GSEA分析

软件和基因集下载

(http://software.broadinstitute.org/gsea/downloads.jsp)

Downloads

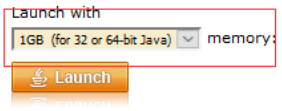
Software

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 8.

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.

javaGSEA Desktop Application

- Easy-to-use graphical user interface.
- Runs on any desktop computer (Windows, macOS, Linux etc.) that supports Java 8. **Oracle Java is recommended as there are known issues when running with OpenJDK.**
- Produces richly annotated reports of enrichment results.
- This release is open source under a BSD-style license. The source is available on our GitHub repository. The changes are noted in the Release Notes.
- **We recommend using a memory configuration smaller than your computer's total memory.**



根据内存大小选择

MSigDB

Use the following links to download individual gene set collections or the complete Molecular Signatures Database (MSigDB). For details on the MSigDB gene set collections refer to the Molecular Signatures Database page.

See the license terms page for details about the license for MSigDB. Please note that the license terms vary for different versions of MSigDB, and that certain gene sets have special access terms.

All gene sets	Current MSigDB gene sets, gene symbols	msigdb.v6.1.symbols.gmt
	Current MSigDB gene sets, Entrez IDs	msigdb.v6.1.entrez.gmt
	Current MSigDB xml file	msigdb_v6.1.xml
h: hallmark gene sets	hallmark gene sets, gene symbols	h.all.v6.1.symbols.gmt
	hallmark gene sets, Entrez IDs	h.all.v6.1.entrez.gmt
c1: positional gene sets	positional gene sets, gene symbols	c1.all.v6.1.symbols.gmt
	positional gene sets, Entrez IDs	c1.all.v6.1.entrez.gmt
c2: curated gene sets	all curated gene sets, gene symbols	c2.all.v6.1.symbols.gmt
	all curated gene sets, Entrez IDs	c2.all.v6.1.entrez.gmt
	chemical and genetic perturbations, gene symbols	c2.cgp.v6.1.symbols.gmt
	chemical and genetic perturbations, Entrez IDs	c2.cgp.v6.1.entrez.gmt
	all canonical pathways, gene symbols	c2.cp.v6.1.symbols.gmt
	all canonical pathways, Entrez IDs	c2.cp.v6.1.entrez.gmt
	BioCarta gene sets, gene symbols	c2.cp.biocarta.v6.1.symbols.gmt
	BioCarta gene sets, Entrez IDs	c2.cp.biocarta.v6.1.entrez.gmt
	KEGG gene sets, gene symbols	c2.cp.kegg.v6.1.symbols.gmt
	KEGG gene sets, Entrez IDs	c2.cp.kegg.v6.1.entrez.gmt
	Reactome gene sets, gene symbols	c2.cp.reactome.v6.1.symbols.gmt
	Reactome gene sets, Entrez IDs	c2.cp.reactome.v6.1.entrez.gmt

输入数据准备

1. 表达矩阵。常见表达矩阵格式，tab键分割，txt格式，第一列为基因名字（名字与注释数据库一致，同为GeneSymbol或EntrezID或其它自定义

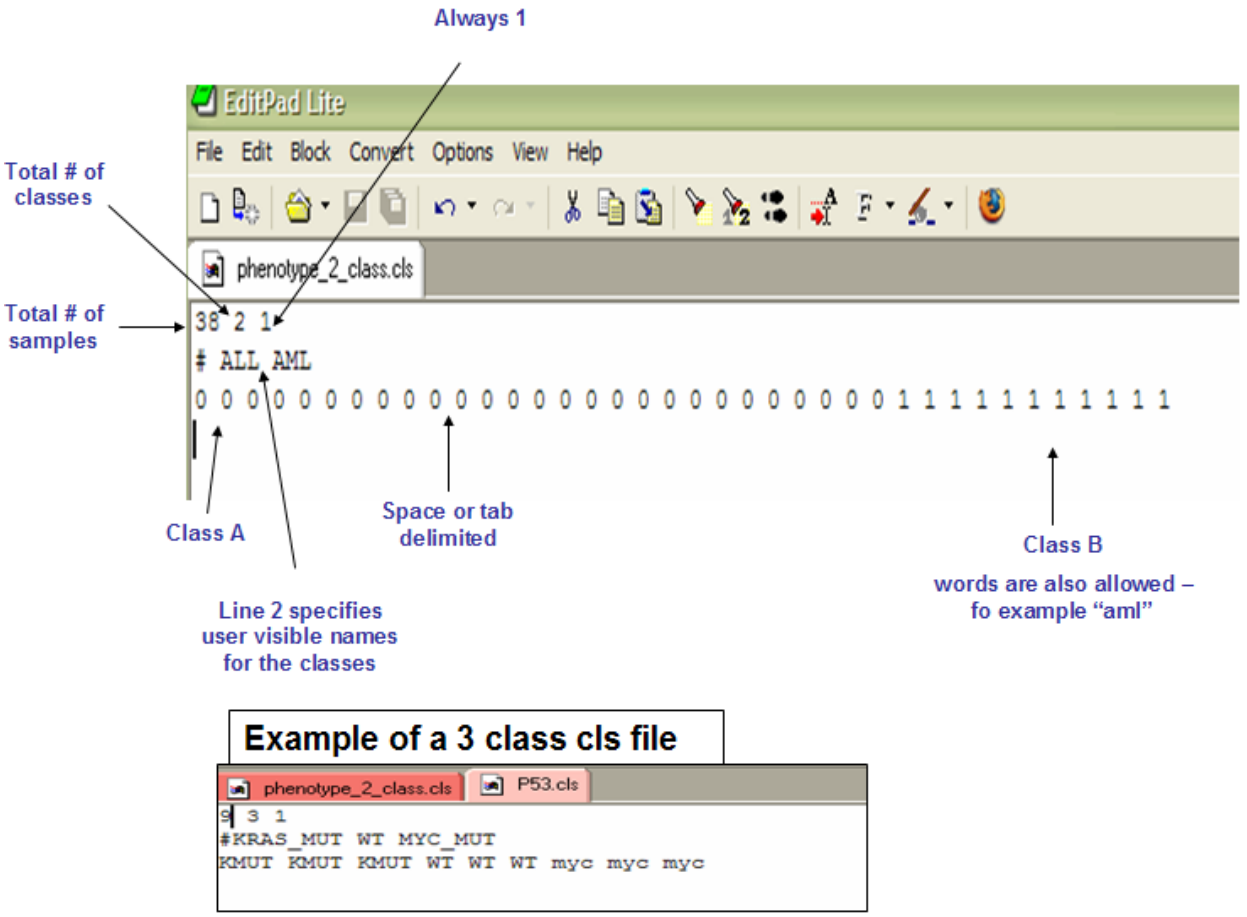
名字），第一行为标题行，含样品信息。也可为gct文件，具体见 http://blog.genesino.com/2014/08/GSEA-usages/

```
ID  f_1 f_2 f_3 Y1_1    Y1_2    Y1_3
255649  42264.8497601085    34960.3491875017    20153.8696454615    391055.501512589    920161.080463168    492945.415693969
9232    31127.7156601786    39178.9864019764    32017.1738625758    437593.259077238    721924.084757642    537023.27861813
4550    39454.5766445383    87503.9075549211    95525.8809782437    197324.703339763    124352.306902701    208507.203045576
2512    42641.9536235372    53883.3480357275    52036.8029831538    74771.6725348972    122966.049447768    168288.868064819
1786    197883.044068384    131242.939503093    144868.978007259    219565.413983912    244966.513907273    178927.788157455
9669    25342.738555255 22621.3624097788    20702.9619129067    298831.629508255    170770.135142098    108888.306611406
359787  35801.7651674336    28297.4459487126    40781.6245173191    112379.425961812    144415.985947764    109704.378213284
7114    8960.80315477245    5962.40407211025    5989.87691453068    189282.367839893    160968.248703679    124648.121444113
801 37528.2893421589    31491.8581680238    29870.9428201612    136639.006348896    131208.396249387    89683.3609961593
808 37528.2893421589    31491.8581680238    29870.9428201612    136639.006348896    131208.396249387    89683.3609961593
805 37528.2893421589    31491.8581680238    29870.9428201612    136639.006348896    131208.396249387    89683.3609961593
4512    22600.4120817441    59471.0190584785    54570.3907208647    93180.7096269669    101271.992138099    115325.094090363
494551  105044.829157163    85849.913406025 76465.3438005081    145714.553558747    159131.89350591 112450.668169716
340168  26203.6225101455    25085.3512261648    22928.4433856742    241690.84298678 217132.382304292    207981.027703163
8115    47837.1538753867    54643.6956748993    52224.4162468464    99296.4006969273    71822.7398211154    100800.291848916
4538    16453.2793746447    47117.0701626324    43538.4073445953    58986.7338024877    48742.4250565132    56956.8904762966
147945  89300.2332591961    75361.4686016353    72006.2940763658    130667.417330536    151160.913140044    102729.601437766
54969   23688.9172875331    19915.9394145861    18159.6700408669    114332.00876101 80508.6454153113    69475.1380440826
7784    12574.2055794828    14775.7717429756    14289.3377993462    95476.2553089761    43201.7545369533    69533.2990491162
126204  20780.1215409771    30218.0378387136    26499.5648187196    115052.519019017    97324.6458316779    84634.258746677
4519    7688.1625490029 21880.0574664002    20165.9998133727    41139.6947116632    47283.1493236865    33145.4115092456
10678   4546.30982018368    6874.41318134942    5881.51408119099    44096.6688105225    43550.498550773 36049.8266981127
402569  98944.5796330845    82604.493965839 83608.3953444597    54753.0155264407    66486.9564096746    51828.7256105967
4549    23072.6412440558    36804.0901980516    45636.9263932259    36767.6384660789    40901.1338707868    47000.4534271015
55246   24106.1096697768    20740.2161039209    19827.9724675817    83056.0994814576    63440.8954139686    54691.3375770959
4514    7510.14234680769    16319.5902947824    13779.8707470775    23333.0041952871    27506.0942649824    32982.7424482921
```

2. 样品分组信息

```
6 2 1
# fresh old
fresh fresh fresh old old old
```

分组信息示例



Example of a 3 class cls file

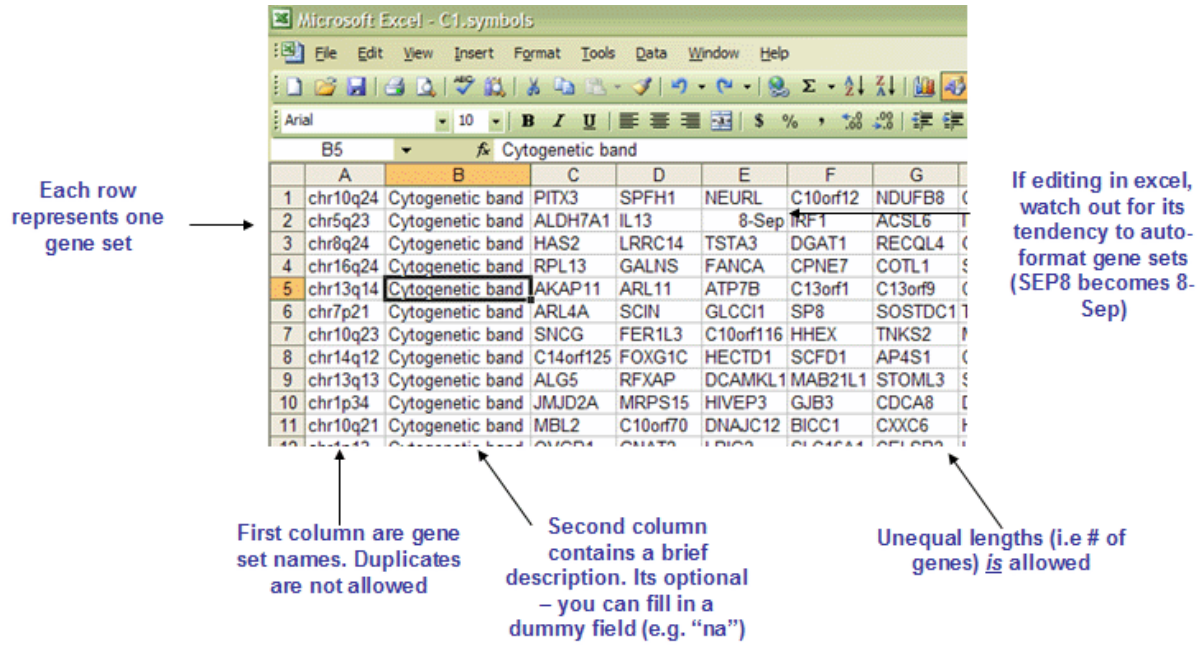
phenotype_2_class.cls | P53.cls

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

3. 基因集信息

```
AAANWWTGC_UNKNOWN   http://www.broadinstitute.org/gsea/msigdb/cards/AAANWWTGC_UNKNOWN   4208   481 6095   10370   351 4216   493 2904   10745   9874   8929   4620   6502   230
AAAYRNCTG_UNKNOWN   http://www.broadinstitute.org/gsea/msigdb/cards/AAAYRNCTG_UNKNOWN   4052   9842   23047   55800   1942   154807 10468   64236   51352   5339   11161   7515
MYOD_01 http://www.broadinstitute.org/gsea/msigdb/cards/MYOD_01 23630   84668   3151   26523   667 6860   5228   1942   154807 127833 9896   146325 84265   714 5905   9730
E47_01  http://www.broadinstitute.org/gsea/msigdb/cards/E47_01  90523   84668   26523   154807 1942   714 6097   5905   29116   5339   9589   10370   961 9815   23230   25837
CMYB_01 http://www.broadinstitute.org/gsea/msigdb/cards/CMYB_01 84668   285668 9218   10468   54482   11161   4601   387521 60509   10370   491 80314   7862   79598   27319   232
AP4_01  http://www.broadinstitute.org/gsea/msigdb/cards/AP4_01  4052   3151   5648   23623   6622   5339   4217   2923   25942   8482   2906   79727   5054   8645   57111
AACTTT_UNKNOWN  http://www.broadinstitute.org/gsea/msigdb/cards/AACTTT_UNKNOWN  4208   84962   100271849   84668   23047   685 6096   6095   154810   23266   351 79877   4216   229
MEF2_01 http://www.broadinstitute.org/gsea/msigdb/cards/MEF2_01 4208   56061   23421   2897   7134   1463   11136   7798   2257   125058 51305   288 2626   1901   162466   444
ELK1_01 http://www.broadinstitute.org/gsea/msigdb/cards/ELK1_01 108 11338   79753   9730   10370   55250   283234 51141   1241   23658   9815   9525   5608   153396 10489   578
SP1_01  http://www.broadinstitute.org/gsea/msigdb/cards/SP1_01  23630   1908   3151   9260   6620   140739 10492   10370   9653   9500   6432   9064   627 8729   1303   201
EVI1_06 http://www.broadinstitute.org/gsea/msigdb/cards/EVI1_06 285195   222663 9099   9586   136259 55421   10370   4893   40 51585   11183   400658 29843   27190   2904   147
ATF_01  http://www.broadinstitute.org/gsea/msigdb/cards/ATF_01  29944   3939   114 1046   27295   9986   154810 4967   11161   23607   64207   51726   56884   51631   53917   170
HOX13 01    http://www.broadinstitute.org/gsea/msigdb/cards/HOX13 01   8334   27327   83607   2770   23432   8543   800 5396   219409 4756   79933   51059   5127   51317   816
```

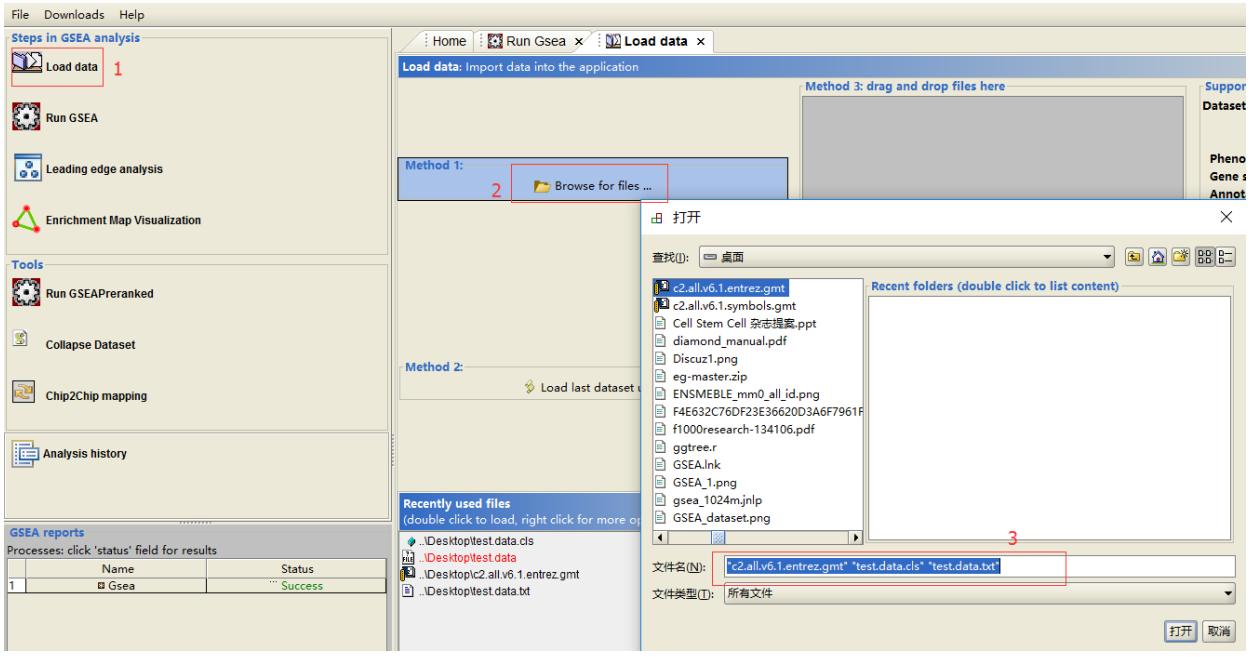
基因集信息示例



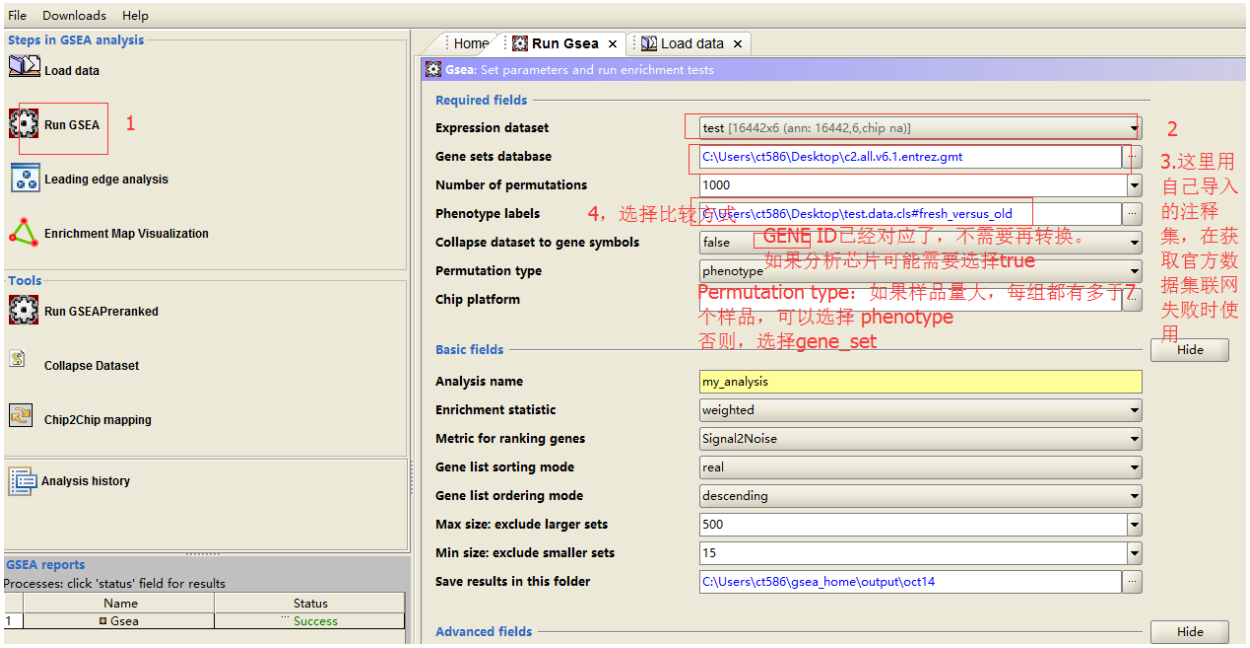
GMT format is convenient to store large databases of gene sets. For a handful of sets (<256) the gmx format offers greater excel-editability

软件运行 (每一步的步骤如有不明确的参考文后第一个链接)

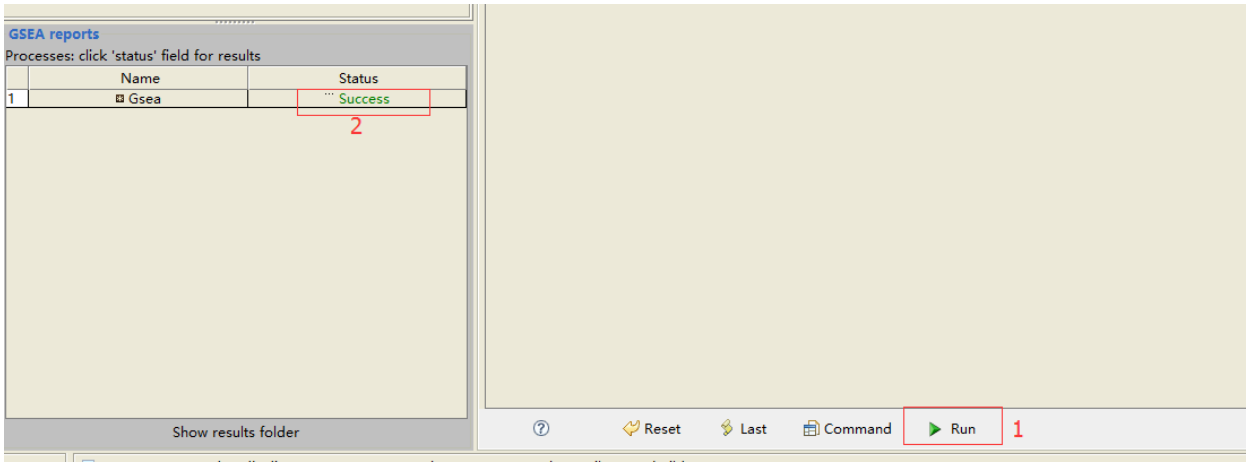
1. 导入数据



2. 运行GSEA (若每组样品都有多于7个样品，则Permutation type选择phenotype，结果理论上更好；否则选择gene_set)



3. 设置好参数后，点击正下方的run，等待运行结束，左侧出现success



4. 点击success，查看结果

Enrichment in phenotype: fresh (3 samples)

- 3013 / 3449 gene sets are upregulated in phenotype **fresh**
- 2188 gene sets are significant at FDR < 25%
- 1723 gene sets are significantly enriched at nominal pvalue < 1%
- 1723 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: old (3 samples)

- 436 / 3449 gene sets are upregulated in phenotype **old**
- 0 gene sets are significantly enriched at FDR < 25%
- 73 gene sets are significantly enriched at nominal pvalue < 1%
- 73 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 16398 features (genes)
- No probe set => gene symbol collapsing was requested, so all 16398 features were used

Gene set details

- Gene set size filters (min=15, max=500) resulted in filtering out 1289 / 4738 gene sets
- The remaining 3449 gene sets were used in the analysis
- List of gene sets used and their sizes (restricted to features in the specified dataset)

顺着网页的导航一步步去查看结果，有耐心就好。主要解释下，最常见的这种图。

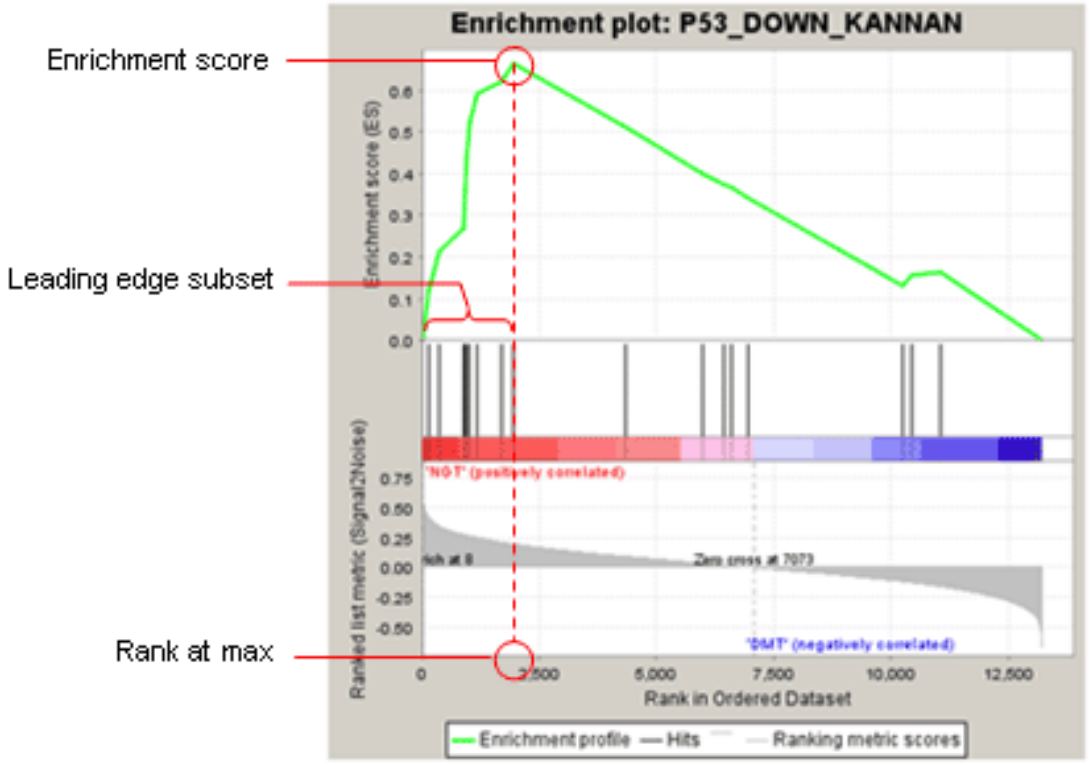
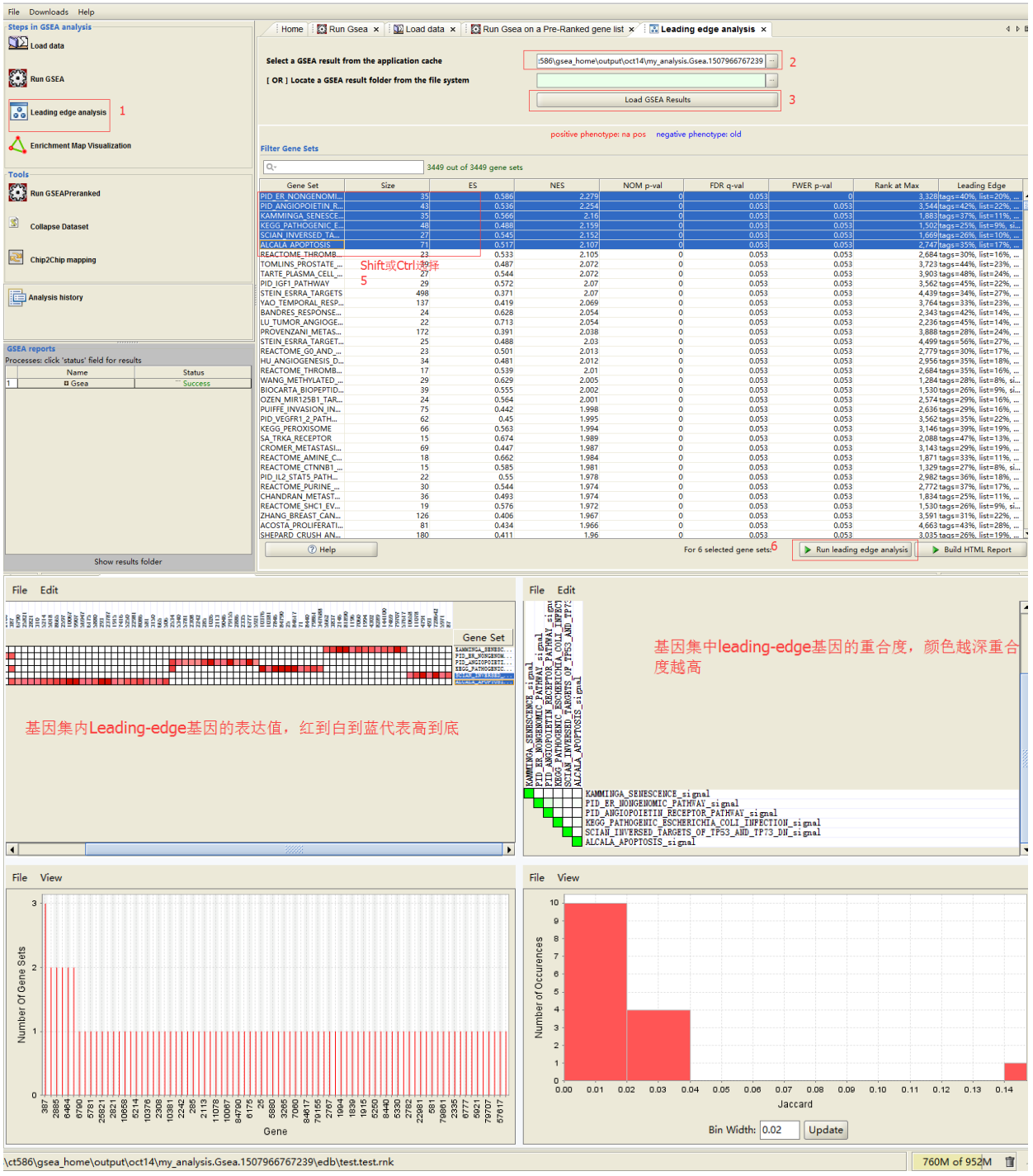


Fig 1: Enrichment plot: P53_DOWN_KANNAN
Profile of the Running ES Score & Positions of GeneSet Members on the Rank Ordered List

1. 图最上面部分展示的是ES的值计算过程，从左至右每到一个基因，计算出一个ES值，连成线。最高峰为富集得分(ES)。在最左侧或最右侧有一个特别明显的峰的基因集通常是感兴趣的基因集。
2. 图中间部分每一条线代表基因集中的一个基因，及其在基因列表中的排序位置。
3. 最下面部分展示的是基因与表型关联的矩阵，红色为与第一个表型(MUT)正相关，在MUT中表达高，蓝色与第二个表型(WT)正相关，在WT中表达高。
4. Leading-edge subset 对富集得分贡献最大的基因成员。若富集得分为正值，则是峰左侧的基因；若富集得分为负值，则是峰右侧的基因。
5. FDR GSEA默认提供所有的分析结果，并且设定FDR<0.25为可信的富集，最可能获得有功能研究价值的结果。但如果样品数目少，而且选择了gene_set作为Permumation type则需要使用更为严格的标准，比如FDR<0.05。

Leading-edge分析

主要对筛选感兴趣的基因有意义；选择一个或多个显著富集的基因集，查看其内Leading-edge基因的表达和重叠状态。



MSigDB

GSEA团队整理好的基因集，可用于注释，也可下载下来搜寻自己感兴趣的方向的基因作为一个补充。每个注释都提供了基于Gene Symbol和Entrez ID的索引表格。

参考

1. 较早记录的一篇GSEA的使用，有脚本可以转换表达矩阵为gct, cls文件作为GSEA的输入。文档为英文，但软件操作步骤还算详细，可配合着看。

http://blog.genesino.com/2014/08/GSEA-usages/

2. 最开始学习的教程，每一步操作都比较详细。

http://www.baderlab.org/Software/EnrichmentMap/Tutorial

3. GSEA软件和数据集下载
http://software.broadinstitute.org/gsea/downloads.jsp
4. 原文对GSEA原理的讲解是很清晰的，可以读下，关键的内容也都摘录在第一个链接里。 https://www.ncbi.nlm.nih.gov/pubmed/16199517