GSEA富集分析

作者邮箱：2740881706@qq.com

GSEA

GSEA(Gene set enrichment analysis) 根据已有的对基因的定位、功能、生物学意义等知识的基础上，首先构建了一个分子标签数据库，数据库中包含了多个功能基因集。通过分析基因表达数据，得到表达状况是否在某种功能上显著富集。

GSEA与其它工具的区别

- David和kobas等分析工具（先找差异基因）

	差异基因	总基因
基因数目	100	20000
某个功能	10	100

- GSEA(无需先找差异基因)

GSEA计算的基本原理是扫描排序序列，当出现一个功能基因集中的基因时，就增加 ES 值，反之，就减少 ES 值，所以在整个扫描过程中，ES是一个动态的值。

GSEA输入文件

- 表达数据集文件(.gct)
- 表型数据文件(.cls)
- 功能基因集文件(.gmt)
- 芯片注释文件(.chip)

GSEA输入文件

- 表达数据集文件(.gct)

#1.2							
22283	32						
NAME	Description	GM10832_0GY	GM10835_0GY	GM7057_0GY	GM13113_0GY	GM10860_0GY	GM03187_0GY
215538_at	na	28.566616	17.435623	16.682034	18.983868	49.374252	28.643951
218987_at	na	152.65285	97.011475	77.56722	120.19234	152.37917	106.619156
205137_x_at	na	47.44099	7.684895	9.738243	13.04413	12.769203	11.245403
210677_at	na	7.2691836	13.469224	6.605069	2.6787052	7.803402	36.91887
205949_at	na	24.103083	12.642892	29.46877	17.586283	57.31954	28.00742
209946_at	na	2.9331794	16.526657	24.133905	57.184532	57.461414	29.068306
221956_at	na	13.518131	10.411794	6.012306	17.702747	17.309364	16.655928

GSEA输入文件

- 表型数据文件(.cls)

```
32 2 1
# MALE FEMALE
m m m m m m m m m m m m m m f f f f f f f f f f f f f f f f
```

GSEA结果

NAME	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX	LEADING EDGE
TRANSLATIONAL_INITIATION	22	0.59631586	1.6403712	0.014141414	1	0.719	3555	tags=45%, list=27%, signal=62%
RIBOSOME	31	0.56243736	1.6205235	0.037109375	1	0.76	3574	tags=58%, list=27%, signal=79%
RNA_CATABOLIC_PROCESS	21	0.5530625	1.579454	0.022088353	1	0.81	1727	tags=43%, list=13%, signal=49%
ER_GOLGI_INTERMEDIATE_COMPARTMENT	18	0.61010396	1.5726844	0.04631579	1	0.823	2958	tags=61%, list=22%, signal=78%

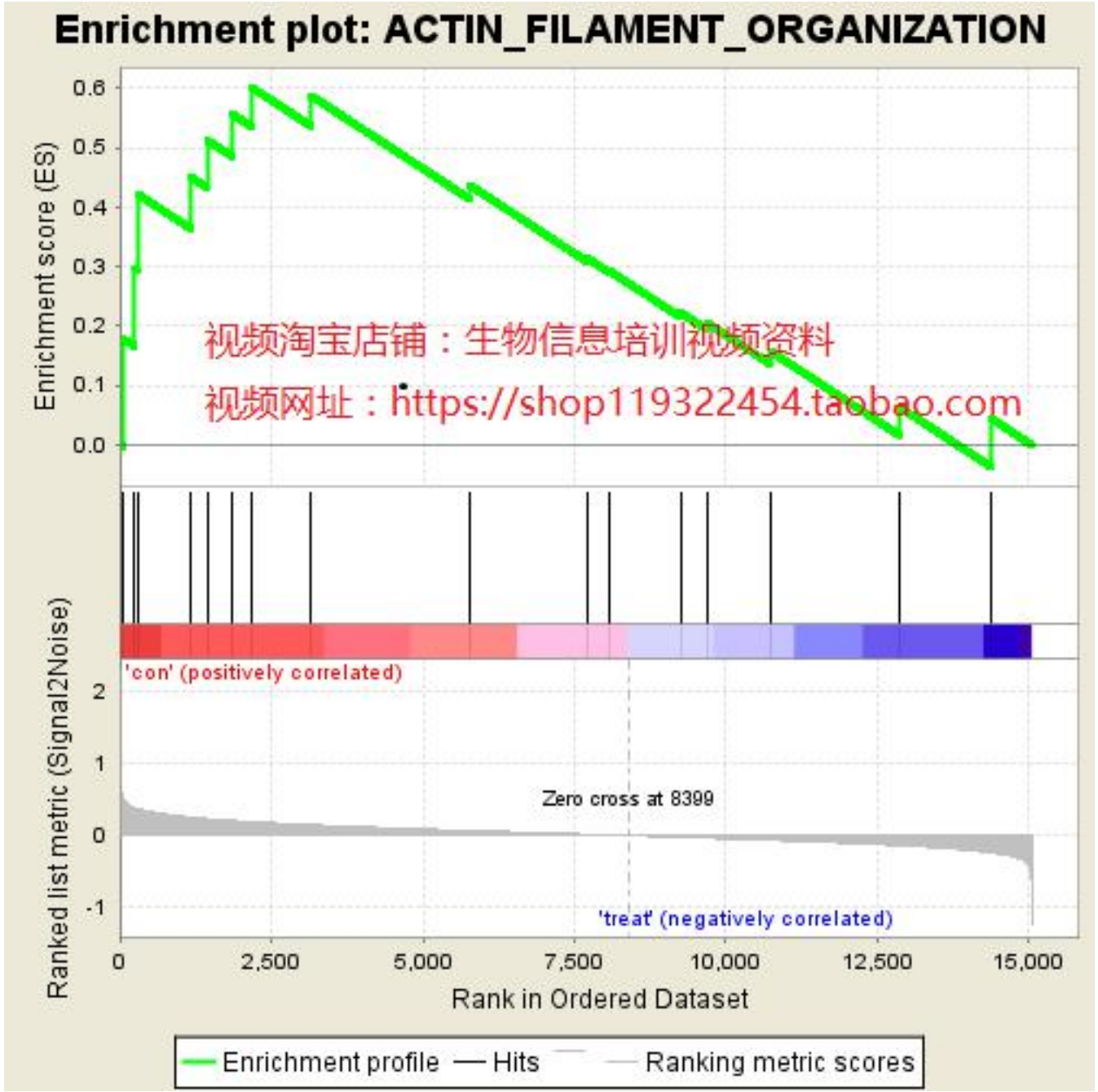
GSEA结果

- Tags. The percentage of gene hits before (for positive ES) or after (for negative ES) the peak in the running enrichment score. This gives an indication of the percentage of genes contributing to the enrichment score.
- List. The percentage of genes in the ranked gene list before (for positive ES) or after (for negative ES) the peak in the running enrichment score. This gives an indication of where in the list the enrichment score is attained.
- Signal. The enrichment signal strength that combines the two previous statistics:

$$(\text{Tag }\%)(1-\text{Gene }\%)\left(\frac{N}{N-Nh}\right)$$

where N is the number of genes in the list and Nh is the number of genes in the gene set. If the gene set is entirely within the first Nh positions in the list, then the signal strength is maximal or 100%. If the gene set is spread throughout the list, then the signal strength decreases towards 0%.

GSEA结果


Enrichment plot: ACTIN_FILAMENT_ORGANIZATION
Enrichment score (ES)
0.6
0.5
0.4
0.3
0.2
0.1
0.0
视频淘宝店铺：生物信息培训视频资料
视频网址：https://shop119322454.taobao.com
'con' (positively correlated)
Ranked list metric (Signal2Noise)
2
1
0
-1
Zero cross at 8399
'treat' (negatively correlated)
0
2,500
5,000
7,500
10,000
12,500
15,000
Rank in Ordered Dataset
Enrichment profile
Hits
Ranking metric scores

GSEA结果