Tools for Microarray Data Analysis

吳漢銘 助理教授
淡江大學 數學系

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

2011/06/16

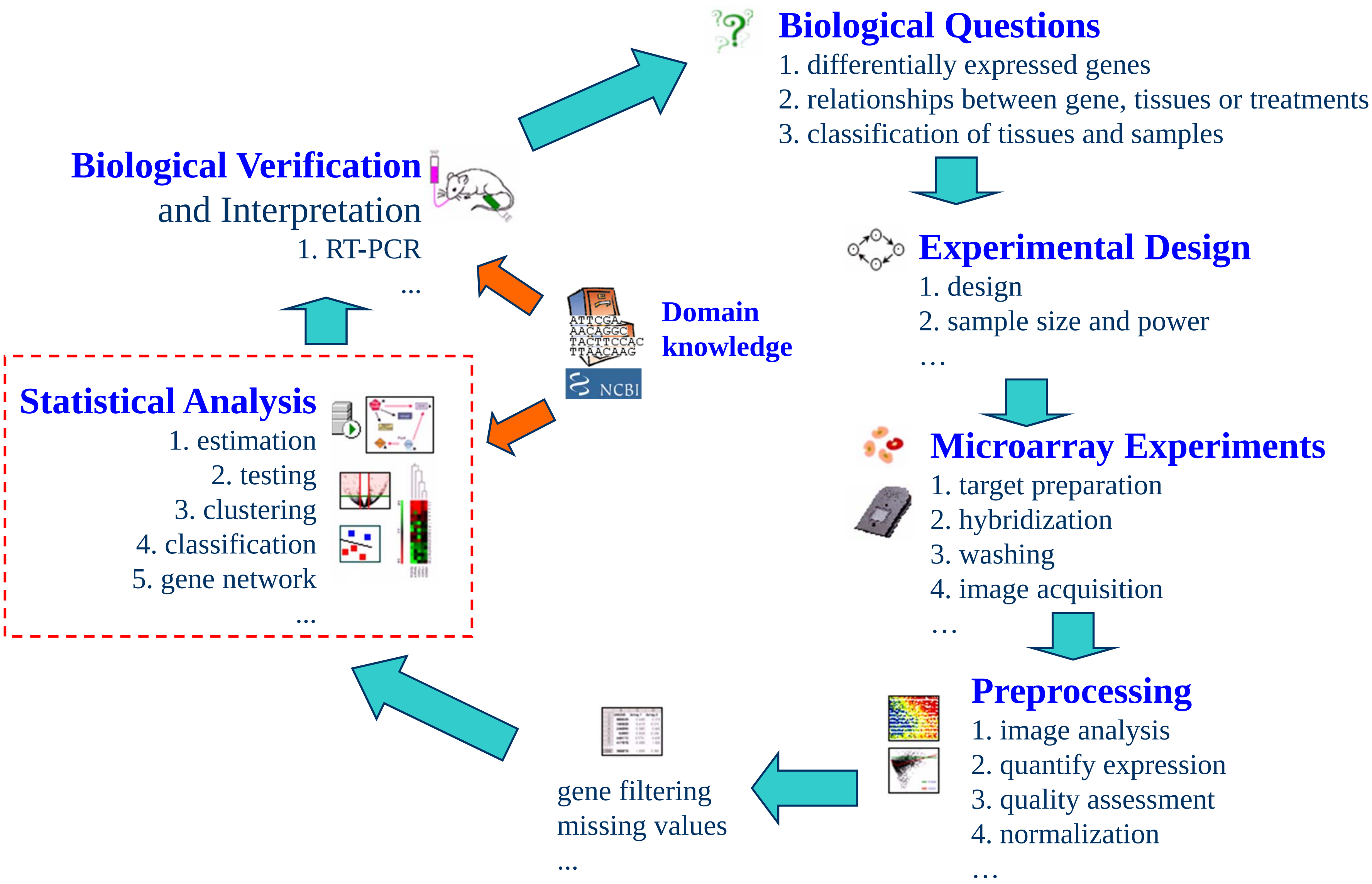
Content

- **Microarray Life Cycle**
- **Statistical Issues and Recent Progress**

- **Finding Differential Expressed Genes**
  - t-test
  - Significance Analysis of Microarrays (SAM)

- **Gene Set Analysis (GSA)**
  - Gene Set Enrichment Analysis (GSEA)

Microarray Life Cycle



Biological Questions
1. differentially expressed genes
2. relationships between gene, tissues or treatments
3. classification of tissues and samples

Experimental Design
1. design
2. sample size and power
…

Microarray Experiments
1. target preparation
2. hybridization
3. washing
4. image acquisition
…

Preprocessing
1. image analysis
2. quantify expression
3. quality assessment
4. normalization
…

gene filtering
missing values
...

Statistical Analysis
1. estimation
2. testing
3. clustering
4. classification
5. gene network
...

Biological Verification and Interpretation
1. RT-PCR
...

Domain knowledge

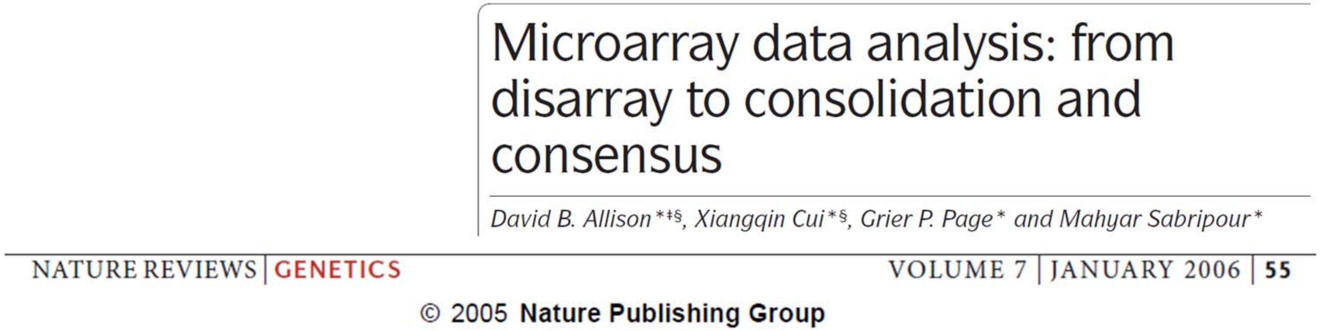
Basic Statistical Issues

- **Data Preprocessing: image processing, normalization**
- **Gene Filtering, Missing Values Imputation**
- ***Finding Differential Expressed Genes***
- ***Visualization (including dimension reduction)***
- ***Clustering***
- **Classification**
- **…**

Advance Statistical Issues

- Experimental Design
- Time Course Microarray Experiments
- Gene Regulatory Networks/Pathway
- Annotations/Databases
- Comparisons, Sample Size, Dye Swap, Replicates, …
- Web Resource, Software Design
- …

Recent Progress



Microarray data analysis: from disarray to consolidation and consensus

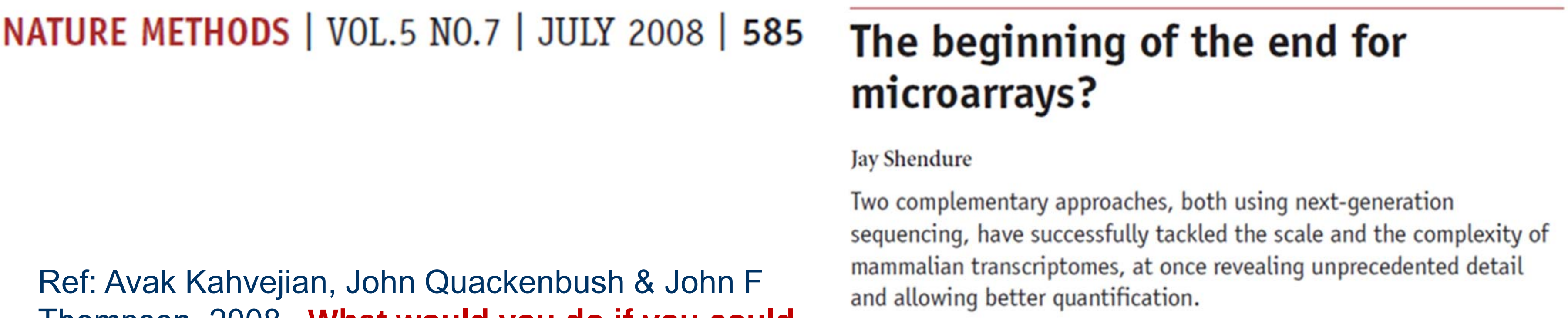
David B. Allison‡§, Xiangqin Cui*§, Grier P. Page* and Mahyar Sabripour**

NATURE REVIEWS | GENETICS VOLUME 7 | JANUARY 2006 | 55

© 2005 **Nature Publishing Group**

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

Recent Progress



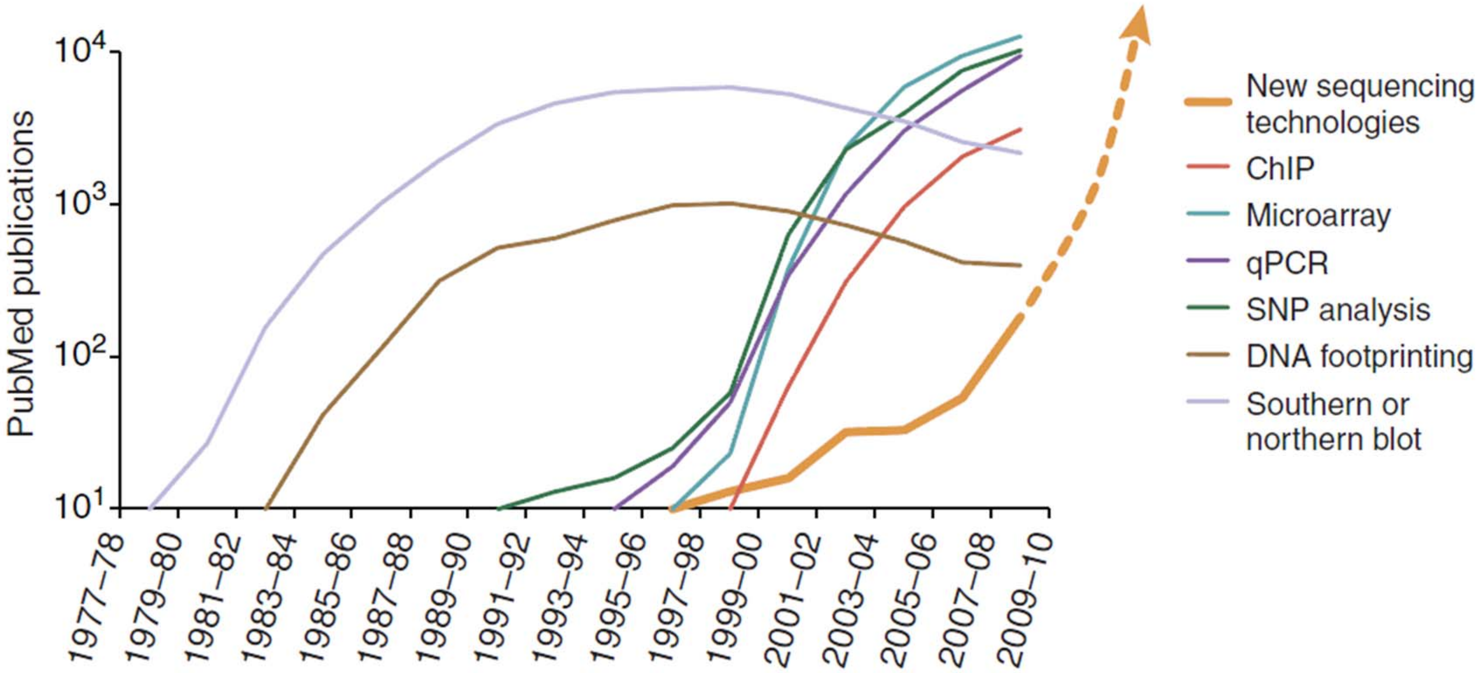
NATURE METHODS | VOL.5 NO.7 | JULY 2008 | 585

The beginning of the end for microarrays?

Jay Shendure

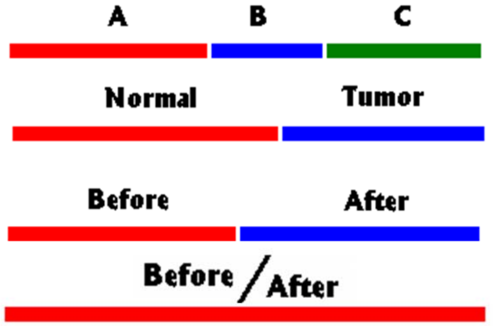
Two complementary approaches, both using next-generation sequencing, have successfully tackled the scale and the complexity of mammalian transcriptomes, at once revealing unprecedented detail and allowing better quantification.

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



Finding Differential Expressed Genes (DEGs)

Finding Differentially Expressed Genes

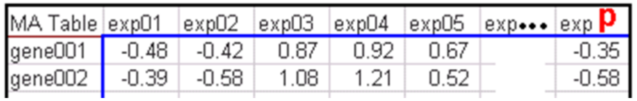


→ More than two samples

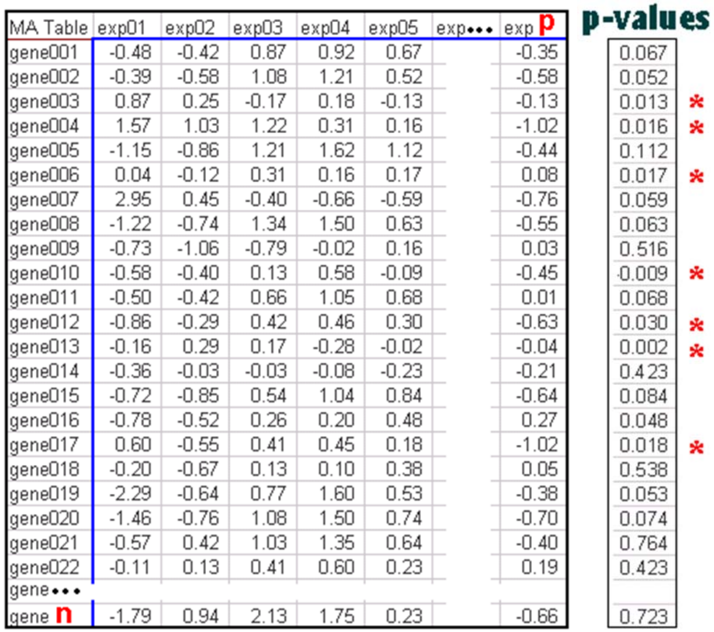
→ Two-sample (independent) ←

→ Paired-sample (dependent)

Cy 5: treatment
Cy 3: control

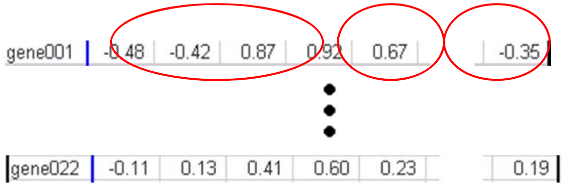


MA Table	exp01	exp02	exp03	exp04	exp05	exp•••	exp p
gene001	-0.48	-0.42	0.87	0.92	0.67		-0.35
gene002	-0.39	-0.58	1.08	1.21	0.52		-0.58



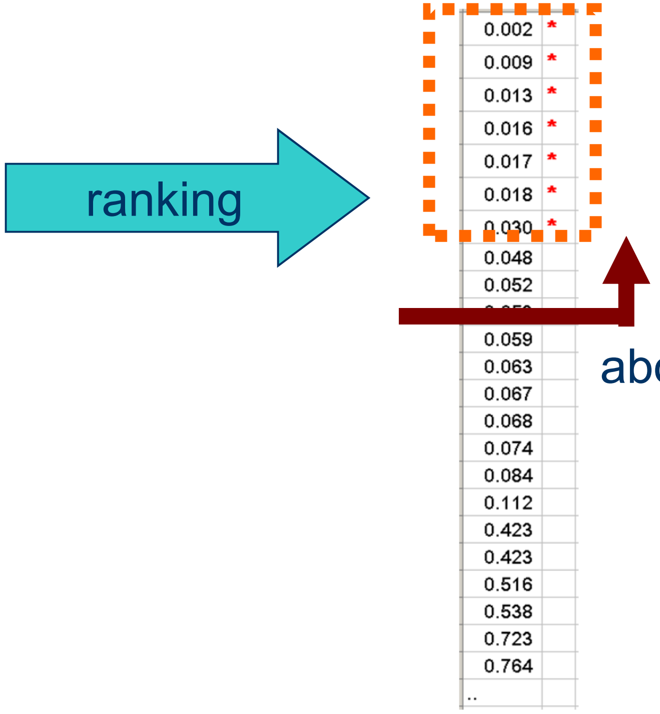
MA Table	exp01	exp02	exp03	exp04	exp05	exp•••	exp p	p-values	
gene001	-0.48	-0.42	0.87	0.92	0.67		-0.35	0.067	
gene002	-0.39	-0.58	1.08	1.21	0.52		-0.58	0.052	
gene003	0.87	0.25	-0.17	0.18	-0.13		-0.13	0.013	*
gene004	1.57	1.03	1.22	0.31	0.16		-1.02	0.016	*
gene005	-1.15	-0.86	1.21	1.62	1.12		-0.44	0.112	
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



gene001	-0.48	-0.42	0.87	0.92	0.67		-0.35
•							
•							
•							
gene022	-0.11	0.13	0.41	0.60	0.23		0.19

p-values or Statistics



0.002	*
0.009	*
0.013	*
0.016	*
0.017	*
0.018	*
0.030	*
0.048	
0.052	
0.053	
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	
..	

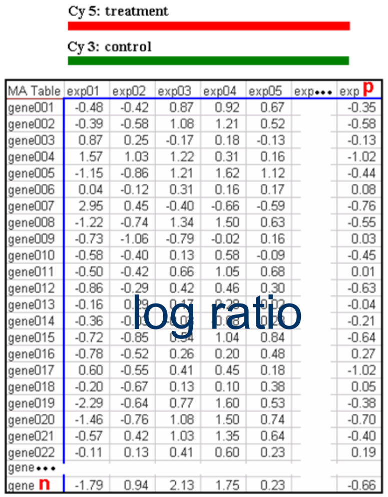
fix number

above some level

Paired Data: Breast Cancer Dataset

cDNA Microarrays Data:

- ***#Samples***: 20 breast cancer patients, before and after a 16 week course of doxorubicin chemotherapy
- ***#Genes***: 9216 genes.



Cy 5: treatment
Cy 3: control
log ratio

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
gene003	0.87	0.25	-0.17	0.18	-0.13		-0.13
gene004	1.57	1.03	1.22	0.31	0.16		-1.02
gene005	-1.15	-0.86	1.21	1.62	1.12		-0.44
gene006	0.04	-0.12	0.31	0.16	0.17		0.08
gene007	2.95	0.45	-0.40	-0.66	-0.59		-0.76
gene008	-1.22	-0.74	1.34	1.50	0.63		-0.55
gene009	-0.73	-1.06	-0.79	-0.02	0.16		0.03
gene010	-0.58	-0.40	0.13	0.58	-0.09		-0.45
gene011	-0.50	-0.42	0.66	1.05	0.68		0.01
gene012	-0.86	-0.29	0.42	0.46	0.30		-0.63
gene013	-0.16	[illegible]	[illegible]	[illegible]	[illegible]		-0.04
gene014	-0.36	[illegible]	[illegible]	[illegible]	[illegible]		-0.21
gene015	-0.72	-0.85	[illegible]	1.04	0.84		-0.64
gene016	-0.78	-0.52	0.26	0.20	0.48		0.27
gene017	0.60	-0.55	0.41	0.45	0.18		-1.02
gene018	-0.20	-0.67	0.13	0.10	0.38		0.05
gene019	-2.29	-0.64	0.77	1.60	0.53		-0.38
gene020	-1.46	-0.76	1.08	1.50	0.74		-0.70
gene021	-0.57	0.42	1.03	1.35	0.64		-0.40
gene022	-0.11	0.13	0.41	0.60	0.23		0.19
gene•••							
gene n	-1.79	0.94	2.13	1.75	0.23		-0.66

reference

9216 x 20

Paired Data:

- Two measurements from each patient, one before treatment and one after treatment.

Interests:

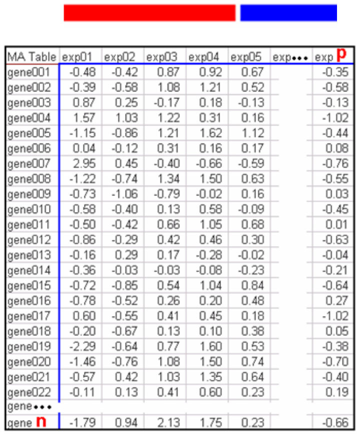
- the difference between the two measurements (the log ratio).
- whether a gene has been up-regulated or down-regulated in breast cancer following that treatment.

Perou CM, et al, (2000), Molecular portraits of human breast tumours. Nature 406:747-752.
Stanford Microarray Database: http://genome-www.stanford.edu/breast_cancer/molecularportraits/

Unpaired Data: Leukemia Dataset

Affymetrix Microarray Data

- ***#Samples***: Bone marrow
  - #ALL (acute lymphoblastic leukemia): 27 patients (急性淋巴細胞白血病)
  - #AML (acute myeloid leukemia): 11 patients (急性骨髓性白血病)
- ***#Genes***: 7070 genes.



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
gene003	0.87	0.25	-0.17	0.18	-0.13		-0.13
gene004	1.57	1.03	1.22	0.31	0.16		-1.02
gene005	-1.15	-0.86	1.21	1.62	1.12		-0.44
gene006	0.04	-0.12	0.31	0.16	0.17		0.08
gene007	2.95	0.45	-0.40	-0.66	-0.59		-0.76
gene008	-1.22	-0.74	1.34	1.50	0.63		-0.55
gene009	-0.73	-1.06	-0.79	-0.02	0.16		0.03
gene010	-0.58	-0.40	0.13	0.58	-0.09		-0.45
gene011	-0.50	-0.42	0.66	1.05	0.68		0.01
gene012	-0.86	-0.29	0.42	0.46	0.30		-0.63
gene013	-0.16	0.29	0.17	-0.28	-0.02		-0.04
gene014	-0.36	-0.03	-0.03	-0.08	-0.23		-0.21
gene015	-0.72	-0.85	0.54	1.04	0.84		-0.64
gene016	-0.78	-0.52	0.26	0.20	0.48		0.27
gene017	0.60	-0.55	0.41	0.45	0.18		-1.02
gene018	-0.20	-0.67	0.13	0.10	0.38		0.05
gene019	-2.29	-0.64	0.77	1.60	0.53		-0.38
gene020	-1.46	-0.76	1.08	1.50	0.74		-0.70
gene021	-0.57	0.42	1.03	1.35	0.64		-0.40
gene022	-0.11	0.13	0.41	0.60	0.23		0.19
gene•••							
gene n	-1.79	0.94	2.13	1.75	0.23		-0.66

7070 x (27+11)

Unpaired Data:

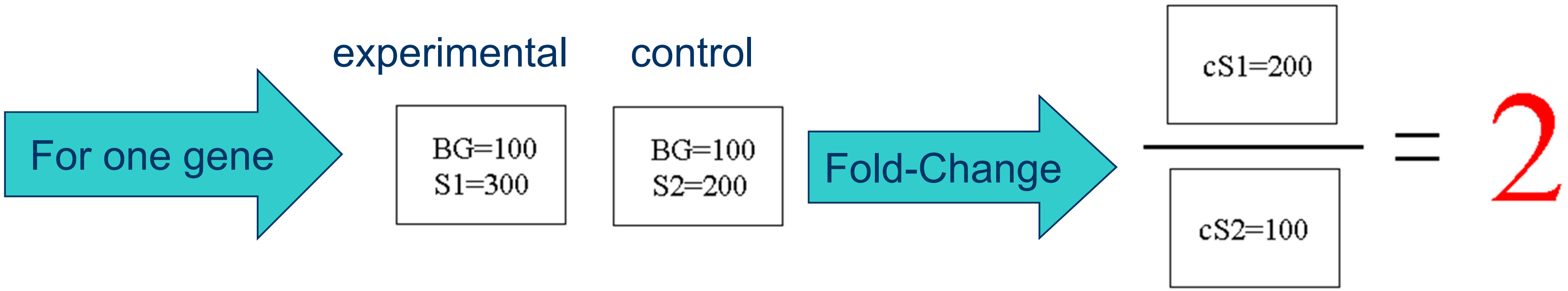
- Two groups of patients (ALL, AML).

Interests:

- To identify the genes that are up- or down-regulated in ALL relative to AML.
- (i.e., differentially expressed between the two groups.)

Golub, T.R et al. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286, 531--537.
Cancer Genomics Program at Whitehead Institute for Genome Research
http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi

Fold-Change Method (1)



For one gene

experimental

BG=100
S1=300

control

BG=100
S2=200

Fold-Change

cS1=200

cS2=100

= 2

1) Calculate fold-change.

2) Rank the genes.

3) Select genes.

Fold-Change Method (2)

Method 1: Select genes based on Numbers

- average differential expression > **FC.**

Problems:

- **FC** is an arbitrary threshold.
- **FC** does not take into account individuals and sample size.

Example:

- s2 (200) close to BG (100), the difference could represent noise.
- credible: a gene is regulated 2-fold with 10000, 5000 units.

Fold-Change Method (3)

Method 2: Select genes based on %

- Choose 5% of genes that have the largest expression ratios.

Problems:

- Possible that no genes have statistically significantly different gene expression.

Hypothesis Testing

null hypothesis:

Biological Question → ***Statistical Formulation***

$\mathbf{H_0}$: 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.

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

The *p*-values

p-values

- Probability of **false positives** (Reject $\mathbf{H_0}$ | $\mathbf{H_0}$ true).
- Probability of observing your data under the assumption that the null hypothesis is true.
- *p-value* = 0.03: only a 3% chance of drawing the sample if the null hypothesis was true.

Decision Rule

- Reject H_0 if *p-value* is less than alpha.
- $P < 0.05$ commonly used. (Reject $\mathbf{H_0}$**,** the test is significant)
- The lower the *p-value*, the more significant.

Use p-value to select genes

- Select differentially expressed genes based on their p-value (not FC).
- The smaller the p-value, the less likely it is that the observed data have occurred by chance, and the more significant the result.

One Sample t-test

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.

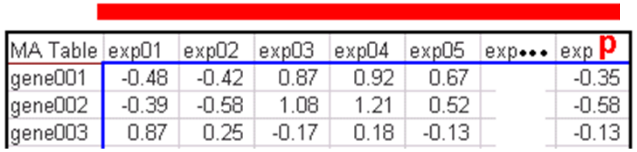
S: sample standard deviation.

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-value* $= P_{H_0}(|\mathbf{T}| > t_0)$, $\mathbf{T} \sim t_{n-1}$.

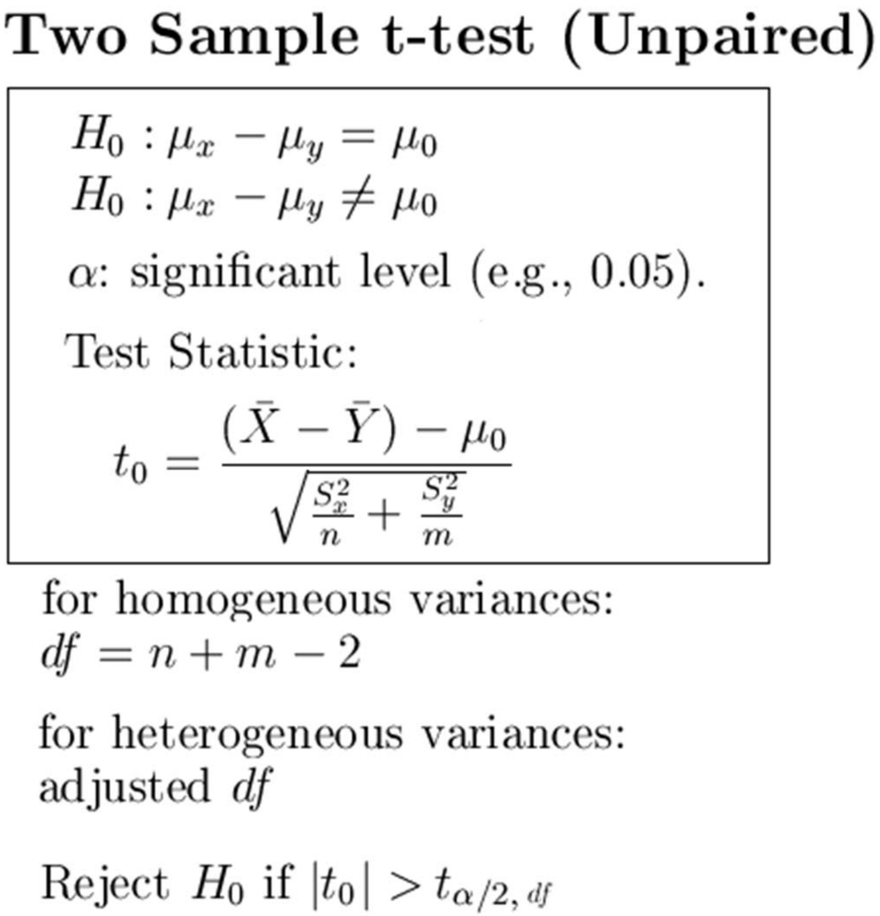
Question

- whether a gene is differentially expressed for a condition with respect to baseline expression?
- H_0: μ=0 (log ratio)



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
gene003	0.87	0.25	-0.17	0.18	-0.13		-0.13

Two Sample t-test (Unpaired)



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

Applied to a Gene From Leukemia Dataset

metallothionein IB

- The gene metallothionein IB is on the Affymetrix array used for the leukemia data.

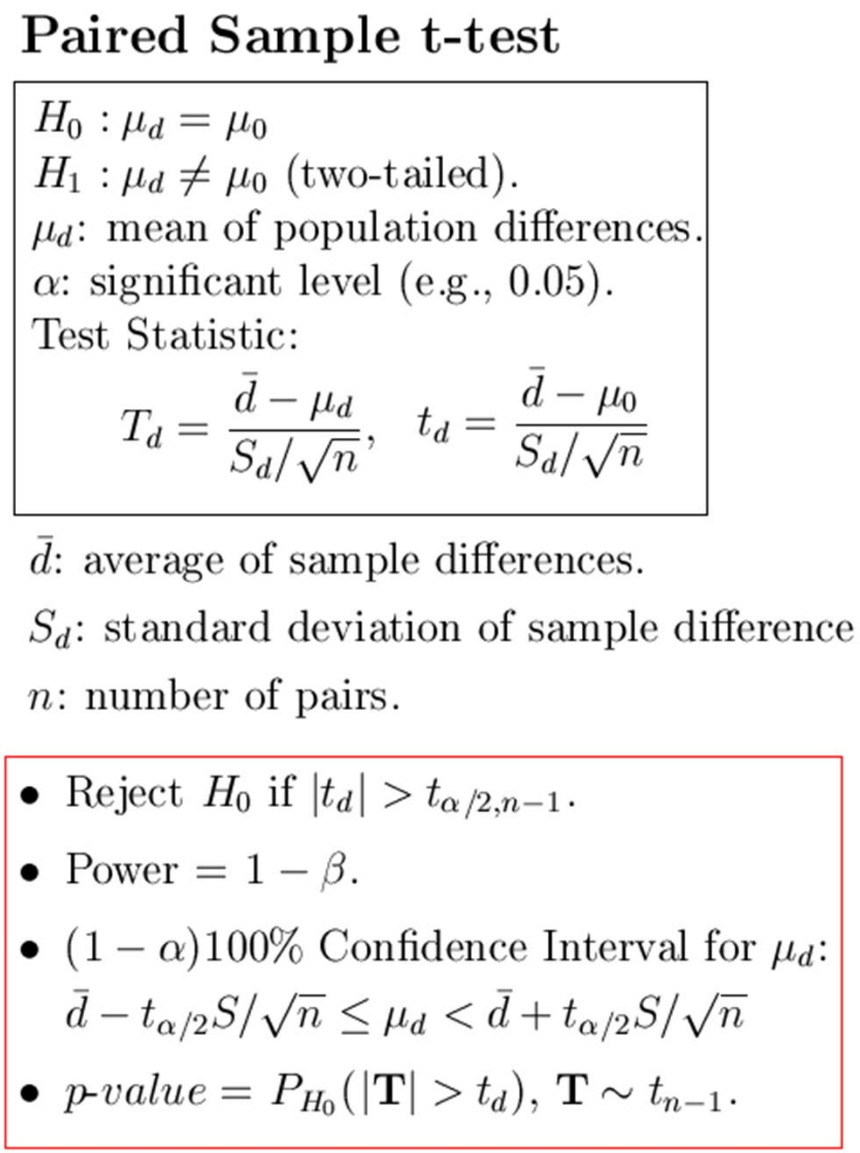
Two-sample t-test

- t=-3.4177, p=0.0016.

Conclusion

- the expression of metallothionein IB is significantly higher in AML than in ALL at the 1% level.

Two Sample t-test (Paired)



Paired Sample t-test

$H_0: \mu_d = \mu_0$

$H_1: \mu_d \neq \mu_0$ (two-tailed).

μ_d: mean of population differences.

α: significant level (e.g., 0.05).

Test Statistic:

$$T_d = \frac{\bar{d} - \mu_d}{S_d/\sqrt{n}}, \quad t_d = \frac{\bar{d} - \mu_0}{S_d/\sqrt{n}}$$

$\bar{d}$: average of sample differences.

S_d: standard deviation of sample difference

n: number of pairs.

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

Applied to a gene From Breast Cancer Data

ACAT2

- The gene acetyl-Coenzyme A acetyltransferase 2 (ACAT2) is on the microarray used for the breast cancer data.

Paired t-test

- t=3.22. (two-tailed)
- p-value = 0.0045, which is significant at a 1% confidence level.

Conclusion

- ACAT2 has been significantly down-regulated following chemotherapy at the 1% level.

Assumptions of t-test

Be Normal

- paired t-test,

  the distribution of the subtracted data that must be normal.
- unpaired t-test,

  the distribution of both data sets must be normal.

How to Detect Normality

- **Plots**: Histogram, Density Plot, QQplot,…
- **Test for Normality**: Jarque-Bera test, Lilliefors test, Kolmogorov-Smirnov test.

Homogeneous

- the variances of the two population are equal.
- Test for equality of the two variances: Variance ratio F-test.

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.

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

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

Lonnstedt, I. and Speed, T.P. Replicated microarray data. *Statistica Sinica* , 12: 31-46, 2002

Penalized t-statistic

Tusher et al (2001, PNAS, SAM)

Efron et al (2001, JASA)

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

General Penalized t-statistic

(Lonnstedt et al 2001)

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

multiple regression model

Penalized two-sample t-statistic

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

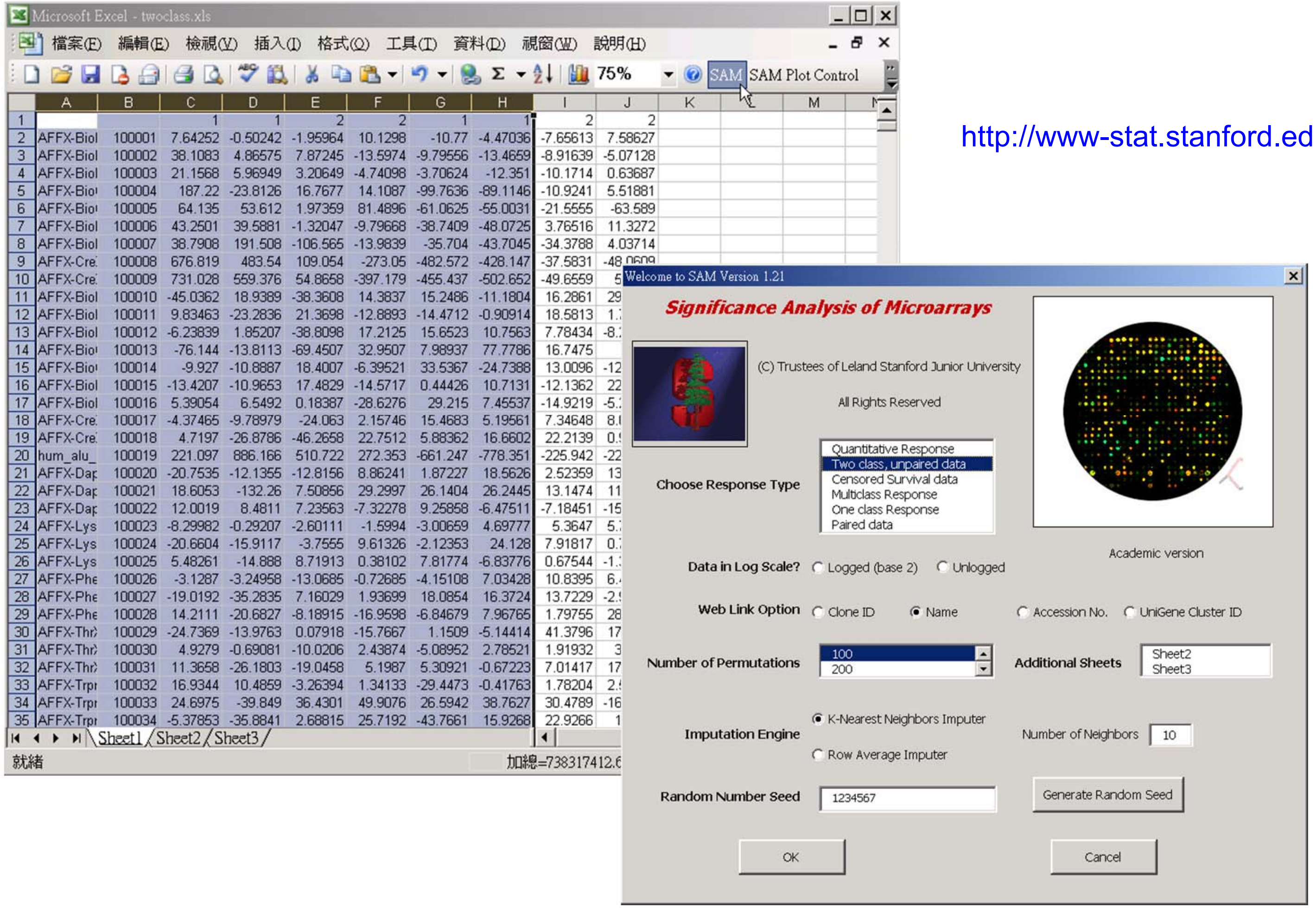
Robust General Penalized t-statistic

Significance Analysis of Microarrays (SAM)

SAM: Significance Analysis of Microarrays

SAM does not do any normalization!

http://www-stat.stanford.edu/~tibs/SAM/



Microsoft Excel - twoclass.xls
檔案(F) 編輯(E) 檢視(V) 插入(I) 格式(O) 工具(T) 資料(D) 視窗(W) 說明(H)
75%
SAM
SAM Plot Control
Sheet1
Sheet2
Sheet3
就緒
加總=738317412.6

	A	B	C	D	E	F	G	H	I	J
1			1	1	2	2	1	1	2	2
2	AFFX-Biol	100001	7.64252	-0.50242	-1.95964	10.1298	-10.77	-4.47036	-7.65613	7.58627
3	AFFX-Biol	100002	38.1083	4.86575	7.87245	-13.5974	-9.79556	-13.4659	-8.91639	-5.07128
4	AFFX-Biol	100003	21.1568	5.96949	3.20649	-4.74098	-3.70624	-12.351	-10.1714	0.63687
5	AFFX-Bio	100004	187.22	-23.8126	16.7677	14.1087	-99.7636	-89.1146	-10.9241	5.51881
6	AFFX-Bio	100005	64.135	53.612	1.97359	81.4896	-61.0625	-55.0031	-21.5555	-63.589
7	AFFX-Biol	100006	43.2501	39.5881	-1.32047	-9.79668	-38.7409	-48.0725	3.76516	11.3272
8	AFFX-Biol	100007	38.7908	191.508	-106.565	-13.9839	-35.704	-43.7045	-34.3788	4.03714
9	AFFX-Cre	100008	676.819	483.54	109.054	-273.05	-482.572	-428.147	-37.5831	-48.0609
10	AFFX-Cre	100009	731.028	559.376	54.8658	-397.179	-455.437	-502.652	-49.6559	5
11	AFFX-Biol	100010	-45.0362	18.9389	-38.3608	14.3837	15.2486	-11.1804	16.2861	29
12	AFFX-Biol	100011	9.83463	-23.2836	21.3698	-12.8893	-14.4712	-0.90914	18.5813	1.
13	AFFX-Biol	100012	-6.23839	1.85207	-38.8098	17.2125	15.6523	10.7563	7.78434	-8.
14	AFFX-Bio	100013	-76.144	-13.8113	-69.4507	32.9507	7.98937	77.7786	16.7475	
15	AFFX-Bio	100014	-9.927	-10.8887	18.4007	-6.39521	33.5367	-24.7388	13.0096	-12
16	AFFX-Biol	100015	-13.4207	-10.9653	17.4829	-14.5717	0.44426	10.7131	-12.1362	22
17	AFFX-Biol	100016	5.39054	6.5492	0.18387	-28.6276	29.215	7.45537	-14.9219	-5.
18	AFFX-Cre	100017	-4.37465	-9.78979	-24.063	2.15746	15.4683	5.19561	7.34648	8.
19	AFFX-Cre	100018	4.7197	-26.8786	-46.2658	22.7512	5.88362	16.6602	22.2139	0.
20	hum_alu_	100019	221.097	886.166	510.722	272.353	-661.247	-778.351	-225.942	-22
21	AFFX-Dap	100020	-20.7535	-12.1355	-12.8156	8.86241	1.87227	18.5626	2.52359	13
22	AFFX-Dap	100021	18.6053	-132.26	7.50856	29.2997	26.1404	26.2445	13.1474	11
23	AFFX-Dap	100022	12.0019	8.4811	7.23563	-7.32278	9.25858	-6.47511	-7.18451	-15
24	AFFX-Lys	100023	-8.29982	-0.29207	-2.60111	-1.5994	-3.00659	4.69777	5.3647	5.
25	AFFX-Lys	100024	-20.6604	-15.9117	-3.7555	9.61326	-2.12353	24.128	7.91817	0.
26	AFFX-Lys	100025	5.48261	-14.888	8.71913	0.38102	7.81774	-6.83776	0.67544	-1.
27	AFFX-Phe	100026	-3.1287	-3.24958	-13.0685	-0.72685	-4.15108	7.03428	10.8395	6.
28	AFFX-Phe	100027	-19.0192	-35.2835	7.16029	1.93699	18.0854	16.3724	13.7229	-2.
29	AFFX-Phe	100028	14.2111	-20.6827	-8.18915	-16.9598	-6.84679	7.96765	1.79755	28
30	AFFX-Thr	100029	-24.7369	-13.9763	0.07918	-15.7667	1.1509	-5.14414	41.3796	17
31	AFFX-Thr	100030	4.9279	-0.69081	-10.0206	2.43874	-5.08952	2.78521	1.91932	3
32	AFFX-Thr	100031	11.3658	-26.1803	-19.0458	5.1987	5.30921	-0.67223	7.01417	17
33	AFFX-Trpı	100032	16.9344	10.4859	-3.26394	1.34133	-29.4473	-0.41763	1.78204	2.
34	AFFX-Trpı	100033	24.6975	-39.849	36.4301	49.9076	26.5942	38.7627	30.4789	-16
35	AFFX-Trpı	100034	-5.37853	-35.8841	2.68815	25.7192	-43.7661	15.9268	22.9266	1

Welcome to SAM Version 1.21
Significance Analysis of Microarrays
(C) Trustees of Leland Stanford Junior University
All Rights Reserved
Academic version
Generate Random Seed
OK
Cancel

Choose Response Type: Quantitative Response / **Two class, unpaired data** (selected) / Censored Survival data / Multiclass Response / One class Response / Paired data

Data in Log Scale? ○ Logged (base 2) ○ Unlogged

Web Link Option ○ Clone ID ● Name ○ Accession No. ○ UniGene Cluster ID

Number of Permutations: **100** (selected) / 200

Additional Sheets: Sheet2 / Sheet3

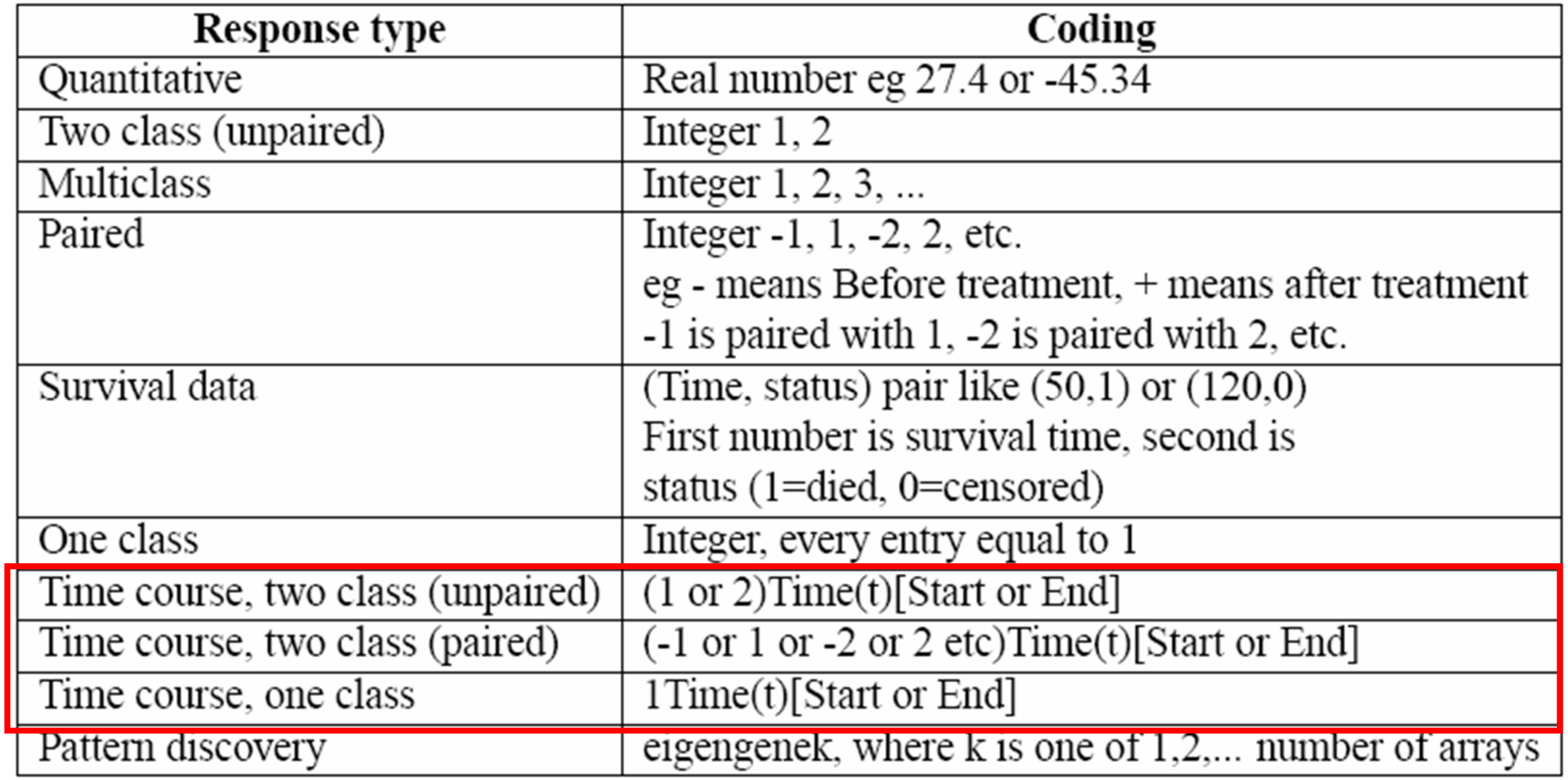
Imputation Engine ● K-Nearest Neighbors Imputer ○ Row Average Imputer

Number of Neighbors: 10

Random Number Seed: 1234567

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.

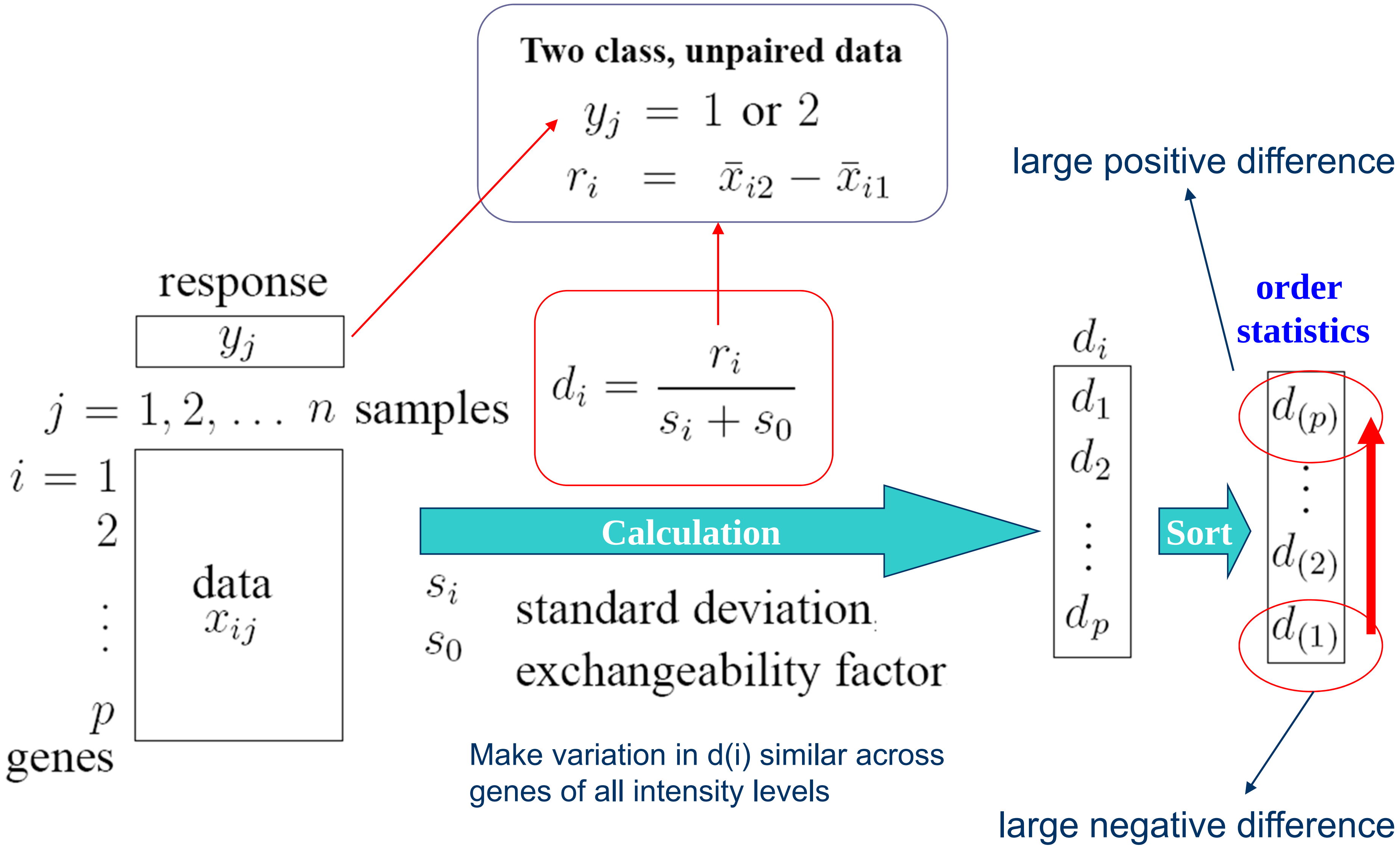
SAM: Response Type



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 Users guide and technical document

SAM: Significance Analysis of Microarrays



Two class, unpaired data

$$y_j = 1 \text{ or } 2$$
$$r_i = \bar{x}_{i2} - \bar{x}_{i1}$$

response y_j

$j = 1, 2, \ldots\ n$ samples

$i = 1, 2, \vdots, p$ genes

data x_{ij}

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

Calculation

s_i standard deviation

s_0 exchangeability factor

Make variation in d(i) similar across genes of all intensity levels

d_i: $d_1, d_2, \vdots, d_p$

Sort

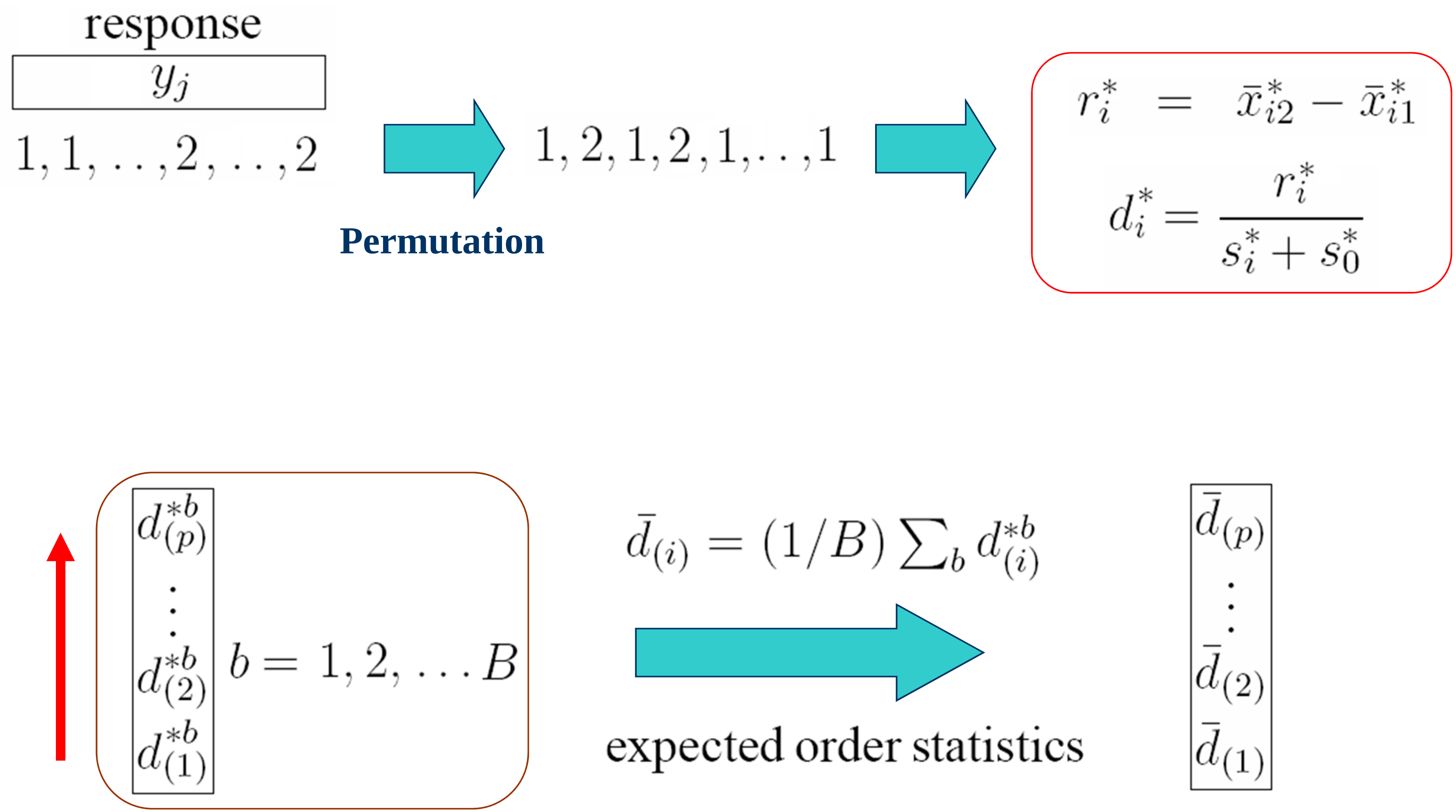
order statistics

$d_{(p)}, \vdots, d_{(2)}, d_{(1)}$

large positive difference

large negative difference

SAM: Expected Test Statistics



response

y_j

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

Permutation

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

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

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

$$\begin{array}{|c|} d^{*b}_{(p)} \\ \vdots \\ d^{*b}_{(2)} \\ d^{*b}_{(1)} \end{array} \quad b = 1, 2, \ldots B$$

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

expected order statistics

$$\begin{array}{|c|} \bar{d}_{(p)} \\ \vdots \\ \bar{d}_{(2)} \\ \bar{d}_{(1)} \end{array}$$

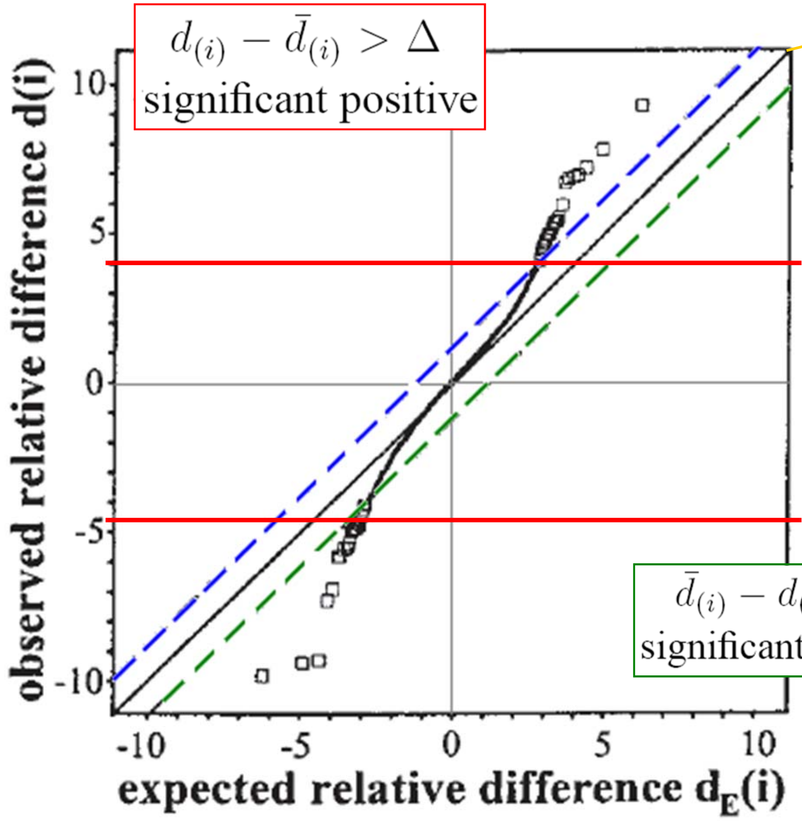
SAM Plot

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

vs

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

Points for genes with evidence of induction



$d_{(i)} - \bar{d}_{(i)} > \Delta$ significant positive

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

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

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

$\bar{d}_{(i)} - d_{(i)} > \Delta$ significant negative

Points for genes with evidence of repression

Gene Set Enrichment Analysis (GSEA)

Gene Sets

- Whether some functionally predefined classes of genes are differentially expressed?

- **A gene set (a gene class)**
  - a group of genes with related functions.
  - sets of genes or pathways, for their association with a phenotype.
  - identified from a **prior** biological knowledge.
  - may better reflect the true underlying biology.
  - may be more appropriate **units** for analysis.

- **Examples:** metabolic pathway, protein complex, or GO (gene ontology) category.

- **Various database**: BioCarta, KEGG, Gene Ontology

Gene Set Analysis

- A statistical test to determine significance of a gene class is referred to as gene class testing (**GCT**) or gene set analysis (**GSA**).
  - The common approach to the GSA is first to identify a list of genes that express differently among two groups of samples.
  - The list of differentially expressed genes is then examined with biologically pre-defined gene sets to determine whether any set is overrepresented in the list compared with the whole list.
- GSA is becoming a powerful alternative to individual-gene analysis.

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)

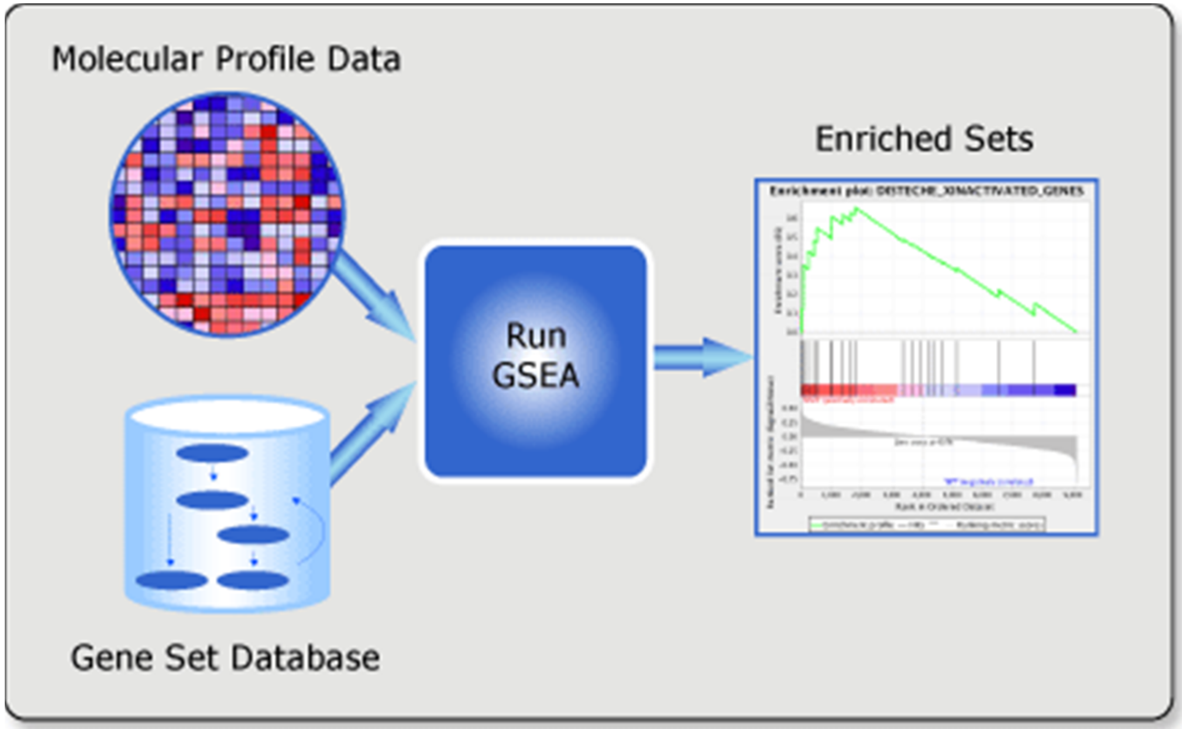
Gene set enrichment analysis: A knowledge-based approach for ...

- [翻譯這個網頁]

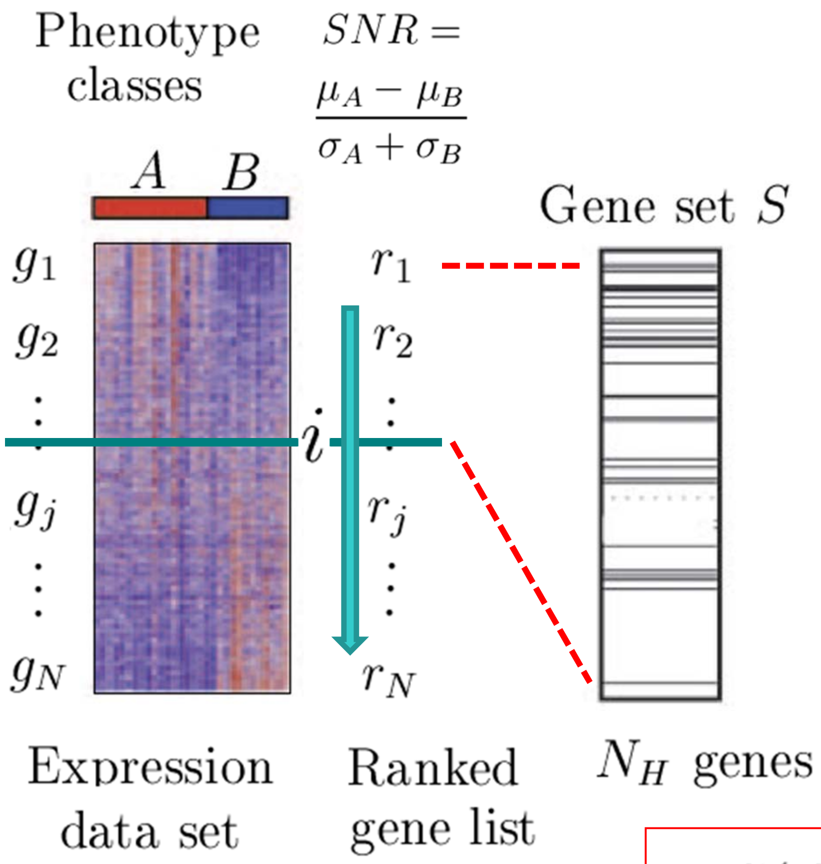
由 A Subramanian 著作 - 2005 - 被引用 2442 次 - 相關文章

25 Oct 2005 ... Overview of GSEA. GSEA considers experiments with genomewide expression profiles from samples belonging to two classes, labeled 1 or 2. ...

www.pnas.org/content/102/43/15545.full - 類似內容



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.

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

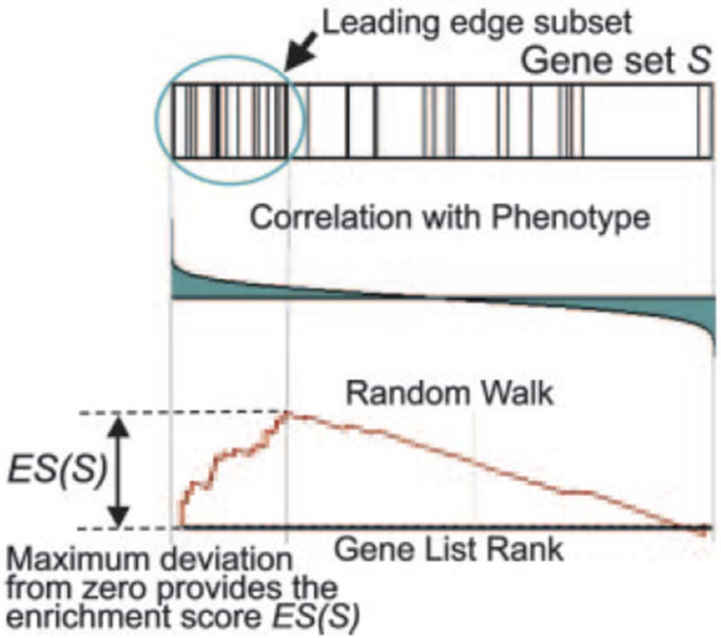
$$P_{\text{miss}}(S,i) = \sum_{g_j \notin S, j \le i} \frac{1}{N - N_H}$$

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

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

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



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

- If p=0

  ES(S) = Kolmogorov-Smirnov statistic.

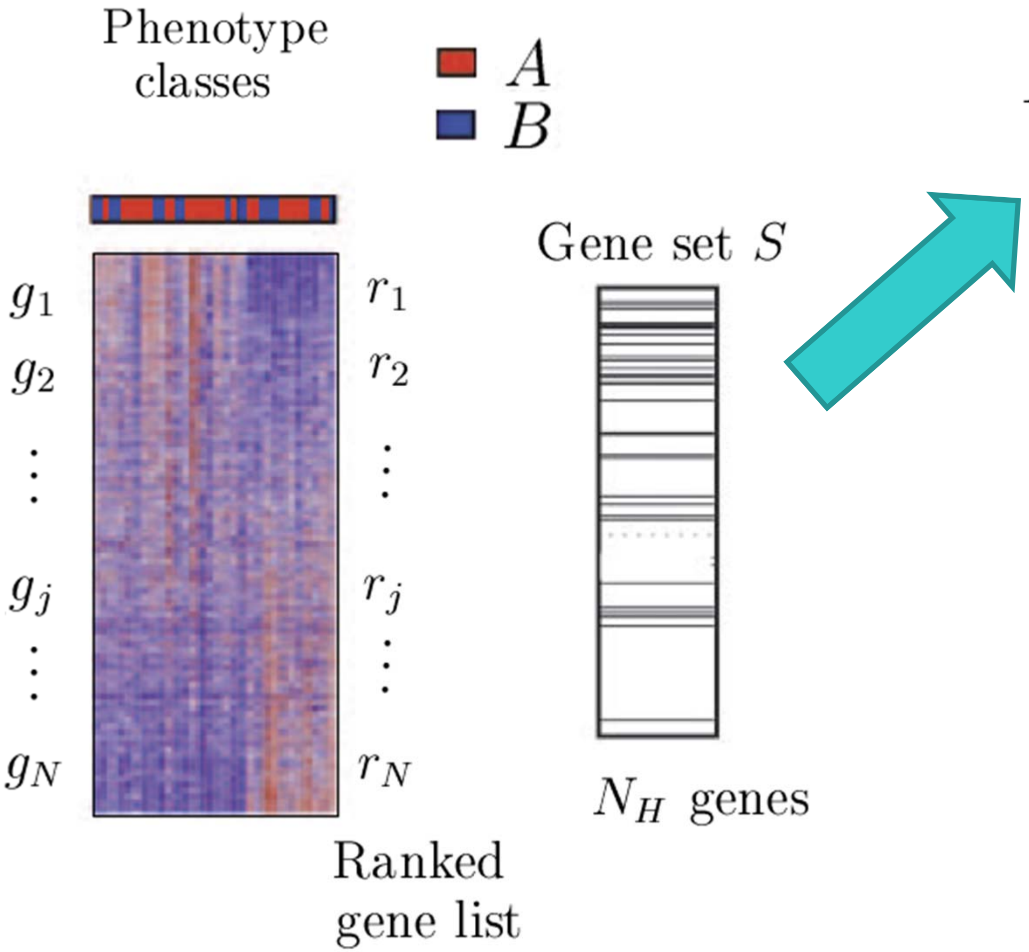
- Set p=1.

- 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

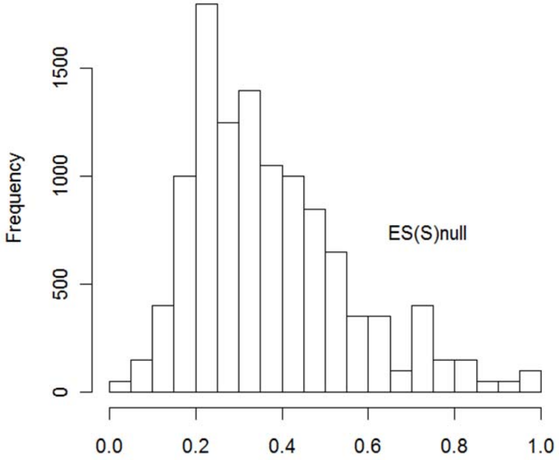
Assess the significance of an observed *ES* by comparing it with the set of score *Esnull* computed with randomly assigned phonotype.



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

- For positive ES
- For negative ES

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



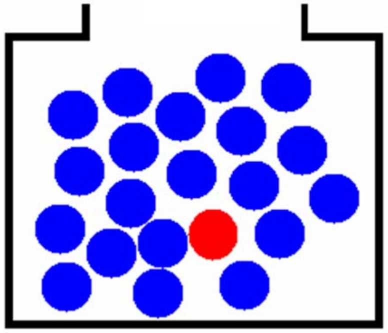
Step 3: Multiple Hypothesis Testing

X: false positive gene

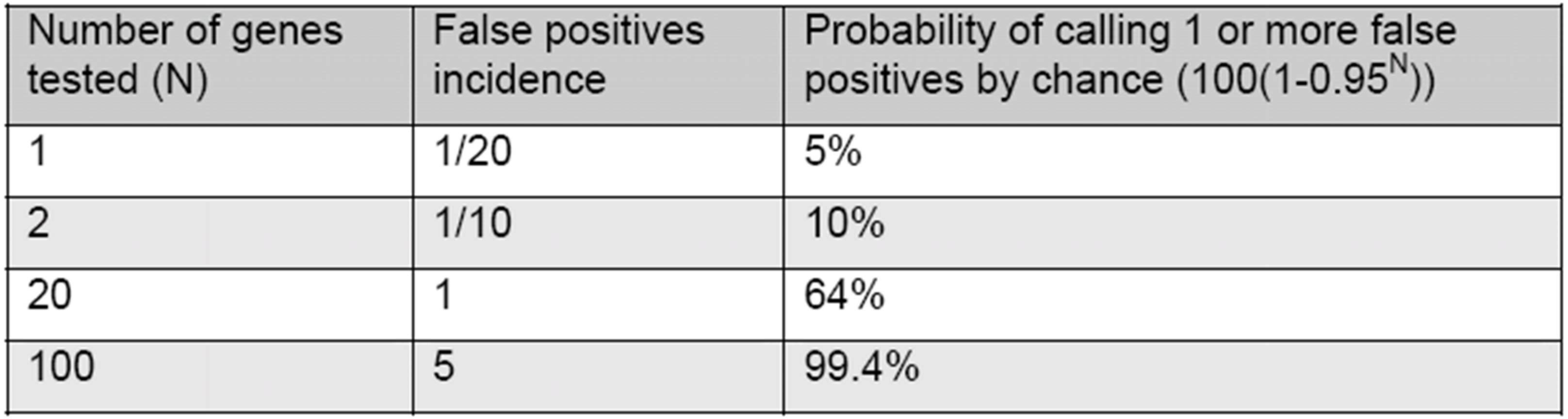
$$P(X \geq 1)$$

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

$$= 1 - 0.95^{n}$$



Population

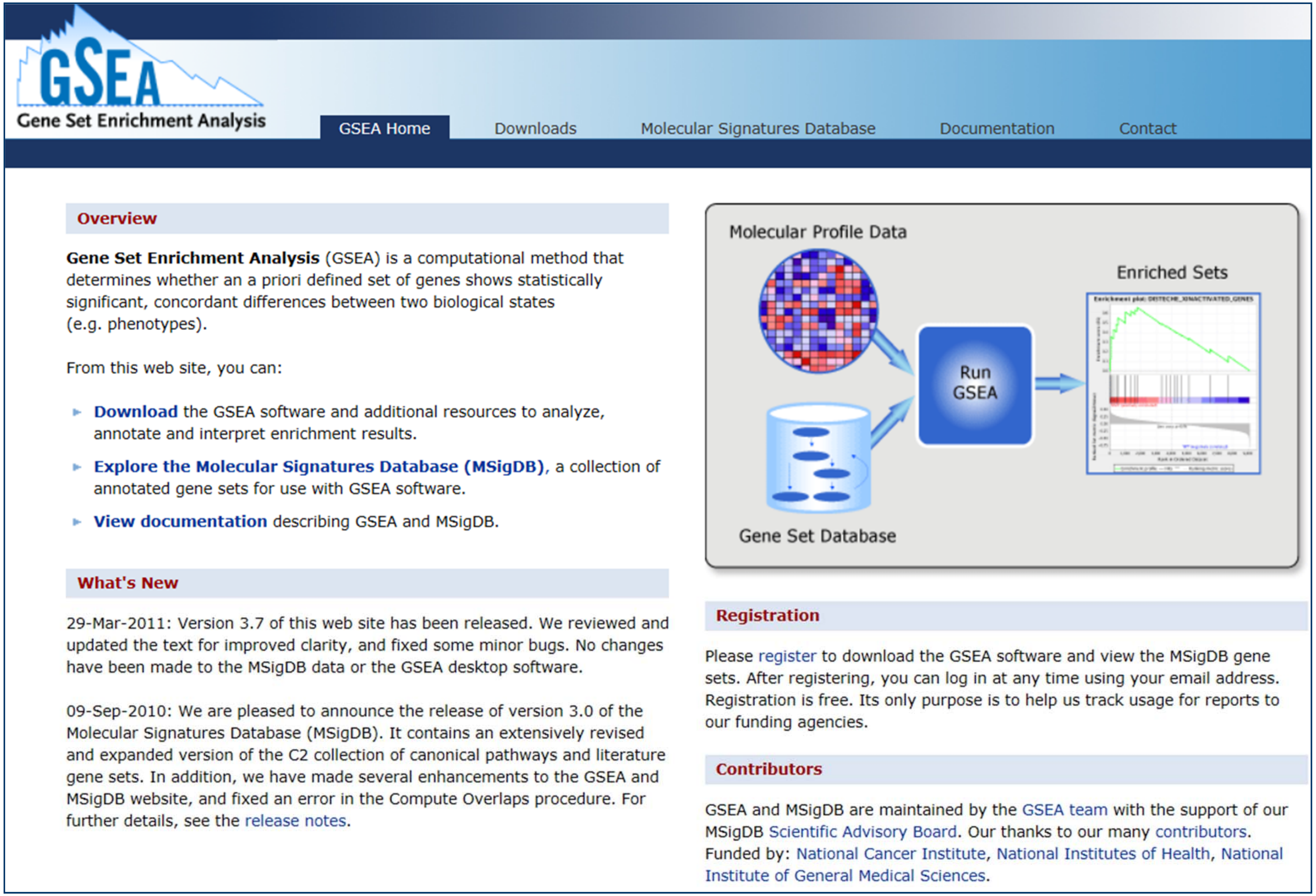


Number of genes tested (N)	False positives incidence	Probability of calling 1 or more false positives by chance (100(1-0.95^{N}))
1	1/20	5%
2	1/10	10%
20	1	64%
100	5	99.4%

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



Overview

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

From this web site, you can:

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

What's New

29-Mar-2011: Version 3.7 of this web site has been released. We reviewed and updated the text for improved clarity, and fixed some minor bugs. No changes have been made to the MSigDB data or the GSEA desktop software.

09-Sep-2010: We are pleased to announce the release of version 3.0 of the Molecular Signatures Database (MSigDB). It contains an extensively revised and expanded version of the C2 collection of canonical pathways and literature gene sets. In addition, we have made several enhancements to the GSEA and MSigDB website, and fixed an error in the Compute Overlaps procedure. For further details, see the release notes.

Registration

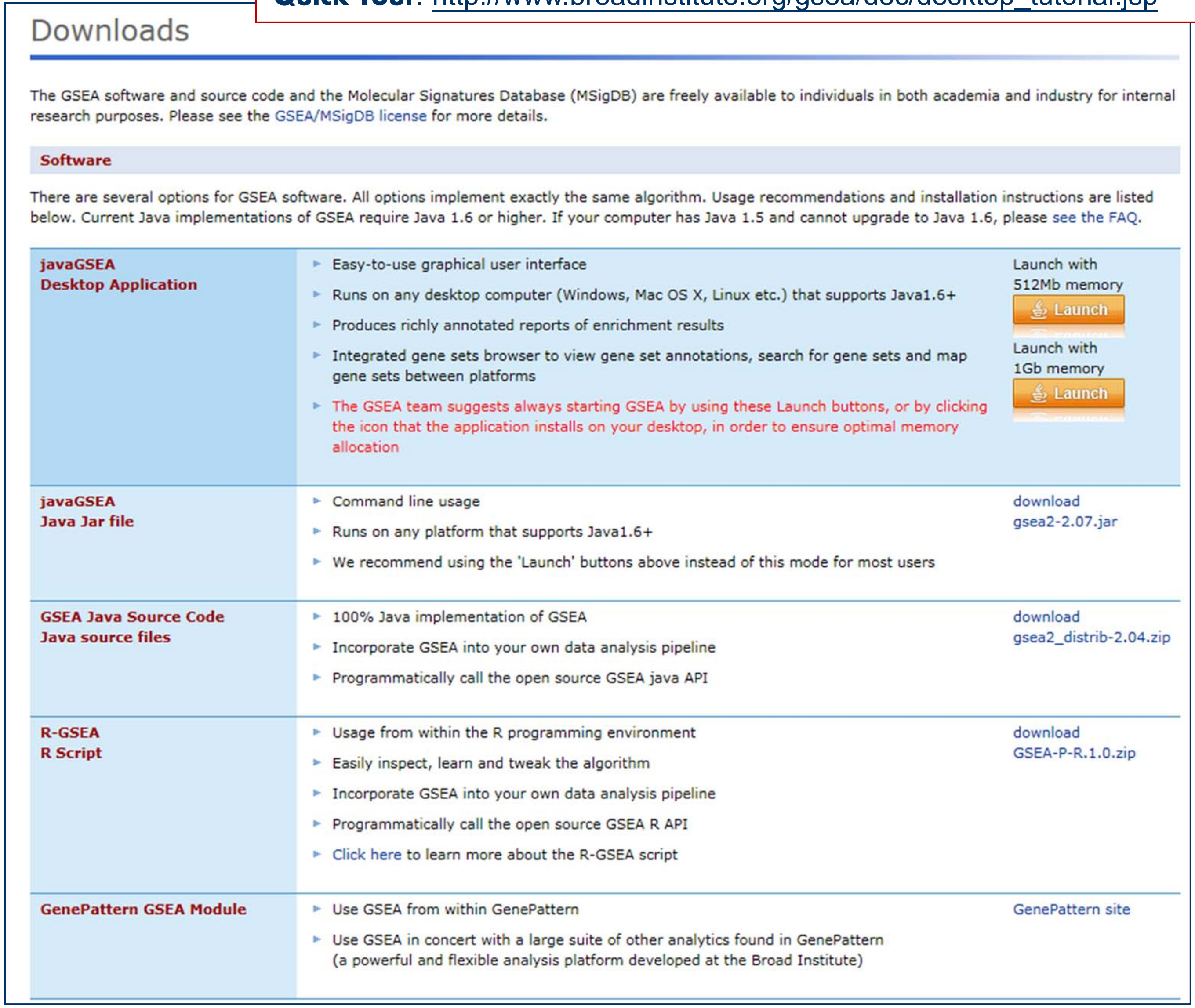
Please register to download the GSEA software and view the MSigDB gene sets. After registering, you can log in at any time using your email address. Registration is free. Its only purpose is to help us track usage for reports to our funding agencies.

Contributors

GSEA and MSigDB are maintained by the GSEA team with the support of our MSigDB Scientific Advisory Board. Our thanks to our many contributors. Funded by: National Cancer Institute, National Institutes of Health, National 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

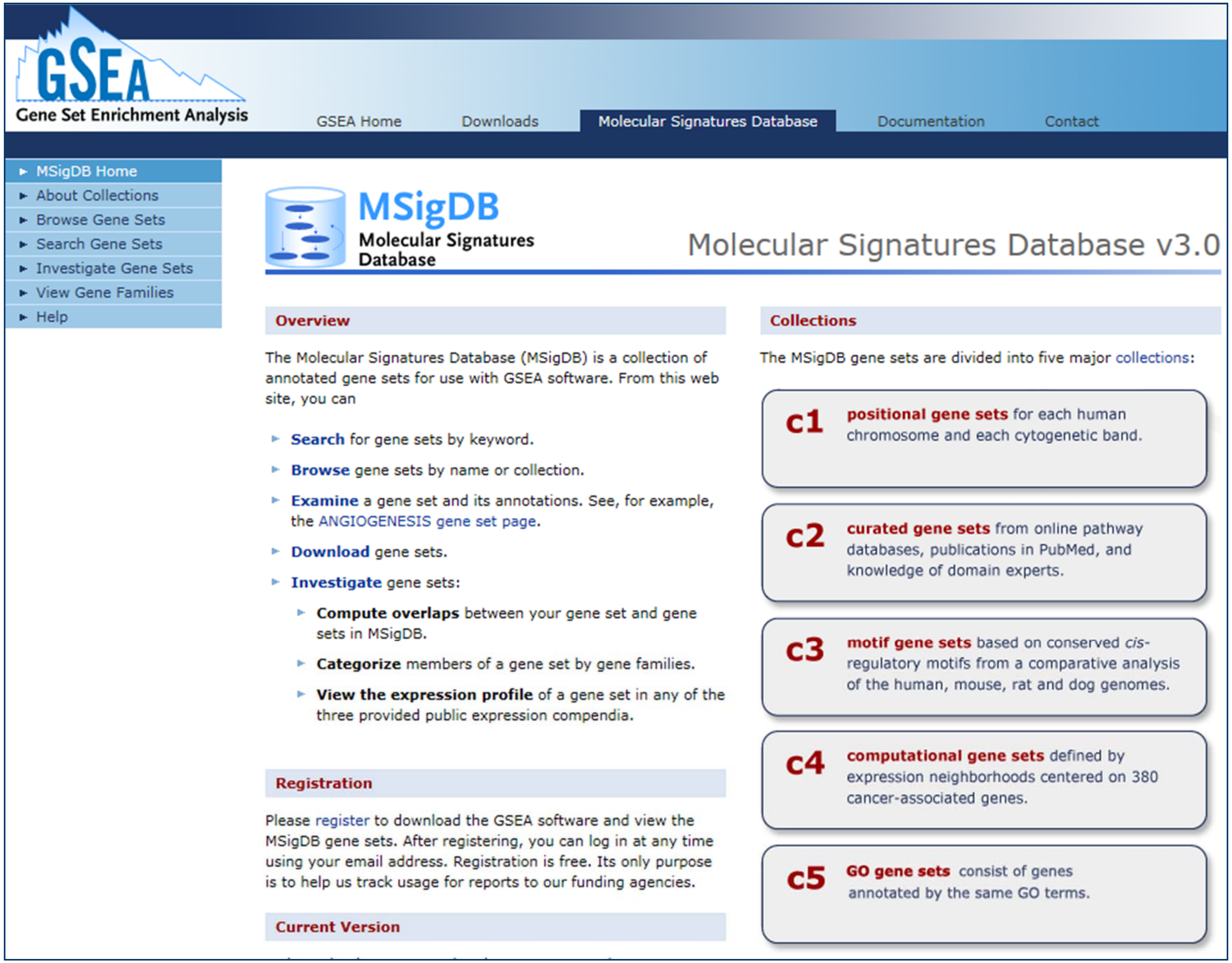
The GSEA software and source code and the Molecular Signatures Database (MSigDB) are freely available to individuals in both academia and industry for internal research purposes. Please see the GSEA/MSigDB license for more details.

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)



Molecular Signatures Database v3.0

Overview

The Molecular Signatures Database (MSigDB) is a collection of annotated gene sets for use with GSEA software. From this web site, you can

- **Search** for gene sets by keyword.
- **Browse** gene sets by name or collection.
- **Examine** a gene set and its annotations. See, for example, the ANGIOGENESIS gene set page.
- **Download** gene sets.
- **Investigate** gene sets:
  - **Compute overlaps** between your gene set and gene sets in MSigDB.
  - **Categorize** members of a gene set by gene families.
  - **View the expression profile** of a gene set in any of the three provided public expression compendia.

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.

Current Version

Collections

The MSigDB gene sets are divided into five major collections:

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

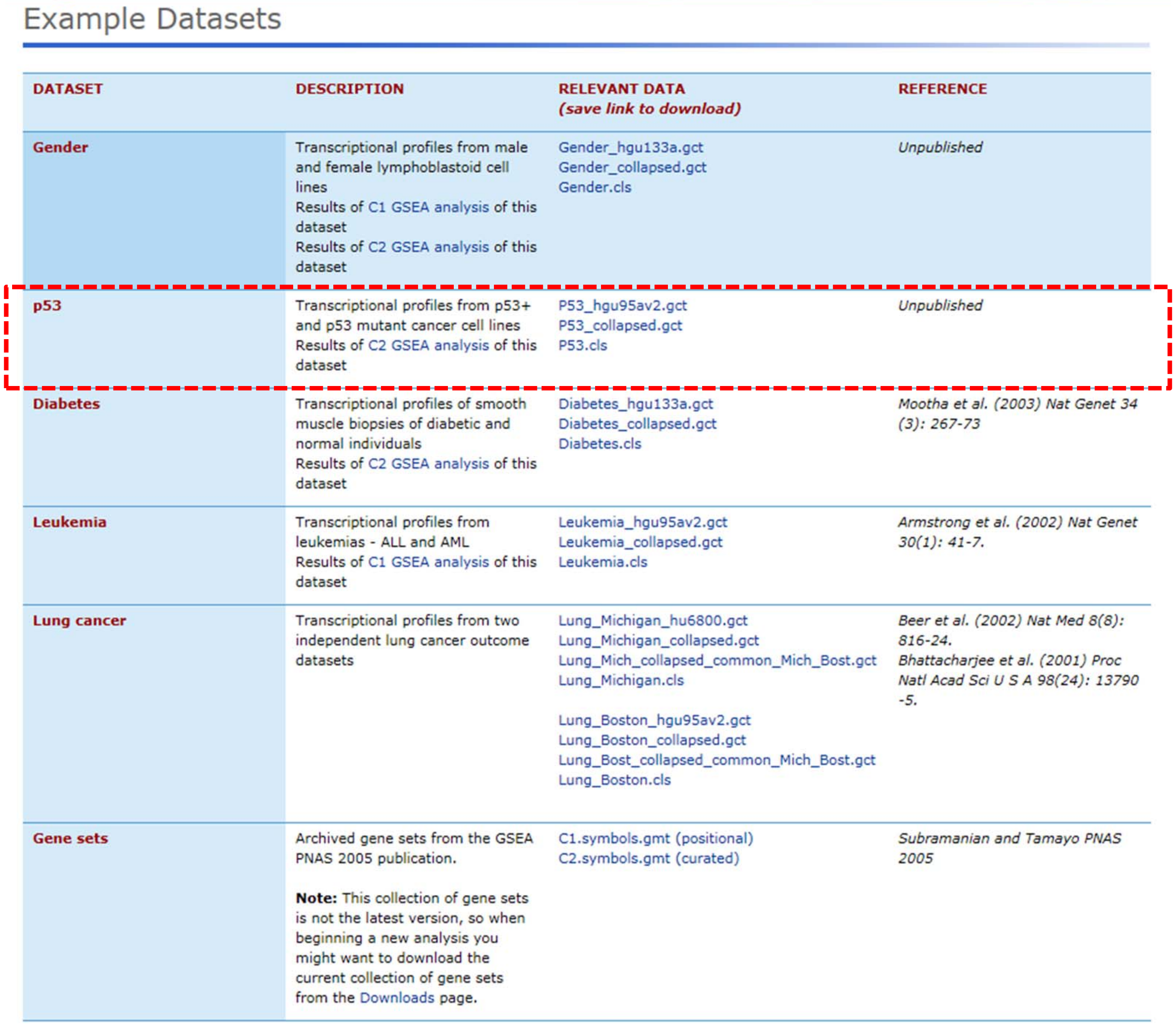
c2 **curated gene sets** from online pathway databases, publications in PubMed, and knowledge of domain experts.

c3 **motif gene sets** based on conserved *cis*-regulatory motifs from a comparative analysis of the human, mouse, rat and dog genomes.

c4 **computational gene sets** defined by expression neighborhoods centered on 380 cancer-associated genes.

c5 **GO gene sets** consist of genes annotated by the same GO terms.

Example Datasets



Example Datasets

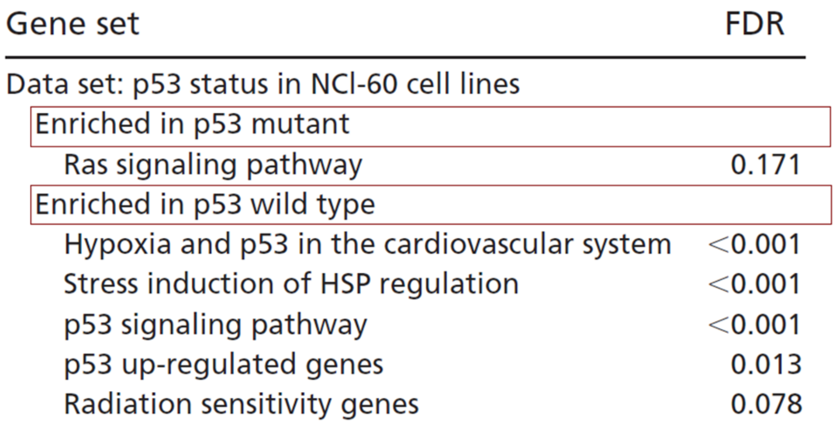
DATASET	DESCRIPTION	RELEVANT DATA *(save link to download)*	REFERENCE
Gender	Transcriptional profiles from male and female lymphoblastoid cell lines Results of C1 GSEA analysis of this dataset Results of C2 GSEA analysis of this dataset	Gender_hgu133a.gct Gender_collapsed.gct Gender.cls	*Unpublished*
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 55 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.

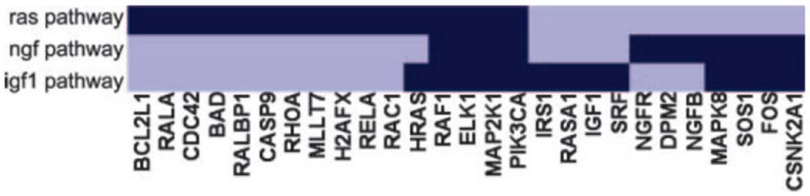


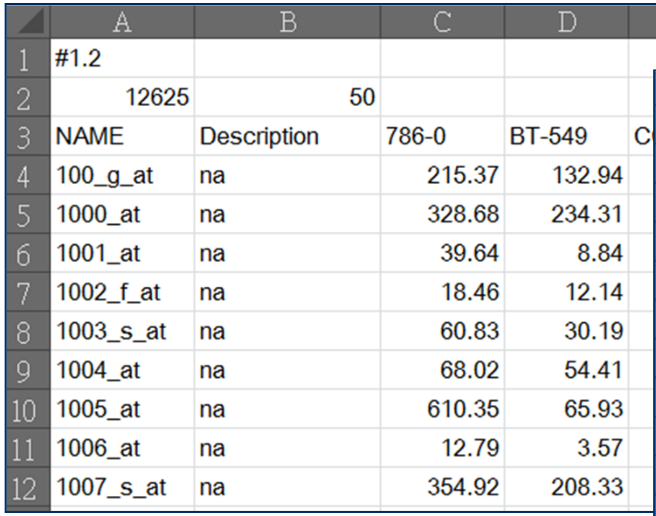
Fig. 3. Leading edge overlap for p53 study. This plot shows the *ras*, *ngf*, and *igf1* gene sets correlated with $P53^{-}$ clustered by their leading-edge subsets indicated in dark blue. A common subgroup of genes, apparent as a dark vertical stripe, consists of MAP2K1, PIK3CA, ELK1, and RAF1 and represents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

Input for GSEA (1)

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

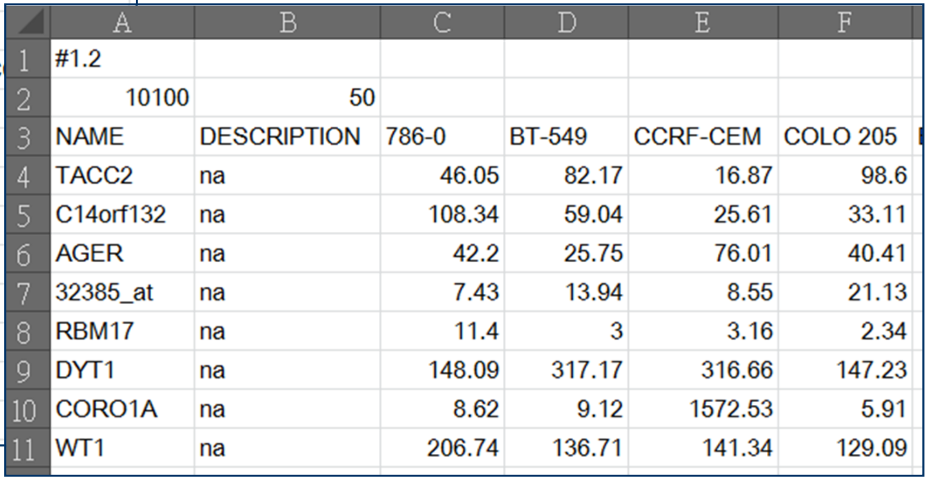
Data File	Content	Format	Source
Expression dataset	Contains features (genes or probes), samples, and an expression value for each feature in each sample. Expression data can come from any source (Affymetrix, Stanford cDNA, and so on).	res, gct, pcl, or txt	You create the file.
Phenotype labels	Contains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

P53_hgu95av2.gct



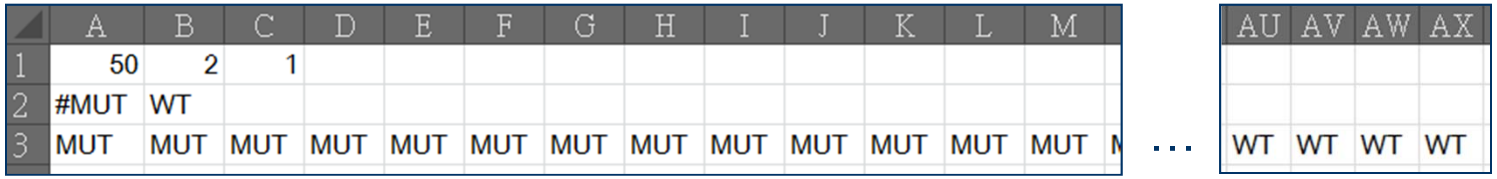
	A	B	C	D	E
1	#1.2				
2	12625	50			
3	NAME	Description	786-0	BT-549	C
4	100_g_at	na	215.37	132.94	
5	1000_at	na	328.68	234.31	
6	1001_at	na	39.64	8.84	
7	1002_f_at	na	18.46	12.14	
8	1003_s_at	na	60.83	30.19	
9	1004_at	na	68.02	54.41	
10	1005_at	na	610.35	65.93	
11	1006_at	na	12.79	3.57	
12	1007_s_at	na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

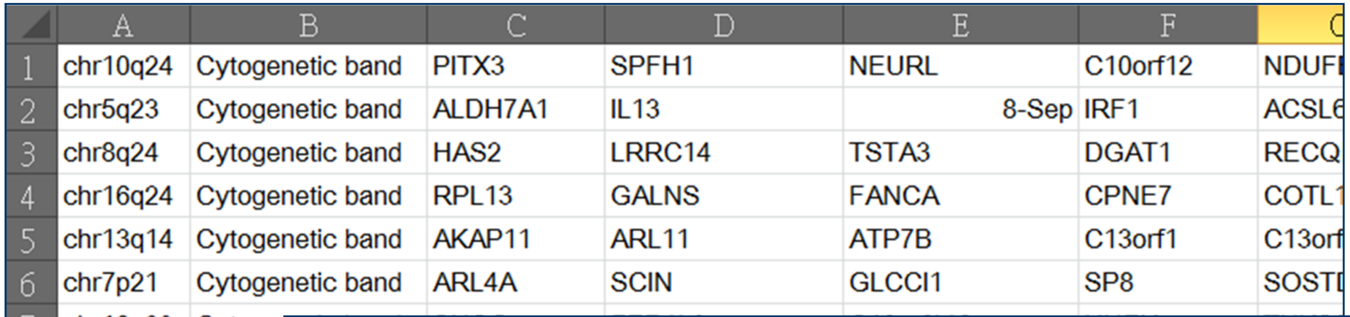


	A	B	C	D	E	F	G	H	I	J	K	L	M	…	AU	AV	AW	AX
1	50	2	1															
2	#MUT	WT																
3	MUT	MUT	MUT	MUT	MUT	MUT	MUT	MUT	MUT	MUT	MUT	MUT	MUT	…	WT	WT	WT	WT

Input for GSEA (2)

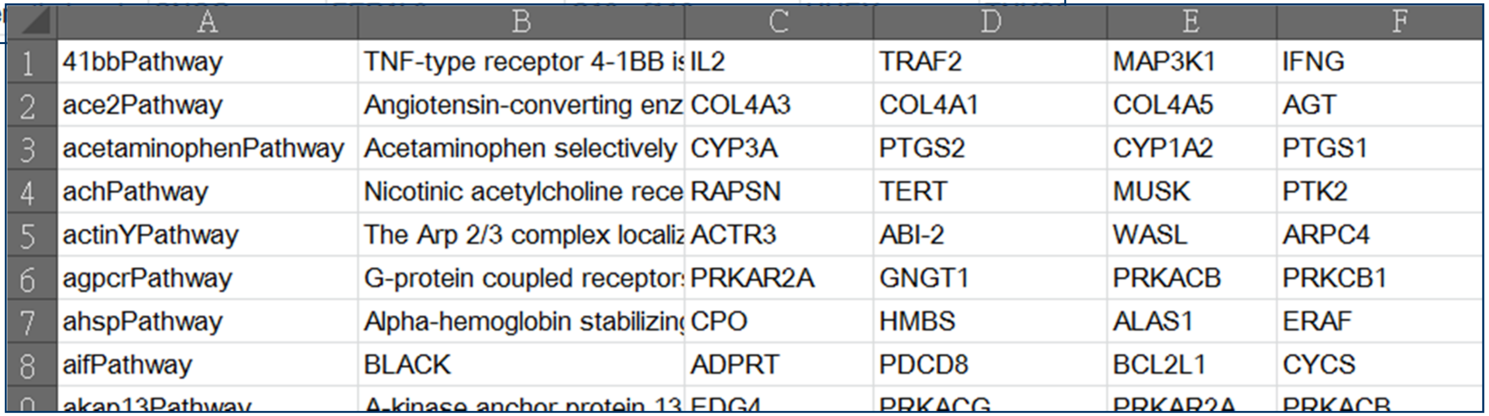
Data File	Content	Format	Source
Gene sets	Contains one or more gene sets. For each gene set, gives the gene set name and list of features (genes or probes) in that gene set.	gmx or gmt	You use the files on the Broad ftp site, export gene sets from the Molecular Signature Database (MSigDb) or create your 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



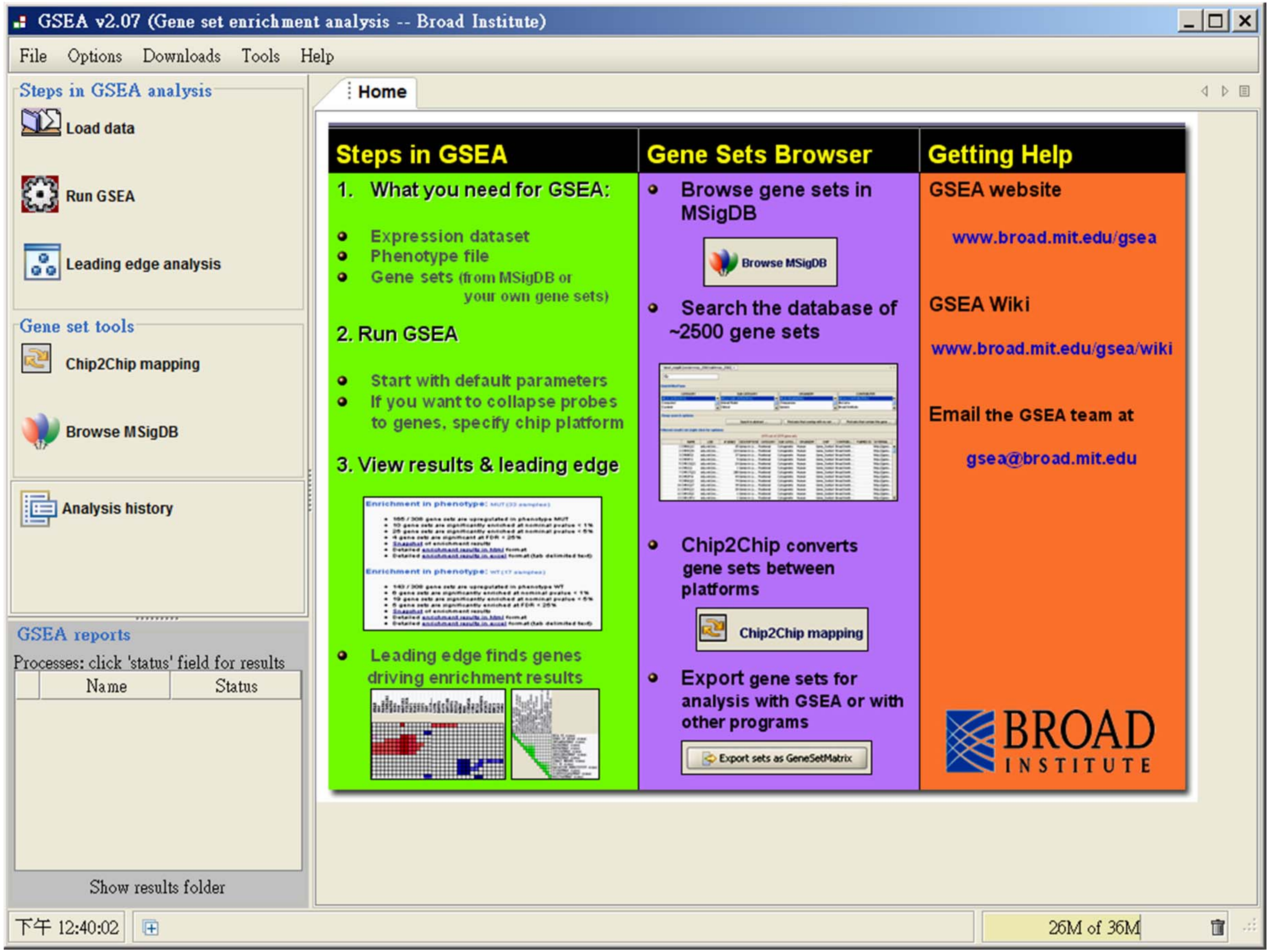
	A	B	C	D	E	F	G
1	chr10q24	Cytogenetic band	PITX3	SPFH1	NEURL	C10orf12	NDUF
2	chr5q23	Cytogenetic band	ALDH7A1	IL13	8-Sep	IRF1	ACSL6
3	chr8q24	Cytogenetic band	HAS2	LRRC14	TSTA3	DGAT1	RECQ
4	chr16q24	Cytogenetic band	RPL13	GALNS	FANCA	CPNE7	COTL1
5	chr13q14	Cytogenetic band	AKAP11	ARL11	ATP7B	C13orf1	C13orf
6	chr7p21	Cytogenetic band	ARL4A	SCIN	GLCCI1	SP8	SOST
7	chr10q23	Cytoge					

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



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

Enrichment in phenotype: MUT (33 samples)

- 165 / 308 gene sets are upregulated in phenotype MUT
- 10 gene sets are significantly enriched at nominal pvalue < 1%
- 25 gene sets are significantly enriched at nominal pvalue < 5%
- 4 gene sets are significant at FDR < 25%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in excel format (tab delimited text)

Enrichment in phenotype: WT (17 samples)

- 143 / 308 gene sets are upregulated in phenotype WT
- 6 gene sets are significantly enriched at nominal pvalue < 1%
- 19 gene sets are significantly enriched at nominal pvalue < 5%
- 5 gene sets are significantly enriched at FDR < 25%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in excel format (tab delimited text)

- Leading edge finds genes driving enrichment results

Gene Sets Browser

- Browse gene sets in MSigDB

- Search the database of ~2500 gene sets

- Chip2Chip converts gene sets between platforms

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

Getting Help

GSEA website

www.broad.mit.edu/gsea

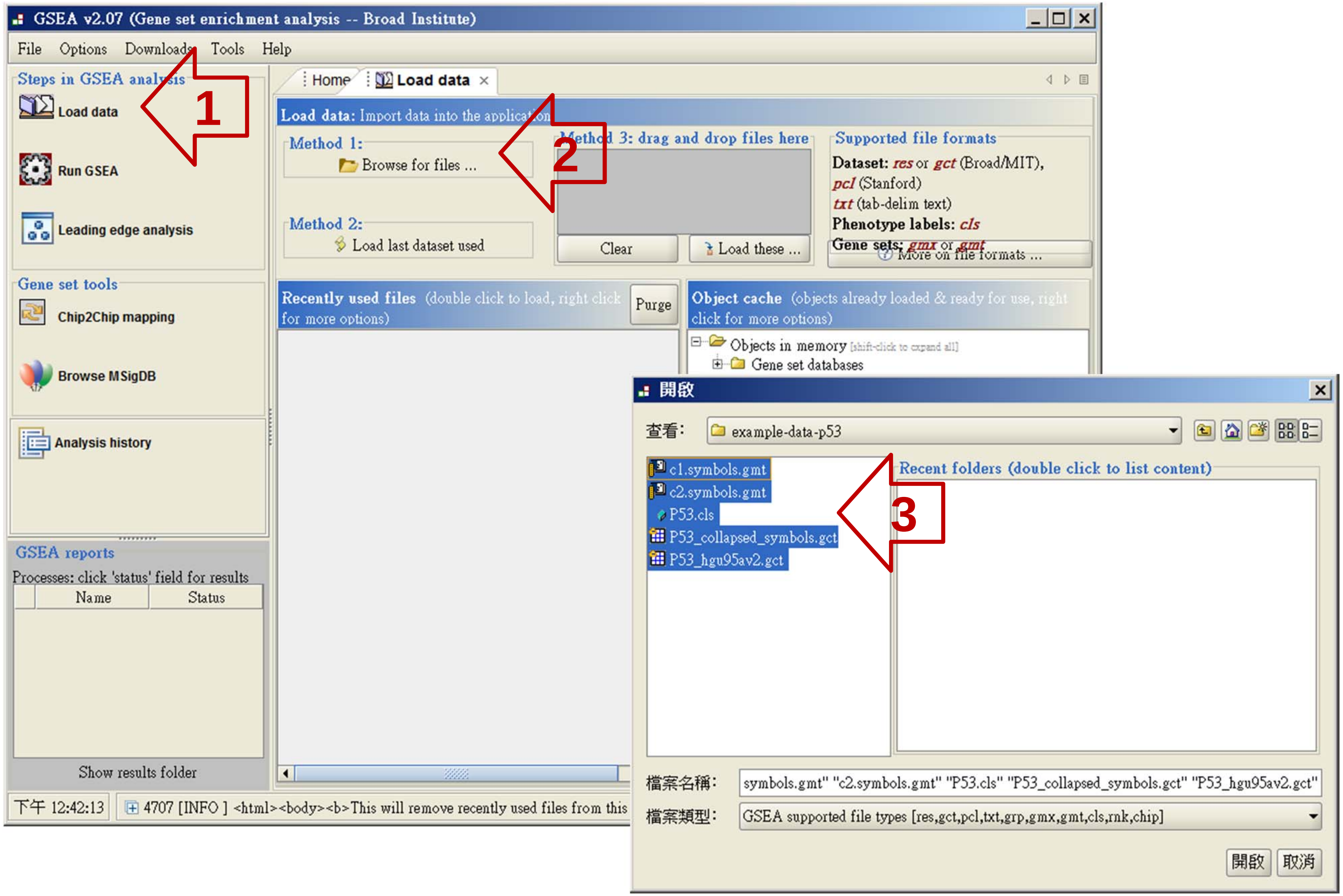
GSEA Wiki

www.broad.mit.edu/gsea/wiki

Email the GSEA team at

gsea@broad.mit.edu

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：
Recent folders (double click to list content)
檔案名稱：
檔案類型：
開啟
取消

下午 12:42:13

4707 [INFO] <html><body><b>This will remove recently used files from this

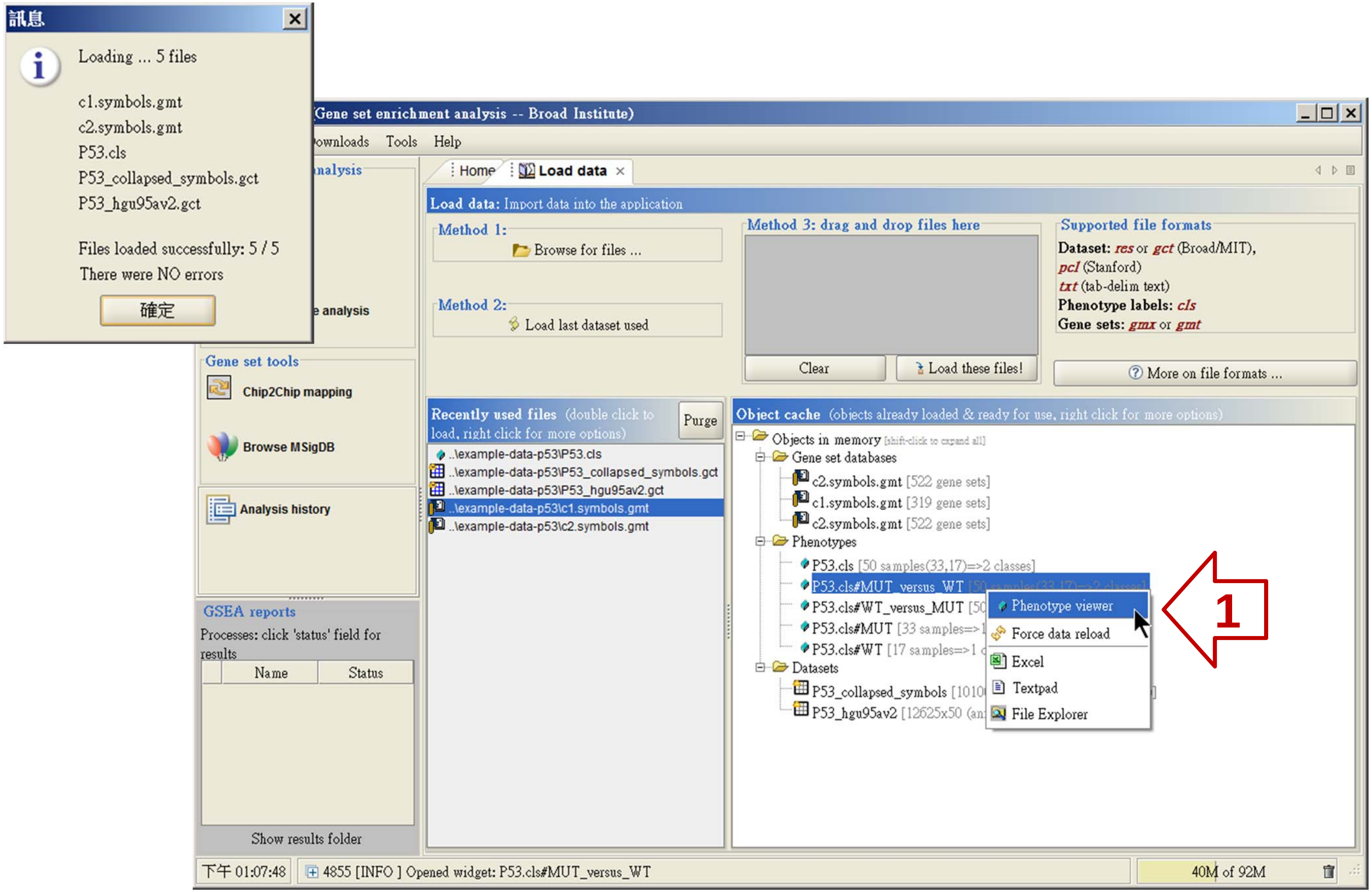
查看： example-data-p53

- c1.symbols.gmt
- c2.symbols.gmt
- P53.cls
- P53_collapsed_symbols.gct
- P53_hgu95av2.gct

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]

Explore Inputs



訊息

Loading ... 5 files

c1.symbols.gmt
c2.symbols.gmt
P53.cls
P53_collapsed_symbols.gct
P53_hgu95av2.gct

Files loaded successfully: 5 / 5
There were NO errors

Gene set enrichment analysis -- Broad Institute)

Load data: Import data into the application

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

Recently used files (double click to load, right click for more options)

- ..\example-data-p53\P53.cls
- ..\example-data-p53\P53_collapsed_symbols.gct
- ..\example-data-p53\P53_hgu95av2.gct
- ..\example-data-p53\c1.symbols.gmt
- ..\example-data-p53\c2.symbols.gmt

Object cache (objects already loaded & ready for use, right click for more options)

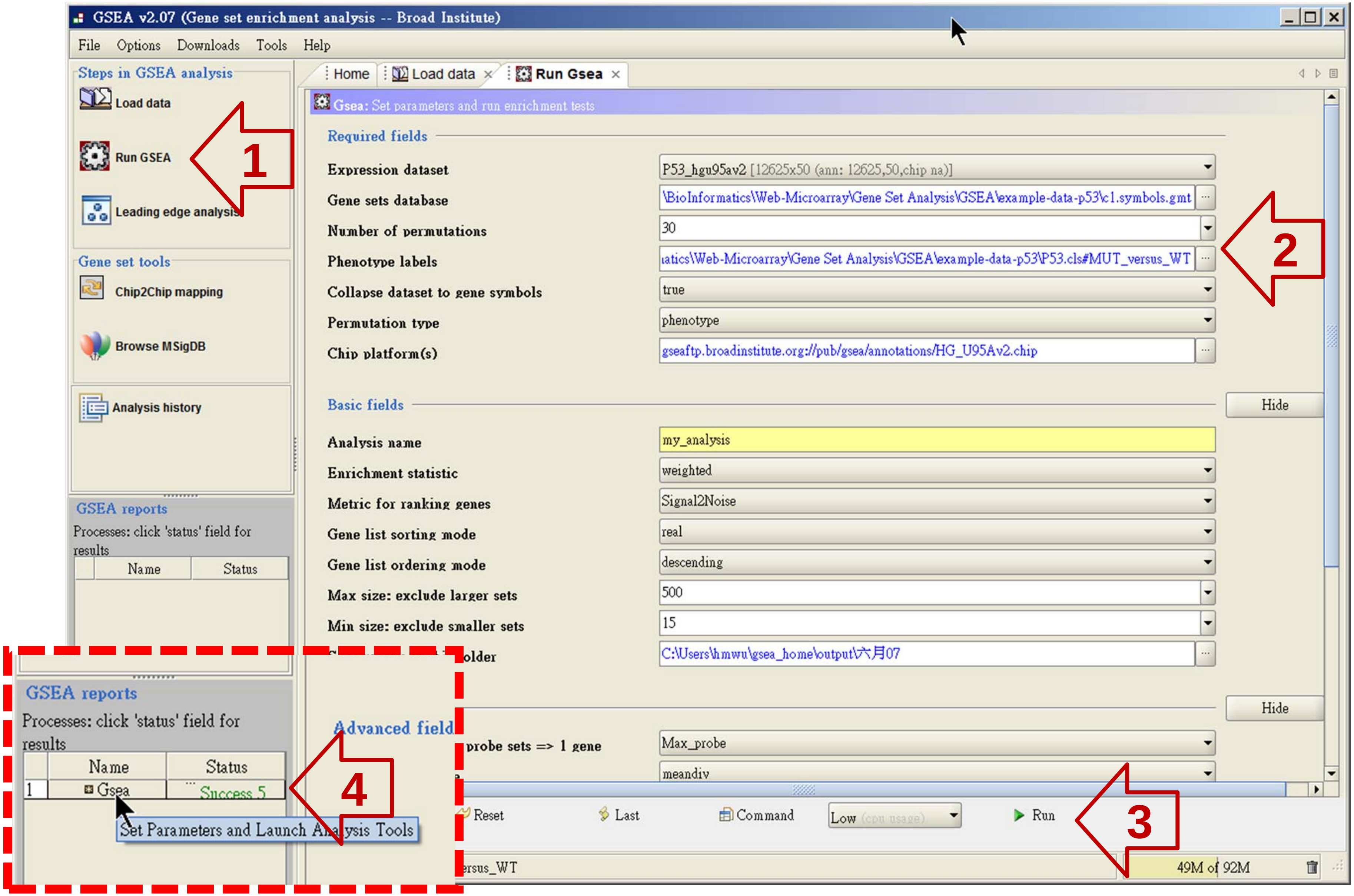
- Objects in memory [shift-click to expand all]
  - Gene set databases
    - c2.symbols.gmt [522 gene sets]
    - c1.symbols.gmt [319 gene sets]
    - c2.symbols.gmt [522 gene sets]
  - Phenotypes
    - P53.cls [50 samples(33,17)=>2 classes]
    - P53.cls#MUT_versus_WT [50 samples(33,17)=>2 classes]
    - P53.cls#WT_versus_MUT [50
    - P53.cls#MUT [33 samples=>1
    - P53.cls#WT [17 samples=>1 c
  - Datasets
    - P53_collapsed_symbols [1010
    - P53_hgu95av2 [12625x50 (an

GSEA reports
Processes: click 'status' field for results

Name	Status

下午 01:07:48 4855 [INFO] Opened widget: P53.cls#MUT_versus_WT 40M of 92M

Run GSEA



GSEA v2.07 (Gene set enrichment analysis -- Broad Institute)

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_U95Av2.chip

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

older: C:\Users\hmwu\gsea_home\output\六月07

Advanced field

probe sets => 1 gene: Max_probe

meandiv

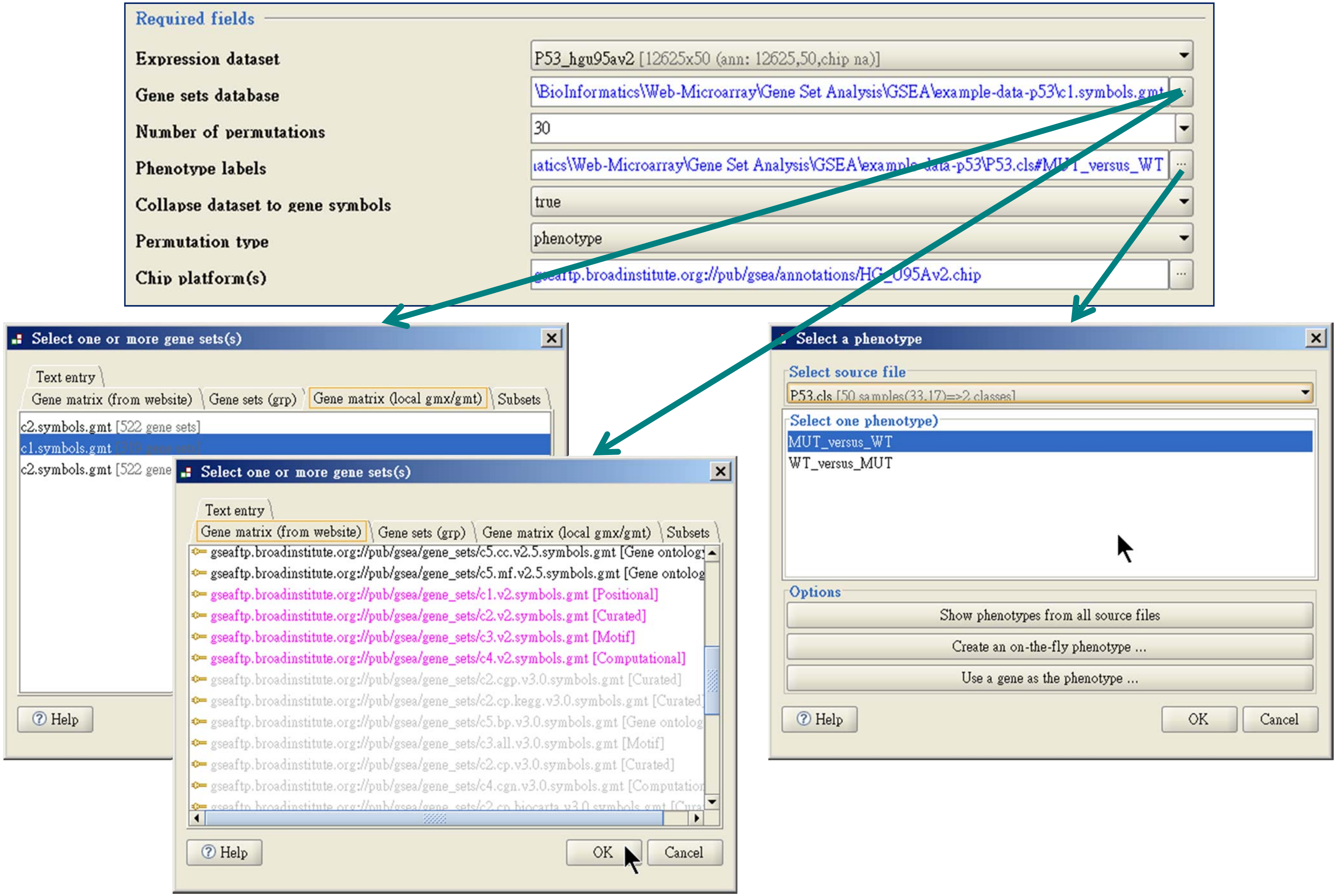
GSEA reports

Processes: click 'status' field for results

	Name	Status
1	Gsea	Success 5

Set Parameters and Launch Analysis Tools

Required Fields



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_U95Av2.chip

Select one or more gene sets(s)

c2.symbols.gmt [522 gene sets]
c1.symbols.gmt [319 gene sets]
c2.symbols.gmt [522 gene

Select one or more gene sets(s)

gseaftp.broadinstitute.org://pub/gsea/gene_sets/c5.cc.v2.5.symbols.gmt [Gene ontology
gseaftp.broadinstitute.org://pub/gsea/gene_sets/c5.mf.v2.5.symbols.gmt [Gene ontolog
gseaftp.broadinstitute.org://pub/gsea/gene_sets/c1.v2.symbols.gmt [Positional]
gseaftp.broadinstitute.org://pub/gsea/gene_sets/c2.v2.symbols.gmt [Curated]
gseaftp.broadinstitute.org://pub/gsea/gene_sets/c3.v2.symbols.gmt [Motif]
gseaftp.broadinstitute.org://pub/gsea/gene_sets/c4.v2.symbols.gmt [Computational]
gseaftp.broadinstitute.org://pub/gsea/gene_sets/c2.cgp.v3.0.symbols.gmt [Curated]
gseaftp.broadinstitute.org://pub/gsea/gene_sets/c2.cp.kegg.v3.0.symbols.gmt [Curated]
gseaftp.broadinstitute.org://pub/gsea/gene_sets/c5.bp.v3.0.symbols.gmt [Gene ontolog
gseaftp.broadinstitute.org://pub/gsea/gene_sets/c3.all.v3.0.symbols.gmt [Motif]
gseaftp.broadinstitute.org://pub/gsea/gene_sets/c2.cp.v3.0.symbols.gmt [Curated]
gseaftp.broadinstitute.org://pub/gsea/gene_sets/c4.cgn.v3.0.symbols.gmt [Computation
gseaftp.broadinstitute.org://pub/gsea/gene_sets/c2.cp.biocarta.v3.0.symbols.gmt [Cura

Select a phenotype

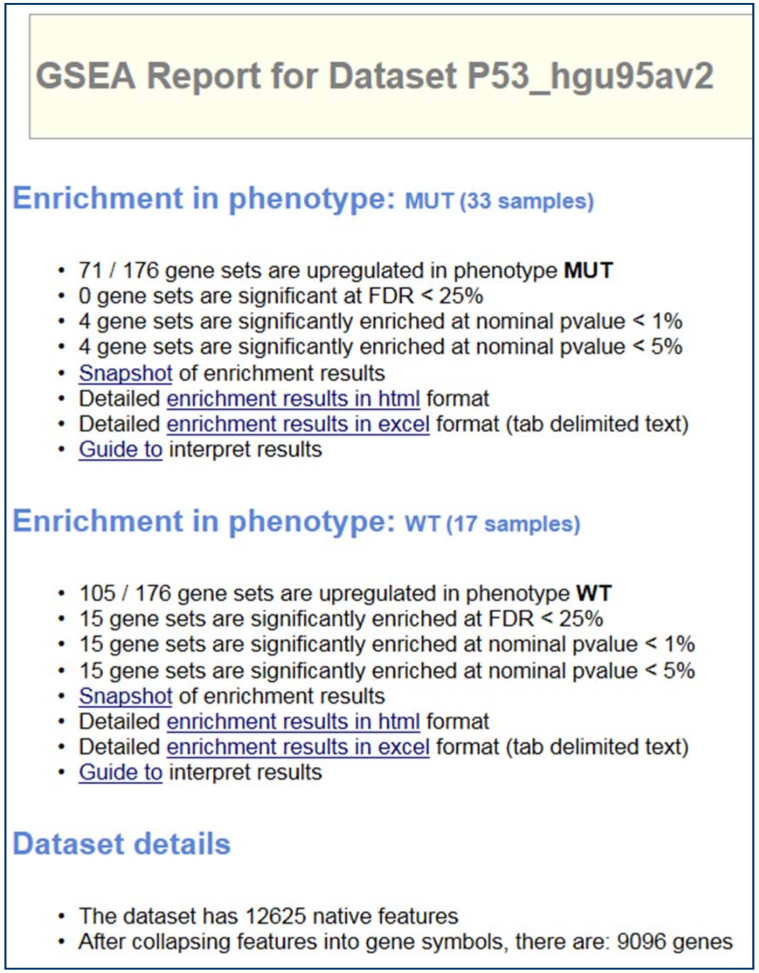
Select source file: P53.cls [50 samples(33,17)=>2 classes]

Select one phenotype)

MUT_versus_WT
WT_versus_MUT

Options
Show phenotypes from all source files
Create an on-the-fly phenotype ...
Use a gene as the phenotype ...
Help
OK
Cancel

Report



GSEA Report for Dataset P53_hgu95av2

Enrichment in phenotype: MUT (33 samples)

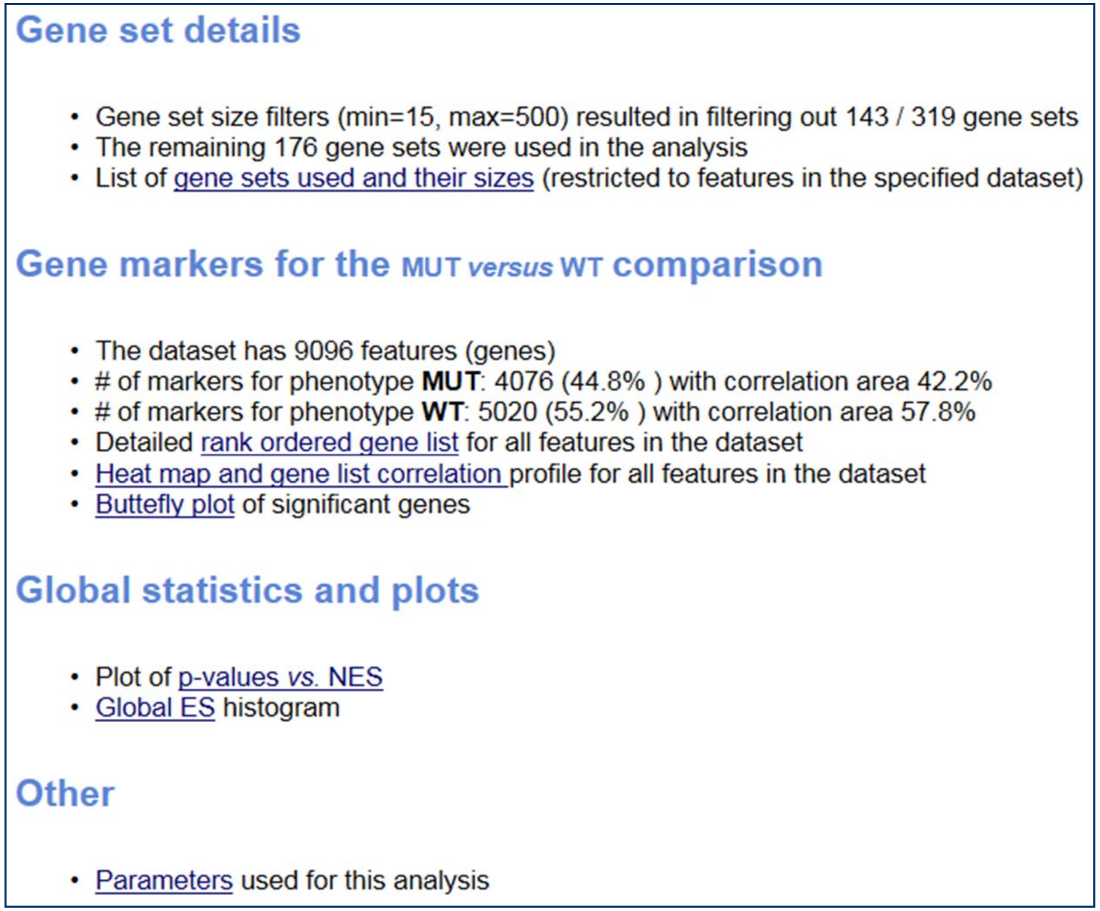
- 71 / 176 gene sets are upregulated in phenotype **MUT**
- 0 gene sets are significant at FDR < 25%
- 4 gene sets are significantly enriched at nominal pvalue < 1%
- 4 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- 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

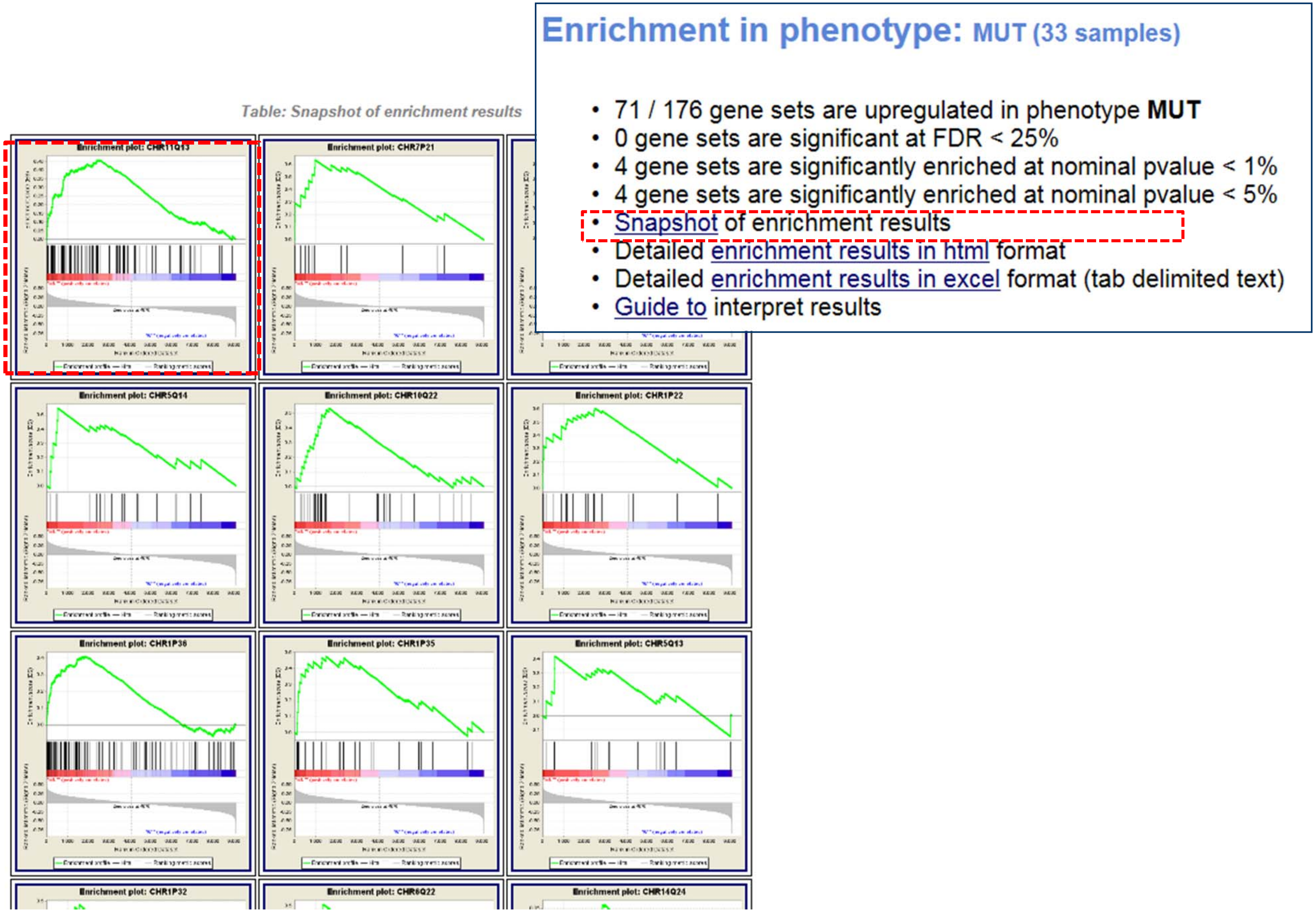


Table: Snapshot of enrichment results

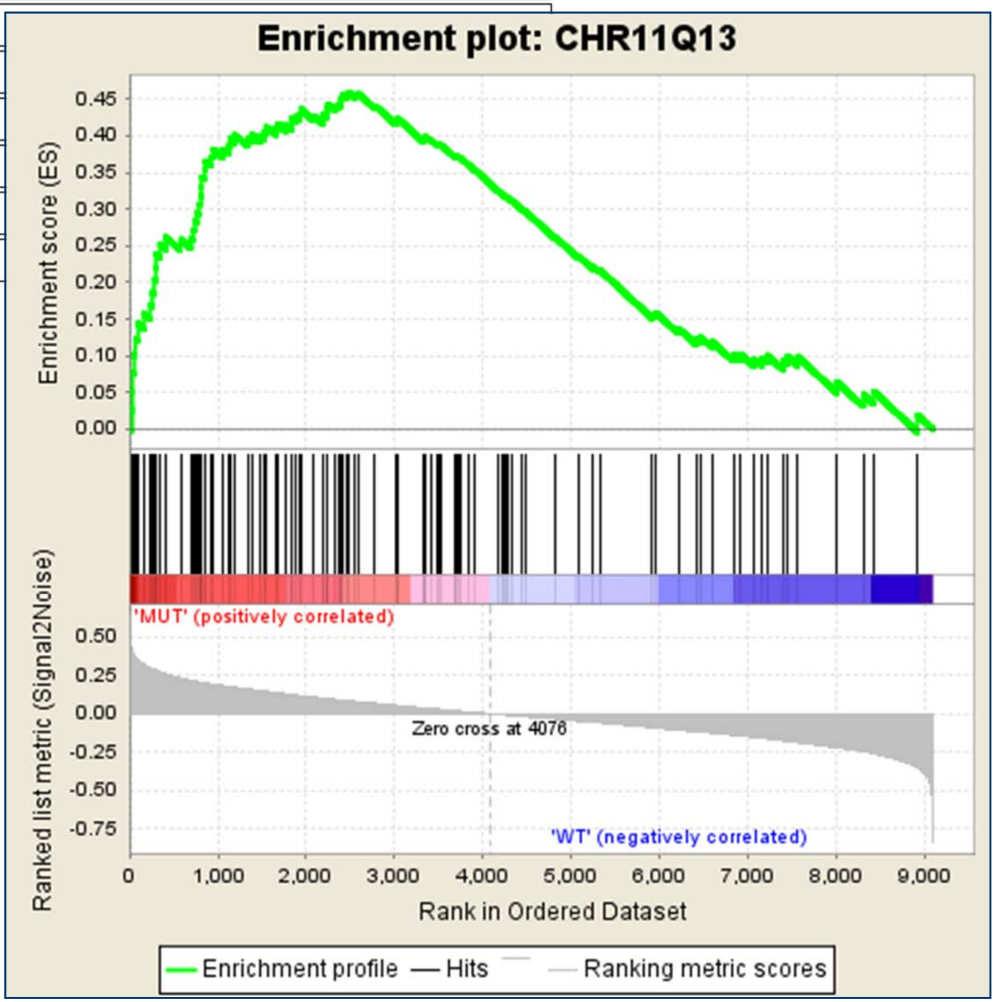
Enrichment in phenotype: MUT (33 samples)

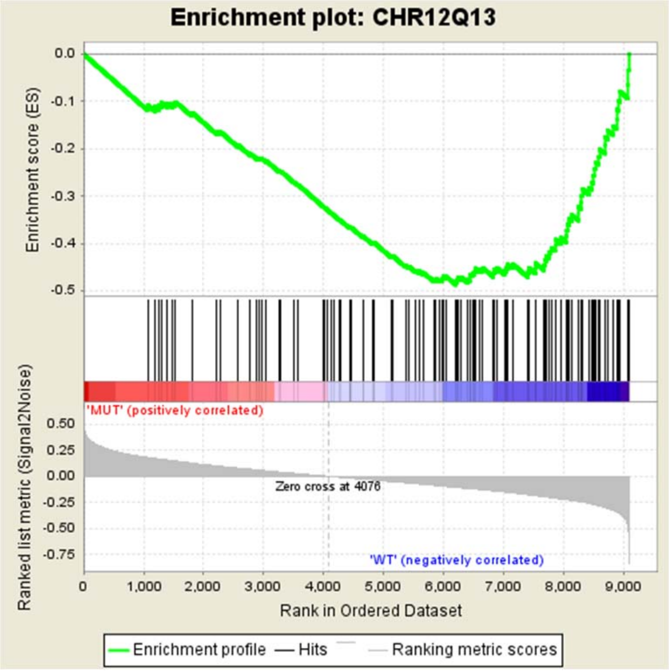
- 71 / 176 gene sets are upregulated in phenotype **MUT**
- 0 gene sets are significant at FDR < 25%
- 4 gene sets are significantly enriched at nominal pvalue < 1%
- 4 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in excel format (tab delimited text)
- Guide to interpret results

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

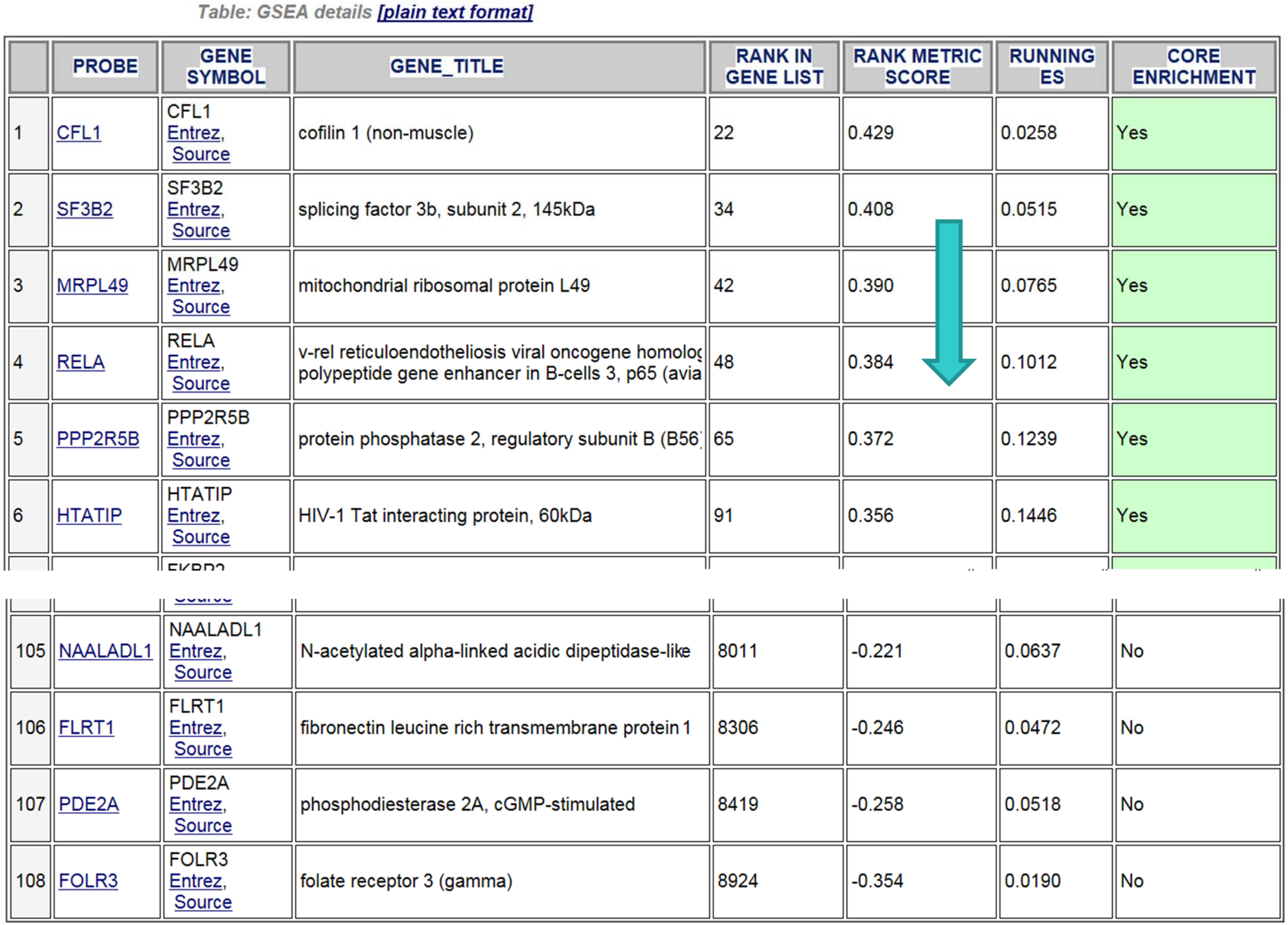
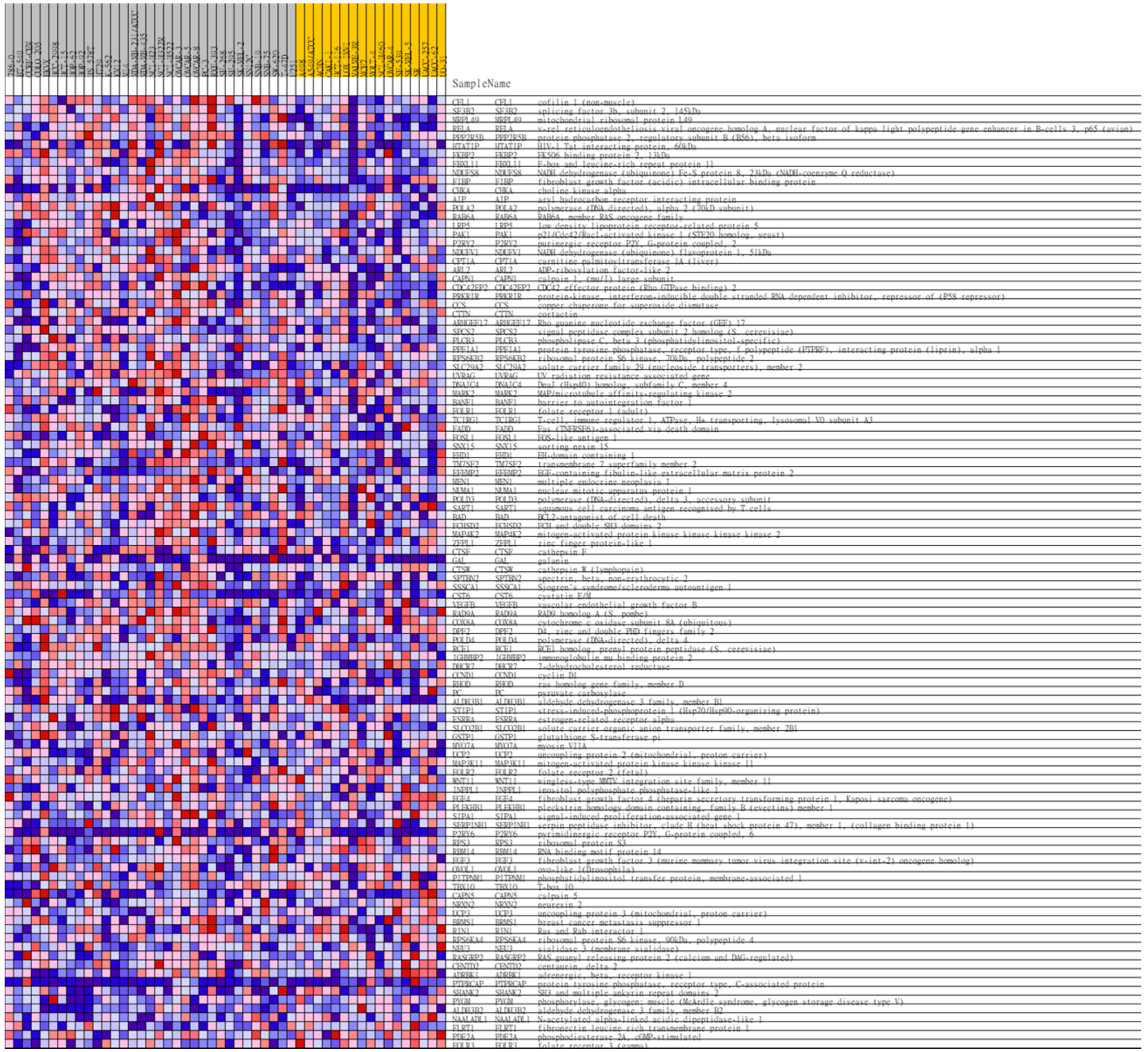


Table: GSEA details [plain text format]

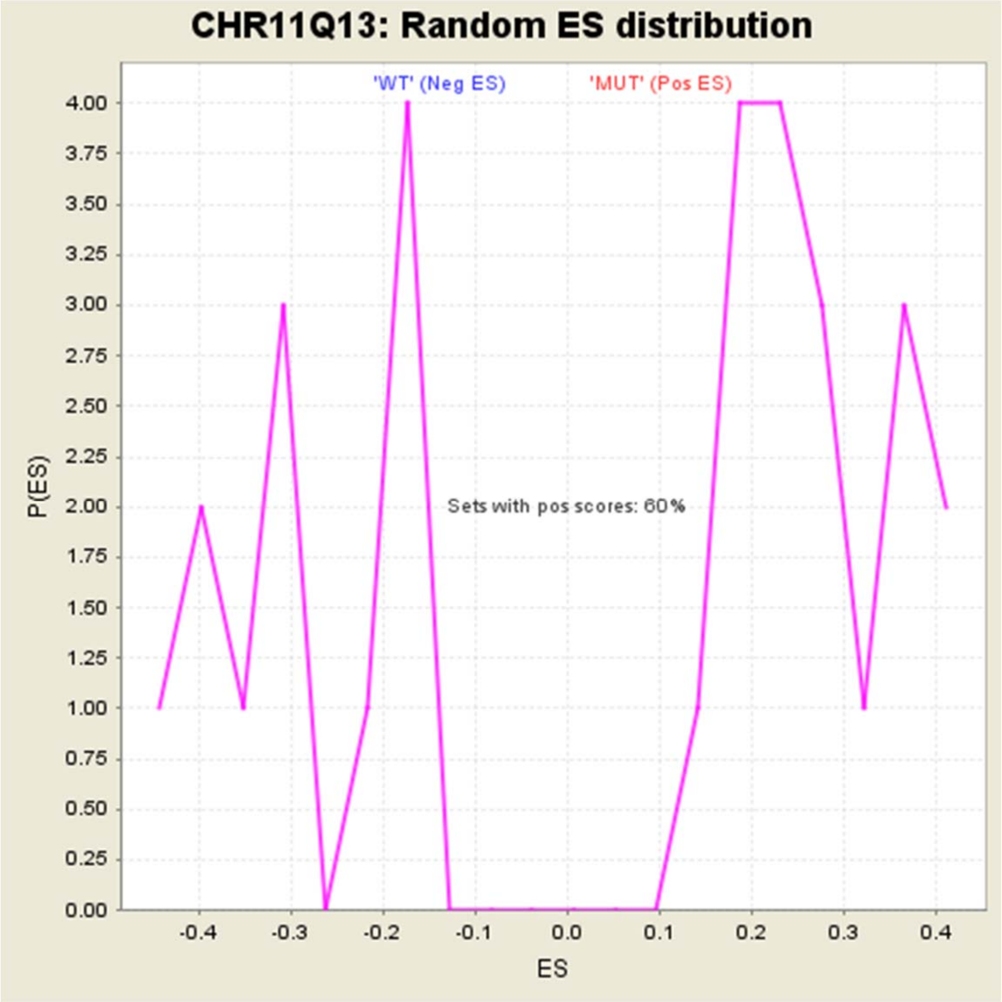
	PROBE	GENE SYMBOL	GENE_TITLE	RANK IN GENE LIST	RANK METRIC SCORE	RUNNING ES	CORE ENRICHMENT
1	CFL1	CFL1 Entrez, Source	cofilin 1 (non-muscle)	22	0.429	0.0258	Yes
2	SF3B2	SF3B2 Entrez, Source	splicing factor 3b, subunit 2, 145kDa	34	0.408	0.0515	Yes
3	MRPL49	MRPL49 Entrez, Source	mitochondrial ribosomal 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
		FKBP2					
		Source					
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



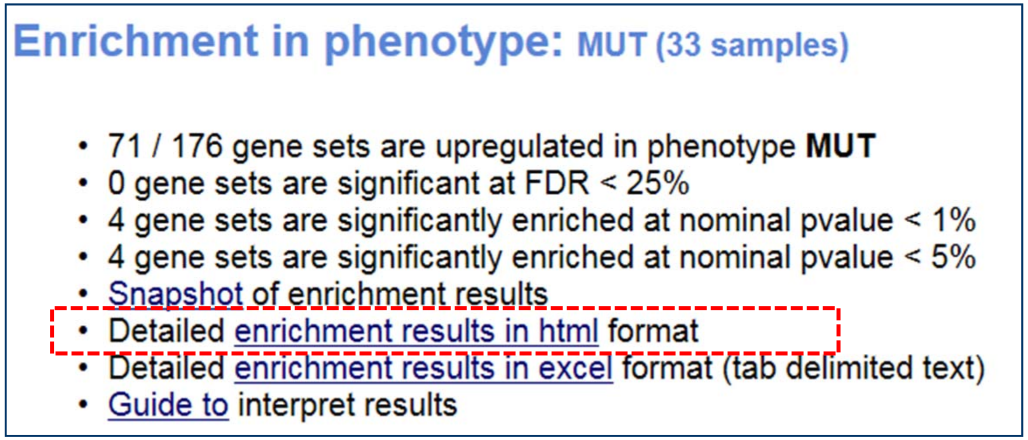
CHR11Q13 : Blue-Pink O' Gram in the Space of the Analyzed GeneSet

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

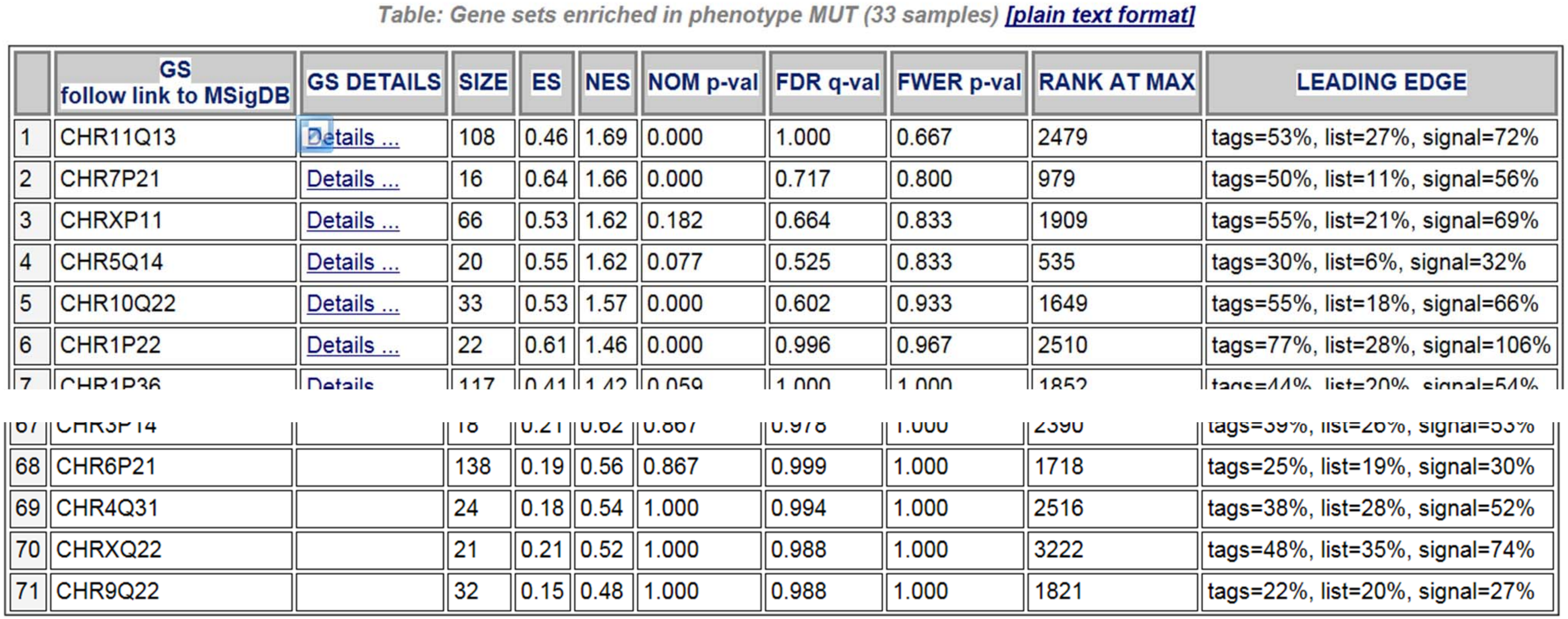
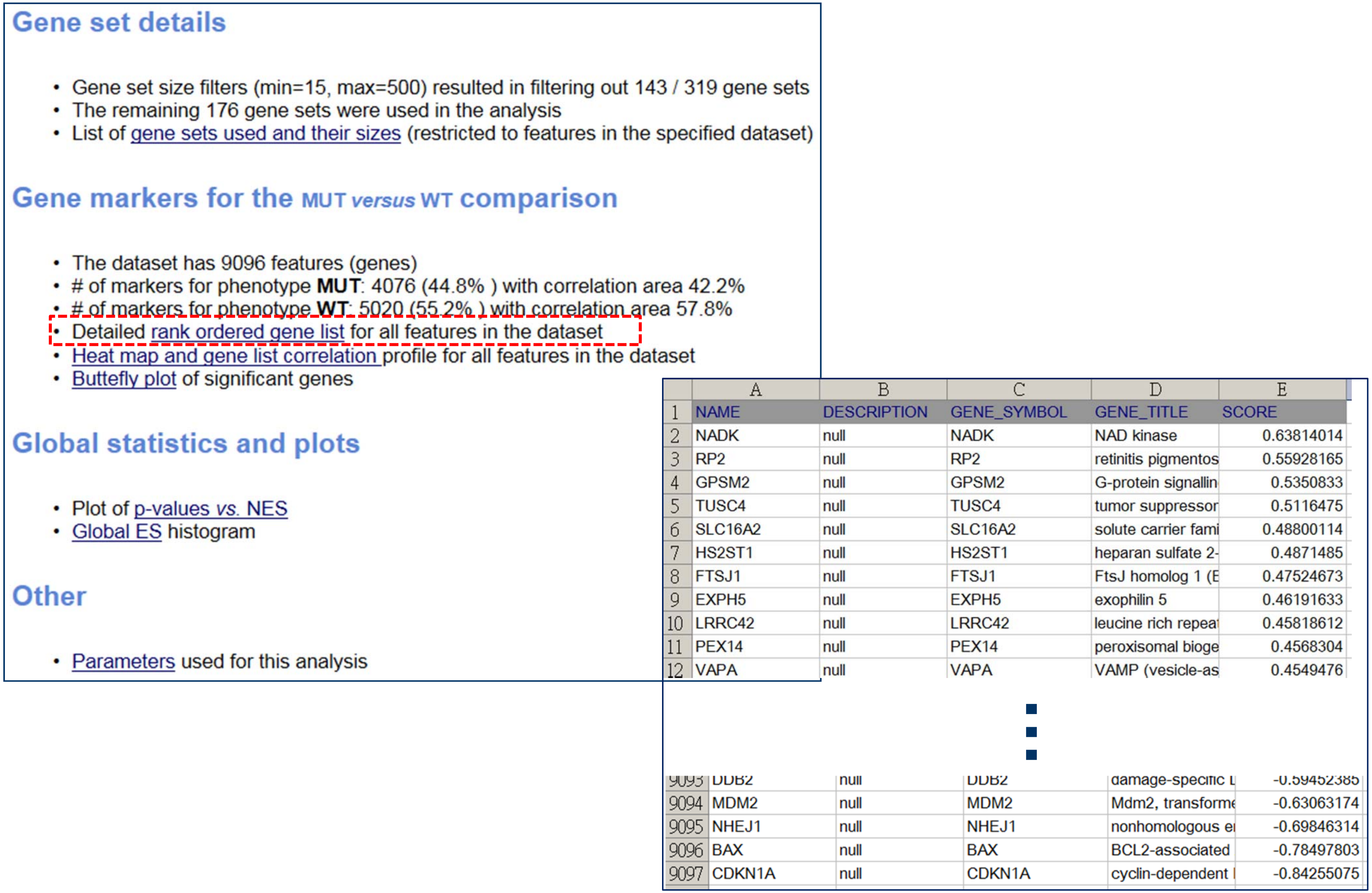


Table: Gene sets enriched in phenotype MUT (33 samples) ***[plain text format]***

	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

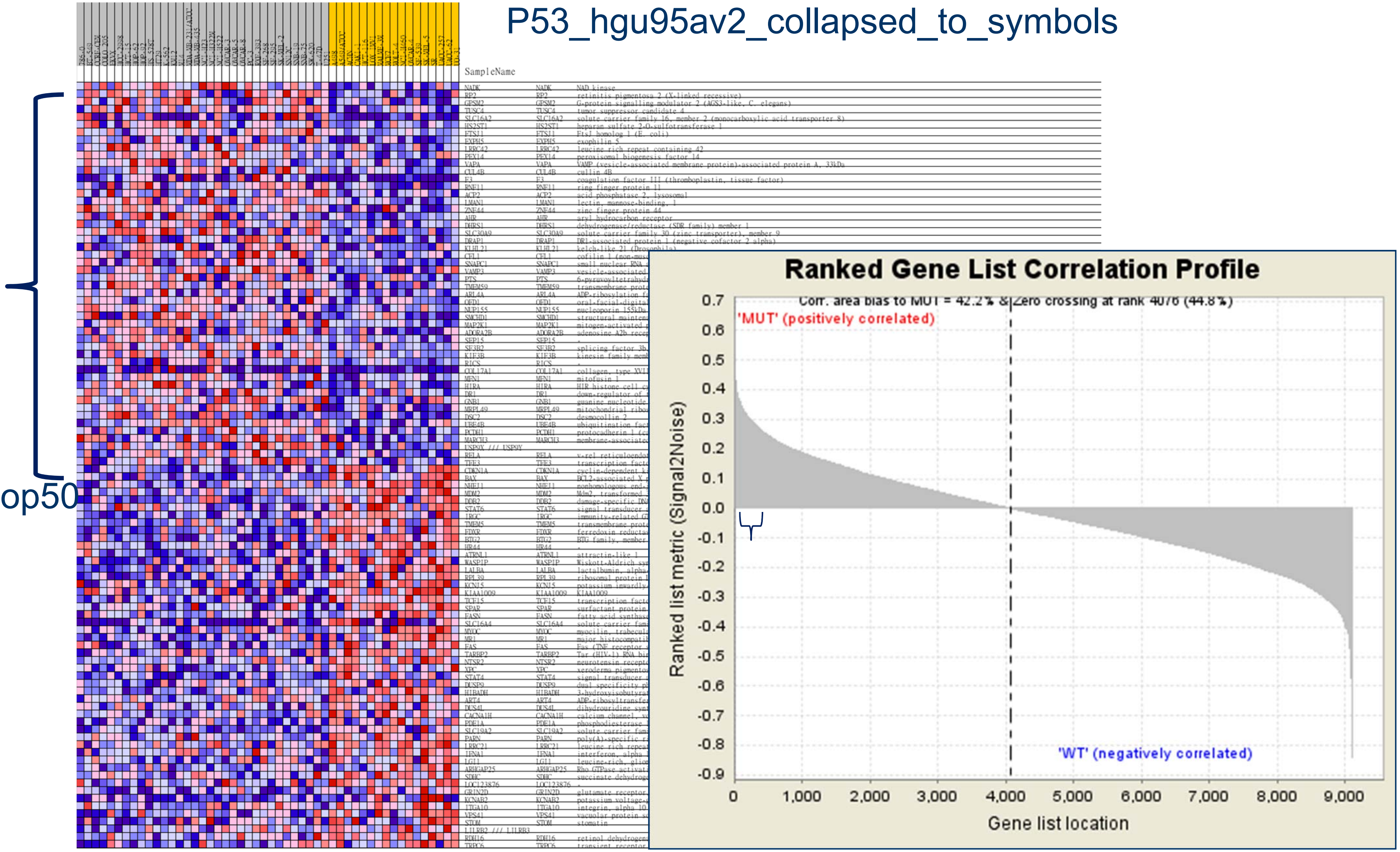
Other

- 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

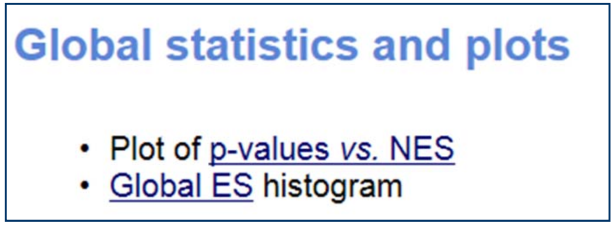
Heat Map and Gene Correlation

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



SampleName		
NADK	NADK	NAD kinase
RP2	RP2	retinitis pigmentosa 2 (X-linked recessive)
GPSM2	GPSM2	G-protein signalling modulator 2 (AGS3-like, C. elegans)
TUSC4	TUSC4	tumor suppressor candidate 4
SLC16A2	SLC16A2	solute carrier family 16, member 2 (monocarboxylic acid transporter 8)
HS2ST1	HS2ST1	heparan sulfate 2-O-sulfotransferase 1
FTSJ1	FTSJ1	FtsJ homolog 1 (E. coli)
EXPH5	EXPH5	exophilin 5
LRRC42	LRRC42	leucine rich repeat containing 42
PEX14	PEX14	peroxisomal biogenesis factor 14
VAPA	VAPA	VAMP (vesicle-associated membrane protein)-associated protein A, 33kDa
CUL4B	CUL4B	cullin 4B
F3	F3	coagulation factor III (thromboplastin, tissue factor)
RNF11	RNF11	ring finger protein 11
ACP2	ACP2	acid phosphatase 2, lysosomal
LMAN1	LMAN1	lectin, mannose-binding, 1
ZNF44	ZNF44	zinc finger protein 44
AHR	AHR	aryl hydrocarbon receptor
DHRS1	DHRS1	dehydrogenase/reductase (SDR family) member 1
SLC30A9	SLC30A9	solute carrier family 30 (zinc transporter), member 9
DRAP1	DRAP1	DR1-associated protein 1 (negative cofactor 2 alpha)
KLHL21	KLHL21	kelch-like 21 (Drosophila)
CFL1	CFL1	cofilin 1 (non-musc
SNAPC1	SNAPC1	small nuclear RNA
VAMP3	VAMP3	vesicle-associated
PTS	PTS	6-pyruvoyltetrahyd
TMEM59	TMEM59	transmembrane prot
ARL4A	ARL4A	ADP-ribosylation f
OFD1	OFD1	oral-facial-digital
NUP155	NUP155	nucleoporin 155kDa
SMCHD1	SMCHD1	structural mainten
MAP2K1	MAP2K1	mitogen-activated
ADORA2B	ADORA2B	adenosine A2b rece
SEP15	SEP15	-
SF3B2	SF3B2	splicing factor 3b,
KIF3B	KIF3B	kinesin family mem
RICS	RICS	-
COL17A1	COL17A1	collagen, type XVII
MFN1	MFN1	mitofusin 1
HIRA	HIRA	HIR histone cell cy
DR1	DR1	down-regulator of
GNB1	GNB1	guanine nucleotide
MRPL49	MRPL49	mitochondrial ribos
DSC2	DSC2	desmocollin 2
UBE4B	UBE4B	ubiquitination fact
PCDH1	PCDH1	protocadherin 1 (c
MARCH3	MARCH3	membrane-associated
USP9X /// USP9Y		
RELA	RELA	v-rel reticuloendo
TFE3	TFE3	transcription fact
CDKN1A	CDKN1A	cyclin-dependent k
BAX	BAX	BCL2-associated X
NHEJ1	NHEJ1	nonhomologous end-
MDM2	MDM2	Mdm2, transformed
DDB2	DDB2	damage-specific DN
STAT6	STAT6	signal transducer
IRGC	IRGC	immunity-related G
TMEM5	TMEM5	transmembrane prot
FDXR	FDXR	ferredoxin reducta
BTG2	BTG2	BTG family, member
HR44	HR44	-
ATRNL1	ATRNL1	attractin-like 1
WASPIP	WASPIP	Wiskott-Aldrich sy
LALBA	LALBA	lactalbumin, alpha-
RPL39	RPL39	ribosomal protein L
KCNJ15	KCNJ15	potassium inwardly-
KIAA1009	KIAA1009	KIAA1009
TCF15	TCF15	transcription fact
SPAR	SPAR	surfactant protein
FASN	FASN	fatty acid synthas
SLC16A4	SLC16A4	solute carrier fam
MYOC	MYOC	myocilin, trabecul
MR1	MR1	major histocompati
FAS	FAS	Fas (TNF receptor
TARBP2	TARBP2	Tar (HIV-1) RNA bi
NTSR2	NTSR2	neurotensin recept
XPC	XPC	xeroderma pigmentos
STAT4	STAT4	signal transducer
DUSP9	DUSP9	dual specificity p
HIBADH	HIBADH	3-hydroxyisobutyra
ART4	ART4	ADP-ribosyltransfe
DUS4L	DUS4L	dihydrouridine syn
CACNA1H	CACNA1H	calcium channel, v
PDE1A	PDE1A	phosphodiesterase
SLC19A2	SLC19A2	solute carrier fam
PARN	PARN	poly(A)-specific r
LRRC21	LRRC21	leucine rich repea
IFNA1	IFNA1	interferon, alpha
LGI1	LGI1	leucine-rich, glio
ARHGAP25	ARHGAP25	Rho GTPase activat
SDHC	SDHC	succinate dehydrog
LOC123876	LOC123876	-
GRIN2D	GRIN2D	glutamate receptor
KCNAB2	KCNAB2	potassium voltage-
ITGA10	ITGA10	integrin, alpha 10
VPS41	VPS41	vacuolar protein s
STOM	STOM	stomatin
LILRB2 /// LILRB3		
RDH16	RDH16	retinol dehydrogen
TRPC6	TRPC6	transient receptor

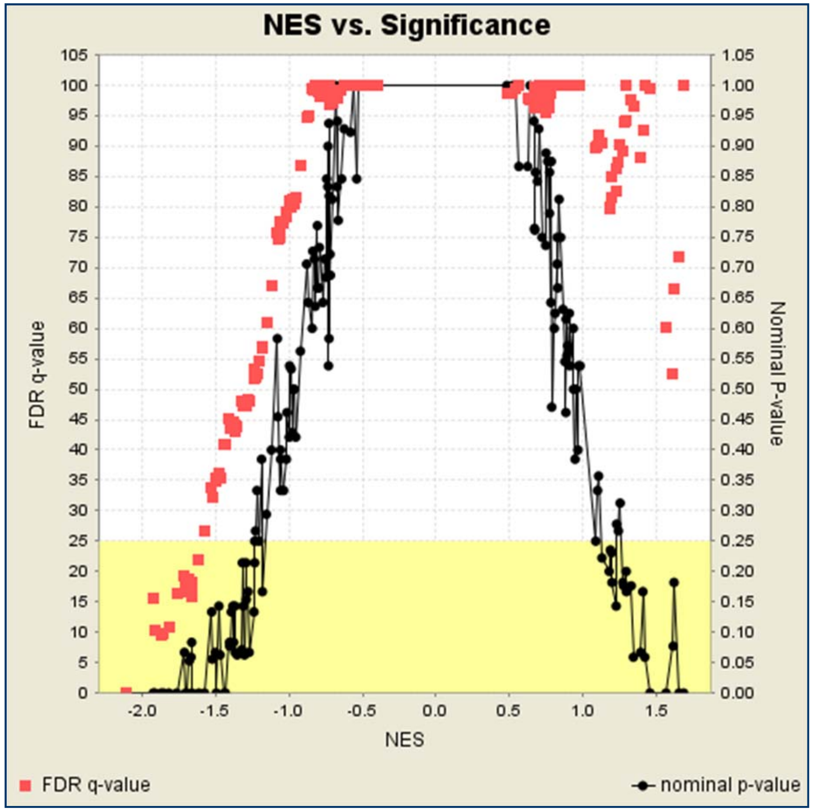
Global Statistics and Plots



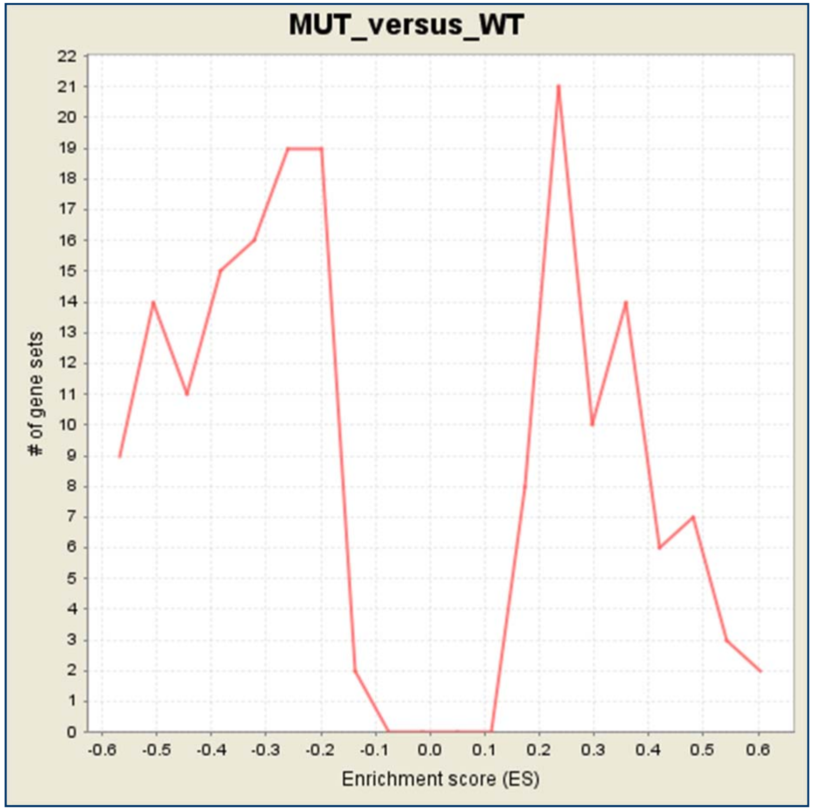
Global statistics and plots

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

Plot of p-values vs. NES

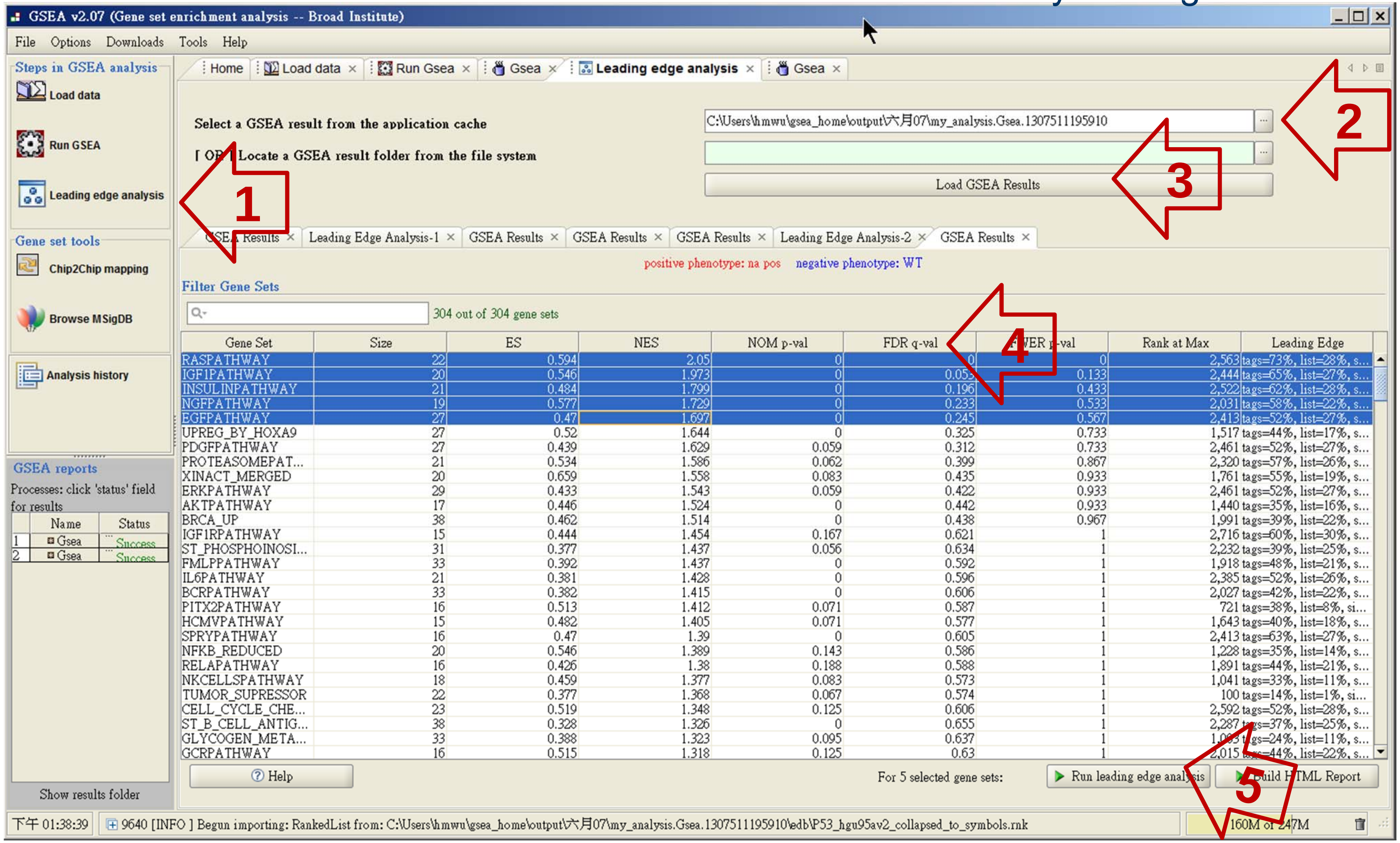


Global ES histogram



Running the Leading Edge Analysis

c2.symbols.gmt



Select a GSEA result from the application cache: C:\Users\hmwu\gsea_home\output\六月07\my_analysis.Gsea.1307511195910

[OR] Locate a GSEA result folder from the file system

positive phenotype: na pos negative phenotype: WT

Filter Gene Sets

304 out of 304 gene sets

Gene Set	Size	ES	NES	NOM p-val	FDR q-val	FWER p-val	Rank at Max	Leading Edge
RASPATHWAY	22	0.594	2.05	0	0	0	2,563	tags=73%, list=28%, s...
IGF1PATHWAY	20	0.546	1.973	0	0.055	0.133	2,444	tags=65%, list=27%, s...
INSULINPATHWAY	21	0.484	1.799	0	0.196	0.433	2,522	tags=62%, list=28%, s...
NGFPATHWAY	19	0.577	1.729	0	0.233	0.533	2,031	tags=58%, list=22%, s...
EGFPATHWAY	27	0.47	1.697	0	0.245	0.567	2,413	tags=52%, list=27%, s...
UPREG_BY_HOXA9	27	0.52	1.644	0	0.325	0.733	1,517	tags=44%, list=17%, s...
PDGFPATHWAY	27	0.439	1.629	0.059	0.312	0.733	2,461	tags=52%, list=27%, s...
PROTEASOMEPAT...	21	0.534	1.586	0.062	0.399	0.867	2,320	tags=57%, list=26%, s...
XINACT_MERGED	20	0.659	1.558	0.083	0.435	0.933	1,761	tags=55%, list=19%, s...
ERKPATHWAY	29	0.433	1.543	0.059	0.422	0.933	2,461	tags=52%, list=27%, s...
AKTPATHWAY	17	0.446	1.524	0	0.442	0.933	1,440	tags=35%, list=16%, s...
BRCA_UP	38	0.462	1.514	0	0.438	0.967	1,991	tags=39%, list=22%, s...
IGF1RPATHWAY	15	0.444	1.454	0.167	0.621	1	2,716	tags=60%, list=30%, s...
ST_PHOSPHOINOSI...	31	0.377	1.437	0.056	0.634	1	2,232	tags=39%, list=25%, s...
FMLPPATHWAY	33	0.392	1.437	0	0.592	1	1,918	tags=48%, list=21%, s...
IL6PATHWAY	21	0.381	1.428	0	0.596	1	2,385	tags=52%, list=26%, s...
BCRPATHWAY	33	0.382	1.415	0	0.606	1	2,027	tags=42%, list=22%, s...
PITX2PATHWAY	16	0.513	1.412	0.071	0.587	1	721	tags=38%, list=8%, si...
HCMVPATHWAY	15	0.482	1.405	0.071	0.577	1	1,643	tags=40%, list=18%, s...
SPRYPATHWAY	16	0.47	1.39	0	0.605	1	2,413	tags=63%, list=27%, s...
NFKB_REDUCED	20	0.546	1.389	0.143	0.586	1	1,228	tags=35%, list=14%, s...
RELAPATHWAY	16	0.426	1.38	0.188	0.588	1	1,891	tags=44%, list=21%, s...
NKCELLSPATHWAY	18	0.459	1.377	0.083	0.573	1	1,041	tags=33%, list=11%, s...
TUMOR_SUPRESSOR	22	0.377	1.368	0.067	0.574	1	100	tags=14%, list=1%, si...
CELL_CYCLE_CHE...	23	0.519	1.348	0.125	0.606	1	2,592	tags=52%, list=28%, s...
ST_B_CELL_ANTIG...	38	0.328	1.326	0	0.655	1	2,287	tags=37%, list=25%, s...
GLYCOGEN_META...	33	0.388	1.323	0.095	0.637	1	1,0[illegible]3	tags=24%, list=11%, s...
GCRPATHWAY	16	0.515	1.318	0.125	0.63	1	2,015	tags=44%, list=22%, s...

For 5 selected gene sets:

GSEA reports

Processes: click 'status' field for results

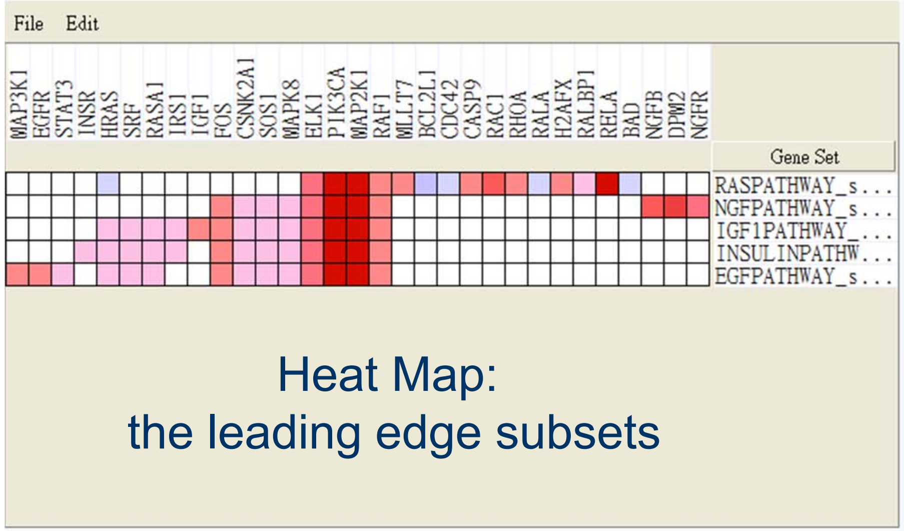
	Name	Status
1	Gsea	Success
2	Gsea	Success

下午 01:38:39

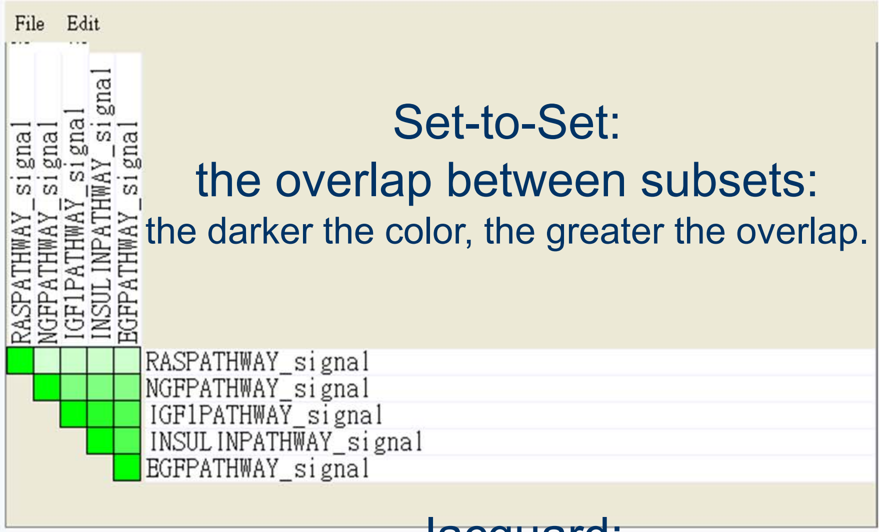
9640 [INFO] Begun importing: RankedList from: C:\Users\hmwu\gsea_home\output\六月07\my_analysis.Gsea.1307511195910\edb\P53_hgu95av2_collapsed_to_symbols.rnk

160M of 247M

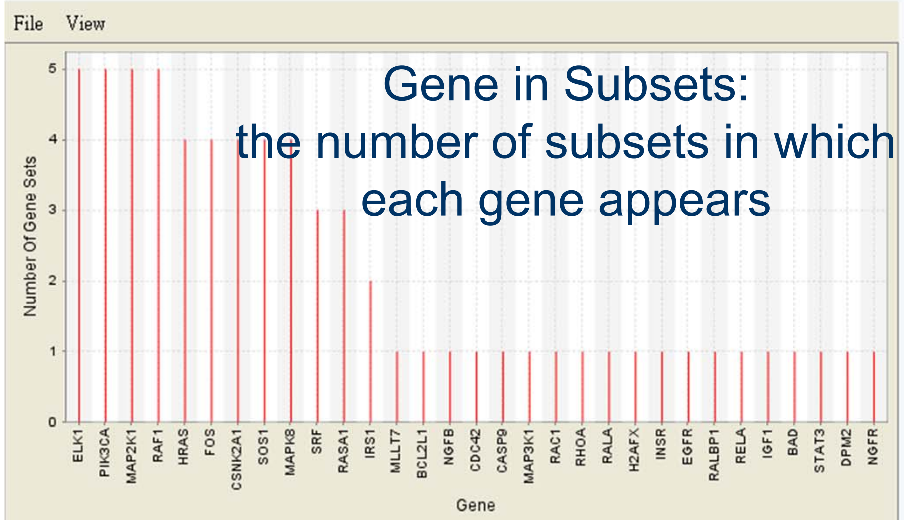
Results



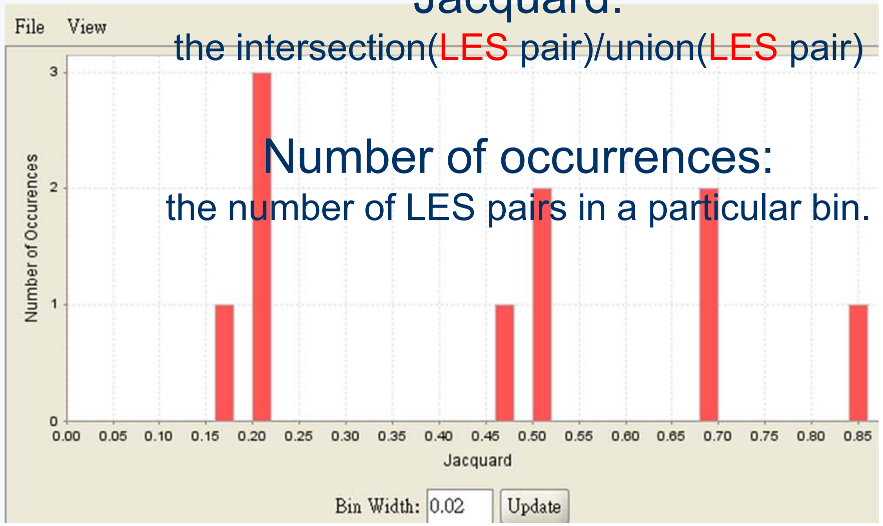
Heat Map:
the leading edge subsets



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



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



Bin Width: 0.02

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

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