实验二 基因表达谱数据分析

1. 背景知识

基因表达谱包含了特定细胞或组织在特定状态下的基因表达种类及表达丰度信息。RNA-seq 是获取转录组信息的常用高通量测序技术，对测序数据进行 QC 后，将 reads 比对到参考转录本，通过对落到各基因区域或转录本上的 reads 数量进行标准化可以估计出基因或转录本的表达丰度，一般用 FPKM（Fragments Per Kilobase per Million mapped reads）或 RPKM（Reads Per Kilobase per Million mapped reads）来表示。通过基因表达谱，可以了解到特定状态条件下基因的表达情况；对比分析不同样本的基因表达谱，可以筛选差异表达基因，即在不同条件下表达有显著性差异的基因。

2. 实验目的

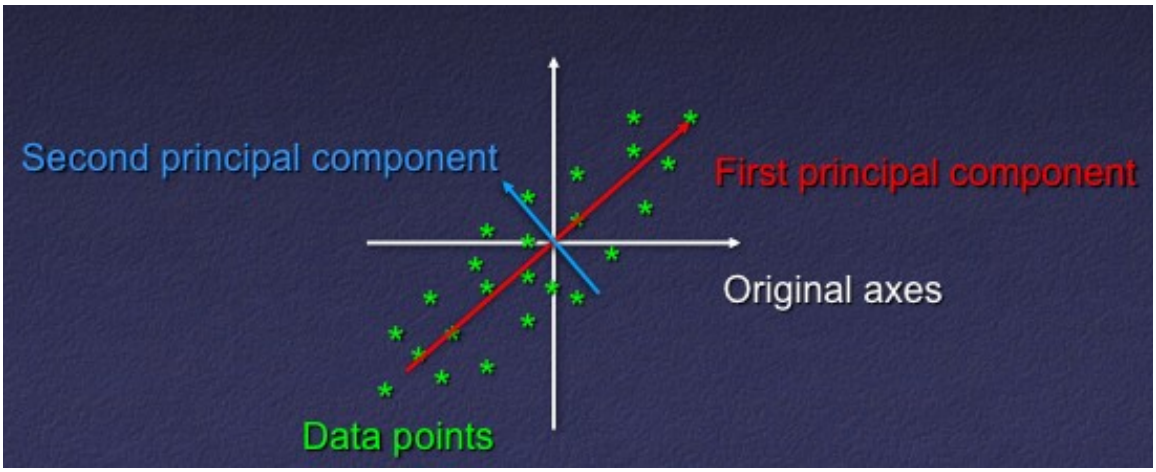
熟练采用 R 语言对基因表达谱数据进行统计分析和可视化。

3. 实验任务

给定多个样本（含重复）的基因表达谱数据文件“data2_exp.txt”和基因表达差异分析结果文件“data2_diff.txt”，进行如下分析。

1）根据样本的基因表达谱数据，对样本进行相关性分析，计算表达谱间的 Pearson 相关系数，并用热图显示。

2）根据样本的基因表达谱数据进行主成分分析（PCA, Principal Component Analysis），并作图显示。主成分分析是一种降维的统计方法，它借助于正交变换，将分量相关的随机向量转化成不相关的新随机向量，实际上是将原坐标系中的向量经过线性组合变换为新坐标系中的向量，以尽可能获取数据的最大方差。如下图所示，第一主成分获取了数据的最大方差；第二主成分与第一主成分正交（即不相关），且获取了数据的次大方差，依次类推。因此，利用前两个或三个主成分分量便可得到原数据中的大部分信息，且通过二维或三维空间作图可以展示变量间的关联。



PCA 示意图

3）比较样本 E 和样本 C，进行差异基因筛选及可视化。将 FPKM 值变化 2 倍及以上，且 q-value < 0.05 的基因定义为差异表达基因。分别用散点图和火山图来展示两个样本中的基因表达谱，用红色和蓝色分别表示在样本 E 中表达显著上调和显著下调的基因，黑色表示非显著差异基因。

散点图（scatter-plot）的横纵坐标分别表示两个样本中基因的表达量（FPKM 值），这里横纵坐标的数值进行对数化处理，每个点代表一个特定的基因或转录本，特定的一个点对应的横坐标值为该基因在样本 C 中的表达量，纵坐标值为该基因在样本 E 的表达量。图中红色点表示显著上调的基因，蓝色点表示显著下调的基因，黑色点为非显著差异基因；将所有基因映射上去后，越接近 0 的点，说明表达量越低；那些偏离了对角线程度越大的点表明该基因在两个样本间表达差异越大。Pearson correlation 是指两个样本基因表达水平的相关性指数，该数值越接近于 1，说明两个样本表达水平越相似；如果两个样本是重复样本的话，说明重复性越好。

火山图（volcano-plot）可以直观展示两个样本间基因差异表达的分布情况。横坐标为 $\log_2$(fold_change)，其中 fold_change 为基因在两个样本间表达水平的比值，通常将表达值都加 1 后再计算比值，以避免表达值为 0 时的影响；纵坐标为 $-\log_{10}$(q-value)，其中 q-value 为基因表达量变化差异的统计学检验值。图中每个点代表一个特定的基因或转录本，红色点表示显著上调的基因，蓝色点表示显著下调的

基因，黑色点为非显著差异基因；将所有基因映射上去之后，可以获知，在左边的蓝色点为表达显著下调的基因，右边的红色点为表达显著上调的基因，越靠两边和上边的点代表基因表达差异越显著。

4）用柱状图展示差异表达基因的数量。利用 3）中同样的差异基因筛选标准，对样本 C、E、F 分别进行两两比较，筛选出上下调基因，并用柱状图进行展示。

5）差异表达基因的功能富集分析。比较样本 E 和样本 C，利用 3）中同样的差异基因筛选标准，筛选出差异表达基因，利用 R 中 clusterProfiler 包进行功能富集分析，并作图展示。

提示：

需要预先安装的 R 包：

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install("clusterProfiler")
BiocManager::install("org.Hs.eg.db")
```

4. 实验报告要求

将实验任务的题目、以及对应的代码及图表结果等信息编辑在一个 word 文件中（注意代码缩进，代码用五号字号，其他文字用小四字号）。