2-DESeq2 analysis

BAI Qiang*

2021-09-24 00:19:34 +0200

Contents

*University Liege, mail qiang.bai@uliege.be

1 Description

RNA-seq data were analyzed using R Bioconductor (3.5.1) and DESeq2 package (version 1.26.0)[1].

2 Load packages and data

```
library(DESeq2)
library(ggplot2)
library(pheatmap)
library(RColorBrewer)
library(EnhancedVolcano)
library(forcats)
```

Counts data are also accessible in NCBI GEO under accession number GSE183973.

```
COUNTS <- read.table("./merged_gene_counts.txt",sep="\t", header=T, row.names = NULL)

dim(COUNTS)
```

```
## [1] 63677    29
```

Make gene names as rownames:

```
Genes <- COUNTS$gene_name
rownames(COUNTS) = make.names(Genes, unique=TRUE)

COUNTS <- COUNTS[,-c(1:2)]
head(COUNTS, 3)
```

```
## # A tibble: 3 x 27
##   X17.non.smoker.1.m~ X25.copd.1.mono_NG~ X10.smoker.1.mono_~ X16.non.
   smoker.1.~
##                 <int>               <int>               <int>
                <int>
## 1                   0                   0                   0
                    0
## 2                  32                  69                 104
                   76
## 3                   0                   0                   0
                    0
## # ... with 23 more variables: ...
```

Arrange the sample order to have the right group order: Healthy, Smoker and COPD.

```
COUNTS <- COUNTS [,c(4,1,15,7,12,21,26,17,27,3,13,24,6,23,18,10,19,20,2,22,9,8,14,5,16,25,11)]
```

3 Make metadata for bulkRNAseq samples

```
colnames(COUNTS) <- c("Healthy_1_Mono", "Healthy_1_cAM", "Healthy_1_sAM", "Healthy_2_Mono", "Healthy_2_cAM", "Healthy_2_sAM", "Healthy_3_Mono", "Helathy_3_cAM", "Healthy_3_sAM", "Smoker_1_Mono", "Smoker_1_cAM", "Smoker_1_sAM", "Smoker_2_Mono", "Smoker_2_cAM", "Smoker_2_sAM", "Smoker_3_Mono", "Smoker_3_cAM", "Smoker_3_sAM", "COPD_1_Mono", "COPD_1_cAM", "COPD_1_sAM","COPD_2_Mono", "COPD_2_cAM", "COPD_2_sAM", "COPD_3_Mono","COPD_3_cAM", "COPD_3_sAM")

SampleSheet <- data.frame(
  "Treatment" = rep(c("Healthy","Smoker","COPD"),each=9),

  "Cells" = rep(c("Monocytes","AFhi␣cAM","AFlo␣AM"),3)
)

SampleSheet
```

```
## # A tibble: 27 x 2
##    Treatment Cells
##    <chr>     <chr>
##  1 Healthy   Monocytes
##  2 Healthy   AFhi cAM
##  3 Healthy   AFlo AM
##  4 Healthy   Monocytes
##  5 Healthy   AFhi cAM
##  6 Healthy   AFlo AM
##  7 Healthy   Monocytes
##  8 Healthy   AFhi cAM
##  9 Healthy   AFlo AM
## 10 Smoker    Monocytes
## # ... with 17 more rows
```

```
rownames(SampleSheet) <- colnames(COUNTS)
SampleSheet
```

```
## # A tibble: 27 x 2
##    Treatment Cells
##    <chr>     <chr>
##  1 Healthy   Monocytes
##  2 Healthy   AFhi cAM
##  3 Healthy   AFlo AM
##  4 Healthy   Monocytes
##  5 Healthy   AFhi cAM
##  6 Healthy   AFlo AM
##  7 Healthy   Monocytes
##  8 Healthy   AFhi cAM
##  9 Healthy   AFlo AM
## 10 Smoker    Monocytes
## # ... with 17 more rows
```

4 DESeq2

```
dds <- DESeqDataSetFromMatrix(
  countData= COUNTS,
  colData= SampleSheet,
  design= ~ Cells + Treatment
)

dds
```

```
## class: DESeqDataSet
## dim: 63677 27
## metadata(1): version
## assays(1): counts
## rownames(63677): DDX11L1 WASH7P ... FAM58CP CTBP2P1
## rowData names(0):
## colnames(27): Healthy_1_Mono Healthy_1_cAM ... COPD_3_cAM COPD_3_sAM
## colData names(2): Treatment Cells
```

4.1 Perform rlog transformation for distances and PCA

```
# keep only genes with more than a single read
dds <- dds[ rowSums(counts(dds)) > 1,]

# perform rlog transformation for distances (for clustering) and PCA
rld<-rlog(dds)
```

```
dds <- dds[ rowSums(counts(dds)) > 1,]
nrow(dds)
```

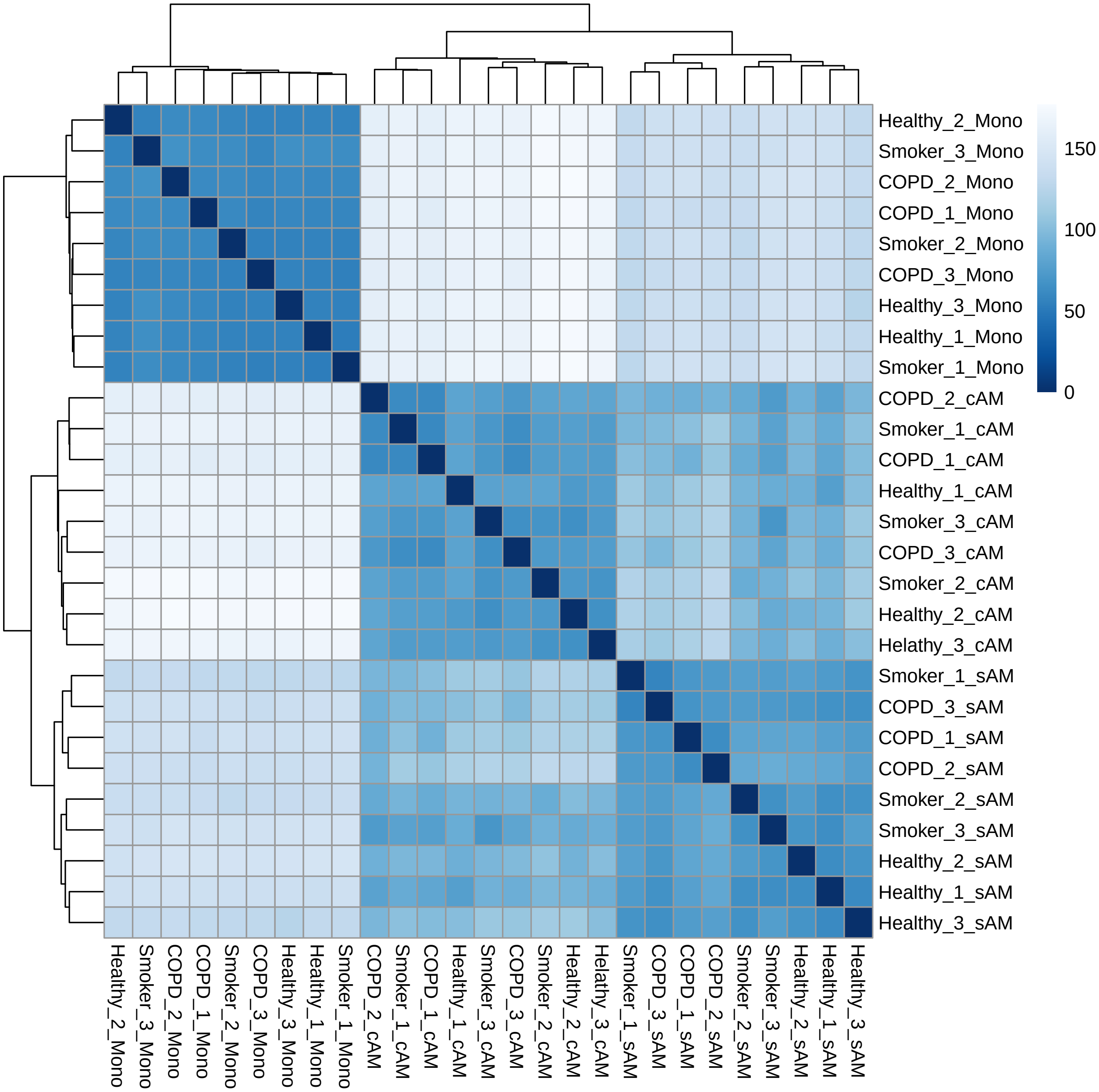
```
## [1] 27596
```

Calculate sample-to-sammple distances

```
sampleDists <- dist( t( assay(rld) ) )
sampleDistMatrix <- as.matrix( sampleDists )
```

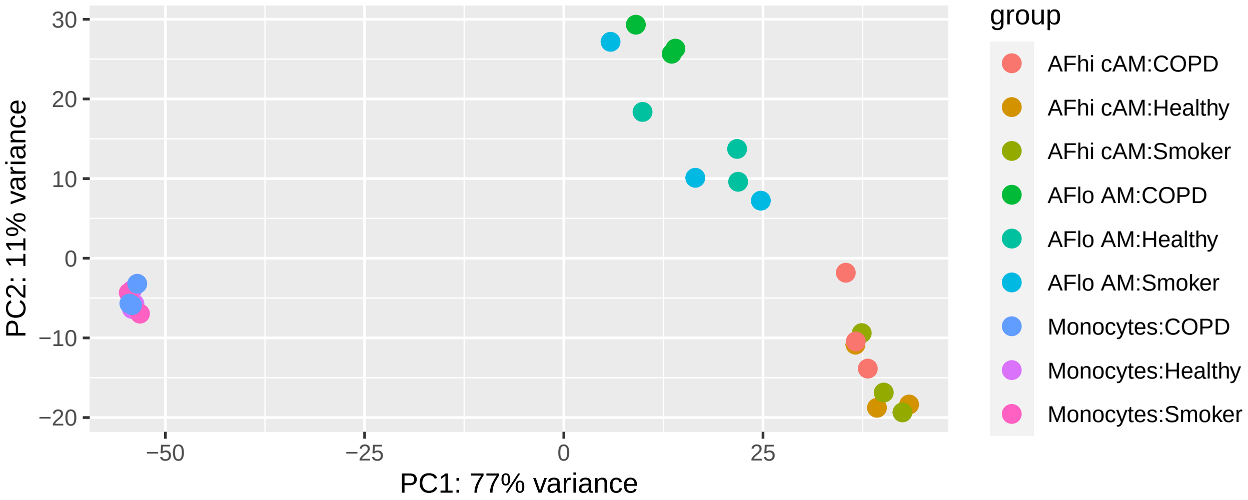
4.2 Heatmap

```
colors <- colorRampPalette( rev(brewer.pal(9, "Blues")) )(255)
heatmap <- pheatmap(sampleDistMatrix,
                    clustering_distance_rows=sampleDists,
                    clustering_distance_cols=sampleDists,
                    col=colors
)
```



4.3 PCA analysis

```
plotPCA <- plotPCA(rld, intgroup = c("Cells","Treatment"))
plotPCA
```



4.4 Differentially expressed (DE) genes in comparing AFlo vs AFhi alveolar macrophages

```
dds1 <- DESeq(dds)
res_AFlo_vs_AFhi<- results(dds1, contrast=c("Cells","AFlo␣AM","AFhi␣cAM"),
    lfcThreshold = 1, alpha = 0.05)
summary(res_AFlo_vs_AFhi)
```

```
##
## out of 27596 with nonzero total read count
## adjusted p-value < 0.05
## LFC > 1.00 (up)     : 438, 1.6%
## LFC < -1.00 (down) : 287, 1%
## outliers [1]        : 60, 0.22%
## low counts [2]      : 8025, 29%
## (mean count < 1)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

```
Res_AFlo_vs_AFhi_Shrunk <- lfcShrink(dds1, contrast=c("Cells","AFlo␣AM","
   AFhi␣cAM"), res=res_AFlo_vs_AFhi, type = "normal")

AFlo_vs_AFhi  <- merge(x=as.data.frame(res_AFlo_vs_AFhi), y=as.data.frame(
   Res_AFlo_vs_AFhi_Shrunk), by=c(0,1))

head(AFlo_vs_AFhi)
```

```
## # A tibble: 6 x 12
##   Row.names baseMean log2FoldChange.x lfcSE.x stat.x pvalue.x padj.x
##   <I<chr>>      <dbl>            <dbl>   <dbl>  <dbl>    <dbl>  <dbl>
## 1 A1BG           3.78            0.128   0.495  0        1        1
## 2 A1BG.AS1     169.             -0.104   0.125  0        1        1
## 3 A2M         3792.              0.159   0.316  0        1        1
## 4 A2M.AS1       40.2            -0.104   0.232  0        1        1
## 5 A3GALT2        1.01            1.49    1.12   0.441    0.659    1
## 6 A4GALT        68.6            -1.75    0.329 -2.29     0.0218   0.388
## # ... with 5 more variables: log2FoldChange.y <dbl>, lfcSE.y <dbl>,
## #   stat.y <dbl>, pvalue.y <dbl>, padj.y <dbl>
```

5 Export DE genes for other analyses

```
Genes2 <- AFlo_vs_AFhi$Row.names
head(Genes2, 3)
```

```
## [1] "A1BG"     "A1BG.AS1" "A2M"
```

```
rownames(AFlo_vs_AFhi) = make.names(Genes2, unique=TRUE)
AFlo_vs_AFhi<- AFlo_vs_AFhi[,-1]
```

Filter

```
AFlo_vs_AFhi <- AFlo_vs_AFhi[!is.na(AFlo_vs_AFhi$padj.y),]
AFlo_vs_AFhi_1 <- subset(AFlo_vs_AFhi, padj.y < 0.05)
dim(AFlo_vs_AFhi_1)
```

```
## [1] 725  11
```

```
AFlo_vs_AFhi_ordered <- AFlo_vs_AFhi_1[order(-AFlo_vs_AFhi_1$
   log2FoldChange.y) , ]
AFlo_vs_AFhi_ordered
```

```
## # A tibble: 725 x 11
##    baseMean log2FoldChange.x lfcSE.x stat.x  pvalue.x    padj.x
   log2FoldChange.y
##       <dbl>            <dbl>   <dbl>  <dbl>     <dbl>     <dbl>
             <dbl>
##  1   3873.              8.31   0.332  22.0  1.06e-107 2.07e-103
              8.13
##  2    351.              8.42   0.856   8.67 4.29e- 18 2.20e- 15
              7.55
##  3    324.              7.87   0.537  12.8  1.89e- 37 9.77e- 34
              7.50
##  4    299.              7.98   0.782   8.93 4.16e- 19 2.32e- 16
              7.27
##  5    659.              7.73   0.564  11.9  7.87e- 33 2.19e- 29
              7.24
##  6    391.              7.45   0.555  11.6  2.76e- 31 4.90e- 28
              7.20
##  7   2002.              7.71   0.576  11.6  2.63e- 31 4.90e- 28
              7.13
##  8    413.              8.04   0.752   9.36 8.16e- 21 5.49e- 18
              7.10
##  9     70.3             7.88   0.761   9.05 1.48e- 19 9.32e- 17
              7.05
## 10    423.              7.61   0.773   8.55 1.25e- 17 6.27e- 15
              6.99
## # ... with 715 more rows, and 4 more variables: lfcSE.y <dbl>, stat.y <
   dbl>,
## #   pvalue.y <dbl>, padj.y <dbl>
```

Save data for other analyses

```
write.table(as.data.frame(AFlo_vs_AFhi_ordered), "Results_Mreg_MA_LFC_9
   patients.txt", sep="\t", row.names=T,col.names=T)
```

6 Session information

```
sessionInfo()
```

```
## R version 4.0.3 (2020-10-10)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 20.04.3 LTS
##
## Matrix products: default
## BLAS:   /usr/lib/x86_64-linux-gnu/openblas-pthread/libblas.so.3
## LAPACK: /usr/lib/x86_64-linux-gnu/openblas-pthread/liblapack.so.3
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_GB.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_GB.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_GB.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_GB.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] parallel  stats4    stats     graphics  grDevices utils
   datasets
## [8] methods   base
##
## other attached packages:
##  [1] forcats_0.5.1               EnhancedVolcano_1.8.0
##  [3] ggrepel_0.9.1               RColorBrewer_1.1-2
##  [5] pheatmap_1.0.12             ggplot2_3.3.5
##  [7] DESeq2_1.30.1               SummarizedExperiment_1.20.0
##  [9] Biobase_2.50.0              MatrixGenerics_1.2.1
## [11] matrixStats_0.60.0          GenomicRanges_1.42.0
## [13] GenomeInfoDb_1.26.7         IRanges_2.24.1
## [15] S4Vectors_0.28.1            BiocGenerics_0.36.1
##
## loaded via a namespace (and not attached):
##  [1] bitops_1.0-7           bit64_4.0.5            ash_1.0-15
##  [4] httr_1.4.2             tools_4.0.3            utf8_1.2.2
##  [7] R6_2.5.0               KernSmooth_2.23-20     vipor_0.4.5
## [10] DBI_1.1.1              colorspace_2.0-2       withr_2.4.2
## [13] tidyselect_1.1.1       ggrastr_0.2.3          ggalt_0.4.0
## [16] bit_4.0.4              compiler_4.0.3         extrafontdb_1.0
## [19] cli_3.0.1              DelayedArray_0.16.3    labeling_0.4.2
## [22] scales_1.1.1           proj4_1.0-10.1         genefilter_1.72.1
## [25] stringr_1.4.0          digest_0.6.27          rmarkdown_2.9
## [28] XVector_0.30.0         pkgconfig_2.0.3        htmltools_0.5.1.1
## [31] extrafont_0.17         fastmap_1.1.0          highr_0.9
## [34] maps_3.3.0             rlang_0.4.11           rstudioapi_0.13
## [37] RSQLite_2.2.7          farver_2.1.0           generics_0.1.0
## [40] BiocParallel_1.24.1    dplyr_1.0.7            RCurl_1.98-1.3
## [43] magrittr_2.0.1         GenomeInfoDbData_1.2.4 Matrix_1.3-4
## [46] Rcpp_1.0.7             ggbeeswarm_0.6.0       munsell_0.5.0
## [49] fansi_0.5.0            lifecycle_1.0.0        stringi_1.7.3
## [52] yaml_2.2.1             MASS_7.3-53            zlibbioc_1.36.0
## [55] grid_4.0.3             blob_1.2.2             crayon_1.4.1
## [58] lattice_0.20-41        splines_4.0.3          annotate_1.68.0
## [61] locfit_1.5-9.4         knitr_1.33             pillar_1.6.2
## [64] geneplotter_1.68.0     XML_3.99-0.6           glue_1.4.2
## [67] evaluate_0.14          vctrs_0.3.8            Rttf2pt1_1.3.9
## [70] gtable_0.3.0           purrr_0.3.4            assertthat_0.2.1
## [73] cachem_1.0.5           xfun_0.24              xtable_1.8-4
## [76] survival_3.2-7         tibble_3.1.3           AnnotationDbi_1.52.0
## [79] beeswarm_0.4.0         memoise_2.0.0          ellipsis_0.3.2
```

References

1. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* 2014; 15: 550.