Differential expression of genes and modules

J. Shah chimeric mouse collaboration

Kim Dill-McFarland, kadm@uw.edu

version May 13, 2020

Contents

Background

The purpose of this workflow is to identify differentially expressed (DE) genes and modules.

Setup

Load packages

```
# Data manipulation and figures
library(tidyverse)
    # Multi-panel figures for ggplot
    library(cowplot)

#Define ggplot colors
logFC.cols <- c("Down, FDR < 0.5"="lightblue",
                "Down, FDR < 0.2"="blue",
```

```
                "Down, FDR < 0.05"="darkblue",
                "NS"="grey",
                "Up, FDR < 0.5"="pink",
                "Up, FDR < 0.2"="red",
                "Up, FDR < 0.05"="darkred")

#Linear models
library(limma)
#Construct networks to ID modules
library(WGCNA)
# Print tty table to knit file
library(knitr)
library(kableExtra)
options(knitr.kable.NA = '')
```

Set seed

```
set.seed(4389)
```

Scripts

```
source("https://raw.githubusercontent.com/kdillmcfarland/R_bioinformatic_scripts/master/RNAseq_module_f
source("https://raw.githubusercontent.com/kdillmcfarland/R_bioinformatic_scripts/master/limma.extract.p
source("scripts/RNAseq_boxplot_fxn.R")

`%notin%` <- Negate(`%in%`)
```

Set variable names and cutoffs for this workflow.

```
#Rdata file WITHIN project directory that holds cleaned data
data.file <- "data_clean/Shah.clean.RData"

#Prefix to give file names
basename <- "Shah_contrast"
#Define variable(s) of interest
#Used in PCA plots and to select significant genes to be used in module building
vars_of_interest <- c("status","cell")

#Maximum fdr for genes to be included in plots and modules
gene.fdr.cutoff <- 0.5
```

Load data

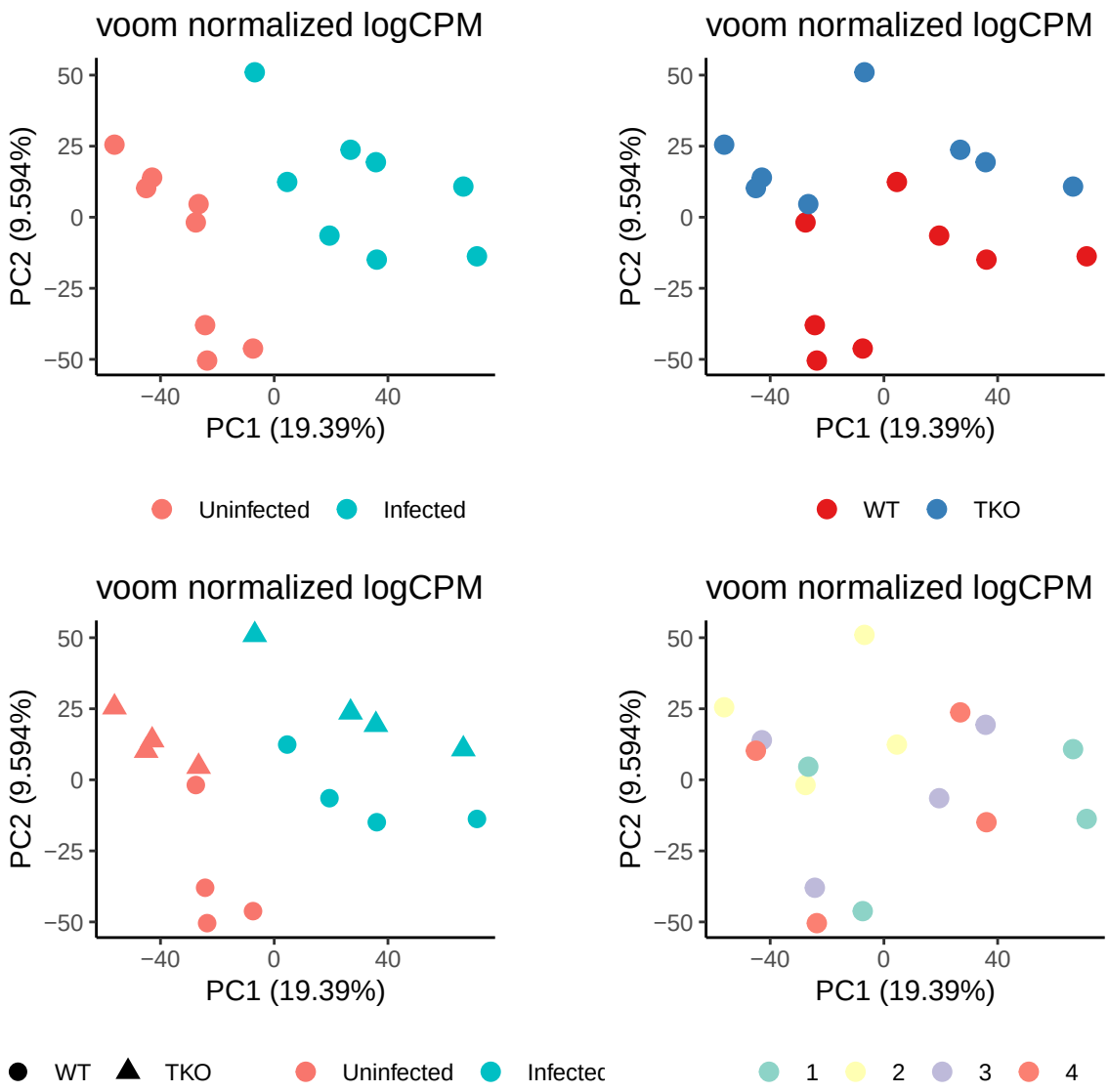
```
#Load data
load(data.file)
```

This includes in the following samples.

status	cell	n
Uninfected	WT	4
Uninfected	TKO	4
Infected	WT	4
Infected	TKO	4

Data exploration

PCA (genes)



Define significant genes

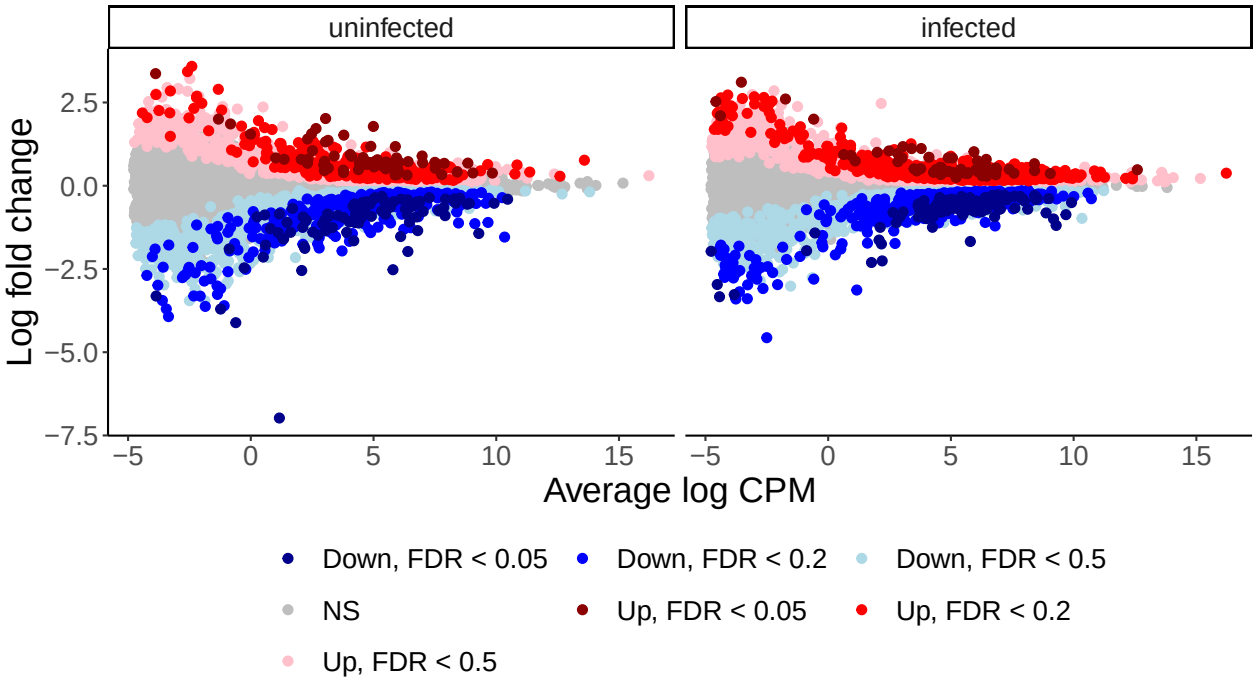
Linear model

```
# Define model
model <- model.matrix(~0 + status:cell, data=dat.voom$targets)
    colnames(model) <- c("Uninfected_WT", "Infected_WT",
                         "Uninfected_TKO", "Infected_TKO")
# Define contrasts of interest
contrasts <- makeContrasts(
  uninfected = Uninfected_TKO-Uninfected_WT,
  infected = Infected_TKO-Infected_WT,
  levels=model)
```

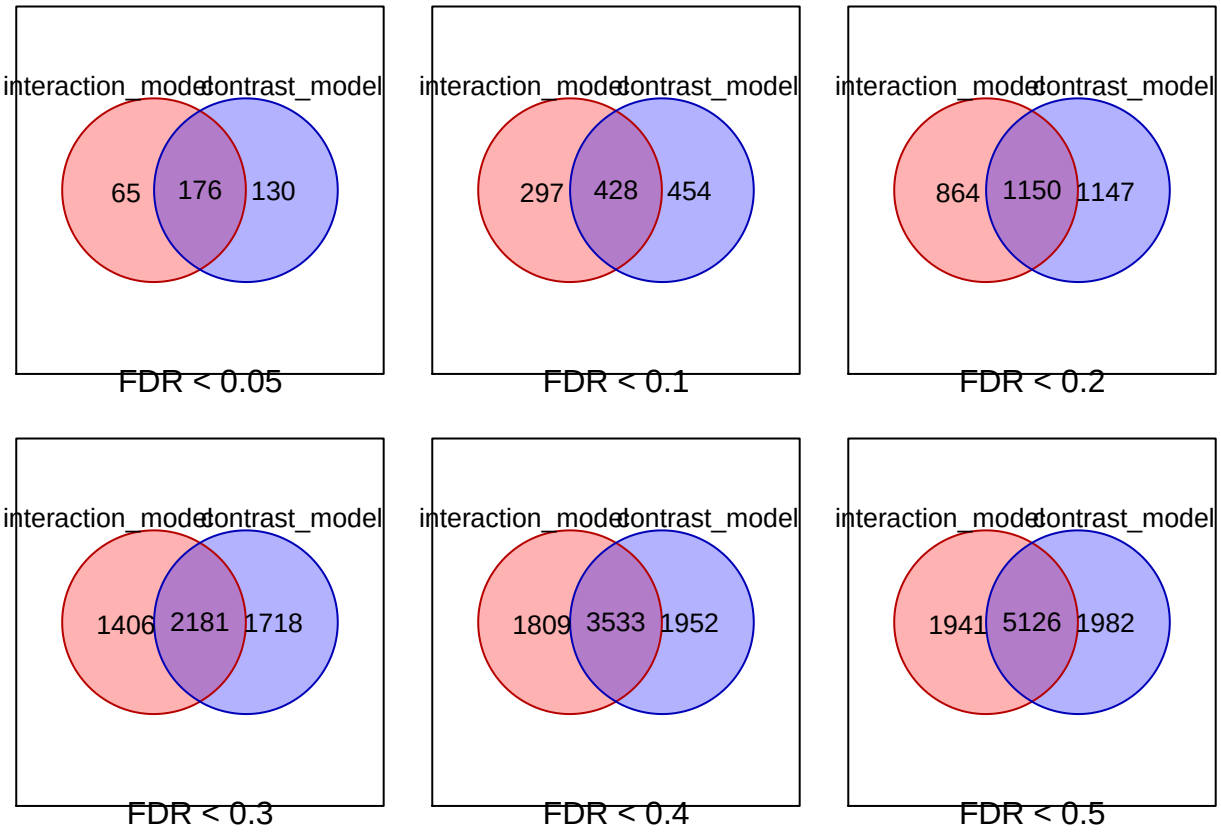
```
#Fit model to transformed count data. Calculate eBayes
  efitQW <- eBayes(
            contrasts.fit(contrasts=contrasts,
            lmFit(dat.voom$E, model)))
```

Summarize gene model

		Genes with FDR <					
Variable	Comparison	0.05	0.1	0.2	0.3	0.4	0.5
uninfected	TKO-WT	157	346	768	1391	2220	3380
infected	TKO-WT	194	645	1817	3084	4292	5391
total (nonredundant)	TKO-WT	306	882	2297	3899	5485	7108



Compare significant genes



FDR	model	genes not signif in other model	variable(s) those gene are signif for in this model
0.05	interaction	65	status
0.05	contrast	130	infected
0.10	interaction	297	status
0.10	contrast	454	infected
0.20	interaction	864	status
0.20	contrast	1147	infected
0.30	interaction	1406	status
0.30	contrast	1718	infected
0.40	interaction	1809	status
0.40	contrast	1952	infected
0.50	interaction	1941	status
0.50	contrast	1982	infected

Roughly half of genes identified as significant in either model are also significant in the other model. Genes unique to the interaction model are only significant for infection `status`. Thus, the contrasts model appears to be successfully removing much of the infection-only signal while retaining all `cell` type and `status:cell` interaction signals.

Interestingly the genes unique to the contrasts model were only significant for cell type within `infected` cells. This may indicate that the contrasts model is better resolving signal from the interaction term `status:cell` since the interaction model defines the interaction from the intercept of `status` = uninfected and `cell` = WT.

Overall, these results indicate that the contrasts model better resolves the signals of interest.

Gene plots

Save in `figs/gene_level_contrast`

Modules: Status and cell

Define customizations for module building.

```
#Set FDR cutoff for gene inclusion in modules
mod.fdr.cutoff <- 0.3
#List variables from which significant genes will be extracted
vars_for_mods <- c("uninfected","infected")
```

In total, 3899 of 14215 genes that significantly differed (FDR $\leq$ 0.3) by one or more variables of interest will be incorporated into gene modules.

```
make.modules(voom.dat = dat.voom,
             genes.signif = genes.signif,
             Rsq.min = 0.8,
             minModuleSize = 50,
             deepSplit = 3,
             nThread = 4,
             basename = basename)
```

```
## Allowing multi-threading with up to 4 threads.
## pickSoftThreshold: will use block size 3899.
##  pickSoftThreshold: calculating connectivity for given powers...
##    ..working on genes 1 through 3899 of 3899
##    Power SFT.R.sq  slope truncated.R.sq mean.k. median.k. max.k.
## 1      1   0.0408 10.500          0.981 1950.00   1950.00 2020.0
## 2      2   0.0484  2.000          0.926 1170.00   1180.00 1350.0
## 3      3   0.0155 -0.520          0.882  785.00    788.00 1060.0
## 4      4   0.0445 -0.601          0.879  560.00    561.00  869.0
## 5      5   0.1020 -0.721          0.892  417.00    414.00  732.0
## 6      6   0.1910 -0.865          0.901  319.00    314.00  627.0
## 7      7   0.2810 -0.947          0.926  250.00    244.00  543.0
## 8      8   0.3660 -1.050          0.939  199.00    191.00  474.0
## 9      9   0.4550 -1.120          0.954  161.00    152.00  417.0
## 10    10   0.5230 -1.190          0.962  132.00    122.00  369.0
## 11    11   0.5790 -1.260          0.969  110.00     99.70  328.0
## 12    12   0.6240 -1.290          0.978   91.60     81.70  293.0
## 13    13   0.6640 -1.320          0.984   77.30     67.40  263.0
## 14    14   0.6960 -1.370          0.986   65.70     56.00  237.0
## 15    15   0.7150 -1.410          0.989   56.20     47.10  216.0
## 16    16   0.7360 -1.440          0.991   48.40     39.60  196.0
## 17    17   0.7580 -1.480          0.991   42.00     33.60  179.0
## 18    18   0.7650 -1.530          0.988   36.50     28.40  164.0
## 19    19   0.7800 -1.560          0.987   32.00     24.30  151.0
## 20    20   0.7890 -1.600          0.990   28.10     20.80  139.0
## 21    21   0.8020 -1.610          0.992   24.80     18.00  128.0
## 22    22   0.8020 -1.650          0.989   22.00     15.40  119.0
## 23    23   0.8130 -1.660          0.991   19.60     13.40  110.0
## 24    24   0.8210 -1.670          0.993   17.50     11.60  102.0
## 25    25   0.8290 -1.680          0.994   15.60     10.10   95.0
## 26    26   0.8290 -1.700          0.993   14.00      8.81   88.5
## 27    27   0.8290 -1.720          0.992   12.60      7.74   82.5
```

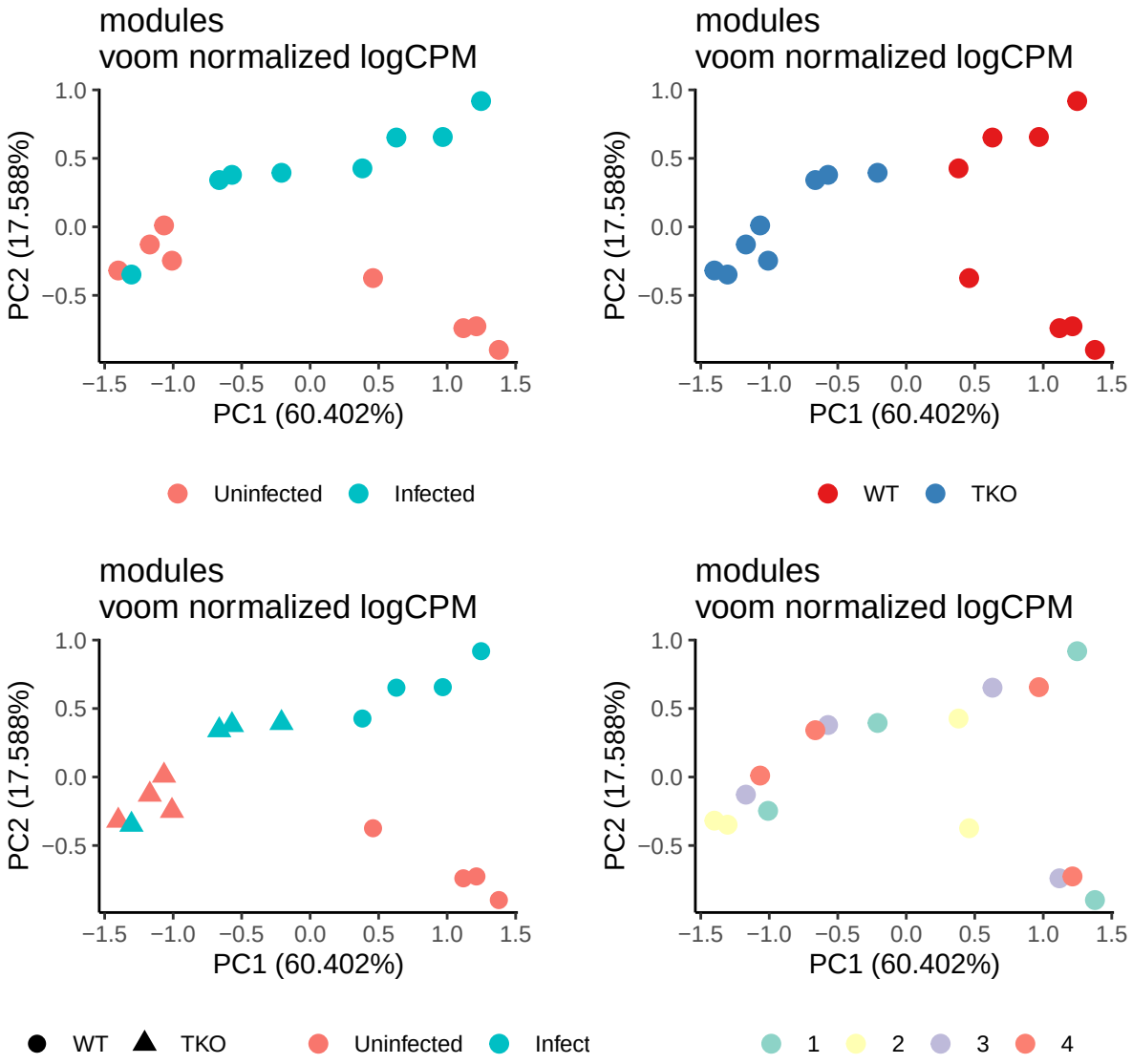
```
## 28    28   0.8350 -1.730          0.993   11.40      6.78   77.1
## 29    29   0.8410 -1.730          0.994   10.30      5.97   72.1
## 30    30   0.8440 -1.720          0.995    9.38      5.28   67.5
```

A power threshold of 21 was used as it achieves high R^2 and sufficient mean connectivity ($R^2 = 0.8015016$, mean k = 24.8126884).

Module	Total genes
00	282
01	583
02	463
03	456
04	314
05	276
06	242
07	219
08	200
09	164
10	151
11	127
12	93
13	82
14	73
15	62
16	58
17	54

This created 17 modules plus 282 (7.2326237%) genes not grouped into any module (*e.g.* in module 0).

PCA (modules)



Linear model

```
# Fit model to transformed count data. Calculate eBayes
  efitQW.mods <- eBayes(
            contrasts.fit(contrasts=contrasts,
            lmFit(voom.mods, model)))
```

Summarize module model

		Modules with FDR <					
Variable	Comparison	0.05	0.1	0.2	0.3	0.4	0.5
uninfected	TKO-WT	13	14	14	17	17	17
infected	TKO-WT	14	15	16	16	17	18
total (nonredundant)	TKO-WT	18	18	18	18	18	18

module	uninfected direction in TKO	infected direction in TKO
Significant in uninfected		
module_Shah_contrast_06	down	
module_Shah_contrast_12	down	
module_Shah_contrast_13	up	
module_Shah_contrast_15	up	
Significant in infected		
module_Shah_contrast_04		down
module_Shah_contrast_08		down
module_Shah_contrast_01		up
module_Shah_contrast_03		up
Significant in both		
module_Shah_contrast_05	down	down
module_Shah_contrast_11	down	down
module_Shah_contrast_17	down	down
module_Shah_contrast_10	down	down
module_Shah_contrast_02	down	down
module_Shah_contrast_14	up	up
module_Shah_contrast_07	up	up
module_Shah_contrast_16	up	up
module_Shah_contrast_09	up	up

Directions are relative to WT. For example, ***down*** *means the module has* ***lower*** *expression in TKO relative to WT.*

Module plots

Create expression plots of modules, and save in `figs/module*`

R session

```
sessionInfo()
```

```
## R version 3.6.1 (2019-07-05)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS Catalina 10.15.4
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/3.6/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  stats     graphics  grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
##  [1] doParallel_1.0.15     iterators_1.0.12      foreach_1.5.0
##  [4] venn_1.9              kableExtra_1.1.0      knitr_1.28
##  [7] WGCNA_1.69            fastcluster_1.1.25    dynamicTreeCut_1.63-1
```

```
## [10] limma_3.40.6          cowplot_1.0.0         forcats_0.5.0
## [13] stringr_1.4.0         dplyr_0.8.5           purrr_0.3.4
## [16] readr_1.3.1           tidyr_1.0.3           tibble_3.0.1
## [19] ggplot2_3.3.0         tidyverse_1.3.0
##
## loaded via a namespace (and not attached):
##  [1] nlme_3.1-147           matrixStats_0.56.0    fs_1.4.1
##  [4] lubridate_1.7.8        bit64_0.9-7           webshot_0.5.2
##  [7] RColorBrewer_1.1-2     httr_1.4.1            tools_3.6.1
## [10] backports_1.1.6        R6_2.4.1              rpart_4.1-15
## [13] Hmisc_4.4-0            DBI_1.1.0             BiocGenerics_0.30.0
## [16] colorspace_1.4-1       nnet_7.3-14           withr_2.2.0
## [19] gridExtra_2.3          tidyselect_1.0.0      preprocessCore_1.46.0
## [22] bit_1.1-15.2           compiler_3.6.1        cli_2.0.2
## [25] rvest_0.3.5            Biobase_2.44.0        htmlTable_1.13.3
## [28] xml2_1.3.2             labeling_0.3          checkmate_2.0.0
## [31] scales_1.1.0           digest_0.6.25         foreign_0.8-76
## [34] rmarkdown_2.1          base64enc_0.1-3       jpeg_0.1-8.1
## [37] pkgconfig_2.0.3        htmltools_0.4.0       dbplyr_1.4.3
## [40] htmlwidgets_1.5.1      rlang_0.4.6           readxl_1.3.1
## [43] impute_1.58.0          rstudioapi_0.11       RSQLite_2.2.0
## [46] farver_2.0.3           generics_0.0.2        jsonlite_1.6.1
## [49] acepack_1.4.1          magrittr_1.5          GO.db_3.8.2
## [52] Formula_1.2-3          Matrix_1.2-18         Rcpp_1.0.4.6
## [55] munsell_0.5.0          S4Vectors_0.22.1      fansi_0.4.1
## [58] lifecycle_0.2.0        stringi_1.4.6         yaml_2.2.1
## [61] grid_3.6.1             blob_1.2.1            crayon_1.3.4
## [64] lattice_0.20-41        haven_2.2.0           splines_3.6.1
## [67] hms_0.5.3              pillar_1.4.4          admisc_0.8
## [70] codetools_0.2-16       stats4_3.6.1          reprex_0.3.0
## [73] glue_1.4.0             evaluate_0.14         latticeExtra_0.6-29
## [76] data.table_1.12.8      modelr_0.1.7          vctrs_0.2.4
## [79] png_0.1-7              cellranger_1.1.0      gtable_0.3.0
## [82] assertthat_0.2.1       xfun_0.13             broom_0.5.6
## [85] viridisLite_0.3.0      survival_3.1-12       AnnotationDbi_1.46.1
## [88] memoise_1.1.0          IRanges_2.18.3        cluster_2.1.0
## [91] ellipsis_0.3.0
```