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

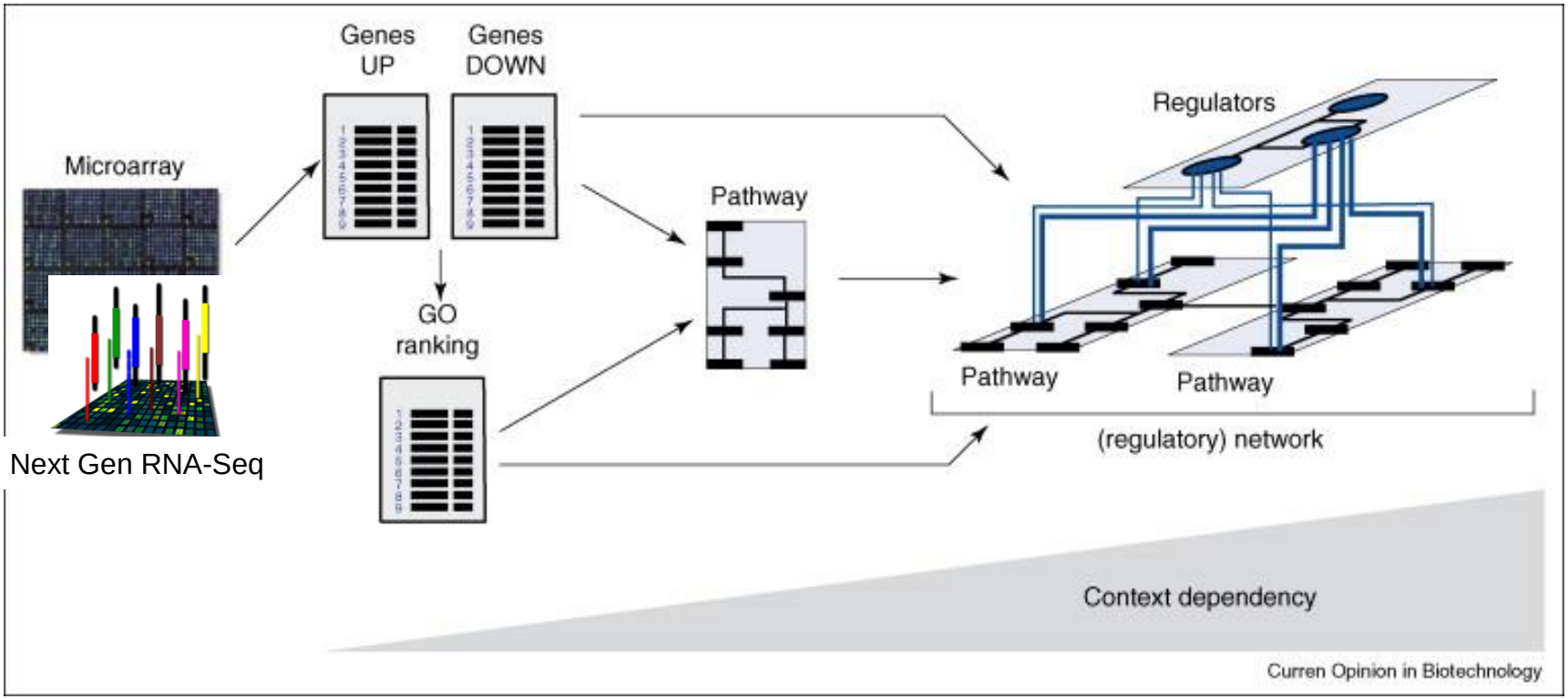
Pathway analysis

Biostat & Bioinfo short course series

7/25/2016

Jianying Li

The Road to Pathway Analysis

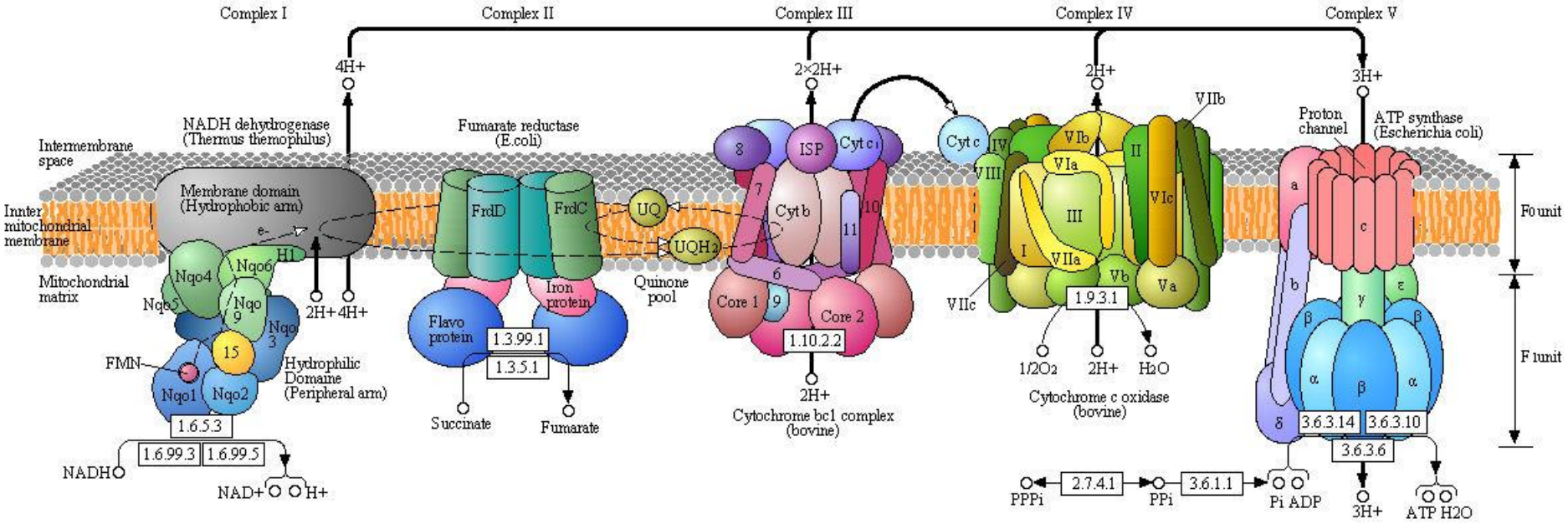


Adapted from Werner (2008), Current Opinion in Biotech., 19:50-64

What are biological pathways??

- Definition in Wikipedia:
  - "... a series of actions among molecules in a cell that leads to a certain product or a change in a cell."
- Pathway types:
  - Metabolic pathway
  - Ontologies
  - Gene regulation pathways
  - Signal transduction pathways
  - ...

The oxidative phosphorylation



http://employees.csbsju.edu/hjakubowski/classes/ch331/oxphos/olcouplingoxphos.html

Why pathway analysis?

- Genes are associated and respond collectively toward perturbation or condition changes
- Signal obtained from a collection of genes working together reflects a bigger picture
- PA offers apparent advantage over single-gene analysis
  - Different samples from a common condition may have different key genes that are all driving changes in the same pathway
  - Pathways may be perturbed by subtle changes in many genes
  - A small list of enriched pathways may be easier to interpret than a large list of associated genes
  - Pathways results from different studies may overlap better than single-gene results

Why pathway analysis –cont.?

- Modern technologies offer large scale (landscape) measurement
  - MicroArray gene expression
  - RNASeq
  - Other genomics measurement
- Well curated knowledgebase and information collection provide rich basis

Well curated knowledgebase

- GeneOntology
  - 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
  - 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,

Molecular Signatures Database (MSigDB)

- 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.

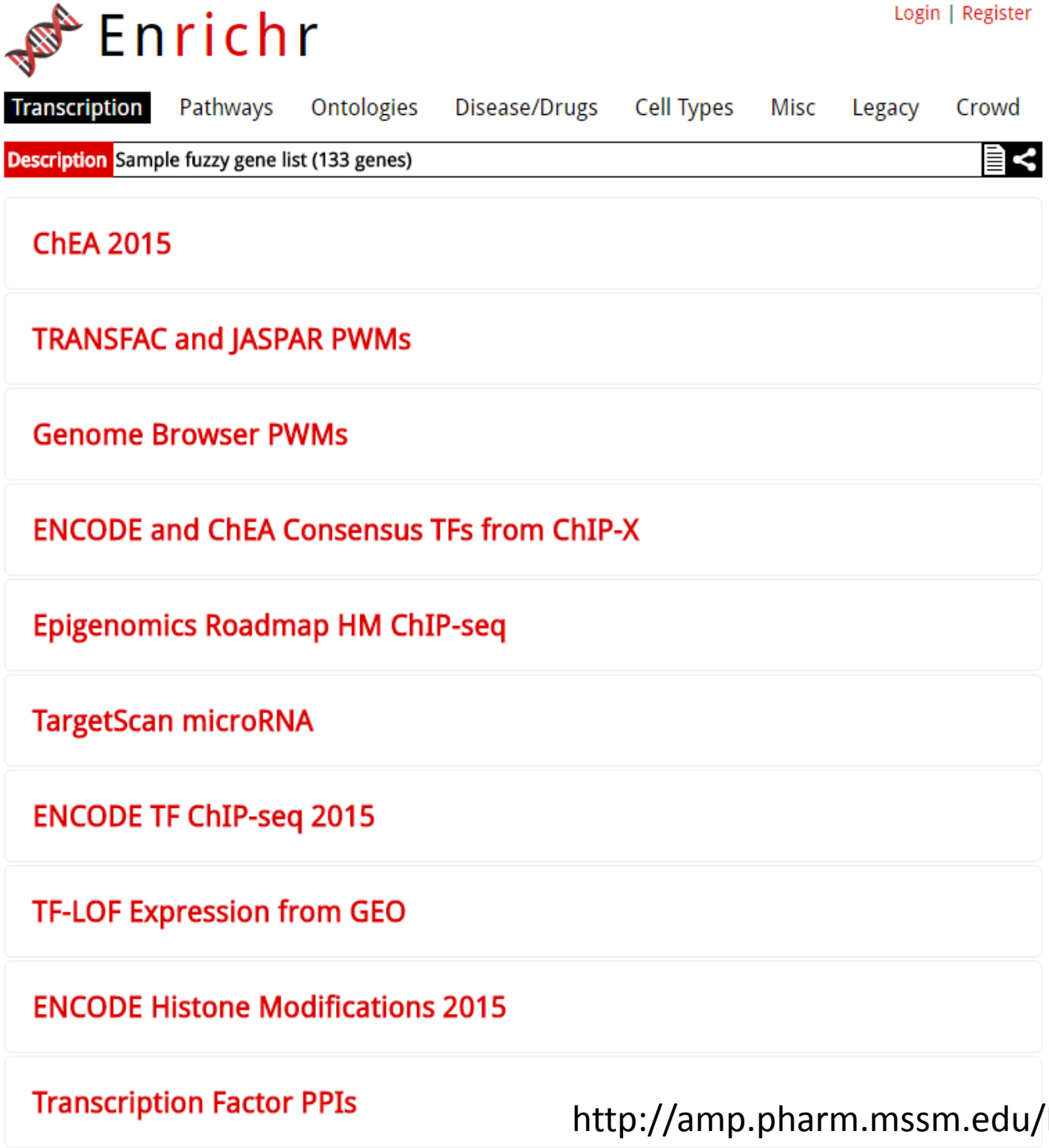
C3 **motif gene sets** based on conserved cis-regulatory motifs from a comparative analysis of the human, mouse, rat, and dog genomes.

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.

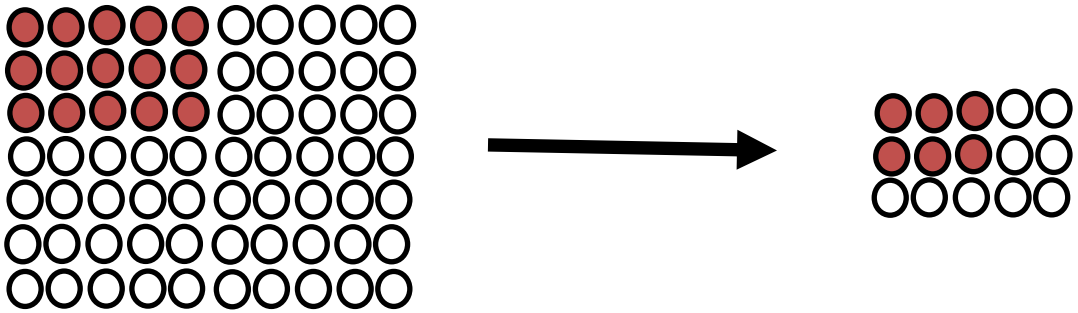


http://amp.pharm.mssm.edu/Enrichr/enrich

Why pathway analysis –cont.?

- 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.

Hypergeometric Test (Right-tailed)



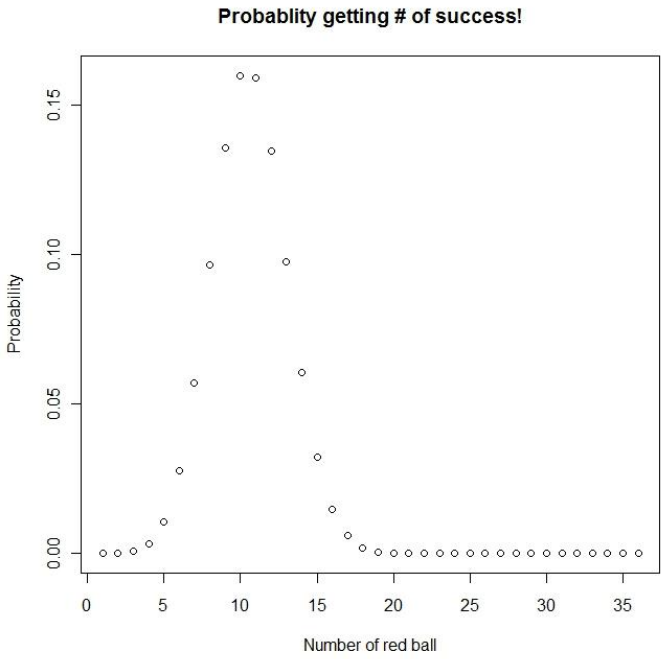
- An urn with two types of marbles:
  - **N total # of marbles**
  - **Of which, m # of red marbles**
  - **Drawing a red marble is a success!**
  - **Drawing a white marble is a failure!**
- n is the # of marbles randomly drawn
- k is the # of successes (red marbles) in the sample
- Hypergeometric distribution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

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

for $k = 0,1,2,\ldots,n$
$k<=m,\ n-k<=N-m$

Hypergeometric distribution

- Number of red ball: 50 ← m
- Number of white ball: 120 } N
- 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)



http://www.aiaccess.net/English/Glossaries/GlosMod/e_gm_hypergeometric.htm

Pathway analysis (an analogy)

- 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
- total N genes – **N**, of which **m** genes belong to a **pathway; k** genes were identified as differentially expressed, **x** genes belong to this "**pathway**"
- Do you think it is significant at $\alpha=0.01$? In other words "is this GO category enriched by/significant this microarray experiment"?
- 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
- The answer is simple, which is to perform a hypergeometric test: ***phyper(4, 20, 500, 40, lower.tail=FALSE) = 0.01371***

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

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

Ingenuity Pathway Analysis (IPA)

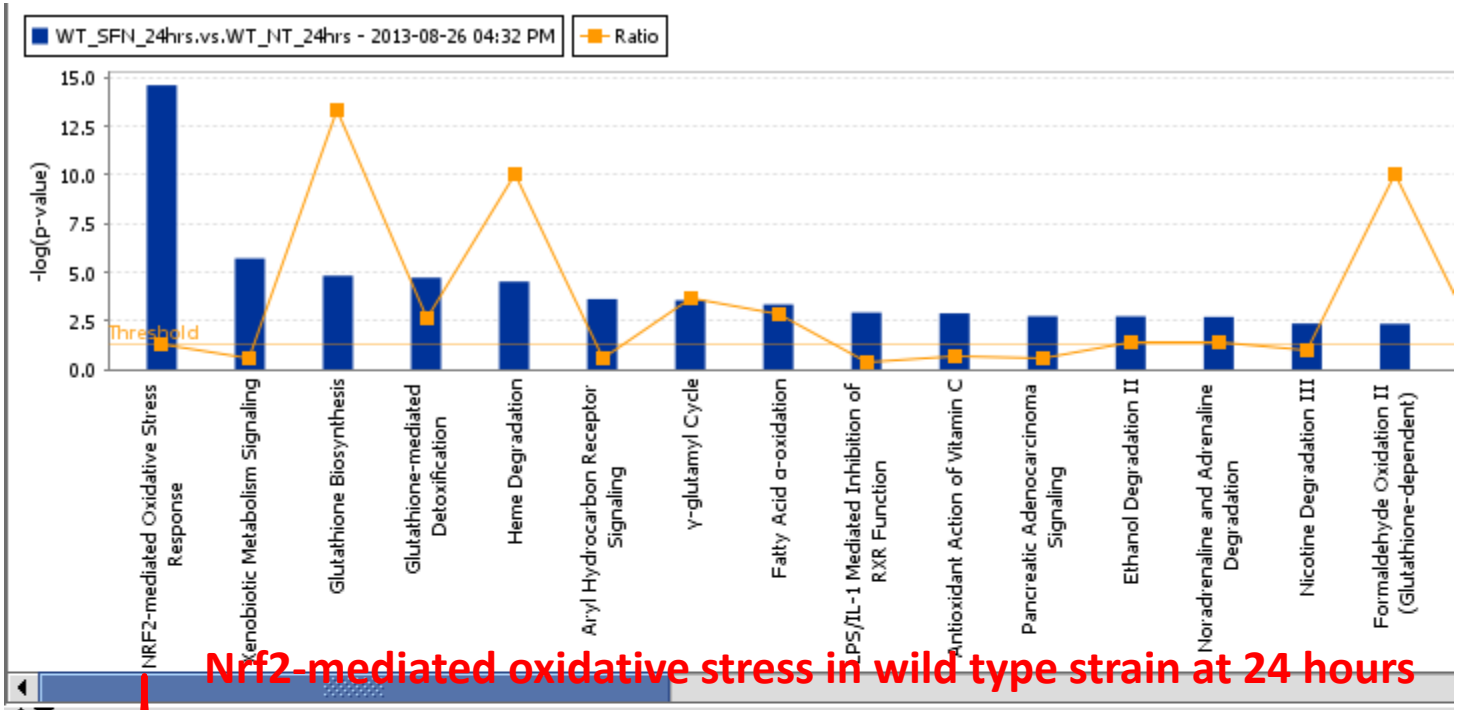
--Knowledgebase

- 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 molecule(s) associated with **NRF2-mediated Oxidative Stress Response** at **WT_SFN_24hrs.vs.WT_NT_24hrs - 2013-08-26 04:32 PM**

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

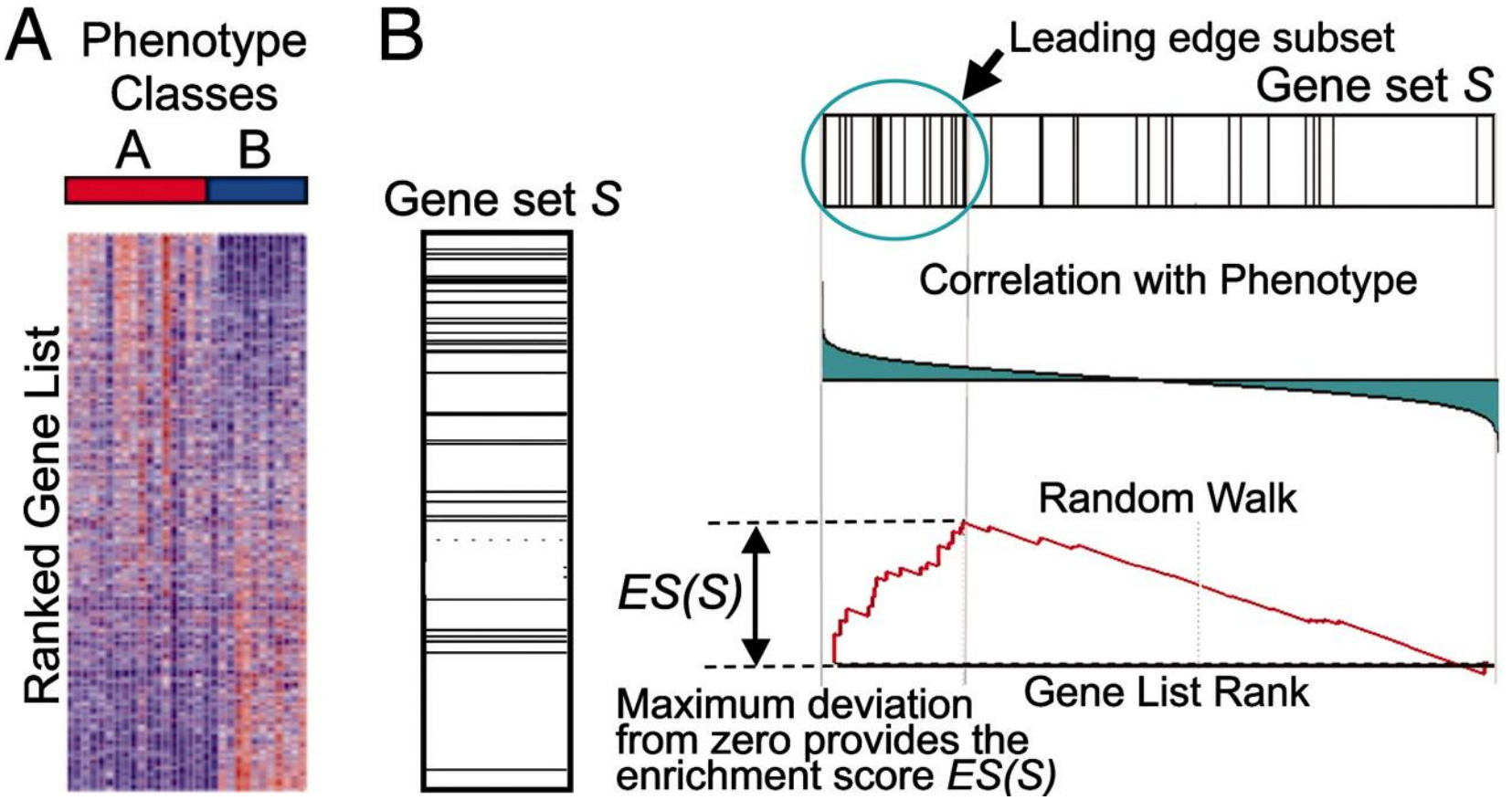
Symbol	Entrez Gene Name	Identifier	Exp Val		
		Affymetrix	p-value	False Discovery Rate	Fold Change
ABCC1	ATP-binding cassette, sub-	1421378_s_at	8.68E-07	6.63E-05	↑2.054
ABCC4	ATP-binding cassette, sub-	1443870_at	3.60E-15	2.68E-12	↑2.962
AOX1	aldehyde oxidase 1	1419435_at	2.15E-11	6.91E-09	↑2.756
EPHX1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

All the DEGs in the pathway are up-regulated

A non-parametric approach

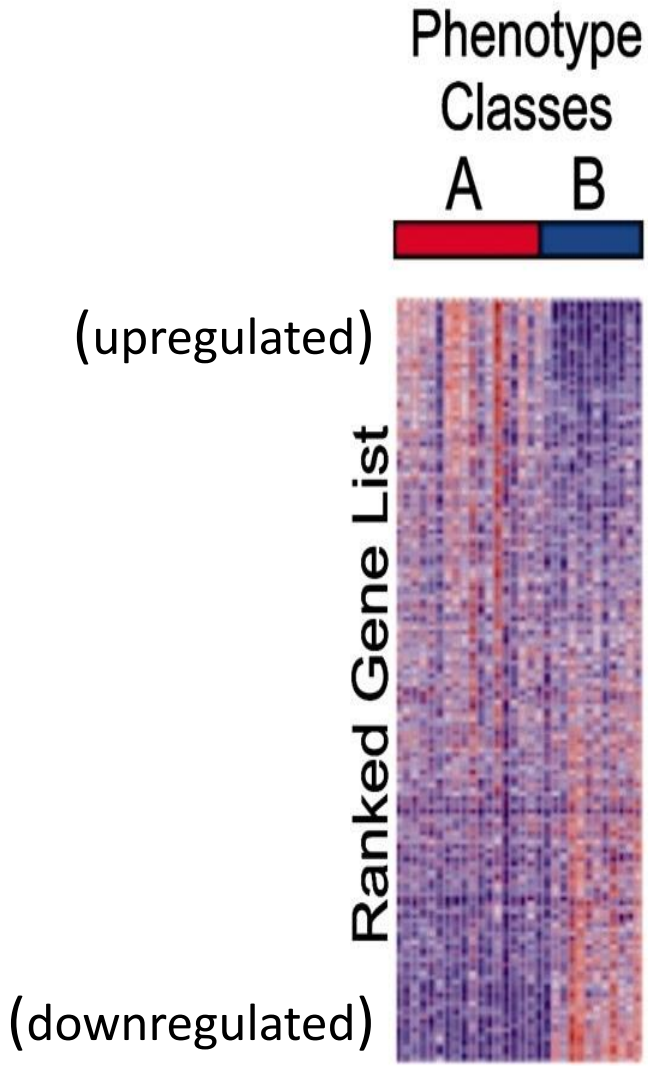
- Based on Kolmogorov-Smirnov (K-S) test
- Compare a sample empirical distribution function to the reference distribution
- Or, compare two sample empirical distribution functions
- Required data:
  - Ranked list of genes sorted by differential expression (includes all genes)
  - A gene set
- If the 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 GSEA overview illustrating the method.

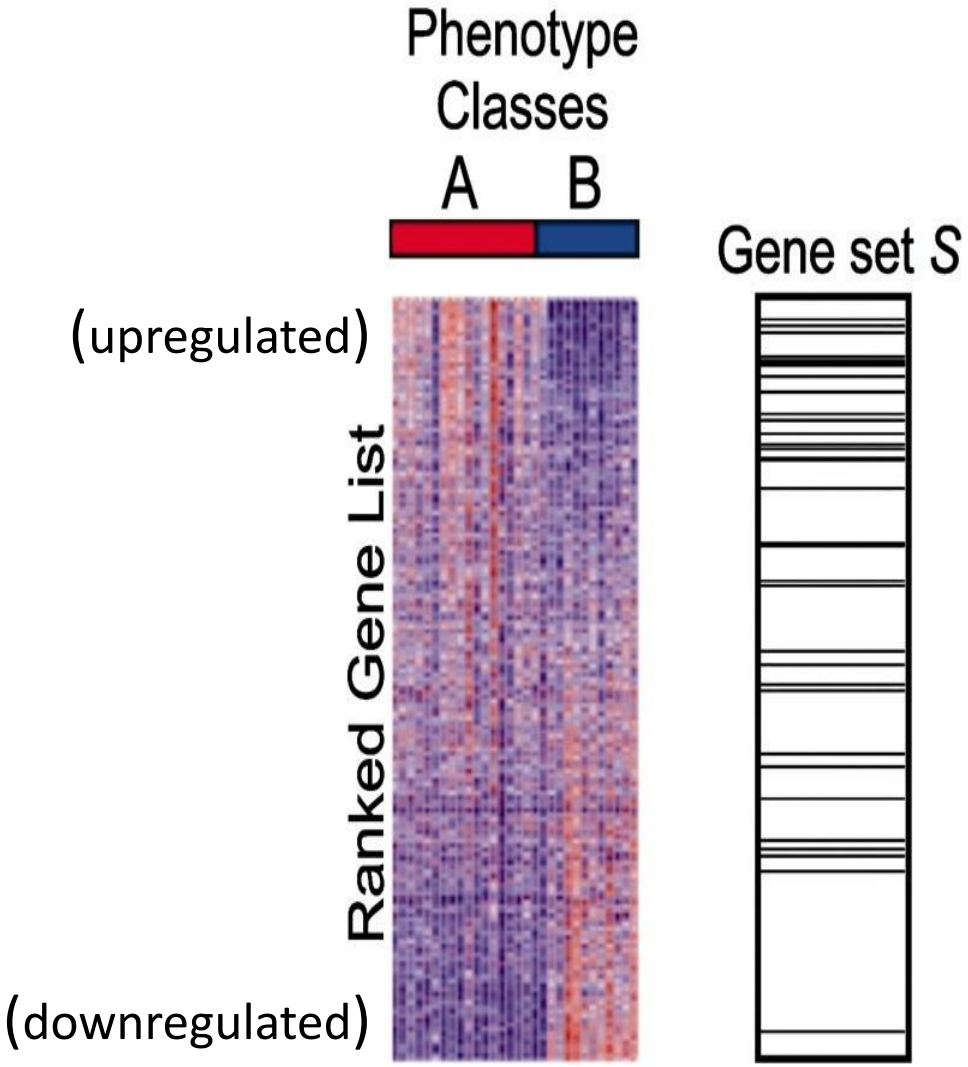


Subramanian A et al. PNAS 2005;102:15545-15550

Kolmogorov–Smirnov (KS) Test

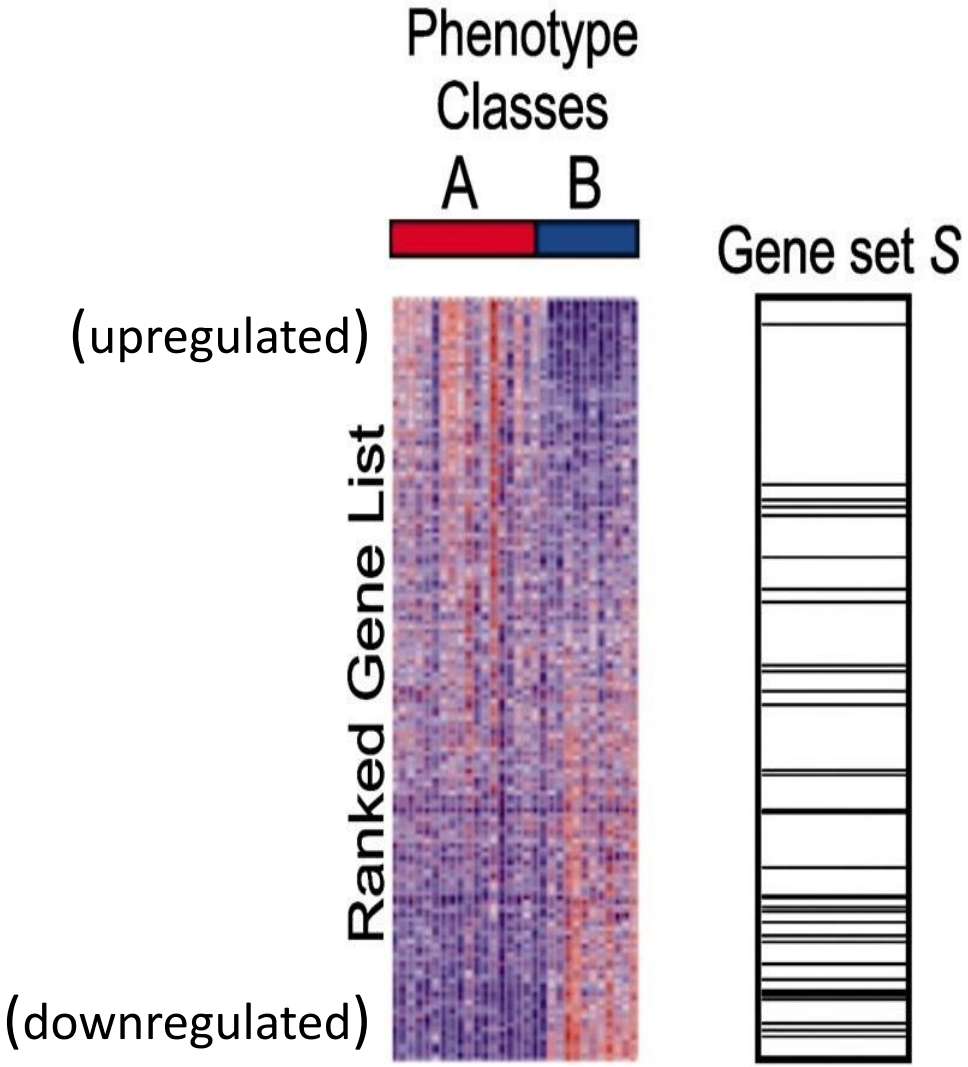


Kolmogorov–Smirnov (KS) Test



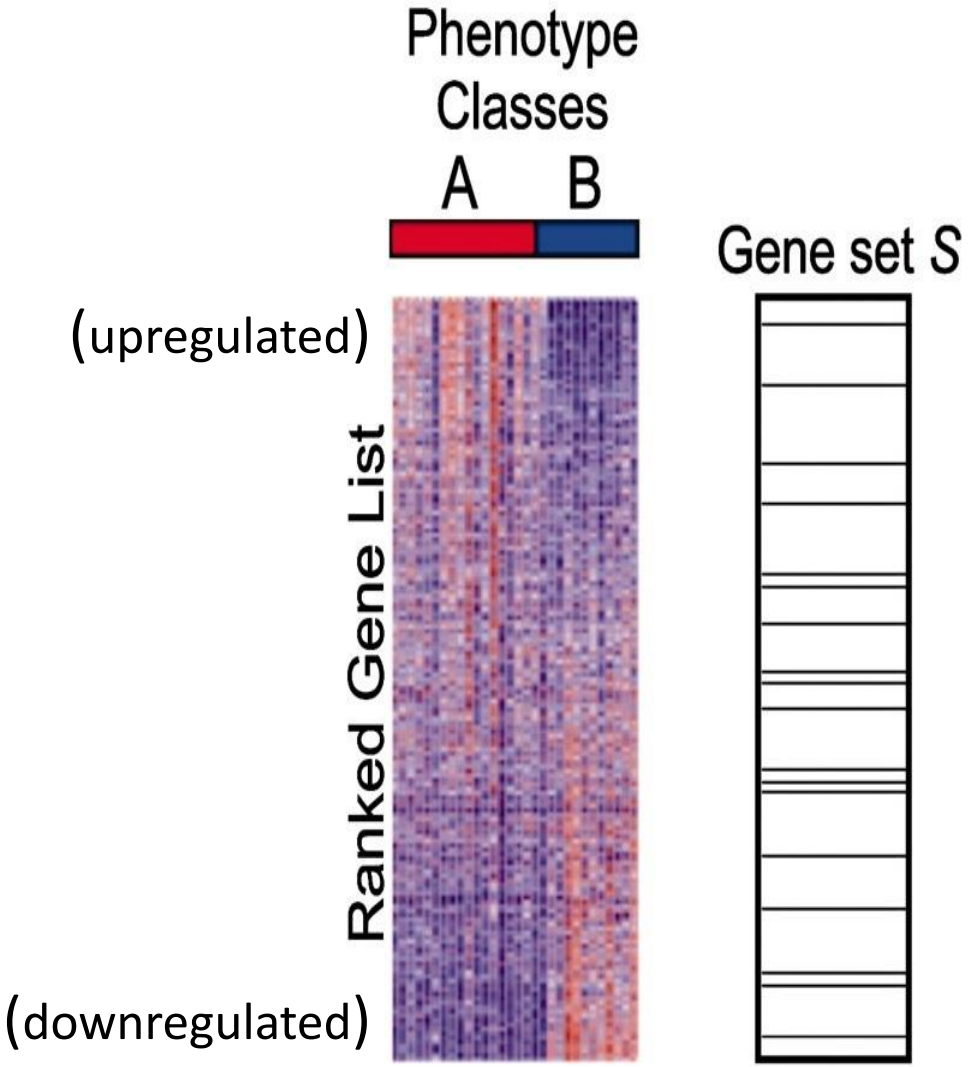
Large, positive enrichment score

Kolmogorov–Smirnov (KS) Test



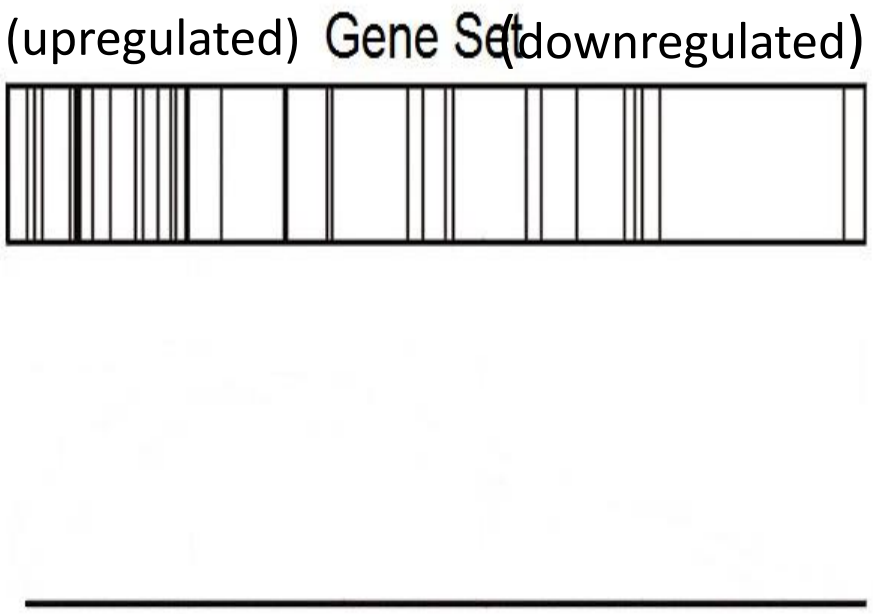
Large, negative enrichment score

Kolmogorov–Smirnov (KS) Test

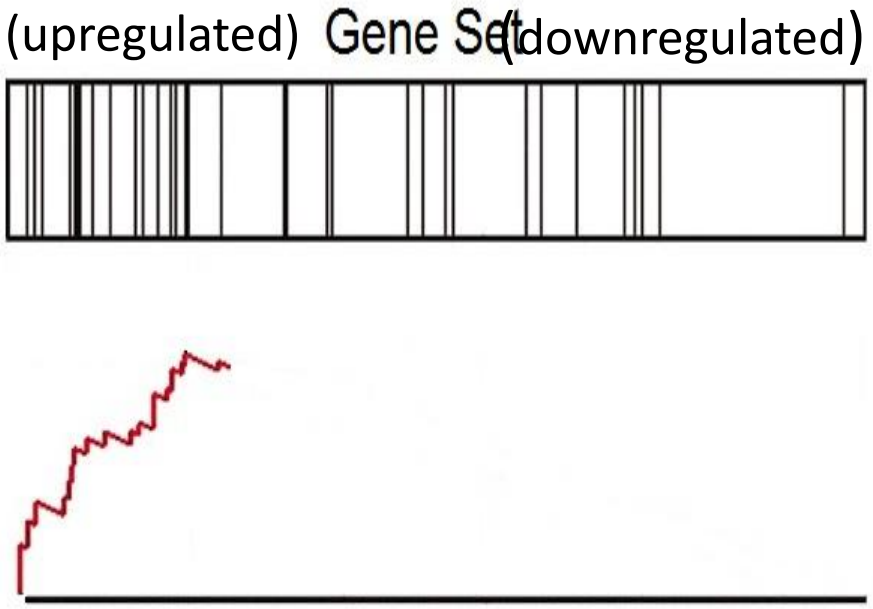


Near zero
enrichment score

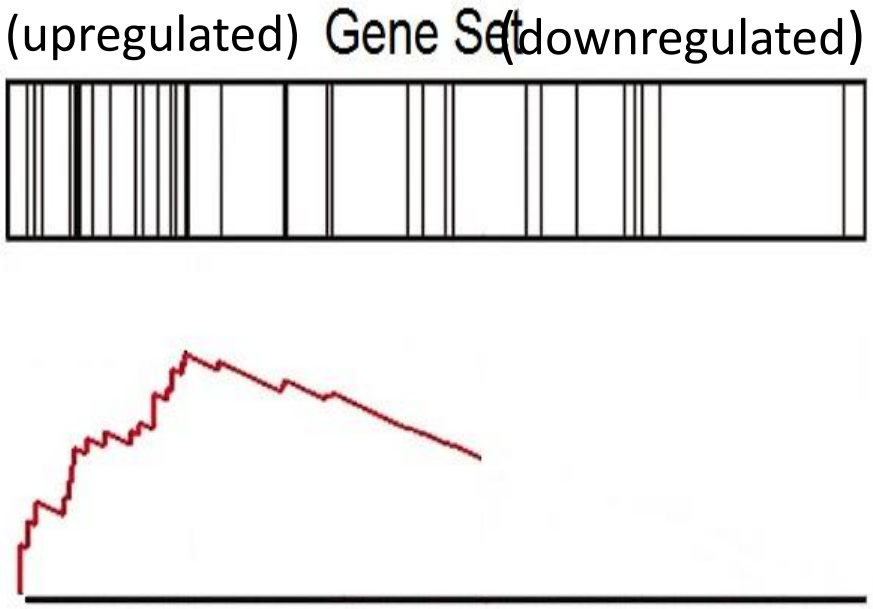
Kolmogorov–Smirnov (KS) Test



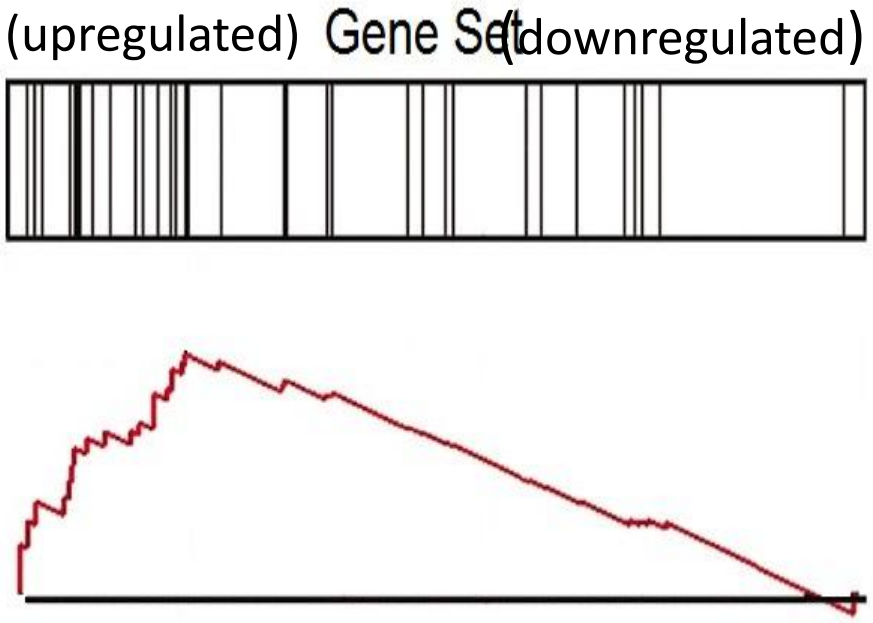
Kolmogorov–Smirnov (KS) Test



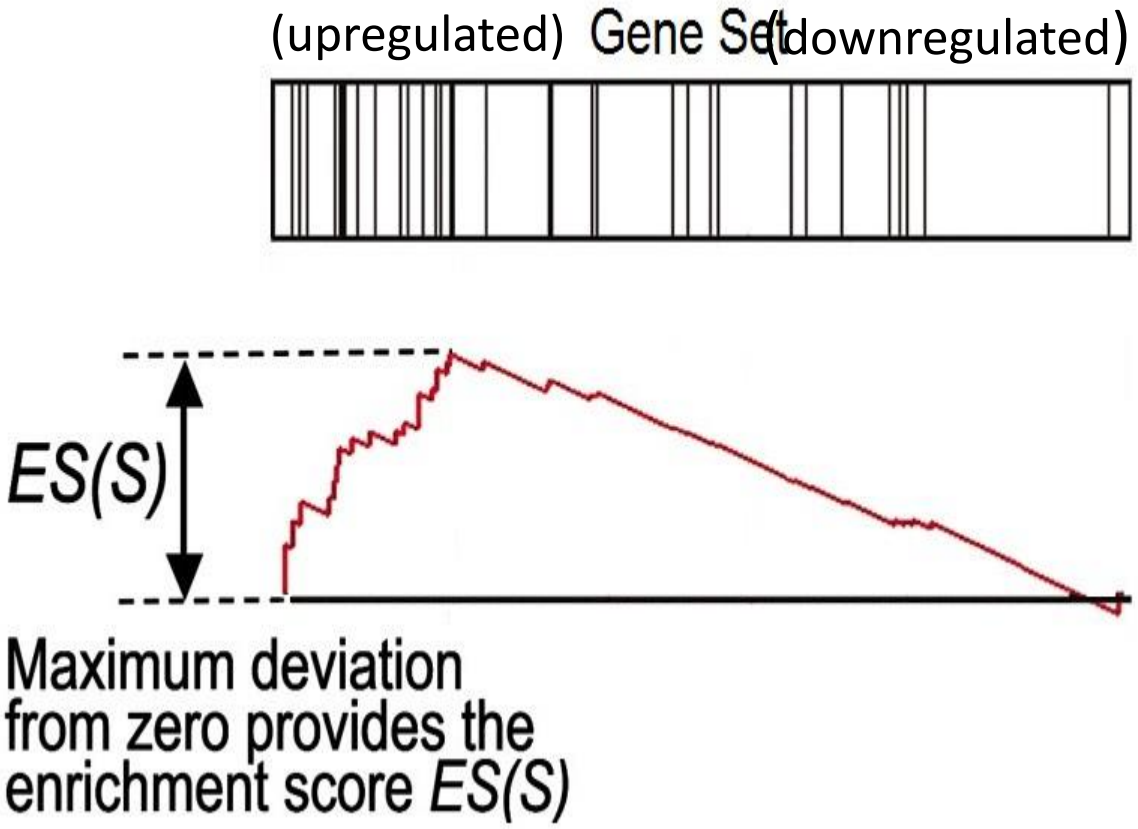
Kolmogorov–Smirnov (KS) Test



Kolmogorov–Smirnov (KS) Test

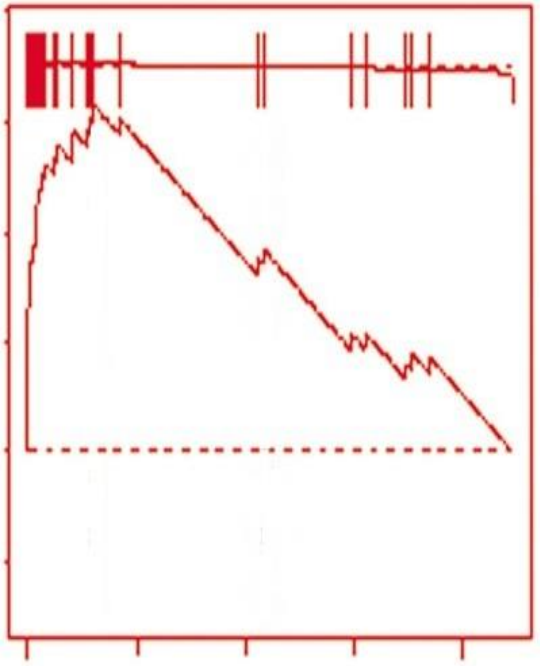


Kolmogorov–Smirnov (KS) Test



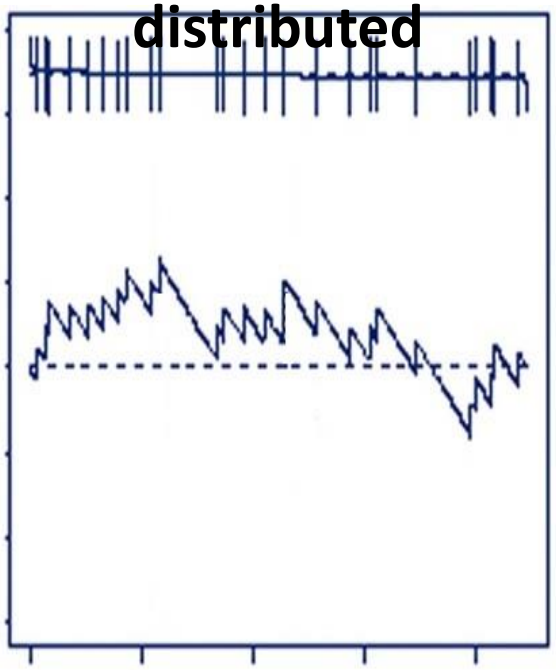
Kolmogorov–Smirnov (KS) Test

Many genes upregulated



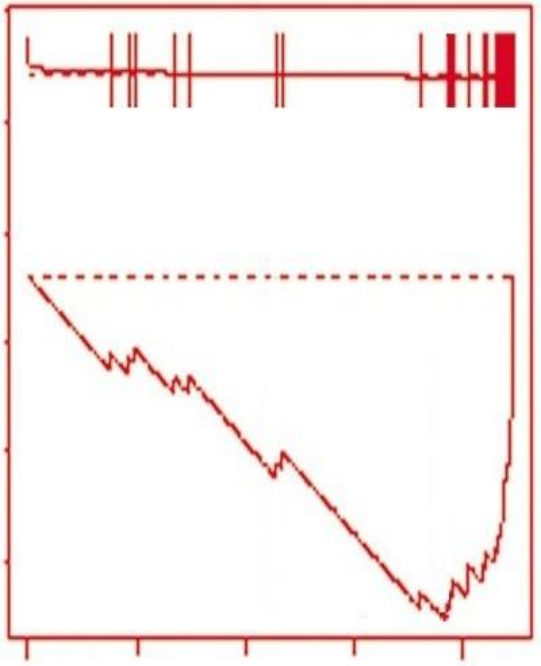
Large positive enrichment score

Genes randomly distributed

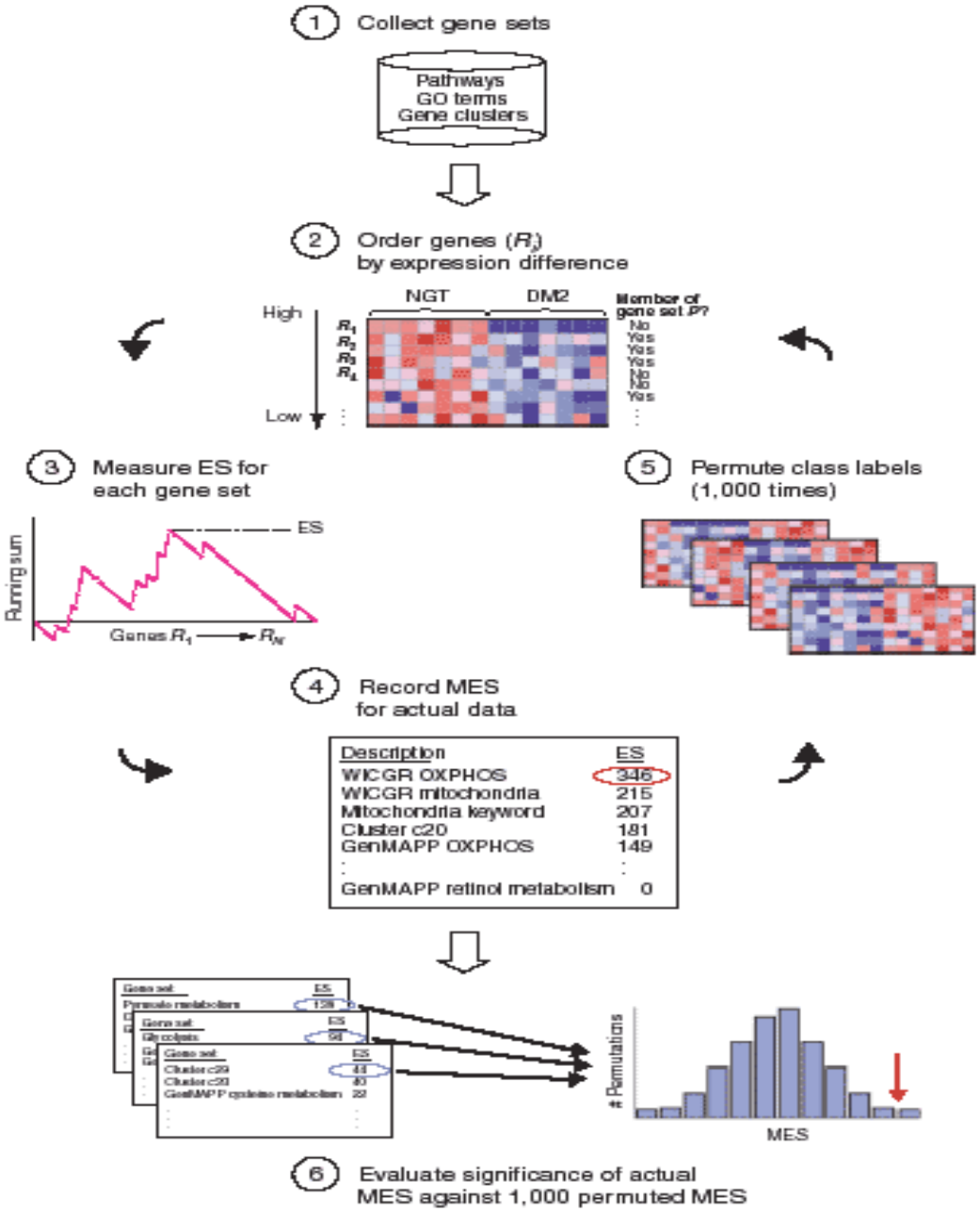


Near zero enrichment score

Many genes downregulated



Large negative enrichment score



Mootha et al., Nature Genetics, 2003, 34(3):267-273

Additional Processing

- Apply a multiple comparison correction (FDR)
- Interpret p-values cautiously
- Filter out gene sets that are too small or too big
- Explore key drivers in significant pathways
- Experimentally validate or follow up on significant gene sets

Other Types of Pathway Analysis

- Other data types (genotype, copy number, etc.)
- Other types of gene lists (genes from proteomics, ChIP targets, etc.)
- Custom gene sets (genes of interest, DEGs from another analysis, etc.)

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

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

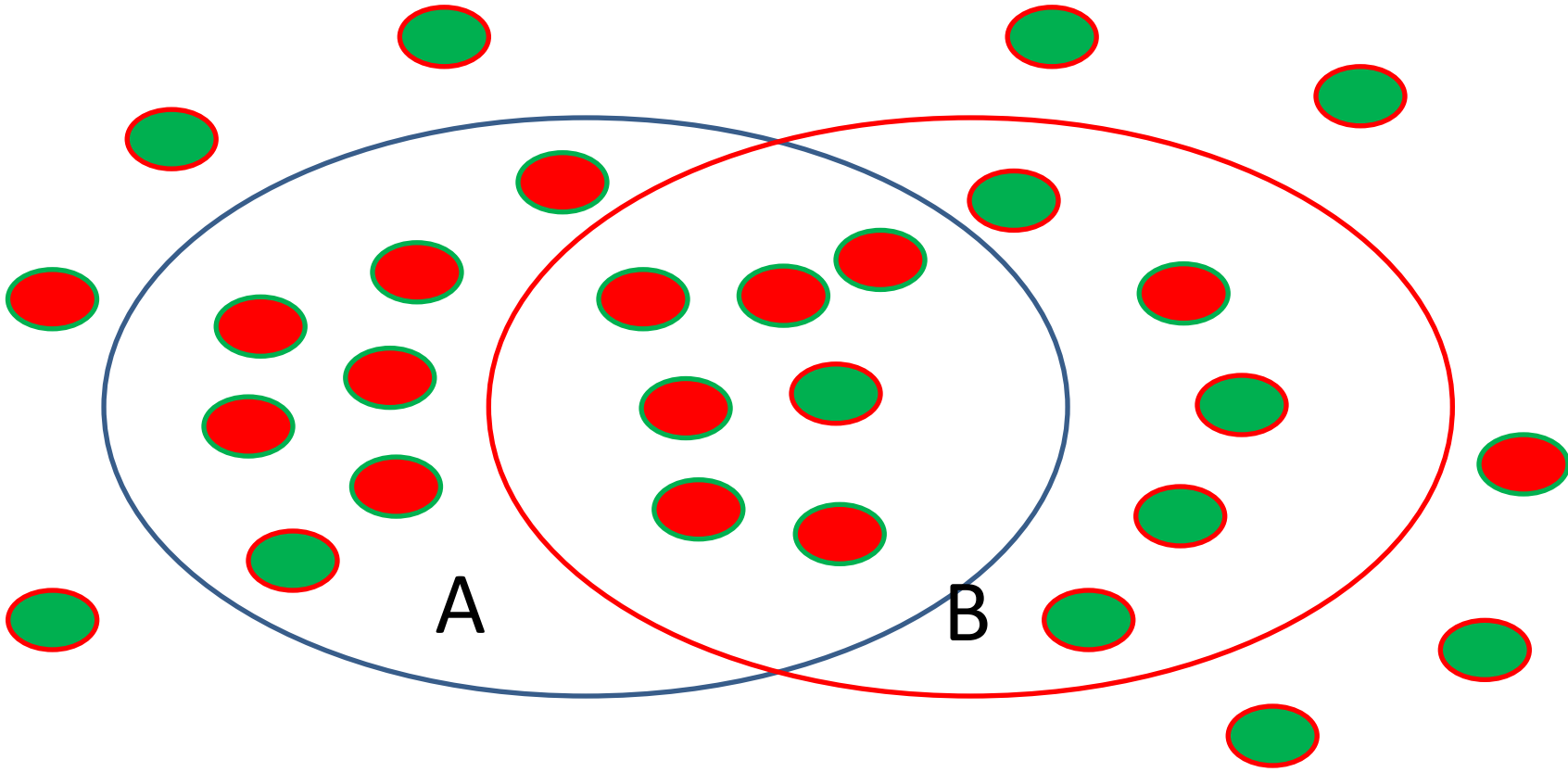
p=0.049

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

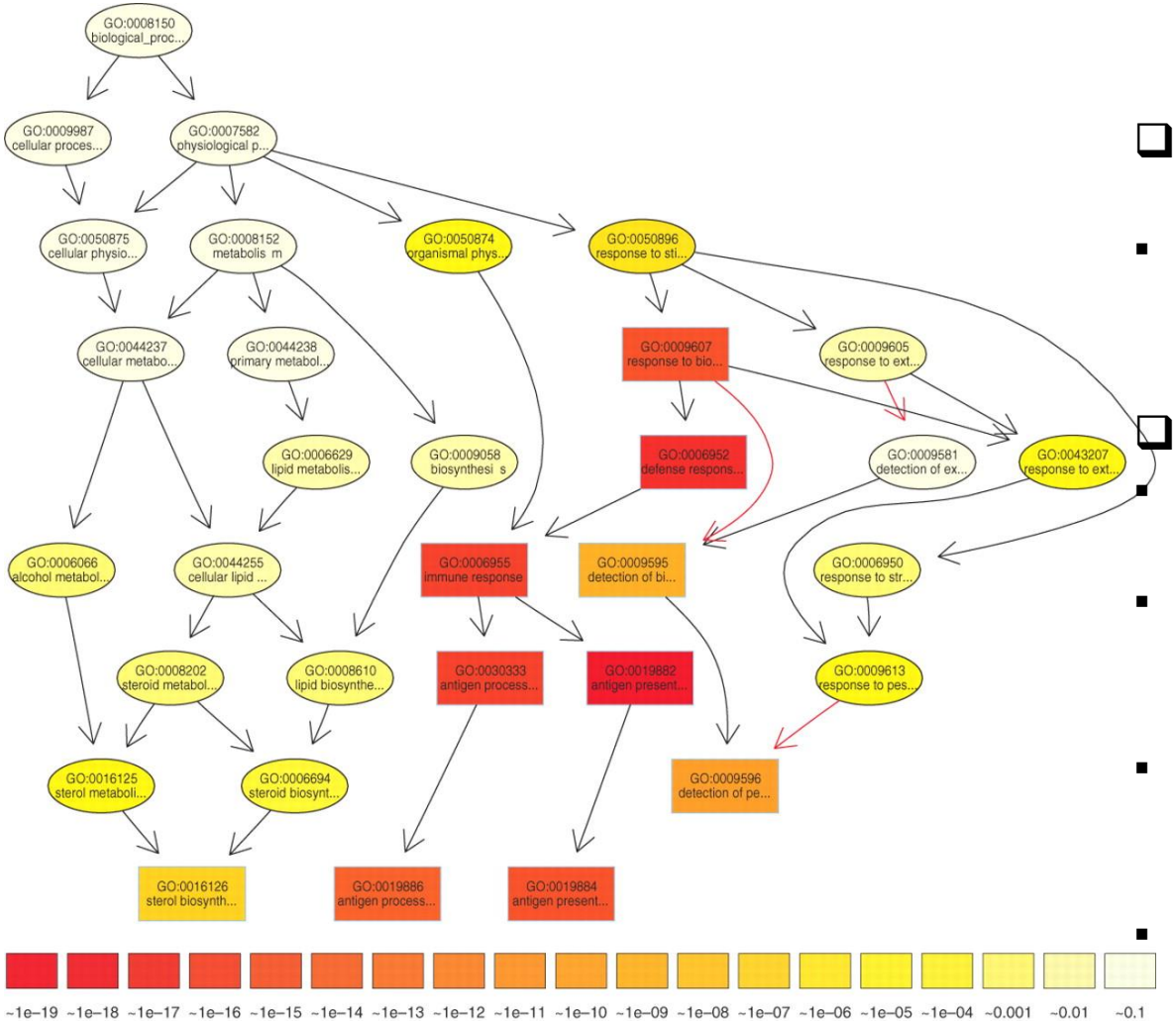
p=0.0048

Sets are overlapped



Set B is enriched only because of it 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.



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

❑ TopGO's elimination algorithm

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

❑ TopGO's weight algorithm

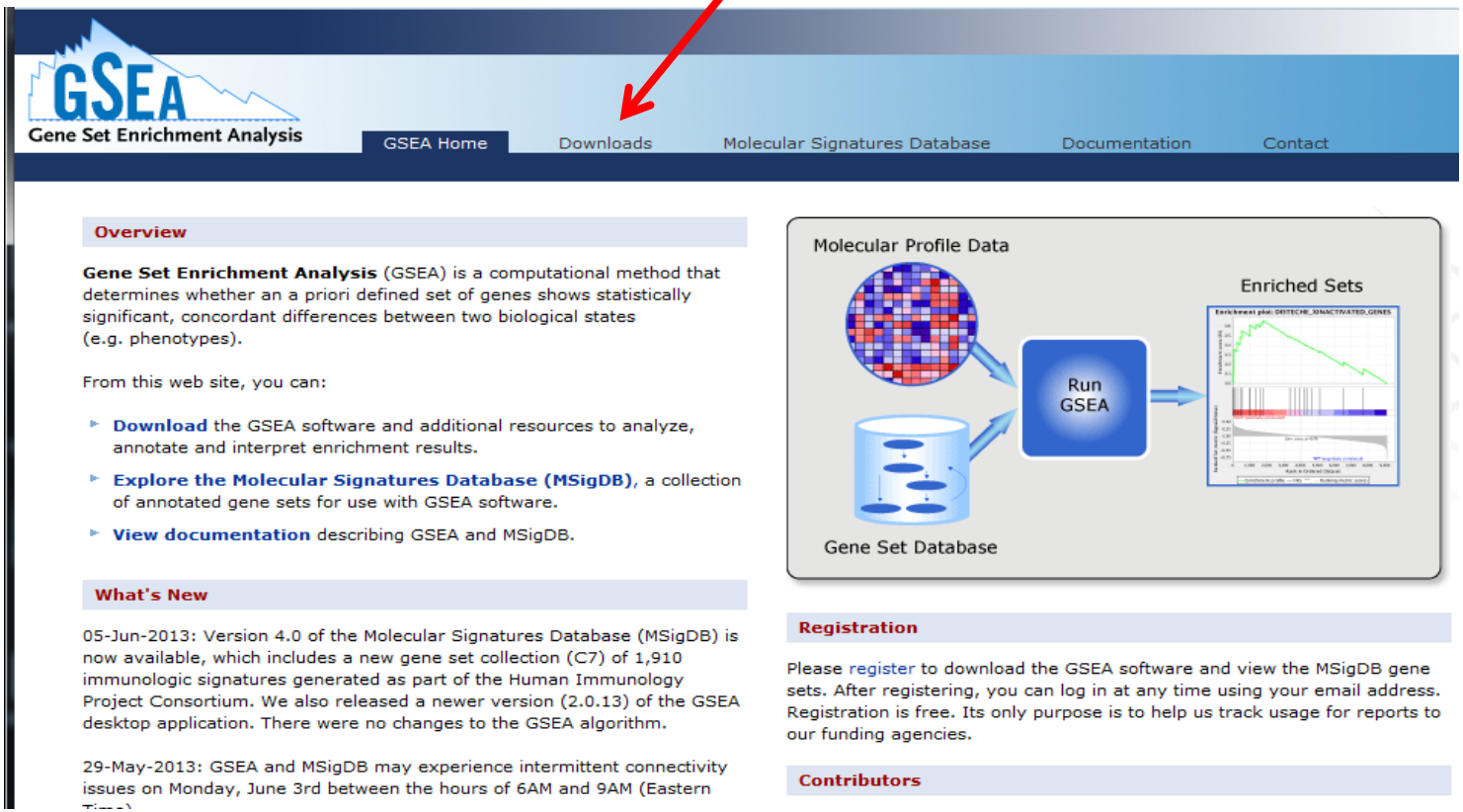
- The genes are weighted by their relevance in the significant nodes.
- The enrichment score of a parent (gene node u) is compared with the scores of its children.
- Children with a better score than u represent the interesting genes better. Therefore, their significance is increased
- Children with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

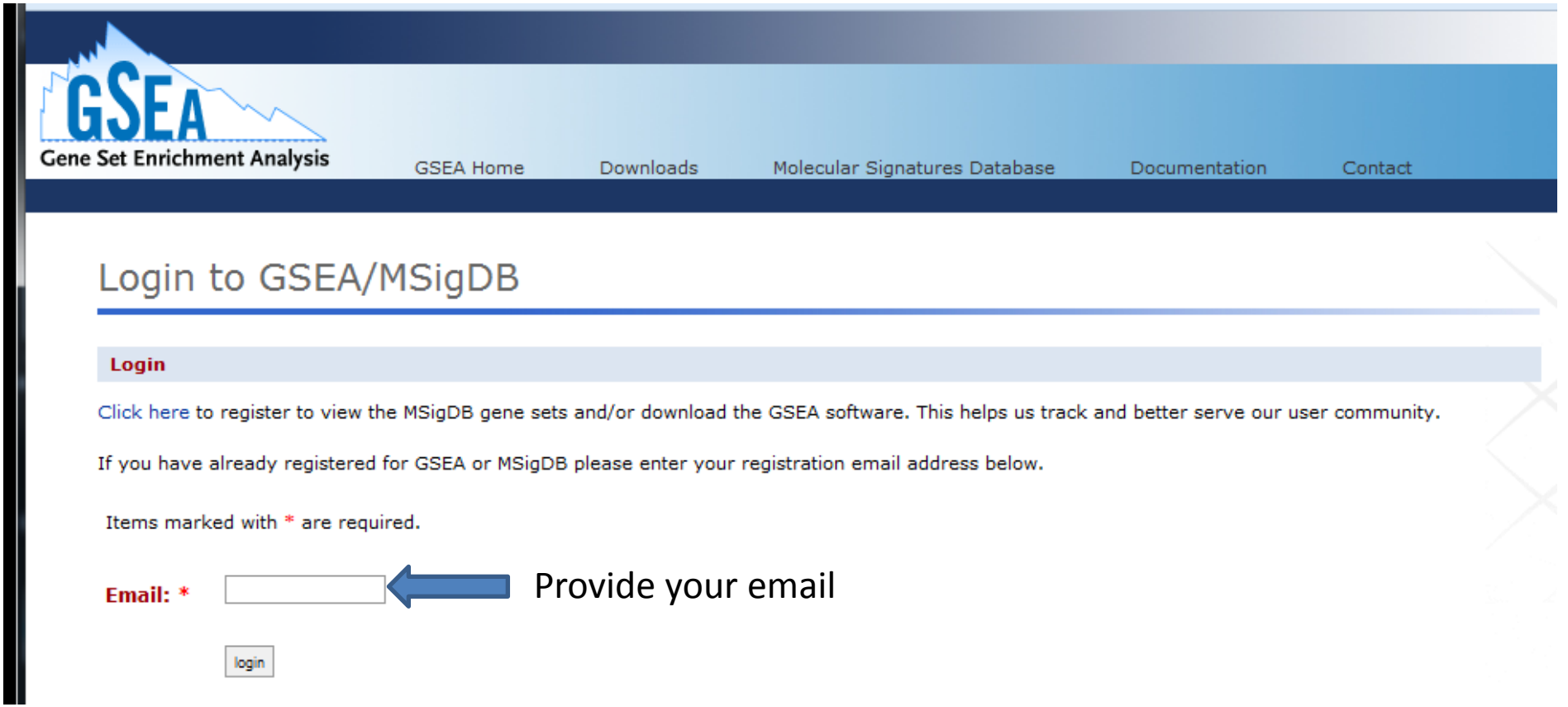
Bioinformatics

Hands on practice on both parametric and non-parametric

NOW, IT IS YOUR TURN

Click to download




GSEA
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.

javaGSEA Desktop Application	▸ Easy-to-use graphical user interface ▸ Runs on any desktop computer (Windows, Mac OS X, Linux etc.) that supports Java 6 or 7 ▸ Produces richly annotated reports of enrichment results ▸ Integrated gene sets browser to view gene set annotations, search for gene sets and ma… gene sets between platforms	Launch with 1GB (for 32 or 64-bit Java) memory: Launch
javaGSEA Java Jar file	▸ Command line usage ▸ Runs on any platform that supports Java 6 or 7 ▸ We recommend using the 'Launch' buttons above instead of this mode for most users	download gsea2-2.2.0.jar
R-GSEA	▸ Usage from within the R programming environment	download

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

Steps in GSEA

1. What you need for GSEA

- Expression data set
- Phenotype annotation
- Gene sets – use MSigDB or your own gene sets

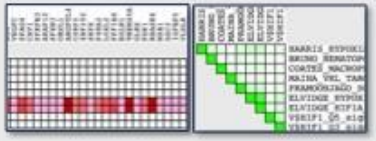
2. Run GSEA

- 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

Getting Help

GSEA web site:
www.broadinstitute.org/gsea

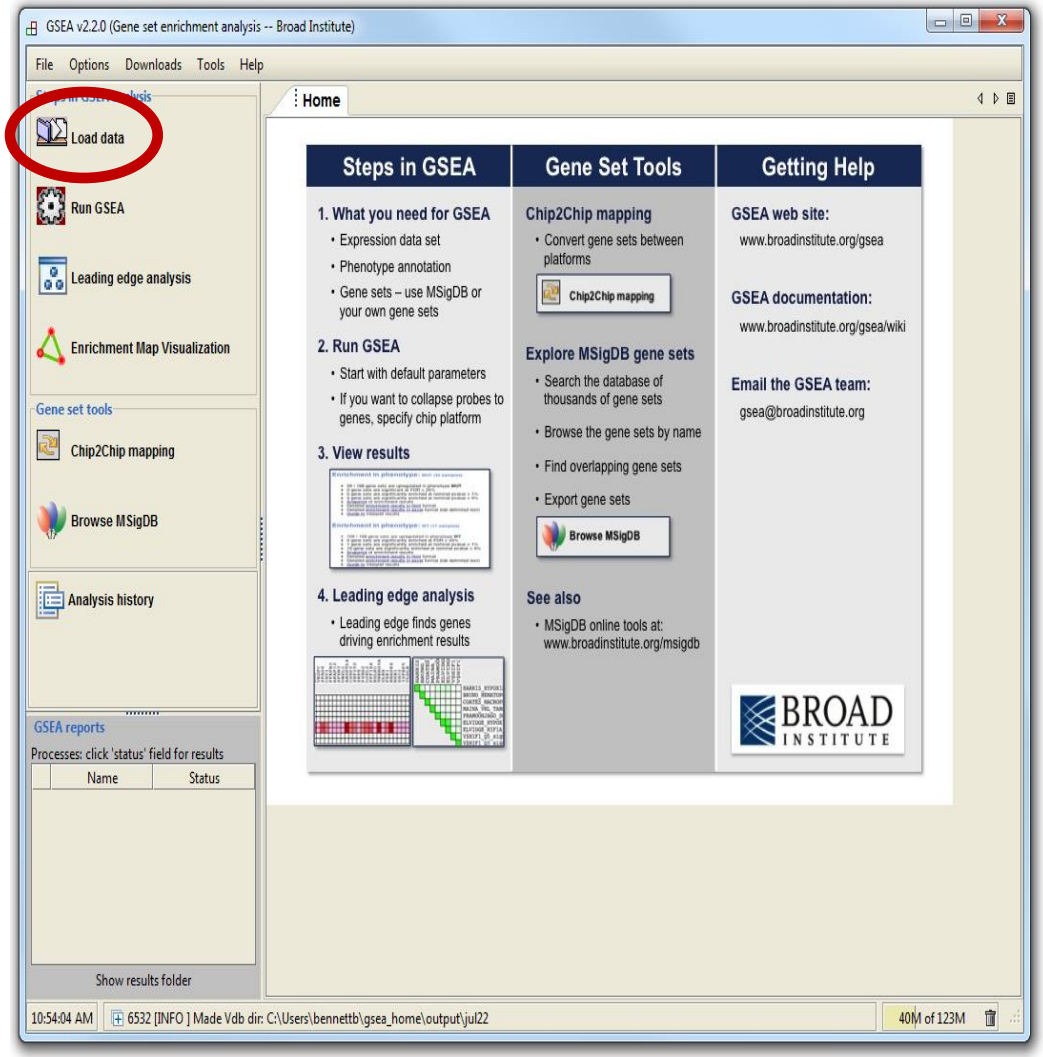
GSEA documentation:
www.broadinstitute.org/gsea/wiki

Email the GSEA team:
gsea@broadinstitute.org

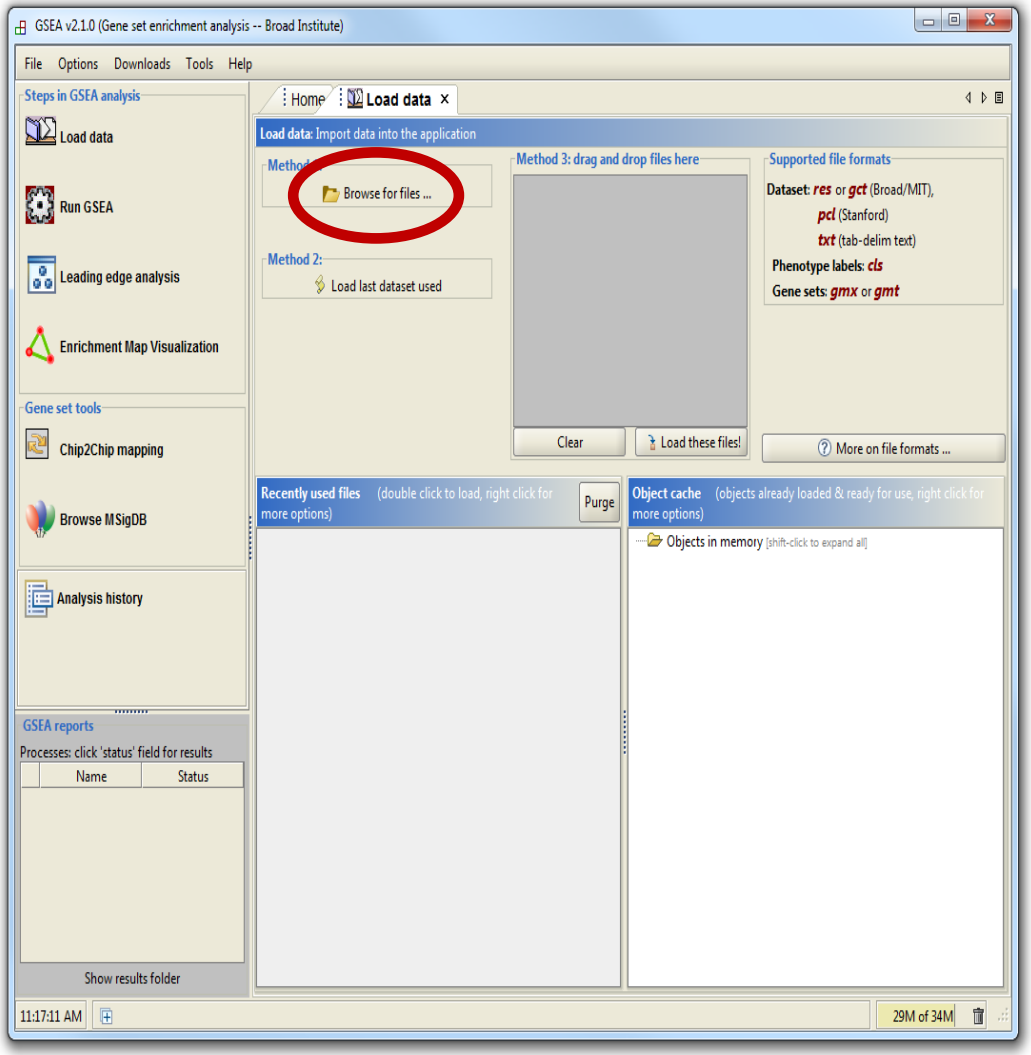
BROAD INSTITUTE

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

GSEA Example



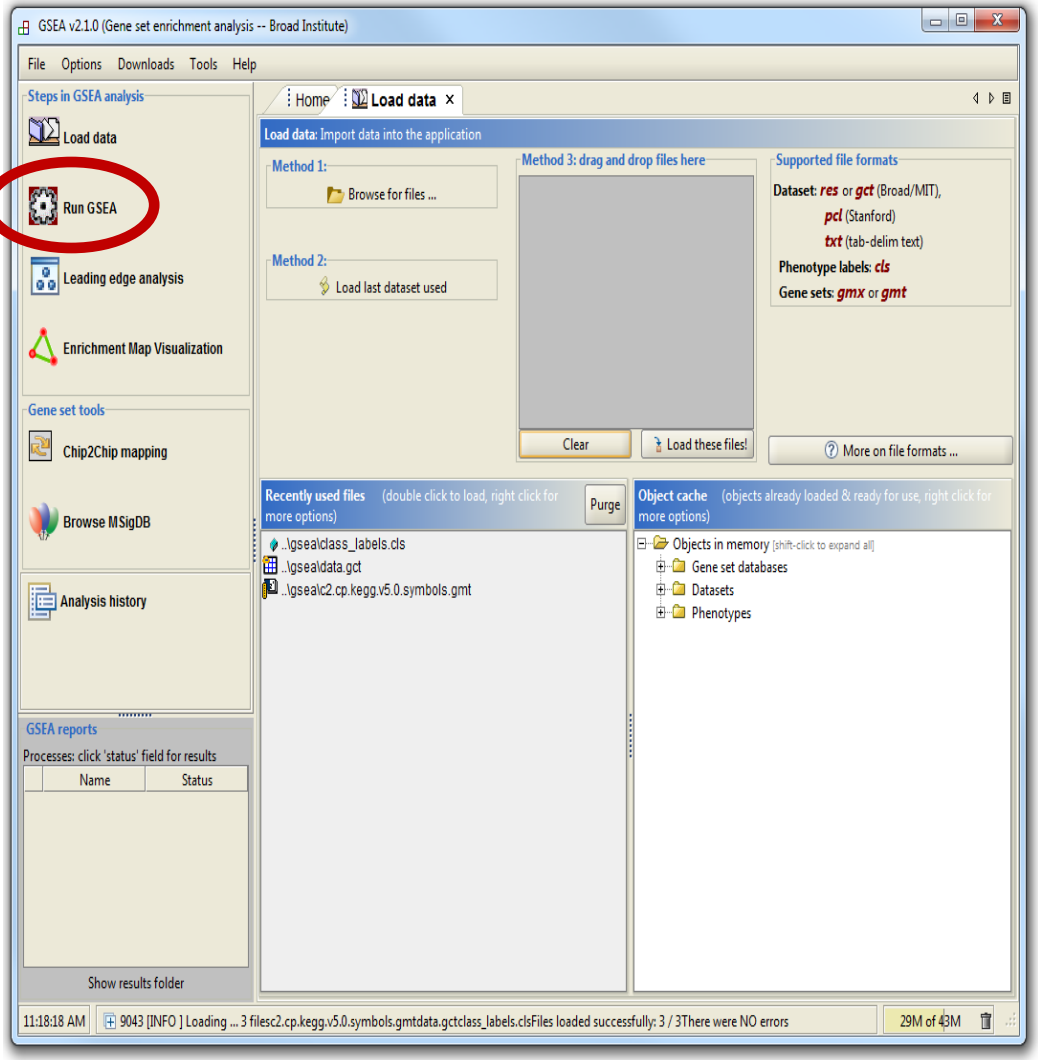
GSEA Example



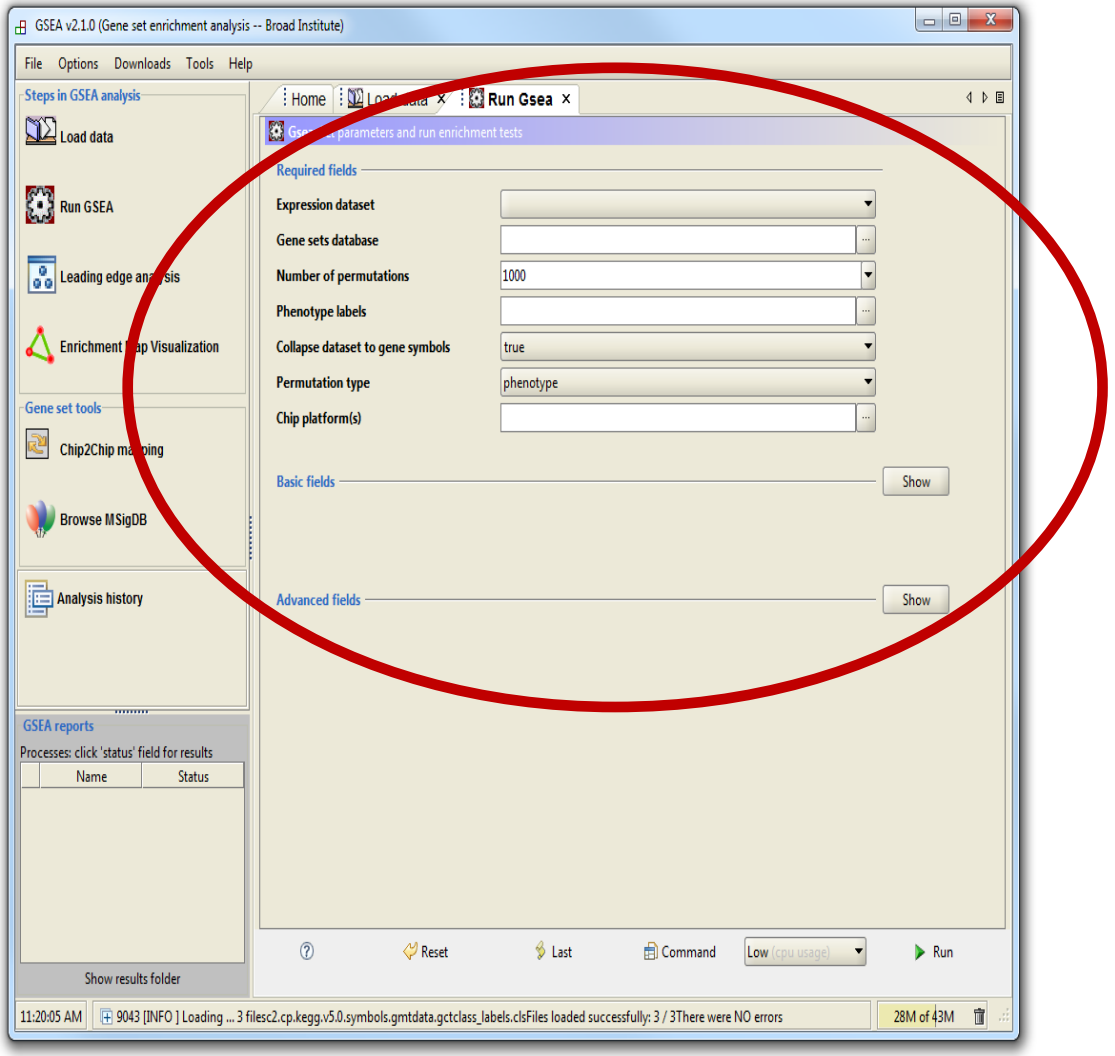
GSEA Example

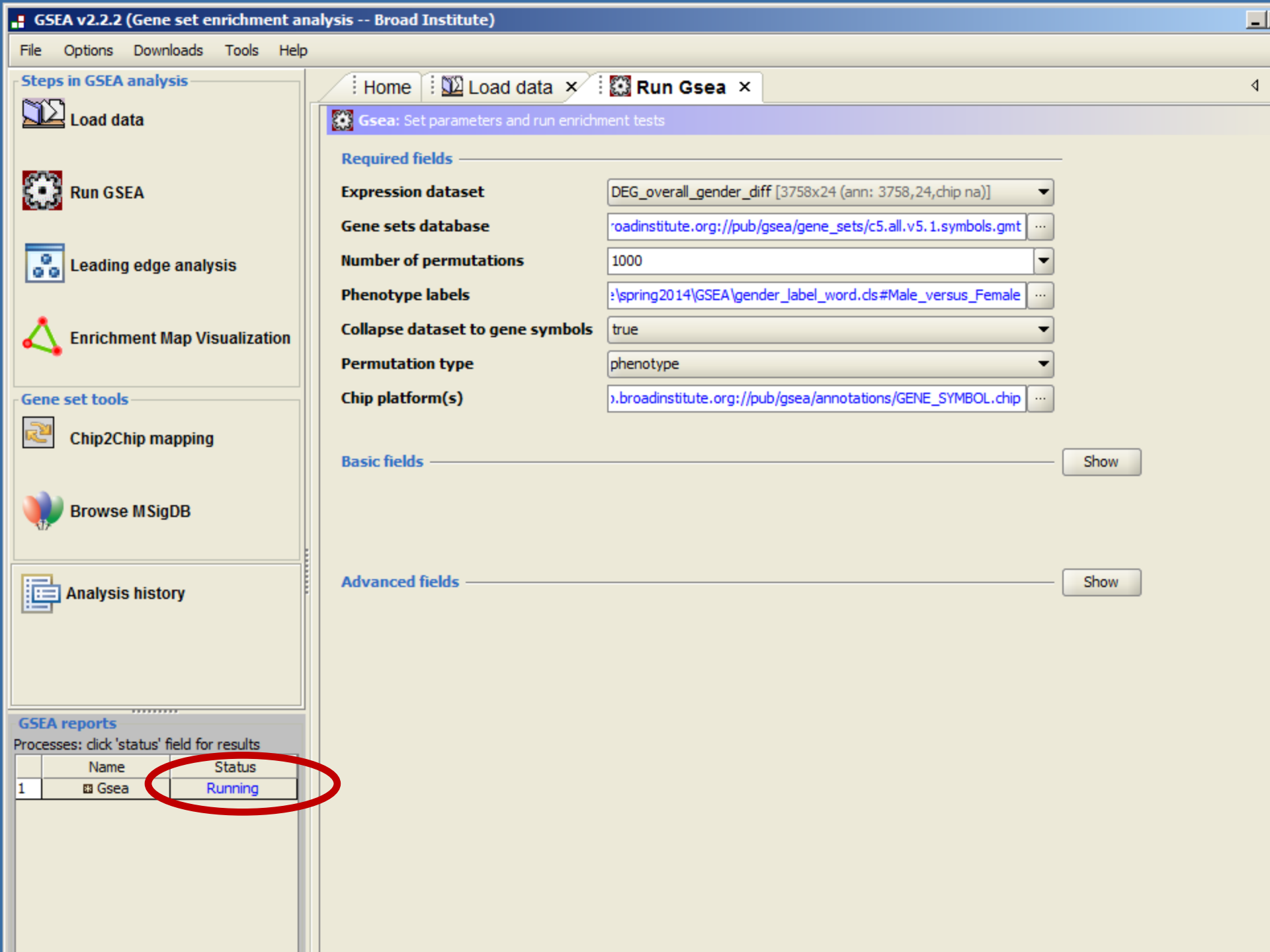
- Supported file types:
  - http://www.broadinstitute.org/cancer/software/gsea/wiki/index.php/Data_formats
- Required:
  - Expression data file
  - Class label file
- Optional:
  - Gene set file

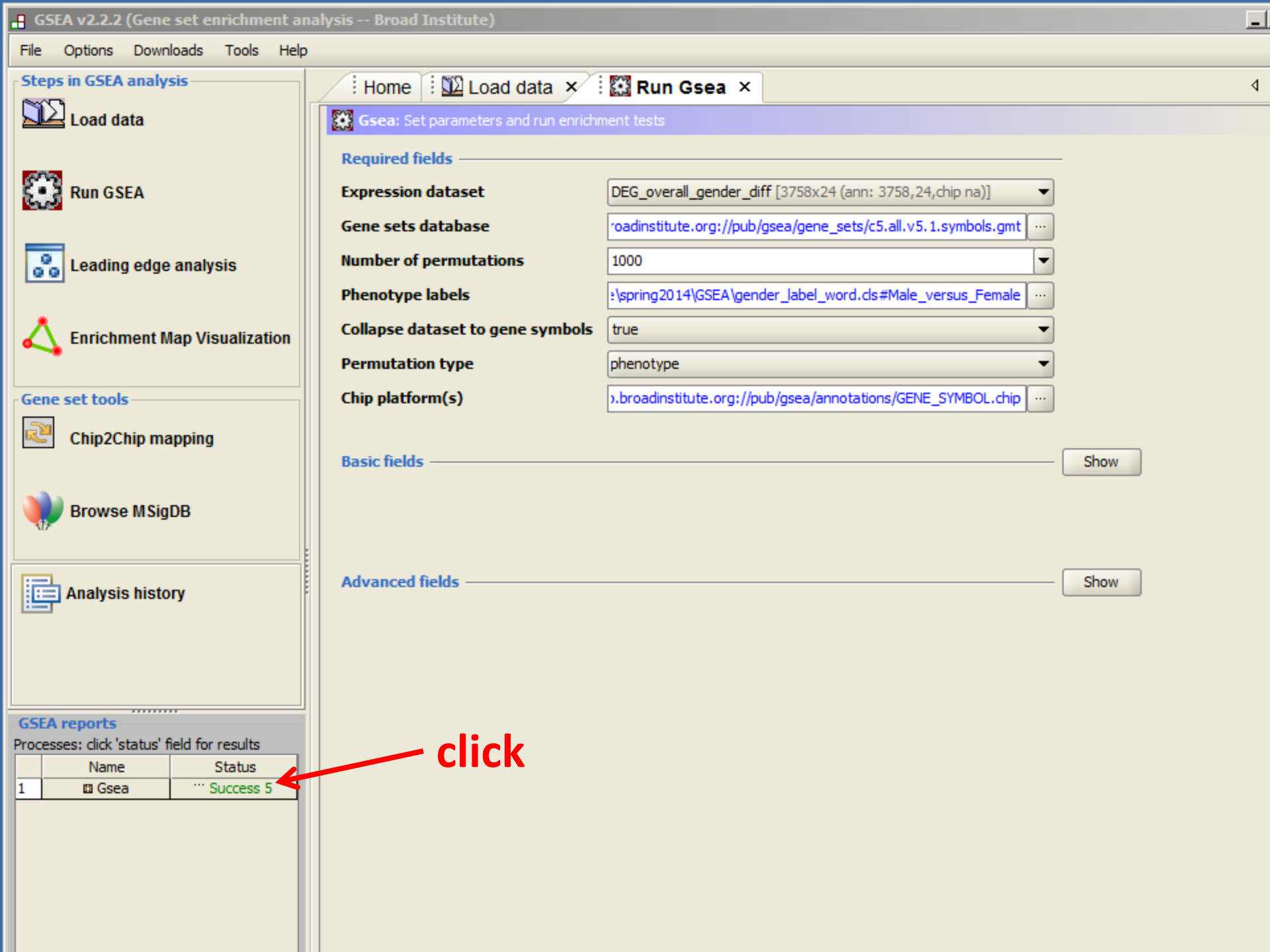
GSEA Example



GSEA Example




GSEA v2.2.2 (Gene set enrichment analysis -- Broad Institute)
File
Options
Downloads
Tools
Help
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
oadinstitute.org://pub/gsea/gene_sets/c5.all.v5.1.symbols.gmt
Number of permutations
1000
Phenotype labels
:\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
Basic fields
Show
Advanced fields
Show


GSEA v2.2.2 (Gene set enrichment analysis -- Broad Institute)
File
Options
Downloads
Tools
Help
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
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
oadinstitute.org://pub/gsea/gene_sets/c5.all.v5.1.symbols.gmt
Number of permutations
1000
Phenotype labels
:\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
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 significantly enriched 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
- Buttefly plot of significant genes

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

GSEA Example

Table: GSEA Results Summary

Dataset	DEG_overall_gender_diff_overall_gender_diff_collapsed_to_symbols.gender_label_word.cls #Male_versus_Female.gender_label_word.cls #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

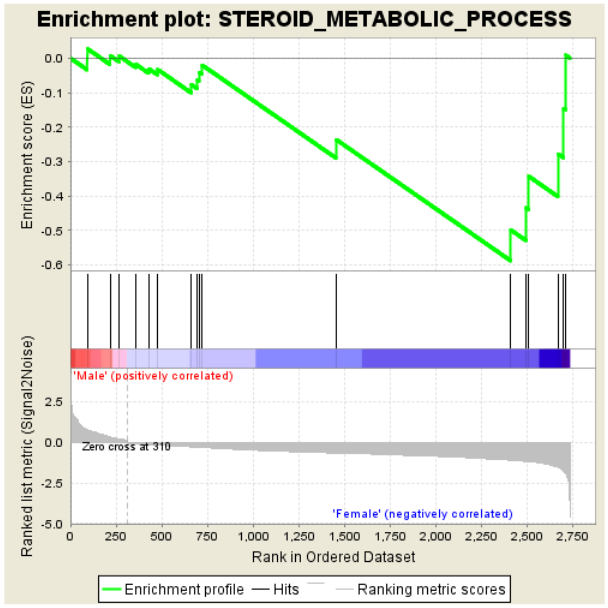
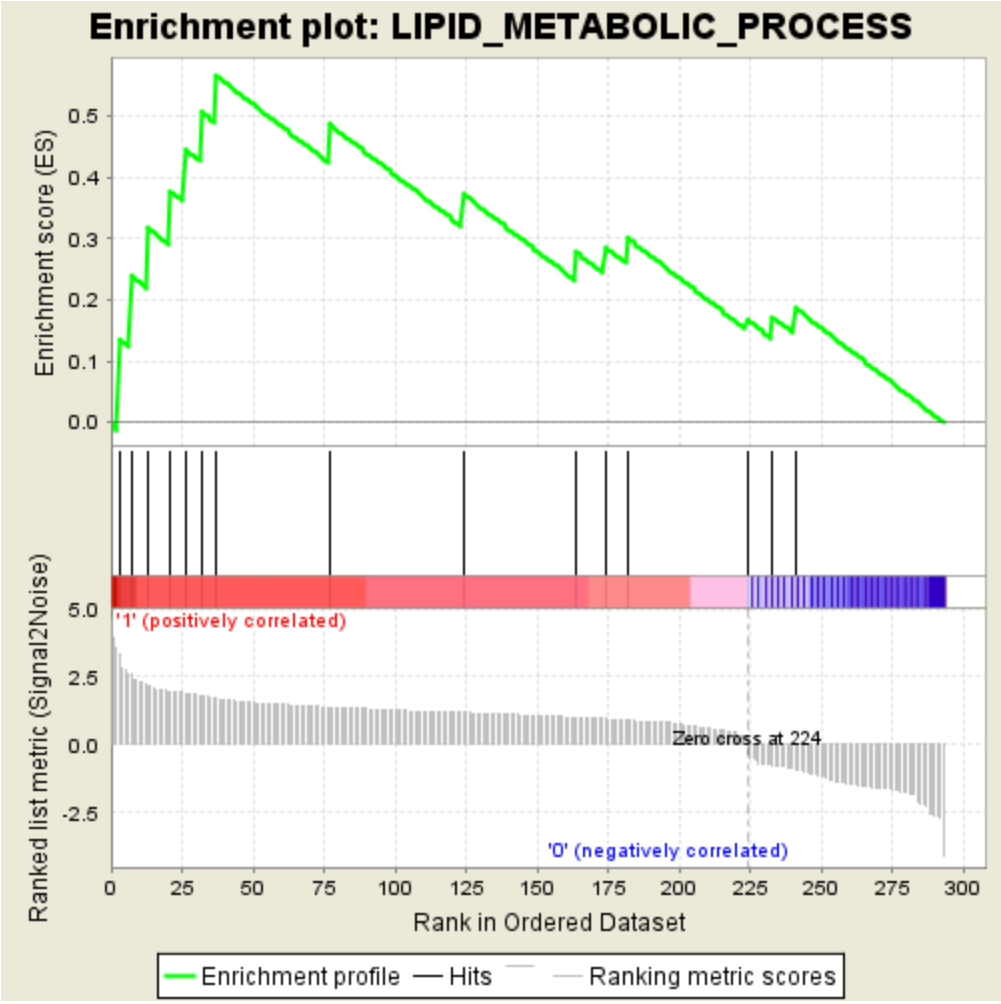
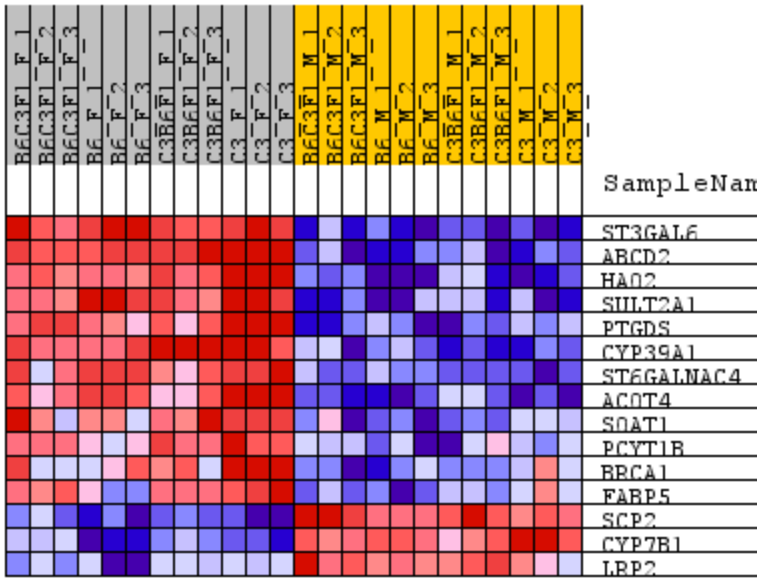


Fig 1: Enrichment plot: STEROID_METABOLIC_PROCESS
Profile of the Running ES Score & Positions of GeneSet Members on the Rank Ordered List

Lipid metabolism enriched -- GSEA





- 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/*