R_homework_Intermediate

dongxu

2019年4月11日

作业-1

根据R包org.Hs.eg.db找到下面ensembl基因ID对应的基因名（symbol）

```
library(org.Hs.eg.db)
```

```
## Loading required package: AnnotationDbi
```

```
## Loading required package: stats4
```

```
## Loading required package: BiocGenerics
```

```
## Loading required package: parallel
```

```
##
## Attaching package: 'BiocGenerics'
```

```
## The following objects are masked from 'package:parallel':
##
##     clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
##     clusterExport, clusterMap, parApply, parCapply, parLapply,
##     parLapplyLB, parRapply, parSapply, parSapplyLB
```

```
## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs
```

```
## The following objects are masked from 'package:base':
##
##     anyDuplicated, append, as.data.frame, basename, cbind,
##     colMeans, colnames, colSums, dirname, do.call, duplicated,
##     eval, evalq, Filter, Find, get, grep, grepl, intersect,
##     is.unsorted, lapply, lengths, Map, mapply, match, mget, order,
##     paste, pmax, pmax.int, pmin, pmin.int, Position, rank, rbind,
##     Reduce, rowMeans, rownames, rowSums, sapply, setdiff, sort,
##     table, tapply, union, unique, unsplit, which, which.max,
##     which.min
```

```
## Loading required package: Biobase
```

```
## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.
```

```
## Loading required package: IRanges
```

```
## Loading required package: S4Vectors
```

```
##
## Attaching package: 'S4Vectors'
```

```
## The following object is masked from 'package:base':
##
##     expand.grid
```

```
##
## Attaching package: 'IRanges'
```

```
## The following object is masked from 'package:grDevices':
##
##     windows
```

```
##
```

```
ensemble_id <- c("ENSG00000000003.13",
                 "ENSG00000000005.5",
                 "ENSG00000000419.11",
                 "ENSG00000000457.12",
                 "ENSG00000000460.15",
                 "ENSG00000000938.11")
## ensembl id的组成主要在小数点前的十五位，小数点后面的为版本号，有版本号的情况下无法进行转换，所以需要删除掉
for(i in 1:length(ensemble_id)){
  ensemble_id[i] <- substr(ensemble_id[i], 1, 15)
}
cols <- c("SYMBOL", "GENENAME")
ens2sym <- select(org.Hs.eg.db, keys = ensemble_id, columns = cols, keytype = "ENSEMBL")
```

```
## 'select()' returned 1:1 mapping between keys and columns
```

```
ens2sym
```

```
##           ENSEMBL   SYMBOL
## 1 ENSG00000000003   TSPAN6
## 2 ENSG00000000005     TNMD
## 3 ENSG00000000419     DPM1
## 4 ENSG00000000457    SCYL3
## 5 ENSG00000000460 C1orf112
## 6 ENSG00000000938      FGR
##                                                    GENENAME
## 1                                             tetraspanin 6
## 2                                               tenomodulin
## 3 dolichyl-phosphate mannosyltransferase subunit 1, catalytic
## 4                                  SCY1 like pseudokinase 3
## 5                       chromosome 1 open reading frame 112
## 6             FGR proto-oncogene, Src family tyrosine kinase
```

```
g2s <- toTable(org.Hs.egSYMBOL)
g2e <- toTable(org.Hs.egENSEMBL)
e_id <- subset(ens2sym, select = c("ENSEMBL"))
g2e_sub <- g2e[which(g2e$ensembl_id %in% e_id$ENSEMBL),]
g2s_sub <- g2s[which(g2s$gene_id %in% g2e_sub$gene_id),]
target <- cbind(ensembl_id = g2e_sub$ensembl_id, g2s_sub)
target
```

```
##            ensembl_id gene_id   symbol
## 1836  ENSG00000000938    2268      FGR
## 5773  ENSG00000000003    7105   TSPAN6
## 6951  ENSG00000000419    8813     DPM1
## 13095 ENSG00000000460   55732 C1orf112
## 13615 ENSG00000000457   57147    SCYL3
## 14229 ENSG00000000005   64102     TNMD
```

作业-2

根据R包hgu133a.db找到探针对应的基因名（symbol）

```
library(hgu133a.db)
```

```
##
```

```
probes <- c("1053_at",
            "117_at",
            "121_at",
            "1255_g_at",
            "1316_at",
            "1320_at",
            "1405_i_at",
            "1431_at",
            "1438_at",
            "1487_at",
            "1494_f_at",
            "1598_g_at",
            "160020_at",
            "1729_at",
            "177_at")
ids <- toTable(hgu133aSYMBOL)
pro2sym <- ids[which(ids$probe_id %in% probes),]
pro2sym
```

```
##     probe_id symbol
## 1    1053_at   RFC2
## 2     117_at  HSPA6
## 3     121_at   PAX8
## 4  1255_g_at GUCA1A
## 5    1316_at   THRA
## 6    1320_at PTPN21
## 7  1405_i_at   CCL5
## 8    1431_at CYP2E1
## 9    1438_at  EPHB3
## 10   1487_at  ESRRA
## 11 1494_f_at CYP2A6
## 12 1598_g_at   GAS6
## 13 160020_at  MMP14
## 14   1729_at  TRADD
## 15    177_at   PLD1
```

作业-3

找到R包CLL内置数据集的表达矩阵里面的TP53基因的表达量，并且绘制在progress.-stable分组的boxplot图

```
library(CLL)
```

```
## Loading required package: affy
```

```
data("sCLLex")
sCLLex
```

```
## ExpressionSet (storageMode: lockedEnvironment)
## assayData: 12625 features, 22 samples
##   element names: exprs
## protocolData: none
## phenoData
##   sampleNames: CLL11.CEL CLL12.CEL ... CLL9.CEL (22 total)
##   varLabels: SampleID Disease
##   varMetadata: labelDescription
## featureData: none
## experimentData: use 'experimentData(object)'
## Annotation: hgu95av2
```

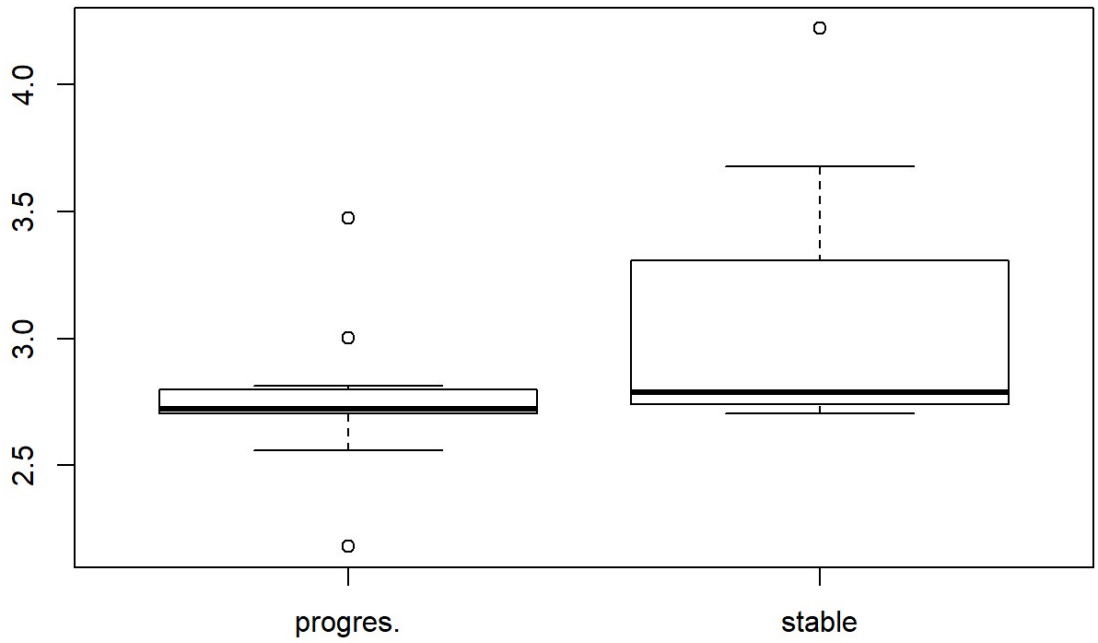
```
exprSet <-exprs(sCLLex)
pd <- pData(sCLLex)
library(hgu95av2.db)
```

```
##
```

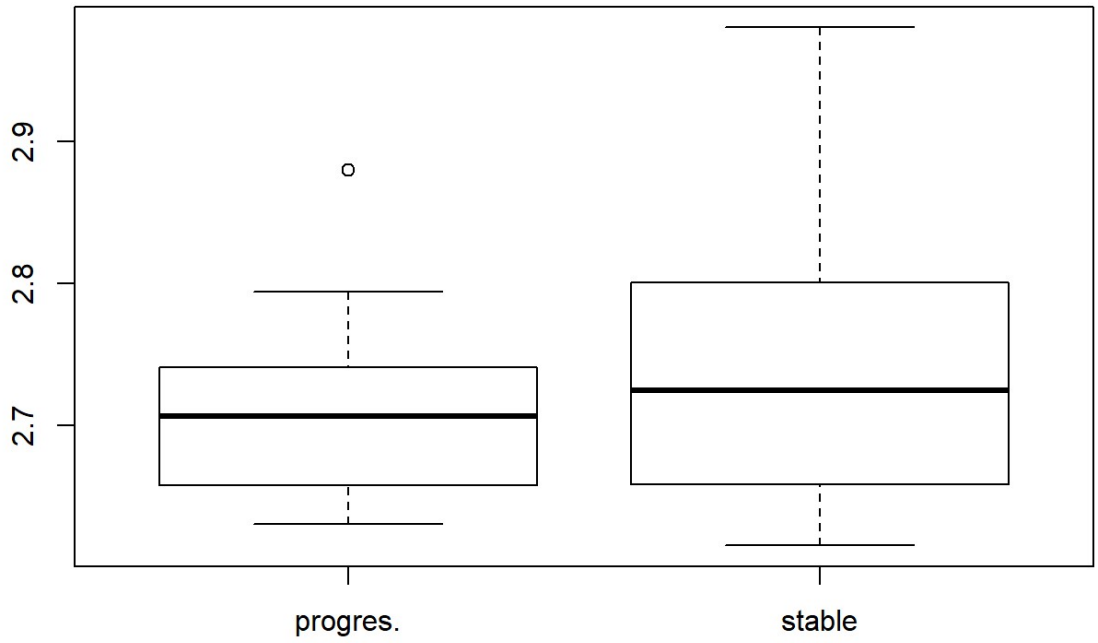
```
CLL_ids <- toTable(hgu95av2SYMBOL)
tp53_prob <- CLL_ids[which(CLL_ids$symbol == "TP53"),]
tp53_prob
```

```
##       probe_id symbol
## 966    1939_at   TP53
## 997  1974_s_at   TP53
## 1420  31618_at   TP53
```

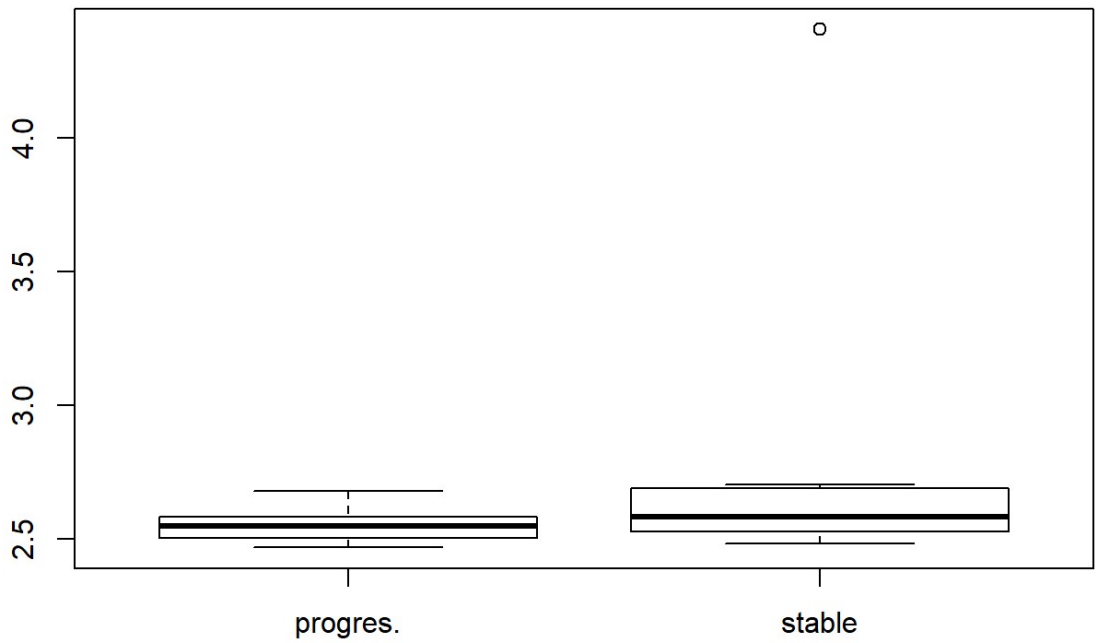
```
boxplot(exprSet[tp53_prob$probe_id[1],]~pd$Disease)
```




```
boxplot(exprSet[tp53_prob$probe_id[2],]~pd$Disease)
```




```
boxplot(exprSet[tp53_prob$probe_id[3],]~pd$Disease)
```



作业-4

找到BRPCA1基因在TCGA数据库的乳腺癌数据集的表达情况

参照简书文章：https://www.jianshu.com/p/d24a47298a14 (https://www.jianshu.com/p/d24a47298a14)

使用该网站进行数据集的选取：http://www.cbioportal.org (http://www.cbioportal.org)

作业-5

找到TP53基因在TCGA数据库的乳腺癌数据集的表达量分组看是否影响生存

参照简书文章：https://www.jianshu.com/p/aa727a67948b (https://www.jianshu.com/p/aa727a67948b)

使用网站：http://www.oncolnc.org/ (http://www.oncolnc.org/)

作业-6

下载数据集GSE17215的表达矩阵并且提取下面的基因画热图

```
#symbol <- c("ACTR3B","ANLN", "BAG1", "BCL2", "BIRC5", "BLVRA", "CCNB1", "CCNE1", "CDC20", "CDC6", "CDCA1", "CDH3", "CENPF", "CEP55", "CXXC5", "EGFR", "ERBB2", "ESR1", "EXO1", "FGFR4", "FOXA1", "FOXC1", "GPR160", "GRB7", "KIF2C", "KNTC2", "KRT14", "KRT17", "KRT5", "MAPT", "MDM2", "MELK", "MIA", "MKI67", "MLPH", "MMP11", "MYBL2", "MYC", "NAT1", "ORC6L", "PGR", "PHGDH", "PTTG1", "RRM2", "SFRP1", "SLC39A6", "TMEM45B", "TYMS", "UBE2C", "UBE2T")##手动输入太蠢了。。。
symbol <- 'ACTR3B ANLN BAG1 BCL2 BIRC5 BLVRA CCNB1 CCNE1 CDC20 CDC6 CDCA1 CDH3 CENPF CEP55 CXXC5 EGFR ERBB2 ESR1 EXO1 FGFR4 FOXA1 FOXC1 GPR160 GRB7 KIF2C KNTC2 KRT14 KRT17 KRT5 MAPT MDM2 MELK MIA MKI67 MLPH MMP11 MYBL2 MYC NAT1 ORC6L PGR PHGDH PTTG1 RRM2 SFRP1 SLC39A6 TMEM45B TYMS UBE2C UBE2T'
symbol <- strsplit(symbol, " ")[[1]]
library(GEOquery)
```

```
## Warning: package 'GEOquery' was built under R version 3.5.2
```

```
## Setting options('download.file.method.GEOquery'='auto')
```

```
## Setting options('GEOquery.inmemory.gpl'=FALSE)
```

```
f <- 'GSE17215_eSet.Rdata'
if(!file.exists(f)){
  gset <- getGEO('GSE17215', destdir = ".",
                 AnnotGPL = F,
                 getGPL = F)
  save(gset, file = f)
}

load('GSE17215_eSet.Rdata')
#class(gset)
#class(gset[[1]])
data <- gset[[1]]
dat <-as.data.frame(exprs(data))
#dat <- cbind(probe_id = rownames(dat),dat)
dim(dat)
```

```
## [1] 22277     6
```

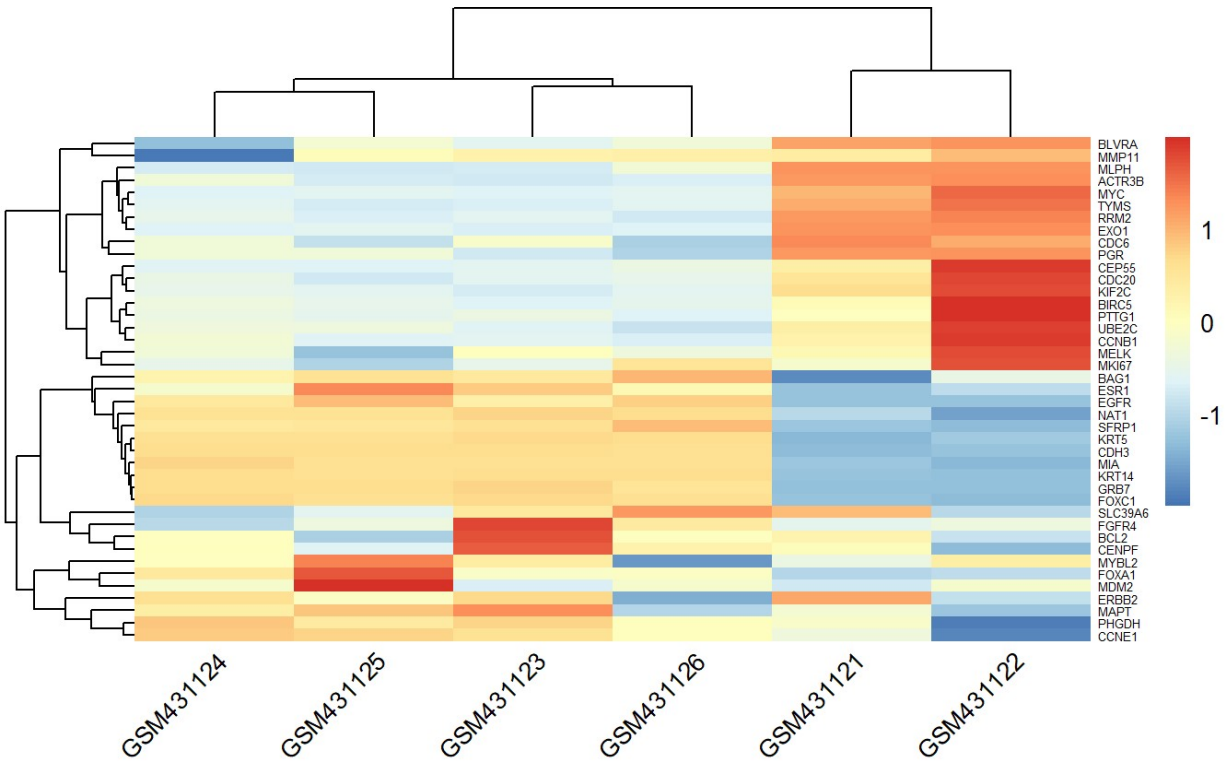
```
library(hgu133a.db)
ids <- toTable(hgu133aSYMBOL)
sym2prob <- ids[ids$symbol %in% symbol,]
#sym2prob_exp <-cbind(symbol = sym2prob$symbol,dat[dat[,1] %in% sym2prob$probe_i
d,])
sym2prob_exp <- dat[rownames(dat) %in% sym2prob$probe_id,]
sym2prob_exp$symbol <- sym2prob$symbol
sym2prob_exp <- sym2prob_exp[!duplicated(sym2prob_exp$symbol),]

hm_data <- sym2prob_exp[,-dim(sym2prob_exp)[2]]
rownames(hm_data) <- sym2prob_exp$symbol
str(hm_data)
```

```
## 'data.frame':    41 obs. of  6 variables:
##  $ GSM431121: num  512.5 42.7 2448.5 585.5 752.2 ...
##  $ GSM431122: num  347.1 46.3 2907.9 702.6 746.6 ...
##  $ GSM431123: num  622.9 46.5 13943.8 85.1 1165.7 ...
##  $ GSM431124: num  639.9 44.5 13080.9 89.6 1216.1 ...
##  $ GSM431125: num  583.4 51.6 13299.1 74 1389.5 ...
##  $ GSM431126: num  534 37.4 13826.8 66.8 1331 ...
```

```
hm_data <- log2(hm_data)
library(pheatmap)
```

```
## Warning: package 'pheatmap' was built under R version 3.5.2
```

```
pheatmap(hm_data, scale = "row",cellheight = 5, fontsize_row = 5, angle_col = 45, b
order_color = NA)
```



作业-7

下载数据集GSE24673的表达矩阵计算样品的相关性并且绘制热图，需要标记上样品分组信息

```
options(stringsAsFactors = F)
GSE_name <- 'GSE24673'
options('download.file.method.GEOquery' = 'libcurl')
gset <- getGEO(GSE_name, getGPL = F)
```

```
## Found 1 file(s)
```

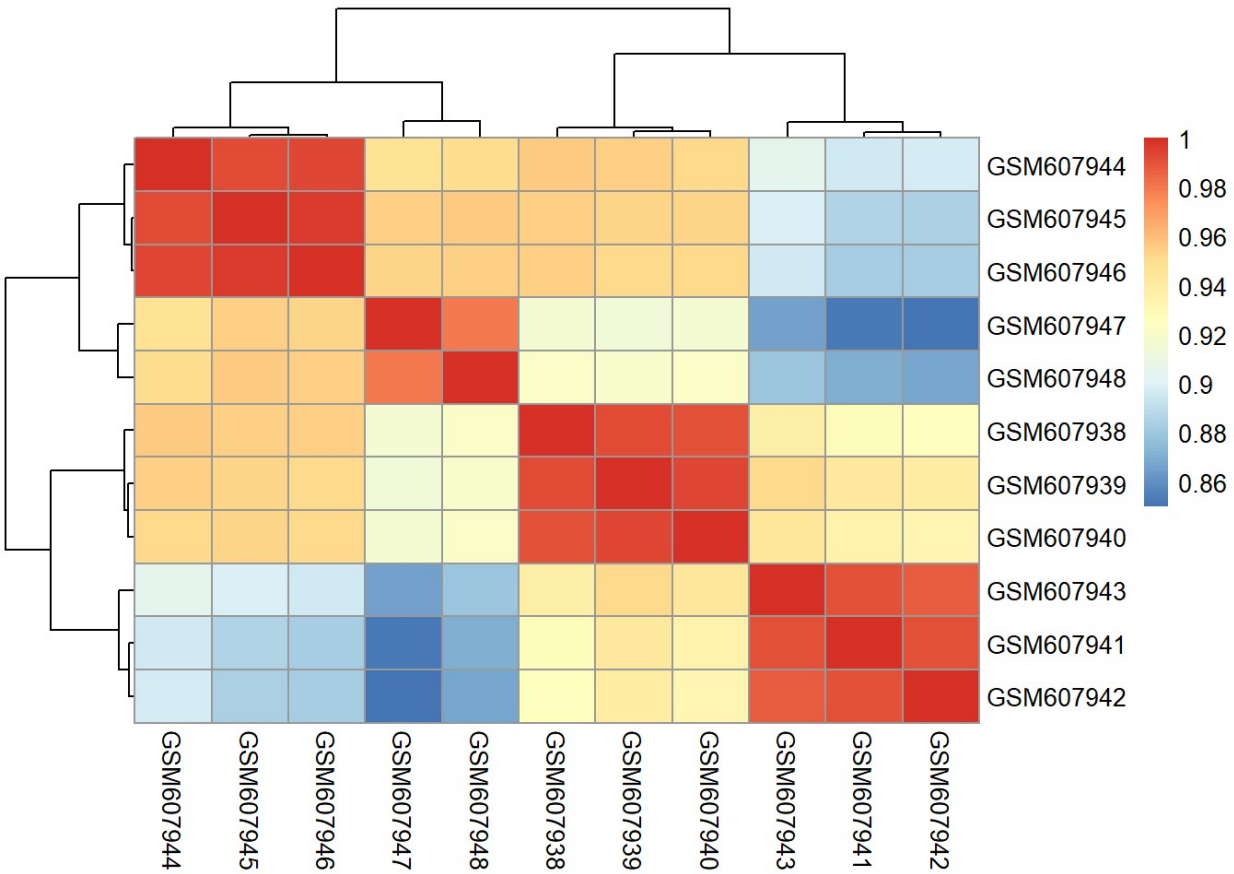
```
## GSE24673_series_matrix.txt.gz
```

```
## Parsed with column specification:
## cols(
##   ID_REF = col_double(),
##   GSM607938 = col_double(),
##   GSM607939 = col_double(),
##   GSM607940 = col_double(),
##   GSM607941 = col_double(),
##   GSM607942 = col_double(),
##   GSM607943 = col_double(),
##   GSM607944 = col_double(),
##   GSM607945 = col_double(),
##   GSM607946 = col_double(),
##   GSM607947 = col_double(),
##   GSM607948 = col_double()
## )
```

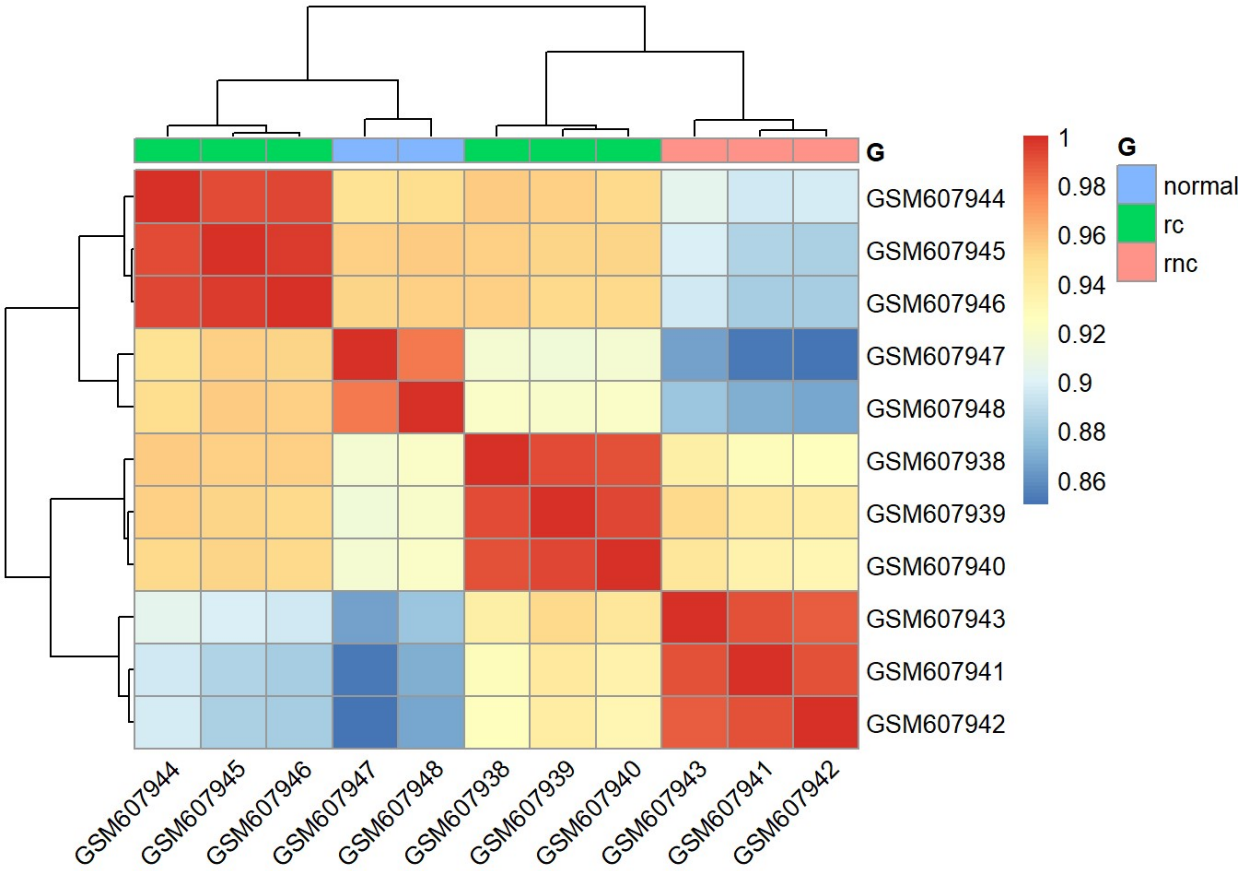
```
save(gset, file = 'GSE24673_gset.Rdata')
load(file = 'GSE24673_gset.Rdata')
data <- gset[[1]]
dat <- exprs(data)
head(dat)
```

```
##         GSM607938 GSM607939 GSM607940 GSM607941 GSM607942 GSM607943
## 7896736     6.231     6.353     6.702     6.663     7.000     7.329
## 7896738     6.607     6.895     6.856     8.136     8.048     8.365
## 7896740     7.191     7.123     7.196     7.845     7.756     8.139
## 7896742     7.447     6.837     7.220     7.391     7.379     7.396
## 7896744     8.305     8.197     8.172     9.544     9.528     9.562
## 7896746    10.554    10.770    10.939    11.040    10.694    11.182
##         GSM607944 GSM607945 GSM607946 GSM607947 GSM607948
## 7896736     7.025     6.111     6.537     6.143     6.851
## 7896738     5.517     5.535     5.591     4.983     5.728
## 7896740     6.123     6.131     6.071     5.478     6.330
## 7896742     6.780     7.130     7.148     7.769     7.197
## 7896744     6.994     6.850     6.735     6.095     7.215
## 7896746     8.057     8.347     7.948     9.994    10.042
```

```
pd <- pData(data)
group_list <- c(rep('rc',3), rep('rnc',3), rep('rc',3), rep('normal',2))
correlation <- cor(dat)
pheatmap(correlation)
```




```
annot <- data.frame(G = group_list)
rownames(annot) <- colnames(correlation)

pheatmap(correlation, annotation_col = annot, angle_col = 45)
```



作业-8 ## 找到GPL6244 platform of Affymetrix Human Gene 1.0 ST Array对应的bioconductor注释包，并且进行安装

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install("hugene10sttranscriptcluster.db", version = "3.8")
```

```
## Bioconductor version 3.8 (BiocManager 1.30.4), R 3.5.1 (2018-07-02)
```

```
## Installing package(s) 'hugene10sttranscriptcluster.db'
```

```
## installing the source package 'hugene10sttranscriptcluster.db'
```

```
## installation path not writeable, unable to update packages: class,
##   cluster, codetools, MASS, Matrix, mgcv, nlme, rpart, survival
```

```
## Update old packages: 'agricolae', 'backports', 'clipr', 'GenomicFeatures',
##   'ggplot2', 'ggthemes', 'labelled', 'plotrix', 'remotes', 'rlang',
##   'Rserve', 'RSpectra', 'shiny', 'spdep', 'tinytex', 'urltools',
##   'usethis', 'WGCNA', 'xfun'
```

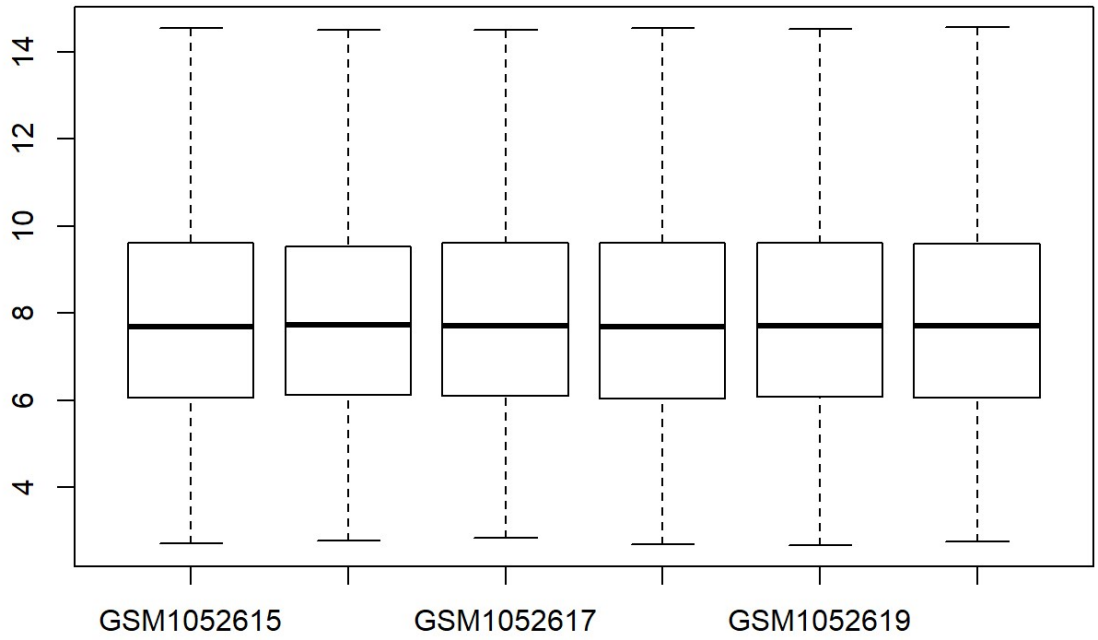
作业-9

下载数据集GSE42872的表达矩阵，并且分别挑出所有样本的最大探针，吧并且找到对应的基因

```
options(stringsAsFactors = F)
f <- 'GSE42872_eSet.Rdata'
if(!file.exists(f)){
  gset <- getGEO('GSE42872', destdir = ".",
                 AnnotGPL = F,
                 getGPL = F)
  save(gset, file = f)
}

load('GSE42872_eSet.Rdata')
#class(gset)
#class(gset[[1]])
data <- gset[[1]]
dat <- exprs(data)
#dat <-as.data.frame(exprs(data))
#dat <- cbind(probe_id = rownames(dat),dat)
dim(dat)
```

```
## [1] 33297     6
```

```
pd <- pData(data)
boxplot(dat)
```




```
mean_max <- which(apply(dat, 1, mean) == max(apply(dat, 1, mean)))

sd_max <- which(apply(dat, 1, sd) == max(apply(dat, 1, sd)))

mad_max <- which(apply(dat, 1, mad) == max(apply(dat, 1, mad)))

target <- dat[c(mean_max, sd_max, mad_max),]
target
```

```
##         GSM1052615 GSM1052616 GSM1052617 GSM1052618 GSM1052619 GSM1052620
## 7978905    14.5467    14.4963   14.51870    14.5476    14.5239    14.5641
## 8133876     4.5461     4.4021    4.49239    10.2506    10.2148    10.3157
## 8133876     4.5461     4.4021    4.49239    10.2506    10.2148    10.3157
```

```
# row_name <- rownames(target)
# row_name_1 <- unlist(lapply(row_name, function(x) strsplit(x, split = "_")[[1]]
[1]))
# row_name_1
library(hugene10sttranscriptcluster.db)
```

```
##
```

```
h10symb <- toTable(hugene10sttranscriptclusterSYMBOL)

prob2symb <- h10symb[h10symb$probe_id %in% rownames(target),]
prob2symb
```

```
##       probe_id symbol
## 16473  8133876   CD36
```

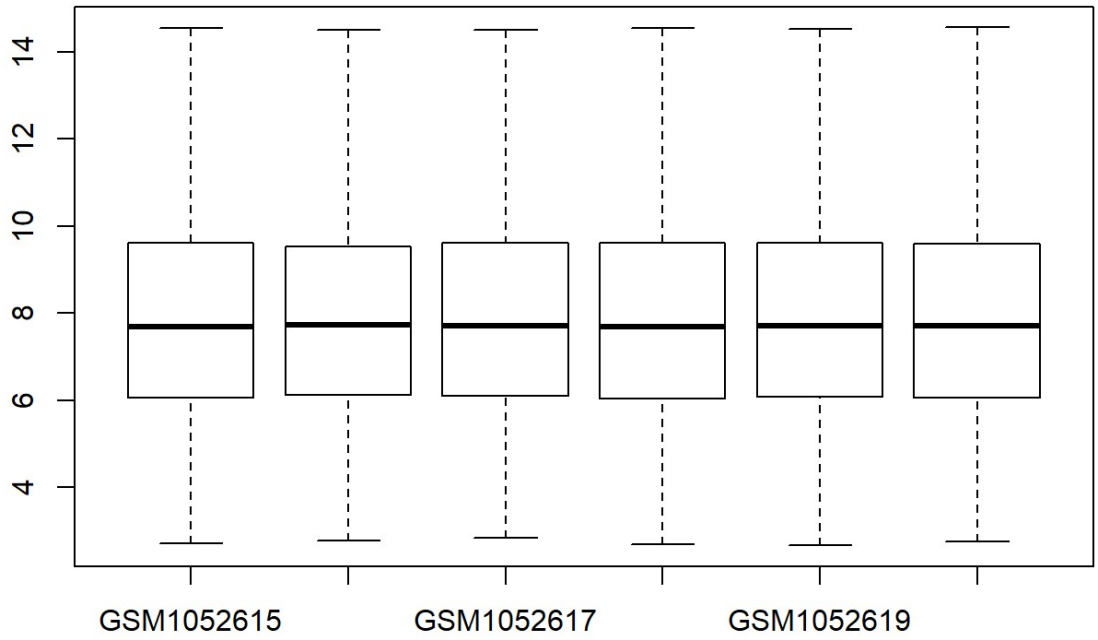
作业-10

下载GSE42872的表达矩阵，根据分组进行limma的差异分析，得到差异结果矩阵

```
options(stringsAsFactors = F)
# f <- "GSE42872_eSet.Rdata"
# if(!file.exists(f)){
#   gset <- getGEO('GSE42872', destdir = ".",
#                  AnnotGPL = F,
#                  getGPL = F)
#   save(gset, file = f)
# }
#
# load('GSE42872_eSet.Rdata')
#class(gset)
#class(gset[[1]])
data <- gset[[1]]
dat <- exprs(data)
#dat <-as.data.frame(exprs(data))
#dat <- cbind(probe_id = rownames(dat),dat)
dim(dat)
```

```
## [1] 33297     6
```

```
pd <- pData(data)
boxplot(dat)
```




```
group_list <- unlist(lapply(pd$title, function(x) strsplit(x, split = ' ')[[1]]
[4]))
exprSet <- dat
head(exprSet)
```

```
##         GSM1052615 GSM1052616 GSM1052617 GSM1052618 GSM1052619 GSM1052620
## 7892501    7.24559    6.80686    7.73301    6.18961    7.05335    7.20371
## 7892502    6.82711    6.70157    7.02471    6.20493    6.76554    6.24252
## 7892503    4.39977    4.50781    4.88250    4.36295    4.18137    4.73492
## 7892504    9.48025    9.67952    9.63074    9.69200    9.91324    9.65897
## 7892505    4.54734    4.45247    5.11753    4.87307    5.15505    3.99340
## 7892506    6.80701    6.90597    6.72472    6.77028    6.77058    6.77685
```

```
library(limma)
```

```
##
## Attaching package: 'limma'
```

```
## The following object is masked from 'package:BiocGenerics':
##
##     plotMA
```

```
design <- model.matrix(~0+factor(group_list))
colnames(design) <- levels(factor(group_list))
rownames(design) <- colnames(exprSet)

contrast.matrix <- makeContrasts(paste0(unique(group_list), collapse = "-"), level
s = design)

fit <- lmFit(exprSet, design)
fit2 <- contrasts.fit(fit, contrast.matrix)
fit2 <- eBayes(fit2)

tempOutput <- topTable(fit2, coef = 1, n=Inf)
nrDEG <- na.omit(tempOutput)
head(nrDEG)
```

```
##             logFC   AveExpr         t      P.Value    adj.P.Val        B
## 8133876 -5.780170  7.370282 -82.94833 3.495205e-12 1.163798e-07 16.32898
## 7965335  4.212683  9.106625  68.40113 1.437468e-11 2.393169e-07 15.71739
## 7972259 -5.633027  8.763220 -57.61985 5.053466e-11 4.431880e-07 15.04752
## 7972217  3.801663  9.726468  57.21112 5.324059e-11 4.431880e-07 15.01709
## 8129573 -3.263063 10.171635 -50.51733 1.324638e-10 8.821294e-07 14.45166
## 8015806  3.843247  9.667077  45.87910 2.681063e-10 1.487856e-06 13.97123
```