GSEA:
Gene Set Enrichment Analysis
軟體操作

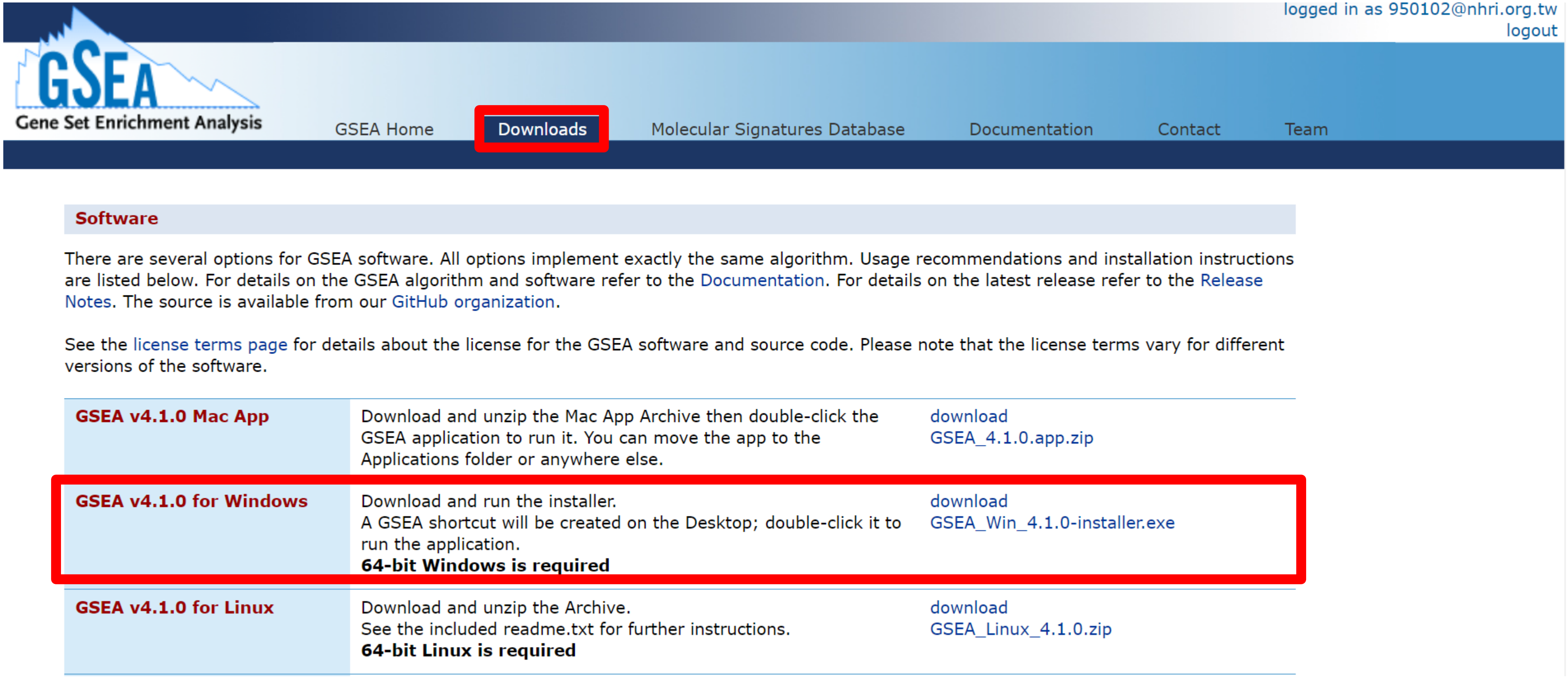
江士昇　蔡芳榆

20210908

Steps

- 1.Download GSEA
- 2.Prepare data
- 3. Loading data
- 4. Running analysis
- 5. Viewing analysis results

http://software.broadinstitute.org/gsea/index.jsp


logged in as 950102@nhri.org.tw
logout
GSEA
Gene Set Enrichment Analysis
GSEA Home
Downloads
Molecular Signatures Database
Documentation
Contact
Team

Software

There are several options for GSEA software. All options implement exactly the same algorithm. Usage recommendations and installation instructions are listed below. For details on the GSEA algorithm and software refer to the Documentation. For details on the latest release refer to the Release Notes. The source is available from our GitHub organization.

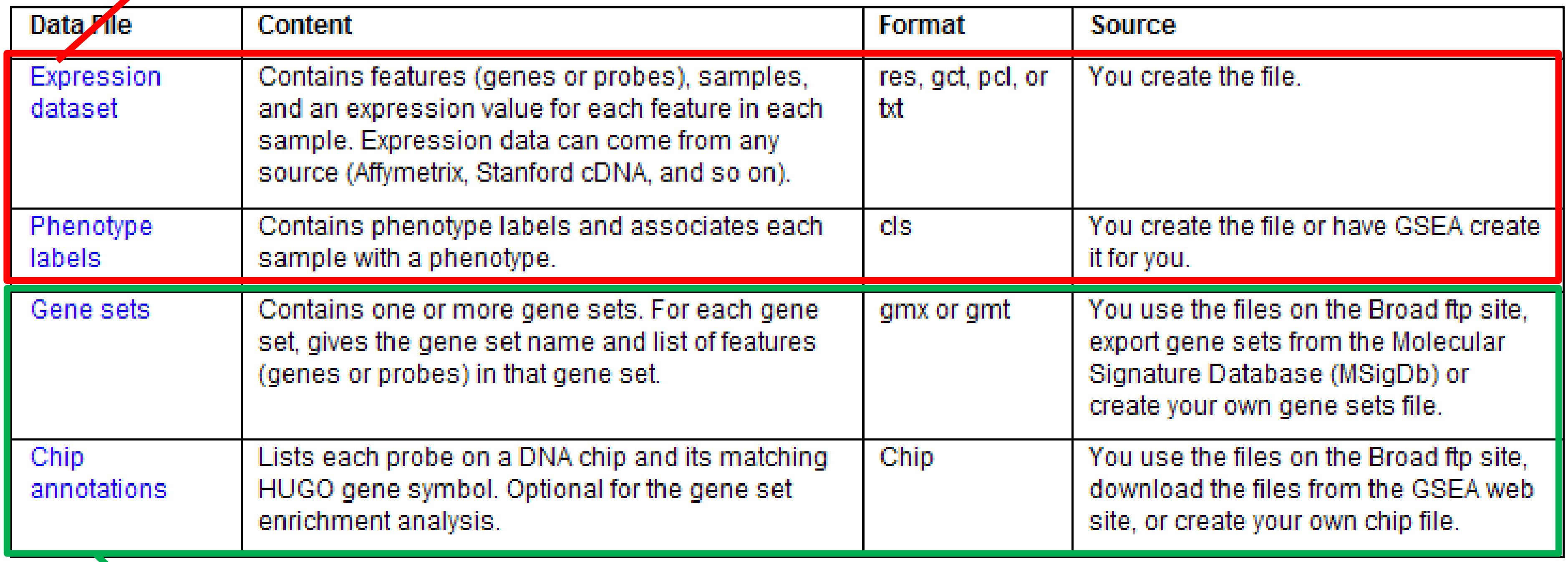
See the license terms page for details about the license for the GSEA software and source code. Please note that the license terms vary for different versions of the software.

GSEA v4.1.0 Mac App	Download and unzip the Mac App Archive then double-click the GSEA application to run it. You can move the app to the Applications folder or anywhere else.	download GSEA_4.1.0.app.zip
GSEA v4.1.0 for Windows	Download and run the installer. A GSEA shortcut will be created on the Desktop; double-click it to run the application. **64-bit Windows is required**	download GSEA_Win_4.1.0-installer.exe
GSEA v4.1.0 for Linux	Download and unzip the Archive. See the included readme.txt for further instructions. **64-bit Linux is required**	download GSEA_Linux_4.1.0.zip

Step2.Prepare data

需準備的實驗數據, 樣本資料

-



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.
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.

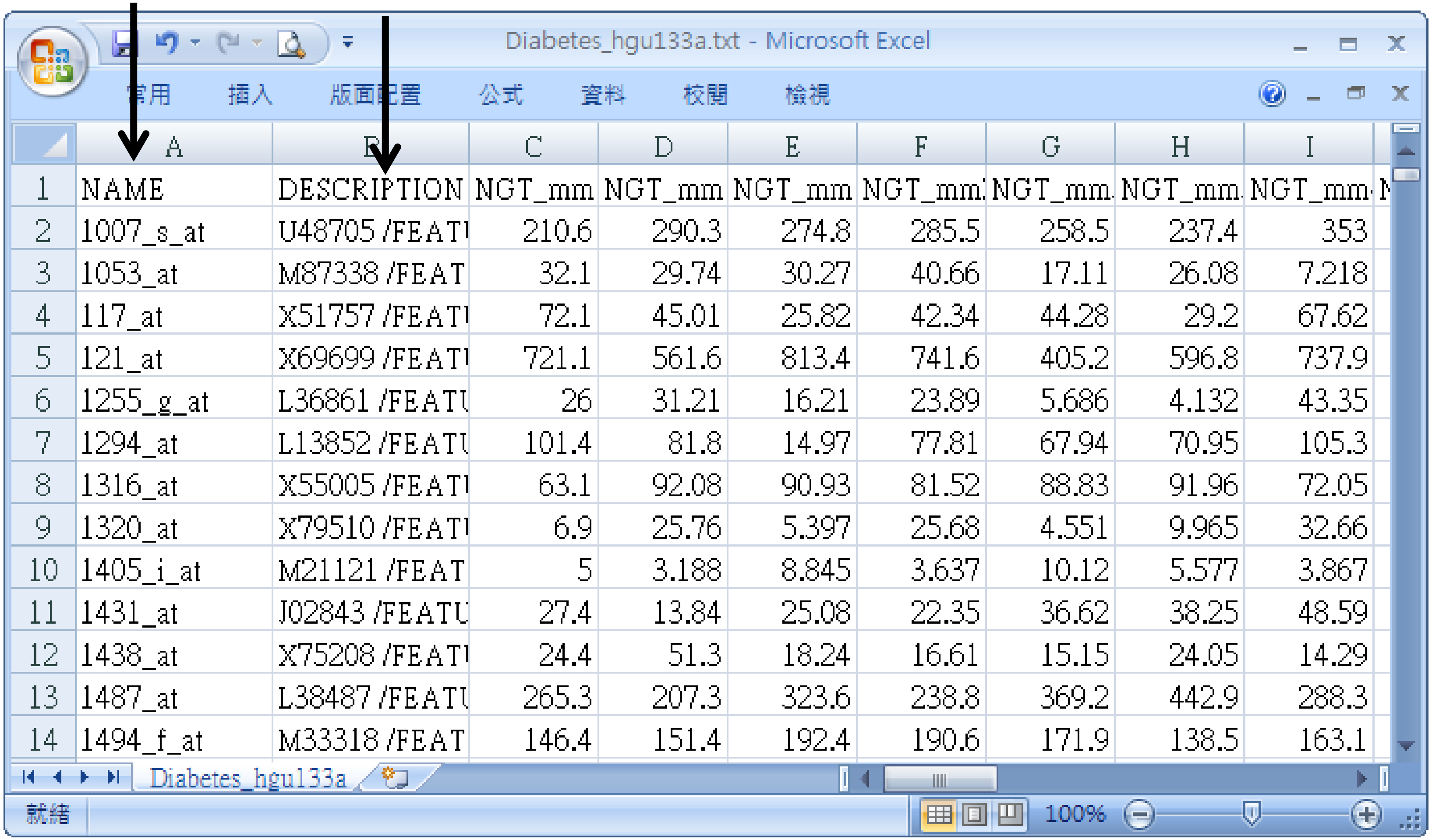
可用GSEA內建的 Gene Set, 實驗平台資訊(probe-gene)

Expression data format: .txt

The **first line** contains the labels **NAME** and **DESCRIPTION** followed by the identifiers for each sample in the dataset.

Probe name

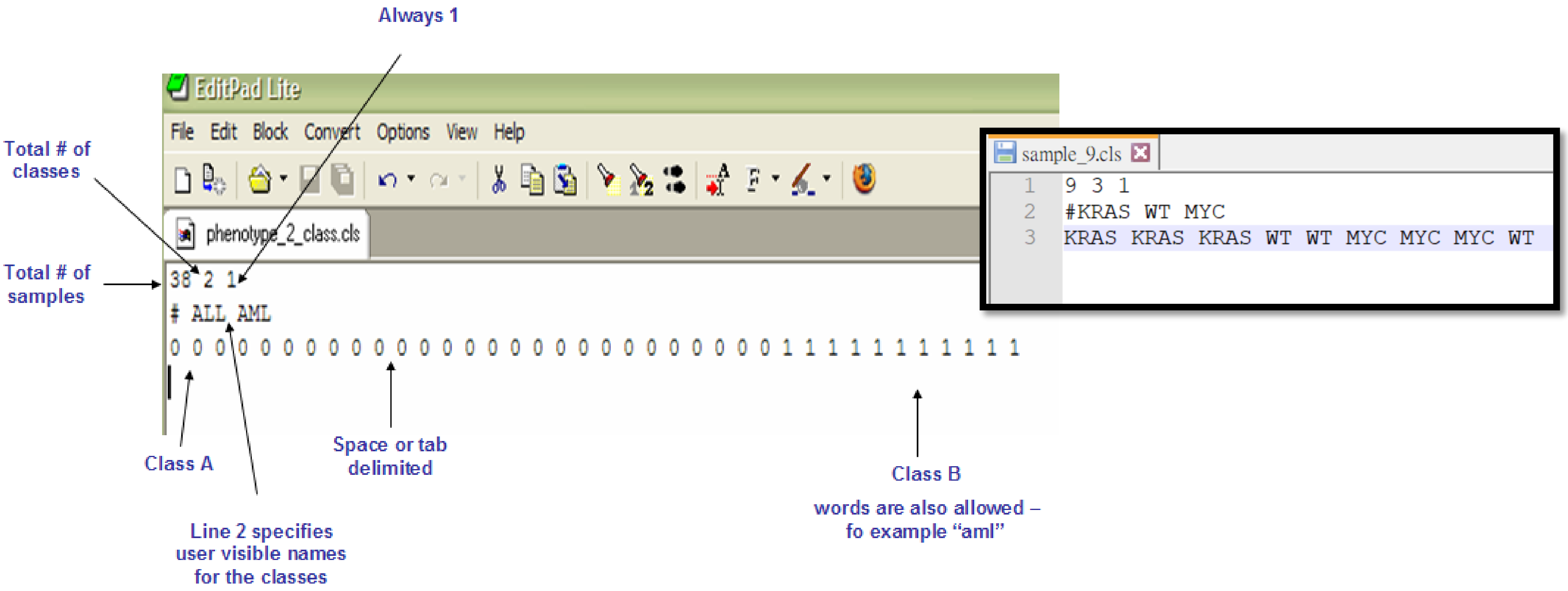
can be ‘na’



	A	B	C	D	E	F	G	H	I
1	NAME	DESCRIPTION	NGT_mm	NGT_mm	NGT_mm	NGT_mm	NGT_mm	NGT_mm	NGT_mm
2	1007_s_at	U48705 /FEAT	210.6	290.3	274.8	285.5	258.5	237.4	353
3	1053_at	M87338 /FEAT	32.1	29.74	30.27	40.66	17.11	26.08	7.218
4	117_at	X51757 /FEAT	72.1	45.01	25.82	42.34	44.28	29.2	67.62
5	121_at	X69699 /FEAT	721.1	561.6	813.4	741.6	405.2	596.8	737.9
6	1255_g_at	L36861 /FEAT	26	31.21	16.21	23.89	5.686	4.132	43.35
7	1294_at	L13852 /FEAT	101.4	81.8	14.97	77.81	67.94	70.95	105.3
8	1316_at	X55005 /FEAT	63.1	92.08	90.93	81.52	88.83	91.96	72.05
9	1320_at	X79510 /FEAT	6.9	25.76	5.397	25.68	4.551	9.965	32.66
10	1405_i_at	M21121 /FEAT	5	3.188	8.845	3.637	10.12	5.577	3.867
11	1431_at	J02843 /FEAT	27.4	13.84	25.08	22.35	36.62	38.25	48.59
12	1438_at	X75208 /FEAT	24.4	51.3	18.24	16.61	15.15	24.05	14.29
13	1487_at	L38487 /FEAT	265.3	207.3	323.6	238.8	369.2	442.9	288.3
14	1494_f_at	M33318 /FEAT	146.4	151.4	192.4	190.6	171.9	138.5	163.1

Phenotype data format: .cls

Categorical class file format (e.g NGT vs DMT; tumor vs normal)



```
38 2 1
# ALL AML
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1
```

```
9 3 1
#KRAS WT MYC
KRAS KRAS KRAS WT WT MYC MYC MYC WT
```

Always 1

Total # of classes

Total # of samples

Class A

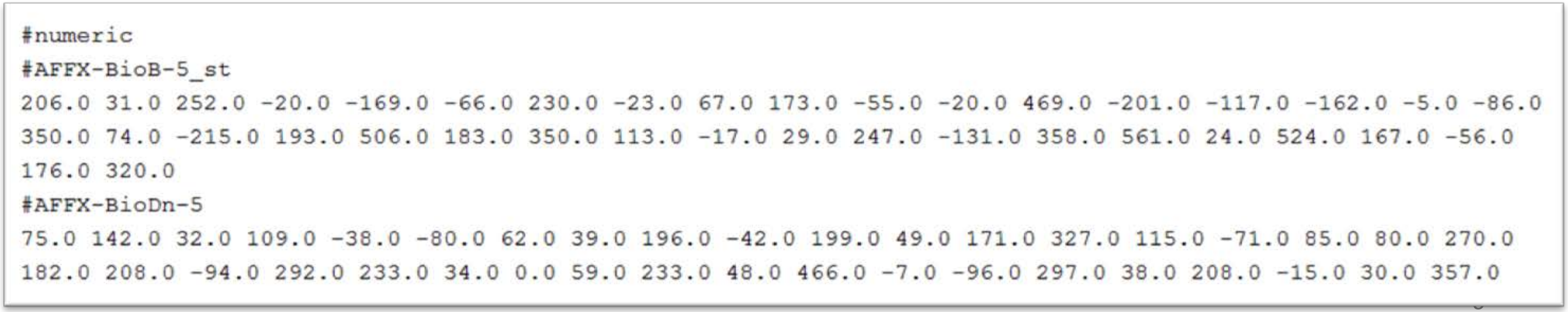
Line 2 specifies user visible names for the classes

Space or tab delimited

Class B

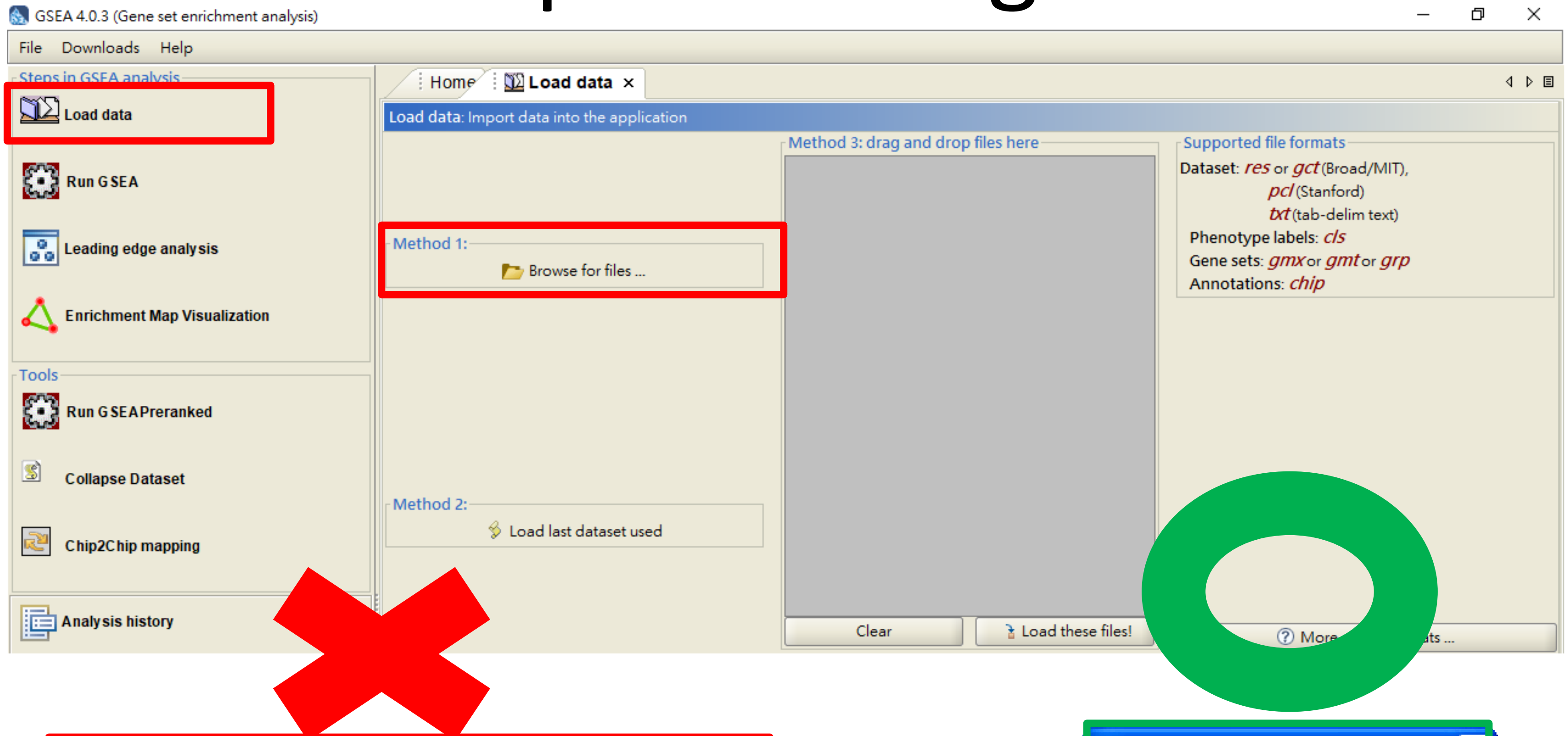
words are also allowed – fo example “aml”

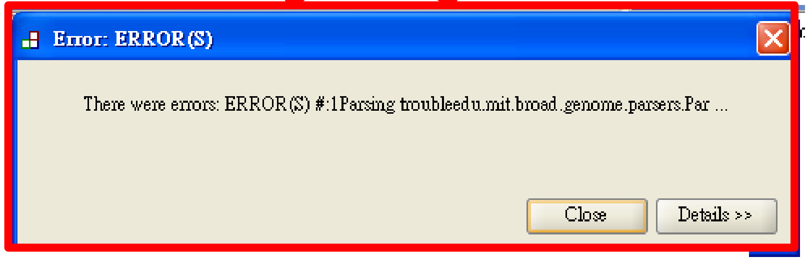
Continuous file format (e.g time-series or gene profile)



```
#numeric
#AFFX-BioB-5_st
206.0 31.0 252.0 -20.0 -169.0 -66.0 230.0 -23.0 67.0 173.0 -55.0 -20.0 469.0 -201.0 -117.0 -162.0 -5.0 -86.0
350.0 74.0 -215.0 193.0 506.0 183.0 350.0 113.0 -17.0 29.0 247.0 -131.0 358.0 561.0 24.0 524.0 167.0 -56.0
176.0 320.0
#AFFX-BioDn-5
75.0 142.0 32.0 109.0 -38.0 -80.0 62.0 39.0 196.0 -42.0 199.0 49.0 171.0 327.0 115.0 -71.0 85.0 80.0 270.0
182.0 208.0 -94.0 292.0 233.0 34.0 0.0 59.0 233.0 48.0 466.0 -7.0 -96.0 297.0 38.0 208.0 -15.0 30.0 357.0
```

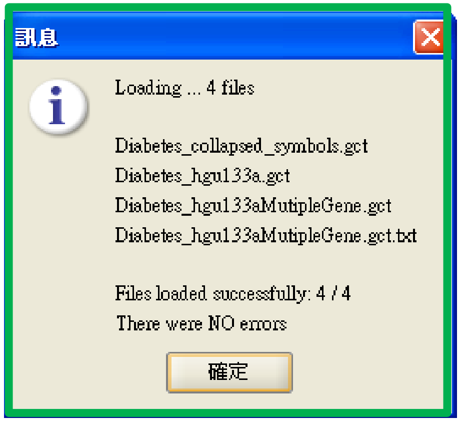
Step3. Loading data





There were errors: ERROR(S) #:1Parsing troubleedu.mit.broad.genome.parsers.Par ...

資料格式有錯誤, 需再檢查

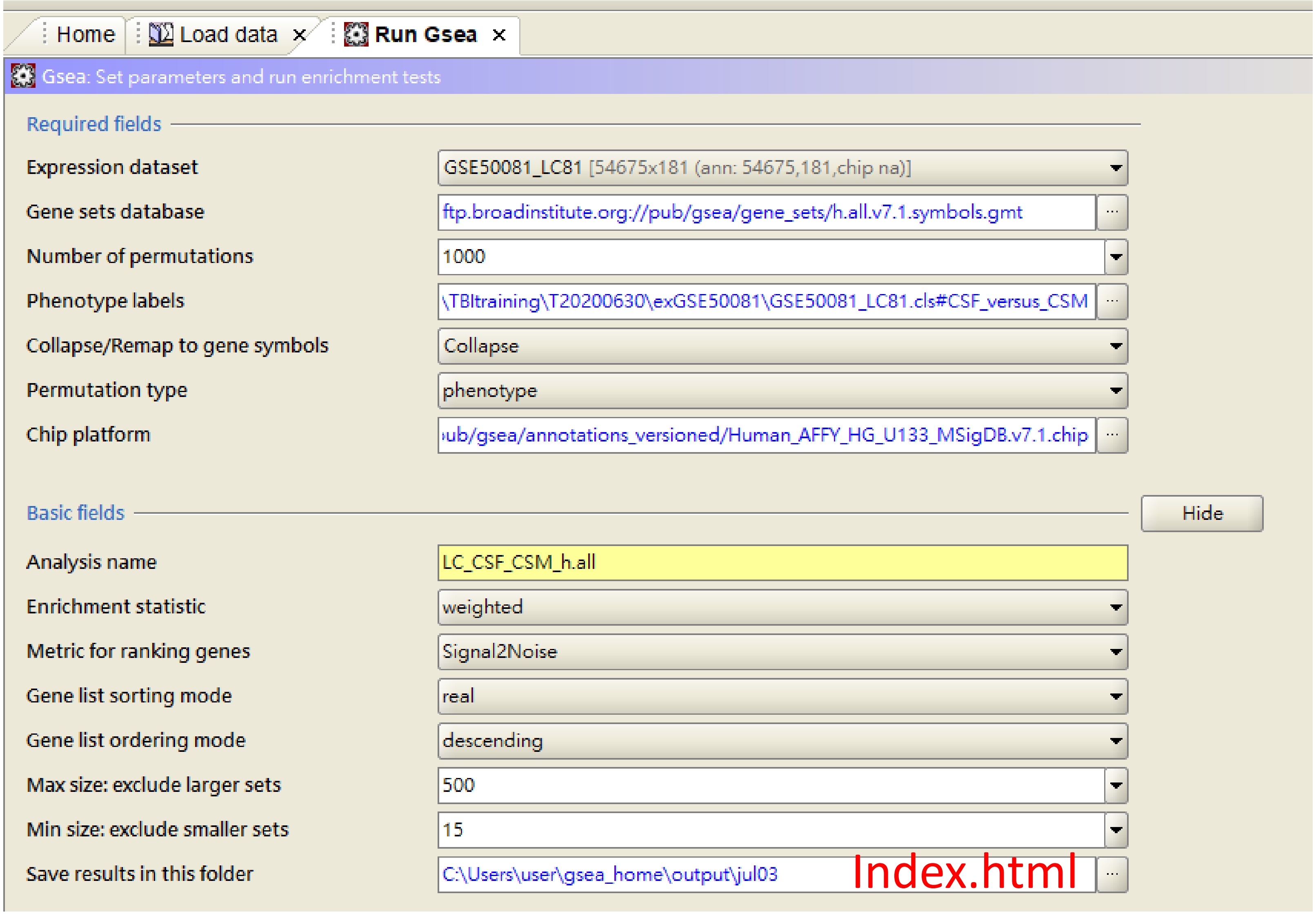


Loading ... 4 files

Diabetes_collapsed_symbols.gct
Diabetes_hgu133a.gct
Diabetes_hgu133aMutipleGene.gct
Diabetes_hgu133aMutipleGene.gct.txt

Files loaded successfully: 4 / 4
There were NO errors

Step4. Running analysis



Gsea: Set parameters and run enrichment tests

A

Required fields

*1	Expression dataset	GSE50081_LC81 [54675x181 (ann: 54675,181,chip na)]
*2	Gene sets database	ftp.broadinstitute.org://pub/gsea/gene_sets/h.all.v7.1.symbols.gmt
*3	Number of permutations	1000
*4	Phenotype labels	\TBltraining\T20200630\exGSE50081\GSE50081_LC81.cls#CSF_versus_CSM
*5	Collapse/Remap to gene symbols	Collapse
*6	Permutation type	phenotype
*7	Chip platform	ub/gsea/annotations_versioned/Human_AFFY_HG_U133_MSigDB.v7.1.chip

B

Basic fields

*1	Analysis name	LC_CSF_CSM_h.all
2	Enrichment statistic	weighted
*3	Metric for ranking genes	Signal2Noise
4	Gene list sorting mode	real
5	Gene list ordering mode	descending
6	Max size: exclude larger sets	500
7	Min size: exclude smaller sets	15
*8	Save results in this folder	C:\Users\user\gsea_home\output\jul03

Index.html

- Metrics for ranking genes

For categorical phenotypes

Signal2Noise

$$\frac{\mu_A - \mu_B}{\sigma_A + \sigma_B}$$

tTest

$$\frac{\mu_A - \mu_B}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}}$$

Ratio_of_Classes

$$\frac{\mu_A}{\mu_B}$$

Diff_of_Classes

$$\mu_A - \mu_B$$

log2_Ratio_of_Classes

$$\log2\left(\frac{\mu_A}{\mu_B}\right)$$

For continuous phenotypes

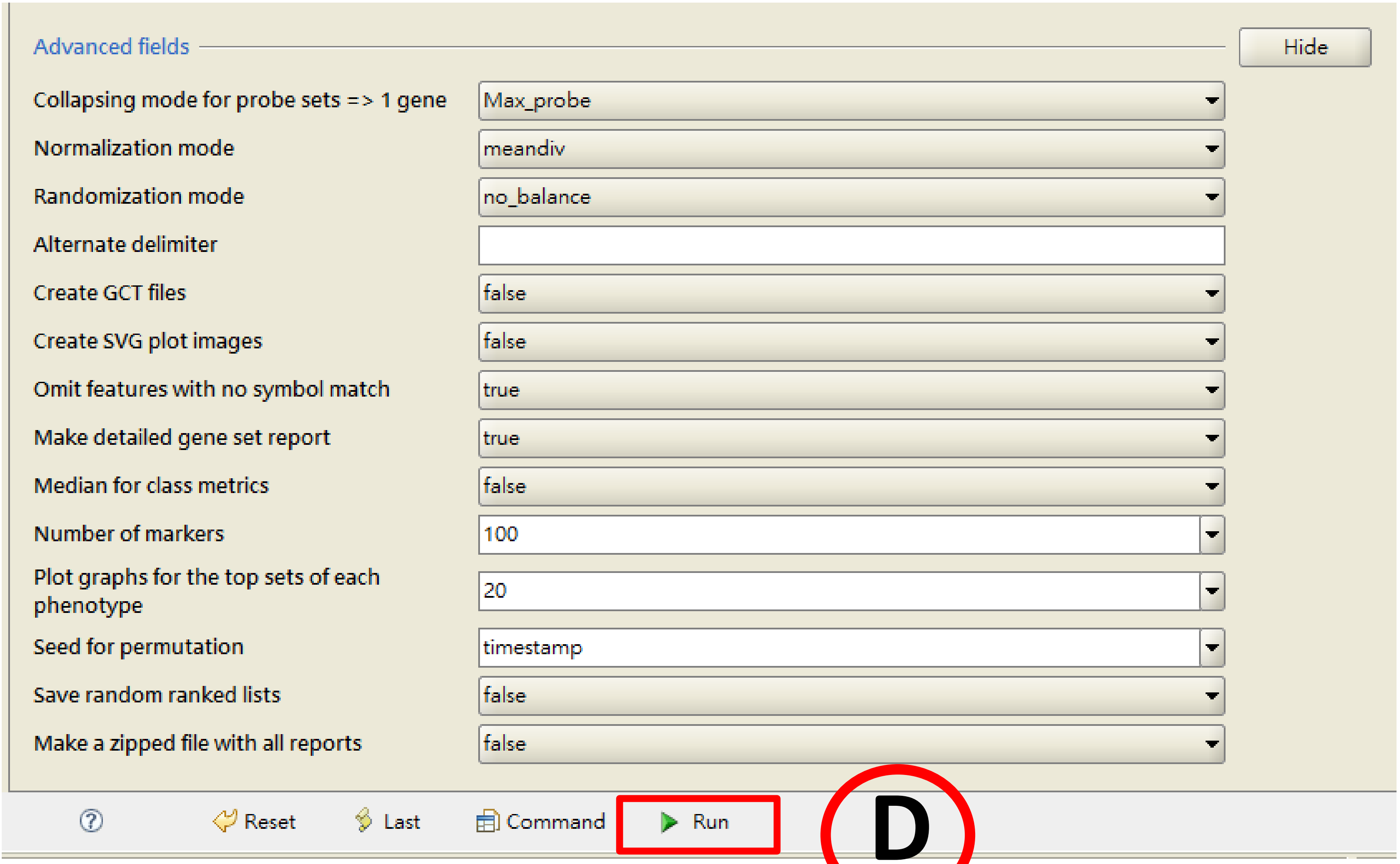
Pearson

Cosine

Manhattan

Euclidean

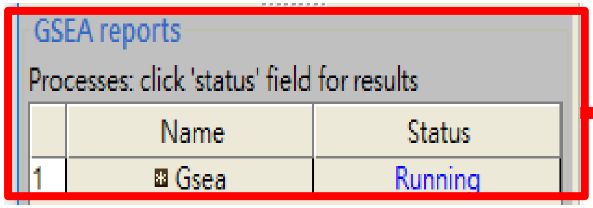
C



Advanced fields

1	Collapsing mode for probe sets => 1 gene	Max_probe
2	Normalization mode	meandiv
3	Randomization mode	no_balance
4	Alternate delimiter	
5	Create GCT files	false
6	Create SVG plot images	false
7	Omit features with no symbol match	true
8	Make detailed gene set report	true
9	Median for class metrics	false
10	Number of markers	100
11	Plot graphs for the top sets of each phenotype	20
12	Seed for permutation	timestamp
13	Save random ranked lists	false
14	Make a zipped file with all reports	false

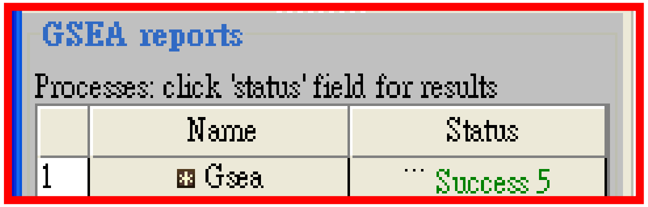
D



GSEA reports

Processes: click 'status' field for results

	Name	Status
1	⊞ Gsea	Running

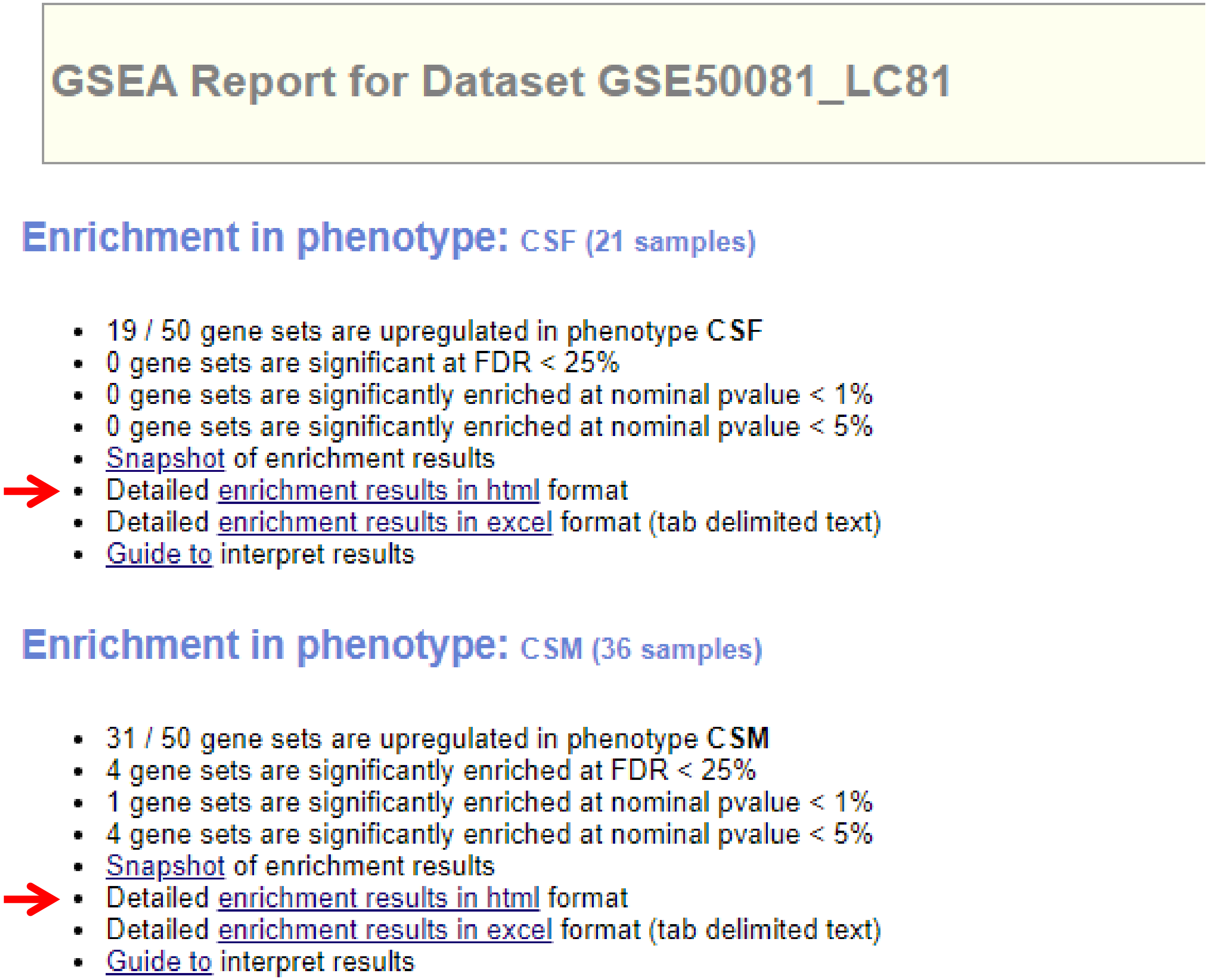


GSEA reports

Processes: click 'status' field for results

	Name	Status
1	⊞ Gsea	⋯ Success 5

Step5. Viewing analysis results



GSEA Report for Dataset GSE50081_LC81

Enrichment in phenotype: CSF (21 samples)

- 19 / 50 gene sets are upregulated in phenotype CSF
- 0 gene sets are significant at FDR < 25%
- 0 gene sets are significantly enriched at nominal pvalue < 1%
- 0 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: CSM (36 samples)

- 31 / 50 gene sets are upregulated in phenotype CSM
- 4 gene sets are significantly enriched at FDR < 25%
- 1 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

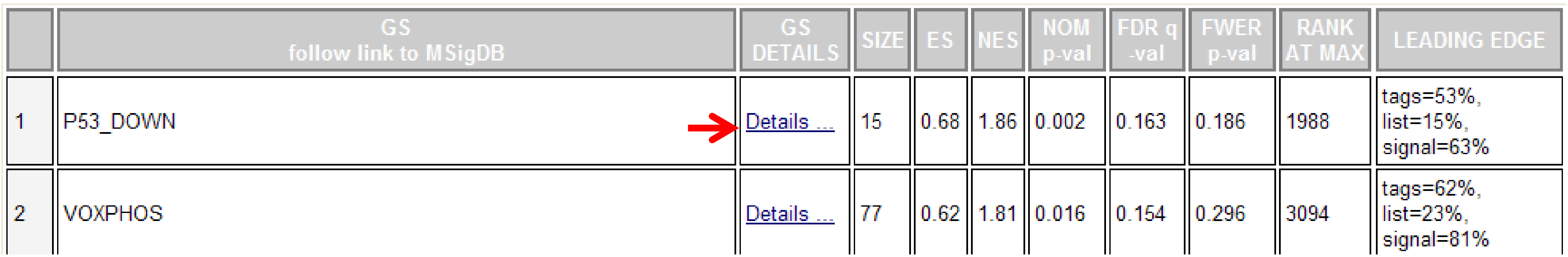
3.SIZE: Number of genes in the gene set after filtering out those genes not in the expression dataset

4.ES: Enrichment score for the gene set; that is, the degree to which this gene set is overrepresented at the top or bottom of the ranked list of genes in the expression dataset.

5.NES: Normalized enrichment score; that is, the enrichment score for the gene set after it has been normalized across analyzed gene sets.

6.NOM p-val: Nominal p value; that is, the statistical significance of the enrichment score. The nominal p value is not adjusted for gene set size or multiple hypothesis testing; therefore, it is of limited use in comparing gene sets.

7.FDR q val: False discovery rate; that is, the estimated probability that the normalized enrichment score represents a false positive finding. (GSEA建議小於0.25)



	1	2	3	4	5	6	7	8	9	10
	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	P53_DOWN	Details ...	15	0.68	1.86	0.002	0.163	0.186	1988	tags=53%, list=15%, signal=63%
2	VOXPHOS	Details ...	77	0.62	1.81	0.016	0.154	0.296	3094	tags=62%, list=23%, signal=81%

Click “GS DETAILS” and you can see ...

GSEA Results Summary

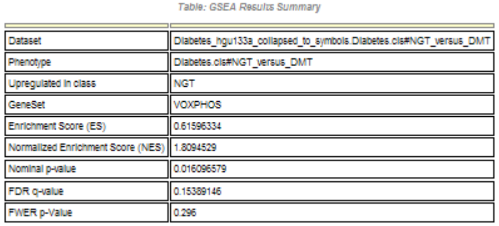
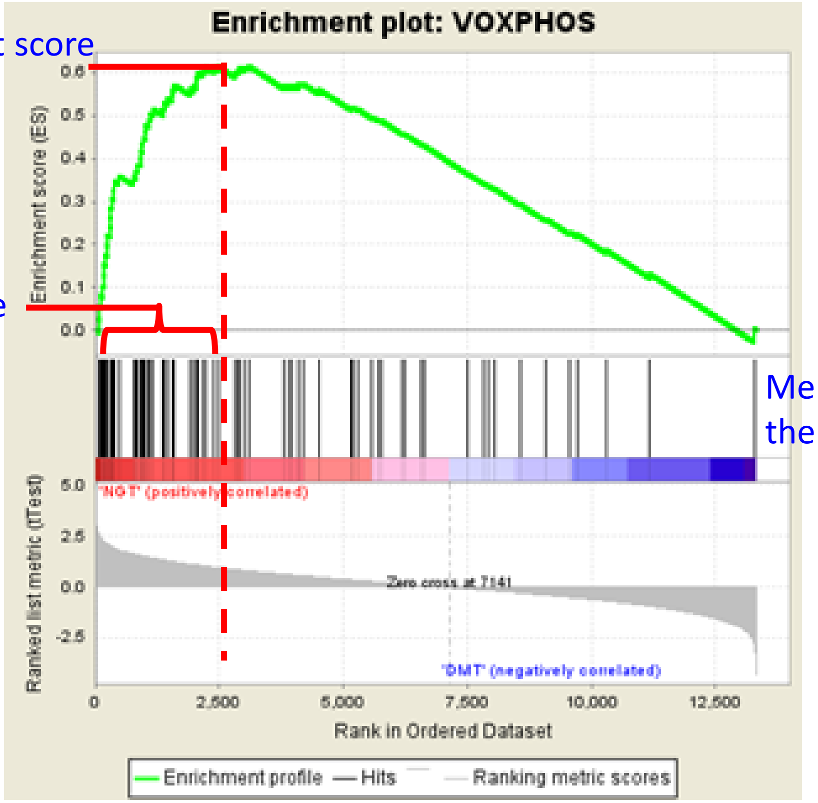


Table: GSEA Results Summary

Dataset	Diabetes_hgu133a_collapsed_to_symbols.Diabetes.cls#NGT_versus_DMT
Phenotype	Diabetes.cls#NGT_versus_DMT
Upregulated in class	NGT
GeneSet	VOXPHOS
Enrichment Score (ES)	0.61596334
Normalized Enrichment Score (NES)	1.8094529
Nominal p-value	0.016096579
FDR q-value	0.15389146
FWER p-Value	0.296

Enrichment plot



Enrichment score

Core gene

Member of the gene set

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

GSEA details

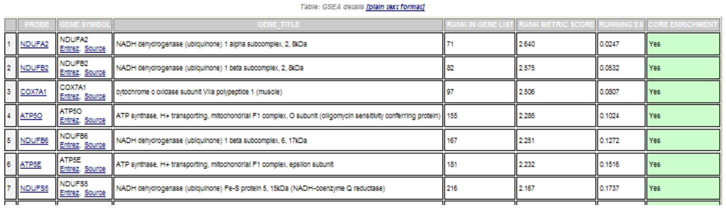
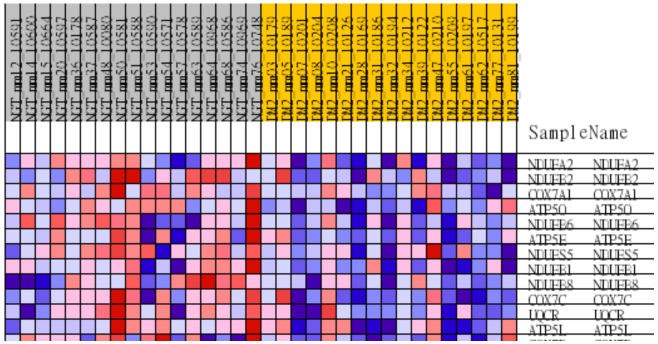


Table: GSEA details [plain text format]

	PROBE	GENE SYMBOL	GENE_TITLE	RANK IN GENE LIST	RANK METRIC SCORE	RUNNING ES	CORE ENRICHMENT
1	NDUFA2	NDUFA2 Entrez, Source	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 2, 8kDa	71	2.640	0.0247	Yes
2	NDUFB2	NDUFB2 Entrez, Source	NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 2, 8kDa	82	2.575	0.0532	Yes
3	COX7A1	COX7A1 Entrez, Source	cytochrome c oxidase subunit VIIa polypeptide 1 (muscle)	97	2.506	0.0807	Yes
4	ATP5O	ATP5O Entrez, Source	ATP synthase, H+ transporting, mitochondrial F1 complex, O subunit (oligomycin sensitivity conferring protein)	155	2.285	0.1024	Yes
5	NDUFB6	NDUFB6 Entrez, Source	NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 6, 17kDa	167	2.251	0.1272	Yes
6	ATP5E	ATP5E Entrez, Source	ATP synthase, H+ transporting, mitochondrial F1 complex, epsilon subunit	181	2.232	0.1516	Yes
7	NDUFS5	NDUFS5 Entrez, Source	NADH dehydrogenase (ubiquinone) Fe-S protein 5, 15kDa (NADH-coenzyme Q reductase)	216	2.167	0.1737	Yes

Heat map



Random ES distribution

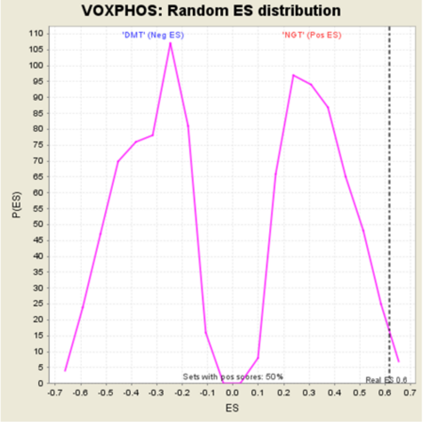


Fig 3: VOXPHOS: Random ES distribution
Gene set null distribution of ES for VOXPHOS

Reference

- *Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP.* (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. PNAS 102 (43) 15545-15550.
- http://software.broadinstitute.org/gsea/index.jsp