R_homework_Advanced

dongxu

2019年4月17日

作业-1

安装一些R packages

```
## 习惯使用R Studio右下的install进行安装
## 也可使用语句install.package( package.name)
## install.package(c(package.names))
library(ALL)
```

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

```
## 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
```

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

```
library(CLL)
```

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

```
library(pasilla)
library(airway)
```

```
## Loading required package: SummarizedExperiment
```

```
## Loading required package: GenomicRanges
```

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

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

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

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

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

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

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

```
## Loading required package: GenomeInfoDb
```

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

```
## Loading required package: DelayedArray
```

```
## Loading required package: matrixStats
```

```
## Warning: package 'matrixStats' was built under R version 3.5.3
```

```
##
## Attaching package: 'matrixStats'
```

```
## The following objects are masked from 'package:Biobase':
##
##     anyMissing, rowMedians
```

```
## Loading required package: BiocParallel
```

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

```
##
## Attaching package: 'DelayedArray'
```

```
## The following objects are masked from 'package:matrixStats':
##
##     colMaxs, colMins, colRanges, rowMaxs, rowMins, rowRanges
```

```
## The following objects are masked from 'package:base':
##
##     aperm, apply
```

```
library(limma)
```

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

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

```
library(DESeq2)
```

```
## Warning: package 'DESeq2' was built under R version 3.5.2
```

```
library(clusterProfiler)
```

```
##
```

```
## clusterProfiler v3.10.1  For help: https://guangchuangyu.github.io/software/clus
terProfiler
##
## If you use clusterProfiler in published research, please cite:
## Guangchuang Yu, Li-Gen Wang, Yanyan Han, Qing-Yu He. clusterProfiler: an R packa
ge for comparing biological themes among gene clusters. OMICS: A Journal of Integra
tive Biology. 2012, 16(5):284-287.
```

```
##
## Attaching package: 'clusterProfiler'
```

```
## The following object is masked from 'package:DelayedArray':
##
##     simplify
```

```
library(reshape2)
library(ggplot2)
```

作业-2

了解ExpressionSet对象，比如CLL包中有data(sCLLex)，找到它包含的元素，提取表达矩阵（使用exprs函数），查看其大小

参考：

http://www.bio-info-trainee.com/bioconductor_China/software/limma.html (http://www.bio-info-trainee.com/bioconductor_China/software/limma.html)

https://github.com/bioconductor-china/basic/blob/master/ExpressionSet.md (https://github.com/bioconductor-china/basic/blob/master/ExpressionSet.md)

```
data("sCLLex")
expSet <- exprs(sCLLex)
dim(expSet)
```

```
## [1] 12625    22
```

```
##提取描述信息
samples <- sampleNames(sCLLex)
samples
```

```
##  [1] "CLL11.CEL" "CLL12.CEL" "CLL13.CEL" "CLL14.CEL" "CLL15.CEL"
##  [6] "CLL16.CEL" "CLL17.CEL" "CLL18.CEL" "CLL19.CEL" "CLL20.CEL"
## [11] "CLL21.CEL" "CLL22.CEL" "CLL23.CEL" "CLL24.CEL" "CLL2.CEL"
## [16] "CLL3.CEL"  "CLL4.CEL"  "CLL5.CEL"  "CLL6.CEL"  "CLL7.CEL"
## [21] "CLL8.CEL"  "CLL9.CEL"
```

```
pdata <- pData(sCLLex)
pdata
```

```
##           SampleID  Disease
## CLL11.CEL    CLL11 progres.
## CLL12.CEL    CLL12   stable
## CLL13.CEL    CLL13 progres.
## CLL14.CEL    CLL14 progres.
## CLL15.CEL    CLL15 progres.
## CLL16.CEL    CLL16 progres.
## CLL17.CEL    CLL17   stable
## CLL18.CEL    CLL18   stable
## CLL19.CEL    CLL19 progres.
## CLL20.CEL    CLL20   stable
## CLL21.CEL    CLL21 progres.
## CLL22.CEL    CLL22   stable
## CLL23.CEL    CLL23 progres.
## CLL24.CEL    CLL24   stable
## CLL2.CEL      CLL2   stable
## CLL3.CEL      CLL3 progres.
## CLL4.CEL      CLL4 progres.
## CLL5.CEL      CLL5 progres.
## CLL6.CEL      CLL6 progres.
## CLL7.CEL      CLL7 progres.
## CLL8.CEL      CLL8 progres.
## CLL9.CEL      CLL9   stable
```

```
## 生成分组信息
group_list <- as.character(pdata$Disease)
group_list
```

```
##  [1] "progres." "stable"   "progres." "progres." "progres." "progres."
##  [7] "stable"   "stable"   "progres." "stable"   "progres." "stable"
## [13] "progres." "stable"   "stable"   "progres." "progres." "progres."
## [19] "progres." "progres." "progres." "stable"
```

作业-3

了解str，head，help函数的作用，用于提取到的表达矩阵

```
str(expSet)
```

```
##  num [1:12625, 1:22] 5.74 2.29 3.31 1.09 7.54 ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : chr [1:12625] "1000_at" "1001_at" "1002_f_at" "1003_s_at" ...
##   ..$ : chr [1:22] "CLL11.CEL" "CLL12.CEL" "CLL13.CEL" "CLL14.CEL" ...
```

```
head(expSet, n = 3)
```

```
##           CLL11.CEL CLL12.CEL CLL13.CEL CLL14.CEL CLL15.CEL CLL16.CEL
## 1000_at    5.743132  6.219412  5.523328  5.340477  5.229904  4.920686
## 1001_at    2.285143  2.291229  2.287986  2.295313  2.662170  2.278040
## 1002_f_at  3.309294  3.318466  3.354423  3.327130  3.365113  3.568353
##           CLL17.CEL CLL18.CEL CLL19.CEL CLL20.CEL CLL21.CEL CLL22.CEL
## 1000_at    5.325348  4.826131  5.212387  5.285830  5.581859  6.251678
## 1001_at    2.350796  2.325163  2.432635  2.256547  2.348389  2.263849
## 1002_f_at  3.502440  3.394410  3.617099  3.279726  3.391734  3.306811
##           CLL23.CEL CLL24.CEL CLL2.CEL CLL3.CEL CLL4.CEL CLL5.CEL CLL6.CEL
## 1000_at    5.480752  5.216033 5.966942 5.397508 5.281720 5.414718 5.460626
## 1001_at    2.264434  2.344079 2.350073 2.406846 2.341961 2.372928 2.356978
## 1002_f_at  3.341444  3.798335 3.427736 3.453564 3.412944 3.411922 3.396466
##           CLL7.CEL CLL8.CEL CLL9.CEL
## 1000_at   5.897821 5.253883 5.214155
## 1001_at   2.222276 2.254772 2.358544
## 1002_f_at 3.247276 3.255148 3.365746
```

```
help()
```

```
## starting httpd help server ... done
```

```
## 或者在console里面输入? str()即可
```

作业-4

安装并了解hgu95av2.db, 使用ls()查看显示结果

```
##安装bioconductor的package使用以下语句
BiocManager::install("hgu95av2.db", version = "3.8")
```

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

```
## Installing package(s) 'hgu95av2.db'
```

```
## installing the source package 'hgu95av2.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'
```

```
library(hgu95av2.db)
```

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

```
## Loading required package: org.Hs.eg.db
```

```
##
```

```
##
```

```
ls("package:hgu95av2.db")
```

```
##  [1] "hgu95av2"              "hgu95av2.db"
##  [3] "hgu95av2_dbconn"       "hgu95av2_dbfile"
##  [5] "hgu95av2_dbInfo"       "hgu95av2_dbschema"
##  [7] "hgu95av2ACCNUM"        "hgu95av2ALIAS2PROBE"
##  [9] "hgu95av2CHR"           "hgu95av2CHRLENGTHS"
## [11] "hgu95av2CHRLOC"        "hgu95av2CHRLOCEND"
## [13] "hgu95av2ENSEMBL"       "hgu95av2ENSEMBL2PROBE"
## [15] "hgu95av2ENTREZID"      "hgu95av2ENZYME"
## [17] "hgu95av2ENZYME2PROBE"  "hgu95av2GENENAME"
## [19] "hgu95av2GO"            "hgu95av2GO2ALLPROBES"
## [21] "hgu95av2GO2PROBE"      "hgu95av2MAP"
## [23] "hgu95av2MAPCOUNTS"     "hgu95av2OMIM"
## [25] "hgu95av2ORGANISM"      "hgu95av2ORGPKG"
## [27] "hgu95av2PATH"          "hgu95av2PATH2PROBE"
## [29] "hgu95av2PFAM"          "hgu95av2PMID"
## [31] "hgu95av2PMID2PROBE"    "hgu95av2PROSITE"
## [33] "hgu95av2REFSEQ"        "hgu95av2SYMBOL"
## [35] "hgu95av2UNIGENE"       "hgu95av2UNIPROT"
```

作业-5

理解head(toTable(hgu95av2SYMBOL))的用法，找到TP53对应的probe id

```
head(toTable(hgu95av2SYMBOL))
```

```
##    probe_id  symbol
## 1   1000_at   MAPK3
## 2   1001_at    TIE1
## 3 1002_f_at CYP2C19
## 4 1003_s_at   CXCR5
## 5   1004_at   CXCR5
## 6   1005_at   DUSP1
```

```
id_prob <- toTable(hgu95av2SYMBOL)
TP53_prob <- id_prob[which(id_prob$symbol == "TP53"),]
TP53_prob
```

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

```
## or in this way
TP53_prob_1 <- id_prob[id_prob$symbol %in% "TP53",]
TP53_prob_1
```

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

作业-6

理解探针与基因的关系，基因总数，基因最多对应多少个探针

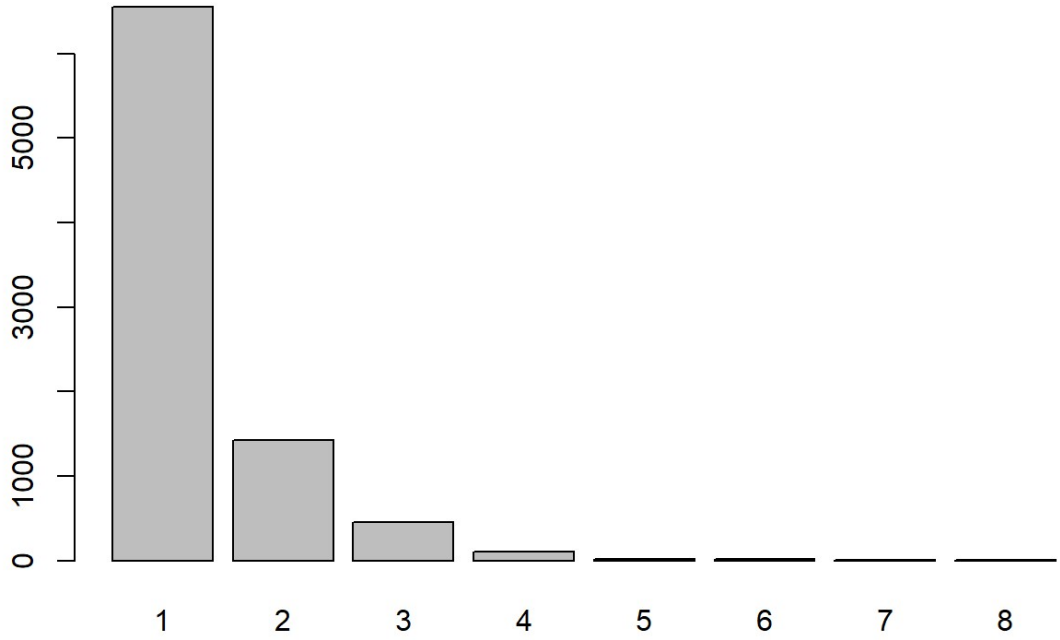
```
dim(id_prob)
```

```
## [1] 11460     2
```

```
length(unique(id_prob$symbol))
```

```
## [1] 8585
```

```
frequency <- table(sort(table(id_prob$symbol)))
barplot(frequency)
```



作业-7

在提取的表达矩阵中找到不存在probe_id

```
in_not <- table(rownames(expSet) %in% id_prob$probe_id)
in_not
```

```
##
## FALSE  TRUE
##  1165 11460
```

作业-8

过滤表达矩阵，删除1165个没有对应基因名字的探针

```
expSet <- expSet[rownames(expSet) %in% id_prob$probe_id,]
dim(expSet)
```

```
## [1] 11460    22
```

作业-9

整合表达矩阵，多个探针对应一个基因的情况下，只保留所有样本中平均表达量最大的那个探针

```
id_prob <- id_prob[match(rownames(expSet), id_prob$probe_id),]
head(id_prob)
```

```
##    probe_id  symbol
## 1   1000_at   MAPK3
## 2   1001_at    TIE1
## 3 1002_f_at CYP2C19
## 4 1003_s_at   CXCR5
## 5   1004_at   CXCR5
## 6   1005_at   DUSP1
```

```
dat <- cbind(subset(id_prob, select = "symbol"), expSet)
rownames(dat) <- rownames(expSet)
head(dat, n = 3)
```

```
##            symbol CLL11.CEL CLL12.CEL CLL13.CEL CLL14.CEL CLL15.CEL
## 1000_at     MAPK3  5.743132  6.219412  5.523328  5.340477  5.229904
## 1001_at      TIE1  2.285143  2.291229  2.287986  2.295313  2.662170
## 1002_f_at CYP2C19  3.309294  3.318466  3.354423  3.327130  3.365113
##           CLL16.CEL CLL17.CEL CLL18.CEL CLL19.CEL CLL20.CEL CLL21.CEL
## 1000_at    4.920686  5.325348  4.826131  5.212387  5.285830  5.581859
## 1001_at    2.278040  2.350796  2.325163  2.432635  2.256547  2.348389
## 1002_f_at  3.568353  3.502440  3.394410  3.617099  3.279726  3.391734
##           CLL22.CEL CLL23.CEL CLL24.CEL CLL2.CEL CLL3.CEL CLL4.CEL
## 1000_at    6.251678  5.480752  5.216033 5.966942 5.397508 5.281720
## 1001_at    2.263849  2.264434  2.344079 2.350073 2.406846 2.341961
## 1002_f_at  3.306811  3.341444  3.798335 3.427736 3.453564 3.412944
##           CLL5.CEL CLL6.CEL CLL7.CEL CLL8.CEL CLL9.CEL
## 1000_at   5.414718 5.460626 5.897821 5.253883 5.214155
## 1001_at   2.372928 2.356978 2.222276 2.254772 2.358544
## 1002_f_at 3.411922 3.396466 3.247276 3.255148 3.365746
```

```
dim(dat)
```

```
## [1] 11460    23
```

```
dat$mean <- apply(dat[,2:dim(dat)[2]], 1, mean)
dat <- dat[order(dat$symbol, dat$mean, decreasing = T),]
dat <- dat[!duplicated(dat$symbol),]
dim(dat)
```

```
## [1] 8585   24
```

```
dat_1 <-data.frame(cbind(rownames(dat), dat))
dat_1 <- dat_1[,-dim(dat_1)[2]]
```

作业-10

更改行名为symbol

```
rownames(dat) <- dat$symbol
dat <- dat[,-c(1,dim(dat)[2])]
head(dat, n=3)
```

```
##       CLL11.CEL CLL12.CEL CLL13.CEL CLL14.CEL CLL15.CEL CLL16.CEL
## ZZZ3   6.645791  7.350613  6.333290   6.60364  6.711462  7.373601
## ZZEF1  5.289264  6.677600  4.447104   7.00826  6.046429  6.413833
## ZYX    3.949769  5.423343  3.540189   5.23442  3.603839  3.687205
##       CLL17.CEL CLL18.CEL CLL19.CEL CLL20.CEL CLL21.CEL CLL22.CEL
## ZZZ3   6.243337  6.730870  7.299798  7.203648  6.519334  6.395689
## ZZEF1  7.369615  6.033872  6.493153  6.631621  6.390880  7.174788
## ZYX    4.191365  3.779226  3.141664  3.648371  6.091596  3.882752
##       CLL23.CEL CLL24.CEL CLL2.CEL CLL3.CEL CLL4.CEL CLL5.CEL CLL6.CEL
## ZZZ3   6.651841  7.338645 5.897972 6.713280 6.529733 6.680138 6.056228
## ZZEF1  4.837948  5.793722 6.998910 6.347929 6.267050 4.822419 5.666789
## ZYX    3.953285  3.554797 6.733884 4.456778 3.652998 3.825987 4.375647
##       CLL7.CEL CLL8.CEL CLL9.CEL
## ZZZ3  6.868983 6.564657 6.607440
## ZZEF1 6.607534 6.553768 6.482294
## ZYX   3.962673 3.618525 4.726375
```

作业-11

对上一题得到的表达矩阵进行探索，画第一个样本的所有基因表达量的boxplot，histgram，density plot，然后花所有样本的这些图

```
dat_melt <- melt(dat_1)
```

```
## Using rownames.dat., symbol as id variables
```

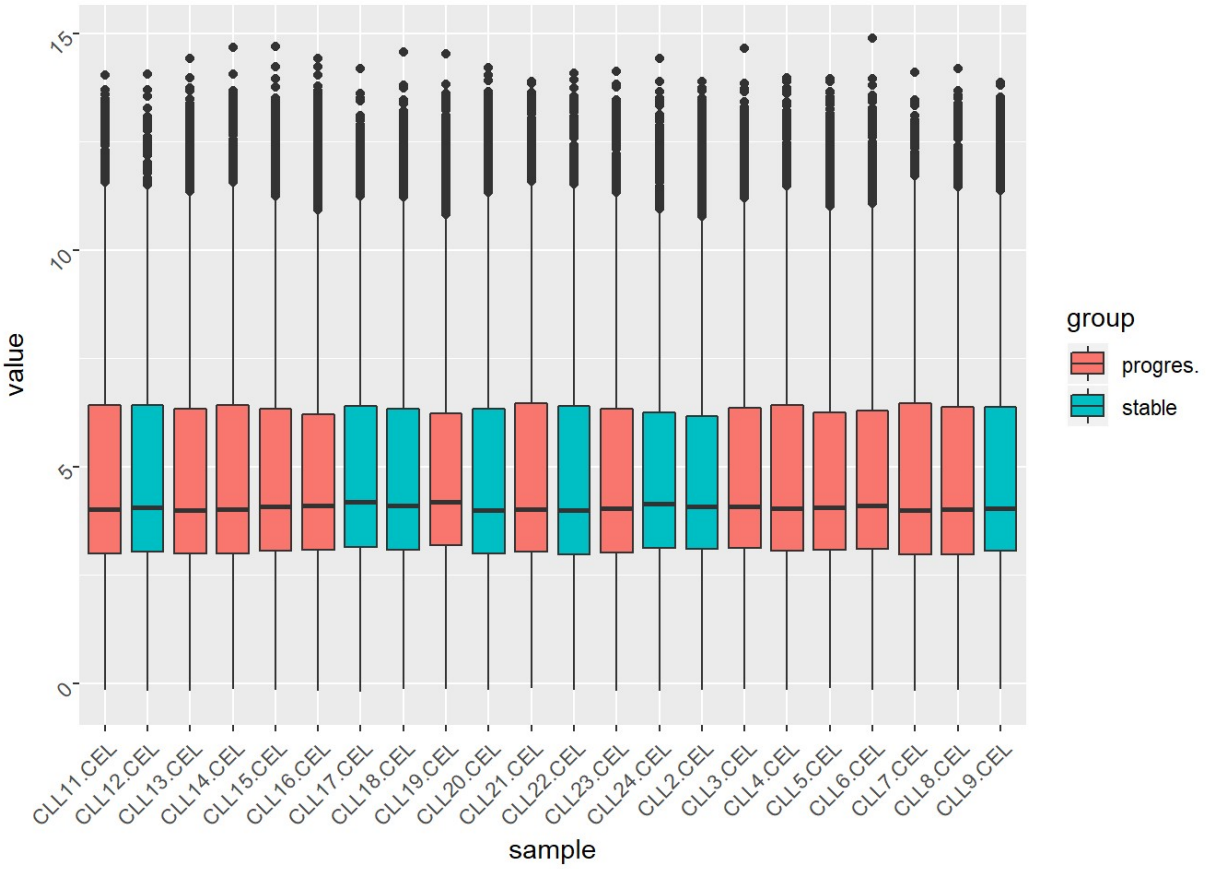
```
colnames(dat_melt) <- c("probe","symbol","sample","value")

##获得分组信息
group_list <- as.character(pdata[,2])
group_list
```

```
##  [1] "progres." "stable"   "progres." "progres." "progres." "progres."
##  [7] "stable"   "stable"   "progres." "stable"   "progres." "stable"
## [13] "progres." "stable"   "stable"   "progres." "progres." "progres."
## [19] "progres." "progres." "progres." "stable"
```

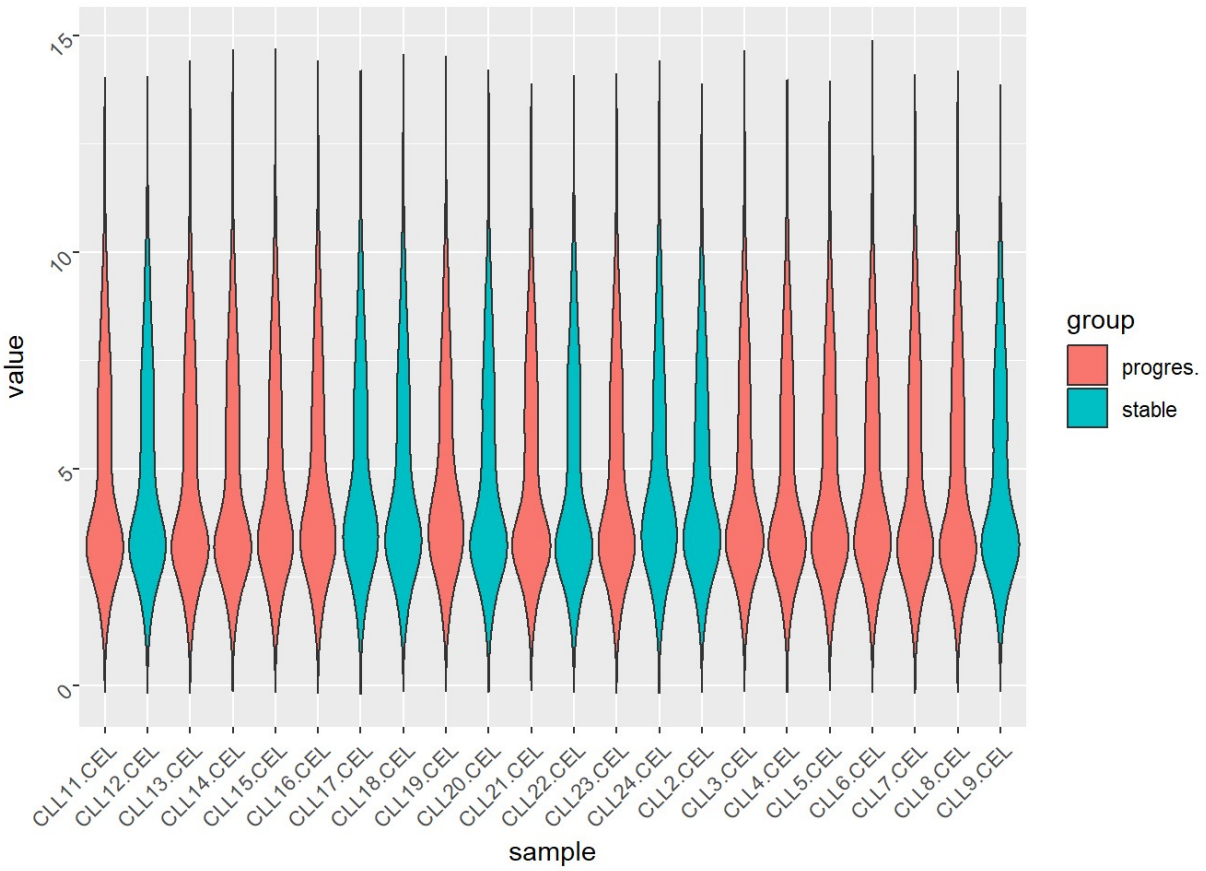
```
dim(dat_1)
```

```
## [1] 8585   24
```

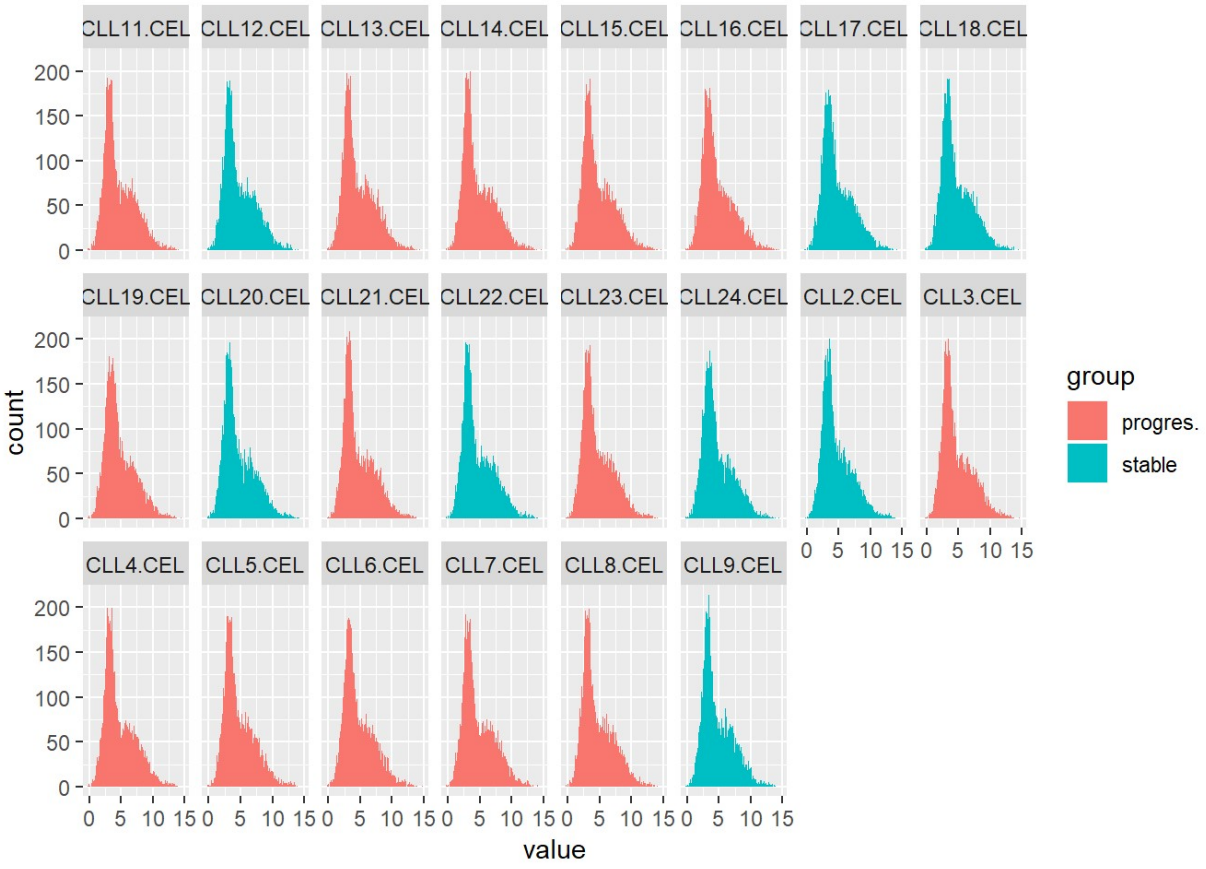
```
dat_melt$group <- rep(group_list, each=nrow(dat_1))

p <- ggplot(dat_melt, aes(x = sample, y = value, fill = group))+
  geom_boxplot()+
  theme(axis.text = element_text(angle = 45, hjust = 1))

print(p)
```



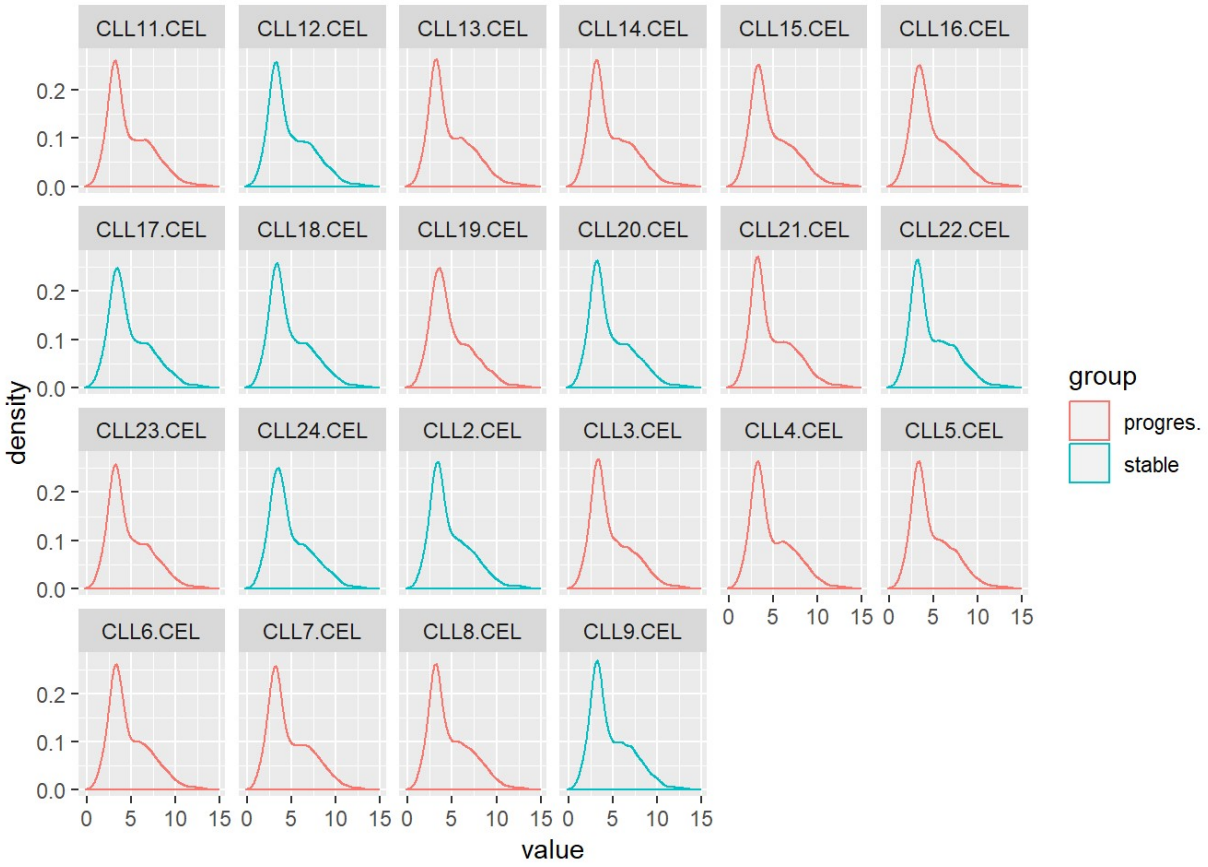

```
p <- ggplot(dat_melt, aes(x = sample, y = value, fill = group))+
  geom_violin()+
  theme(axis.text = element_text(angle = 45, hjust = 1))
print(p)
```



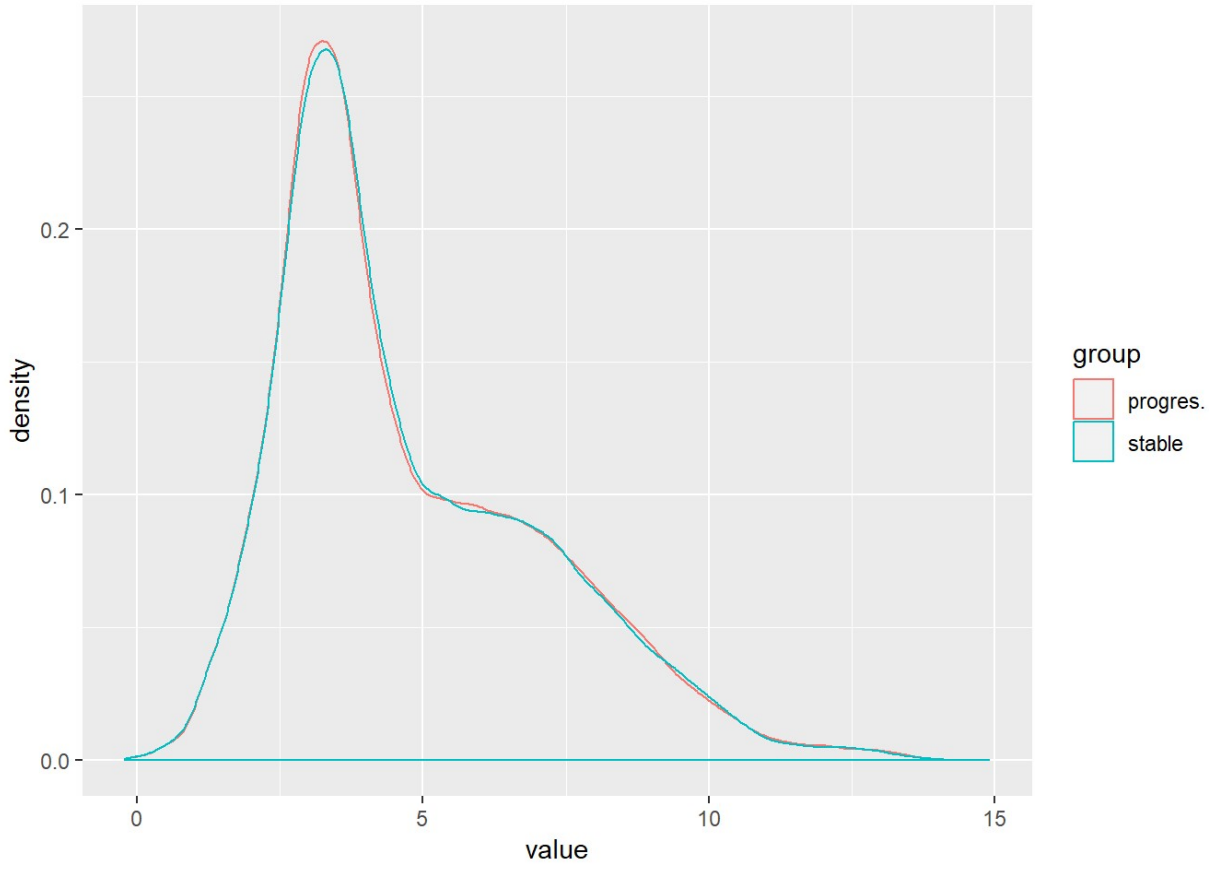

```
p <- ggplot(dat_melt, aes(value, fill = group))+
  geom_histogram(bins = 200)+
  facet_wrap(~sample, nrow = 3)
print(p)
```



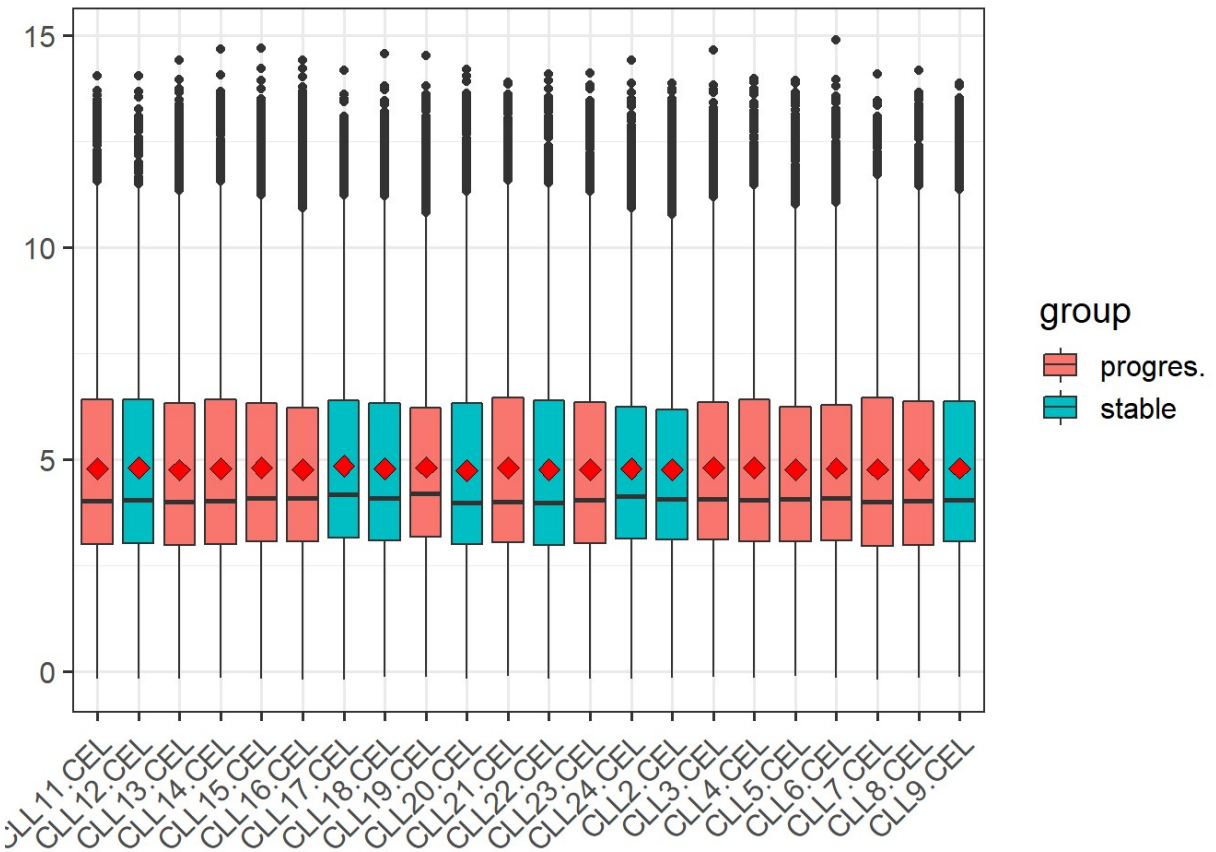

```
p <- ggplot(dat_melt, aes(value, col = group))+
  geom_density()+
  facet_wrap(~sample, nrow = 4)
print(p)
```




```
p <- ggplot(dat_melt, aes(value, col = group))+
  geom_density()
print(p)
```




```
p <- ggplot(dat_melt, aes(x = sample, y = value, fill = group))+
  geom_boxplot()
p <- p + stat_summary(fun.y = "mean", geom = "point", shape = 23, size = 3, fill =
"red")
p <- p+theme_set(theme_set(theme_bw(base_size = 15)))
p <- p+theme(axis.text.x = element_text(angle = 45, hjust = 1), axis.title = elemen
t_blank())
print(p)
```



作业-12

计算出每个gene在所有样本中的mean，median，max，min，sd，var，mad值，最后按照mad进行排序，取top50 mad值的基因，得到列表

```
head(dat)
```

```
##        CLL11.CEL CLL12.CEL CLL13.CEL CLL14.CEL CLL15.CEL CLL16.CEL
## ZZZ3    6.645791  7.350613  6.333290  6.603640  6.711462  7.373601
## ZZEF1   5.289264  6.677600  4.447104  7.008260  6.046429  6.413833
## ZYX     3.949769  5.423343  3.540189  5.234420  3.603839  3.687205
## ZWINT   4.316881  2.705329  3.131087  2.821306  2.963397  2.876353
## ZW10    4.382004  4.355469  4.336743  4.304551  4.482850  4.474894
## ZSWIM8  4.091876  4.050844  4.113627  4.041756  4.101077  4.203981
##        CLL17.CEL CLL18.CEL CLL19.CEL CLL20.CEL CLL21.CEL CLL22.CEL
## ZZZ3    6.243337  6.730870  7.299798  7.203648  6.519334  6.395689
## ZZEF1   7.369615  6.033872  6.493153  6.631621  6.390880  7.174788
## ZYX     4.191365  3.779226  3.141664  3.648371  6.091596  3.882752
## ZWINT   2.905329  2.885641  3.002759  3.127091  4.853690  2.810114
## ZW10    4.660235  4.537073  4.869948  4.307655  4.433605  4.398306
## ZSWIM8  4.131596  4.128203  4.192091  4.089017  4.170105  4.132040
##        CLL23.CEL CLL24.CEL CLL2.CEL CLL3.CEL CLL4.CEL CLL5.CEL CLL6.CEL
## ZZZ3    6.651841  7.338645 5.897972 6.713280 6.529733 6.680138 6.056228
## ZZEF1   4.837948  5.793722 6.998910 6.347929 6.267050 4.822419 5.666789
## ZYX     3.953285  3.554797 6.733884 4.456778 3.652998 3.825987 4.375647
## ZWINT   2.730719  2.879867 2.922759 2.762910 2.926378 2.907199 2.747928
## ZW10    4.420724  4.610161 4.460409 4.446617 4.381458 4.778824 4.483773
## ZSWIM8  4.133988  4.129267 4.165340 4.211647 4.158063 4.246774 4.182852
##        CLL7.CEL CLL8.CEL CLL9.CEL
## ZZZ3   6.868983 6.564657 6.607440
## ZZEF1  6.607534 6.553768 6.482294
## ZYX    3.962673 3.618525 4.726375
## ZWINT  2.800924 2.896882 3.273290
## ZW10   4.296971 4.588888 4.378410
## ZSWIM8 4.068309 4.072161 4.154959
```

```
g_mean <- sort(apply(dat, 1, mean), decreasing = T)
g_median <- sort(apply(dat, 1, median), decreasing = T)
g_max <- sort(apply(dat, 1, max), decreasing = T)
g_min <- sort(apply(dat, 1, min), decreasing = T)
g_sd <- sort(apply(dat, 1, sd), decreasing = T)
g_var <- sort(apply(dat, 1, var), decreasing = T)
g_mad <- sort(apply(dat, 1, mad), decreasing = T)

top50_mad <- g_mad[1:50]
top50_mean <- g_mean[1:50]
top50_median <- g_median[1:50]
top50_max <- g_max[1:50]
top50_min <- g_min[1:50]
top50_sd <- g_sd[1:50]
top50_var <- g_var[1:50]

names(top50_mad)
```

```
##  [1] "FAM30A"   "IGF2BP3"  "DMD"      "TCF7"     "SLAMF1"   "FOS"
##  [7] "LGALS1"   "IGLC1"    "ZAP70"    "FCN1"     "LHFPL2"   "HBB"
## [13] "S100A8"   "GUSBP11"  "COBLL1"   "VIPR1"    "PCDH9"    "IGH"
## [19] "ZNF804A"  "TRIB2"    "OAS1"     "CCL3"     "GNLY"     "CYBB"
## [25] "VAMP5"    "RNASE6"   "RGS2"     "PLXNC1"   "CAPG"     "RBM38"
## [31] "VCAN"     "APBB2"    "ARF6"     "TGFBI"    "NR4A2"    "S100A9"
## [37] "ZNF266"   "TSPYL2"   "CLEC2B"   "FLNA"     "H1FX"     "DUSP5"
## [43] "DUSP6"    "ANXA4"    "LPL"      "THEMIS2"  "P2RY14"   "ARHGAP44"
## [49] "TNFSF9"   "PFN2"
```

作业-13

根据作业12中得到的基因列表来去表达矩阵的子集，并且绘制热图 五种热图包的绘制情况

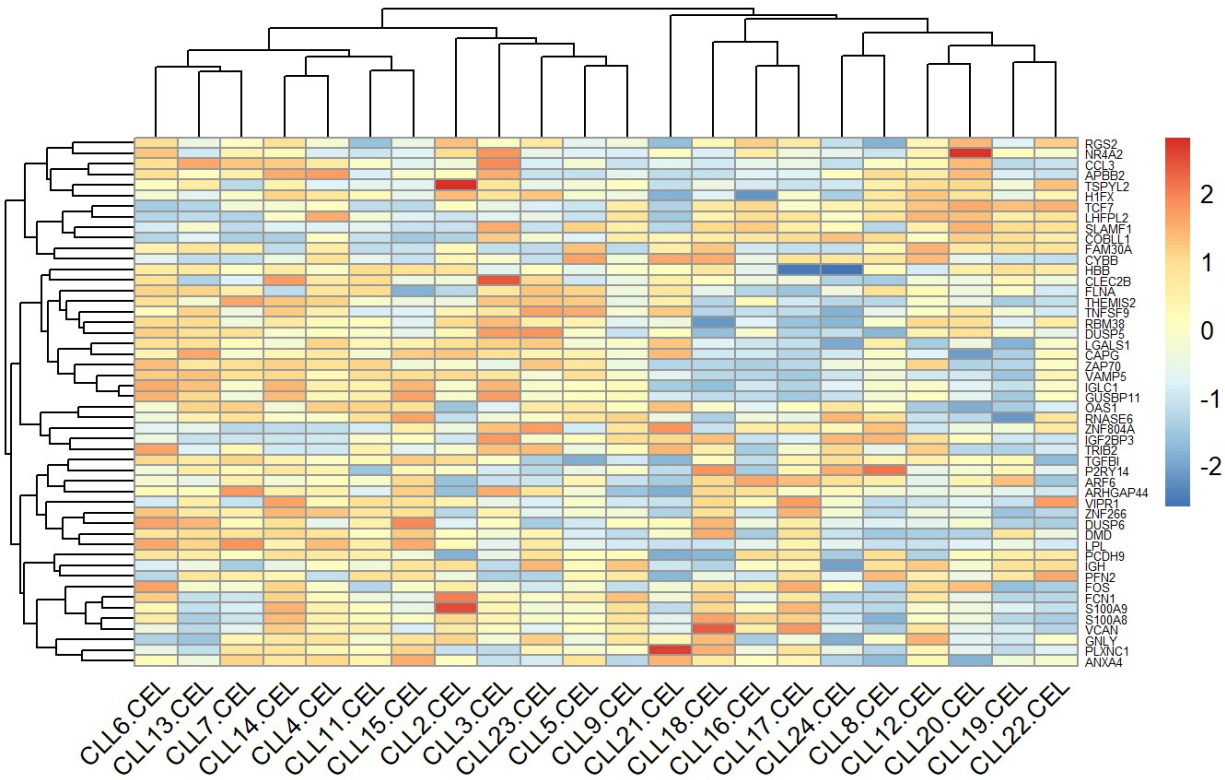
```
## 多种方法绘制热图：http://www.sohu.com/a/210713199_688647
top_50 <- dat[names(top50_mad),]
head(top_50)
```

```
##         CLL11.CEL CLL12.CEL CLL13.CEL CLL14.CEL CLL15.CEL CLL16.CEL
## FAM30A   2.470149  9.901761  7.939790  2.422295  2.683739  2.984042
## IGF2BP3  4.752641  7.007690  2.352665  2.465527  5.141380  2.335350
## DMD      6.966708  3.246325  9.157620  9.405685  9.808787  3.098357
## TCF7     5.186734 10.410150  4.939448  7.301564  4.973025  9.910380
## SLAMF1   4.088991  6.383560  5.082827  3.884222  3.302885  7.776638
## FOS      4.436873  9.559263  6.917740  9.599603  7.174250  8.503131
##         CLL17.CEL CLL18.CEL CLL19.CEL CLL20.CEL CLL21.CEL CLL22.CEL
## FAM30A   2.630670  9.210898  8.194040  7.740176  7.363434  8.072097
## IGF2BP3  2.946122  7.850253  2.639384  5.038179  7.214619  2.462464
## DMD      9.058481 10.653183  8.736892  3.456012  6.546933  5.690320
## TCF7     7.596952  9.123323 10.585619 10.976634  4.754913 10.900729
## SLAMF1   6.447888  7.410069  7.347221  8.282893  3.025406  7.088338
## FOS     10.571769  9.059870  3.581185 10.150864  6.693852  4.105843
##         CLL23.CEL CLL24.CEL CLL2.CEL CLL3.CEL CLL4.CEL CLL5.CEL  CLL6.CEL
## FAM30A   2.432329  2.520234 6.381942 2.533053 7.449871 9.460453  6.823559
## IGF2BP3  4.565494  8.040716 2.576302 8.899667 2.605947 5.257853  3.846314
## DMD      7.723079  4.115776 7.105942 6.367917 8.965992 4.613248  8.950720
## TCF7     5.872403  8.521424 7.875395 6.579674 7.798986 5.434597  4.547559
## SLAMF1   3.265837  5.702919 3.345209 8.385662 3.914609 7.547004  3.682875
## FOS      5.096457  7.384896 8.578894 6.350829 7.101665 6.894014 10.831309
##         CLL7.CEL CLL8.CEL CLL9.CEL
## FAM30A  7.408566 6.475890 2.477134
## IGF2BP3 2.489988 8.037834 7.108131
## DMD     8.896737 3.060286 3.782691
## TCF7    7.099306 9.214168 9.042241
## SLAMF1  3.134628 3.016306 6.370420
## FOS     6.726206 4.402530 4.782527
```

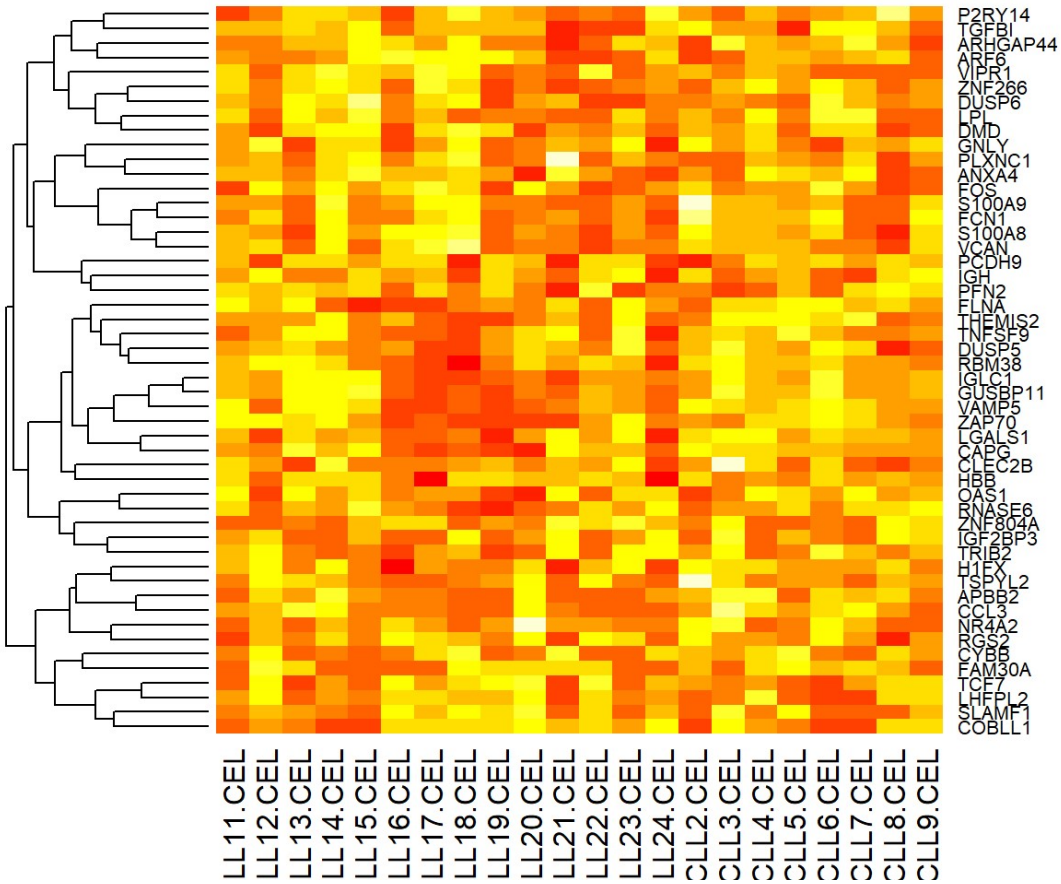
```
## pheatmap
library(pheatmap)
```

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

```
top_50_mat <- t(scale(t(top_50)))
pheatmap(top_50_mat,
         angle_col = 45,
         fontsize_row = 5,
         cellheight = 4.3)
```




```
## heatmap
heatmap(top_50_mat, cexRow = .8, cexCol = 1.2, Colv = NA)
```




```
## ggplot2
hc <- hclust(dist(top_50_mat))
row_order <- hc$order
top_50_mat_1 <- top_50_mat[row_order,]
top_50_mat_1 <- melt(top_50_mat_1)
colnames(top_50_mat_1)
```

```
## [1] "Var1"  "Var2"  "value"
```

```
p<-ggplot(top_50_mat_1,aes(x=Var2,y=Var1,fill=value))+
  xlab("")+
  ylab("")+
  labs(title="")+
  geom_tile(colour="white",size=0)+
  scale_fill_gradient(low="green",high="red")+
  geom_text(aes(label=round(value,2)),angel = 45,size = 3)
```

```
## Warning: Ignoring unknown parameters: angel
```

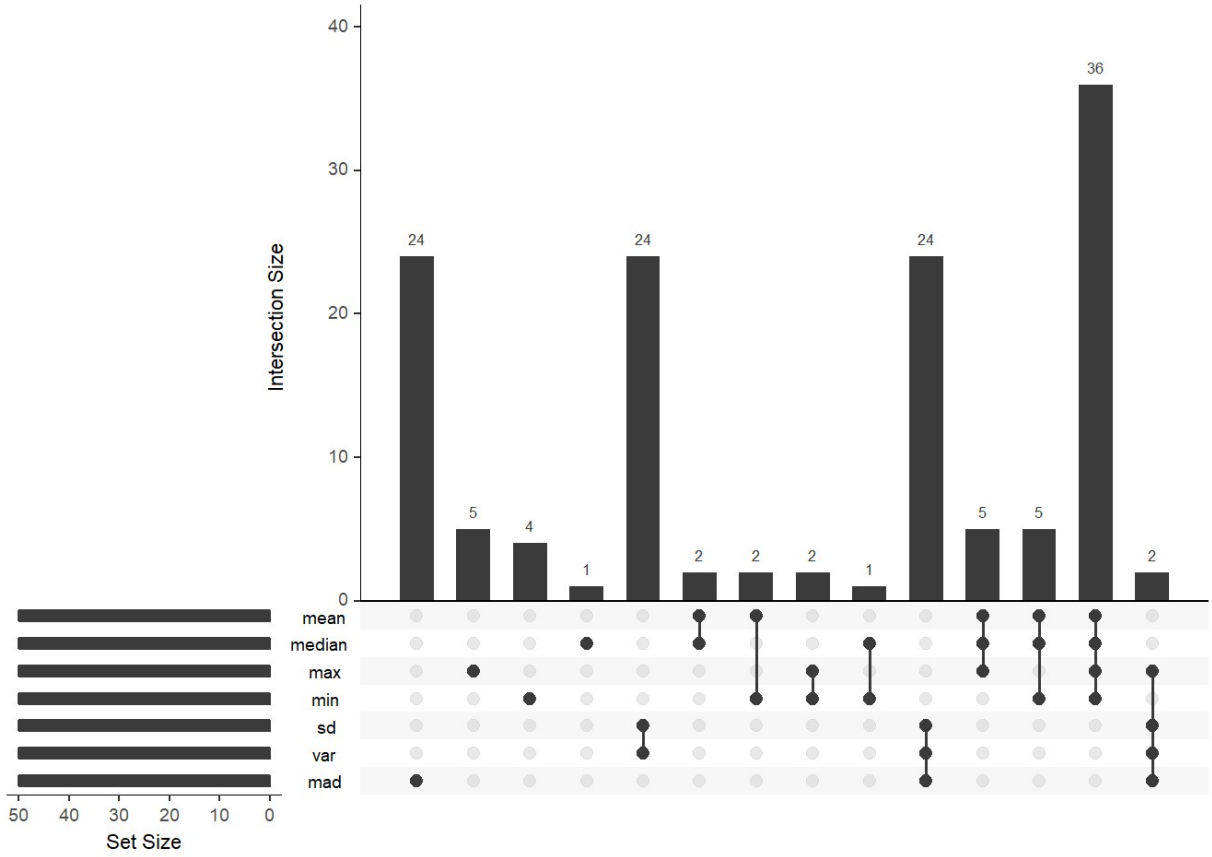
作业-14

取不同统计指标mean，median，max，min，sd，var，mad的各top50基因列表，使用UpSet包来查看它们之间的overlap情况

```
library(UpSetR)
```

```
## Warning: package 'UpSetR' was built under R version 3.5.3
```

```
g_all <- unique(c(names(top50_mean), names(top50_median), names(top50_max), names(t
op50_min), names(top50_sd), names(top50_var), names(top50_mad)))

g_dat <- data.frame(g_all = g_all,
                    mean = ifelse(g_all %in% names(top50_mean), 1, 0),
                    median = ifelse(g_all %in% names(top50_median), 1, 0),
                    max = ifelse(g_all %in% names(top50_max), 1, 0),
                    min = ifelse(g_all %in% names(top50_min), 1, 0),
                    sd = ifelse(g_all %in% names(top50_sd), 1, 0),
                    var = ifelse(g_all %in% names(top50_var), 1, 0),
                    mad = ifelse(g_all %in% names(top50_mad), 1, 0))

upset(g_dat, nsets = 7)
```



作业-15

在第二步的基础上面提取CLL包里面的data(sCLLex)数据对象的表性数据

```
pdata <- pData(sCLLex)
pdata
```

```
##           SampleID  Disease
## CLL11.CEL    CLL11 progres.
## CLL12.CEL    CLL12   stable
## CLL13.CEL    CLL13 progres.
## CLL14.CEL    CLL14 progres.
## CLL15.CEL    CLL15 progres.
## CLL16.CEL    CLL16 progres.
## CLL17.CEL    CLL17   stable
## CLL18.CEL    CLL18   stable
## CLL19.CEL    CLL19 progres.
## CLL20.CEL    CLL20   stable
## CLL21.CEL    CLL21 progres.
## CLL22.CEL    CLL22   stable
## CLL23.CEL    CLL23 progres.
## CLL24.CEL    CLL24   stable
## CLL2.CEL      CLL2   stable
## CLL3.CEL      CLL3 progres.
## CLL4.CEL      CLL4 progres.
## CLL5.CEL      CLL5 progres.
## CLL6.CEL      CLL6 progres.
## CLL7.CEL      CLL7 progres.
## CLL8.CEL      CLL8 progres.
## CLL9.CEL      CLL9   stable
```

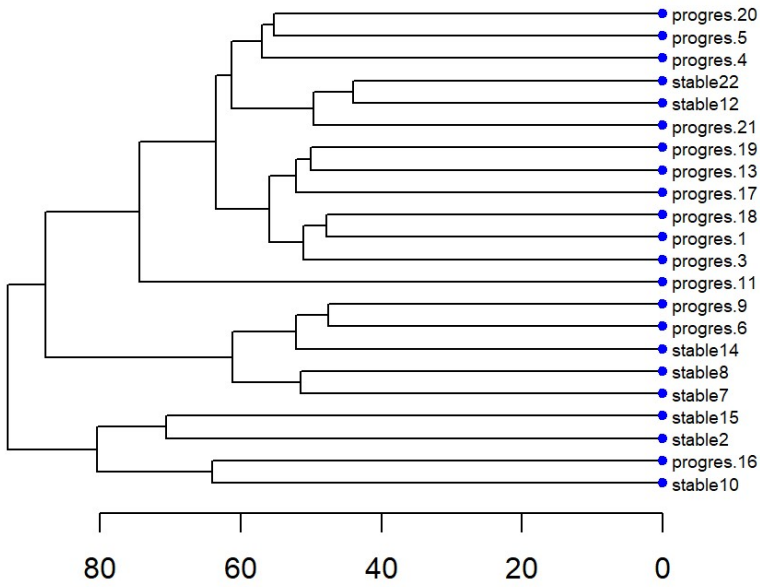
```
group_list <- as.character(pdata[,2])
dim(expSet)
```

```
## [1] 11460    22
```

作业-16

对所有样本的表达矩阵进行聚类并且绘图，然后添加样品的临床表型数据信息

```
colnames(expSet) <- paste(group_list, 1:22,sep = "")
nodePar <- list(lab.cex = .6, pch = c(NA, 19), cex = .7, col = "blue")
hc <- hclust(dist(t(expSet)))
par(mar=c(5,5,5,10))
plot(as.dendrogram(hc), nodePar = nodePar, horiz = T)
```



作业-17

对所有样本进行PCA分析并且绘图，同样添加表型信息

```
df <- as.data.frame(t(expSet))
library(FactoMineR)
```

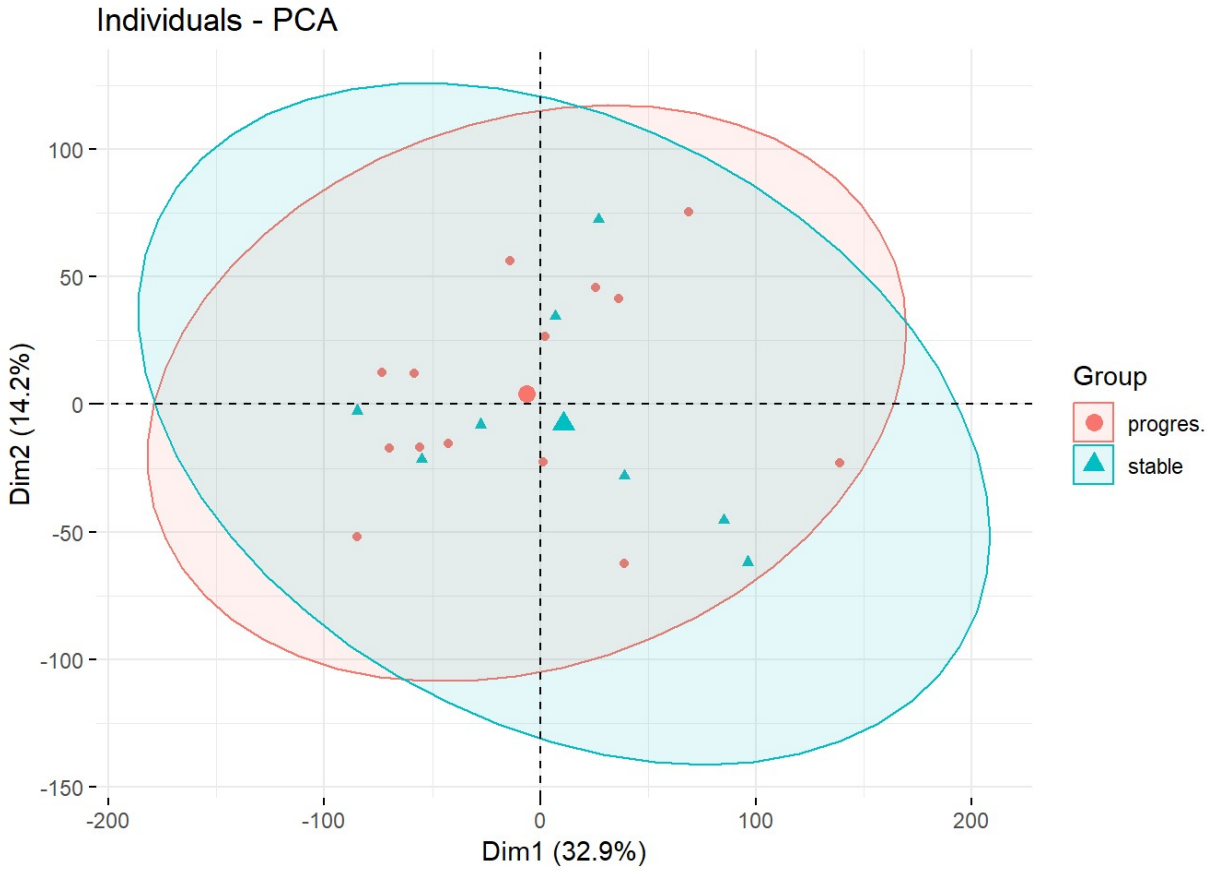
```
## Warning: package 'FactoMineR' was built under R version 3.5.3
```

```
library(factoextra)
```

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

```
## Welcome! Related Books: `Practical Guide To Cluster Analysis in R` at https://go
o.gl/13EFCZ
```

```
pca <- PCA(df, graph = F)
fviz_pca_ind(pca,
             geom.ind = "point",
             col.ind = group_list,
             addEllipses = T,
             legend.title = "Group")
```



作业-18

根据表达矩阵及样本分组信息进行批量T检验，得到检验结果表格

```
dat_2 <- dat
group_list <- as.factor(group_list)
group1 <- which(group_list == levels(group_list)[1])
group2 <- which(group_list == levels(group_list)[2])
dat1 <- dat_2[,group1]
dat2 <- dat_2[,group2]

dat_3 <- cbind(dat1, dat2)
dim(dat)
```

```
## [1] 8585   22
```

```
pvals <- apply(dat, 1, function(x){
  t.test(as.numeric(x)~group_list)$p.value
})
p.adj <- p.adjust(pvals, method = "BH")
avg_1 <- rowMeans(dat1)
avg_2 <- rowMeans(dat2)
log2FC <- avg_2 - avg_1
DEG_t.test <- cbind(avg_1, avg_2,log2FC,pvals, p.adj)
DEG_t.test <- DEG_t.test[order(DEG_t.test[,4]),]
DEG_t.test <- as.data.frame(DEG_t.test)

head(DEG_t.test)
```

```
##           avg_1    avg_2     log2FC        pvals     p.adj
## SGSM2  7.875615 8.791753  0.9161377 1.629755e-05 0.1399145
## PDE8A  6.622749 7.965007  1.3422581 4.058944e-05 0.1656600
## DLEU1  7.616197 5.786041 -1.8301554 6.965416e-05 0.1656600
## LDOC1  4.456446 2.152471 -2.3039752 8.993339e-05 0.1656600
## USP6NL 5.988866 7.058738  1.0698718 9.648226e-05 0.1656600
## COMMD4 4.157971 3.407405 -0.7505660 2.454557e-04 0.2586989
```

作业-19

使用limma包对表达矩阵及样本分组信息进行差异分析，得到差异分析表格，重点看logFC和P值，画火山图

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

```
##           progres. stable
## CLL11.CEL        1      0
## CLL12.CEL        0      1
## CLL13.CEL        1      0
## CLL14.CEL        1      0
## CLL15.CEL        1      0
## CLL16.CEL        1      0
## CLL17.CEL        0      1
## CLL18.CEL        0      1
## CLL19.CEL        1      0
## CLL20.CEL        0      1
## CLL21.CEL        1      0
## CLL22.CEL        0      1
## CLL23.CEL        1      0
## CLL24.CEL        0      1
## CLL2.CEL         0      1
## CLL3.CEL         1      0
## CLL4.CEL         1      0
## CLL5.CEL         1      0
## CLL6.CEL         1      0
## CLL7.CEL         1      0
## CLL8.CEL         1      0
## CLL9.CEL         0      1
## attr(,"assign")
## [1] 1 1
## attr(,"contrasts")
## attr(,"contrasts")$`factor(group_list)`
## [1] "contr.treatment"
```

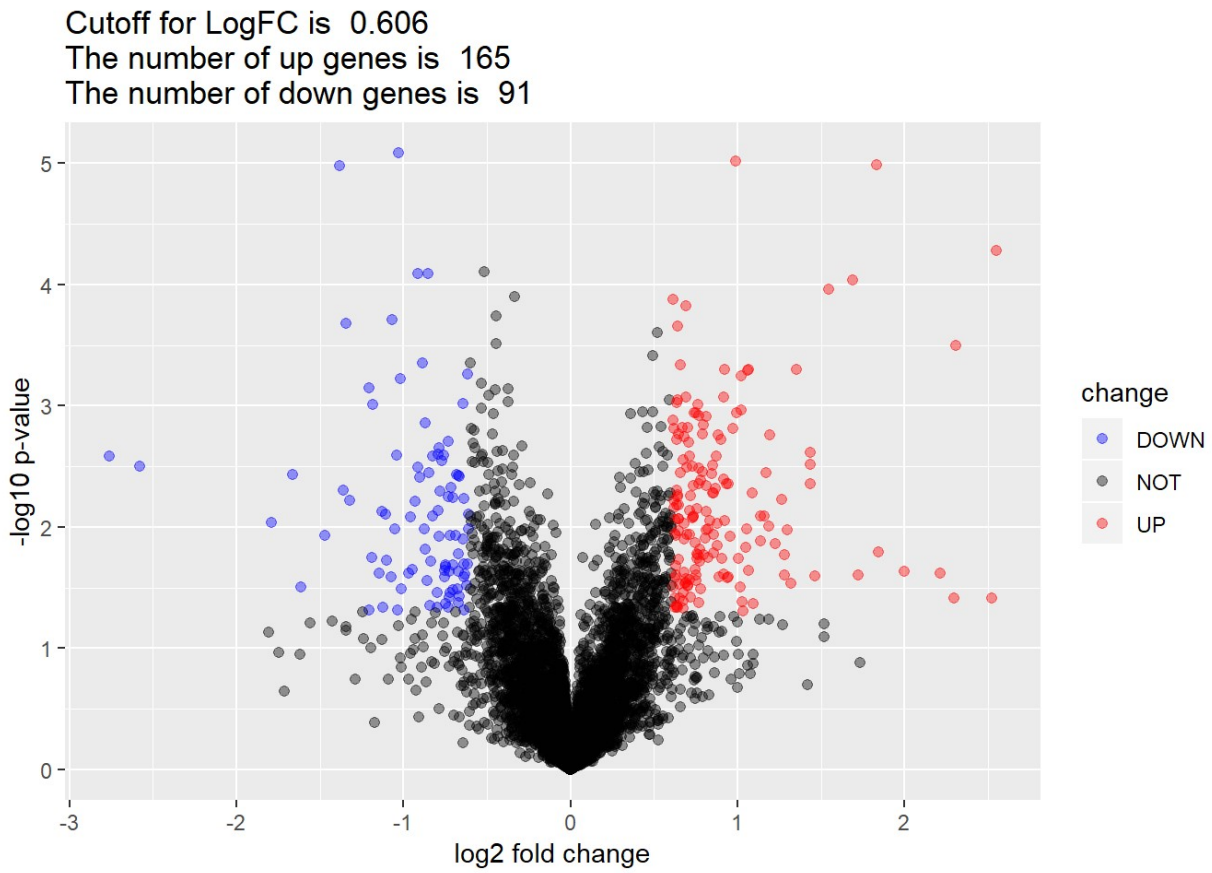
```
contrast.matrix <- makeContrasts(paste0(unique(group_list), collapse = "-"), levels
= design)
contrast.matrix
```

```
##           Contrasts
## Levels     progres.-stable
##   progres.               1
##   stable                -1
```

```
fit <- lmFit(dat, 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
## TBC1D2B -1.0284628 5.620700 -5.837398 8.240961e-06 0.02236713 3.351813
## CLIC1    0.9888221 9.954273  5.772843 9.560006e-06 0.02236713 3.230775
## DLEU1    1.8301554 6.950685  5.740883 1.029092e-05 0.02236713 3.170615
## SH3BP2  -1.3835699 4.463438 -5.735418 1.042149e-05 0.02236713 3.160313
## GPM6A    2.5471980 6.915045  5.043180 5.268833e-05 0.08731397 1.821657
## YTHDC2  -0.5187135 7.602354 -4.873724 7.881207e-05 0.08731397 1.485027
```

```
## volcano plot
DEG <- nrDEG
logFC_cutoff <- with(DEG, mean(abs(logFC)) + 2*sd(abs(logFC)))

DEG$change <- as.factor(ifelse(DEG$P.Value < 0.05 & abs(DEG$logFC) > logFC_cutoff,
ifelse(DEG$logFC > logFC_cutoff, "UP", "DOWN"), "NOT"))

title <- paste0('Cutoff for LogFC is  ', round(logFC_cutoff,3),'\nThe number of up
genes is  ', nrow(DEG[DEG$change == "UP",]), '\nThe number of down genes is  ', nro
w(DEG[DEG$change == "DOWN",]))

g <- ggplot(data = DEG, aes(x = logFC, y = -log10(P.Value), color = change))+
  geom_point(alpha = .4, size = 1.75)+
  xlab("log2 fold change")+
  ylab("-log10 p-value")+
  ggtitle(title)+
  #theme(element_text(size = 15, hjust = .5))+
  scale_colour_manual(values = c('blue', 'black', 'red'))

print(g)
```



作业-20

对T检验结果的P值和limma包差异分析的P值画散点图，看看哪些基因相差很大

```
head(nrDEG)
```

```
##              logFC  AveExpr         t      P.Value  adj.P.Val        B
## TBC1D2B -1.0284628 5.620700 -5.837398 8.240961e-06 0.02236713 3.351813
## CLIC1    0.9888221 9.954273  5.772843 9.560006e-06 0.02236713 3.230775
## DLEU1    1.8301554 6.950685  5.740883 1.029092e-05 0.02236713 3.170615
## SH3BP2  -1.3835699 4.463438 -5.735418 1.042149e-05 0.02236713 3.160313
## GPM6A    2.5471980 6.915045  5.043180 5.268833e-05 0.08731397 1.821657
## YTHDC2  -0.5187135 7.602354 -4.873724 7.881207e-05 0.08731397 1.485027
```

```
head(DEG_t.test)
```

```
##           avg_1    avg_2     log2FC        pvals     p.adj
## SGSM2  7.875615 8.791753  0.9161377 1.629755e-05 0.1399145
## PDE8A  6.622749 7.965007  1.3422581 4.058944e-05 0.1656600
## DLEU1  7.616197 5.786041 -1.8301554 6.965416e-05 0.1656600
## LDOC1  4.456446 2.152471 -2.3039752 8.993339e-05 0.1656600
## USP6NL 5.988866 7.058738  1.0698718 9.648226e-05 0.1656600
## COMMD4 4.157971 3.407405 -0.7505660 2.454557e-04 0.2586989
```

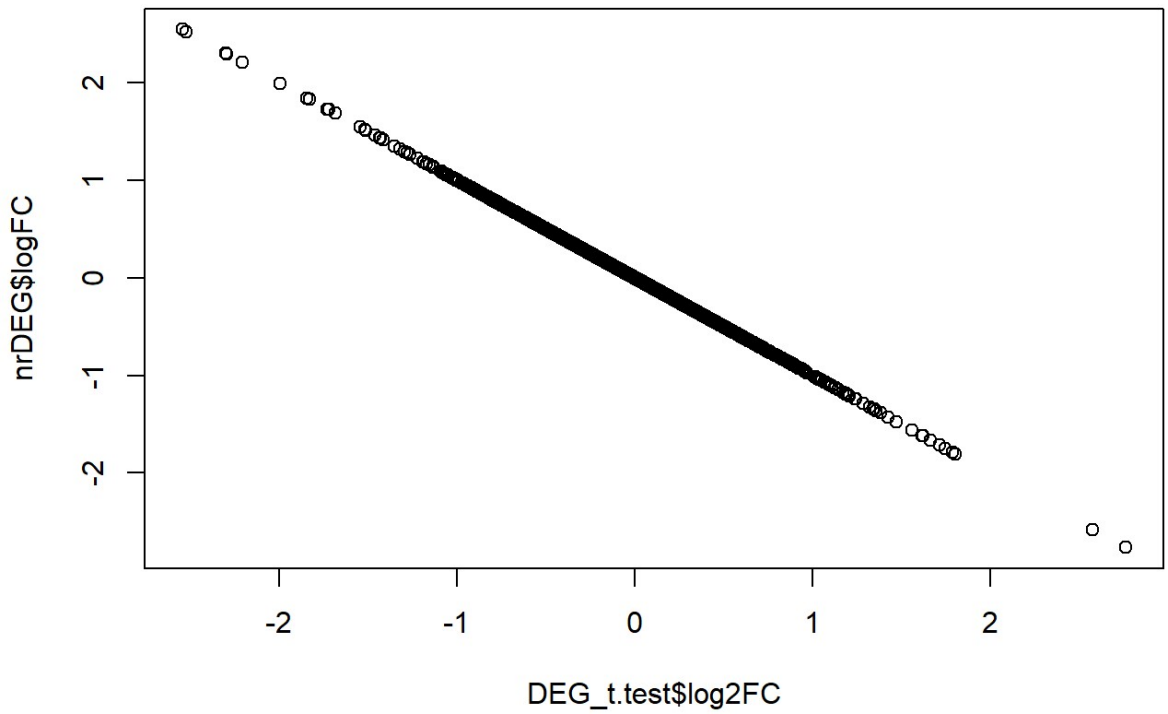
```
dim(nrDEG)
```

```
## [1] 8585    6
```

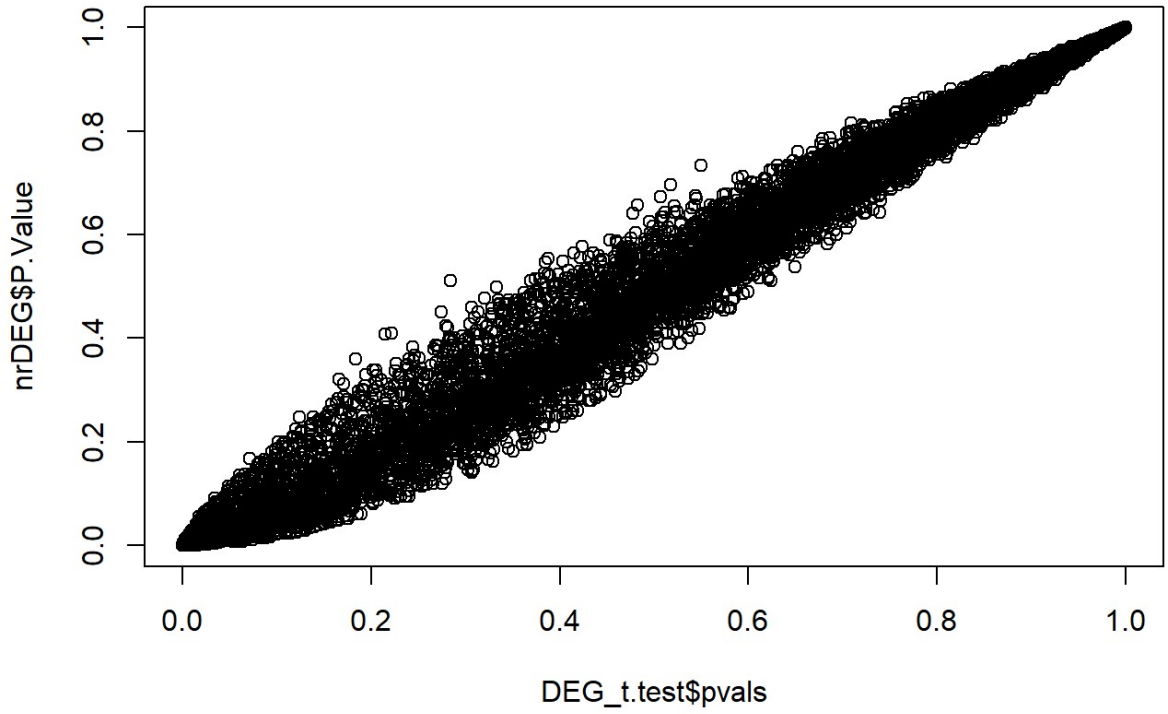
```
dim(DEG_t.test)
```

```
## [1] 8585    5
```

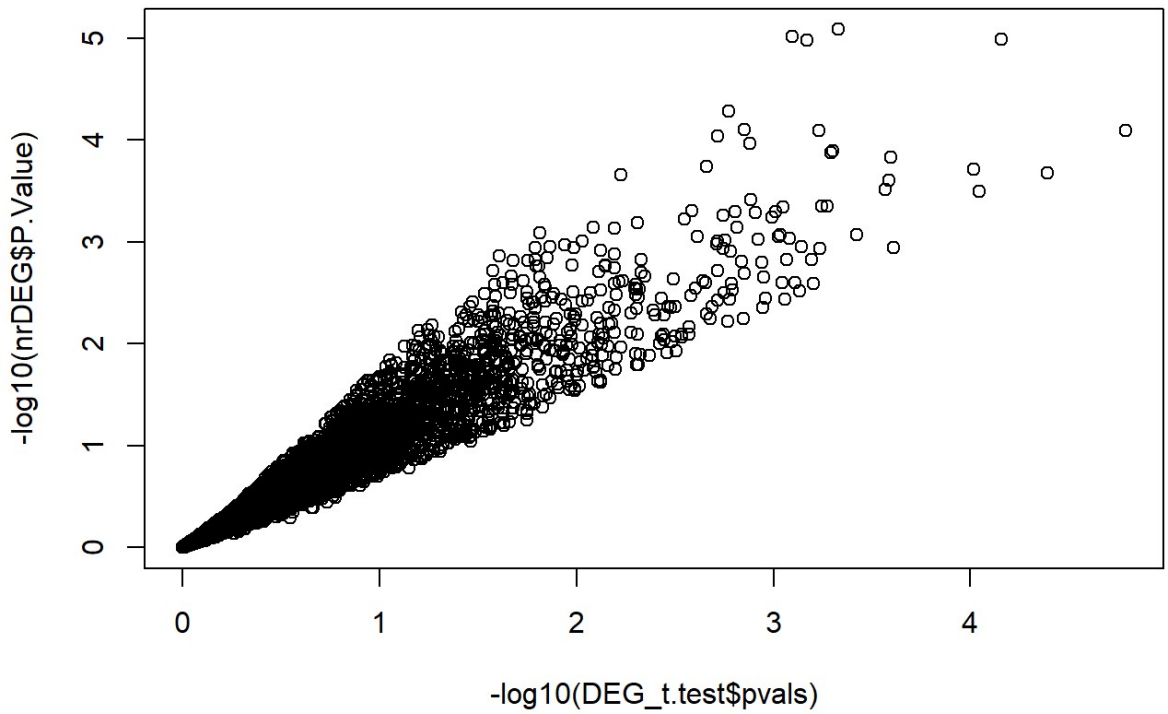
```
## 排好位置
DEG_t.test <- DEG_t.test[rownames(nrDEG),]

plot(DEG_t.test$log2FC, nrDEG$logFC)
```




```
plot(DEG_t.test$pvals, nrDEG$P.Value)
```




```
plot(-log10(DEG_t.test$pvals), -log10(nrDEG$P.Value))
```



```
library(ggpubr)
```

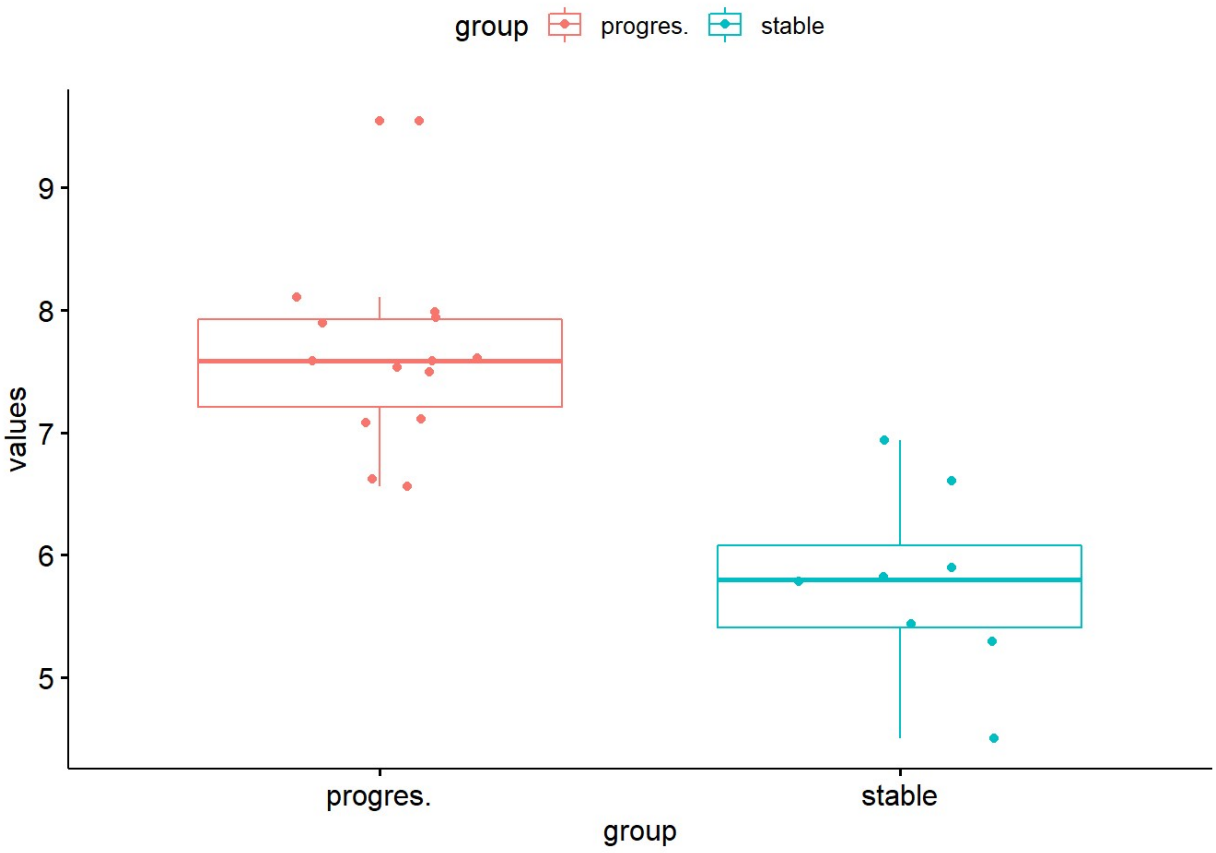
```
## Warning: package 'ggpubr' was built under R version 3.5.3
```

```
## Loading required package: magrittr
```

```
## Warning: package 'magrittr' was built under R version 3.5.3
```

```
comp <- list(c('stable', 'progres'))
dat_4 <- data.frame(group = group_list,
                    sampleID = names(dat['DLEU1',]),
                    values = as.numeric(dat['DLEU1',]))

ggboxplot(dat_4, x='group', y = 'values',
          color = "group", add = "jitter")+
  stat_compare_means(comparisons = comp, method = "t.test")
```

```
## Warning: Computation failed in `stat_signif()`:
## not enough 'y' observations
```



```
choose_gene <- head(rownames(nrDEG),50)
choose_value <- dat[choose_gene,]
pheatmap(choose_value,scale = "row", angle_col = 45, fontsize = 4.5, cellheight =
4)
```