edgeR_pipeline

AXIE

December 11, 2018

packages

```
library(Rsubread)
library(ggplot2)
library(org.Mm.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, cbind, colMeans,
##     colnames, colSums, 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
```

```
##
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:AnnotationDbi':
##
##     select
```

```
## The following objects are masked from 'package:IRanges':
##
##     collapse, desc, intersect, setdiff, slice, union
```

```
## The following objects are masked from 'package:S4Vectors':
##
##     first, intersect, rename, setdiff, setequal, union
```

```
## The following object is masked from 'package:Biobase':
##
##     combine
```

```
## The following objects are masked from 'package:BiocGenerics':
##
##     combine, intersect, setdiff, union
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(limma)
```

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

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

```
library(edgeR)

setwd("~/Desktop/set18_RNAseq/sample/featureCounts")
```

Count Reads using Feature Counts

```
##write the counts table into txt file
#fls = dir(".","bam")
#x = featureCounts(files = fls, annot.inbuilt = "mm10", GTF.featureType =
"gene",GTF.attrType = "gene_id")

#write.table(x = data.frame(x$annotation, x$counts,
stringsAsFactors=FALSE),file="test-counts.txt", quote=FALSE, sep="\t",
row.names = FALSE)

## read the counts table and change the colnames
set18 = read.table("test-counts.txt", header = TRUE, quote = '\t', skip =
1)

names(set18) =
c("GeneID","Chr","Start","End","Strand","Length","KO1","KO2","WT1","WT2")
head(set18)
```

```
##      GeneID                                              Chr
## 1 100503874                                       chr1;chr1
## 2 100038431                                            chr1
## 3     19888                  chr1;chr1;chr1;chr1;chr1;chr1
## 4     20671                       chr1;chr1;chr1;chr1;chr1
## 5     27395                       chr1;chr1;chr1;chr1;chr1
## 6     18777 chr1;chr1;chr1;chr1;chr1;chr1;chr1;chr1;chr1
##
Start
## 1
3647309;3658847
## 2
3680155
## 3
4290846;4343507;4351910;4352202;4360200;4409170
## 4
4490928;4493100;4493772;4495136;4496291
## 5
4773198;4777525;4782568;4783951;4785573
## 6
4807893;4808455;4828584;4830268;4832311;4837001;4839387;4840956;4844963
##
End
## 1
3650509;3658904
## 2
3681788
## 3
4293012;4350091;4352081;4352837;4360314;4409241
## 4
4492668;4493466;4493863;4495942;4496413
## 5
4776801;4777648;4782733;4784105;4785726
## 6
4807982;4808486;4828649;4830315;4832381;4837074;4839488;4841132;4846735
##              Strand Length KO1  KO2  WT1  WT2
## 1               -;-   3259  12    4   15   10
## 2                 +   1634   0    0    0    0
## 3       -;-;-;-;-;-   9747   0    3    0    0
## 4         -;-;-;-;-   3130   0    2    0    0
## 5         -;-;-;-;-   4203 969 1047 1055 1178
## 6 +;+;+;+;+;+;+;+;+   2433  54   53   42   52
```

```
#anyDuplicated(set18$GeneID)
```

Generate Count Matrix

```
countMatrix = set18[7:10]

rownames(countMatrix) = set18$GeneID
head(countMatrix)
```

```
##           KO1  KO2  WT1  WT2
## 100503874  12    4   15   10
## 100038431   0    0    0    0
## 19888       0    3    0    0
## 20671       0    2    0    0
## 27395     969 1047 1055 1178
## 18777      54   53   42   52
```

Generate DGEList Object

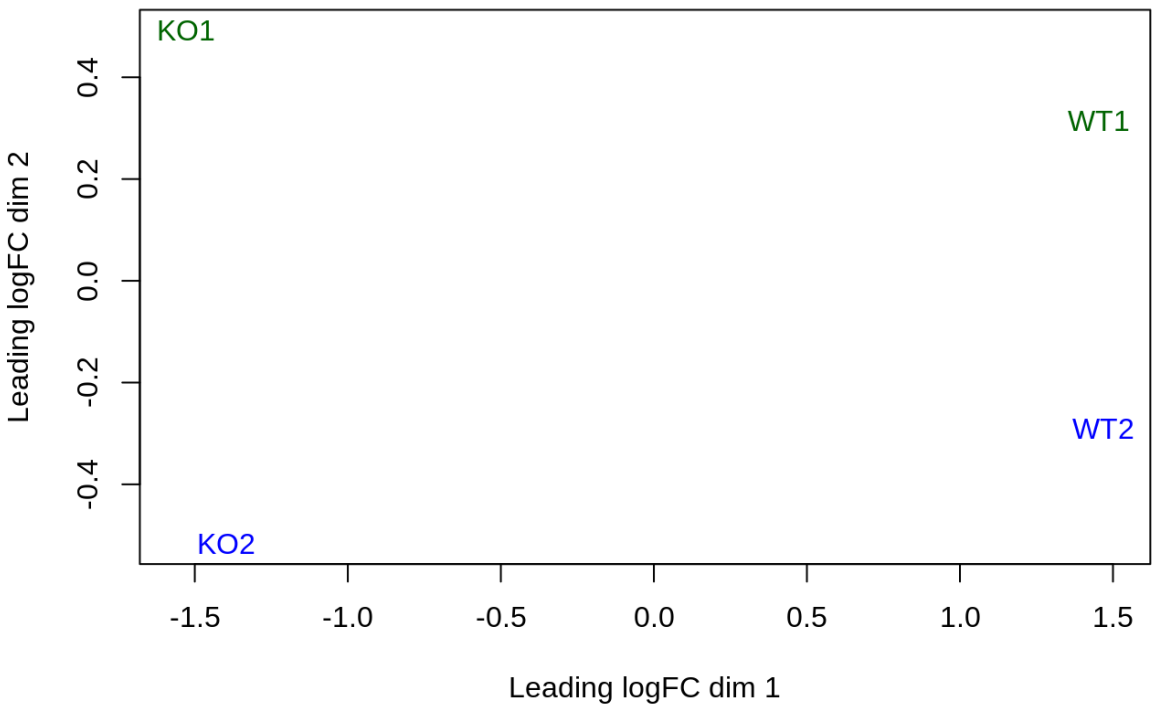
```
group = c("KO","KO","WT","WT")
y = DGEList(counts = countMatrix,  genes = set18[,1], group = group)
## filter weakly expressed features
keep <- rowSums(cpm(y)>1) >= 2
y <- y[keep, , keep.lib.sizes=FALSE]
```

Estimate Normalization Factors

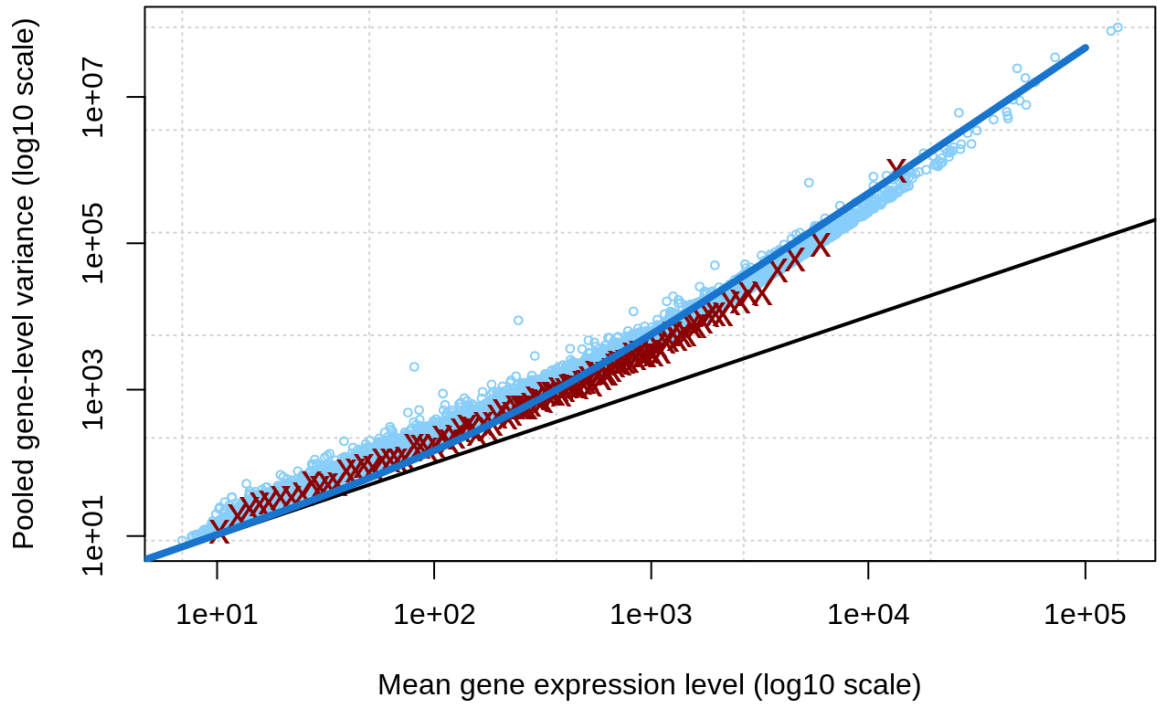
```
y = calcNormFactors(y)
```

Inspect the relationships between samples using multidimensional scaling plot

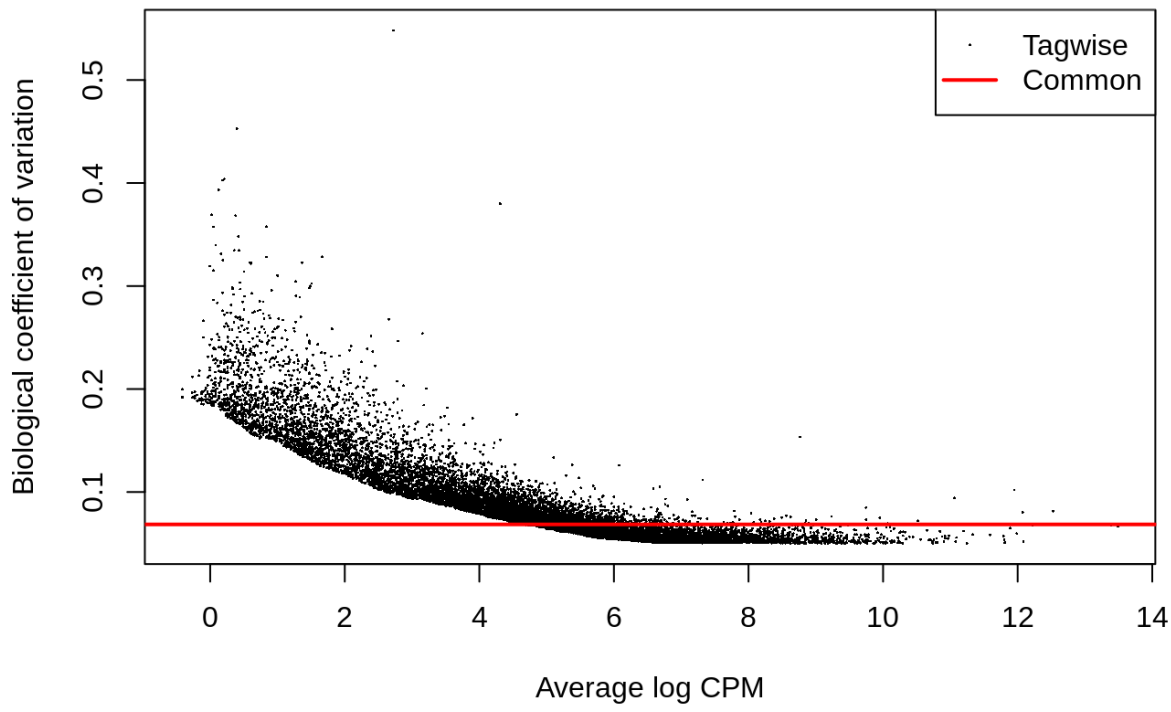
```
plotMDS(y, labels  = y$group, col = c("darkgreen","blue"))
```




```
d= estimateCommonDisp(y)
d = estimateTagwiseDisp(d)

plotMeanVar(d, show.tagwise.vars = TRUE, NBline = TRUE)
```




```
plotBCV(d)
```



Differential gene expression

```
de = exactTest(d, pair = c("KO","WT"))
tt= topTags(de)
head(tt$table)
```

```
##               genes     logFC   logCPM        PValue           FDR
## 270711       270711 -8.101527 6.686458  0.000000e+00  0.000000e+00
## 100628626 100628626 -4.718906 7.963937  0.000000e+00  0.000000e+00
## 19791         19791 -4.527570 9.747693  0.000000e+00  0.000000e+00
## 102436       102436 -3.631765 8.954038  0.000000e+00  0.000000e+00
## 20296         20296 -6.948003 5.848351 2.109844e-301 5.473779e-298
## 67900         67900  3.871182 6.737123 2.851520e-288 6.164986e-285
```

```
nrow(de)
```

```
## [1] 12972
```

```
## jump to the next de gene expression to plot and write the table
```

Differential expression analysis using design matrix (2)

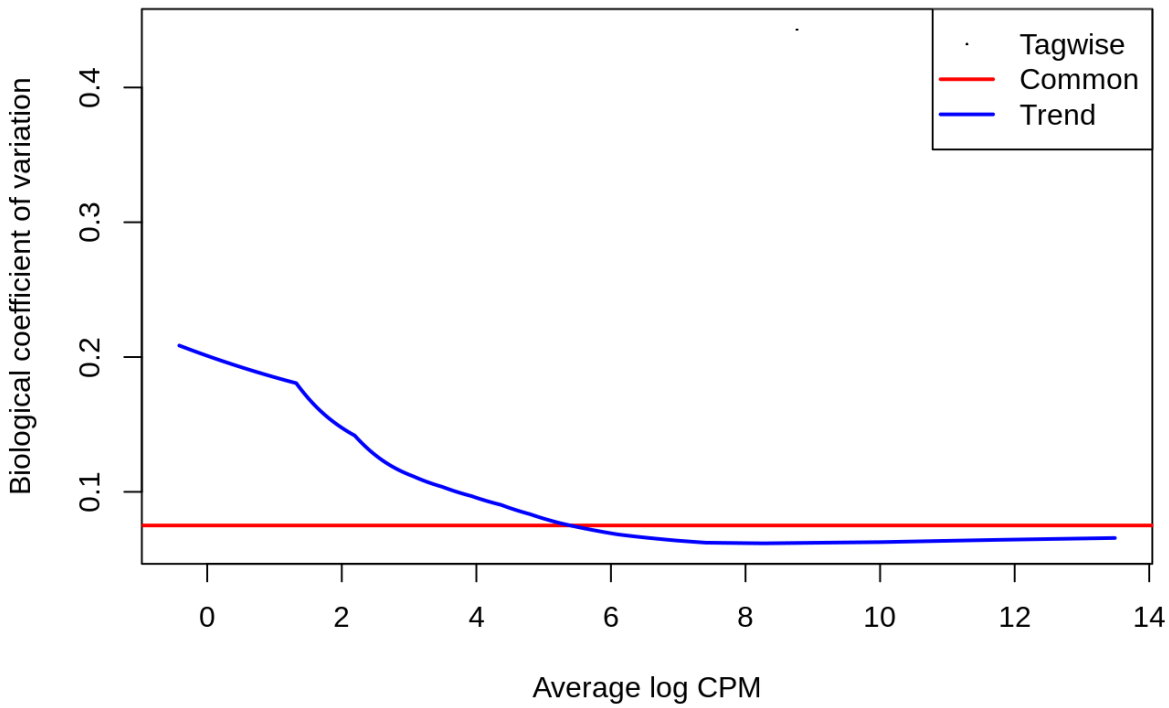
```
#design matrix
replicate = factor(c("1","2","1","2"))
group = factor(c("WT","WT","KO","KO"))

design = model.matrix(~ 0+ group + replicate)
rownames(design) = colnames(y)
design
```

```
##     groupKO groupWT replicate2
## KO1       0       1          0
## KO2       0       1          1
## WT1       1       0          0
## WT2       1       0          1
## attr(,"assign")
## [1] 1 1 2
## attr(,"contrasts")
## attr(,"contrasts")$group
## [1] "contr.treatment"
##
## attr(,"contrasts")$replicate
## [1] "contr.treatment"
```

```
d2 = estimateDisp(y, design, robust = TRUE)
#d2  = estimateGLMTagwiseDisp(d2,design)
d2$common.dispersion
```

```
## [1] 0.005636757
```

```
#plotMeanVar(d2, show.tagwise.vars = TRUE, NBline = TRUE)
plotBCV(d2)
```



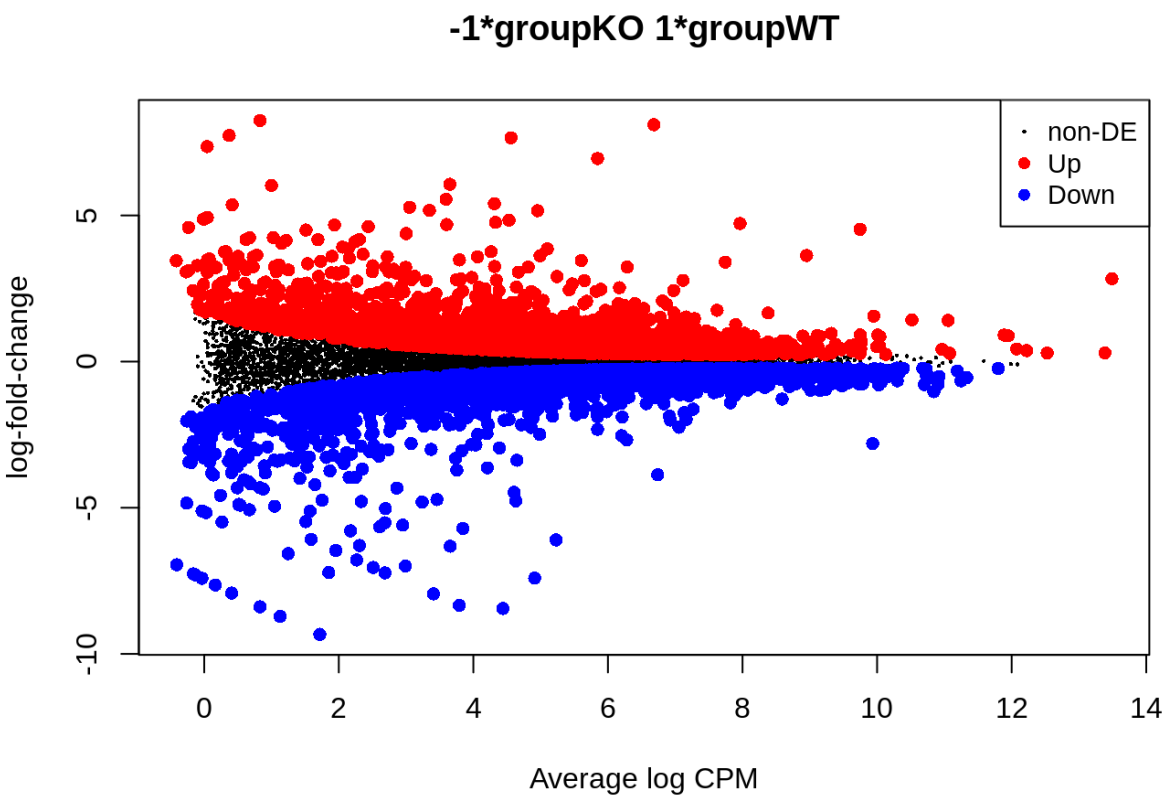

```
fit = glmQLFit(d2, design)
qlf = glmQLFTest(fit,contrast = c(-1,1,0))
tt1 = topTags(qlf)
#cpm(y)
#colnames(design)

head(tt1$table)
```

```
##                genes     logFC   logCPM         F        PValue
## 270711        270711  8.105483 6.684813 1704.6363  0.000000e+00
## 19791          19791  4.526399 9.747464 1592.4947  0.000000e+00
## 100628626 100628626  4.725147 7.963140 1532.5894 5.914519e-317
## 20296          20296  6.947695 5.846061 1147.2491 4.732713e-241
## 102436        102436  3.630457 8.953682 1127.0621 5.115450e-237
## 67900          67900 -3.870547 6.738606  951.9605 8.813391e-202
##                      FDR
## 270711     0.000000e+00
## 19791      0.000000e+00
## 100628626 2.557438e-313
## 20296     1.534819e-237
## 102436    1.327152e-233
## 67900     1.905455e-198
```

```
summary(decideTests(qlf))
```

```
##        -1*groupKO 1*groupWT
## Down                   2828
## NotSig                 7382
## Up                     2762
```

```
#qlf
plotMD(qlf)
```



Get the differential expressed genes and write to table

```
threshold = as.factor(ifelse(qlf$table$PValue<0.05 & abs(qlf$table$logFC)
> 1.5, ifelse(qlf$table$logFC > 1.5,"up","down"),"not"))

# select the logFC > 1.5 or <-1.5 and the pValue < 0.05 genes , generate
the table and plots
qlfTable = qlf$table[qlf$table$PValue<0.05 & (qlf$table$logFC> 1.5 |
qlf$table$logFC < -1.5),]
head(qlfTable)
```

```
##            logFC    logCPM         F       PValue
## 240690 -3.488098 2.0765053  76.43226 2.556889e-18
## 77673  -2.133951 0.5693629  12.93300 3.240385e-04
## 21419  -2.630148 1.6272212  37.45917 9.603730e-10
## 14859   4.102830 2.2415184  95.51185 1.757559e-22
## 280645 -1.710785 1.2892250  14.05445 1.783647e-04
## 214854 -1.792026 4.8200011 134.45366 6.176489e-31
```

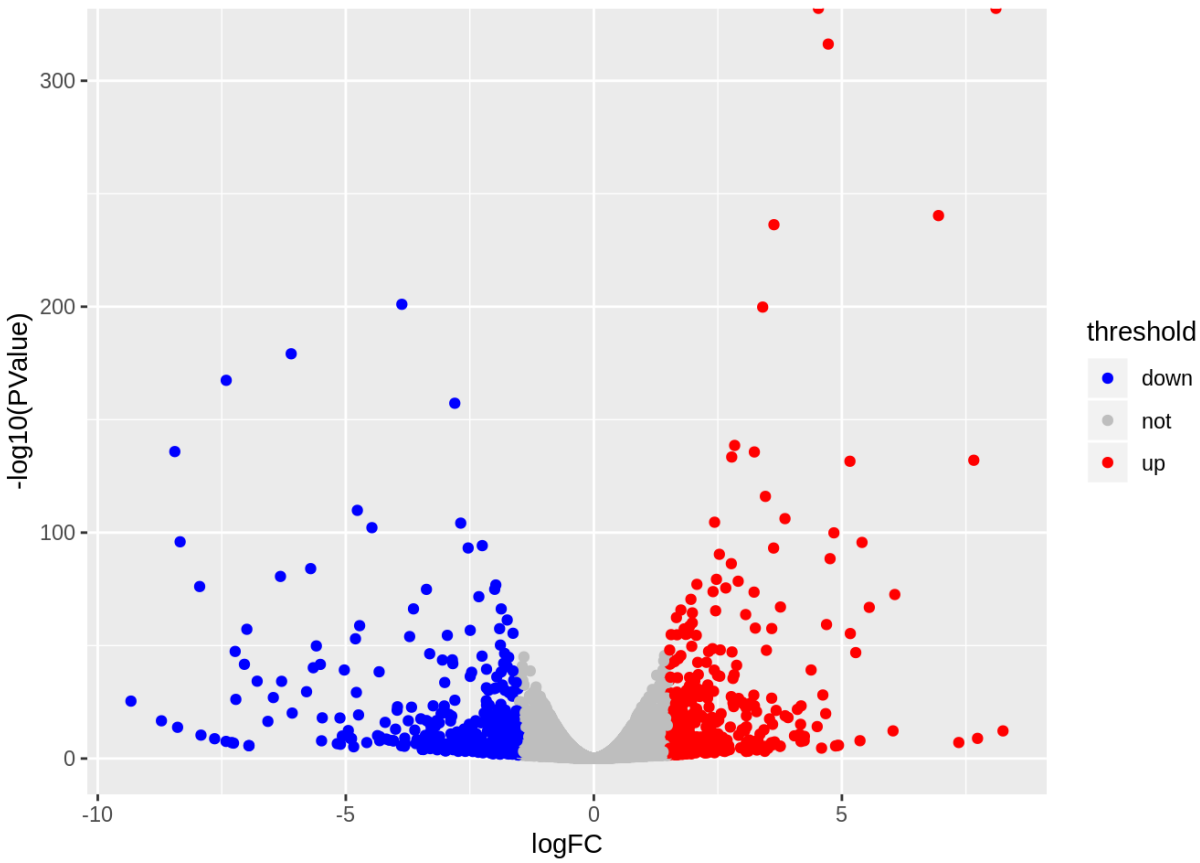
```
# add the CPM of each gene in the table
deGeneID = rownames(qlf$table)
deCPM = cpm(y)[deGeneID,]

deGeneTable= merge(qlfTable, deCPM,by = 0)




## rename the first col as "gene_id"
colnames(deGeneTable)[1] = c("gene_id")

## get the gene_id , gene_name table
egGENENAME = toTable(org.Mm.egSYMBOL)

## merge above table and the deGeneTable, so that the gene name is added
into the table.
DEnameAndId=  inner_join(egGENENAME,deGeneTable,by = "gene_id")
```

```
## Warning: Column `gene_id` has different attributes on LHS and RHS of
join
```

```
head(DEnameAndId)
```

```
##   gene_id symbol     logFC    logCPM         F       PValue        KO1
## 1   11304  Abca4  3.241519 0.7267794  24.44871 7.726927e-07   1.736019
## 2   11409  Acads -1.532244 5.5122855 137.35270 1.456236e-31  21.658899
## 3   11433   Acp5  2.462324 0.8326890  18.57305 1.647213e-05   2.976032
## 4   11475  Acta2  1.984385 6.1534583 273.41466 8.556080e-61 109.451840
## 5   11496 Adam22 -8.339986 3.7913194 442.67951 1.154453e-96   0.000000
## 6   11548 Adra1b  3.603237 1.8975355  65.24997 7.177579e-16   6.613404
##           KO2        WT1        WT2
## 1   3.3787073  0.3673853  0.1617809
## 2  25.0886989 77.0590582 59.5353874
## 3   2.3722839  0.4592316  0.4853428
## 4 115.9543591 30.0337450 27.0174176
## 5   0.1437748 26.7272776 29.3632413
## 6   6.1104281  0.8266168  0.2426714
```

```
#class(egGENENAME$gene_id)
#class(deGeneTable$gene_id)
write.csv(DEnameAndId, file = "differential_expressed_genes.csv")

## make vocano plot

ggplot(qlf$table, aes(x = logFC, y = -log10(PValue),colour = threshold)) +
geom_point() + scale_color_manual(values = c("blue","grey","red"))
```



GO analysis

```
go = goana(qlf, species = "Mm")
topGO(go,sort = "down")
```

```
##                                                    Term Ont     N   Up
## GO:0005840                                    ribosome  CC   213   18
## GO:0022626                            cytosolic ribosome  CC   103    5
## GO:0005737                                     cytoplasm  CC  7729 1642
## GO:0044445                                cytosolic part  CC   206   24
## GO:0022625            cytosolic large ribosomal subunit  CC    58    3
## GO:0044444                             cytoplasmic part  CC  5964 1277
## GO:0044391                            ribosomal subunit  CC   179   15
## GO:0098800 inner mitochondrial membrane protein complex  CC   114    9
## GO:0044429                            mitochondrial part  CC   697  113
## GO:0003735           structural constituent of ribosome  MF   148   14
## GO:0098798                mitochondrial protein complex  CC   137   11
## GO:0044455                  mitochondrial membrane part  CC   180   20
## GO:0097458                                neuron part  CC  1125  245
## GO:0015934                    large ribosomal subunit  CC   112    9
## GO:0005740                     mitochondrial envelope  CC   529   88
## GO:0005746             mitochondrial respiratory chain  CC    73    6
## GO:0031966                      mitochondrial membrane  CC   491   78
## GO:0005622                                intracellular  CC  9626 2069
## GO:0036477                  somatodendritic compartment  CC   630  133
## GO:0005623                                        cell  CC 10272 2277
##            Down         P.Up       P.Down
## GO:0005840   99 9.999999e-01 8.727071e-16
## GO:0022626   59 9.999995e-01 5.366176e-15
## GO:0005737 1864 6.540709e-01 1.454854e-14
## GO:0044445   94 9.999114e-01 1.934836e-14
## GO:0022625   40 9.998798e-01 1.995192e-14
## GO:0044444 1478 4.500921e-01 6.654316e-14
## GO:0044391   83 9.999993e-01 2.207631e-13
## GO:0098800   60 9.999735e-01 5.292295e-13
## GO:0044429  230 9.998111e-01 1.993184e-12
## GO:0003735   70 9.999664e-01 5.233960e-12
## GO:0098798   66 9.999936e-01 7.488100e-12
## GO:0044455   80 9.999012e-01 9.164586e-12
## GO:0097458  338 3.708241e-01 1.209252e-11
## GO:0015934   57 9.999626e-01 1.219744e-11
## GO:0005740  181 9.976751e-01 1.798224e-11
## GO:0005746   41 9.993030e-01 1.776445e-10
## GO:0031966  166 9.992321e-01 3.613387e-10
## GO:0005622 2228 2.620933e-01 5.110499e-10
## GO:0036477  201 5.771427e-01 1.378388e-09
## GO:0005623 2356 3.324676e-06 1.695411e-09
```