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)
```

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")
```

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"
#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")

#Gene model for use in limma
#Model MUST have intercept
model_gene <- as.formula("~ status*cell")
#Names for variables in model
#Recommend exactly match variable names in model. Have not tested other values
model_gene_names <- c("status","cell","status:cell")
#Maximum fdr for genes to be included in plots and modules
gene.fdr.cutoff <- 0.5
```

Load data

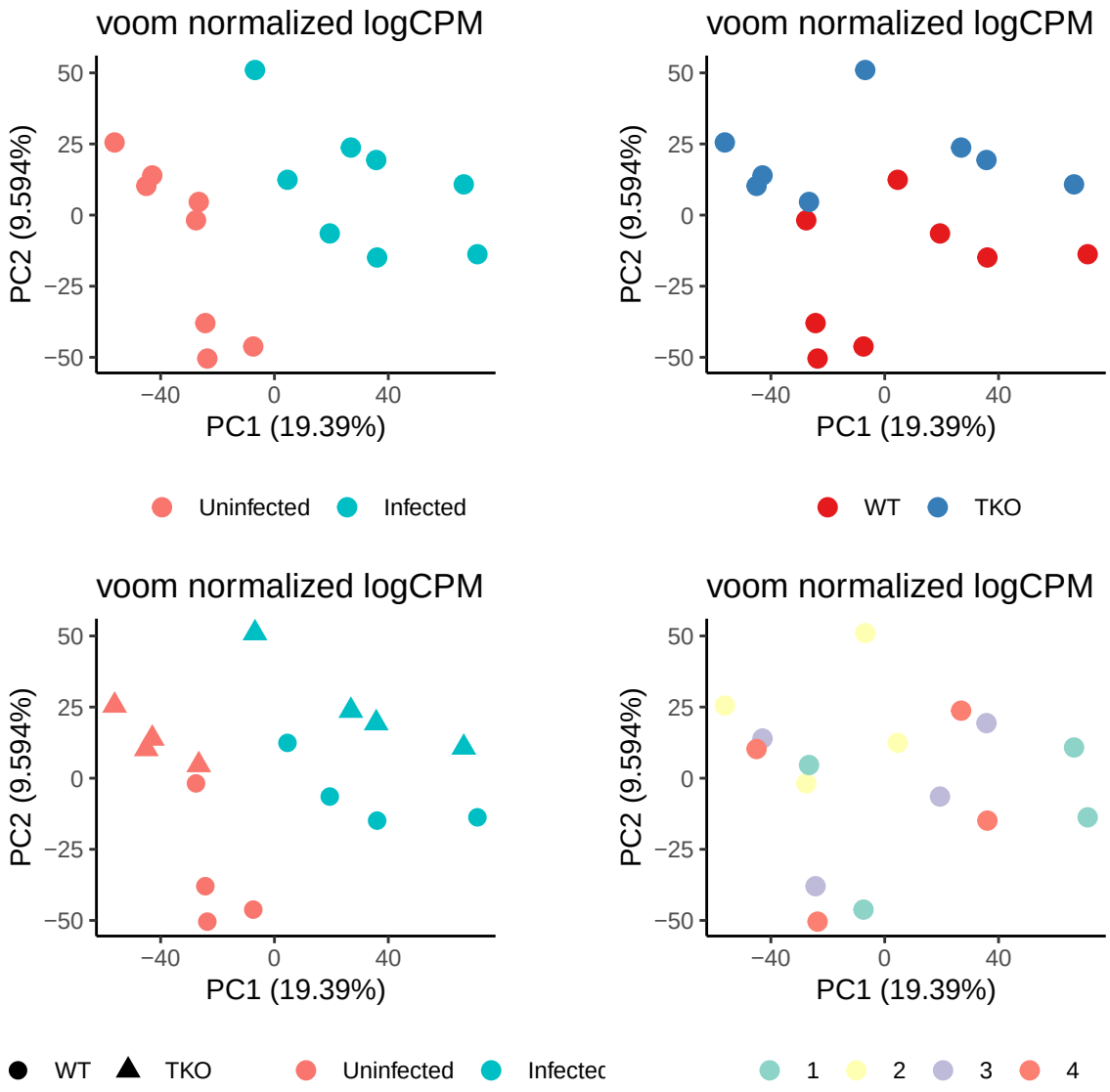
```
#Load data
load(data.file)
```

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

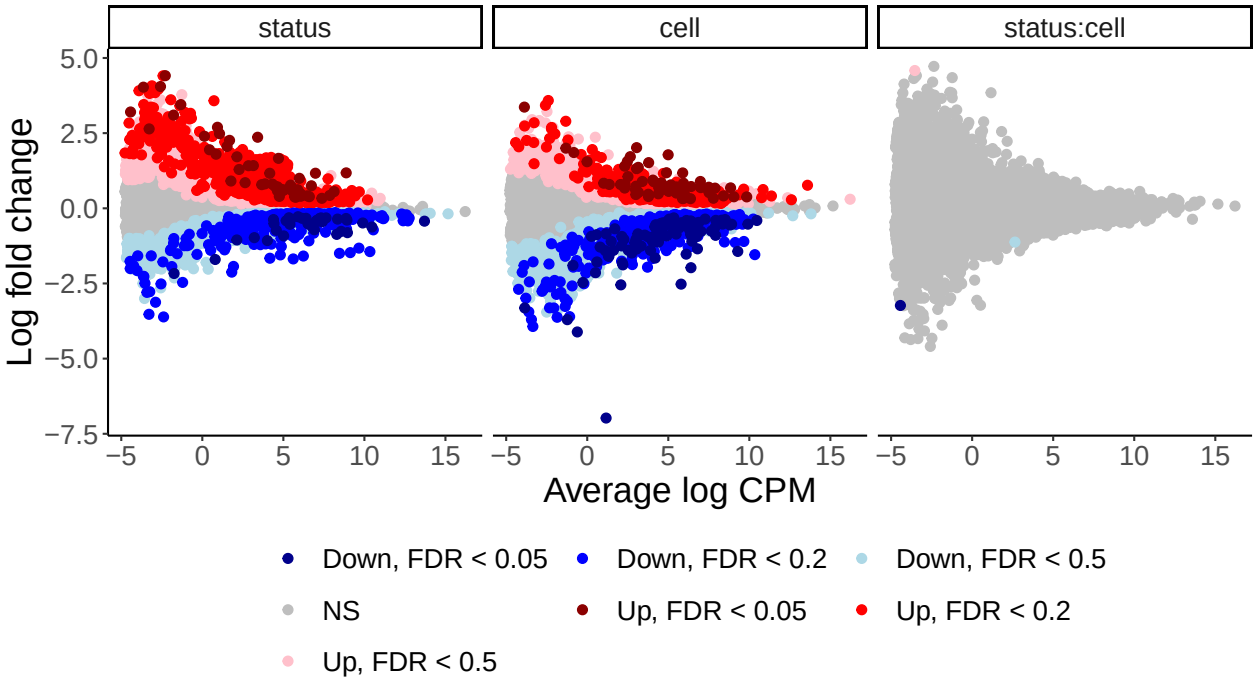
Linear model

```
# Define model
model <- model.matrix(model_gene, data=dat.voom$targets)
  colnames(model) <- c("(Intercept)", model_gene_names)
```

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

Summarize gene model

	Genes with FDR <					
Variable	0.05	0.1	0.2	0.3	0.4	0.5
status	90	435	1427	2591	3898	5111
cell	157	346	768	1391	2220	3380
status:cell	1	1	1	3	3	3
total (nonredundant)	241	725	2014	3587	5342	7067



Highlight genes most significant for interaction term.

geneName	logFC	AveExpr	adj.P.Val	group
ENSMUSG00000020473	-3.231882	-4.426071	0.0201904	status:cell
ENSMUSG00000030431	-1.124078	2.641619	0.2886665	status:cell
ENSMUSG00000024935	4.583051	-3.537011	0.2886665	status:cell

Gene plots

Create expression plots of genes with at least one variable FDR < 0.5. Save in `figs/gene_level`

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("status","cell","status:cell")
#Module model for use in limma
#Model MUST have intercept
model_mod <- model_gene
#Names for variables in model
#Recommend exactly match variable names in model. Have not tested other values
model_mod_names <- model_gene_names
```

In total, 3587 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 3587.
##  pickSoftThreshold: calculating connectivity for given powers...
##    ..working on genes 1 through 3587 of 3587
##    Power SFT.R.sq   slope truncated.R.sq mean.k. median.k. max.k.
## 1      1 2.08e-02  2.5800          0.771 1800.00   1800.00 1900.0
## 2      2 1.18e-02 -1.1600          0.965 1070.00   1070.00 1260.0
## 3      3 4.56e-05 -0.0423          0.970  700.00    701.00  936.0
## 4      4 1.54e-03 -0.1650          0.951  490.00    489.00  734.0
## 5      5 9.97e-03 -0.3030          0.931  359.00    357.00  593.0
## 6      6 2.04e-02 -0.3500          0.911  272.00    268.00  487.0
## 7      7 4.09e-02 -0.3870          0.924  211.00    206.00  406.0
## 8      8 8.27e-02 -0.4470          0.945  167.00    162.00  343.0
## 9      9 1.89e-01 -0.6450          0.950  135.00    128.00  299.0
## 10    10 3.00e-01 -0.8140          0.960  110.00    103.00  265.0
## 11    11 4.24e-01 -0.9920          0.961   90.90     83.70  237.0
## 12    12 5.07e-01 -1.1000          0.971   75.90     68.90  214.0
## 13    13 5.76e-01 -1.1700          0.985   64.10     57.20  194.0
## 14    14 6.45e-01 -1.3000          0.990   54.50     47.70  179.0
## 15    15 6.96e-01 -1.3900          0.989   46.80     40.20  166.0
## 16    16 7.39e-01 -1.4700          0.988   40.40     34.00  155.0
## 17    17 7.73e-01 -1.5200          0.995   35.10     28.90  145.0
## 18    18 7.98e-01 -1.5700          0.990   30.70     24.60  136.0
## 19    19 8.25e-01 -1.6100          0.989   27.00     21.10  128.0
## 20    20 8.46e-01 -1.6500          0.988   23.80     18.10  121.0
## 21    21 8.57e-01 -1.6800          0.977   21.10     15.60  114.0
## 22    22 8.73e-01 -1.7000          0.976   18.80     13.50  108.0
## 23    23 8.86e-01 -1.7200          0.978   16.90     11.70  103.0
## 24    24 8.96e-01 -1.7400          0.976   15.20     10.20   98.3
## 25    25 9.06e-01 -1.7500          0.973   13.70      8.90   93.8
## 26    26 9.18e-01 -1.7600          0.973   12.40      7.81   89.7
## 27    27 9.25e-01 -1.7500          0.969   11.20      6.90   85.9
## 28    28 9.30e-01 -1.7500          0.966   10.20      6.08   82.4
```

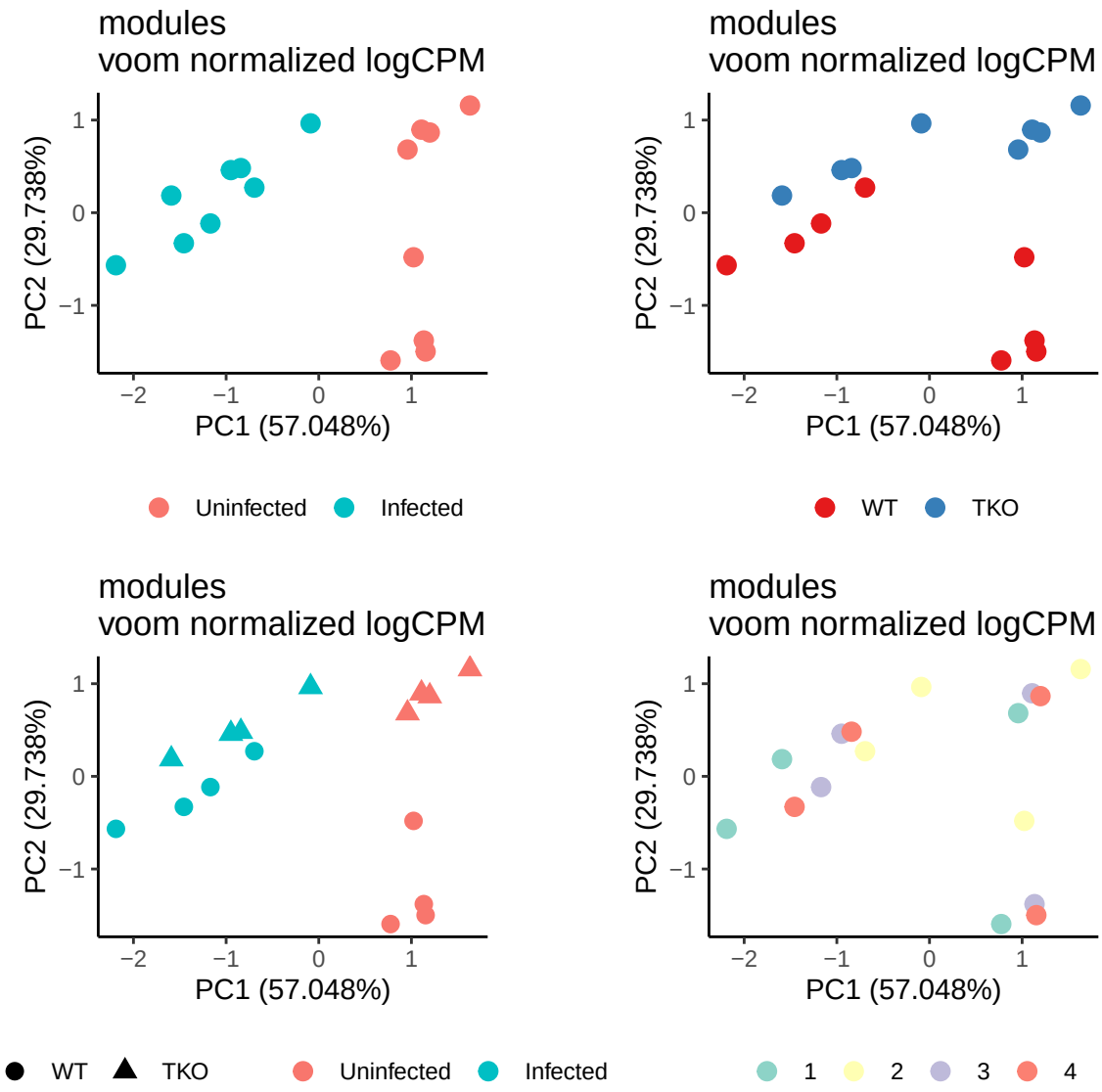
```
## 29    29 9.36e-01 -1.7600          0.966    9.31      5.37   79.3
## 30    30 9.40e-01 -1.7600          0.964    8.52      4.73   76.5
```

A power threshold of 19 was used as it achieves high R^2 and sufficient mean connectivity ($R^2 = 0.825248$, mean k = 26.9721425).

Module	Total genes
00	43
01	438
02	349
03	259
04	233
05	229
06	228
07	181
08	177
09	173
10	172
11	166
12	161
13	150
14	145
15	136
16	133
17	121
18	93

This created 18 modules plus 43 (1.1987733%) genes not grouped into any module (*e.g.* in module 0).

PCA (modules)



Linear model

```
# Define model
model_2 <- model.matrix(model_mod, data=dat.voom$targets)
    colnames(model_2) <- c("(Intercept)",model_mod_names)

# Fit model to transformed count data. Calculate eBayes
  efitQW.mods <- eBayes(
            lmFit(voom.mods, model_2))
```

Summarize module model

	Modules with FDR <					
Variable	0.05	0.1	0.2	0.3	0.4	0.5
status	14	14	15	16	16	17
cell	12	12	13	14	15	15
status:cell	8	14	15	15	16	17
total (nonredundant)	19	19	19	19	19	19

Module plots

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

Annotate results

Add MGI symbols to results files

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] kableExtra_1.1.0      knitr_1.28            WGCNA_1.69
##  [7] fastcluster_1.1.25    dynamicTreeCut_1.63-1 limma_3.40.6
## [10] cowplot_1.0.0         forcats_0.5.0         stringr_1.4.0
## [13] dplyr_0.8.5           purrr_0.3.4           readr_1.3.1
## [16] tidyr_1.0.3           tibble_3.0.1          ggplot2_3.3.0
## [19] 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          codetools_0.2-16
## [70] stats4_3.6.1          reprex_0.3.0          glue_1.4.0
## [73] evaluate_0.14         latticeExtra_0.6-29   data.table_1.12.8
## [76] modelr_0.1.7          vctrs_0.2.4           png_0.1-7
## [79] cellranger_1.1.0      gtable_0.3.0          assertthat_0.2.1
## [82] xfun_0.13             broom_0.5.6           viridisLite_0.3.0
## [85] survival_3.1-12       AnnotationDbi_1.46.1  memoise_1.1.0
## [88] IRanges_2.18.3        cluster_2.1.0         ellipsis_0.3.0
```
