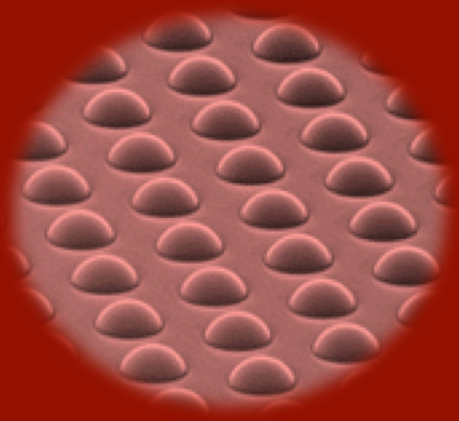
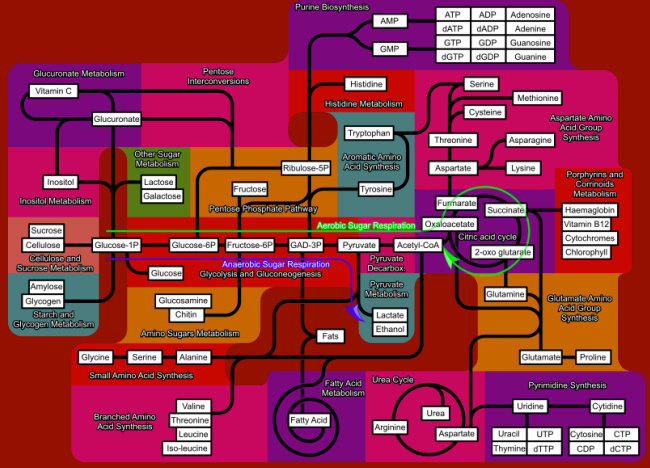
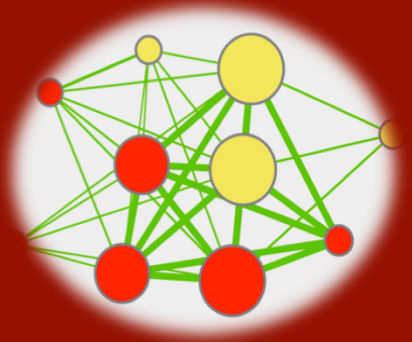




http://en.wikipedia.org/wiki/File:Metabolism_790px.png

# 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)**

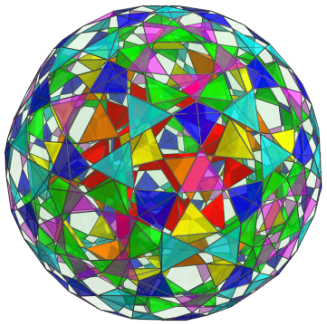
1.1 What is Pathway and Network analysis ?
1.2 Brief description of a few databases
1.3 How to manipulate gene identifiers
1.4 Presentation of two enrichment analysis techniques: GSEA and DAVID

# 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

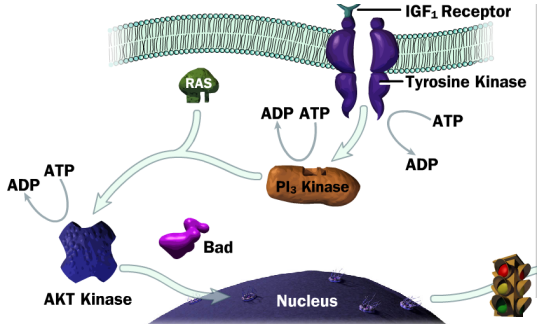


http://eusebeia.dyndns.org/4d/rect120cell

a cell

?

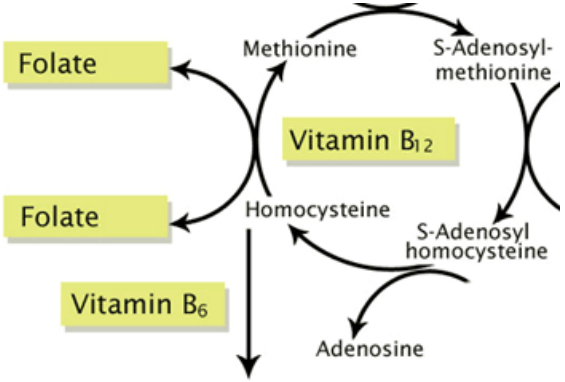
?




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

signaling pathway

active
or
inactive
?




http://proventigen.com/bvitamins

metabolic pathway

active
or
inactive
?

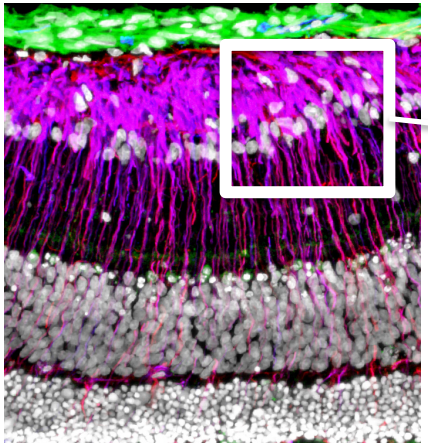
# Pathway analysis is an alternative to traditional analysis

- gene by gene basis
- requires literature searching

- time-consuming

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

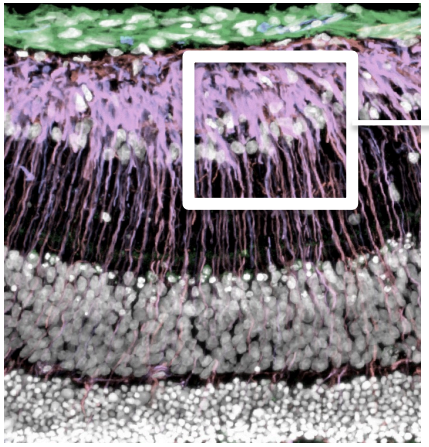
# Gene expression data (arrays, RNA seq)



RNA

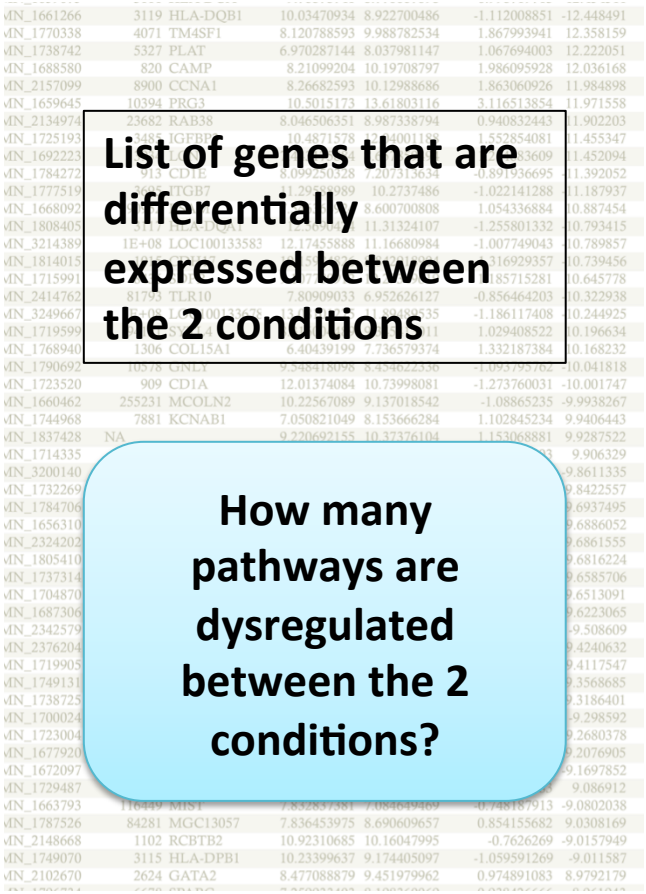

http://www.neuroscience.cam.ac.uk/directory/profile.php?thomasvjohnson

Wild-type



RNA


mutant

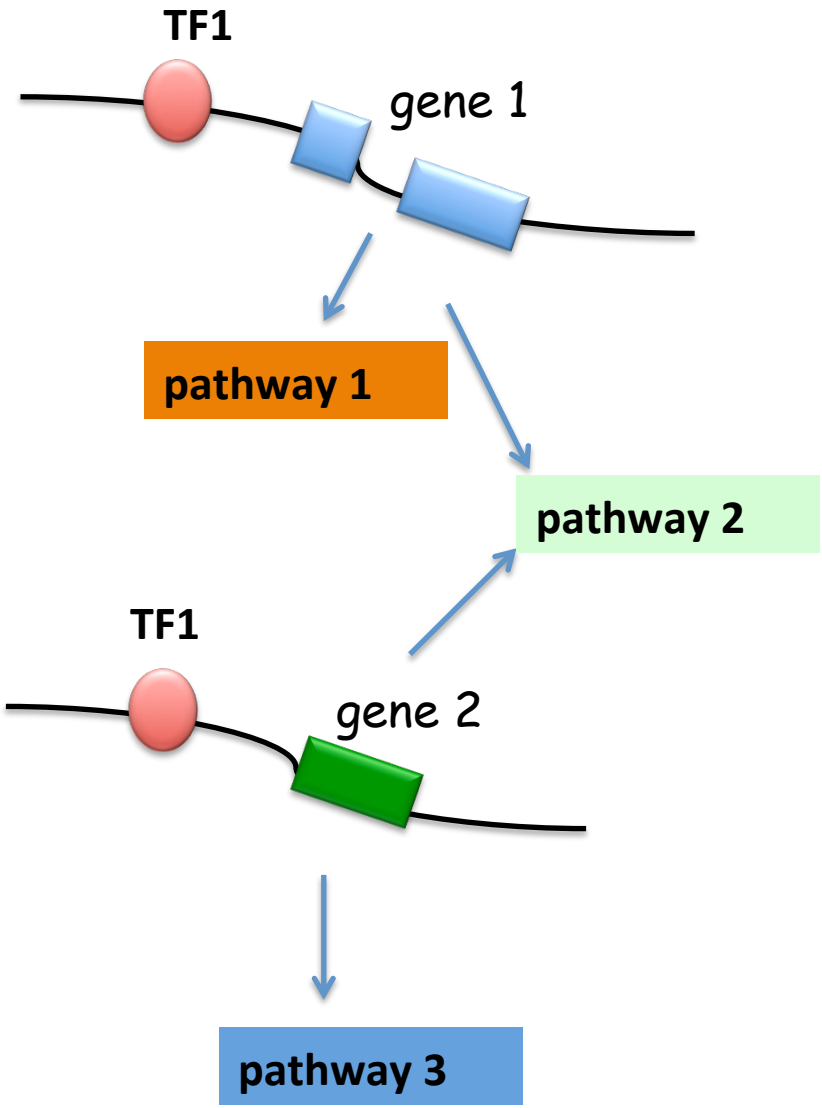


**List of genes that are differentially expressed between the 2 conditions**

**How many pathways are dysregulated between the 2 conditions?**

Enrichment analysis works with any gene lists.

# CHIP chip + gene expression




CHIP-chip

Cell population 1

Cell population 2

Gene expression

Cell population 1

Cell population 2

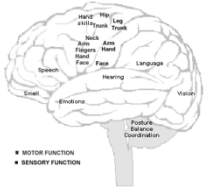
Genes closed to bound sites ↔ Genes differentially expressed

Genes closed to the bound sites and differentially expressed

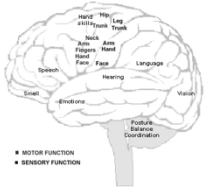
**In which pathways is the transcription factor involved (repression/ activation)**

# CNA CNV SNPs

DNA obtained from blood or buccal swabs




Patient (autism)




reference

SNPs arrays,
Deep sequencing,
arrayCGH,…

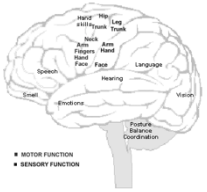
Regions of the genome are altered when comparing the patient with the reference

list of genes that are within the altered regions

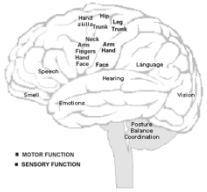
**Do these genes belong to same functional pathways?**

Pathway and network analysis

# CNA CNV SNPs



patient (brain cancer)



patient (normal tissue)

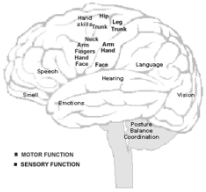
SNPs arrays,

List1: Non silent somatic mutations within genes

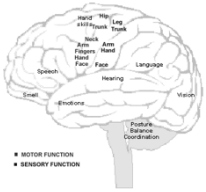
Pathway and network analysis

List 2: Genes differentially expressed in cancer versus normal

gene expression



patient (brain cancer)

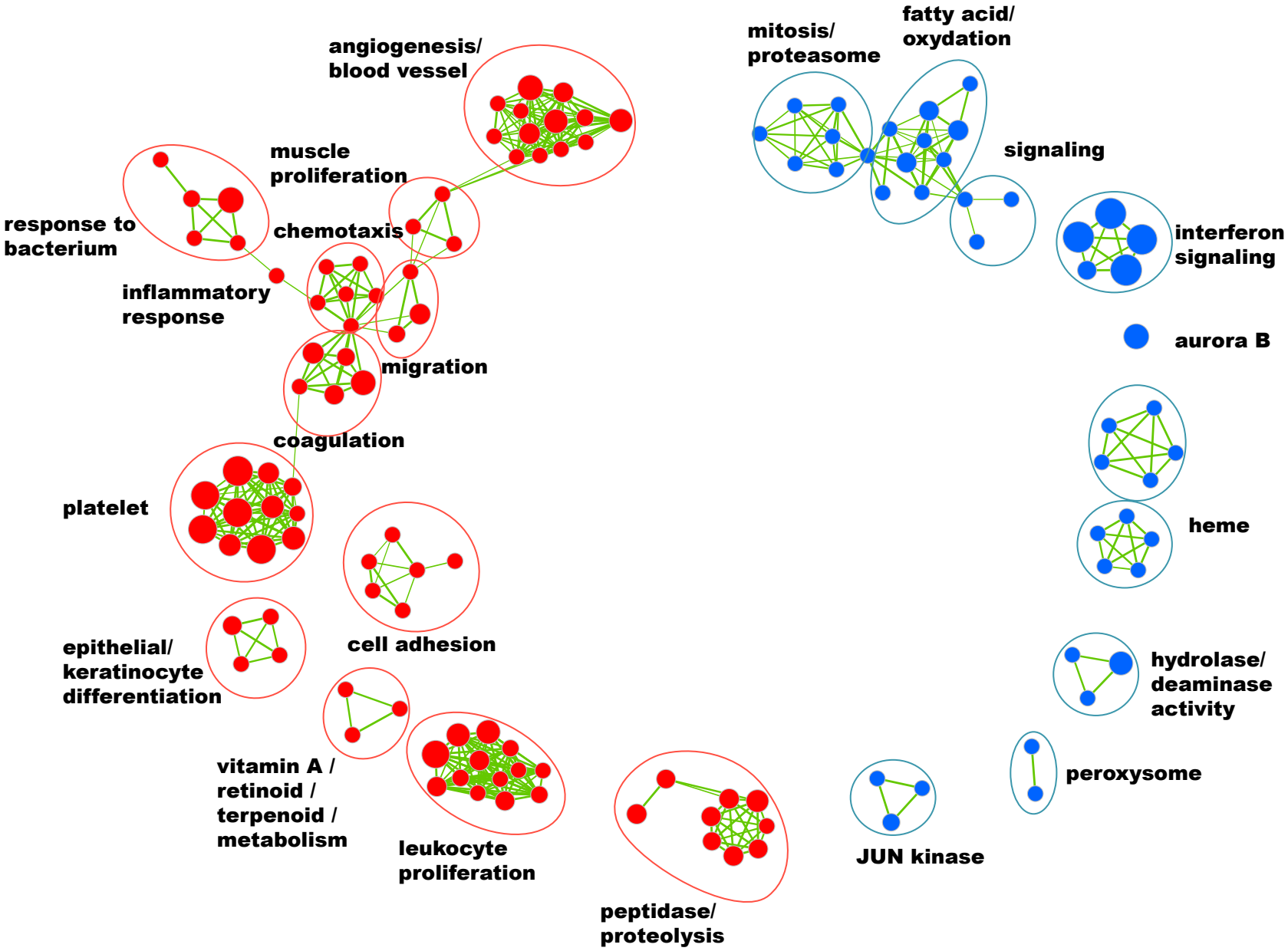


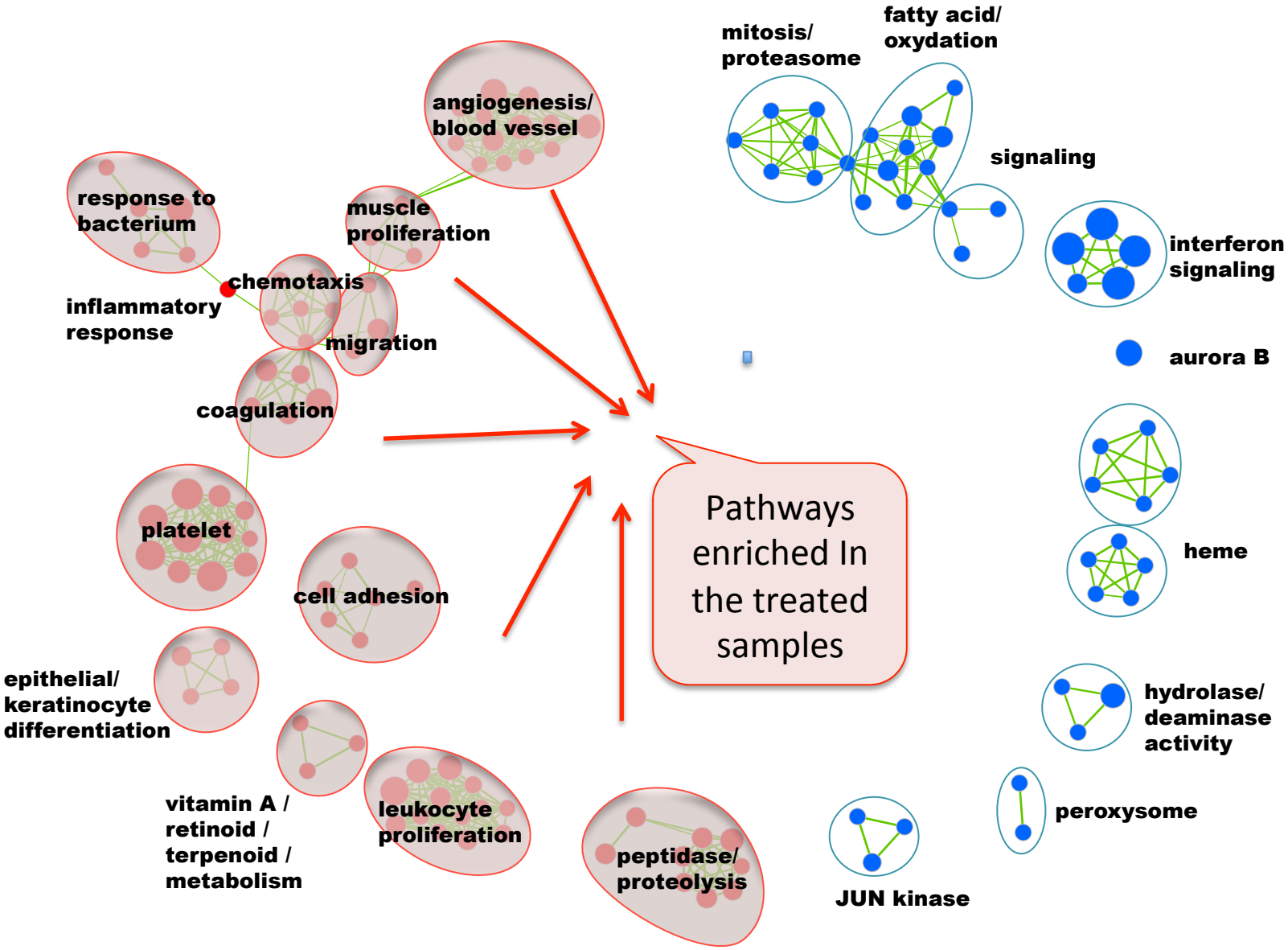
patient (normal tissue)

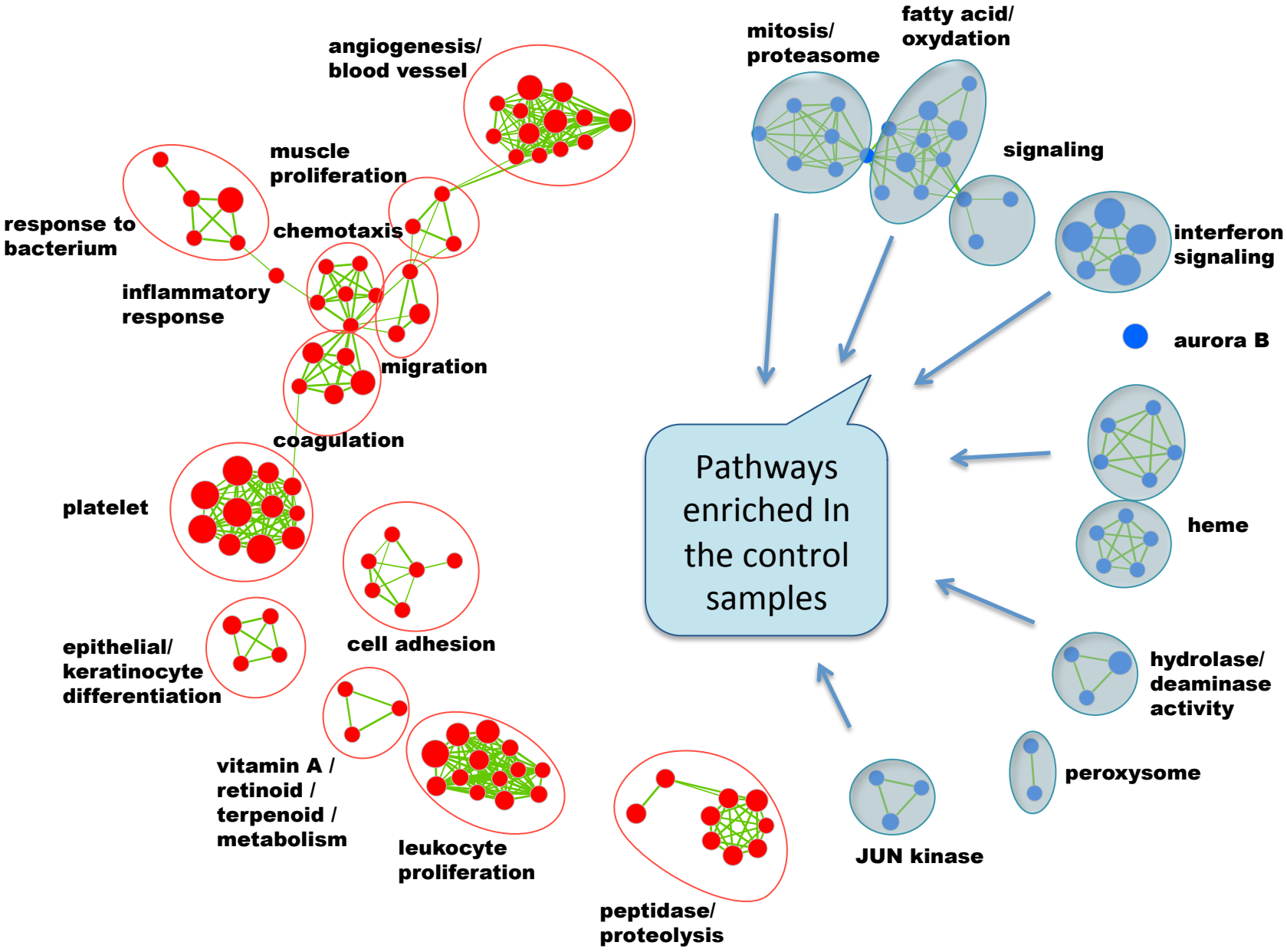
**Can we correlate at the pathway level the initiation events (somatic non silent mutations) and the resulting phenotype (gene expression)**

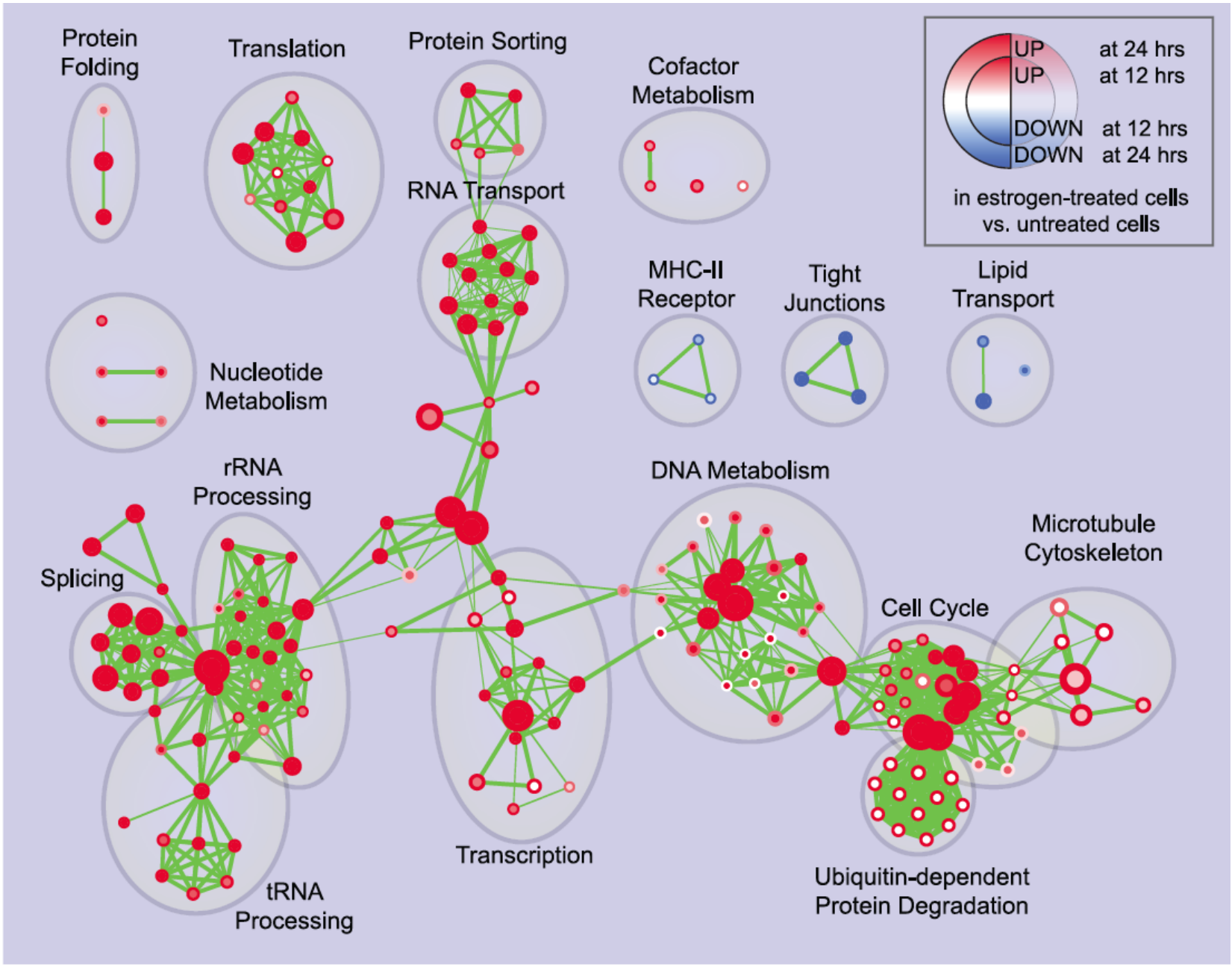
etc….

Some examples

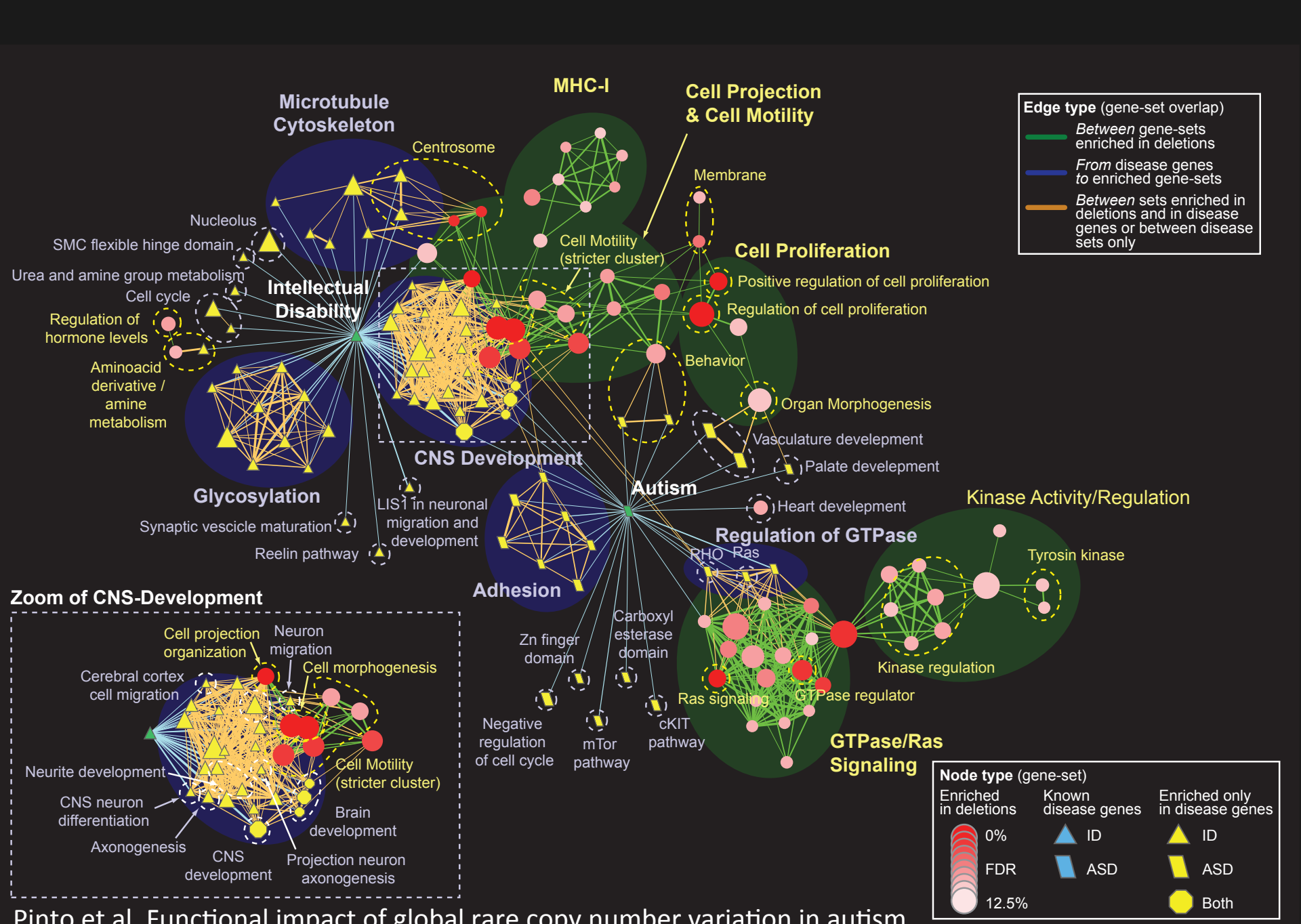

angiogenesis/
blood vessel
muscle
proliferation
response to
bacterium
chemotaxis
inflammatory
response
migration
coagulation
platelet
epithelial/
keratinocyte
differentiation
cell adhesion
vitamin A /
retinoid /
terpenoid /
metabolism
leukocyte
proliferation
peptidase/
proteolysis
JUN kinase
peroxysome
hydrolase/
deaminase
activity
heme
aurora B
interferon
signaling
signaling
mitosis/
proteasome
fatty acid/
oxydation


mitosis/
proteasome
fatty acid/
oxydation
signaling
interferon
signaling
aurora B
heme
hydrolase/
deaminase
activity
peroxysome
JUN kinase
peptidase/
proteolysis
leukocyte
proliferation
vitamin A /
retinoid /
terpenoid /
metabolism
epithelial/
keratinocyte
differentiation
cell adhesion
platelet
coagulation
migration
inflammatory
response
chemotaxis
response to
bacterium
muscle
proliferation
angiogenesis/
blood vessel
Pathways
enriched In
the treated
samples

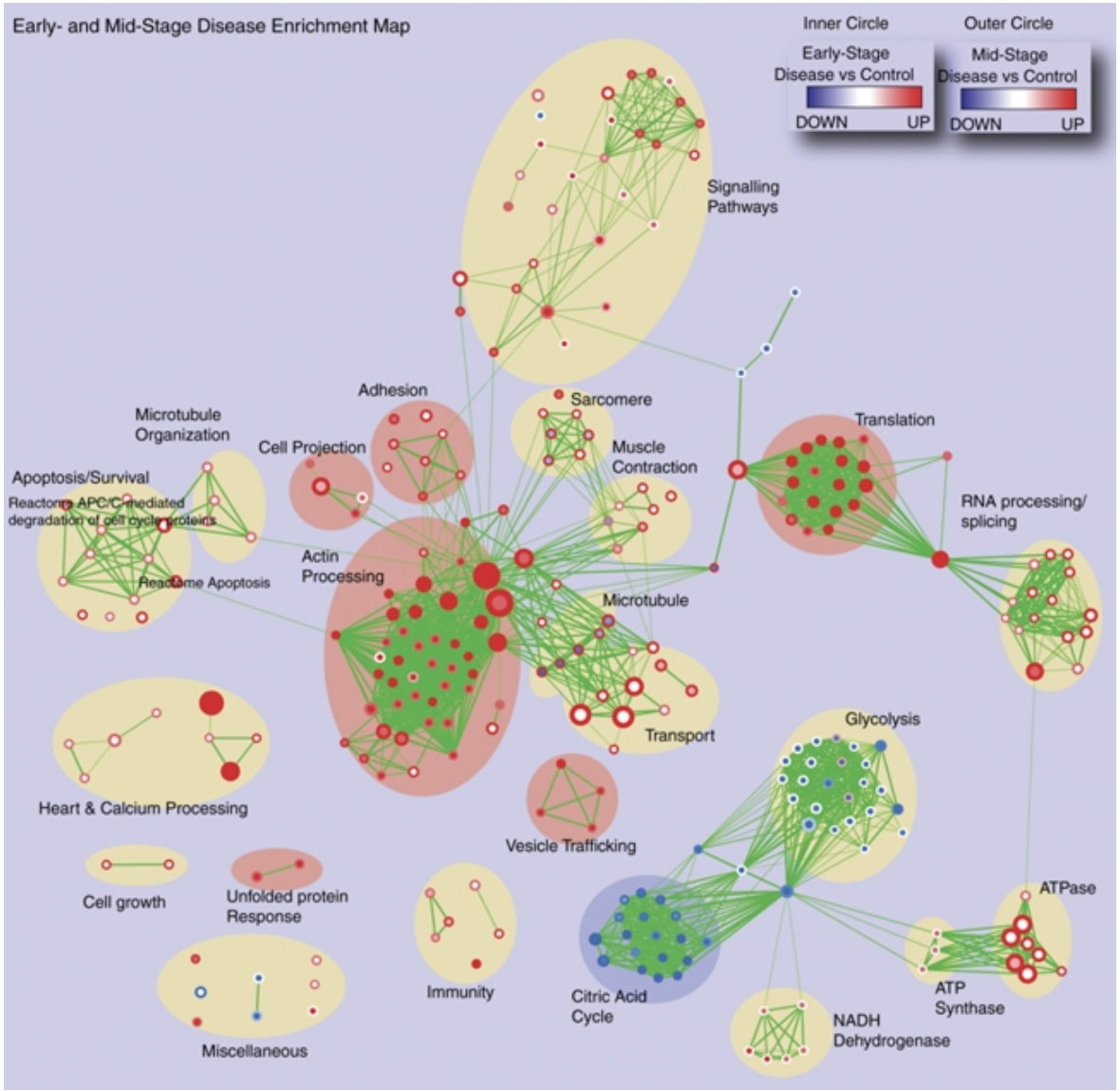

angiogenesis/
blood vessel
muscle
proliferation
response to
bacterium
chemotaxis
inflammatory
response
migration
coagulation
platelet
epithelial/
keratinocyte
differentiation
cell adhesion
vitamin A /
retinoid /
terpenoid /
metabolism
leukocyte
proliferation
peptidase/
proteolysis
mitosis/
proteasome
fatty acid/
oxydation
signaling
interferon
signaling
aurora B
heme
hydrolase/
deaminase
activity
peroxysome
JUN kinase
Pathways
enriched In
the control
samples




Merico et al. PLoS One, 2010



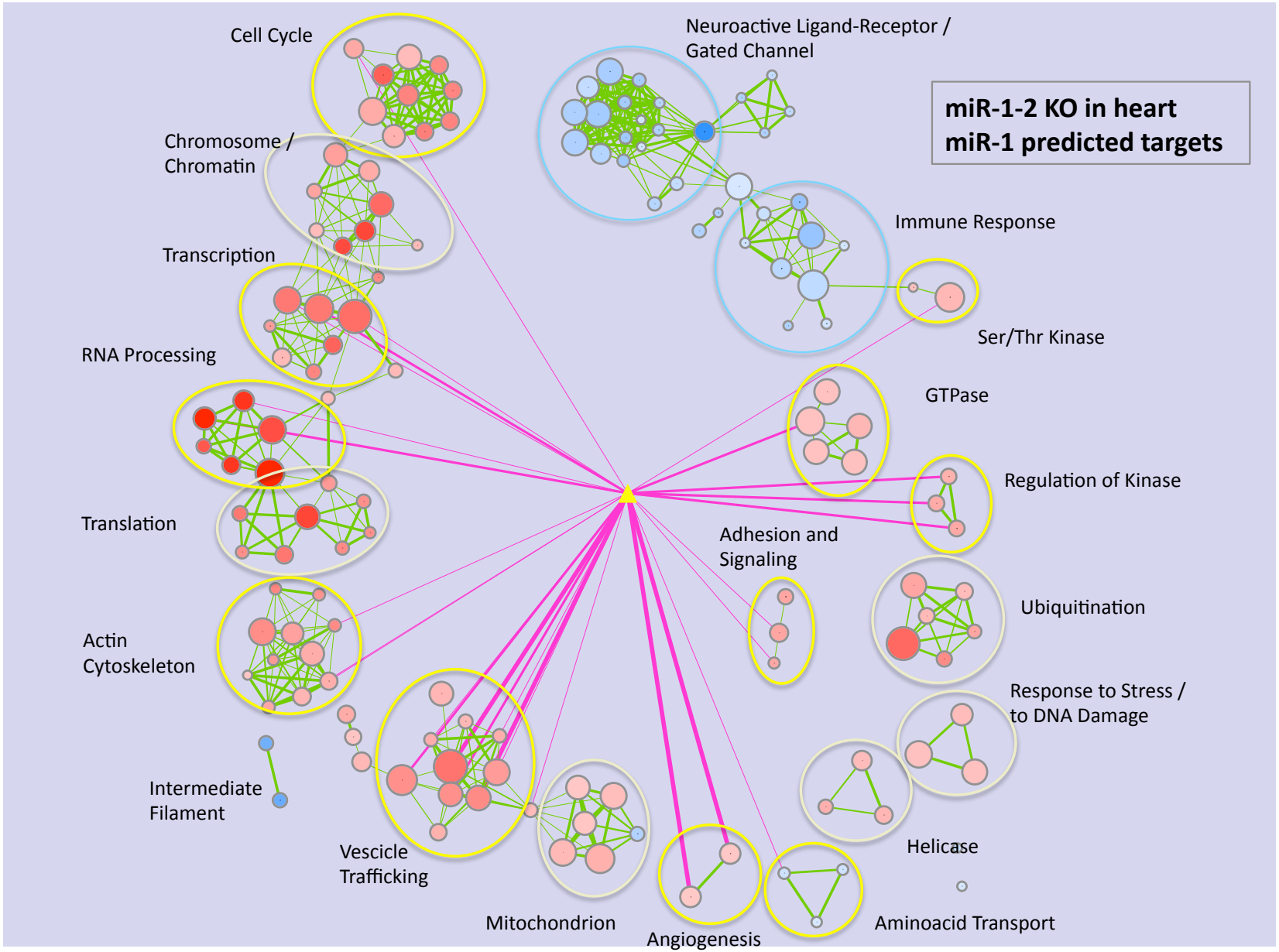

Pinto et al. Functional impact of global rare copy number variation in autism spectrum disorders. Nature. 2010 Jun 9.




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

Isserlin et al. , Proteomics 2010

# microRNA knock-out



Comparison of the gene list with databases containing annotated pathways

# DATABASES containing annotated pathways

Gene Ontology


MSigDB-c2


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

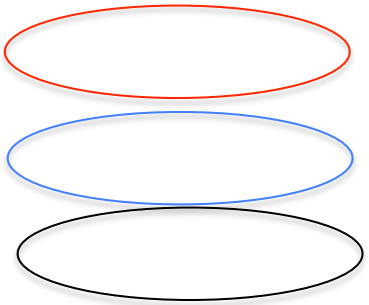
In this study, we report the isolation and molecular characterization of the B. napus **PERK1** cDNA, that is predicted to encode a novel receptor-like kinase. We have shown that like other plant RLKs, the kinase domain of PERK1 has serine/threonine kinase activity, In addition, the location of a PERK1-GTP 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

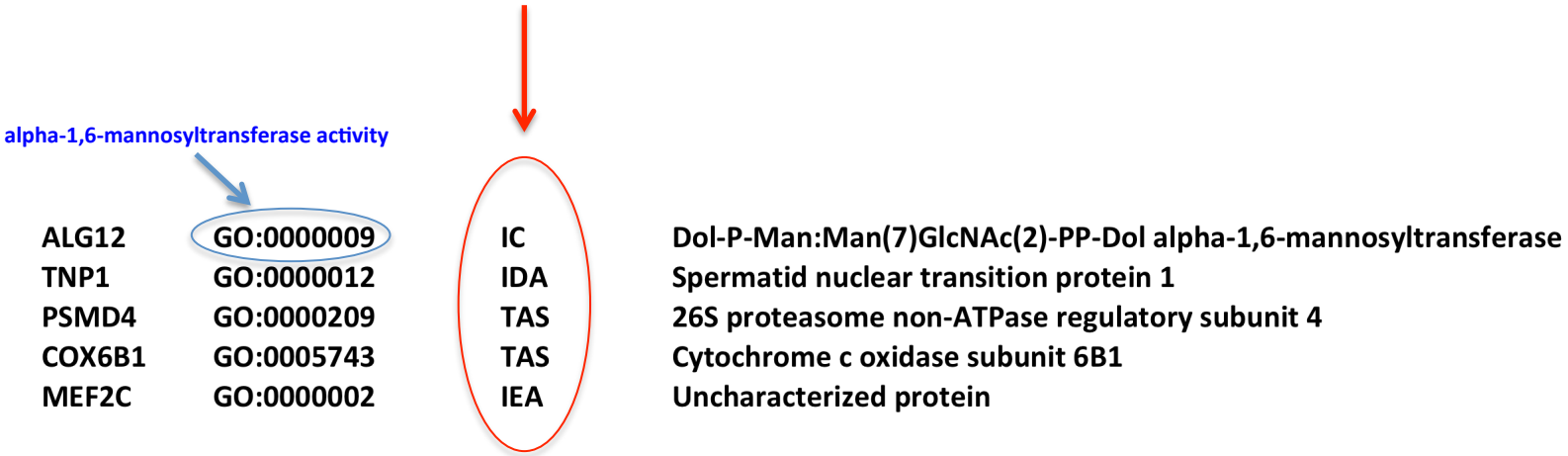


Molecular function

Molecular component

Biological process

GO Evidence Codes: information about how the annotation was created



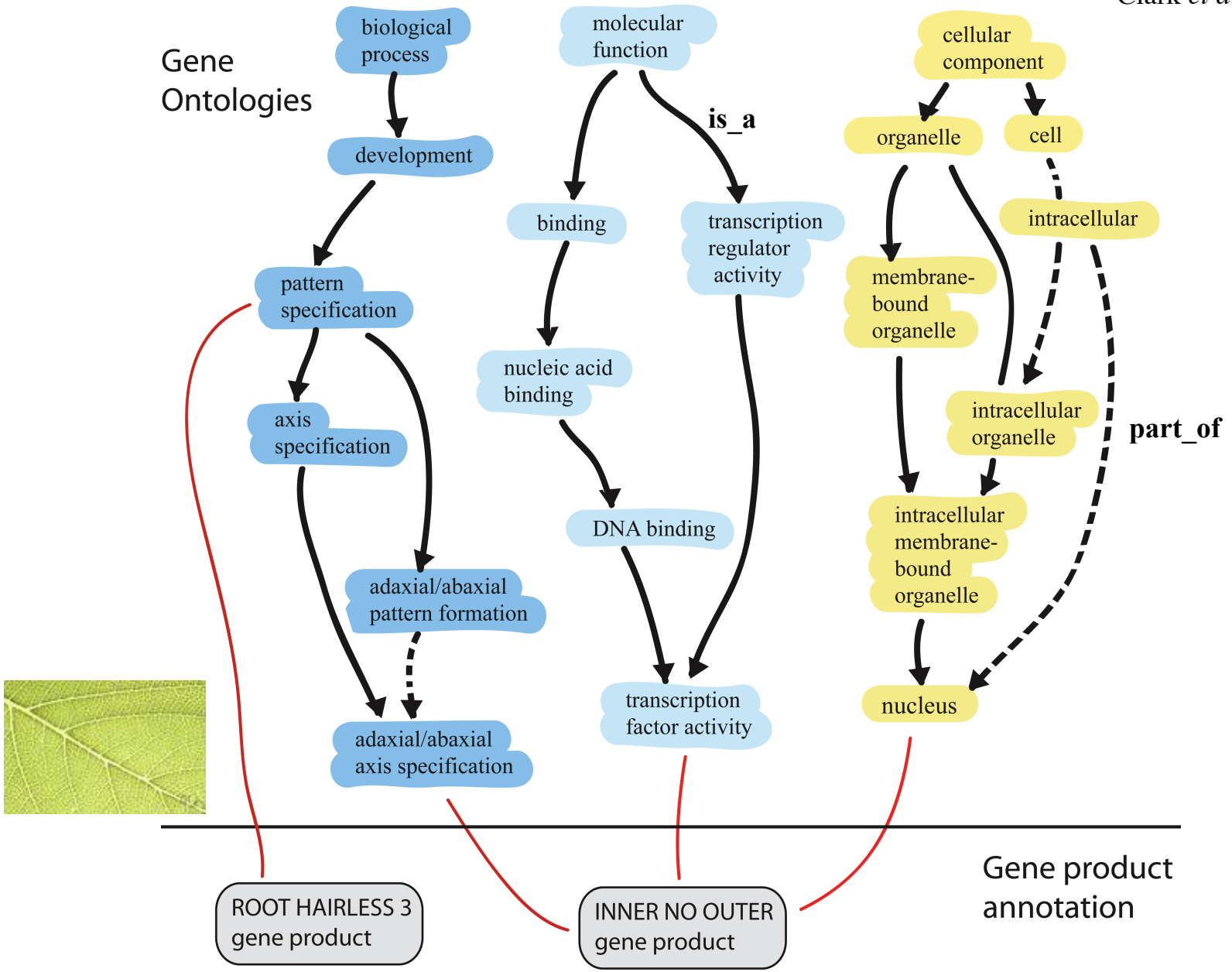

IC: inferred by curator
IDA: Inferred from direct assay
TAS: Traceable Author Statement
IEA: Inferred by electronic annotations

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

GO covers 3 domains:

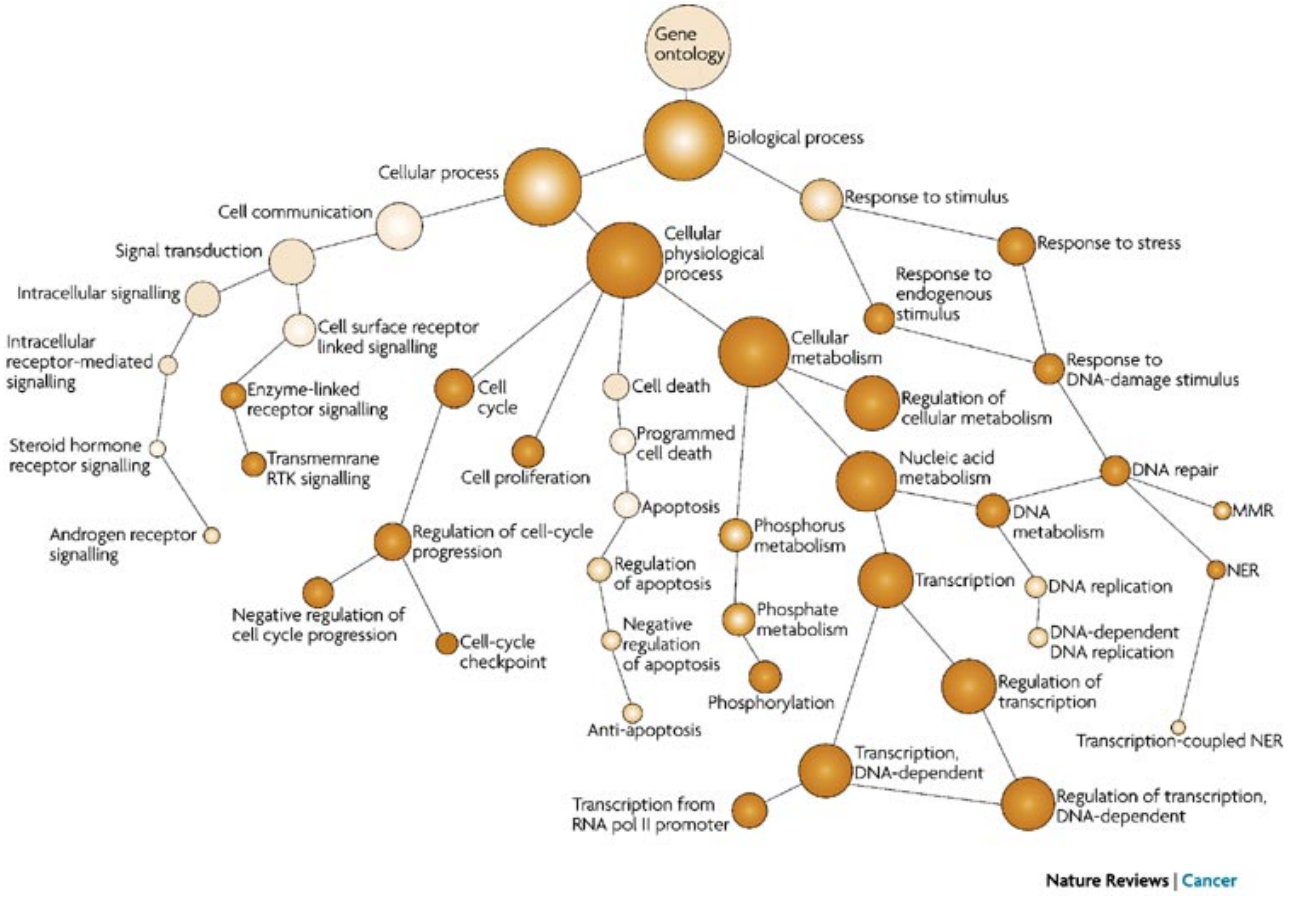
- **cellular component**: the parts of a cell or its extracellular environment;
- **molecular function**: the elemental activities of a gene product at the molecular level, such as binding or catalysis;
- **biological process**: operations or sets of molecular events with a defined beginning and end, pertinent to the functioning of integrated living units: cells, tissues, organs, and organisms.

Clark *et al.*, 2005




http://www.ncbi.nlm.nih.gov/pubmed/16010001

# BiNGO




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

nature REVIEWS CANCER

# KEGG: Kyoto Encyclopedia of Genes and Genomes

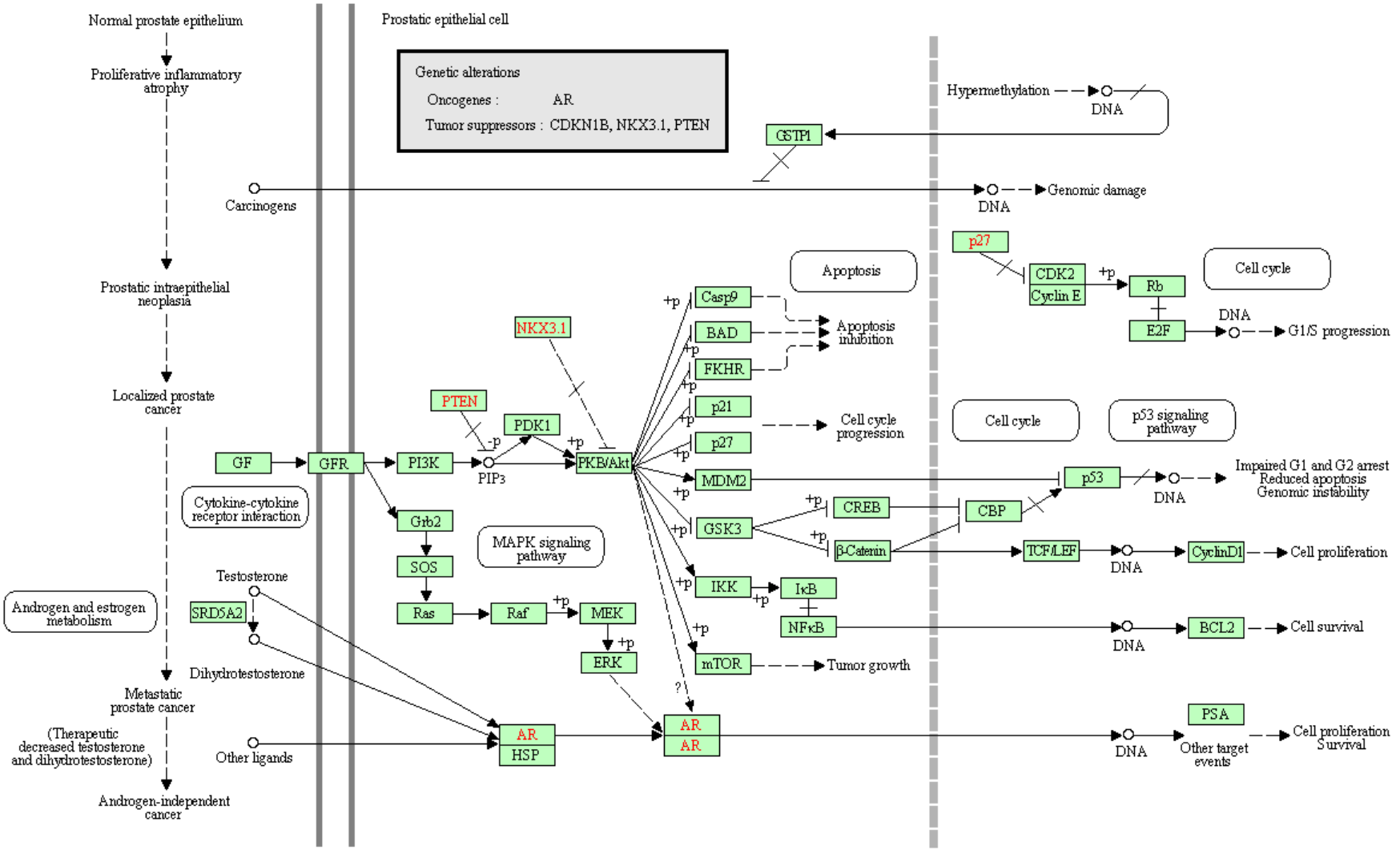
- Pathway maps for metabolism and other cellular processes, as well as human diseases; manually created from published materials
- 413 pathways
- Homo sapiens and Mus musculus model organism (Mus musculus inferred from Homo sapiens).

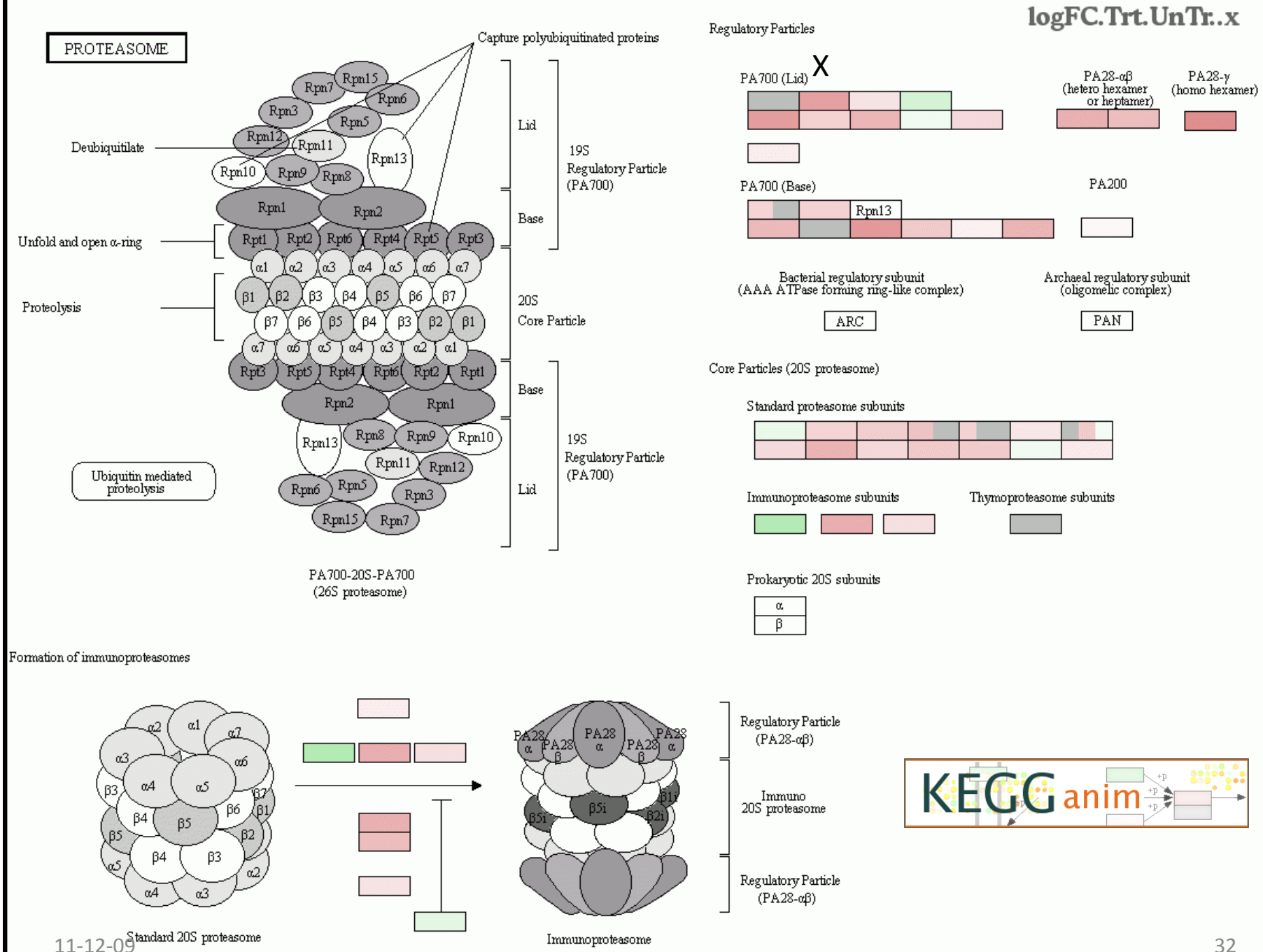
**Example of 2 categories of pathway**:

- Cellular Process
  - Cell Growth and Death
    - P53 signaling pathway

- Metabolism
  - Metabolism of cofactors and vitamins
    - Retinol metabolism

# KEGG prostate cancer pathway (42 genes)



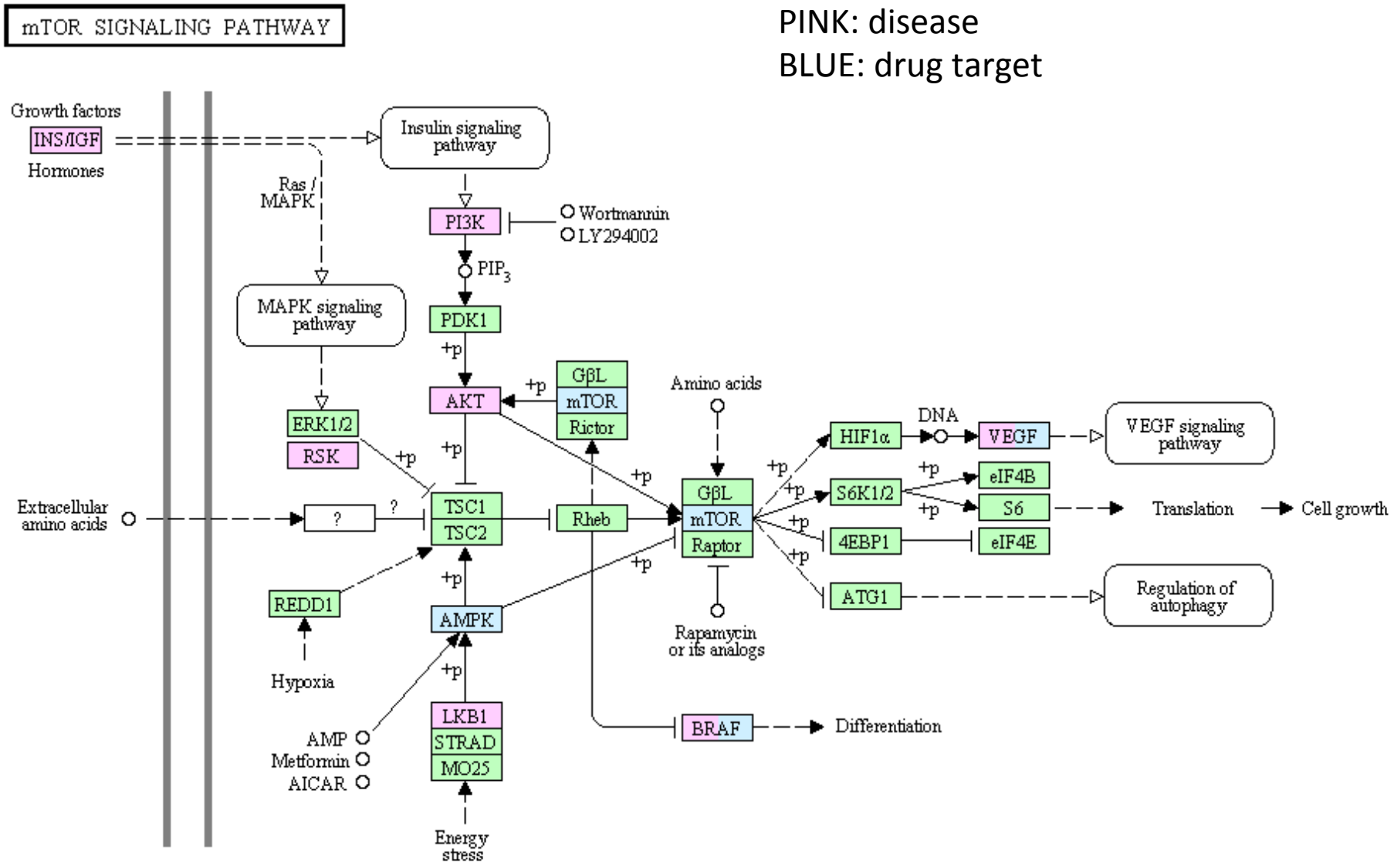





03050 3/31/10
(c) Kanehisa Laboratories

X 66997 Psmd12 ::proteasome (prosome, macropain) 26S subunit, non-ATPase, 12 0.37 -0.03 -0.01 0.879245654 0.96317355 0.99511344

PINK: disease
BLUE: drug target



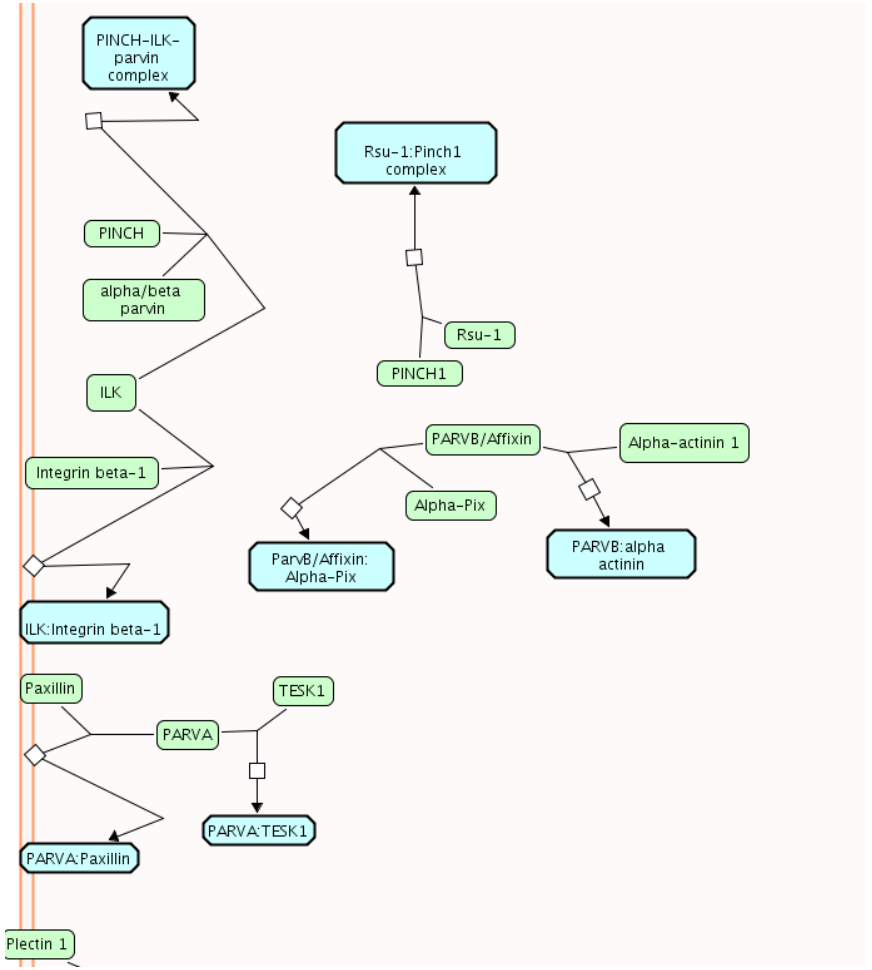

04150 3/31/09
(c) Kanehisa Laboratories

# 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



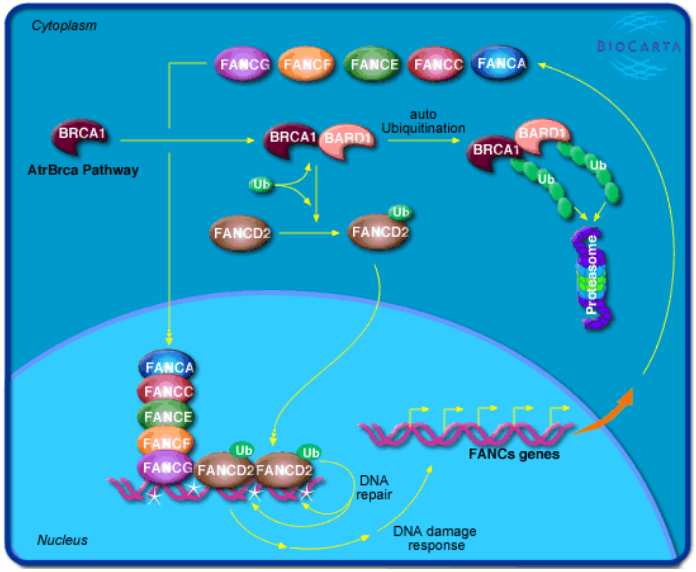
# Netpath

- 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 |

# Biocarta pathways

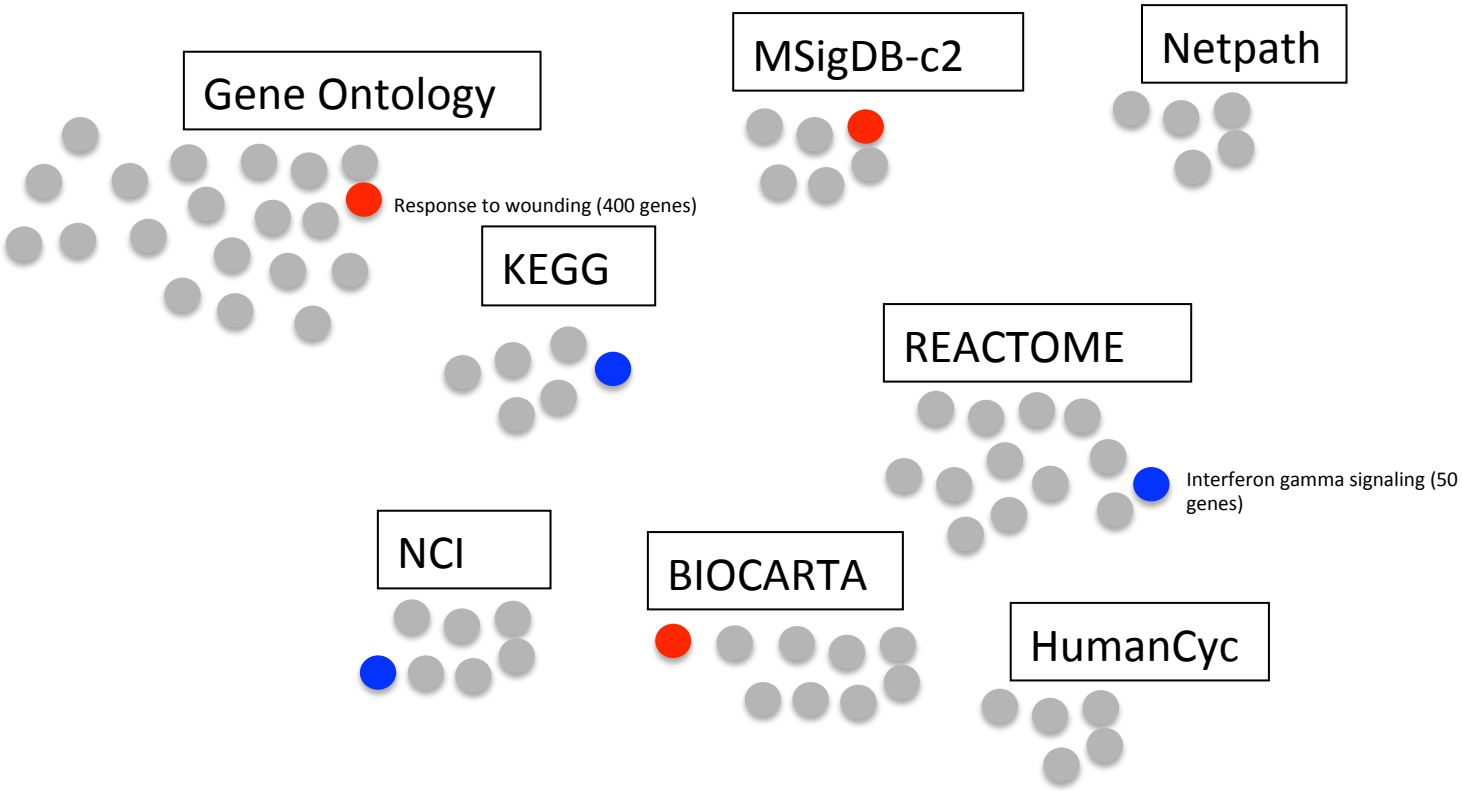
- company which develops reagents and assays for biopharmaceutical research
- expert-curated interactive graphic models of many pathways from diverse fields like apoptosis, cell cycle, cell signalling, development, immunology, neuroscience, adhesion, and metabolism.
- organisms: Homo sapiens and Mus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

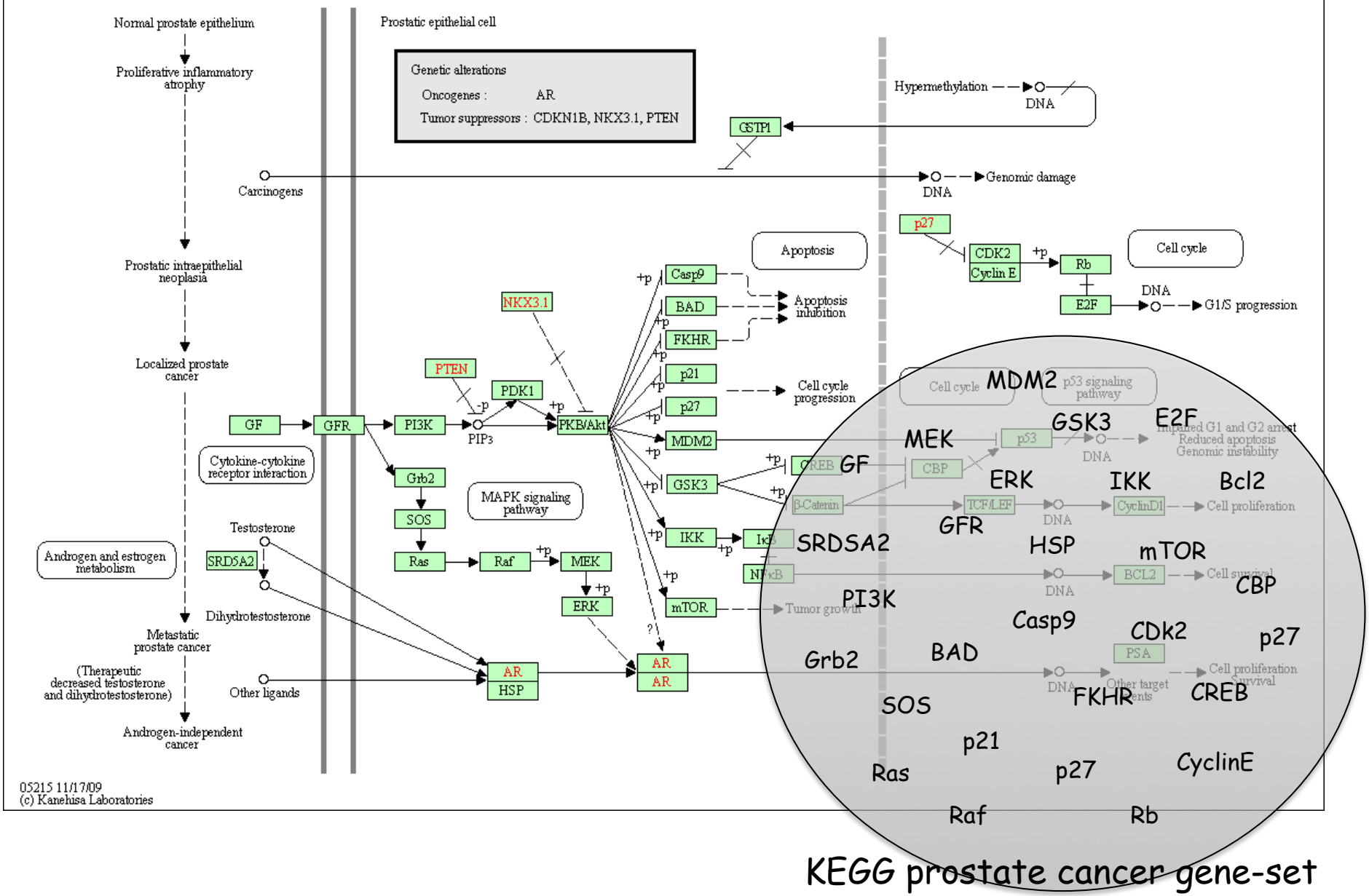
# Gene-sets

# DATABASES containing annotated pathways

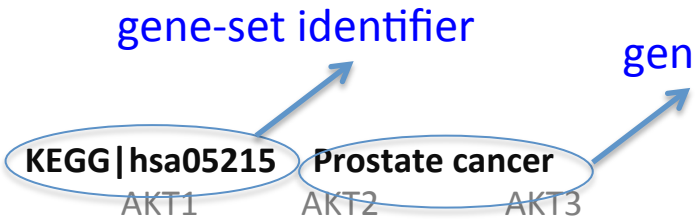



- a gene-set : group of genes with “similar function or annotation”

# KEGG prostate cancer pathway (42 genes)




KEGG prostate cancer gene-set (42 genes)




| 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 | EP300ERBB2 | | FGFR1 | FGFR2 | FOXO1 |
| GRB2 | GSK3B | GSTP1 | HRAS | HSP90AA1 | | HSP90AB1 | HSP90B1 | IGF1 |
| IGF1R | IKBKB | IKBKG INS | 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 | INS | 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** |

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.

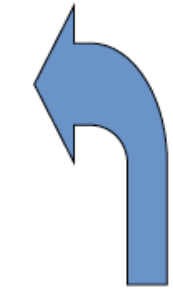
# Fisher's exact test

a.k.a., the hypergeometric test

(or modified Fisher's exact test)

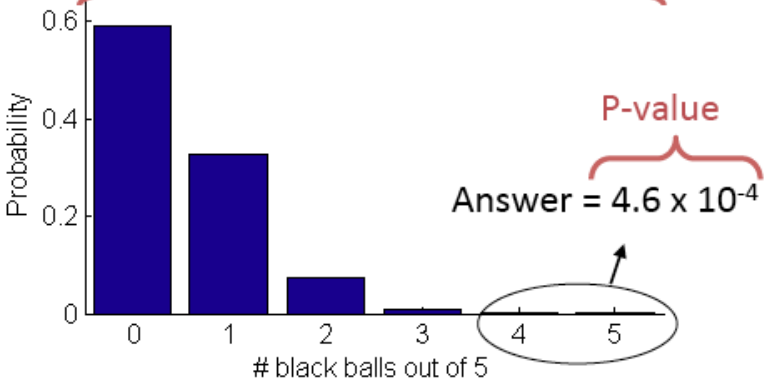
Gene list

- RRP6
- MRD1
- RRP7
- RRP43
- RRP42



Genes from all gene-sets

Null distribution




P-value

Answer = $4.6 \times 10^{-4}$

Background population:
500 black genes,
4500 red genes

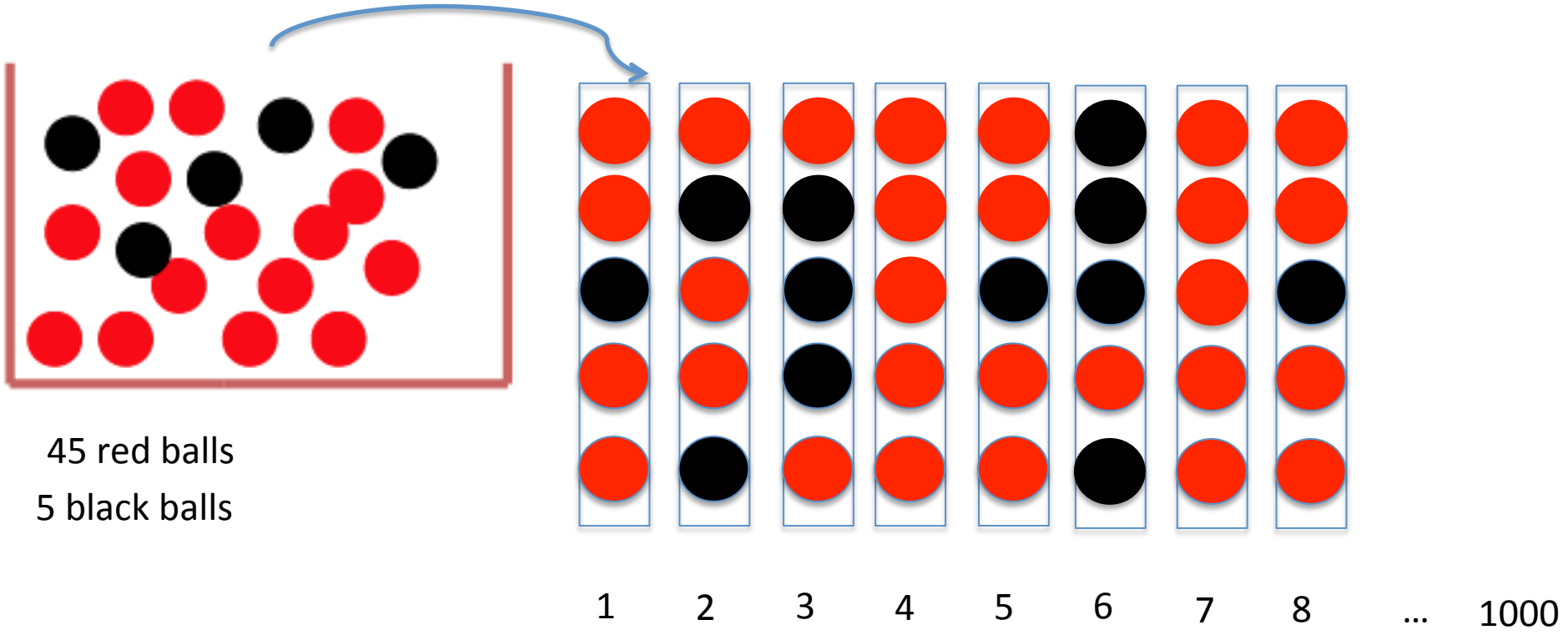


Slide: modified from Quaid Morris, CBW bioinformatics workshop, Pathway and Network analysis, Toronto, July 2011 , http://bioinformatics.ca/workshops/2011/course-content

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)

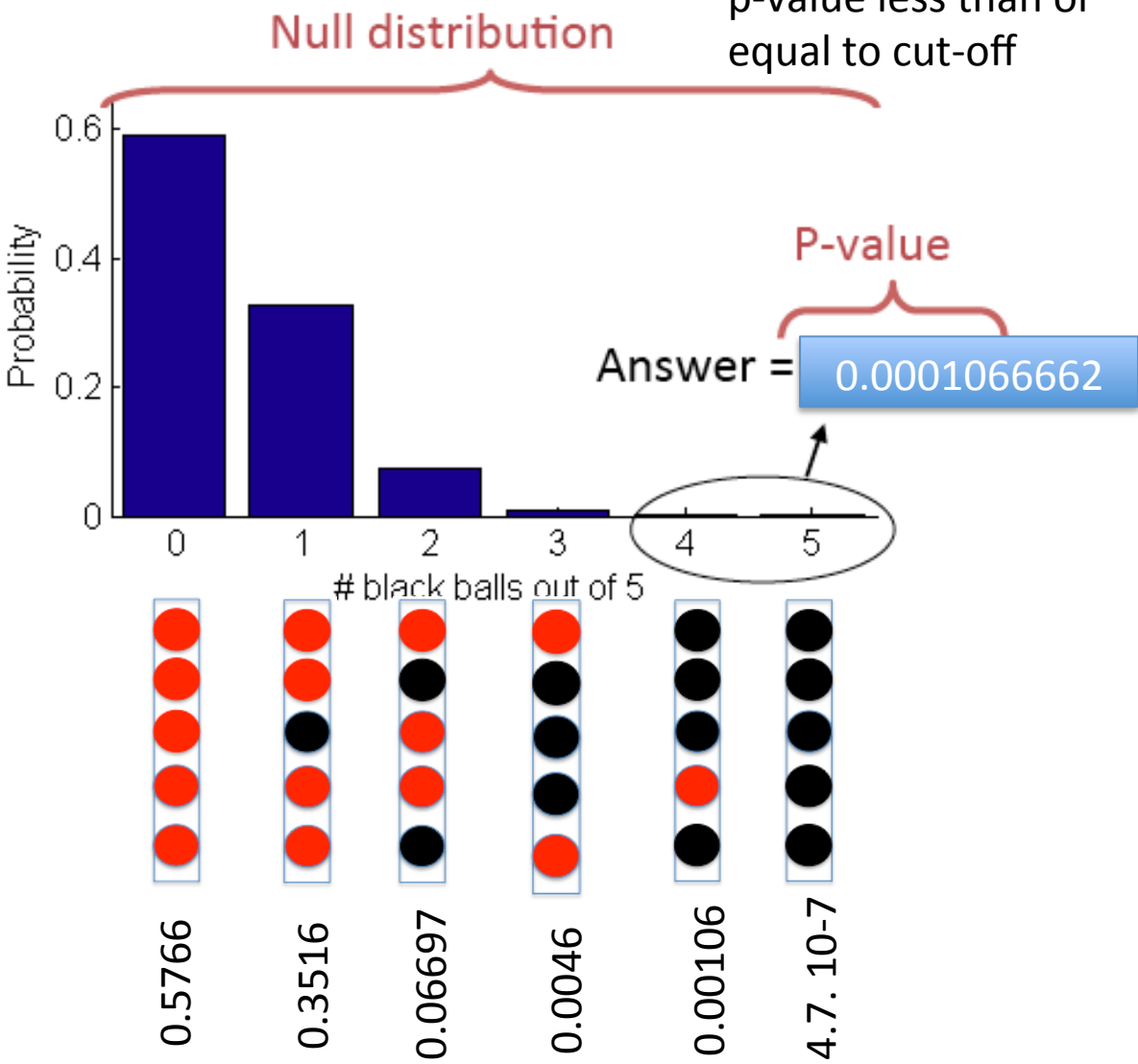
50 balls

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

1 gene-set (prostate cancer)

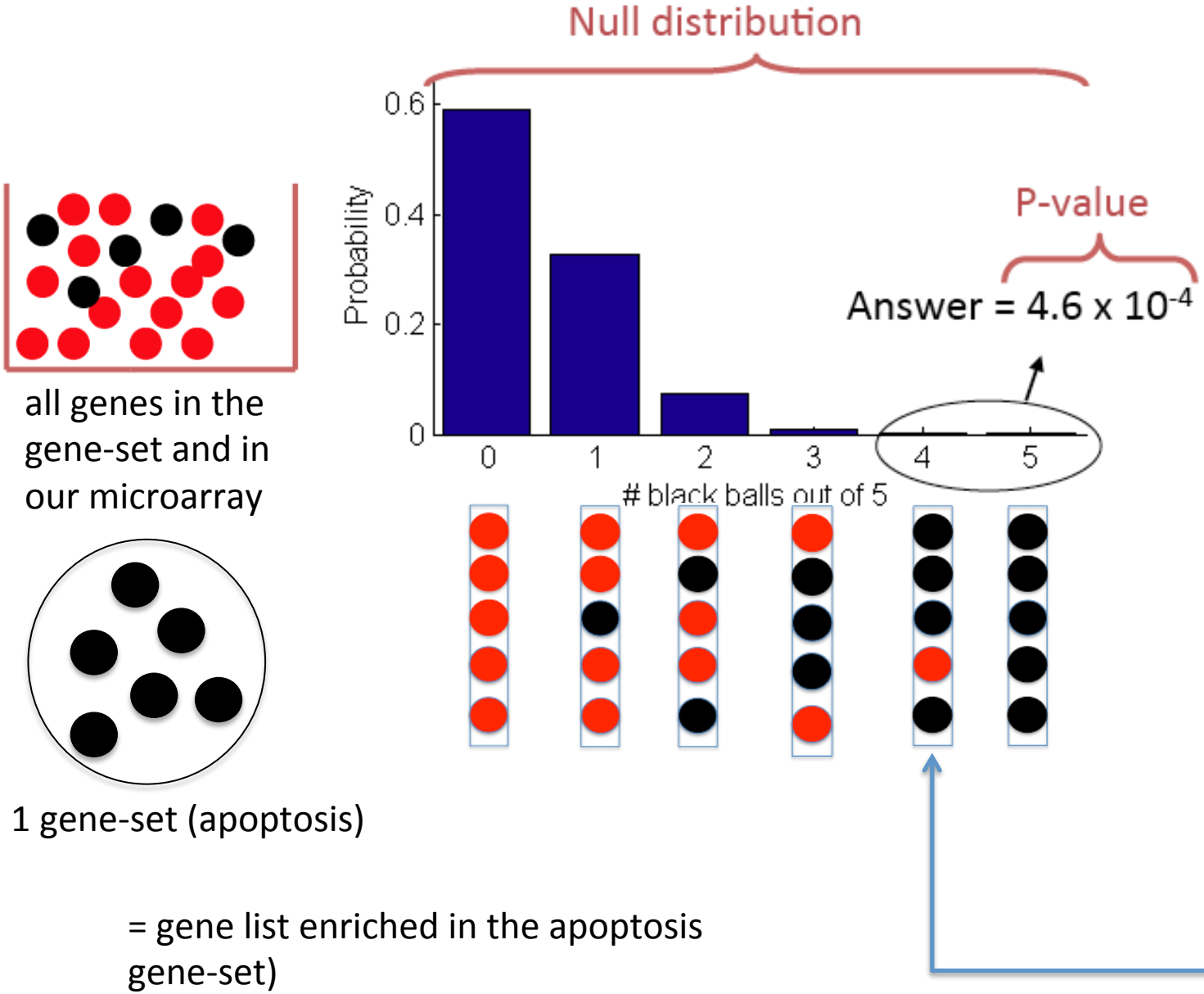
Cut-off: my case = 0.000106

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



Fisher's exact test

**The null distribution**




= gene list enriched in the apoptosis gene-set)

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 sensitity 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 contigency table is built on above numbers:

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

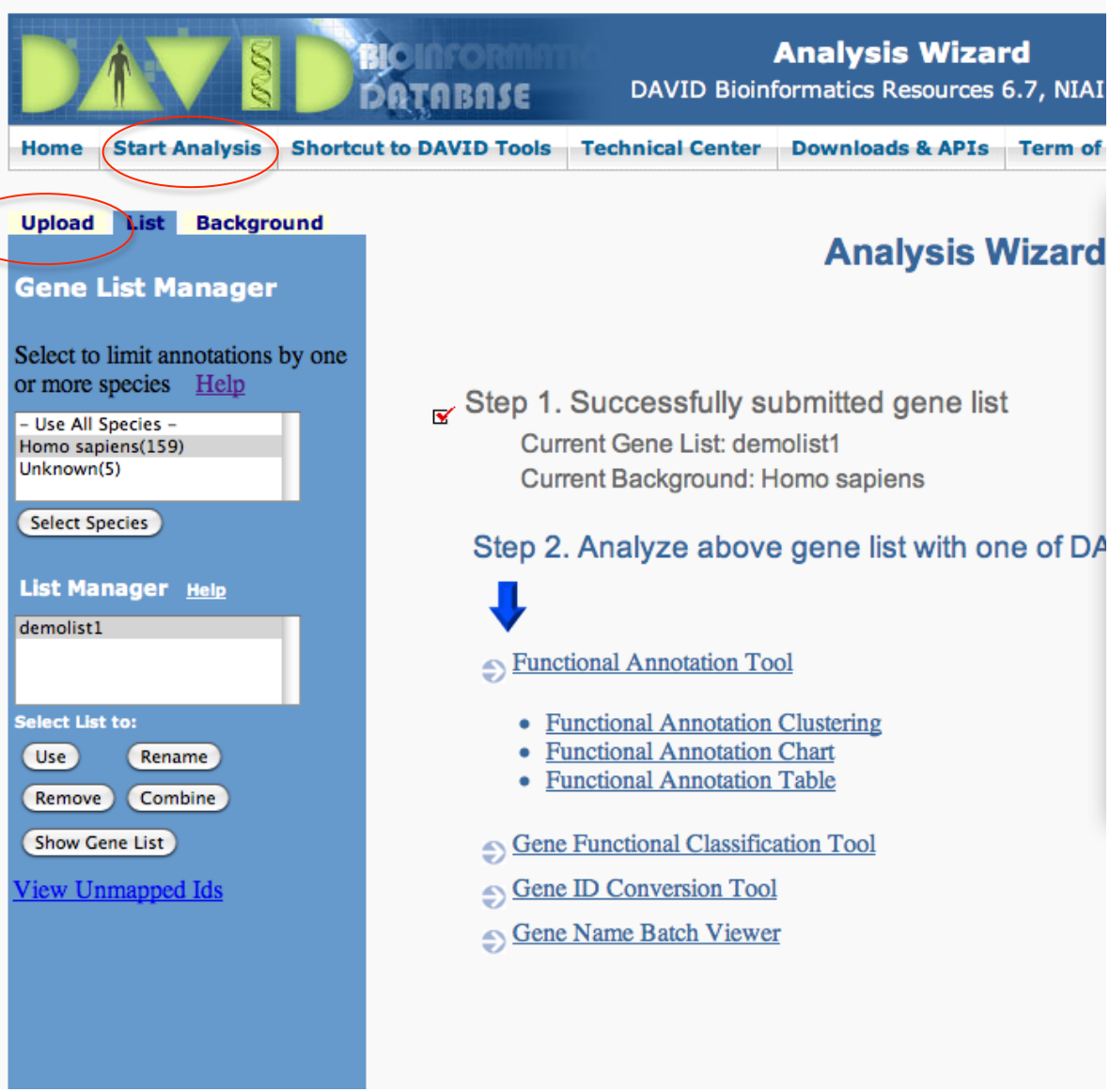
k-1: modified test ( to be more conservative)

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

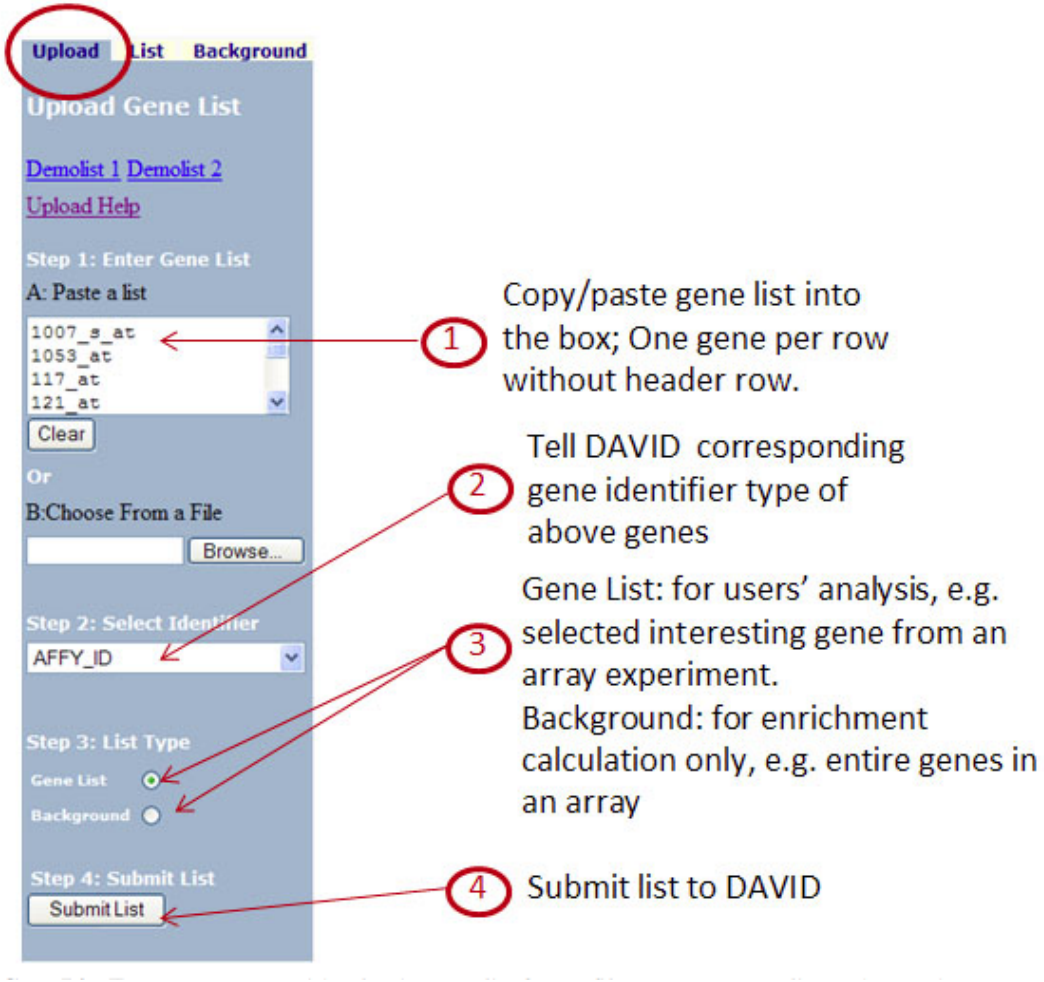
However, EASE Score is more conservative to exame the situation. EASE Score = 0.06 (using 3-1 instead of 3). Since P-Value > 0.01, this user gene list is specifically associated (enriched) in p53 signalling pathway no mmore than random chance

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

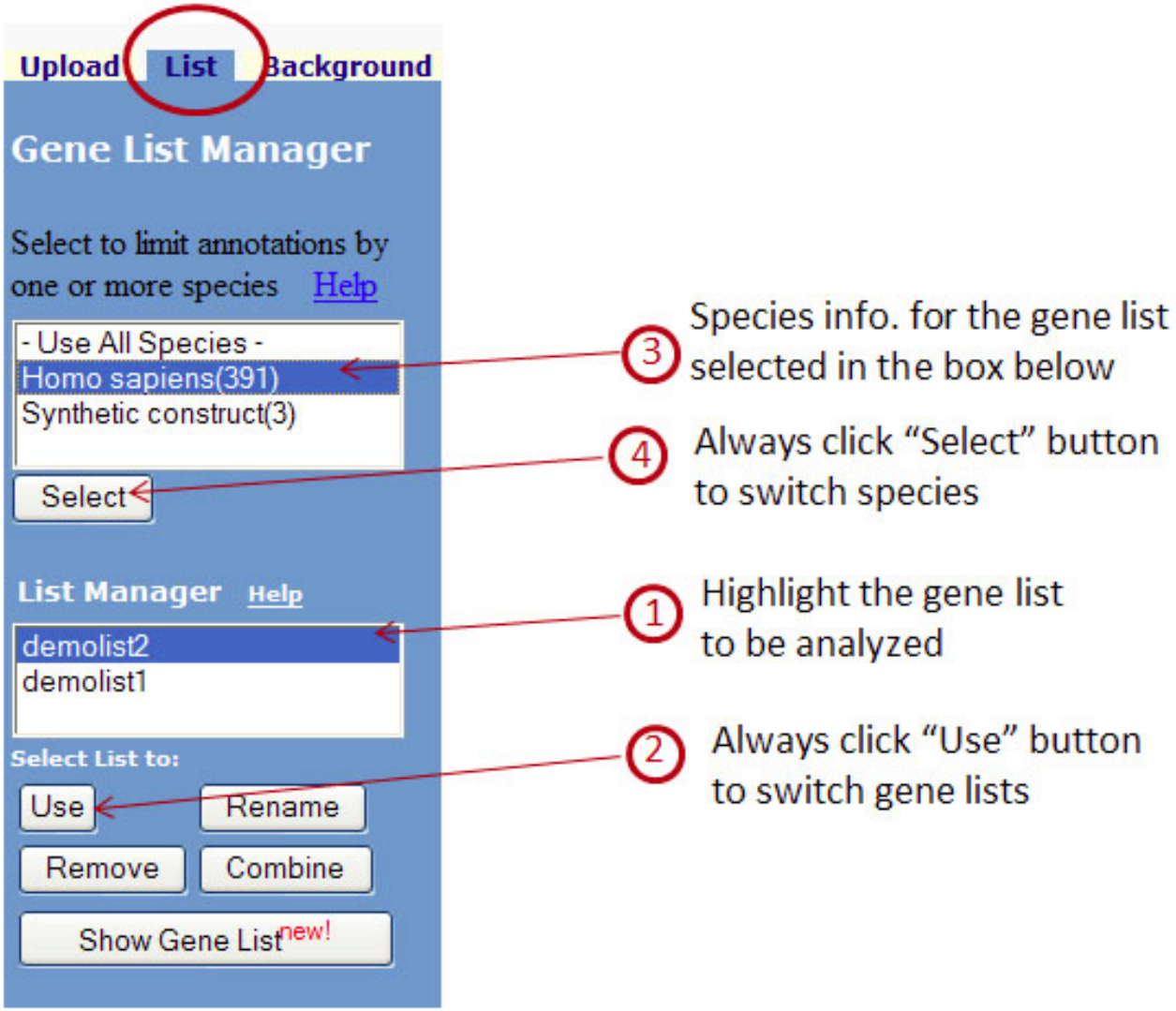
# 1. Start the analysis



## 2. Upload Tab


Upload
List
Background
Upload Gene List
Demolist 1 Demolist 2
Upload Help
Step 1: Enter Gene List
A: Paste a list
1007_s_at
1053_at
117_at
121_at
Clear
Or
B:Choose From a File
Browse...
Step 2: Select Identifier
AFFY_ID
Step 3: List Type
Gene List
Background
Step 4: Submit List
Submit List
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

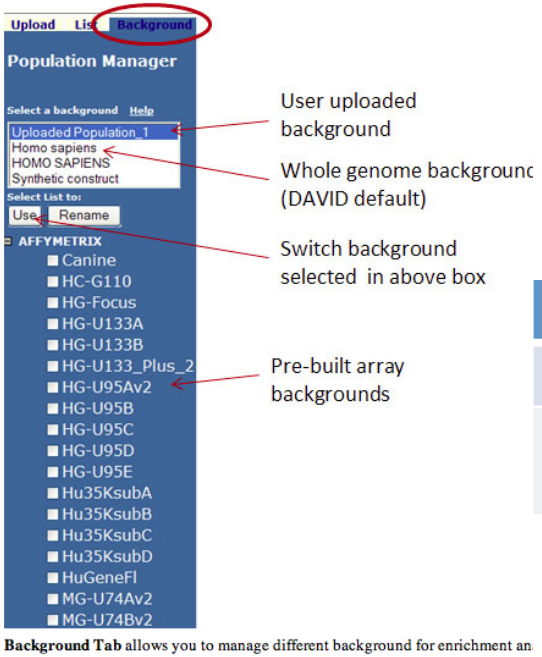
### 3. List Tab


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**



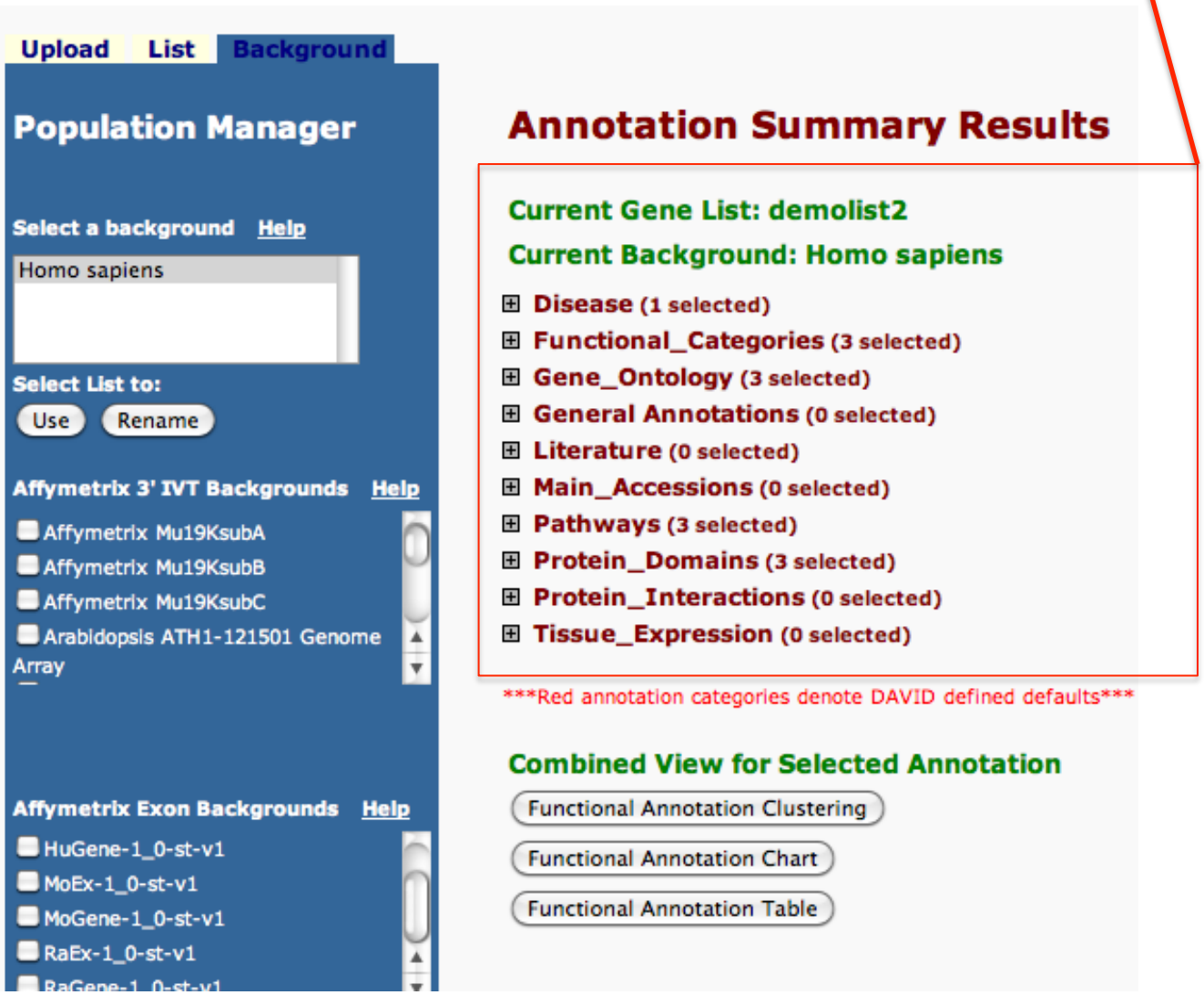

**Background Tab** allows you to manage different background for enrichment an

= 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** |

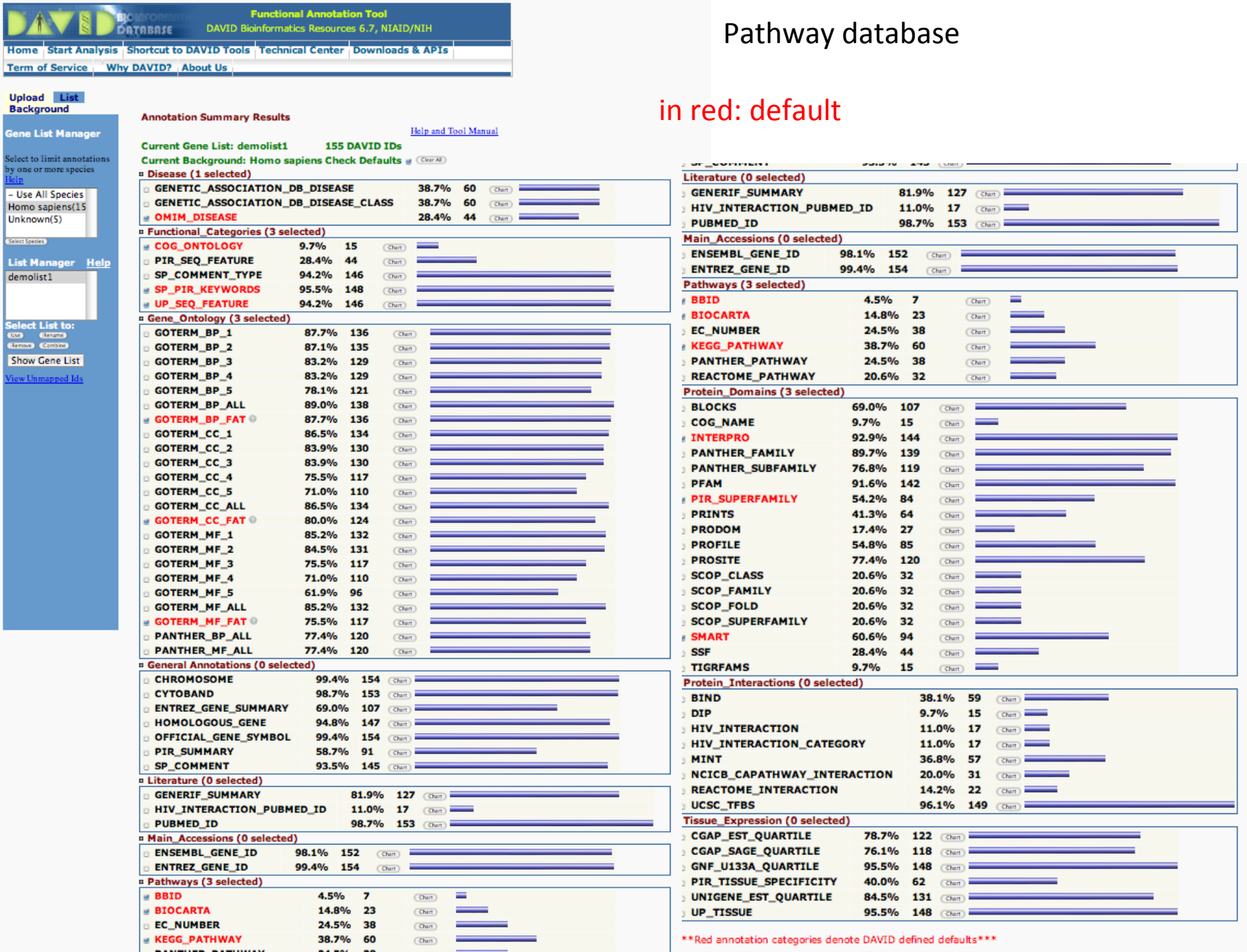
- DAVID has an automatic procedure to 'guess' the background.
- Most of the studies are genome-wide or close to genome-wide studies.

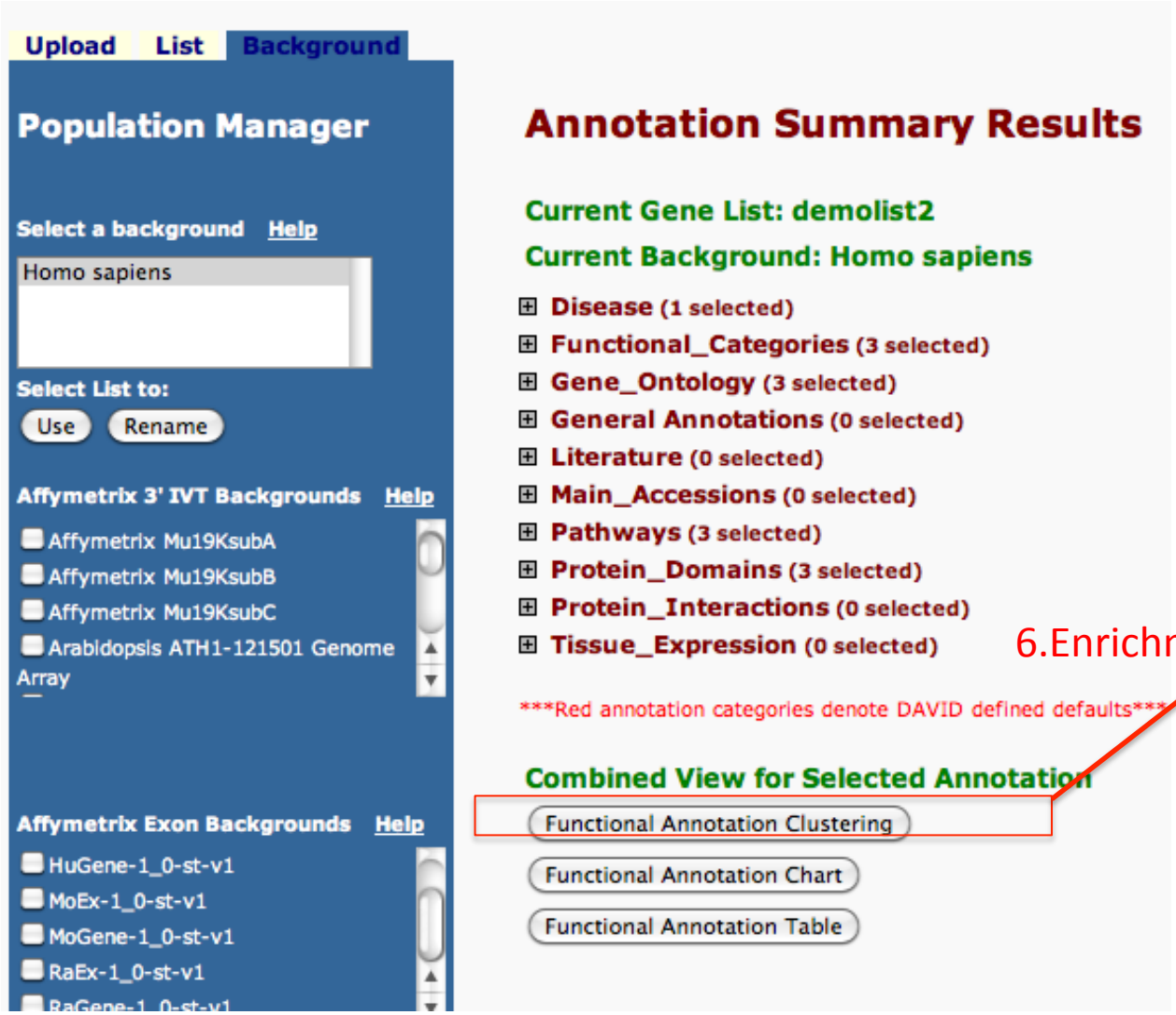
5 .Select the databases



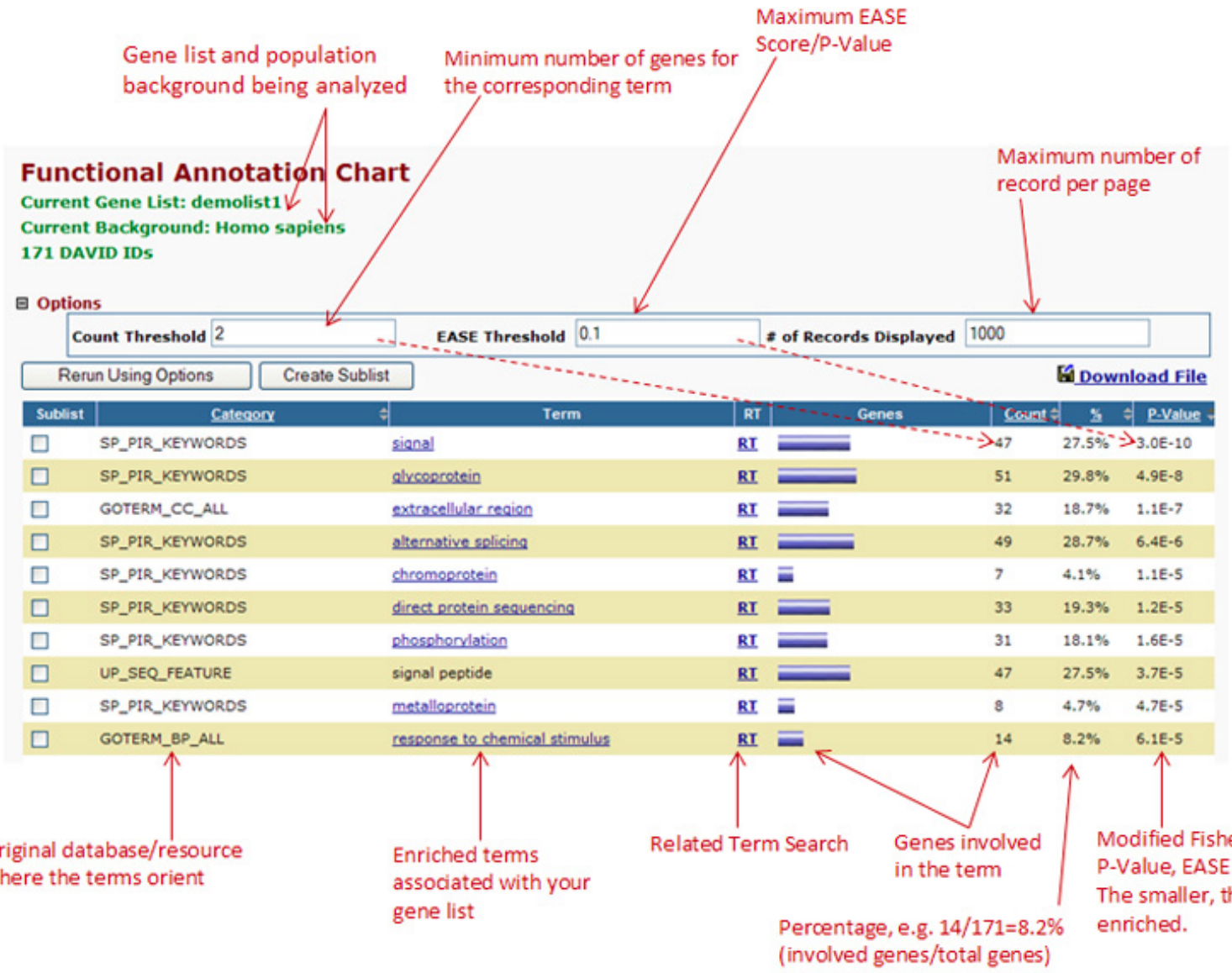
# Pathway database

in red: default

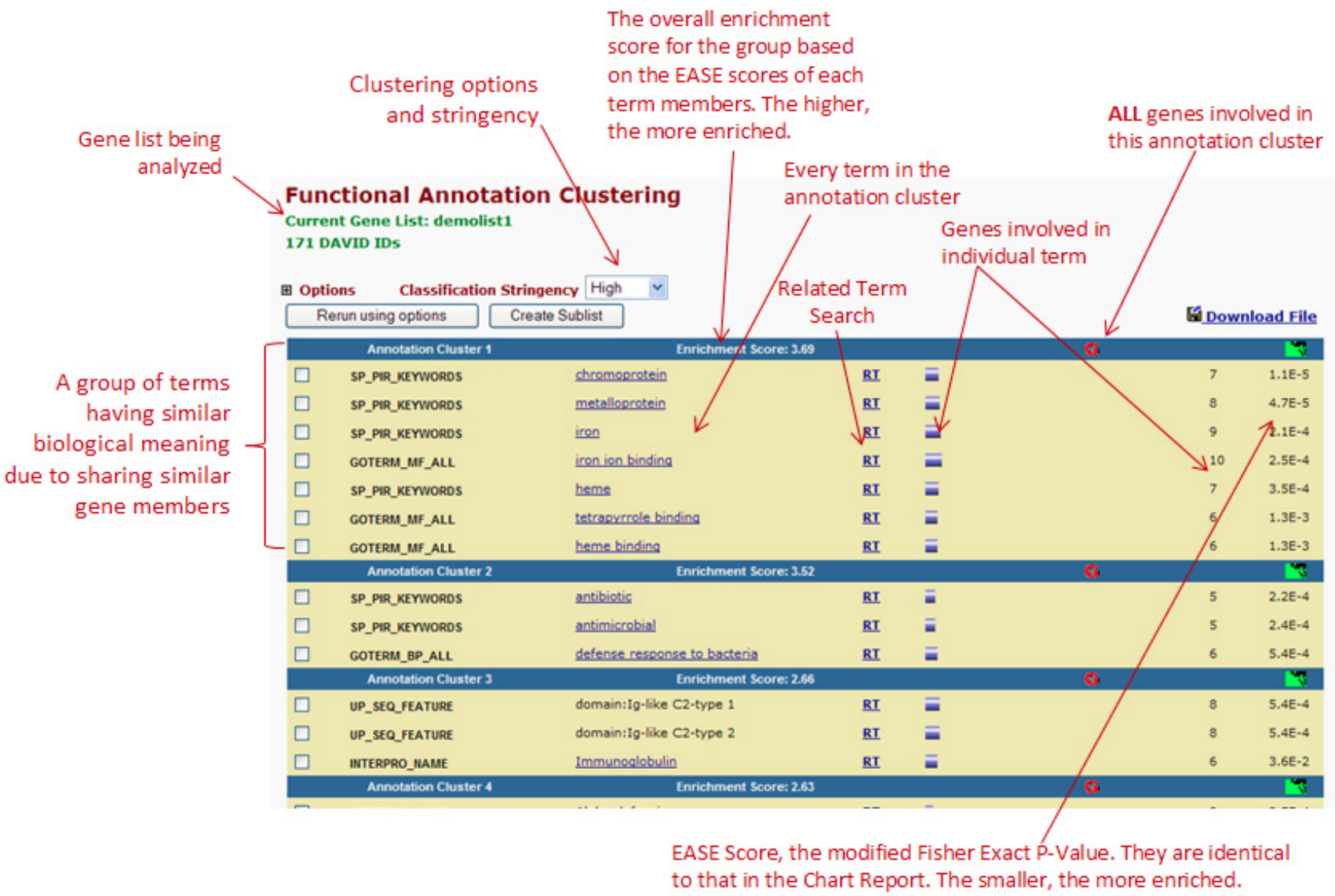



Upload
List
Background
Population Manager
Select a background
Help
Homo sapiens
Select List to:
Use
Rename
Affymetrix 3' IVT Backgrounds
Help
Affymetrix Mu19KsubA
Affymetrix Mu19KsubB
Affymetrix Mu19KsubC
Arabidopsis ATH1-121501 Genome Array
Affymetrix Exon Backgrounds
Help
HuGene-1_0-st-v1
MoEx-1_0-st-v1
MoGene-1_0-st-v1
RaEx-1_0-st-v1
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)



# 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)

Slide from Daniele Merico/ Oliver Stuecker / Bader lab

# The GSEA algorithm

## Rank the genes using a differential expression value

| 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) (CD | 8.906 | 12.154 | 3.[illegible] | 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 | 6.264 | 8.473 | 2.413 | 20.497 | 2.01E-12 | 1.41E-08 |
| 466855 | ILMN_2319077 | FAS | Fas (TNF receptor superfamily, member 6) (FAS), tra | 7.558 | 9.604 | 1.949 | 16.849 | 3.44E-11 | 1.57E-07 |
| 7458 | ILMN_2089875 | TNFSF4 | tumor necrosis factor (ligand) superfamily, member | 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), | 7.772 | 9.445 | 1.967 | 15.897 | 7.91E-11 | 2.07E-07 |
| 12544 | ILMN_1907834 | | cDNA: FLJ21679 fis, clone COL09221 | 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 co | 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), t | 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 (LOC | 6.422 | 6.036 | -0.003 | -0.026 | 9.79E-01 | 9.96E-01 |
| 85465 | ILMN_2400292 | MAPK9 | mitogen-activated protein kinase 9 (MAPK9), transc | 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 v | 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 | 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 | LOC100133316 | similar to breakpoint cluster region (LOC100133163) | 9.851 | 9.451 | -0.507 | -5.661 | 4.44E-05 | 6.01E-03 |
| 12544 | ILMN_1815184 | ASPM | asp (abnormal spindle) homolog, microcephaly assoc | 10.494 | 9.725 | -0.671 | -5.784 | 3.54E-05 | 4.95E-03 |
| 1005687 | ILMN_1651433 | DCK | deoxycytidine kinase (DCK), mRNA. | 10.316 | 9.932 | -0.533 | -5.930 | 2.71E-05 | 3.94E-03 |
| 85465 | ILMN_1802951 | CDCA1 | cell division cycle associated 1 (CDCA1), transcript va | 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 varia | 9.163 | 8.248 | -0.724 | -6.417 | 1.13E-05 | 1.93E-03 |
| 1235 | ILMN_1783621 | CMPK2 | cytidine monophosphate (UMP-CMP) kinase 2, mitoc | 8.833 | 8.327 | -0.706 | -6.447 | 1.07E-05 | 1.88E-03 |
| 1458 | ILMN_3239771 | DLGAP5 | discs, large (Drosophila) homolog-associated protein | 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) | 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 | 9.078 | 8.143 | -0.879 | -6.713 | 6.77E-06 | 1.37E-03 |
| 9854777 | ILMN_1749829 | DLGAP5 | discs, large (Drosophila) homolog-associated protein | 10.167 | 9.184 | -1.006 | -7.167 | 3.15E-06 | 7.70E-04 |

Up-regulated in treated

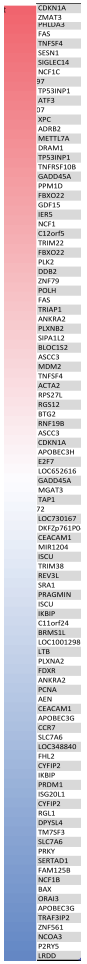
Down-regulated in treated

UP

Ranked list

> 40 000 probes

DOWN



# Enrichment score

**Enrichment score from Mootha et al. 2003:**

N: gene list length (20000)
G: # of genes in the gene-set (50)

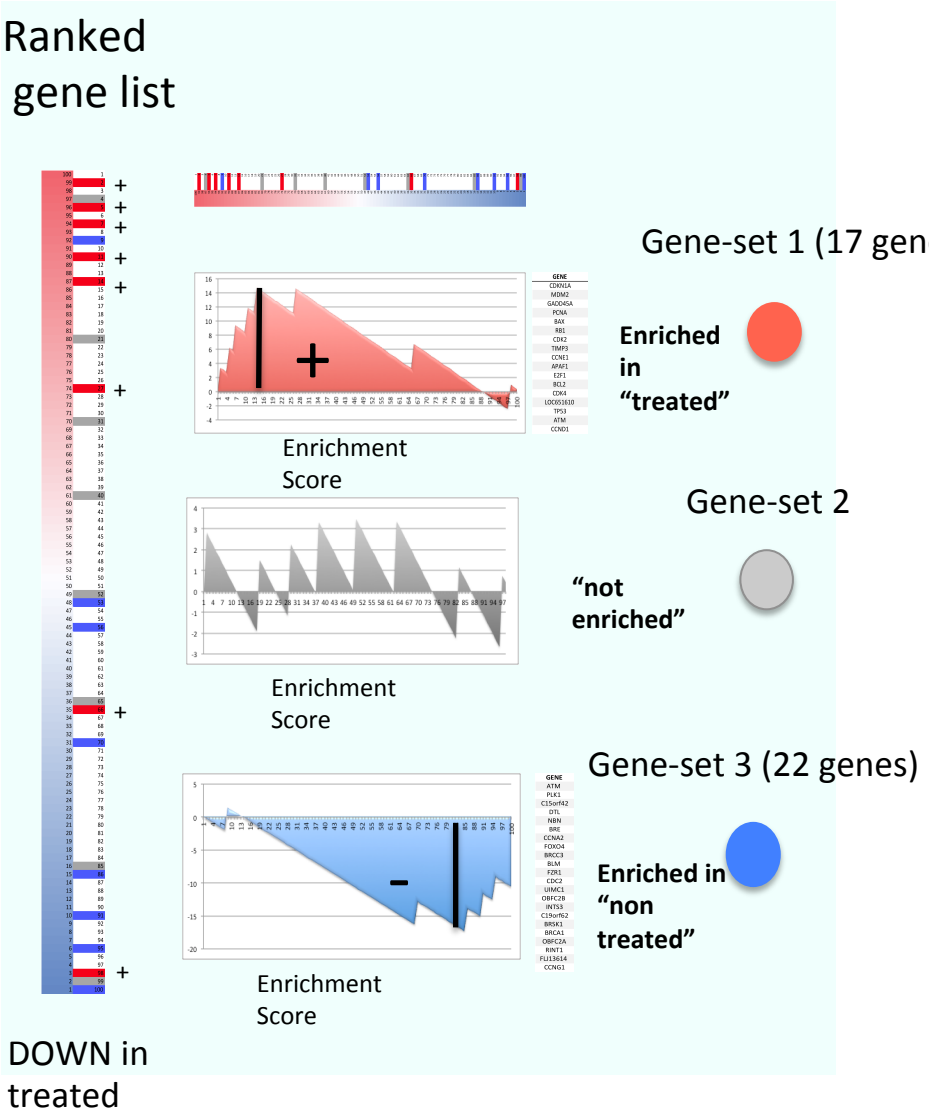
if the gene i is in the gene-set:

$X_i = \sqrt{((N-G) / G)}$
$X_i = \sqrt{((20000-50) / 50)}$
$X_i = 20$

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

$X_i = -\sqrt{(G / N-G)}$
$X_i = -\sqrt{( 50 / (20000-50))}$
$X_i = -0.05$

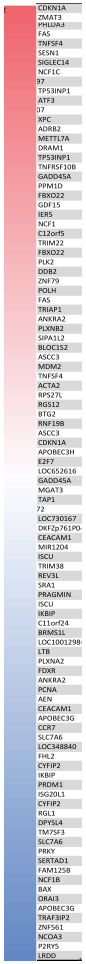
The running sum of X is calculated until we reach the last gene in the ranked list.



# Calculate the enrichment score for each gene-set

Ranked list

UP



> 40 000 probes

DOWN

Pathway information from multiple source

Gene Ontology

PFAM

KEGG

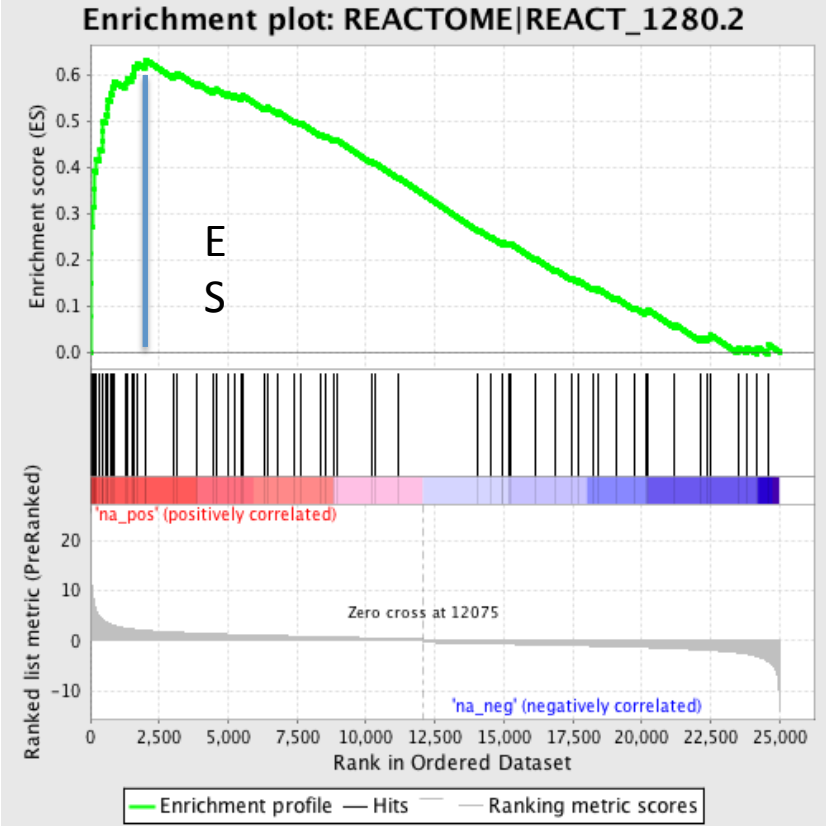
HOMOLOGY DATA

NCI

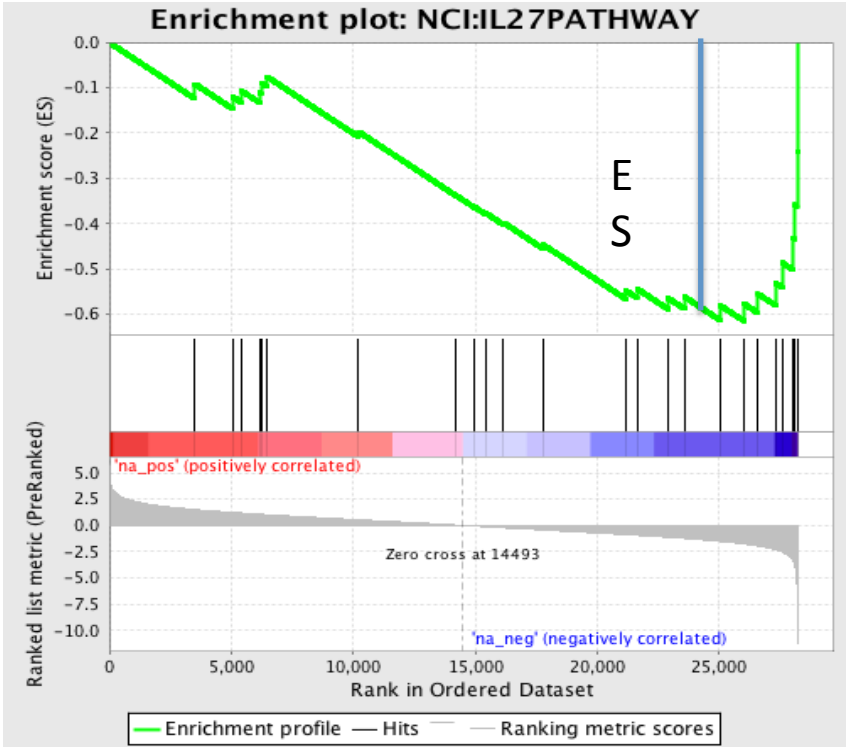
BIOCARTA

- 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

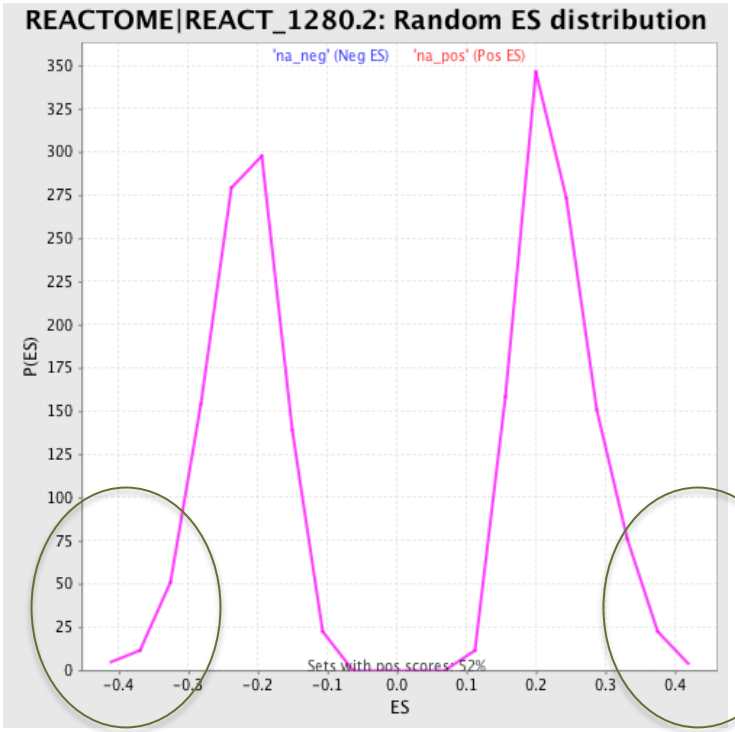
## positive enriched geneset




## negative enriched geneset



1. Null distribution




P -value

**Significant enrichment p-value**

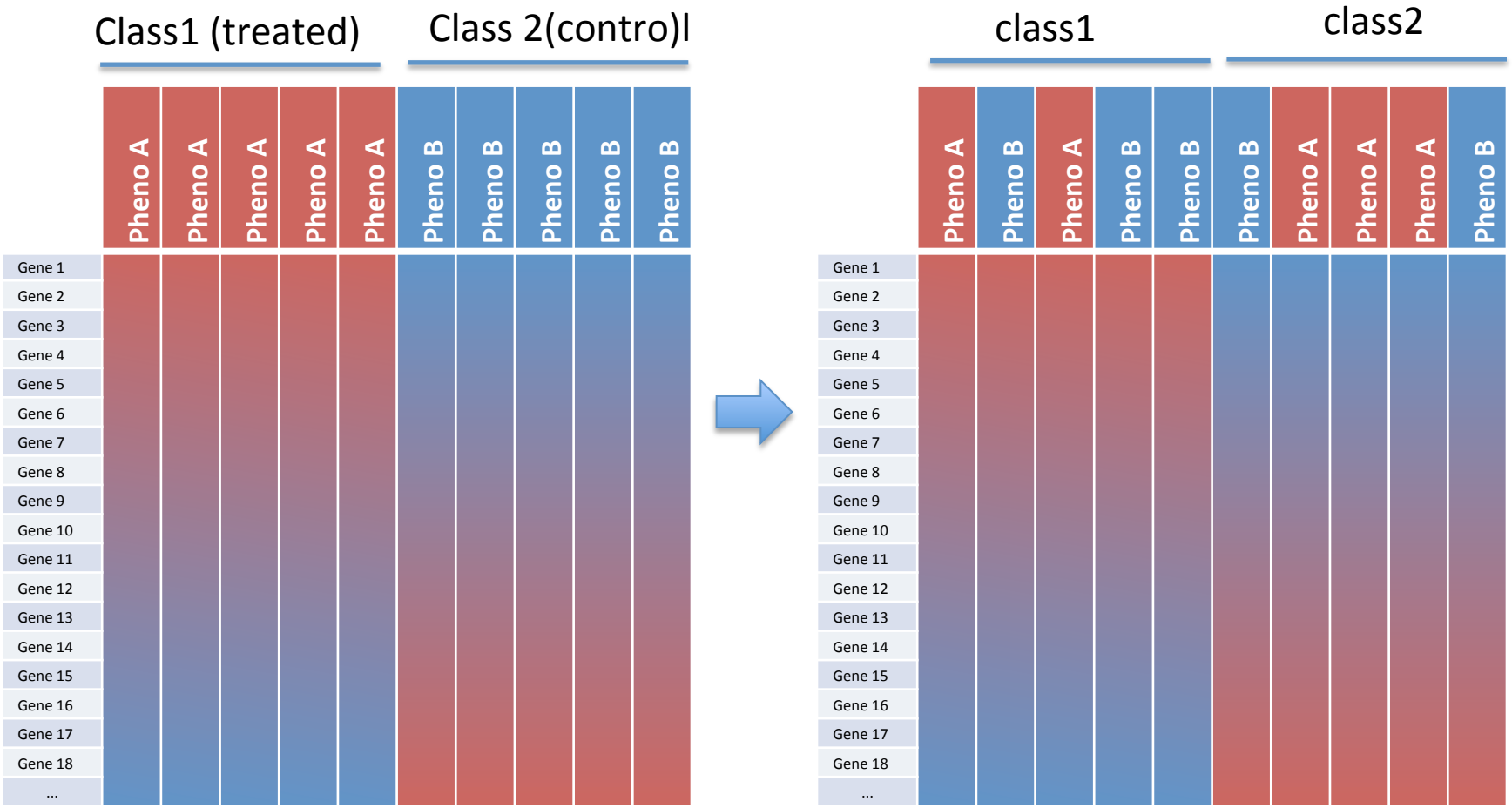
**Significant enrichment p-value**

ES

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

# Phenotype Permutation

## shuffling phenotype labels (1000-2000 times)

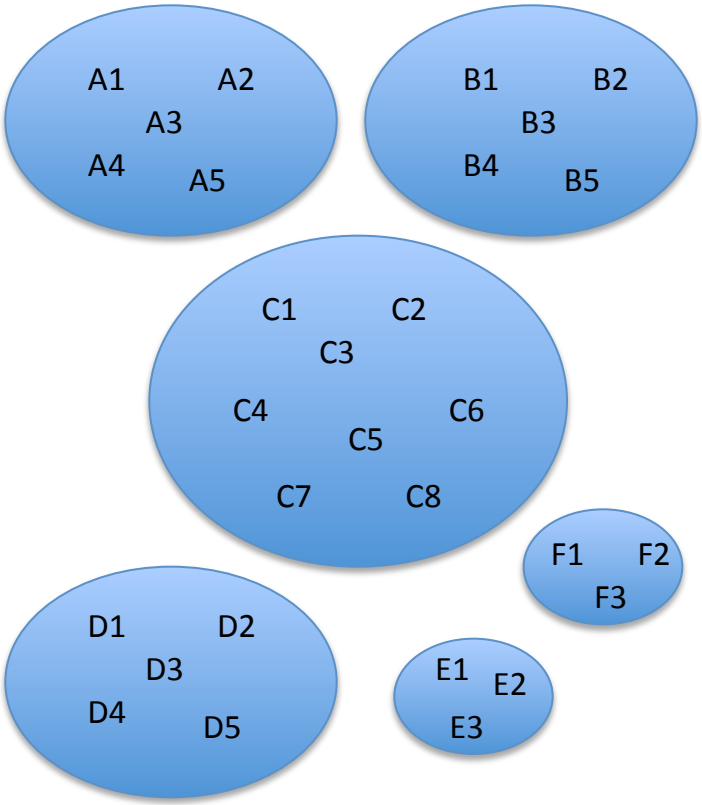



Slide from Daniele Merico/ Oliver Stuecker / Bader lab

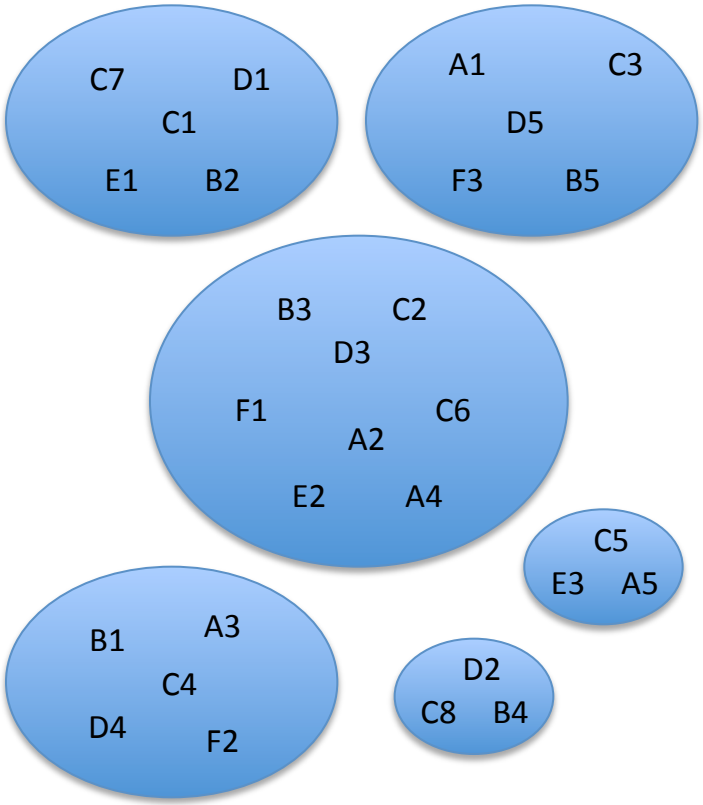
# Geneset Permutation

1000-2000 times

**original geneset**

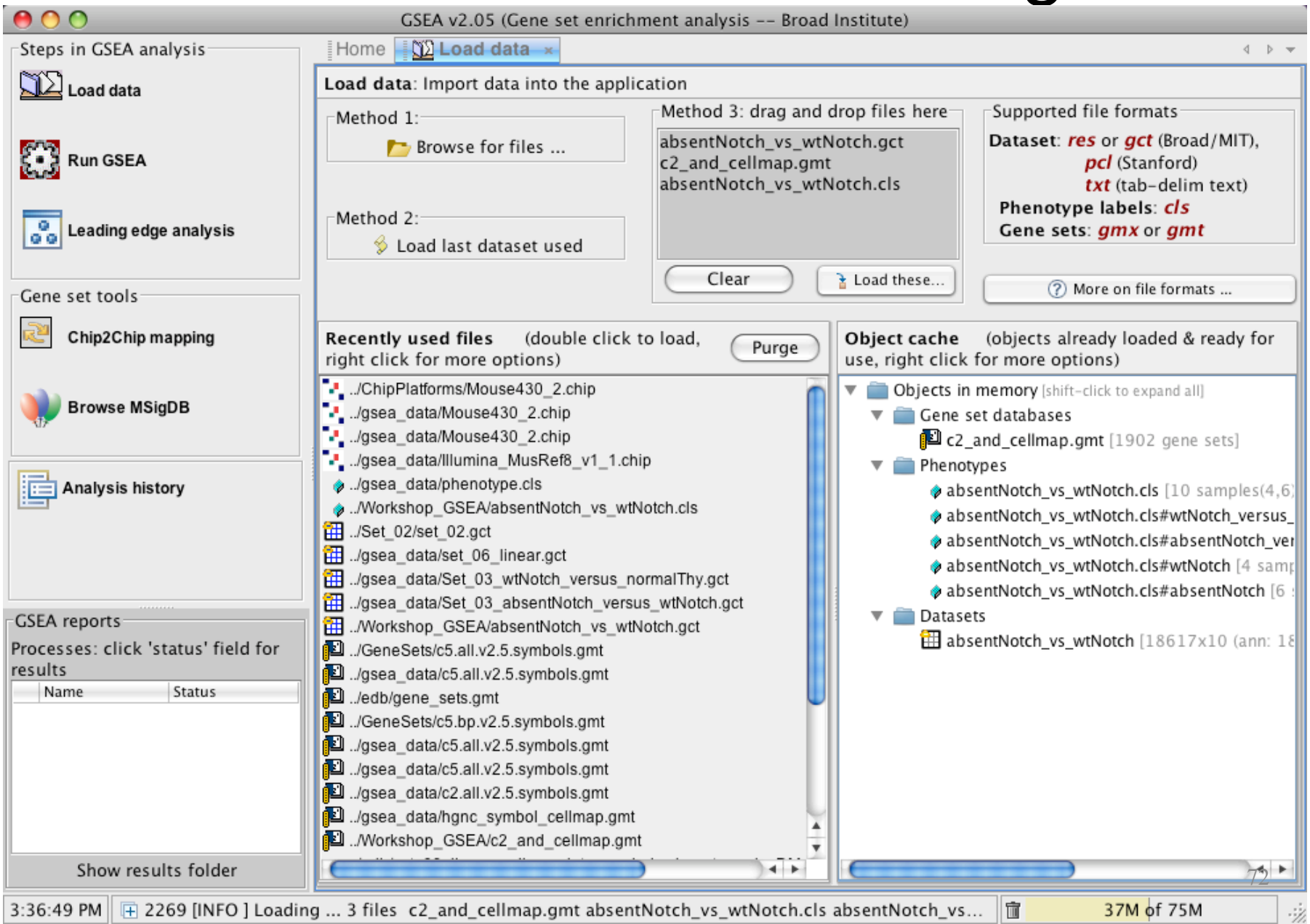



**random geneset**

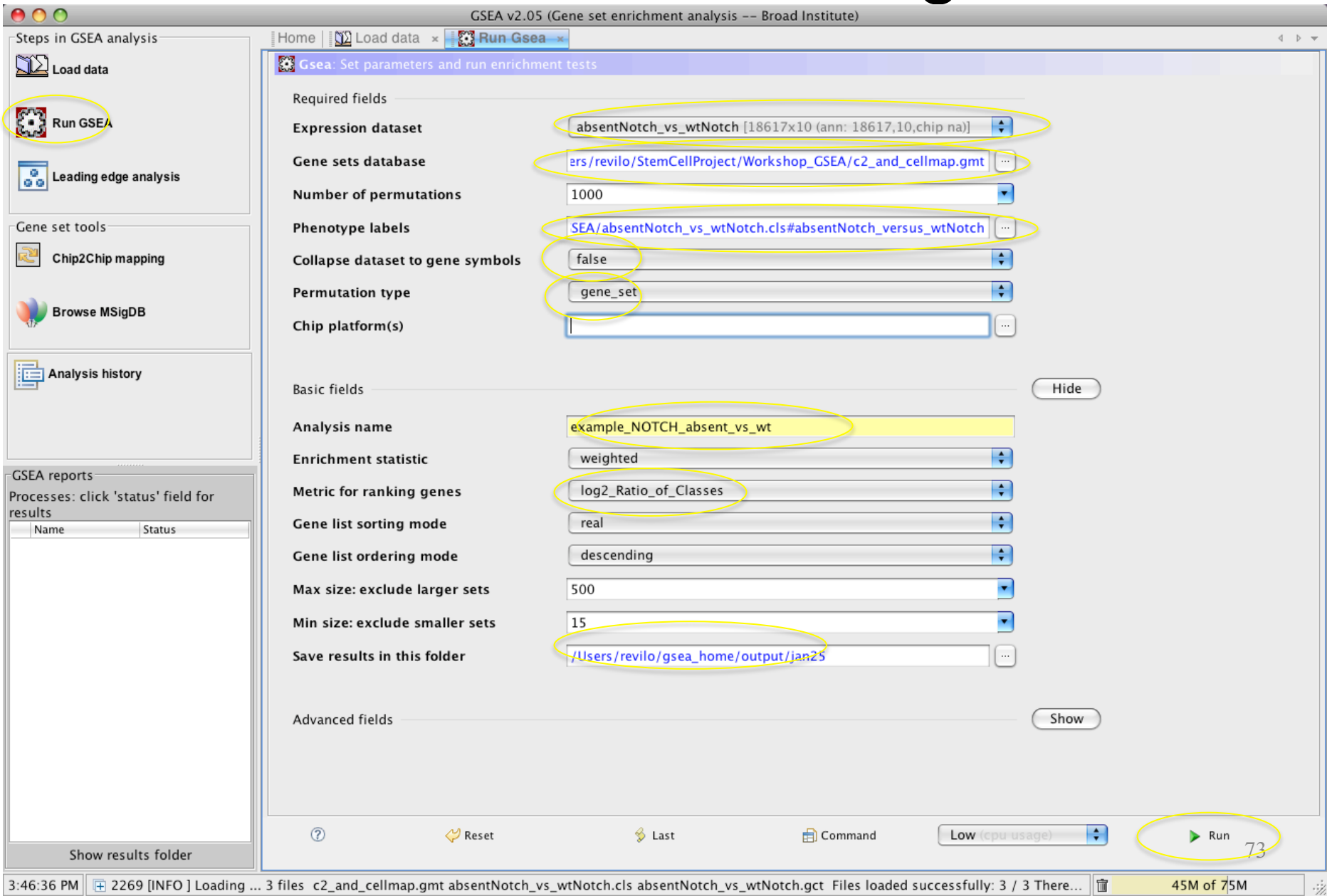



Slide from Daniele Merico/ Oliver Stuecker / Bader lab

# GSEA Load Data dialog



# Run GSEA dialog



# Expression dataset (.txt)

| Entrez_Gene_ID | DESCRIPTION | treated1 | control1 | treated2 | control2 | treated3 | control3 |
|---|---|---|---|---|---|---|---|
| 285741 | LOC285741 ::PREDICTED: misc_RNA (LOC28 | 11.941118 | 12.611188 | 8.806047 | 12.565253 | 10.38005 | 9.248268 |
| 100008589 | LOC100008589 :: 28S ribosomal RNA (LOC1( | 12.402481 | 12.600485 | 11.416079 | 11.966387 | 12.392032 | 11.706466 |
| 442727 | LOC442727 ::PREDICTED: misc_RNA (LOC44 | 15.125171 | 15.756729 | 13.360757 | 15.476214 | 14.241237 | 13.291751 |
| 7915 | ALDH5A1 :: aldehyde dehydrogenase 5 fami | 9.550562 | 9.901636 | 9.46661 | 9.802373 | 9.546112 | 9.56268 |
| 29094 | HSPC159 :: galectin-related protein (HSPC15 | 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 | 7.6873574 | 7.912635 | 7.979784 | 8.256096 | 8.202624 | 8.424054 |
| 841 | CASP8 :: caspase 8, apoptosis-related cystei | 7.217502 | 7.244122 | 7.308674 | 7.0536304 | 7.543329 | 7.321113 |
| 9595 | PSCDBP :: pleckstrin homology, Sec7 and co | 8.920641 | 9.557003 | 9.01201 | 9.311855 | 8.998903 | 8.831511 |
| 132 | ADK :: adenosine kinase (ADK), transcript va | 7.3660574 | 7.609104 | 7.6317477 | 7.130101 | 7.468771 | 7.3359246 |
| 641975 | LOC641975 ::PREDICTED: hypothetical prot | 8.157521 | 8.624233 | 8.089965 | 8.85631 | 8.054138 | 8.382164 |
| 647089 | LOC647089 ::PREDICTED: hypothetical LOC( | 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 E2L | 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-myo-inositol hexakisphosphate biosyı | 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 delimitated 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 * absolute(μ), where μ=0 is adjusted to μ=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 * absolute(μ), where μ=0 is adjusted to μ=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.

http://www.broadinstitute.org/gsea/doc/GSEAUserGuideFrame.html

# 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:

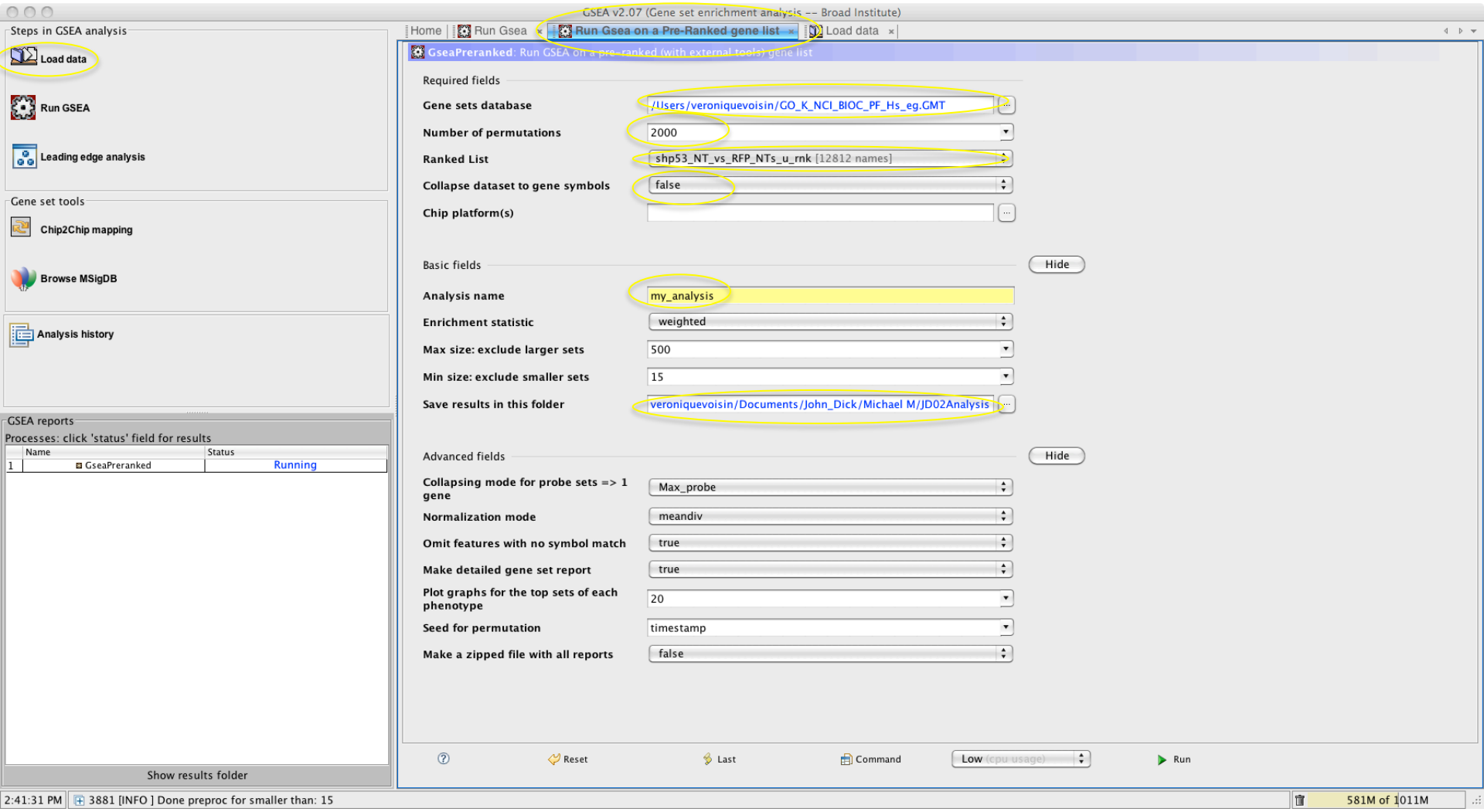
- SAM
- ANOVA
- any other algorithm that can score and rank the genes according a 1- or 2-class model (e.g. TREAT)

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 pre-ranked

- Gene set databse (.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

Ranked 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 feat

### Gene set details

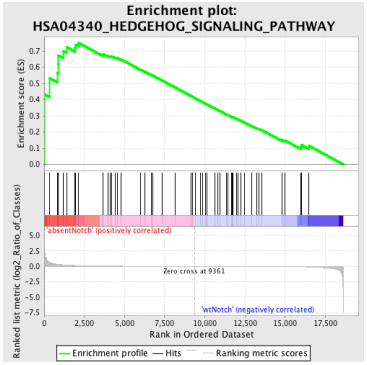
- Gene set size filters (min=10, max=500) resulted in filtering out 279 / 1902
- The remaining 1623 gene sets were used in the analysis
- List of gene sets used and their sizes (restricted to features in the specifie

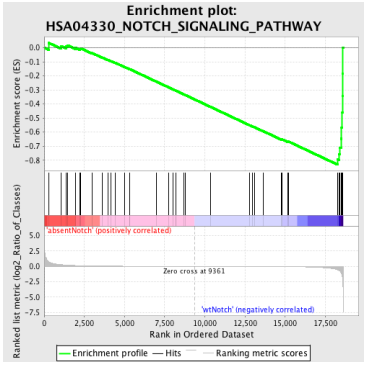
### Gene markers for the absentNotch *versus* wtNotch compariso

- The dataset has 18617 features (genes)
- # of markers for phenotype **absentNotch**: 9361 (50.3% ) with correlation a
- # of markers for phenotype **wtNotch**: 9256 (49.7% ) with correlation area
- 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<br> 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_TRAFFICKINGPATHW | NCI:ARF6_TRAFFICKINGPATHWAY | Details ... | 41 | 0.5889985 | 1.9386604 | 0 | 0.05997115 | 0.6645 | 3384 | tags=51%, list=26%, signal=69% |

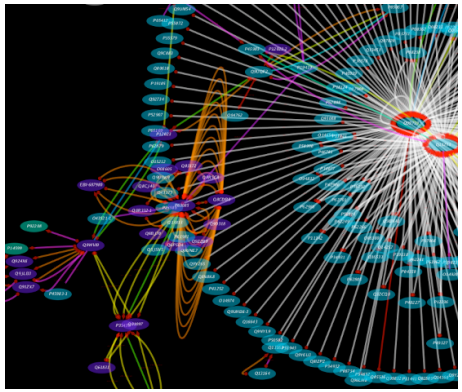


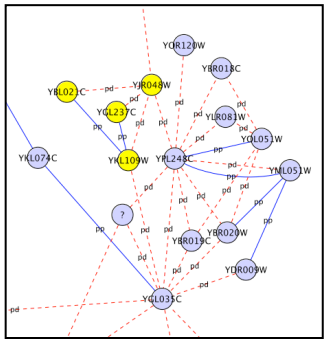



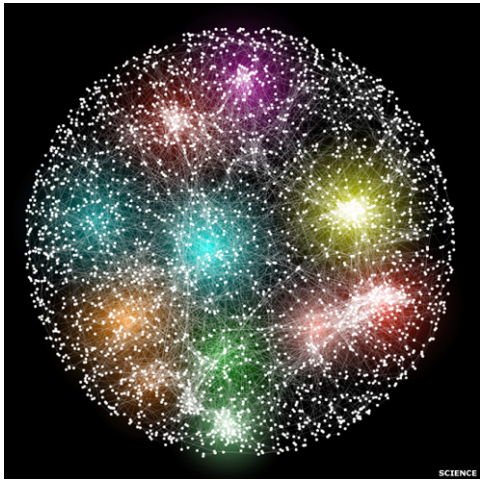
# Cytoscape

is an

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





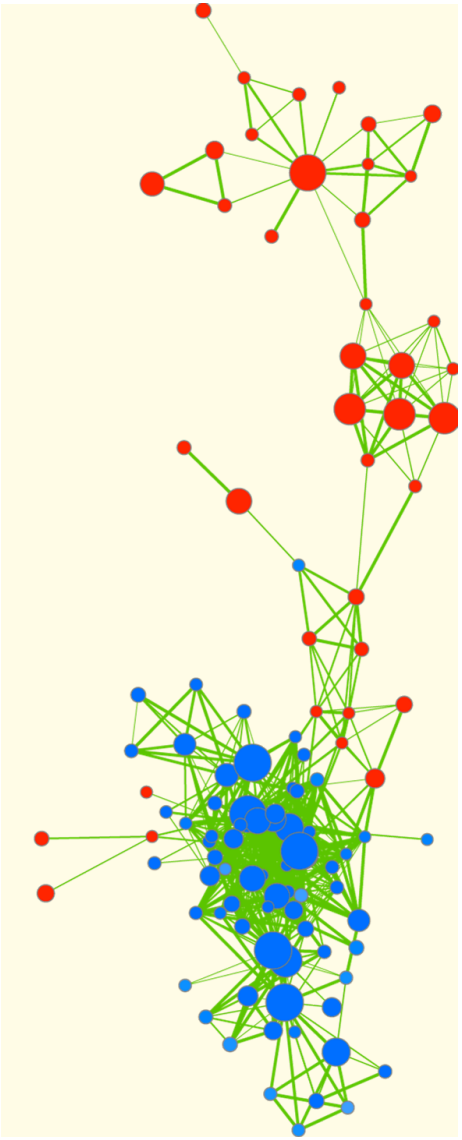
# 5) Enrichment Map (EM)

http://baderlab.org/Software/EnrichmentMap

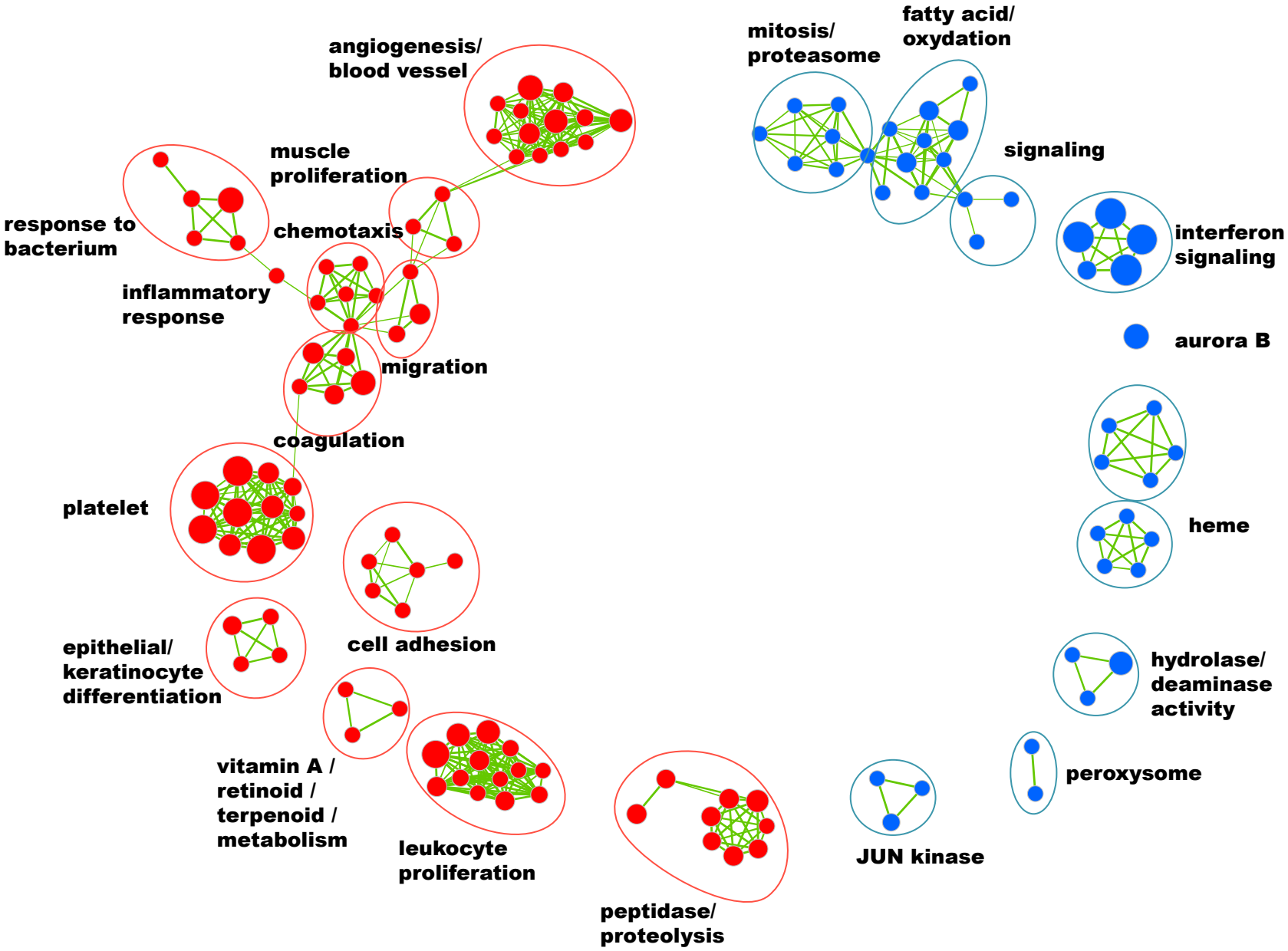
Enriched in…

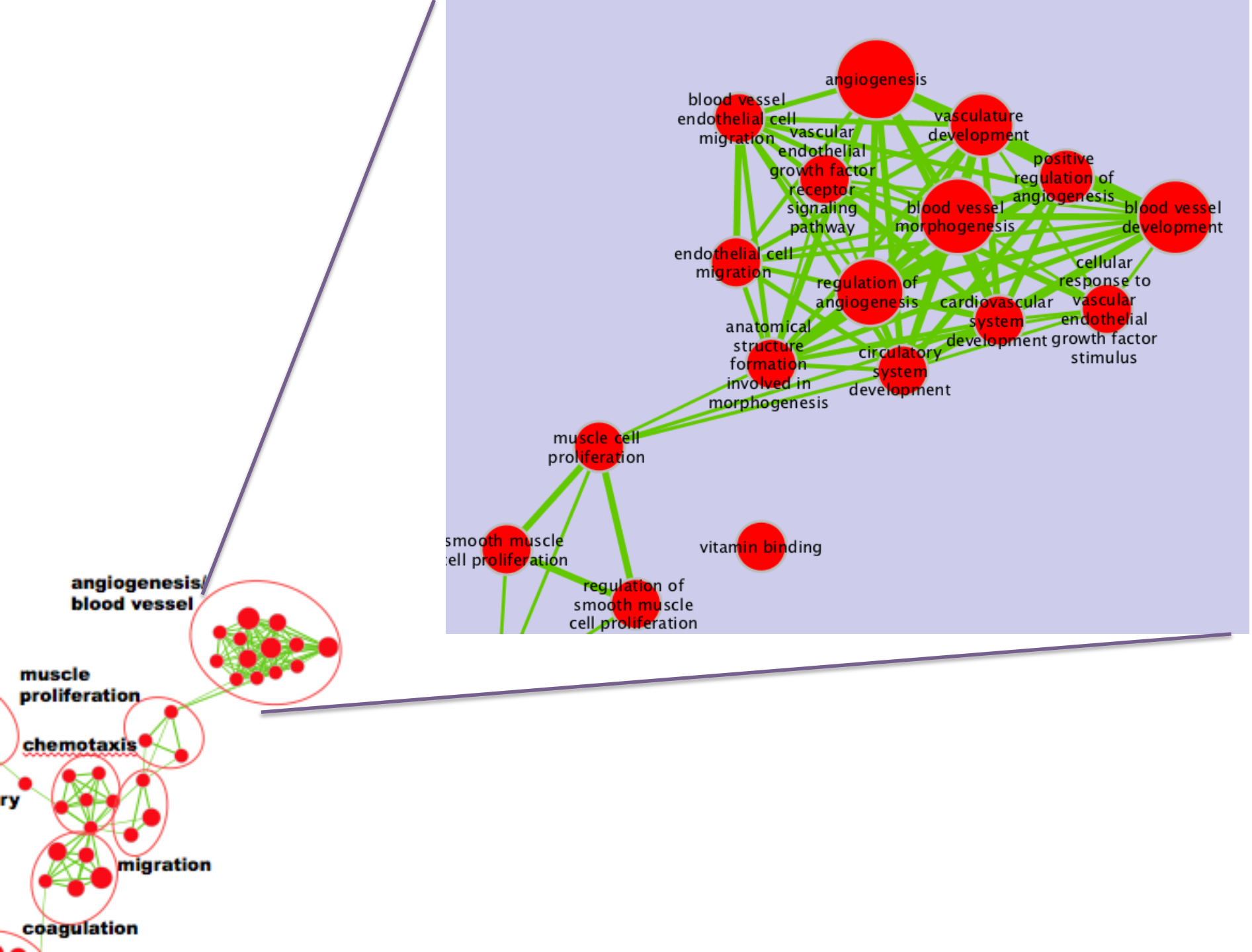
- treated
- non-treated
- gene overlap between 2 gene-sets

color intensity
depends on the
nerichment p-value



gene overlap

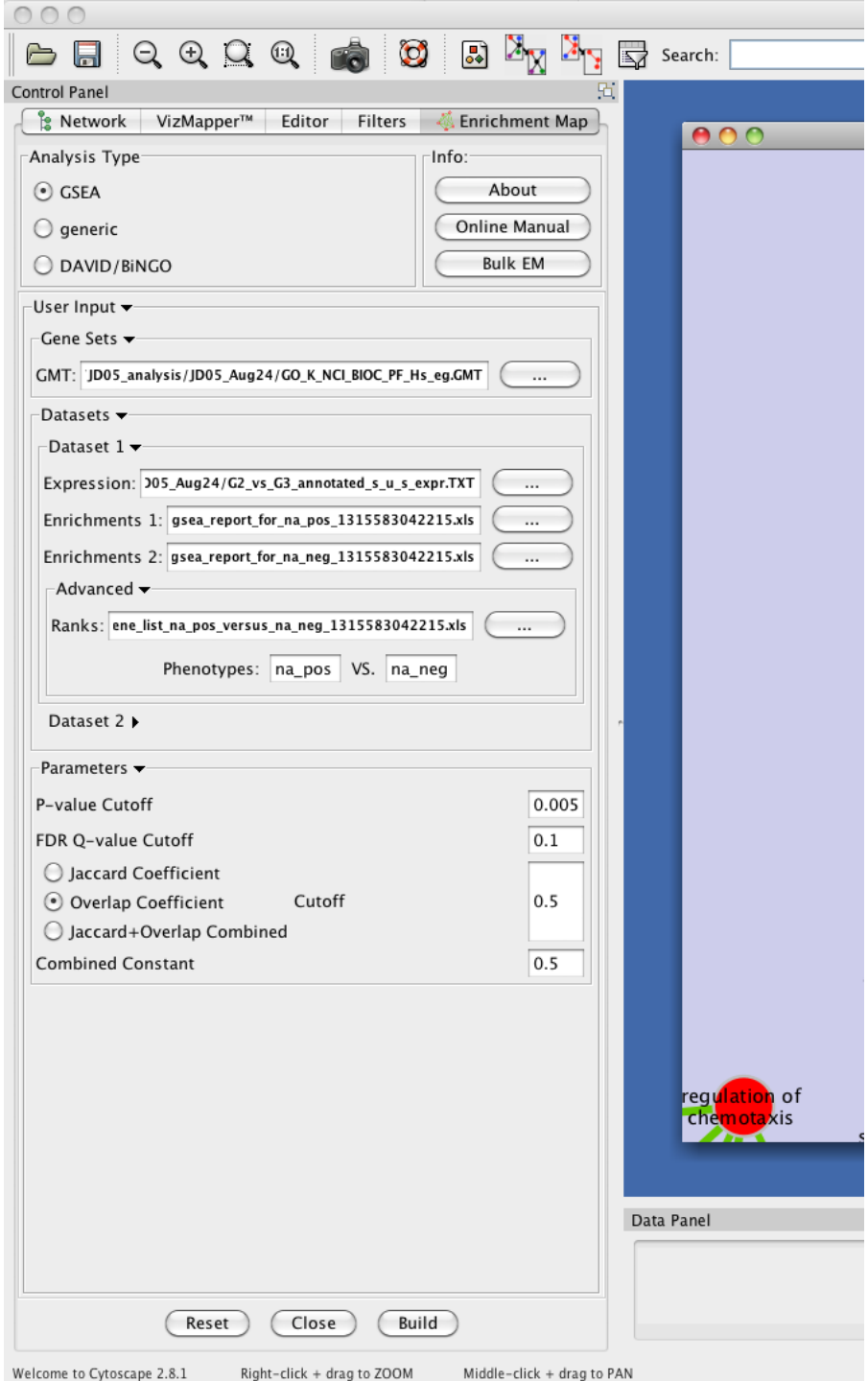

angiogenesis/
blood vessel
muscle
proliferation
response to
bacterium
chemotaxis
inflammatory
response
migration
coagulation
platelet
epithelial/
keratinocyte
differentiation
cell adhesion
vitamin A /
retinoid /
terpenoid /
metabolism
leukocyte
proliferation
peptidase/
proteolysis
JUN kinase
peroxysome
hydrolase/
deaminase
activity
heme
aurora B
interferon
signaling
signaling
mitosis/
proteasome
fatty acid/
oxydation

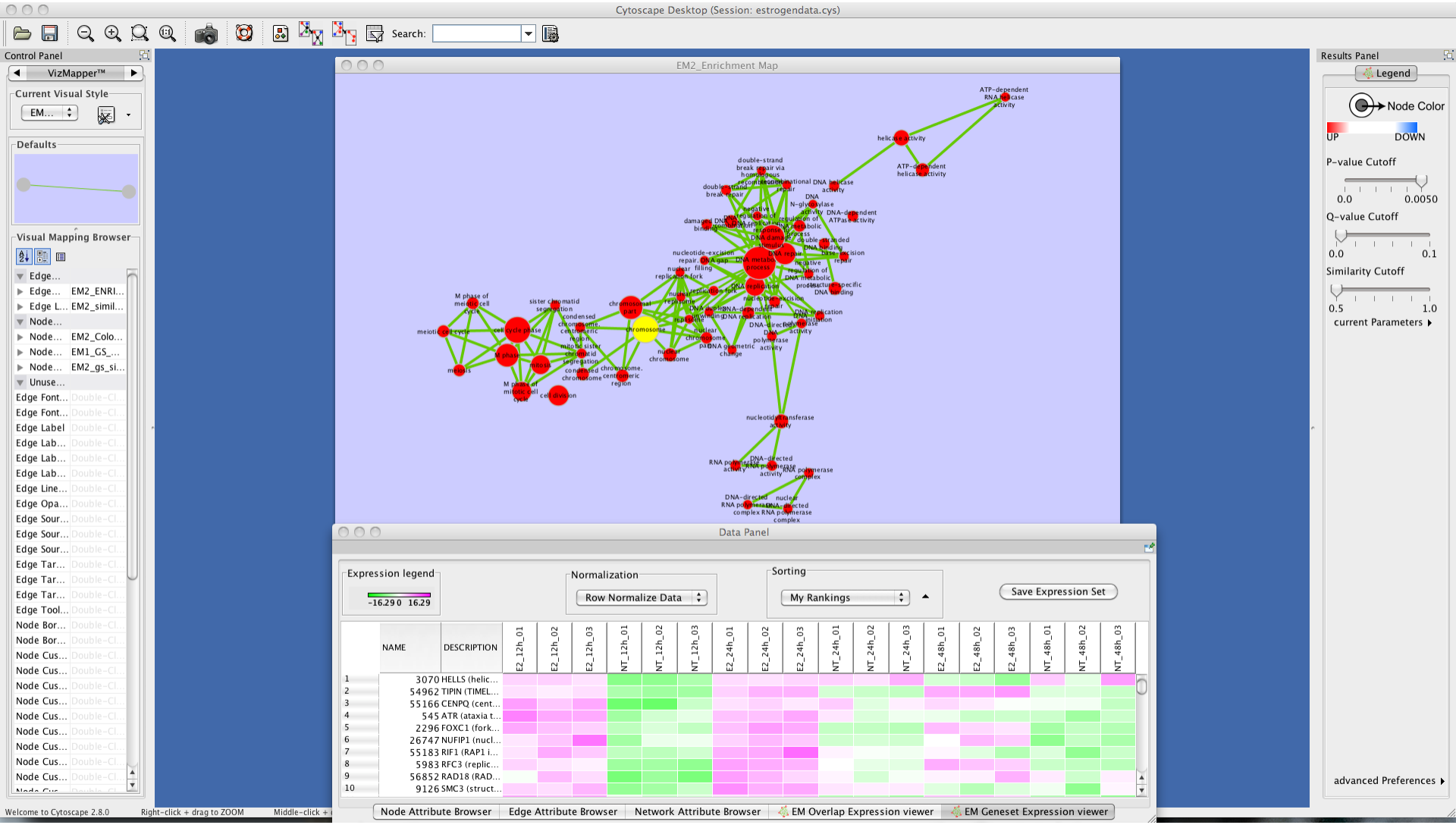

angiogenesis
blood vessel endothelial cell migration
vascular endothelial growth factor receptor signaling pathway
vasculature development
positive regulation of angiogenesis
blood vessel morphogenesis
blood vessel development
endothelial cell migration
regulation of angiogenesis
cellular response to vascular endothelial growth factor stimulus
cardiovascular system development
anatomical structure formation involved in morphogenesis
circulatory system development
muscle cell proliferation
smooth muscle cell proliferation
vitamin binding
regulation of smooth muscle cell proliferation
angiogenesis
blood vessel
muscle proliferation
chemotaxis
migration
coagulation

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

To run enrichment map , you need:

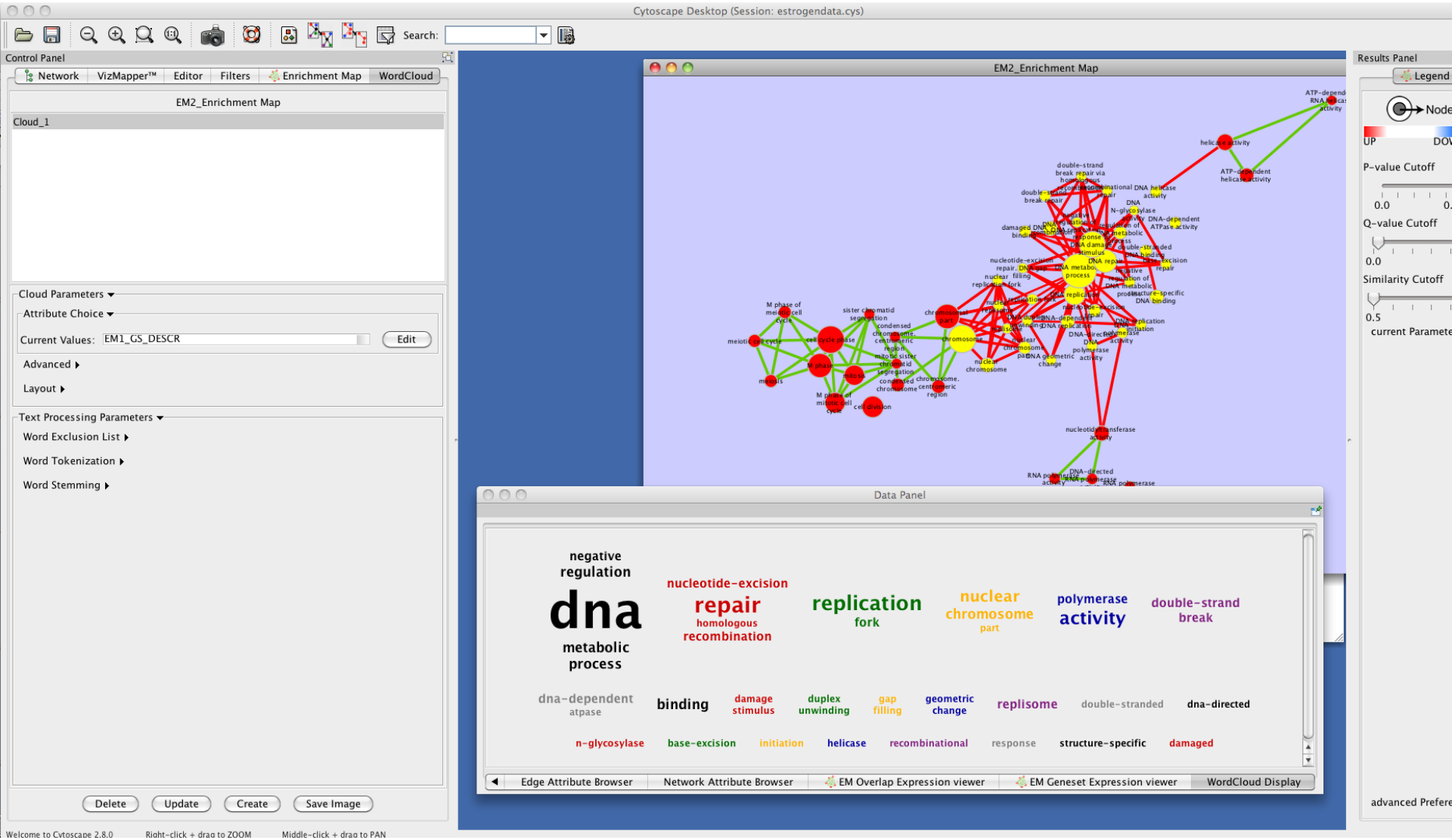
1) GMT file .GMT : provided
2) Expression file .TXT
3) GSEA reports (use the .rpt file in the GSEA folder to upload the rank file and the enrichment files)





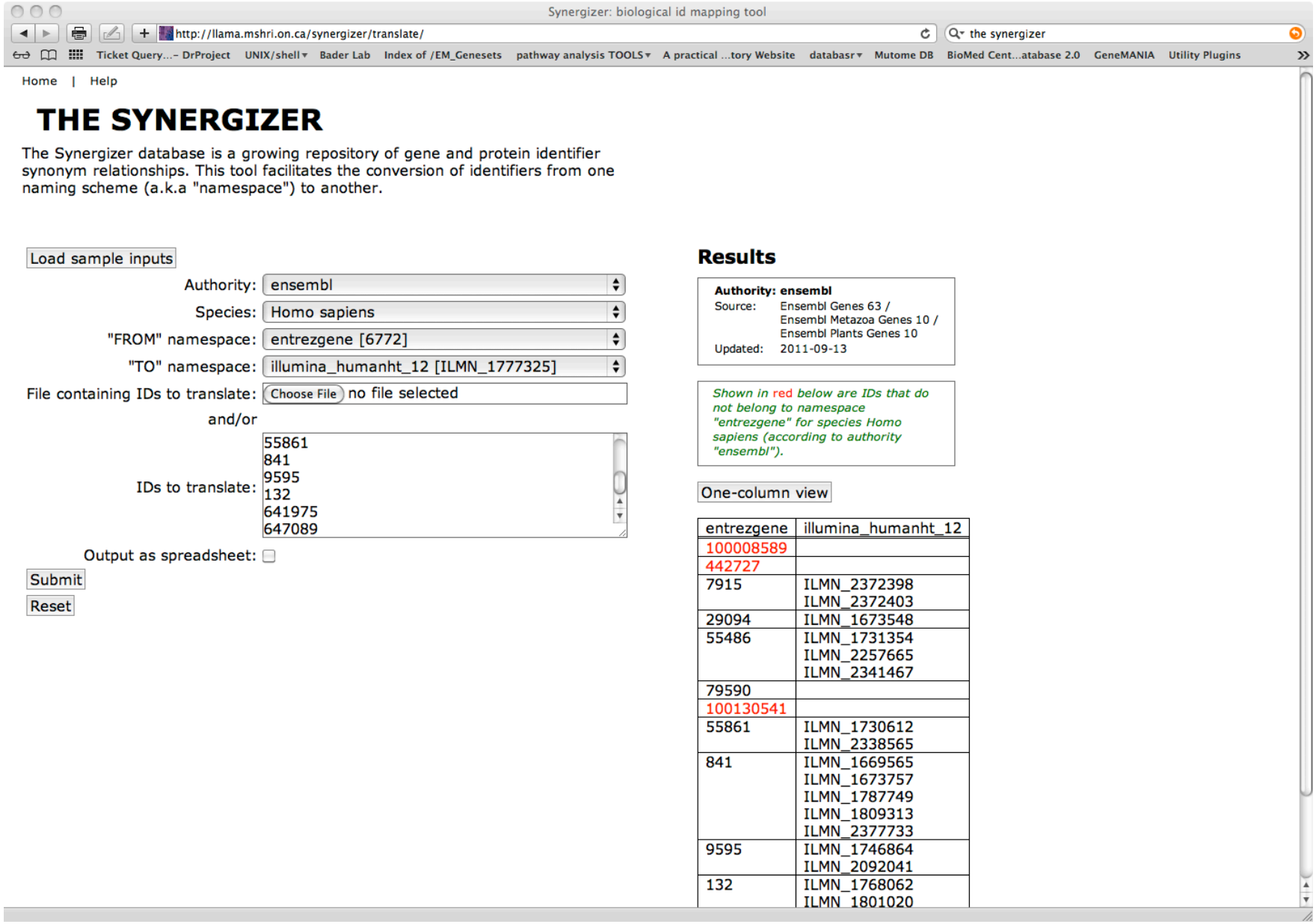

Save your map: .cys file

# WordCloud: is a visual representation for text data

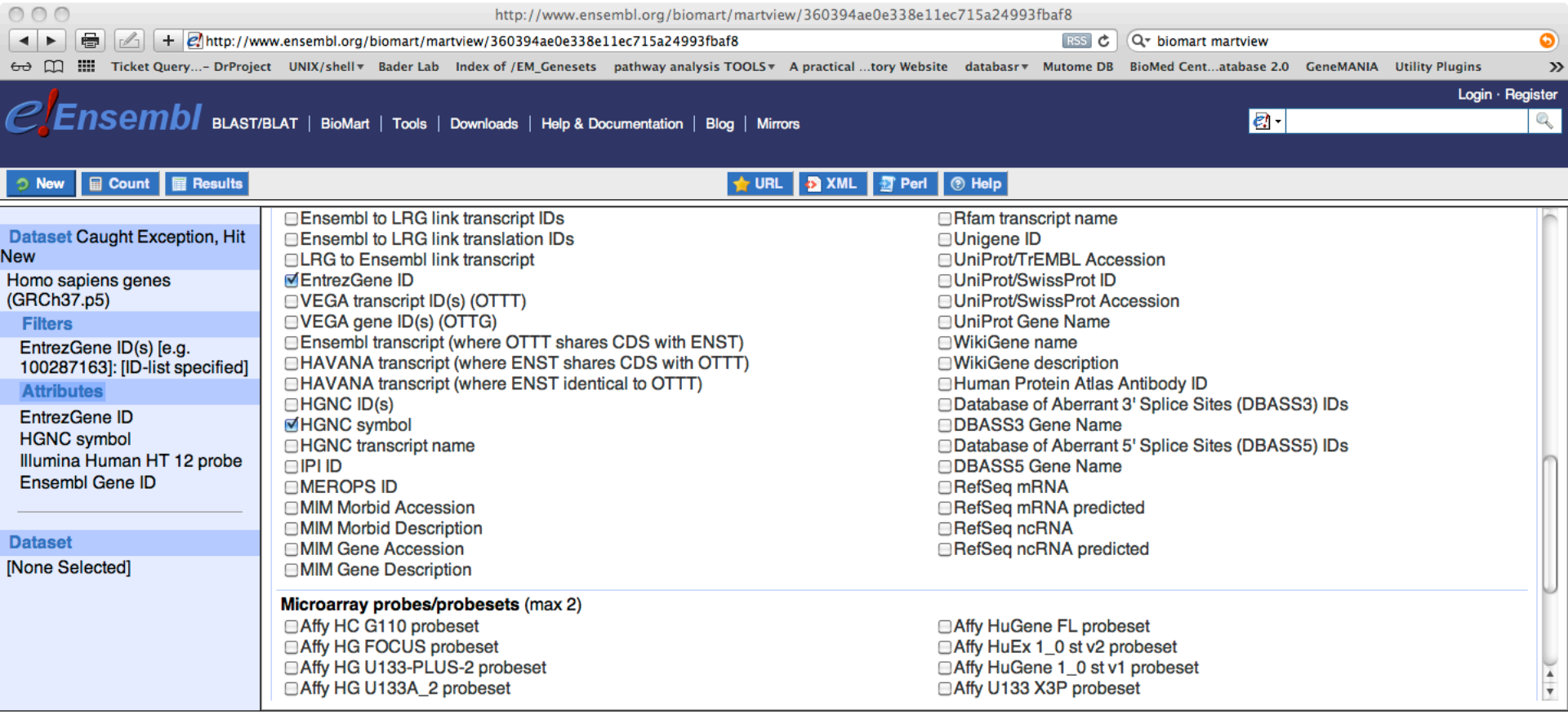



http://baderlab.org/WordCloud

# How to convert between different gene identifiers



# 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
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
Microarray probes/probesets (max 2)
Affy HC G110 probeset
Affy HG FOCUS probeset
Affy HG U133-PLUS-2 probeset
Affy HG U133A_2 probeset
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**