

GSEA 分析軟體操作與分析步驟-焜慕



1. 註冊與下載軟體

GSEA home page 以信箱進行註冊

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

註冊後至 download 頁面安裝 GSEA 軟體

<http://software.broadinstitute.org/gsea/downloads.jsp>

目前是 GSEA v4.0.2 版本

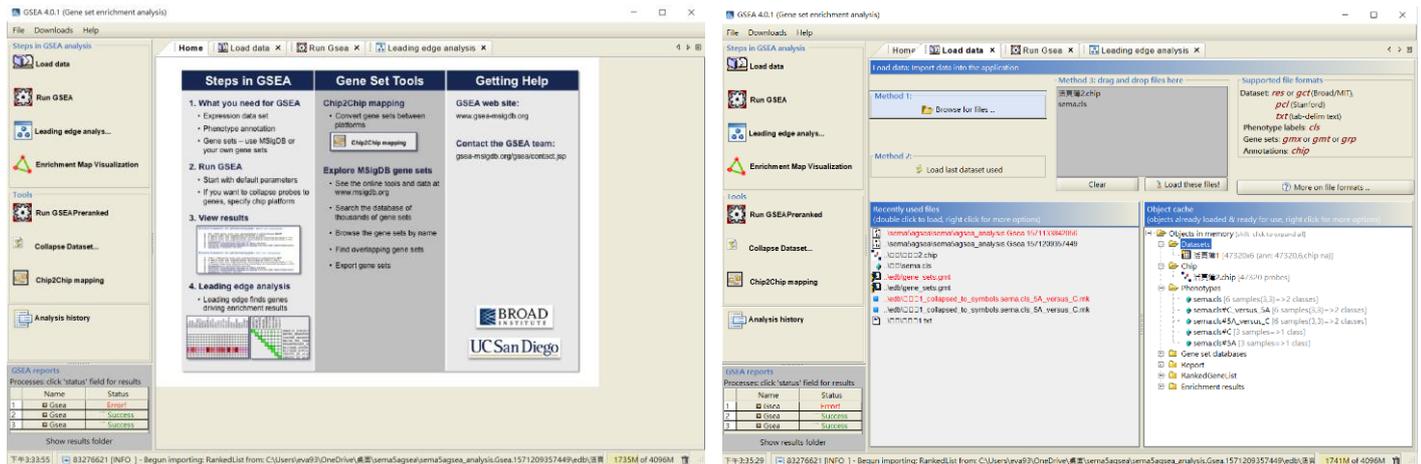
2. 準備分析檔案

軟體中 Home 頁面左側有分析所需的功能列表，基本上是由上往下依序進行

Load data->Run GSEA->Leading edge analysis

擊點 Load data 出現分頁，右上小方框有 GSEA 分析所需的特定格式

須將檔案依照規定存成 dataset/phenotype/annotation 三種不同的檔案上傳



Load data

分析檔案格式參考說明：<https://www.jianshu.com/p/aab52528c1e2>

① Dataset：一般分析的 dataset 用.txt 檔最簡單

第一列基因名(必須大寫字母)或是 probe ID (避免基因重複認到)

第二列 DESCRIPTION 都打 na 就好，後面幾列是不同組別分析結果

以 overexpress sema5a 與 control 組做 microarray 的 data 為例

另開一個 excel 活頁簿貼上紅框中要選取的数据範圍再轉成.txt 檔

Ex.

NAME	DESCRIPTION	TACC2	BT-549	CCRFL-DEM	COLO 205	EVKX	HCC-2998	HCC
TACC2	na	46.05	82.17	16.87	98.6	141.02	114.32	12.0288204
AGEF1	na	108.34	59.04	25.61	33.11	42.53	9.12	12.9224918
AGEF2	na	42.2	25.75	76.01	40.41	32.17	48.26	12.9224918
AGEF3	na	7.43	13.94	8.65	21.13	15.09	19.05	12.9224918
AGEF4	na	11.4	3	3.16	2.34	4.43	1.56	12.9224918
AGEF5	na	148.09	317.17	316.66	147.23	125.76	261.39	12.9224918
AGEF6	na	8.62	9.12	1672.83	5.91	5.31	11.96	12.9224918
AGEF7	na	206.74	136.71	141.34	129.09	138.01	138.16	12.9224918
AGEF8	na	7.94	36.68	7.8	1.97	7.75	4.73	12.9224918
AGEF9	na	10.45	8.5	4.05	4.77	2.35	3.72	12.9224918
AGEF10	na	4.22	5.16	3.95	37.56	110.36	203.25	12.9224918
AGEF11	na	209.99	253.07	90.36	61.83	360.49	145.01	12.9224918
AGEF12	na	3.35	5.20	2.90	4.82	4.36	1.45	12.9224918

② Phenotype : 用來做類別分類 · 需存成.cls 檔

先在記事本裡面打完存檔再改附檔名成.cls 檔

第一行 total sample 數(空格)類別數目(空格)1(打 1 就對了)

第二行#(空格)分類 0 的名稱(空格)分類 1 的名稱

第三行跟 dataset 相對應的分類分組 Ex.分類 0 有三重複就打 0 0 0
以 overexpress sema5a 與 control 組做 microarray 的 data 為例

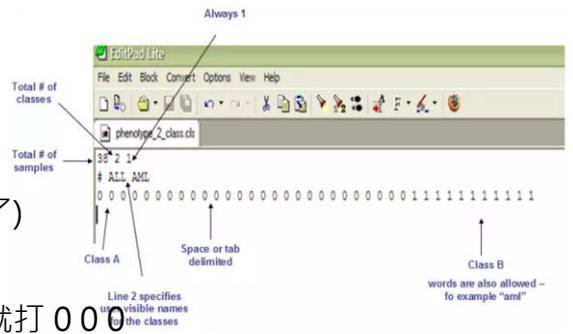
第一行 Total_6 組(control 3 組+overexpress 5a 3 組)/分 control 跟 overexpress 2 組/就打 1

第二行#/control 組 C /overexpress 組 5A

第三行 0 代表 control 組 1 代表 overexpress 5a 組 依照 dataset 排序所以打 0 0 0 1 1 1

Ex.

```
sema.cls - 記事本
檔案(F) 編輯(E) 格式(O) 檢視(V) 說明(H)
6 2 1
# C 5A
0 0 0 1 1 1
```



③ Gene sets : 連網路的話 GSEA 軟體裡面就有 · 不用弄

④ Annotations : 對 dataset 的說明 · 存成.chip 檔

先在 excel 中打完轉記事本存檔再改成.chip 附檔名

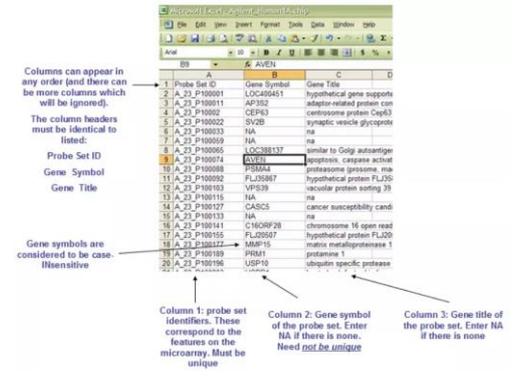
第一行是每一行的名稱

必須是 Probe Set ID/Gene Symbol/Gene Title 三列

第一列依 dataset/第二列打 na 就好/第三列沒用不用打

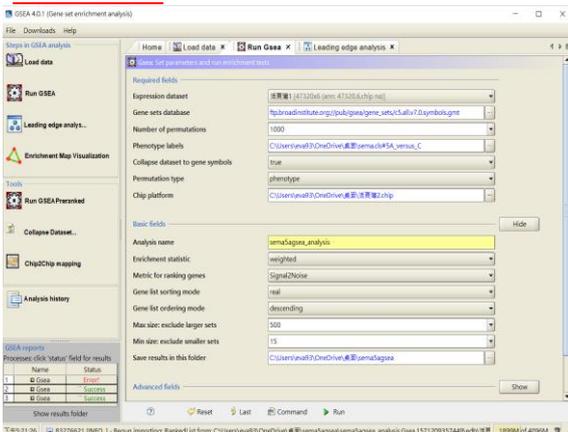
Ex.

```
活頁簿2.chip - 記事本
檔案(F) 編輯(E) 格式(O) 檢視(V) 說明(H)
Probe Set ID Gene Symbol Gene Title
2600747 IFIT2 na
1690066 MX1 na
5360156 IFITM1 na
```



⑤ 三個不同格式的檔案存好之後上傳到 GSEA 軟體 · 軟體自己會認哪個檔是幹嘛的

Run GSEA



Expression dataset:基本上軟體自己會選

Gene set database: c5 的都可以選

Permutations: data 打亂重排次數 >100

Phenotype:選讓誰當基準

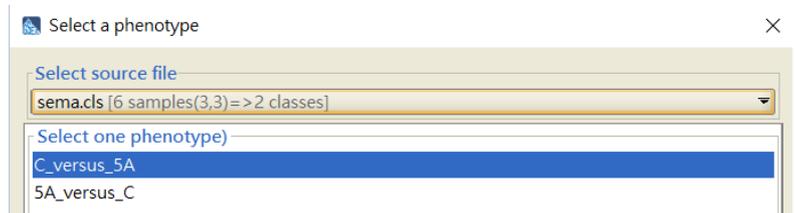
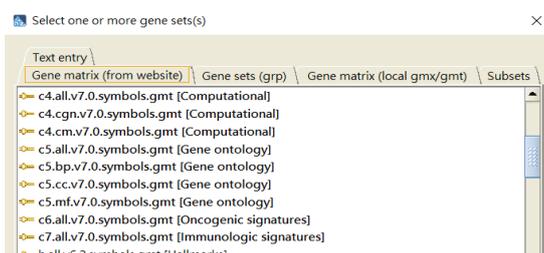
Chip: 選剛才存的 chip 檔

Max/Min size:設定要分析的 gene set 的大小

數字代表基因數目

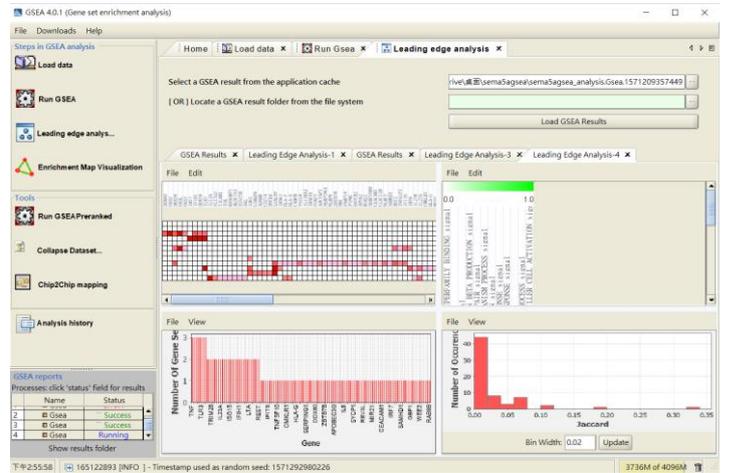
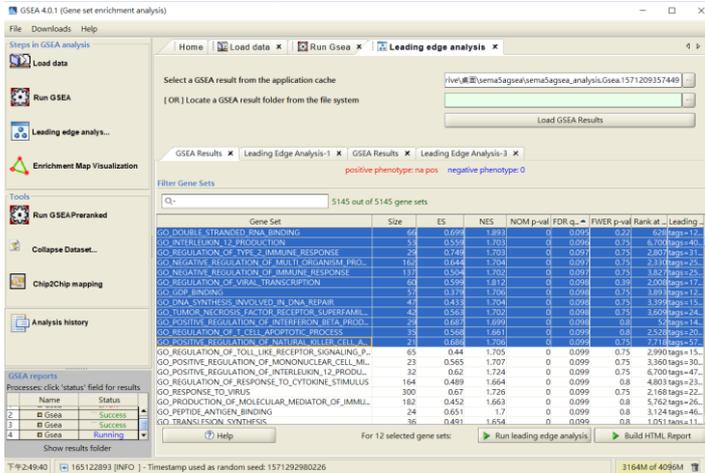
分析檔命名跟存檔位置記得設定

→Run

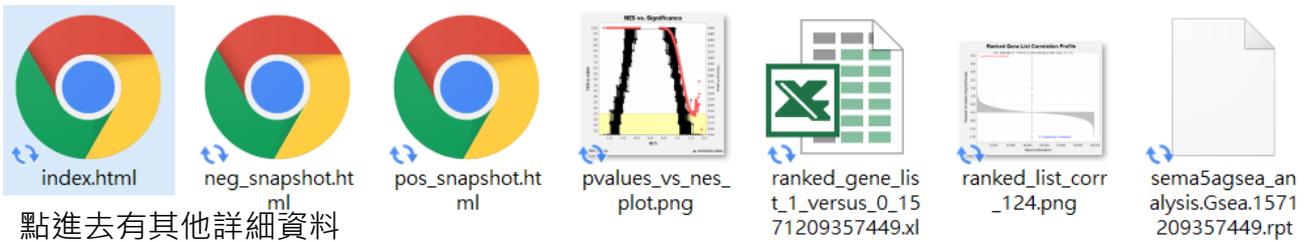


Leading edge analysis

選取想看的 gene set -> Run
看細部基因跟 gene set 分析結果



分析結果的資料夾中只有前 20 名的 GSEA 圖
裡面有一個 index.html 連結



點進去有其他詳細資料



Enrichment in phenotype: 1 (3 samples)

- 3192 / 5145 gene sets are upregulated in phenotype 1
- 115 gene sets are significant at FDR < 25%
- 779 gene sets are significantly enriched at nominal pvalue < 1%
- 779 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: 0 (3 samples)

- 1953 / 5145 gene sets are upregulated in phenotype 0
- 0 gene sets are significantly enriched at FDR < 25%
- 124 gene sets are significantly enriched at nominal pvalue < 1%
- 124 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 47320 native features
- After collapsing features into gene symbols, there are: 34691 genes

Gene set details

- Gene set size filters (min=15, max=500) resulted in filtering out 4851 / 9996 gene sets
- The remaining 5145 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 1 versus 0 comparison

