**聚类分析**

基地一班 刘利生 201730152009

**Hierarchical clustering（层次分析）**

1. **agglomerative herarchical clustering（凝聚层次聚类）**
2. **divisive herarchical clustering（分裂层次聚类）**

**类间距离计算方法**

**Single-linkage 最小距离**

**Complete-linkage 最大距离**

**Average-linkage 平均距离**

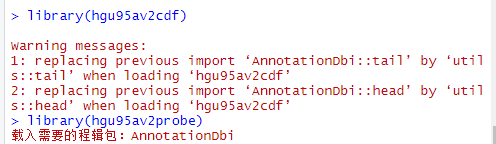
**Centroid 重心**

**Ward 离差平方和**

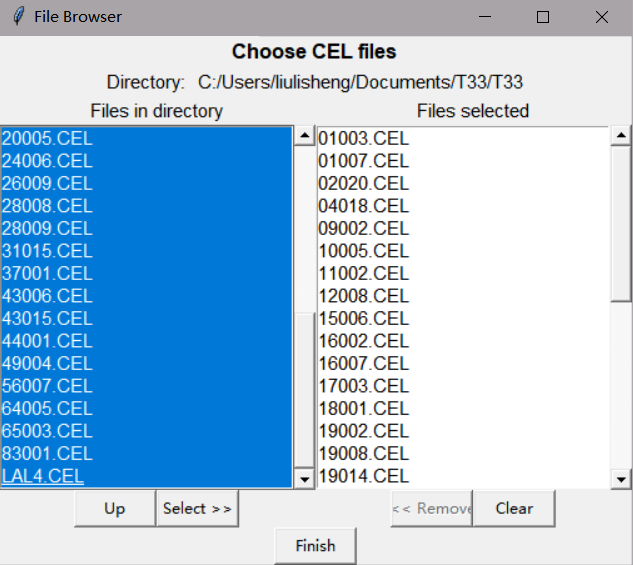
**非特异性筛选**：

导入数据，对数据进行预处理









library(affy)

library(hgu95av2cdf)

library(hgu95av2probe)

#载入affy、hgu95av2cdf、hgu95av2probe程辑包

data.raw<-ReadAffy(widget = T) #导入数据

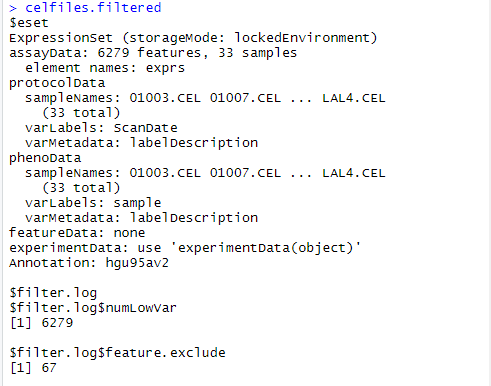


ALLRMA<-rma(data.raw)

#使用RMA算法对data.raw数据进行预处理







library(genefilter)

#载入genefilter程辑包

celfiles.filtered<-nsFilter(ALLRMA,require.entrez = FALSE,remove.dupEntrez=FALSE)

#非特异性筛选

查看celfiles.filtered数据类型，可见其为列表（list）形式，包含有eset（类型为expressionset对象）与filter.log。

**1.用agnes函数进行凝聚层次聚类**

cluster重要的参数

x，函数矩阵或数据框

diss，若x为相异度矩阵（Dissimilarity Matrix），则此参数为TRUE，否则为FALSE

method，选择聚类的方法，默认为“average”（可根据帮助文档设置）



library(cluster) #载入cluster包

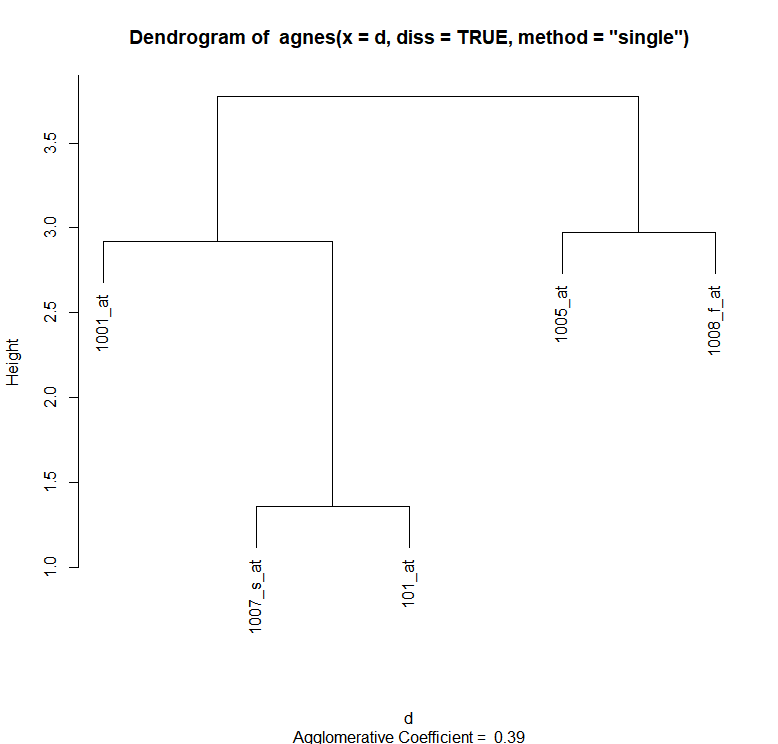
d<-dist(exprs(celfiles.filtered$eset)[1:5,1:5],"euclidean")

#用dist函数计算距离矩阵，使用欧式几何（euclidean）方法处理，求eset矩阵中第一到五行与一到五列的距离矩阵命名为d，对应的diss=T

**最小距离法**







agnSingle<-agnes(d,diss = TRUE,method = "single")

# 最小距离法（single）对d距离矩阵进行处理，diss参数类型为TRUE

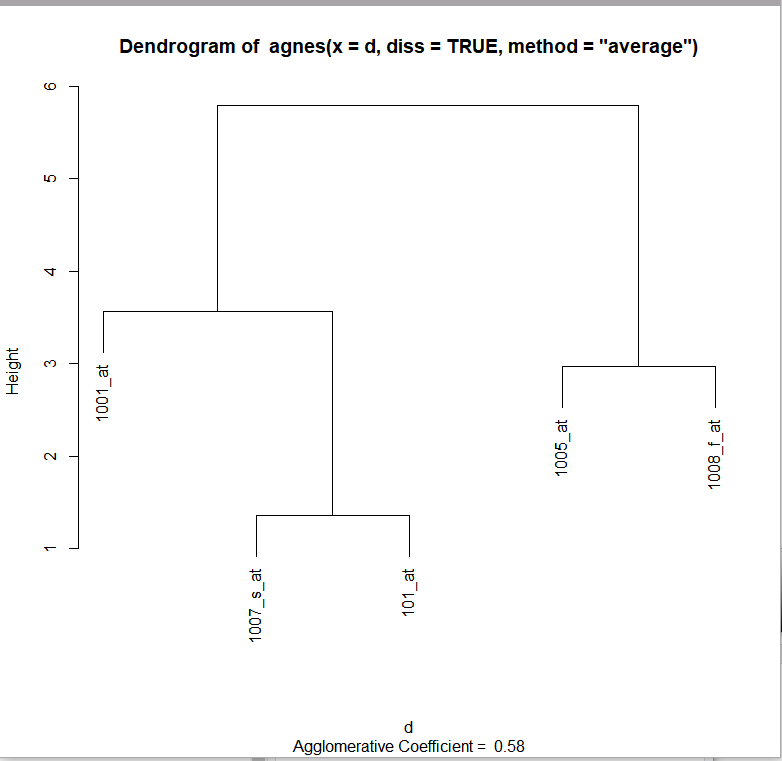
plot（agnSingle）

#查看五行数据的聚类分析处理结果，即将每个基因在五个病人中表达进行区分。

可见五个基因表达联系的紧密程度为**((**(1007-s-at,101-at),1001-at**)**,**(**1005-at,1008-t-at**))**

**平均距离法**





agAverage<-agnes(d,diss=TRUE,method="average")

# 用平均距离法（average）对d距离矩阵，diss参数类型为TRUE进行计算

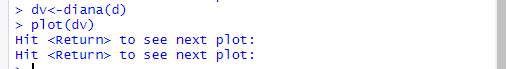
plot(agAverage)

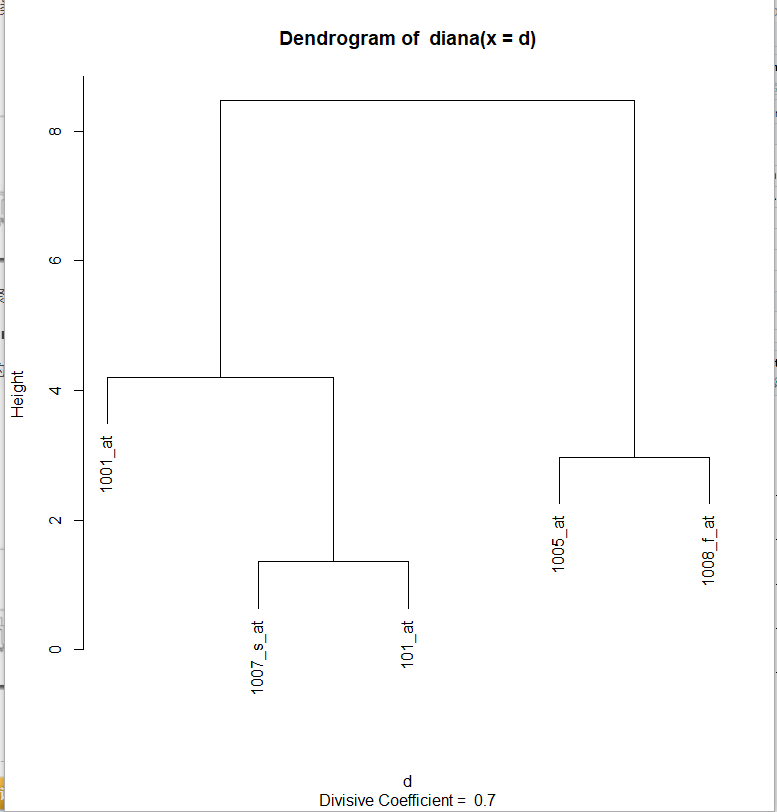
#查看五行数据采取平均距离法处理后的聚类分析处理结果，

可见五个基因表达联系的紧密程度也为**((**(1007-s-at,101-at),1001-at**)**,**(**1005-at,1008-t-at**))**，与最小距离法的处理结果相同。

**2.分裂层次聚类**

采用cluster包中的DIANA(Divisive Analysis)层次聚类算法





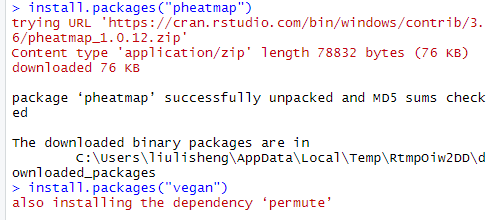
dv<-diana(d) #使用DIANA算法处理d矩阵数据，进行聚类分析

plot(dv) #查看数据采取DIANA算法处理后的聚类分析处理结果，

可见五个基因表达联系的紧密程度为**((**(1007-s-at,101-at),1001-at**)**,**(**1005-at,1008-t-at**))**，与上述凝聚层次聚类两种处理结果相同。

**聚类热图**

**安装程序包**



install.packages(“pheatmap”)

install.packages(“vegan”)

#安装相关程序包

**加载所需软件包**





library(pheatmap)

library(gplots)

library(vegan)

library(permute)

library(lattice)

#载入相关包

**计算出行与列的距离矩阵**

**1.用vegdist函数计算**



data<-exprs(celfiles.filtered$eset)

#提取出celfiles.filtered中eset的表达矩阵(data)

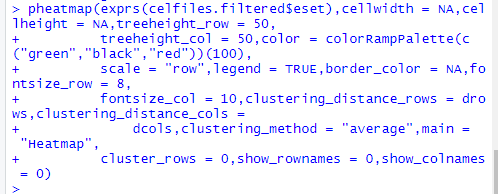
drows<-vegdist(data,method = "euclidean")

#使用vegdist函数计算出data矩阵行的欧式几何距离

dcols<-vegdist(t(data),method = "euclidean")

#t(data)将data矩阵行列倒换，从而计算列间欧式几何距离，命名为dcols

**2.用dist()函数计算变量之间的距离**



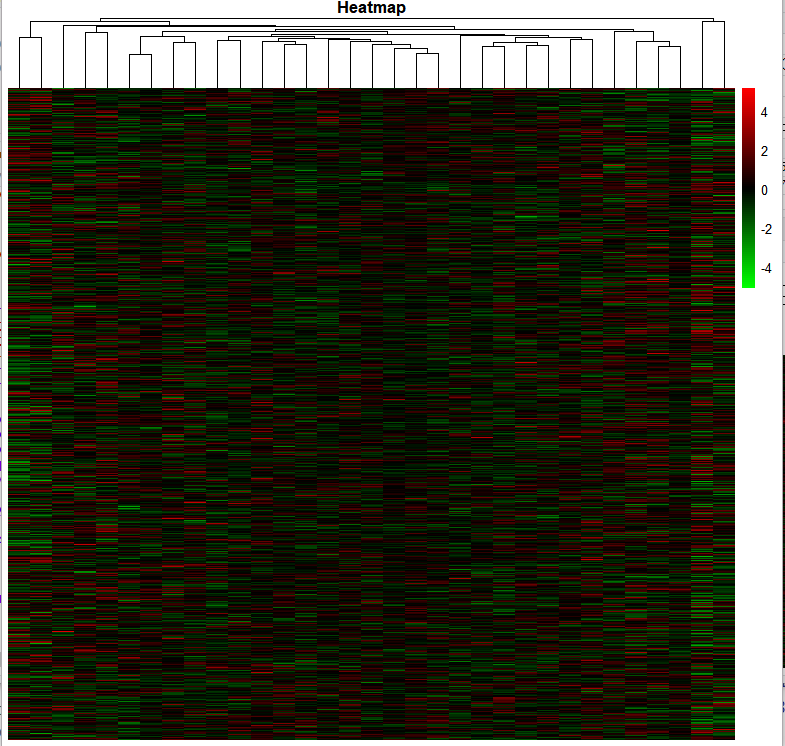
pheatmap(exprs(celfiles.filtered$eset),cellwidth = NA,cellheight = NA,treeheight\_row = 50,

treeheight\_col = 50,color = colorRampPalette(c("green","black","red"))(100),

scale = "row",legend = TRUE,border\_color = NA,fontsize\_row = 8,

fontsize\_col = 10,clustering\_distance\_rows = drows,clustering\_distance\_cols =

dcols,clustering\_method = "average",main = "Heatmap",

cluster\_rows = 0,show\_rownames = 0,show\_colnames = 0) 

#图为对列（病人）进行聚类分析的结果

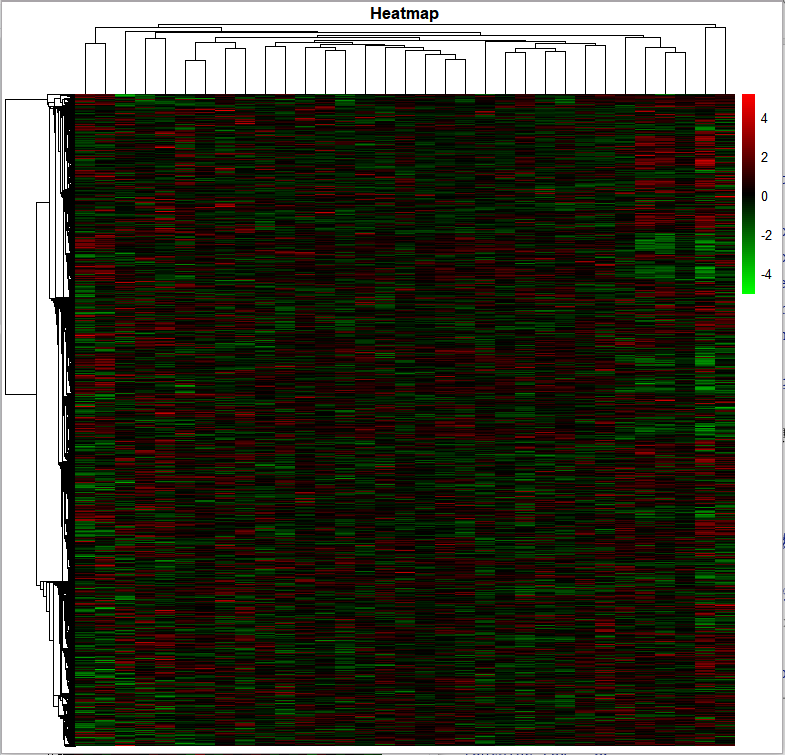
exprs(celfiles.filtered$eset) #选取的表达矩阵

cellwidth = NA #代表色块的宽度

cellheight = NA #为色块长度

cluster\_rows = 0 #为布尔值（boolean），表示是否对行进行聚类，0代表FULSE

将相关代码修改为cluster\_rows = 1后，图表如下



可见图表左边对矩阵的列（探针信息）也进行了聚类分析

clustering\_distance\_rows = drows

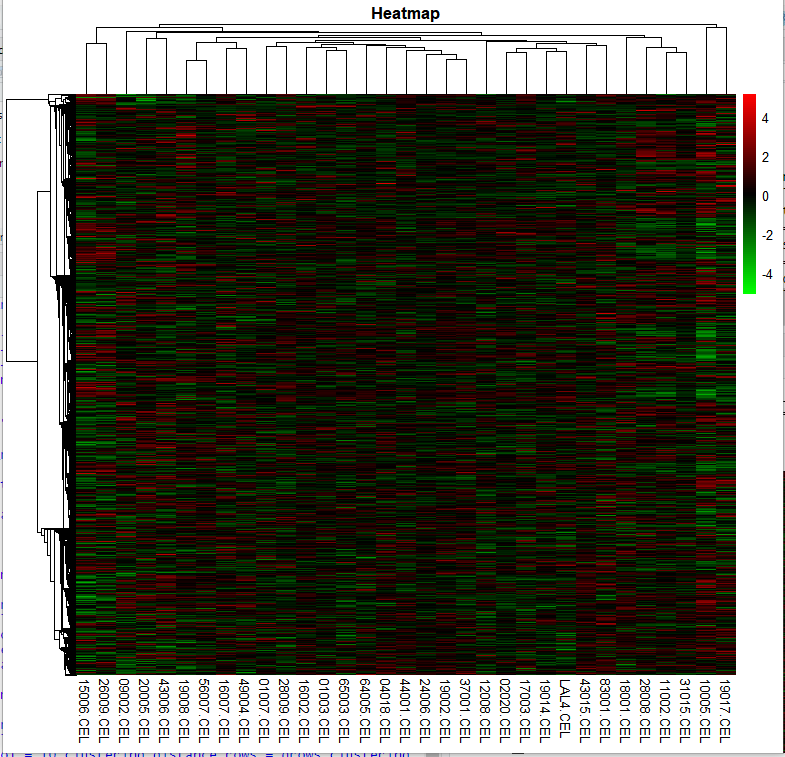
#参数clustering\_distance\_rows设置为d距离矩阵行（row）的格式

clustering\_method = "average" #聚类方法为平均值法

show\_rownames = 0 #布尔值（boolean），是否在图中显示矩阵行名（即探针ID）

show\_colnames = 0 #布尔值（boolean），是否在图中显示矩阵列名（即病人编号）

将相关代码修改为show\_colnames = 1后，图表如下



可见图表下方显示除了列（病人编号）的注释信息

color = colorRampPalette(c("green","black","red"))

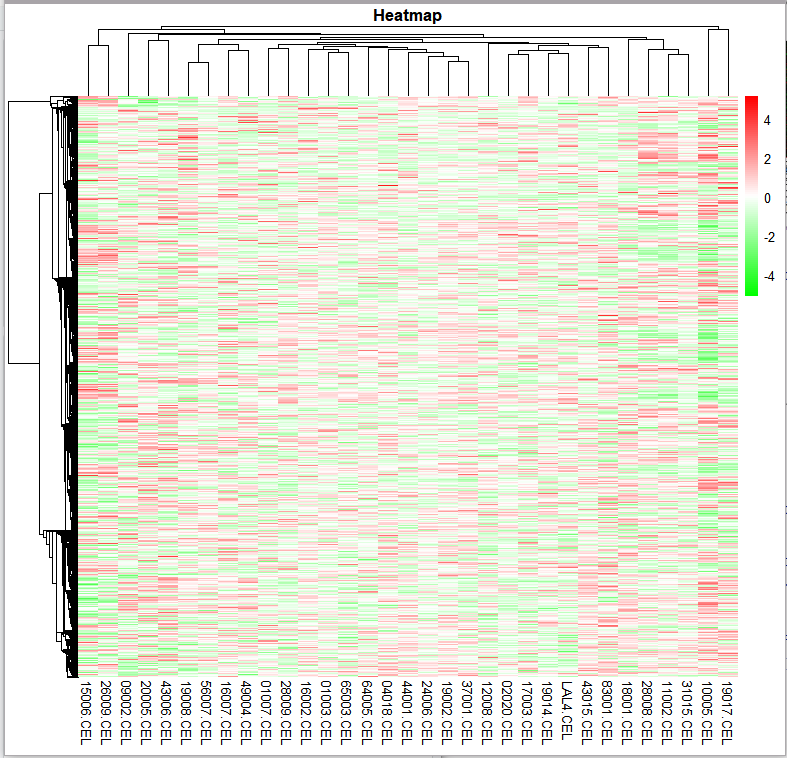
#设置颜色，向量c("green","black","red")第一个参数green为下调，表达正常颜色为black，red代表基因表达上调。

将相关代码修改为color = colorRampPalette(c("green","white","red"))后，

最终代码如下：

pheatmap(exprs(celfiles.filtered$eset),cellwidth = NA,cellheight = NA,treeheight\_row = 50,treeheight\_col = 50,color = colorRampPalette(c("green","white","red"))(100),scale = "row",legend = TRUE,border\_color = NA,fontsize\_row = 8,fontsize\_col = 10,clustering\_distance\_rows = drows,clustering\_distance\_cols = dcols,clustering\_method = "average",main = "Heatmap",cluster\_rows = 1,show\_rownames = 0,show\_colnames = 1)

图表如下



可见图表中表达无明显差异的基因颜色变为白色。

main = "Heatmap" #设置图表名称，为Heatmap

fontsize\_row = 8 #行字体大小

fontsize\_col = 10 #列字体大小