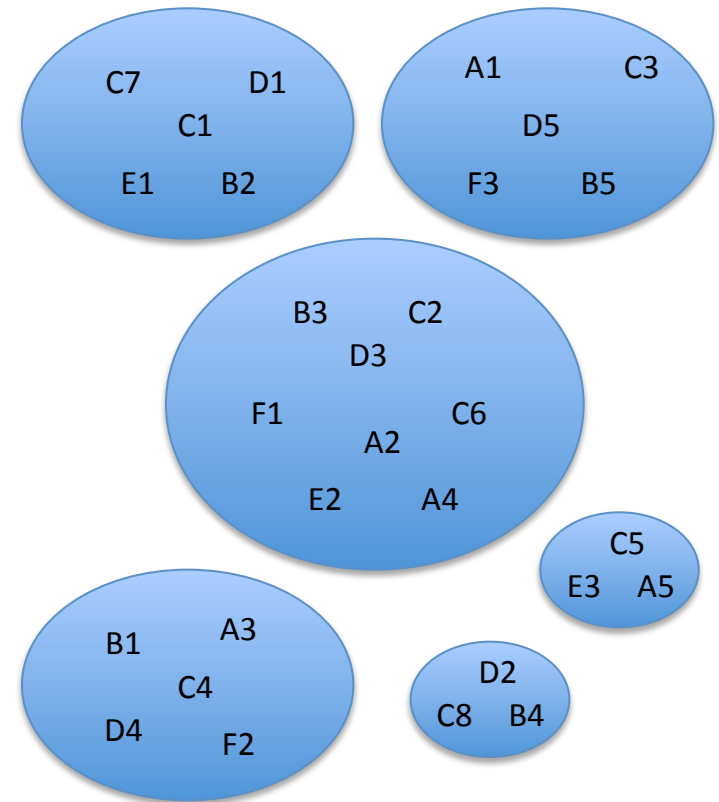
Geneset Permutation

1000-2000 times

original geneset



random geneset



GSEA Load Data dialog

GSEA v2.05 (Gene set enrichment analysis -- Broad Institute)

Steps in GSEA analysis

- Load data
- Run GSEA
- Leading edge analysis

Gene set tools

- Chip2Chip mapping
- Browse MSigDB

Analysis history

GSEA reports

Processes: click 'status' field for results

Name	Status

Show results folder

Home Load data

Load data: Import data into the application

Method 1:

Method 2:

Method 3: drag and drop files here

absentNotch_vs_wtNotch.gct
c2_and_cellmap.gmt
absentNotch_vs_wtNotch.cls

Supported file formats

Dataset: *res* or *gct* (Broad/MIT),
pcl (Stanford)
txt (tab-delim text)

Phenotype labels: *cls*

Gene sets: *gmx* or *gmt*

Recently used files (double click to load, right click for more options)

- ../ChipPlatforms/Mouse430_2.chip
- ../gsea_data/Mouse430_2.chip
- ../gsea_data/Mouse430_2.chip
- ../gsea_data/Illumina_MusRef8_v1_1.chip
- ../gsea_data/phenotype.cls
- ../Workshop_GSEA/absentNotch_vs_wtNotch.cls
- ../Set_02/set_02.gct
- ../gsea_data/set_06_linear.gct
- ../gsea_data/Set_03_wtNotch_versus_normalThy.gct
- ../gsea_data/Set_03_absentNotch_versus_wtNotch.gct
- ../Workshop_GSEA/absentNotch_vs_wtNotch.gct
- ../GeneSets/c5.all.v2.5.symbols.gmt
- ../gsea_data/c5.all.v2.5.symbols.gmt
- ../edb/gene_sets.gmt
- ../GeneSets/c5.bp.v2.5.symbols.gmt
- ../gsea_data/c5.all.v2.5.symbols.gmt
- ../gsea_data/c5.all.v2.5.symbols.gmt
- ../gsea_data/c2.all.v2.5.symbols.gmt
- ../gsea_data/hgnc_symbol_cellmap.gmt
- ../Workshop_GSEA/c2_and_cellmap.gmt

Object cache (objects already loaded & ready for use, right click for more options)

- Objects in memory [shift-click to expand all]
 - Gene set databases
 - c2_and_cellmap.gmt [1902 gene sets]
 - Phenotypes
 - absentNotch_vs_wtNotch.cls [10 samples(4,6)]
 - absentNotch_vs_wtNotch.cls#wtNotch_versus_
 - absentNotch_vs_wtNotch.cls#absentNotch_ver
 - absentNotch_vs_wtNotch.cls#wtNotch [4 samp
 - absentNotch_vs_wtNotch.cls#absentNotch [6 :
 - Datasets
 - absentNotch_vs_wtNotch [18617x10 (ann: 18

3:36:49 PM 2269 [INFO] Loading ... 3 files c2_and_cellmap.gmt absentNotch_vs_wtNotch.cls absentNotch_vs... 37M of 75M

Run GSEA dialog

GSEA v2.05 (Gene set enrichment analysis -- Broad Institute)

Steps in GSEA analysis

- Load data
- Run GSEA**
- Leading edge analysis

Gene set tools

- Chip2Chip mapping
- Browse MSigDB

Analysis history

GSEA reports

Processes: click 'status' field for results

Name	Status
------	--------

Show results folder

Gsea: Set parameters and run enrichment tests

Required fields

- Expression dataset: absentNotch_vs_wtNotch [18617x10 (ann: 18617,10,chip na)]
- Gene sets database: rs/revilo/StemCellProject/Workshop_GSEA/c2_and_cellmap.gmt
- Number of permutations: 1000
- Phenotype labels: SEA/absentNotch_vs_wtNotch.cls#absentNotch_versus_wtNotch
- Collapse dataset to gene symbols: false
- Permutation type: gene_set
- Chip platform(s):

Basic fields

- Analysis name: example_NOTCH_absent_vs_wt
- Enrichment statistic: weighted
- Metric for ranking genes: log2_Ratio_of_Classes
- Gene list sorting mode: real
- Gene list ordering mode: descending
- Max size: exclude larger sets: 500
- Min size: exclude smaller sets: 15
- Save results in this folder: /Users/revilo/gsea_home/output/jan25

Advanced fields

Reset Last Command Low (cpu usage) Run

3:46:36 PM 2269 [INFO] Loading ... 3 files c2_and_cellmap.gmt absentNotch_vs_wtNotch.cls absentNotch_vs_wtNotch.gct Files loaded successfully: 3 / 3 There... 45M of 75M

Expression dataset (.txt)

Entrez_Gene_ID	DESCRIPTION	treated1	control1	treated2	control2	treated3	control3
285741	LOC285741 ::PREDICTED: misc_RNA (LOC285741)	11.941118	12.611188	8.806047	12.565253	10.38005	9.248268
100008589	LOC100008589 :: 28S ribosomal RNA (LOC100008589)	12.402481	12.600485	11.416079	11.966387	12.392032	11.706466
442727	LOC442727 ::PREDICTED: misc_RNA (LOC442727)	15.125171	15.756729	13.360757	15.476214	14.241237	13.291751
7915	ALDH5A1 :: aldehyde dehydrogenase 5 family class A member 1	9.550562	9.901636	9.46661	9.802373	9.546112	9.56268
29094	HSPC159 :: galectin-related protein (HSPC159)	8.320405	7.979131	8.538786	8.406312	8.245906	8.062359
55486	PARL :: presenilin associated, rhomboid-like	9.578815	9.654142	9.898891	10.751403	9.794695	9.681501
79590	MRPL24 :: mitochondrial ribosomal protein	8.762275	9.094272	9.400059	9.782166	8.980708	9.095544
100130541	LOC100130541 ::PREDICTED: similar to hCG	7.9757347	7.5204177	7.883453	7.1043377	8.2067995	8.135658
55861	DBNDD2 :: dysbindin (dystrobrevin binding protein 2)	7.6873574	7.912635	7.979784	8.256096	8.202624	8.424054
841	CASP8 :: caspase 8, apoptosis-related cysteine protease	7.217502	7.244122	7.308674	7.0536304	7.543329	7.321113
9595	PSCDBP :: pleckstrin homology, Sec7 and coiled-coil domain containing	8.920641	9.557003	9.01201	9.311855	8.998903	8.831511
132	ADK :: adenosine kinase (ADK), transcript variant 1	7.3660574	7.609104	7.6317477	7.130101	7.468771	7.3359246
641975	LOC641975 ::PREDICTED: hypothetical protein	8.157521	8.624233	8.089965	8.85631	8.054138	8.382164
647089	LOC647089 ::PREDICTED: hypothetical LOC647089	8.280826	9.240892	8.153017	9.265228	8.254154	8.593111
3306	HSPA2 :: heat shock 70kDa protein 2 (HSPA2)	6.955488	7.772578	7.289512	7.399042	7.492552	7.3760567
572558	LOC572558 :: hypothetical locus LOC572558	9.316893	8.592318	8.541199	8.780995	8.681224	8.508745
7332	UBE2L3 :: ubiquitin-conjugating enzyme E2L3	8.6934805	8.669022	8.129649	9.024807	8.188721	8.083305

tab delimited file (.txt)

with column names

First column: entrez gene ID or official gene symbol

Second column: gene name /description

Additional columns: log2 normalized data of each replicate

Gene Set Database (.GMT)

- MSigDB from the GSEA website
- or .GMT file available for the Bader lab (<http://baderlab.org/GeneSets>)
- or your custom .GMT file

HUMANCYC MGLDLCTANA-PWY	methylglyoxal degradation VI	197257							
HUMANCYC ARGININE-SYN4-PWY	arginine biosynthesis IV	5009	4942	2746	2747	445	435		
HUMANCYC SERDEG-PWY	L-serine degradation	113675	10993						
HUMANCYC NONOXIPENT-PWY	pentose phosphate pathway (non-oxidative brai	7086	8277	729020	84076	6120	6888	22934	
HUMANCYC PWY-6405	Rapoport-Luebering glycolytic shunt	100290936	5224	669	9562	5223	441531		
HUMANCYC PWY-6342	noradrenaline and adrenaline degradation	125	5409	224	217	4129	4128	1312	
HUMANCYC PWY66-201	nicotine degradation II	2328	1548	2329	29785	11185	260293	316	22952
HUMANCYC HYDROXYPRODEG-PWY	4-hydroxyproline degradation I	8659							
HUMANCYC PWY66-341	cholesterol biosynthesis I	6713	4047	1595	10682	1718	1717	51478	6307
HUMANCYC LIPASYN-PWY	phospholipases	89869	81579	5338	50640	5335	5336	26279	5337
HUMANCYC DETOX1-PWY	superoxide radicals degradation	7306	847	6649	6648	6647			
HUMANCYC GLUTAMATE-SYN2-PWY	glutamate biosynthesis II	2746	2747						
HUMANCYC PWY-6554	1D-<i>myo</i>-inositol hexakisphosphate biosyn	3705	64768	253430					
HUMANCYC PWY66-5	superpathway of cholesterol biosynthesis	6713	4047	1595	10682	1718	1717	51478	6307
HUMANCYC PWY66-4	cholesterol biosynthesis III (via desmosterol)	6713	4047	1595	10682	1718	1717	51478	6307
HUMANCYC PWY66-3	cholesterol biosynthesis II (via 24,25-dihydrolan	6713	4047	1595	10682	1718	1717	51478	6307
HUMANCYC TYRFUMCAT-PWY	tyrosine degradation I	6898	3242	2954	3081	2184			

Tab delimited file (.txt) with no column names

First columns: gene-set name

Second columns: gene-set description

Additional columns: Entrez Gene Id or Official Gene Symbol

Phenotype label .cls

```
6 2 1
# treated control
1 0 1 0 1 0
```

Create a tab delimited file, save as .txt and then change to .cls

6: number of replicates

2: number of samples

1: always 1 , don't change

names of the 2 classes

Next line: corresponds to the columns in the expression dataset file: treated
control treated control....

Chip Annotation file (chip)

- necessary if probeset-IDs (AffyID, Illumina, etc.) are used in the expression file (or ranked gene list)
- not necessary if identifiers in the expression dataset and in the .GMT file are Entrez Gene IDs or Official Gene Symbol.
- maps probeset-ID to HUGO gene symbols
(MSigDB Genesets use HUGO gene symbols)
e.g: 1418633_at → NOTCH1
1421205_at → ATM
1418102_at → HES1
- use parameter “collapse = true” to collapse expression values for multiple probesets that match to the same gene

Important GSEA Parameters

- permutation type:
 - “phenotype” only if ≥ 7 samples per class are available
 - “gene_set” works also with fewer samples
- collapse only if chip-annotation file is used
- collapsing mode
- scoring scheme: weighted
- metric
 - Ratio_of_Classes  use with log2 expression data
 - log2_Ratio_of_Classes  use with linear expression data
 - t-Test
 - Signal2Noise
- Min/Max size of Gene Sets

Metrics for ranking genes

For categorical phenotypes, GSEA determines a gene's mean expression value for each phenotype and then uses one of the following metrics to calculate the gene's differential expression with respect to the two phenotypes. To use median rather than mean expression values, set the *Median for class metrics* parameter to True, as described above.

- Signal2Noise (default) uses the difference of means scaled by the standard deviation. **Note:** You must have at least three samples for each phenotype to use this metric.

$$\frac{\mu_A - \mu_B}{\sigma_A + \sigma_B}$$

where μ is the mean and σ is the standard deviation; σ has a minimum value of $.2 * \text{absolute}(\mu)$, where $\mu=0$ is adjusted to $\mu=1$. The larger the signal-to-noise ratio, the larger the differences of the means (scaled by the standard deviations); that is, the more distinct the gene expression is in each phenotype and the more the gene acts as a "class marker."

- tTest uses the difference of means scaled by the standard deviation and number of samples. **Note:** You must have at least three samples for each phenotype to use this metric.

$$\frac{\mu_A - \mu_B}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}}$$

where μ is the mean, n is the number of samples, and σ is the standard deviation; σ has a minimum value of $.2 * \text{absolute}(\mu)$, where $\mu=0$ is adjusted to $\mu=1$. The larger the tTest ratio, the more distinct the gene expression is in each phenotype and the more the gene acts as a "class marker."

- Ratio_of_Classes (also referred to as fold change) uses the ratio of class means to calculate fold change for natural scale data:

$$\frac{\mu_A}{\mu_B}$$

where μ is the mean. The larger the fold change, the more distinct the gene expression is in each phenotype and the more the gene acts as a "class marker."

- Diff_of_Classes uses the difference of class means to calculate fold change for log scale data:

$$\mu_A - \mu_B$$

where μ is the mean. The larger the fold change, the more distinct the gene expression is in each phenotype and the more the gene acts as a "class marker."

- log2_Ratio_of_Classes uses the log2 ratio of class means to calculate fold change for natural scale data:

$$\log_2 \left(\frac{\mu_A}{\mu_B} \right)$$

where μ is the mean. This is the recommended statistic for calculating fold change for natural scale data.

GSEAPreranked

Why?

- GSEA only has a limited number of ranking statistics: (Signal2Noise, Ratio_of_Classes, log2_Ratio_of_Classes, t-Test, ...)
- GSEAPreranked starts with a user-ranked gene list

Examples:

- SAM
- ANOVA
- any other algorithm that can score and rank the genes according a 1- or 2-class model (e.g. TREAT)

McCarthy DJ, Smyth GK (2009) *Testing significance relative to a fold-change threshold is a TREAT*. Bioinformatics (Oxford, England) 25: 765-71. Available: <http://www.ncbi.nlm.nih.gov/pubmed/19176553>.

Use the GSEA pre-ranked option to use GSEA with your ranked gene list (recommended)

- 1) Load Data -> Browse for files
- 2) Tools (menu bar) -> GSEAPreRanked

GSEA v2.07 (Gene set enrichment analysis -- Broad Institute)

Home Run Gsea Run Gsea on a Pre-Ranked gene list Load data

GseaPreranked: Run GSEA on a pre-ranked (with external tool) gene list

Steps in GSEA analysis

- Load data
- Run GSEA
- Leading edge analysis

Gene set tools

- Chip2Chip mapping
- Browse MSigDB
- Analysis history

GSEA reports

Processes: click 'status' field for results

Name	Status
1 GseaPreranked	Running

Show results folder

Required fields

- Gene sets database: /Users/veroniquevoisin/GO_K_NCI_BIOC_PF_Hs_eg.GMT
- Number of permutations: 2000
- Ranked List: shp53_NT_vs_RFP_NT_s_u_rnk [12812 names]
- Collapse dataset to gene symbols: false
- Chip platform(s):

Basic fields

- Analysis name: my_analysis
- Enrichment statistic: weighted
- Max size: exclude larger sets: 500
- Min size: exclude smaller sets: 15
- Save results in this folder: veroniquevoisin/Documents/John_Dick/Michael M/JD02Analysis

Advanced fields

- Collapsing mode for probe sets => 1 gene: Max_probe
- Normalization mode: meandiv
- Omit features with no symbol match: true
- Make detailed gene set report: true
- Plot graphs for the top sets of each phenotype: 20
- Seed for permutation: timestamp
- Make a zipped file with all reports: false

Reset Last Command Low (cpu usage) Run

2:41:31 PM 3881 [INFO] Done preproc for smaller than: 15 581M of 1011M

- Gene set database (.GMT file)
- Ranked file (.RNK):
Tab delimited file,
No columns names,
The first column corresponds to EntrezGeneIDs
The second column contains ranking values

Optional but recommended: no duplicate Entrez
Gene IDs or Official Gene Symbol

Example from gene expression data:

- 1) Choose Entrez Gene IDs as identifier
- 2) Calculate the t statistic between the class1 (treated) and the class 2 (control)
- 3) Rank the list by the absolute t value and remove the duplicates
- 4) Rank the unique list by the t value in decreasing order
- 5) Save the tab delimited file as a .RNK file

1647	19.01220882
8493	16.27436286
9518	14.94607615
330	14.21269431
51278	13.87147359
7508	13.69382804
23612	13.68238497
57103	13.6756894
22954	13.59165777
27244	13.38443774
355	12.84444568
7292	12.45805175
467	11.76706505
51499	11.76223998
7128	11.35098834
8795	11.31987394
1026	10.93251033
50650	10.48532767
25840	10.43577658
94241	10.43395545
55332	10.29659906
4193	10.27230371
154	10.21171612
4814	10.03746692
127544	9.928558704
64782	9.87041321
4033	9.687226741
57763	9.366840962
55924	9.328431674
144455	9.290170914
8793	9.281064381
220001	9.055312974
7633	9.019313933
26263	8.817529242
639	8.816647111
5616	8.697830946
29970	8.687011833
51313	8.621949683
282991	8.615689952

GSEA results

GSEA Report for Dataset Set_03_absentNotch

Enrichment in phenotype: **absentNotch** (6 samples)

- 1016 / 1623 gene sets are upregulated in phenotype **absentNotch**
- 16 gene sets are significant at FDR < 25%
- 34 gene sets are significantly enriched at nominal pvalue < 1%
- 104 gene sets are significantly enriched at nominal pvalue < 5%
- [Snapshot](#) of enrichment results
- Detailed [enrichment results in html](#) format
- Detailed [enrichment results in excel](#) format (tab delimited text)
- [Guide to](#) interpret results

Enrichment in phenotype: **wtNotch** (4 samples)

- 607 / 1623 gene sets are upregulated in phenotype **wtNotch**
- 9 gene sets are significantly enriched at FDR < 25%
- 14 gene sets are significantly enriched at nominal pvalue < 1%
- 43 gene sets are significantly enriched at nominal pvalue < 5%
- [Snapshot](#) of enrichment results
- Detailed [enrichment results in html](#) format
- Detailed [enrichment results in excel](#) format (tab delimited text)
- [Guide to](#) interpret results

Dataset details

- The dataset has 18617 features (genes)
- No probe set => gene symbol collapsing was requested, so all 18617 features are represented

Gene set details

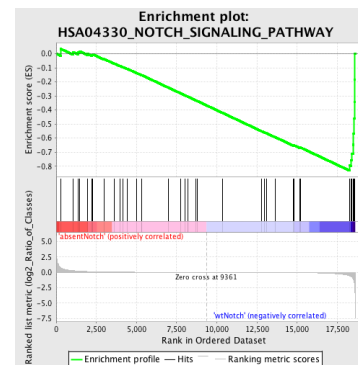
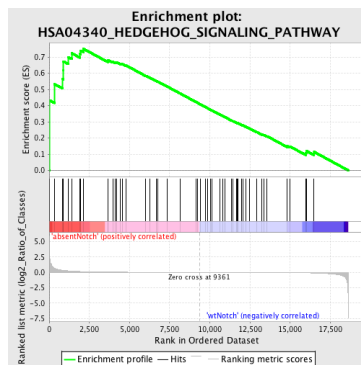
- Gene set size filters (min=10, max=500) resulted in filtering out 279 / 1902 gene sets
- The remaining 1623 gene sets were used in the analysis
- List of [gene sets used and their sizes](#) (restricted to features in the specified dataset)

Gene markers for the **absentNotch versus wtNotch** comparison

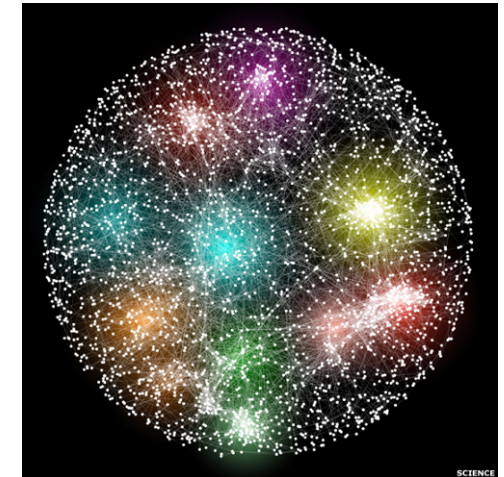
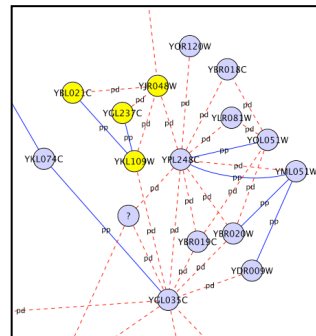
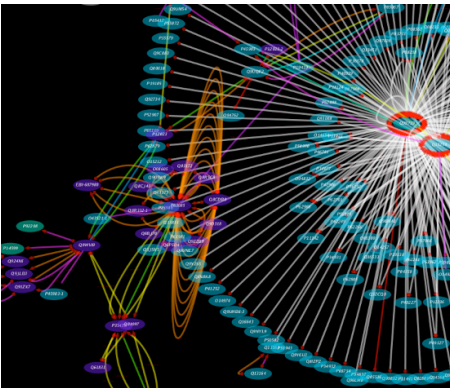
- The dataset has 18617 features (genes)
- # of markers for phenotype **absentNotch**: 9361 (50.3%) with correlation area > 0.1
- # of markers for phenotype **wtNotch**: 9256 (49.7%) with correlation area > 0.1
- Detailed [rank ordered gene list](#) for all features in the dataset
- [Heat map and gene list correlation](#) profile for all features in the dataset

GSEA results

NAME	GS follow link to MSigDB	GS DETAILS	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MA	LEADING EDGE	
NCI:P53DOWNSTREAMPATHW	NCI:P53DOWNSTREAMPATHWAY	Details ...	112	0.7174656	2.8111715	0	0	0	1437	tags=42%, list=11%, signal=47%	
KEGG:04115	KEGG:04115	Details ...	59	0.73530716	2.5667064	0	0	0	193	tags=31%, list=2%, signal=31%	
BIOC:P53PATHWAY	BIOC:P53PATHWAY	Details ...	16	0.81812257	2.2041845	0	0.00409315	0.011	163	tags=38%, list=1%, signal=38%	
PF00020	PF00020	Details ...	18	0.7858835	2.1676812	0	0.00475014	0.017	417	tags=33%, list=3%, signal=34%	
NCI:S1P_S1P3_PATHWAY	NCI:S1P_S1P3_PATHWAY	Details ...	25	0.7025801	2.0786188	0	0.02628367	0.112	1477	tags=48%, list=12%, signal=54%	
GO:0005125	GO:0005125	Details ...	99	0.5381393	2.0726626	0	0.02404444	0.123	1996	tags=32%, list=16%, signal=38%	
PF00531	PF00531	Details ...	26	0.6951619	2.06757	0	0.02228651	0.131	112	tags=23%, list=1%, signal=23%	
GO:0042787	GO:0042787	Details ...	16	0.774162	2.0471377	8.70E-04	0.02971941	0.1945	44	tags=19%, list=0%, signal=19%	
GO:0030330	GO:0030330	Details ...	31	0.65348965	2.0385714	0	0.03064614	0.223	1239	tags=42%, list=10%, signal=46%	
GO:0009411	GO:0009411	Details ...	55	0.5804456	2.03297	0	0.03021042	0.2385	1239	tags=31%, list=10%, signal=34%	
KEGG:04060	KEGG:04060	Details ...	160	0.48535216	2.0152059	0	0.03550203	0.304	1043	tags=21%, list=8%, signal=22%	
GO:0012502	GO:0012502	Details ...	271	0.45233214	1.9946653	0	0.04509291	0.388	1979	tags=25%, list=15%, signal=29%	
GO:0006917	GO:0006917	Details ...	271	0.45233214	1.9861351	0	0.0470538	0.424	1979	tags=25%, list=15%, signal=29%	
GO:0007050	GO:0007050	Details ...	95	0.51430047	1.9729838	0	0.05305111	0.4865	1391	tags=26%, list=11%, signal=29%	
GO:0010942	GO:0010942	Details ...	369	0.43251514	1.9597678	0	0.06030389	0.55	1627	tags=21%, list=13%, signal=24%	
GO:0043068	GO:0043068	Details ...	365	0.43386024	1.956914	0	0.05905404	0.5675	1627	tags=21%, list=13%, signal=24%	
GO:0000718	GO:0000718	Details ...	21	0.68914753	1.9563942	0.00166667	0.05584443	0.57	1527	tags=33%, list=12%, signal=38%	
GO:0002831	GO:0002831	Details ...	29	0.63304216	1.9546326	0	0.05398624	0.58	3455	tags=62%, list=27%, signal=85%	
GO:0043065	GO:0043065	Details ...	364	0.4323633	1.9530113	0	0.0524402	0.5895	1627	tags=21%, list=13%, signal=24%	
NCI:ARF6_TRAFFICKINGPATH	NCI:ARF6_TRAFFICKINGPATHWAY	Details ...	41	0.5889985	1.9386604	0	0.05997115	0.6645	3384	tags=51%, list=26%, signal=69%	



- open source bioinformatics software platform
- for molecular interaction networks
- visualization and analysis



5) Enrichment Map (EM)

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

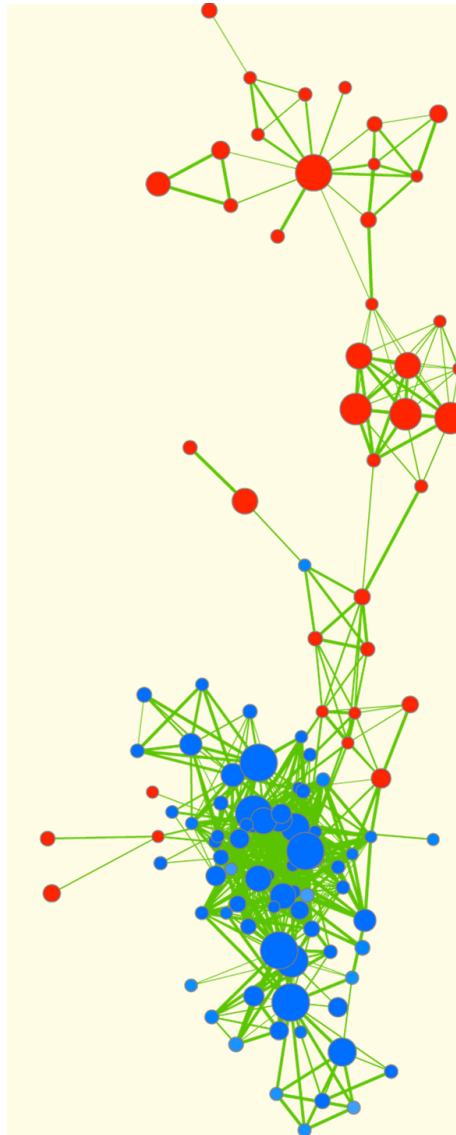
Enriched in...

● treated

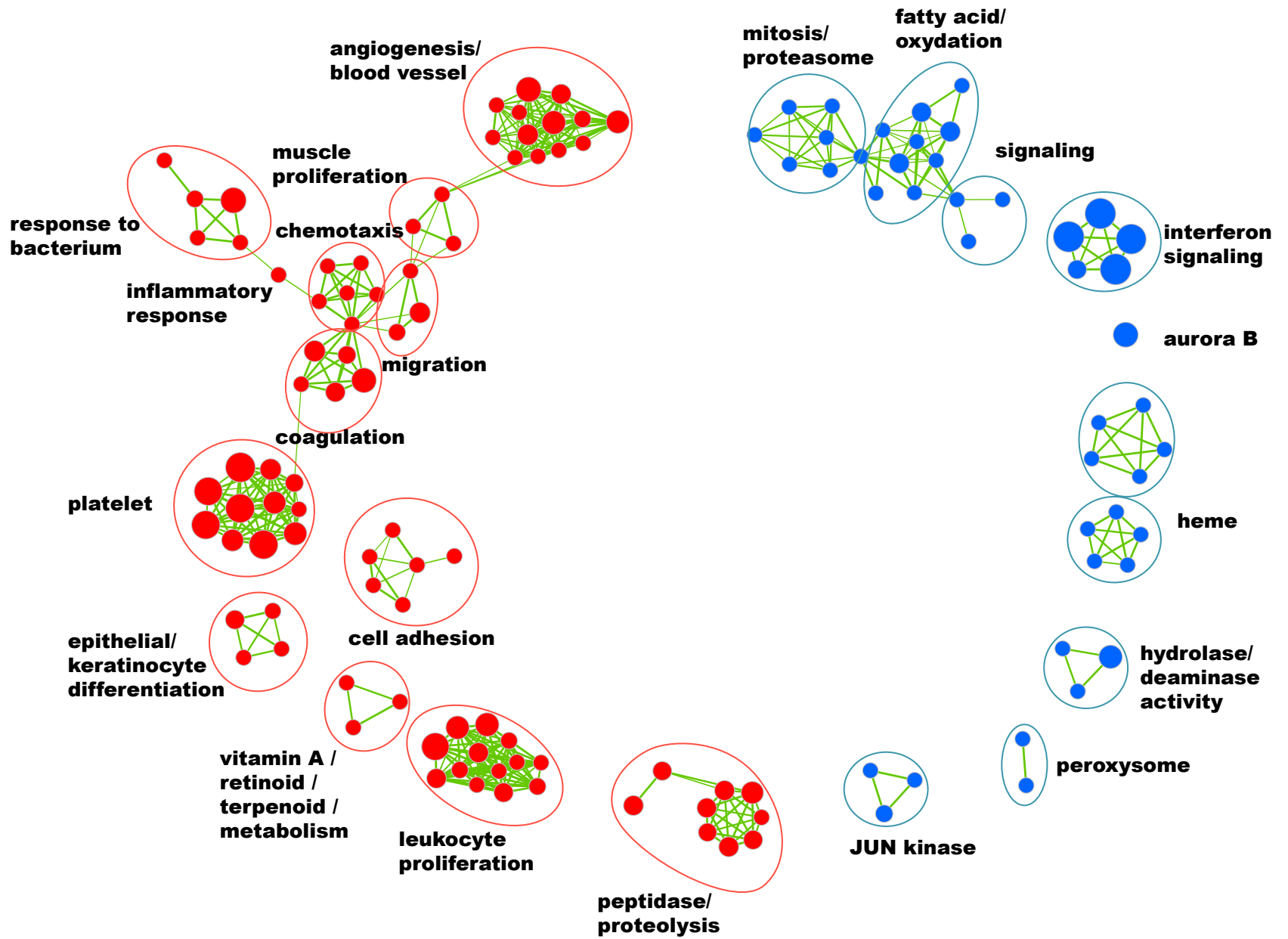
● non-treated

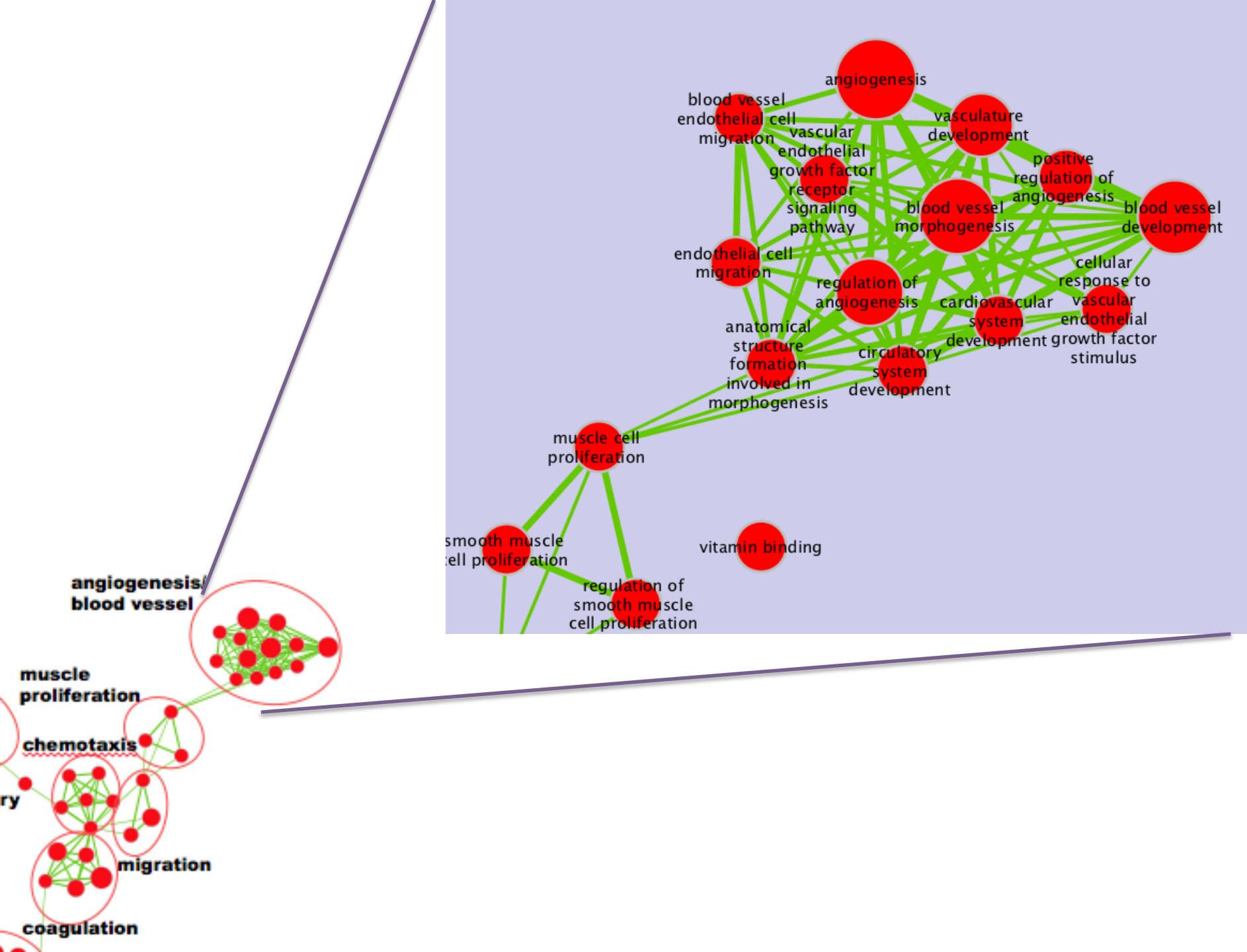
— gene overlap between 2 gene-sets

color intensity
depends on the
enrichment p-value



gene overlap



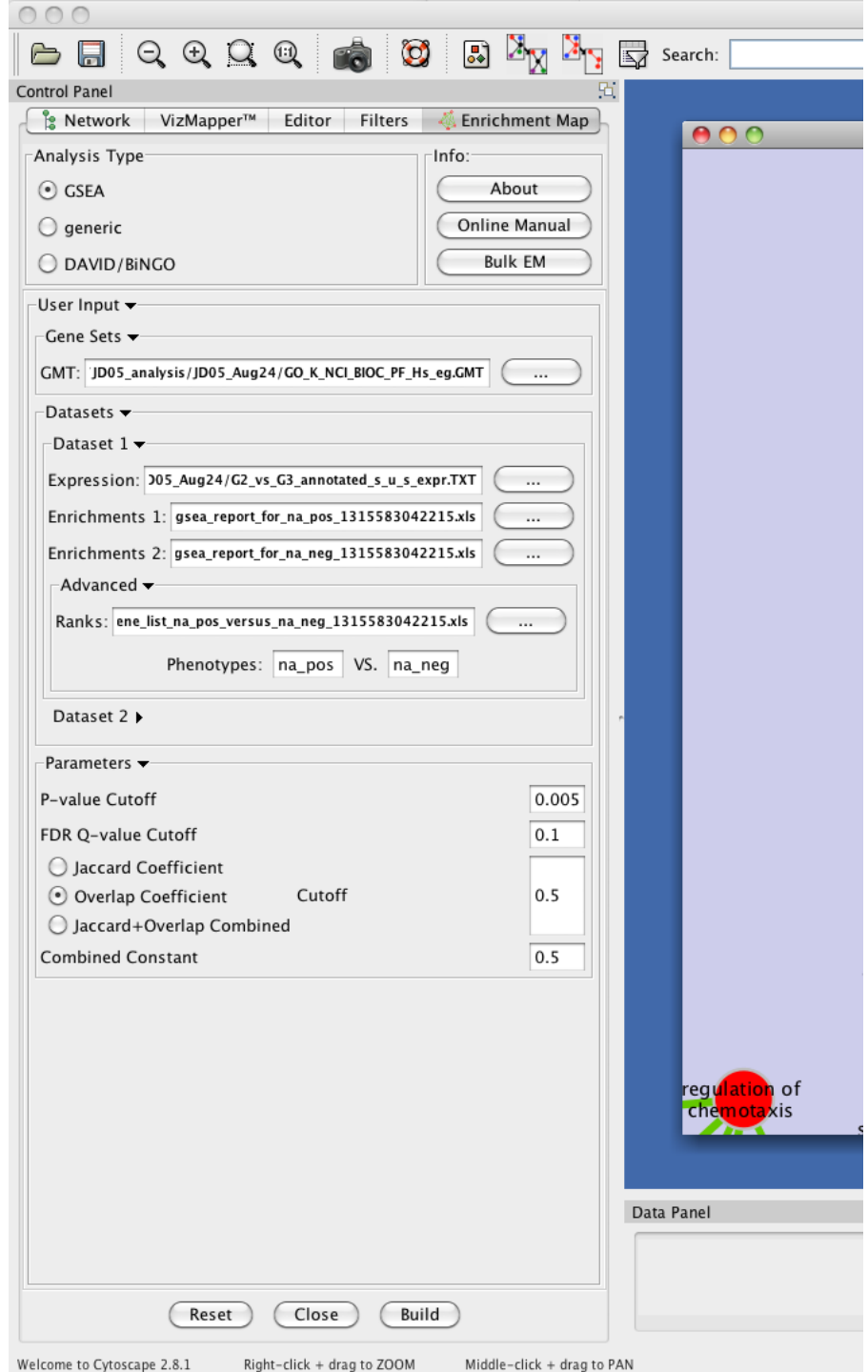


Cytoscape

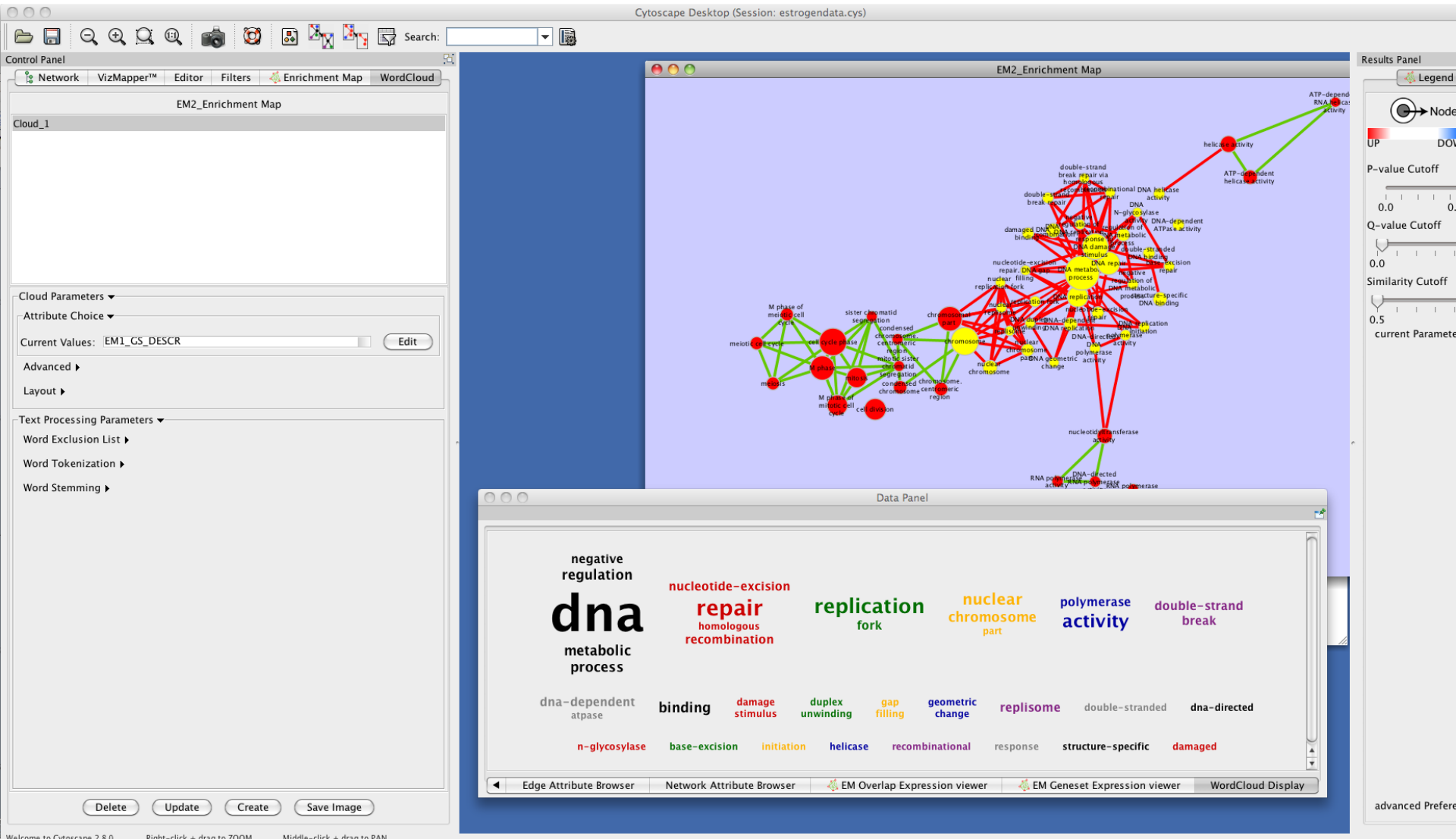
- > Plugins (menu bar)
- > Enrichment Map
- > Control Panel
- > Load Enrichment Results

To run enrichment map , you need:

- 1) GMT file .GMT : provided
- 2) Expression file .TXT
- 3) GSEA reports (use the .rpt file in the GSEA folder to upload the rank file and the enrichment files)



WordCloud: is a visual representation for text data



<http://baderlab.org/WordCloud>

How to convert between different gene identifiers

Synergizer: biological id mapping tool

http://llama.mshri.on.ca/synergizer/translate/ the synergizer

Ticket Query... DrProject UNIX/shell Bader Lab Index of /EM_Genesets pathway analysis TOOLS A practical ...tory Website databasr Mutome DB BioMed Cent...atabase 2.0 GeneMANIA Utility Plugins

Home | Help

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

Authority:

Species:

"FROM" namespace:

"TO" namespace:

File containing IDs to translate: no file selected

and/or

IDs to translate:

55861
841
9595
132
641975
647089

Output as spreadsheet: ☐

Results

Authority: ensembl
Source: Ensembl Genes 63 /
Ensembl Metazoa Genes 10 /
Ensembl Plants Genes 10
Updated: 2011-09-13

Shown in red below are IDs that do not belong to namespace "entrezgene" for species Homo sapiens (according to authority "ensembl").

One-column view

entrezgene	illumina_humanht_12
100008589	
442727	
7915	ILMN_2372398 ILMN_2372403
29094	ILMN_1673548
55486	ILMN_1731354 ILMN_2257665 ILMN_2341467
79590	
100130541	
55861	ILMN_1730612 ILMN_2338565
841	ILMN_1669565 ILMN_1673757 ILMN_1787749 ILMN_1809313 ILMN_2377733
9595	ILMN_1746864 ILMN_2092041
132	ILMN_1768062 ILMN_1801020

BIOMART

http://www.ensembl.org/biomart/martview/360394ae0e338e11ec715a24993fbaf8

http://www.ensembl.org/biomart/martview/360394ae0e338e11ec715a24993fbaf8

biomart martview

Ticket Query...- DrProject UNIX/shell Bader Lab Index of /EM_Genesets pathway analysis TOOLS A practical ...tory Website databasr Mutome DB BioMed Cent...atabase 2.0 GeneMANIA Utility Plugins

Ensembl

BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors

Login · Register

New Count Results

URL XML Perl Help

Dataset Caught Exception, Hit New

Homo sapiens genes (GRCh37.p5)

Filters

EntrezGene ID(s) [e.g. 100287163]: [ID-list specified]

Attributes

EntrezGene ID

HGNC symbol

Illumina Human HT 12 probe

Ensembl Gene ID

Dataset

[None Selected]

☐ Ensembl to LRG link transcript IDs

☐ Ensembl to LRG link translation IDs

☐ LRG to Ensembl link transcript

☒ EntrezGene ID

☐ VEGA transcript ID(s) (OTTT)

☐ VEGA gene ID(s) (OTTG)

☐ Ensembl transcript (where OTTT shares CDS with ENST)

☐ HAVANA transcript (where ENST shares CDS with OTTT)

☐ HAVANA transcript (where ENST identical to OTTT)

☐ HGNC ID(s)

☒ HGNC symbol

☐ HGNC transcript name

☐ IPI ID

☐ MEROPS ID

☐ MIM Morbid Accession

☐ MIM Morbid Description

☐ MIM Gene Accession

☐ MIM Gene Description

Microarray probes/probesets (max 2)

☐ Affy HC G110 probeset

☐ Affy HG FOCUS probeset

☐ Affy HG U133-PLUS-2 probeset

☐ Affy HG U133A_2 probeset

☐ Rfam transcript name

☐ Unigene ID

☐ UniProt/TrEMBL Accession

☐ UniProt/SwissProt ID

☐ UniProt/SwissProt Accession

☐ UniProt Gene Name

☐ WikiGene name

☐ WikiGene description

☐ Human Protein Atlas Antibody ID

☐ Database of Aberrant 3' Splice Sites (DBASS3) IDs

☐ DBASS3 Gene Name

☐ Database of Aberrant 5' Splice Sites (DBASS5) IDs

☐ DBASS5 Gene Name

☐ RefSeq mRNA

☐ RefSeq mRNA predicted

☐ RefSeq ncRNA

☐ RefSeq ncRNA predicted

☐ Affy HuGene FL probeset

☐ Affy HuEx 1_0 st v2 probeset

☐ Affy HuGene 1_0 st v1 probeset

☐ Affy U133 X3P probeset



Aknowledgments

**Gary Bader
Ruth Isserlin
Daniele Merico**

**OICR Cancer Stem Cell Program
John Dick
Jayne Danska
Melissa Cooper
Shaheena Bashir**