



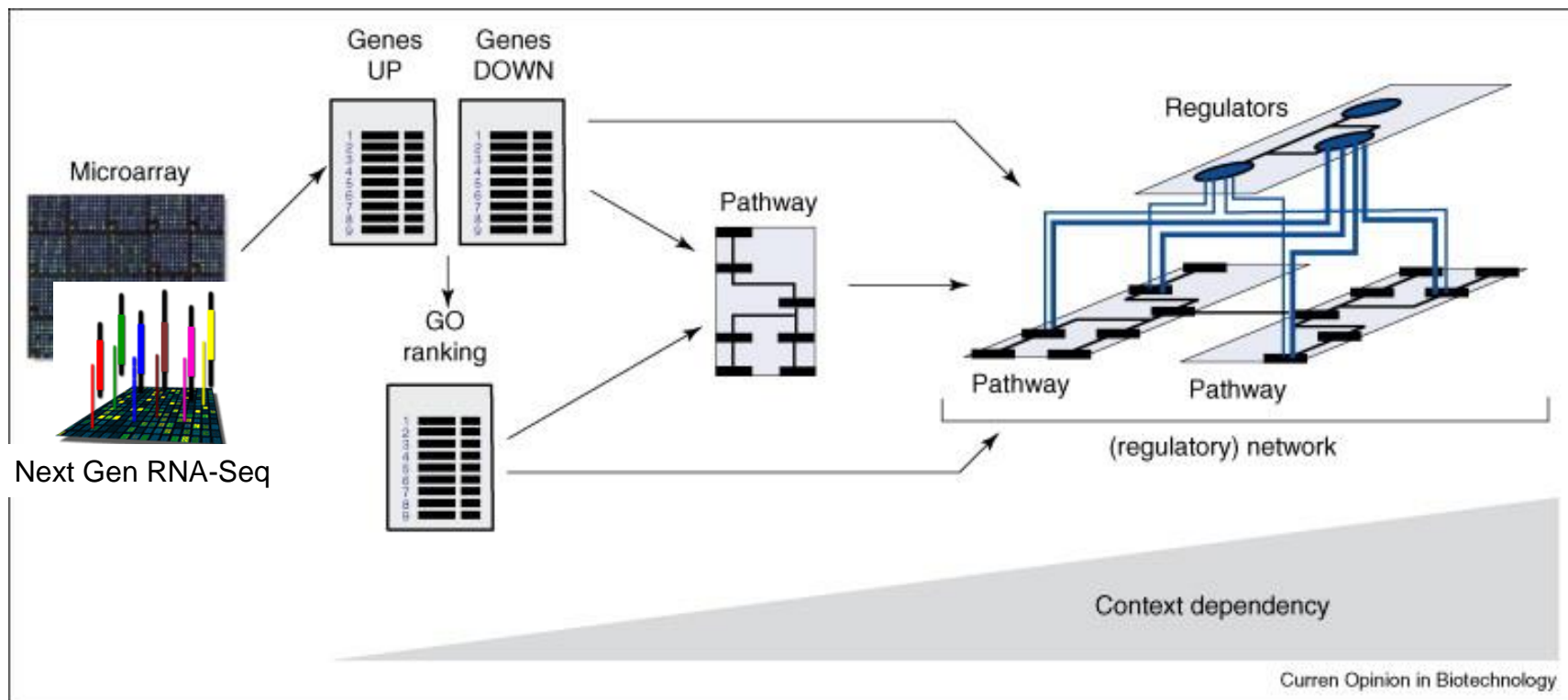
National Institute of Environmental Health Sciences  
*Your Environment. Your Health.*

# Pathway analysis

Biostat & Bioinfo short course series

Jianying Li

# The Road to Pathway Analysis

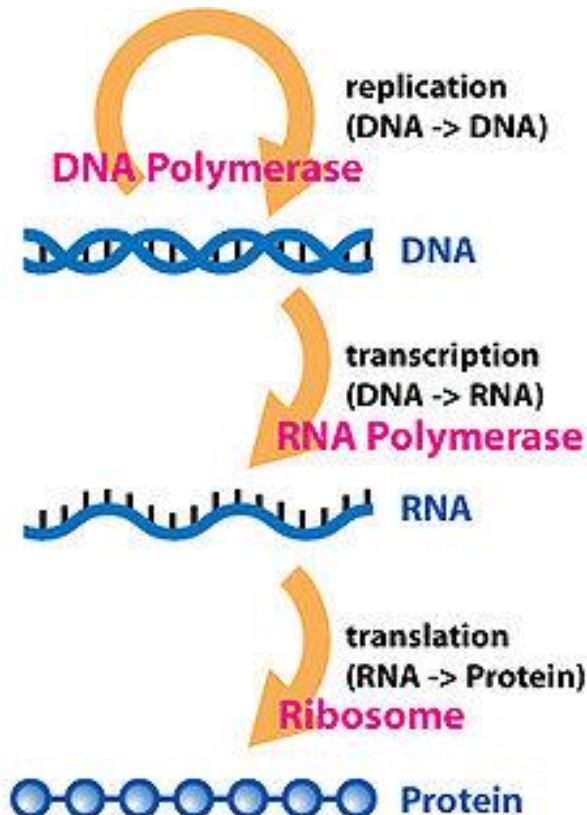


# Outline of the pathway analysis

- Knowledge based pathway analysis
  - Available knowledge bases
- Two major statistical approaches
  - Parametric: Hypergeometric/Fisher Exact Test (FET)
  - Non-parametric: Kolmogorov-Smirnov/ranking statistics
- Hands on practice
  - Parametric --> IPA
  - Non-parametric --> GSEA
- The proceedings in pathway analysis research

# What do we mean by pathway?

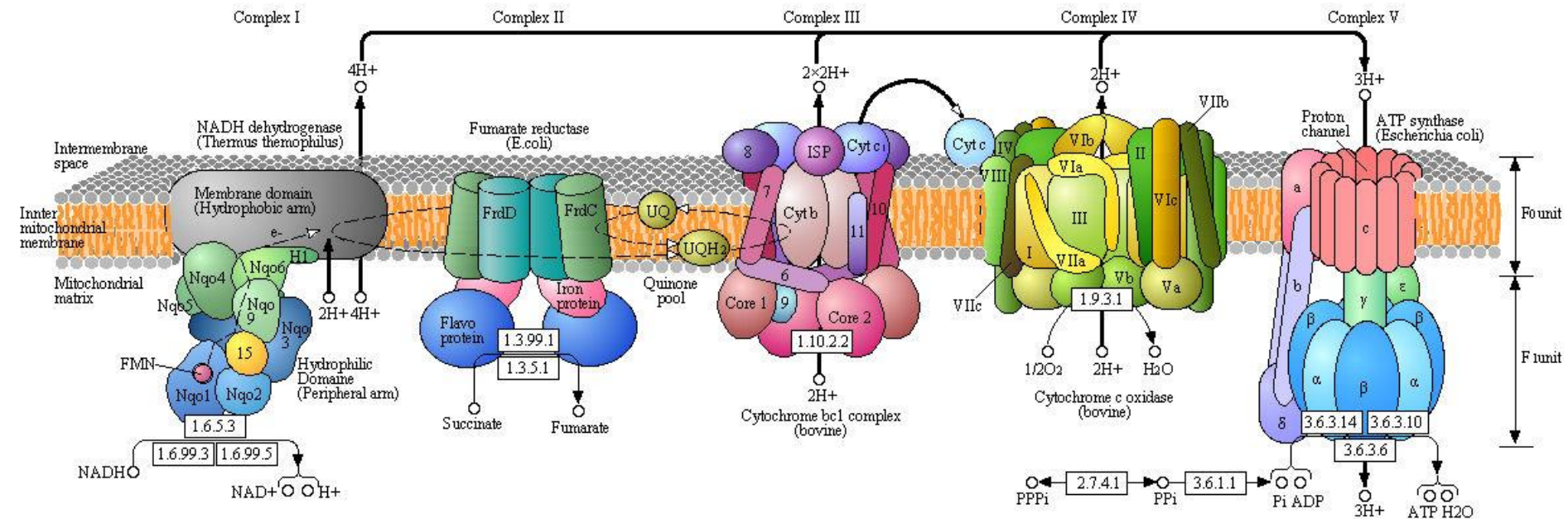
## Central Dogma



## Involvement of Gene Products

- Biological process or molecular function
- Signaling cascades
- Etc.
- Genes are categorized based on criteria

# The oxidative phosphorylation



# Gene sets (e.g. function driven)

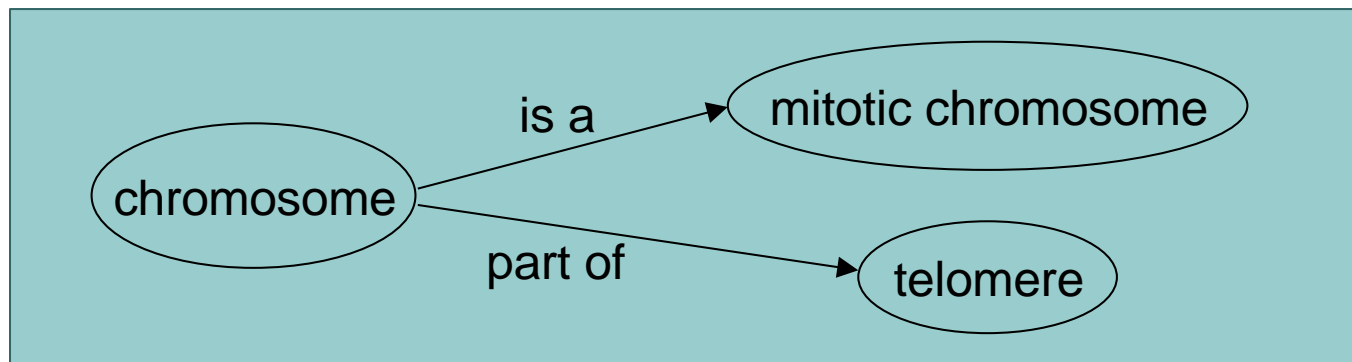
- Genes (sets) have something in common
  - On the same cytogenetic band
  - Coding for proteins that are part of the same cellular component
  - Can be part of the same biochemical pathway
  - Co-expressed under certain conditions
  - Putative targets of the same regulatory factor
  - ....

# Gene Ontology

- **Gene Ontology (GO)** Consortium was established in 1998 to develop shared, structured vocabulary (an ontology) for the annotation of molecular characteristics across different organisms.
  - a collaborative effort to address the need for consistent descriptions of gene and gene products in different databases
  - Original members of the consortium: SGD, FlyBase and MGD
- Two primary purposes for an ontology:
  1. to facilitate communication between people and organizations
  2. to improve upon the interoperability between systems

# GO structure

- The ontologies are structured vocabularies in the form of directed acyclic graphs (DAGs)
- The DAG represents a network (not a tree) in which each term may be a child of one or more than one parent
- The relationships of child to parent can be of the “is a” type or the “part of” type





# Ontologies within GO

- **molecular function** describing activities, such as catalytic or binding activities, at the molecular level
- **biological process** referring to a biological objective to which the gene product contributes
- **cellular component** referring to the place in the cell (i.e. the location) where a gene product is found



# National Institute of Environmental Health Sciences

*Your Environment. Your Health.*

[Term]  
id: GO:0004363  
name: glutathione synthase activity  
namespace: molecular\_function  
def: "Catalysis of the reaction: L-gamma-glutamyl-L-cysteine + ATP + glycine = ADP + glutathione + 2 H(+) + phosphate."  
subset: gosubset\_prok  
synonym: "gamma-L-glutamyl-L-cysteine:glycine ligase (ADP-forming)" EXACT [EC:6.3.2.3]  
synonym: "glutathione synthetase activity" EXACT [EC:6.3.2.3]  
synonym: "GSH synthetase activity" EXACT [EC:6.3.2.3]  
xref: EC:6.3.2.3  
xref: KEGG:R00497  
xref: MetaCyc:GLUTATHIONE-SYN-RXN  
xref: Reactome:REACT\_100394 "gamma-glutamylcysteine combines with glycine to form glutathione, Caenorhabditis elegans"  
xref: Reactome:REACT\_104397 "gamma-glutamylcysteine combines with glycine to form glutathione, Saccharomyces cerevisiae"  
xref: Reactome:REACT\_105605 "gamma-glutamylcysteine combines with glycine to form glutathione, Bos taurus"  
xref: Reactome:REACT\_105891 "gamma-glutamylcysteine combines with glycine to form glutathione, Taeniopygia guttata"  
xref: Reactome:REACT\_108106 "gamma-glutamylcysteine combines with glycine to form glutathione, Sus scrofa"  
xref: Reactome:REACT\_108848 "gamma-glutamylcysteine combines with glycine to form glutathione, Rattus norvegicus"  
xref: Reactome:REACT\_6886 "gamma-glutamylcysteine combines with glycine to form glutathione, Homo sapiens"  
xref: Reactome:REACT\_80709 "gamma-glutamylcysteine combines with glycine to form glutathione, Arabidopsis thaliana"  
xref: Reactome:REACT\_81099 "gamma-glutamylcysteine combines with glycine to form glutathione, Canis familiaris"  
xref: Reactome:REACT\_82957 "gamma-glutamylcysteine combines with glycine to form glutathione, Drosophila melanogaster"  
xref: Reactome:REACT\_84994 "gamma-glutamylcysteine combines with glycine to form glutathione, Mus musculus"  
xref: Reactome:REACT\_86625 "gamma-glutamylcysteine combines with glycine to form glutathione, Schizosaccharomyces pombe"  
xref: Reactome:REACT\_86921 "gamma-glutamylcysteine combines with glycine to form glutathione, Oryza sativa"  
xref: Reactome:REACT\_87951 "gamma-glutamylcysteine combines with glycine to form glutathione, Xenopus tropicalis"  
xref: Reactome:REACT\_90100 "gamma-glutamylcysteine combines with glycine to form glutathione, Dictyostelium discoideum"  
xref: Reactome:REACT\_93797 "gamma-glutamylcysteine combines with glycine to form glutathione, Gallus gallus"  
xref: Reactome:REACT\_99493 "gamma-glutamylcysteine combines with glycine to form glutathione, Danio rerio"  
xref: Reactome:REACT\_99980 "gamma-glutamylcysteine combines with glycine to form glutathione, Plasmodium falciparum"  
xref: RHEA:13560  
is\_a: GO:0016881 ! acid-amino acid ligase activity  
relationship: **part of** GO:0006750 ! glutathione biosynthetic process



# National Institute of Environmental Health Sciences

*Your Environment. Your Health.*

```
[Term]
id: GO:0010778
name: meiotic DNA repair synthesis involved in reciprocal meiotic recombination
namespace: biological_process
def: "The synthesis of DNA proceeding from the broken 3' single-strand DNA end that uses the homologous intact duplex
is_a: GO:0000711 ! meiotic DNA repair synthesis
intersection_of: GO:0000711 ! meiotic DNA repair synthesis
intersection_of: part_of GO:0007131 ! reciprocal meiotic recombination
relationship: part_of GO:0007131 ! reciprocal meiotic recombination

[Term]
id: GO:0010779
name: meiotic DNA repair synthesis involved in meiotic gene conversion
namespace: biological_process
def: "The synthesis of DNA proceeding from the broken 3' single-strand DNA end that uses the homologous intact duplex
is_a: GO:0000711 ! meiotic DNA repair synthesis
intersection_of: GO:0000711 ! meiotic DNA repair synthesis
intersection_of: part_of GO:0006311 ! meiotic gene conversion
relationship: part_of GO:0006311 ! meiotic gene conversion

--
```



# National Institute of Environmental Health Sciences

*Your Environment. Your Health.*

```
[Term]
id: GO:0038142
name: EGFR:ERBB2 complex
namespace: cellular_component
def: "A heterodimeric complex between the tyrosine kinase receptor ERBB2 and a ligand-activated epidermal growth factor"
synonym: "EGF:EGFR:ERBB2 complex" NARROW [Reactome:REACT_116379.1]
synonym: "EGFR:ERBB2 heterodimer" RELATED [Reactome:REACT_116379.1]
xref: Reactome:REACT_116379.1 "ERBB2 heterodimers"
is_a: GO:0043235 ! receptor complex
is_a: GO:0044459 ! plasma membrane part
created_by: rfoulger
creation_date: 2012-03-30T02:10:38Z
```

# KEGG: Kyoto Encyclopedia of Genes and Genomes

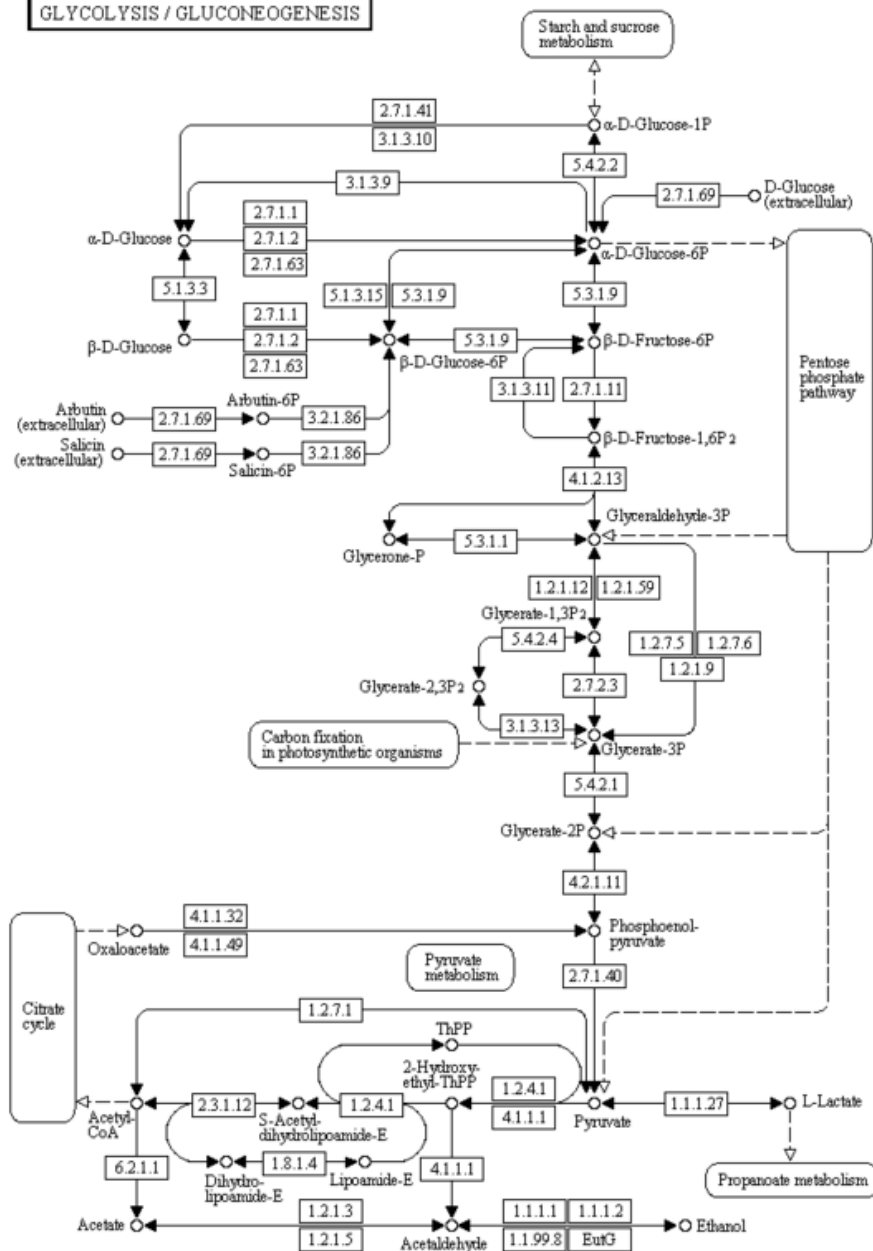
## About KEGG

- **Kyoto Encyclopedia of Genes and Genomes (KEGG)** knowledgebase was developed in 1996 consisting of genetic building blocks of genes and proteins.
- A collection of manually drawn pathway maps representing current knowledge on the **molecular interaction** and **reaction networks**
- Manually curated based on published literature
- Constructed as wiring diagrams with enzymes and proteins, processes and reactions and substrates, co-factors, intermediates, metabolites and end products

## Category in KEGG

- Metabolism: carbohydrates, energy, lipid, nucleotides, amino acid, xenobiotics
- Genetic information processing
- Environmental information processing
- Cellular processes
- Human diseases
- Drug development: the structure relationships

## GLYCOLYSIS / GLUCONEOGENESIS



# Transfac and Transpath databases

## --Signaling Pathways

### Experimentally validated

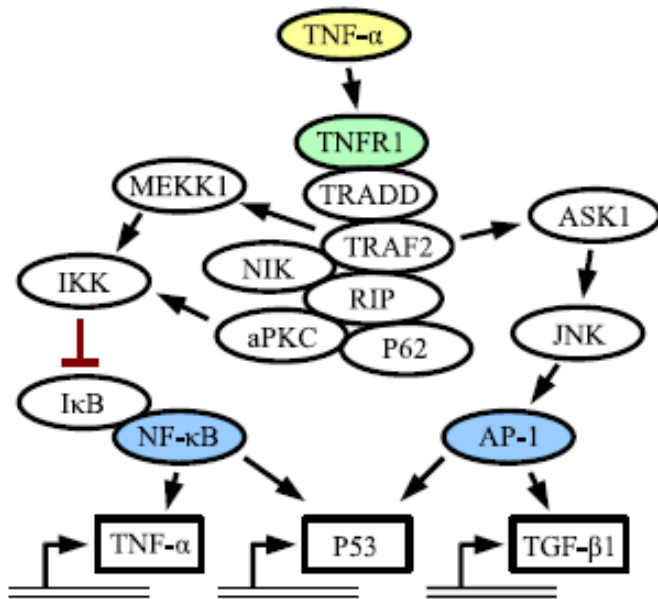


Figure 1

**Example pathway.** A simplified and partial view of the TNF- $\alpha$  pathway. A ligand (yellow) binds to a receptor (green) on the cell surface, triggering a cascade of events. Eventually, transcription factors (blue) activate or repress the expression of genes.

- **Transfac** - data on **transcription factors**, their experimentally-proven **binding sites**, and **regulated genes**.

The compilation of binding sites allows the derivation of positional weight matrices.

- **Transpath** – data about protein-protein interactions and directed modification of proteins involved in signal transduction pathways,

With a particular focus on **signaling cascades** that affect the activity of transcription factors.

# More on gene set collections

- Gene Ontology (GO)
  - Cellular components (CC)
  - Biological processes (BP)
  - Molecular functions (MF)
- Well curated pathway databases
  - KEGG pathway
  - Biocarta
  - GenMAPP
  - IPA pathway database
- Gene set collections
  - MSigDB (GSEA)
  - Gazer
  - Customized collections(GSA)



# A parametric approach

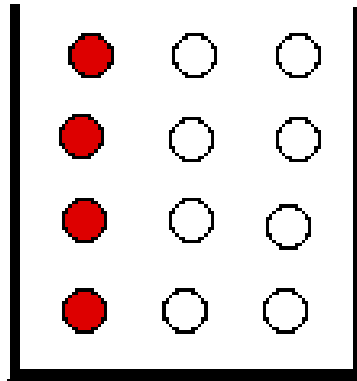
- Hypergeometric distribution and testing
  - A little theoretical background
  - Distribution density
  - Formulating the test
- Fisher's exact test
  - Two by two contingent table
  - One-tailed test formulation
- Watch out for a common error

# Hypergeometric Distribution

$$P(X = k) = \frac{\binom{m}{k} \binom{N-m}{n-k}}{\binom{N}{n}} \quad \text{for } k = 0, 1, 2, \dots, n$$

$k \leq m, n-k \leq N-m$

A discrete probability distribution that describes the number of successes ( $k$ ) in a sequence of  $n$  draws from a finite population without replacement



$$P(k=2, n, m, N) = 0.339$$

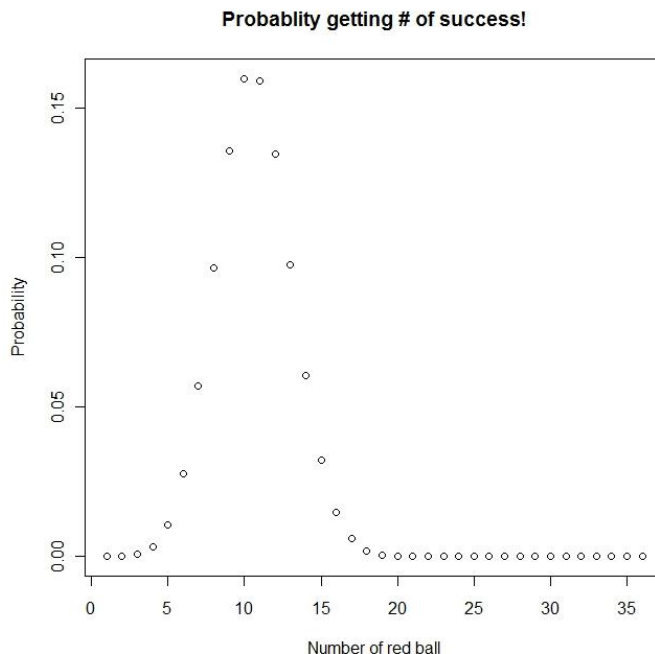
Cumulative dist.:

$$P(k \leq 2) = 0.933$$

- An urn with two types of marbles:
  - **N** total # of marbles
  - Of which, **m** # of **red** marbles
  - Drawing a red marble is a **success!**
  - Drawing a white marble is a **failure!**
- $n$  is the # of marbles randomly drawn
- $k$  is the # of successes (**red marbles**) in the sample
- Hypergeometric distribution gives the probability

# Hypergeometric distribution

- Number of red ball: 50 ←  $m$
- Number of white ball: 120
- Number of ball drawn (without replacement): 36 ←  $n$
- Possible number of success ??
  - (0,1, 2, ....36),
- Probability to get 20 red balls is: 0.0001494571  $p(k=20, 36, 50, 170)$



# Rationale (an analogy)

- A study with gene expression microarray, can be RNAseq also (total # of genes --  $N$ )
- Experiment samples from two or more conditions (control vs. treated, wild type vs knock out, etc.)
- Several biological replicates at each condition
- Differentially Expressed Genes (DEGs) obtained from statistical models (total # of genes drawn --  $n$ )
- Of which,  $k$  genes belong to a “pathway”
- There are totally  $m$  genes (out of  $N$ ) belong to this pathway
- Is this pathway (the biological process) significantly enriched? (Perturbed? Turned on? Involved?)

[illegible]

$$P(x = k) = \frac{\binom{m}{k} \binom{N-m}{n-k}}{\binom{N}{n}}$$

$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{m}{i} \binom{N-m}{n-i}}{\binom{N}{n}}$$

The p-value indicates the probability that the biological process category (particular pathway) is enriched by this microarray experiment **by chance**. The smaller the p-val, the higher the significance

# Pathway analysis

- Gene expression analysis results
  - Total number of gene on a chip: 520
  - Of which 20 genes are in a GO
  - Total number of DEG: 40
  - Of which 5 genes are in GO
- Do you think it is significant at  $\alpha=0.01$ ? In other words “is this GO category enriched by/significant this microarray experiment”?
- The answer is simple, which is to perform a hypergeometric test: *phyper(4, 20, 500, 40, lower.tail=FALSE) = 0.01371*
- Sometime people report: ~~*phyper(5, 20, 500, 40, lower.tail=FALSE) = 0.002459*~~

Watch out!!

# Fisher's Exact Test (FET)

*--right-tailed test*

## Contingency table

	DEG	Not DEGs	totals
In a GO category	x	m-x	m
Not in GO category	k-x	n-k+x	n
totals	k	m+n -k	m + n (genes on array)

So, now you are probably given something like the following:

```
N { x <- 5    #num_of_DEG in GO
    m <- 20  #num_of_gene on chip in GO
    n <- 500 #num_of_gene on chip NOT in GO
    k <- 40  #num_of_DEG
```

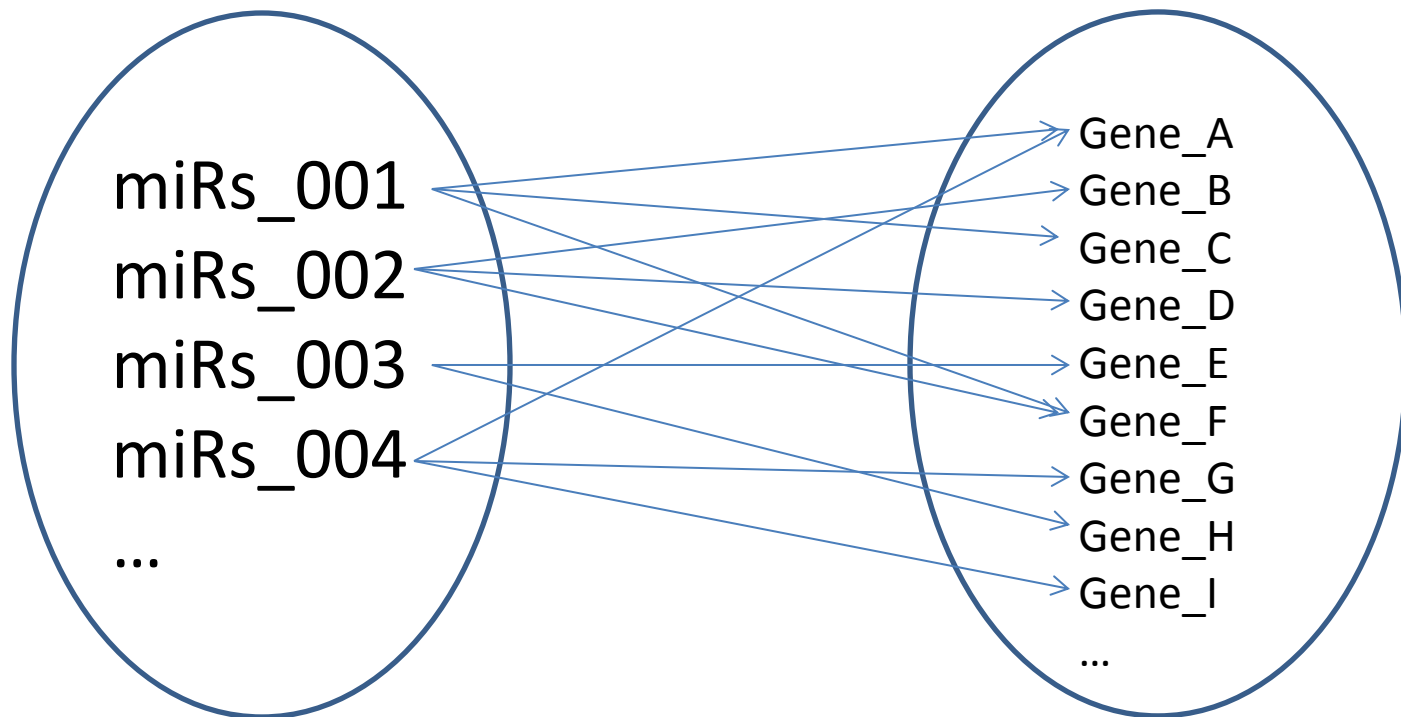
Comparing hypergeometric test vs. FET, watch out!!

***phyper((x-1), m, n, k, lower.tail=FALSE)***

***(fisher.test(matrix(c(x,(k-x), (m-x), (n-k+x)),2,2), alternative='greater'))\$p.value***

# Define your own “gene set”

--miRs vs. their targets





# A parametric approach

- To associate the microRNA's "expression" to their target genes
- Hypergeometric distribution
  - All genes on the affymetrix array (total # of genes --  $N$ )
  - Of which, some are "predicted" as microRNA targets ( $m$ ) <-- a **microRNA target set**
  - DEG lists obtained from 17 contrasts ( $n$ )
  - Of which,  $k$  genes belong to **this microRNA target set**
- In practice, a Fisher's exact test was used

$$P(x = k) = \frac{\binom{m}{k} \binom{N-m}{n-k}}{\binom{N}{n}}$$

Hypergeometric distribution density

$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{m}{i} \binom{N-m}{n-i}}{\binom{N}{n}}$$

Hypergeometric test

# Ingenuity Pathway Analysis (IPA)

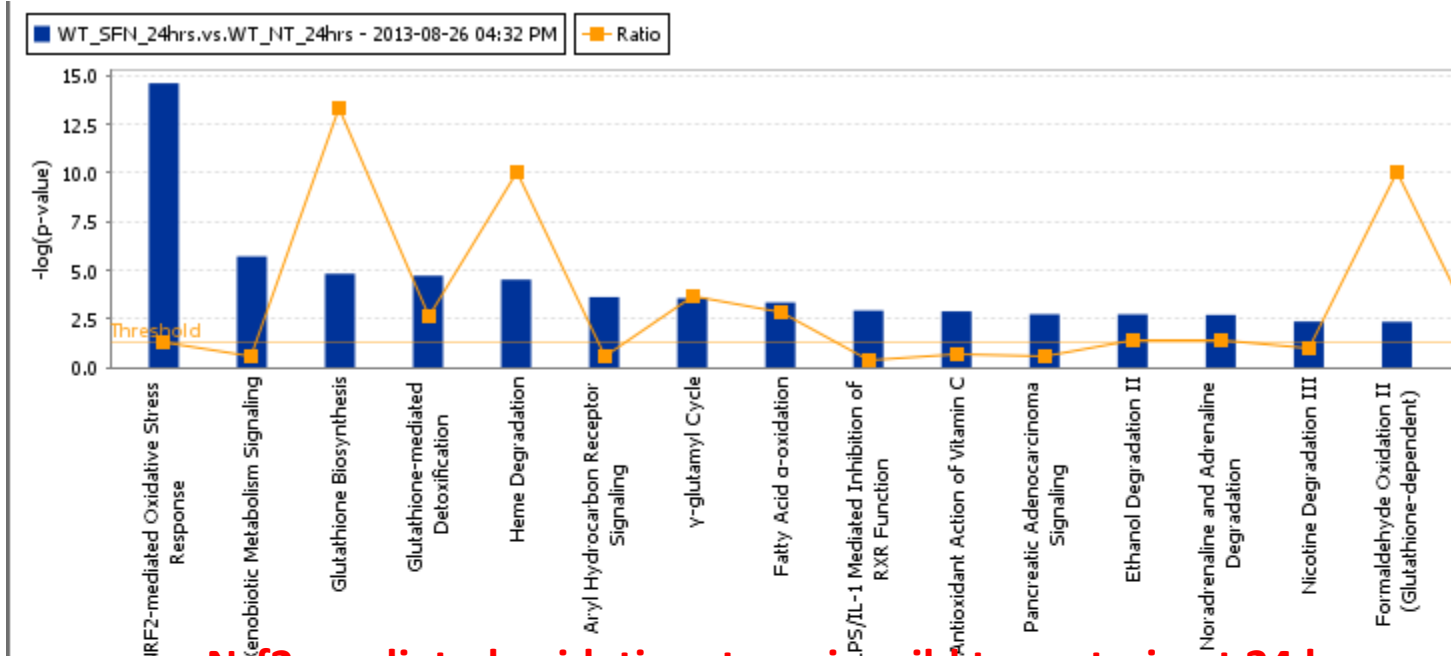
## --Knowledgebase

- Desktop Java application utilizing a remote server for data, analysis and file management
- IPA Ontology: Curation of the scientific literature and content extraction of the IPA repository of molecular interactions, regulatory events, biological processes, gene-to-phenotype associations, and chemical knowledge

Ingenuity® Expert Findings	Experimentally demonstrated Findings that are manually curated for accuracy and contextual details from the full-text of articles in top journals.
Ingenuity® ExpertAssist Findings	Manually reviewed, automatically extracted Findings from the abstracts of a broad range of recently published journal articles.
Ingenuity® Expert Knowledge	Knowledge modeled by Ingenuity experts such as pathways, toxicity lists, and more.
Ingenuity® Supported Third Party Information	Manually reviewed content from selected sources and databases such as BIND, Argonaute 2, etc.

# IPA Enrichment Analysis

- Uses the Fisher's exact test to determine the significance of a functional group or pathway
  - # molecules in a list that are associated with a function/pathway (**k**)
  - total # of molecules that are associated with a function/pathway (**m**)
  - # of molecules in all possible functions/pathways (**N**)
  - # of molecules in a list (**n**)

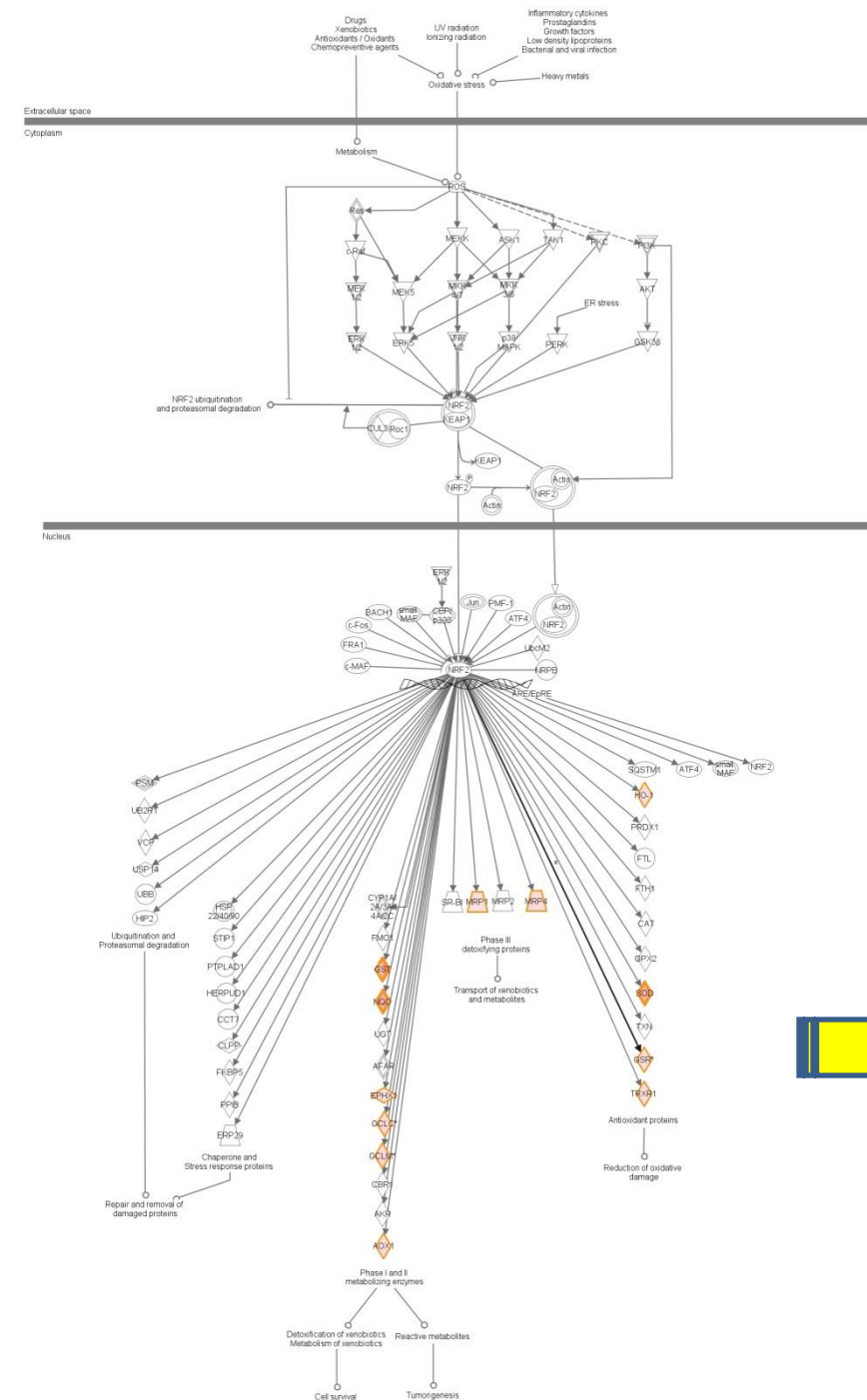


**Nrf2-mediated oxidative stress in wild type strain at 24 hours**

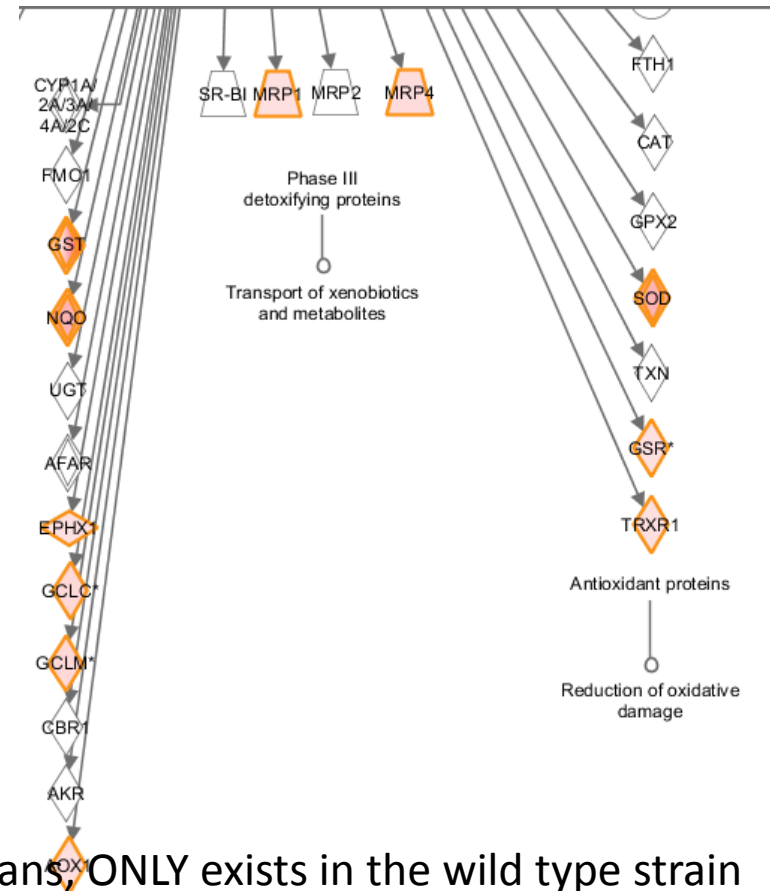
12 molecules associated with **NRF2-mediated Oxidative Stress Response** at WT\_SF\_N\_24hrs.vs.WT\_NT\_24hrs - 2013-08-26 04:32 PM

<a href="#">ADD TO MY PATHWAY</a> <a href="#">ADD TO MY LIST</a> <a href="#">CREATE DATASET</a> <a href="#">CUSTOMIZE TABLE</a>						
<input type="checkbox"/>	Symbol	Entrez Gene Name	Identifier	Exp Val		
			Affymetrix	p-value	False Discovery Rate	Fold Change
<input type="checkbox"/>	ABCC1	ATP-binding cassette, sub-	1421378_s_at	8.68E-07	6.63E-05	↑2.054
<input type="checkbox"/>	ABCC4	ATP-binding cassette, sub-	1443870_at	3.60E-15	2.68E-12	↑2.962
<input type="checkbox"/>	AOX1	aldehyde oxidase 1	1419435_at	2.15E-11	6.91E-09	↑2.756
<input type="checkbox"/>	EPHX1	epoxide hydrolase 1,	1422438_at	7.19E-17	6.69E-14	↑2.347
<input type="checkbox"/>	GCLC*	glutamate-cysteine ligase,	1424296_at*	1.80E-17	2.24E-14	↑3.268
<input type="checkbox"/>	GCLM*	glutamate-cysteine ligase,	1418627_at*	1.24E-21	4.63E-18	↑3.343
<input type="checkbox"/>	GSR*	glutathione reductase	1421816_at*	8.32E-12	3.04E-09	↑2.604
<input type="checkbox"/>	GSTA5	glutathione S-transferase	1421040_a_at	1.58E-24	1.47E-20	↑17.533
<input type="checkbox"/>	GSTM5*	glutathione S-transferase mu	1416416_x_at*	4.93E-17	4.83E-14	↑2.879
<input type="checkbox"/>	HMOX1	heme oxygenase (decycling)	1448239_at	2.28E-08	3.05E-06	↑2.507
<input type="checkbox"/>	NQO1	NAD(P)H dehydrogenase,	1423627_at	1.49E-26	2.78E-22	↑6.419
<input type="checkbox"/>	TXNRD1	thioredoxin reductase 1	1421529_a_at	2.11E-19	3.57E-16	↑2.240

All the DEGs in the pathway are up-regulated



WT\_SF\_N\_24hr vs. WT\_NT\_24 hr






By all means, ONLY exists in the wild type strain

# Three upstream regulators

WT\_SFNT24hrs.vs.WT\_NT24hrs - 2013-08-26 04:32 PM

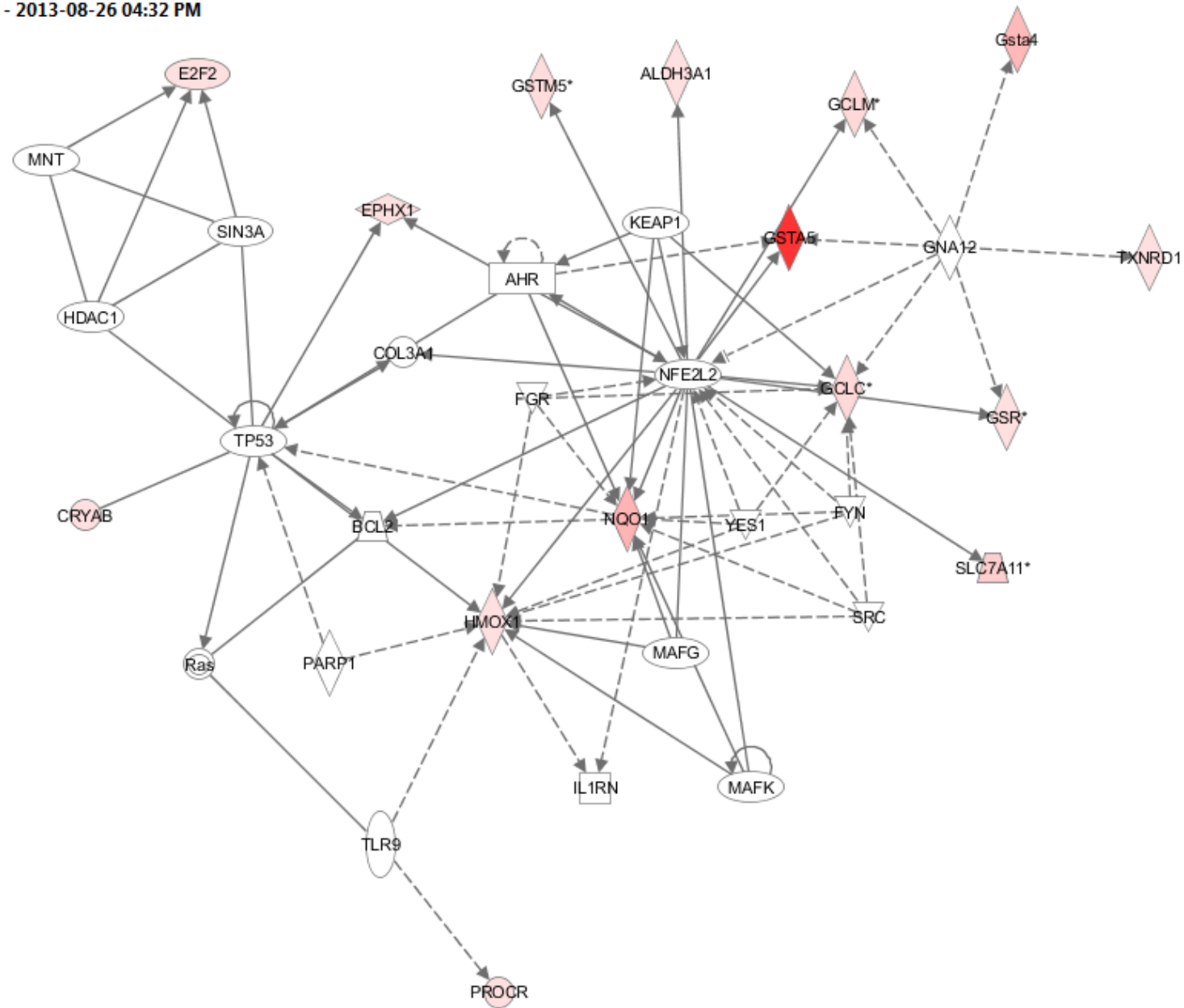
Summary \ Functions \ Canonical Pathways \ **Upstream Analysis** \ Networks \ Molecules \ Lists \ My Pathways

ADD TO MY PATHWAY ADD TO MY LIST CUSTOMIZE TABLE DISPLAY AS NETWORK MECHANISTIC NETWORKS    p-value of over... 8.53E-19 - 2.98E-02 (p1 of 2) ▼

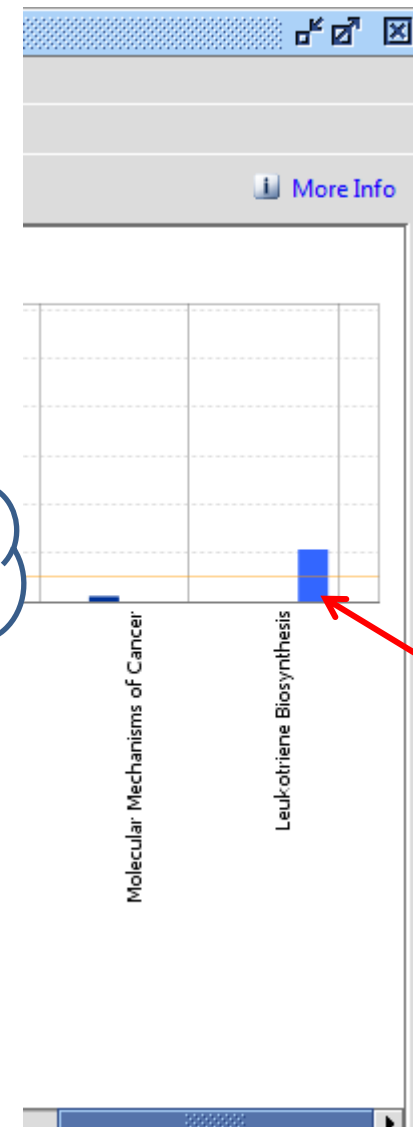
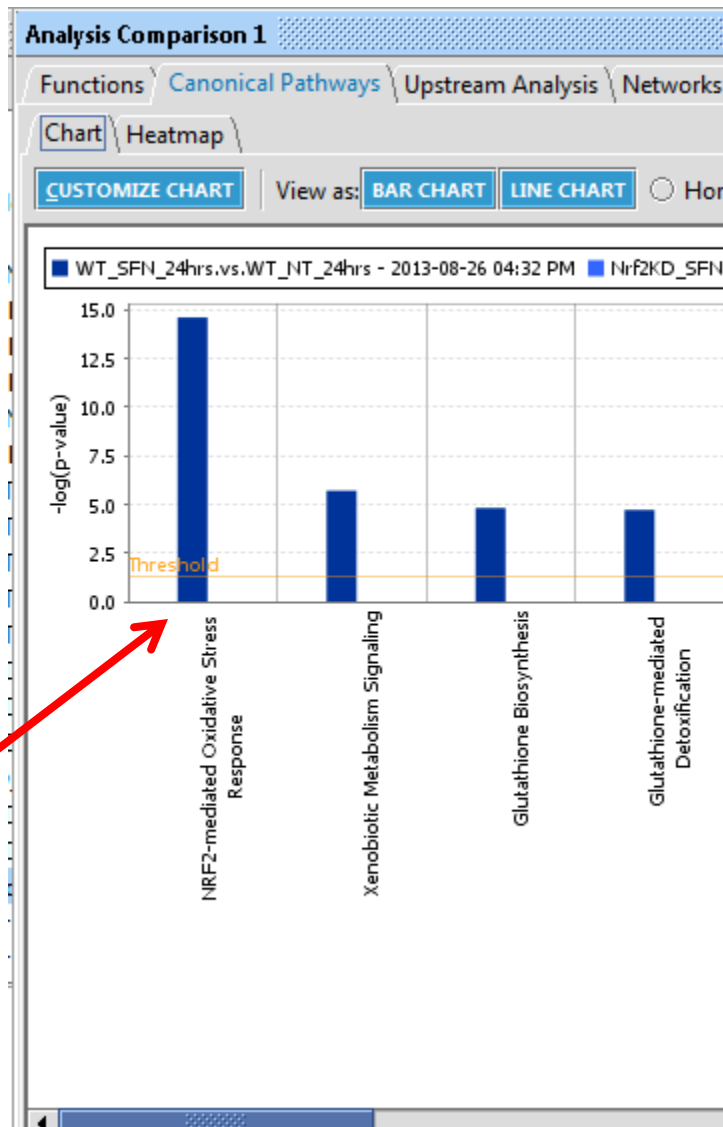
<input type="checkbox"/>	Upstream Regulator	Fold Change	Molecule Type ▼	Predicted Activation St...	Activation z-score	Notes	▲ p-value of overlap	Target molecules in da...
<input type="checkbox"/>	NFE2L2		transcription regulator	Activated	2.273	bias	8.53E-19	↑ALDH3A1, ↑E... ...all 10
<input type="checkbox"/>	GNA12		enzyme	Inhibited	-2.433	bias	5.62E-15	↑GCLC, ↑GCLM, . ...all 6
<input type="checkbox"/>	FGR		kinase				9.30E-09	↑GCLC, ↑HMOX1 ...all 3
<input type="checkbox"/>	YES1		kinase				9.28E-08	↑GCLC, ↑HMOX1 ...all 3
<input type="checkbox"/>	SRC		kinase				5.17E-07	↑GCLC, ↑HMOX1 ...all 3
<input type="checkbox"/>	FYN		kinase				2.02E-06	↑GCLC, ↑HMOX1 ...all 3
<input type="checkbox"/>	TLR3		transmembrane receptor	Activated	2.200	bias	3.72E-06	↑CXCL3, ↑HMOX1 ...all 5

# Cell death and survival, cancer, hematological disease

Overlay: WT\_SF2N\_24hrs.vs.WT\_NT\_24hrs - 2013-08-26 04:32 PM



# Comparison with Nrf2 KD – 24 hrs





# Parametric approach summary

- Statistics distribution based test (hyper geometric/FET)
- And, there are a few violations (will be addressed shortly)
- Knowledgebase used matters
- Hands on practice on IPA

# A non-parametric approach

- Based on Kolmogorov-Smirnov (K-S) test
- Compare a sample empirical distribution function to the reference distribution
- Or, compare two sample empirical distribution functions
- The test is performed in relevant to different null hypothesis setting

# Gene set enrichment analysis (GSEA)

- Given an *a priori* defined set of genes (i.e. genes encoding products in a metabolic pathway, or sharing the same GO category)
- The goal is to determine whether the members in a set  $S$  are randomly distributed throughout all sorted genes ( $L$ )
- Mootha *et al.* [2003] suggest to use the Kolmogorov-Smirnov statistic:
  - First, sort all genes by some criteria
  - Go through the list, increasing a running sum for each gene in the gene set by  $(N-n)$ 
    - $X_i = \sqrt{\frac{N-G}{G}}$ , if  $R_i$  is a member of  $S$
  - And decreasing it for each gene not in the gene set by  $n$ . [ $N$ : number of genes,  $n$ : size of gene set]
    - $X_i = -\sqrt{\frac{G}{N-G}}$ , if  $R_i$  is not a member of  $S$
- The maximum value of the running sum is the enrichment score (ES)
  - $ES = \max_{1 \leq j \leq N} \sum_{i=1}^j X_i$
- Statistical significance determined by permutation GSEA-P

Just FYI

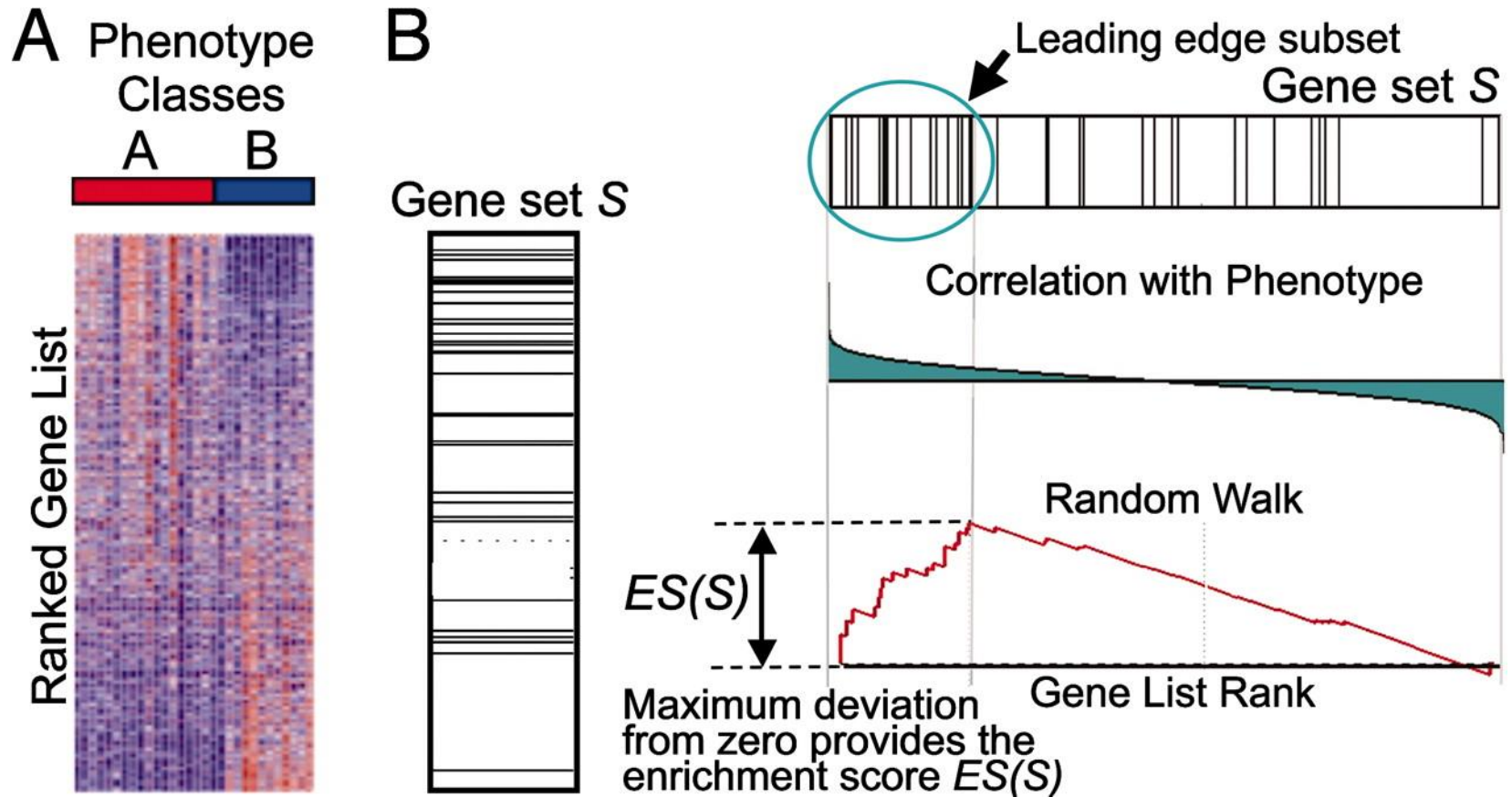


# Improved KS statistics tests

- Subramanian *et al.*, PNAS **102** (2005) proposed an improvement strategy
- $P_{hit}(S, i) = \sum_{j \leq i} g_{j \in S} \frac{|r_j|^p}{N_R}$ , where  $N_R = \sum_{j \in S} |r_j|^p$
- $P_{miss}(S, i) = \sum_{j \leq i} g_{j \notin S} \frac{1}{(N - N_H)}$
- The ES is the maximum deviation from zero of  $p_{hit}$ - $p_{miss}$
- If  $p = 0$ , ES reduces to Kolmogorov-Smirnov statistics, in the paper, author used  $p = 1$

Another **FYI** page

## A GSEA overview illustrating the method.



Subramanian A et al. PNAS 2005;102:15545-15550



# GSEA web-portal

<http://www.broadinstitute.org/gsea/msigdb/index.jsp>

**c1** **positional gene sets** for each human chromosome and cytogenetic band.

**c2** **curated gene sets** from online pathway databases, publications in PubMed, and knowledge of domain experts.

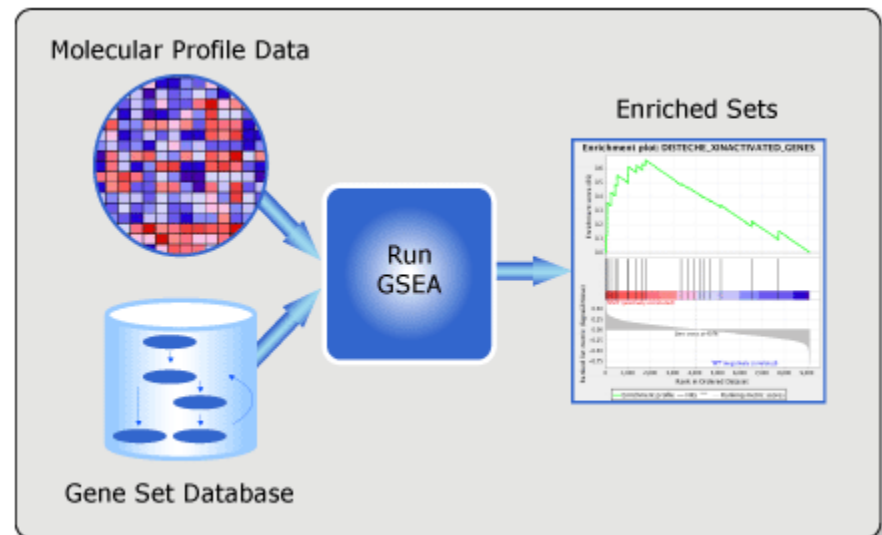
**c3** **motif gene sets** based on conserved cis-regulatory motifs from a comparative analysis of the human, mouse, rat, and dog genomes.

**c4** **computational gene sets** defined by mining large collections of cancer-oriented microarray data.

**c5** **GO gene sets** consist of genes annotated by the same GO terms.

**c6** **oncogenic signatures** defined directly from microarray gene expression data from cancer gene perturbations.

**c7** **immunologic signatures** defined directly from microarray gene expression data from immunologic studies.



Hands on practice on both parametric and non-parametric

**NOW, IT IS YOUR TURN**



# Ingenuity Pathway Analysis

Discover the Biology

INGENUITY<sup>®</sup>  
SYSTEMS

Welcome! Please login

Email

li11@niehs.nih.gov

Password

••••••••

☐ Remember my password

LOG IN

[Sign Up](#) | [Forgot Password](#)

Contact Customer Support

Live Chat:

Need Help?

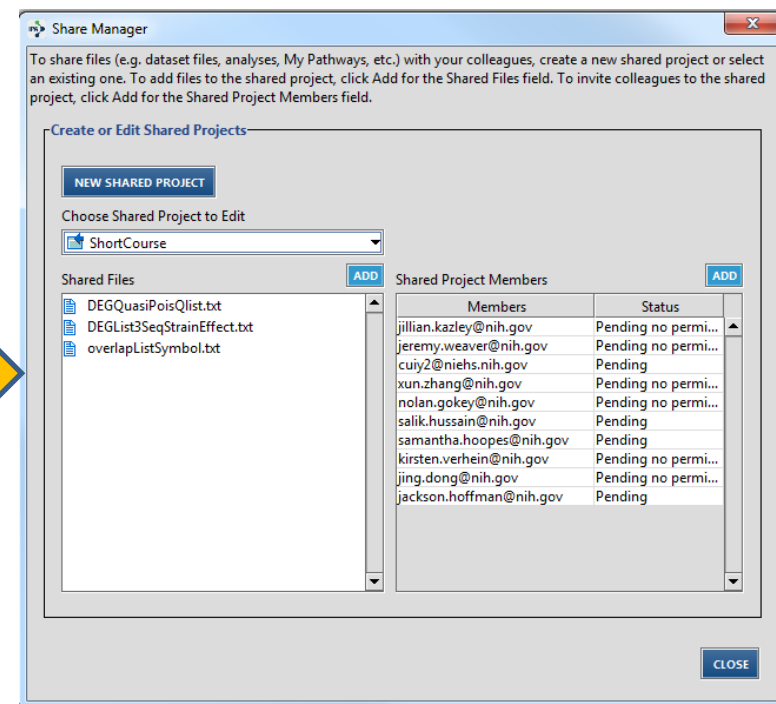
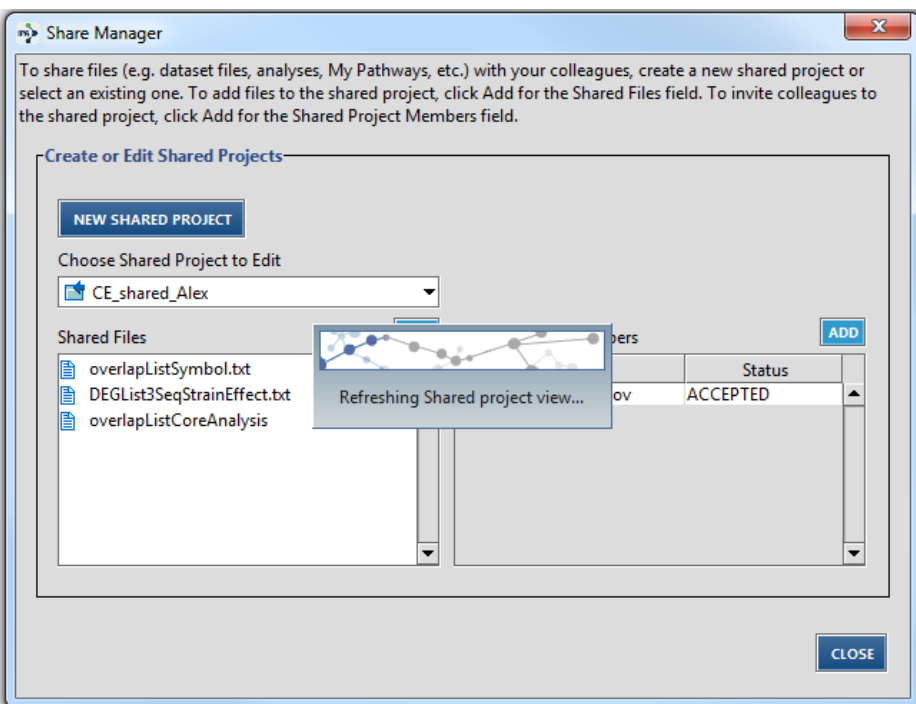
Customer Support

Phone: 650.381.5111  
Hours: 6am - 5pm (PST)  
Monday - Friday (excluding holidays)  
[support@ingenuity.com](mailto:support@ingenuity.com)

For Product and Sales related inquiries contact:

650.381.5056  
[sales@ingenuity.com](mailto:sales@ingenuity.com)

# Sharing a project



# Bio function associated -- IPA

Top Bio Functions		
<b>Diseases and Disorders</b>		
Name	p-value	# Molecules
Inflammatory Response	4.52E-12 - 6.18E-04	106
Infectious Disease	2.80E-07 - 5.75E-04	41
Connective Tissue Disorders	3.01E-07 - 5.03E-05	55
Inflammatory Disease	3.01E-07 - 4.16E-04	72
Skeletal and Muscular Disorders	3.01E-07 - 4.16E-04	56
<b>Molecular and Cellular Functions</b>		
Name	p-value	# Molecules
Cell Morphology	2.16E-13 - 7.75E-04	77
Cell Death and Survival	4.69E-12 - 9.35E-04	135
Lipid Metabolism	1.04E-10 - 9.56E-04	89
Small Molecule Biochemistry	1.04E-10 - 9.56E-04	102
Cell-To-Cell Signaling and Interaction	3.11E-10 - 8.78E-04	93
<b>Physiological System Development and Function</b>		
Name	p-value	# Molecules
Hematological System Development and Function	1.81E-12 - 9.52E-04	98
Tissue Morphology	1.81E-12 - 8.39E-04	91
Tissue Development	2.38E-12 - 8.87E-04	69
Immune Cell Trafficking	4.52E-12 - 9.52E-04	65
Humoral Immune Response	2.18E-07 - 5.16E-04	38

# Other pathways (IPA – cont.)

## Top Canonical Pathways

Name	p-value	Ratio
Nicotine Degradation II	3.17E-14	16/83 (0.193)
Nicotine Degradation III	1.53E-11	13/71 (0.183)
LPS/IL-1 Mediated Inhibition of RXR Function	8.06E-10	22/239 (0.092)
Bupropion Degradation	1.28E-09	9/33 (0.273)
Acetone Degradation I (to Methylglyoxal)	1.88E-09	9/36 (0.25)

## Top Upstream Regulators

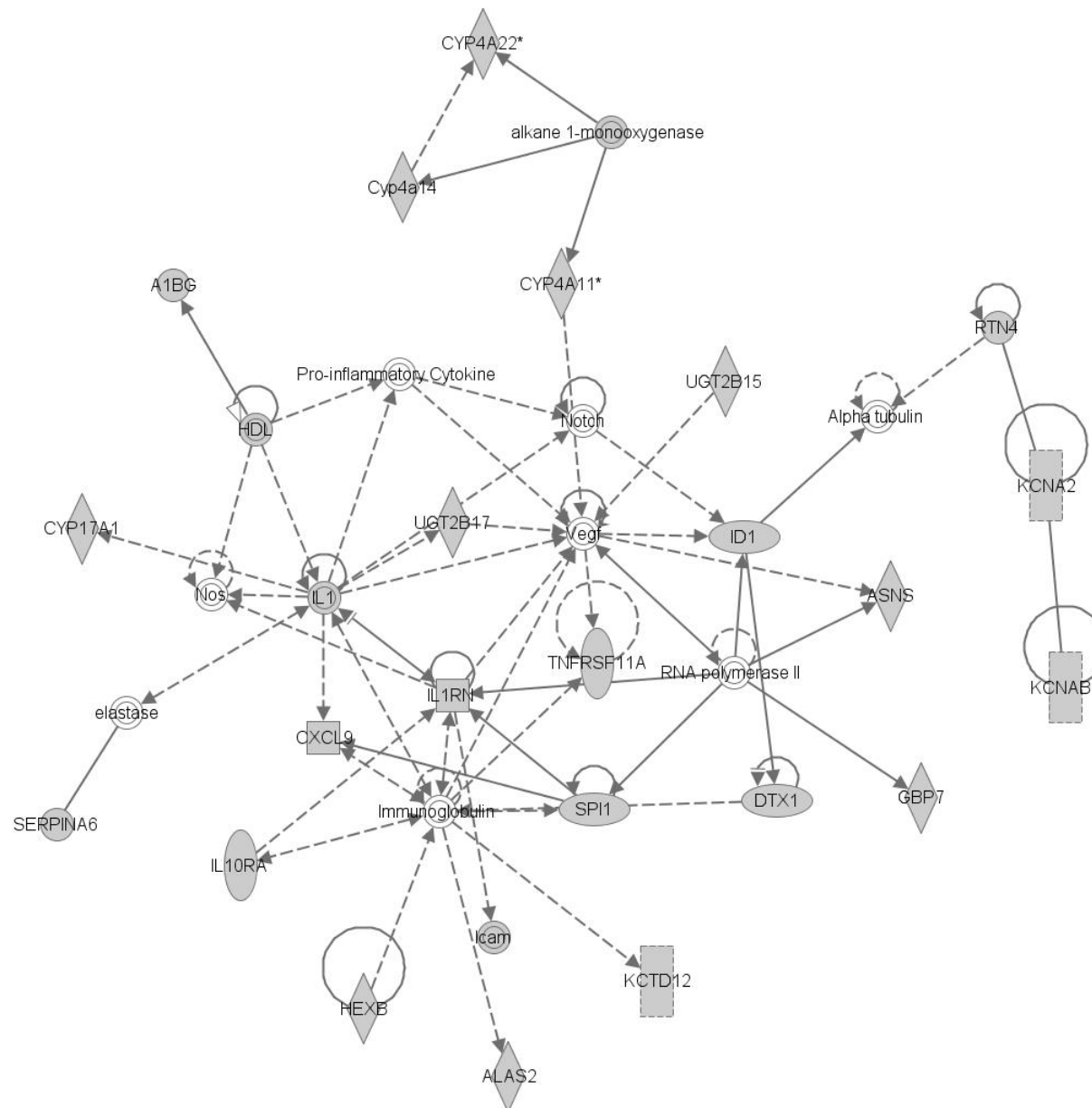
Upstream Regulator	p-value of overlap	Predicted Activat.
RORC	3.13E-31	
RORA	2.09E-29	
GPD1	7.95E-23	
SLC25A13	1.45E-22	
NR1B	5.71E-17	

## Top Tox Lists


Name	p-value	Ratio
LPS/IL-1 Mediated Inhibition of RXR Function	1.22E-12	27/247 (0.109)
Cytochrome P450 Panel - Substrate is a Xenobiotic (Mouse)	8.49E-10	9/25 (0.36)
NRF2-mediated Oxidative Stress Response	2.4E-09	22/232 (0.095)
CAR/RXR Activation	3.87E-09	9/29 (0.31)
Cytochrome P450 Panel - Substrate is a Xenobiotic (Rat)	2.22E-08	8/25 (0.32)

# Lipid metabolism, endocrine system development and function

Network 2 : overlapListCoreAnalysis : overlapListSymbol.txt : overlapListCoreAnalysis



Click to download

**Gene Set Enrichment Analysis**

[GSEA Home](#) [Downloads](#) [Molecular Signatures Database](#) [Documentation](#) [Contact](#)

### Overview

**Gene Set Enrichment Analysis (GSEA)** is a computational method that determines whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states (e.g. phenotypes).

From this web site, you can:

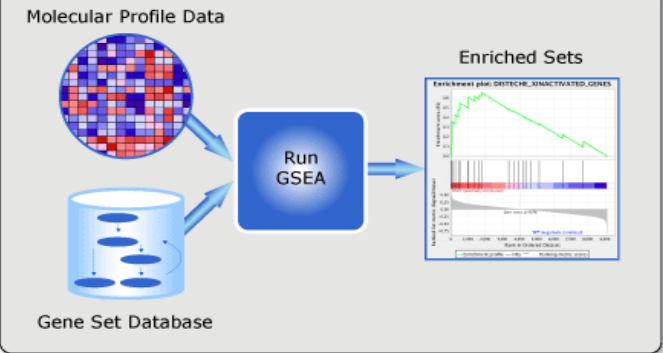
- ▶ **Download** the GSEA software and additional resources to analyze, annotate and interpret enrichment results.
- ▶ **Explore the Molecular Signatures Database (MSigDB)**, a collection of annotated gene sets for use with GSEA software.
- ▶ **View documentation** describing GSEA and MSigDB.

### What's New

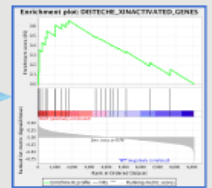
05-Jun-2013: Version 4.0 of the Molecular Signatures Database (MSigDB) is now available, which includes a new gene set collection (C7) of 1,910 immunologic signatures generated as part of the Human Immunology Project Consortium. We also released a newer version (2.0.13) of the GSEA desktop application. There were no changes to the GSEA algorithm.

29-May-2013: GSEA and MSigDB may experience intermittent connectivity issues on Monday, June 3rd between the hours of 6AM and 9AM (Eastern Time).

**Molecular Profile Data**



**Enriched Sets**



**Registration**

Please [register](#) to download the GSEA software and view the MSigDB gene sets. After registering, you can log in at any time using your email address. Registration is free. Its only purpose is to help us track usage for reports to our funding agencies.

### Contributors



Gene Set Enrichment Analysis

[GSEA Home](#)

[Downloads](#)

[Molecular Signatures Database](#)

[Documentation](#)

[Contact](#)

## Login to GSEA/MSigDB

### Login

[Click here](#) to register to view the MSigDB gene sets and/or download the GSEA software. This helps us track and better serve our user community.

If you have already registered for GSEA or MSigDB please enter your registration email address below.

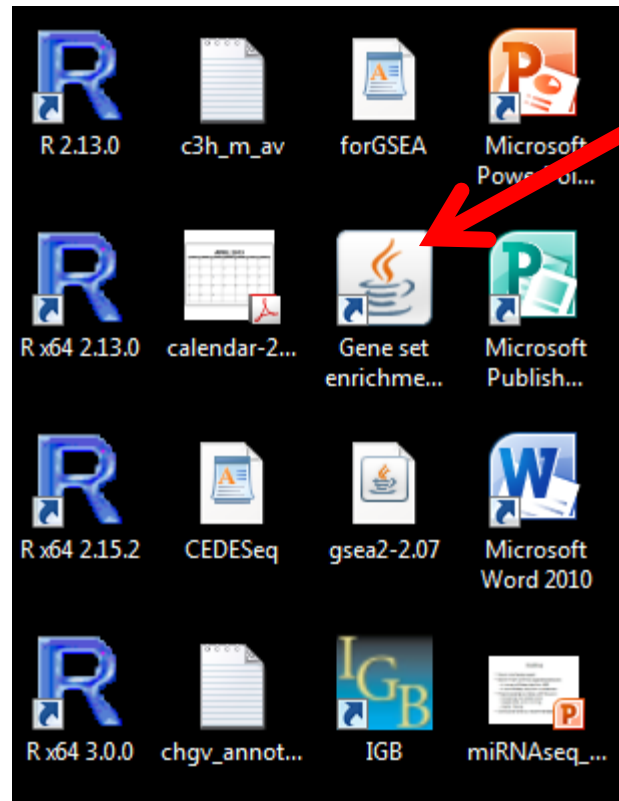
Items marked with \* are required.

Email: \*



Provide your email

login



This is the icon you need to click to launch



Step 1

Step 2

Step 3

GSEA v2.0.12 (Gene set enrichment analysis -- Broad Institute)

File Options Downloads Tools Help

Steps in GSEA analysis

Load data

Run GSEA

Leading edge analysis

Gene set tools

Chip2Chip mapping

Browse MSigDB

Analysis history

GSEA reports


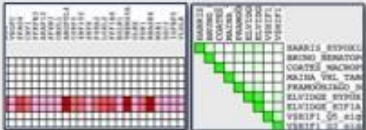
Processes: click 'status' field for results

Name	Status

Show results folder

Home Load data x


### Steps in GSEA

- 1. What you need for GSEA**
  - Expression data set
  - Phenotype annotation
  - Gene sets – use MSigDB or your own gene sets
- 2. Run GSEA**
  - Start with default parameters
  - If you want to collapse probes to genes, specify chip platform
- 3. View results**
- 4. Leading edge analysis**
  - Leading edge finds genes driving enrichment results

### Gene Set Tools


**Chip2Chip mapping**

- Convert gene sets between platforms

 Chip2Chip mapping

**Explore MSigDB gene sets**

- Search the database of thousands of gene sets
- Browse the gene sets by name
- Find overlapping gene sets
- Export gene sets

 Browse MSigDB

**See also**

- MSigDB online tools at: [www.broadinstitute.org/msigdb](http://www.broadinstitute.org/msigdb)

### Getting Help

**GSEA web site:**


[www.broadinstitute.org/gsea](http://www.broadinstitute.org/gsea)

**GSEA documentation:**

[www.broadinstitute.org/gsea/wiki](http://www.broadinstitute.org/gsea/wiki)

**Email the GSEA team:**

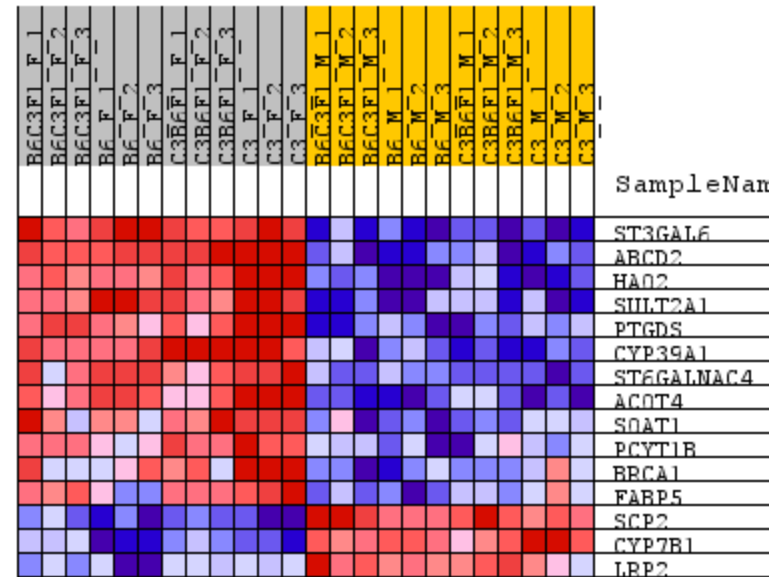
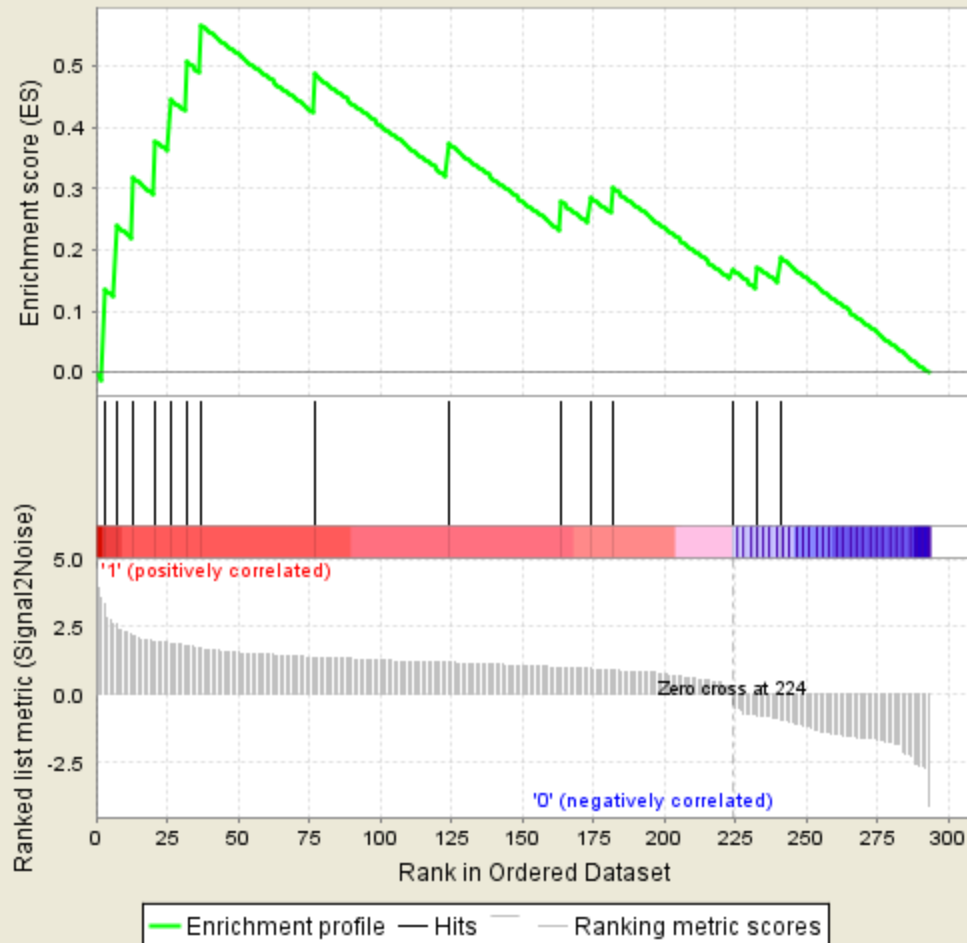
[gsea@broadinstitute.org](mailto:gsea@broadinstitute.org)



4:03:18 PM 6510 [INFO] Made Vdb dir: C:\Users\li11\gsea\_home\output\jul11 14M of 61M

# Lipid metabolism enriched -- GSEA

Enrichment plot: LIPID\_METABOLIC\_PROCESS



# Genes involved (GSEA – cont.)

PROBE	GENE SYMBOL	GENE_TITLE	RANK IN GENE LIST	RANK METRIC SCORE	RUNNING ES	CORE ENRICHMENT
<a href="#">ST3GAL6</a>	<a href="#">ST3GAL6</a> <a href="#">Entrez</a> , <a href="#">Source</a>	ST3 beta-galactoside alpha-2,3-sialyltransferase 6	3	3.348	0.1357	Yes
<a href="#">ABCD2</a>	<a href="#">ABCD2</a> <a href="#">Entrez</a> , <a href="#">Source</a>	ATP-binding cassette, sub-family D (ALD), member 2	7	2.603	0.2389	Yes
<a href="#">HAO2</a>	<a href="#">HAO2</a> <a href="#">Entrez</a> , <a href="#">Source</a>	hydroxyacid oxidase 2 (long chain)	13	2.192	0.3169	Yes
<a href="#">SULT2A1</a>	<a href="#">SULT2A1</a> <a href="#">Entrez</a> , <a href="#">Source</a>	sulfotransferase family, cytosolic, 2A, dehydroepiandrosterone (DHEA)-preferring, member 1	21	1.965	0.3777	Yes
<a href="#">PTGDS</a>	<a href="#">PTGDS</a> <a href="#">Entrez</a> , <a href="#">Source</a>	prostaglandin D2 synthase 21kDa (brain)	26	1.886	0.4459	Yes
<a href="#">CYP39A1</a>	<a href="#">CYP39A1</a> <a href="#">Entrez</a> , <a href="#">Source</a>	cytochrome P450, family 39, subfamily A, polypeptide 1	32	1.783	0.5060	Yes
<a href="#">ST6GALNAC4</a>	<a href="#">ST6GALNAC4</a> <a href="#">Entrez</a> , <a href="#">Source</a>	ST6 (alpha-N-acetyl-neuraminy1-2,3-beta-galactosyl-1,3)-N-acetylgalactosaminide alpha-2,6-sialyltransferase 4	37	1.693	0.5658	Yes
<a href="#">ACOT4</a>	<a href="#">ACOT4</a> <a href="#">Entrez</a> , <a href="#">Source</a>	acyl-CoA thioesterase 4	77	1.374	0.4861	No
<a href="#">SOAT1</a>	<a href="#">SOAT1</a> <a href="#">Entrez</a> , <a href="#">Source</a>	sterol O-acyltransferase (acyl-Coenzyme A: cholesterol acyltransferase) 1	124	1.179	0.3728	No
<a href="#">PCYT1B</a>	<a href="#">PCYT1B</a> <a href="#">Entrez</a> , <a href="#">Source</a>	phosphate cytidyltransferase 1, choline, beta	164	1.006	0.2770	No
<a href="#">BRCA1</a>	<a href="#">BRCA1</a> <a href="#">Entrez</a> , <a href="#">Source</a>	breast cancer 1, early onset	174	0.946	0.2862	No
<a href="#">FABP5</a>	<a href="#">FABP5</a> <a href="#">Entrez</a> , <a href="#">Source</a>	fatty acid binding protein 5 (psoriasis-associated)	182	0.899	0.3004	No
<a href="#">SCP2</a>	<a href="#">SCP2</a> <a href="#">Entrez</a> , <a href="#">Source</a>	sterol carrier protein 2	224	-0.319	0.1674	No
<a href="#">CYP7B1</a>	<a href="#">CYP7B1</a> <a href="#">Entrez</a> , <a href="#">Source</a>	cytochrome P450, family 7, subfamily B, polypeptide 1	233	-0.761	0.1720	No
<a href="#">LRP2</a>	<a href="#">LRP2</a> <a href="#">Entrez</a> , <a href="#">Source</a>	low density lipoprotein-related protein 2	241	-0.901	0.1864	No

# Some caveat in pathway analysis

- Concerns with GO based pathway analysis
  - Ignore GO hierarchy – treat each term independently
  - Ignore gene (expression) level/rank
  - Ignore correlation – assuming genes (in a category) are uncorrelated
  - Sampling over genes (observation)
  - All these make  $p$  values quite unclear
- Facts on the non-parametric approach
  - *Pros*
    - Relax on “strong assumption”
    - Customized gene set is allowed
  - *Cons*
    - Computationally intensive
    - Some concerns on the statistical power

# The universe matters

- It is important to choose the universe correctly

Case 1: universe is all genes in the genome

	DEGs	Non-DEGs	Total
In GO	10	30	40
Not in GO	390	3570	3960
Total	400	3600	4000

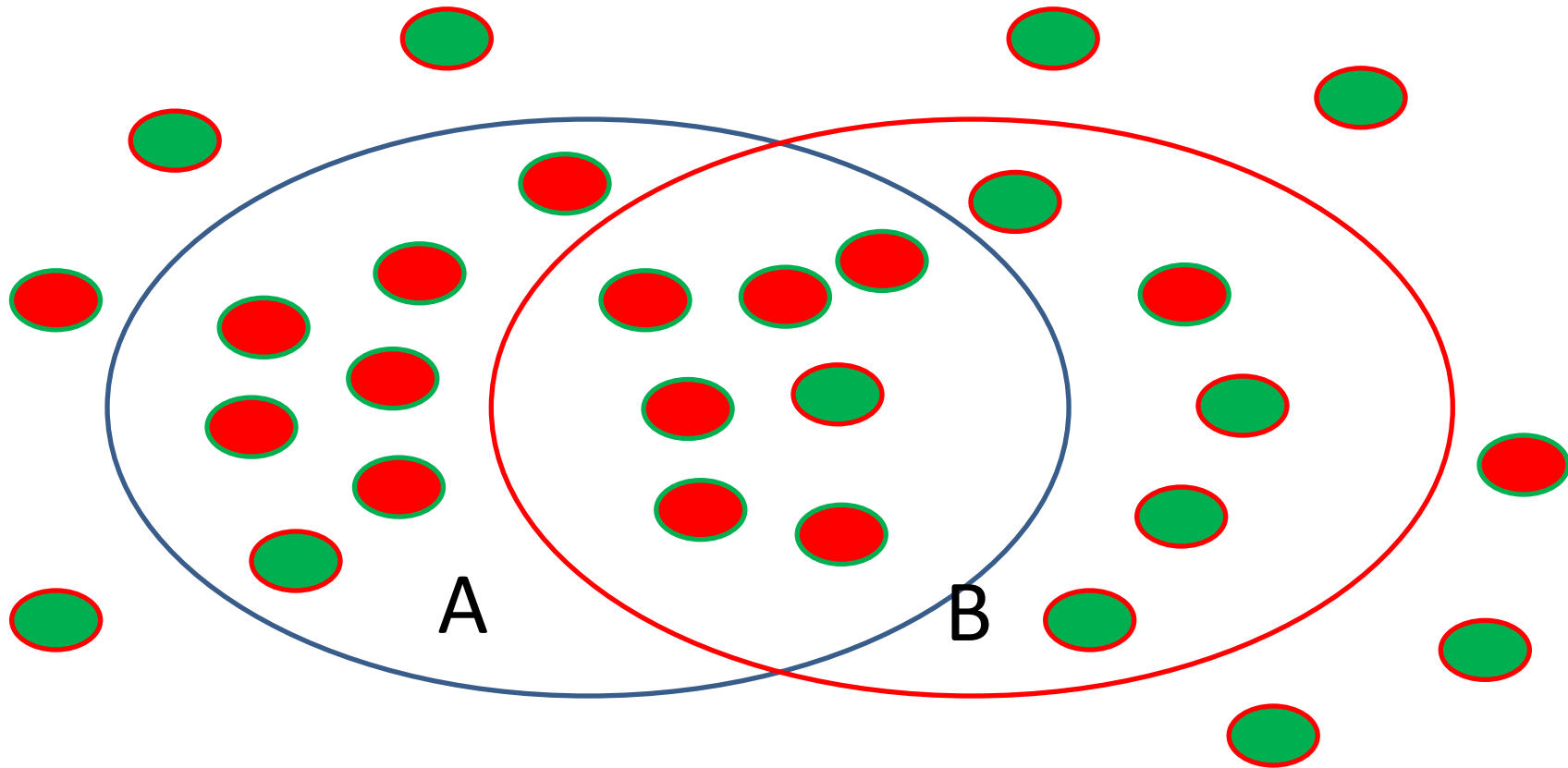
$p=0.049$

Case 2: universe is only expressed genes

	DEGs	Non-DEGs	Total
In GO	10	30	40
Not in GO	390	570	960
Total	400	600	1000

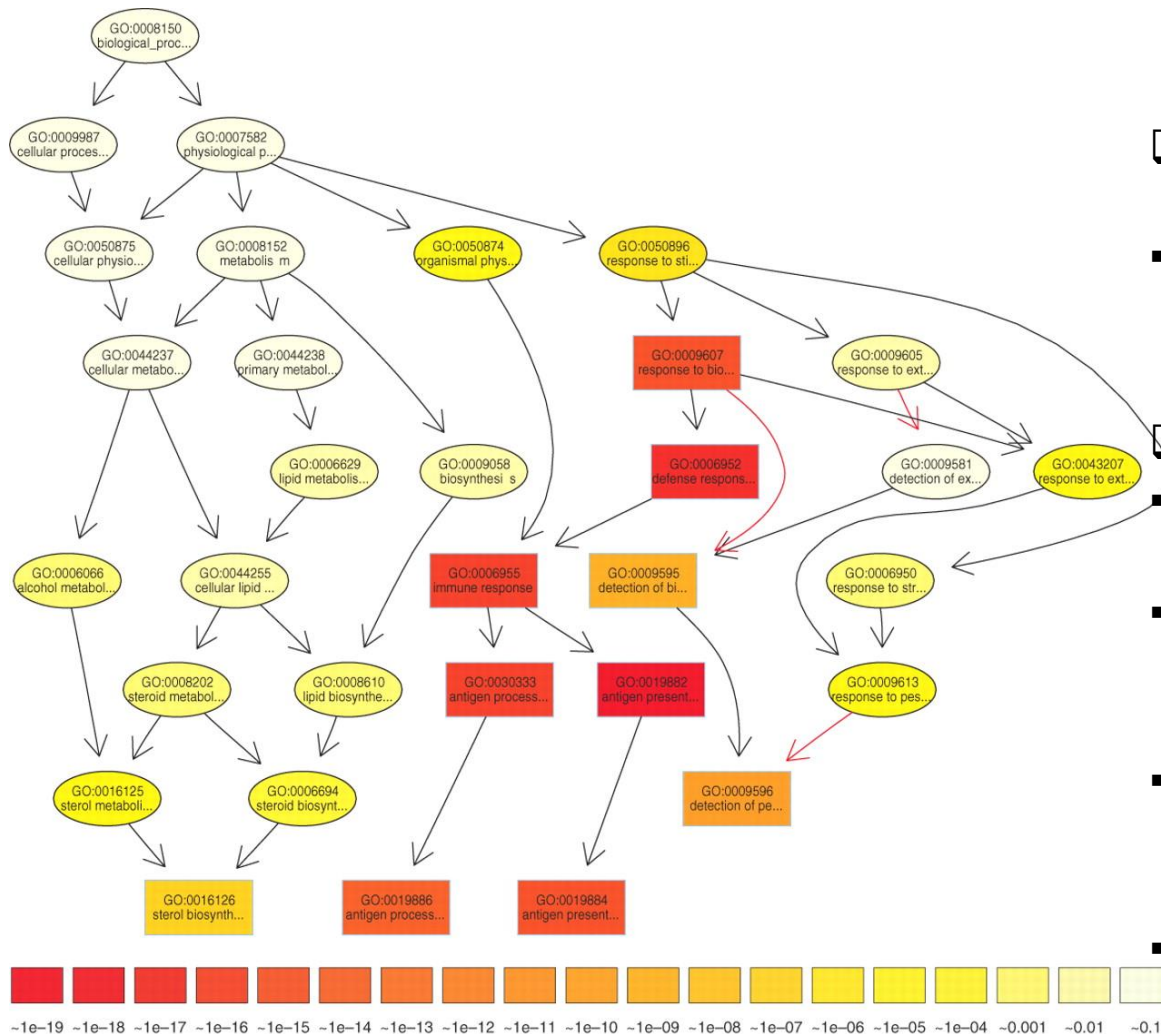
$p=0.0048$

# Sets are overlapped



Set B is enriched only because of its overlap with set A

**The subgraph induced by the 10 most significant GO terms identified by a current state-of-the-art method for scoring GO terms for enrichment.**



## ❑ TopGO's elimination algorithm

- Test the leaf sets first. If significant, remove its “genes” before testing its ancestor sets

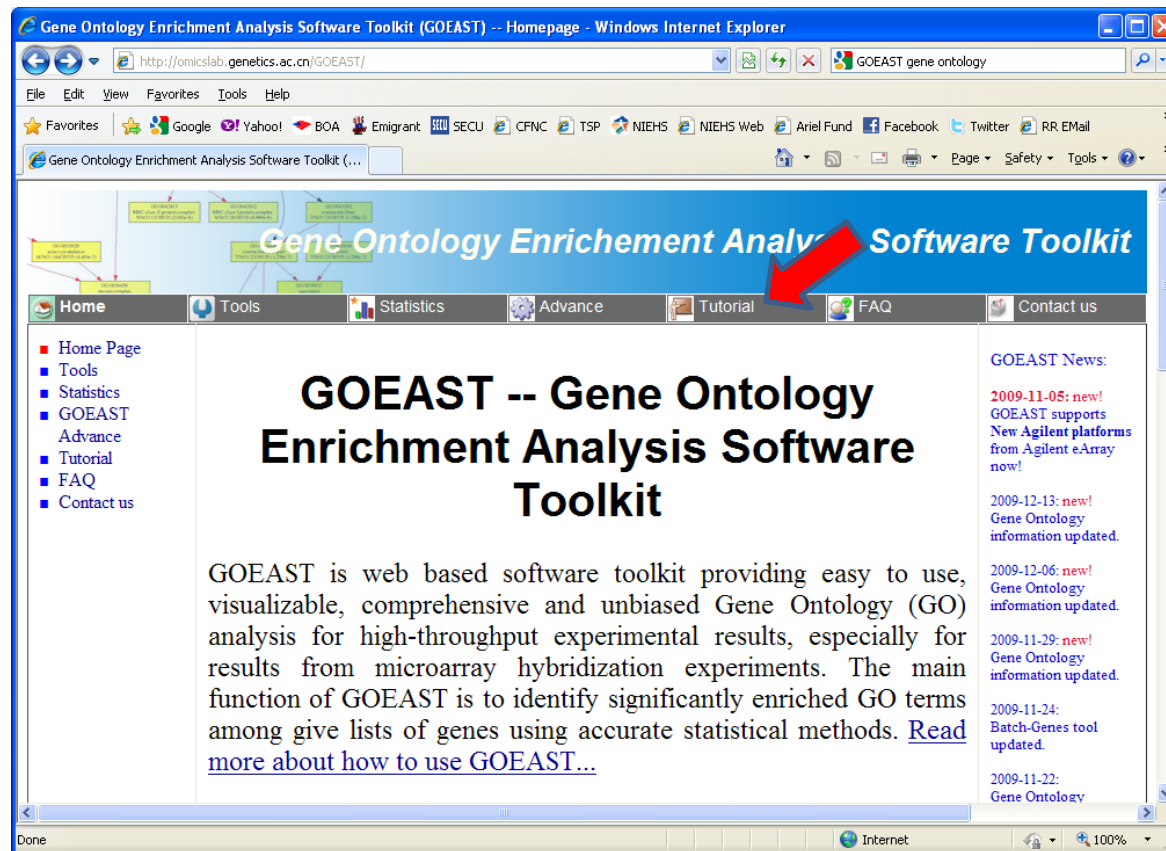
- ❑ **TopGO's weight algorithm**

- The genes are weighted by their relevance in the significant nodes.
- The enrichment score of a parent (gene node  $u$ ) is compared with the scores of its children.
- Children with a better score than  $u$  represent the interesting genes better. Therefore, their significance is increased
- Children with a lower score than  $u$  have their significance reduced.

Alexa A et al. *Bioinformatics* 2006;22:1600-1607

# GOEAST – modification incorporated

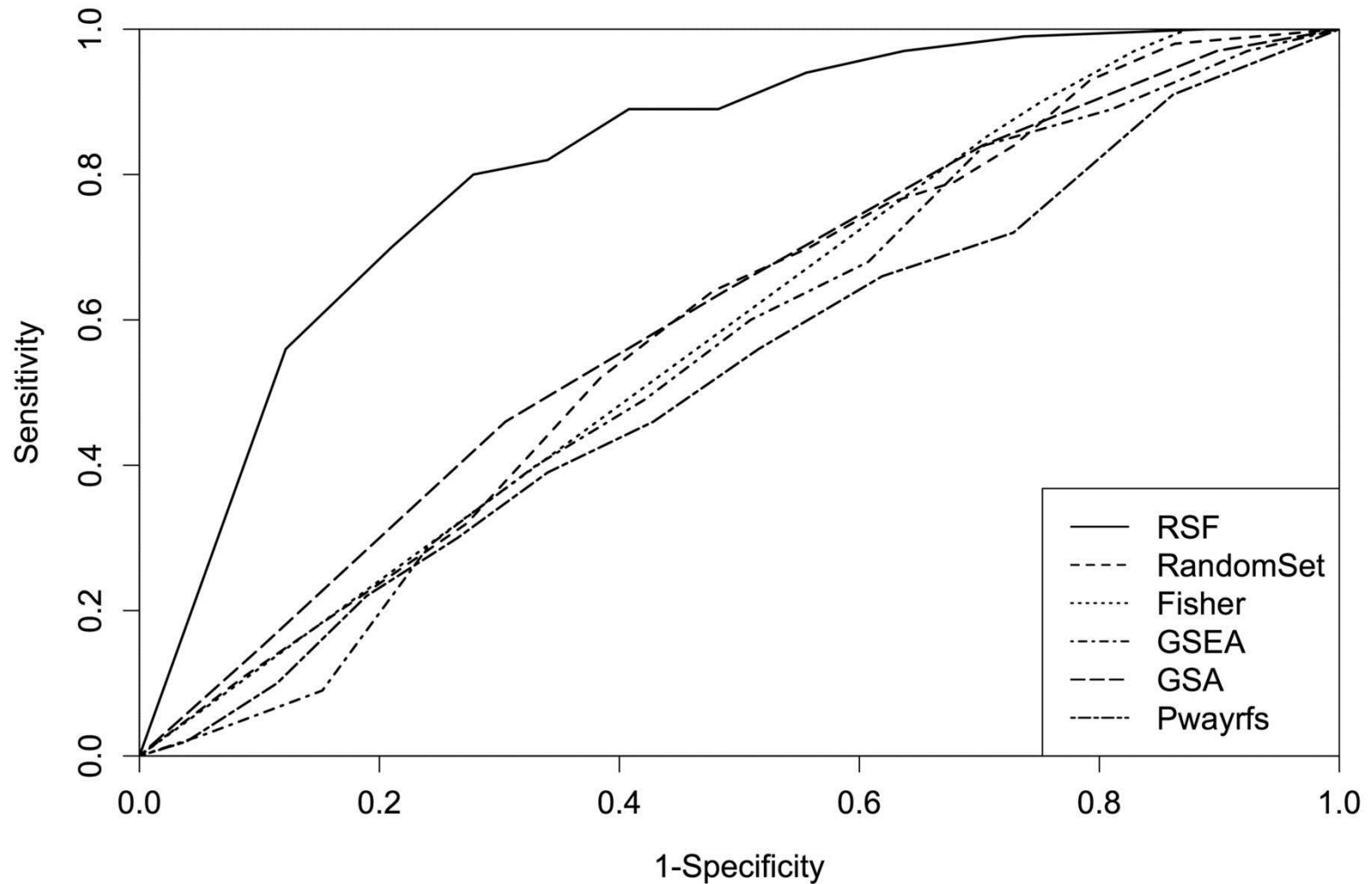
Open web browser and navigate to: <http://omicslab.genetics.ac.cn/GOEAST/>



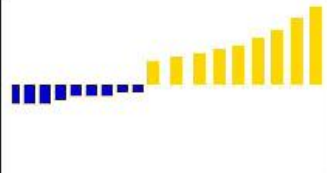
At the header tool bar, click on Tutorial



## Comparison of performances of RSF, random-set, Fisher's exact test, GSEA, GSA and Pwayrfsurvival using simulated expression data.



Chen X , and Ishwaran H *Bioinformatics* 2013;29:99-105



Home Document Supplementary Document Contact



DAVID Bioinformatics Resources 6.7  
National Institute of Allergy and Infectious Diseases (NIAID), NIH

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

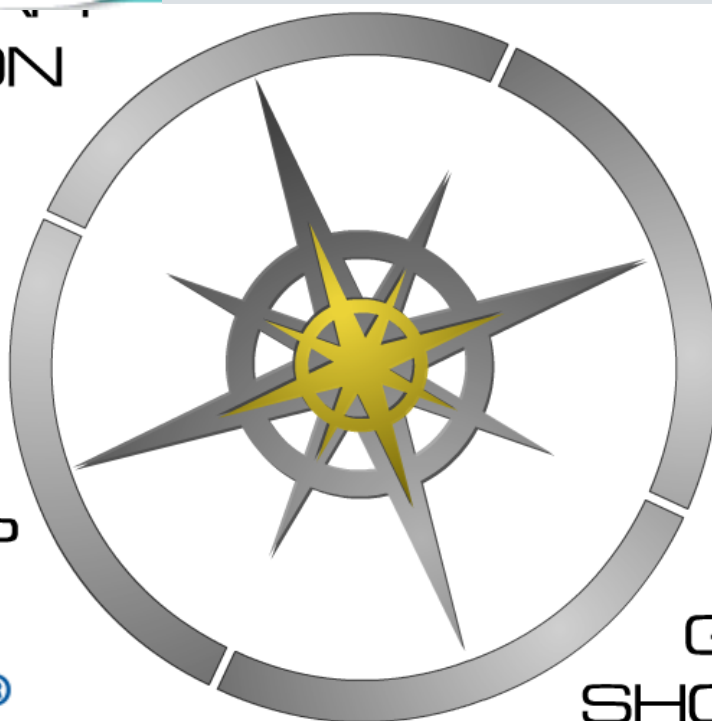


ExPASy  
Bioinformatics Resource Portal

INTRODUCTION



DOWNLOAD  
GenMAPP



BIOBASE  
BIOLOGICAL DATABASES  
NEWS ARCHIVE  
TRANSFAC®  
ExPlain™



GenMAPP  
SHOWCASE



- <http://www-stat.stanford.edu/~tibs/GSA/>
- <http://www.netsci.org/Resources/Software/Bioinform/pathwayanalysis.html>
- <http://www.broadinstitute.org/gsea/index.jsp>
- <http://david.abcc.ncifcrf.gov/>
- <http://www.biocarta.com/>
- <http://web.expasy.org/pathways/>
- <http://www.genmapp.org/>
- <http://www.genome.jp/kegg/>
- <http://www.ingenuity.com/>
- <http://www.genego.com/metacore.php>
- <http://www.geneontology.org/>
- <http://omicslab.genetics.ac.cn/GOEAST/tutorial.php>
- <http://expressome.kobic.re.kr/GAzer/document.jsp>
- <http://www.biobase-international.com/products>
- <http://jaspar.genereg.net/>

# Pathway analysis summary

- Two major statistics approaches
  - Parametric vs. non-parametrics
- Three major databases
  - Gene Ontology
  - KEGG
  - Transfac & transpath
- A few popular applications
  - DAVID (free online)
  - GSEA (online and desktop java application)
  - GSA (R –package)
  - Ingenuity Pathway Analysis (licensed based)
  - Biobase/JASPER (license based)
  - GeneGO (license based)