Functional Analysis of Microarray Data Using GSEA

Aiguo Li, Ph.D. &
Alan Berger, Ph.D.

CIT Course 445

Aiguo Li contact info:
301-435-1454
liai@mail.nih.gov

Alan Berger contact info:
301-588-1469
aberger3527@comcast.net

May 15, 2008

Notes

- **Sign attendance list**
- **Ask questions by all means**
- **Fill out course evaluation forms after class is finished**
- **Version 13 file of class slides will be available from the CIT Course 445 description page**

Outline

- **Functional Analysis of Microarray Data – Analysis at the Level of *Gene Sets***
- **Introduction to GSEA (Gene Set Enrichment Analysis)**
- **<break>**
- **Installing GSEA: Desktop**
- **Running GSEA: Required Input Files & Parameter Selection; Broad Institute Utilities**
- **<break>**
- **Understanding the GSEA Outputs**
- **Live Demonstration Running Desktop GSEA**

Background

- **Genome-wide expression profiling with microarrays has become an effective frequently used technique in molecular biology**
- **Interpreting the results to gain insights into biological mechanisms remains a major challenge**
- **For a typical study (e.g., experimental condition vs. control, disease state vs. normal, tumor type A vs. tumor type B), a standard approach has been to produce a list of differentially expressed genes (DEGs) based on one or more criteria:**
  - statistically significant differential expression (e.g., by t-test)
  - sufficient level of fold change (up-regulated and/or down-regulated genes
  - sufficient expression level in at least one of the two classes

Challenges in Interpreting Gene Microarray Data

- **Even with DEG list(s) of up and/or down-regulated genes, still need to accurately extract valid biological inferences. Cutoff for inclusion in DEG lists is somewhat arbitrary. Must address multiple hypothesis testing.**
- **May obtain a long list of statistically significant genes without any obvious unifying biological theme**
- **May have few individual genes meeting the threshold for statistical significance**
- **When different groups study the same biological system, the lists of statistically significant genes from the two studies may show limited overlap because the number of samples were small, or the platforms were different**

An Existing Way to Study Enrichment of Gene Categories

- **Statistical procedures such as Fisher's exact test based on the hypergeometric distribution are used to test if members of a list of differentially expressed genes are overrepresented in given GO categories or in predefined gene sets compared with the distribution of the whole set of genes represented on the chip.**
- **Tools developed along this line include: GOMINER, GENMAPP, ONTO-TOOLs, CHIPINFO, GOSTAT.**

Example

Suppose 100 of 10,000 genes on a chip are in some pathway S while 5 (or more) of 50 genes in a particular differentially expressed gene (DEG) list are found to be in S: what is the probability P_S that this event occurred just by random chance. Note must correct for *multiple hypothesis testing* if examining multiple pathways. <P_S = 0.000134, using the hypergeometric distribution>.

One way to view this is think of there being 10,000 candies in a bin, 100 of which are Ghirardelli chocolates, and being given a random batch of 50 candies from the bin. If you got 5 or more of the chocolates, were you unusually lucky? <Indeed yes!>

Fisher Exact Viewpoint: 2×2 Contingency Table

	in pathway S	not in pathway S	
in DEG list	5	45	50
not in DEG list	95	9855	9950
Totals	100	9900	10000

Limitations with Category Enrichment Methods[1]

- **No further use made of information contained in expression values for the non-DEG list genes**
- **The level of differential expression of the genes in the significant gene list is not taken into consideration: simply counting the number of the differentially expressed genes that are contained in each category being considered does not make full use of available information**
- **The correlation structure of the expression data is not considered at all.**

[1]cf. the discussion in Tian et al. ref. [4] three pages below

Approach of Gene Set Enrichment Methods

- **These methods formulate a statistic reflecting the difference in expression level between the two phenotypes under consideration for the ensemble of genes in each gene set being considered**
- **The levels of differential expression for all the genes in the chip are utilized**
- **Can be applied to gene sets from, e.g., pathways in BioCarta & KEGG; genes co-located in cytobands; genes having common transcription factor motifs; genes changed in response to some disease state or experimental condition**
- **But note: results depend on the collection of gene sets examined, and still must address multiple testing error control (though much less severe than for DEG lists)**

Several Leading Gene Set Analytical Tools

Method	Type	Statistics	Implementation	References
GSEA	Non-parametric	Kolmogorov-Smirnov	Java & R package	Subramanian et al., PNAS, 2005
GSA	Parametric	Maxmean / t-statistics	R package	Efron & Tibshirani, Annals Appl. Stat. 2007
PAGE	Parametric	Z score	Python or JMP	Kim & Volsky, BMC Bioinformatics, 2005
SigPathway	Parametric	t-statistics	R package	Tian et al., PNAS 2005

Gene Set Analysis Literature

1. **Gene Set Enrichment Analysis (GSEA): *A knowledge-based approach for interpreting genome-wide expression profiles*, Subramanian et al., PNAS 2005, 102:15545; note Lamb et al., The Connectivity Map…, Science 2006, 313:1929. (see Broad Institute web page for this and other software)**

2. ***On testing the significance of sets of genes*, Efron and Tibshirani, The Annals of Applied Statistics 2007, 1:107 (Gene-Set Analysis (GSA)) (*maxmean* statistic; *restandardization* – use of gene permutations in conjunction with sample label permutation to improve statistical behavior).**

3. ***Parametric Analysis of Gene-Set Enrichment (PAGE), Kim and Volsky, BMC Bioinformatics 2005, 6:144 (uses the average of the measure of differential expression (DE) of genes in a gene set, and values of DE over the chip to get a *gene set z-score*).**

4. **Discovering statistically Significant Pathways in expression profiling studies, Tian et al., PNAS 2005, 102:13544 (uses both row (gene) and column (phenotype) permutations and statistical procedures to account for correlations among the gene expression profiles). see http://www.chip.org/~ppark/Supplements/PNAS05.html for SigPathway software**

Functional Annotation: Gene Network

- Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. stat Appl Genet Mol Biol. 2005;4:17.

- Aoki K, Ogata Y, Shibata D. Approaches for extracting practical information from gene co-expression networks in plant biology.
  Plant Cell Physiol. 2007 Mar;48(3):381-90. Epub 2007 Jan 23. Review
- Dong J, Horvath S. Understanding Network Concepts in Modules. BMC Syst Biol. 2007 Jun 4;1(1):24

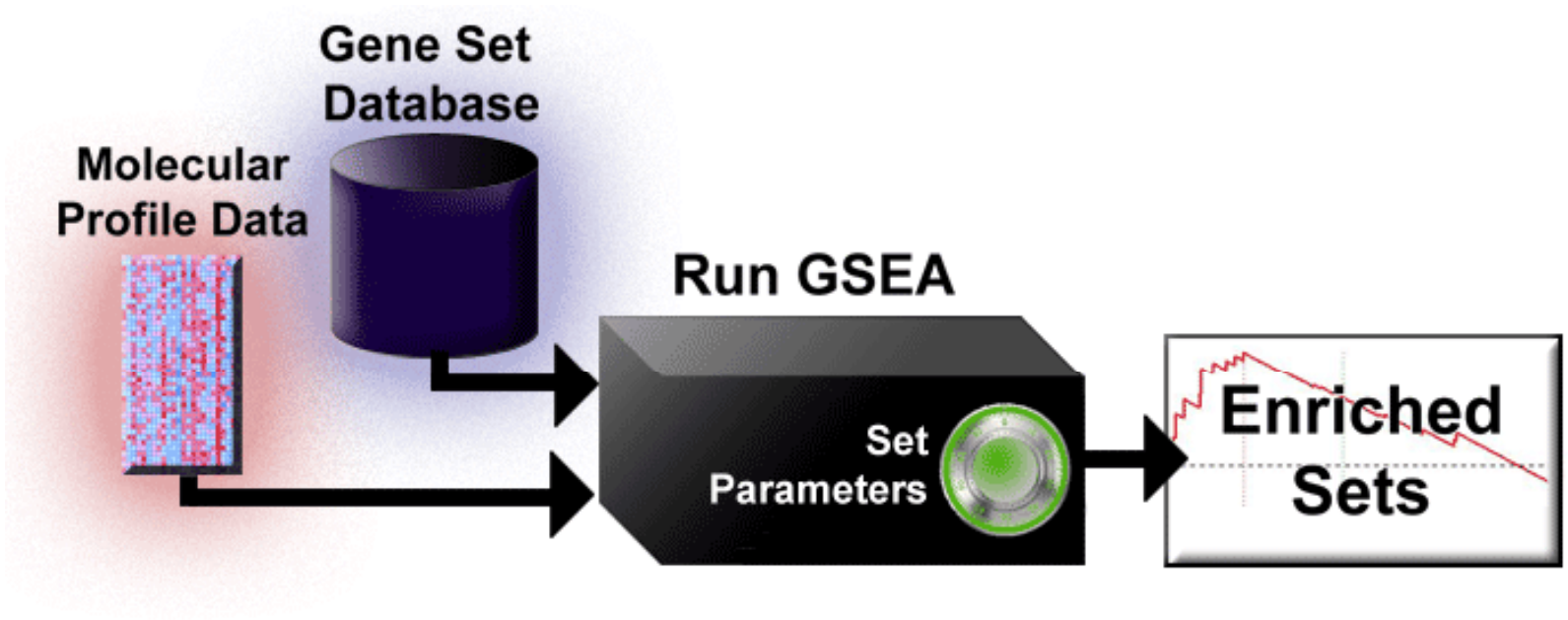
- Ghazalpour A, Doss S, Zhang B, Wang S, Plaisier C, Castellanos R, Brozell A, Schadt EE, Drake TA, Lusis AJ, Horvath S. Integrating genetic and network analysis to characterize genes related to mouse weight. PLoS Genet. 2006 Aug 18;2(8):e130.

Outline

- **Functional Analysis of Microarray Data – Analysis at the Level of *Gene Sets***
- **Introduction to GSEA (Gene Set Enrichment Analysis)**
- **<break>**
- **Installing GSEA: Desktop**
- **Running GSEA: Required Input Files & Parameter Selection; Broad Institute Utilities**
- **<break>**
- **Understanding the GSEA Outputs**
- **Live Demonstration Running Desktop GSEA**

GSEA Overview -- Workflow

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).



text and figure from the Broad Institute web pages for GSEA : http://www.broad.mit.edu/gsea/index.html

Three Main Components in GSEA

- Algorithm
- Software implementation (Broad Institute)
- Database: Molecular signature database (MSigDB at Broad Institute) containing gene sets of interest; also, utilities mapping chip features to genes (e.g., Affymetrix probe set IDs to HUGO gene symbols)

GSEA Summary

Gene Set Enrichment Analysis

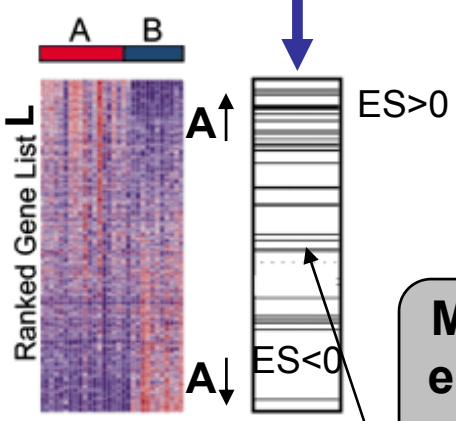
Get ranked list L of all the genes on the chip based on a chosen measure of the difference of their expression levels between the phenotypes A & B under study, e.g., tumor vs. normal

For each gene set S: find the location of each gene s in S within L

Generate enrichment score ES for S based on running-sum statistic: "reward" presence of s toward top or bottom of L

Analyze significance of this Kolmogorov-Smirnov type statistic by permutation testing

Multiple hypothesis testing (MHT) error control for multiple S's using the false discovery rate (FDR)



locations in L of genes from S

Genes at top of L are upreg. in class A vs. class B; if genes of S are conc. there, then ES(S) will be > 0.

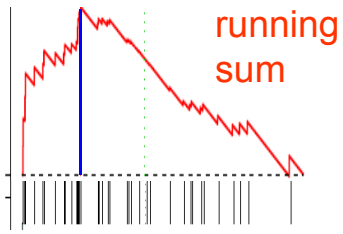
GSEA Algorithm: Three Elements

(Subramanian et al 2005)

- **Calculate an enrichment score (ES) for each gene set S: walking from top to bottom down the ranked gene list L, increase a running-sum statistic each time encounter a gene g in S and decrease it when encounter genes not in S. The size of each positive increment depends on g's degree of differential expression. The enrichment score is the maximum deviation from zero encountered in the walk down L; it corresponds to a weighted Kolmogorov-Smirnov-like statistic.**
- **Estimation of the Significance Level of each ES: Permute the phenotype labels and re-compute the ES for all the gene sets for the permuted data, which generates a null distribution for the ES. This is used to calculate normalized enrichment scores (NES) for the gene sets S, and an empirical null distribution for NES and each ES(S). The empirical, nominal p value for each ES(S) is then calculated relative to the null distr. for ES(S). (Calcs. split for +, - ES's, NES's)**
- **Adjustment for multiple hypothesis testing (MHT) of the entire database of gene sets being considered: GSEA uses the NES(S) values for the gene sets under consideration, and the empirical null distribution for NES to compute an estimated false discovery rate (FDR) for each gene set S. (FDR = expected fraction of false discoveries among the gene sets declared to be significantly differentially expressed between the two phenotypes under study, e.g., 40 declared, 10 falsely $\longrightarrow$ FDR = 25%) e.g., declare all S with NES $\geq \gamma$ or NES $\leq -\gamma$**

Enrichment Score Calculation Example

Schematic Example: N = 1020 genes in chip; 20 genes in a pathway S. Suppose the sum Σ over the N_H = 20 genes g in S of the |measure of differential expression DE(g)| between the phenotypes A & B = 100. Color locations of genes in S in red, locations of genes not in S in black.



ranked list	|DE|	contribution to running sum for ES	running sum for ES
	15	+ 0.15	+ 0.15
	12	+ 0.12	+ 0.27
	10	- 0.001	0.27 - 0.001
	9	+ 0.09	0.36 - 0.001
	8	+ 0.08	0.44 - 0.001
	6	- 0.001	0.44 - 0.002
...			

$+|DE| / \Sigma$

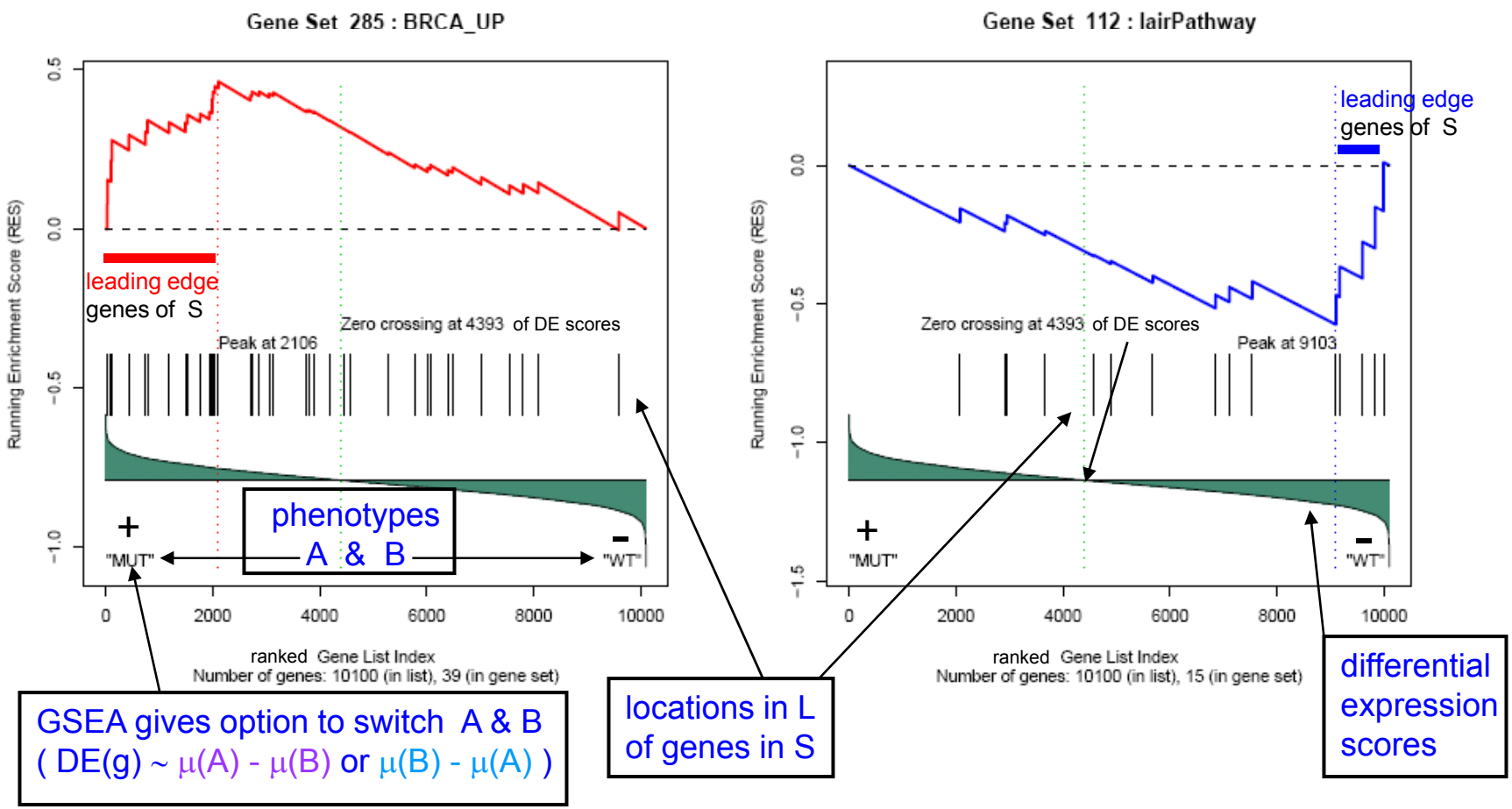
or

$-1/(N-N_H)$

ES(S) ≡ value of max deviation from 0 (*extr*) of the running sum

Annotated examples of ES calculations using plots from GSEA analysis of the P53 NCI-60 data set

(files from the Broad GSEA site)



from P53 data analysis from GSEA–R distribution from Broad Institute GSEA web page

GSEA Algorithm: Definition of Enrichment Scores

w_k **= measure of differential expression of gene k; order genes in ranked list so** w_k **decreases from the top (k=1) to the bottom (k=N) of the list**

$$p_{hit}(S,i) = \sum_{\substack{g_j \in S \\ j \le i}} \frac{|w_j|^p}{N_R} \quad \text{where} \quad N_R = \sum_{g_k \in S} |w_k|^p$$

for GSEA default is *p* = 1, for Kolmogorov-Smirnov *p* = 0

$$P_{miss}(S,i) = \sum_{\substack{g_i \notin S \\ j \le i}} \frac{1}{(N - N_H)},$$

$N_H =$ # genes in S

N = # genes in chip

$ES(S,i) = P_{hit}(S,i) - P_{miss}(S,i)$; $ES(S) = \text{extr}_i(ES(S,i))$

How NES(S) is calculated from ES(S)

Given a gene set S, one calculates ES(S) as just shown. To generate an empirical null distribution for ES(S), one uses a random number generator to create a set of $\mathcal{N}$ permutations π of the class labels for the expression data. (If insufficient # of samples for phenotype perm. then do gene_set perm.) For each of these permutations one calculates ES(S, π) as before (except now using the permuted phenotype labels). Then:

$$\mathbf{NES(S)} \equiv \frac{\textbf{original ES(S)}}{\textbf{mean}_{\pi}\textbf{(the ES(S, }\pi\textbf{) values with the same sign as ES(S))}}$$

and for use in calculating FDRs

$$\mathbf{NES(S, \pi_p)} \equiv \frac{\textbf{original ES(S, }\pi_p\textbf{)}}{\textbf{mean}_{\pi}\textbf{(the ES(S, }\pi\textbf{) values with the same sign as ES(S, }\pi_p\textbf{))}}$$

Outline

- **Functional Analysis of Microarray Data – Analysis at the Level of *Gene Sets***
- **Introduction to GSEA (Gene Set Enrichment Analysis)**
- **<break>**
- **Installing GSEA: Desktop**
- **Running GSEA: Required Input Files & Parameter Selection; Broad Institute Utilities**
- **<break>**
- **Understanding the GSEA Outputs**
- **Live Demonstration Running Desktop GSEA**

Three Main Components in GSEA

- Algorithm
- Software implementation (Broad Institute)
- Database: Molecular signature database (MSigDB at Broad Institute) containing gene sets of interest; also, utilities mapping chip features to genes (e.g., Affymetrix probe set IDs to HUGO gene symbols)

GSEA - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites

Address http://www.broad.mit.edu/gsea/ Go Links

BROAD INSTITUTE

GSEA

Gene Set Enrichment Analysis

Search

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).

What's New

A new release of the Molecular Signatures Database (MSigDB) is now available. The release includes new gene sets based on KEGG pathways, GO annotations, and the module map for cancer compiled by Segal et al. (Nature Genetics 36, 1090 - 1098, 2004). If you have questions, please contact us at: gsea@broad.mit.edu.

Free, will ask for email address

Getting Started

A quick tutorial to get you up and running.

Tools and Information

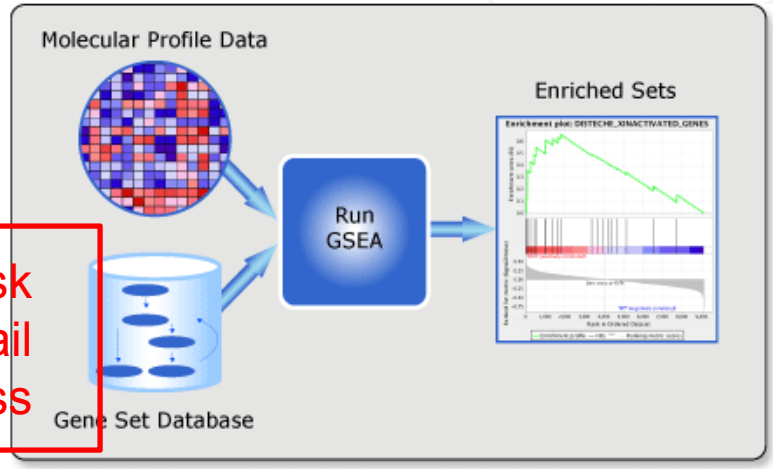
Downloads: Implementations of GSEA plus additional resources to analyze, annotate and interpret enrichment results.

Molecular Signatures Database: A collection of gene sets for use with GSEA software and tools for exploring them.

Documentation: Information on the GSEA software, the GSEA algorithm.

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

GSEA is maintained by the GSEA team. Our thanks to our many contributors. Funded by: National Cancer Institute, National Institutes of Health, National Institute of General Medical Sciences.

Citing GSEA

To cite your use of the GSEA software, please reference Subramanian, Tamayo, et al. (2005, PNAS 102, 15545-15550) and Mootha, Lindgren, et al. (2003, Nat Genet 34, 267-273).

Right click on appropriate Launch (JAVA) icon, save .jnlp file to Desktop, run, accept certificate – it installs for the current user (do not need Administrator privileges)

GSEA | Downloads - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites

Address http://www.broad.mit.edu/gsea/downloads.jsp Go Links

logged in as alanberger@alum.mit.edu
logout

BROAD INSTITUTE

GSEA
Gene Set Enrichment Analysis

Search

GSEA Home | Downloads | Molecular Signatures Database | Documentation | Contact

Downloads

Software

There are several options for GSEA software. All options implement exactly the same algorithm. Usage recommendations and installation instructions are listed below.

R-GSEA
R Package

- Usage from within the R programming environment
- Easily inspect, learn and tweak the algorithm
- Incorporate GSEA into your own data analysis pipeline
- Programmatically call the open source GSEA R API

download GSEA-P-R.1.0.zip

Many firewalls block direct click here and download

javaGSEA
Desktop Application

- Easy-to-use graphical user interface
- Runs on any desktop computer (Windows, Mac OSX, Linux etc.)
- 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
- The GSEA team suggests always starting GSEA by using these Launch buttons, or by clicking the icon that the application installs on your desktop, in order to ensure optimal memory allocation.

Launch with 512Mb memory
Launch

Launch with 1Gb memory
Launch

Release Notes - GeneSetEnrichmentAnalysisWiki - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites

Address http://www.broad.mit.edu/cancer/software/gsea/wiki/index.php/Release_Notes Go Links

Log in

GSEA
Gene Set Enrichment Analysis

- Third Fake Nav Item

navigation

- Documentation Home
- Tutorial ^
- User Guide ^
- Data Formats
- FAQ
- Known Issues
- Algorithm
- PNAS 2005 Examples ^
- Papers that use GSEA

msigdb

- Gene Set Pages
- Document Type Definition (DTD) for MSigDB
- Release Notes
- Mapping between v1 and v2 gene sets
- MSigDB License
- MSigDB Acknowledgements

software

- Release Notes
- R-GSEA Readme
- Java Source Code
- JavaDoc^
- Software License
- Software Acknowledgements
- For Broad Users

internal only

- Wiki How-to-Use
- Bugs and Enhancements
- Context Sensitive

Release Notes

GSEA Home | Downloads | Molecular Signatures Database | Documentation | Contact

GSEA Software Release Notes

[edit]

Date	Release	Description	Release Notes
Feb 2007	2.0.1*	Fixed error running leading edge analysis	wiki
Jan 2007	2.0	Major release with significant enhancements	html
Mar 2005	1.0	Initial release	Not available

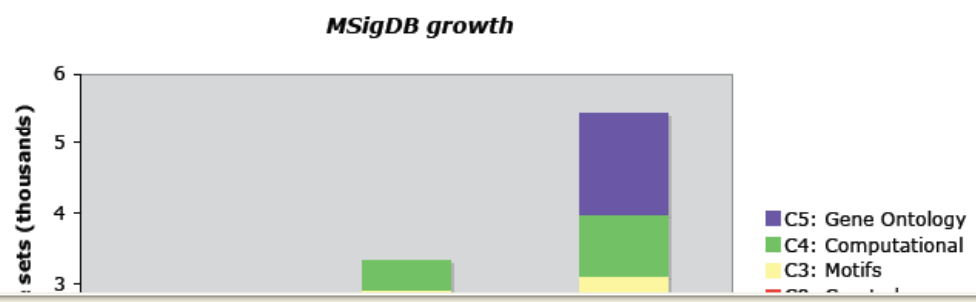
MSigDB Release Notes

[edit]

Date	Release	Description	Release Notes
April 2008	2.5*	C1 (+0); C2 (+205); C3 (+0); C4 (+456); C5 (+1454)	wiki
Feb 2007	2.1	Minor updates to MSigDB v2.0 annotations	
Jan 2007	2.0	C1 (updated); C2 (+269); C3 (+214); C4 (+0)	wiki
Nov 2005	1.1	C1 (updated); C2 (+350); C3 (+566); C4 (+0)	pdf
March 2005	1.0	Initial release	pdf

* Current release

MSigDB Statistics



Done Internet

GSEA | Downloads - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites

Address http://www.broad.mit.edu/gsea/downloads.jsp Go Links

For details on the GSEA algorithm and software refer to the Documentation.
For details on the latest release refer to the Release Notes.

Broad Gene Sets Data Base

MSigDB

Use the following links to download individual gene set collections or the complete Molecular Signatures Database (MSigDB).

c1: positional gene sets	C1 gene sets file	c1.all.v2.5.symbols.gmt
c2: curated gene sets	C2 gene sets file	c2.all.v2.5.symbols.gmt
	canonical pathway gene sets gene sets file	c2.cp.v2.5.symbols.gmt
	chemical and genetic perturbations gene sets file	c2.cgp.v2.5.symbols.gmt
	BioCarta gene sets file	c2.biocarta.v2.5.symbols.gmt
	GenMAPP gene sets file	c2.genmapp.v2.5.symbols.gmt
	KEGG gene sets file	c2.kegg.v2.5.symbols.gmt
c3: motif gene sets	C3 gene sets file	c3.all.v2.5.symbols.gmt
	transcription factor targets gene sets file	c3.tft.v2.5.symbols.gmt
	microRNA targets gene sets file	c3.mir.v2.5.symbols.gmt
c4: computational gene sets	C4 gene sets file	c4.all.v2.5.symbols.gmt
	cancer gene neighborhoods gene sets file	c4.cgn.v2.5.symbols.gmt
	cancer modules gene sets file	c4.cm.v2.5.symbols.gmt
c5: gene ontology gene sets	C5 gene sets file	c5.all.v2.5.symbols.gmt
	GO biological process gene sets file	c5.bp.v2.5.symbols.gmt
	GO cellular component gene sets file	c5.cc.v2.5.symbols.gmt
	GO molecular function gene sets file	c5.mf.v2.5.symbols.gmt
XML database file (all gene sets)	Current MSigDB xml file	msigdb_v2.5.xml
Archived release	MSigDB version 2.1	February 2007
	MSigDB version 1.0 - described first in the PNAS paper Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles, Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee,	March 2005

Internet

MSigDB: C2 Curated Gene Sets

- **BioCarta** - http://www.biocarta.com
- **Signaling pathway database** - http://www.grt.kyushu-u.ac.jp/spad/menu.html
- **Signaling gateway** - http://www.signaling - gateway.org/
- **Signal transduction knowledge environment-** **http://stke.sciencemag.org/**
- **Human protein reference database** - http://www.hprd.org/
- **GenMAPP** - http://www.genmapp.org/
- **KEGG** - http://www.genome.jp/kegg/
- **Gene ontology** - http://www.geneontology.org
- **Sigma-Aldrich pathways** - http://www.sigmaaldrich.com/Area of_Interest/Biochemicals/Enzyme_Explorer/KeyResources.html
- **Gene arrays, BioScience Corp** - http://www.superarray.com/
- **Human cancer genome anatomy consortium** - [http://cgap.nci.nih.gov/ http://cgap.nci.nih.gov/]
- **NetAffx** - http://www.affymetrix.com/index.affx

C3 Transcription factor & miRNA Targets

- **TFT – Transcription Factor Targets**
  - **Each of these gene sets is annotated by a TRANSFAC record.**
  - **Gene sets containing genes that share a transcription factor binding site defined in the TRANSFAC (version 7.4, http://www.gene -regulation.com/) database.**
- **MIR: Gene sets that contain genes that share a 3'-UTR microRNA binding motif.**

MSigDB: C4 Computed Gene Sets

- **Brentani et al., 2003 PNAS, 100:13418-13423, The generation and utilization of a cancer-oriented representation of the human transcriptome by using expressed sequence tags**
- **380 cancer associated genes curated from this paper and then neighborhoods were defined around these genes by Pearson correlation with a cutoff of R >= 0.85 by using four large gene expression data sets. Therefore, a given oncogene may have up to four "types" of neighborhoods according to the correlation present in each compendium**
- **C4 includes 427 gene sets (neighborhoods with <25 genes were omitted)**

Annotation Tools - Functions

Annotations

1) Browsing function

2) Searching function

Explore gene set annotations to gain further insight into the biology behind a gene set in question:

- compute overlaps with other gene sets in MSigDB (details)
- categorize members of the gene set by gene families (details)
- display the gene set expression profile based on a selected compendium of expression profiles (details)

3) Advanced analysis function

Genes to annotate

ANK1
BACH1
EPB49
FTL
GSTT2
GYPA
HBA2
HBB
HBZ
HEBP1
HEBP2
HMBS
KLF1
MAFB
MAFF
MAFG

Compute overlaps with

- [ ] C1: Chromosomal locations
- [ ] C2: Curated gene sets
  - [ ] CP: Canonical pathways
  - [ ] CGP: Chemical and genetic perturbations
- [x] C3: Motif gene sets
  - [ ] TFT: Transcription factor targets
  - [x] MIR: miRNA targets
- [ ] C4: Computed gene sets

overlaps: view download

show top 10 genesets [x] clustered

Compendia expression profiles

- ( ) Human tissue compendium (Novartis)
- ( ) Global Cancer Map (Broad Institute)
- ( ) NCI-60 cell lines (National Cancer Institute)

display expression profile

Gene families

show gene families

Chip format

GENE SYMBOL

Outline

- **Functional Analysis of Microarray Data – Analysis at the Level of *Gene Sets***
- **Introduction to GSEA (Gene Set Enrichment Analysis)**
- **<break>**
- **Installing GSEA: Desktop (JAVA)**
- **Running GSEA: Required Input Files & Parameter Selection; Broad Institute Utilities**
- **<break>**
- **Understanding the GSEA Outputs**
- **Live Demonstration Running Desktop GSEA**

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

File Options Downloads Tools Help

Menu bar

Help information

Steps in GSEA analysis

Load data

Run GSEA

Leading edge analysis

Action Buttons

Gene set tools

Chip2Chip mapping

Browse MSigDB

Analysis history

GSEA reports

Processes: click 'status' field for results

Name	Status

Show results folder

Home

Steps in GSEA

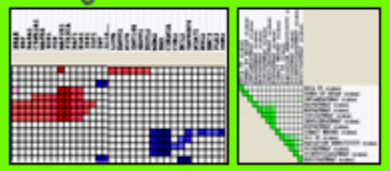
1. What you need for GSEA:

- Expression dataset
- Phenotype file
- Gene sets (from MSigDB or your own gene sets)

- Start with default parameters
- If you want to collapse probes to genes, specify chip platform

3. View results & leading edge

- Leading edge finds genes driving enrichment results

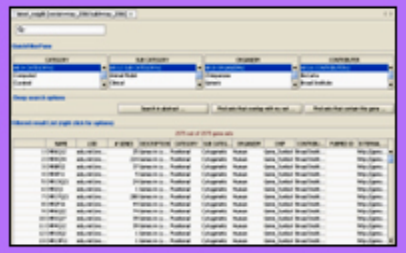


Gene Sets Browser

- Browse gene sets in MSigDB

Browse MSigDB

- Search the database of ~2500 gene sets



- Chip2Chip converts gene sets between platforms

Chip2Chip mapping

- Export gene sets for analysis with GSEA or with other programs

Export sets as GeneSetMatrix

Getting Help

GSEA website

www.broad.mit.edu/gsea

GSEA Wiki

www.broad.mit.edu/gsea/wiki

Email the GSEA team at

gsea@broad.mit.edu

BROAD INSTITUTE

11:05:38 AM | 5156 [INFO] Made Vdb dir: C:\Documents and Settings\liai\gsea_home\output\aug02 | 57M of 63M

Outline

- **Functional Analysis of Microarray Data**
- **Introduction to GSEA (Gene Set Enrichment Analysis)**
- **<break>**
- **Running desktop GSEA:**
  - **Required Input Files - GSEA file formats**
  - **Broad Institute Utilities – Gene Set & Chip Files**
  - **Parameter Selection**
  - **Running GSEA**
- **<break>**
- **Understanding the GSEA Outputs**
- **Live Demonstration Running Desktop GSEA**

Main Data Files for Input

- **Resource: http://www.broad.mit.edu/cancer/software/gsea/wiki/index.php/Data_formats**
- **Gene sets database files (several options)**
  - **GeneMatrix (gmt) from Broad ftp site**
  - **GeneSets (grp): single gene set in a simple newline-delimited text format**
  - **GeneMatrix (gmt) from local machine**
- **Data files (gene expr. data or ranked list, several format options)**
  - **Gene Cluster Text file: *.gct**
  - **Ranked list file format: *.rnk**
  - **ExpRESsion (with P and A calls) file (*.res)**
  - **Stanford cDNA file format (*.pcl)**
  - **Text file format for expression dataset (*.txt)**
- **Phenotype variables (specify information for each sample)**
  - **Categorical (e.g tumor vs normal) class file format (*.cls)**
  - **Continuous (e.g time-series or gene profile) file format (*.cls)**
- **The Broad GSEA web site has good documentation and tutorials**
  - **http://www.broad.mit.edu/cancer/software/gsea/doc/GSEAUserGuideFrame.html**
  - **http://www.broad.mit.edu/cancer/software/gsea/wiki/index.php/Main_Page see the "navigation" panel**

gct & res Expression Data File Formats (tab delimited text files, displayed here using Excel):

	A	B	C	D	E	AX	AY	AZ	B
1	#1.2								
2	10100	50							
3	NAME	DESCRIPTION	786-0	BT-549	CCRF-CEM	UACC-257	UACC-62	UO-31	
4	TACC2	na	46.05	82.17	16.87	32.16	45.7	48.13	
5	C14orf132	na	108.3	59.04	25.61	102.7	62.16	73.44	
6	AGER	na	42.2	25.75	76.01	56.57	50.4	36.75	

*.gct file: gives feature identifiers in column 1 and gene expression data

	A	B	C	D	E	F	G	H
1	Description	Accession	ALL_19769		ALL_23953		ALL_28373	
2		CH1999021515AA		CH199902151 1AA/scale fa		CH1999021507AA/sc		CH199902
3	1000							
4	Semaphorin E	AB000220_at	36	A	39	A	39	A
5	MNK1	AB000409_at	-299	A	-11	A	237	P
6	VRK1	AB000449_at	57	A	274	P	311	P
7	VRK2	AB000450_at	186	P	245	P	186	P
8	mRNA, clone RES4-	AB000460_at	1647	P	2128	P	1608	P
9	SH3 binding protein,	AB000462_at	137	A	-82	A	204	P
10	mRNA, clone RES4-	AB000464_at	803	P	1489	P	322	P
11	mRNA, clone RES4-	AB000466_at	-894	A	-969	A	-444	A
12	mRNA, clone RES4-	AB000467_at	-632	A	-909	A	-254	P

*.res file: ExpRESsign with P, A, M calls and gene expression data

Screen Image of P53.gct file (gene cluster text tab delimited file)

required, always the same

features or genes

NO Dashes (–) allowed in GSEA file names (Java)

samples (chips)

sample identifiers, must be unique

chip feature IDs (unique)

DESCRIPTION

can be a non-blank filler

Expression levels

	A	B	C	D	E	AX	AY	AZ
1	#1.2							
2	10100	50						
3	NAME	DESCRIPTION	786-0	BT-549	CCRF-CEM	UACC-257	UACC-62	UO-31
4	TACC2	na	46.05	82.17	16.87	32.16	45.7	48.13
5	C14orf132	na	108.3	59.04	25.61	102.7	62.16	73.44
6	AGER	na	42.2	25.75	76.01	56.57	50.4	36.75
7	32385_at	na	7.43	13.94	8.55	4.5	14.59	11.33
10098	CRYBA1	na	7.28	5.92	19.17	18.13	4.66	9.49
10099	CTNND2	na	9.27	6.57	10.86	5.6	11.9	6.22
10100	SEZ6L	na	96.07	76.24	49.7	90.68	84.98	66.05
10101	EIF3S2	na	545.5	508.67	625.43	1000	1224.8	274.1
10102	AMACR	na	64.63	28.45	27.48	35.21	89.32	48.3
10103	LDLR	na	148.7	716.39	222.09	281.4	283.19	549.9
10104								
10105								

Case Sensitive

expression levels (or logs), for a missing value leave cell empty

P53.gct tab delimited text file (displayed here in **Excel**) from the R-GSEA distribution from the Broad Institute: http://www.broad.mit.edu/gsea/index.html

Gene Set File Formats

- A gene sets file is a tab-delimited text file in gmx or gmt format. The specific gene set file formats can be found at http://www.broad.mit.edu/gsea/wiki/index.php/Data_formats
- For desktop version you can
  - Select gene sets file from Broad Inst. ftp web site
  - Export gene sets from MSigDB using browse MSigDB function
- Create gene sets using a text editor or Excel

Schematic of a **.gmt** Gene Matrix Transposed Gene Sets file (Each Row is 1 Gene Set) (tab delimited text file)

Gene Set Names, one per row, names must be unique

Gene Set Description, could be just "na"

Gene Identifiers, number of genes per gene set can vary, **Case Sensitive**

	A	B	C	D	E	F	G	H
1	41bbPathway	TNF-type recept	IL2	TRAF2	MAP3K1	IFNG	CHUK	NFKBIA
2	ace2Pathway	Angiotensin-con	COL4A3	COL4A1	COL4A5	AGT	COL4A6	AGTR1
3	acetaminophenPath	Acetaminophen	CYP3A	PTGS2	CYP1A2	PTGS1	NR1I3	CYP2E1
4	achPathway	Nicotinic acetylc	RAPSN	TERT	MUSK	PTK2		
5	actinYPathway	The Arp 2/3 com	ACTR3	ABI-2	WASL	ARPC4	NTRK1	ARPC3
6	agpcrPathway	G-protein couple	PRKAR2A	GNGT1	PRKACB	PRKCB1	PRKCA	PRKAR2B
7	ahspPathway	Alpha-hemoglobi	CPO	HMBS	ALAS1	ERAF	HBA2	ALAD
8	aifPathway	aif	ADPRT	PDCD8	BCL2L1	CYCS		
9	akap13Pathway	A-kinase anchor	EDG4	PRKACG	PRKAR2A	PRKACB	PRKAG1	GNA12
10	akap96Pathway	BioCarta	CDC2	DDX5	PRKACG			
11	akapCentrosomePa	Anchoring protei	PRK	AKAP9	PRKACB	PRKAR2B	PCNT2	PRKCL1
12	aktPathway	Second messen	FOXO1A	CASP9	PDPK1	GHR	CHUK	NFKBIA

Gene Identifiers can be probe set IDs or gene symbols but MUST BE CONSISTENT WITH column 1 of the .gct file (the chip feature IDs). If using a **.chip** file within Java-GSEA then col 1 of the .gct and .chip files must correspond & the gene symbols here should be those used within the Gene Symbol column of the .chip file.

Subset of the HG_U133A chip file from the Broad Web Page

HG_U133A.chip

	A	B	C
1	Probe Set	Gene Sym	Gene Title
2	1007_s_at	DDR1	discoidin domain receptor family, member 1
3	1053_at	RFC2	replication factor C (activator 1) 2, 40kDa
4	117_at	HSPA6 ///	heat shock 70kDa protein 6 (HSP70B') /// similar to heat shock 70kDa protein 6 (HSP70B)
5	121_at	PAX8	paired box gene 8
6	1255_g_at	GUCA1A	guanylate cyclase activator 1A (retina)
7	1294_at	UBE1L	ubiquitin-activating enzyme E1-like
8	1316_at	THRA	thyroid hormone receptor, alpha (erythroblastic leukemia viral (v-erb-a) oncogene homolog, avian)
9	1320_at	PTPN21	protein tyrosine phosphatase, non-receptor type 21
10	1405_i_at	CCL5	chemokine (C-C motif) ligand 5
11	1431_at	CYP2E1	cytochrome P450, family 2, subfamily E, polypeptide 1
12	1438_at	EPHB3	EPH receptor B3
13	1487_at	ESRRA	estrogen-related receptor alpha
14	1494_f_at	CYP2A6	cytochrome P450, family 2, subfamily A, polypeptide 6
15	1598_g_at	GAS6	growth arrest-specific 6
16	160020_at	MMP14	matrix metallopeptidase 14 (membrane-inserted)
17	1729_at	TRADD	TNFRSF1A-associated via death domain
...	...	...	...
22282	AFFX-ThrX	---	---
22283	AFFX-Trpr	---	---
22284	AFFX-Trpr	---	---
22285	AFFX-Trpr	---	---
22286			

GSEA v2 (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 | MSigDB gene sets browser

File path or URL to the MSigDB database: ftp://ftp.broad.mit.edu/pub/gsea/msigdb_v2.xml ... Load database

msigdb_v2 [version=V2 build=Nov1_2006]

Selection

QuickFilterPane

COLLECTION	ORGANISM	CHIP	CONTRIBUTER
Computational	All (1 ORGANISM)	All (1 CHIP)	All (1 CONTRIBUTER)
Curated	Human,Mouse,Rat,Dog	GENE_SYMBOL	Xiaohui Xie
Motif			
Positional			

Deep search options

Find sets that overlap with my set ... | Find sets that contain this gene ...

Filtered result List (right click for options)

837 out of 3337 gene sets

	NAME	# GENES	DESCRIPTION	COLLECTION	ORGANISM	CHIP	CONTRIBUTER	PUBMED ID	EXTERN.
2074	RGAGGAAR...	522	Genes with ...	Motif	Human,Mou...	GENE_SYMBOL	Xiaohui Xie		
2075	KRCTCNNN...	66	Genes with ...	Motif	Human,Mou...	GENE_SYMBOL	Xiaohui Xie		
2076	AAAYWAAC...	264	Genes with ...	Motif	Human,Mou...	GENE_SYMBOL	Xiaohui Xie		
2077	YYCATTCA...	194	Genes with ...	Motif	Human,Mou...	GENE_SYMBOL	Xiaohui Xie		
2078	CYTAGCAAY...	153	Genes with ...	Motif	Human,Mou...	GENE_SYMBOL	Xiaohui Xie		
2079	CAGCTG_V$...	1561	Genes with ...	Motif	Human,Mou...	GENE_SYMBOL	Xiaohui Xie		
2080	GGCNKCCA...	122	Genes with ...	Motif	Human,Mou...	GENE_SYMBOL	Xiaohui Xie		
2081	RRAGTTGT_...	258	Genes with ...	Motif	Human,Mou...	GENE_SYMBOL	Xiaohui Xie		
2082	GATTGGY_V...	1190	Genes with ...	Motif	Human,Mou...	GENE_SYMBOL	Xiaohui Xie		
2083	CATTGTYY_...	370	Genes with ...	Motif	Human,Mou...	GENE_SYMBOL	Xiaohui Xie		
2084	TAAWWATA...	175	Genes with ...	Motif	Human,Mou...	GENE_SYMBOL	Xiaohui Xie		
2085	AACYNNNN...	101	Genes with ...	Motif	Human,Mou...	GENE_SYMBOL	Xiaohui Xie		
2086	GCANCTGN...	953	Genes with ...	Motif	Human,Mou...	GENE_SYMBOL	Xiaohui Xie		
2087	YCATTAA_U...	575	Genes with ...	Motif	Human,Mou...	GENE_SYMBOL	Xiaohui Xie		
2088	KMCATNNW...	97	Genes with ...	Motif	Human,Mou...	GENE_SYMBOL	Xiaohui Xie		
2089	GATAAGR_V...	304	Genes with ...	Motif	Human,Mou...	GENE_SYMBOL	Xiaohui Xie		

Help

Export → Export sets as GeneSetMatrix

1:53:42 PM | 3148 [INFO] Begun importing: MSigDB from: ftp://ftp.broad.mit.edu/pub/gsea/msigdb_v2.xml | 72M of 90M

Sample categorical .cls (class) files:

Specify phenotype of each sample, e.g., tumor type 1, tumor type 2; treatment works, does not work, same order from left to right as the samples in the expression file (the .gct file)

Categorical class files: 3 line, space delimited text files

```
12  2  1
# DEAD ALIVE
D  D  A  D  D  D  A  D  A  D  A  D
```

samples in .gct file, # of phenotypes, 1

the 1 is required, does not change

class names used by GSEA in output data

Symbols corresponding to the classes of the samples in the .gct file

```
11  2  1
# ALL AML
ALL ALL ALL AML AML ALL AML AML ALL AML ALL
```

line 3 can be tab delimited

1 1 1 0 0 1 0 0 1 0 1 alternate line 3 for this class file

Example of a “Numeric” Class File

Identifies this as a “numeric” .cls file

```
#numeric
#ALLOx24AML1x24
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

arbitrary text used in some of the GSEA output file names

vector V of numbers, one for each sample; genes will be ranked by a measure of their correlation with V

A “numeric” .cls file, of the form one would have in order to use the “Pearson correlation” gene ranking metric (this would NOT be a normal choice for data with 2 phenotypes). If one had time series data, so, e.g., each sample was expression data of some system at a sequence of time points, the numeric values in line 3 (one for each sample, ordered as the samples are ordered in the .gct file) could be an expression pattern over time one was looking to have gene sets match. These values could be the expression levels of a gene one was looking to match, or a measure of disease severity for each sample.

Illustration of **.chip** Description File

tab delimited text file (optionally used by Java GSEA to convert feature IDs to gene symbols)

	A	B	C
1	Probe Set ID	Gene Symbol	Gene Title
2	1007_s_at	DDR1	discoidin domain receptor family, member 1
3	1053_at	RFC2	replication factor C (activator 1) 2, 40kDa
4	117_at	HSPA6 /// LOC652878	heat shock 70kDa protein 6 (HSP70B') /// similar to heat shock 70kDa protein 6 (HSP70B)
5	121_at	PAX8	paired box gene 8
6	201736_s_at	****Excel****did not import column as "text" 6-Mar should be MARCH6	membrane-associated ring finger (C3HC4) 6
7	201307_at	****Excel****did not import column as "text" 11-Sep should be SEPT11	septin 11
8	207923_x_at	PAX8	paired box gene 8

GSEA may optionally combine duplicate expr. values

column headers MUST be as displayed

IDs & symbols are **case sensitive**

Use either these probe Set IDs in the gene set .gmt file or use these gene symbols

IDs should corr. to col 1 of .gct file

for more info see http://www.broad.mit.edu/gsea/doc/GSEAUserGuideFrame.html?_Preparing_Data_Files and http://www.broad.mit.edu/cancer/software/gsea/wiki/index.php/Data_formats#Microarray_Chip_Annotation_Formats

UNIQUE ID | if none enter --- or **null** or **na** | if none enter --- or **null** or **na**

see http://discover.nci.nih.gov/symbolmutation/ for info on proper text file import into Excel

Excel display of a modified section of: HG_U133A.chip. Chip files are available from the Broad Institute web page http://www.broad.mit.edu/gsea/resources/resources_index.html (click on Array Annotations for ftp site)

Sample input choices for a test run for Desktop Java GSEA

GSEA v2 (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 | Run Gsea

Gsea: Set parameters and run enrichment tests

Required fields

Expression dataset	P53 [10100x50 (ann: 10100,50,chip na)]
Gene sets database	C:\berger\GSEA_java\C2.gmt
Number of permutations	10
Phenotype labels	C:\berger\GSEA_java\P53.cls#MUT_versus_WT
Collapse dataset to gene symbols	false
Permutation type	phenotype
Chip platform(s)	

Basic fields Hide

Analysis name	P53test10perms
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:\berger\GSEA_java\P53_testJ_Aug3

Advanced fields Show

Reset Last Command Low (cpu usage) Run

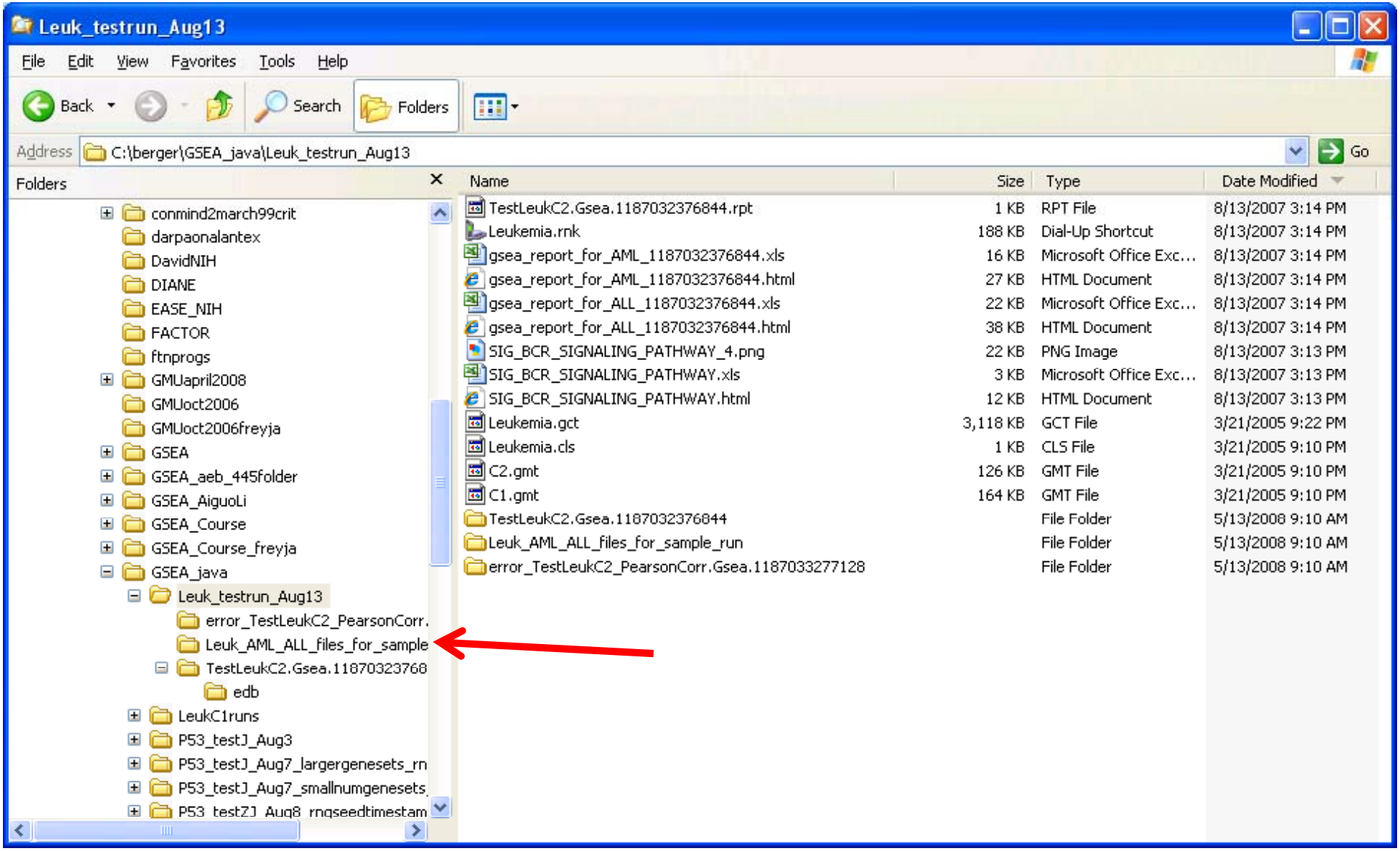
P53.gct, C2.gmt, P53.cls were previously loaded

small # for test run

P53.gct already has gene symbols in col 1, so no .chip file needed

Click when have finished selecting parameters

Once a test run checks out, use, e.g., 1000 permutations; can check stability of NES & FDR results by varying #permutations, varying the initial random number generator seed (see next slide)


Leuk_testrun_Aug13
File Edit View Favorites Tools Help
Back Search Folders
Address C:\berger\GSEA_java\Leuk_testrun_Aug13
Go
Folders
conmind2march99crit
darpaonalantex
DavidNIH
DIANE
EASE_NIH
FACTOR
ftnprogs
GMUapril2008
GMUoct2006
GMUoct2006freyja
GSEA
GSEA_aeb_445folder
GSEA_AiguoLi
GSEA_Course
GSEA_Course_freyja
GSEA_java
Leuk_testrun_Aug13
error_TestLeukC2_PearsonCorr.
Leuk_AML_ALL_files_for_sample
TestLeukC2.Gsea.11870323768
edb
LeukC1runs
P53_testJ_Aug3
P53_testJ_Aug7_largergenesets_rn
P53_testJ_Aug7_smallnumgenesets
P53_testJ_Aug8_rngseedtimestam
Name Size Type Date Modified
TestLeukC2.Gsea.1187032376844.rpt 1 KB RPT File 8/13/2007 3:14 PM
Leukemia.rnk 188 KB Dial-Up Shortcut 8/13/2007 3:14 PM
gsea_report_for_AML_1187032376844.xls 16 KB Microsoft Office Exc... 8/13/2007 3:14 PM
gsea_report_for_AML_1187032376844.html 27 KB HTML Document 8/13/2007 3:14 PM
gsea_report_for_ALL_1187032376844.xls 22 KB Microsoft Office Exc... 8/13/2007 3:14 PM
gsea_report_for_ALL_1187032376844.html 38 KB HTML Document 8/13/2007 3:14 PM
SIG_BCR_SIGNALING_PATHWAY_4.png 22 KB PNG Image 8/13/2007 3:13 PM
SIG_BCR_SIGNALING_PATHWAY.xls 3 KB Microsoft Office Exc... 8/13/2007 3:13 PM
SIG_BCR_SIGNALING_PATHWAY.html 12 KB HTML Document 8/13/2007 3:13 PM
Leukemia.gct 3,118 KB GCT File 3/21/2005 9:22 PM
Leukemia.cls 1 KB CLS File 3/21/2005 9:10 PM
C2.gmt 126 KB GMT File 3/21/2005 9:10 PM
C1.gmt 164 KB GMT File 3/21/2005 9:10 PM
TestLeukC2.Gsea.1187032376844 File Folder 5/13/2008 9:10 AM
Leuk_AML_ALL_files_for_sample_run File Folder 5/13/2008 9:10 AM
error_TestLeukC2_PearsonCorr.Gsea.1187033277128 File Folder 5/13/2008 9:10 AM

Sample input choices for Desktop GSEA: Advanced Fields

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

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 | Run Gsea

Analysis name	P53test
Enrichment statistic	weighted
Metric for ranking genes	Signal2Noise
Gene list sorting mode	
Gene list ordering mode	
Max size: exclude larger sets	
Min size: exclude smaller sets	
Save results in this folder	

Signal2Noise
tTest
Cosine
Euclidean
Manhatten
Pearson
Ratio_of_Classes
Diff_of_Classes

Choices for measures of differential expression

Advanced fields — Hide

Collapsing mode for probe sets => 1 gene	Max_probe
Normalization mode	meandiv
Randomization mode	no_balance
Omit features with no symbol match	true
Make detailed gene set report	true
Median for class metrics	false
Number of markers	100
Plot graphs for the top sets of each phenotype	20
Seed for permutation	timestamp
Save random ranked lists	false
Make a zipped file with all reports	false

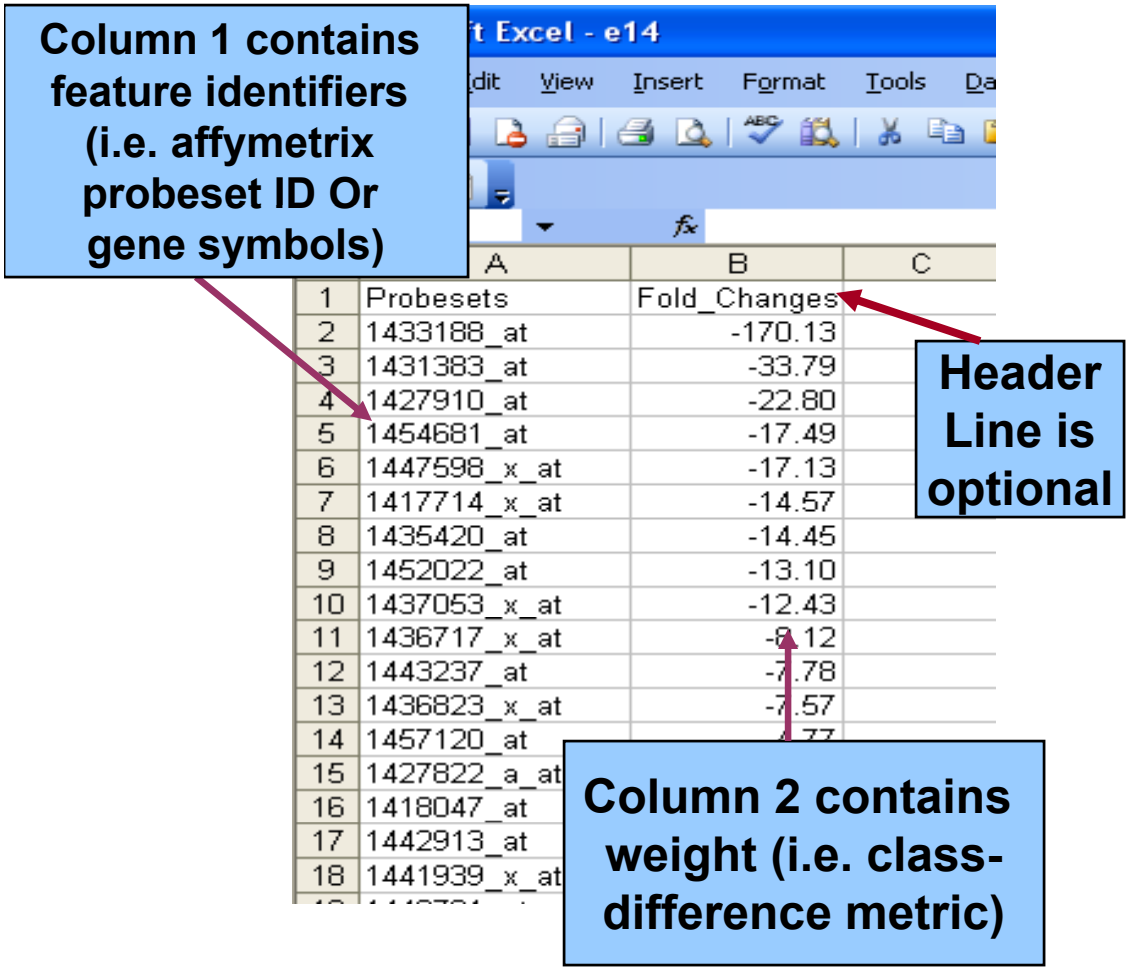
this means using NES

choose an explicit integer seed for random # generator so can easily reproduce results

Reset Last Command Low (cpu usage) Run

39:40 AM 3256 [INFO] Loading ... 3 filesC2.gmtP53.clsP53.gctFiles loaded successfully: 3 / 3There were NO errors 50M of 63M

File Format for Preranked Genes



- A *.rnk file contains a single, rank ordered gene list (*not* gene set) in a simple tab delimited text format.
- It is used when you have a pre-ordered ranked list that you want to analyze with GSEA.
- Column B is a differential expression score between the two phenotypes
- List need not be sorted
- lose ability to generate NES empirical null distribution by class label permutation

Running Desktop GSEA using a pre-ranked gene list

GSEA v2 (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
1	GseaPrera...	Running

Show results folder

Home | Run Gsea on a Pre-Ranked gene list | Load data

GseaPreranked: Run GSEA on a pre-ranked (with external tools) gene list

Required fields

Gene sets database	ftp.broad.mit.edu://pub/gsea/gene_sets/c2.v2.symbols.gmt
Number of permutations	1000
Ranked List	e18 [30824 names]
Collapse dataset to gene symbols	true
Chip platform(s)	D:\TSC\diff\mouse\data\Mouse430_2.chip

Basic fields

Analysis name	my_analysis
Enrichment statistic	weighted
Max size: exclude larger sets	500
Min size: exclude smaller sets	15
Save results in this folder	C:\Documents and Settings\liai\gsea_home\output\aug02

Advanced fields

Hide

Collapsing mode for probe sets => 1 gene	Max_probe
Normalization mode	meandiv
Omit features with no symbol match	true
Make detailed gene set report	true
Plot graphs for the top sets of each phenotype	20
Seed for permutation	timestamp
Make a zipped file with all reports	false

Reset Last Command Low (cpu usage) Run

1:30:54 PM 8633 [INFO] Loading ... 1 filesMouse430_2.chipFiles loaded successfully: 1 / 1There were NO errors 45M of 63M

Required fields

.rnk file

Basic fields

Advanced fields

GSEA Parameters and Defaults

Parameters	Default	Option
Collapse dataset to gene symbols	True	True or False (if true need to supply a .chip file)
Permutation type	Phenotype	Phenotype or Gene_set
Enrichment statistics	Weighted	Classic, weighted <p=1>, weighted_p2, weighted_p1.5
Metric for ranking genes	Signal2Noise	Signal2Noise, tTest, Pearson, Ratio_of_classes, Diff_of_classes, Log2_ratio_of_classes, Euclidean, Manhatten, cosine
Gene list sorting mode	Real	Real or Abs
Gene list ordering mode	Descending	Descending or Ascending
Maximum size	500	User defined fields
Minimum size	15	User defined fields
Collapsing for probe set	Max_probe	Max_probe or Median of probes
Number of permutations	1000	User defined fields
Normalization mode	meandiv	Meandiv (use NES) or None (use ES)
Seed for permutation	timestamp	We recommend putting in a user chosen positive integer < 2^32

Measures of Differential Expression

Let the expression data consist of samples from two phenotypes A and B. For a given gene g: let μ_A be the mean of the expression levels for g from the subset of samples having phenotype A & similarly for μ_B; and likewise with standard deviations σ_A and σ_B. Then the signal2noise (GSEA default) measure of differential expression of g between A and B used as the gene ranking metric is:

$$\textbf{signal2noise(g)} \equiv \frac{\mu_A - \mu_B}{\sigma_A + \sigma_B}$$

A number of other options are available from the Desktop GSEA, including tTest, log2_Ratio_of_Classes, Ratio_of_classes, and several measures of correlation for continuous phenotypes; see "Metrics for Ranking Genes" in http://www.broad.mit.edu/cancer/software/gsea/doc/GSEAUserGuideFrame.html

GSEA User Guide - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites

Address http://www.broad.mit.edu/gsea/doc/GSEAUserGuideFrame.html Go Links

logged in as alanberger@alum.mit.edu
logout

BROAD INSTITUTE

GSEA
Gene Set Enrichment Analysis

Search

GSEA Home | Downloads | Molecular Signatures Database | Documentation | Contact

them. (Does not display the graphs.)

If you are building an HTML report: track analysis progress, as described in Tracking Analysis Progress, and view analysis results, as described in Viewing Analysis Results.

6. Interpret analysis results, as described in Interpreting Leading Edge Analysis Results.

Running Other GSEA Analyses

GSEA also provides the following analyses:

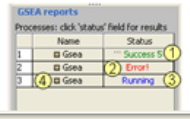
- Chip2Chip. Converts the genes in a gene set from HUGO gene symbols to the probe identifiers for a selected target chip. For example, if you have a dataset that uses probes from the HG_U95Av2 chip to identify genes, you can use this utility to convert MSigDB gene sets from HUGO gene symbols to probe identifiers for the HG_U95Av2 chip.
- GSEAPreranked. Runs the gene set enrichment analysis against a ranked list of genes, which you supply. When you use the Run GSEA icon to run the gene set enrichment analysis, GSEA ranks the genes in your expression dataset (based on the metric that you select using the *metric for ranking genes* parameter) and then analyzes that ranked list of genes. Alternatively, you can create your own ranked list of genes and use GSEAPreranked to analyze that ranked list of genes.
- CollapseDataset. Creates a new dataset by collapsing each probe set into a single vector for the gene, which is identified by its HUGO gene symbol. When you run the gene set enrichment analysis with the *Collapse dataset to gene symbols* parameter set to True, GSEA runs this analysis as part of the gene set enrichment analysis.

To run one of these analyses:

1. Select the Chip2Chip icon on the GSEA main page, or select an analysis from the Tools menu. The GSEA page for the selected analysis appears.
2. Enter values for the analysis parameters. For parameter descriptions, click *Help*, which displays the Chip2Chip, GSEAPreranked, or CollapseDataset page of this guide.
3. Click *Run* to start the analysis.
4. Track analysis progress, as described in Tracking Analysis Progress.
5. View analysis results, as described in Viewing Analysis Results.

Tracking Analysis Progress

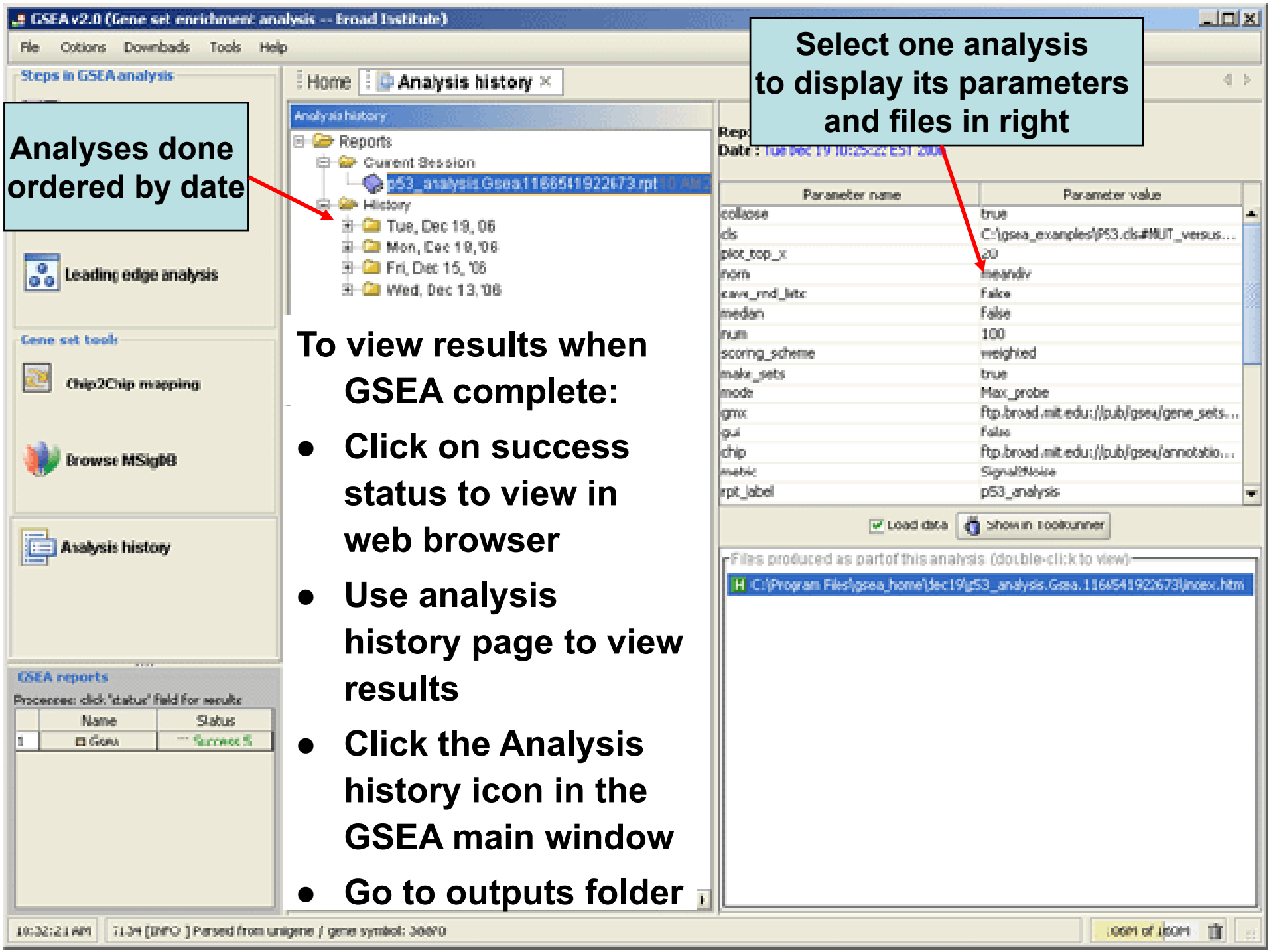
When you start an analysis, the gene set enrichment analysis or any other analysis, the Processes area shows the analysis as running (blue). When the analysis is finished, it shows the analysis has succeeded (green). If an error occurs, it shows an error message (red). When you exit from GSEA, the Processes area is cleared.



Internet

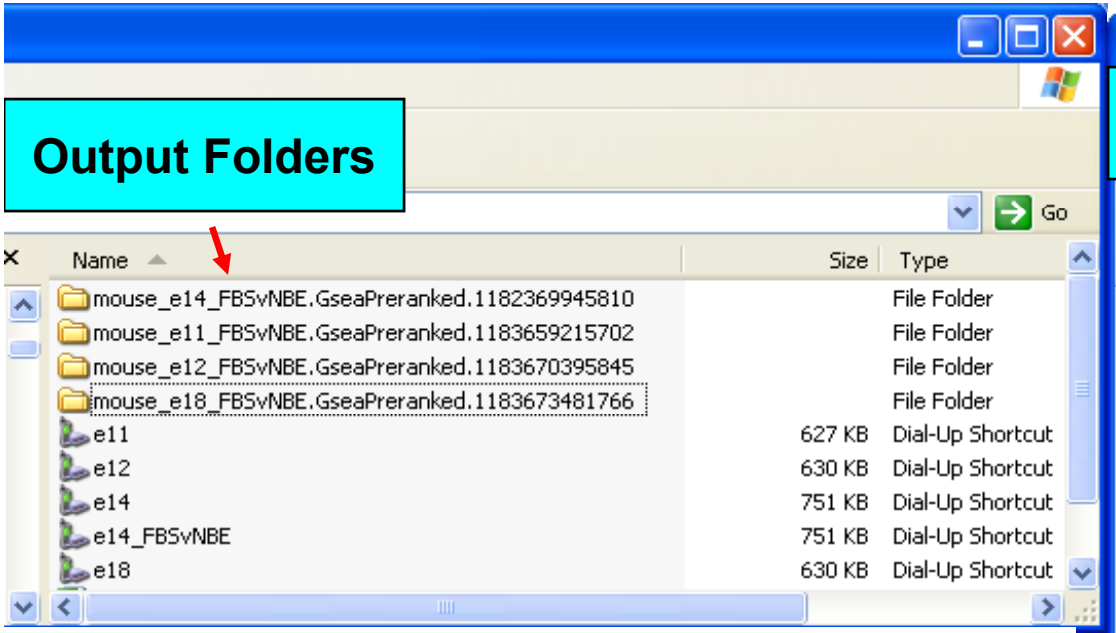
Outline

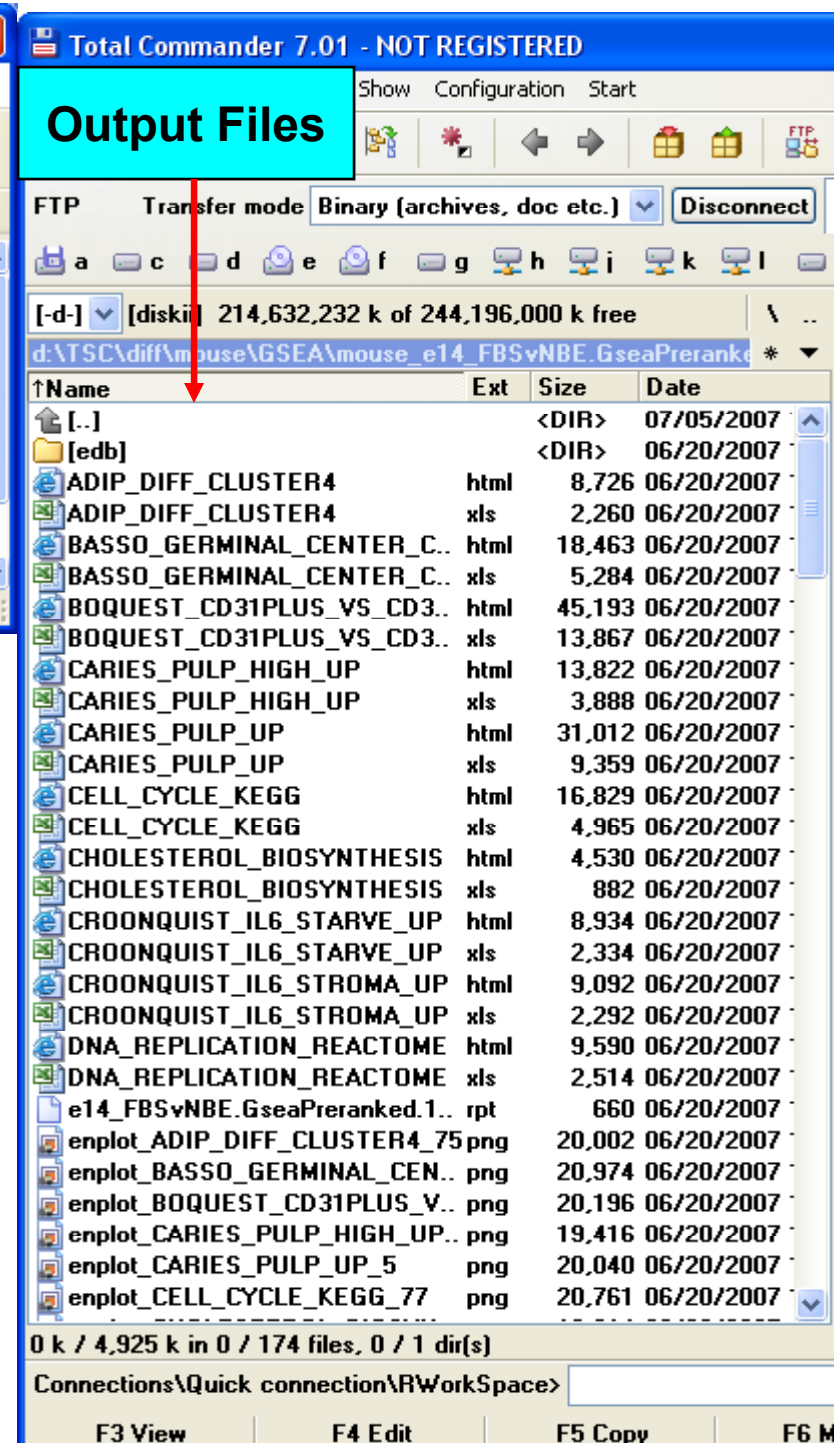
- **Functional Analysis of Microarray Data – Analysis at the Level of *Gene Sets***
- **Introduction to GSEA (Gene Set Enrichment Analysis)**
- **<break>**
- **Installing GSEA: Desktop**
- **Running GSEA: Required Input Files & Parameter Selection; Broad Institute Utilities**
- **<break>**
- **Understanding the GSEA Outputs**
- **Live Demonstration Running Desktop GSEA**



To view results when GSEA complete:

- Click on success status to view in web browser
- Use analysis history page to view results
- Click the Analysis history icon in the GSEA main window
- Go to outputs folder

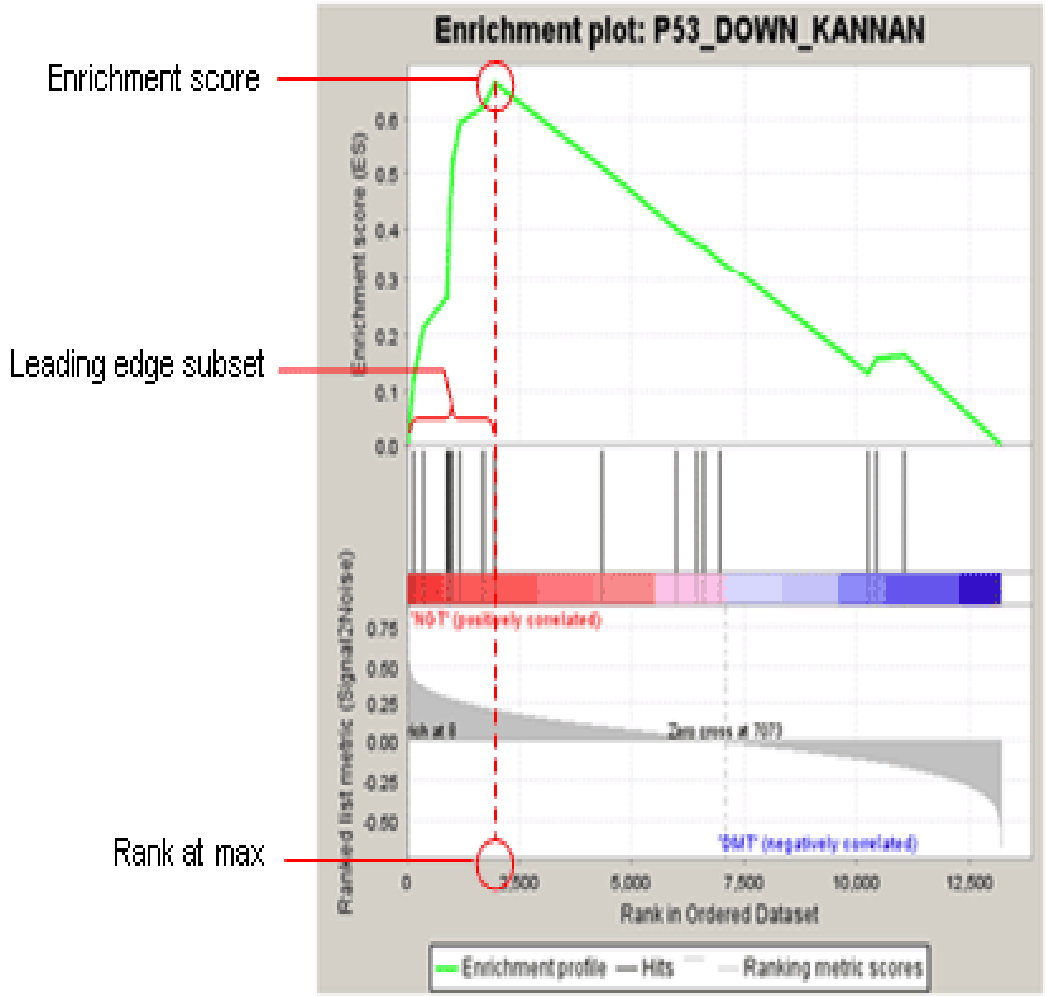




OUTPUT FILE SUMMARY:

- Enrichment plot
  - Running enrichment score
  - A heat map of the genes in the gene set
  - A histogram of distribution values for the gene set
- Random ES distribution plot
- Heat-map plot for a given gene set
- Detailed information for the gene set in Excel format

Desktop GSEA Output Example



	A	B	C	D	E	F	G	
1	NAME	PROBE	GEN	GENI	RANK IN	RANK METRIC SC	RUNNING ES	CO
2	row_0	HTRA1	null	null	1	179.7689972	0.22774598	Ye
3	row_1	GFAP	null	null	4	128.8899994	0.39095533	Ye
4	row_2	RAMP1	null	null	76	14.10369968	0.404457	Ye
5	row_3	DHRS3	null	null	106	11.23560047	0.41690975	Ye
6	row_4	TST	null	null	115	10.74059963	0.43002802	Ye
7	row_5	CAV2	null	null	130	9.910790443	0.44172534	Ye
8	row_6	CSRP1	null	null	137	9.633099556	0.45356327	Ye
9	row_7	CYP1B1	null	null	146	9.454489708	0.46505174	Ye
10	row_8	TGM2	null	null	158	9.138770103	0.47595543	Ye
11	row_9	OLIG2	null	null	187	8.207819939	0.48463285	Ye
12	row_10	EFHD1	null	null	226	7.341119766	0.49159637	Ye
13	row_11	CALD1	null	null	227	7.325870037	0.5008799	Ye
14	row_12	ITGB1	null	null	236	7.145329952	0.50944215	Ye
15	row_13	PDLIM3	null	null	280	6.404119968	0.51491046	Ye
16	row_14	EFEMP1	null	null	315	6.008480072	0.52043146	Ye
17	row_15	WIPI1	null	null	321	5.8604002	0.5275501	Ye
18	row_16	4-Sep	null	null	347	5.629469872	0.53314483	Ye
19	row_17	CEBPD	null	null	380	5.385200024	0.53799915	Ye
20	row_18	PLP1	null	null	399	5.256130219	0.54355174	Ye
21	row_19	NTRK2	null	null	407	5.219820023	0.5497355	Ye
22	row_20	SLC4A4	null	null	423	5.115940094	0.5552951	Ye
23	row_21	SSPN	null	null	456	4.941669941	0.55958736	Ye
24	row_22	MYO10	null	null	460	4.92798996	0.56564754	Ye
25	row_23	METTL7A	null	null	468	4.870259762	0.5713883	Ye
26	row_24	ZHX2	null	null	472	4.846690178	0.5773455	Ye
27	row_25	LPP	null	null	507	4.588429928	0.58106697	Ye
28	row_26	WWTR1	null	null	589	4.242750168	0.581457	Ye
29	row_27	MFAP3L	null	null	604	4.146800041	0.5858501	Ye

GSEA Outputs

- **Basic outputs: 2 files for each of the gene sets**
  - **Graphic file: genes in the gene set as "hits" against the ranked list of genes, running enrichment score (RES), histogram of null distribution values for the gene sets, and heat map**
  - **Gene files: a summary of the gene list**
- **Summary outputs:**
- **Two text files for gene sets summary**
- **7 summary graphic files for leading gene sets**

- **Desktop Java GSEA and R-GSEA output files are similar**

GSEA Outputs: Thresholds

- **Nominal p values – obtained from the empirical null distribution of the gene set enrichment scores**
- **FDR q value – (False Discovery Rate) is the estimated fraction of false positives in a collection of gene sets, here of the form**

  $$\{\text{all } S \mid NES(S) \geq \gamma\} \quad \text{or} \quad \{\text{all } S \mid NES(S) \leq -\gamma\}.$$

  It estimates the probability that a gene set at or beyond a given NES is a false positive finding and it is computed by comparing the tails of the observed and null distribution of the NES (separate calculations for positive & negative tails for p, q values).
- **FWER p value – stands for FamilyWise-Error Rate. It is a very conservative correction that seeks to ensure that a list of reported results is not likely to include even a single false-positive gene set; calculated using the empirical NES null distr.**

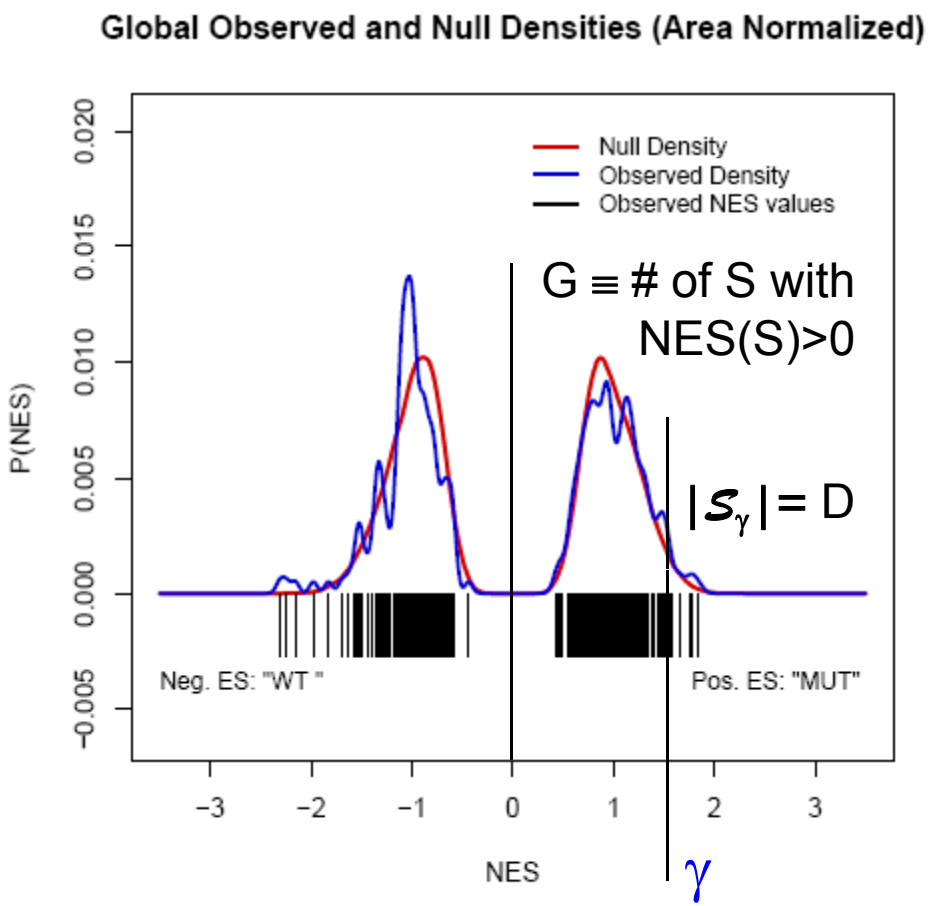


Figure extracted from testP53_C2.global.plots.pdf from Broad Institute GSEA R distribution

Simple Definition of FDR

for sets of the form $\mathcal{S}_\gamma$ = {all S with NES ≥ γ} using NES(S) & NES(S, π) values. Similarly for $\mathcal{S}_{-\gamma}$ = {S | NES(S) ≤ -γ} (+, - NES values are treated separately)

Use fraction $\mathcal{F}$ of nonnegative NES(S, π) values that are ≥ γ to estimate number F of gene sets that are by random chance in $\mathcal{S}_\gamma$ (F ≈ $\mathcal{F}$ G). If actual size of $\mathcal{S}_\gamma$ is D then estimate FDR by F / D, e.g., if have G = 500 gene sets with NES(S) > 0, and the fraction $\mathcal{F}$ is 0.01, then estimate F = 5; so if D were, say, 20, then estimate FDR by 5/20.

GSEA outputs on P53 data files, top pane is output from GSEA R Broad distribution. Bottom pane is output from Java GSEA, test run with just 10 permutations. Note ES is same: the perm. based quantities **NES** & significance measures are of course different, due to the diff. # of perm. used for the 2 runs (also different random # generator seeds).

P53_C2.SUMMARY.RESULTS.REPORT.MUT.txt

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	GS	SIZE	SOURCE	ES	NES	NOM p-val	FDR q-val	FWER p-v:	Tag %	Gene %	Signal	FDR (medi	glob.p.val
2	rasPathway	22	BioCarta	0.60308	1.8903	0.002024	0.1714	0.136	0.727	0.27	0.532	0	0.063
3	ngfPathway	19	BioCarta	0.58879	1.7927	0.001898	0.29535	0.366	0.579	0.212	0.457	0	0.083
4	UPREG_BY_HOXA9	29	Manually (	0.5832	1.7605	0.01111	0.27522	0.461	0.517	0.171	0.43	0	0.063
5	igf1Pathway	20	BioCarta	0.55742	1.7382	0.007505	0.25669	0.517	0.65	0.258	0.483	0.19977	0.045
6	XINACT_MERGED	16	na	0.60528	1.6472	0.03462	0.4874	0.785	0.5	0.185	0.408	0.38372	0.124
7	egfPathway	27	The epider	0.48185	1.5981	0.02381	0.60866	0.907	0.519	0.254	0.388	0.48962	0.184
8	insulinPathway	21	BioCarta	0.48505	1.562	0.02295	0.69687	0.95	0.571	0.258	0.425	0.59077	0.229
9	MAPK_Cascade	21	GO	0.49233	1.5537	0.01625	0.64789	0.957	0.476	0.172	0.395	0.55743	0.181
10	BRCA_UP	39	Welcsh_et	0.46066	1.5482	0.03636	0.5991	0.96	0.436	0.209	0.346	0.50926	0.15
11	ST_ERK1_ERK2_MAP	28	Signalling	0.44366	1.5479	0.0239	0.54051	0.961	0.464	0.235	0.356	0.46122	0.122
12	NFKB_REDUCED	21	Hinata_et_	0.57548	1.5198	0.05051	0.60061	0.976	0.381	0.129	0.333	0.52053	0.149
13	gcrPathway	18	BioCarta	0.50218	1.4997	0.05323	0.63534	0.984	0.389	0.211	0.307	0.56857	0.159

P53_C2.SUMMARY.RESULTS.REPORT.M

gsea_report_for_MUT_1186181691156.xls

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	NAME	GS fo	GS DETAI	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-v:	RANK AT	LEADING EDGE		
2	XINACT_MERGED	XINACT_M	Details ...	16	0.605281	1.976712	0	0.089189	0.1	1863	tags=50%, list=18%, signal=61%		
3	RASPATHWAY	RASPATH	Details ...	22	0.603078	1.976159	0	0.044595	0.1	2728	tags=73%, list=27%, signal=99%		
4	NGFPATHWAY	NGFPATH	Details ...	19	0.588791	1.950186	0	0.02973	0.1	2143	tags=58%, list=21%, signal=73%		
5	IGF1PATHWAY	IGF1PATH	Details ...	20	0.557419	1.907802	0	0.022297	0.1	2605	tags=65%, list=26%, signal=87%		
6	UPREG_BY_HOXA9	UPREG_B	Details ...	29	0.583197	1.872689	0	0.035676	0.1	1725	tags=52%, list=17%, signal=62%		
7	EGFPATHWAY	EGFPATH	Details ...	27	0.48185	1.770715	0	0.074087	0.2	2566	tags=52%, list=25%, signal=69%		
8	PDGFPATHWAY	PDGFPAT	Details ...	27	0.450959	1.703785	0	0.074266	0.3	2624	tags=52%, list=26%, signal=70%		
9	INSULINPATHWAY	INSULINP/	Details ...	21	0.485053	1.643465	0	0.211563	0.7	2605	tags=57%, list=26%, signal=77%		
10	NFKB_REDUCED	NFKB_RE	Details ...	21	0.575481	1.636601	0	0.188056	0.7	1297	tags=38%, list=13%, signal=44%		
11	MAPK_CASCADE	MAPK_CA	Details ...	21	0.492333	1.615752	0	0.197927	0.7	1738	tags=48%, list=17%, signal=57%		
12	ST_PHOSPHOINOSITID	ST_PHOS	Details ...	32	0.397832	1.565562	0	0.272537	0.8	2371	tags=41%, list=23%, signal=53%		
13	SA_B_CELL_RECEPTC	SA_B_CEI	Details ...	23	0.448227	1.546039	0.142857	0.282459	0.8	2132	tags=48%, list=21%, signal=60%		

From page 30 of the Broad GSEA User Guide

HTML Report

The HTML Report for the leading edge analysis contains the following sections:

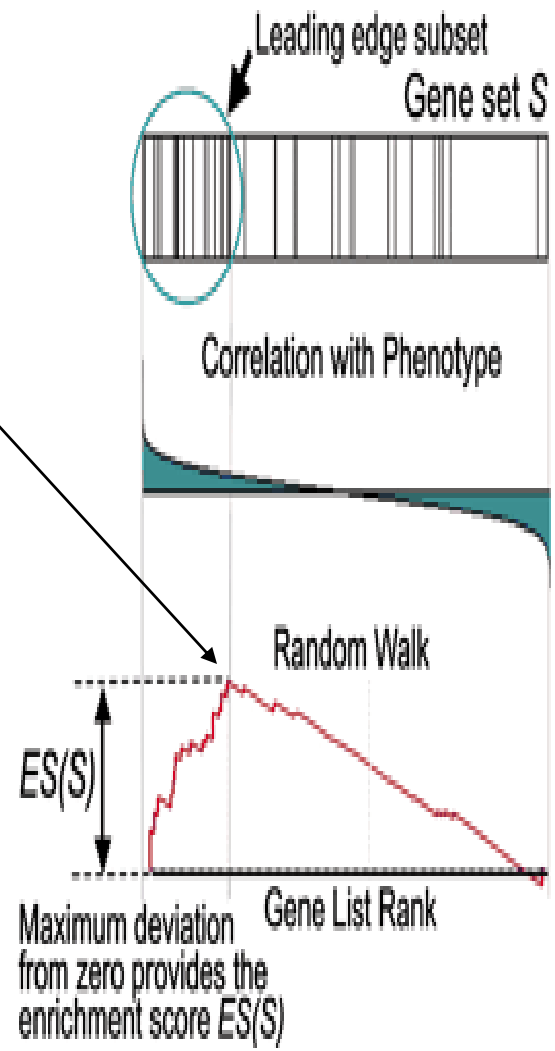
- **Clustered results**. Provides the number of gene sets analyzed and a heat map of the leading edge subsets after clustering.
- **Details of gene sets**. Provides the following information for each of the analyzed gene sets and its leading edge subset:
  - # members. Number of genes in the gene set.
  - # members in signal. Number of genes in the leading edge subset.
  - Tag %. The percentage of gene hits before (for positive ES) or after (for negative ES) the peak in the running enrichment score. This gives an indication of the percentage of genes contributing to the enrichment score.
  - List %. The percentage of genes in the ranked gene list before (for positive ES) or after (for negative ES) the peak in the running enrichment score. This gives an indication of where in the list the enrichment score is attained.
  - Signal strength. The enrichment signal strength that combines the two previous statistics:

    $$(\text{Tag }\%)(1-\text{Gene }\%)\left(\frac{N}{N-Nh}\right)$$

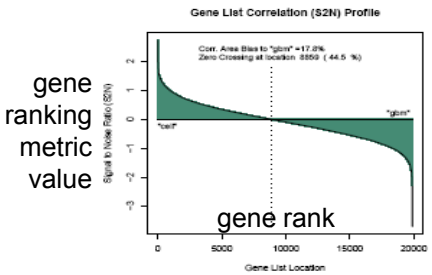
    where *N* is the number of genes in the list and *Nh* is the number of genes in the gene set. If the gene set is entirely within the first *Nh* positions in the list, then the signal strength is maximal or 100%. If the gene set is spread throughout the list, then the signal strength decreases towards 0%.
- **Other files made**. Provides a heat map of the (unclustered) leading edge subsets and tabular ways of examining the leading edge subsets:
  - Clustered dataset (gct) uses the expression dataset format to describe the clustered leading edge subsets: each row is a gene set, each column is a gene, and an "expression value" of 1 indicates the gene is in the leading edge subset for the gene set.
  - GeneMatrix (gms) provides a gene set file for the leading edge subsets, which lists each gene set and the genes in its leading edge subset.
  - Dataset (gct) uses the expression dataset format to describe the leading edge subsets (not clustered): each row is a gene set, each column is a gene, and an "expression value" of 1 indicates the gene is in the leading edge subset for the gene set.
  - Heat map shows a heat map of the leading edge subsets (not clustered).
- **Other**. Lists the analysis parameters. Knowing the parameters used to produce the analysis is critical for reproducible research.

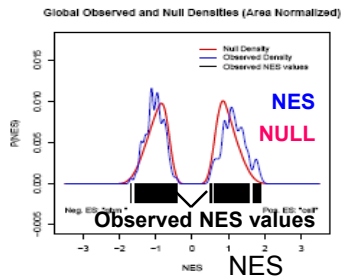
Leading-edge Subsets

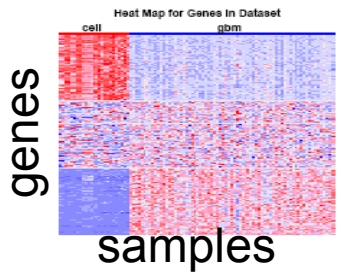
- **The Leading-edge subset is defined as those genes in S that appear in the ranked list L between the point where the running sum reaches its maximum deviation from zero and the adjacent end of L; it is the core of a gene set that accounts for the enrichment signal**
- **Examination of the leading-edge subset can reveal a biologically important subset within a gene set**
- **Grouping high scoring gene sets according to the leading edge subsets of genes may reveal which of those gene sets correspond to the same/related biological processes and which represent distinct processes.**

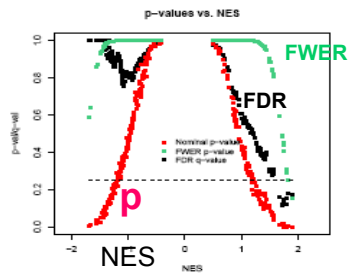


Summary Outputs

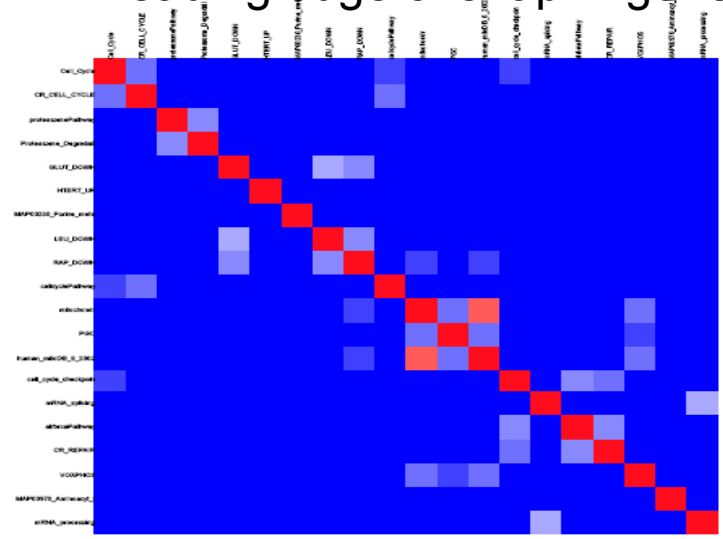




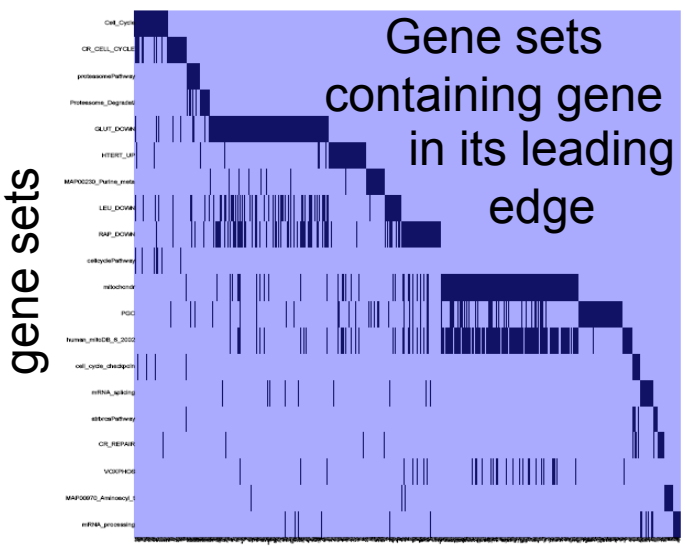


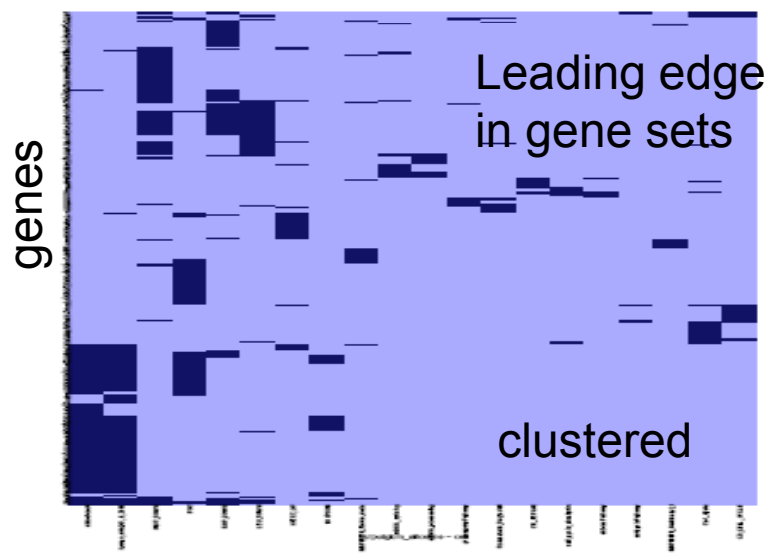


Leading edge overlap in gene sets



one plot per pheno-type



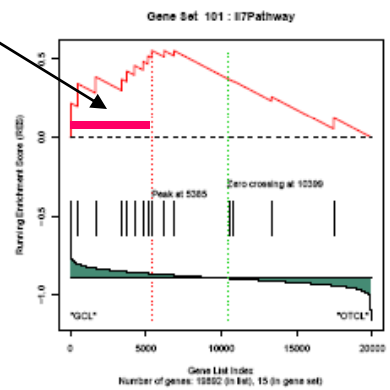


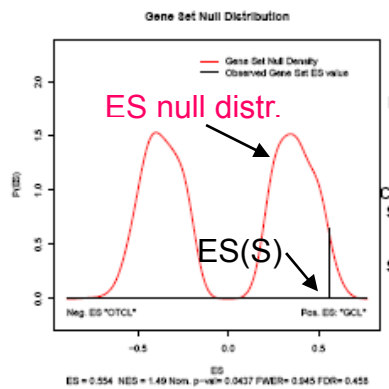
one plot per pheno-type

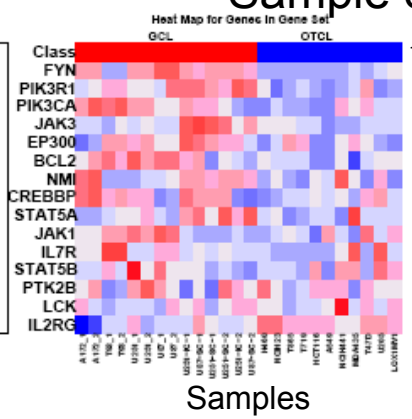
Sample Outputs: S = IL7 Pathway

Leading edge subset

Sample class Indicator







#	GENE	SYMBOL	DESC	LIST LOCATION	S2N	RES	CORE_ENRICHMENT
1	FYN	FYN	FYN	44	1.42	0.218	YES
2	PIK3R1	PIK3R1	PIK3R1	481	0.943	0.343	YES
3	PIK3CA	PIK3CA	PIK3CA	1669	0.666	0.386	YES
4	JAK3	JAK3	JAK3	3408	0.466	0.371	YES
5	EP300	EP300	EP300	3746	0.436	0.422	YES
6	BCL2	BCL2	BCL2	4243	0.396	0.458	YES
7	NMI	NMI	NMI	4825	0.352	0.483	YES
8	CREBBP	CREBBP	CREBBP	5144	0.328	0.518	YES
9	STAT5A	STAT5A	STAT5A	5385	0.31	0.554	YES
10	JAK1	JAK1	JAK1	6212	0.256	0.552	NO

RES = running enrichment score

GSEA

- **Provides a systematic way to examine whether the expression levels of a gene set is correlated with the experimental or observed conditions for the samples**
- **Generally look at two phenotypes at one time**
- **10+ samples (balanced across the 2 phenotypes) are recommended by the GSEA manual (or should do gene set permutation). The output analysis provided by GSEA is rich but complex**
- **You can use GSEA to analyze a dataset that contains a preranked list of genes but then can not do permutation testing on phenotype labels**

Keep in Mind

- In the preprocessing step, GSEA excludes genes in a gene set that are not in the expression dataset
- The GSEA software does not preprocess the expression dataset.
- Normalizing the enrichment score across gene set size is done automatically
- Avoid analyzing gene sets with same set of genes, but a different gene set name.

Outline

- **Functional Analysis of Microarray Data – Analysis at the Level of *Gene Sets***
- **Introduction to GSEA (Gene Set Enrichment Analysis)**
- **<break>**
- **Installing GSEA: Desktop**
- **Running GSEA: Required Input Files & Parameter Selection; Broad Institute Utilities**
- **<break>**
- **Understanding the GSEA Outputs**
- **Live Demonstration Running Desktop GSEA**

QUESTIONS?

- Please fill out course evaluations
- Updated class slides will be available at the CIT Course 445 web page (or send us an email request)

Supplementary Slides

- Information on setting up and running R-version of GSEA
- Sketches of several other gene set methods

R Resources

- R installation resources:
  - http://cran.r-project.org/bin/windows/base/rw-FAQ.html
  - http://www.r-project.org/
- R programming basics: http://www.faculty.ucr.edu/~tgirke/Documents/R_BioCond/R_Programming.html

R Installation

- Go to Comprehensive R Archive Network (CRAN) at the URL: http://cran.r-project.org/
- Get the R-x.y.z-win32.exe binaries for base distribution from the 'bin/windows' directory of a CRAN site. The contrib link contains a large number of add-on packages.
- Make sure that the package version will run using the R version you have
- Your file system must allow long file names
- For Installation, just double-click on the icon and follow the instructions, e.g., 'R-2.5.1-win32.exe'.
- Uninstall can be done from the Control Panel.
- The Bioconductor R package is available via: from your R session, type:
  source("http://www.bioconductor.org/getBioC.R")
  getBioC()

Example of a file to run GSEA – R, sets input parameters & starts run. Run.P53_C2.R run file from the Broad Institute web page (with some annotations added)

```
# GSEA 1.0 -- Gene Set Enrichment Analysis / Broad Institute
# R script to run GSEA Analysis of the P53 vs C2 example (cut and paste into R console)

GSEA.program.location <- "d:/CGP2005/GSEA/GSEA-P-R/GSEA.1.0.R"
#  R source program (###### change pathnames to the right location in local machine #####)
source(GSEA.program.location, verbose=T, max.deparse.length=9999)
GSEA(                                                          # Input/Output Files :-----------------------------
 input.ds =  "d:/CGP2005/GSEA/GSEA-P-R/Datasets/P53.gct",      # Input gene expression Affy dataset file in RES or GCT format
 input.cls = "d:/CGP2005/GSEA/GSEA-P-R/Datasets/P53.cls",      # Input class vector (phenotype) file in CLS format
 gs.db =     "d:/CGP2005/GSEA/GSEA-P-R/GeneSetDatabases/C2.gmt",  # Gene set database in GMT format
 output.directory      = "d:/CGP2005/GSEA/GSEA-P-R/P53_C2/",   # Directory where to store output and results (default: "")
#  Program parameters :-------------------------------------------------------------------------------------------
 doc.string            = "P53_C2",        # Documentation string used as a prefix to name result files (default: "GSEA.analysis")
 non.interactive.run   = F,               # Run in interactive (i.e. R GUI) or batch (R command line) mode (default: F)
 reshuffling.type      = "sample.labels", # Type of permutation reshuffling: "sample.labels" or "gene.labels" (default: "sample.labels"
#                                          always use "sample.labels" unless not enough samples to get enough permutations
 nperm                 = 1000,            # Number of random permutations (default: 1000)
 weighted.score.type   =  1,              # Enrichment correlation-based weighting: 0=no weight (KS), ** 1 = weigthed **, 2 = over-weighted
 nom.p.val.threshold   = -1,              # Significance threshold for nominal p-vals for gene sets (default: -1, no threshold)
 fwer.p.val.threshold  = -1,              # Significance threshold for FWER p-vals for gene sets (default: -1, no threshold)
 fdr.q.val.threshold   = 0.25,            # Significance threshold for FDR q-vals for gene sets (default: 0.25)
 topgs                 = 20,              # Besides those passing test, number of top scoring gene sets used for detailed reports (def. 10)
 adjust.FDR.q.val      = F,               # Adjust the FDR q-vals (default: F)
 gs.size.threshold.min = 15,              # Minimum size (in genes) for database gene sets to be considered (default: 25)
 gs.size.threshold.max = 500,             # Maximum size (in genes) for database gene sets to be considered (default: 500)
 reverse.sign          = F,               # Reverse direction of gene list (pos. enrichment becomes negative, etc.: switch A&B) (default: F)
 preproc.type          = 0,               # Preproc.normalization: ** 0=none **, 1=col(z-score)., 2=col(rank) and row(z-score)., 3=col(rank)
 random.seed           = 760435,          # Random number generator seed. (default: 123456)
 perm.type             = 0,               # For experts only. Permutation type: 0 = unbalanced, 1 = balanced (default: 0)
 fraction              = 1.0,             # For experts only. Subsampling fraction. Set to 1.0 (no resampling) (default: 1.0)
 replace               = F,               # For experts only, Resampling mode (replacement or not replacement) (default: F)
 save.intermediate.results = F,           # For experts only, save intermediate results (e.g. matrix of random perm. scores) (default: F)
 OLD.GSEA              = F,               # Use original (old) version of GSEA (default: F)
 use.fast.enrichment.routine = T          # Use faster routine to compute enrichment for random permutations (default: T)
)
#-------------------------------------------------------------------------------------------------------------------

# Overlap and leading gene subset assignment analysis of the GSEA results

GSEA.Analyze.Sets(
   directory           = "d:/CGP2005/GSEA/GSEA-P-R/P53_C2/",          # Directory where to store output and results (default: "")
   topgs = 20,                                                        # number of top scoring gene sets used for analysis
```

if set non.interactive.run = T
get pdf output graphics files

make these two output locations the same

Memory Issues for R version of GSEA

Known Issues - GeneSetEnrichmentAnalysisWiki

Known Issues

From GeneSetEnrichmentAnalysisWiki

Jump to: navigation, search
GSEA Home | Software | MSigDB | Documentation | Resources

Error in memory.size when running GSEA-R

[edit]

Problem: When running the example programs provided for R, the following error occurs:

[1] " *** Running GSEA Analysis..."
Error in memory.size(size) : don't be silly!: your machine has a 4Gb address limit

Solution: This is produced by the following line early in the GSEA.1.R file:

```
memory.limit(6000000000)
```

This line set the memory limit to a large size as a work around to a platform problem with an earlier R version.
The easiest fix is just to comment out that line:

```
#       memory.limit(6000000000)
```

This will allocate the default amount of memory. If after this change the program runs out of memory, change the line to:

```
memory.limit(max. size in Mbytes available)
```

still requires large amount of RAM to run appropriate number of permutations

PAGE
Parametric Analysis of Gene set Enrichment

Input the fold changes between two experimental groups for all the genes on the chip

μ = mean of all the fold changes
δ = standard deviation for all the fold changes

for a given gene set S containing m genes, let
ave(S) = mean(fold changes) for the genes of S,

then the
PAGE z-score for S = $(\text{ave}(S) - \mu) / (\delta / m^{1/2})$

---

can use difference measures other than fold change

SigPathway Algorithm

- **Two Null Hypothesis tests**
  - **Q1 – The genes in the gene set have the same association pattern with phenotypes as the rest of the genes (Row/gene permutation (for NTk))**
  - **Q2 – There are no genes in the gene set whose expression is correlated with the phenotypes (Column/phenotype permutation (for NEk))**

- **Summary statistics Tk & Ek are defined for each gene set $\mathcal{S}_k$ in terms of averages of the t-scores t_i for each of their constituent genes.**

- **A standardization step is introduced to obtain Normalized gene set scores NTk & NEk: empirical null distributions are obtained by row or column permutations as indicated above.**

Tian et al., (2005) PNAS 102:13544-13549 note uses Affymetrix Probe Set IDs as gene identifiers

GSA: Gene Set Analysis

- **Compute a summary statistic (e.g. t-statistic[1]) z_i for each gene, and let $z_{\mathcal{S}}$ be the vector of z_i values for the genes in each gene set $\mathcal{S}$.**

- **Compute a summary statistic $\mathcal{s}(z_{\mathcal{S}})$ (e.g. maxmean[2]) for each gene set $\mathcal{S}$: denote it by $S = \mathcal{s}(z_{\mathcal{S}})$.**

- **Standardize S: $S' = (S - \text{mean}_S) / \text{stdev}_S$ (mean_S and stdev_S from row permutations)**

- **Re-compute S' on multiple column permuted datasets. Estimate p-values and FDR from resulting empirical null distribution.**

- **GSA can handle multiple classes, survival times and quantitative outcomes.**

---

[1]or z-value having the same p-value. [2]maxmean = ave(pos. parts) or ave(neg. parts) whichever has the largest magnitude. pos. part(x) = max(x,0); neg. part(x) = min(x,0)