NIH National Institute of Environmental Health Sciences

Your Environment. Your Health.

Pathway Analysis

Biostatistics and Bioinformatics Short Course

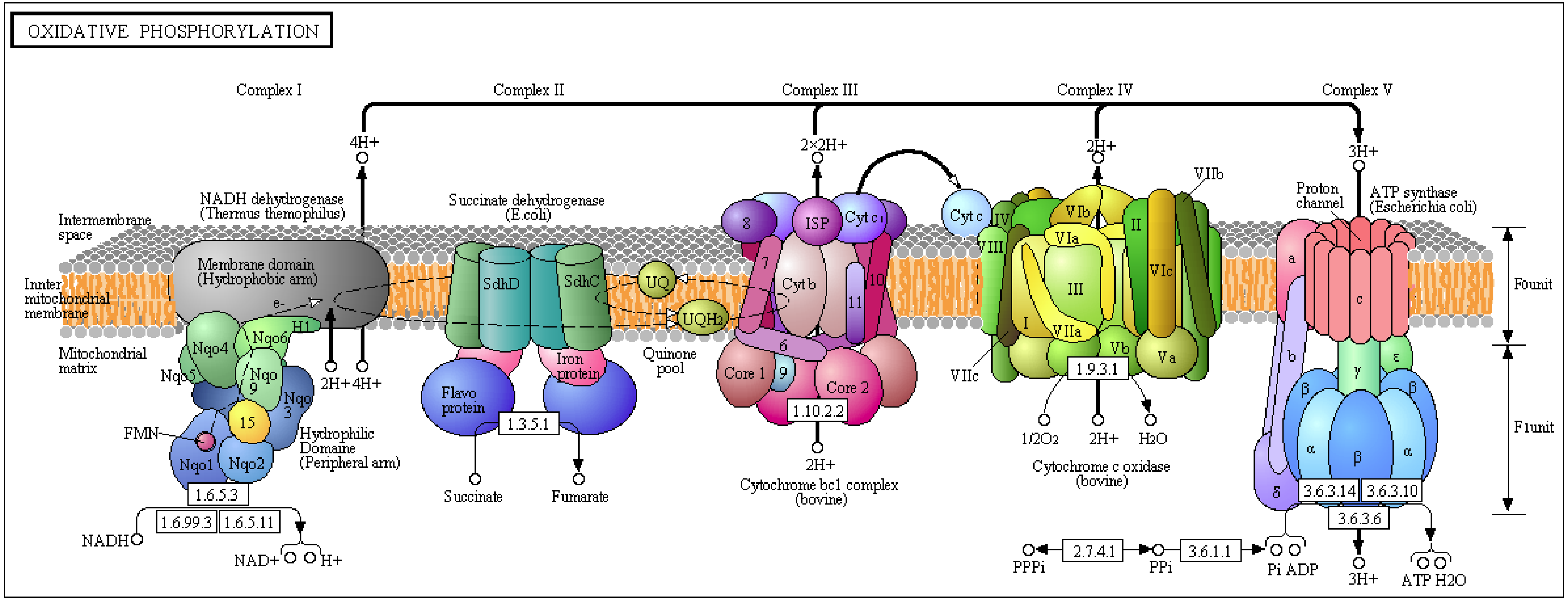
02/12/16

Brian Bennett, Ph.D.

(brian.bennett@nih.gov)

What is a Biological Pathway?

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



What is Pathway Analysis?

- Identifying pathways whose genes are associated with a particular biological condition
- Examines the combined signal from multiple genes, as opposed to looking at individual genes separately
- Recently, this definition has been extended to include any set of genes that have some sort biological connection (gene sets)
    - Mechanistic and signaling cascades (KEGG, Biocarta)
    - Functional and biological processes (GO)
    - Associated with a disease or condition
    - Chromosomal proximity
    - Computationally derived

Advantages over Single-Gene Analysis

- Different samples from a common condition may have different key genes that are all driving changes in the same pathway

CELEBRATING OUR TENTH YEAR

HISTORICAL PERSPECTIVE

nature medicine

http://www.nature.com/naturemedicine

Cancer genes and the pathways they control

Bert Vogelstein & Kenneth W Kinzler

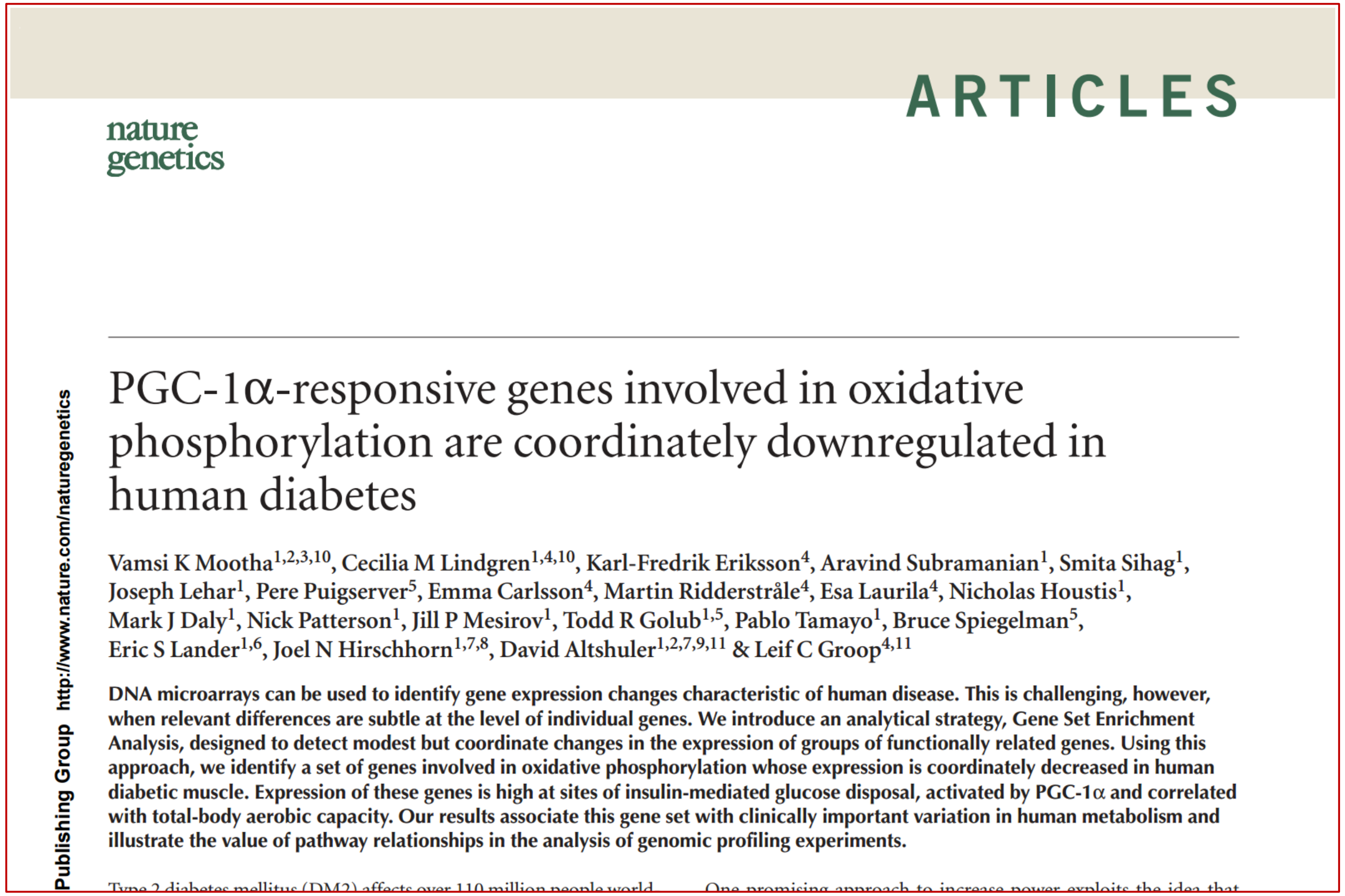
The revolution in cancer research can be summed up in a single sentence: cancer is, in essence, a genetic disease. In the last decade, many important genes responsible for the genesis of various cancers have been discovered, their mutations precisely identified, and the pathways through which they act characterized. The purposes of this review are to highlight examples of progress in these areas, indicate where knowledge is scarce and point out fertile grounds for future investigation.

What we know

product. Such inactivations arise from missense mutations at

Advantages over Single-Gene Analysis

- Pathways may be perturbed by subtle changes in many genes



ARTICLES

nature genetics

Publishing Group http://www.nature.com/naturegenetics

PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes

Vamsi K Mootha[1,2,3,10], Cecilia M Lindgren[1,4,10], Karl-Fredrik Eriksson[4], Aravind Subramanian[1], Smita Sihag[1], Joseph Lehar[1], Pere Puigserver[5], Emma Carlsson[4], Martin Ridderstråle[4], Esa Laurila[4], Nicholas Houstis[1], Mark J Daly[1], Nick Patterson[1], Jill P Mesirov[1], Todd R Golub[1,5], Pablo Tamayo[1], Bruce Spiegelman[5], Eric S Lander[1,6], Joel N Hirschhorn[1,7,8], David Altshuler[1,2,7,9,11] & Leif C Groop[4,11]

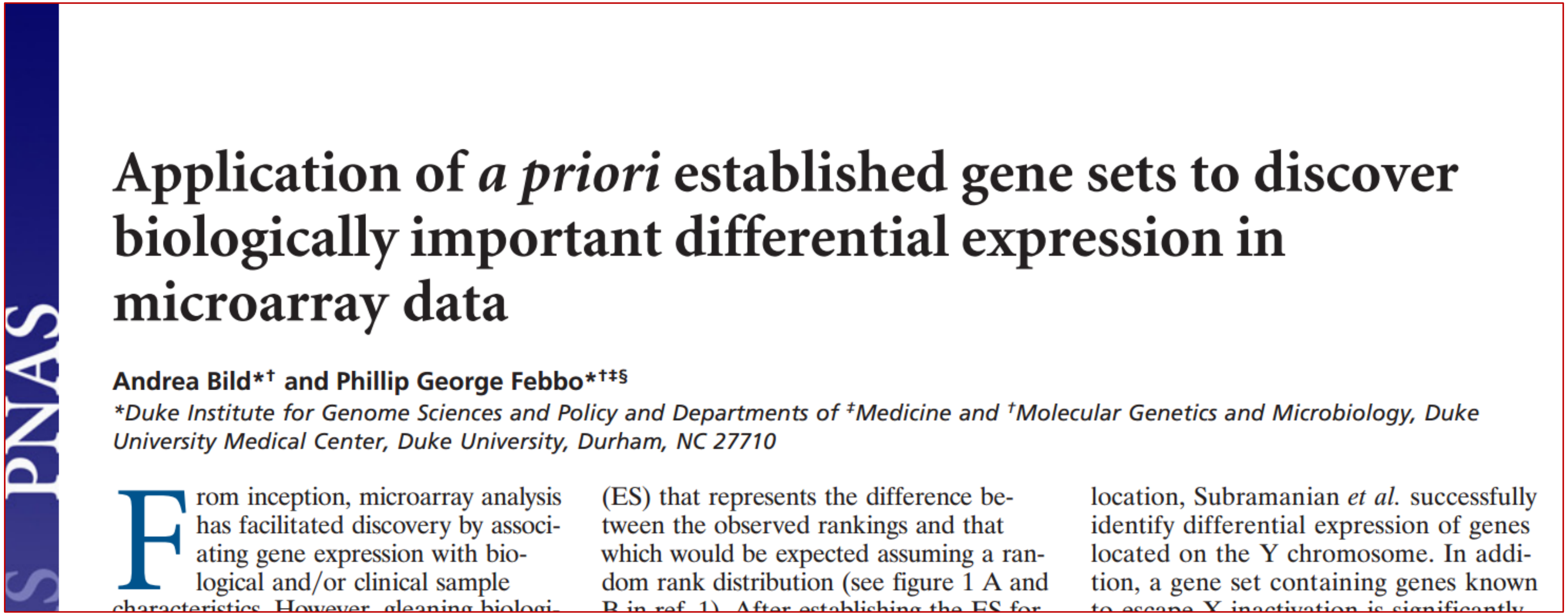
DNA microarrays can be used to identify gene expression changes characteristic of human disease. This is challenging, however, when relevant differences are subtle at the level of individual genes. We introduce an analytical strategy, Gene Set Enrichment Analysis, designed to detect modest but coordinate changes in the expression of groups of functionally related genes. Using this approach, we identify a set of genes involved in oxidative phosphorylation whose expression is coordinately decreased in human diabetic muscle. Expression of these genes is high at sites of insulin-mediated glucose disposal, activated by PGC-1α and correlated with total-body aerobic capacity. Our results associate this gene set with clinically important variation in human metabolism and illustrate the value of pathway relationships in the analysis of genomic profiling experiments.

Type 2 diabetes mellitus (DM2) affects over 110 million people world

One promising approach to increase power exploits the idea that

Advantages over Single-Gene Analysis

- A small list of enriched pathways may be easier to interpret than a large list of associated genes



Application of *a priori* established gene sets to discover biologically important differential expression in microarray data

Andrea Bild[*,†] and Phillip George Febbo[*,†,‡,§]

[*]Duke Institute for Genome Sciences and Policy and Departments of [‡]Medicine and [†]Molecular Genetics and Microbiology, Duke University Medical Center, Duke University, Durham, NC 27710

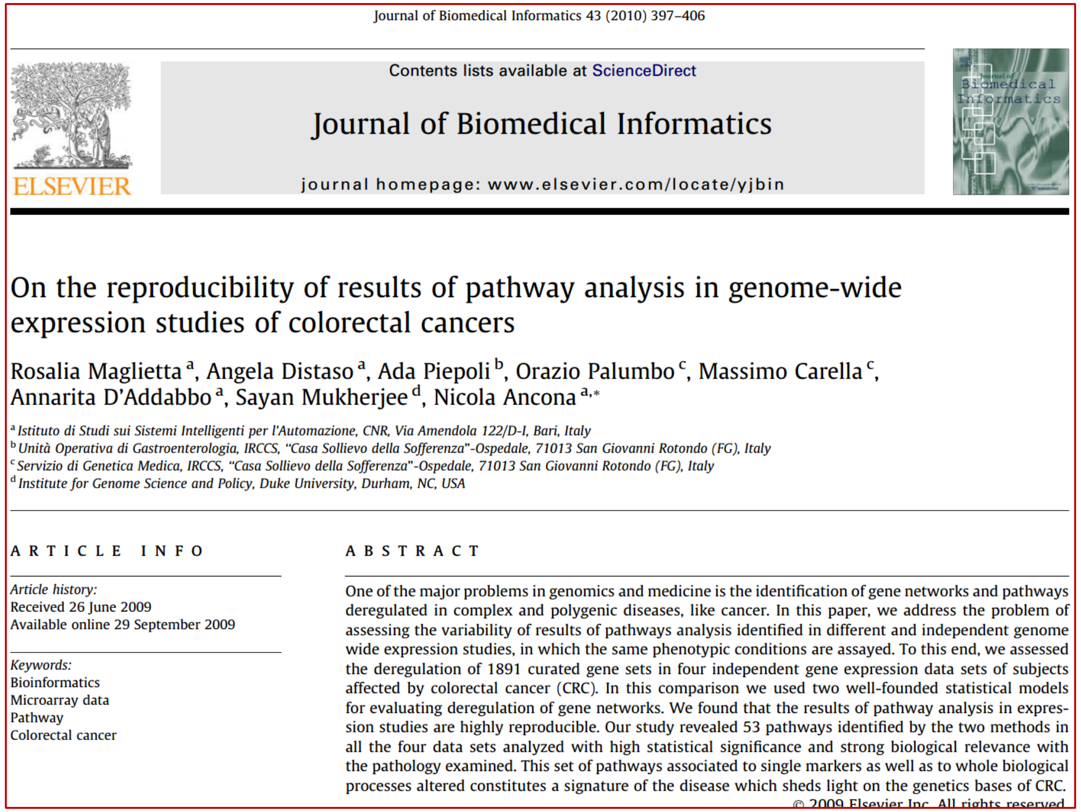
From inception, microarray analysis has facilitated discovery by associating gene expression with biological and/or clinical sample characteristics. However, gleaning biologi-

(ES) that represents the difference between the observed rankings and that which would be expected assuming a random rank distribution (see figure 1 A and B in ref. 1). After establishing the ES for

location, Subramanian *et al.* successfully identify differential expression of genes located on the Y chromosome. In addition, a gene set containing genes known to escape X inactivation is significantly

Advantages over Single-Gene Analysis

- Pathways results from different studies may overlap better than single-gene results



Journal of Biomedical Informatics 43 (2010) 397–406

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

On the reproducibility of results of pathway analysis in genome-wide expression studies of colorectal cancers

Rosalia Maglietta[a], Angela Distaso[a], Ada Piepoli[b], Orazio Palumbo[c], Massimo Carella[c], Annarita D'Addabbo[a], Sayan Mukherjee[d], Nicola Ancona[a,*]

[a] *Istituto di Studi sui Sistemi Intelligenti per l'Automazione, CNR, Via Amendola 122/D-I, Bari, Italy*
[b] *Unità Operativa di Gastroenterologia, IRCCS, "Casa Sollievo della Sofferenza"-Ospedale, 71013 San Giovanni Rotondo (FG), Italy*
[c] *Servizio di Genetica Medica, IRCCS, "Casa Sollievo della Sofferenza"-Ospedale, 71013 San Giovanni Rotondo (FG), Italy*
[d] *Institute for Genome Science and Policy, Duke University, Durham, NC, USA*

ARTICLE INFO

Article history:
Received 26 June 2009
Available online 29 September 2009

Keywords:
Bioinformatics
Microarray data
Pathway
Colorectal cancer

ABSTRACT

One of the major problems in genomics and medicine is the identification of gene networks and pathways deregulated in complex and polygenic diseases, like cancer. In this paper, we address the problem of assessing the variability of results of pathways analysis identified in different and independent genome wide expression studies, in which the same phenotypic conditions are assayed. To this end, we assessed the deregulation of 1891 curated gene sets in four independent gene expression data sets of subjects affected by colorectal cancer (CRC). In this comparison we used two well-founded statistical models for evaluating deregulation of gene networks. We found that the results of pathway analysis in expression studies are highly reproducible. Our study revealed 53 pathways identified by the two methods in all the four data sets analyzed with high statistical significance and strong biological relevance with the pathology examined. This set of pathways associated to single markers as well as to whole biological processes altered constitutes a signature of the disease which sheds light on the genetics bases of CRC.

Gene Expression Pathway Analysis

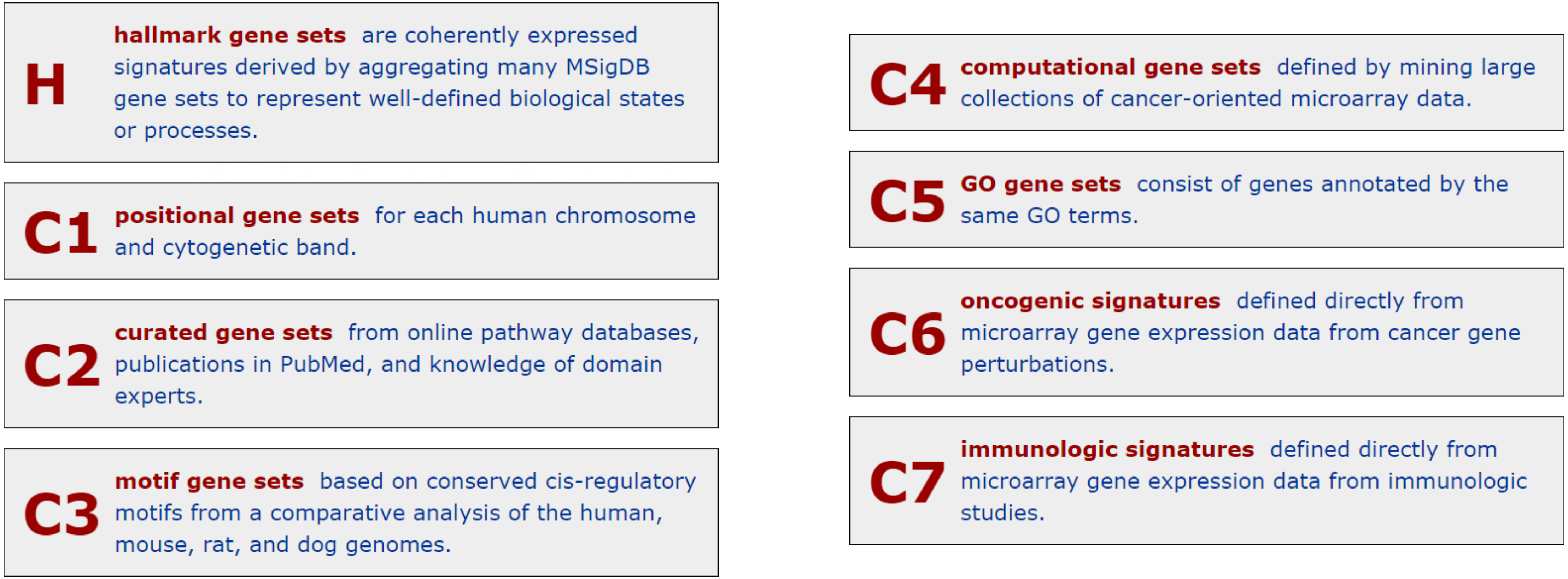
- Data: Gene expression data for samples in two groups (disease vs. control, treatment vs. no treatment, etc.)
- Biological Question: Which pathways are impacted by the condition or treatment?
- Statistical Question: Which pathways have more differentially expressed genes than expected by chance?

Required Data

1. Gene expression data set
2. Collection of gene sets

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.

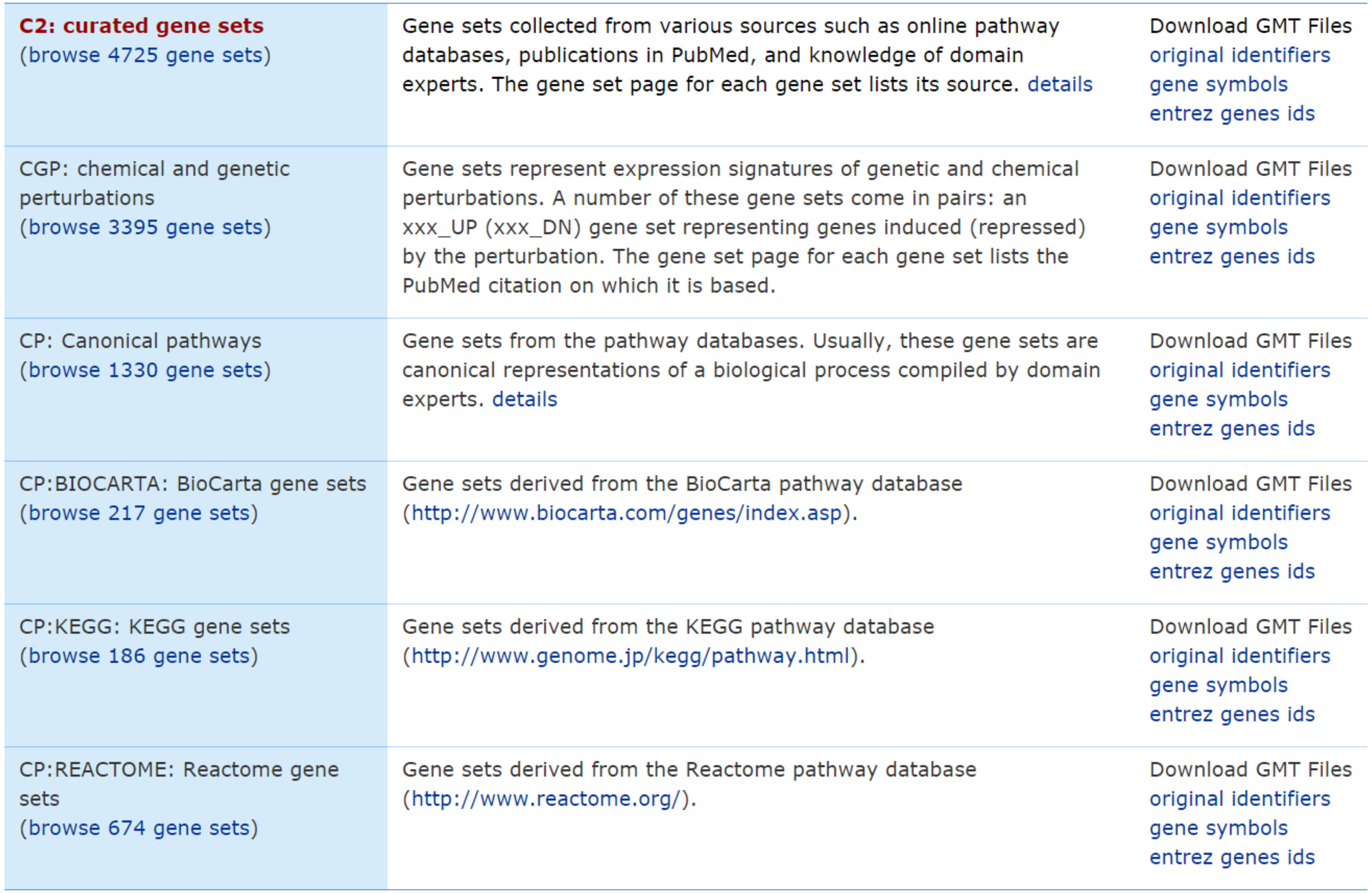
C4 **computational gene sets** defined by mining large collections of cancer-oriented microarray data.

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.

Molecular Signatures Database (MSigDB)

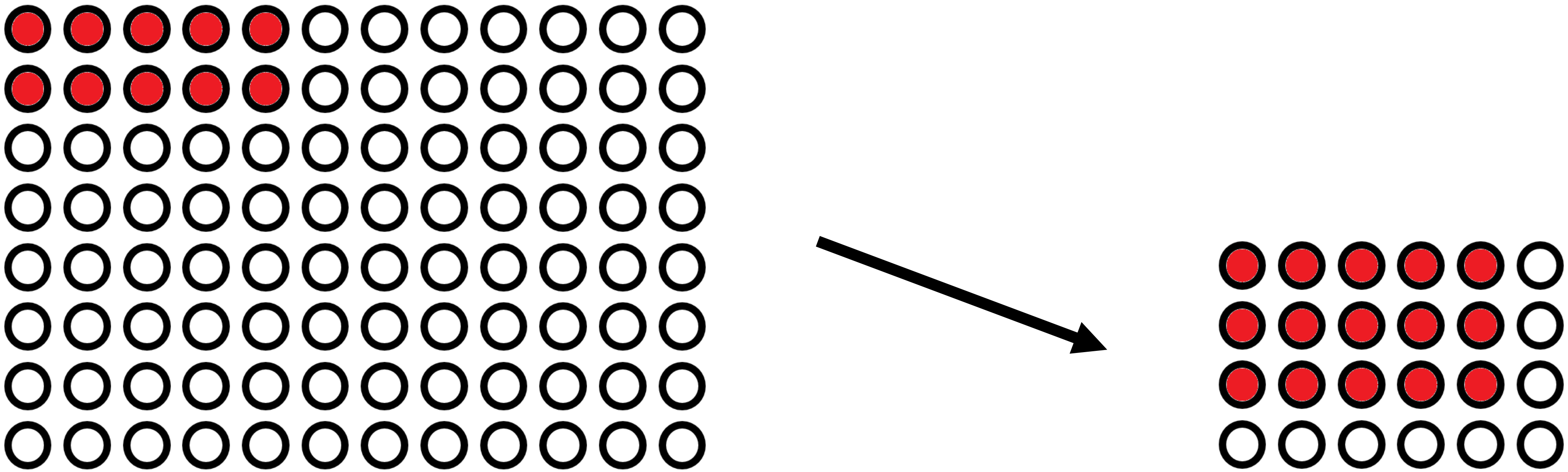


C2: curated gene sets (browse 4725 gene sets)	Gene sets collected from various sources such as online pathway databases, publications in PubMed, and knowledge of domain experts. The gene set page for each gene set lists its source. details	Download GMT Files original identifiers gene symbols entrez genes ids
CGP: chemical and genetic perturbations (browse 3395 gene sets)	Gene sets represent expression signatures of genetic and chemical perturbations. A number of these gene sets come in pairs: an xxx_UP (xxx_DN) gene set representing genes induced (repressed) by the perturbation. The gene set page for each gene set lists the PubMed citation on which it is based.	Download GMT Files original identifiers gene symbols entrez genes ids
CP: Canonical pathways (browse 1330 gene sets)	Gene sets from the pathway databases. Usually, these gene sets are canonical representations of a biological process compiled by domain experts. details	Download GMT Files original identifiers gene symbols entrez genes ids
CP:BIOCARTA: BioCarta gene sets (browse 217 gene sets)	Gene sets derived from the BioCarta pathway database (http://www.biocarta.com/genes/index.asp).	Download GMT Files original identifiers gene symbols entrez genes ids
CP:KEGG: KEGG gene sets (browse 186 gene sets)	Gene sets derived from the KEGG pathway database (http://www.genome.jp/kegg/pathway.html).	Download GMT Files original identifiers gene symbols entrez genes ids
CP:REACTOME: Reactome gene sets (browse 674 gene sets)	Gene sets derived from the Reactome pathway database (http://www.reactome.org/).	Download GMT Files original identifiers gene symbols entrez genes ids

Hypergeometric Test (Right-tailed)

- Parametric method
- Can also use Fisher’s Exact (FE) test
- Required data:
  - List of DEGs (along with the number of total genes)
  - A gene set
- If the proportion of DEGs within the gene set if sufficiently higher than the proportion within the entire set, the gene set is considered significantly enriched

Hypergeometric Test (Right-tailed)



Full data

k = total DEGs = 1000

N = total overall genes = 10000

(10% are DEGs)

Gene set

x = total DEGs in gene set = 60

m = total genes in gene set = 100

(60% are DEGs)

Hypergeometric Test (Right-tailed)

- x = total DEGs in gene set = 60
- m = total genes in gene set = 100
- k = total DEGs = 1000
- N = total overall genes = 10000

- R code:
  - `phyper( x-1, m, N-m, k, lower.tail=FALSE )`
  - `fisher.test( matrix(c(x,k-x,m-x,N-m-k+x),2,2), alternative='greater' )$p.value`

- P-value = 5.4×10^{-35}

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

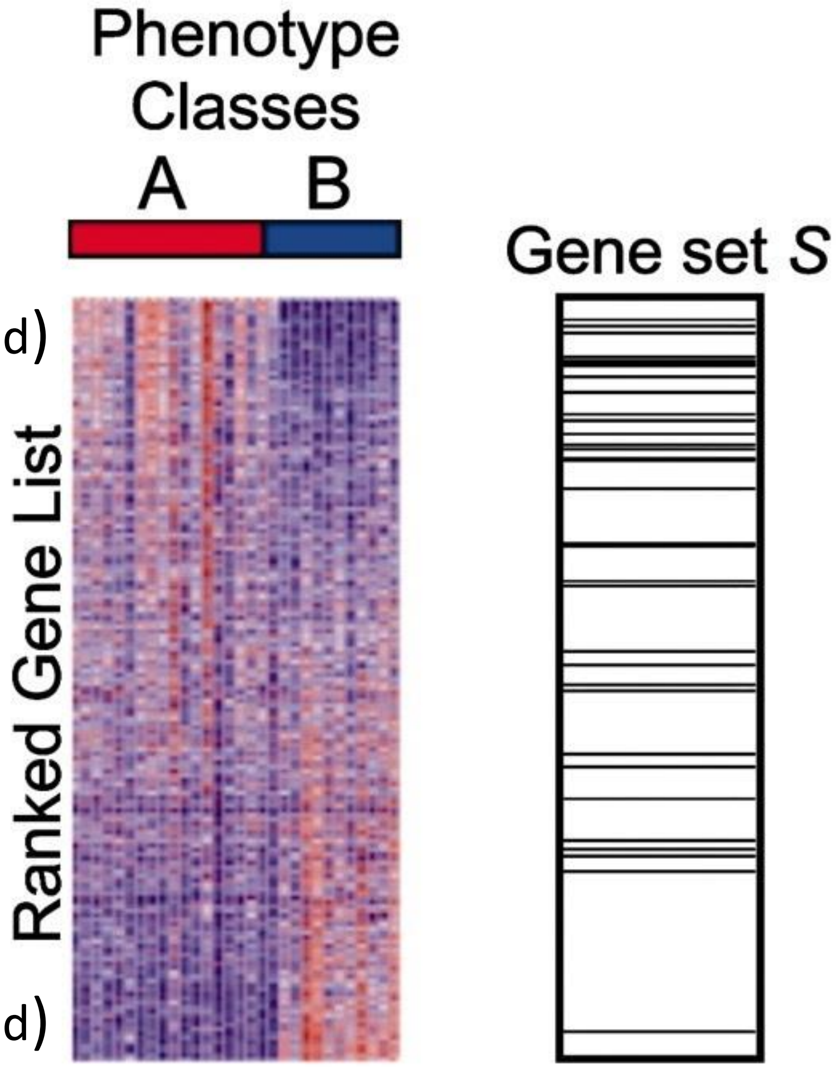
Hypergeometric Test (Right-tailed)

- 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

Kolmogorov–Smirnov (KS) Test

- Nonparametric method
- 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

Kolmogorov–Smirnov (KS) Test

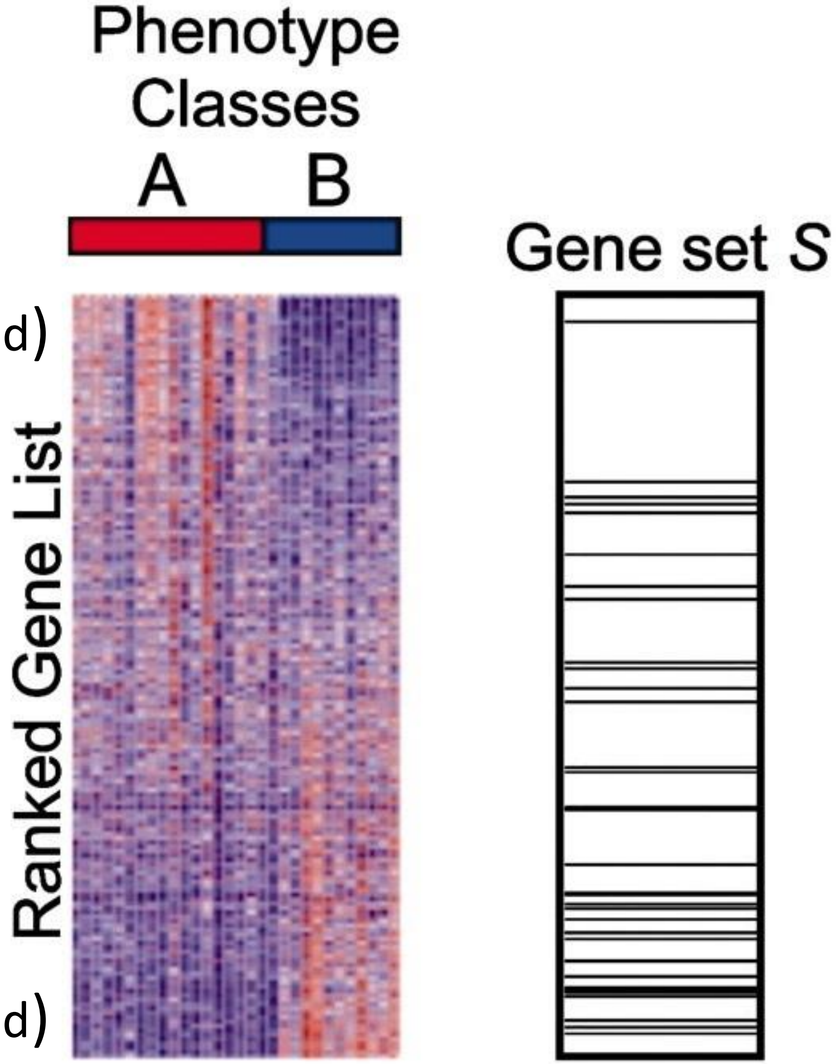


(upregulated)

(downregulated)

Large, positive enrichment score

Kolmogorov–Smirnov (KS) Test

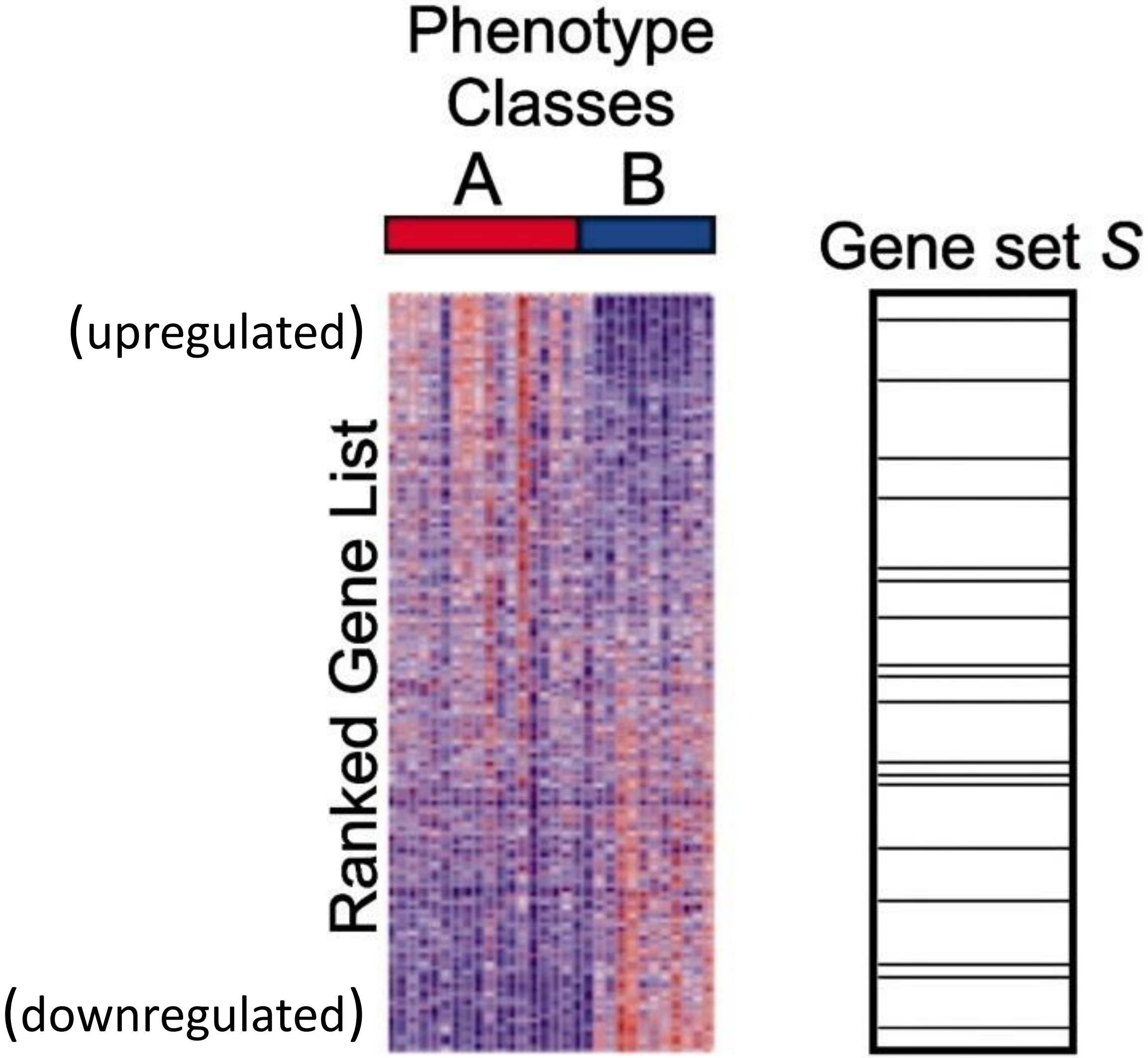


(upregulated)

(downregulated)

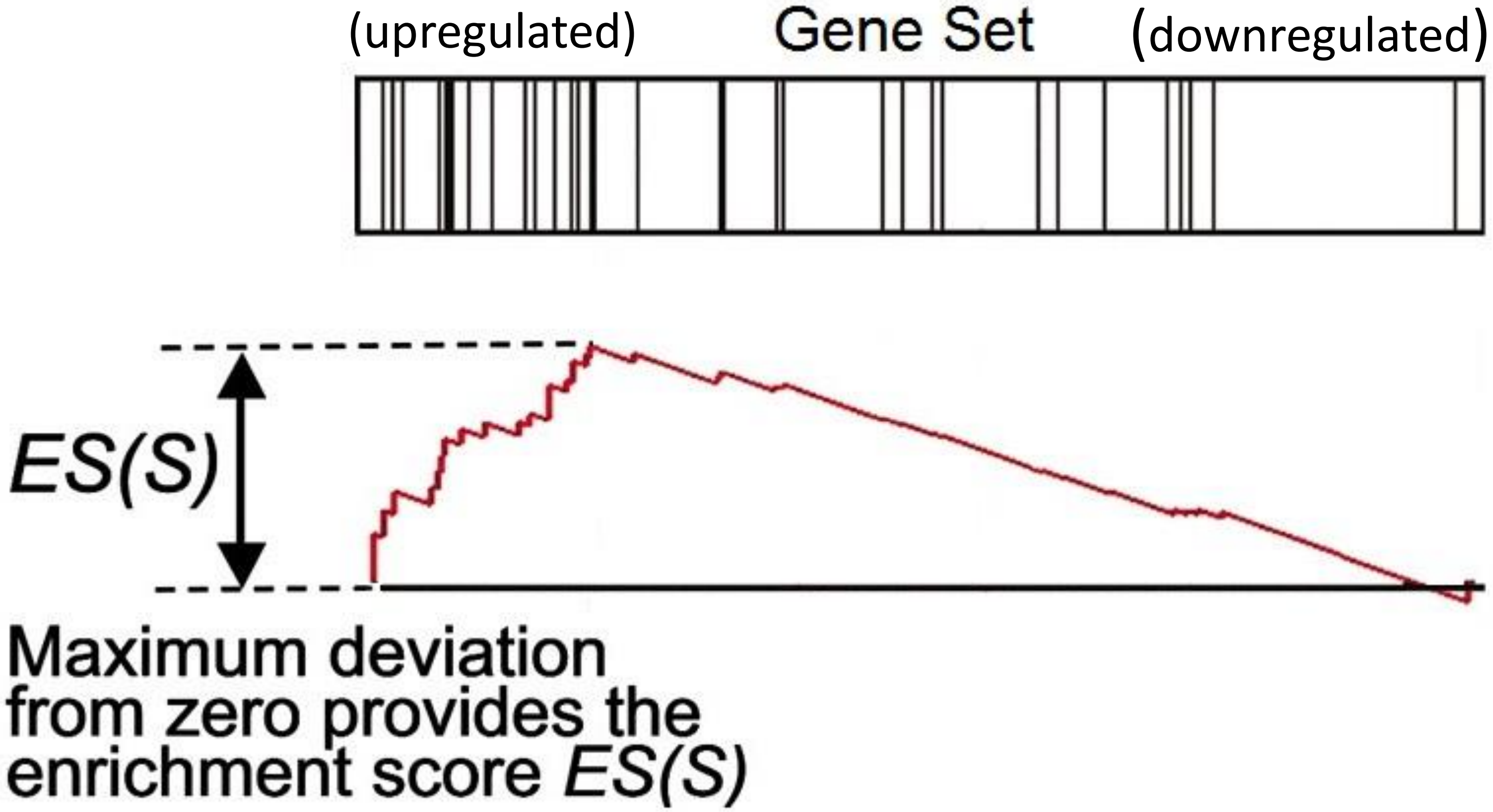
Large, negative enrichment score

Kolmogorov–Smirnov (KS) Test



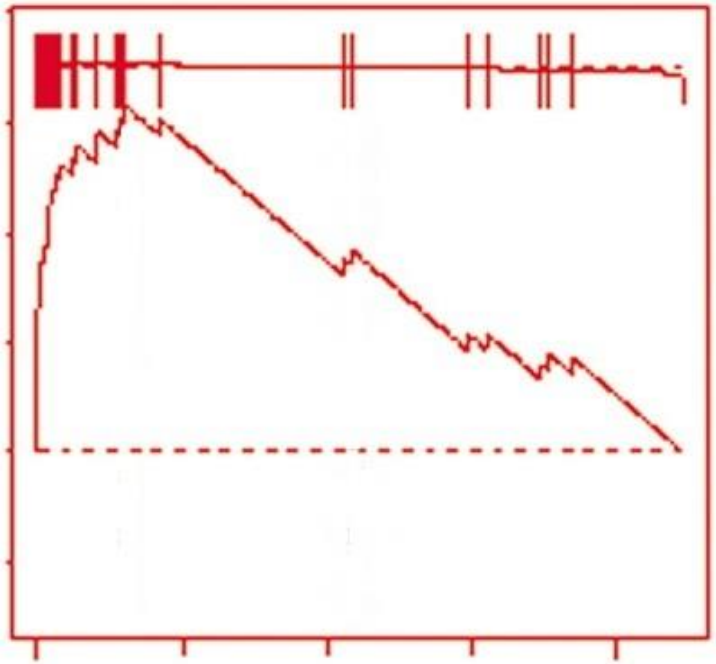
Near zero enrichment score

Kolmogorov–Smirnov (KS) Test



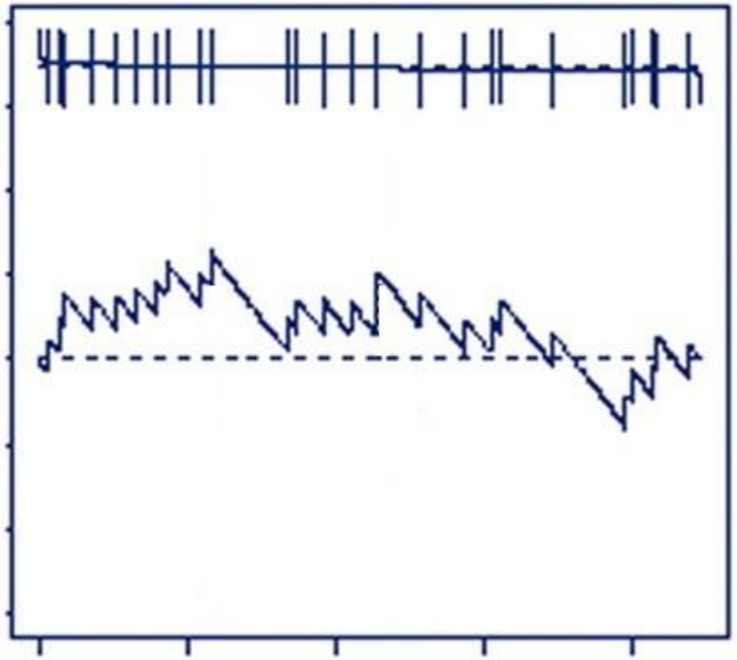
Kolmogorov–Smirnov (KS) Test

Many genes upregulated



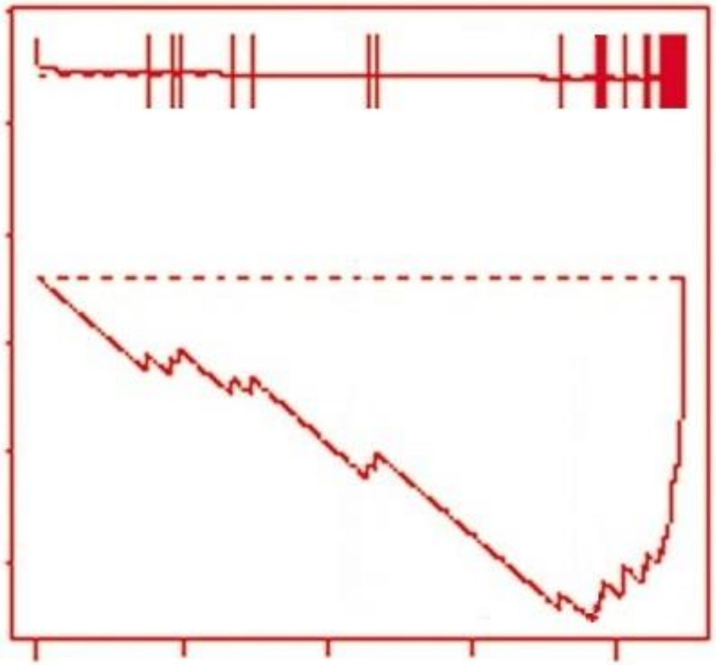
Large positive enrichment score

Genes randomly distributed



Near zero enrichment score

Many genes downregulated



Large negative enrichment score

Kolmogorov–Smirnov (KS) Test

- Generate a null distribution of permuted ESs by shuffling the class labels
- Calculate normalized enrichment score (NES)
  - This adjusts for gene set size and correlation bias
  - Divide the original ES by the mean of permuted ESs with the same sign
- Calculate p-value
  - Compare the original ES to the distribution of permuted ESs (one-tailed)
  - Calculate the percentage of permuted ESs that are higher than the original ES

Kolmogorov–Smirnov (KS) Test

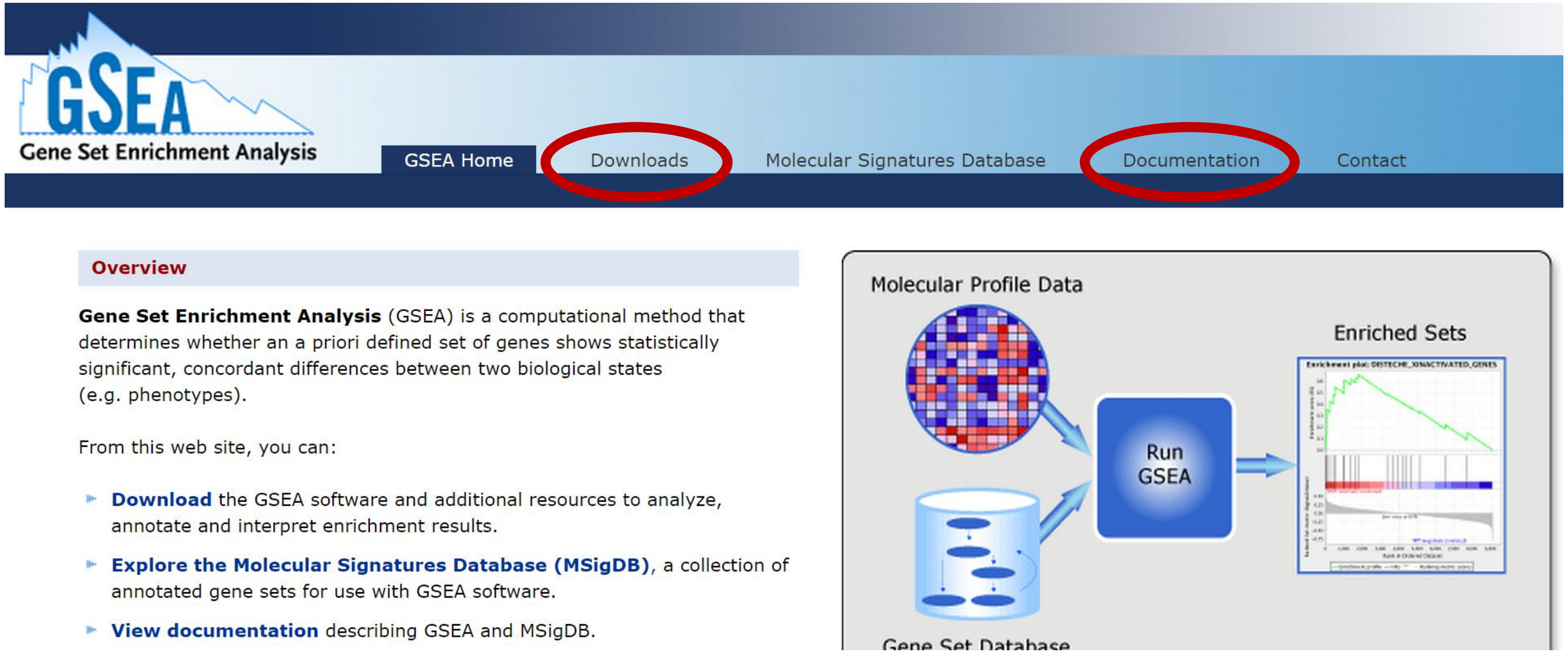
- Used in Gene Set Enrichment Analysis (GSEA)
  - Free software (http://www.broadinstitute.org/gsea)
  - Pros:
    - Large collection of gene sets
    - Uses more information than methods that only use DEG list
    - Enrichment plot improves interpretability
  - Cons:
    - Permutation-based p-values
    - Gene sets are not as clean

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

GSEA Example

- http://www.broadinstitute.org/gsea



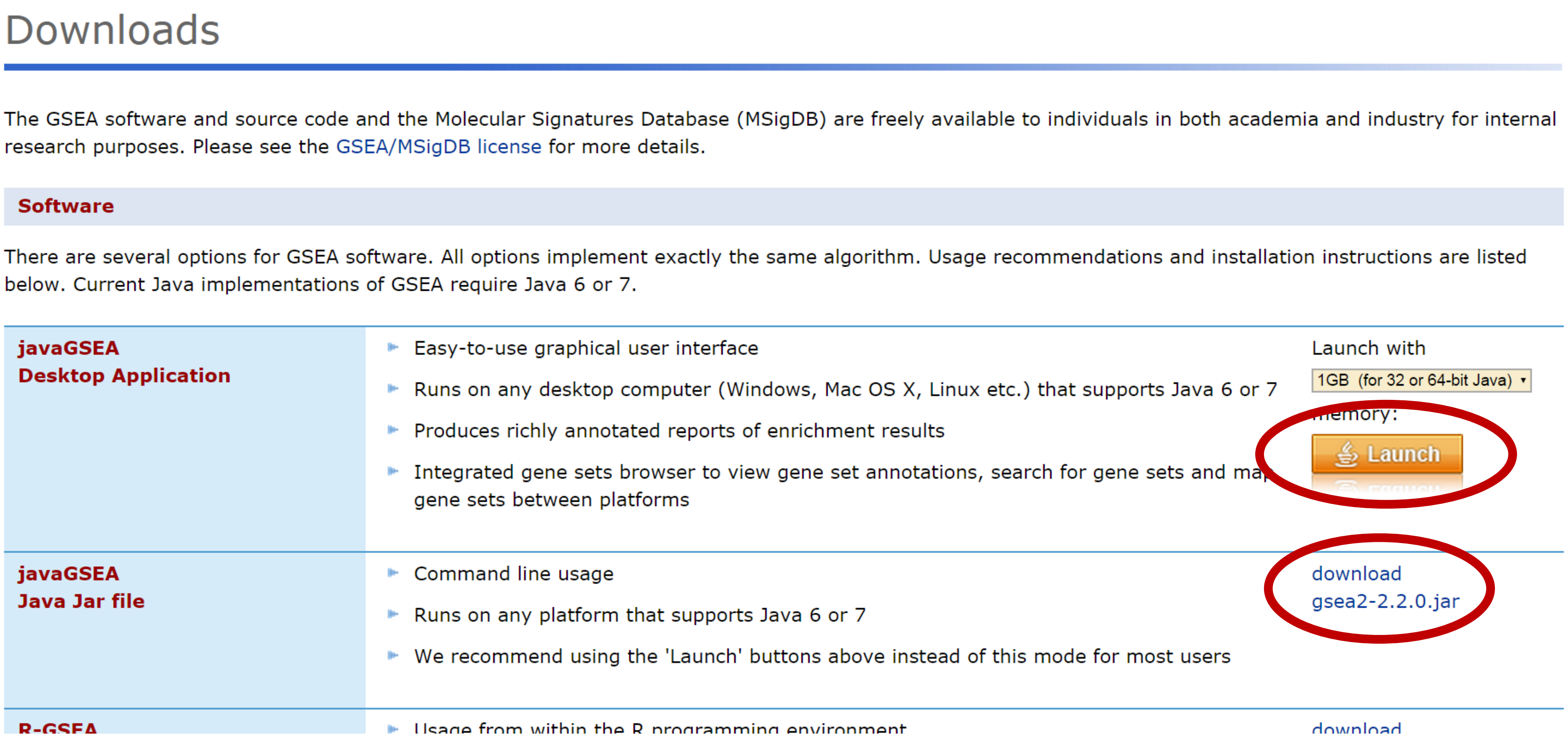
Overview

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.

GSEA Example



Downloads

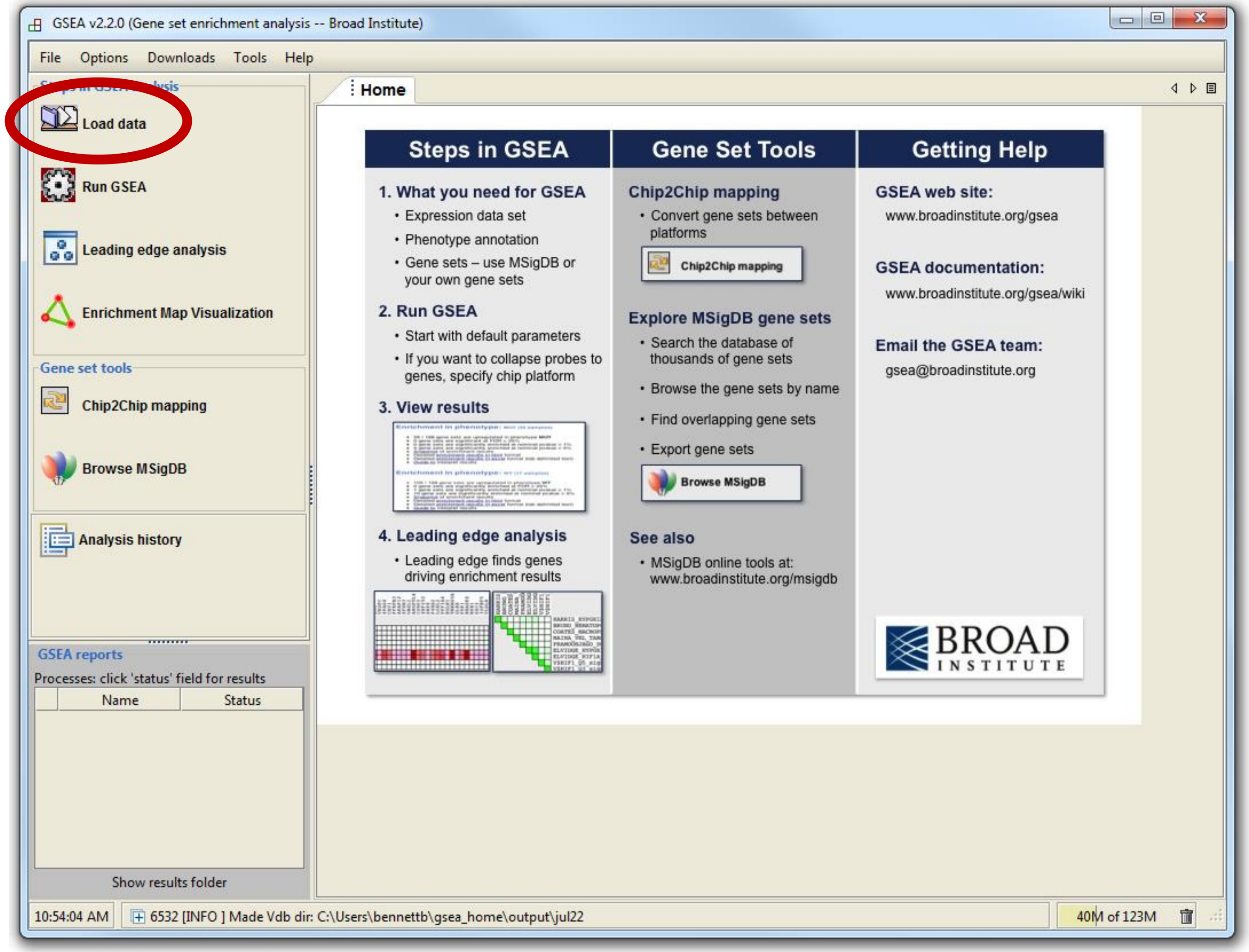
The GSEA software and source code and the Molecular Signatures Database (MSigDB) are freely available to individuals in both academia and industry for internal research purposes. Please see the GSEA/MSigDB license for more details.

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 map 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

GSEA Example



Processes: click 'status' field for results

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

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

See also

- MSigDB online tools at: www.broadinstitute.org/msigdb

Getting Help

GSEA web site:

www.broadinstitute.org/gsea

GSEA documentation:

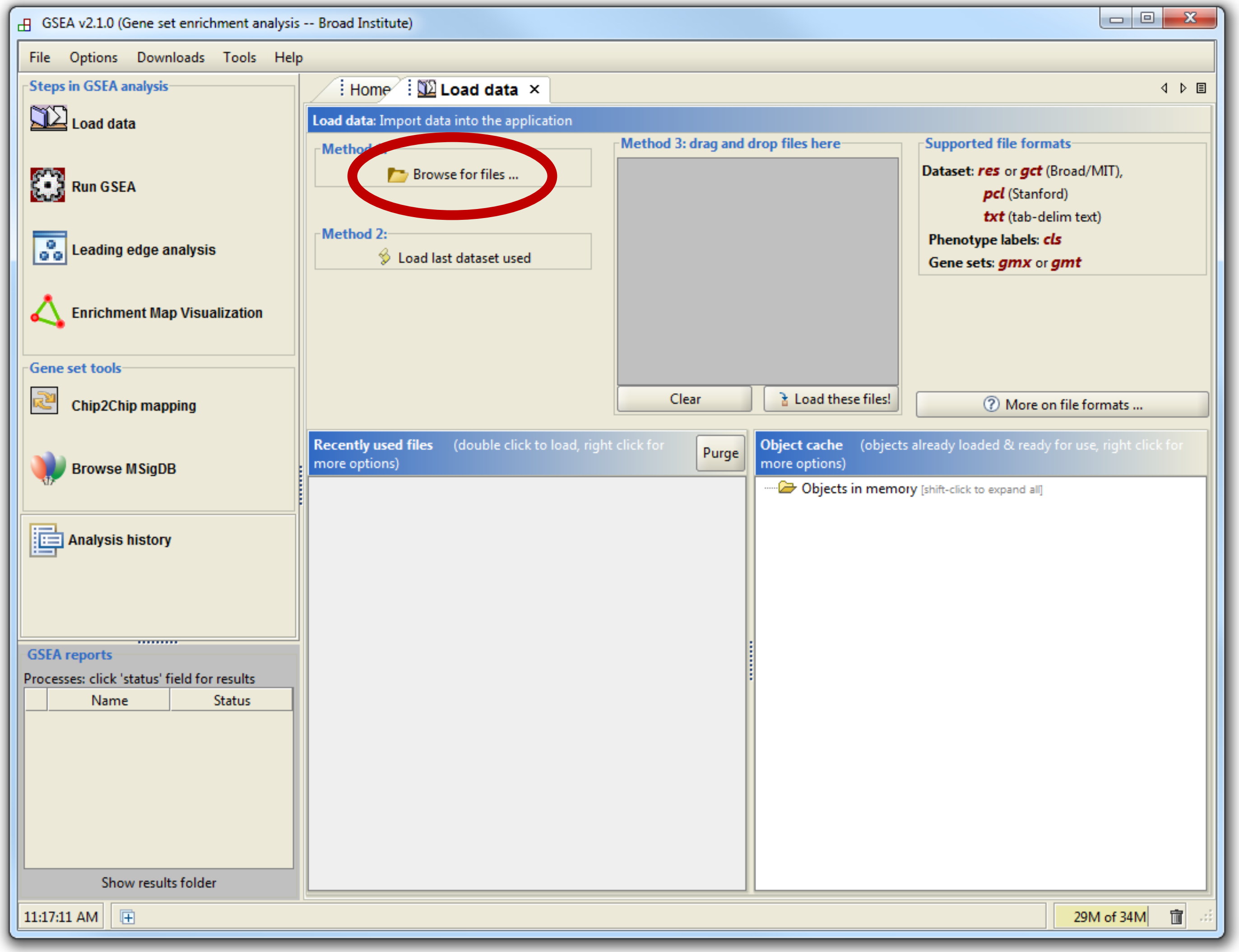
www.broadinstitute.org/gsea/wiki

Email the GSEA team:

gsea@broadinstitute.org

```
10:54:04 AM   6532 [INFO ] Made Vdb dir: C:\Users\bennettb\gsea_home\output\jul22   40M of 123M
```

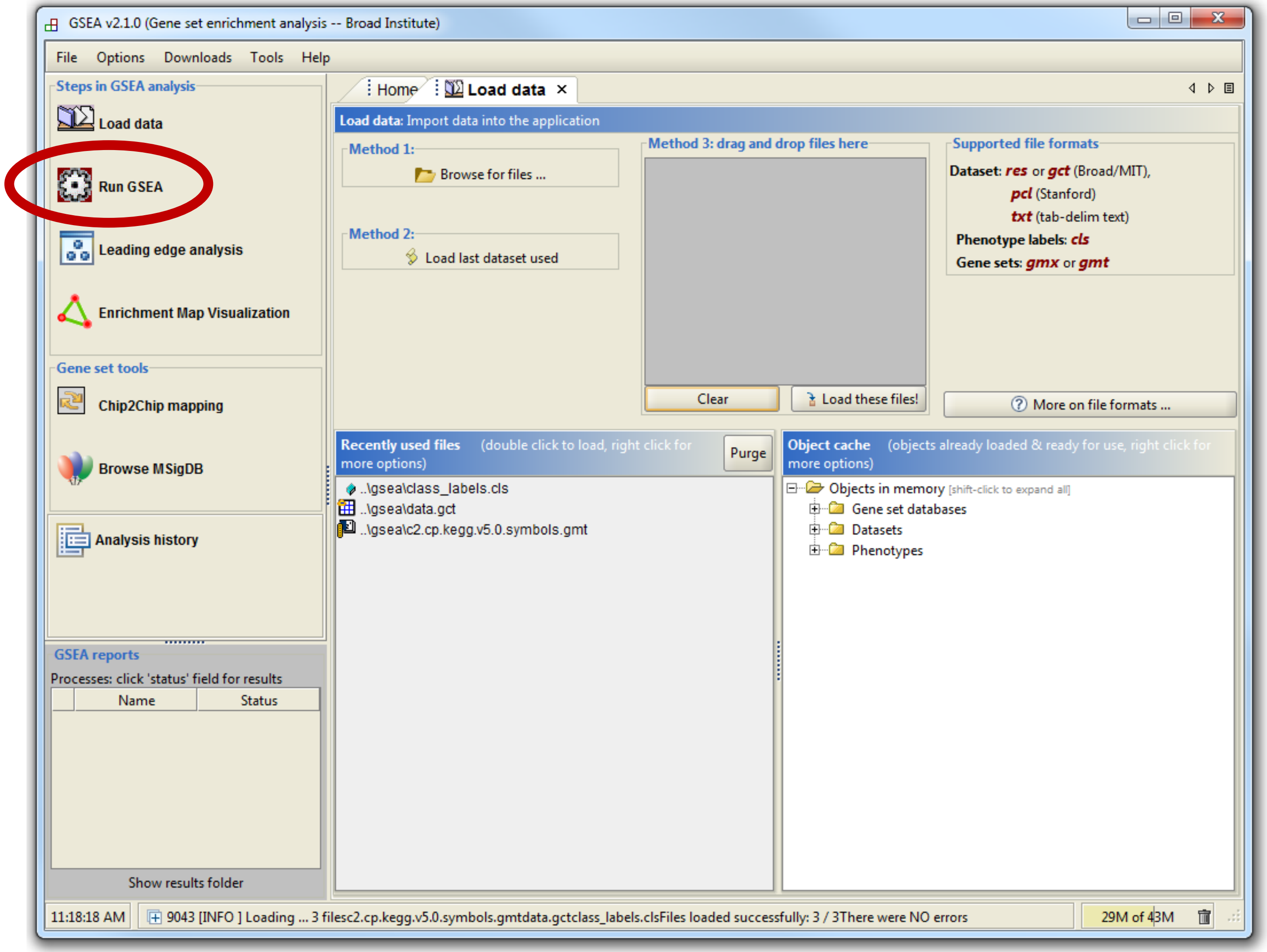
GSEA Example


GSEA v2.1.0 (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
Show results folder
Home
Load data
Load data: Import data into the application
Method
Browse for files ...
Method 2:
Load last dataset used
Method 3: drag and drop files here
Clear
Load these files!
Supported file formats
Dataset: res or gct (Broad/MIT),
pcl (Stanford)
txt (tab-delim text)
Phenotype labels: cls
Gene sets: gmx or gmt
More on file formats ...
Recently used files (double click to load, right click for more options)
Purge
Object cache (objects already loaded & ready for use, right click for more options)
Objects in memory [shift-click to expand all]
11:17:11 AM
29M of 34M

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

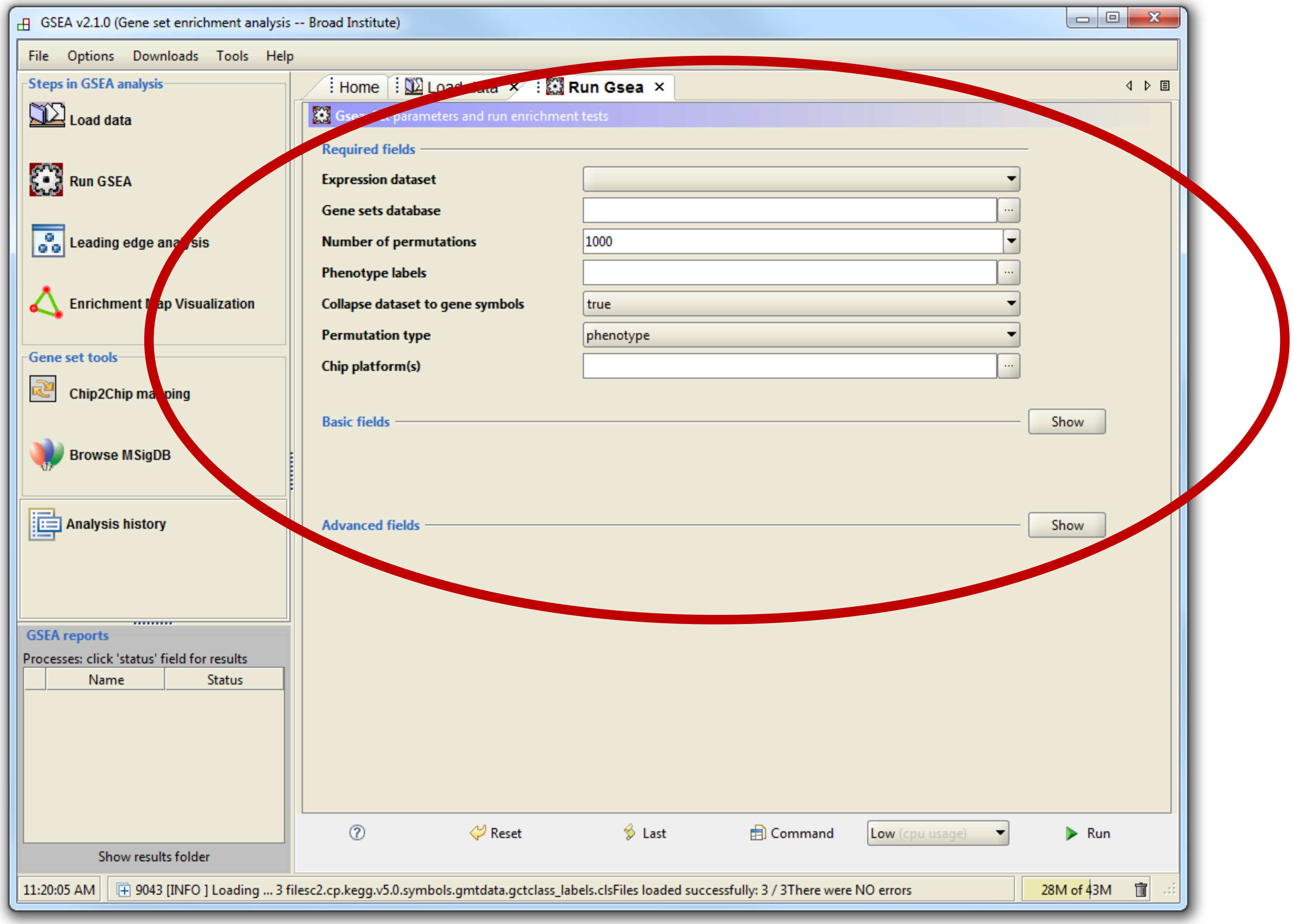

GSEA v2.1.0 (Gene set enrichment analysis -- Broad Institute)
File
Options
Downloads
Tools
Help
Steps in GSEA analysis
Load data
Run GSEA
Leading edge analysis
Enrichment Map Visualization
Gene set tools
Chip2Chip mapping
Browse MSigDB
Analysis history
GSEA reports
Processes: click 'status' field for results
Name
Status
Show results folder
Home
Load data
Load data: Import data into the application
Method 1:
Browse for files ...
Method 2:
Load last dataset used
Method 3: drag and drop files here
Clear
Load these files!
Supported file formats
More on file formats ...
Recently used files (double click to load, right click for more options)
Purge
Object cache (objects already loaded & ready for use, right click for more options)
Objects in memory [shift-click to expand all]
Gene set databases
Datasets
Phenotypes
29M of 43M

Dataset: *res* or *gct* (Broad/MIT),
pcl (Stanford)
txt (tab-delim text)
Phenotype labels: *cls*
Gene sets: *gmx* or *gmt*

..\gsea\class_labels.cls
..\gsea\data.gct
..\gsea\c2.cp.kegg.v5.0.symbols.gmt

```
11:18:18 AM   9043 [INFO ] Loading ... 3 filesc2.cp.kegg.v5.0.symbols.gmtdata.gctclass_labels.clsFiles loaded successfully: 3 / 3There were NO errors
```

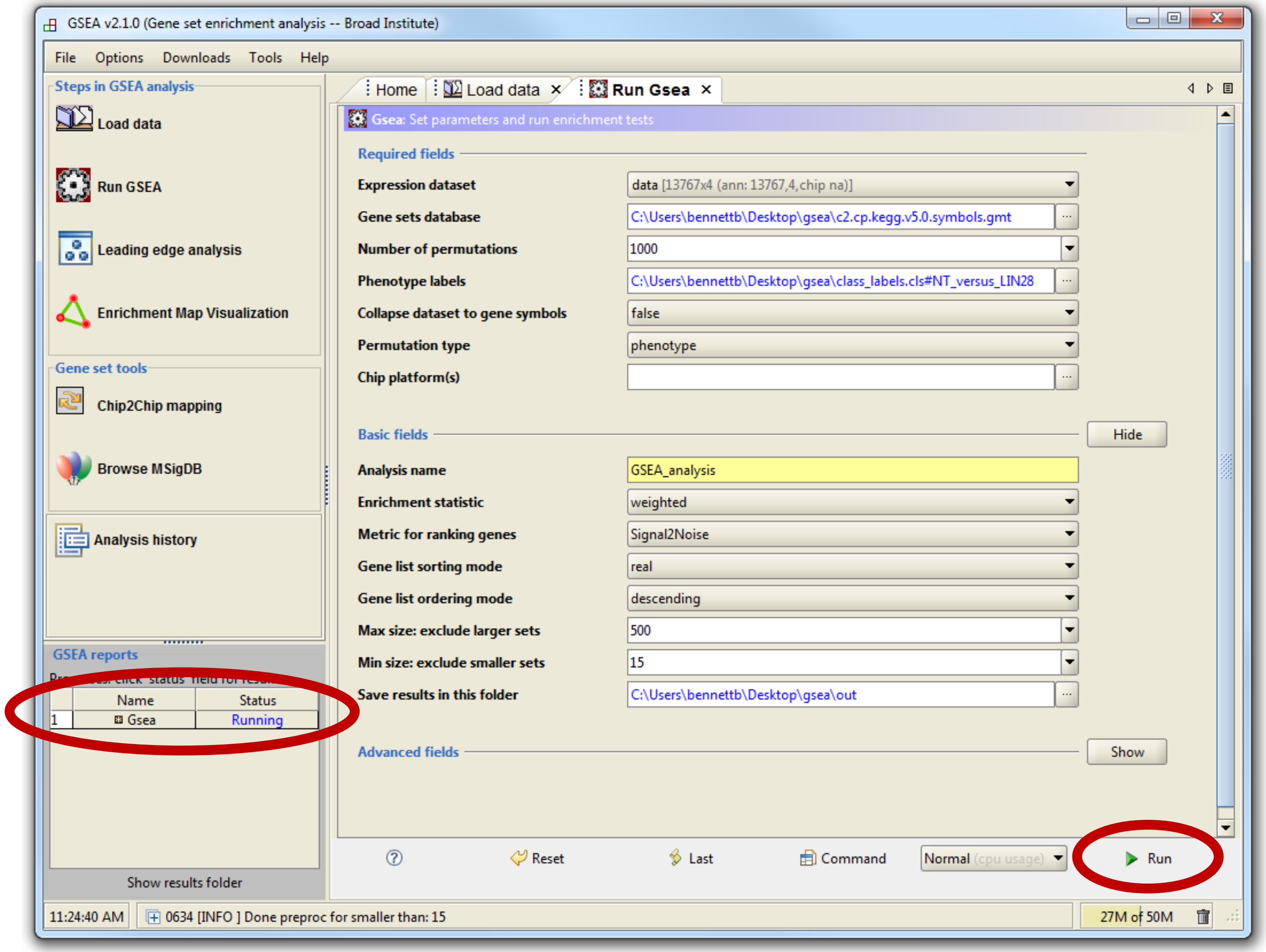
GSEA Example



Expression dataset	
Gene sets database	
Number of permutations	1000
Phenotype labels	
Collapse dataset to gene symbols	true
Permutation type	phenotype
Chip platform(s)	

```
11:20:05 AM  9043 [INFO ] Loading ... 3 filesc2.cp.kegg.v5.0.symbols.gmtdata.gctclass_labels.clsFiles loaded successfully: 3 / 3There were NO errors  28M of 43M
```

GSEA Example



GSEA v2.1.0 (Gene set enrichment analysis -- Broad Institute)

Gsea: Set parameters and run enrichment tests

Required fields

Expression dataset	data [13767x4 (ann: 13767,4,chip na)]
Gene sets database	C:\Users\bennettb\Desktop\gsea\c2.cp.kegg.v5.0.symbols.gmt
Number of permutations	1000
Phenotype labels	C:\Users\bennettb\Desktop\gsea\class_labels.cls#NT_versus_LIN28
Collapse dataset to gene symbols	false
Permutation type	phenotype
Chip platform(s)	

Basic fields Hide

Analysis name	GSEA_analysis
Enrichment statistic	weighted
Metric for ranking genes	Signal2Noise
Gene list sorting mode	real
Gene list ordering mode	descending
Max size: exclude larger sets	500
Min size: exclude smaller sets	15
Save results in this folder	C:\Users\bennettb\Desktop\gsea\out

Advanced fields Show

GSEA reports

Processes: click 'status' field for results

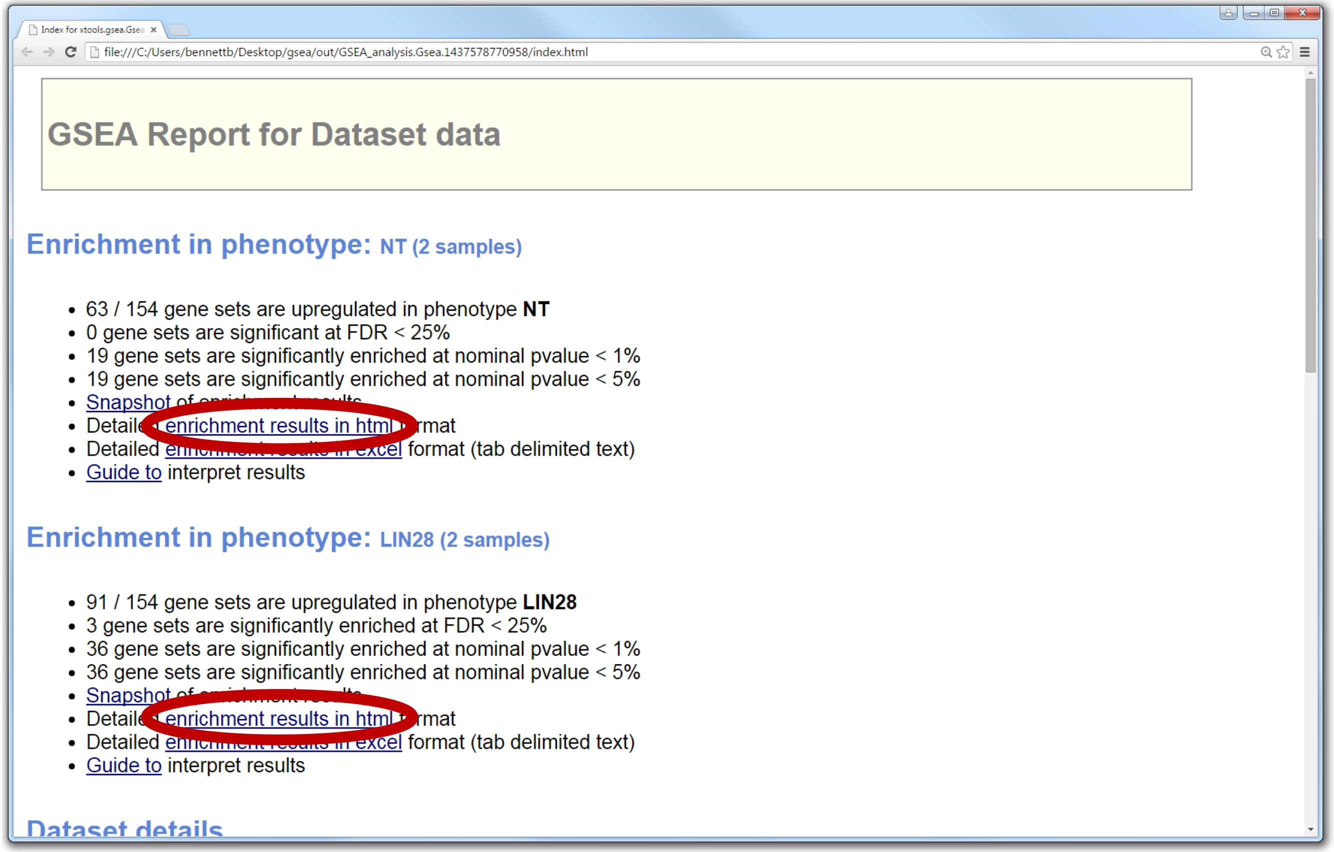
	Name	Status
1	Gsea	Running

11:24:40 AM 0634 [INFO] Done preproc for smaller than: 15

27M of 50M

GSEA Example

- Within output directory: index.html



Index for xtools.gsea.Gsea
file:///C:/Users/bennettb/Desktop/gsea/out/GSEA_analysis.Gsea.1437578770958/index.html

GSEA Report for Dataset data

Enrichment in phenotype: NT (2 samples)

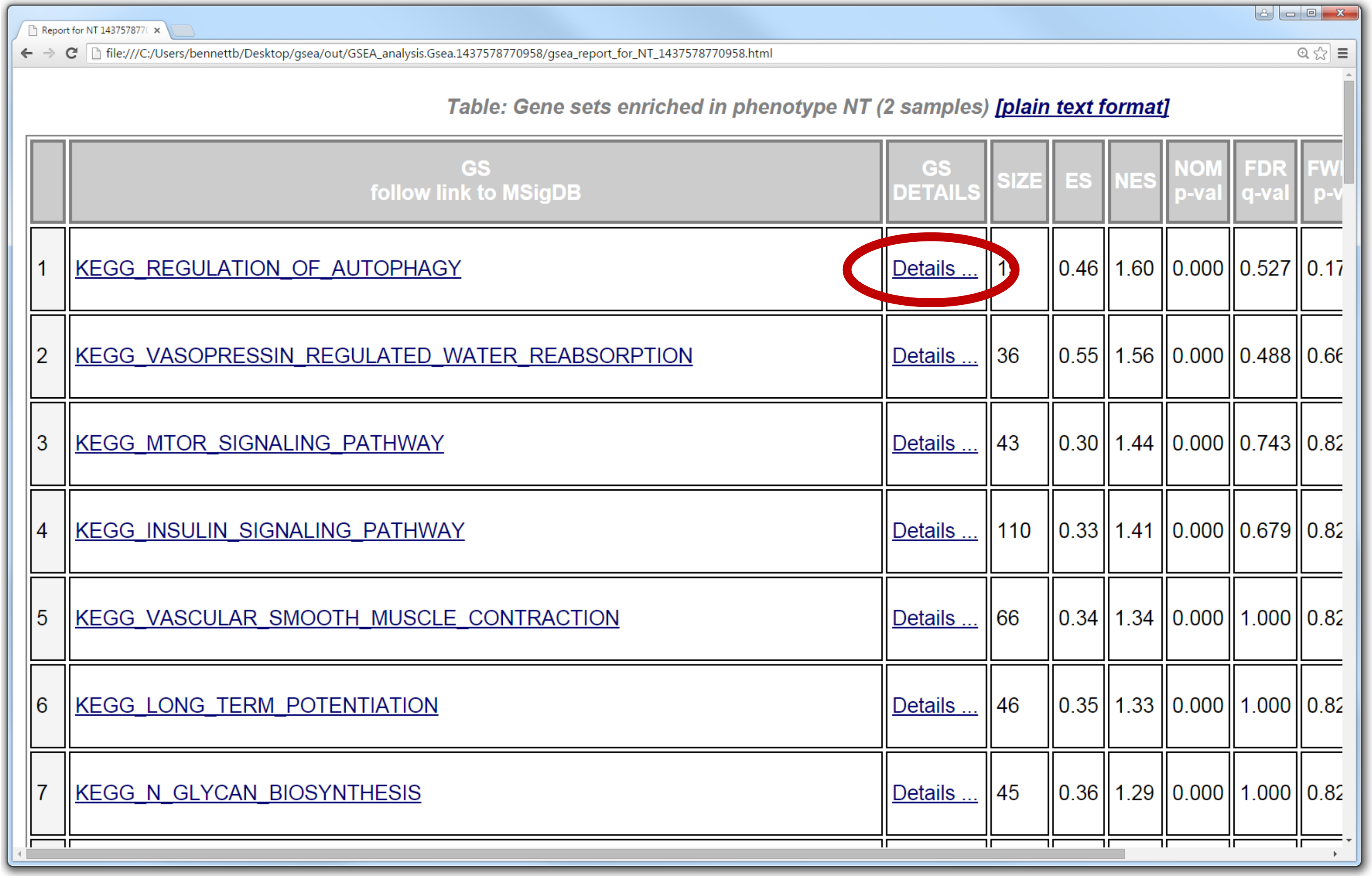
- 63 / 154 gene sets are upregulated in phenotype **NT**
- 0 gene sets are significant at FDR < 25%
- 19 gene sets are significantly enriched at nominal pvalue < 1%
- 19 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in excel format (tab delimited text)
- Guide to interpret results

Enrichment in phenotype: LIN28 (2 samples)

- 91 / 154 gene sets are upregulated in phenotype **LIN28**
- 3 gene sets are significantly enriched at FDR < 25%
- 36 gene sets are significantly enriched at nominal pvalue < 1%
- 36 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

GSEA Example



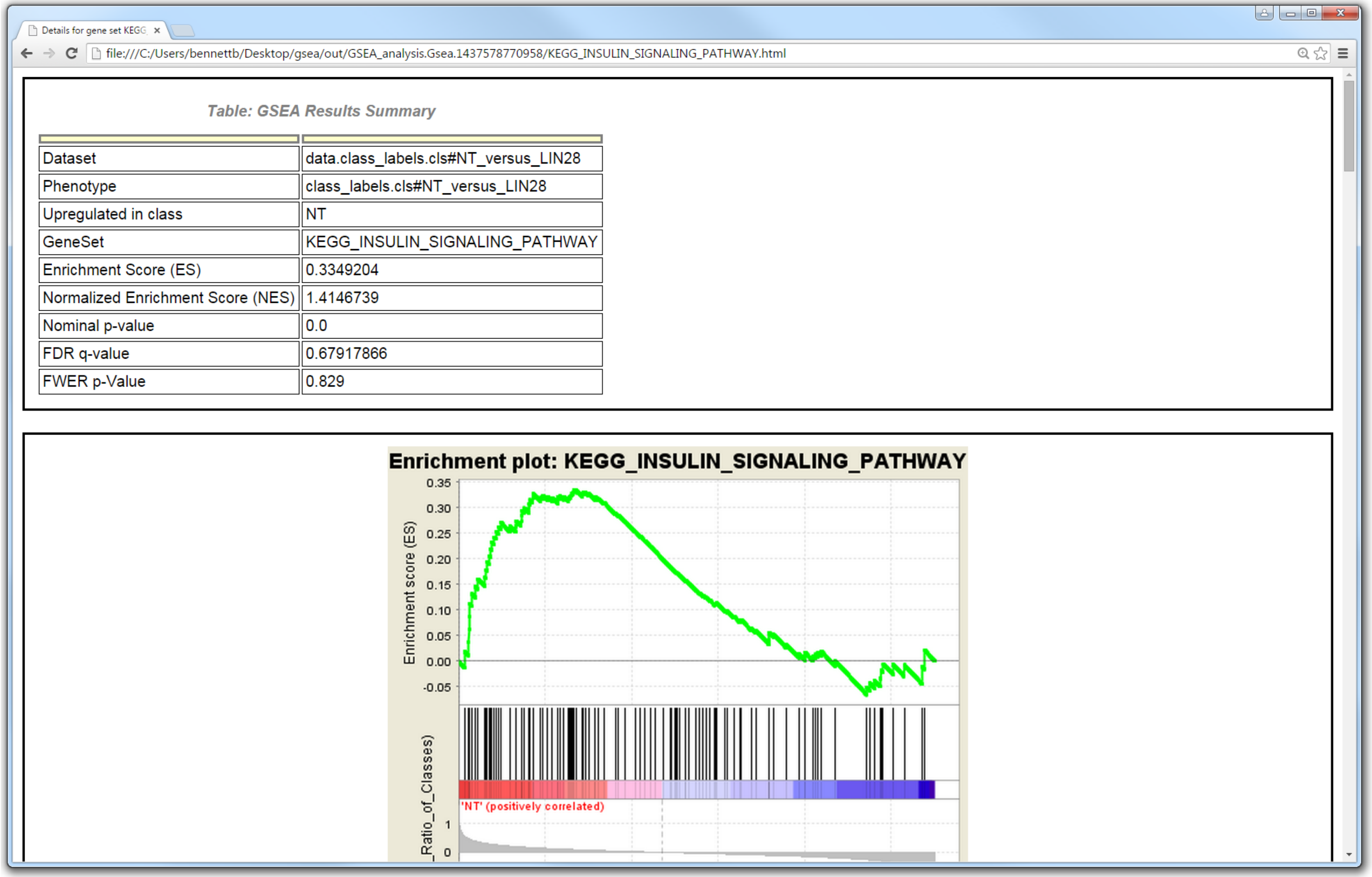
Report for NT 143757877(

file:///C:/Users/bennettb/Desktop/gsea/out/GSEA_analysis.Gsea.1437578770958/gsea_report_for_NT_1437578770958.html

Table: Gene sets enriched in phenotype NT (2 samples) ***[plain text format]***

	GS follow link to MSigDB	GS DETAILS	SIZE	ES	NES	NOM p-val	FDR q-val	FWI p-v
1	KEGG_REGULATION_OF_AUTOPHAGY	Details ...	1	0.46	1.60	0.000	0.527	0.17
2	KEGG_VASOPRESSIN_REGULATED_WATER_REABSORPTION	Details ...	36	0.55	1.56	0.000	0.488	0.66
3	KEGG_MTOR_SIGNALING_PATHWAY	Details ...	43	0.30	1.44	0.000	0.743	0.82
4	KEGG_INSULIN_SIGNALING_PATHWAY	Details ...	110	0.33	1.41	0.000	0.679	0.82
5	KEGG_VASCULAR_SMOOTH_MUSCLE_CONTRACTION	Details ...	66	0.34	1.34	0.000	1.000	0.82
6	KEGG_LONG_TERM_POTENTIATION	Details ...	46	0.35	1.33	0.000	1.000	0.82
7	KEGG_N_GLYCAN_BIOSYNTHESIS	Details ...	45	0.36	1.29	0.000	1.000	0.82

GSEA Example



Details for gene set KEGG
Enrichment plot: KEGG_INSULIN_SIGNALING_PATHWAY
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
-0.05
Enrichment score (ES)
_Ratio_of_Classes)
'NT' (positively correlated)
1
0

file:///C:/Users/bennettb/Desktop/gsea/out/GSEA_analysis.Gsea.1437578770958/KEGG_INSULIN_SIGNALING_PATHWAY.html

Table: GSEA Results Summary

Dataset	data.class_labels.cls#NT_versus_LIN28
Phenotype	class_labels.cls#NT_versus_LIN28
Upregulated in class	NT
GeneSet	KEGG_INSULIN_SIGNALING_PATHWAY
Enrichment Score (ES)	0.3349204
Normalized Enrichment Score (NES)	1.4146739
Nominal p-value	0.0
FDR q-value	0.67917866
FWER p-Value	0.829

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.)

Resources

- Pathway analysis tools
    - Gene Set Enrichment Analysis (GSEA)
        - http://www.broadinstitute.org/gsea
    - Ingenuity Pathway Analysis (IPA)
        - http://www.ingenuity.com/products/ipa
    - Database for Annotation, Visualization and Integrated Discovery (DAVID)
        - https://david.ncifcrf.gov
    - Gene Ontology Enrichment Analysis Software Toolkit (GOEAST)
        - http://omicslab.genetics.ac.cn/GOEAST

Resources

- Gene set and pathway databases
  - Molecular Signatures Database (MSigDB)
    - http://www.broadinstitute.org/gsea/msigdb
  - Kyoto Encyclopedia of Genes and Genomes (KEGG)
    - http://www.genome.jp/kegg
  - BioCarta
    - http://www.biocarta.com
  - Gene Ontology (GO)
    - http://geneontology.org
  - PANTHER GO-slim
    - http://www.pantherdb.org/panther/ontologies.jsp