



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

# Pathway analysis

Biostat & Bioinfo short course series

02/08/2019

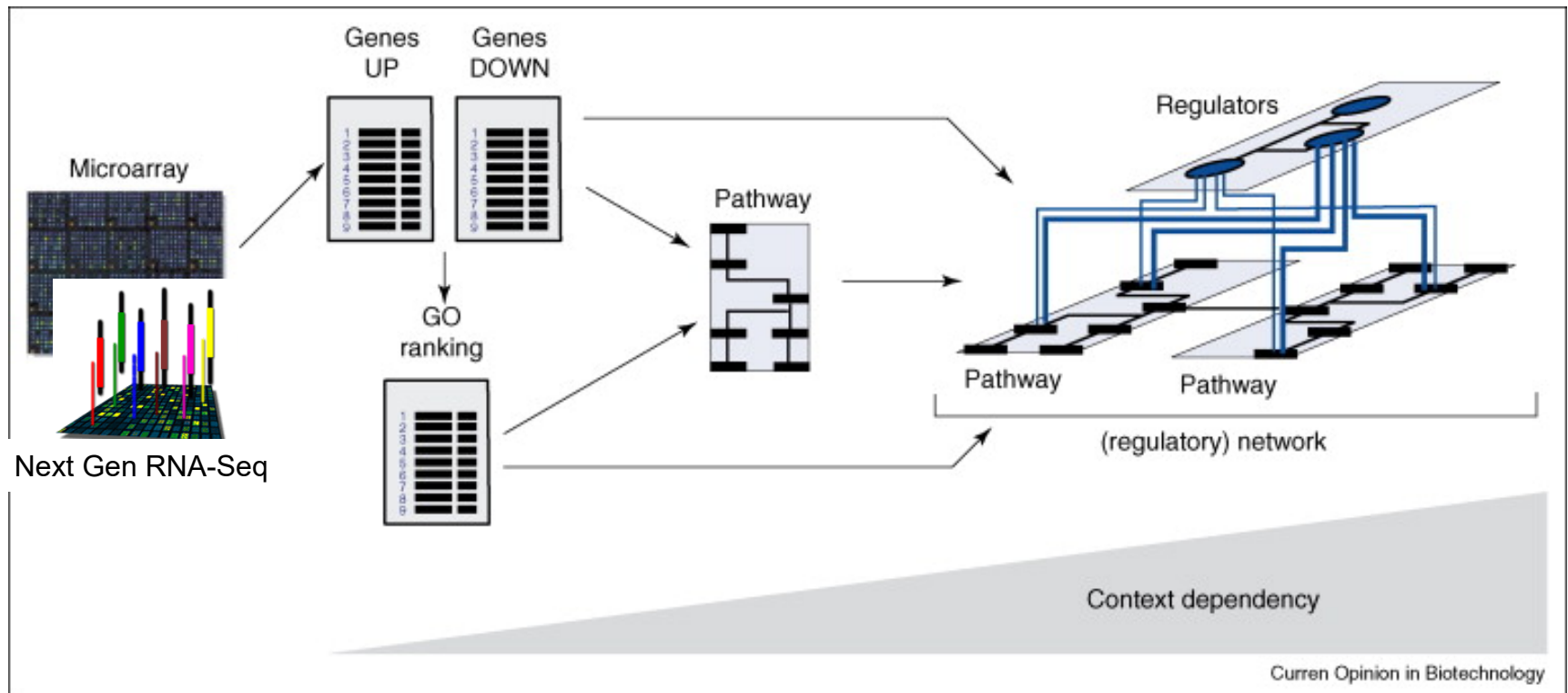
Jianying Li

# An outline of the pathway analysis course

- Background of pathway analysis
- Random variable distribution and its usage in the pathway analysis
- A parametric approach
  - Hypergeometric distribution in details
  - An example: IPA
- Non-parametric approach
  - Experimental layout
  - An example: GSEA
- Pathway analysis summary
  - Making a right choice
  - Pros and cons
- Hands on practice
  - R and basic distribution
  - GSEA



# The Road to Pathway Analysis



# Why pathway analysis?

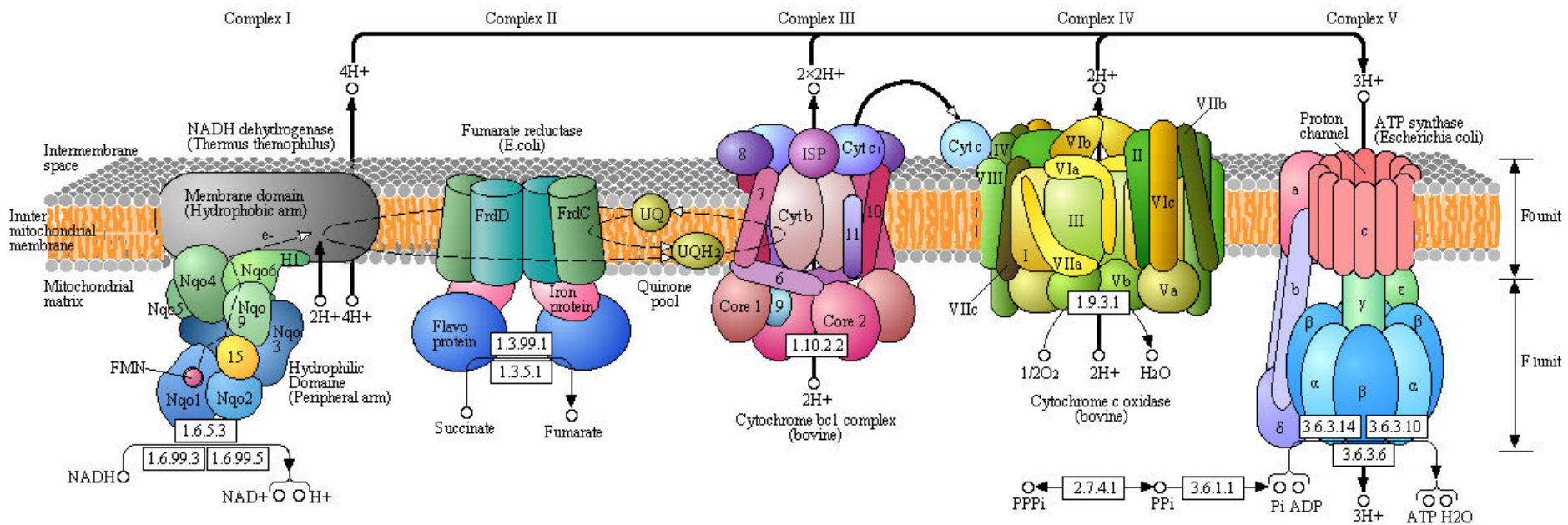
- Biological responses are systematic and collaborative
- Modern technologies provide high-throughput measurement
- Rich knowledge-based accumulation available
- Statistical frame works provide the analytical component

# Common experimental rationale

- An established measurement of gene expression profile:
  - Microarray
  - RNAseq
  - Nano-string
  - Etc.
- Samples obtained from two conditions (e.g. treated vs. control)
- Several biological replicates at each condition
- Our focus is “how would a study object respond differently at two conditions?”
- Gene-centered analysis often
  - Assess statistical difference between conditions via some testing, t-test, limma etc.
  - Correct for multiple testing issues
  - Obtain a set of differentially expressed genes (DEGs)
- Where is the biology??

# The oxidative phosphorylation

- A set of molecules in a cell that work together through a series of actions to achieve a particular outcome



# Rich knowledge-based accumulation available

- GeneOntology (originated since 1998)
  - 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
- KEGG pathway (originated since 1996)
  - Metabolism: carbohydrates, energy, lipid, nucleotides, amino acid, xenobiotics
  - Human diseases
  - Genetic information processing
- Transfac/Transpath
  - Data on transcription factors, their experimentally-proven binding sites, and regulated genes
  - Protein-protein interactions and directed modification of proteins involved in signal transduction pathways,
- It is a knowledge based/driven approach
- Pathway --- gene sets are interexchange

# Gene sets – generalized definition

Gene sets are sets of genes that have something in common, e. g., that they are

- part of the same pathway
- coding for proteins that are part of the same cellular component
- co-expressed under certain conditions
- putative targets of the same regulatory factor
- on the same cytogenetic band
- have come up as hits in some published assay
- Etc.



# Molecular Signatures Database (MSigDB v5.2)

- Free database (<http://www.broadinstitute.org/gsea/msigdb>)

**H** **hallmark gene sets** are coherently expressed signatures derived by aggregating many MSigDB gene sets to represent well-defined biological states or processes.

**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.

**Description** overlap between up genes of Gene-set Example and dn genes of CGP-60474 (12 genes)


### ChEA 2016

STAT3\_1855785\_ChIP-Seq\_MESCs\_Mouse  
 PKCTHETA\_26484144\_ChIP-Seq\_BREAST\_Hur  
 CLOCK\_20551151\_ChIP-Seq\_293T\_Human  
 EZF7\_22180533\_ChIP-Seq\_HELA\_Human  
 ERA\_27197147\_ChIP-Seq\_ENDOMETRIOID-A

### TRANSFAC and JASPAR PWMs

Zfx (mouse)  
 ARNT (human)  
 SP4 (human)  
 ELK1 (human)  
 ATF4 (human)

### Genome Browser PWMs

V\$IK1\_01  
 V\$NFKAPPAB\_01  
 V\$BACH2\_01  
 V\$NFKB\_Q6\_01  
 V\$CREL\_01

### ENCODE and ChEA Consensus TFs from ChIP-X

NFIC\_ENCODE  
 RELA\_ENCODE  
 STAT3\_ENCODE  
 EGR1\_CHEA  
 BRCA1\_ENCODE

### Epigenomics Roadmap HM ChIP-seq

H2BK120ac H9  
 H3K27ac Brain Cingulate Gyrus  
 H3K27me3 IPS DF 19.11  
 H3K4me2 IMR90  
 H3K4me2 H9

### TargetScan microRNA

CTTTGCA,MIR-527  
 GTTAAAG,MIR-302B  
 AAGCACT,MIR-520F  
 TTTGTAG,MIR-520D  
 CACTGTG,MIR-128A,MIR-128B

### ENCODE TF ChIP-seq 2015

POLR2A\_bone marrow macrophage\_mm9  
 RELA\_GM18526\_hg19  
 GATA2\_endothelial cell of umbilical vein\_hg1  
 RELA\_GM12891\_hg19  
 GATA3\_SK-N-SH\_hg19

### TF-LOF Expression from GEO

stat3\_000000000\_a549\_lof\_human\_gpl571\_gs  
 ezh2\_22267199\_heart\_lof\_mouse\_gpl6246\_g  
 emx2\_20962046\_e10dot5\_urogenital\_epithe  
 rent1\_15448691\_hela\_lof\_human\_gpl8300\_g  
 foxa1\_21151129\_mcfdash7\_lof\_human\_gpl1

### ENCODE Histone Modifications 2015

H3K4me3\_GM12864\_hg19  
 H3K27ac\_osteoblast\_hg19  
 H3K27ac\_fibroblast of dermis\_hg19  
 H3K36me3\_spleen\_mm9  
 H3K36me3\_skeletal muscle myoblast\_hg19

### Transcription Factor PPIs

PPARG  
 RXRA  
 PPARD  
 YY1  
 RARA

# The statistical framework

- Well-established statistical and computation algorithms
- Parametric method
  - Hypergeometric test
  - Fisher's exact test
- Non-parametric approaches
  - GSEA
  - GSA
  - Etc.
- Other algorithms
  - **Auto expand** : Draws sub-networks around the selected objects
  - **Shortest paths**: Uses Dijkstra's shortest paths algorithm to find the shortest directed paths between the selected objects.
  - **Self regulation** : Finds the shortest directed paths containing transcription factors between the selected objects
  - Etc.

# A good place to start

- A simple example:
  - 5 patients with the disease  $D$  and 5 healthy control subjects
  - Checked for elevated levels of the blood constituent  $C$ .
  - 4 of the patients, but only 2 of the healthy subjects show an elevated level of  $C$ .
- May we infer that the concentration of  $C$  is elevated in patients with disease  $D$  more often than in healthy subjects?
- Or could our result have been mere coincidence?

# 2x2 contingency table

|                              | Patient with disease D | Healthy control subject | Total |
|------------------------------|------------------------|-------------------------|-------|
| Elevated level of compound C | 4                      | 2                       | 6     |
| Normal level of compound C   | 1                      | 3                       | 4     |
| Total                        | 5                      | 5                       | 10    |

$$E_{r,c} = \frac{(\text{Sum of row } r) \times (\text{Sum of column } c)}{\text{Sample size}}$$

|                              | Patient with disease D | Healthy control subject | Total |
|------------------------------|------------------------|-------------------------|-------|
| Elevated level of compound C | 4<br>(3)               | 2<br>(3)                | 6     |
| Normal level of compound C   | 1<br>(2)               | 3<br>(2)                | 4     |
| Total                        | 5                      | 5                       | 10    |

# Hypergeometric distribution

Probability to get this 2×2 table without an association between *D* and *C*:

$$\frac{\begin{array}{l} \text{Number of ways to} \\ \text{choose 4 out of 5} \\ \text{patients to} \\ \text{have elevated C} \end{array} \times \begin{array}{l} \text{Number of ways to} \\ \text{choose 2 out of 5} \\ \text{controls to} \\ \text{have elevated C} \end{array}}{\begin{array}{l} \text{nuber of ways to} \\ \text{choose 6 our of 10} \\ \text{persons to have} \\ \text{elevated C} \end{array}} = \frac{\binom{5}{4} \binom{5}{2}}{\binom{10}{6}}$$

in R:

```
> dhyper( 4, 5, 5, 6 )  
[1] 0.2380952
```



# Hypergeometric distribution

Under the null hypothesis, i.e., the assumption that there is no association between elevated levels of compound C and presence of disease D, the probability that all 5 patients have elevated levels of C would be,

|                              | Patient with disease D | Healthy control subject | Total |
|------------------------------|------------------------|-------------------------|-------|
| Elevated level of compound C | 5                      | 1                       | 6     |
| Normal level of compound C   | 0                      | 4                       | 4     |
| Total                        | 5                      | 5                       | 10    |

in R:

```
> dhyper( 5, 5, 5, 6 )  
[1] 0.02380952
```

# Hypergeometric distribution

Under the null hypothesis, i.e., the assumption that there is no association between elevated levels of compound *C* and presence of disease *D*, the probability that 4 or even more of the patients have elevated levels of *C*,

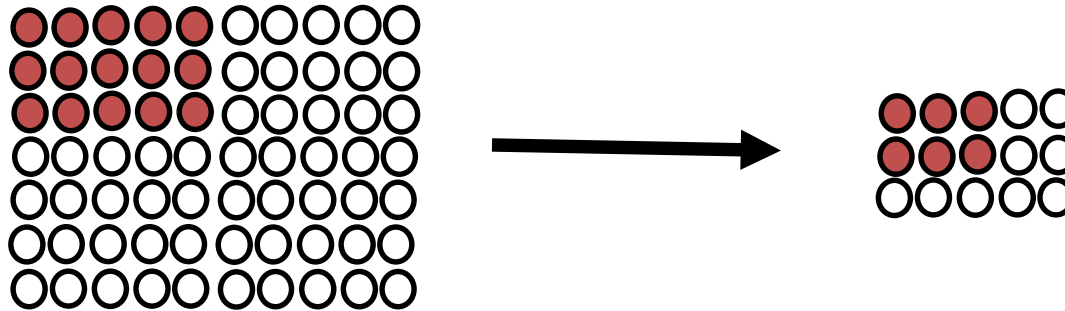
$$p = \frac{\binom{5}{4}\binom{5}{2}}{\binom{10}{6}} + \frac{\binom{5}{5}\binom{5}{1}}{\binom{10}{6}} = 0.26$$

**This is insignificant**

```
in R:  
> 1 - phyper(3, 5, 5, 6 )  
[1] 0.2619048
```



# Hypergeometric Test (Right-tailed)



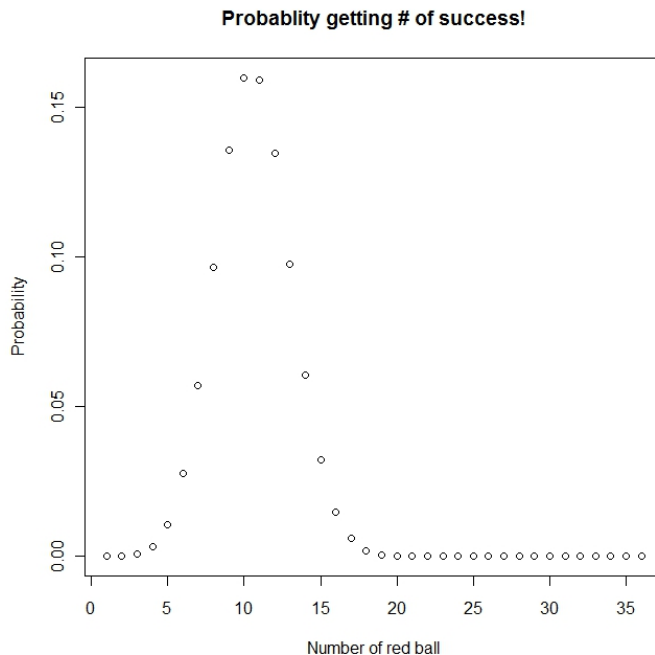
- An urn with two types of marbles:
  - N total # of marbles
  - Of which, m # of **red** marbles
  - Drawing a red marble is a success!
  - Drawing a white marble is a failure!
- n is the # of marbles randomly drawn
- k is the # of successes (**red marbles**) in the sample
- Hypergeometric distribution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

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

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

# 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  $p(k=20, 36, 50, 170)$



# Pathway analysis (behind the scene)

|                    | <u>Contingency table</u> |          |                        |
|--------------------|--------------------------|----------|------------------------|
|                    | DEG                      | Not DEGs | totals                 |
| In a GO category   | <b>x</b>                 | m-x      | <b>m</b>               |
| 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

```

Hypergeometric test vs. FET, we shall get same result

- *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*

## Explore



## Datasets

Annotate and filter datasets and use them directly for hypothesis generation when exploring pathways and gene lists.

[› Annotate Datasets](#) [› Filter datasets](#)



## Compare

Identify the union, unique, and common molecules across lists, pathways, biomarkers, and analyses.

[› Compare data](#)



## Pathways

Create pathways from your datasets, targets, biomarkers, diseases and biological functions. Communicate pathways and network results through visually enhanced representations.

[› Build pathways](#) [› Design pathways](#)

## Analyze



## Core

Interpret your data in the context of biological processes, pathways, and networks.

[› Analyze dataset](#) [› Compare analyses](#)



## IPA-Tox

Assess toxicity and safety of test compounds in the context of toxicological processes, pathways, and networks.

[› Analyze dataset](#) [› Compare analyses](#)



## IPA-Biomarker

Filter your datasets and identify and prioritize potential biomarker candidates.

[› Analyze dataset](#) [› Compare analyses](#)



## IPA-Metabolomics

Explore genotype-phenotype relationships and environmental influences via metabolite data.

[› Analyze dataset](#) [› Compare analyses](#)

# Ingenuity Pathway Analysis (IPA)

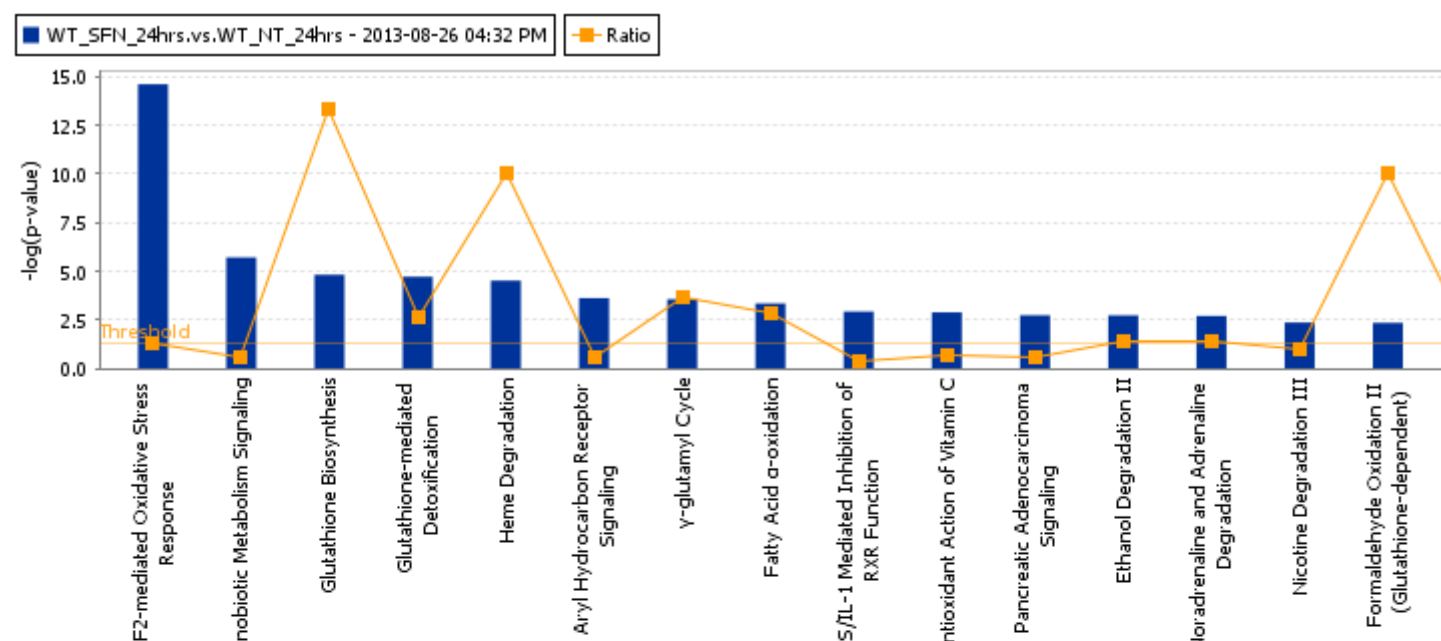
## --Knowledgebase

- 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

ADD TO MY PATHWAY ADD TO MY LIST CREATE DATASET CUSTOMIZE TABLE

|                          | 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

# Hypergeometric Test (Right-tailed)

- Used in Ingenuity Pathway Analysis (IPA)
  - Commercial software  
(<http://www.ingenuity.com/products/ipa>)
  - Pros:
    - Great source of clean, expertly curated gene sets
  - Cons:
    - Not free
    - Throws away information by only using DEG list



# Hypergeometric Test (Right-tailed)

- Used in Database for Annotation, Visualization and Integrated Discovery (DAVID)
  - Free software (<https://david.ncifcrf.gov>)
  - Pros:
    - Easy to use (web-based)
    - Large collection of gene sets
  - Cons:
    - Gene sets are not as clean
    - Throws away information by only using DEG list

# Sampling over genes

- Hypergeometric testing for gene sets has been criticized on the ground of it sampling over genes (observation) instead of over microarrays (subjects)
- Hence, the meaning of the  $p$  values is quite unclear.
- Especially: Correlations between genes inflate the apparent sample size, causing potentially severe over-estimation of significance.
- Increasing the number of replicates influences significance only indirectly.

# Sampling over subjects

- Instead of using the hypergeometric distribution to get a  $p$  value from our statistic, we should better use subject permutation:
  - Let  $L_0$  be the list of differential expressed genes and  $m=|L_0|$  its size.
  - For  **$N$  permutations**  $\sigma_i$  ( $i=1,\dots,N$ ) of the *subject* labels, calculate the DE statistic and let  $L_i$  be the list of the  $m$  top ranking genes.
  - Let  $k_i$  be the number of differentially expressed genes in the gene set, i.e. the size of the intersection  $L_i \cap S$ .
  - The  $p$  value for gene set  $S$  is now the fraction of permutation that had a larger gene set than the correct sample assignment, i.e.,

$$p = \frac{|\{i | k_i > k_0\}|}{N}$$

# 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  |

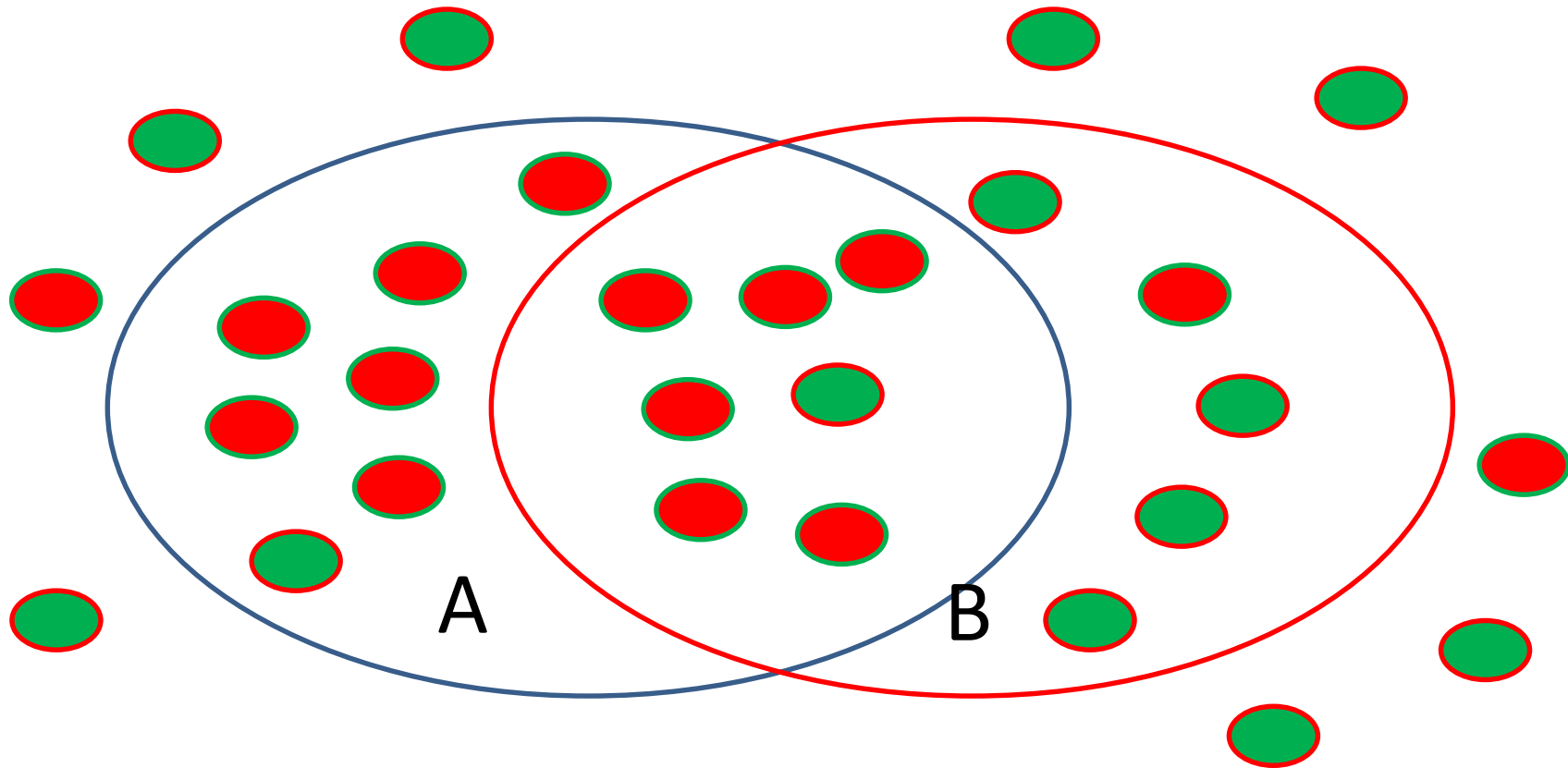
$p=0.049$

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  |

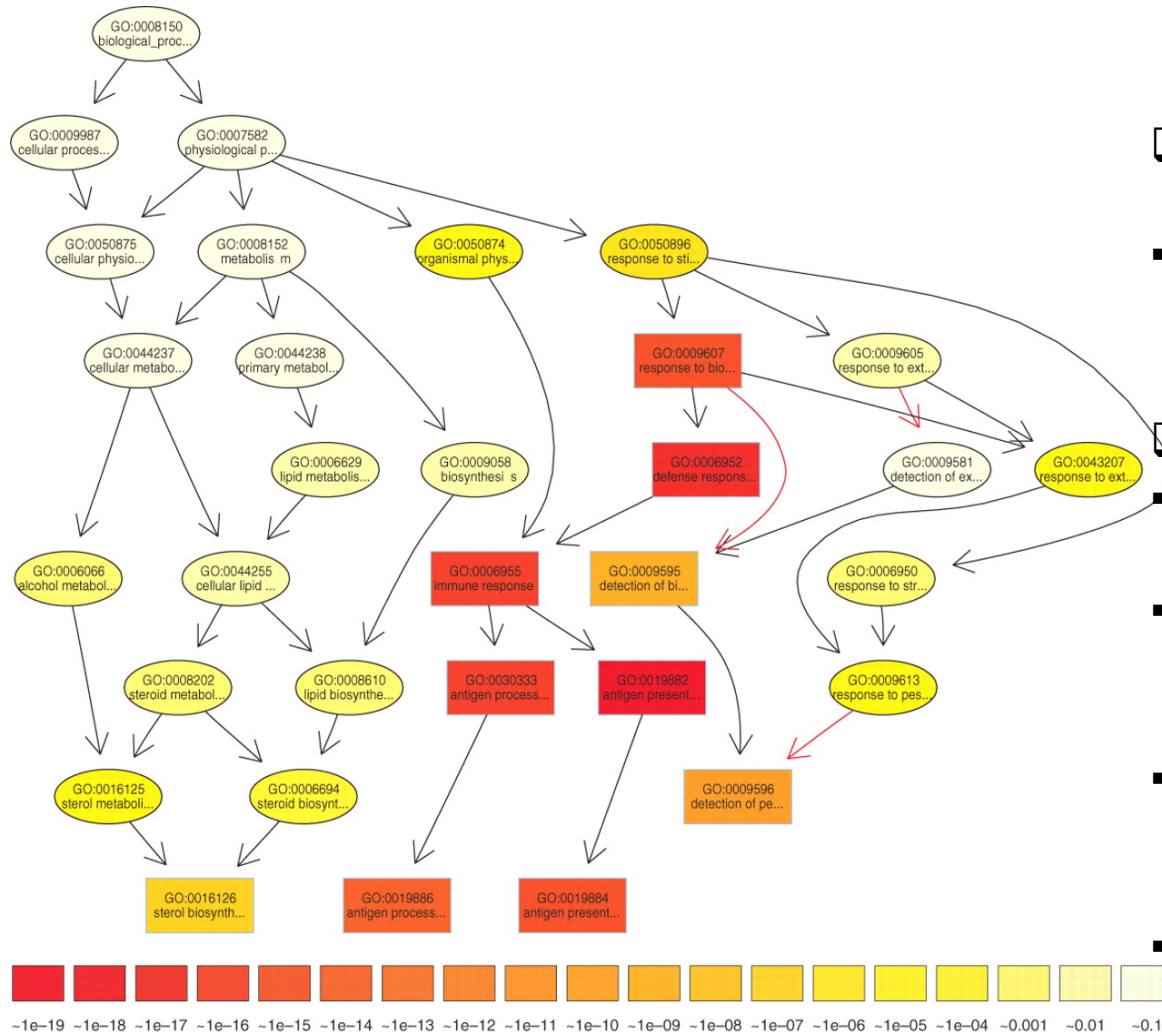
$p=0.0048$

# 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

# A non-parametric approach

- Compare a sample empirical distribution function to the reference distribution
- Or, compare two sample empirical distribution functions
- Required data:
  - Ranked list of genes sorted by differential expression (includes all genes)
  - A gene set
- If the genes in the gene set tend to fall near either end of the ranked list, the gene set is considered significantly enriched
- The test significance is obtained from empirical distribution
  - Permutation based
  - Bootstrapping
  - Etc.

# A non-parametric approach (GSEA)

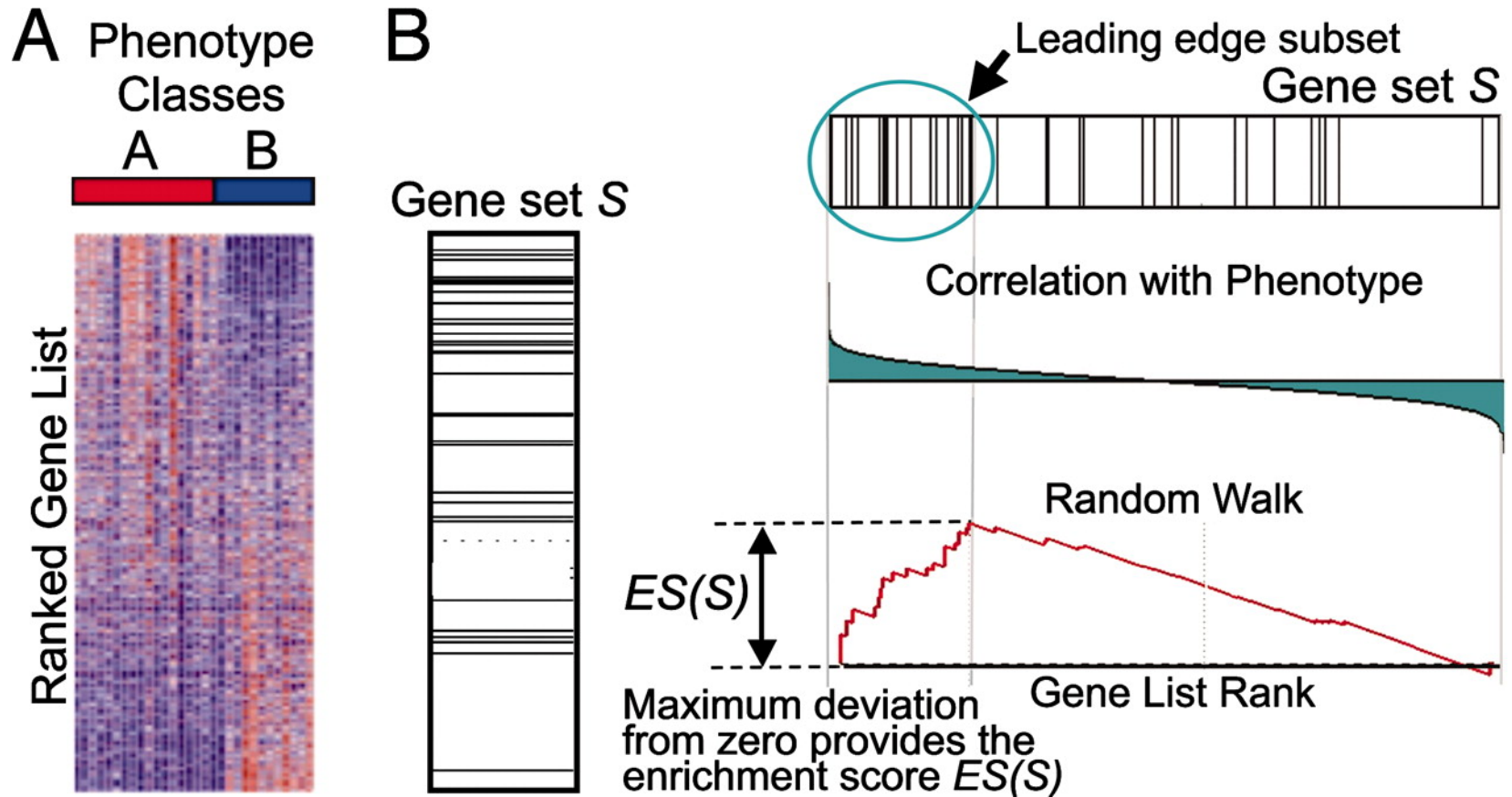
- Mootha *et al.* [2003] suggest to use Kolmogorov-Smirnov (K-S) test
  - Sort all genes by LFC.
  - Go through the list, increasing a running sum for each gene in the gene set by  $(N-n)$ , and decreasing it for each gene not in the gene set by  $n$ .  
  
[ $N$ : number of genes,  $n$ : size of gene set]
  - The maximum value of the running sum is the enrichment score (ES).



# A non-parametric approach

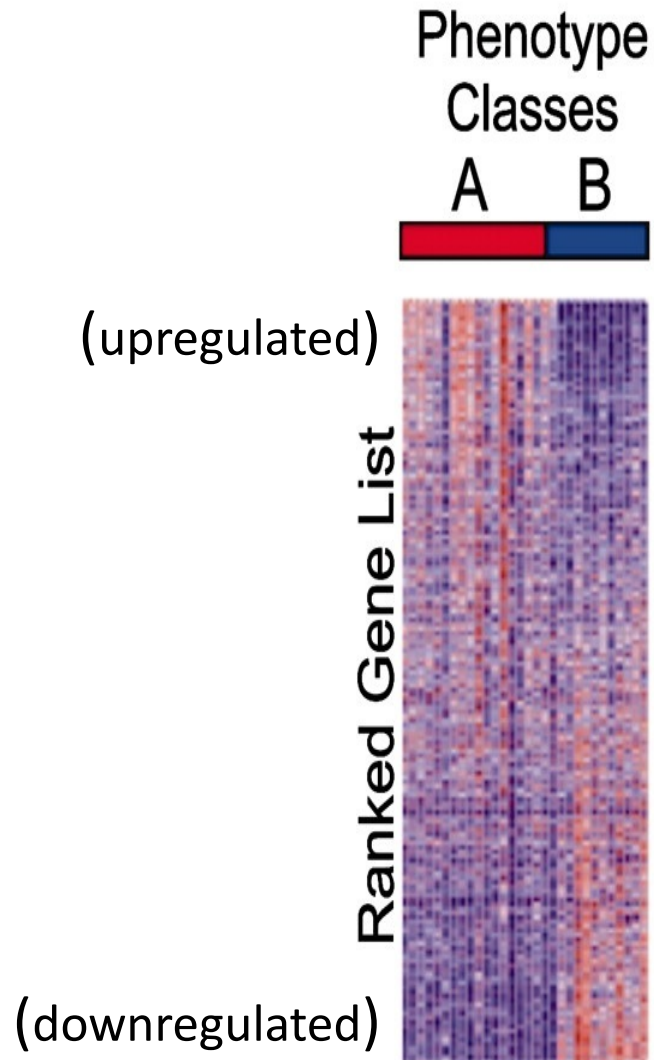
- Assessing significance
  - To get  $p$  values, we do not use the KS distribution but rather estimate the null by subject permutation
- Improved enrichment score
  - The KS statistic tests whether distributions are different, but this difference may not have a clear direction, making biological interpretation difficult
  - The updated GSEA algorithm [Subramanian *et al.*, PNAS **102** (2005) 15545] weights the in-/decrements of the running sum by the LFC.

A GSEA overview illustrating the method.

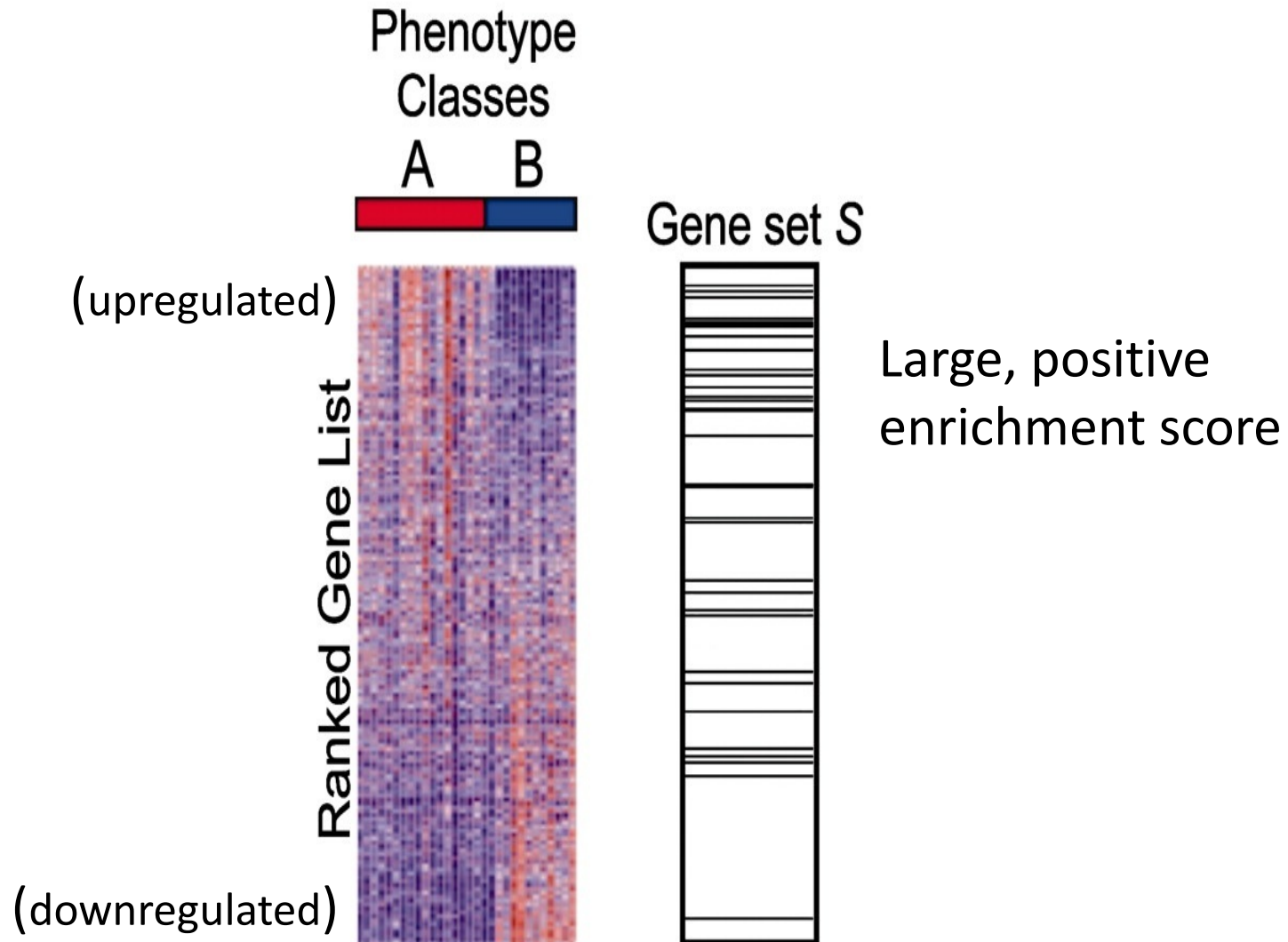


Subramanian A et al. PNAS 2005;102:15545-15550

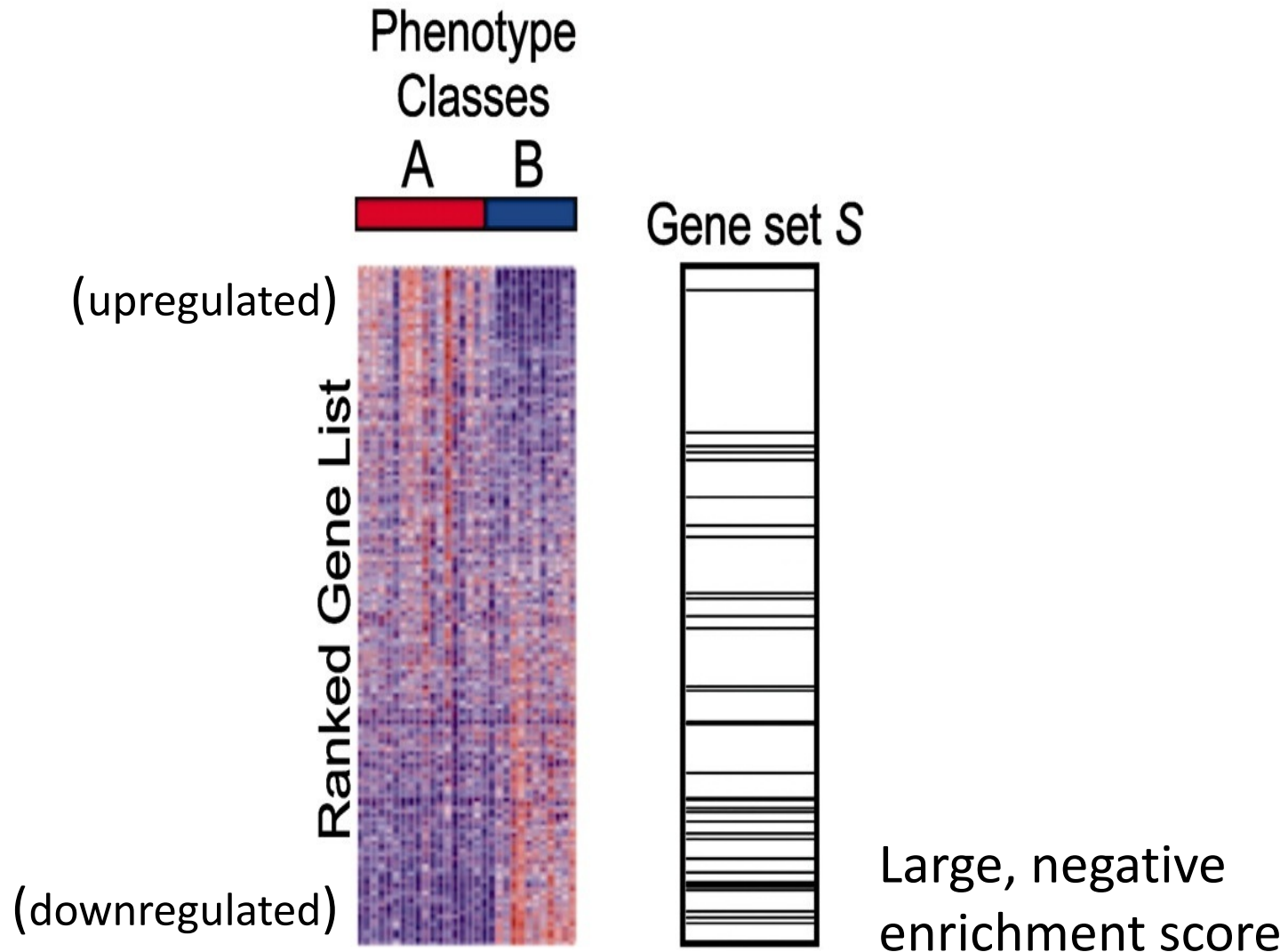
# Kolmogorov–Smirnov (KS) Test



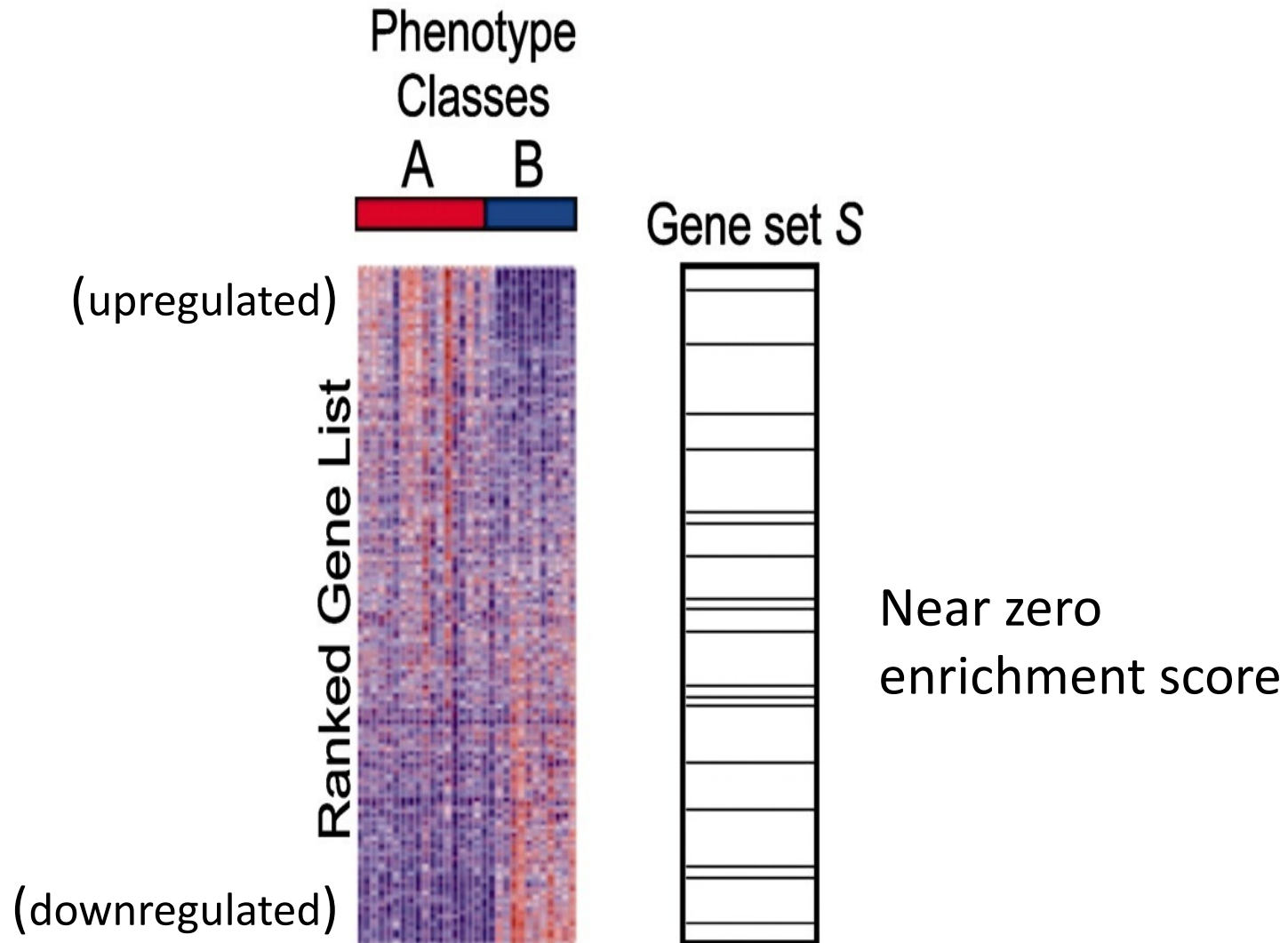
# Kolmogorov–Smirnov (KS) Test



# Kolmogorov–Smirnov (KS) Test

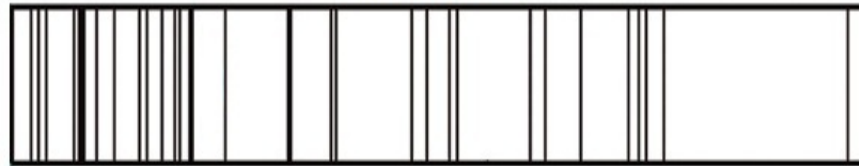


# Kolmogorov–Smirnov (KS) Test



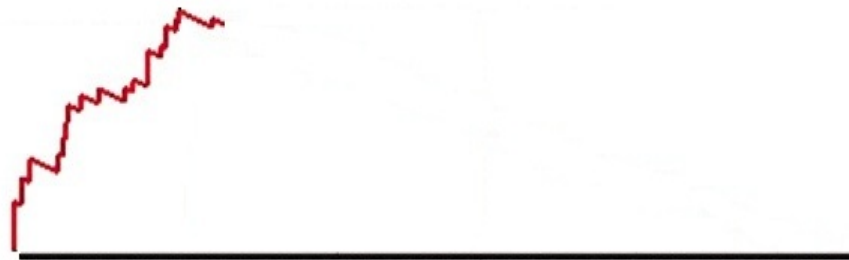
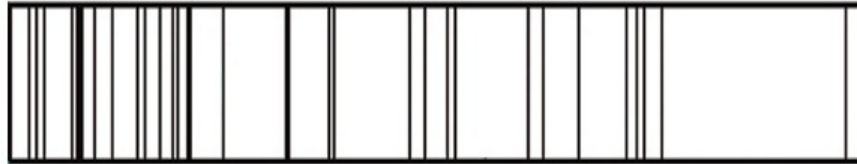
# Kolmogorov–Smirnov (KS) Test

(upregulated) Gene Set (downregulated)



# Kolmogorov–Smirnov (KS) Test

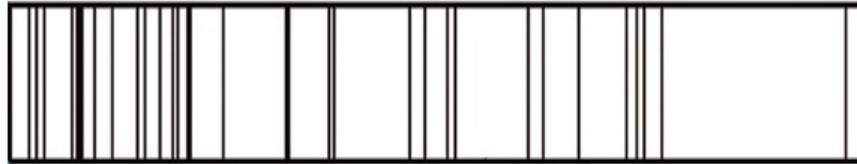
(upregulated) Gene Set (downregulated)





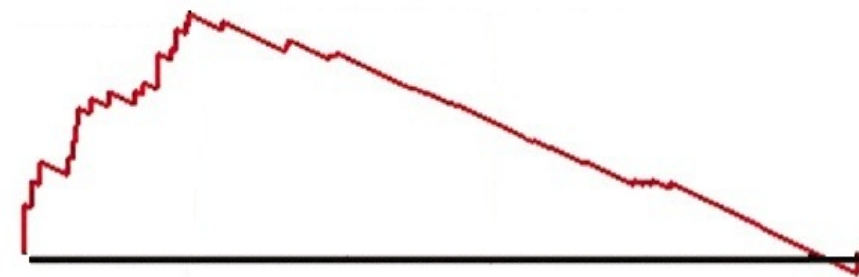
# Kolmogorov–Smirnov (KS) Test

(upregulated) Gene Set (downregulated)

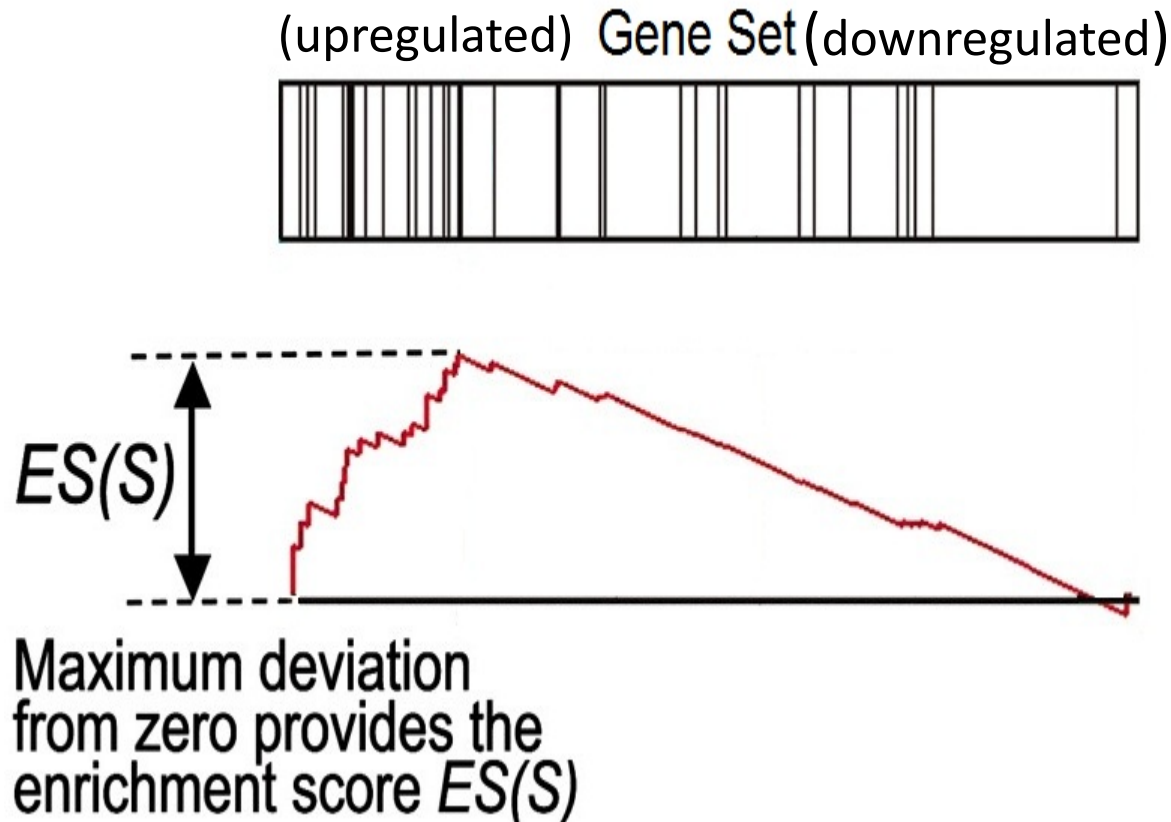


# Kolmogorov–Smirnov (KS) Test

(upregulated) Gene Set (downregulated)

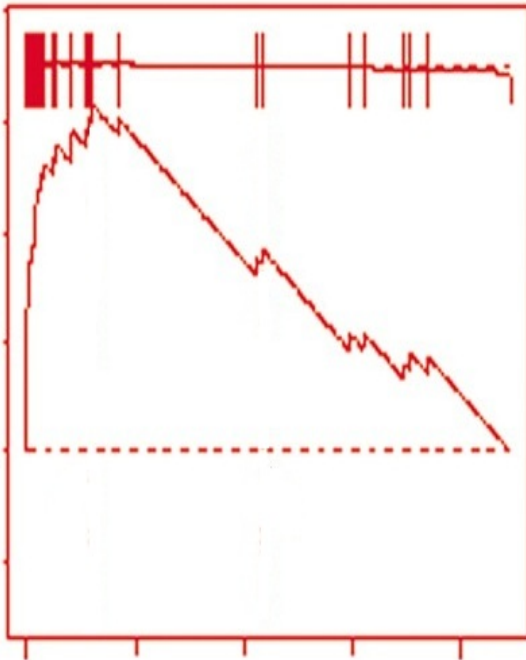


# Kolmogorov–Smirnov (KS) Test



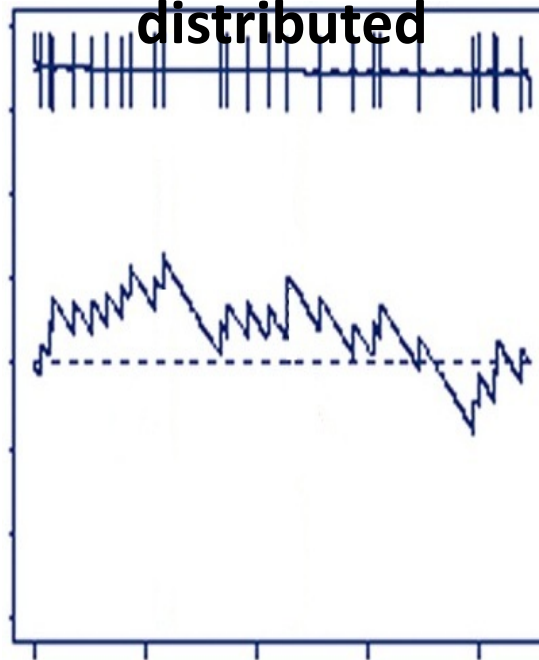
# Kolmogorov–Smirnov (KS) Test

**Many genes  
upregulated**



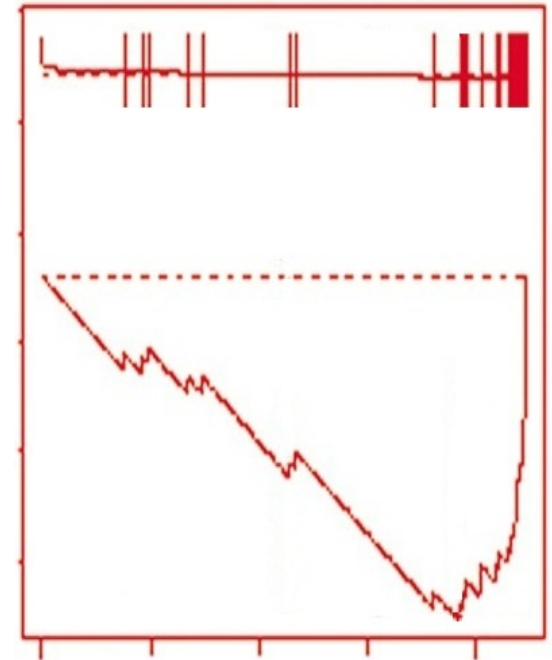
Large positive  
enrichment score

**Genes  
randomly  
distributed**

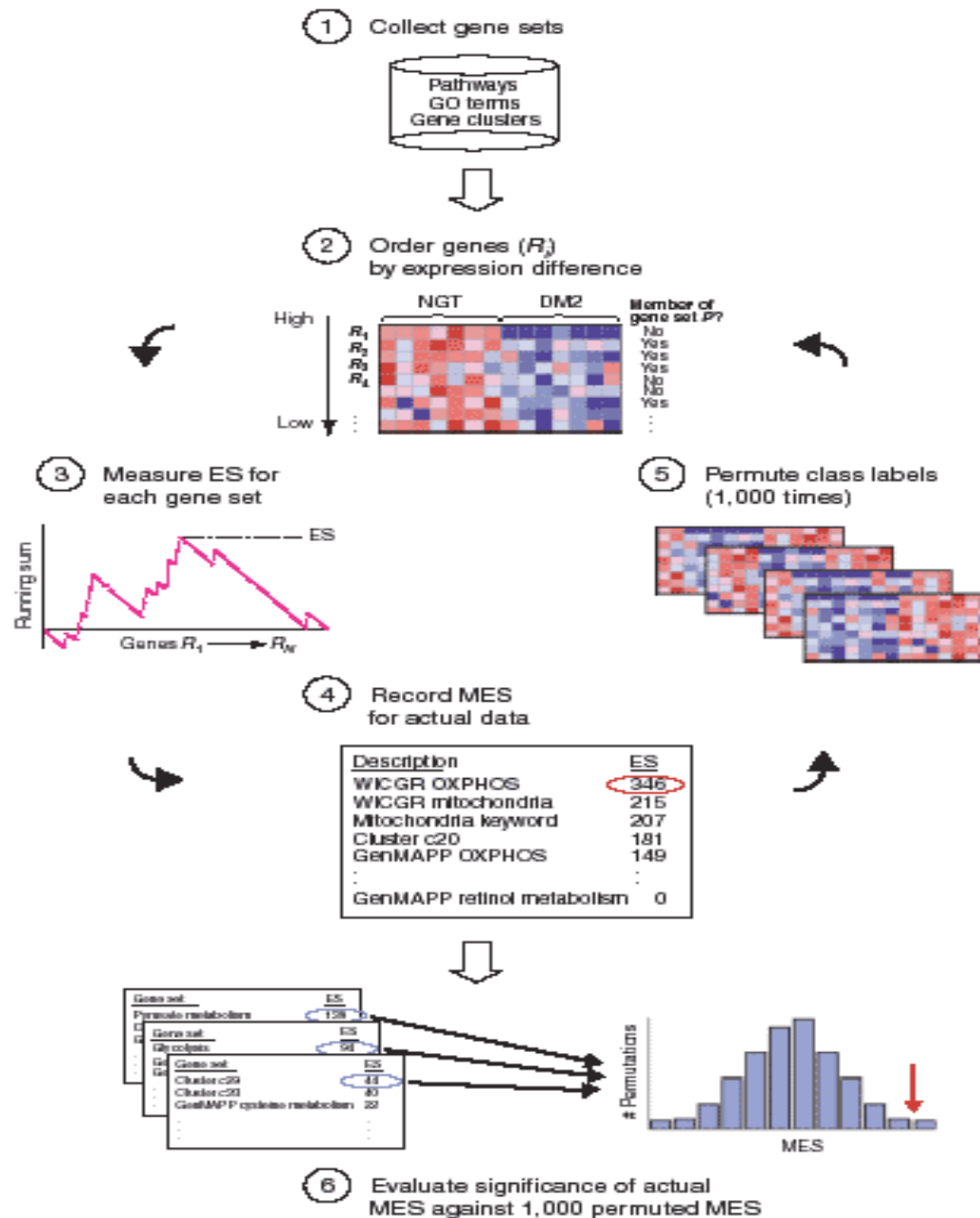


Near zero  
enrichment score

**Many genes  
downregulated**



Large negative  
enrichment score



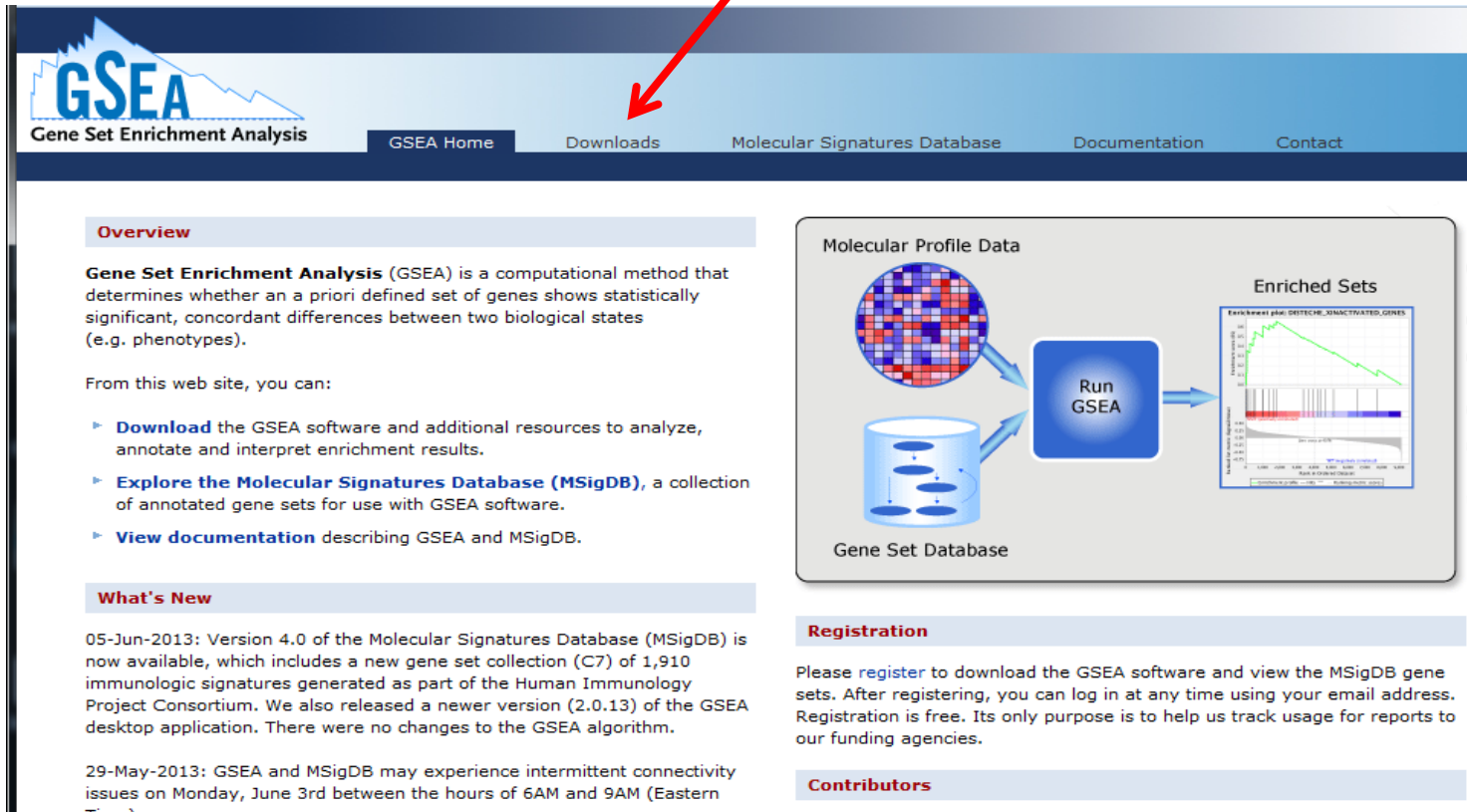
# Kolmogorov–Smirnov (KS) Test

- Used in Gene Set Enrichment Analysis (GSEA)
  - Free software  
(<http://www.broadinstitute.org/gsea>)
  - Pros:
    - Large collection of gene sets
    - Uses more information than methods that only use DEG list
    - Enrichment plot improves interpretability
  - Cons:
    - Permutation-based p-values

Hands on practice on both parametric and non-parametric

**NOW, IT IS YOUR TURN**

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:

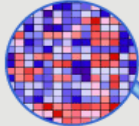
- ▶ **Download** the GSEA software and additional resources to analyze, annotate and interpret enrichment results.
- ▶ **Explore the Molecular Signatures Database (MSigDB)**, a collection of annotated gene sets for use with GSEA software.
- ▶ **View documentation** describing GSEA and MSigDB.

### 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**



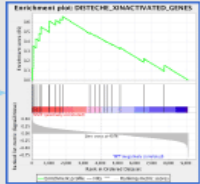
**Gene Set Database**



**Run GSEA**



**Enriched Sets**



The diagram illustrates the GSEA workflow. It starts with 'Molecular Profile Data' (represented by a heatmap) and 'Gene Set Database' (represented by a database icon). Both inputs feed into a central 'Run GSEA' button. The output of the 'Run GSEA' button is an 'Enrichment plot' showing a green curve peaking at the top of the plot, indicating a set of genes that is enriched in the 'ON' state. The plot is titled 'Enrichment plot: DRITCHEB\_XINACTIVATED\_GENES'.

### 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


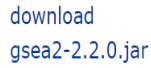

# GSEA Example

## Downloads

The GSEA software and source code and the Molecular Signatures Database (MSigDB) are freely available to individuals in both academia and industry for internal research purposes. Please see the [GSEA/MSigDB license](#) for more details.

### Software

There are several options for GSEA software. All options implement exactly the same algorithm. Usage recommendations and installation instructions are listed below. Current Java implementations of GSEA require Java 6 or 7.

|   |   |  |
|---|---|--|
| <b>javaGSEA<br/>Desktop Application</b> | <ul style="list-style-type: none"><li>▶ Easy-to-use graphical user interface</li><li>▶ Runs on any desktop computer (Windows, Mac OS X, Linux etc.) that supports Java 6 or 7</li><li>▶ Produces richly annotated reports of enrichment results</li><li>▶ Integrated gene sets browser to view gene set annotations, search for gene sets and map gene sets between platforms</li></ul> | Launch with<br>1GB (for 32 or 64-bit Java) ▾<br>memory:<br> |
| <b>javaGSEA<br/>Java Jar file</b>       | <ul style="list-style-type: none"><li>▶ Command line usage</li><li>▶ Runs on any platform that supports Java 6 or 7</li><li>▶ We recommend using the 'Launch' buttons above instead of this mode for most users</li></ul>   |   |
| <b>R-GSEA</b>                           | <ul style="list-style-type: none"><li>▶ Usage from within the R programming environment</li></ul>   |   |

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**

Enrichment in phenotype: best hit example

Enrichment in phenotype: set of examples
- 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)

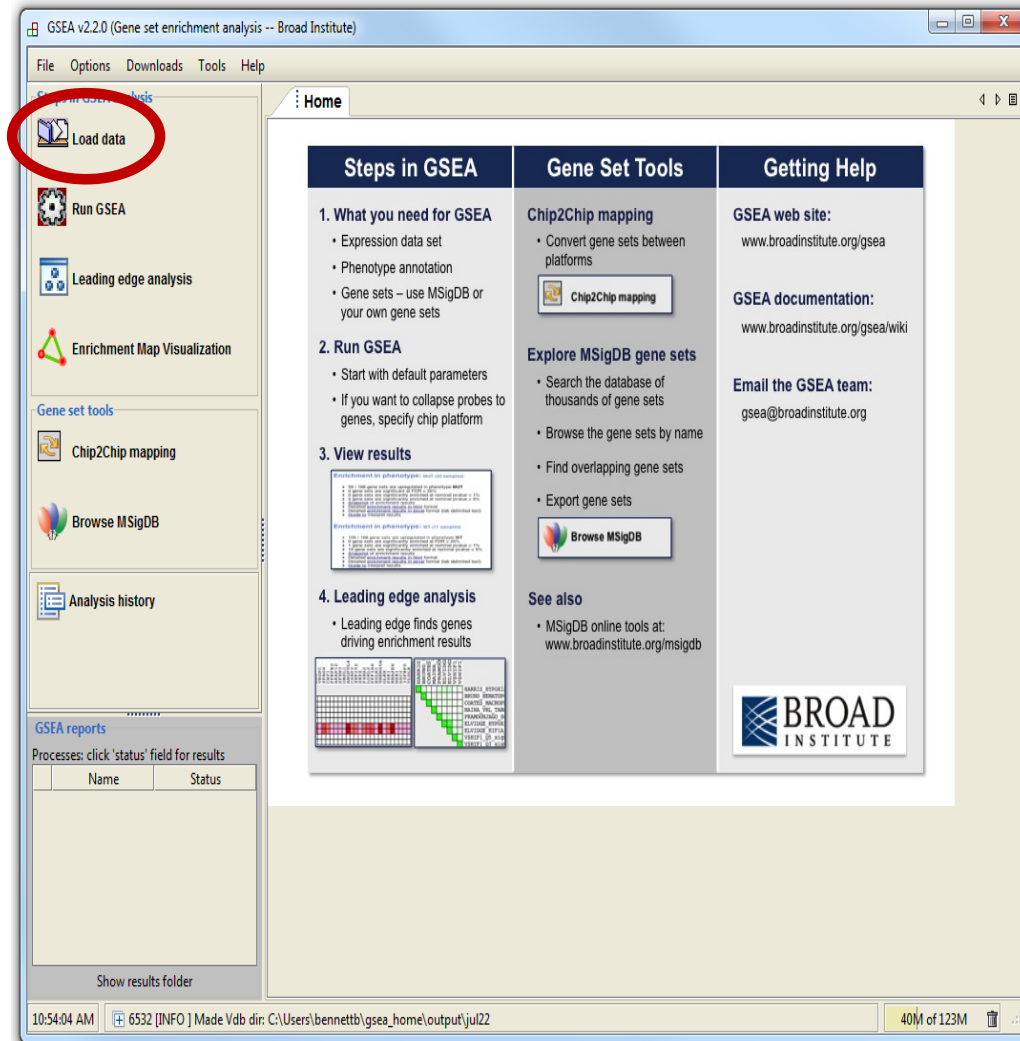
**Email the GSEA team:**

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

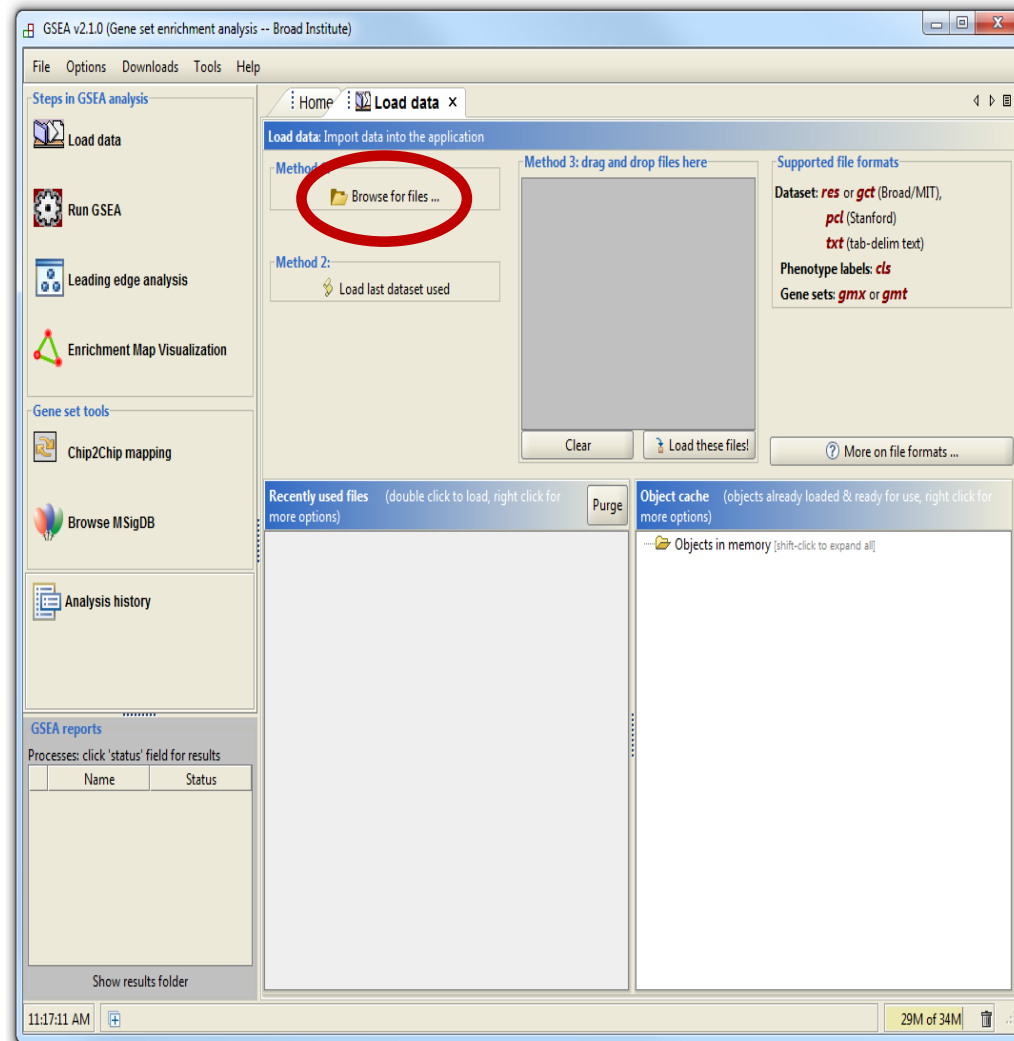
**BROAD INSTITUTE**

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

# GSEA Example



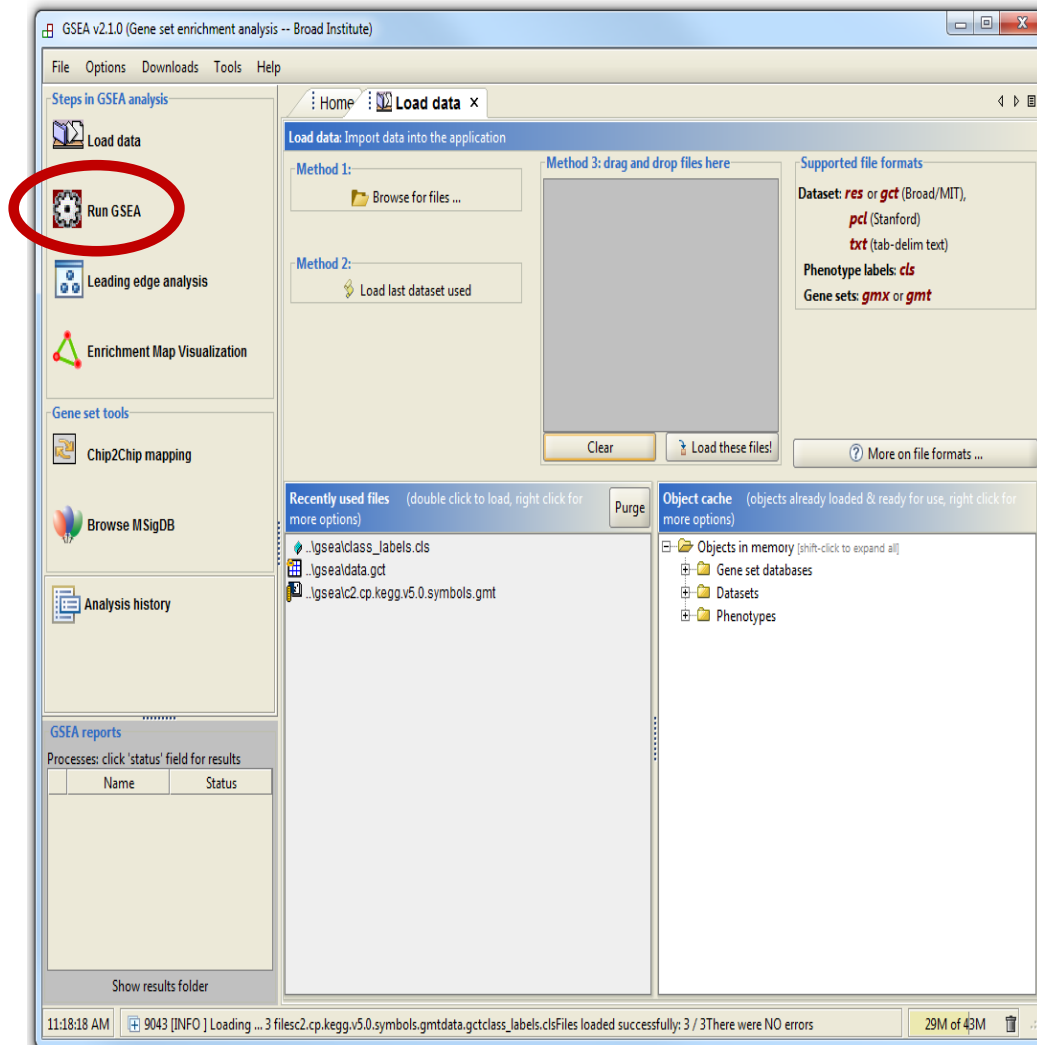
# GSEA Example



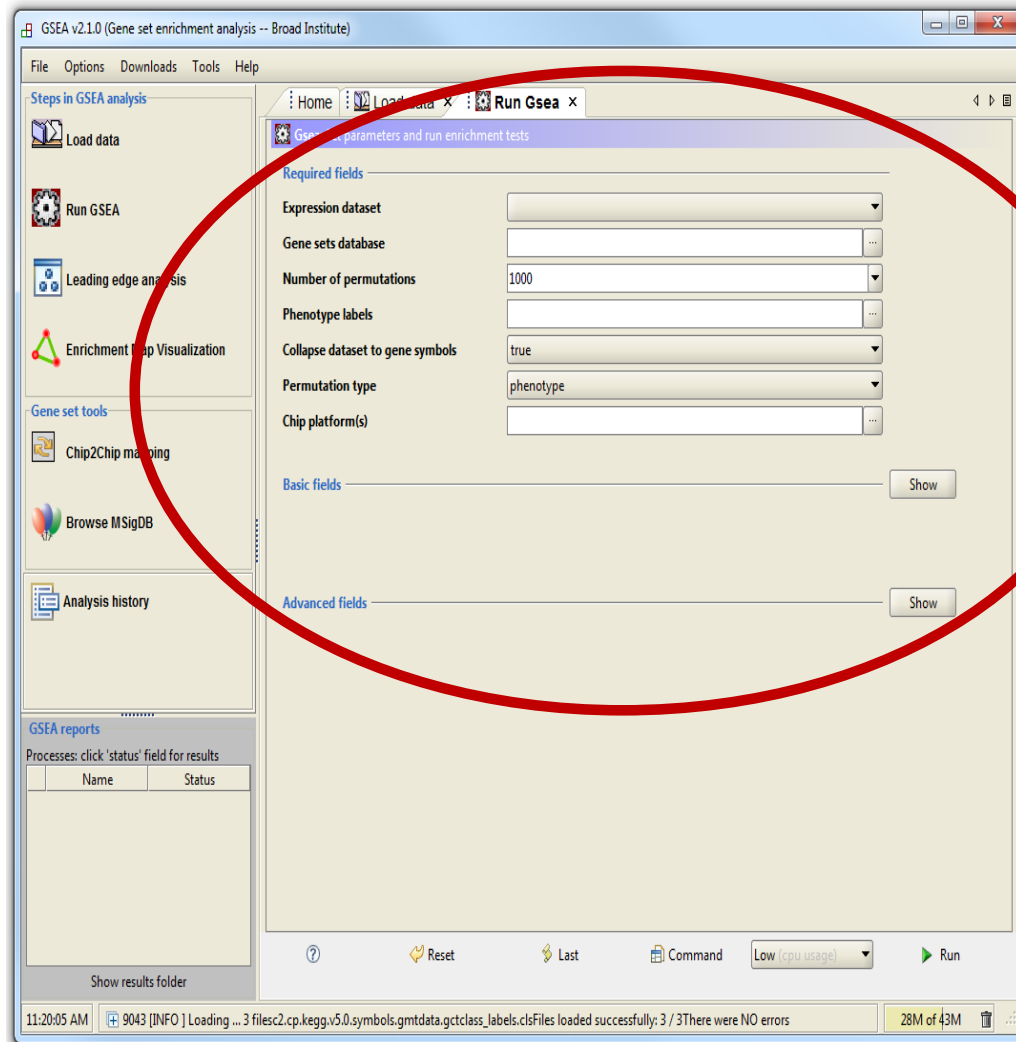
# GSEA Example

- Supported file types:
  - [http://www.broadinstitute.org/cancer/software/gsea/wiki/index.php/Data\\_formats](http://www.broadinstitute.org/cancer/software/gsea/wiki/index.php/Data_formats)
- Required:
  - Expression data file
  - Class label file
- Optional:
  - Gene set file

# GSEA Example



# GSEA Example





## Steps in GSEA analysis



## Load data



Run GSEA



### Leading edge analysis



### Enrichment Map Visualization

## Gene set tools



## Chip2Chip mapping



[Browse MSigDB](#)



### Analysis history

## GSEA reports

Processes: click 'status' field for results

|   | Name   | Status  |
|---|--------|---------|
| 1 | * Gsea | Running |

Home Load data × Run Gsea ×



### Gsea: Set parameters and run enrichment tests

### Required fields

### Expression dataset

DEG overall gender diff [3758x24 (ann: 3758,24,chip na)]

Gene sets database

[roadinstitute.org/pub/qsea/gene\\_sets/c5.all.v5.1.symbols.gmt](http://roadinstitute.org/pub/qsea/gene_sets/c5.all.v5.1.symbols.gmt)

### Number of permutations

1000

### Phenotype labels

e:\spring2014\GSEA\gender\_label\_word.ds#Male versus Female

### Collapse dataset to gene symbols

true

Permutation type

phenotype

Chip platform(s)

[broadinstitute.org/pub/qsea/annotations/GENE\\_SYMBOL.chip](http://broadinstitute.org/pub/qsea/annotations/GENE_SYMBOL.chip)

## Basic fields

Show

### Advanced fields

Show

### Steps in GSEA analysis



### Load data



Run GSEA



### Leading edge analysis



### Enrichment Map Visualization

## Gene set tools



## Chip2Chip mapping



[Browse M SigDB](#)



### Analysis history

## GSEA reports

Processes: click 'status' field for results

|   | Name | Status    |
|---|------|-----------|
| 1 | Gsea | Success 5 |

**click**

Home Load data × Run Gsea ×



## Gsea: Set parameters and run enrichment tests

### Required fields

## Expression dataset

DEG\_overall\_gender\_diff [3758x24 (ann: 3758,24,chip na)]

Gene sets database

[roadinstitute.org://pub/gsea/gene\\_sets/c5.all.v5.1.symbols.gmt](http://roadinstitute.org://pub/gsea/gene_sets/c5.all.v5.1.symbols.gmt)

### Number of permutations

1000

## Phenotype labels

```
e:\spring2014\GSEA\gender_label_word.cls#Male_versus_Female
```

### Collapse dataset to gene symbols

true

Permutation type

phenotype

Chip platform(s)

[broadinstitute.org/pub/gsea/annotations/GENE\\_SYMBOL.chip](https://broadinstitute.org/pub/gsea/annotations/GENE_SYMBOL.chip)

## Basic fields

Show

### Advanced fields

Show

## GSEA Report for Dataset DEG\_overall\_gender\_diff

### Enrichment in phenotype: Male (12 samples)

- 31 / 378 gene sets are upregulated in phenotype **Male**
- 0 gene sets are significant at FDR < 25%
- 0 gene sets are significantly enriched at nominal pvalue < 1%
- 1 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: Female (12 samples)

- 347 / 378 gene sets are upregulated in phenotype **Female**
- 4 gene sets are significant at FDR < 25%
- 4 gene sets are significantly enriched at nominal pvalue < 1%
- 5 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 3750 native features
- After collapsing features into gene symbols, there are: 2736 genes

### Gene set details

- Gene set size filters (min=15, max=500) resulted in filtering out 1076 / 1454 gene sets
- The remaining 378 gene sets were used in the analysis
- List of [gene sets used and their sizes](#) (restricted to features in the specified dataset)

### Gene markers for the Male *versus* Female comparison

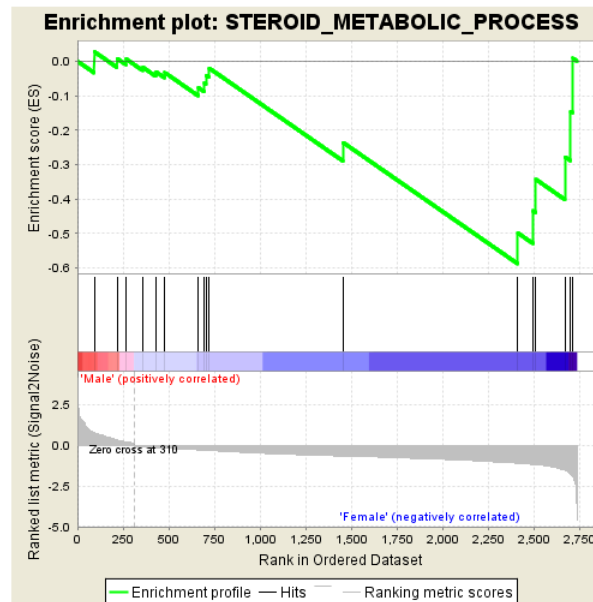
- The dataset has 2736 features (genes)
- # of markers for phenotype **Male**: 310 (11.3% ) with correlation area 10.8%
- # of markers for phenotype **Female**: 2426 (88.7% ) with correlation area 89.2%
- Detailed [rank ordered gene list](#) for all features in the dataset
- [Heat map and gene list correlation](#) profile for all features in the dataset
- [Butterfly plot](#) of significant genes

|    | GS<br>follow link to MSigDB                                    | GS<br>DETAILS               | SIZE | ES    | NES   | NOM<br>p-val | FDR<br>q-val | FWER<br>p-val | RANK<br>AT MAX | LEADING EDGE                         |
|----|--|-----------------------------|------|-------|-------|--------------|--------------|---------------|----------------|--------------------------------------|
| 1  | <a href="#">STEROID_METABOLIC_PROCESS</a>                      | <a href="#">Details ...</a> | 1    | -0.59 | -1.86 | 0.000        | 0.247        | 0.093         | 326            | tags=35%,<br>list=12%,<br>signal=40% |
| 2  | <a href="#">MONOCARBOXYLIC_ACID_METABOLIC_PROCESS</a>          | <a href="#">Details ...</a> | 27   | -0.47 | -1.84 | 0.004        | 0.146        | 0.107         | 644            | tags=37%,<br>list=24%,<br>signal=48% |
| 3  | <a href="#">CELLULAR_LIPID_METABOLIC_PROCESS</a>               | <a href="#">Details ...</a> | 64   | -0.45 | -1.81 | 0.000        | 0.136        | 0.138         | 575            | tags=31%,<br>list=21%,<br>signal=39% |
| 4  | <a href="#">FATTY_ACID_METABOLIC_PROCESS</a>                   | <a href="#">Details ...</a> | 19   | -0.59 | -1.76 | 0.008        | 0.184        | 0.217         | 644            | tags=47%,<br>list=24%,<br>signal=62% |
| 5  | <a href="#">LIPID_METABOLIC_PROCESS</a>                        | <a href="#">Details ...</a> | 78   | -0.39 | -1.65 | 0.014        | 0.410        | 0.415         | 675            | tags=29%,<br>list=25%,<br>signal=38% |
| 6  | <a href="#">CARBOXYLIC_ACID_METABOLIC_PROCESS</a>              | <a href="#">Details ...</a> | 51   | -0.32 | -1.45 | 0.090        | 1.000        | 0.696         | 696            | tags=24%,<br>list=25%,<br>signal=31% |
| 7  | <a href="#">ORGANIC_ACID_METABOLIC_PROCESS</a>                 | <a href="#">Details ...</a> | 51   | -0.32 | -1.45 | 0.090        | 1.000        | 0.696         | 696            | tags=24%,<br>list=25%,<br>signal=31% |
| 8  | <a href="#">LIPID_BIOSYNTHETIC_PROCESS</a>                     | <a href="#">Details ...</a> | 22   | -0.46 | -1.40 | 0.150        | 1.000        | 0.747         | 575            | tags=36%,<br>list=21%,<br>signal=46% |
| 9  | <a href="#">GOLGI_APPARATUS</a>                                | <a href="#">Details ...</a> | 30   | -0.43 | -1.39 | 0.076        | 1.000        | 0.753         | 379            | tags=37%,<br>list=14%,<br>signal=42% |
| 10 | <a href="#">GTPASE_ACTIVITY</a>                                | <a href="#">Details ...</a> | 15   | -0.50 | -1.38 | 0.092        | 1.000        | 0.761         | 195            | tags=33%,<br>list=7%,<br>signal=36%  |
| 11 | <a href="#">CELL_MIGRATION</a>                                 | <a href="#">Details ...</a> | 22   | -0.48 | -1.35 | 0.121        | 1.000        | 0.781         | 969            | tags=64%,<br>list=35%,<br>signal=98% |
| 12 | <a href="#">OXIDOREDUCTASE_ACTIVITY</a>                        | <a href="#">Details ...</a> | 66   | -0.27 | -1.34 | 0.119        | 1.000        | 0.783         | 284            | tags=12%,<br>list=10%,<br>signal=13% |
| 13 | <a href="#">TRANSMEMBRANE_RECEPTOR_PROTEIN_KINASE_ACTIVITY</a> | <a href="#">Details ...</a> | 15   | -0.53 | -1.34 | 0.105        | 1.000        | 0.794         | 795            | tags=67%,<br>list=29%,<br>signal=93% |

# GSEA Example

Table: GSEA Results Summary

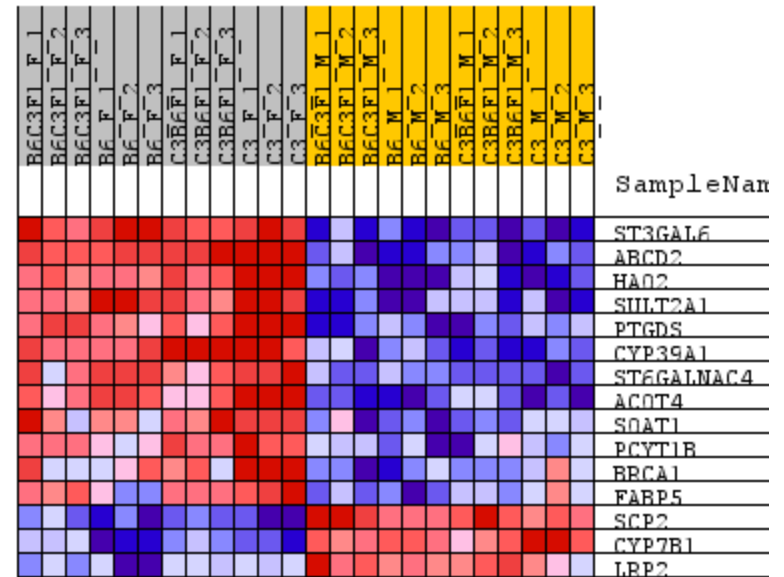
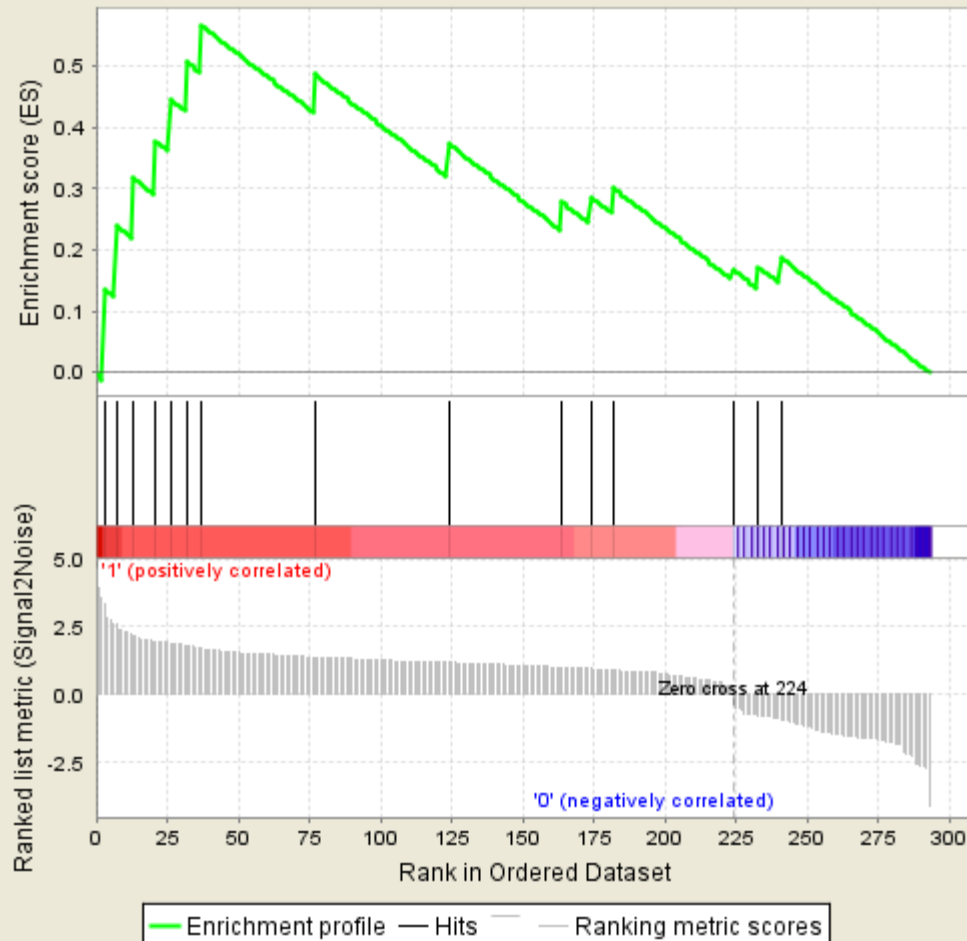
|                                   |  |
|-----------------------------------|--|
| Dataset                           | DEG_overall_gender_diff_overall_gender_diff_collapsed_to_symbols.gender_label_word.cls<br>#Male_versus_Female.gender_label_word.cls<br>#Male_versus_Female_repos |
| Phenotype                         | gender_label_word.cls#Male_versus_Female_repos   |
| Upregulated in class              | Female   |
| GeneSet                           | STEROID_METABOLIC_PROCESS  |
| Enrichment Score (ES)             | -0.5875557   |
| Normalized Enrichment Score (NES) | -1.8563647   |
| Nominal p-value                   | 0.0  |
| FDR q-value                       | 0.24662763   |
| FWER p-Value                      | 0.093  |



**Fig 1: Enrichment plot: STERIOD\_METABOLIC\_PROCESS**  
Profile of the Running ES Score & Positions of GeneSet Members on the Rank Ordered List

# Lipid metabolism enriched -- GSEA

Enrichment plot: LIPID\_METABOLIC\_PROCESS



- <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/>