利用DESeq2进行差异表达分析

Hello 大家好! 我是**林更新**, 今天为大家介绍如何用R语言做差异表达分析。

安装R包

首先我们来安装分析所需的R包，分别可以从Bioconductor和R语言官网下载。当然最方便的方法是在R或Rstudio中输入以下代码进行安装：

```
source("https://bioconductor.org/biocLite.R")
biocLite("DESeq2")
install.packages("pheatmap")
```

导入数据

DESeq2的输入文件为二代测序产生的counts文件,不能输入标准化后的数据（基因表达量不是整数）。

Symbol	KO1	KO2	KO3	WT1	WT2	WT3
Adora2b	17	21	37	13	16	17
Adora2a	1	6	7	4	11	3
Gm4342	39	38	34	30	30	33

我们可以将数据整理成如上格式，保存成制表符分隔的txt文件格式。然后输入以下代码将counts文件导入到R中。

```
cd <- read.table("data.txt",header = T, row.names = 1)
```

查看数据

用dim()函数查看数据中包含的样本数和基因数

```
dim(cd)
```

创建Metadata

为了方便后续分析，这一步需建立包含样本名称及分组信息的元数据。name = c(……)括号里面填所有样本的名称，condition = c(……)括号里填每个样本对应的分组。下面代码中 rep("KO",3) = c("KO","KO","KO"),对

应"KO1","KO2","KO3"。

```
colData <- data.frame(name = c("KO1","KO2","KO3","WT1","WT2","WT3"), condition = c
(rep("KO",3),rep("WT",3))
rownames(colData) <- colnames(cd)
```

数据读入及质控

调用R包DESeq2。利用DESeqDataSetFromMatrix()函数读取数据及metadata，并存入变量dds（后续计算产生的数据都会存到这个变量里）。然后删去在所有样本中表达量均为0的基因。

```
library(DESeq2)
dds <- DESeqDataSetFromMatrix(cd,colData = colData, design = ~condition)
dds <- dds[rowSums(counts(dds))>1,]
```

设置对照组

此处，我们告诉软件将哪些样本作为对照来分析。那么后续得出的表达量的变化都是相对这些样本而言的。如下，ref="WT"表示以WT作为对照，分析时根据实际数据更改。

```
dds$condition <- relevel(dds$condition, ref="WT")
```

差异表达分析

现在我们可以进行差异表达分析了，代码灰常的简单。

```
dds <- DESeq(dds)
```

输出分析结果

提取分析结果并按q值排序

```
res <- results(dds,alpha = 0.05)
res
summary(res)
sum(res$padj<0.05,na.rm = T)

# 将所有基因按q值从小到大排序
resOrdered <- res[order(res$padj),]
```

提取表达量改变2倍以上（log2foldchange>1或<-1）且q值小于0.05的基因,并保存为csv文件。根据实际情

况，padj的值可调，也可将padj改为pvalue。log2FoldChange的值也可调整。

```
res_fc2_padj0.05 <- subset(resOrdered, (padj < 0.05)&(abs(log2FoldChange)>1))
write.csv(as.data.frame(res_fc2_padj0.05),file="results_fc2_padj0.05.csv")
```

输出的csv文件包含了根据设置的条件筛选出的差异表达基因，可以导入到DAVID等网站进行GO分析。