GEO 数据库挖掘 (2) 之数据整合

阎俊安 2022-01-17

目录

本节来介绍如何使用 **GEO** 数据库进行数据挖掘，请参考前文一**文搞定 GEO 数据下载**

1 安装并加载 R 包

```
package.list=c("tidyverse","GEOquery","magrittr")

for (package in package.list) {
  if (!require(package,character.only=T, quietly=T)) {
    install.packages(package)
    library(package, character.only=T)
  }
}
```

2 导入数据

```
load(file ="GSE33126.rdata")

gset <- gset[[1]]
class(gset)
```

3 样本信息表

```
sampleinfo <- pData(gset) %>%# 提取样本信息表
  select(source_name_ch1,characteristics_ch1.1) %>%
  rename(group = source_name_ch1,patient=characteristics_ch1.1) %>%
  mutate_at(vars(patient),~str_split(.," ",simplify = T)[,2])
```

4 基因表达矩阵

```
gene_exp <- exprs(gset) %>% as.data.frame()
```

5 检查探针编号

```
tail(gene_exp[,1:3])
```

6 同步表达矩阵与样本信息

```
gene_exp <- gene_exp[,which(
  colnames(gene_exp) %in% rownames(sampleinfo)
)]
```

- 查看数据

```
summary(gene_exp)
gene_exp <- log2(gene_exp)
```

- 绘制箱线图

```
boxplot(gene_exp,outline=FALSE)
```

若箱线图中位数差异较大可以执行下面代码对数据进行标准化处理

- 对数据进行标准化

```
library(limma)
p <- as.data.frame(
  normalizeBetweenArrays(gene_exp)
)
```

经过上面的步骤我们得到了样本信息表 & 基因表达矩阵信息表，接下来从 **GEO** 数据库下载基因信息表就可进行后续分析

7 整合基因信息表

```
gene_info <- read_delim("GPL6947-13512.txt", "\t",escape_double = FALSE, comment = "#",
                        trim_ws = TRUE) %>%
  dplyr::select(ID,Gene_Symbol = Symbol,Entrez_Gene_ID,
                Gene_Title = Definition) %>% drop_na()
```

现在我们得到了分析需要的 3 张表，**save** 将其保存后续分析直接加载即可

```
save(gene_exp,sampleinfo,gene_info, file='GSE33126-info.rdata')
```