RNAseqReanalysis01132021a

Xin-Qiao Zhang

1/13/2021

Introduction

```
library(airway)
```

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

```
## Loading required package: MatrixGenerics
```

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

```
##
## Attaching package: 'MatrixGenerics'
```

```
## The following objects are masked from 'package:matrixStats':
##
##     colAlls, colAnyNAs, colAnys, colAvgsPerRowSet, colCollapse,
##     colCounts, colCummaxs, colCummins, colCumprods, colCumsums,
##     colDiffs, colIQRDiffs, colIQRs, colLogSumExps, colMadDiffs,
##     colMads, colMaxs, colMeans2, colMedians, colMins, colOrderStats,
##     colProds, colQuantiles, colRanges, colRanks, colSdDiffs, colSds,
##     colSums2, colTabulates, colVarDiffs, colVars, colWeightedMads,
##     colWeightedMeans, colWeightedMedians, colWeightedSds,
##     colWeightedVars, rowAlls, rowAnyNAs, rowAnys, rowAvgsPerColSet,
##     rowCollapse, rowCounts, rowCummaxs, rowCummins, rowCumprods,
##     rowCumsums, rowDiffs, rowIQRDiffs, rowIQRs, rowLogSumExps,
##     rowMadDiffs, rowMads, rowMaxs, rowMeans2, rowMedians, rowMins,
##     rowOrderStats, rowProds, rowQuantiles, rowRanges, rowRanks,
##     rowSdDiffs, rowSds, rowSums2, rowTabulates, rowVarDiffs, rowVars,
##     rowWeightedMads, rowWeightedMeans, rowWeightedMedians,
##     rowWeightedSds, rowWeightedVars
```

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

```
## 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, colnames,
##     dirname, do.call, duplicated, eval, evalq, Filter, Find, get, grep,
##     grepl, intersect, is.unsorted, lapply, Map, mapply, match, mget,
##     order, paste, pmax, pmax.int, pmin, pmin.int, Position, rank,
##     rbind, Reduce, rownames, sapply, setdiff, sort, table, tapply,
##     union, unique, unsplit, which.max, which.min

## Loading required package: S4Vectors

##
## Attaching package: 'S4Vectors'

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

## Loading required package: IRanges

## Loading required package: GenomeInfoDb

## 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")'.

##
## Attaching package: 'Biobase'

## The following object is masked from 'package:MatrixGenerics':
##
##     rowMedians

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

```
library(annotables)
library(AnnotationDbi)
library(AnnotationHub)
```

```
## Loading required package: BiocFileCache

## Loading required package: dbplyr

##
## Attaching package: 'AnnotationHub'

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

```
library(apeglm)
library(ashr)
```

```r
library(base)
library(Biobase)
library(BiocGenerics)
library(biomaRt)
library(clusterProfiler)
```

```
##
## clusterProfiler v3.18.0  For help: https://guangchuangyu.github.io/software/clusterProfiler
##
## If you use clusterProfiler in published research, please cite:
## Guangchuang Yu, Li-Gen Wang, Yanyan Han, Qing-Yu He. clusterProfiler: an R package for comparing biol
##
## Attaching package: 'clusterProfiler'
## The following object is masked from 'package:biomaRt':
##
##     select
## The following object is masked from 'package:AnnotationDbi':
##
##     select
## The following object is masked from 'package:IRanges':
##
##     slice
## The following object is masked from 'package:S4Vectors':
##
##     rename
## The following object is masked from 'package:stats':
##
##     filter
```

```r
library(data.table)
```

```
##
## Attaching package: 'data.table'
## The following object is masked from 'package:SummarizedExperiment':
##
##     shift
## The following object is masked from 'package:GenomicRanges':
##
##     shift
## The following object is masked from 'package:IRanges':
##
##     shift
## The following objects are masked from 'package:S4Vectors':
##
##     first, second
```

```r
library(datasets)
library(dbplyr)
library(DEGreport)
```

```
library(DESeq2)
library(devtools)
```

```
## Loading required package: usethis
```

```
library(DOSE)
```

```
## DOSE v3.16.0  For help: https://guangchuangyu.github.io/software/DOSE
##
## If you use DOSE in published research, please cite:
## Guangchuang Yu, Li-Gen Wang, Guang-Rong Yan, Qing-Yu He. DOSE: an R/Bioconductor package for Disease
```

```
library(dplyr)
```

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

```
## The following objects are masked from 'package:data.table':
##
##     between, first, last
```

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

```
## The following objects are masked from 'package:dbplyr':
##
##     ident, sql
```

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

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

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

```
## The following object is masked from 'package:GenomeInfoDb':
##
##     intersect
```

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

```
## The following object is masked from 'package:matrixStats':
##
##     count
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
library(EnhancedVolcano)
## Loading required package: ggplot2
## Loading required package: ggrepel
## Registered S3 methods overwritten by 'ggalt':
##   method                  from
##   grid.draw.absoluteGrob  ggplot2
##   grobHeight.absoluteGrob ggplot2
##   grobWidth.absoluteGrob  ggplot2
##   grobX.absoluteGrob      ggplot2
##   grobY.absoluteGrob      ggplot2
library(EnsDb.Hsapiens.v86)
## Loading required package: ensembldb
## Loading required package: GenomicFeatures
## Loading required package: AnnotationFilter
##
## Attaching package: 'ensembldb'
## The following object is masked from 'package:dplyr':
##
##     filter
## The following object is masked from 'package:clusterProfiler':
##
##     filter
## The following object is masked from 'package:stats':
##
##     filter
library(ensembldb)
library(forcats)
library(genefilter)
##
## Attaching package: 'genefilter'
## The following objects are masked from 'package:MatrixGenerics':
##
##     rowSds, rowVars
## The following objects are masked from 'package:matrixStats':
##
##     rowSds, rowVars
library(geneplotter)
```

```
## Loading required package: lattice
##
## Attaching package: 'lattice'
## The following object is masked from 'package:clusterProfiler':
##
##     dotplot
## Loading required package: annotate
## Loading required package: XML
```

```
library(ggplot2)
library(ggrepel)
library(goseq)
```

```
## Loading required package: BiasedUrn
## Loading required package: geneLenDataBase
##
## Attaching package: 'geneLenDataBase'
## The following object is masked from 'package:S4Vectors':
##
##     unfactor
```

```
library(gplots)
```

```
##
## Attaching package: 'gplots'
## The following object is masked from 'package:IRanges':
##
##     space
## The following object is masked from 'package:S4Vectors':
##
##     space
## The following object is masked from 'package:stats':
##
##     lowess
```

```
library(IHW)
```

```
##
## Attaching package: 'IHW'
## The following object is masked from 'package:ggplot2':
##
##     alpha
```

```
library(IRanges)
library(knitr)
library(locfit)
```

```
## locfit 1.5-9.4    2020-03-24
```

```
library(magrittr)
```

```
##
```

```
## Attaching package: 'magrittr'

## The following object is masked from 'package:AnnotationFilter':
##
##     not
```

```
library(matrixStats)
library(MatrixGenerics)
library(methods)
library(pbapply)
library(pheatmap)
library(RColorBrewer)
library(Rcpp)
library(readr)
```

```
##
## Attaching package: 'readr'

## The following object is masked from 'package:genefilter':
##
##     spec
```

```
library(rmarkdown)
library(Rsubread)
library(S4Vectors)
library(stats)
library(stats4)
library(SummarizedExperiment)
library(testthat)
```

```
##
## Attaching package: 'testthat'

## The following objects are masked from 'package:magrittr':
##
##     equals, is_less_than, not

## The following object is masked from 'package:AnnotationFilter':
##
##     not

## The following object is masked from 'package:dplyr':
##
##     matches

## The following object is masked from 'package:devtools':
##
##     test_file
```

```
library(tibble)
library(tidyr)
```

```
##
## Attaching package: 'tidyr'

## The following object is masked from 'package:testthat':
##
##     matches

## The following object is masked from 'package:magrittr':
```

```
##
##     extract
```

```
## The following object is masked from 'package:S4Vectors':
##
##     expand
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v purrr   0.3.4     v stringr 1.4.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x IHW::alpha()             masks ggplot2::alpha()
## x dplyr::between()         masks data.table::between()
## x dplyr::collapse()        masks IRanges::collapse()
## x dplyr::combine()         masks Biobase::combine(), BiocGenerics::combine()
## x dplyr::count()           masks matrixStats::count()
## x dplyr::desc()            masks IRanges::desc()
## x testthat::equals()       masks magrittr::equals()
## x tidyr::expand()          masks S4Vectors::expand()
## x tidyr::extract()         masks magrittr::extract()
## x ensembldb::filter()      masks dplyr::filter(), clusterProfiler::filter(), stats::filter()
## x dplyr::first()           masks data.table::first(), S4Vectors::first()
## x dplyr::ident()           masks dbplyr::ident()
## x testthat::is_less_than() masks magrittr::is_less_than()
## x purrr::is_null()         masks testthat::is_null()
## x dplyr::lag()             masks stats::lag()
## x dplyr::last()            masks data.table::last()
## x tidyr::matches()         masks testthat::matches(), dplyr::matches()
## x purrr::none()            masks locfit::none()
## x testthat::not()          masks magrittr::not(), AnnotationFilter::not()
## x ggplot2::Position()      masks BiocGenerics::Position(), base::Position()
## x purrr::reduce()          masks GenomicRanges::reduce(), IRanges::reduce()
## x dplyr::rename()          masks clusterProfiler::rename(), S4Vectors::rename()
## x ensembldb::select()      masks dplyr::select(), clusterProfiler::select(), biomaRt::select(), Anno
## x purrr::set_names()       masks magrittr::set_names()
## x purrr::simplify()        masks clusterProfiler::simplify()
## x dplyr::slice()           masks clusterProfiler::slice(), IRanges::slice()
## x readr::spec()            masks genefilter::spec()
## x dplyr::sql()             masks dbplyr::sql()
## x purrr::transpose()       masks data.table::transpose()
```

```
library(tximeta)
library(tximport)
library(tximportData)
library(vsn)
library(XML)
```

setup

setup

```
setwd("~/Desktop/XinqiaoB2020Lungca/3LungCastrandno/3LungcaAndBladderca")
listMarts()
```

```
##                biomart                version
## 1 ENSEMBL_MART_ENSEMBL      Ensembl Genes 102
## 2   ENSEMBL_MART_MOUSE      Mouse strains 102
## 3     ENSEMBL_MART_SNP  Ensembl Variation 102
## 4 ENSEMBL_MART_FUNCGEN Ensembl Regulation 102
```

```
humanmart <- useEnsembl(biomart="ensembl", dataset="hsapiens_gene_ensembl", version="101")
```

input data

```
counts <- read.csv("PRJNA382834strandnocount.csv", stringsAsFactors = FALSE)
counts <- counts[, c(1:6, 8,9,13)]
counts <- data.frame(counts[,-1], row.names = counts[,1])
head(counts, n=6)
```

```
##                 HT1197 HT1376  J82  T24 x253JP RT112   RT4  UC3
## ENSG00000000003   1460    740 1421 1011   2530  3650  7688 1250
## ENSG00000000005      0      0    0    0      0     0     0    0
## ENSG00000000457    414    567  294  477    391   677   709  405
## ENSG00000000460    794   1842  903 1182    879  1449   609  731
## ENSG00000000938      1      5    2    0      1    27    81    0
## ENSG00000000971      5     22   52   23     63  1409 16905  293
```

```
samples <- read.csv("conditionPRJNA382834strandnoA.csv", stringsAsFactors = FALSE)
samples <- samples[1:8,]
samples <- data.frame(samples[,-1], row.names = samples[,1])
head(samples, n=8)
```

```
##        condition replicate
## HT1197         R         1
## HT1376         R         2
## J82            R         3
## T24            R         4
## x253JP         S         1
## RT112          S         2
## RT4            S         3
## UC3            S         4
```

```
colnames(counts) <- c(rownames(samples))
all(rownames(samples) == colnames(counts))
```

```
## [1] TRUE
```

DESeq2 analysis

```
dds <- DESeqDataSetFromMatrix(countData = counts, colData = samples, design = ~condition)
```

```
## Warning in DESeqDataSet(se, design = design, ignoreRank): some variables in
## design formula are characters, converting to factors
```

```
dds <- estimateSizeFactors(dds)
sizeFactors(dds)
```

```
##    HT1197    HT1376       J82       T24    x253JP     RT112       RT4       UC3
## 0.9845122 1.0230076 0.9010706 1.1620333 1.1918203 0.9425694 0.9545667 1.0372552
```

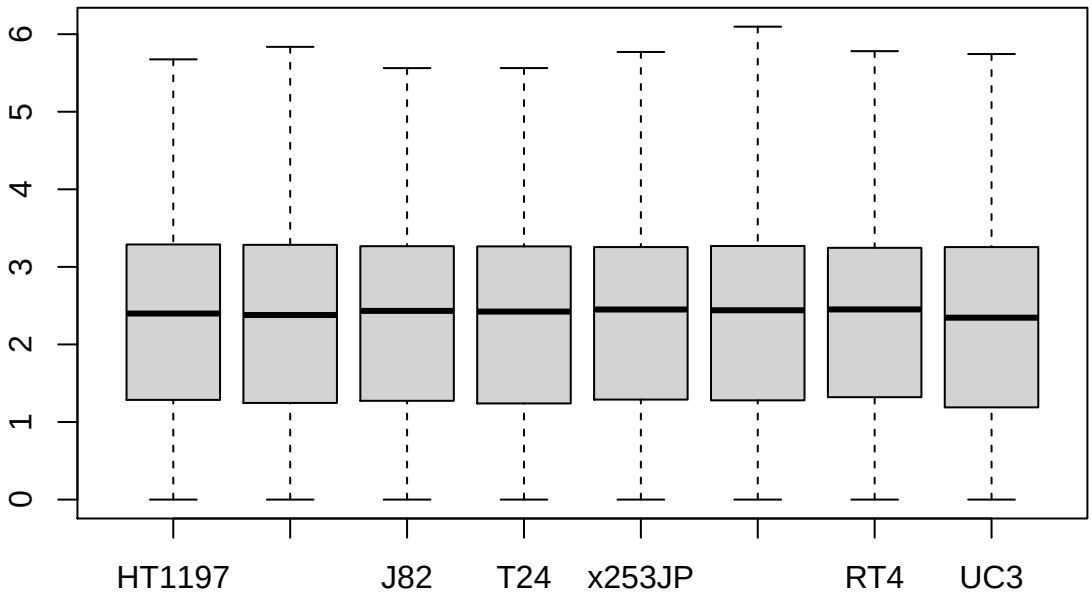
```
colData(dds)
```

```
## DataFrame with 8 rows and 3 columns
##         condition replicate sizeFactor
##          <factor> <integer>  <numeric>
## HT1197         R         1   0.984512
## HT1376         R         2   1.023008
## J82            R         3   0.901071
## T24            R         4   1.162033
## x253JP         S         1   1.191820
## RT112          S         2   0.942569
## RT4            S         3   0.954567
## UC3            S         4   1.037255
```

```
keep <- rowSums(counts(dds) >= 5) >= 4
table(keep)
```

```
## keep
## FALSE  TRUE
## 36026 19414
```

```
dds <-dds[keep,]
normalized_counts <- counts(dds, normalized=TRUE)
boxplot(log10(counts(dds, normalized=TRUE)+1))
```




```
vsd <- vst(dds)
```

```
## -- note: fitType='parametric', but the dispersion trend was not well captured by the
##    function: y = a/x + b, and a local regression fit was automatically substituted.
##    specify fitType='local' or 'mean' to avoid this message next time.
```

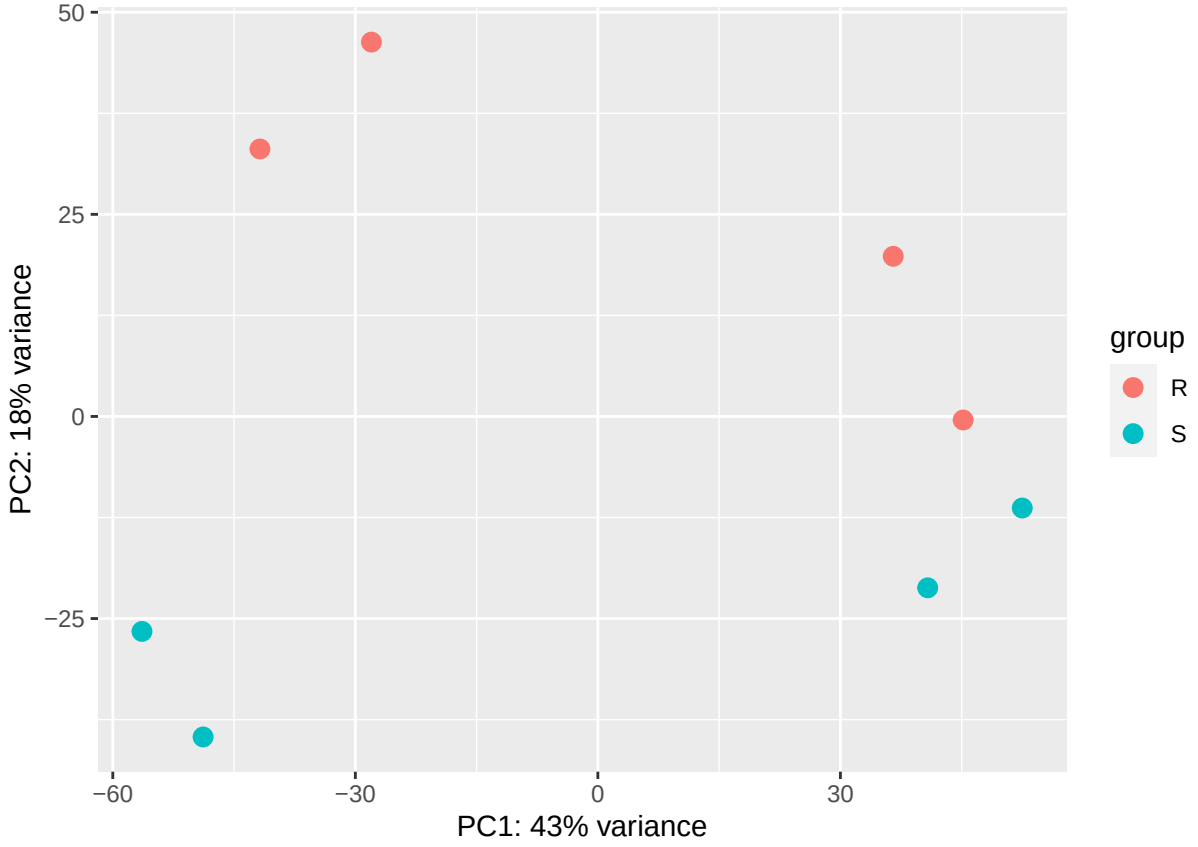
```
class(vsd)
```

```
## [1] "DESeqTransform"
## attr(,"package")
## [1] "DESeq2"
```

```
assay(vsd)[1:3, 1:8]
```

```
##                    HT1197    HT1376       J82       T24    x253JP     RT112
## ENSG00000000003 10.651548  9.848954 10.725604 10.043988 11.094746 11.902333
## ENSG00000000457  9.321348  9.581646  9.096379  9.299386  9.101097  9.841628
## ENSG00000000460  9.962911 10.887914 10.198737 10.215316  9.869072 10.694782
##                       RT4       UC3
## ENSG00000000003 12.975876 10.407563
## ENSG00000000457  9.876405  9.254231
## ENSG00000000460  9.720845  9.822061
```

```
plotPCA(vsd, "condition")
```

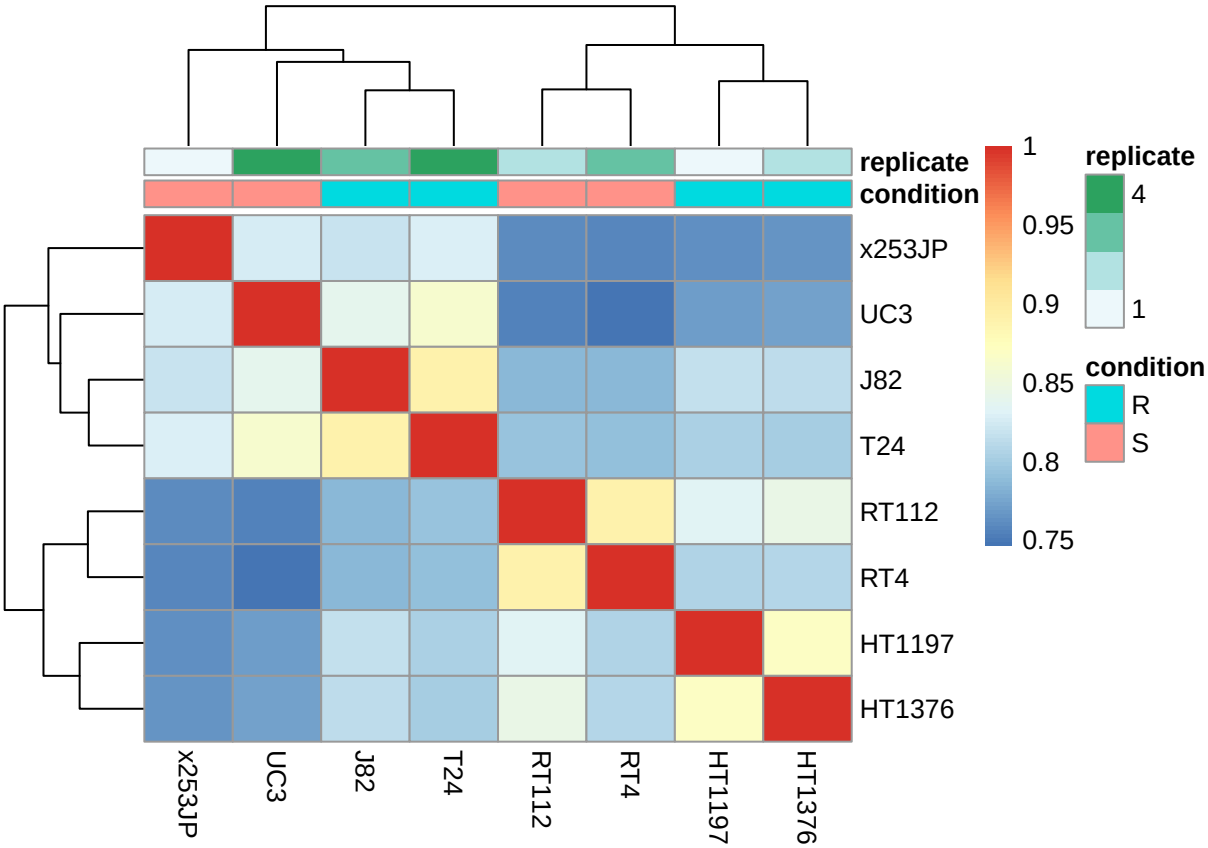



```
plotPCA(vsd, "condition", returnData=TRUE)
```

```
##              PC1         PC2 group condition   name
## HT1197 -28.00838  46.3037010     R         R HT1197
## HT1376 -41.79575  33.0875849     R         R HT1376
## J82     36.53862  19.8098940     R         R    J82
## T24     45.18675  -0.4378793     R         R    T24
## x253JP  40.80287 -21.1969259     S         S x253JP
## RT112  -56.39084 -26.5916600     S         S  RT112
## RT4    -48.82928 -39.6456937     S         S    RT4
## UC3     52.49602 -11.3290211     S         S    UC3
```

```
vsd %>%
  assay() %>%
  cor() %>%
```

```
pheatmap(annotation=samples[,c("condition", "replicate")])
```




```
dds <- DESeq(dds)
```

```
## using pre-existing size factors
## estimating dispersions
## gene-wise dispersion estimates
## mean-dispersion relationship
## final dispersion estimates
## fitting model and testing
```

```
res <- results(dds)
resultsNames(dds)
```

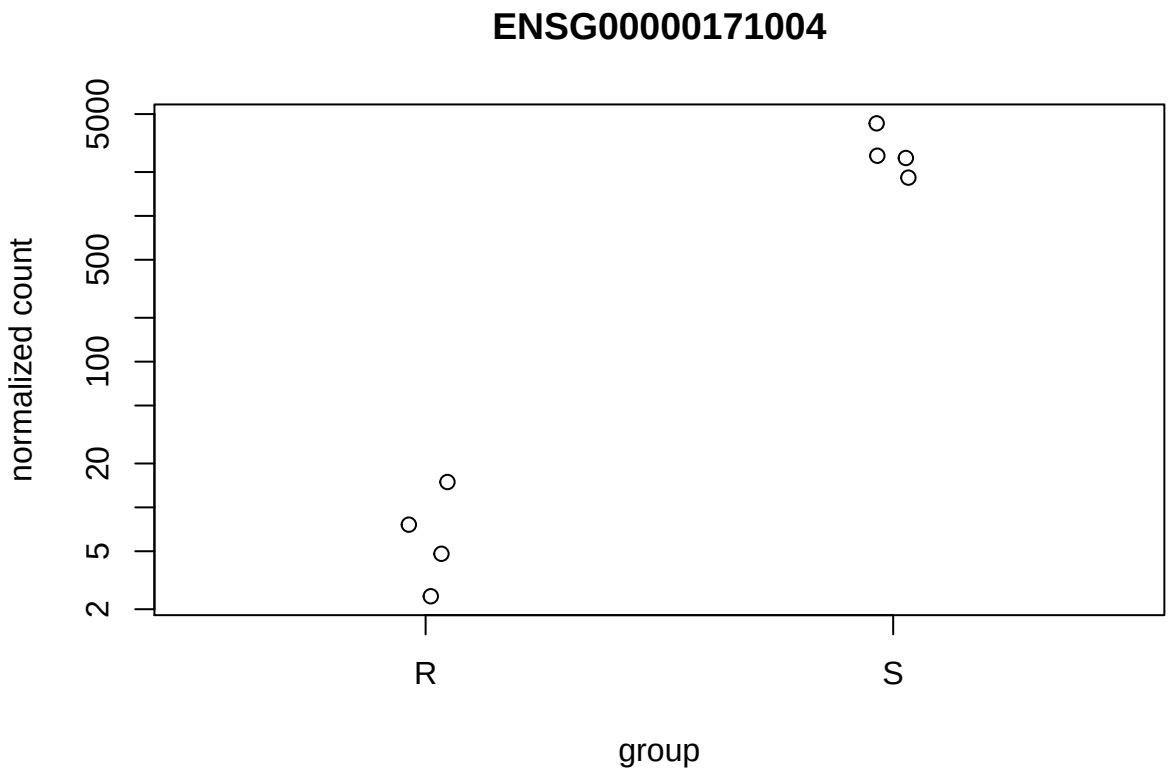
```
## [1] "Intercept"        "condition_S_vs_R"
```

```
head(res[order(res$padj), ], n=60)
```

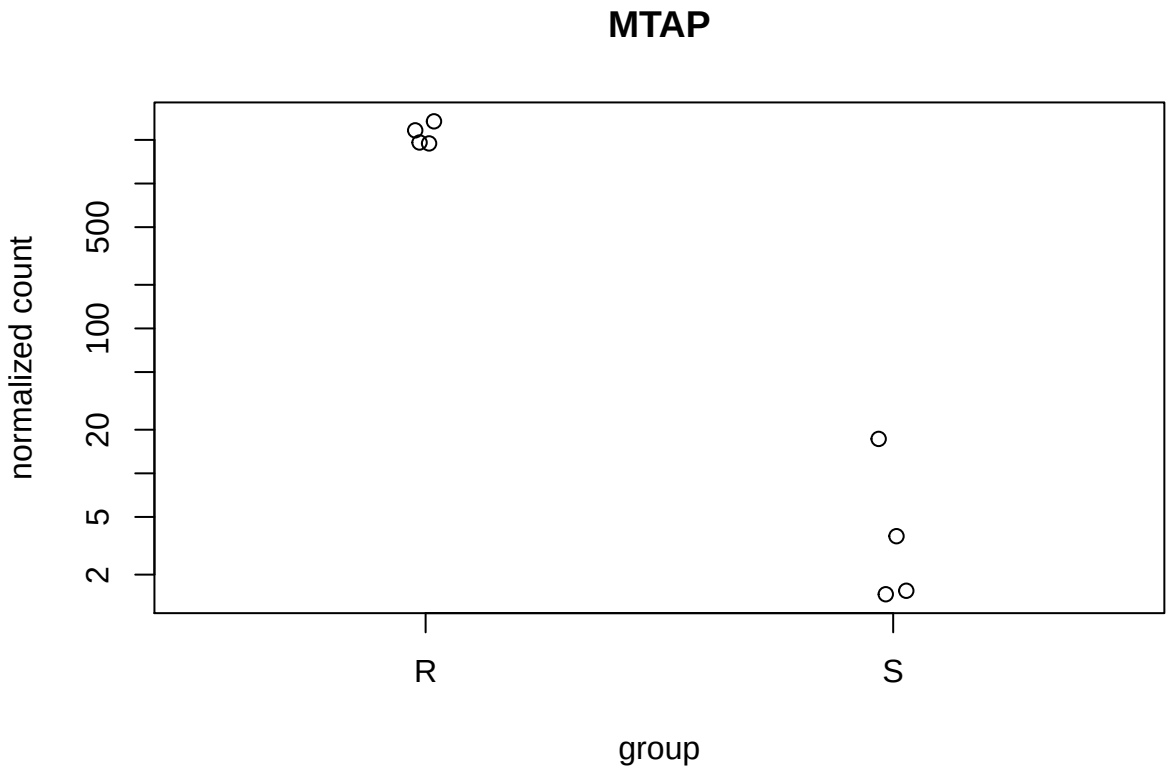
```
## log2 fold change (MLE): condition S vs R
## Wald test p-value: condition S vs R
## DataFrame with 60 rows and 6 columns
##                   baseMean log2FoldChange     lfcSE      stat      pvalue
##                  <numeric>      <numeric> <numeric> <numeric>   <numeric>
## ENSG00000171004   1407.34        8.68073  0.619538  14.01163 1.32336e-44
## ENSG00000149582    537.20       -5.64295  0.505749 -11.15762 6.57283e-29
## ENSG00000099810   1105.68       -8.61934  0.856420 -10.06438 7.93861e-24
```

```
## ENSG00000165949   1949.55       -6.90973  0.688394 -10.03747 1.04321e-23
## ENSG00000166147  16350.49       -8.11368  0.840921  -9.64856 4.98529e-22
## ...                   ...            ...       ...       ...         ...
## ENSG00000074410  5928.905        7.30609  1.358011   5.37999 7.44892e-08
## ENSG00000064787  2653.990        9.41571  1.753535   5.36956 7.89286e-08
## ENSG00000076706   865.793       -7.70499  1.436826  -5.36251 8.20741e-08
## ENSG00000143369  2258.291       -5.18343  0.968615  -5.35138 8.72857e-08
## ENSG00000130758  1061.319        1.46541  0.274099   5.34630 8.97725e-08
##                       padj
##                  <numeric>
## ENSG00000171004 2.26599e-40
## ENSG00000149582 5.62733e-25
## ENSG00000099810 4.46570e-20
## ENSG00000165949 4.46570e-20
## ENSG00000166147 1.70726e-18
## ...                     ...
## ENSG00000074410 2.27764e-05
## ENSG00000064787 2.37104e-05
## ENSG00000076706 2.42303e-05
## ENSG00000143369 2.53321e-05
## ENSG00000130758 2.54505e-05
```

```
plotCounts(dds, which.min(res$padj), "condition")
```




```
plotCounts(dds, gene = "ENSG00000099810", "condition", main="MTAP")
```




```
summary(res)
```

```
##
## out of 19414 with nonzero total read count
## adjusted p-value < 0.1
## LFC > 0 (up)       : 277, 1.4%
## LFC < 0 (down)     : 350, 1.8%
## outliers [1]       : 785, 4%
## low counts [2]     : 1506, 7.8%
## (mean count < 9)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

```
sum(res$padj <0.1, na.rm = TRUE)
```

```
## [1] 627
```

```
sum(res$padj <0.05, na.rm = TRUE)
```

```
## [1] 446
```

```
sum(res$padj <0.01, na.rm = TRUE)
```

```
## [1] 233
```

Top 60 Gene