百迈客生物科技
BIOMARKER TECHNOLOGIES

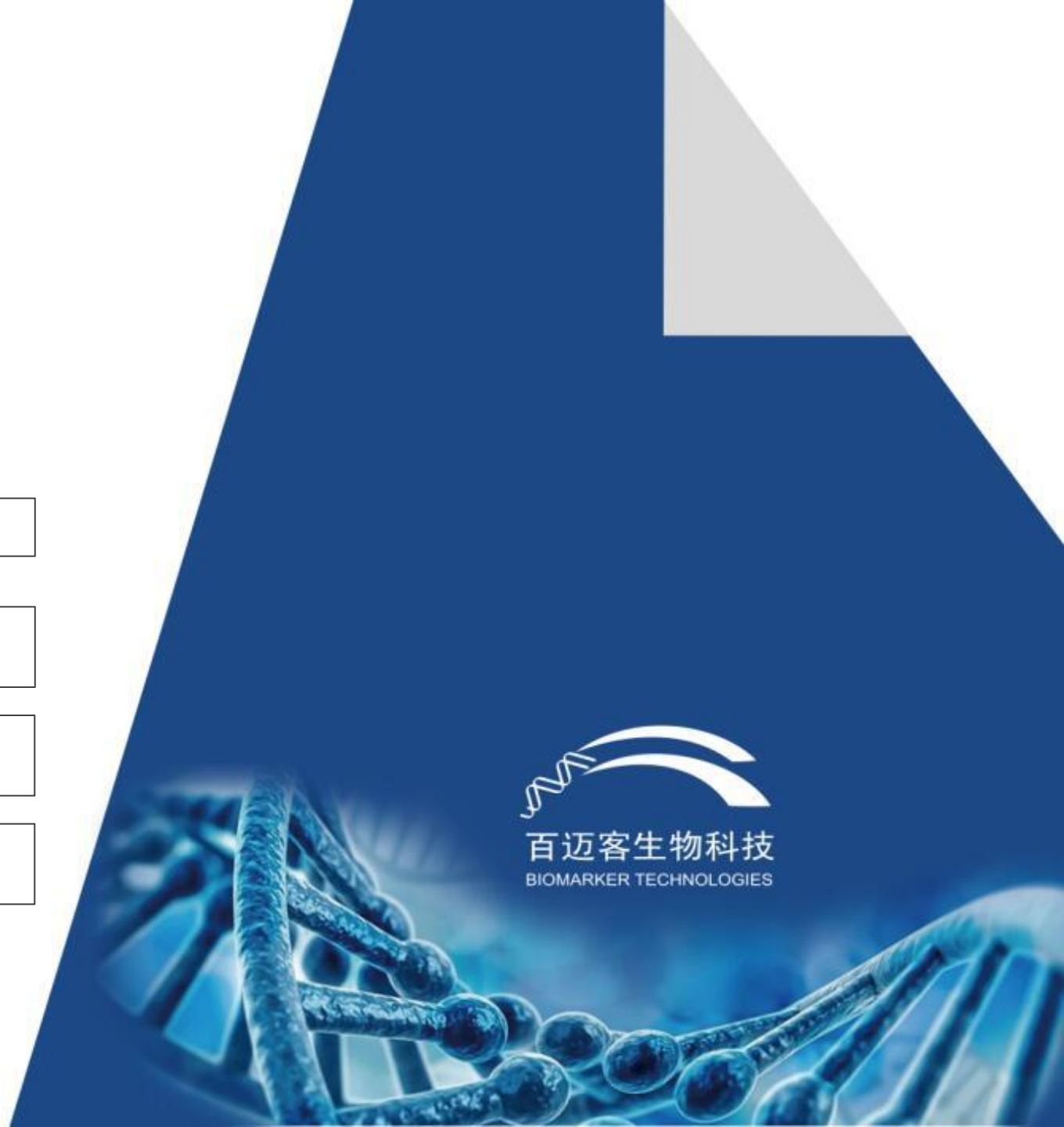
为 世 界 创 造 新 的 可 能

R语言基础

百迈客 培训事业部

原正文：18519006131

目录
Contents



百迈客生物科技
BIOMARKER TECHNOLOGIES

1 R语言介绍

1、R语言简介

2、R安装及环境搭建

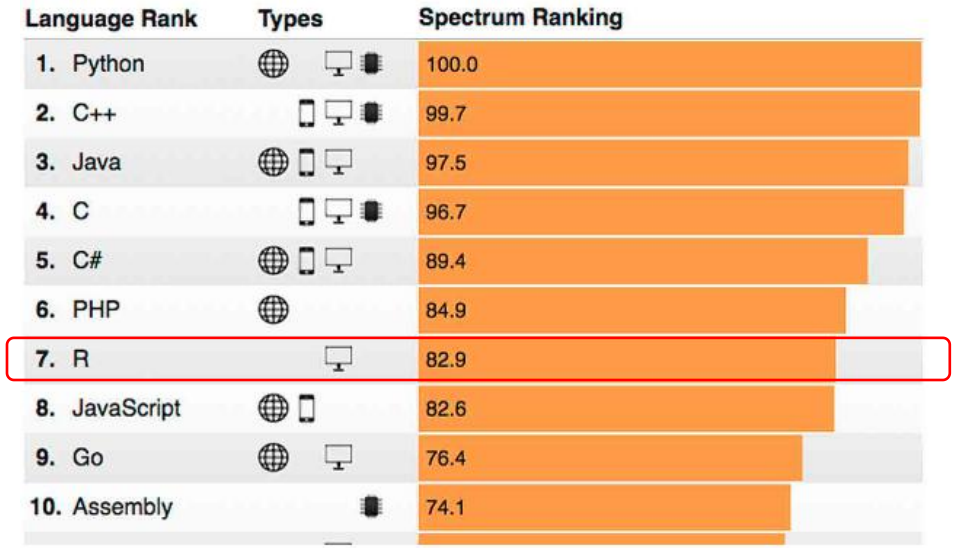
1. R语言简介

R是什么？

数据挖掘、统计分析、作图的解释型语言

能做什么？

统计分析，二维作图



2018年度IEEE编程语言排行榜前10

2. R安装及环境搭建

百迈客生物科技
BIOMARKER TECHNOLOGIES

R软件下载安装

https://www.r-project.org/

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- Download R for Linux
- Download R for (Mac) OS X
- Download R for Windows

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

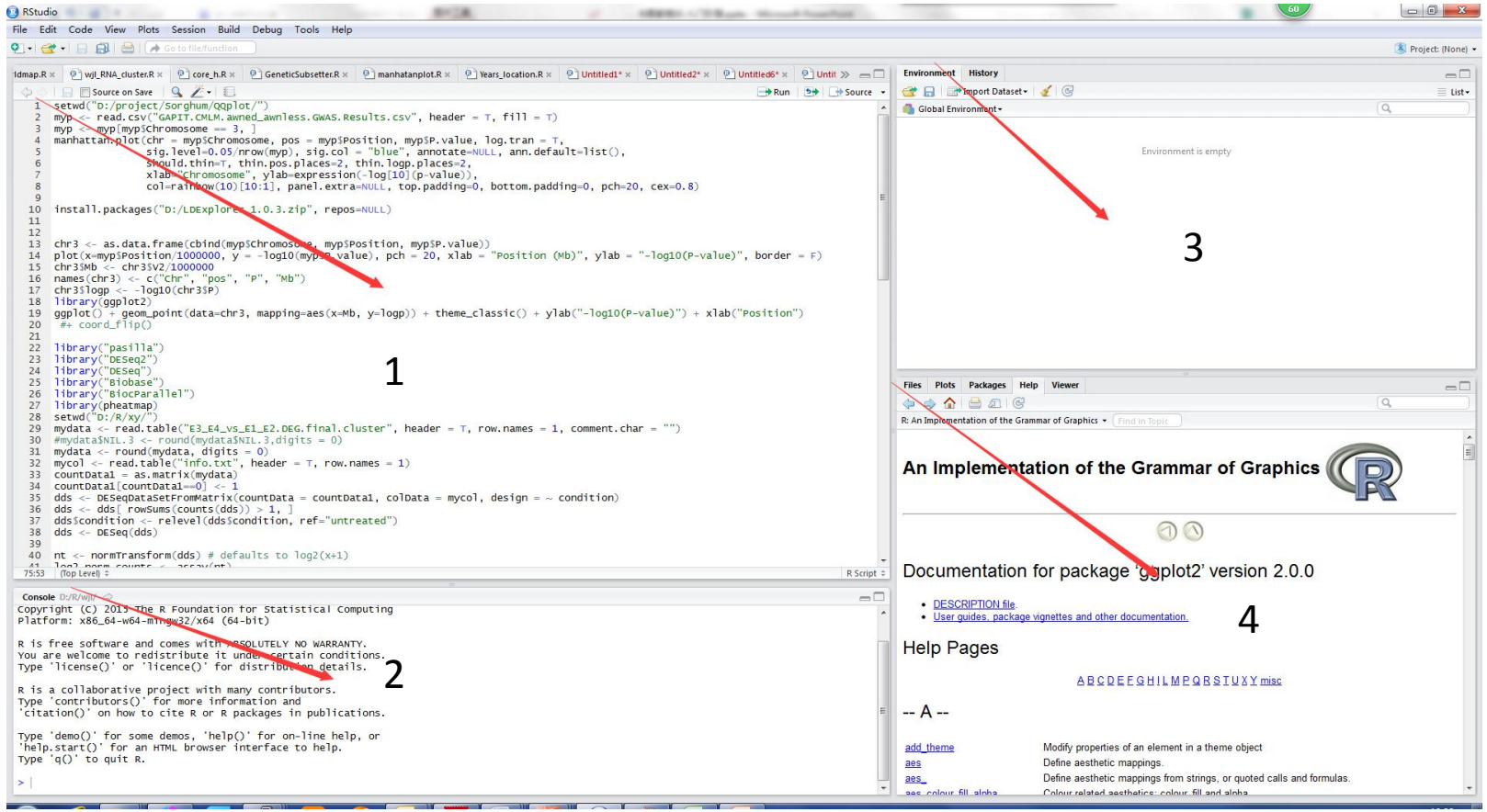
- The latest release (2018-12-20, Eggshell Igloo) R-3.5.2.tar.gz, read what's new in the latest version.
- Sources of R alpha and beta releases (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are available here. Please read about new features and bug fixes before filing corresponding feature requests or bug reports.
- Source code of older versions of R is available here.
- Contributed extension packages

Questions About R

- If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

R语言编程环境

Rstudio https://www.rstudio.com/products/rstudio/download/



Rstudio IDE是目前使用频次最高的R开发工具之一，其在便捷性、协同等工作场景下非常优秀。

1：Text Edictor 编辑器

2：Console 控制台

3：Global Environment 显示环境变量的区间

4：展示台，可以浏览文件目录，展示绘图结果，显示已安装和加载的软件包，显示帮助文档等

百迈客生物科技
BIOMARKER TECHNOLOGIES

② R语言常用数据结构

向量、矩阵、数组、数据框、列表、因子

常用数据结构及其类型

百迈客生物科技
BIOMARKER TECHNOLOGIES

对象	描述	模式	是否允许同一个对象中有多种模式
向量	一系列元素的组合	数值型,字符型,复数型,逻辑型	否
因子	因子是一个分类变量	**数值型,字符型**	否
数组	数组是k维的数据表（**k in 1:n, n** 为正整数）	数值型,字符型,复数型,逻辑型	否
矩阵	二维的数据表，是数组的一个特例	数值型,字符型,复数型,逻辑型	否
数据框	是由一个或几个向量和（或）因子构成，它们必须是等长的，但可以是不同的数据类型。	数值型,字符型,复数型,逻辑型	是
列表	列表可以包含任何类型的对象。	数值型,字符型,复数型,逻辑型,函数,表达式,...	是

数值型 **Numeric** 如 **100, 0, -4.335**

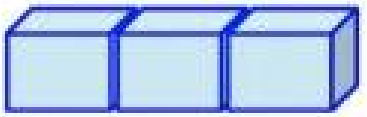
字符型 **Character** 如 **"China"，"Japan"**

逻辑型 **Logical** 如 **TRUE, FALSE，T，F**

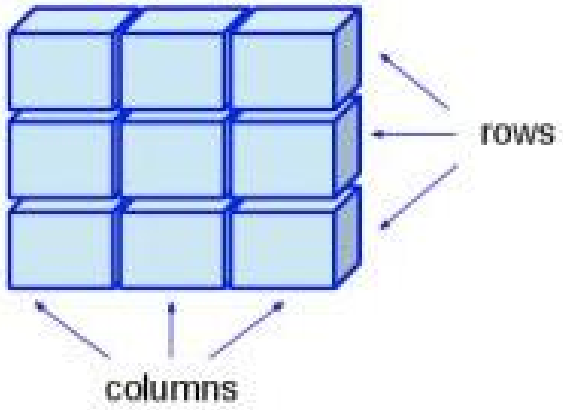
因子型 **Factor** 表示不同类别

复数型 **Complex** 如 **2 + 3i**

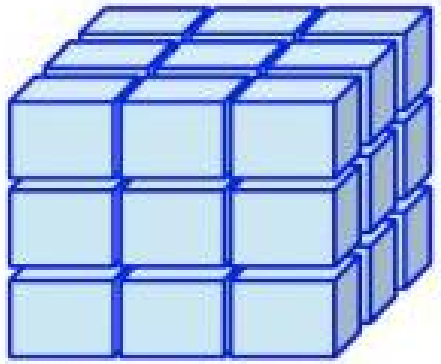
Vector



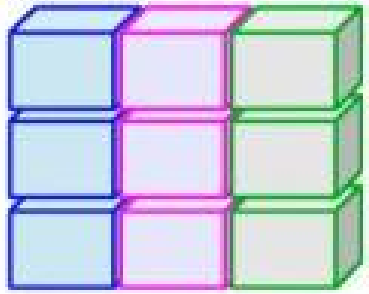
Matrix



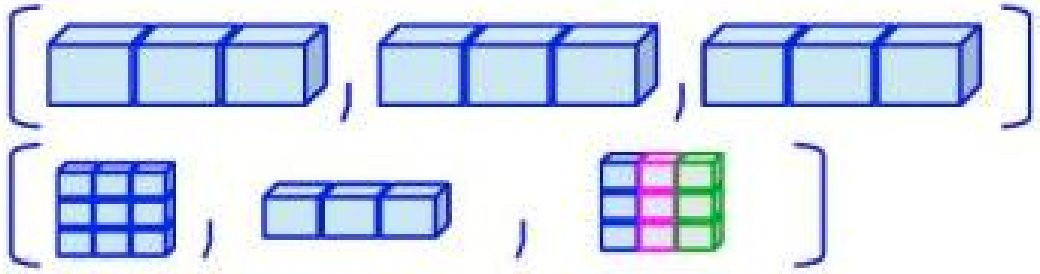
Array



Data Frame
(Table)



Lists



百迈客生物科技
BIOMARKER TECHNOLOGIES

3 R语言基础操作

3.1 数据读写

3.2 包的安装及管理

3.1 数据读写

百迈客生物科技
BIOMARKER TECHNOLOGIES

3.1.1 从键盘读取并将内容输出到屏幕上

……

3.1.2 常见文件格式的读取（csv，txt）

数据准备

- 尽量使用简洁的行名和列名，如需要连接不同的单词作为行名或者列名，应使用"."或下划线"_"连接，例如Rice.Grain.Length或Rice_Grain_Length ，尽量避免使用Rice-Grain-Length或Rice/Grain/Length
- 行名与列名中都不能有重复项
- 以下符号避免出现在行名或列名中：？$ % ^ & * ( ) - # ，,, < > / | \ [ ] { }
- 确保所有的缺失值都用NA表示，软件包有特殊要求的请按其要求表示
- 尽量不要读取Excel文件，推荐读取csv或txt文件，知道文件里数据的分隔符是什么最好

百迈客生物科技
BIOMARKER TECHNOLOGIES

数据读写

读取数据框格式数据

read.table() ###读取常规的数据框（是最常用的一种数据读取形式，分割符号默认是空格）

read.csv() ###数据分割符号默认是逗号

read.csv2() ###数据分割符号默认是分号

read.delim() ###数据分割符号默认是制表符

read.delim2() ###数据分割符号默认是制表符，但小数点标示符是逗号

write.table() ###常见的数据框写出函数

sink() ###不规则数据写出函数

```
read.table(file, header = FALSE, sep = "", quote = "\"'",
           dec = ".", numerals = c("allow.loss", "warn.loss", "no.loss"),
           row.names, col.names, as.is = !stringsAsFactors,
           na.strings = "NA", colClasses = NA, nrows = -1,
           skip = 0, check.names = TRUE, fill = !blank.lines.skip,
           strip.white = FALSE, blank.lines.skip = TRUE,
           comment.char = "#",
           allowEscapes = FALSE, flush = FALSE,
           stringsAsFactors = default.stringsAsFactors(),
           fileEncoding = "", encoding = "unknown", text, skipNul = FALSE)
```

规则输入

file: 数据文件名
header: T/F,是否有表头
sep:字段分隔符
quote:字符串标示符
dec：小数点标示符
row.names:指定行名
col.names:指定列名
na.strings:缺失值标示符
comment.char:注释标示符，读取时跳过该注释行
stringsAsFactors:因子转换
fileEncoding:声明文件编码
check.names:是否检查表头行

备注：读取后检查：str()函数看数据类型与结构，head()查看表头

```
write.table(x, file = "", append = FALSE, quote = TRUE, sep = " ",
            eol = "\n", na = "NA", dec = ".", row.names = TRUE,
            col.names = TRUE, qmethod = c("escape", "double"),
            fileEncoding = "")
```

x:需要保存的对象
file:输出的文件名
append:追加文件
quote:字符段标示符
sep:字段分隔符
row.names:是否保存行名
col.names:是否保存列名

规则输出

百迈客生物科技
BIOMARKER TECHNOLOGIES

```
sink(file = NULL, append = FALSE, type = c("output", "message"),
     split = FALSE)
```

file:输出文件的名字
append:追加

非规则输出

注意：sink()函数一般成对使用

```
sink(file = "data.txt",append = T)      #输出重定向
print("输出内容")                        #输出内容
cat("输出内容")
sink()                                  #结束重定向
```

3.2 包的安装及管理

众多的扩展包是支撑R语言**蓬勃发展**的主要原因，基础包以外的所有包都是扩展包

- **CRAN上的扩展包的安装（在线安装）**

install.packages("包的名字")

install.packages(c("包的名字","包的名字"))

- **Bioconductor上的扩展包的安装（在线安装）**

source("http://bioconductor.org/biocLite.R")

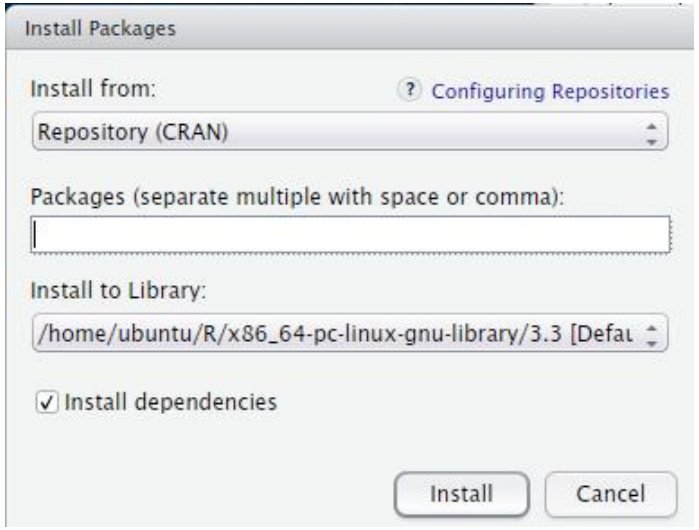
biocLite("包的名字")

- **文本型的扩展包（或称为函数）的安装（调用）**

source("http://bioconductor.org/biocLite.R")

source("http://zzlab.net/GAPIT/gapit_functions.txt")

- **Rstudio安装**



扩展包的调用

library(ggplot2)　require(ggplot2)

关闭加载的扩展包

detach(package:ggplot2)

帮助

- 对于基础包，每次启动R时会自动加载基础包中的函数，因此输入？函数名或者help(函数名)即可获取函数的帮助文档
- 对于扩展包，首先需要加载到R环境中，再用上述命令

ibrary(ggplot2) help(geom_abline)或?geom_abline

- 模糊搜索：两个问号+搜索主题

??heatmap

上述帮助查询基本都会返回一个帮助文档

帮助文档的构成：

描述（Description）

用法（Usage）

各个参数的具体类型与含义（Arguments）

某些参数的细节（Details）

返回值（Value）

注意事项（Note）

作者（Authors）

参考文献（Reference）

其他类似函数（See also）

举例（Examples）等

以上红色加粗部分是应当重点阅读的部分

百迈客生物科技
BIOMARKER TECHNOLOGIES

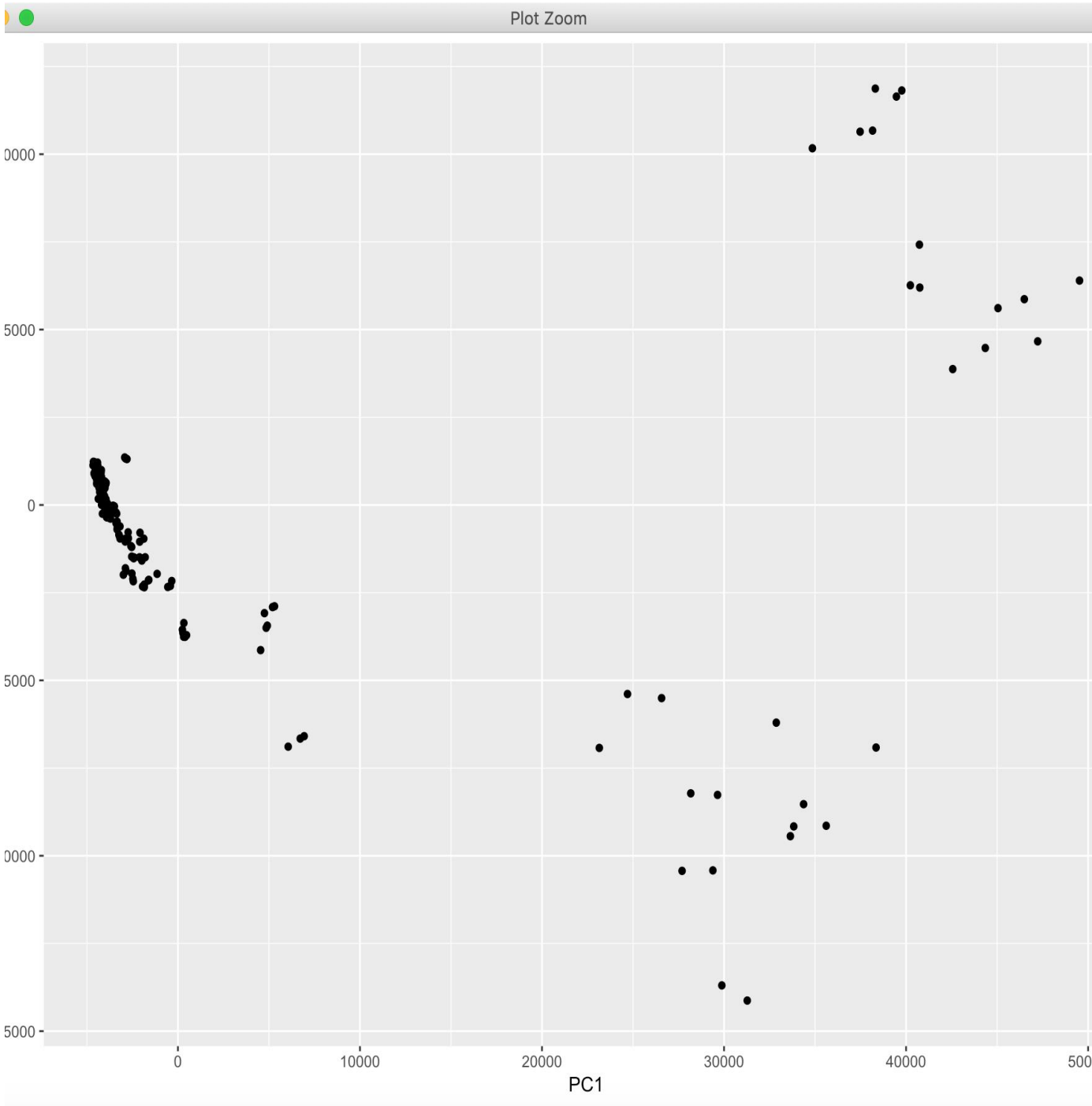
4 绘图基础

4.1 PCA图

4.2 聚类热图

PCA图

- PCA通过线性变换将原始数据变换为一组各维度线性无关的表示，可用于提取数据的主要特征分量，常用于高维数据的降维。在生物信息分析分析中，PCA常用于分析不同样本之间的相互关系，可以基于表达量或者SNP突变类型进行分析。



PCA图

- autoplot(pam(cleandata, 3), frame = TRUE, frame.type = 'norm') +

#添加图片标题

- labs(title="PCA Analysis") +

#图片标题居中，设置字体大小为20

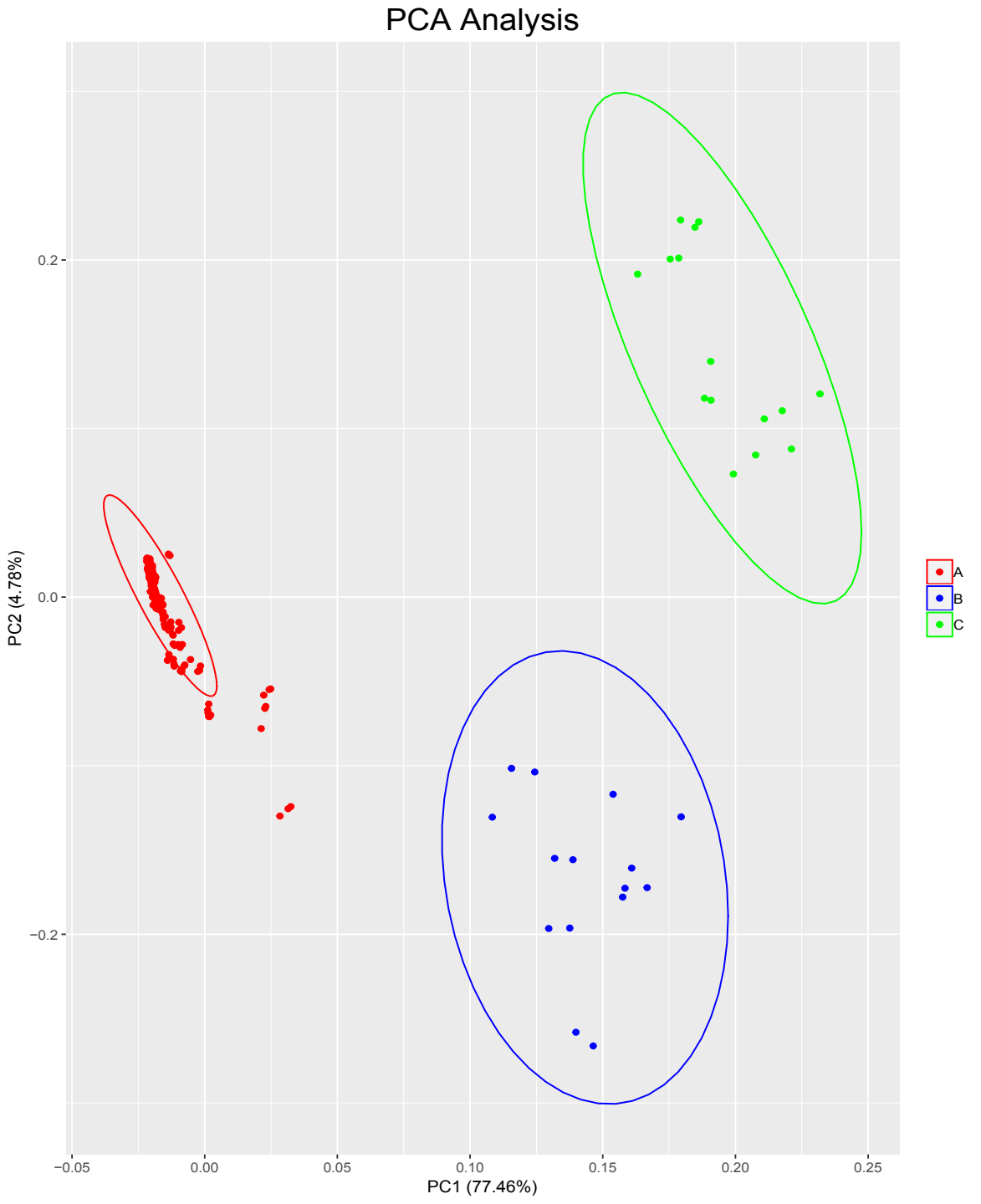
- theme(plot.title = element_text(hjust = 0.5 ,size = 20),

#删除图例标题

- legend.title = element_blank()) +

#手动指定图列的颜色与名称

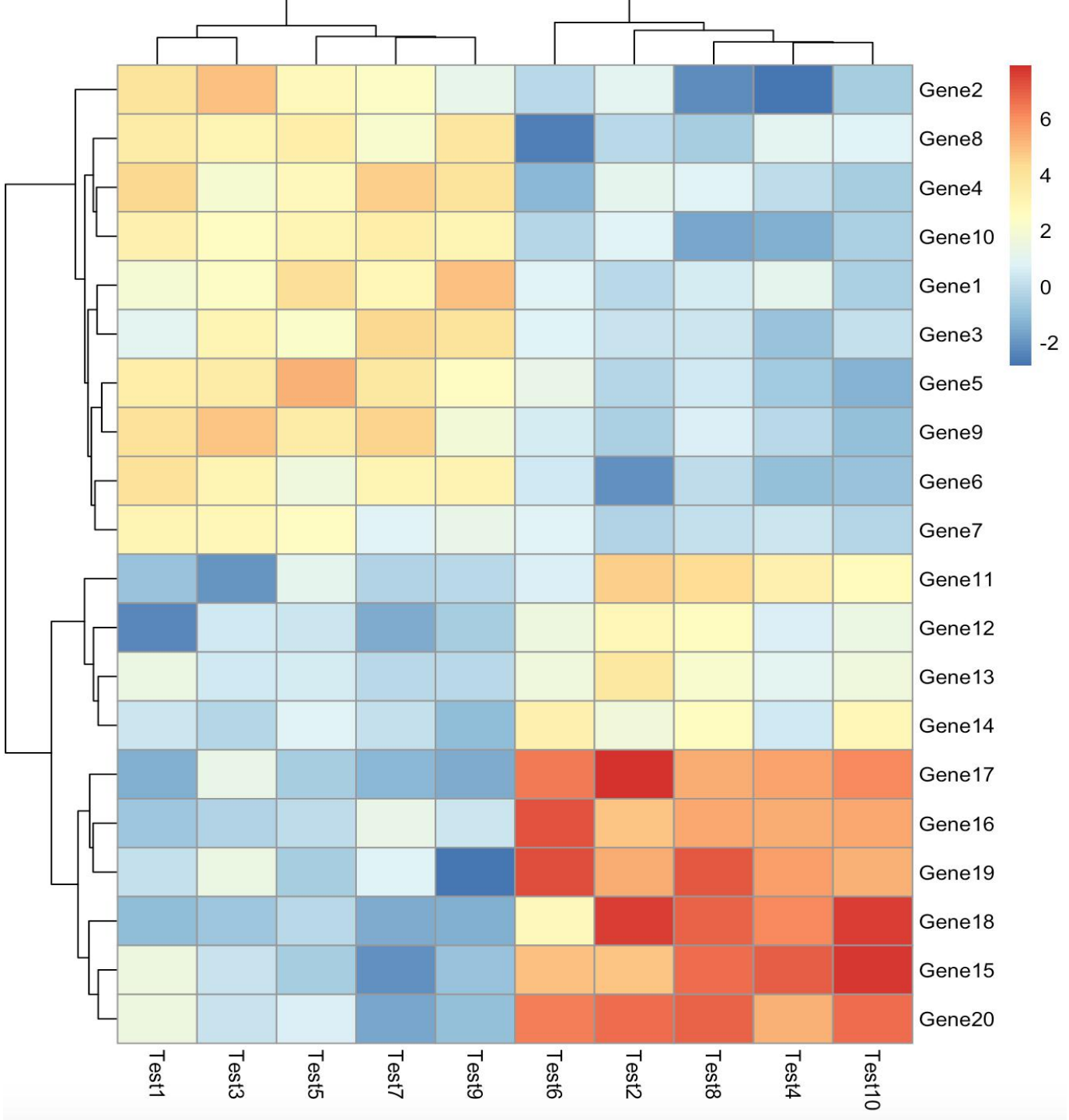
- scale_fill_manual(values = c('red','blue','green'),labels = c('A','B','C')) +
- scale_color_manual(values = c('red','blue','green'),labels = c('A','B','C'))



聚类热图

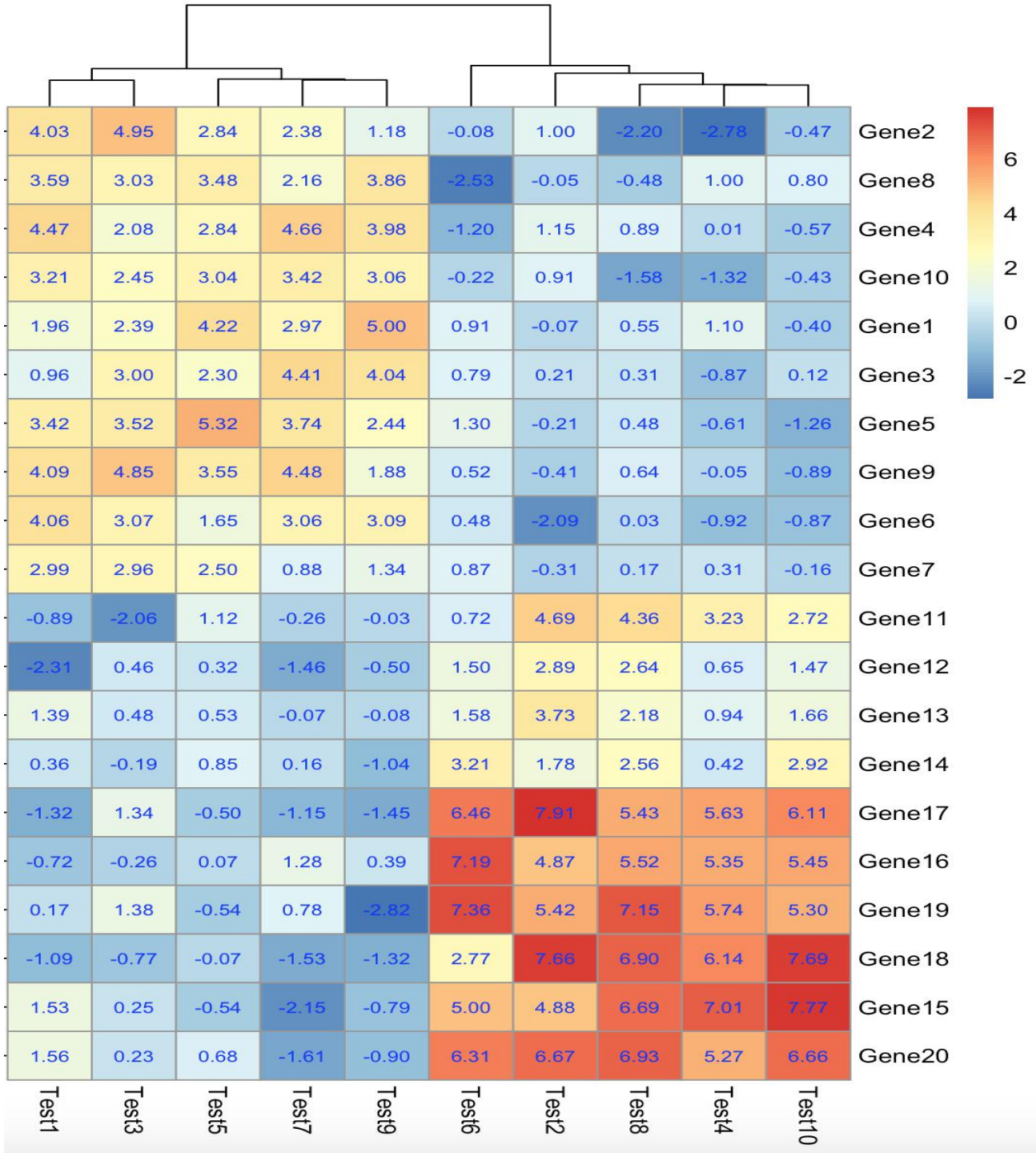
#绘图

- pheatmap(test)



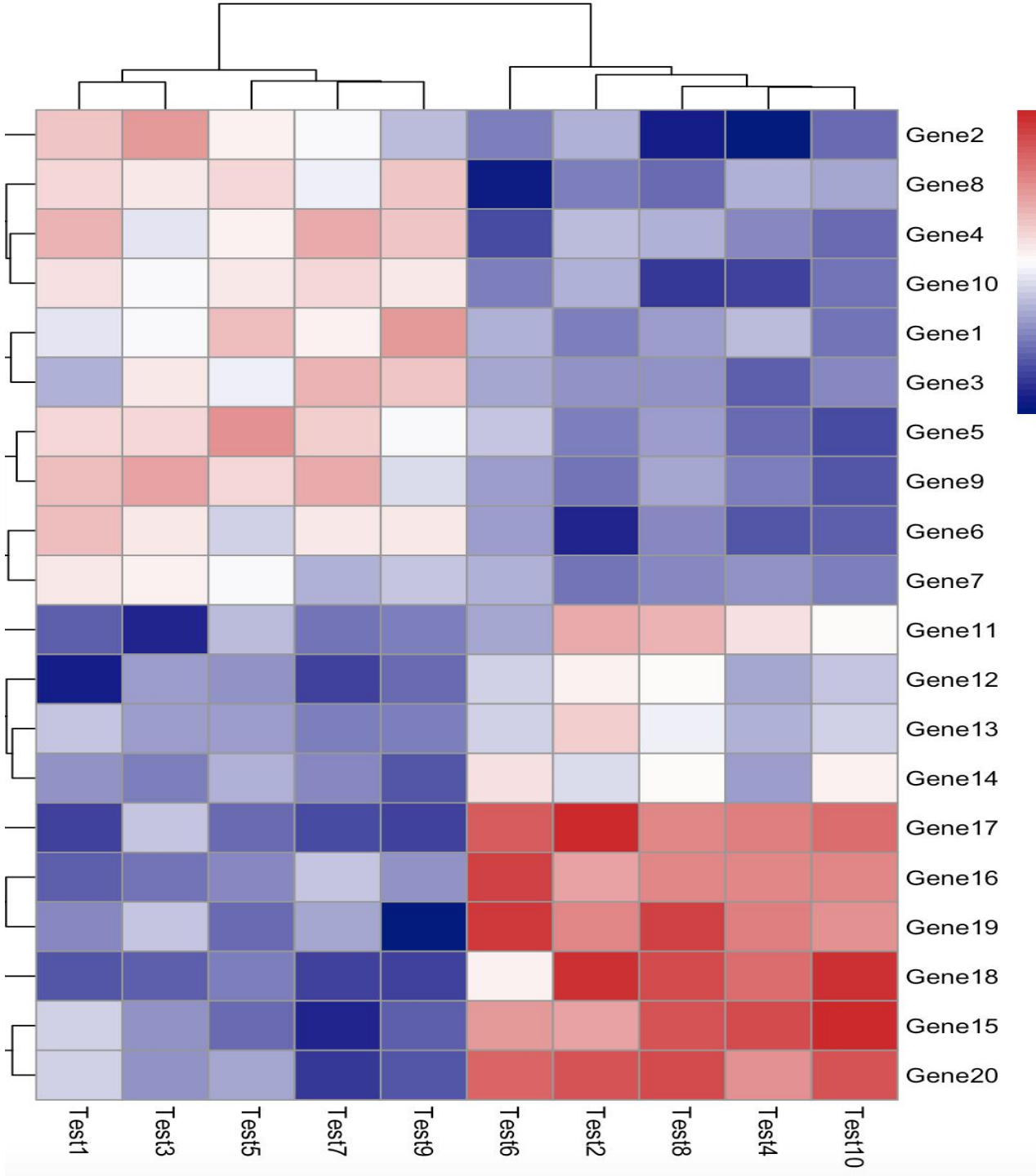
聚类热图

- pheatmap(test, display_numbers = TRUE,number_color = "blue")



聚类热图

- pheatmap(test, color = colorRampPalette(c("navy", "white", "firebrick3"))(50))
- colorRampPalette函数可以根据给定的向量生成渐变色。三个参数分别指定了最大值，中间值和最小值的颜色。



百迈客生物科技
BIOMARKER TECHNOLOGIES

谢谢！

百 迈 客 基 因　为 世 界 创 造 新 的 可 能