RNAseqReanalysis01132021a

Xin-Qiao Zhang

1/14/2021

Introduction

Bladder cancer

Library for rstudio

Setup

setup working directory

```
setwd("/Users/xinqiaozhang/Desktop/PRJNA382834/01142021")
listMarts()
```

```
## Ensembl site unresponsive, trying asia mirror
```

```
##                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
```

```
if(interactive()){
  mart <- useEnsembl("ensembl")
  humanmart <- useEnsembl(biomart = "ensembl", mirror = "useast")
}
humanmart = useEnsembl(biomart = "ensembl", dataset = "hsapiens_gene_ensembl", mirror = "useast")
```

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
```

```r
samples <- read.csv("conditionPRJNA382834strandno.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
## x5637          S         2
## RT112          S         3
## RT4            S         4
```

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

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

DESeq2 analysis

```r
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
```

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

```
##    HT1197    HT1376       J82       T24    x253JP     x5637     RT112       RT4
## 0.9845122 1.0230076 0.9010706 1.1620333 1.1918203 0.9425694 0.9545667 1.0372552
```

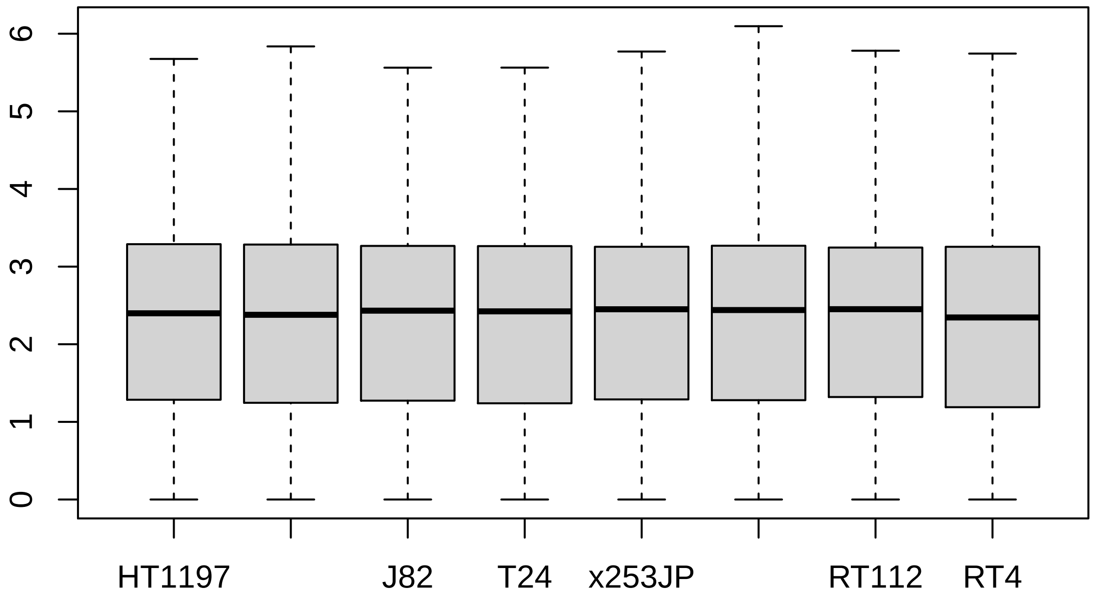
```r
colData(dds)
```

```
## 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
## x5637          S         2   0.942569
## RT112          S         3   0.954567
## RT4            S         4   1.037255
```

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

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

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



```r
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.
```

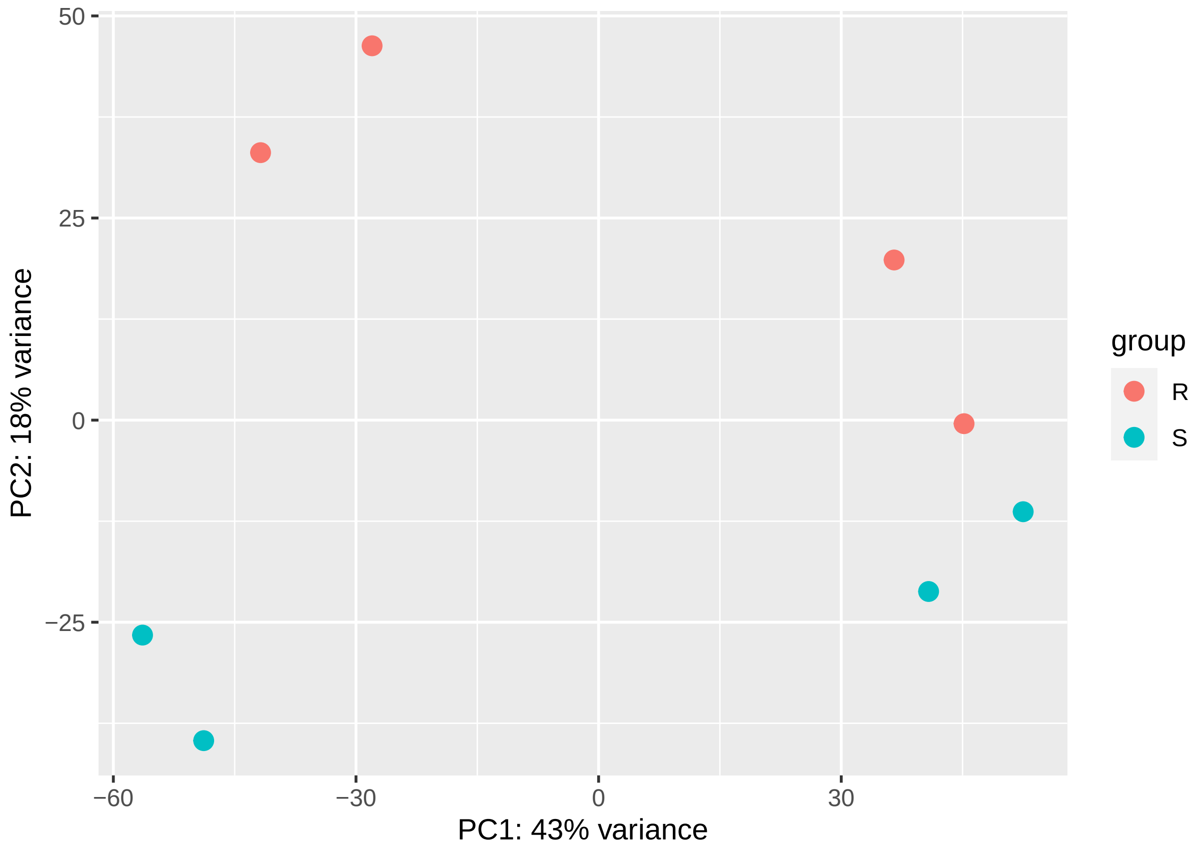
```r
class(vsd)
```

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

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

```
##                    HT1197    HT1376       J82       T24    x253JP     x5637
## 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
##                     RT112       RT4
## ENSG00000000003 12.975876 10.407563
## ENSG00000000457  9.876405  9.254231
## ENSG00000000460  9.720845  9.822061
```

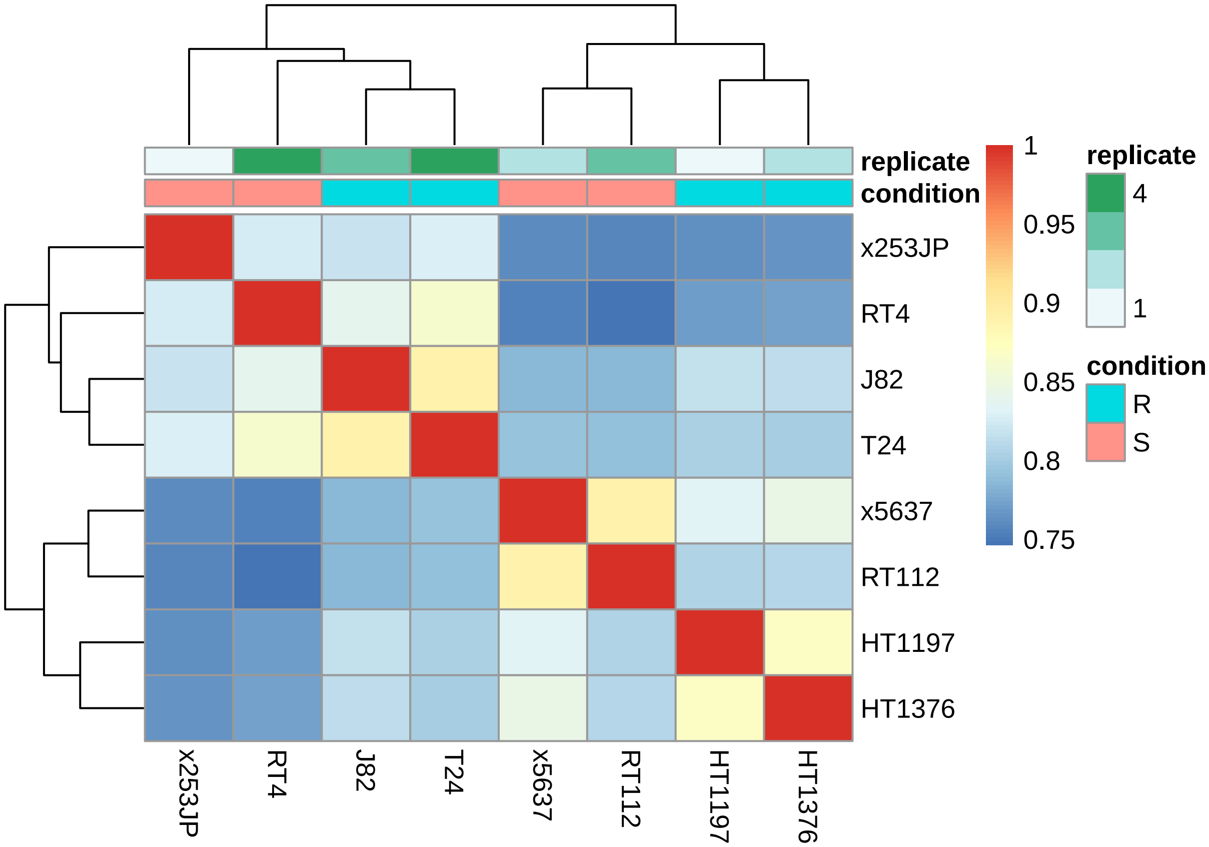
```r
plotPCA(vsd, "condition")
```



```r
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
## x5637  -56.39084 -26.5916600     S         S  x5637
## RT112  -48.82928 -39.6456937     S         S  RT112
## RT4     52.49602 -11.3290211     S         S    RT4
```

```r
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
```

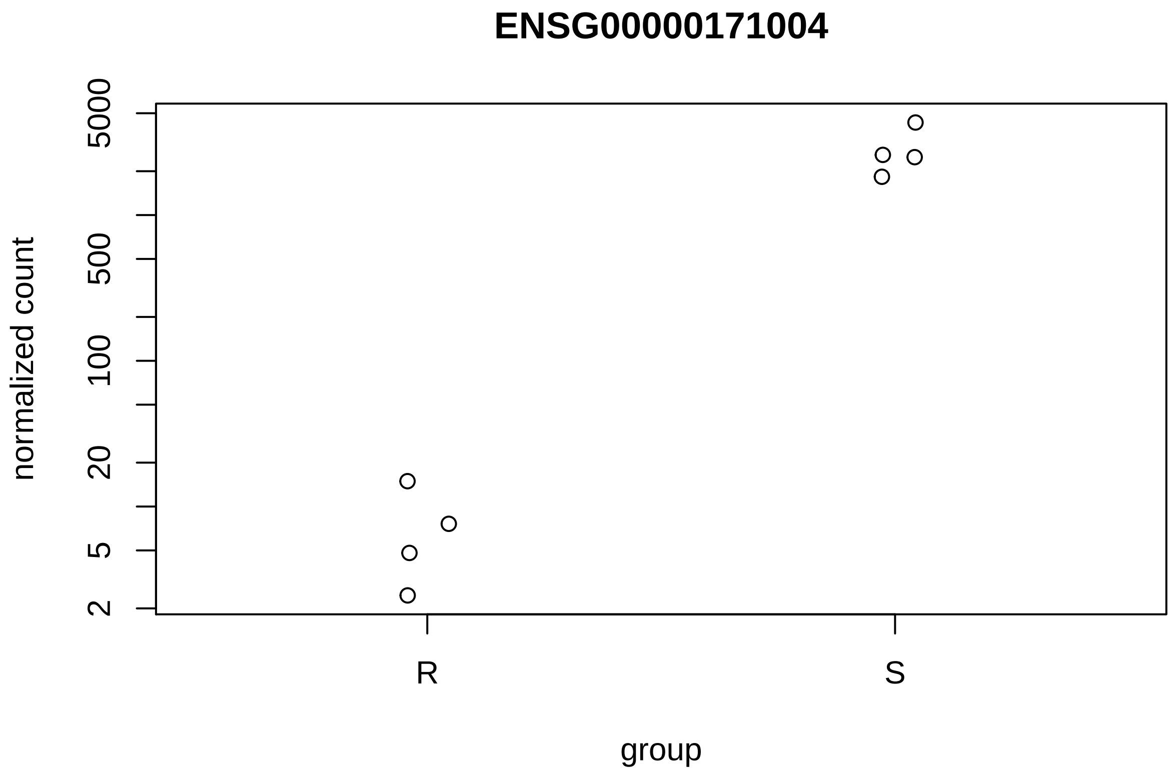
```
res <- results(dds)
resultsNames(dds)
```

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

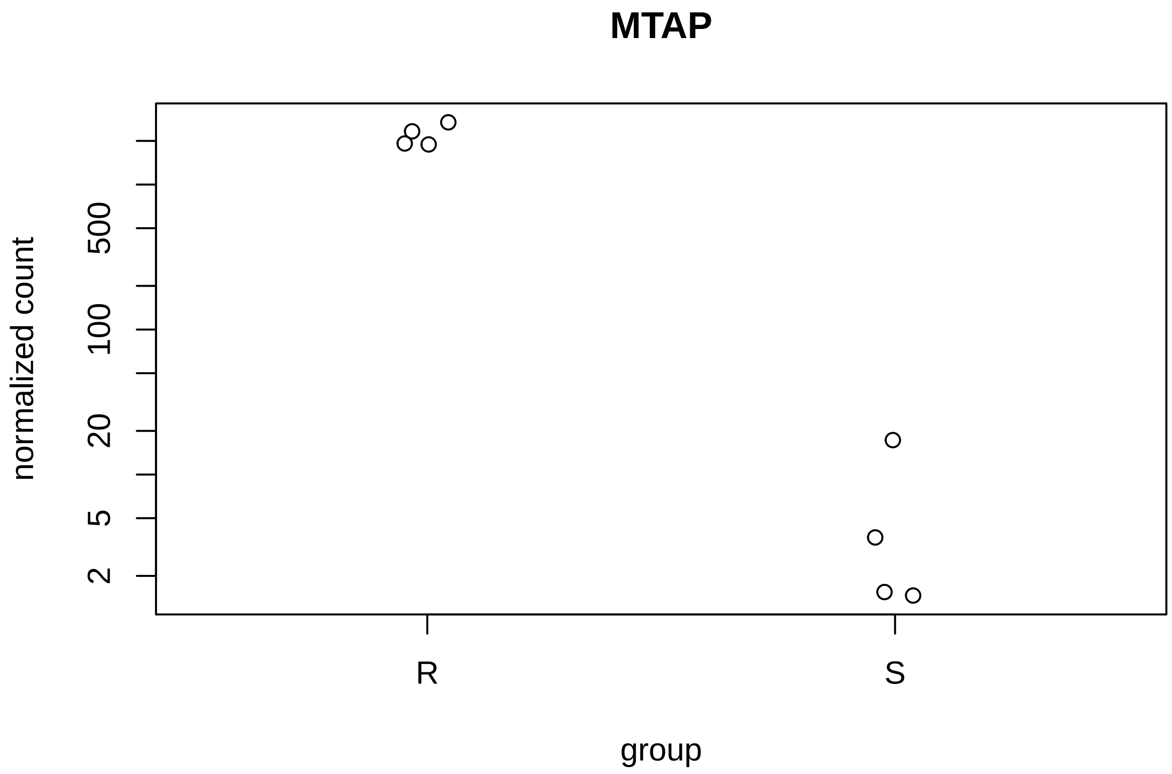
```
head(res[order(res$padj), ], n=60)
```

```
## 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
```

```r
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
```

```r
res1 <- results(dds, name = "condition_S_vs_R")
res1 <- results(dds, contrast = c("condition", "S", "R"))
resLFC <- lfcShrink(dds, coef = "condition_S_vs_R", type="apeglm")
```

```
## using 'apeglm' for LFC shrinkage. If used in published research, please cite:
##     Zhu, A., Ibrahim, J.G., Love, M.I. (2018) Heavy-tailed prior distributions for
##     sequence count data: removing the noise and preserving large differences.
##     Bioinformatics. https://doi.org/10.1093/bioinformatics/bty895
```

```r
resLFC
```

```
## log2 fold change (MAP): condition S vs R
## Wald test p-value: condition S vs R
## DataFrame with 19414 rows and 5 columns
##                  baseMean log2FoldChange     lfcSE      pvalue        padj
##                 <numeric>      <numeric> <numeric>   <numeric>   <numeric>
## ENSG00000000003  2488.448      0.8658578  1.236514 9.13667e-03 0.164410405
## ENSG00000000457   486.381      0.1202203  0.247501 3.79943e-01 0.811809642
## ENSG00000000460  1030.492     -0.0991357  0.251677 4.32892e-01 0.843179196
## ENSG00000000971  2455.485      6.9794622  1.514278 1.14249e-07 0.000031553
## ENSG00000001036  3683.040     -0.2075753  0.308397 1.66731e-01 0.636387722
## ...                   ...            ...       ...         ...         ...
## ENSG00000283041 6073.4798      0.0149682  0.227995   0.9064768    0.984796
## ENSG00000283050  407.4790     -0.0724409  0.248527   0.5447904    0.887176
## ENSG00000283064   11.6368     -0.1445268  0.317970   0.0796848    0.492223
## ENSG00000283078   36.2228     -0.0643776  0.259935   0.5313379    0.882368
## ENSG00000283103  150.3640      0.2732784  0.424769   0.0514922    0.402398
```

```r
resLFCa <- resLFC[order(resLFC$padj),]
resLFCb <- resLFCa[1:60,]
resLFCb
```

```
## log2 fold change (MAP): condition S vs R
## Wald test p-value: condition S vs R
## DataFrame with 60 rows and 5 columns
##                  baseMean log2FoldChange     lfcSE      pvalue        padj
##                 <numeric>      <numeric> <numeric>   <numeric>   <numeric>
## ENSG00000171004   1407.34        8.59263  0.624035 1.32336e-44 2.26599e-40
## ENSG00000149582    537.20       -5.54878  0.509533 6.57283e-29 5.62733e-25
## ENSG00000099810   1105.68       -8.44030  0.860692 7.93861e-24 4.46570e-20
## ENSG00000165949   1949.55       -6.76511  0.695394 1.04321e-23 4.46570e-20
## ENSG00000166147  16350.49       -7.92542  0.850526 4.98529e-22 1.70726e-18
## ...                   ...            ...       ...         ...         ...
## ENSG00000074410  5928.905       6.773197  1.436804 7.44892e-08 2.27764e-05
## ENSG00000064787  2653.990       8.711776  1.878490 7.89286e-08 2.37104e-05
## ENSG00000076706   865.793      -0.117002  0.313760 8.20741e-08 2.42303e-05
## ENSG00000143369  2258.291      -4.780574  1.018371 8.72857e-08 2.53321e-05
## ENSG00000130758  1061.319       1.364209  0.283355 8.97725e-08 2.54505e-05
```

Top 60 Gene

```
meta <- samples
meta <- meta %>%
  rownames_to_column(var = "samplename") %>%
  as_tibble()

normalized_counts1 <- normalized_counts
normalized_counts1 <- normalized_counts1 %>%
  data.frame() %>%
  rownames_to_column(var = "gene")

resA <- res
resA <- data.frame(resA)
res_table_tb <- resA
res_table_tb <- data.frame(res_table_tb) %>%
  rownames_to_column(var = "gene") %>%
  as_tibble()
top30_sig_genes <- res_table_tb %>%
  arrange(padj) %>%
  pull(gene) %>%
  head(n=30)

top30_sig_genes
```

```
##  [1] "ENSG00000171004" "ENSG00000149582" "ENSG00000099810" "ENSG00000165949"
##  [5] "ENSG00000166147" "ENSG00000114270" "ENSG00000147889" "ENSG00000141574"
##  [9] "ENSG00000147883" "ENSG00000128591" "ENSG00000139910" "ENSG00000142619"
## [13] "ENSG00000115414" "ENSG00000151640" "ENSG00000117525" "ENSG00000131737"
## [17] "ENSG00000239697" "ENSG00000255874" "ENSG00000162734" "ENSG00000184489"
## [21] "ENSG00000265194" "ENSG00000103064" "ENSG00000079931" "ENSG00000185885"
## [25] "ENSG00000168386" "ENSG00000101335" "ENSG00000221968" "ENSG00000151388"
## [29] "ENSG00000177096" "ENSG00000170848"
```

Session information

```
sessionInfo()
```

```
## R version 4.0.3 (2020-10-10)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS Big Sur 10.16
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRblas.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
```

```
## [1] parallel  stats4    stats     graphics  grDevices utils     datasets 
## [8] methods   base     
## 
## other attached packages:
##  [1] vsn_3.58.0                  txImportData_1.18.0        
##  [3] tximport_1.18.0             tximeta_1.8.3              
##  [5] tidyverse_1.3.0             tidyr_1.1.2                
##  [7] tibble_3.0.4                testthat_3.0.1             
##  [9] stringr_1.4.0               Rsubread_2.4.2             
## [11] rmarkdown_2.6               readr_1.4.0                
## [13] Rcpp_1.0.5                  RColorBrewer_1.1-2         
## [15] purrr_0.3.4                 pheatmap_1.0.12            
## [17] pbapply_1.4-3               magrittr_2.0.1             
## [19] locfit_1.5-9.4              knitr_1.30                 
## [21] IHW_1.18.0                  httr_1.4.2                 
## [23] gplots_3.1.1                goseq_1.42.0               
## [25] geneLenDataBase_1.26.0      geneplotter_1.68.0         
## [27] annotate_1.68.0             XML_3.99-0.5               
## [29] lattice_0.20-41             genefilter_1.72.0          
## [31] forcats_0.5.0               EnsDb.Hsapiens.v86_2.99.0  
## [33] ensembldb_2.14.0            GenomicFeatures_1.42.1     
## [35] EnhancedVolcano_1.8.0       ggrepel_0.9.0              
## [37] ggplot2_3.3.2               dplyr_1.0.2                
## [39] DOSE_3.16.0                 devtools_2.3.2             
## [41] usethis_2.0.0               DESeq2_1.30.0              
## [43] DEGreport_1.26.0            data.table_1.13.4          
## [45] clusterProfiler_3.18.0      biomaRt_2.46.0             
## [47] BiasedUrn_1.07              ashr_2.2-47                
## [49] apeglm_1.12.0               AnnotationHub_2.22.0       
## [51] BiocFileCache_1.14.0        dbplyr_2.0.0               
## [53] AnnotationFilter_1.14.0     AnnotationDbi_1.52.0       
## [55] annotables_0.1.91           airway_1.10.0              
## [57] SummarizedExperiment_1.20.0 Biobase_2.50.0             
## [59] GenomicRanges_1.42.0        GenomeInfoDb_1.26.2        
## [61] IRanges_2.24.1              S4Vectors_0.28.1           
## [63] BiocGenerics_0.36.0         MatrixGenerics_1.2.0       
## [65] matrixStats_0.57.0         
## 
## loaded via a namespace (and not attached):
##   [1] rappdirs_0.3.1                rtracklayer_1.50.0           
##   [3] coda_0.19-4                   bit64_4.0.5                  
##   [5] irlba_2.3.3                   DelayedArray_0.16.0          
##   [7] RCurl_1.98-1.2                generics_0.1.0               
##   [9] preprocessCore_1.52.0         callr_3.5.1                  
##  [11] cowplot_1.1.0                 RSQLite_2.2.1                
##  [13] shadowtext_0.0.7              bit_4.0.4                    
##  [15] enrichplot_1.10.1             lubridate_1.7.9.2            
##  [17] xml2_1.3.2                    httpuv_1.5.4                 
##  [19] assertthat_0.2.1              viridis_0.5.1                
##  [21] xfun_0.19                     hms_0.5.3                    
##  [23] evaluate_0.14                 promises_1.1.1               
##  [25] fansi_0.4.1                   progress_1.2.2               
##  [27] readxl_1.3.1                  caTools_1.18.0               
##  [29] igraph_1.2.6                  DBI_1.1.0                    
##  [31] tmvnsim_1.0-2                 reshape_0.8.8
##  [33] ellipsis_0.3.1                backports_1.2.1
##  [35] vctrs_0.3.6                   remotes_2.2.0
##  [37] Cairo_1.5-12.2                withr_2.3.0
##  [39] ggforce_0.3.2                 lasso2_1.2-21.1
##  [41] bdsmatrix_1.3-4               GenomicAlignments_1.26.0
##  [43] fdrtool_1.2.15                prettyunits_1.1.1
##  [45] mnormt_2.0.2                  cluster_2.1.0
##  [47] lazyeval_0.2.2                crayon_1.3.4
##  [49] labeling_0.4.2                slam_0.1-48
##  [51] edgeR_3.32.0                  pkgconfig_2.0.3
##  [53] tweenr_1.0.1                  nlme_3.1-151
##  [55] vipor_0.4.5                   pkgload_1.1.0
##  [57] ProtGenerics_1.22.0           rlang_0.4.9
##  [59] lifecycle_0.2.0               downloader_0.4
##  [61] affyio_1.60.0                 extrafontdb_1.0
##  [63] modelr_0.1.8                  invgamma_1.1
##  [65] cellranger_1.1.0              ggrastr_0.2.1
##  [67] rprojroot_2.0.2               polyclip_1.10-0
##  [69] Matrix_1.3-0                  lpsymphony_1.18.0
##  [71] reprex_0.3.0                  beeswarm_0.2.3
##  [73] GlobalOptions_0.1.2           processx_3.4.5
##  [75] png_0.1-7                     viridisLite_0.3.0
##  [77] rjson_0.2.20                  bitops_1.0-6
##  [79] ConsensusClusterPlus_1.54.0   KernSmooth_2.23-18
##  [81] Biostrings_2.58.0             blob_1.2.1
##  [83] shape_1.4.5                   mixsqp_0.3-43
##  [85] SQUAREM_2020.5                qvalue_2.22.0
##  [87] scales_1.1.1                  memoise_1.1.0
##  [89] plyr_1.8.6                    zlibbioc_1.36.0
##  [91] compiler_4.0.3                scatterpie_0.1.5
##  [93] bbmle_1.0.23.1                ash_1.0-15
##  [95] clue_0.3-58                   affy_1.68.0
##  [97] Rsamtools_2.6.0               cli_2.2.0
##  [99] XVector_0.30.0                ps_1.5.0
## [101] mgcv_1.8-33                   MASS_7.3-53
## [103] tidyselect_1.1.0              stringi_1.5.3
## [105] proj4_1.0-10                  emdbook_1.3.12
## [107] yaml_2.2.1                    GOSemSim_2.16.1
## [109] askpass_1.1                   grid_4.0.3
## [111] fastmatch_1.1-0               tools_4.0.3
## [113] circlize_0.4.11               logging_0.10-108
## [115] gridExtra_2.3                 farver_2.0.3
## [117] ggraph_2.0.4                  digest_0.6.27
## [119] rvcheck_0.1.8                 BiocManager_1.30.10
## [121] shiny_1.5.0                   broom_0.7.3
## [123] ggalt_0.4.0                   BiocVersion_3.12.0
## [125] later_1.1.0.1                 Nozzle.R1_1.1-1
## [127] ggdendro_0.1.22               ComplexHeatmap_2.6.2
## [129] psych_2.0.12                  colorspace_2.0-0
## [131] rvest_0.3.6                   fs_1.5.0
## [133] truncnorm_1.0-8               splines_4.0.3
## [135] graphlayouts_0.7.1            sessioninfo_1.1.1
## [137] xtable_1.8-4                  jsonlite_1.7.2
## [139] tidygraph_1.2.0               R6_2.5.0
## [141] pillar_1.4.7                  htmltools_0.5.0
## [143] mime_0.9                      glue_1.4.2
## [145] fastmap_1.0.1                 BiocParallel_1.24.1
## [147] interactiveDisplayBase_1.28.0 maps_3.3.0
## [149] fgsea_1.16.0                  pkgbuild_1.2.0
## [151] mvtnorm_1.1-1                 numDeriv_2016.8-1.1
## [153] curl_4.3                      ggbeeswarm_0.6.0
## [155] gtools_3.8.2                  GO.db_3.12.1
## [157] openssl_1.4.3                 Rttf2pt1_1.3.8
## [159] survival_3.2-7                limma_3.46.0
## [161] desc_1.2.0                    munsell_0.5.0
## [163] DO.db_2.9                     GetoptLong_1.0.5
## [165] GenomeInfoDbData_1.2.4        haven_2.3.1
## [167] reshape2_1.4.4                gtable_0.3.0
## [169] extrafont_0.17
```