# R语言：基于单基因批量相关性分析的GSEA

作者：果子 来源：果子学生信

本文原地址：https://www.iikx.com/news/statistics/6109.html



**R语言：基于单基因批量相关性分析的GSEA的使用场景**

1.已经确定研究的基因，但是想探索他潜在的功能，可以通过跟这个基因表达最相关的基因来反推他的功能，这种方法在英语中称为guilt of association，协同犯罪。

2.我们的注释方法依赖于TCGA大样本，既然他可以注释基因，那么任何跟肿瘤相关的基因都可以被注释，包括长链非编码RNA。

这个方法以前阐述过：

单基因批量相关性分析的妙用

但是这个方法有个小缺陷，并不知道最后富集的通路是正向影响还是反向影响，也就是无法判断方向。判断方向的工具也不是没有，GSEA就是一个。所以，我想能不能把批量相关性分析和GSEA结合起来。

GSEA需要的gene set是现成的没有问题，但是genelist没有，这里我们可以把所有基因跟单个基因的相关性系数当做LogFC，有正有负，就解决了geneList的问题。这个想法不是我的，是我的一个学员的，不过他要解决的是microRNA把基因的问题。

下面来实战一下：

**1.首先加载数据**

这个数据是我下载了TPM数据，然后提取出乳腺癌的数据得来的。

```
load(file = "BRCA_mRNA_exprSet.Rdata")
exprSet <- mRNA_exprSet
test <- exprSet[1:10,1:10]
```

| | TCGA-C8-A1HL-01 | TCGA-EW-A2FS-01 | TCGA-E2-A153-11 | TCGA-A2-A3XX-01 | TCGA-BH-A0BQ-11 | TCGA-Z7-A8R5-01 | TCGA-B6-A0RL-01 | TCGA-BH-A204-11 | TCGA-E2-A2P6-01 | TCGA-AO-A1KT-01 |
|---|---|---|---|---|---|---|---|---|---|---|
| OR4F5 | -9.9658 | -9.9658 | -9.9658 | -9.9658 | -9.9658 | -9.9658 | -9.9658 | -9.9658 | -9.9658 | -9.9658 |
| SAMD11 | 3.6543 | 1.2147 | 0.3685 | 0.9191 | -0.4719 | 4.2297 | -3.3076 | -1.6394 | -1.4305 | -1.4305 |
| NOC2L | 4.6271 | 5.3252 | 5.6685 | 6.6980 | 5.0427 | 5.4809 | 5.8040 | 3.9617 | 5.5970 | 4.8490 |
| KLHL17 | 1.5216 | 2.8582 | 2.4934 | 3.1179 | 1.7097 | 2.0360 | 2.8877 | -0.5125 | 3.7773 | 0.6880 |
| PLEKHN1 | 0.5271 | 1.1250 | 1.4808 | 2.8422 | 1.2696 | 2.0325 | 1.5902 | -3.8160 | 0.4016 | -0.1031 |
| PERM1 | 0.3685 | 1.2875 | 1.1117 | 2.5137 | 0.9115 | 2.5828 | 1.7316 | -3.1714 | 0.7999 | -0.1993 |
| HES4 | 2.9224 | 3.4129 | 2.3222 | 2.0325 | 2.5137 | 4.4108 | 4.1740 | 0.5271 | 4.0260 | 1.2023 |
| ISG15 | 6.5717 | 9.2399 | 4.4848 | 7.4406 | 4.6821 | 5.4380 | 7.8918 | 3.6793 | 6.7473 | 5.6017 |
| AGRN | 5.6955 | 6.0135 | 5.9656 | 5.9213 | 5.4848 | 5.6525 | 6.0119 | 3.1129 | 6.5838 | 4.3148 |
| RNF223 | 1.9527 | 1.3109 | 0.3231 | 0.4657 | 0.5170 | 0.0990 | 2.1988 | -3.1714 | 2.1798 | 0.6517 |

## 2.写一个函数批量计算相关性

这个函数只要输入一个基因，他就会批量计算这个基因跟其他编码基因的相关性，返回相关性系数和p值。

```
batch_cor <- function(gene){
  y <- as.numeric(exprSet[gene,])
  rownames <- rownames(exprSet)
  do.call(rbind,future_lapply(rownames, function(x){
    dd  <- cor.test(as.numeric(exprSet[x,]),y,type="spearman")
    data.frame(gene=gene,mRNAs=x,cor=dd$estimate,p.value=dd$p.value )
  }))
}
```

## 3.并行化运行函数

以PCDC1这个基因为例

```
library(future.apply)
plan(multiprocess)
system.time(dd <- batch_cor("PDCD1"))
```

这是返回的结果

| | gene | mRNAs | cor | p.value |
|---|---|---|---|---|
| cor | PDCD1 | OR4F5 | -0.0458526436 | 1.111925e-01 |
| cor1 | PDCD1 | SAMD11 | 0.1945074379 | 9.194344e-12 |
| cor2 | PDCD1 | NOC2L | 0.2193508721 | 1.258030e-14 |
| cor3 | PDCD1 | KLHL17 | 0.2558680312 | 1.645638e-19 |
| cor4 | PDCD1 | PLEKHN1 | 0.2126125777 | 8.172526e-14 |
| cor5 | PDCD1 | PERM1 | 0.1752551763 | 8.654398e-10 |
| cor6 | PDCD1 | HES4 | 0.2224512530 | 5.208031e-15 |
| cor7 | PDCD1 | ISG15 | 0.3290383929 | 6.805207e-32 |
| cor8 | PDCD1 | AGRN | 0.1825116678 | 1.652243e-10 |
| cor9 | PDCD1 | RNF223 | 0.0417828178 | 1.466835e-01 |
| cor10 | PDCD1 | C1orf159 | 0.2781130153 | 6.775857e-23 |
| cor11 | PDCD1 | TTLL10 | 0.1465681422 | 3.111806e-07 |
| cor12 | PDCD1 | TNFRSF18 | 0.2016891091 | 1.488654e-12 |
| cor13 | PDCD1 | TNFRSF4 | 0.5756661775 | 1.509789e-107 |

**4.制作genelist gene**

```
gene <- dd$mRNAs
## 转换
library(clusterProfiler)
gene = bitr(gene, fromType="SYMBOL", toType="ENTREZID", OrgDb="org.Hs.eg.db")
## 去重
gene <- dplyr::distinct(gene,SYMBOL,.keep_all=TRUE)

gene_df <- data.frame(logFC=dd$cor,
                      SYMBOL = dd$mRNAs)
gene_df <- merge(gene_df,gene,by="SYMBOL")

## geneList 三部曲
## 1.获取基因logFC
geneList <- gene_df$logFC
## 2.命名
names(geneList) = gene_df$ENTREZID
## 3.排序很重要
geneList = sort(geneList, decreasing = TRUE)
```

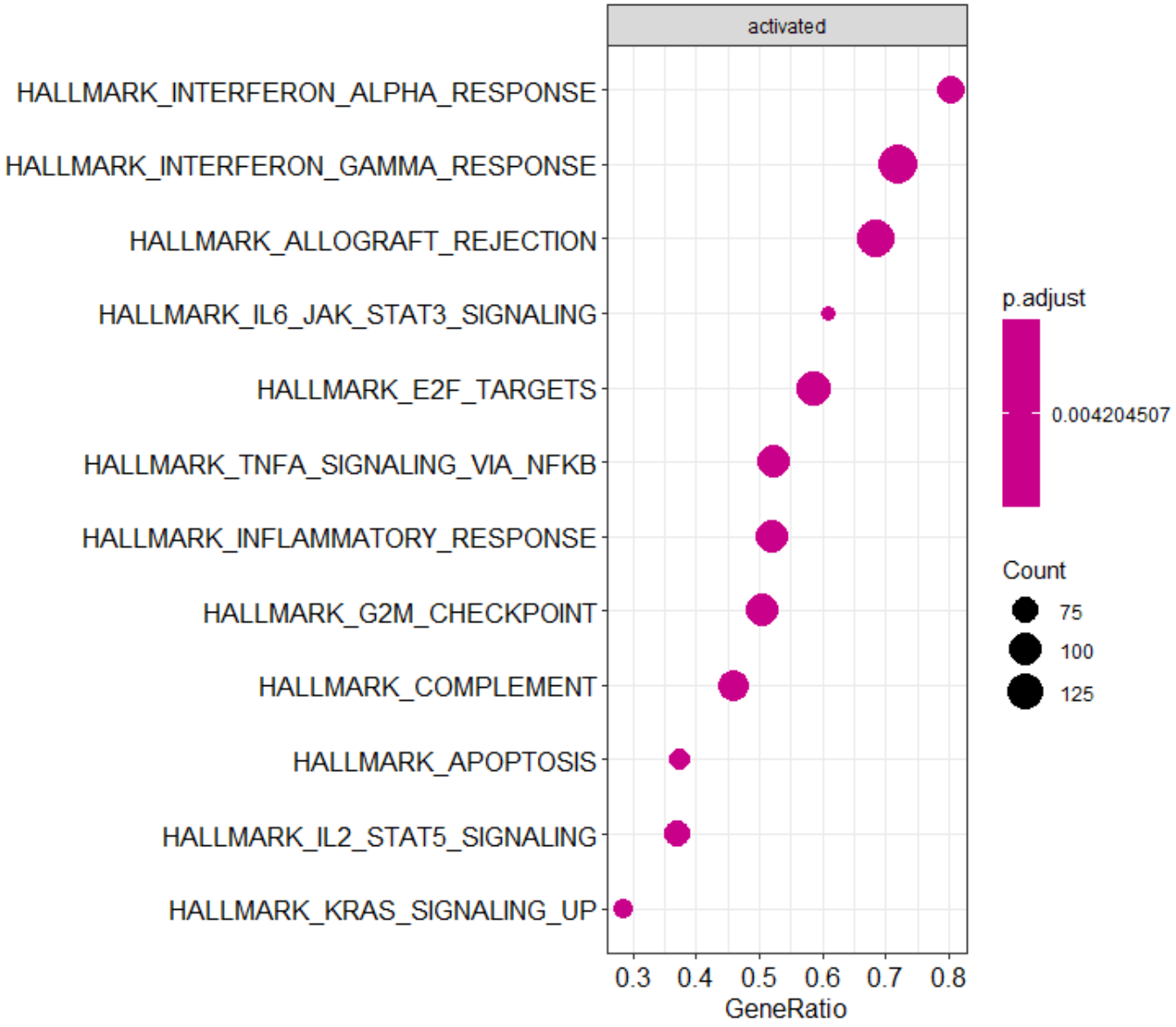
**5.运行GSEA分析**

```
library(clusterProfiler)
## 读入hallmarks gene set，从哪来？
hallmarks <- read.gmt("h.all.v6.2.entrez.gmt")
# 需要网络
y <- GSEA(geneList,TERM2GENE =hallmarks)
```

作图看整体分布

```
### 看整体分布
library(ggplot2)
dotplot(y,showCategory=12,split=".sign")+facet_grid(~.sign)
```

本次结果中全是激活的




**6.特定通路作图**

```
yd <- data.frame(y)
library(enrichplot)
gseaplot2(y,"HALLMARK_INTERFERON_ALPHA_RESPONSE",color = "red",pvalue_table
= T)
```

PCDC1跟阿拉法干扰素正相关，这个事情没什么好说的吧。

更多 统计方法 请访问 https://www.iikx.com/news/statistics/