rna-seq-analysis

Victoria Latynina

2023-03-18

RNA-seq analysis

Pipeline steps:

Combine counts into count matrix

Run VST for PCA, run PCA

Run differential expression

Find pathways for DE

```
library(ggplot2)
library(DESeq2)
library(apeglm)
library(ggrepel)
library(dplyr)
library(org.Mm.eg.db)
library(PCAtools)
```

Combine counts into count matrix

```
countFiles <- list.files("~/Code/sys_bio/Part_2/materials/GSE137888_RAW", full.names = T)
countFiles
```

```
## [1] "/Users/victoria_latynina/Code/sys_bio/Part_2/materials/GSE137888_RAW/SRR10166256_Control.fc.txt
## [2] "/Users/victoria_latynina/Code/sys_bio/Part_2/materials/GSE137888_RAW/SRR10166257_Control.fc.txt
## [3] "/Users/victoria_latynina/Code/sys_bio/Part_2/materials/GSE137888_RAW/SRR10166258_Control.fc.txt
## [4] "/Users/victoria_latynina/Code/sys_bio/Part_2/materials/GSE137888_RAW/SRR10166259_Prdm16.fc.txt"
## [5] "/Users/victoria_latynina/Code/sys_bio/Part_2/materials/GSE137888_RAW/SRR10166260_Prdm16.fc.txt"
## [6] "/Users/victoria_latynina/Code/sys_bio/Part_2/materials/GSE137888_RAW/SRR10166261_Prdm16.fc.txt"
```

```
counts <- lapply(countFiles, function(countsFile) {
  read.table(countsFile, sep="\t", header=1, skip = 1, row.names = 1, stringsAsFactors = F, comment.cha
})

head(counts[[1]])
```

```
##                                                Chr
## ENSMUSG00000102693.1                          chr1
```

```
## ENSMUSG00000064842.1                                  chr1
## ENSMUSG00000051951.5 chr1;chr1;chr1;chr1;chr1;chr1;chr1
## ENSMUSG00000102851.1                                  chr1
## ENSMUSG00000103377.1                                  chr1
## ENSMUSG00000104017.1                                  chr1
##                                                                                  Start
## ENSMUSG00000102693.1                                                           3073253
## ENSMUSG00000064842.1                                                           3102016
## ENSMUSG00000051951.5 3205901;3206523;3213439;3213609;3214482;3421702;3670552
## ENSMUSG00000102851.1                                                           3252757
## ENSMUSG00000103377.1                                                           3365731
## ENSMUSG00000104017.1                                                           3375556
##                                                                                    End
## ENSMUSG00000102693.1                                                           3074322
## ENSMUSG00000064842.1                                                           3102125
## ENSMUSG00000051951.5 3207317;3207317;3215632;3216344;3216968;3421901;3671498
## ENSMUSG00000102851.1                                                           3253236
## ENSMUSG00000103377.1                                                           3368549
## ENSMUSG00000104017.1                                                           3377788
##                                Strand Length Count
## ENSMUSG00000102693.1                +   1070     0
## ENSMUSG00000064842.1                +    110     0
## ENSMUSG00000051951.5 -;-;-;-;-;-;-   6094     0
## ENSMUSG00000102851.1                +    480     0
## ENSMUSG00000103377.1                -   2819     0
## ENSMUSG00000104017.1                -   2233     0
```

```
counts <- lapply(counts, function(countsTable) countsTable[, "Count", drop=F])
counts <- do.call(cbind, counts)
colnames(counts) <- gsub(".*(SRR\\d+).*", "\\1", countFiles)
head(counts)
```

```
##                      SRR10166256 SRR10166257 SRR10166258 SRR10166259
## ENSMUSG00000102693.1           0           0           0           0
## ENSMUSG00000064842.1           0           0           0           0
## ENSMUSG00000051951.5           0           1           0           1
## ENSMUSG00000102851.1           0           0           0           0
## ENSMUSG00000103377.1           0           0           0           0
## ENSMUSG00000104017.1           0           0           0           0
##                      SRR10166260 SRR10166261
## ENSMUSG00000102693.1           0           0
## ENSMUSG00000064842.1           0           0
## ENSMUSG00000051951.5           1           0
## ENSMUSG00000102851.1           0           0
## ENSMUSG00000103377.1           0           0
## ENSMUSG00000104017.1           0           0
```

```
coldata <- data.frame(
  srr=gsub(".*(SRR\\d+).*", "\\1", countFiles),
  condition=gsub(".*(SRR\\d+)_(Control|Prdm16).*", "\\2", countFiles),
  row.names =gsub(".*(SRR\\d+).*", "\\1", countFiles)
)
```

```
write.table(coldata, file="~/Code/sys_bio/Part_2/materials/GSE137888_coldata.tsv", sep="\t", quote=F, co
coldata
```

```
##                      srr condition
## SRR10166256 SRR10166256   Control
## SRR10166257 SRR10166257   Control
## SRR10166258 SRR10166258   Control
## SRR10166259 SRR10166259    Prdm16
## SRR10166260 SRR10166260    Prdm16
## SRR10166261 SRR10166261    Prdm16
```

Run differential expression

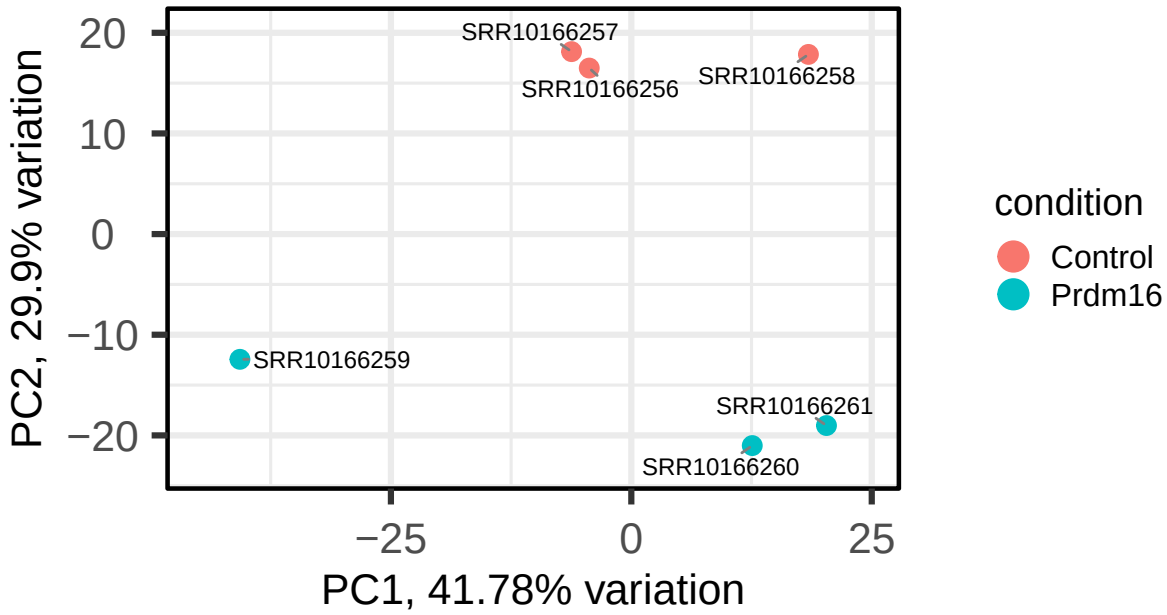
```
dds <- DESeqDataSetFromMatrix(countData = counts[rowSums(counts) > 10, ],
                              colData = coldata,
                              design= ~ condition)
dds <- DESeq(dds)
resultsNames(dds) # lists the coefficients
```

```
## [1] "Intercept"                    "condition_Prdm16_vs_Control"
```

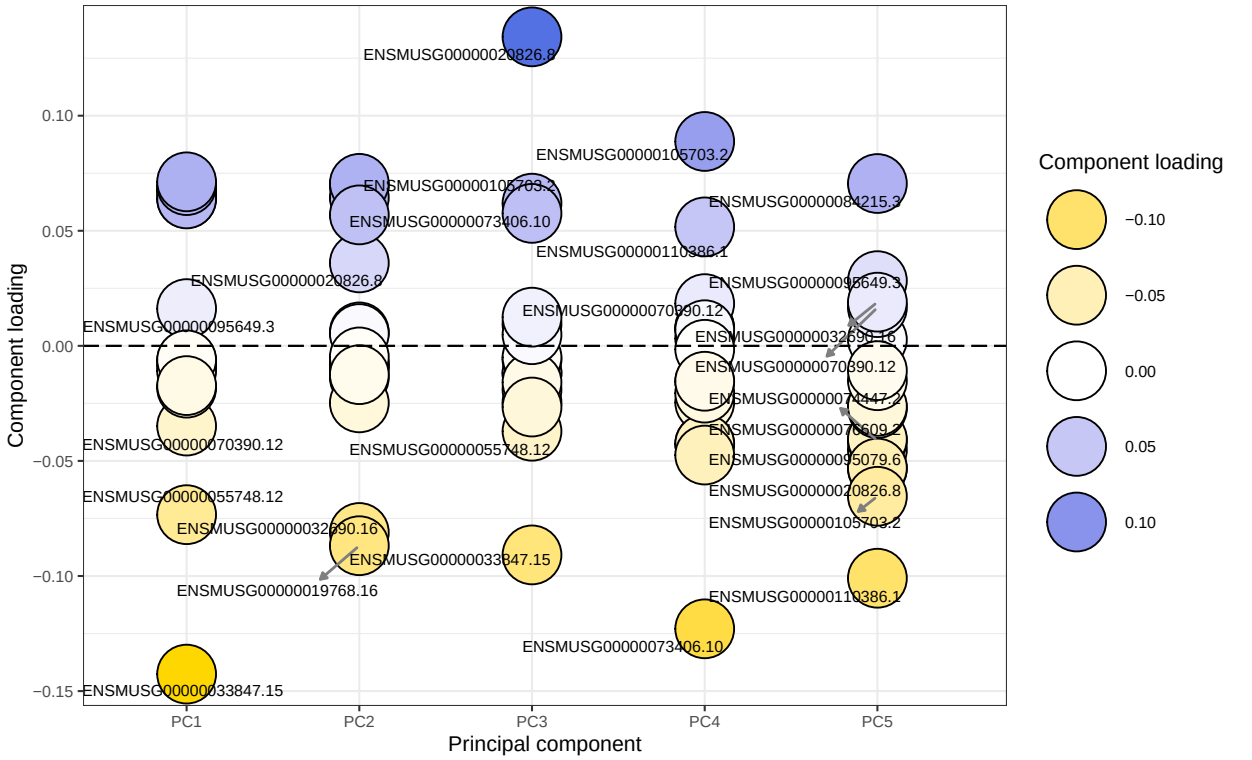
```
rlog <- rlogTransformation(dds)
write.table(assay(rlog), file="~/Code/sys_bio/Part_2/materials/GSE137888_rlog.tsv", sep="\t", quote=F, c
```

Run VST for PCA, run PCA

```
vst <- varianceStabilizingTransformation(dds)
# plotPCA(vst, intgroup=c("condition"), ntop=nrow(vst)) + theme_bw()
pcaData <- pca(assay(vst), metadata=coldata)
biplot(pcaData, colby="condition", legendPosition = "right")
```



```
plotloadings(pcaData) + theme_bw(base_size=8)
```




```
head(results(dds))
```

```
## log2 fold change (MLE): condition Prdm16 vs Control
## Wald test p-value: condition Prdm16 vs Control
## DataFrame with 6 rows and 6 columns
##                         baseMean log2FoldChange     lfcSE       stat    pvalue
##                        <numeric>      <numeric> <numeric>  <numeric> <numeric>
## ENSMUSG00000103147.1     2.05487     -1.2346017  1.664859 -0.7415654  0.458351
## ENSMUSG00000098104.1    18.91484     -0.2730375  0.482898 -0.5654148  0.571792
## ENSMUSG00000103922.1    15.18818     -0.6780091  0.571090 -1.1872201  0.235141
## ENSMUSG00000033845.13  445.20248     -0.0216442  0.213024 -0.1016049  0.919070
## ENSMUSG00000102275.1     3.11991     -0.1135539  1.145577 -0.0991238  0.921040
## ENSMUSG00000025903.14 1500.43596      0.0785121  0.165100  0.4755418  0.634401
##                             padj
##                        <numeric>
## ENSMUSG00000103147.1          NA
## ENSMUSG00000098104.1    0.972583
## ENSMUSG00000103922.1    0.835081
## ENSMUSG00000033845.13   0.997337
## ENSMUSG00000102275.1          NA
## ENSMUSG00000025903.14   0.979834
```

Get volcano plots

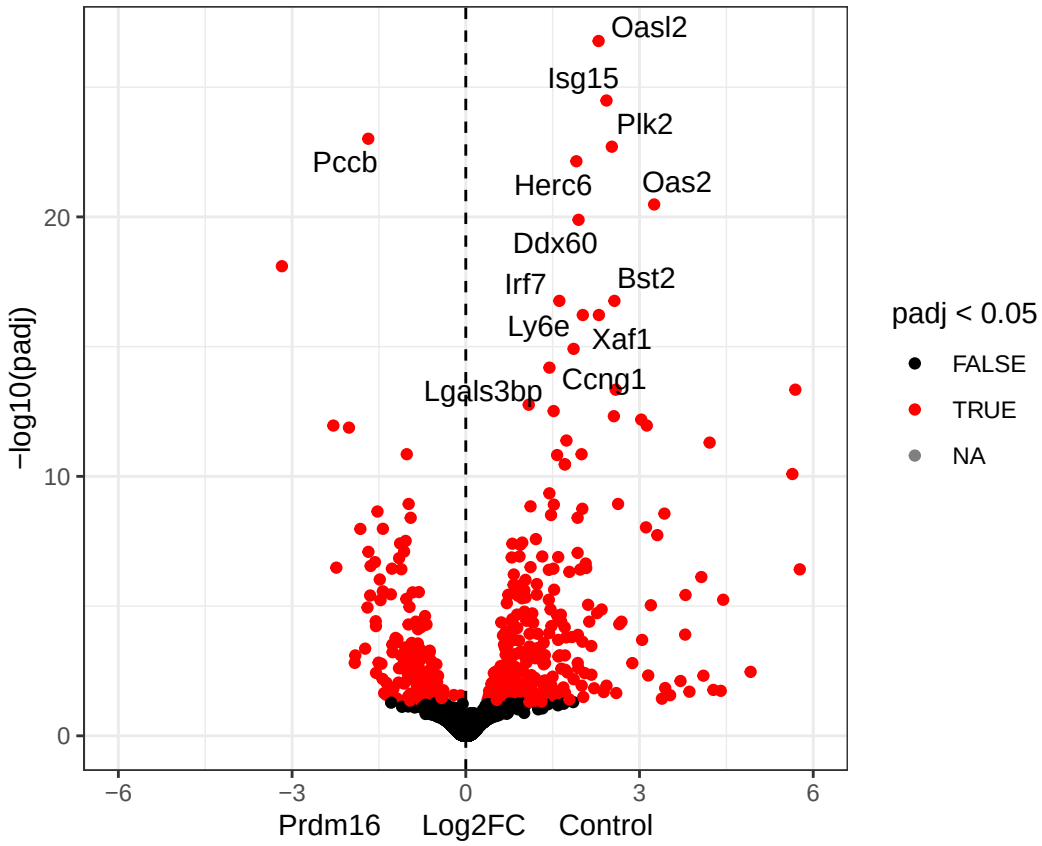
```
res <- lfcShrink(dds, coef="condition_Prdm16_vs_Control", type="apeglm")
head(res)
```

```
## log2 fold change (MAP): condition Prdm16 vs Control
## Wald test p-value: condition Prdm16 vs Control
## DataFrame with 6 rows and 5 columns
##                          baseMean log2FoldChange     lfcSE    pvalue      padj
##                         <numeric>      <numeric> <numeric> <numeric> <numeric>
## ENSMUSG00000103147.1     2.05487    -0.01156381  0.161490  0.458351        NA
## ENSMUSG00000098104.1    18.91484    -0.02776637  0.156325  0.571792  0.972583
## ENSMUSG00000103922.1    15.18818    -0.05241754  0.167191  0.235141  0.835081
## ENSMUSG00000033845.13  445.20248    -0.00776160  0.128871  0.919070  0.997337
## ENSMUSG00000102275.1     3.11991    -0.00217361  0.160042  0.921040        NA
## ENSMUSG00000025903.14 1500.43596     0.03956795  0.118090  0.634401  0.979834
```

```
# keytypes(org.Mm.eg.db)
res$Gene.symbol <- mapIds(org.Mm.eg.db, gsub("\\.\\d+", "", rownames(res)), column="SYMBOL", keytype="E
```

```
resDF <- as.data.frame(res)
volcanoPlot <- ggplot(resDF, aes(x=log2FoldChange, y=-log10(padj), color=padj < 0.05)) +
  geom_point() + theme_bw() + scale_color_manual(values=c("black", "red"))
volcanoPlot <- volcanoPlot +
  geom_text_repel(data=resDF %>% dplyr::filter(padj < 1e-14), aes(label=Gene.symbol), color="black")

volcanoPlot <- volcanoPlot +
  xlim(c(-6, 6)) +
  xlab("Prdm16     Log2FC    Control") +
  geom_vline(xintercept = 0, lty=2) +
  theme(aspect.ratio = 1)
volcanoPlot
```


Oasl2
Isg15
Plk2
Pccb
Herc6
Oas2
Ddx60
Irf7
Bst2
Ly6e
Xaf1
Lgals3bp
Ccng1
−log10(padj)
20
10
0
−6
−3
0
3
6
Prdm16
Log2FC
Control
padj < 0.05
FALSE
TRUE
NA

Pathway analysis