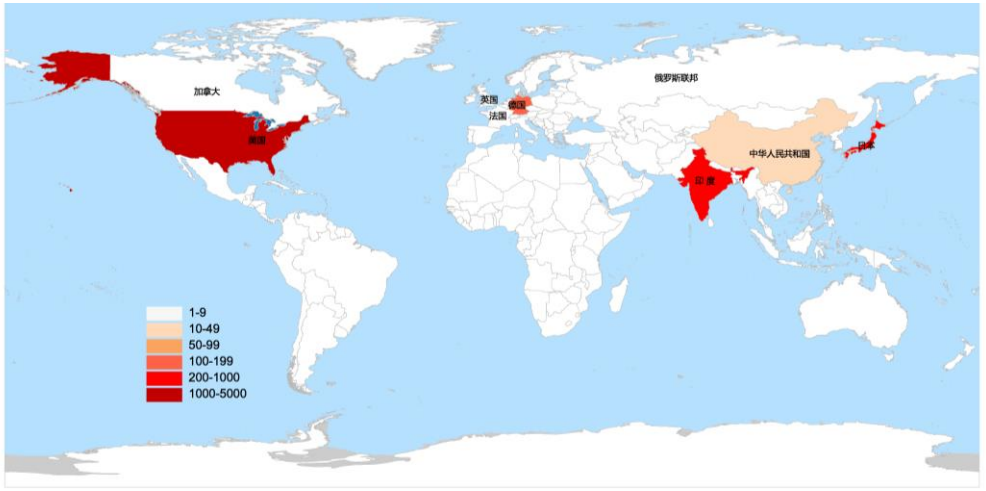
微生信
Wei Sheng Xin

基因集富集分析
Gene Set Enrichment Analysis

陈明杰 202411

基因集（gene set）概念

- 一群具有共同生物学属性或功能的基因集合
- 符合某个具体的标准：在同一条染色体上、同一个通路里边、某miRNA的靶基因，某细胞系中高表达的基因等。例如：铁死亡相关基因，铜死亡相关基因，乳酸化相关基因等等
- 用有限的基因，创造出无限的基因集（可能性）
- 以“跨国组织”类比，200个国家/地区，7000个跨国组织



基因集1：亚太经合组织
基因集2：东盟
基因集3：欧盟
基因集4：北约
基因集5：非盟
…
…

20世纪以来若干年份政府间国际组织（IGOs）数量表

年份	数量
1909	37
1951	123
1962	163
1970	242
1981	1039
1989	4068
1992	4878
1996	5885
2000	6556
2005	7350

数据来源：Union of International Associations

GSEA官网

https://www.gsea-msigdb.org/gsea/index.jsp

GSEA
Gene Set Enrichment Analysis

下载

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

UC San Diego

BROAD INSTITUTE

Overview

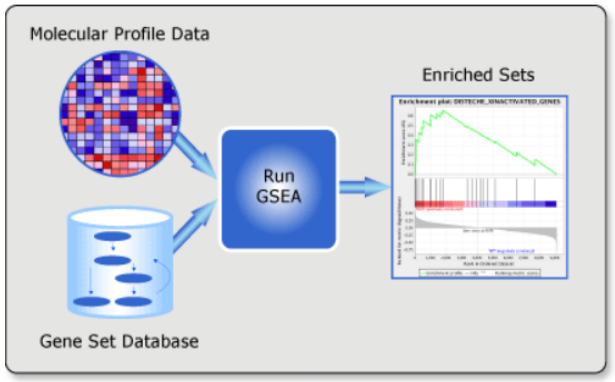
Gene Set Enrichment Analysis (GSEA) is a computational method that determines whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states (e.g. phenotypes).

- **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.
- View guidelines for **using RNA-seq datasets with GSEA**.
- Use the **GenePattern** platform to run analyses, including classical GSEA and a variation designed for single-sample analysis (**ssGSEA**).

What's New

9-Aug-2024: MSigDB 2024.1 provides collection updates for GO, Reactome, WikiPathways, and more along with numerous new set additions for Human and Mouse Databases. Additionally, gene data has been updated to Ensembl 112. See the release notes for details.

6-Feb-2024: GSEA 4.3.3 released. This is a minor release to update the OpenJDK distribution included in the platform bundles, to pick up recent OpenJDK improvements, bug fixes, and security patches. See the release



License Terms

GSEA and MSigDB are available for use under these license terms.

Please register to download the GSEA software, access our web tools, 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.

Citing GSEA

To cite your use of the GSEA software, a joint project of UC San Diego and Broad Institute, please reference Subramanian, Tamayo, et al. (2005, PNAS) and Mootha, Lindgren, et al. (2003, Nature Genetics).

Human Collections

H **hallmark gene sets** are coherently expressed signatures derived by aggregating many MSigDB gene sets to represent well-defined biological states or processes.

C1 **positional gene sets** corresponding to human chromosome cytogenetic bands.

C2 **curated gene sets** from online pathway databases, publications in PubMed, and knowledge of domain experts.

C3 **regulatory target gene sets** based on gene target predictions for microRNA seed sequences and predicted transcription factor binding sites.

C4 **computational gene sets** defined by mining large collections of cancer-oriented expression data.

C5 **ontology gene sets** consist of genes annotated by the same ontology term.

C6 **oncogenic signature gene sets** defined directly from microarray gene expression data from cancer gene perturbations.

C7 **immunologic signature gene sets** represent cell states and perturbations within the immune system.

C8 **cell type signature gene sets** curated from cluster markers identified in single-cell sequencing studies of human tissue.

Mouse Collections

MH **mouse-ortholog hallmark gene sets** are versions of gene sets in the MSigDB Hallmarks collection mapped to their mouse orthologs.

M1 **positional gene sets** corresponding to mouse chromosome cytogenetic bands.

M2 **curated gene sets** from online pathway databases, publications in PubMed, and knowledge of domain experts.

M3 **regulatory target gene sets** based on gene target predictions for microRNA seed sequences and predicted transcription factor binding sites.

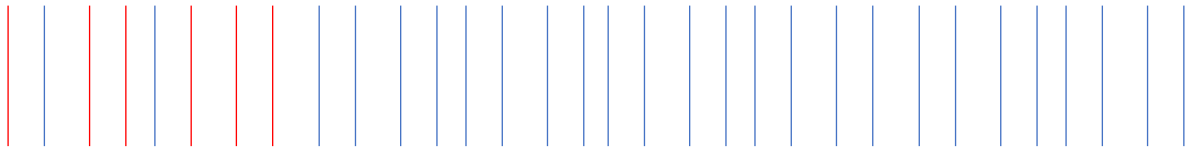
M5 **ontology gene sets** consist of genes annotated by the same ontology term.

M8 **cell type signature gene sets** curated from cluster markers identified in single-cell sequencing studies of mouse tissue.

sinobio@gmail.com

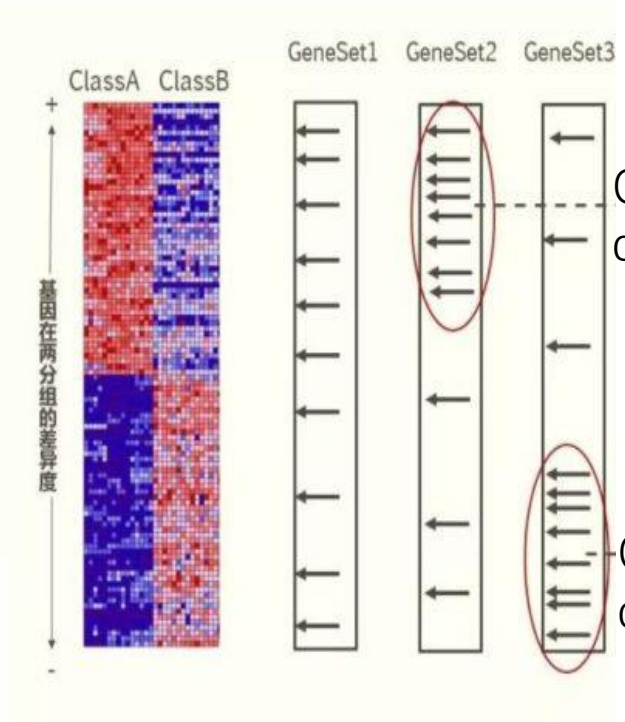
基因排排队

- 使用p值和fold change进行一刀切获得差异基因，然后进行富集分析的分析方法，往往富集不到我们感兴趣的结果
- 基因集富集分析（Gene set enrichment analysis，GSEA），它使用全部基因作为输入，找出具有协同差异 (concordant differences)的基因集，兼顾了差异较小的基因（因为在某些条件下，1.5倍的差异可能就算很大的了）
- 以学生成绩类比，按照成绩从高到低排序，红色表示学生会成员，他们的分数排序都在前面，说明学生会成员的分数普遍高（在高分处富集）



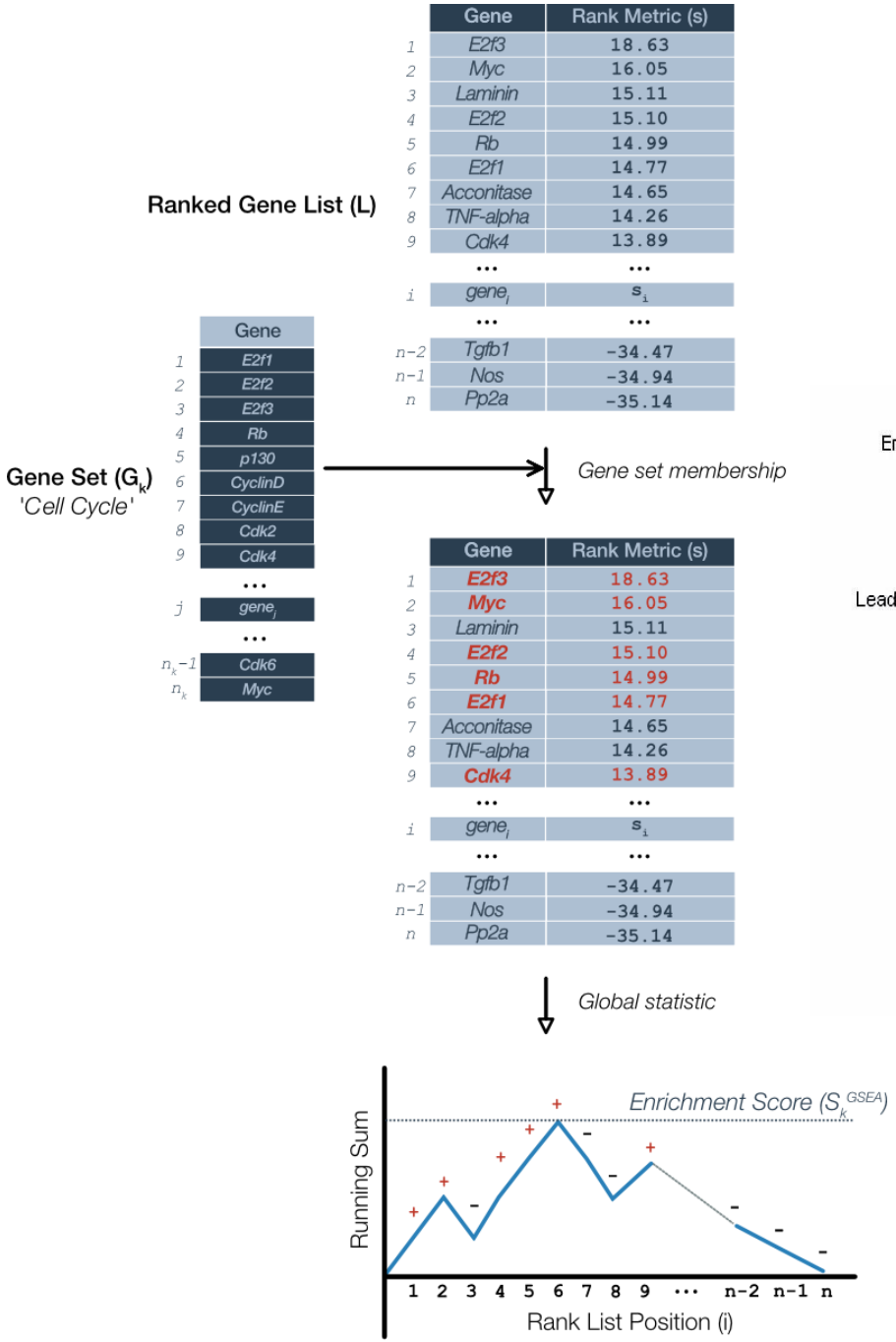
分数高　　分数低

基本原理



Geneset2在classA富集

Geneset3在classB富集



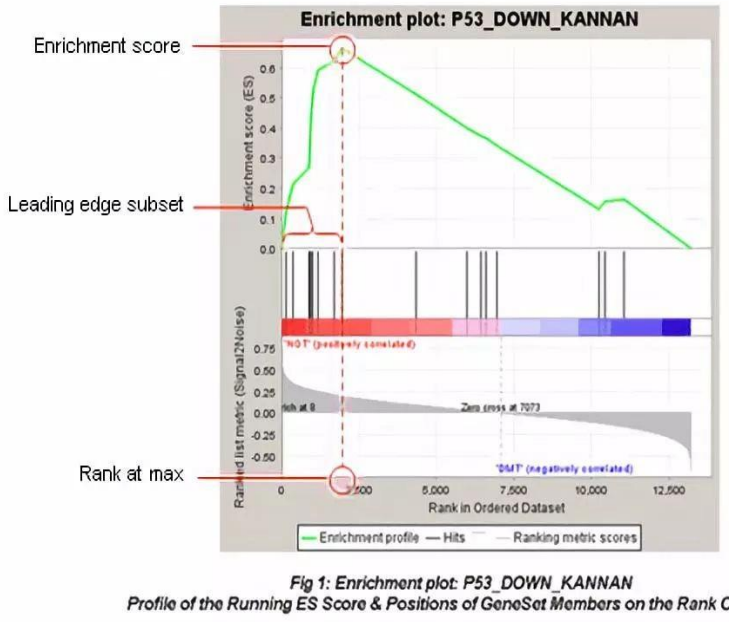
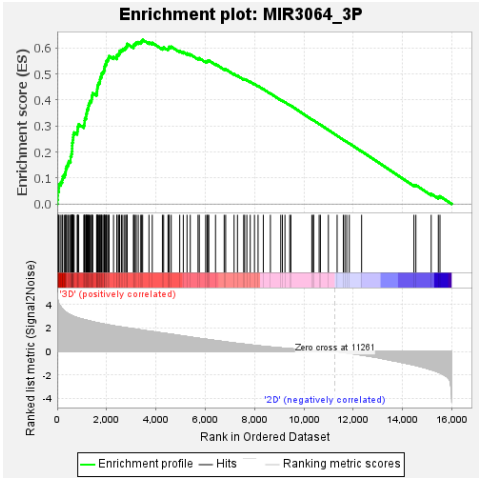


Fig 1: Enrichment plot: P53_DOWN_KANNAN
Profile of the Running ES Score & Positions of GeneSet Members on the Rank O

https://www.pathwaycommons.org/guide/primers/data_analysis/gsea/

官网GSEA分析

- 两种模式：表达谱（至少3 vs 3）、prerank（1vs1，2vs2，排序数据，与clusterProfile一样）
- 输入：
  - 1）表达谱（tpm，fpkm等，不用log2）
  - 2）表型信息（顺序问题）
  - 3）基因集（官网，或自定义）
- 输出：



	A	B	C	D	E	F	G
1	**gene_id**	**cancer-1**	**cancer-2**	**cancer-3**	**control-1**	**control-2**	**control-3**
2	TSPAN6	8.730898	9.890362	9.390107	2.222694	4.743941	3.45907
3	TNMD	0	0	0	0	0	0
4	DPM1	5.147331	3.756595	4.574771	1.331211	2.187636	1.630465
5	SCYL3	1.22566	1.035075	1.08619	0.201576	0.515936	0.258509
6	C1orf112	3.660326	3.202013	3.148393	1.076804	1.869494	1.185398

group.cls

```
6 2 1
# cancer control
cancer cancer cancer control control control
```

	A	B	C	D	E
1	EGFR tyro	https://www.genor	AKT3	BCL2L11	P3R3URF-
2	Long-tern	https://www.genor	RAPGEF3	ADCY1	ADCY8
3	Aminoacy	https://www.genor	FARSB	WARS2	FARS2
4	Biosynthe	https://www.genor	GNE	NANP	FCSK
5	Morphine	https://www.genor	GNB5	ADCY1	ADCY2
6	Parkinson	https://www.genor	NDUFC2-	PPIF	TRAP1
7	Galactose	https://www.genor	GALM	AKR1B1	G6PC1
8	Yersinia in	https://www.genor	AKT3	ARPC5	ARPC4
9	Starch an	https://www.genor	TREH	LOC1249(	LOC1249(

名字　来源　具体…基因

GSEA 4.2.3 (Gene set enrichment analysis)

File Downloads Help

Steps in GSEA analysis

- Load data
- Run GSEA 表达谱模式
- Leading edge analysis
- Enrichment Map Visualization

Tools

- Run GSEAPreranked Prerank模式
- Collapse Dataset
- Chip2Chip mapping

- Analysis history

GSEA reports

Processes: click 'status' field for results

Name	Status

Show results folder

Home

Steps in GSEA

1. What you need for GSEA

- Expression data set
- Phenotype annotation
- Gene sets – use MSigDB or your own gene sets

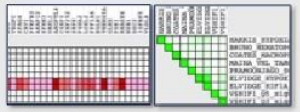
2. Run GSEA

- Start with default parameters
- If you want to collapse probes to genes, specify chip platform

3. View results

4. Leading edge analysis

- Leading edge finds genes driving enrichment results



Gene Set Tools

Chip2Chip mapping

- Convert gene sets between platforms

Explore MSigDB gene sets

- See the online tools and data at www.msigdb.org
- Search the database of thousands of gene sets
- Browse the gene sets by name
- Find overlapping gene sets
- Export gene sets

Getting Help

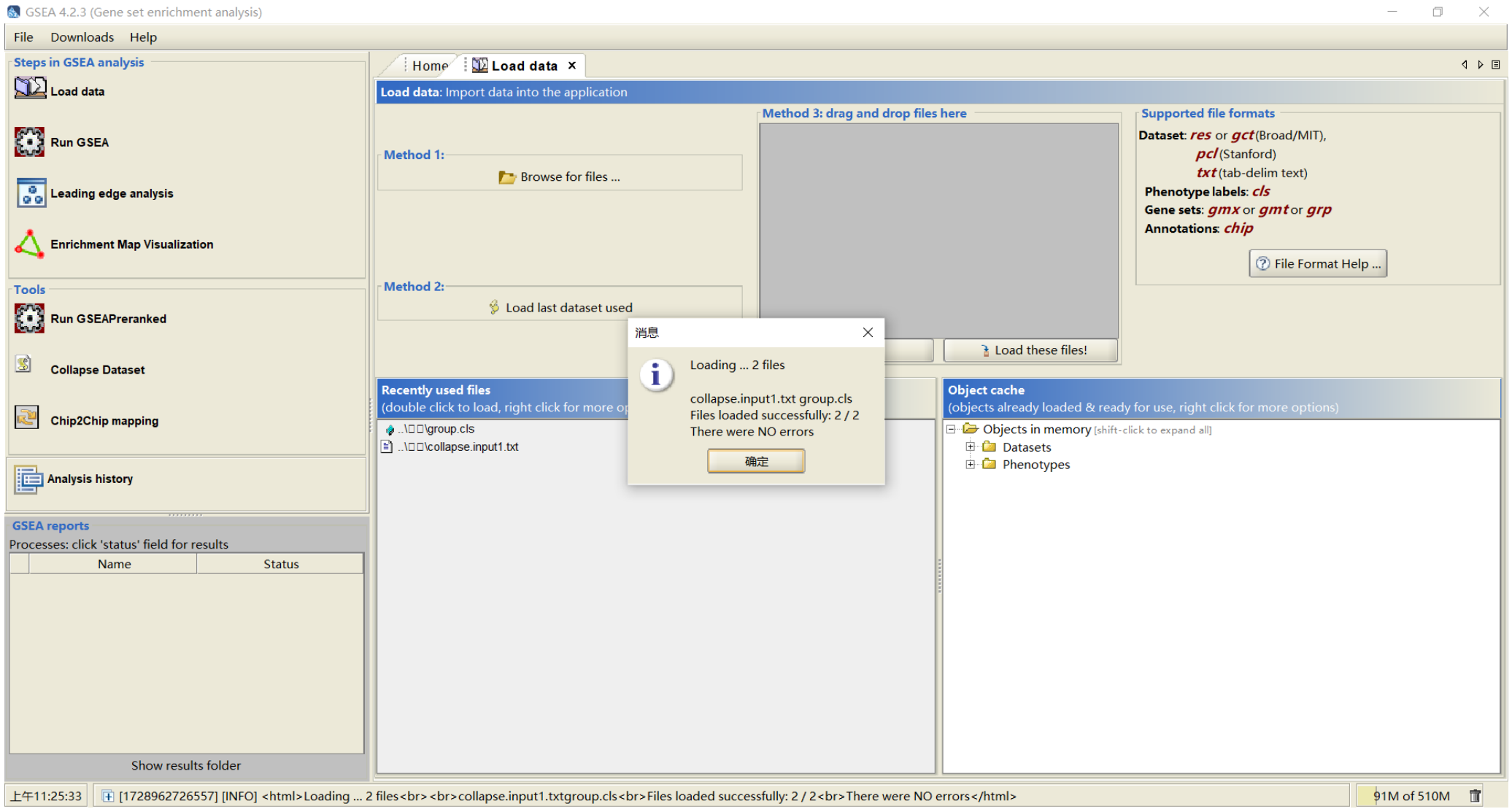
GSEA web site:
www.gsea-msigdb.org

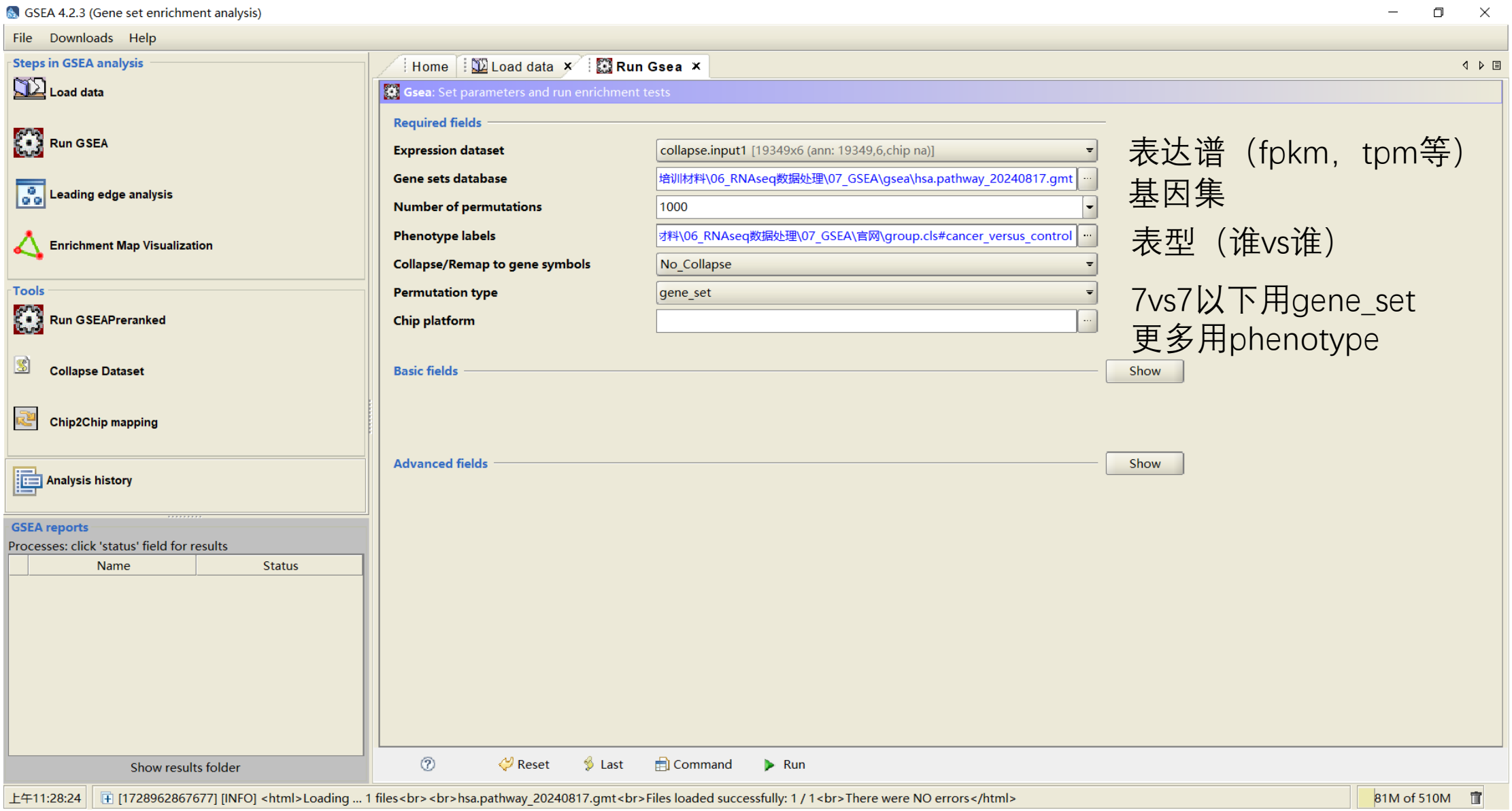
Contact the GSEA team:
gsea-msigdb.org/gsea/contact.jsp

BROAD INSTITUTE

UC San Diego

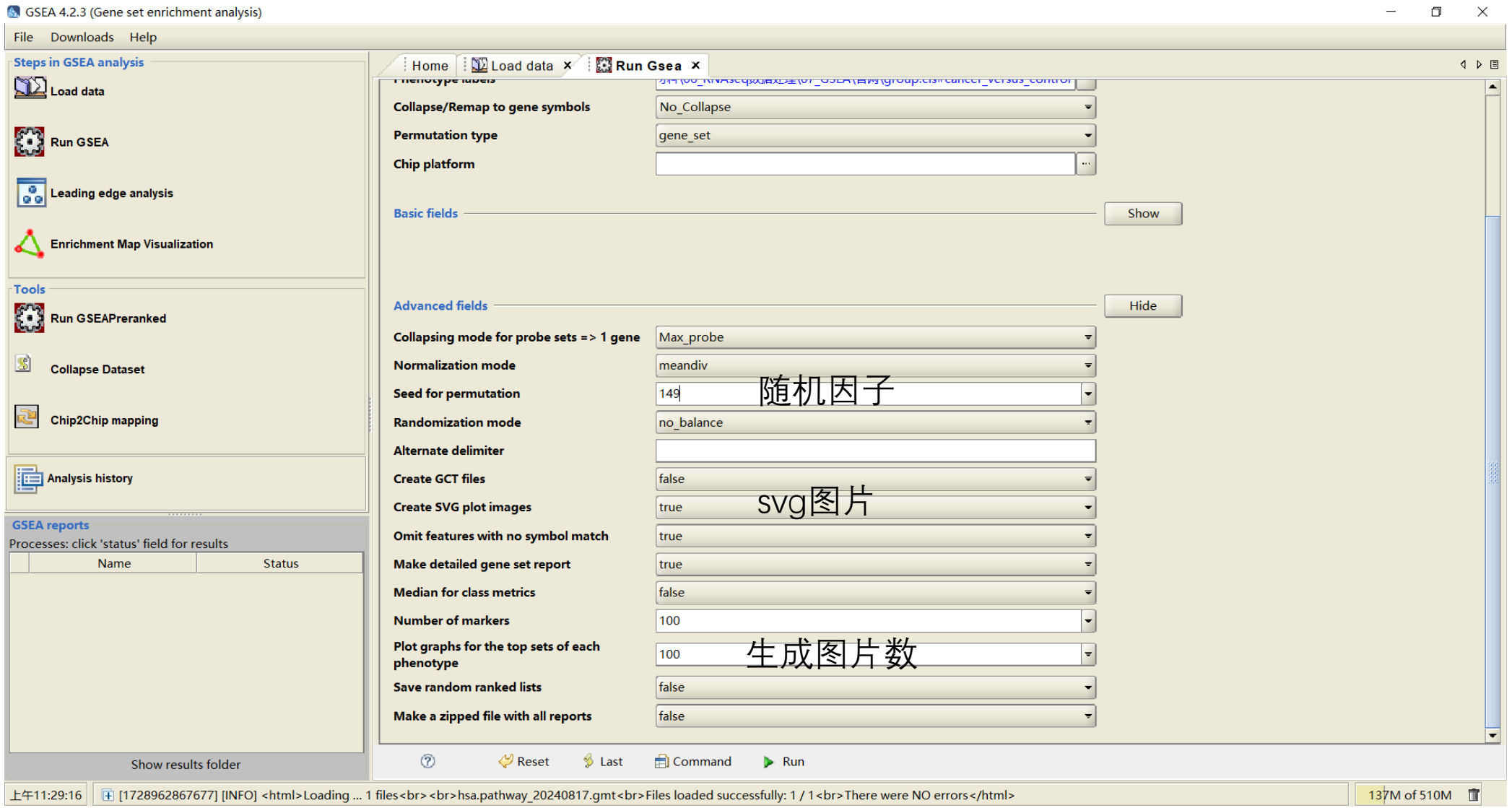
上午11:18:22 [1728962279710] [信息] Made Vdb dir: C:\Users\uuu\gsea_home\output\oct15 39M of 510M





表达谱（fpkm，tpm等）
基因集
表型（谁vs谁）

7vs7以下用gene_set
更多用phenotype



GSEA Report for Dataset collapse.input1

Enrichment in phenotype: cancer (3 samples)

- 327 / 336 gene sets are upregulated in phenotype **cancer**
- 0 gene sets are significant at FDR < 25%
- 18 gene sets are significantly enriched at nominal pvalue < 1%
- 45 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in TSV format (tab delimited text)
- Guide to interpret results

Table: Gene sets enriched in phenotype cancer (3 samples) [plain text format]

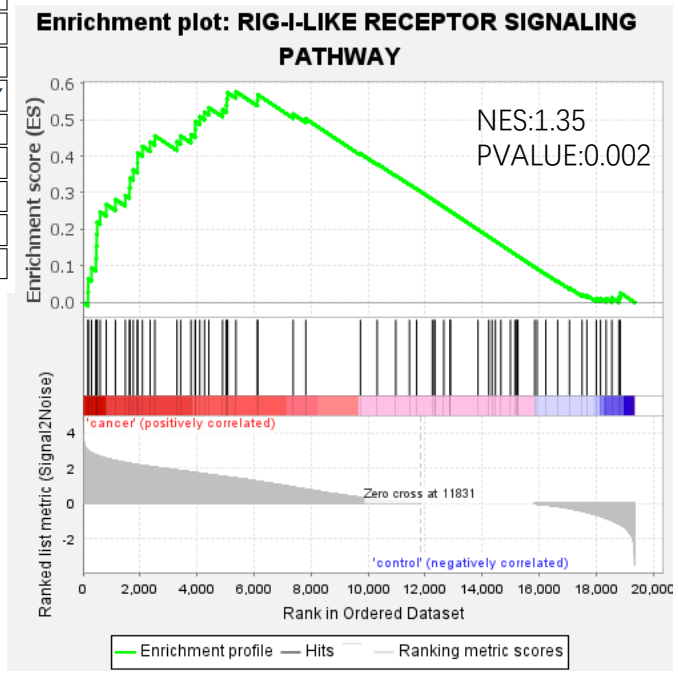
	GS follow link to MSigDB	GS DETAILS	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX	LEADING EDGE
1	RIG-I-LIKE RECEPTOR SIGNALING PATHWAY	Details ...	71	0.58	1.35	0.002	1.000	0.796	5365	tags=45%, list=28%, signal=62%
2	PANTOTHENATE AND COA BIOSYNTHESIS	Details ...	21	0.61	1.30	0.053	1.000	0.969	6630	tags=57%, list=34%, signal=87%
3	UBIQUITIN MEDIATED PROTEOLYSIS	Details ...	141	0.53	1.28	0.001	1.000	0.991	5812	tags=57%, list=30%, signal=82%
4	CIRCADIAN RHYTHM	Details ...	33	0.57	1.28	0.030	1.000	0.992	5157	tags=55%, list=27%, signal=74%
5	AUTOPHAGY - ANIMAL	Details ...	168	0.53	1.28	0.000	0.948	0.992	6508	tags=60%, list=34%, signal=90%
6	TOLL-LIKE RECEPTOR SIGNALING PATHWAY	Details ...	107	0.53	1.27	0.005	0.914	0.997	5365	tags=42%, list=28%, signal=58%
7	T CELL RECEPTOR SIGNALING PATHWAY	Details ...	122	0.53	1.27	0.000	0.787	0.997	6625	tags=48%, list=34%, signal=73%
8	VALINE, LEUCINE AND ISOLEUCINE DEGRADATION	Details ...	48	0.55	1.27	0.028	0.740	0.999	6630	tags=58%, list=34%, signal=89%
9	CHAGAS DISEASE	Details ...	102	0.54	1.26	0.004	0.665	1.000	6132	tags=51%, list=32%, signal=74%
10	CELL CYCLE	Details ...	158	0.52	1.26	0.001	0.640	1.000	6839	tags=64%, list=35%, signal=98%
11	ADHERENS JUNCTION	Details ...	92	0.53	1.26	0.005	0.586	1.000	6926	tags=62%, list=36%, signal=96%
12	NOD-LIKE RECEPTOR SIGNALING PATHWAY	Details ...	183	0.51	1.25	0.000	0.646	1.000	6722	tags=48%, list=35%, signal=73%
13	NICOTINATE AND NICOTINAMIDE METABOLISM	Details ...	38	0.55	1.25	0.045	0.604	1.000	4017	tags=37%, list=21%, signal=46%
14	TNF SIGNALING PATHWAY	Details ...	119	0.52	1.25	0.001	0.570	1.000	5798	tags=54%, list=30%, signal=76%

Enrichment in phenotype: control (3 samples)

- 9 / 336 gene sets are upregulated in phenotype **control**
- 5 gene sets are significantly enriched at FDR < 25%
- 2 gene sets are significantly enriched at nominal pvalue < 1%
- 2 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in TSV format (tab delimited text)
- Guide to interpret results

Table: GSEA Results Summary

Dataset	collapse.group.cls#cancer_versus_control
Phenotype	group.cls#cancer_versus_control
Upregulated in class	cancer
GeneSet	RIG-I-LIKE RECEPTOR SIGNALING PATHWAY
Enrichment Score (ES)	0.5763697
Normalized Enrichment Score (NES)	1.3535781
Nominal p-value	0.002
FDR q-value	1.0
FWER p-Value	0.759



Dataset details

- The dataset has 19349 features (genes)
- No probe set => gene symbol collapsing was requested, so all 19349 features were used

Gene set details

- Gene set size filters (min=15, max=500) resulted in filtering out 23 / 359 gene sets
- The remaining 336 gene sets were used in the analysis
- List of gene sets used and their sizes (restricted to features in the specified dataset)

clusterProfiler GSEA

07_gsea.R

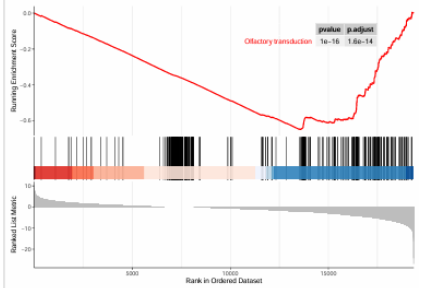
```r
# 载入R包
library(clusterProfiler)
library(enrichplot)
#set.seed(20170312)

# 读取输入差异基因（symbol，log2fc两列）
mydata = read.table("07_gsea_input.txt", sep="\t", header=TRUE, quote="")
GSEA_input<-mydata$logFC
names(GSEA_input) = as.character(mydata$symbol)
GSEA_input = sort(GSEA_input, decreasing = TRUE)

# 读取基因集
kegg_gmt <- read.gmt("hsa.pathway_20240817.gmt")

# 富集分析
gsea.KEGG <- GSEA(GSEA_input, minGSSize = 10, maxGSSize =500, TERM2GENE = kegg_gmt, eps=1e-16, pvalueCutoff=1.0)

# 导出结果
write.table(gsea.KEGG, file='07_gsea_result.xls', sep='\t', row.names=F, quote=F)

# 绘图
pdf('07_gsea_plot.pdf', width=10, height=7, family="ArialMT")
gseaplot2(gsea.KEGG, c("Olfactory transduction"), color=c('#FE0000'), pvalue_table = T, base_size=12)
dev.off()
```

07_gsea_input.txt

	symbol	logFC
1	BARX1	10.11489103
2	SBSN	9.160176804
3	STUB1	9.000828135
4	ALDH3A1	8.994223861
5	FTH1	8.839088785
6	PRG4	8.81770757
7	IGFBP3	8.570776189
8	INSYN1	8.28441162
9	INHBB	8.15523049
10	ARL2BP	8.122468625

	A	B	C	D	E
1	EGFR tyro	https://www.genor	AKT3	BCL2L11	P3R3URF-
2	Long-tern	https://www.genor	RAPGEF3	ADCY1	ADCY8
3	Aminoacy	https://www.genor	FARSB	WARS2	FARS2
4	Biosynthe	https://www.genor	GNE	NANP	FCSK
5	Morphine	https://www.genor	GNB5	ADCY1	ADCY2
6	Parkinson	https://www.genor	NDUFC2-	PPIF	TRAP1
7	Galactose	https://www.genor	GALM	AKR1B1	G6PC1
8	Yersinia in	https://www.genor	AKT3	ARPC5	ARPC4
9	Starch an	https://www.genor	TREH	LOC12490	LOC12490

	A	B	C	D	E	F	G	H	I	J
1	Descriptio	setSize	enrichmer	NES	pvalue	p.adjust	qvalue	rank	leading_e	core_enrich
2	Olfactory	428	-0.6484	-1.88154	1.00E-16	1.60E-14	1.27E-14	5773	tags=41%,	OR3A1/OR
3	Neuroacti	368	-0.60952	-1.76231	1.00E-16	1.60E-14	1.27E-14	4141	tags=47%,	CRH/FSHR/
4	Neutrophi	184	-0.60121	-1.71325	6.78E-12	7.24E-10	5.76E-10	4590	tags=48%,	H3C10/H2A
5	Systemic l	127	-0.62761	-1.76785	2.07E-10	1.65E-08	1.32E-08	2951	tags=43%,	H2BW2/IFN
6	Taste tran	85	-0.67421	-1.86208	3.44E-10	2.20E-08	1.75E-08	3108	tags=47%,	ASIC2/ADC
7	Hematopo	96	-0.64865	-1.80198	2.10E-09	1.12E-07	8.93E-08	3554	tags=50%,	IL1R2/CD24
8	Cell adhes	154	-0.56567	-1.60445	1.03E-07	4.72E-06	3.76E-06	4169	tags=45%,	ITGA6/CLDI



GeneSet，ID，Description：基因集的名字及描述信息
setSize：基因集中包含的基因数
enrichmentScore：富集分数ES
NES：标准化以后的富集分数，Normalized Enrichment Score
pvalue：富集的P值
p.adjust：校正的P值
qvalues：FDR（false discovery rate）错误发现率
rank：当富集分数ES最大时，对应基因所在排序好的基因列表中所处的位置
leading_edge：tags表示核心基因占该基因集的百分比；list表示核心基因占所有基因的百分比；signal，将前2个统计值放一起计算出的富集信号强度
core_enrichment：核心基因列表

GeneSet	ID	Description	setSize	enrichmentScore	NES	pvalue	p.adjust	qvalues	rank	leading_edge	core_enrichment
HALLMARK_	HALLMAR	HALLMARK_BILE	112	-0.831280444	-2.615133614	1E-16	1E-15	4.21E-16	2264	tags=63%, list=11%, signal=57%	HSD17B11/IDI1/NEDI
HALLMARK_	HALLMAR	HALLMARK_E2F_	196	0.733123923	3.047759875	1E-16	1E-15	4.21E-16	2931	tags=63%, list=14%, signal=55%	PTTG1/TOP2A/DLGAP5
HALLMARK_	HALLMAR	HALLMARK_FATT	155	-0.729026212	-2.361585203	1E-16	1E-15	4.21E-16	2264	tags=57%, list=11%, signal=51%	HSD17B11/OSTC/HADH
HALLMARK_	HALLMAR	HALLMARK_G2M_	192	0.705013268	2.902961942	1E-16	1E-15	4.21E-16	2766	tags=55%, list=13%, signal=48%	PTTG1/TOP2A/CDC6/M

其他注意事项

- 来源一样，基因不完全一样。例如：GSEA官网的MAPK signaling pathway包含267个基因，而KEGG官网里边包含300个基因
- 数据库更新频率不同
- GSEA官网的只有Human和Mouse，其他物种需要自定义
- 可用于xx相关基因检索，即自定义基因集
- tpm，fpkm，normalized counts在GSEA官网版里的结果基本一样
- 官网版本跟clusterProfiler版本结果大体一致，但不完全一样，挑选个最佳结果
- GSEA官网版（Broad Institute开发，更权威）更受老外青睐，clusterProfiler更受中国人青睐（Y叔开发）
- 算法通用：适用于代谢组等