GSEA:
Gene Set Enrichment Analysis
軟體操作

蔡芳榆

2013.09.11

Steps

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

Step1.Download GSEA

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

GSEA
Gene Set Enrichment Analysis

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

Downloads

The GSEA software and source code and the Molecular Signatures Database (MSigDB) are freely available to individuals in both academia and industry 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 a listed below. Current Java implementations of GSEA require Java 6 or 7.

javaGSEA Desktop Application

- Easy-to-use graphical user interface
- Runs on any desktop computer (Windows, Mac OS X, Linux etc.) that supports Java 6 or 7
- Produces richly annotated reports of enrichment results
- Integrated gene sets browser to view gene set annotations, search for gene sets and map gene sets between platforms

Launch with
1GB (for 32 or 64-bit Java)
memory:
Launch

Step2.Prepare data

需準備

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

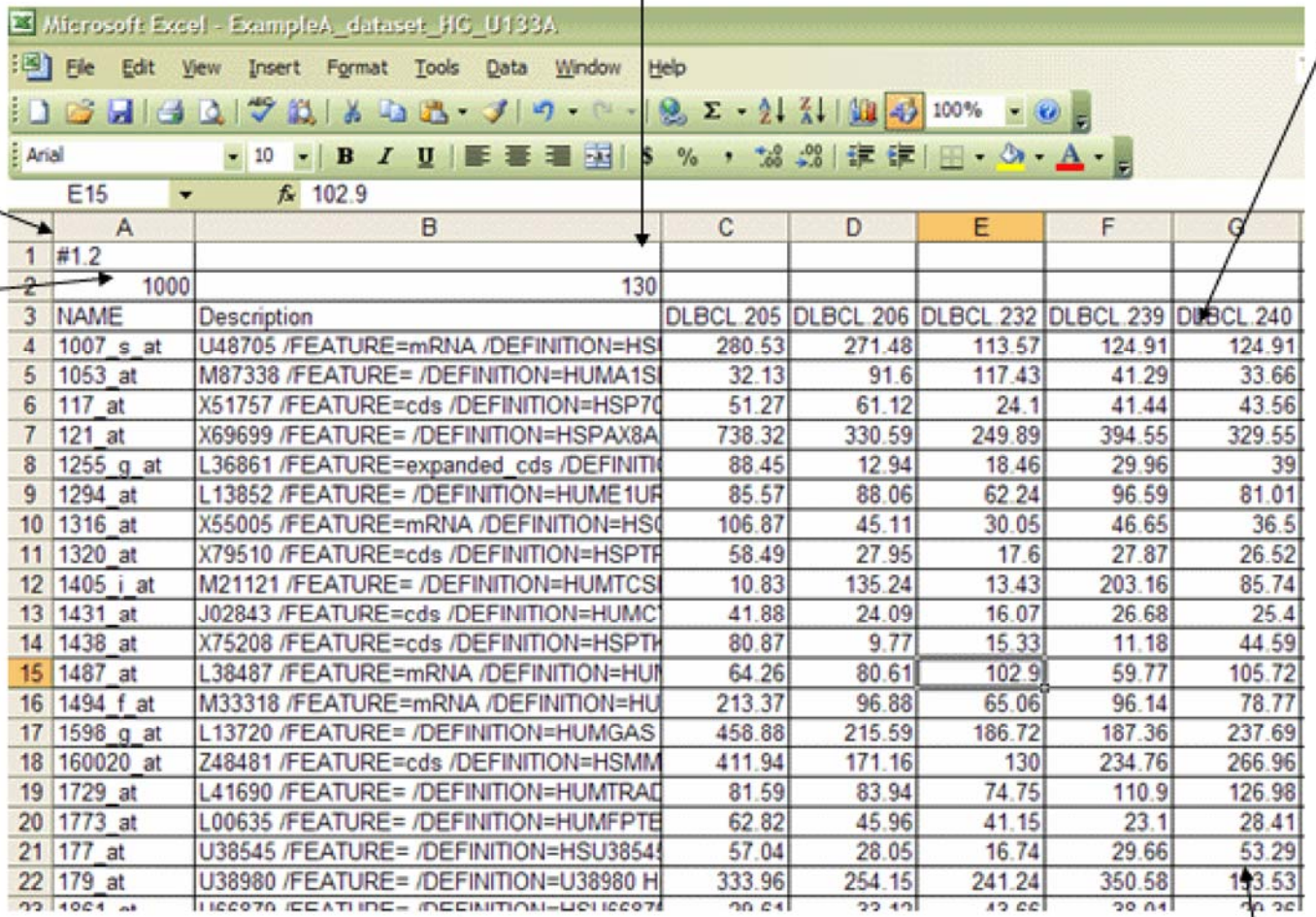
Expression data format: .gct

of samples

Third column onwards are sample names. These must be UNIQUE

Always "#1.2"

The # of rows (i.e probe sets)



	A	B	C	D	E	F	G
1	#1.2						
2	1000	130					
3	NAME	Description	DLBCL 205	DLBCL 206	DLBCL 232	DLBCL 239	DLBCL 240
4	1007_s_at	U48705 /FEATURE=mRNA /DEFINITION=HS	280.53	271.48	113.57	124.91	124.91
5	1053_at	M87338 /FEATURE= /DEFINITION=HUMA1S	32.13	91.6	117.43	41.29	33.66
6	117_at	X51757 /FEATURE=cds /DEFINITION=HSP7	51.27	61.12	24.1	41.44	43.56
7	121_at	X69699 /FEATURE= /DEFINITION=HSPAX8A	738.32	330.59	249.89	394.55	329.55
8	1255_g_at	L36861 /FEATURE=expanded_cds /DEFINITI	88.45	12.94	18.46	29.96	39
9	1294_at	L13852 /FEATURE= /DEFINITION=HUME1UR	85.57	88.06	62.24	96.59	81.01
10	1316_at	X55005 /FEATURE=mRNA /DEFINITION=HS	106.87	45.11	30.05	46.65	36.5
11	1320_at	X79510 /FEATURE=cds /DEFINITION=HSPTF	58.49	27.95	17.6	27.87	26.52
12	1405_i_at	M21121 /FEATURE= /DEFINITION=HUMTCS	10.83	135.24	13.43	203.16	85.74
13	1431_at	J02843 /FEATURE=cds /DEFINITION=HUMC	41.88	24.09	16.07	26.68	25.4
14	1438_at	X75208 /FEATURE=cds /DEFINITION=HSPTH	80.87	9.77	15.33	11.18	44.59
15	1487_at	L38487 /FEATURE=mRNA /DEFINITION=HU	64.26	80.61	102.9	59.77	105.72
16	1494_f_at	M33318 /FEATURE=mRNA /DEFINITION=HU	213.37	96.88	65.06	96.14	78.77
17	1598_g_at	L13720 /FEATURE= /DEFINITION=HUMGAS	458.88	215.59	186.72	187.36	237.69
18	160020_at	Z48481 /FEATURE=cds /DEFINITION=HSMM	411.94	171.16	130	234.76	266.96
19	1729_at	L41690 /FEATURE= /DEFINITION=HUMTRAD	81.59	83.94	74.75	110.9	126.98
20	1773_at	L00635 /FEATURE= /DEFINITION=HUMFPTE	62.82	45.96	41.15	23.1	28.41
21	177_at	U38545 /FEATURE= /DEFINITION=HSU3854	57.04	28.05	16.74	29.66	53.29
22	179_at	U38980 /FEATURE= /DEFINITION=U38980 H	333.96	254.15	241.24	350.58	1[illegible]3.53

Data starts on line 4

Column 1: Row identifiers. Typically probe set ids or clone ids. These must be UNIQUE

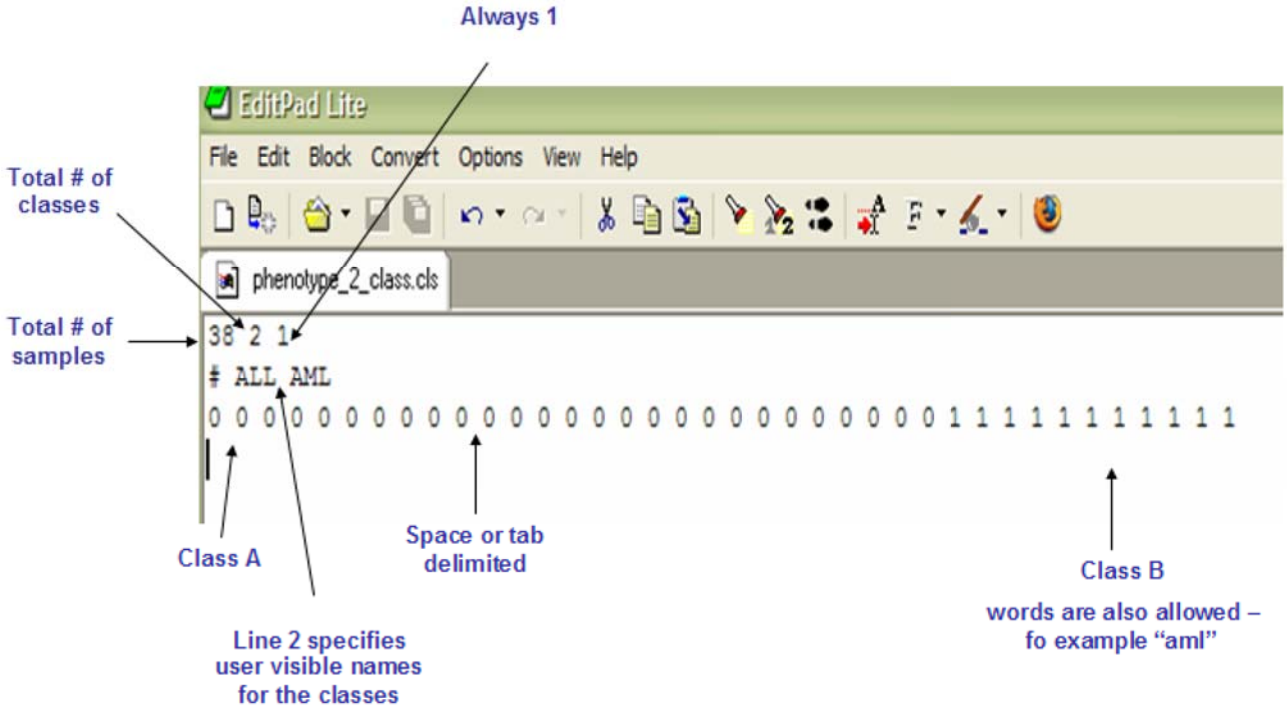
Column 2: Row descriptions. Ignored by the program – can be dummy values (e.g. "na")

Each column contains expression values from 1 sample. Missing values are allowed (leave empty).

If editing, in Excel, make sure to save your data as "tab delimited text"

Phenotype data format: .cls

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



Example of a 3 class cls file

phenotype_2_class.cls | P53.cls

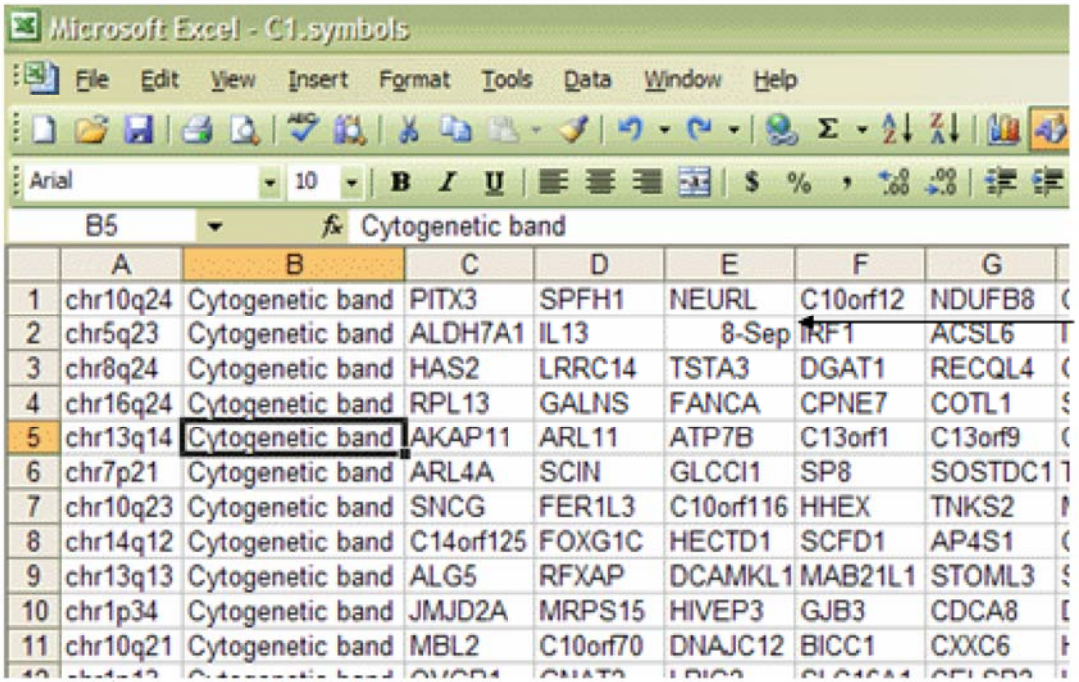
```
9 3 1
#KRAS_MUT WT MYC_MUT
KMUT KMUT KMUT WT WT WT myc myc myc
```

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
```

Gene set database formats : .gmt

Each row represents one gene set



	A	B	C	D	E	F	G
1	chr10q24	Cytogenetic band	PITX3	SPFH1	NEURL	C10orf12	NDUFB8
2	chr5q23	Cytogenetic band	ALDH7A1	IL13	8-Sep	IRF1	ACSL6
3	chr8q24	Cytogenetic band	HAS2	LRRC14	TSTA3	DGAT1	RECQL4
4	chr16q24	Cytogenetic band	RPL13	GALNS	FANCA	CPNE7	COTL1
5	chr13q14	Cytogenetic band	AKAP11	ARL11	ATP7B	C13orf1	C13orf9
6	chr7p21	Cytogenetic band	ARL4A	SCIN	GLCCI1	SP8	SOSTDC1
7	chr10q23	Cytogenetic band	SNCG	FER1L3	C10orf116	HHEX	TNKS2
8	chr14q12	Cytogenetic band	C14orf125	FOXG1C	HECTD1	SCFD1	AP4S1
9	chr13q13	Cytogenetic band	ALG5	RFXAP	DCAMKL1	MAB21L1	STOML3
10	chr1p34	Cytogenetic band	JMJD2A	MRPS15	HIVEP3	GJB3	CDCA8
11	chr10q21	Cytogenetic band	MBL2	C10orf70	DNAJC12	BICC1	CXXC6

If editing in excel, watch out for its tendency to auto-format gene sets (SEP8 becomes 8-Sep)

First column are gene set names. Duplicates are not allowed

Second column contains a brief description. Its optional – you can fill in a dummy field (e.g. "na")

Unequal lengths (i.e # of genes) *is* allowed

GMT format is convenient to store large databases of gene sets. For a handful of sets (<256) the gmx format offers greater excel-editability

Microarray chip annotation formats : .chip

Columns can appear in any order (and there can be more columns which will be ignored).

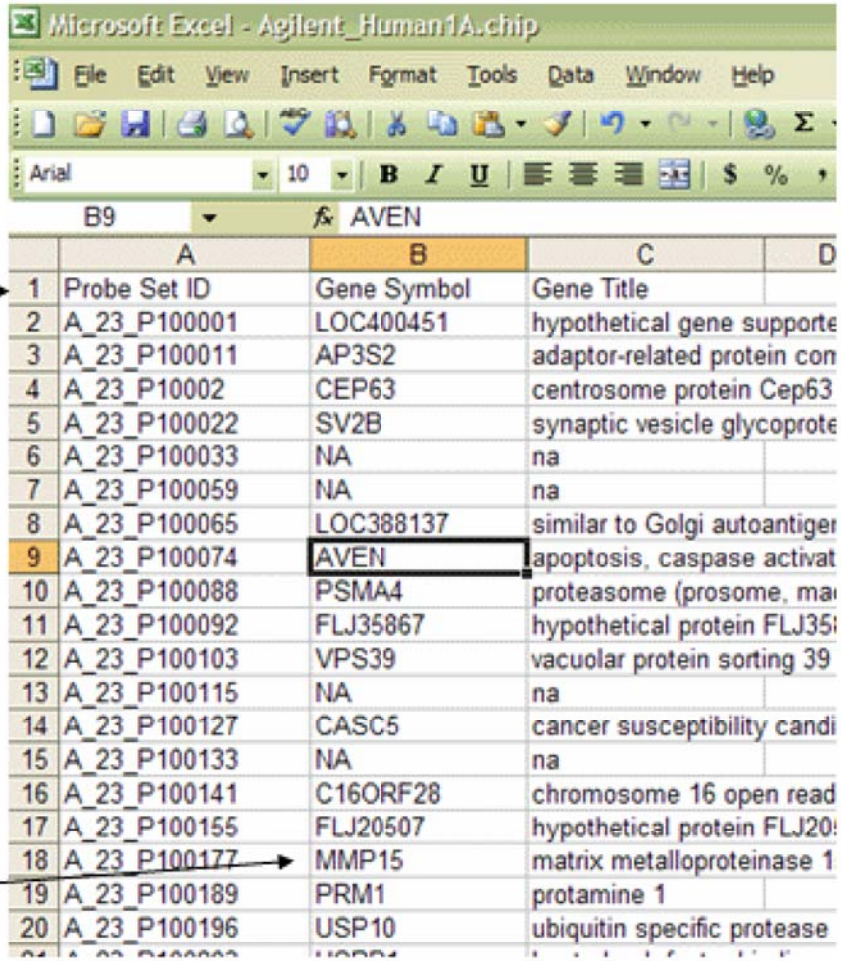
The column headers must be identical to listed:

Probe Set ID

Gene Symbol

Gene Title

Gene symbols are considered to be case-INsensitive



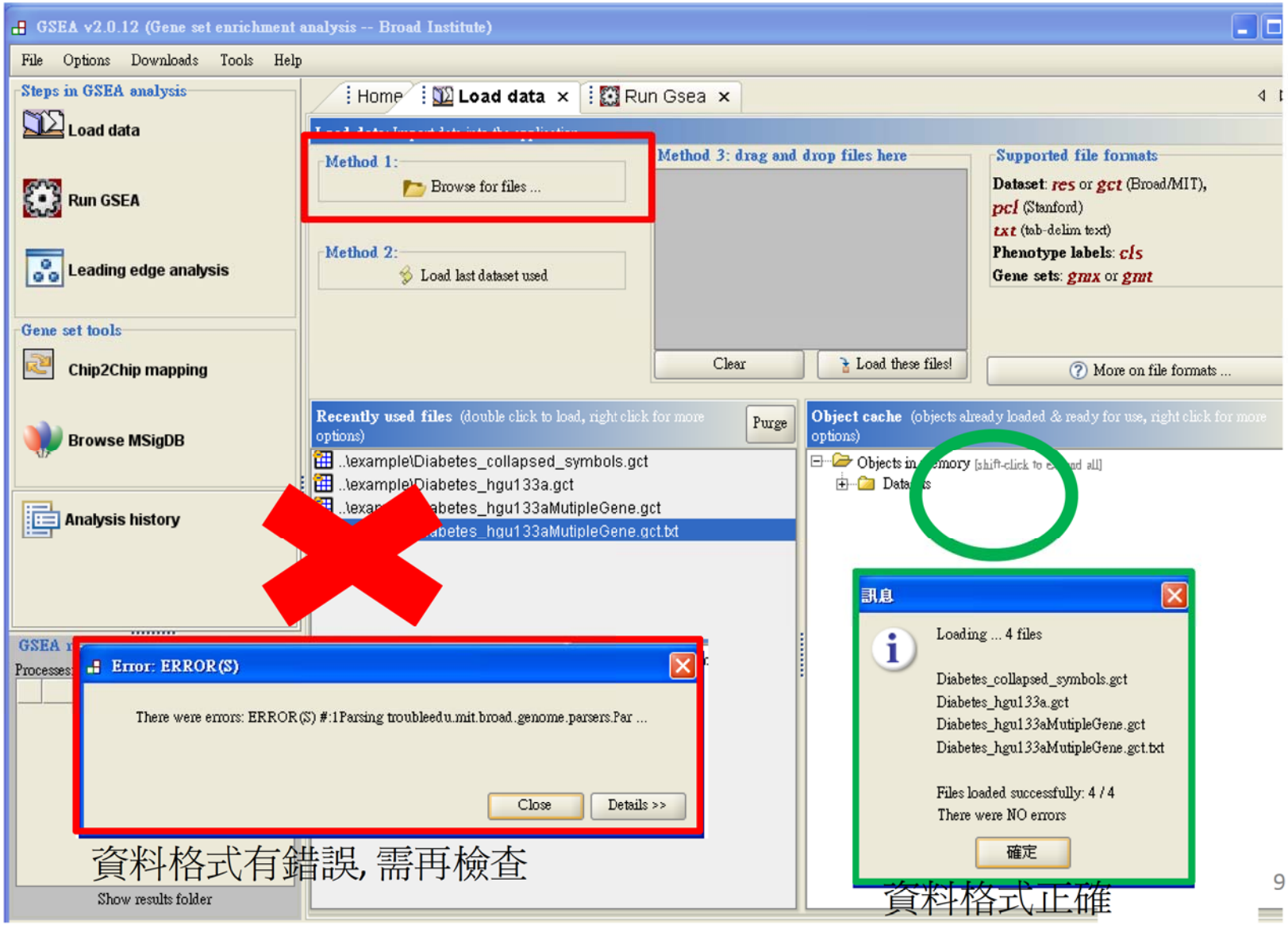
	A	B	C
1	Probe Set ID	Gene Symbol	Gene Title
2	A_23_P100001	LOC400451	hypothetical gene supporte
3	A_23_P100011	AP3S2	adaptor-related protein con
4	A_23_P10002	CEP63	centrosome protein Cep63
5	A_23_P100022	SV2B	synaptic vesicle glycoprote
6	A_23_P100033	NA	na
7	A_23_P100059	NA	na
8	A_23_P100065	LOC388137	similar to Golgi autoantiger
9	A_23_P100074	AVEN	apoptosis, caspase activat
10	A_23_P100088	PSMA4	proteasome (prosome, ma
11	A_23_P100092	FLJ35867	hypothetical protein FLJ35
12	A_23_P100103	VPS39	vacuolar protein sorting 39
13	A_23_P100115	NA	na
14	A_23_P100127	CASC5	cancer susceptibility candi
15	A_23_P100133	NA	na
16	A_23_P100141	C16ORF28	chromosome 16 open read
17	A_23_P100155	FLJ20507	hypothetical protein FLJ20
18	A_23_P100177	MMP15	matrix metalloproteinase 1
19	A_23_P100189	PRM1	protamine 1
20	A_23_P100196	USP10	ubiquitin specific protease

Column 1: probe set identifiers. These correspond to the features on the microarray. Must be unique

Column 2: Gene symbol of the probe set. Enter NA if there is none. Need *not be unique*

Column 3: Gene title of the probe set. Enter NA if there is none

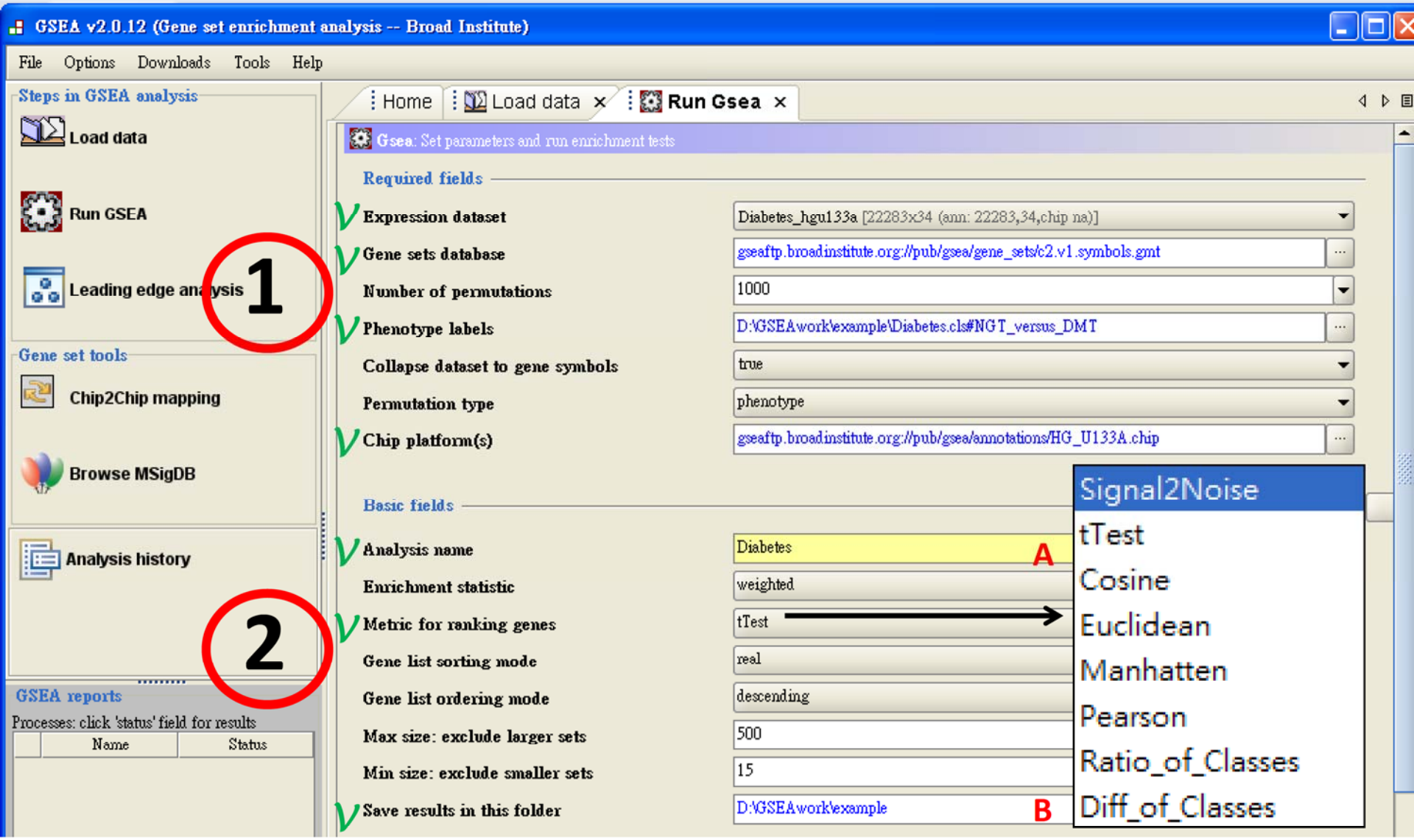
Step3. Loading data

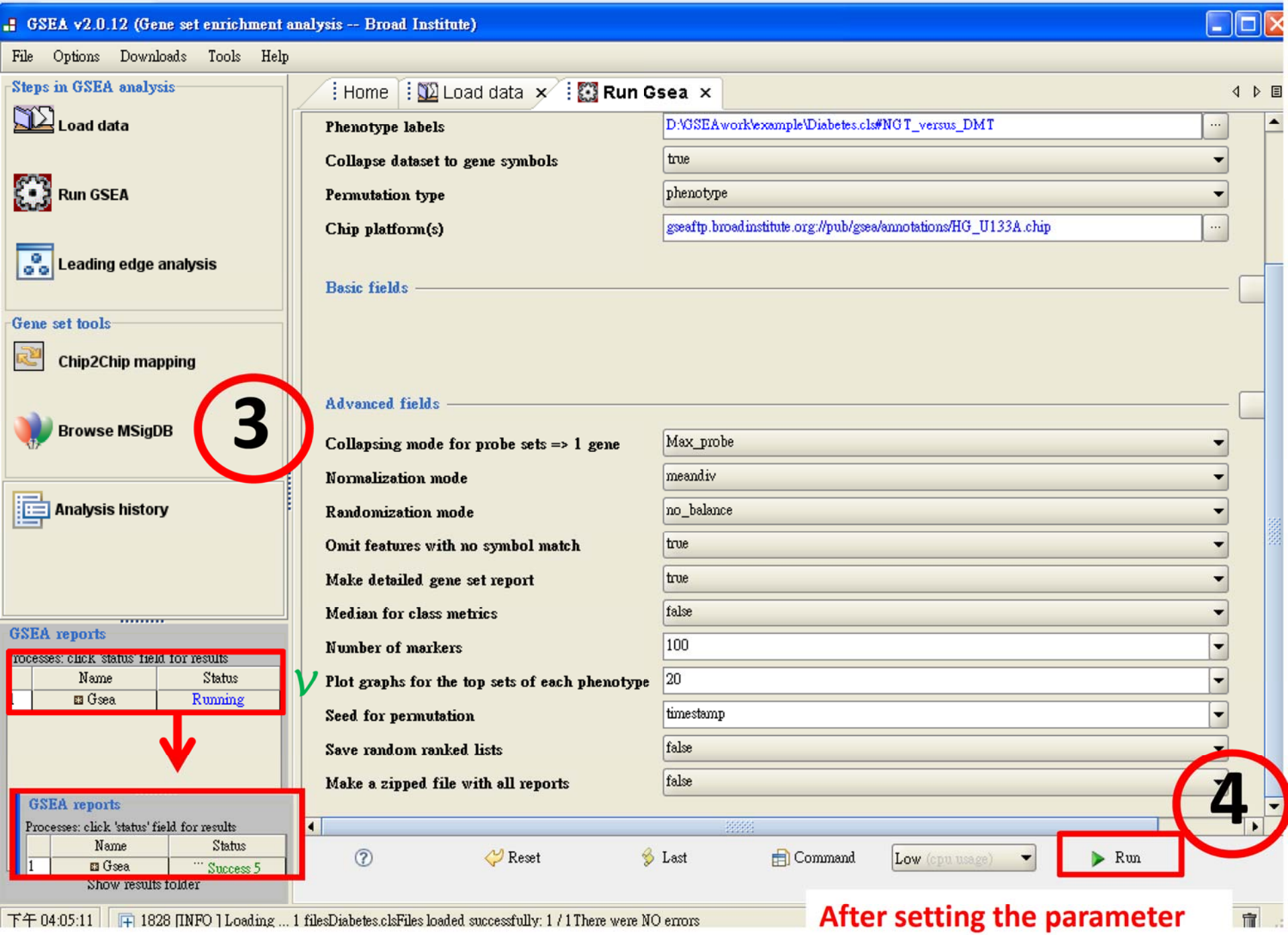


資料格式有錯誤, 需再檢查

資料格式正確

Step4. Running analysis




GSEA v2.0.12 (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
3
Analysis history
GSEA reports
rocesses: click 'status' field for results
Name
Status
Gsea
Running
GSEA reports
Processes: click 'status' field for results
Name
Status
1
Gsea
Success 5
Show results folder
Home
Load data
Run Gsea
Phenotype labels
D:\GSEAwork\example\Diabetes.cls#NGT_versus_DMT
Collapse dataset to gene symbols
true
Permutation type
phenotype
Chip platform(s)
gseaftp.broadinstitute.org://pub/gsea/annotations/HG_U133A.chip
Basic fields
Advanced fields
Collapsing mode for probe sets => 1 gene
Max_probe
Normalization mode
meandiv
Randomization mode
no_balance
Omit features with no symbol match
true
Make detailed gene set report
true
Median for class metrics
false
Number of markers
100
Plot graphs for the top sets of each phenotype
20
Seed for permutation
timestamp
Save random ranked lists
false
Make a zipped file with all reports
false
4
Reset
Last
Command
Low (cpu usage)
Run
下午 04:05:11
1828 [INFO] Loading ... 1 filesDiabetes.clsFiles loaded successfully: 1 / 1 There were NO errors
After setting the parameter

Step5. Viewing analysis results

GSEA Report for Dataset Diabetes_hgu133a

Enrichment in phenotype: NGT (17 samples)

- 109 / 305 gene sets are upregulated in phenotype **NGT**
- 3 gene sets are significant at FDR < 25%
- 1 gene sets are significantly enriched at nominal pvalue < 1%
- 3 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: DMT (17 samples)

- 196 / 305 gene sets are upregulated in phenotype **DMT**
- 0 gene sets are significantly enriched at FDR < 25%
- 2 gene sets are significantly enriched at nominal pvalue < 1%
- 7 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

[檔案位置 D:\GSEAwork\example\Diabetes.Gsea.1376985908703**index.html**]

B A GSEA產生

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

•**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.

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

•**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.

•**FDR q val**: False discovery rate; that is, the estimated probability that the normalized enrichment score represents a false positive finding.

Table: Gene sets enriched in phenotype NGT (17 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	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%

[檔案位置 D:\GSEAwork\example\Diabetes.Gsea.1376985908703**gsea_report_for**_NGT_1376985908703.**html** or **gsea_report_for**_NGT_1376985908703.**xls**]

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

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	0.18389146
FWER p-Value	0.296

GSEA Results Summary

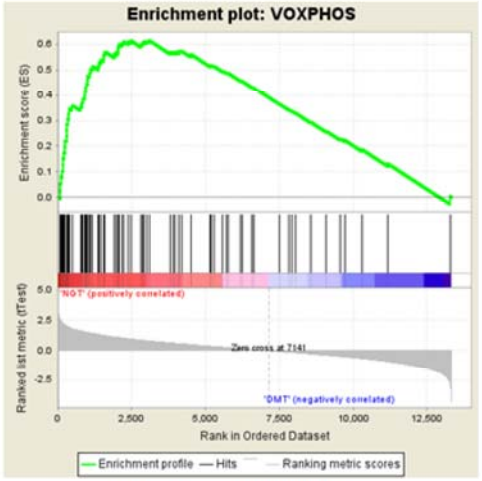
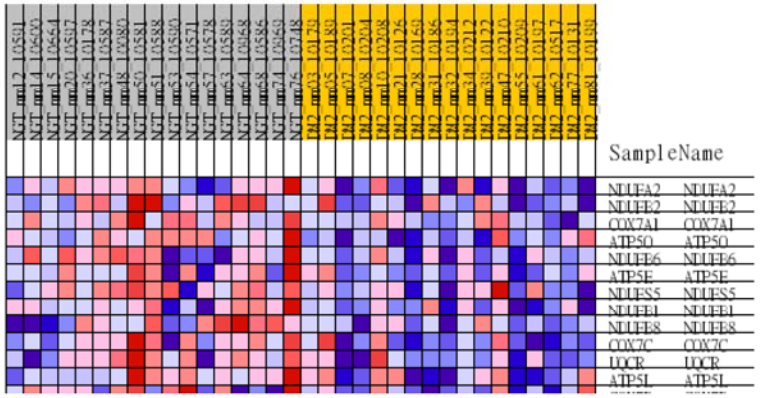


Fig 1: Enrichment plot: VOXPHOS
Profile of the Running ES Score & Positions of GeneSet Members on the Rank Ordered List

Enrichment plot



Heat map

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	0.0632	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	2.286	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

GSEA details

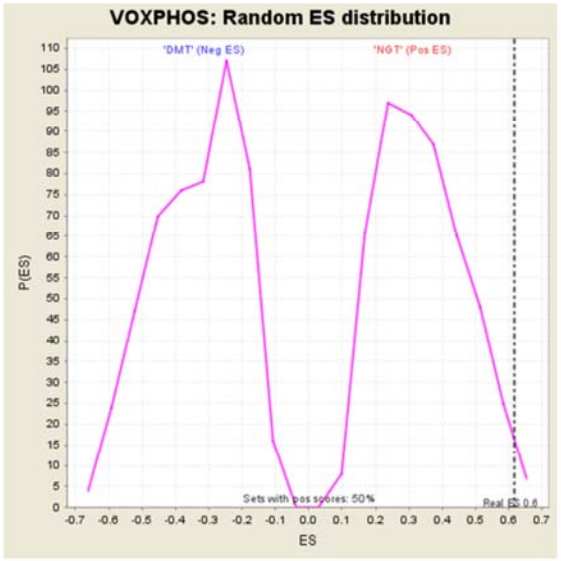
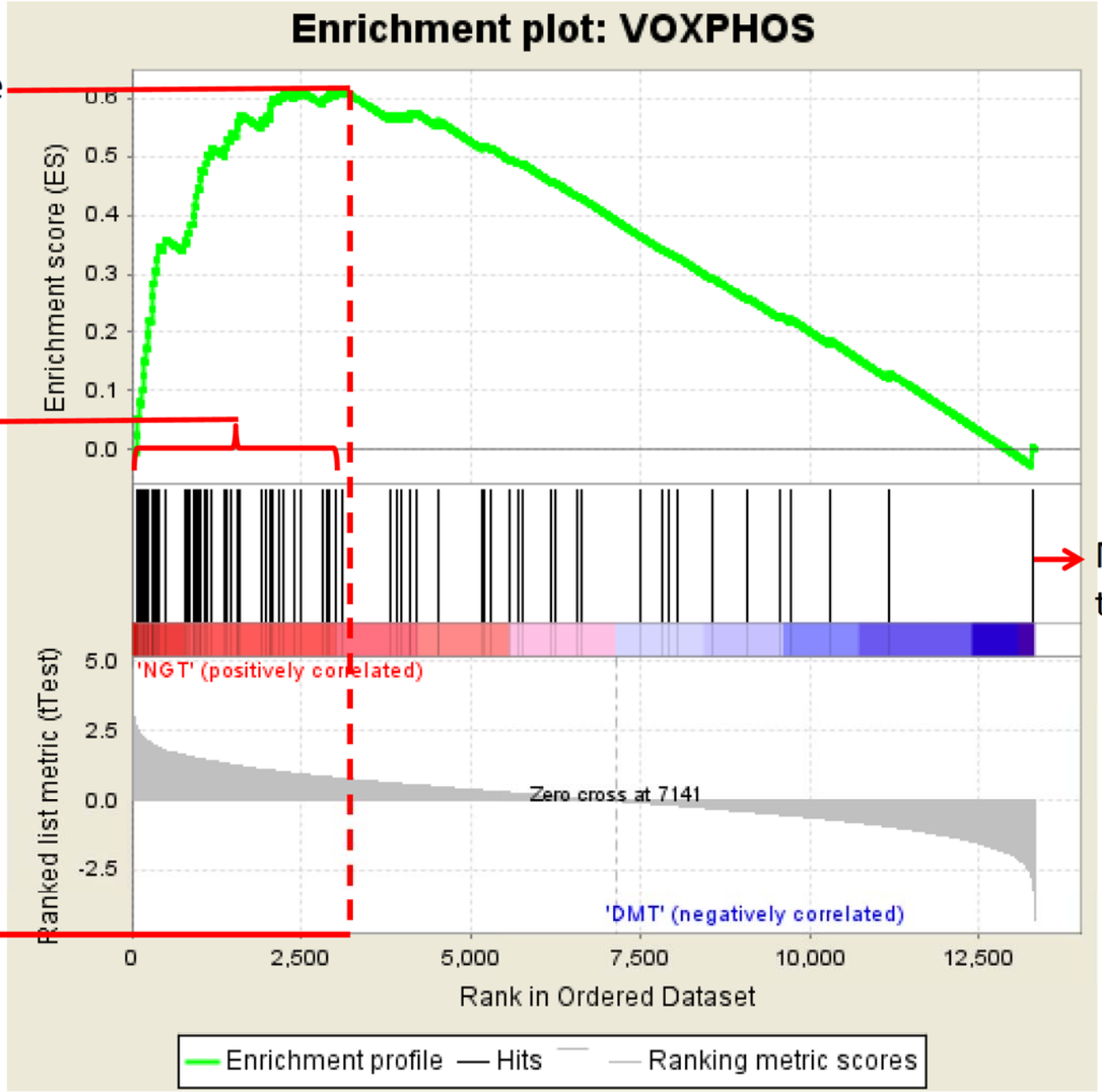


Fig 3: VOXPHOS: Random ES distribution
Gene set null distribution of ES for VOXPHOS

Random ES distribution


Enrichment plot: VOXPHOS
Enrichment score
Core gene
Rank at max
Member of the gene set
Enrichment score (ES)
0.6
0.5
0.4
0.3
0.2
0.1
0.0
'NGT' (positively correlated)
Ranked list metric (tTest)
5.0
2.5
0.0
-2.5
Zero cross at 7141
'DMT' (negatively correlated)
0
2,500
5,000
7,500
10,000
12,500
Rank in Ordered Dataset
Enrichment profile
Hits
Ranking metric scores

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

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

ES(S)=P_{hit}-P_{miss}

N: the number of genes in the gene list L={g_1,...,g_N}

N_H: the number of genes in the gene set S

p=1: weighted by correlation

r_j=r(g_j)

從0開始,
如果有遇到交集的gene, 往上爬,
爬的高度是

$$\frac{\text{該gene的 score絕對值}}{\text{所有交集 gene的score絕對值總和}}\quad \text{ex}:\ \frac{1.4}{5}$$

否則爲往下跌, 跌的高度是

$$\frac{1}{\text{Input gene數-交集gene數}}\quad \text{ex:}\ \frac{1}{25-5}=0.05$$

取最高或最低的高度爲Enrichment score (ES)

粗黑線代表input 與 gene set 兩邊都有的gene=交集

假設全部input 25 genes, 對應的t檢定量(score)由大到小排序如下

1.5 1.4 1.3 1.2 1.1 1.0 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1 0.0 -0.1 -0.2 -0.3 -0.4 -0.5 -0.6 -0.7 -0.8 -0.9 -1.0 -1.1 -1.2 -1.3 -1.4 -1.5

Sample 1 2 3 ... 32 33 34

Data 1 2 3 ... 32 33 34 ➡ **ES**

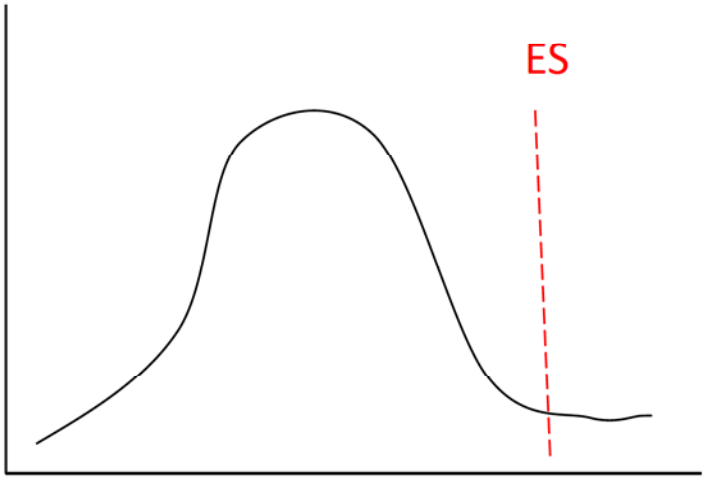
Random permutation:

Data' 1 2 8 32 ... 12 9 20 ➡ ES'1 ➡ NES'1

Data' 2 31 18 22 ... 25 9 21 ➡ ES'2 ➡ NES'2

⋮

Data' 1000 2 17 19 ... 34 15 18 ➡ES'1000➡NES'1000



Distribution of 1000 ES'

NES (Normalized enrichment score)= $\frac{\text{Actual ES}}{\text{mean(ES')}}$

Nominal p value= $\frac{\text{ES'大於(小於)ES的數目}}{\text{ES'與ES同號的數目}}$

FDR(False discovery rate):根據所有gene set 的NES及NES'而來代表此gene set 的NES 是個錯誤發現(false discovery)的機率

GSEA建議值是FDR<0.25

Reference

- Subramanian *et al.* (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. PNAS 102 (43) 15545-15550.
- Mootha *et al.* (2003) PGA-1α -responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet 34, 267-273.
- GSEA http://www.broadinstitute.org/gsea/index.jsp