P259: Model selection

Dendritic cells (pDC)

Kim Dill-McFarland, kadm@uw.edu

version May 18, 2021

Contents

Background

The purpose of this workflow is to identify differentially expressed genes in pDC. Several models are compared in order to determine the best way to extract the effects of Anti-IL5 therapy and eosinophil (EOS) supernatant over the much larger human rhinovirus (RV) signal.

Setup

Load packages

```r
# Data manipulation and figures
library(tidyverse)
    # Multi-panel figures for ggplot
    library(cowplot)
    # Add p-values to plots
    library(ggpubr)
#Venn diagrams
library(venn)
#Linear models
library(limma)
#Define ggplot colors
group.cols <- c("none:none"="#dadaeb",
                "none:AntiIL5"="#9e9ac8",
                "none:EOS.supp"="#54278f",
                "HRV:none"="#c7e9c0",
                "HRV:AntiIL5"="#74c476",
                "HRV:EOS.supp"="#006d2c",
                "flu:none"="#fdae6b",
                "flu:AntiIL5"="#e6550d")
samp.cols <- c("AC1"="#969696",
               "AC2"="#a6cee3",
               "AC3"="#1f78b4",
               "AC4"="#b2df8a",
               "AC5"="#33a02c",
               "AT1"="#fb9a99",
               "AT2"="#e31a1c",
               "AT3"="#fdbf6f",
               "AT4"="#ff7f00",
               "donor1"="#cab2d6",
               "donor2"="#6a3d9a",
               "donor3"="#ffff99",
               "donor4"="#b15928")

# Print tty table to knit file
library(knitr)
library(kableExtra)
```

Set seed

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

Scripts and functions

```r
#To extract and format p-value from lmFit
source("https://raw.githubusercontent.com/kdillmcfarland/R_bioinformatic_scripts/master/limma.extract.p
#opposite of %in%
`%notin%` <- Negate(`%in%`)
```

Load data

Load normalized edgeR data objects. RNA-seq data was cleaned and normalized in data cleaning. This included filtering for median CV coverage, mapped duplicate reads, total aligned counts, PCA outliers, and

experiment	asthma	IL5	virus	n
P259_1	healthy	none	none	4
		none	HRV	4
		EOS.supp	none	3
		EOS.supp	HRV	4
P259_2	asthma	none	none	3
		none	HRV	6
		AntiIL5	none	4
		AntiIL5	HRV	8

rare genes.

```
load("data_clean/P259_pDC_clean.RData")
```

This includes in the following pDC samples.

Variables of interest

Main effects

- `donorID`: character for each donor (N = 11)
- `IL5`: factor of patient AntiIL5 therapy or sample EOS supernatant treatment (none, AntiIL5, EOS.supp)
- `virus`: factor of viral conditions (none, HRV)

Fixed effects

- `Sex`: NA, Male, Female
- `Age`: age, years (11 - 32)
  - Neither available for 2 of 4 donors in P259.1

Explore fixed effects and possible confounding variables
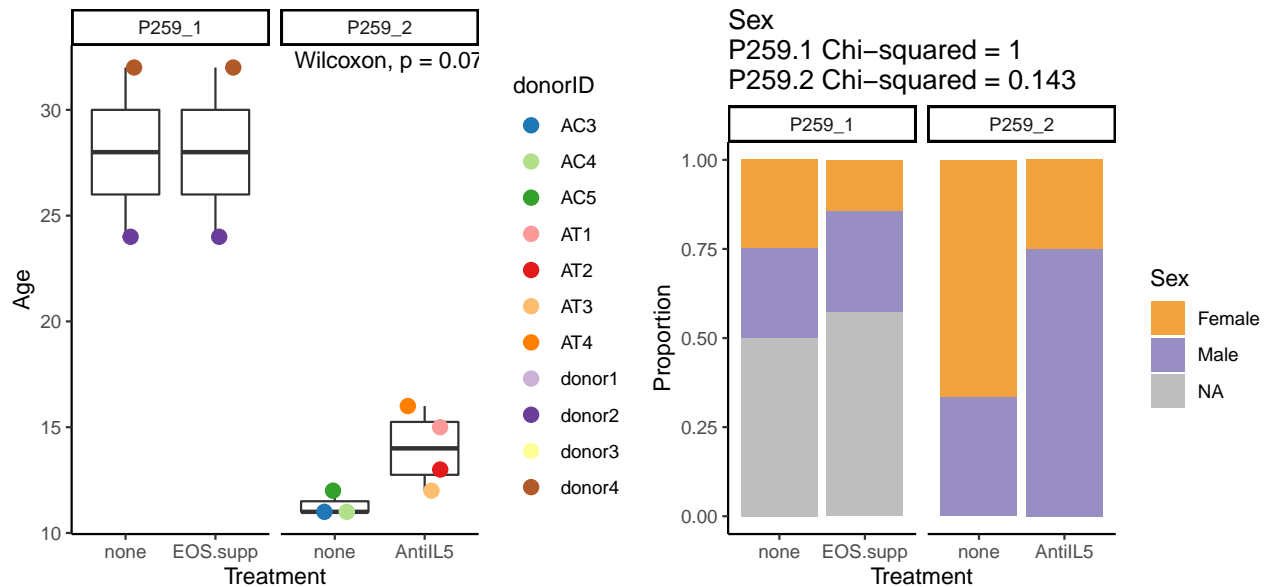
Age and sex appear to be different between treatment groups but are technically not significantly different, likely as the result of the small sample size. Due to this, the small sample size, and missing data in P259.1, sex and age will not be used as co-variates.

PCA
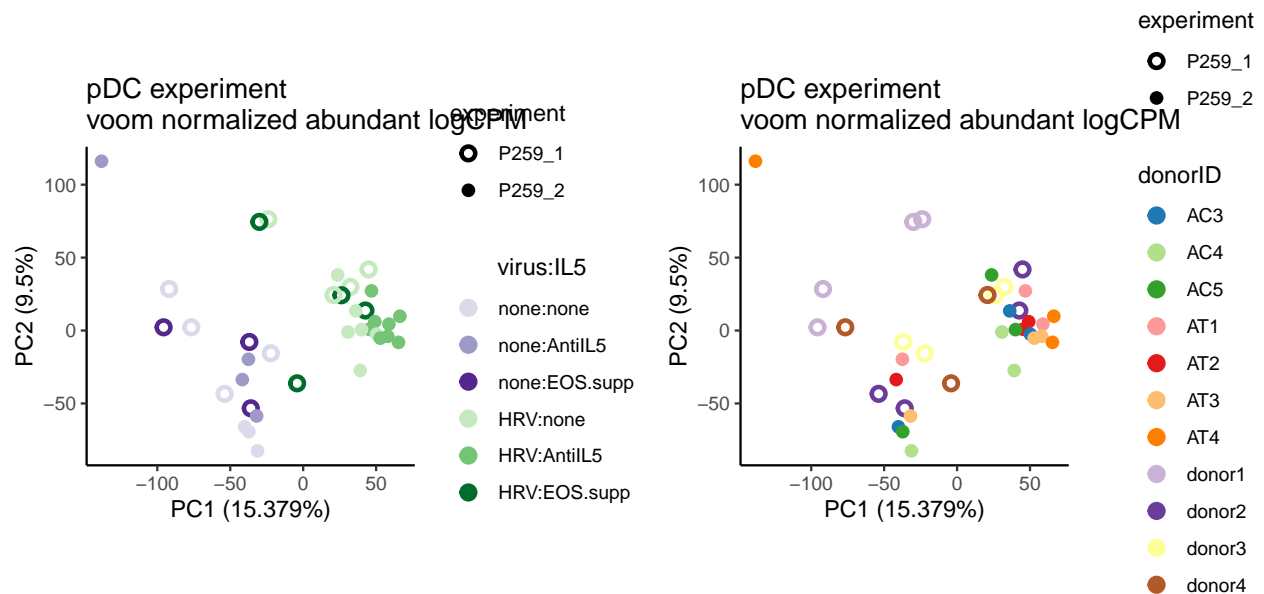
P259.1 model comparison

Run models

Interaction model

gene expression ~ EOS.supp * virus + (1|donor)

```r
# Define model
model_1.interact <- model.matrix(~ IL5*virus,
                          data=dat.pDC.voom_1$targets)
    colnames(model_1.interact) <- c("(Intercept)", "EOS.supp",
                              "HRV", "EOS.supp:HRV")
#Block by donor
consensus.corr <- duplicateCorrelation(
  dat.pDC.voom_1$E,
  model_1.interact,
  block=dat.pDC.voom_1$targets$donorID)$consensus.correlation

consensus.corr
```

```
## [1] 0.265985
```

```r
# Fit model to transformed count data. Calculate eBayes
efitQW_1.interact <- eBayes(
              lmFit(dat.pDC.voom_1$E,
                  model_1.interact,
                  block=dat.pDC.voom_1$targets$donorID,
                  correlation=consensus.corr))
```

Summary

	Total genes at FDR					
Variable	<0.05	<0.1	<0.2	<0.3	<0.4	<0.5
EOS.supp	2	2	5	10	10	10
HRV	1011	1697	2937	4012	5075	6290
EOS.supp:HRV	0	0	0	3	3	3
total (nonredundant)	1011	1697	2939	4013	5076	6291

Contrasts model

```
gene expression ~ EOS.supp:virus + (1|donor)
# Define model
dat.pDC.voom_1$targets <- dat.pDC.voom_1$targets %>%
  mutate(contrast=paste(IL5, virus, sep="_")) %>%
  mutate(contrast=factor(contrast,
                  levels = c("none_none","none_HRV",
                          "EOS.supp_none","EOS.supp_HRV")))

model_1.contrast<- model.matrix(~ 0 + contrast,
                      data=dat.pDC.voom_1$targets)
    colnames(model_1.contrast) <- c(
                          "none_none","none_HRV",
                          "EOS.supp_none","EOS.supp_HRV")
#Block by donor
consensus.corr <- duplicateCorrelation(
  dat.pDC.voom_1$E,
  model_1.contrast,
  block=dat.pDC.voom_1$targets$donorID)$consensus.correlation

consensus.corr
```

```
## [1] 0.265985
```

```
#Fit model
fitQW_1.contrast <- lmFit(dat.pDC.voom_1$E,
                  model_1.contrast,
                  block=dat.pDC.voom_1$targets$donorID,
                  correlation=consensus.corr)
#Get contrasts
contrast.matrix <- makeContrasts(
  none_HRV-none_none,
  EOS.supp_HRV-EOS.supp_none,
  EOS.supp_none-none_none,
  EOS.supp_HRV-none_HRV,
  levels=model_1.contrast)

efitQW_1.contrast <- eBayes(contrasts.fit(fitQW_1.contrast,
                          contrast.matrix))
```

Summary

Variable	Total genes at FDR					
	<0.05	<0.1	<0.2	<0.3	<0.4	<0.5
none_HRV - none_none	1011	1697	2937	4012	5075	6290
EOS.supp_HRV - EOS.supp_none	79	394	1050	1817	2739	3838
EOS.supp_none - none_none	2	2	5	10	10	10
EOS.supp_HRV - none_HRV	0	0	0	0	0	0
total (nonredundant)	1027	1781	3206	4522	5875	7427

Contrast selected genes Genes of interest are those that change with virus (*e.g.* significant for virus in untreated and/or treated samples) AND with EOS supernatant in media OR HRV.
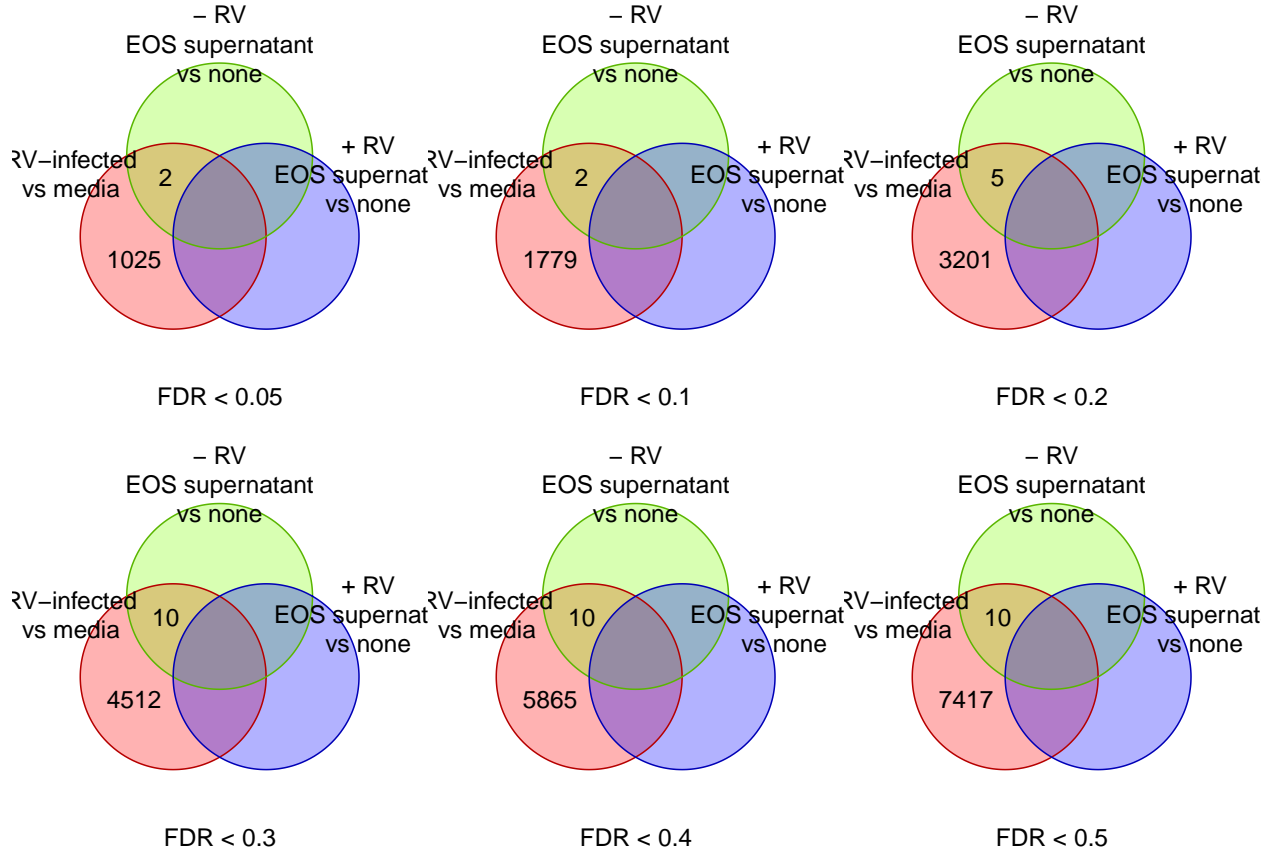
Delta model

`RV - media gene expression ~ EOS.supp + (1|donor)`

Calculate HRV - media counts.

Fit model

```
# Define model
model_1.delta <- model.matrix(~ IL5,
                    data=dat.pDC.voom_1.delta.tgts)
    colnames(model_1.delta) <- c("(Intercept)","EOS.supp")

#Block by donor
consensus.corr <- duplicateCorrelation(
  dat.pDC.voom_1.delta.E,
```

```
  model_1.delta,
  block=dat.pDC.voom_1.delta.tgts$donorID)$consensus.correlation
```

```
## Warning in glmgam.fit(dx, dy, coef.start = start, tol = tol, maxit = maxit, :
## Too much damping - convergence tolerance not achievable
```

```
## Warning in glmgam.fit(dx, dy, coef.start = start, tol = tol, maxit = maxit, :
## Too much damping - convergence tolerance not achievable
```

```
## Warning in glmgam.fit(dx, dy, coef.start = start, tol = tol, maxit = maxit, :
## Too much damping - convergence tolerance not achievable
```

```
consensus.corr
```

```
## [1] 0.1272602
```

```
# Fit model to transformed count data. Calculate eBayes
efitQW_1.delta <- eBayes(
              lmFit(dat.pDC.voom_1.delta.E,
                    model_1.delta,
                    block=dat.pDC.voom_1.delta.tgts$donorID,
                    correlation=consensus.corr))
```
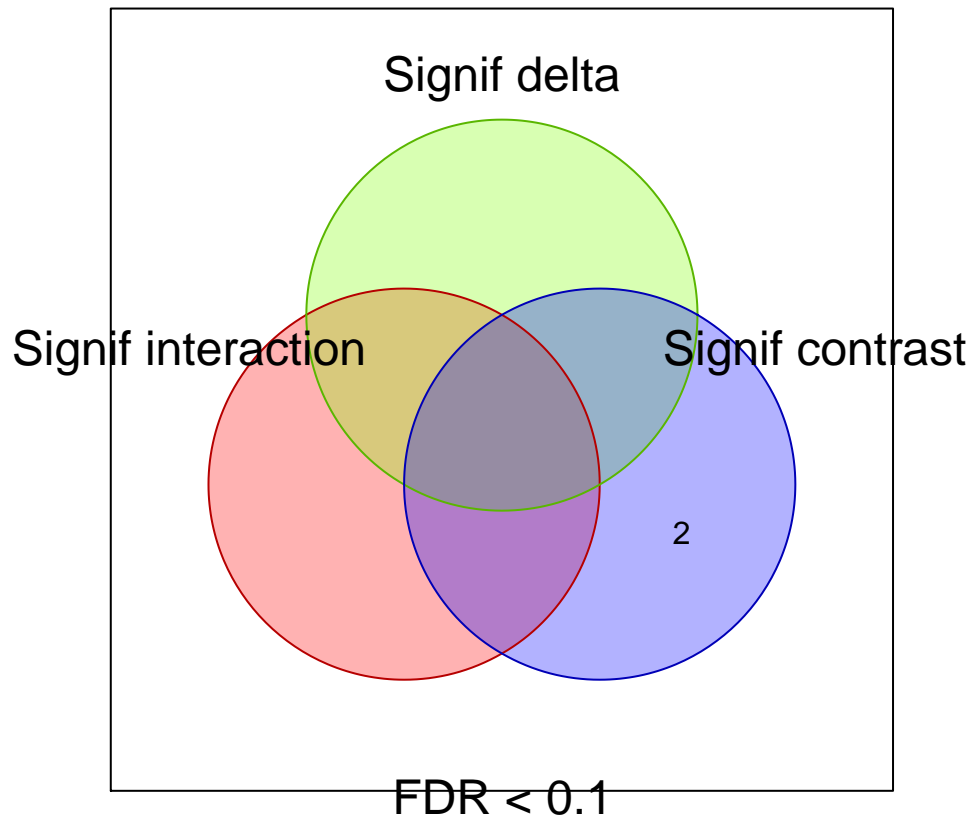
Summary

| Variable | Total genes at FDR | | | | | |
	<0.05	<0.1	<0.2	<0.3	<0.4	<0.5
EOS.supp	0	0	0	1	11	12
total (nonredundant)	0	0	0	1	11	12

Compare models

Total genes significant in one or more models at FDR < 0.1

Save results

```r
dir.create("results/gene_level/P259.1_model.selection/",
           showWarnings = FALSE, recursive = TRUE)

save(pval_1.interact, pval_1.contrast, pval_1.delta,
     file="results/gene_level/P259.1_model.selection/P259.1_gene_pval.RData")

#gene lists
save(venn.list, file="results/gene_level/P259.1_model.selection/P259.1_gene_lists.RData")

#pval csv
pval_1.interact %>%
  mutate(model = "interaction") %>%
  bind_rows(mutate(pval_1.contrast,
                   model="contrasts")) %>%
  bind_rows(mutate(pval_1.delta,
                   model="delta")) %>%
  #Add HGNC symbol
  left_join(dat.pDC.voom_1$genes) %>%
  write_csv(path = "results/gene_level/P259.1_gene_pval.csv")
```

```
## Warning: The `path` argument of `write_csv()` is deprecated as of readr 1.4.0.
## Please use the `file` argument instead.

## Joining, by = "geneName"
```

P259.2 model comparison

Run models

Interaction model

gene expression ~ AntiIL5 * virus + (1|donor)

```
# Define model
model_2.interact <- model.matrix(~ IL5*virus,
                        data=dat.pDC.voom_2$targets)
    colnames(model_2.interact) <- c("(Intercept)", "AntiIL5",
                            "HRV", "AntiIL5:HRV")
#Block by donor
consensus.corr <- duplicateCorrelation(
  dat.pDC.voom_2$E,
  model_2.interact,
  block=dat.pDC.voom_2$targets$donorID)$consensus.correlation

consensus.corr
```

```
## [1] 0.1119666
```

```
# Fit model to transformed count data. Calculate eBayes
efitQW_2.interact <- eBayes(
                lmFit(dat.pDC.voom_2$E,
                    model_2.interact,
                    block=dat.pDC.voom_2$targets$donorID,
                    correlation=consensus.corr))
```

Summary

Variable	Total genes at FDR					
	<0.05	<0.1	<0.2	<0.3	<0.4	<0.5
AntiIL5	18	34	89	168	309	536
HRV	1031	1538	2271	3095	4062	5221
AntiIL5:HRV	14	23	45	90	1160	2321
total (nonredundant)	1042	1556	2315	3184	4816	6543

Contrasts model

gene expression ~ AntiIL5:virus + (1|donor)

```
# Define model
dat.pDC.voom_2$targets <- dat.pDC.voom_2$targets %>%
  mutate(contrast=paste(IL5,virus, sep="_")) %>%
  mutate(contrast=factor(contrast,
                    levels = c("none_none","none_HRV",
                        "AntiIL5_none","AntiIL5_HRV")))

model_2.contrast<- model.matrix(~ 0 + contrast,
                        data=dat.pDC.voom_2$targets)
    colnames(model_2.contrast) <- c(
                            "none_none","none_HRV",
                            "AntiIL5_none","AntiIL5_HRV")
#Block by donor
consensus.corr <- duplicateCorrelation(
```

```
    dat.pDC.voom_2$E,
    model_2.contrast,
    block=dat.pDC.voom_2$targets$donorID)$consensus.correlation

consensus.corr
```

```
## [1] 0.1119666
```

```
#Fit model
fitQW_2.contrast <- lmFit(dat.pDC.voom_2$E,
                    model_2.contrast,
                    block=dat.pDC.voom_2$targets$donorID,
                    correlation=consensus.corr)

#Get contrasts
contrast.matrix <- makeContrasts(
  none_HRV-none_none,
  AntiIL5_HRV-AntiIL5_none,
  AntiIL5_none-none_none,
  AntiIL5_HRV-none_HRV,
  levels=model_2.contrast)

efitQW_2.contrast <- eBayes(contrasts.fit(fitQW_2.contrast,
                            contrast.matrix))
```

Summary

	Total genes at FDR					
Variable	<0.05	<0.1	<0.2	<0.3	<0.4	<0.5
none_HRV - none_none	1031	1538	2271	3095	4062	5221
AntiIL5_HRV - AntiIL5_none	2577	3673	5274	6538	7646	8667
AntiIL5_none - none_none	18	34	89	168	309	536
AntiIL5_HRV - none_HRV	0	3	4	4	4	4
total (nonredundant)	2848	4119	5948	7466	8836	10109

Contrast selected genes

Genes of interest are those that change with virus (*e.g.* significant for virus in untreated and/or treated donors) AND are different between untreated and treated donors in media OR RV.

FDR < 0.05 FDR < 0.1 FDR < 0.2

FDR < 0.3 FDR < 0.4 FDR < 0.5
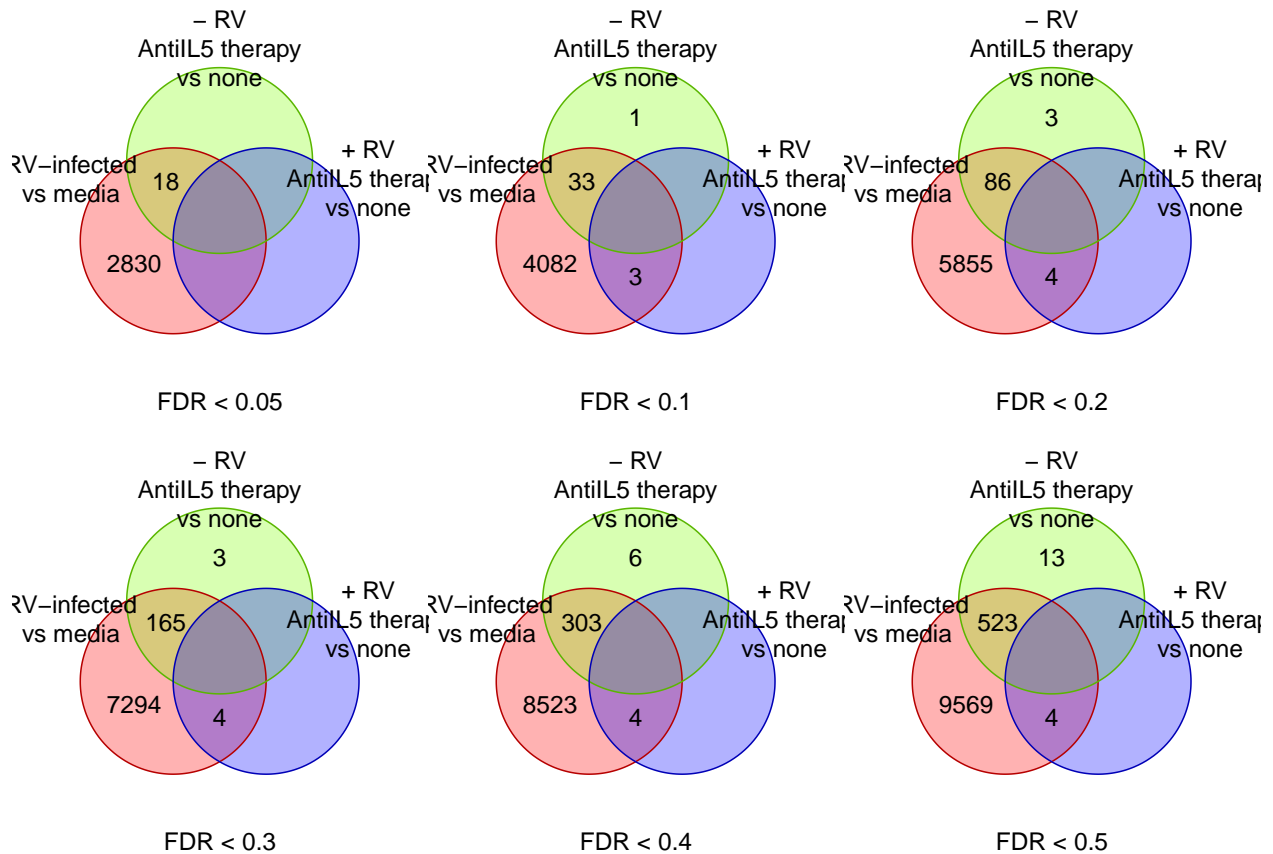
Delta model

RV - media gene expression ~ AntiIL5

Calculate HRV - media counts.

Fit model

```r
# Define model
model_2.delta <- model.matrix(~ IL5,
                          data=dat.pDC.voom_2.delta.tgts)
    colnames(model_2.delta) <- c("(Intercept)","AntiIL5")

#Block by donor
consensus.corr <- duplicateCorrelation(
  dat.pDC.voom_2.delta.E,
  model_2.delta,
  block=dat.pDC.voom_2.delta.tgts$donorID)$consensus.correlation

consensus.corr
```

```
## [1] 0.6068913
```

```r
# Fit model to transformed count data. Calculate eBayes
efitQW_2.delta <- eBayes(
             lmFit(dat.pDC.voom_2.delta.E,
                   model_2.delta,
                   block=dat.pDC.voom_2.delta.tgts$donorID,
                   correlation=consensus.corr))
```

Summary

	Total genes at FDR					
Variable	<0.05	<0.1	<0.2	<0.3	<0.4	<0.5
AntiIL5	1	7	9	17	29	55
total (nonredundant)	1	7	9	17	29	55

Compare models

Total genes significant in one or more models at FDR < 0.1
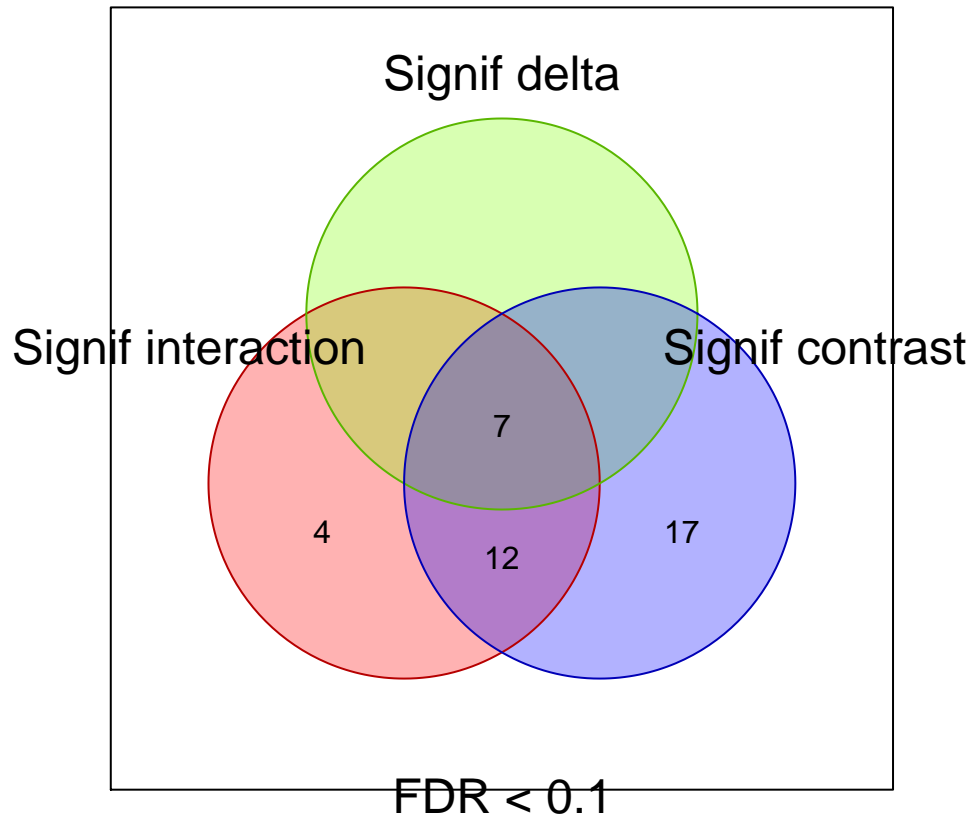
Save results

```
dir.create("results/gene_level/P259.2_model.selection/",
           showWarnings = FALSE, recursive = TRUE)

save(pval_2.interact, pval_2.contrast, pval_2.delta,
     file="results/gene_level/P259.2_model.selection/P259.2_gene_pval.RData")

#gene lists
save(venn.list, file="results/gene_level/P259.2_model.selection/P259.2_gene_lists.RData")

#pval csv
pval_2.interact %>%
  mutate(model = "interaction") %>%
  bind_rows(mutate(pval_2.contrast,
                   model="contrasts")) %>%
```

```
bind_rows(mutate(pval_2.delta,
                 model="delta")) %>%
#Add HGNC symbol
left_join(dat.pDC.voom_2$genes) %>%
write_csv(path = "results/gene_level/P259.2_gene_pval.csv")
```

Joining, by = "geneName"

R session

```
sessionInfo()
```

```
## R version 4.0.2 (2020-06-22)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS  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] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] kableExtra_1.3.4.9000 knitr_1.33           limma_3.44.3
##  [4] venn_1.10             ggpubr_0.4.0         cowplot_1.1.1
##  [7] forcats_0.5.1         stringr_1.4.0        dplyr_1.0.6
## [10] purrr_0.3.4           readr_1.4.0          tidyr_1.1.3
## [13] tibble_3.1.1          ggplot2_3.3.3        tidyverse_1.3.1
##
## loaded via a namespace (and not attached):
##  [1] httr_1.4.2       jsonlite_1.7.2   viridisLite_0.4.0 carData_3.0-4
##  [5] modelr_0.1.8     assertthat_0.2.1 statmod_1.4.36    highr_0.9
##  [9] cellranger_1.1.0 yaml_2.2.1       pillar_1.6.0      backports_1.2.1
## [13] glue_1.4.2       digest_0.6.27    ggsignif_0.6.1    rvest_1.0.0
## [17] colorspace_2.0-1 htmltools_0.5.1.1 pkgconfig_2.0.3  broom_0.7.6
## [21] haven_2.4.1      scales_1.1.1     webshot_0.5.2     svglite_2.0.0
## [25] openxlsx_4.2.3   rio_0.5.26       farver_2.1.0      generics_0.1.0
## [29] admisc_0.12      car_3.0-10       ellipsis_0.3.2    withr_2.4.2
## [33] cli_2.5.0        magrittr_2.0.1   crayon_1.4.1      readxl_1.3.1
## [37] evaluate_0.14    fs_1.5.0         fansi_0.4.2       rstatix_0.7.0
## [41] xml2_1.3.2       foreign_0.8-81   tools_4.0.2       data.table_1.14.0
## [45] hms_1.0.0        lifecycle_1.0.0  munsell_0.5.0     reprex_2.0.0
## [49] zip_2.1.1        compiler_4.0.2   systemfonts_1.0.2 rlang_0.4.11
## [53] grid_4.0.2       rstudioapi_0.13  labeling_0.4.2    rmarkdown_2.8
## [57] gtable_0.3.0     abind_1.4-5      DBI_1.1.1         curl_4.3.1
## [61] R6_2.5.0         lubridate_1.7.10 utf8_1.2.1        stringi_1.6.1
## [65] Rcpp_1.0.6       vctrs_0.3.8      dbplyr_2.1.1      tidyselect_1.1.1
## [69] xfun_0.22
```