Statistical Methods for Gene Set Analysis of Gene Expression Data

吳漢銘 助理教授
淡江大學 數學系

hmwu@mail.tku.edu.tw
http://www.hmwu.idv.tw

2011/12/13

Content

- **Recent Progress**
- **Finding Differential Expressed Genes**
  - t-test and Significance Analysis of Microarrays (SAM)
- **Gene Set Analysis (GSA)**
  - Gene Set Enrichment Analysis (GSEA)
  - SAM-GS
  - Maxmean
- **Null Hypothesis, Statistical Significance Of Gene Set Scores (P-values)**
- **Gene Set Analysis Tools, Methods Comparison**

Recent Progress

Microarray data analysis: from disarray to consolidation and consensus

David B. Allison†§, Xiangqin Cui*§, Grier P. Page* and Mahyar Sabripour**

NATURE REVIEWS | GENETICS VOLUME 7 | JANUARY 2006 | 55

- **Incorporating biological knowledge into analysis.**
- **Meta-analysis: pooling**
- **Well-curated publicly data set.**
- **Development of standardized testing platforms (e.g., AffyComp).**
- **Quality-control assessment.**
- **Gene set analysis (GSA)**

Recent Progress

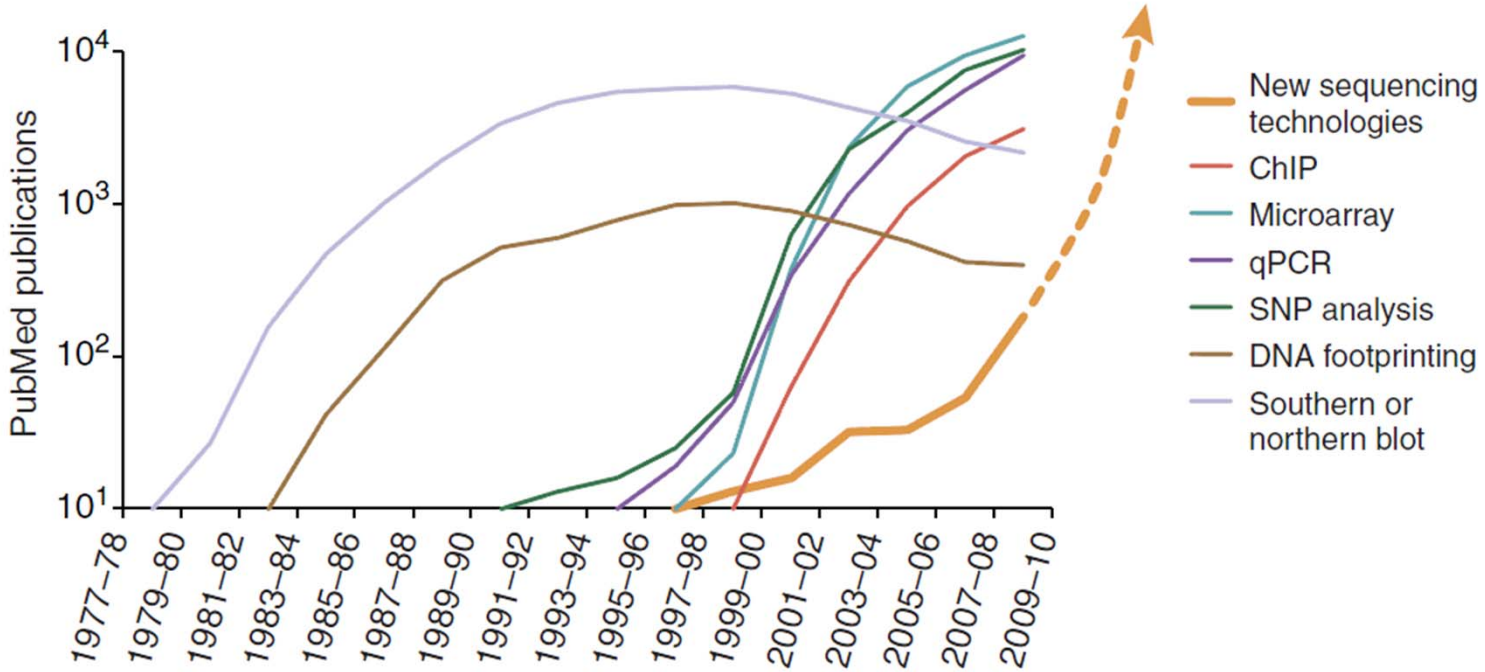
NATURE METHODS | VOL.5 NO.7 | JULY 2008 | 585

The beginning of the end for microarrays?

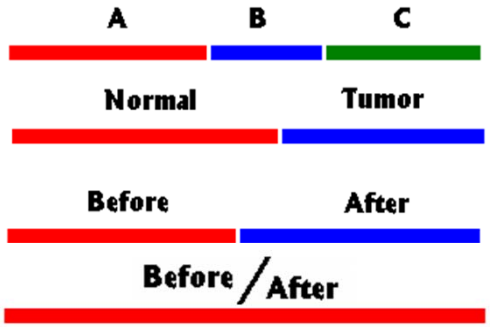
Jay Shendure

Two complementary approaches, both using next-generation sequencing, have successfully tackled the scale and the complexity of mammalian transcriptomes, at once revealing unprecedented detail and allowing better quantification.

Ref: Avak Kahvejian, John Quackenbush & John F Thompson, 2008, **What would you do if you could sequence everything?** Nature Biotechnology 26, 1125 - 1133



Finding Differentially Expressed Genes (DEGs)



→ More than two samples

→ Two-sample (independent)

Cy 5: treatment

Cy 3: control

MA Table	exp01	exp02	exp03	exp04	exp05	exp•••	exp p
gene001	-0.48	-0.42	0.87	0.92	0.67		-0.35
gene002	-0.39	-0.58	1.08	1.21	0.52		-0.58

→ Paired-sample (dependent)

MA Table	exp01	exp02	exp03	exp04	exp05	exp•••	exp p	p-values	
gene001	-0.48	-0.42	0.87	0.92	0.67		-0.35	0.067	
gene002	-0.39	-0.58	1.08	1.21	0.52		-0.58	0.052	
gene003	0.87	0.25	-0.17	0.18	-0.13		-0.13	0.013	*
gene004	1.57	1.03	1.22	0.31	0.16		-1.02	0.016	*
gene005	-1.15	-0.86	1.21	1.62	1.12		-0.44	0.112	
gene006	0.04	-0.12	0.31	0.16	0.17		0.08	0.017	*
gene007	2.95	0.45	-0.40	-0.66	-0.59		-0.76	0.059	
gene008	-1.22	-0.74	1.34	1.50	0.63		-0.55	0.063	
gene009	-0.73	-1.06	-0.79	-0.02	0.16		0.03	0.516	
gene010	-0.58	-0.40	0.13	0.58	-0.09		-0.45	0.009	*
gene011	-0.50	-0.42	0.66	1.05	0.68		0.01	0.068	
gene012	-0.86	-0.29	0.42	0.46	0.30		-0.63	0.030	*
gene013	-0.16	0.29	0.17	-0.28	-0.02		-0.04	0.002	*
gene014	-0.36	-0.03	-0.03	-0.08	-0.23		-0.21	0.423	
gene015	-0.72	-0.85	0.54	1.04	0.84		-0.64	0.084	
gene016	-0.78	-0.52	0.26	0.20	0.48		0.27	0.048	
gene017	0.60	-0.55	0.41	0.45	0.18		-1.02	0.018	*
gene018	-0.20	-0.67	0.13	0.10	0.38		0.05	0.538	
gene019	-2.29	-0.64	0.77	1.60	0.53		-0.38	0.053	
gene020	-1.46	-0.76	1.08	1.50	0.74		-0.70	0.074	
gene021	-0.57	0.42	1.03	1.35	0.64		-0.40	0.764	
gene022	-0.11	0.13	0.41	0.60	0.23		0.19	0.423	
gene•••									
gene n	-1.79	0.94	2.13	1.75	0.23		-0.66	0.723	

Microarray Data Matrix

ranking

p-values or Statistics

0.002 *
0.009 *
0.013 *
0.016 *
0.017 *
0.018 *
0.030 *
0.048
0.052
0.059
0.063
0.067
0.068
0.074
0.084
0.112
0.423
0.423
0.516
0.538
0.723
0.764
..

(1) fixed number

(2) above some level

Individual Gene Analysis (IGA)

- IGA evaluates the significance of individual genes between two groups of samples compared and yields a list of altered genes.
- The list is investigated with biologically defined gene sets derived from Gene Ontology or some pathway databases to assess the enrichment of specific biological themes in the list.

Gene	Description	Unigene Id	Score(d)	Fold Change
SDCCAG33	serologically defined colon cancer antigen 33	Hs.284217	0.9968087	4.65098
PTGER3	prostaglandin E receptor 3 (subtype EP3)	Hs.27860	0.9779319	4.26760
BMPR1A	bone morphogenetic protein receptor, type IA	Hs.2534	0.9773508	3.99990
AUTS2	autism susceptibility candidate 2	Hs.296720	0.9760794	3.38653
				3.11846
				2.61334
				2.98120
				4.59277
				2.82651
				3.30323
				5.57371
				2.78782
				4.98333
				2.54623
				3.17894
				1.89733
				3.67324
				3.29672
				3.43921
				2.99402
				2.26245
				2.62416

Gene	Description	Unigene Id	Score(d)	Fold Change
GJA12	gap junction protein, alpha 12, 47kDa	Hs.100072	0.9350071	4.59277
LOC83690	CocoaCrisp	Hs.436542	0.9044242	2.82651

Gene	Description	Unigene Id	Score(d)	Fold Change
TM4SF12	transmembrane 4 superfamily member 12	Hs.16529	1.4343708	6.11278
DIO2	deiodinase, iodothyronine, type II	Hs.436020	1.3718215	7.33664
CAPNS2	calpain small subunit 2	Hs.13359	1.3074648	[illegible]
CCKBR	cholecystokinin B receptor	Hs.203	1.2563882	[illegible]
LCN2	lipocalin 2 (oncogene 24p3)	Hs.204238	1.2002127	[illegible]
BLVRB	**biliverdin reductase B (flavin reductase (NADPH))	Hs.76289	1.1751826	5.45960
EHD4	EH-domain containing 4	Hs.55058	1.1646133	5.67989
MGC29898	hypothetical protein MGC29898	Hs.388749	1.1500615	5.20681
FLRT2	fibronectin leucine rich transmembrane protein 2	Hs.48998	1.1337755	6.33282
GUCY1A3	guanylate cyclase 1, soluble, alpha 3	Hs.433488	1.1045856	3.29975
TM4SF9	transmembrane 4 superfamily member 9	Hs.8037	1.1039144	4.61288
PPP2R3A	protein phosphatase 2 (formerly 2A), regulatory subunit B", alpha	Hs.133234	1.1008696	8.09162
ROBO2	roundabout, axon guidance receptor, homolog 2 (Drosophila)	Hs.31141	1.0970799	5.14947
GUCY1A3	guanylate cyclase 1, soluble, alpha 3	Hs.433488	1.0831366	4.01820
CD34	CD34 antigen	Hs.374990	1.0823162	2.66837
B3GAT1	beta-1,3-glucuronyltransferase 1 (glucuronosyltransferase P)	Hs.381050	1.0821947	4.39357
COL8A1	collagen, type VIII, alpha 1	Hs.114599	1.0747639	3.39593
WWP2	Nedd-4-like ubiquitin-protein ligase	Hs.315485	1.0729466	5.10464

Gene Lists

Biological Relevance

Hypothesis Testing and P-Values

Biological Question → ***Statistical Formulation***

Null Hypothesis:

$\mathbf{H_0}$: no differential expressed.

$\mathbf{H_0}$: no difference in the mean gene expression in the group tested.

$\mathbf{H_0}$: the gene will have equal means across every group.

$\mathbf{H_0}$: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ ($\ldots = \mu_n$)

P-Values:

Probability of observing your data under the assumption that the null hypothesis is true.

Decision Rule:

Reject H_0 if *p-value* is less than alpha. ($p < 0.05$ commonly used).

The lower the *p-value*, the more significant.

t-test Statistics

One sample t-test

$H_0 : \mu = \mu_0$
$H_1 : \mu \neq \mu_0$ (two-tailed).
μ: population mean.
α: significant level (e.g., 0.05).
Test Statistic:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}, \quad t_0 = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

$\bar{X}$: sample mean.

S: sample standard deviation.

n: number of observations in the sample.

- Reject H_0 if $|t_0| > t_{\alpha/2,n-1}$.
- Power $= 1 - \beta$.
- $(1-\alpha)100\%$ Confidence Interval for μ:
  $\bar{X} - t_{\alpha/2} S/\sqrt{n} \leq \mu < \bar{X} + t_{\alpha/2} S/\sqrt{n}$
- $p\text{-}value = P_{H_0}(|\mathbf{T}| > t_0)$, $\mathbf{T} \sim t_{n-1}$.

Two Sample t-test (Unpaired)

$H_0 : \mu_x - \mu_y = \mu_0$
$H_0 : \mu_x - \mu_y \neq \mu_0$
α: significant level (e.g., 0.05).
Test Statistic:

$$t_0 = \frac{(\bar{X} - \bar{Y}) - \mu_0}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}}$$

for homogeneous variances:
$df = n + m - 2$

for heterogeneous variances:
adjusted df

Reject H_0 if $|t_0| > t_{\alpha/2,\, df}$

Other t-Statistics for Microarray Data

B-statistic

Lonnstedt and Speed, Statistica Sinica 2002: parametric empirical Bayes approach.

- B-statistic is an estimate of the posterior log-odds that each gene is DE.
- B-statistic is equivalent for the purpose of ranking genes to the penalized t-statistic $t = \frac{\bar{M}}{\sqrt{(a+s^2)/n}}$, where a is estimated from the mean and standard deviation of the sample variances s^2.

Penalized t-statistic

Tusher et al (2001, PNAS, SAM)
Efron et al (2001, JASA)

$$t = \frac{\bar{M}}{(a+s)/\sqrt{n}}$$

$$M_{gj}|\mu_g, \sigma_g \sim N(\mu_g, \sigma_g^2)$$

$$B_g = \log \frac{P(\mu_g \neq 0|M_{gj})}{P(\mu_g = 0|M_{gj})}$$

General Penalized t-statistic

(Lonnstedt et al 2001)

$$t = \frac{b}{s^* \times SE}$$

multiple regression model

Penalized two-sample t-statistic

$$t = \frac{\bar{M}_A - \bar{M}_B}{s^* \times \sqrt{1/n_A + 1/n_B}}, \quad \text{where } s^* = \sqrt{a + s^2}$$

Robust General Penalized t-statistic

The Main Problems of IGA

- The final result of IGA is significantly affected by the selected threshold (which is normally chosen arbitrarily).
  - **Example: Pan et al.** showed that different choices of the threshold value severely alter the biological conclusions (enrichment of specific function categories in the gene list).

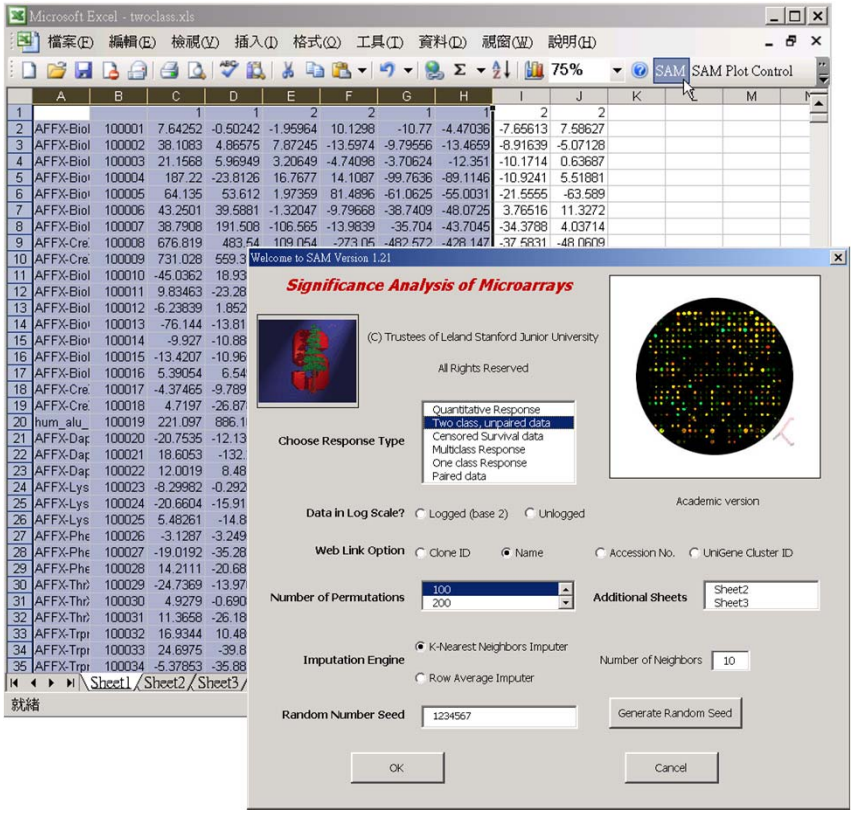
- Many genes with moderate but meaningful expression changes are discarded by the strict cutoff value, which leads to a reduction in statistical power.

- All the statistical methods applied are based on the wrong assumption of independent gene (or gene group) sampling, which increases false positive predictions.

Significance Analysis of Microarrays (SAM)

SAM: Significance Analysis of Microarrays

SAM: Supervised learning software for genomic expression data mining
http://www-stat.stanford.edu/~tibs/SAM/

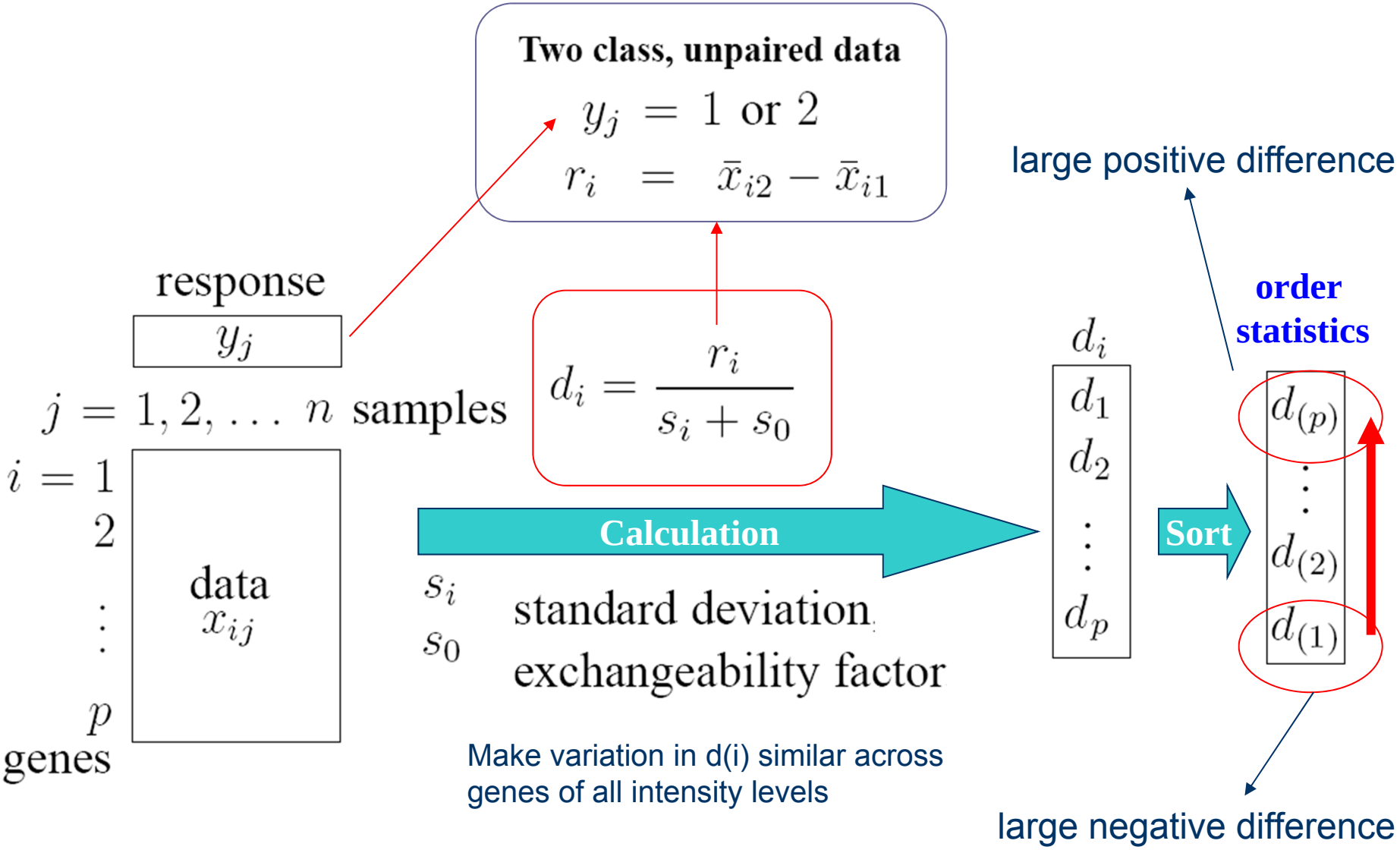


Response type	Coding
Quantitative	Real number eg 27.4 or -45.34
Two class (unpaired)	Integer 1, 2
Multiclass	Integer 1, 2, 3, ...
Paired	Integer -1, 1, -2, 2, etc. eg - means Before treatment, + means after treatment -1 is paired with 1, -2 is paired with 2, etc.
Survival data	(Time, status) pair like (50,1) or (120,0) First number is survival time, second is status (1=died, 0=censored)
One class	Integer, every entry equal to 1
Time course, two class (unpaired)	(1 or 2)Time(t)[Start or End]
Time course, two class (paired)	(-1 or 1 or -2 or 2 etc)Time(t)[Start or End]
Time course, one class	1Time(t)[Start or End]
Pattern discovery	eigengenek, where k is one of 1,2,... number of arrays

- SAM has facilities for Gene Set Analysis (**GSA**).
- GSA uses the "**maxmean**" statistic.

- Tusher VG, Tibshirani R, Chu G.(2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci* 98(9):5116-21.
- B. Efron and R. Tibshirani. On testing the significance of sets of genes. Annals of applied statistics, 2007

SAM: Significance Analysis of Microarrays

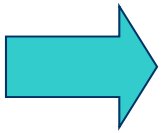


SAM: Expected Test Statistics

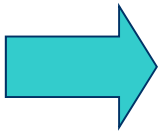
response

y_j

$1, 1, \ldots, 2, \ldots, 2$



$1, 2, 1, 2, 1, \ldots, 1$



$$r_i^* = \bar{x}_{i2}^* - \bar{x}_{i1}^*$$

$$d_i^* = \frac{r_i^*}{s_i^* + s_0^*}$$

Permutation

$$\begin{bmatrix} d_{(p)}^{*b} \\ \vdots \\ d_{(2)}^{*b} \\ d_{(1)}^{*b} \end{bmatrix} \quad b = 1, 2, \ldots B$$

$$\bar{d}_{(i)} = (1/B) \sum_b d_{(i)}^{*b}$$

expected order statistics

$$\begin{bmatrix} \bar{d}_{(p)} \\ \vdots \\ \bar{d}_{(2)} \\ \bar{d}_{(1)} \end{bmatrix}$$

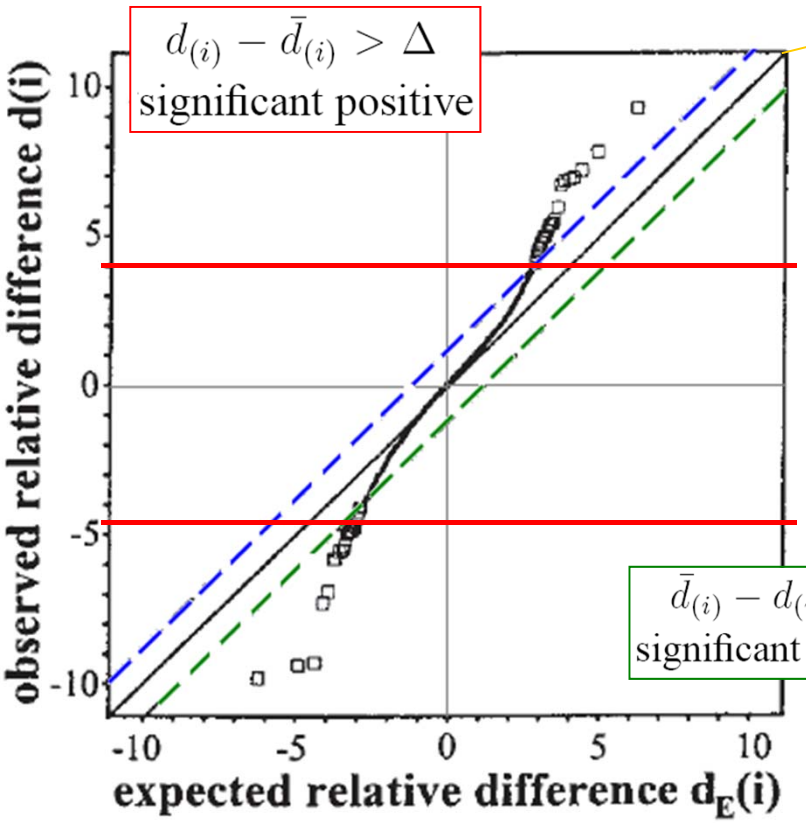
SAM Plot

Points for genes with evidence of induction

$$\begin{bmatrix} d_{(p)} \\ \vdots \\ d_{(2)} \\ d_{(1)} \end{bmatrix}$$

vs

$$\begin{bmatrix} \bar{d}_{(p)} \\ \vdots \\ \bar{d}_{(2)} \\ \bar{d}_{(1)} \end{bmatrix}$$



$d_{(i)} = \bar{d}_{(i)}$

upper cut-point $\mathrm{cut}_{up}(\Delta)$

lower cut-point $\mathrm{cut}_{low}(\Delta)$

Points for genes with evidence of repression

Gene Set Analysis

(GSA)

Gene Sets

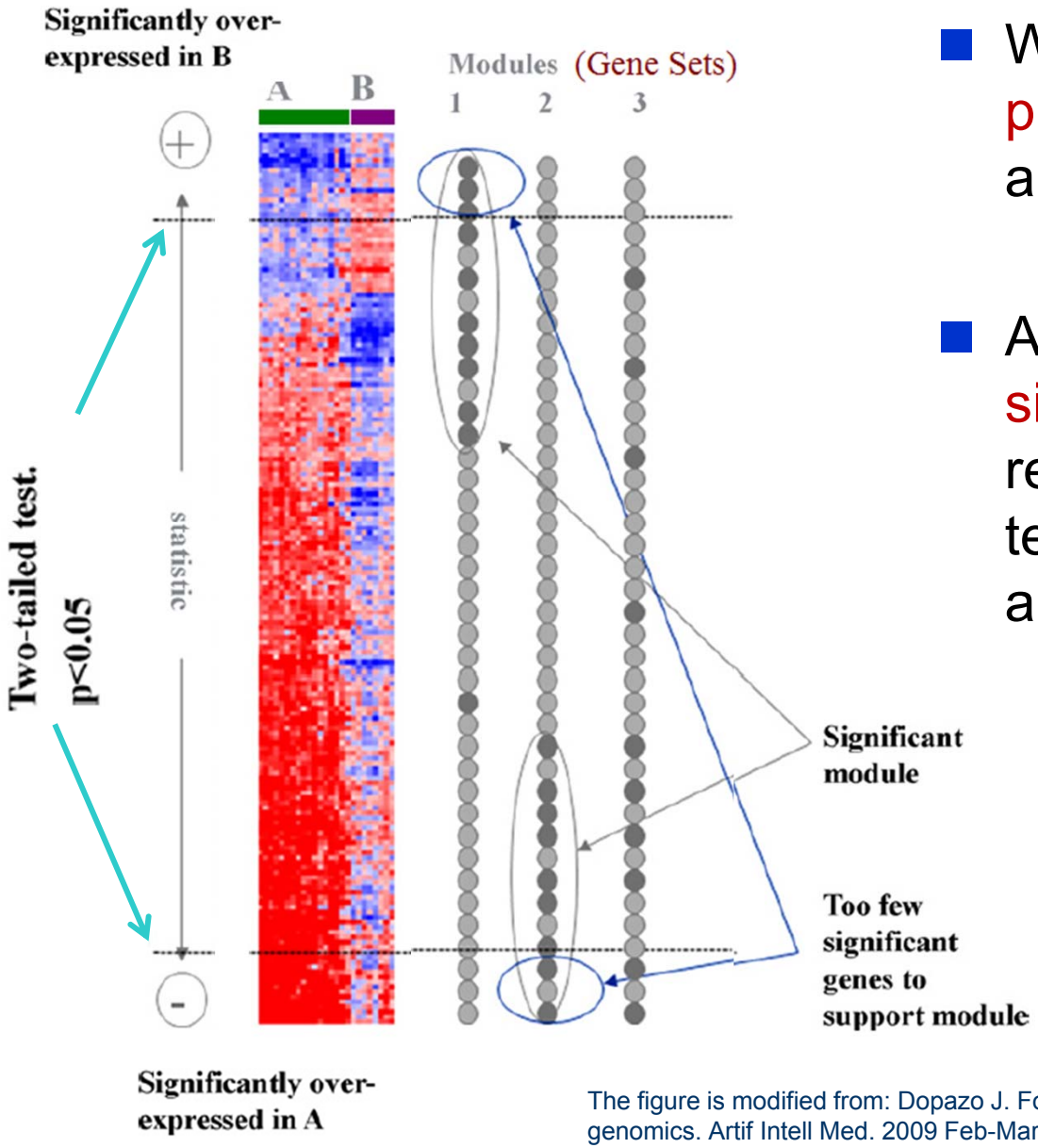
- **A gene set**
  - a group of genes with related functions.
  - sets of genes or pathways, for their association with a phenotype.
  - **Examples:** metabolic pathway, protein complex, or GO (gene ontology) category.
- Identified from a **prior** biological knowledge.
- May better reflect the true underlying biology.
- May be more appropriate **units** for analysis.

c2.symbols.gmt

M gene sets

	A	B	C	D	E	F	
1	41bbPathway	TNF-type receptor 4-1BB	IL2	TRAF2	MAP3K1	IFNG	m1
2	ace2Pathway	Angiotensin-converting enz	COL4A3	COL4A1	COL4A5	AGT	m2
3	acetaminophenPathway	Acetaminophen selectively	CYP3A	PTGS2	CYP1A2	PTGS1	m3
4	achPathway	Nicotinic acetylcholine rece	RAPSN	TERT	MUSK	PTK2	...
5	actinYPathway	The Arp 2/3 complex localiz	ACTR3	ABI-2	WASL	ARPC4	...
6	agpcrPathway	G-protein coupled receptor	PRKAR2A	GNGT1	PRKACB	PRKCB1	...
7	ahspPathway	Alpha-hemoglobin stabilizin	CPO	HMBS	ALAS1	ERAF	
8	aifPathway	BLACK	ADPRT	PDCD8	BCL2L1	CYCS	
9	akap13Pathway	A-kinase anchor protein 13	EDG4	PRKACG	PRKAR2A	PRKACB	

Gene Set Analysis

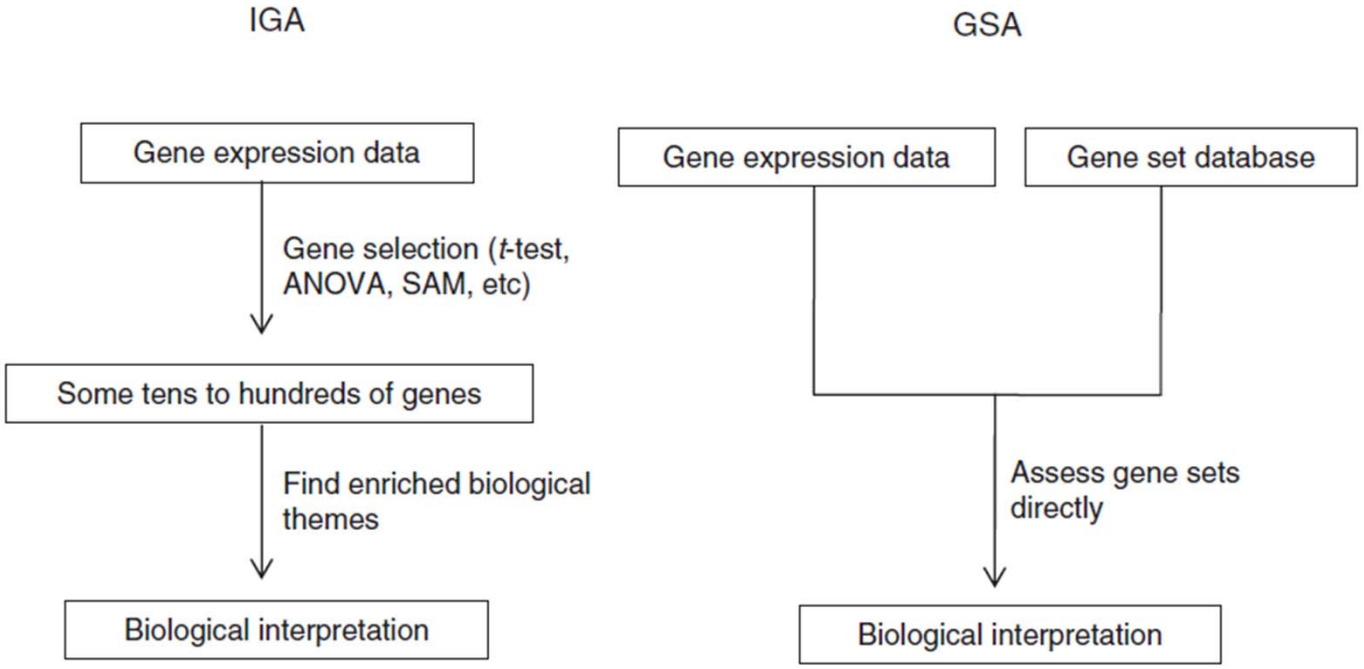


- Whether some functionally predefined classes of genes are differentially expressed?
- A statistical test to determine significance of a gene class is referred to as gene class testing (**GCT**) or gene set analysis (**GSA**).

The figure is modified from: Dopazo J. Formulating and testing hypotheses in functional genomics. Artif Intell Med. 2009 Feb-Mar;45(2-3):97-107.

IGA and GSA

- GSA aims to identify gene sets with 'subtle but coordinated' expression changes that cannot be detected by IGA methods.
  - even weak expression changes in individual genes gathered to a large gene set can show a significant pattern.
- Results of GSA are not affected by arbitrarily chosen cutoffs.
- GSA does not provide information as detailed as IGA.



Source: Dougu Nam and Seon-Young Kim, Gene-set approach for expression pattern analysis, Briefings In Bioinformatics. 2008, 9(3). 189-197.

Several benefits of GSA

From a statistical point of view:

GSA typically increases power and reduces the dimensionality of the underlying statistical problem.

From the biological perspective:

GSA help to understand the functional mechanism in a cell.

(1) is a certain pathway activated in a given tissue under some treatment x?

(2) is the pathway more active than other pathway?

Literature Review

- **Global Test** (global model with random effects): Goeman et al., **2004**
- **ANCOVA Global Test**: Mansmann and Meister, **2005**
- **GSEA**: Subramanian et al., **2005**
- Principal component analysis (**PCA**): Kong et al., **2006**
- Significance analysis of microarray for gene sets (**SAM-GS**): Dinu et al., **2007**
- Gene list analysis with prediction accuracy (**GLAPA**): Maglietta et al., **2007**
- **Maxmean**: Efron and Tibshirani, **2007**
- **exSAM-GS**: Adewale et al. **2008**
- Multivariate analysis of variance test (**MANOVA**, modified Hotelling's T2): Tsai and Chen, **2009**
- Linear combination Test (**LCT**): Wang, Dinu, Liu and Yasui, **2011**

- **Review:** Allison et al. 2006, Goeman and Buhlmann 2007, Nam and Kim **2008.**

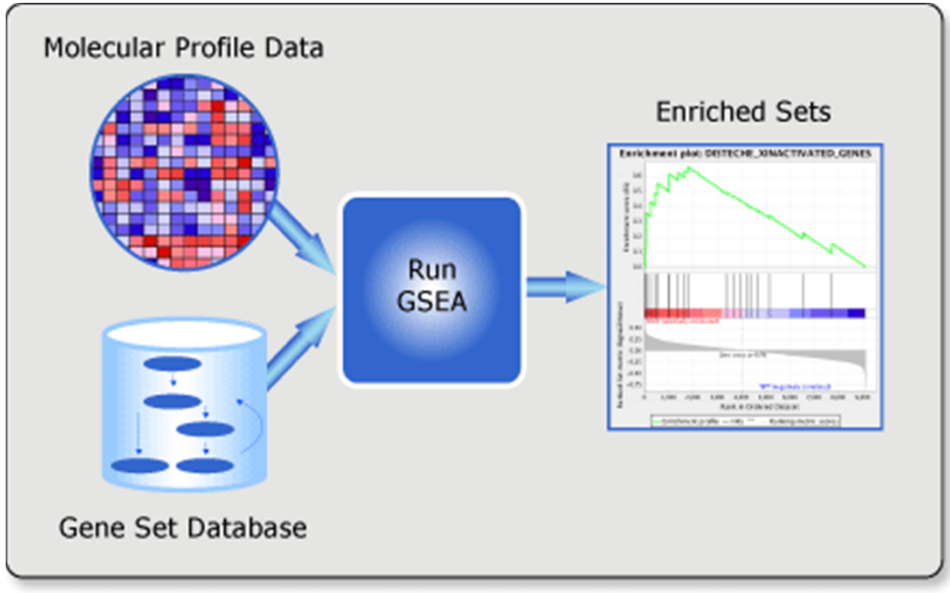
Gene Set Enrichment Analysis (GSEA)

GSEA (Subramanian et al., *PNAS*, 2005)

GSEA: Gene Set Enrichment Analysis

- GSEA was introduced by Mootha et al. 2003, and was used to identify pre-defined gene sets which exhibited significant differences in expression between samples from normal and diabetic patients.
- The methodology was subsequently refined by Subramanian et al. 2005.

Gene set enrichment analysis: a knowledge-based approac..
www.ncbi.nlm.nih.gov/pubmed/16199517 - 翻譯這個網頁
由 A Subramanian 著作 - 2005 - 被引用 2934 次 - 相關文章
Proc Natl Acad Sci U S A. 2005 Oct 25;102(43):15545-50. Epub 2005 Sep 30. Gene set enrichment analysis: a knowledge-based approach for interpreting ...

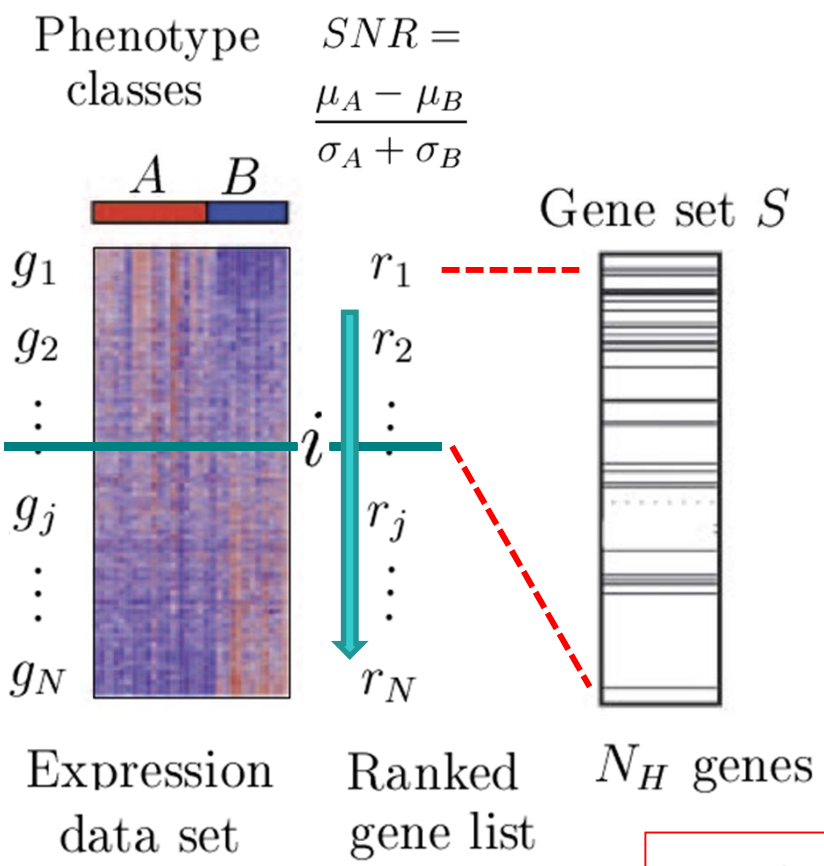


Subramanian, A., Kuehn, H., Gould, J., Tamayo, P., Mesirov, J.P. (2007) GSEA-P: A desktop application for Gene Set Enrichment Analysis. Bioinformatics, doi: 10.1093/bioinformatics/btm369.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S. & Mesirov, J. P. (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA 102, 15545-15550.

Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., et al. (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet 34, 267-273.

Step 1: Enrichment Score (ES)



Evaluate the fraction of genes in S ("hits") weighted by their correlation and the fraction of genes not in S ("misses") present up to a given position i in L.

$$P_{\text{hit}}(S,i) = \sum_{g_j \in S, j \le i} \frac{|r_j|^p}{N_R} \qquad N_R = \sum_{g_j \in S} |r_j|^p$$

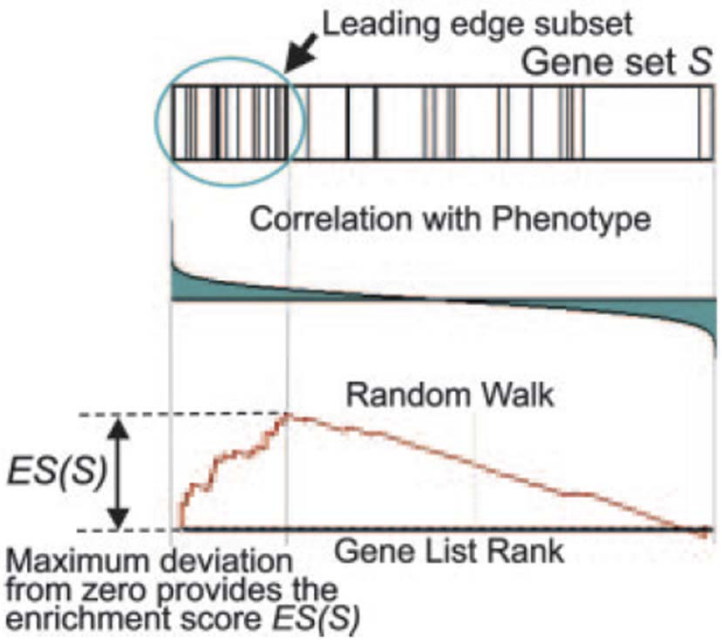
$$P_{\text{miss}}(S,i) = \sum_{g_j \notin S, j \le i} \frac{1}{N - N_H}$$

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

ES(S) > 0: gene set enrichment at the top of the ranked list.
ES(S) < 0: gene set enrichment at the bottom of the ranked list.

Enrichment Plot

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



Subramanian et al., PNAS 102(43), 15545–15550 (2005).

- If p=0

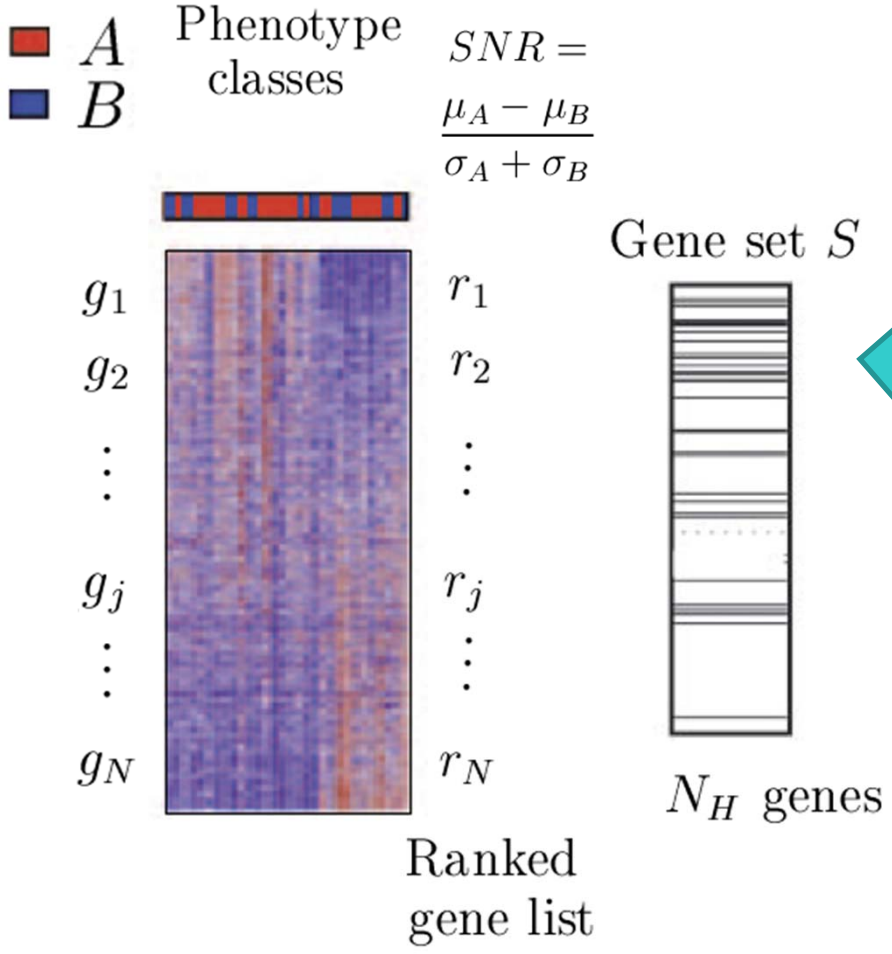
  ES(S) = Kolmogorov-Smirnov statistic.
- Set p=1.

- For a randomly distribued S, $ES(S)$ will be relatively small.
- It is concentrated at the top or bottom of hte list,

  or nonrandomly distributed, then $ES(S)$ will be corresponding high.

Step 2: Estimating Significance

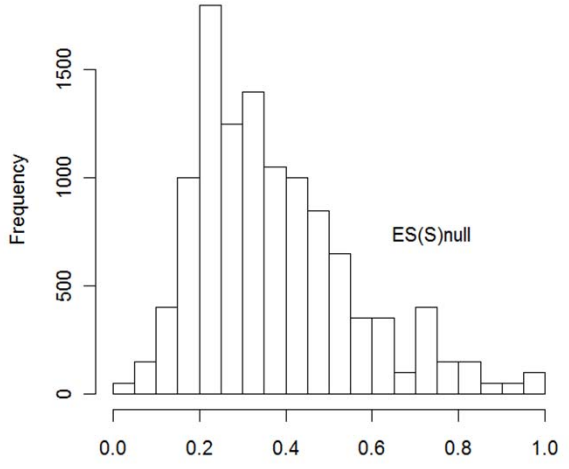
Assess the significance of an observed *ES* by comparing it with the set of score *ES(null)* computed with randomly assigned phonotype.



$$ES^{(b)}(S), b = 1, \cdots, 1000$$

- For positive ES
- For negative ES

$$\text{p-value} \approx \frac{\#\{ES^{(b)} > ES_{obs}\}}{\#permutations}$$



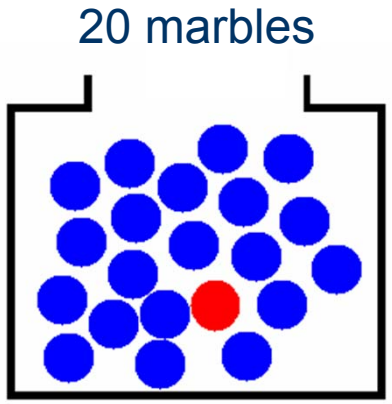
Step 3: Multiple Hypothesis Testing

X: false positive gene

$$P(X \geq 1)$$

$$= 1 - P(X = 0)$$

$$= 1 - 0.95^n$$



Population

Number of genes tested (N)	False positives incidence	Probability of calling 1 or more false positives by chance ($100(1-0.95^N)$)
1	1/20	5%
2	1/10	10%
20	1	64%
100	5	99.4%

Step 3: Multiple Hypothesis Testing

- When an entire database of gene sets is evaluated, we adjust the estimated significance level to account for multiple hypothesis testing.
  - Normalize ES for each gene set to account for the size of the set (**NES**).
  - Control the proportion of false positives by calculating the false discovery rate (**FDR**) corresponding to each NES.
- **FDR**
  - It is the estimated probability that a set with a given NES represents a false positive finding.
  - it is computed by comparing the tails of the observed and null distributions for the NES.

GSEA Software

GSEA
Gene Set Enrichment Analysis

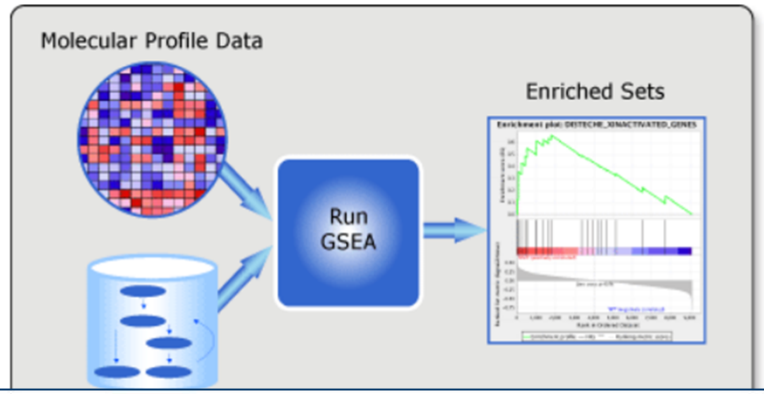
GSEA Home | Downloads | Molecular Signatures Database | Documentation | Contact

Overview

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

From this web site, you can:

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



What's New

29-Mar-2011: Version 3.7 of this web s
updated the text for improved clarity, a
have been made to the MSigDB data o

09-Sep-2010: We are pleased to annou
Molecular Signatures Database (MSigD
and expanded version of the C2 collect
gene sets. In addition, we have made s
MSigDB website, and fixed an error in t
further details, see the release notes.

BIOINFORMATICS APPLICATIONS NOTE
Vol. 23 no. 23 2007, pages 3251–3253
doi:10.1093/bioinformatics/btm369

Gene expression

GSEA-P: a desktop application for Gene Set Enrichment Analysis

Aravind Subramanian, Heidi Kuehn, Joshua Gould, Pablo Tamayo and Jill P. Mesirov*
Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA

Received on June 18, 2007; revised on June 18, 2007; accepted on July 9, 2007
Advance Access publication July 20, 2007
Associate Editor: Olga Troyanskaya

Institute of General Medical Sciences.

Downloads (register first!)

User Guide: http://www.broadinstitute.org/gsea/doc/GSEAUserGuideFrame.html
Quick Tour: http://www.broadinstitute.org/gsea/doc/desktop_tutorial.jsp

Downloads

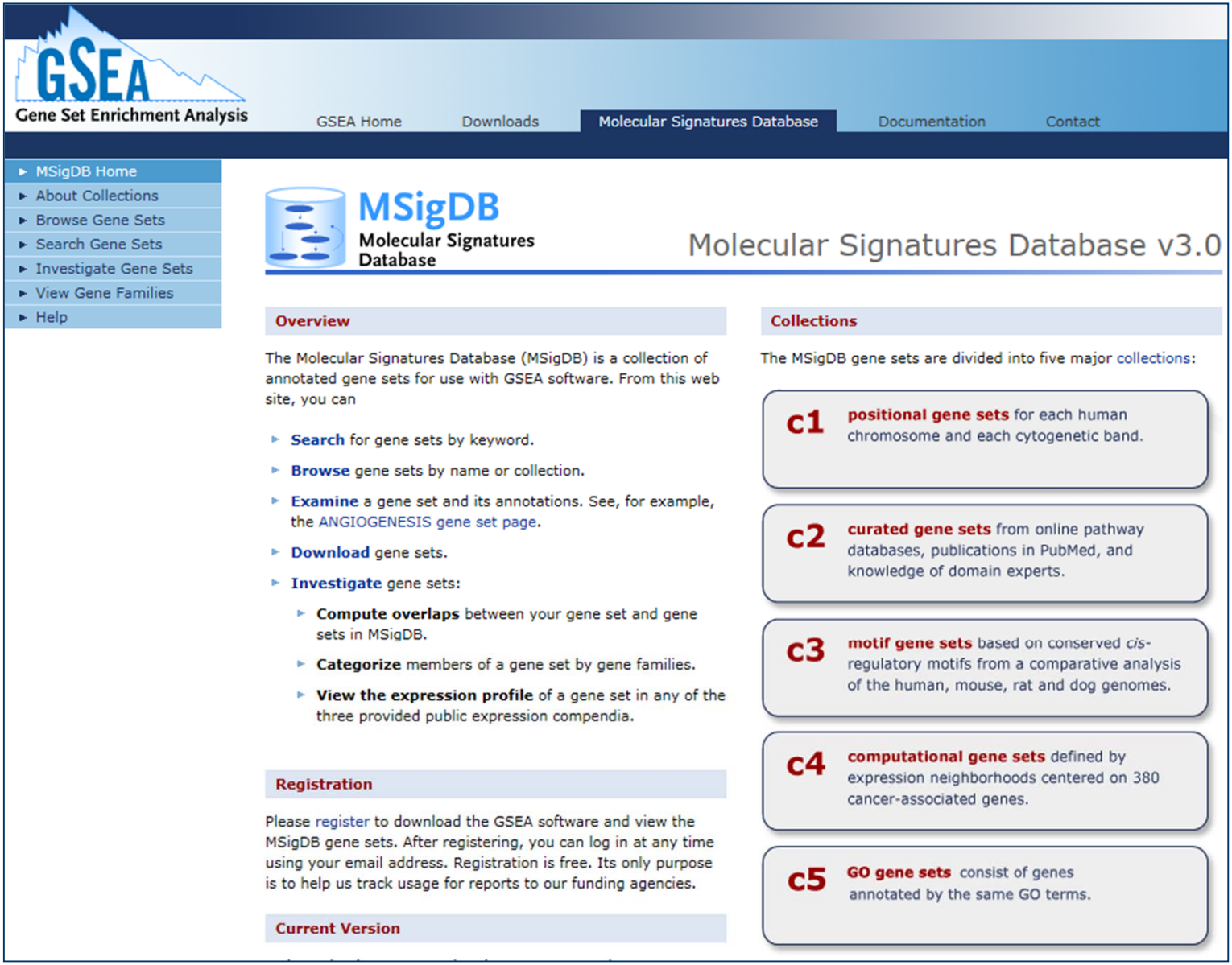
The GSEA software and source code and the Molecular Signatures Database (MSigDB) are freely available to individuals in both academia and industry for internal research purposes. Please see the GSEA/MSigDB license for more details.

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 1.6 or higher. If your computer has Java 1.5 and cannot upgrade to Java 1.6, please see the FAQ.

javaGSEA Desktop Application	▸ Easy-to-use graphical user interface ▸ Runs on any desktop computer (Windows, Mac OS X, Linux etc.) that supports Java1.6+ ▸ 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]
javaGSEA Java Jar file	▸ Command line usage ▸ Runs on any platform that supports Java1.6+ ▸ We recommend using the 'Launch' buttons above instead of this mode for most users	download gsea2-2.07.jar
GSEA Java Source Code Java source files	▸ 100% Java implementation of GSEA ▸ Incorporate GSEA into your own data analysis pipeline ▸ Programmatically call the open source GSEA java API	download gsea2_distrib-2.04.zip
R-GSEA R Script	▸ 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 ▸ Click here to learn more about the R-GSEA script	download GSEA-P-R.1.0.zip
GenePattern GSEA Module	▸ Use GSEA from within GenePattern ▸ Use GSEA in concert with a large suite of other analytics found in GenePattern (a powerful and flexible analysis platform developed at the Broad Institute)	GenePattern site

Molecular Signatures database (MsigDB)



Example Datasets

Example Datasets

DATASET	DESCRIPTION	RELEVANT DATA *(save link to download)*	REFERENCE
Gender	Transcriptional profiles from male and female lymphoblastoid cell lines Results of C1 GSEA analysis of this dataset Results of C2 GSEA analysis of this dataset	Gender_hgu133a.gct Gender_collapsed.gct Gender.cls	*Unpublished*
p53	Transcriptional profiles from p53+ and p53 mutant cancer cell lines Results of C2 GSEA analysis of this dataset	P53_hgu95av2.gct P53_collapsed.gct P53.cls	*Unpublished*
Diabetes	Transcriptional profiles of smooth muscle biopsies of diabetic and normal individuals Results of C2 GSEA analysis of this dataset	Diabetes_hgu133a.gct Diabetes_collapsed.gct Diabetes.cls	*Mootha et al. (2003) Nat Genet 34 (3): 267-73*
Leukemia	Transcriptional profiles from leukemias - ALL and AML Results of C1 GSEA analysis of this dataset	Leukemia_hgu95av2.gct Leukemia_collapsed.gct Leukemia.cls	*Armstrong et al. (2002) Nat Genet 30(1): 41-7.*
Lung cancer	Transcriptional profiles from two independent lung cancer outcome datasets	Lung_Michigan_hu6800.gct Lung_Michigan_collapsed.gct Lung_Mich_collapsed_common_Mich_Bost.gct Lung_Michigan.cls Lung_Boston_hgu95av2.gct Lung_Boston_collapsed.gct Lung_Bost_collapsed_common_Mich_Bost.gct Lung_Boston.cls	*Beer et al. (2002) Nat Med 8(8): 816-24.* *Bhattacharjee et al. (2001) Proc Natl Acad Sci U S A 98(24): 13790 -5.*
Gene sets	Archived gene sets from the GSEA PNAS 2005 publication. **Note:** This collection of gene sets is not the latest version, so when beginning a new analysis you might want to download the current collection of gene sets from the Downloads page.	C1.symbols.gmt (positional) C2.symbols.gmt (curated)	*Subramanian and Tamayo PNAS 2005*

P53 Status in Cancer Cell Lines

- NCI-60 collection of cancer cell lines.
  - Past usage: to identify targets of the transcription factor p53, which regulates gene expression in response to various signals of cellular stress.
  - The mutational status of the p53 gene has been reported for 50 of the NCI-60 cell lines: 17 normal, and 33 mutations.

GSEA: to identify functional gene sets (C2) correlated with p53 status.

- (p53+ > p53-): five gene sets.
- (p53- > p53+): one sig. gene set + two gene sets.

Gene set	FDR
Data set: p53 status in NCI-60 cell lines	
Enriched in p53 mutant	
Ras signaling pathway	0.171
Enriched in p53 wild type	
Hypoxia and p53 in the cardiovascular system	<0.001
Stress induction of HSP regulation	<0.001
p53 signaling pathway	<0.001
p53 up-regulated genes	0.013
Radiation sensitivity genes	0.078

LES: (p53- > p53+) whether three gene sets reflect a common biological function.

- resulting 16, 11, 13 genes.
- 4 gene in common: MAPK pathway.

Fig. 3. Leading edge overlap for p53 study. This plot shows the *ras*, *ngf*, and *igf1* gene sets correlated with P53⁻ clustered by their leading-edge subsets indicated in dark blue. A common subgroup of genes, apparent as a dark vertical stripe, consists of MAP2K1, PIK3CA, ELK1, and RAF1 and represents a subsection of the MAPK pathway.

Input for GSEA (1)

Demo Dataset: Transcriptional profiles from p53+ and p53 mutant cancer cell lines

Data File	Content	Format	Source
Expression dataset	Contains features (genes or probes), samples, and an expression value for each feature in each sample. Expression data can come from any source (Affymetrix, Stanford cDNA, and so on).	res, gct, pcl, or txt	You create the file.
Phenotype labels	Contains phenotype labels and associates each sample with a phenotype.	cls	You create the file or have GSEA create it for you.

P53_hgu95av2.gct

	A	B	C	D	E
1	#1.2				
2	12625	50			
3	NAME	Description	786-0	BT-549	C
4	100_g_at	na	215.37	132.94	
5	1000_at	na	328.68	234.31	
6	1001_at	na	39.64	8.84	
7	1002_f_at	na	18.46	12.14	
8	1003_s_at	na	60.83	30.19	
9	1004_at	na	68.02	54.41	
10	1005_at	na	610.35	65.93	
11	1006_at	na	12.79	3.57	
12	1007_s_at	na	354.92	208.33	

P53_collapsed_symbols.gct

	A	B	C	D	E	F
1	#1.2					
2	10100	50				
3	NAME	DESCRIPTION	786-0	BT-549	CCRF-CEM	COLO 205
4	TACC2	na	46.05	82.17	16.87	98.6
5	C14orf132	na	108.34	59.04	25.61	33.11
6	AGER	na	42.2	25.75	76.01	40.41
7	32385_at	na	7.43	13.94	8.55	21.13
8	RBM17	na	11.4	3	3.16	2.34
9	DYT1	na	148.09	317.17	316.66	147.23
10	CORO1A	na	8.62	9.12	1572.53	5.91
11	WT1	na	206.74	136.71	141.34	129.09

P53.cls

	A	B	C	D	E	F	G	H	I	J	K	L	M	...	AU	AV	AW	AX
1	50	2	1															
2	#MUT	WT																
3	MUT	MUT	MUT	MUT	MUT	MUT	MUT	MUT	MUT	MUT	MUT	MUT	MUT		WT	WT	WT	WT

Input for GSEA (2)

Data File	Content	Format	Source
Gene sets	Contains one or more gene sets. For each gene set, gives the gene set name and list of features (genes or probes) in that gene set.	gmx or gmt	You use the files on the Broad ftp site, export gene sets from the Molecular Signature Database (MSigDb) or create your own gene sets file.
Chip annotations	Lists each probe on a DNA chip and its matching HUGO gene symbol. Optional for the gene set enrichment analysis.	Chip	You use the files on the Broad ftp site, download the files from the GSEA web site, or create your own chip file.

c1.symbols.gmt

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

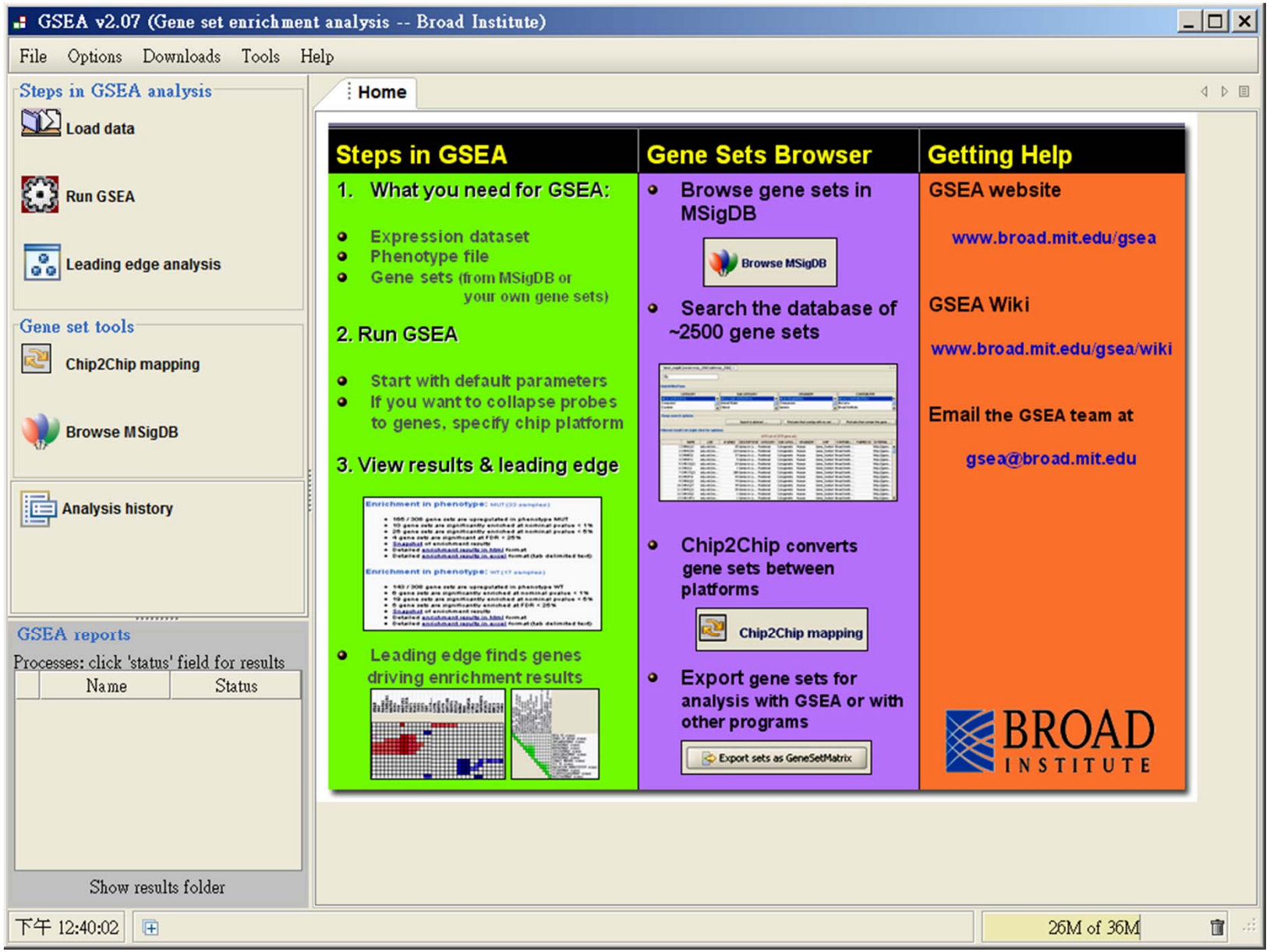
	A	B	C	D	E	F	G
1	chr10q24	Cytogenetic band	PITX3	SPFH1	NEURL	C10orf12	NDUF
2	chr5q23	Cytogenetic band	ALDH7A1	IL13	8-Sep	IRF1	ACSL6
3	chr8q24	Cytogenetic band	HAS2	LRRC14	TSTA3	DGAT1	
4	chr16q24	Cytogenetic band	RPL13	GALNS	FANCA	CPNE7	
5	chr13q14	Cytogenetic band	AKAP11	ARL11	ATP7B	C13orf1	C13orf
6	chr7p21	Cytogenetic band	ARL4A	SCIN	GLCCI1	SP8	SOSTD
7	chr10q23	C					

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

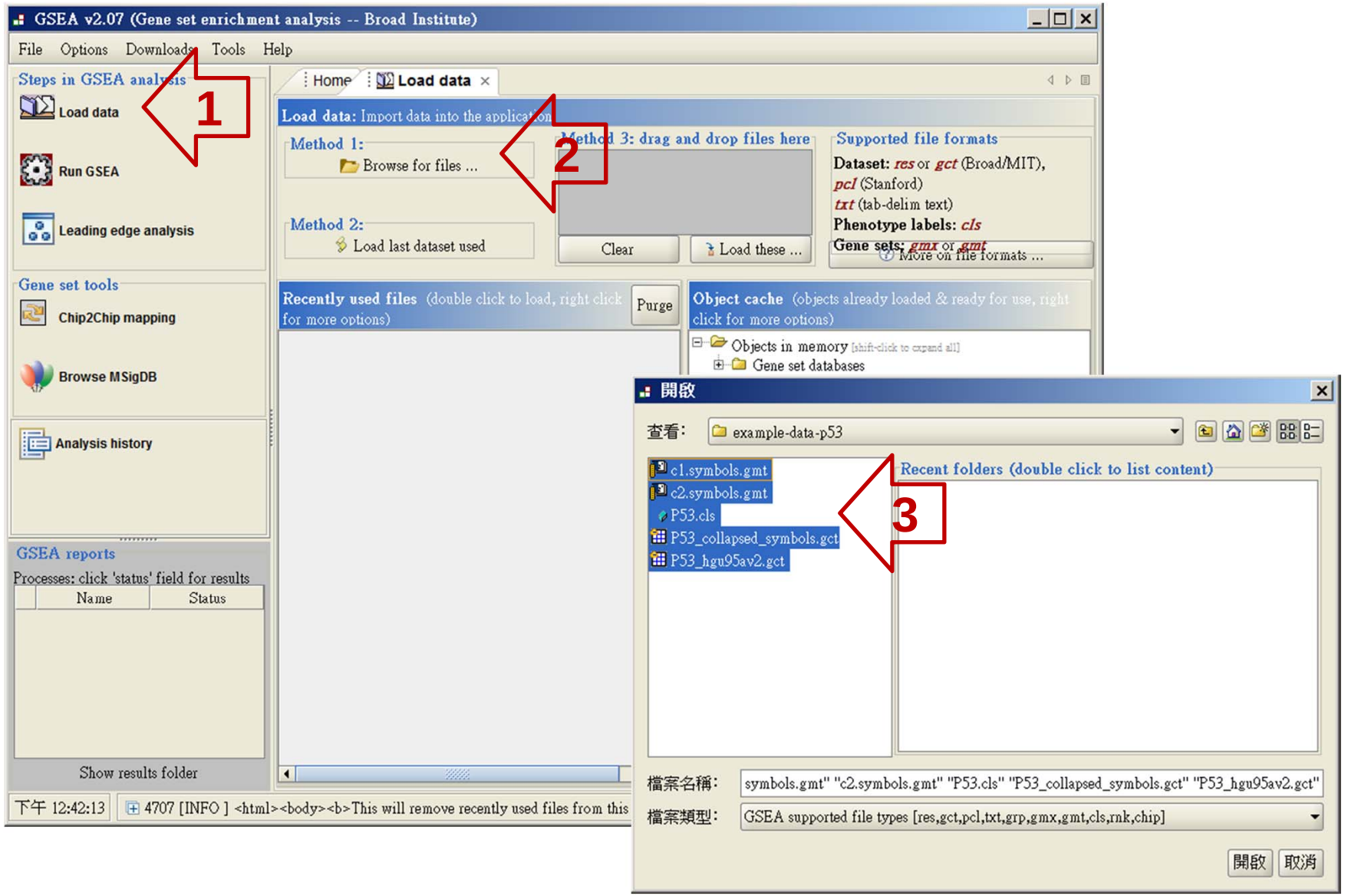
c2.symbols.gmt

	A	B	C	D	E	F
1	41bbPathway	TNF-type receptor 4-1BB is	IL2	TRAF2	MAP3K1	IFNG
2	ace2Pathway	Angiotensin-converting enz	COL4A3	COL4A1	COL4A5	AGT
3	acetaminophenPathway	Acetaminophen selectively	CYP3A	PTGS2	CYP1A2	PTGS1
4	achPathway	Nicotinic acetylcholine rece	RAPSN	TERT	MUSK	PTK2
5	actinYPathway	The Arp 2/3 complex localiz	ACTR3	ABI-2	WASL	ARPC4
6	agpcrPathway	G-protein coupled receptor	PRKAR2A	GNGT1	PRKACB	PRKCB1
7	ahspPathway	Alpha-hemoglobin stabilizin	CPO	HMBS	ALAS1	ERAF
8	aifPathway	BLACK	ADPRT	PDCD8	BCL2L1	CYCS
9	akap13Pathway	A-kinase anchor protein 13	EDG4	PRKACG	PRKAR2A	PRKACB

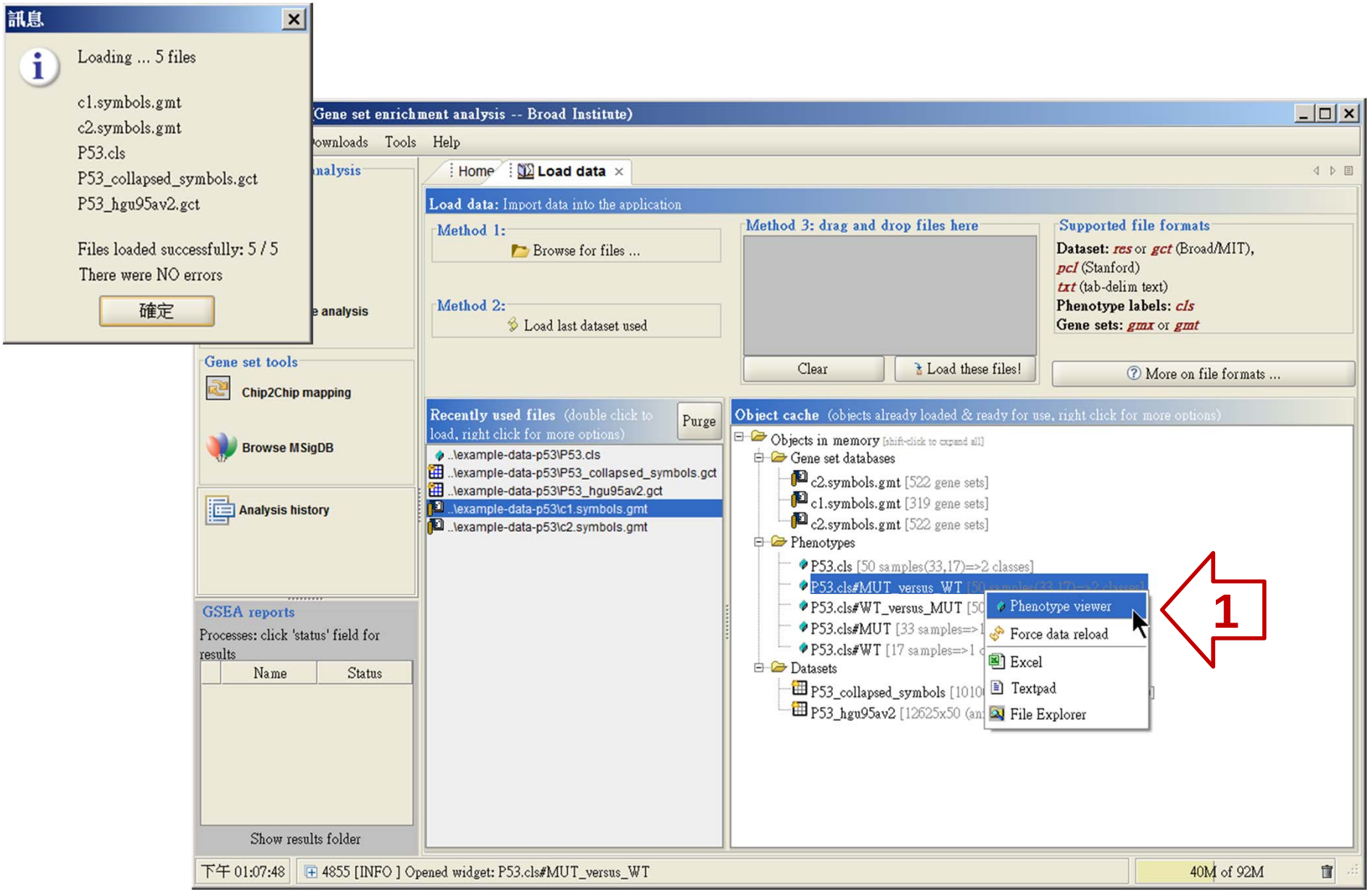
Launch GSEA


GSEA v2.07 (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
Steps in GSEA
1. What you need for GSEA:
Expression dataset
Phenotype file
Gene sets (from 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 & 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
下午 12:40:02
26M of 36M

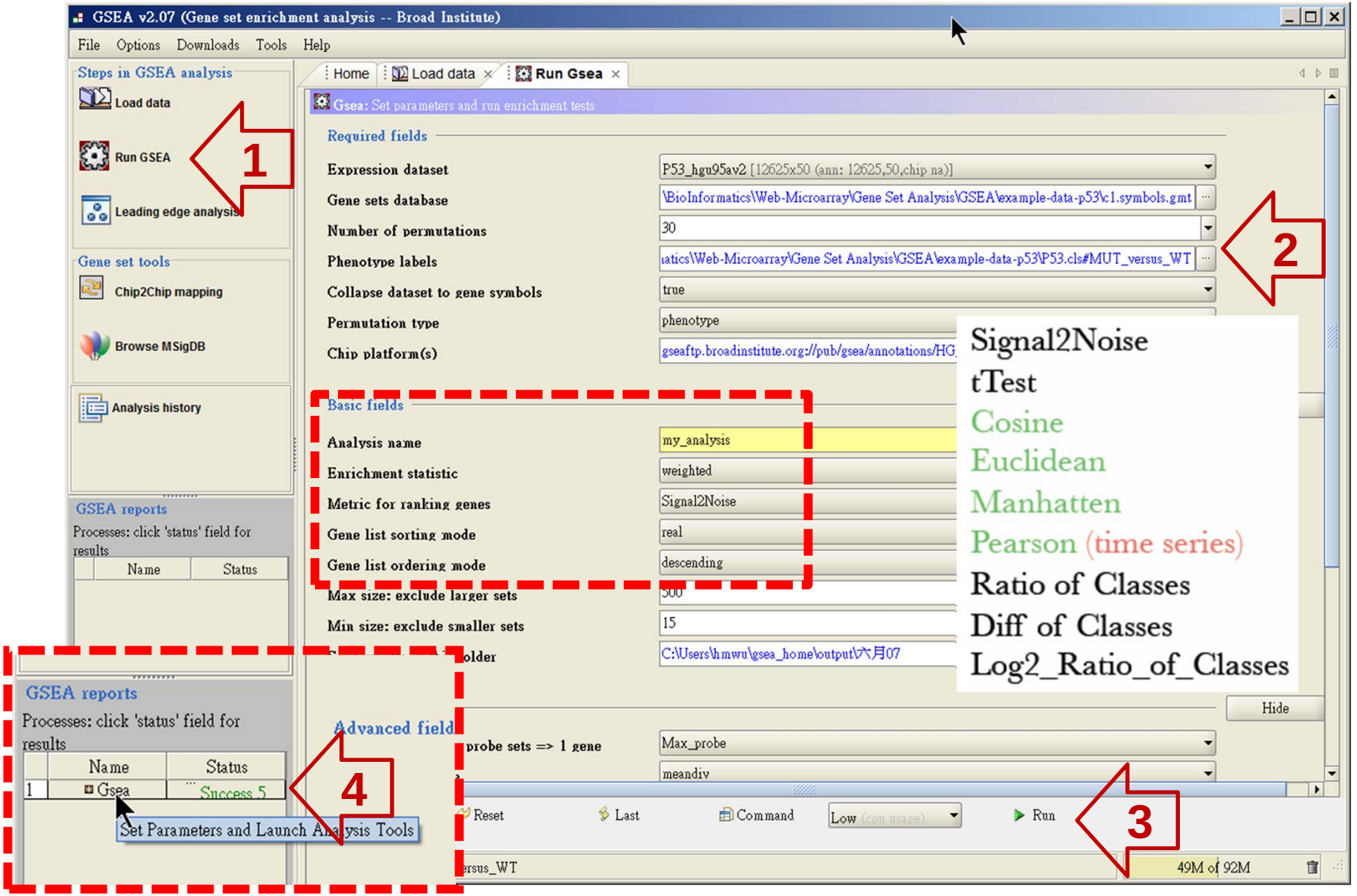
Load Data


GSEA v2.07 (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
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 ...
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]
Gene set databases
1
2
3
開啟
查看:
example-data-p53
c1.symbols.gmt
c2.symbols.gmt
P53.cls
P53_collapsed_symbols.gct
P53_hgu95av2.gct
Recent folders (double click to list content)
檔案名稱:
symbols.gmt" "c2.symbols.gmt" "P53.cls" "P53_collapsed_symbols.gct" "P53_hgu95av2.gct"
檔案類型:
GSEA supported file types [res,gct,pcl,txt,grp,gmx,gmt,cls,rnk,chip]
開啟
取消
下午 12:42:13
4707 [INFO] <html><body><b>This will remove recently used files from this

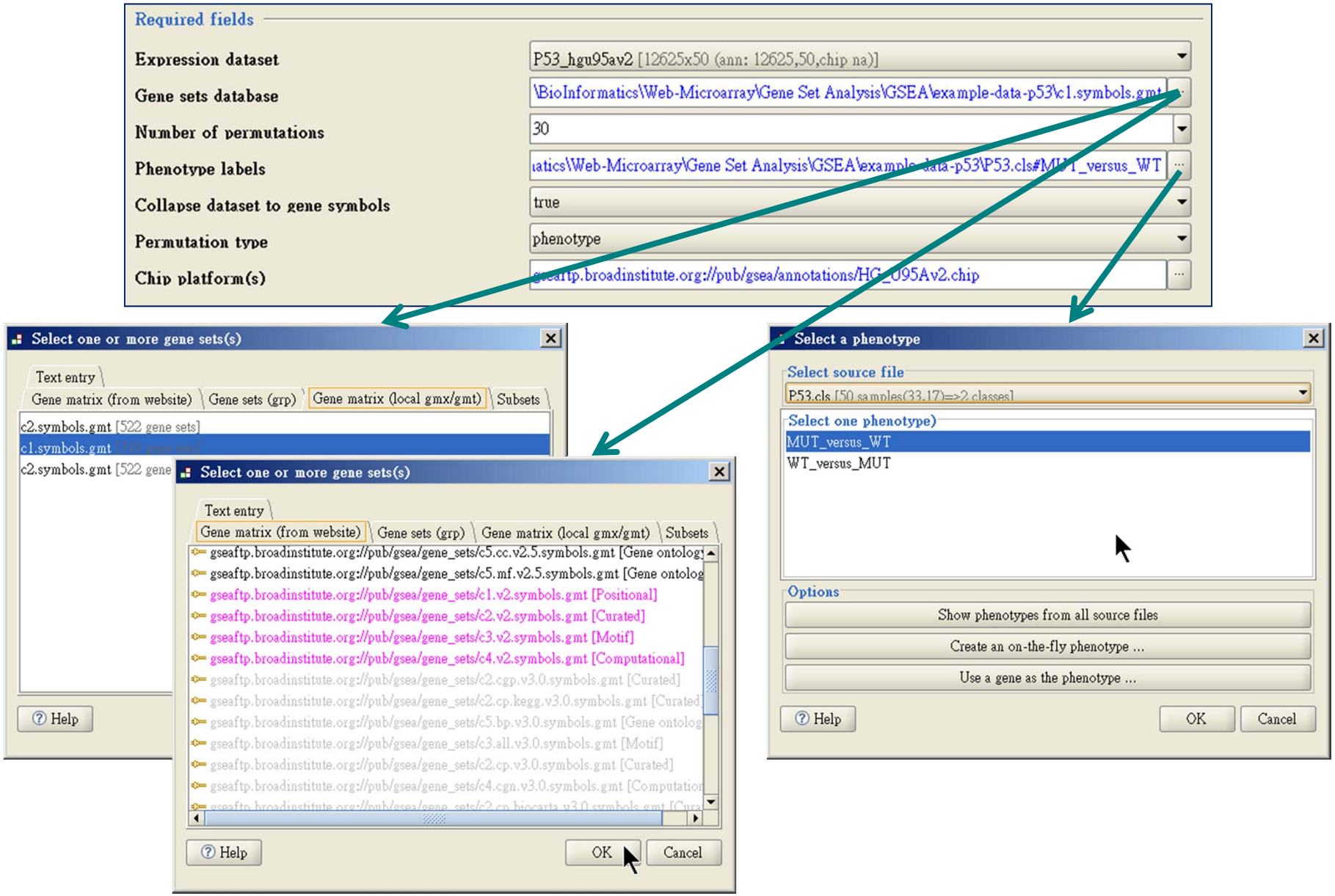
Explore Inputs



Run GSEA


GSEA v2.07 (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
Home Load data Run Gsea
Gsea: Set parameters and run enrichment tests
Required fields
Expression dataset
P53_hgu95av2 [12625x50 (ann: 12625,50,chip na)]
Gene sets database
\BioInformatics\Web-Microarray\Gene Set Analysis\GSEA\example-data-p53\c1.symbols.gmt
Number of permutations
30
Phenotype labels
atics\Web-Microarray\Gene Set Analysis\GSEA\example-data-p53\P53.cls#MUT_versus_WT
Collapse dataset to gene symbols
true
Permutation type
phenotype
Chip platform(s)
gseaftp.broadinstitute.org://pub/gsea/annotations/HG
Basic fields
Analysis name
my_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
C:\Users\hmwu\gsea_home\output\六月07
Signal2Noise
tTest
Cosine
Euclidean
Manhatten
Pearson (time series)
Ratio of Classes
Diff of Classes
Log2_Ratio_of_Classes
Hide
Advanced field
probe sets => 1 gene
Max_probe
meandiv
Reset
Last
Command
Low (cpu usage)
Run
ersus_WT
49M of 92M
1
2
3
4
Set Parameters and Launch Analysis Tools

Required Fields



Report

GSEA Report for Dataset P53_hgu95av2

Enrichment in phenotype: MUT (33 samples)

- 71 / 176 gene sets are upregulated in phenotype **MUT**
- 0 gene sets are significant at FDR < 25%
- 4 gene sets are significantly enriched at nominal pvalue < 1%
- 4 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: WT (17 samples)

- 105 / 176 gene sets are upregulated in phenotype **WT**
- 15 gene sets are significantly enriched at FDR < 25%
- 15 gene sets are significantly enriched at nominal pvalue < 1%
- 15 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in excel format (tab delimited text)
- Guide to interpret results

Dataset details

- The dataset has 12625 native features
- After collapsing features into gene symbols, there are: 9096 genes

Gene set details

- Gene set size filters (min=15, max=500) resulted in filtering out 143 / 319 gene sets
- The remaining 176 gene sets were used in the analysis
- List of gene sets used and their sizes (restricted to features in the specified dataset)

Gene markers for the MUT *versus* WT comparison

- The dataset has 9096 features (genes)
- # of markers for phenotype **MUT**: 4076 (44.8%) with correlation area 42.2%
- # of markers for phenotype **WT**: 5020 (55.2%) with correlation area 57.8%
- Detailed rank ordered gene list for all features in the dataset
- Heat map and gene list correlation profile for all features in the dataset
- Buttefly plot of significant genes

Global statistics and plots

- Plot of p-values *vs.* NES
- Global ES histogram

Other

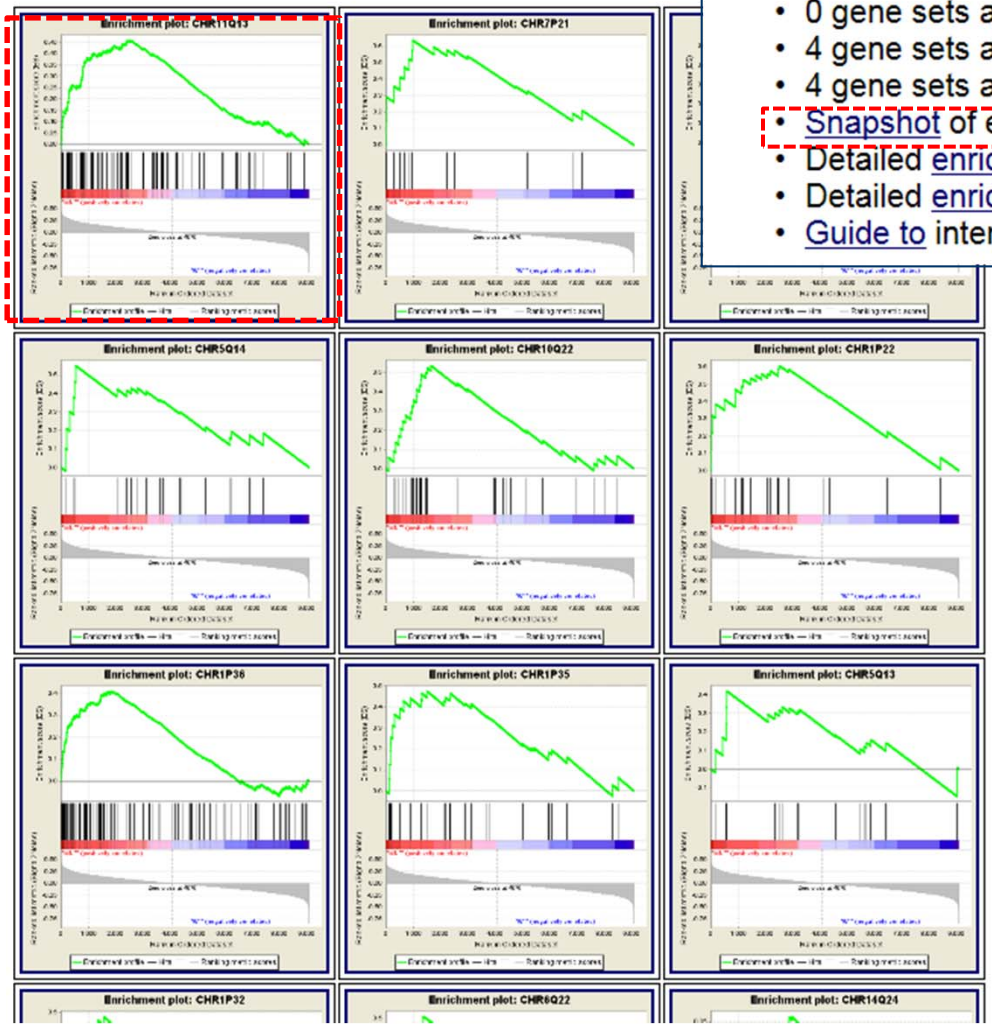
- Parameters used for this analysis

Interpretation

Enrichment in phenotype: MUT (33 samples)

- 71 / 176 gene sets are upregulated in phenotype **MUT**
- 0 gene sets are significant at FDR < 25%
- 4 gene sets are significantly enriched at nominal pvalue < 1%
- 4 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

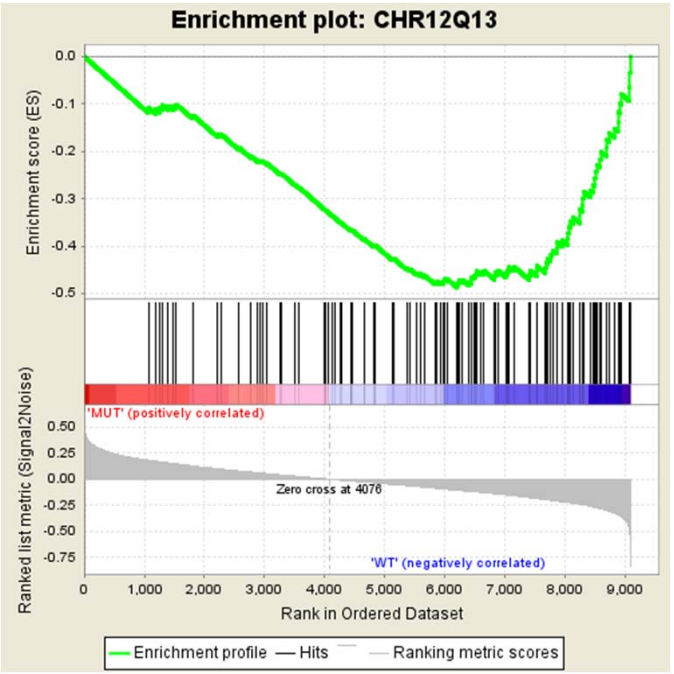
Table: Snapshot of enrichment results

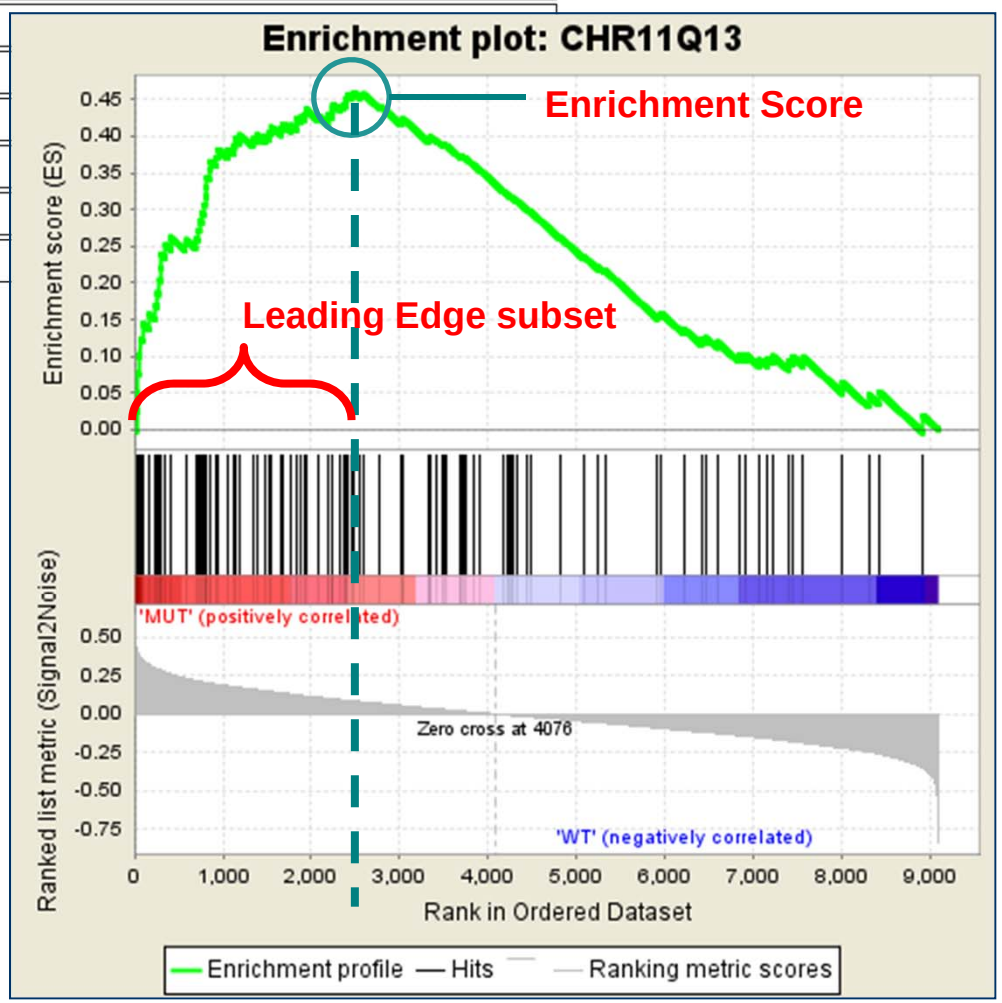


Enrichment plot

Table: GSEA Results Summary

Dataset	P53_hgu95av2_collapsed_to_symbols.P53.cls#MUT_versus_WT
Phenotype	P53.cls#MUT_versus_WT
Upregulated in class	MUT
GeneSet	CHR11Q13
Enrichment Score (ES)	0.45963296
Normalized Enrichment Score (NES)	1.6873256
Nominal p-value	0.0
FDR q-value	1.0
FWER p-Value	0.6666667



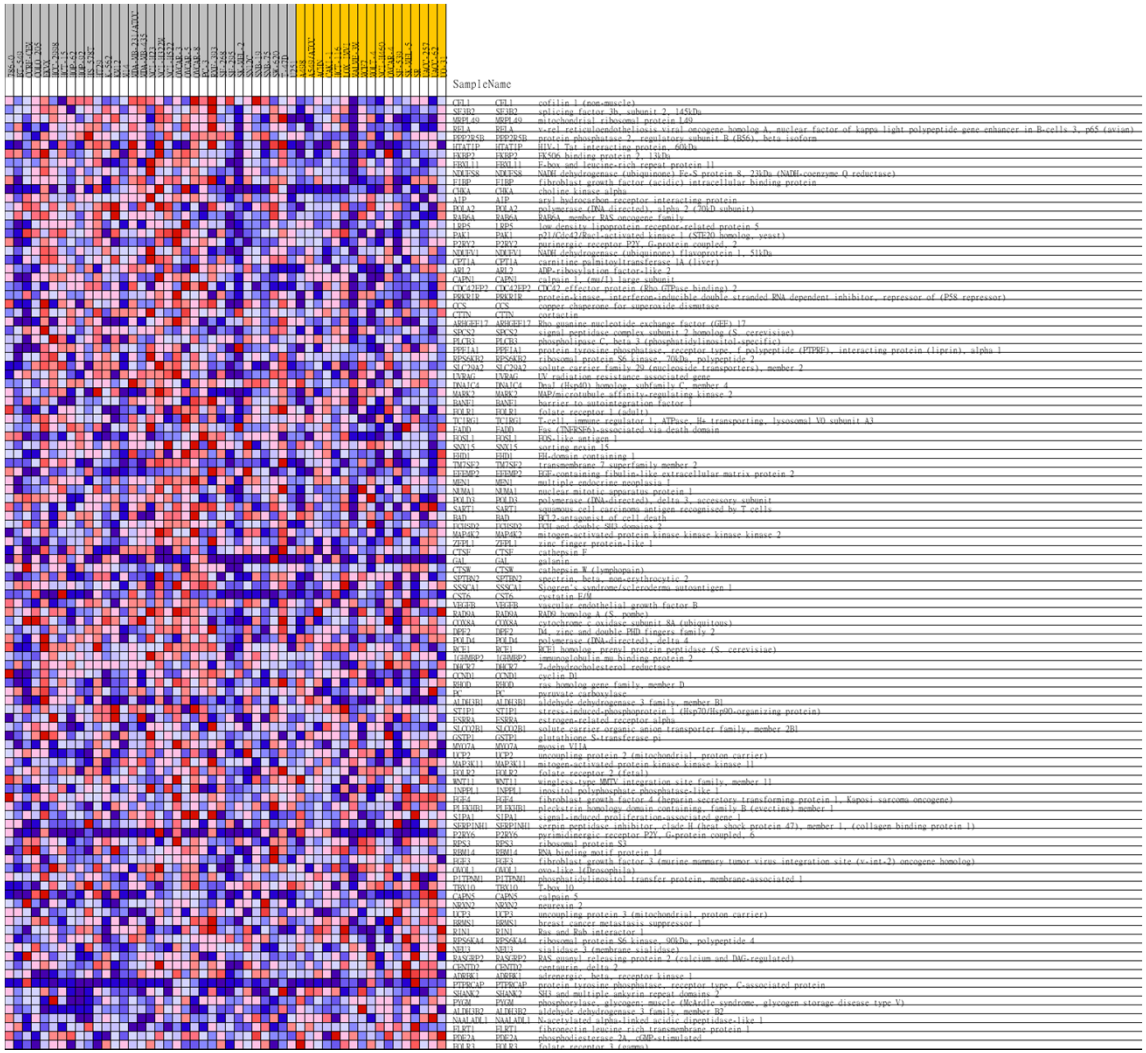


Hits

Table: GSEA details ***[plain text format]***

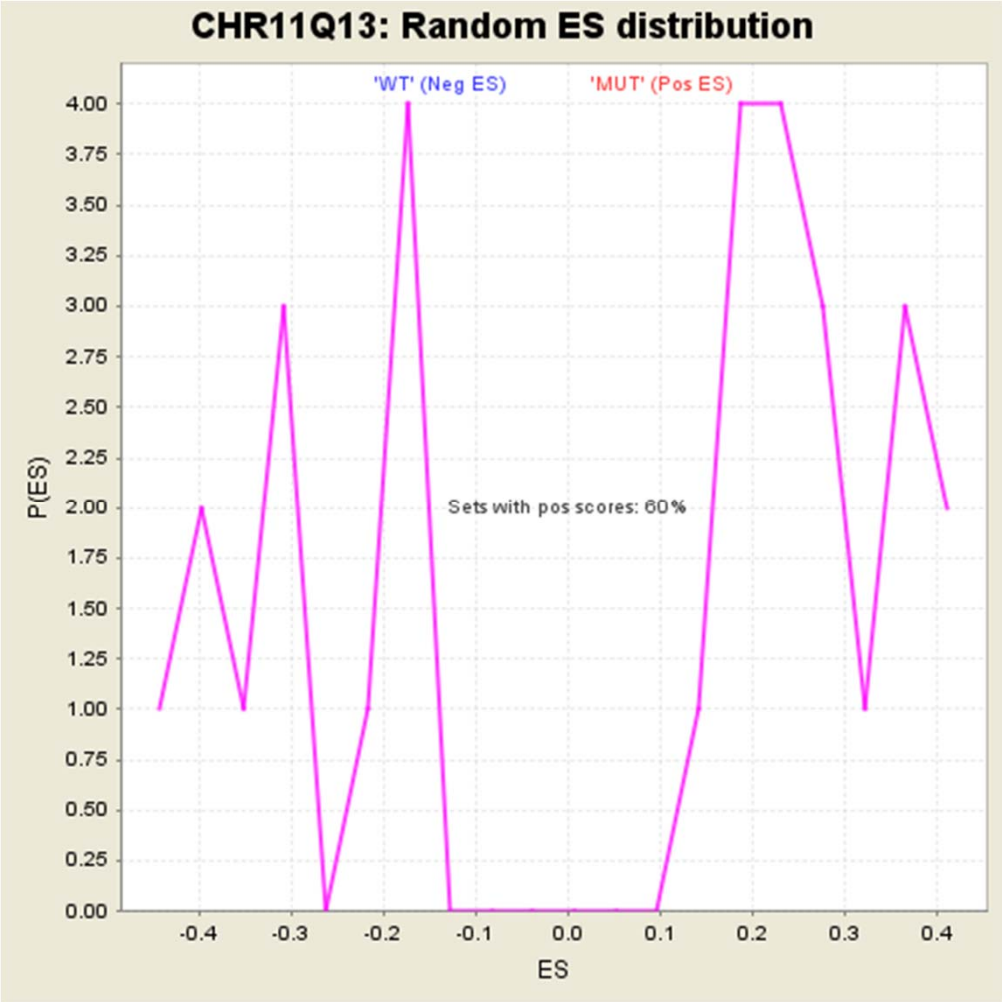
	PROBE	GENE SYMBOL	GENE_TITLE	RANK IN GENE LIST	RANK METRIC SCORE	RUNNING ES	CORE ENRICHMENT
1	CFL1	CFL1 Entrez, Source	cofilin 1 (non-muscle)	22	0.429	0.0258	Yes
2	SF3B2	SF3B2 Entrez, Source	splicing factor 3b, subunit 2, 145kDa	34	0.408	0.0515	Yes
3	MRPL49	MRPL49 Entrez, Source	mitochondrial ribosomal protein L49	42	0.390	0.0765	Yes
4	RELA	RELA Entrez, Source	v-rel reticuloendotheliosis viral oncogene homolog polypeptide gene enhancer in B-cells 3, p65 (avia	48	0.384	0.1012	Yes
5	PPP2R5B	PPP2R5B Entrez, Source	protein phosphatase 2, regulatory subunit B (B56'	65	0.372	0.1239	Yes
6	HTATIP	HTATIP Entrez, Source	HIV-1 Tat interacting protein, 60kDa	91	0.356	0.1446	Yes
105	NAALADL1	NAALADL1 Entrez, Source	N-acetylated alpha-linked acidic dipeptidase-like	8011	-0.221	0.0637	No
106	FLRT1	FLRT1 Entrez, Source	fibronectin leucine rich transmembrane protein 1	8306	-0.246	0.0472	No
107	PDE2A	PDE2A Entrez, Source	phosphodiesterase 2A, cGMP-stimulated	8419	-0.258	0.0518	No
108	FOLR3	FOLR3 Entrez, Source	folate receptor 3 (gamma)	8924	-0.354	0.0190	No

Heat Map for Hits



CHR11Q13 : Blue-Pink O' Gram in the Space of the Analyzed GeneSet

Gene Set Null Distribution of ES



CHR11Q13: Random ES distribution.
Gene set null distribution of ES for CHR11Q13

Detailed Enrichment Results

Enrichment in phenotype: MUT (33 samples)

- 71 / 176 gene sets are upregulated in phenotype **MUT**
- 0 gene sets are significant at FDR < 25%
- 4 gene sets are significantly enriched at nominal pvalue < 1%
- 4 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

Table: Gene sets enriched in phenotype MUT (33 samples) ***[plain text format]***

	GS follow link to MSigDB	GS DETAILS	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX	LEADING EDGE
1	CHR11Q13	Details ...	108	0.46	1.69	0.000	1.000	0.667	2479	tags=53%, list=27%, signal=72%
2	CHR7P21	Details ...	16	0.64	1.66	0.000	0.717	0.800	979	tags=50%, list=11%, signal=56%
3	CHRXP11	Details ...	66	0.53	1.62	0.182	0.664	0.833	1909	tags=55%, list=21%, signal=69%
4	CHR5Q14	Details ...	20	0.55	1.62	0.077	0.525	0.833	535	tags=30%, list=6%, signal=32%
5	CHR10Q22	Details ...	33	0.53	1.57	0.000	0.602	0.933	1649	tags=55%, list=18%, signal=66%
6	CHR1P22	Details ...	22	0.61	1.46	0.000	0.996	0.967	2510	tags=77%, list=28%, signal=106%
7	CHR1P36	Details ...	117	0.41	1.42	0.059	1.000	1.000	1852	tags=44%, list=20%, signal=54%
67	CHR3P14		18	0.21	0.62	0.867	0.978	1.000	2390	tags=39%, list=26%, signal=53%
68	CHR6P21		138	0.19	0.56	0.867	0.999	1.000	1718	tags=25%, list=19%, signal=30%
69	CHR4Q31		24	0.18	0.54	1.000	0.994	1.000	2516	tags=38%, list=28%, signal=52%
70	CHRXQ22		21	0.21	0.52	1.000	0.988	1.000	3222	tags=48%, list=35%, signal=74%
71	CHR9Q22		32	0.15	0.48	1.000	0.988	1.000	1821	tags=22%, list=20%, signal=27%

Gene Markers for the MUT versus WT Comparison

Gene set details

- Gene set size filters (min=15, max=500) resulted in filtering out 143 / 319 gene sets
- The remaining 176 gene sets were used in the analysis
- List of gene sets used and their sizes (restricted to features in the specified dataset)

Gene markers for the MUT *versus* WT comparison

- The dataset has 9096 features (genes)
- # of markers for phenotype **MUT**: 4076 (44.8%) with correlation area 42.2%
- # of markers for phenotype **WT**: 5020 (55.2%) with correlation area 57.8%
- Detailed rank ordered gene list for all features in the dataset
- Heat map and gene list correlation profile for all features in the dataset
- Buttefly plot of significant genes

Global statistics and plots

- Plot of p-values *vs.* NES
- Global ES histogram

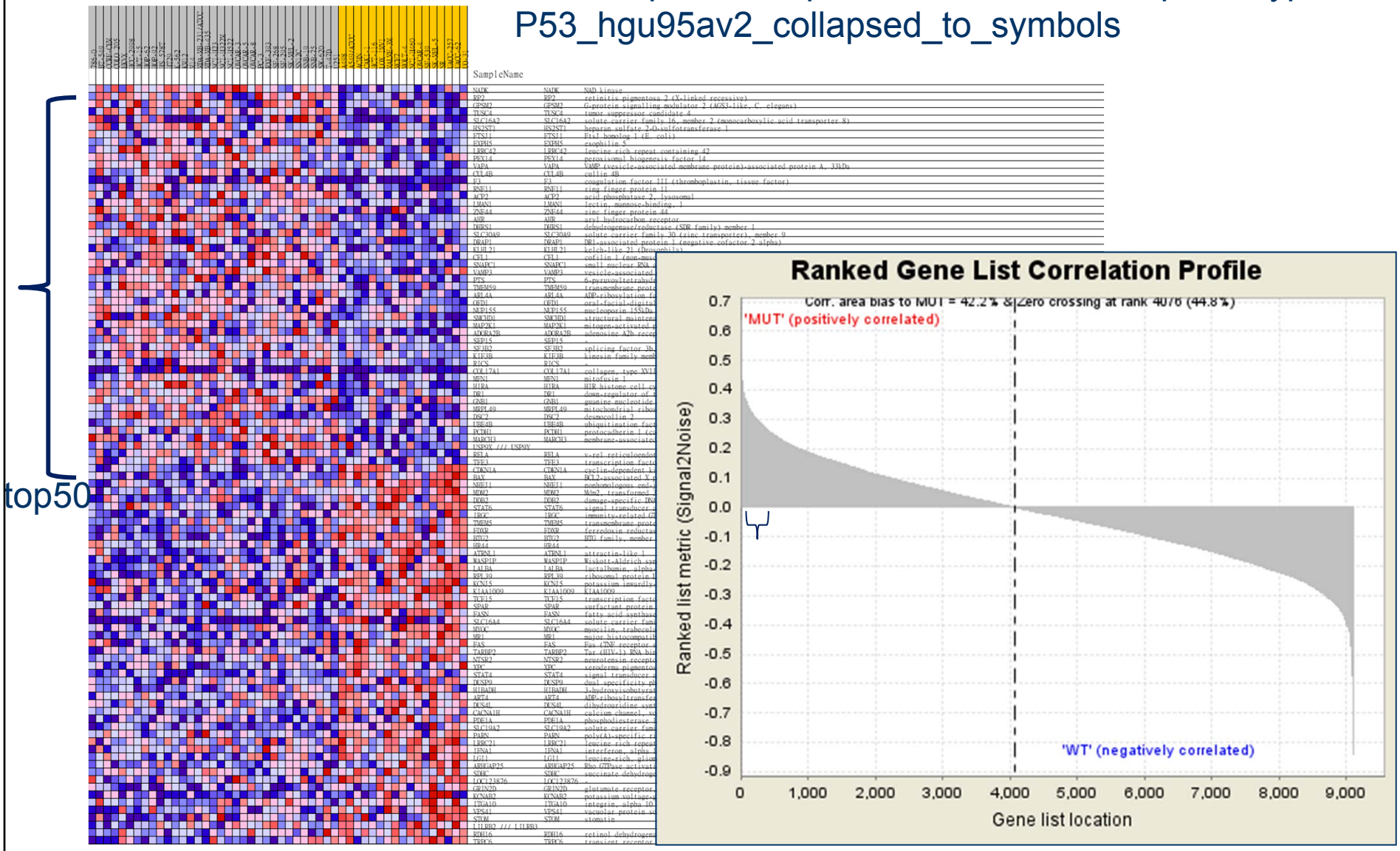
Other

- Parameters used for this analysis

	A	B	C	D	E
1	NAME	DESCRIPTION	GENE_SYMBOL	GENE_TITLE	SCORE
2	NADK	null	NADK	NAD kinase	0.63814014
3	RP2	null	RP2	retinitis pigmentos	0.55928165
4	GPSM2	null	GPSM2	G-protein signallin	0.5350833
5	TUSC4	null	TUSC4	tumor suppressor	0.5116475
6	SLC16A2	null	SLC16A2	solute carrier fami	0.48800114
7	HS2ST1	null	HS2ST1	heparan sulfate 2-	0.4871485
8	FTSJ1	null	FTSJ1	FtsJ homolog 1 (E	0.47524673
9	EXPH5	null	EXPH5	exophilin 5	0.46191633
10	LRRC42	null	LRRC42	leucine rich repeat	0.45818612
11	PEX14	null	PEX14	peroxisomal bioge	0.4568304
12	VAPA	null	VAPA	VAMP (vesicle-as	0.4549476
⋮					
9093	DDB2	null	DDB2	damage-specific D	-0.59452385
9094	MDM2	null	MDM2	Mdm2, transforme	-0.63063174
9095	NHEJ1	null	NHEJ1	nonhomologous e	-0.69846314
9096	BAX	null	BAX	BCL2-associated	-0.78497803
9097	CDKN1A	null	CDKN1A	cyclin-dependent	-0.84255075

Heat Map and Gene Correlation

Heat Map of the top 50 features for each phenotype in P53_hgu95av2_collapsed_to_symbols

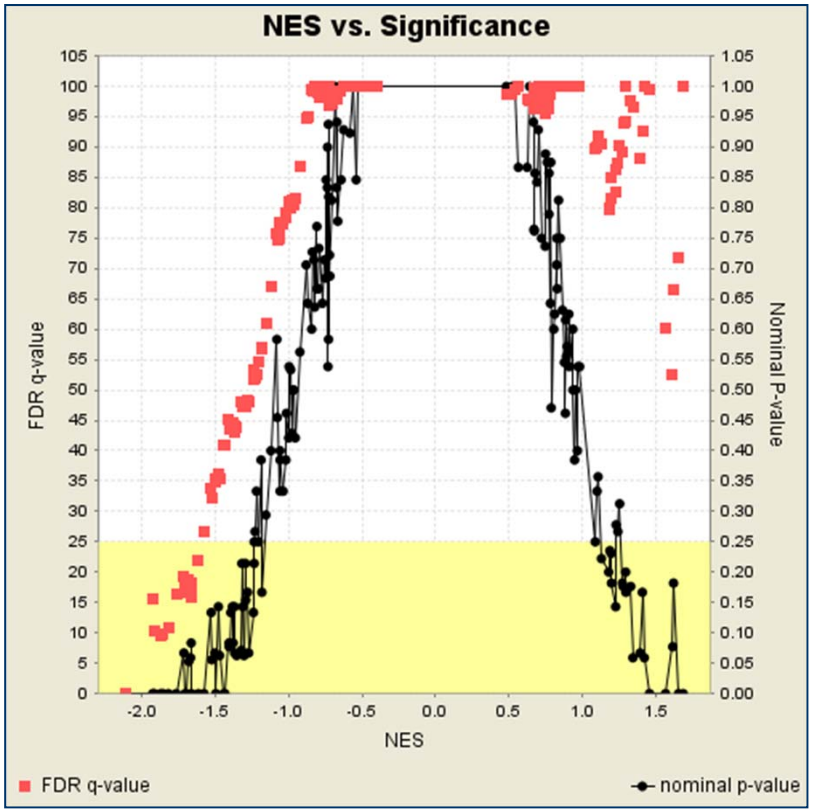


Global Statistics and Plots

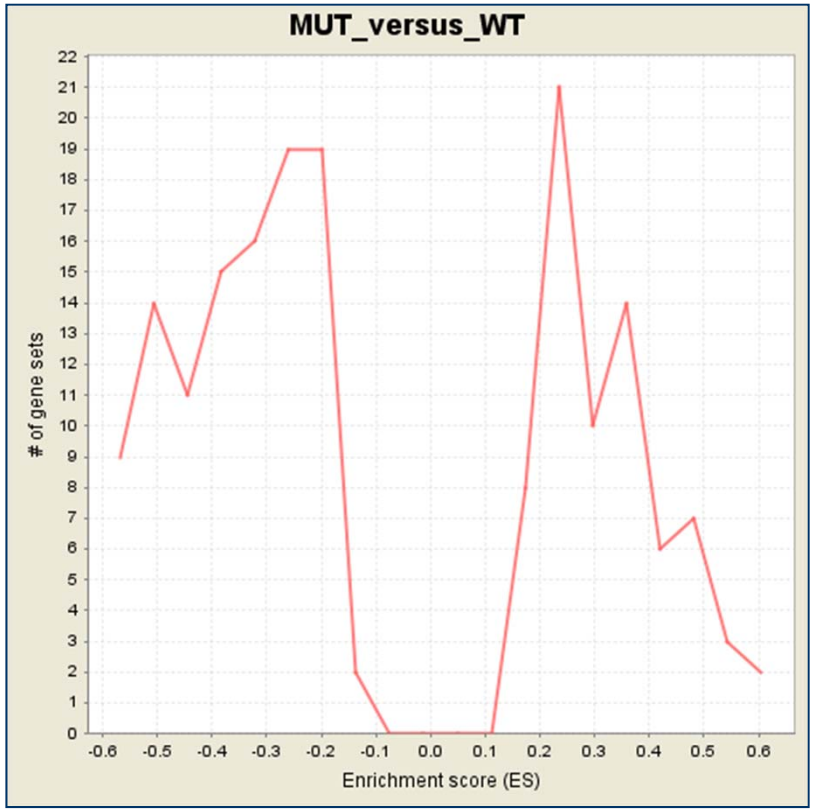
Global statistics and plots

- Plot of p-values *vs.* NES
- Global ES histogram

Plot of p-values vs. NES

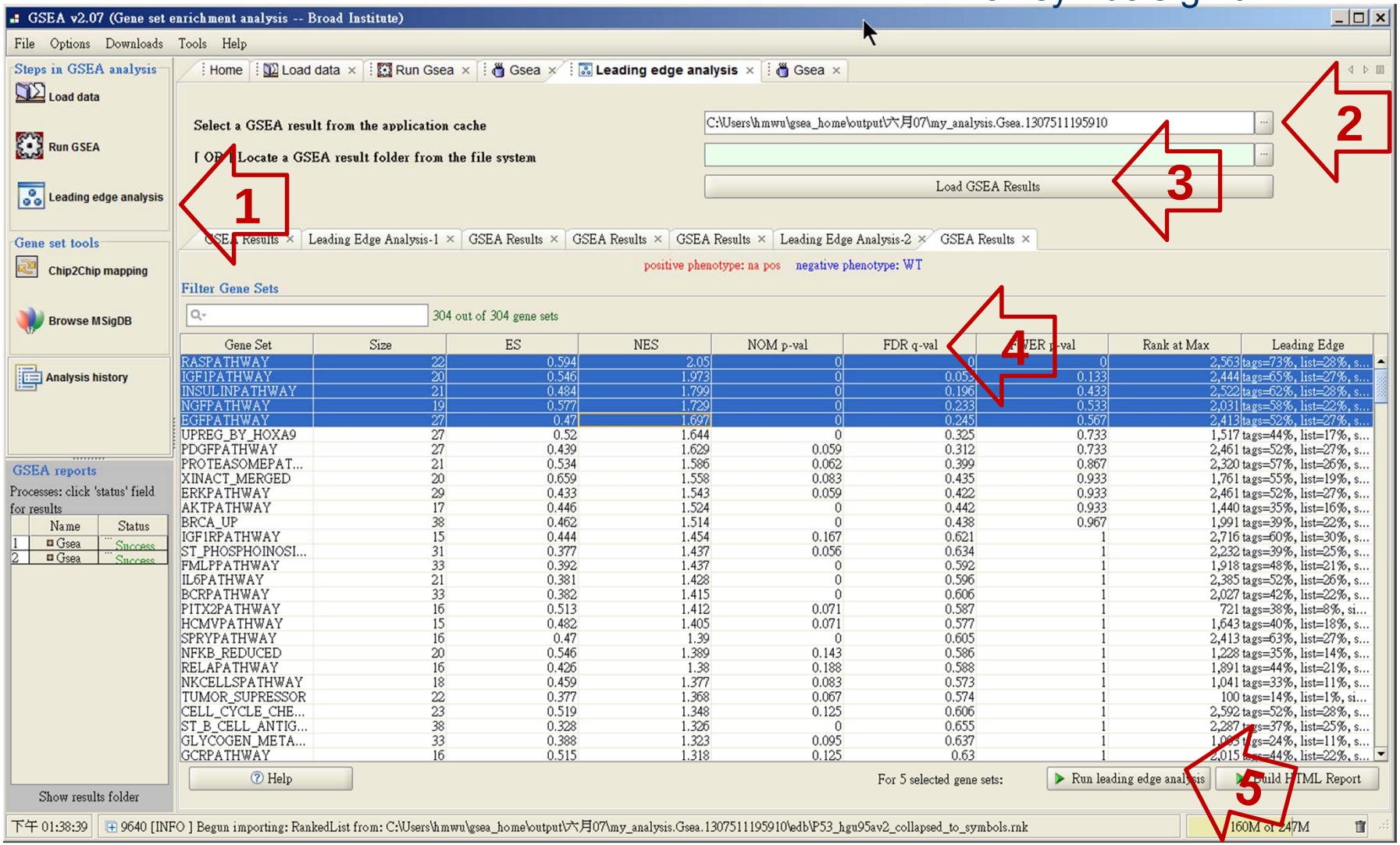


Global ES histogram

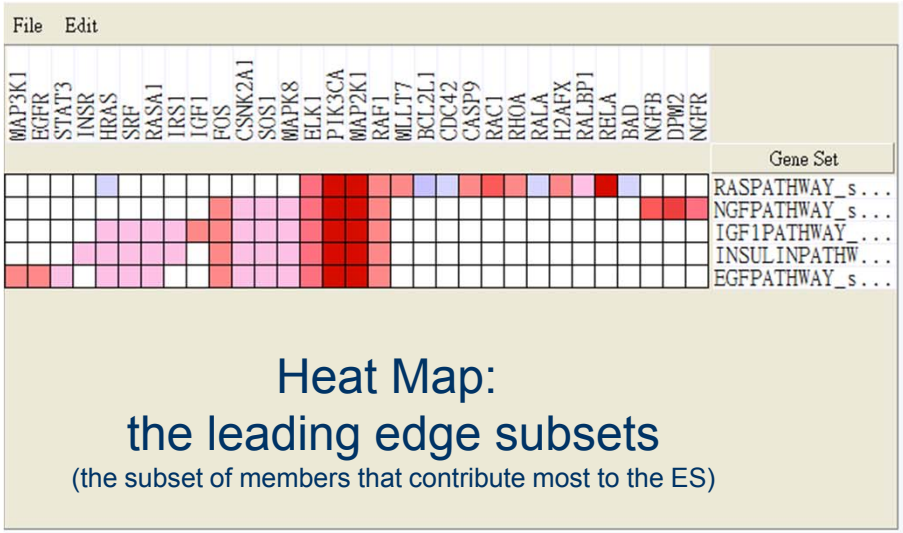


Running the Leading Edge Analysis

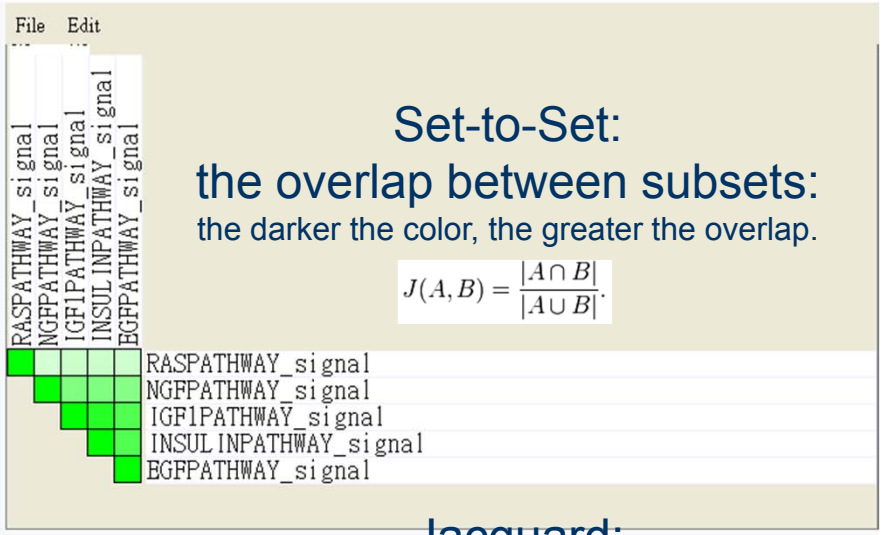
c2.symbols.gmt



Results

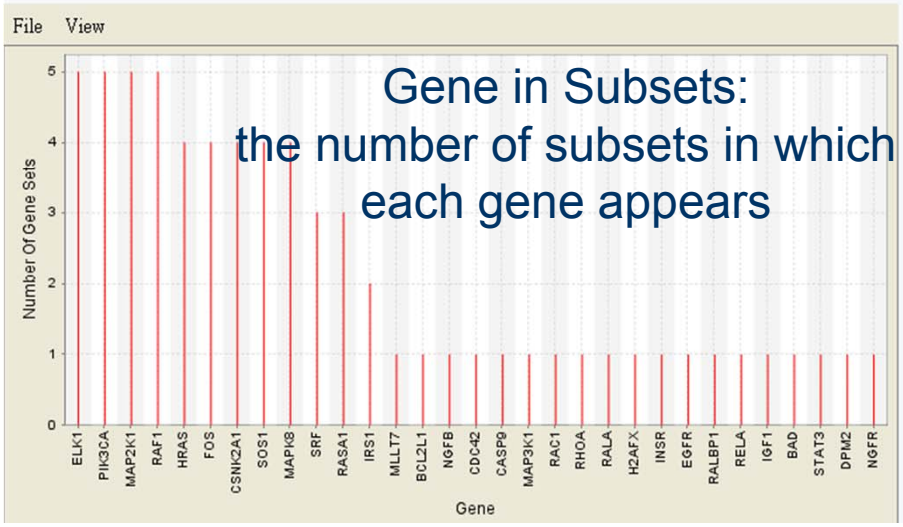


Heat Map:
the leading edge subsets
(the subset of members that contribute most to the ES)

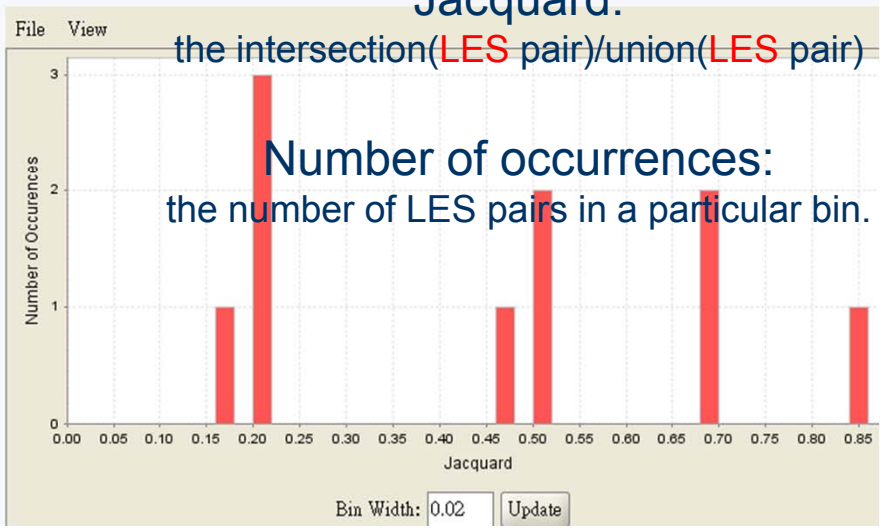


Set-to-Set:
the overlap between subsets:
the darker the color, the greater the overlap.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}.$$



Gene in Subsets:
the number of subsets in which each gene appears



Jacquard:
the intersection(LES pair)/union(LES pair)

Number of occurrences:
the number of LES pairs in a particular bin.

Gene Set Analysis:

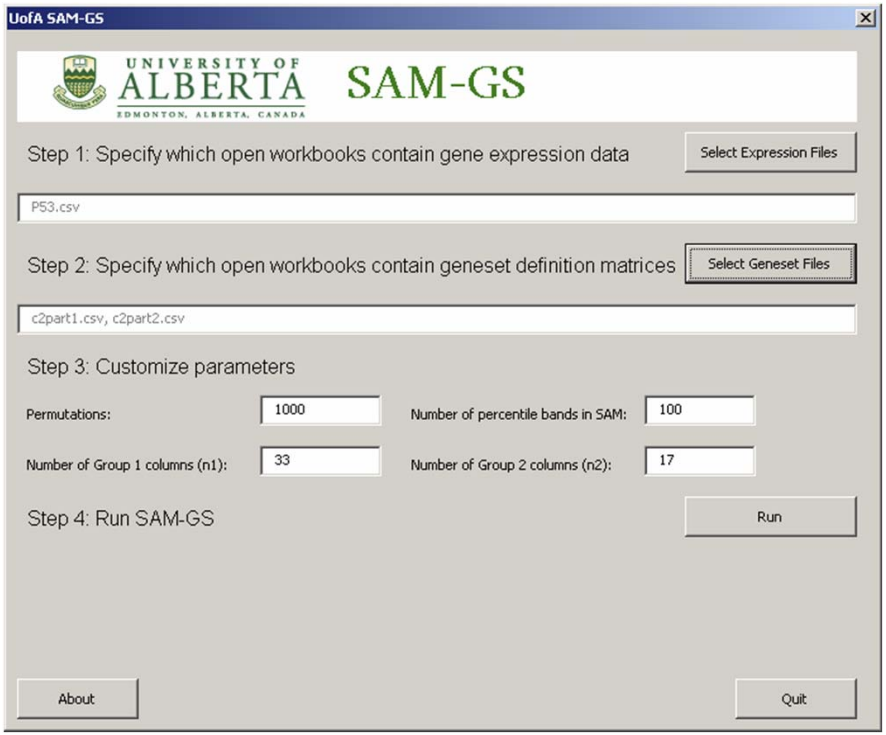
SAM-GS, Maxmean

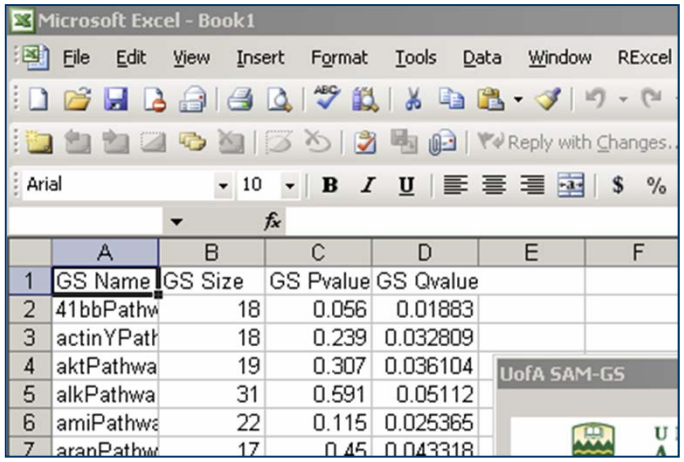
SAM-GS

- SAM-GS:Significance Analysis of Microarray for Gene Sets (Liu et al, 2007)

$$\text{SAM-GS} = \sum_{g_j \in S} d_j^2.$$

- Significance is assessed by permuting sample labels.





- Liu Q, Dinu I, Adewale A, Potter J, Yasui Y: Comparative evaluation of gene-set analysis methods. BMC Bioinformatics 2007, 8:431.
- Qi Liu Irina Dinu Yutaka Yasui, SAM-GS Excel Add-in, version 1.0.2, May 23, 2007

The Maxmean Statistic and Restandardization

The Annals of Applied Statistics
2007, Vol. 1, No. 1, 107–129
DOI: 10.1214/07-AOAS101

ON TESTING THE SIGNIFICANCE OF SETS OF GENES

BY BRADLEY EFRON[1] AND ROBERT TIBSHIRANI[2]

Stanford University

4. Computational issues and software. The developments in the previous two sections lead to our *Gene Set Analysis procedure*, which we summarize here:

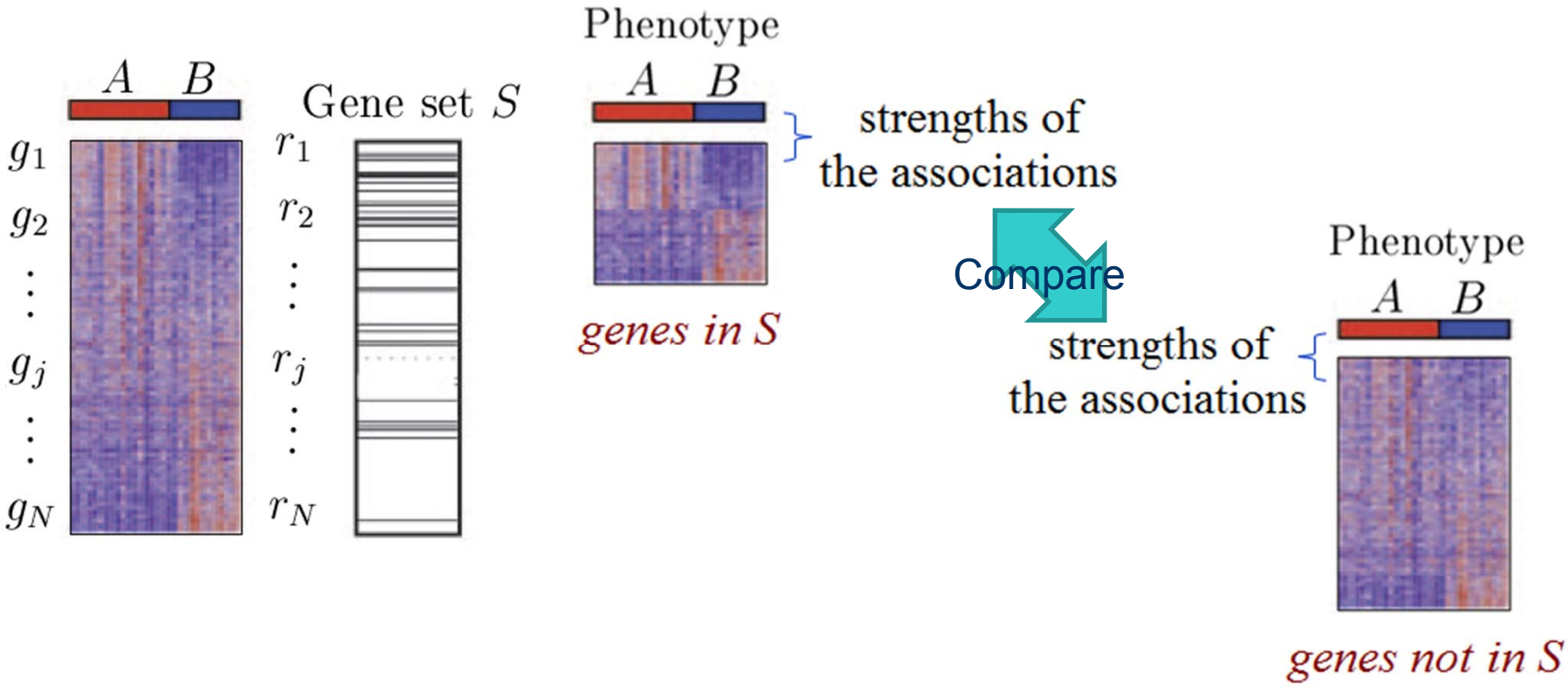
1. Compute a summary statistic z_i for each gene, for example, the two-sample t-statistic for two-class data. Let $\mathbf{z}_{\mathcal{S}}$ be the vector of z_i values for genes in a gene-set $\mathcal{S}$.
2. For each gene-set $\mathcal{S}$, choose a summary statistic $S = s(\mathbf{z})$: choices include the average of z_i or $|z_i|$ for genes in $\mathcal{S}$, the GSEA statistic or, our recommended choice, the *maxmean* statistic (3.11).
3. Standardize S by its randomization mean and standard deviation as in (2.14): $S' = (S - \text{mean}_s)/\text{stdev}_s$. For summary statistics such as the mean, mean absolute value or maxmean (but not GSEA), this can be computed from the genewise means and standard deviations, without having to draw random sets of genes. Note formula (3.13) for the maxmean statistic.
4. Compute permutations of the outcome values (e.g., the class labels in the two-class case) and recompute S' on each permuted dataset, yielding permutation values $S'^{*1}, S'^{*2}, \ldots, S'^{*B}$.

Null Hypothesis, Statistical Significance Of Gene Set Scores (P-values)

Null Hypothesis of GSA: Q1

Q1: competitive null hypothesis

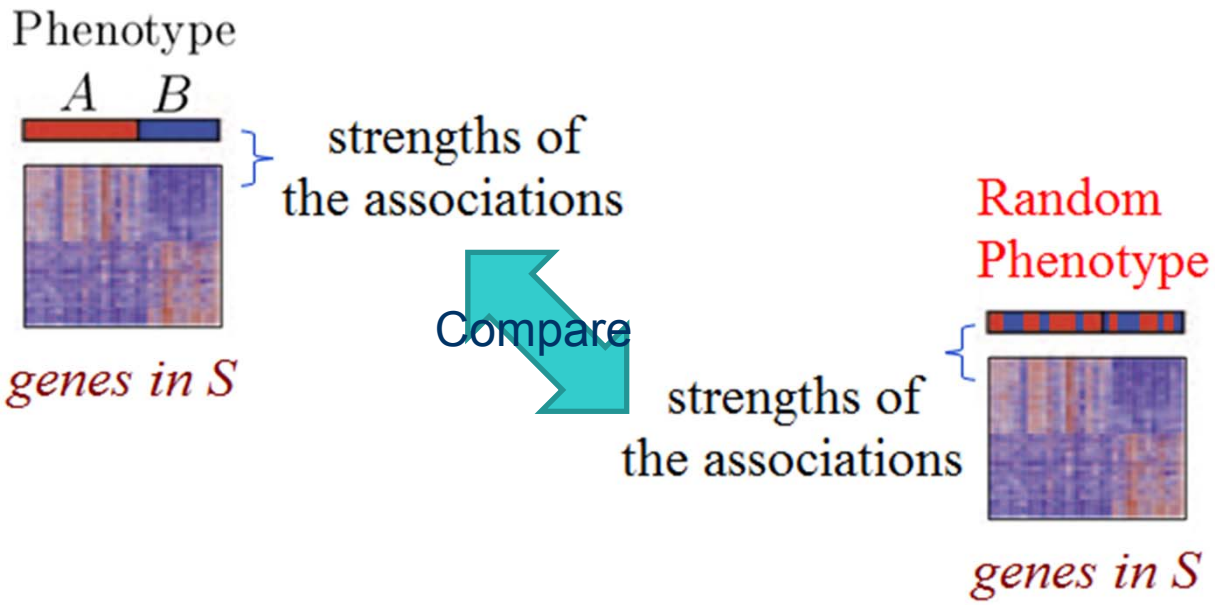
- Compares the association between the genes in the set and the phenotype is compared with the association of the remainder of the genes and the phenotype.
- Genes are the sampling units and the association between the samples and the phenotypes is fixed.



Null Hypothesis of GSA: Q2

Q2: self-contained null hypothesis

- Compares the association of the gene set and the phenotype with that of random phenotypes.
- Phenotypes (Samples) are the sampling units while the gene set membership is fixed.



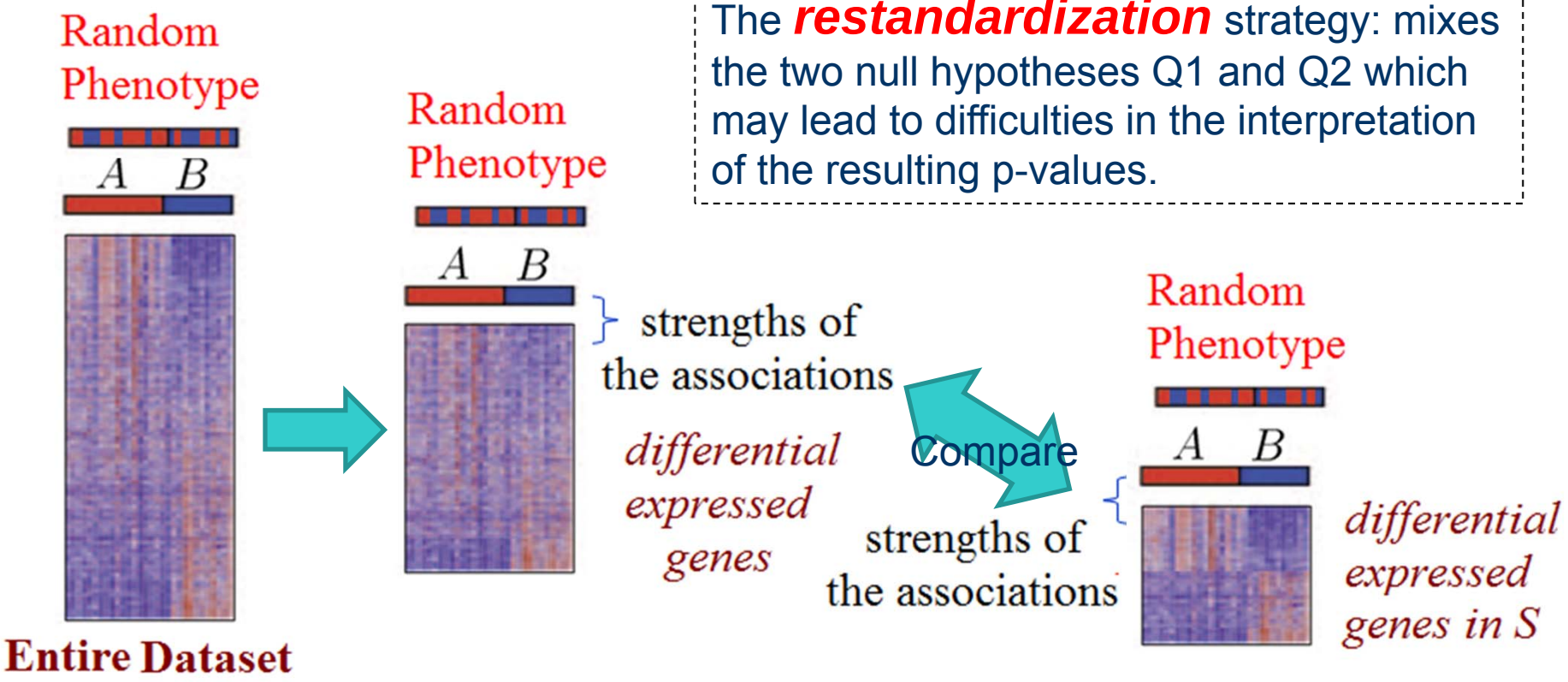
Tian et al (PNAS, 2005) defined Q1, Q2
Geoman and Buhlmann (Bioinformatics, 2007) defined competitive and self-contained.

Null Hypothesis of GSA: Q3

Q3: the "nested null hypothesis"

- Where differential expression of the genes in the gene set is compared to the differential expression of all genes under consideration (both inside and outside the gene set).
- H0: none of the gene sets considered is associated with the phenotype.

The ***restandardization*** strategy: mixes the two null hypotheses Q1 and Q2 which may lead to difficulties in the interpretation of the resulting p-values.



Statistical Significance Of Gene Set Scores (P-values)

- One suggests that both gene and sample randomizations should be used because they test two different but complementary null hypotheses.
- The other insists that only sample randomization should be used to avoid inherent problems of the gene randomization method.
- The interpretation of a P-value greatly depends on the sampling scheme on which the test is based and is related to hypothetical replications of the experiment performed.

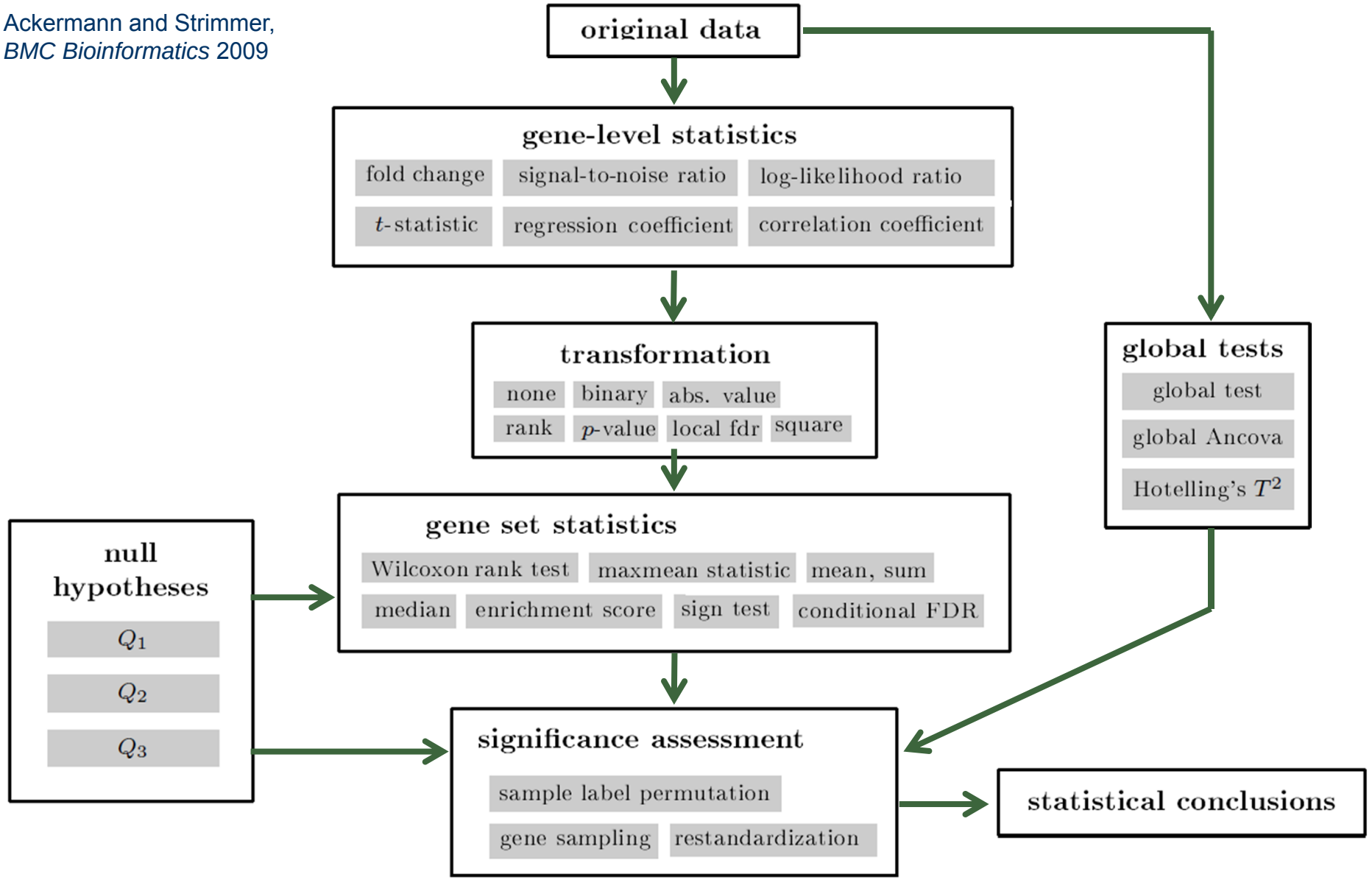
GSEA

- Subramanian et al. (2005) included the gene randomization option in their GSEA program and suggested using gene randomization to generate hypothesis when the number of samples is small.
- GSEA utilizes a competitive statistic (Kolmogorov–Smirnov statistic) as a 'score function', if not as a test statistic, to represent the relative enrichment of DEGs in each gene set.
- GSEA tests the significance of the entire dataset by applying sample permutation to the scores.
- GSEA is considered a competitive method relative to individual gene sets, but is considered a self-contained method relative to the entire dataset (set of gene sets).

Gene Set Analysis Tools

A General Modular Framework for Gene Set Enrichment Analysis

Ackermann and Strimmer,
BMC Bioinformatics 2009



Ackermann and Strimmer, *BMC Bioinformatics* 2009

- **Gene-level statistic:** in general, the choice of the gene-level statistic *does not* seem to have a great impact on the results of the enrichment analysis.

- **Transformation:** The choice of a transformation has quite a *substantial effect* on the detection rates. The rank squared transformation was most accurate.

- **Gene set statistics**: Overall, the mean or the median work very well. (less sensitive with regard to outliers)
  - The median and the rank-based Wilcoxon test may lead to a smaller number of significant findings.
  - The GSEA (the enrichment score) is not as reliable as the other gene set test statistics.

- **Significance assessment**: Depending on the data structure, both approaches can yield quite different results.

Gene Set Analysis Methods

Table I: Cutoff-free gene set analysis methods

Dougu Nam and Seon-Young Kim, Gene-set approach for expression pattern analysis, *Briefings In Bioinformatics*. 2008, VOL 9. NO 3. 189-197.

Authors	Year	Name	Statistical test	Self-contained versus competitive	Gene versus ample randomization	Reference
Virtaneva *et al.*	2001		sample randomization	self-contained	sample	[8]
Pavlidis *et al.*	2002		gene randomization	competitive	gene	[9]
Mootha *et al.*	2003	GSEA	sample randomization	mixed	sample	[7]
Breslin *et al.*	2004	Catmap	gene randomization	competitive	gene	[3]
Goeman *et al.*	2004	globaltest	sample randomization	self-contained	sample	[17]
Smid *et al.*	2004	GO-Mapper	z-test	competitive	gene	[38]
Volinia *et al.*	2004	GOAL	gene randomization	competitive	gene	[39]
Barry *et al.*	2005	SAFE	sample randomization	competitive	sample	[19]
Beh-Shaul *et al.*	2005		Kolmogorov–Smirnov test	competitive	gene	[5]
Boorsma e al.	2005	T-profiler	*t*-test	competitive	gene	[15]
Kim *et al.*	2005	PAGE	z-test	competitive	gene	[14]
Lee *et al.*	2005	ErmineJ	sample randomization	competitive	gene	[16]
Subramanian *et al.*	2005	GSEA	sample randomization	mixed	gene	[25]
Tian *et al.*	2005	Q1, Q2	gene or sample randomization	competitive or self-contained	gene or sample	[10]
Tomfohr *et al.*	2005	PLAGE	sample randomization	self-contained	sample	[20]
Edelman *et al.*	2006	ASSESS	sample randomization	competitive	sample	[28]
Kong *et al.*	2006		Hotelling's T squared	self-contained	sample	[21]
Nam *et al.*	2006	ADGO	z-test	competitive	gene	[29]
Saxena *et al.*	2006	AE	sample randomization	competitive	sample	[31]
Scheer *et al.*	2006	JProGO	Fisher's exact test, Kolmogorov–Smirnov test, *t*-test, unpaired Wilcoxon's test	competitive	gene	[40]
Al-Shahrour *et al.*	2007	Fatiscan	Fisher's exact test, hypergeometric test	competitive	gene	[41]
Backes *et al.*	2007	GeneTrail	Fisher's exact test, hypergeometric test, sample randomization	competitive	gene or sample	[42]
Cavalieri *et al.*	2007	Eu.Gene Analyzer	Fisher's exact test, sample randomization	competitive	gene or sample	[43]
Dinu *et al.*	2007	SAM-GS	sample randomization	self-contained	sample	[22]
Efron *et al.*	2007	GSA	sample randomization	mixed	sample	[26]
Newton *et al.*	2007	Random set	z-test	competitive	gene	[44]

Gene Set Analysis Tools (1)

Table 2: Gene set analysis tools

Dougu Nam and Seon-Young Kim, Gene-set approach for expression pattern analysis, *Briefings In Bioinformatics*. 2008, VOL 9. NO 3. 189-197.

Name	Organism[a]	Application Type	URL	Reference
ADGO	H, M, R, Y	Web server	http://array.kobic.re.kr/ADGO	[29]
ASSESS	H, M, R	Octave/Java standalone	http://people.genome.duke.edu/~jhg9/assess/	[28]
Babelomics	H, M, R, DM, S, C	Web server	http://www.babelomics.org	[45]
Catmap	H	Perl script	http://bioinfo.thep.lu.se/catmap.html	[3]
ErmineJ	H, M, R	Java standalone	http://www.bioinformatics.ubc.ca/ermineJ/	[16]
Eu.Gene Analyzer	H, M, R, Y	Windows/Unix standalone	http://www.ducciocavalieri.org/bio/Eugene.htm	[43]
FatiScan	H, M, R, Y, B, D, G, C, A, S, DM	Web server	http://fatiscan.bioinfo.cipf.es/	[41]
GAZER	H, M, R, Y	Web server	http://integromics.kobic.re.kr/GAzer/index.faces;	[13]
GeneTrail	H, M, R, Y, SA, CG, AT	Web server	http://genetrail.bioinf.uni-sb.de/	[42]
Global test	NA	R package	http://bioconductor.org/packages/2.0/bioc/html/globaltest.html	[17]
GOAL	H, M	Web server	http://microarrays.unife.it	[39]
GO-Mapper	H, M, R, Z, DM, Y	Windows standalone, Perl script	http://www.gatcplatform.nl/	[38]
GSA	H	R package	http://www-stat.stanford.edu/~tibs/GSA/	[26]
GSEA	H	Java standalone, R package	http://www.broad.mit.edu/gsea/	[25]
JProGO	Various prokaryotes	Web server	http://www.jprogo.de/	[40]
MEGO	H	Windows standalone	http://www.dxy.cn/mego/	[46]
PAGE	H, M, R, Y	Python script	From the author (kimsy@kribb.re.kr)	[14]
PLAGE	H, M	Web server	http://dulci.biostat.duke.edu/pathways/	[20]
SAFE	NA	R package	http://bioconductor.org/packages/2.0/bioc/html/safe.html	[19]
SAM-GS	NA	Windows Excel Add-In	http://www.ualberta.ca/~yyasui/homepage.html	[22]
T-profiler	Y, CA	Web server	http://www.t-profiler.org/	[15]

[a]H: *Homo sapiens*; M: *Mus musculus*; R: *Rattus norvegicus*; Y: *Sacchaomyces cerevisiae*; B: *Bos Taurus*; D: *Daniel rerio*; G: *Gallus gallus*; C: *Caenorhabditis elegans*; A: *Arabidopsis thaliana*; DM: *Drosophila melanogaster*; Z: *Zebra fish*; CA: *Candida albicans*; SA: *Staphylococcus aureus*; CG: *Corynebacterium glutamicum*; AT: *Arabidopsis thaliana*.

Gene Set Analysis Tools (2)

Table 2 Tools available for functional profiling by gene-set analysis

Tool	Application type or URL for web servers	References	Test[a]	Citations[b]
GSEA	http://www.broad.mit.edu/gsea/	[21,61]	GS, C	1013
Babelomics (FatiGO + FatiScan)	http://www.babelomics.org	[26,27,37,40,66]	FE/GS, C	402
FuncAssociate	http://llama.med.harvard.edu/Software.html	[87]	FE/GS, C	91
Global test	R package	[64]	GS, SC	89
PAGE	Python script	[65]	GS, C	42
ErmineJ	Java	[105]	GS, C	35
FatiScan	http://www.babelomics.org	[66]	GS, C	34
GO-mapper	Windows, Perl script	[63]	GS, C	33
SAFE	R package	[62]	GS, C	27
GOAL	http://microarrays.unife.it	[106]	GS, C	25
Catmap	Perl script.	[107]	GS, C	19
PLAGE	http://dulci.biostat.duke.edu/pathways/	[108]	GS, SC	18
GODist	Mathlab program	[109]	GS, SC	17
t-profiler	http://www.t-profiler.org/	[110]	GS, C	12
JProGO	http://www.jprogo.de/	[111]	GS, C	7
ADGO	http://array.kobic.re.kr/ADGO	[112]	GS, C	3
GeneTrail	http://genetrail.bioinf.uni-sb.de/	[113]	GS, C	3
ASSESS	Java	[114]	GS, C	2
DEA	R package	[115]	GS, C	1
GlobalANCOVA	R package	[67]	GS, SC	1
GAZER	http://integromics.kobic.re.kr/GAzer/index.faces	[116]	GS, C	–
SAM-GS	Windows excel add-in	[117]	GS, SC	–

[a] Type of test: GS: gene set; C: competitive, SC: self-contained; FE: functional enrichment.
[b] Citations are taken from Scholar Google (as of January 2008). Scholar Google is taken as an indirect estimation of the citation in papers but gives an idea on the impact in the scientific community.

Dopazo J. Formulating and testing hypotheses in functional genomics. Artif Intell Med. 2009 Feb-Mar;45(2-3):97-107.

Gene Set Databases

- **The limitations of the information on current annotation databases for IGA:** **(1)** incomplete knowledge, **(2)** time-delayed curation, **(3)** imprecise or incorrect electronic annotations, **(4)** inability to predict new functions and **(5)** semantic misclassification of annotations.
- The problems are all shared by GSA except for finding more relevant categories among overlapping gene sets for which GSA is able to assign different scores.
- Using overlapping gene sets in GSA substantially improves the analysis.

Table 3: Gene set databases

Dougu Nam and Seon-Young Kim, Gene-set approach for expression pattern analysis, *Briefings In Bioinformatics*. 2008, VOL 9. NO 3. 189-197.

Name	Organism[a]	Gene sets	Web address	Reference
ASSESS	H	Cytogenetic, pathway, motif	http://people.genome.duke.edu/~jhg9/assess/genesets.shtml	[28]
ErmineJ	H, M, R	GO	http://www.bioinformatics.ubc.ca/microannots/	[16]
GAzer	H, M, R, Y	GO, composite GO, InterPro, Pathways, TFBS	http://integromics.kobic.re.kr/GAzer/document.jsp	[13]
GSA	H	Tissue, cellular processes, cytobands, chromosome arms, 5Mb chromosomal tiles, cancer module	http://www-stat.stanford.edu/~tibs/GSA/	[26]
MSigDb	H	Cytobands, curated pathways, motif, computed	http://www.broad.mit.edu/gsea/msigdb/msigdb.index.html	[25]
PLAGE	H, M	KEGG and BioCarta pathways	http://dulci.biostat.duke.edu/pathways/misc.html	[20]

[a]H: *Homo sapiens*; M: *Mus musculus*; R: *Rattus norvegicus*; Y: *Saccharomyces cerevisiae*.

Methods Comparison

Goeman and Buhlmann, *Bioinformatics*, 2007

- Goeman and Buhlmann (2007) has investigated methodological issues in methods that test for differential expression of gene sets.

- It has revealed some methodological aspects of popular methods that are inefficient or even incorrect from a statistical point of view.

Jelle J. Goeman and Peter Buhlmann, Analyzing gene expression data in terms of gene sets: methodological issues, Bioinformatics (2007) 23 (8): 980-987.

Liu et al., *BMC Bioinformatics*, 2007

- **Compare:** the three "self-contained" gene set methods: (1) Global Test, (2) ANCOVA Global Test and (3) SAM-GS.
- After the standardization of gene expression, the three methods gave very similar results, with slightly higher statistical significance given by SAM-GS.
- Global Test and ANCOVA Global Test were able to analyze continuous and survival phenotypes and to adjust for covariates.
- The framework of the competitive hypothesis testing via. gene sampling is subject to serious errors in calculating and interpreting statistical significance of gene sets, because of its implicit or explicit untenable assumption of probabilistic independence across genes.

Liu Q, Dinu I, Adewale A, Potter J, Yasui Y: Comparative evaluation of gene-set analysis methods. *BMC Bioinformatics* **2007**, 8:431.

Efron and Tibshirani, *Annals of Applied Statistics*, 2007

- Efron and Tibshirani (2007) introduced five test statistics for a GSEA algorithm: (1) mean, (2) mean.abs, (3) maxmean, (4) GSEA and (5) GSEA.abs.

- They concluded that the maxmean statistic is the only method with consistently low P-values in all situations.

- Some criticisms of GSEA: the enrichment score can be influenced by the size of a gene set and by the presence or absence of lower-ranking sets.

- The appropriate method is chosen depending on either the number of the samples or the property of the DEG sets the user wants to find.

Efron B, Tibshirani R: On testing the significance of sets of genes. *Annals of Applied Statistics* **2007**, 1:107-129.

Song and Black, *BMC Bioinformatics*, 2008

Methods were selected for which software packages were available through the Bioconductor project

Table 1: Summary of measurements used for each gene set method.

Method	Measurement of gene set	Permutation
GSEA-Category	The **sum** of the **T-statistics**	Sample
GSEA-limma	The **mean** of the **T-statistics**	Gene
SAFE	The **sum** of the **ranks** of the **T-statistics**	Sample
Globaltest	The **mean** of the **Q-statistics**	Sample
PCOT2	Hotelling's $\mathbf{T^2}$ (multivariate T-statistic)	Sample
sigPathway	The **mean** of the **T-statistics**	Sample

correlation-based methods

Globaltest ~ PCOT2 > GSEA-Category

poor: SAFE, sigPathway

Sarah Song and Michael A Black. Microarray-based gene set analysis: a comparison of current methods, BMC Bioinformatics 2008, 9:502.

Dinu et al., *Cancer Informatics*, 2008

- **Compare:** the five self-contained gene set methods and the competitive GSEA method:
  (1) SAM-GS, (2) Global Test, (3) Global ANCOVA Test, (4) Tian, (5) Tomfohr, (6) GSEA-P, (7) GSEA-FDR.

- The self-contained methods of SAM-GS, Global test and ANCOVA Global outperformed GSEA.

- General conclusions regarding the relative performance of the investigated methods could not be made, as no simulation studies were completed.

Dinu et al., A Biological Evaluation of Six Gene Set Analysis Methods for Identification of Differentially Expressed Pathways in Microarray Data, *Cancer Informatics* **2008**, 6, 357-368.

Nam and Kim, *Briefings In Bioinformatics*, 2008

The criterion for choosing a statistical method for GSA:

- If the purpose is to find gene sets relatively enriched with DEGs, a competitive method based on Q1 should be used.
- If the purpose is to find gene sets clearly separated between the two sample groups, a self-contained method based on Q2 should be selected.
- Prefers the mixed approach (Q3) to avoid the clear drawbacks of the other methods, but recommends using all the methods simultaneously, if possible, with biological analyses.
- Sample randomization provides statistically sound P-values.

Dougu Nam and Seon-Young Kim, Gene-set approach for expression pattern analysis, *Briefings In Bioinformatics*. 2008, VOL 9. NO 3. 189-197.

Which One To Use?
A Practical Guideline

- Selecting an optimal tool, which of course depends on the type of experimental data. The first thing to consider is the type of organism.

- For a human gene expression dataset with an enough number of samples (more than 10), GSEA is highly recommended because it is a statistically sound method based on sample randomization and provides a user-friendly, standalone program.
  - GSA (Efron and Tibshirani, 2007) and SAFE provide potentially better **statistical properties** than GSEA (R packages).

- For mouse, rat or yeast datasets for which the GSEA program is not available, web servers such as Babelomics, GAzer or GeneTrail are recommended.
  - When the number of samples is small, gene randomization-based tools such as ErmineJ or GAzer are highly recommended.

Dougu Nam and Seon-Young Kim, Gene-set approach for expression pattern analysis, *Briefings In Bioinformatics*. 2008, VOL 9. NO 3. 189-197.

Tsai and Chen, *Bioinformatics*, 2009

- Tsai and Chen (2009) proposed using a MANOVA test for gene-set analysis.
- They compared it to several methods including (1) principal component analysis, (2) SAM-GS, (3) GSEA, (4) Maxmean, (5) ANCOVA, and (6) Global Test.
- They found the MANOVA approach appeared to perform best, but concluded that most methods, except GSEA and maxmean, were generally comparable in terms of power.
- A limitation of the MANOVA method is that it is only applicable to categorical outcomes data.

Tsai CA, Chen JJ (2009) Multivariate analysis of variance test for gene set analysis. Bioinformatics 25: 897-903.

Fridley, Jenkins, and Biernacka, *PLoS ONE*, 2010

- The self-contained have been reported to be more powerful than competitive methods.
- Compare: (1) Kolmogorov–Smirnov test (KS), (2) Fisher's method (FM), (3) Stouffer's method (SM), (4) tail strength (TS), (5) a novel modified tail strength statistic (MTS), (6) global model with fixed effects (GMFE), (7) global model with random effects (GMRE), and (8) principal component analysis (PCA).
- The simulation scenarios varied according to: (1) number of genes in a gene set, (2) number of genes associated with the phenotype, (3) effect sizes, (4) correlation between expression of genes within a gene set, and (5) the sample size.
- Over a variety of scenarios, the **FM** with empirical p-values or the **GMRE** were the most powerful analytical approaches for a self-contained gene set analysis.
- The analysis based on the first principal component and Kolmogorov-Smirnov test tended to have lowest power.

Fridley BL, Jenkins GD, Biernacka JM, Self-Contained Gene-Set Analysis of Expression Data: An Evaluation of Existing and Novel Methods. *PLoS ONE*, **2010**, 5(9): e12693.

Reference

- Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci* USA **2005**, 102:15545-15550.
- Goeman JJ, Bühlmann P: Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics* **2007**, 23:980-987
- Efron B, Tibshirani R: On testing the significance of sets of genes. *Annals of Applied Statistics* **2007**, 1:107-129.
- Jiang Z, Gentleman R: Extensions to gene set enrichment. Bioinformatics **2007**, 23:306-313.
- Liu Q, Dinu I, Adewale A, Potter J, Yasui Y: Comparative evaluation of gene-set analysis methods. *BMC Bioinformatics* **2007**, 8:431.
- Chen JJ, Lee T, Delongchamp RR, Chen T, Tsai CA: Significance analysis of groups of genes in expression profiling studies. *Bioinformatics* **2007**, 23:2104-2112.
- Song S, Black MA: Microarray-based gene set analysis: a comparison of current methods. *BMC Bioinformatics* **2008**, 9:502
- Dougu Nam and Seon-Young Kim, Gene-set approach for expression pattern analysis, *Briefings In Bioinformatics*. **2008**, VOL 9. NO 3. 189-197.
- Dinu et al., A Biological Evaluation of Six Gene Set Analysis Methods for Identification of Differentially Expressed Pathways in Microarray Data, *Cancer Informatics* **2008**, 6, 357-368.
- Dopazo J. Formulating and testing hypotheses in functional genomics. *Artif Intell Med*. **2009** Feb-Mar;45(2-3):97-107.
- Marit Ackermann and Korbinian Strimmer, A general modular framework for gene set enrichment analysis, *BMC Bioinformatics* **2009**, 10:47
- Fridley BL, Jenkins GD, Biernacka JM, Self-Contained Gene-Set Analysis of Expression Data: An Evaluation of Existing and Novel Methods. *PLoS ONE*, **2010**, 5(9): e12693.

Further Reading

Stiglic *et al. BMC Bioinformatics* 2010, **11**:176
http://www.biomedcentral.com/1471-2105/11/176

METHODOLOGY ARTICLE Open Access

Gene set enrichment meta-learning analysis: next-generation sequencing versus microarrays

Gregor Stiglic*, Mateja Bajgot and Peter Kokol

Massa *et al. BMC Systems Biology* 2010, **4**:121
http://www.biomedcentral.com/1752-0509/4/121

METHODOLOGY ARTICLE Open Access

Gene set analysis exploiting the topology of a pathway

Maria Sofia Massa[1], Monica Chiogna[1*], Chiara Romualdi[2]

Han-Ming Wu

http://www.hmwu.idv.tw

Thank You!