Lupus Mini Project

Aditi

Table of contents

Background

Lupus arises when the immune system, often responsible for safeguarding the body against infections and diseases, erroneously targets its own tissues. This attack induces inflammation and, in certain instances, irreversible tissue damage, potentially impacting several systems, including the skin, joints, heart, lungs, kidneys, circulating blood cells, and brain.

```
library(GEOquery)
library(dplyr)
library(ggplot2)
library(DESeq2)
```

```
gse <- getGEO("GSE149050")
```

Extract metadata for the 'gse' object

```
#gse list has only one entry
metadata <- pData(phenoData(gse[[1]]))
dim(metadata)
```

```
[1] 288  64
```

Exploratory Analysis

```r
table(metadata$characteristics_ch1)
```

```
                   disease state: healthy control 
                                               85 
disease state: systemic lupus erythematosus (SLE) 
                                              203 
```

Q. How many different cell types are there?

```r
table(metadata$characteristics_ch1.2)
```

```
cell type: B cells     cell type: cDC     cell type: cMo     cell type: pDC 
                32                 32                119                 33 
    cell type: PMN cell type: T cells 
                36                 36 
```

Q. How many male and female patients?

```r
table(metadata$characteristics_ch1.7)
```

```
gender: Female   gender: Male 
           281              7 
```

Q. Why are few males affected?

Lupus affects women significantly more than men, with a 9:1 female-to-male ratio, and while the exact reasons are still being researched, hormonal differences, particularly estrogen, and genetics are thought to play a role.

Q. Break down of race by gender?

```
table(metadata$characteristics_ch1.10, metadata$characteristics_ch1.7)
```

```
                                    gender: Female gender: Male
  ethnicity: Hispanic                           24            0
  ethnicity: Hispanic/Latino                    53            6
  ethnicity: Non Hispanic                       37            0
  ethnicity: Non-Hispanic/Latino               137            1
  ethnicity: Not available                      23            0
  ethnicity: not listed                          1            0
  ethnicity: Pacific Islander                    6            0
```

Setup for DESeq2

```
metadata.tc <- filter(metadata, characteristics_ch1.2 == "cell type: T cells" & characterist
head(metadata.tc[,1:3])
```

```
                    title geo_accession                status
GSM4489145 001_L0038_HC_T    GSM4489145 Public on Feb 01 2021
GSM4489147 003_L0140_HC_T    GSM4489147 Public on Feb 01 2021
GSM4489148 004_T4631_HC_T    GSM4489148 Public on Feb 01 2021
GSM4489149 005_T5210_HC_T    GSM4489149 Public on Feb 01 2021
GSM4489150 006_T5466_HC_T    GSM4489150 Public on Feb 01 2021
GSM4489151 007_T5502_HC_T    GSM4489151 Public on Feb 01 2021
```

Q. How were these samples processed (alignment/mapping software version and genome build used)?

```
metadata.tc$data_processing[1]
```

```
[1] "Bulk RNA-seq data (FASTQ files) were mapped against the hg38 genome (GRCh38.p7) referen
```

```
metadata.subset <- metadata.tc %>%
  select(title,
         disease_state = characteristics_ch1,
         ifn_status = characteristics_ch1.3,
         patient_id = characteristics_ch1.4) %>%
  mutate(disease_state = gsub("disease state: ","", disease_state)) %>%
  mutate(ifn_status = gsub("ifn status: ","", ifn_status)) %>%
  mutate(patient_id = gsub("patientuid: ","", patient_id)) %>%
  mutate(state = recode(disease_state,
                        "healthy control" = "control",
                        "systemic lupus erythematosus (SLE)" = "lupus"))
table(metadata.subset$state)
```

```
control   lupus
     10      23
```

```
#Read count data
counts.all <- read.delim("GSE149050_Bulk_Human_RawCounts.txt.gz",
                         check.names=FALSE, row.names = 1)
head(counts.all[,1:3])
```

```
          001_L0038_HC_T 002_L0088fresh_HC_T 003_L0140_HC_T
5S_rRNA                2                   0              2
5_8S_rRNA              0                   0              0
7SK                    1                   1              2
A1BG                  53                  56            105
A1BG-AS1               7                  24             46
A1CF                   0                   2              2
```

```
head(metadata.subset$title)
```

```
[1] "001_L0038_HC_T" "003_L0140_HC_T" "004_T4631_HC_T" "005_T5210_HC_T"
[5] "006_T5466_HC_T" "007_T5502_HC_T"
```

```
counts.subset <- counts.all %>% select(metadata.subset$title)
dim(counts.subset)
```

```
[1] 56269    33
```

```
all(colnames(counts.subset) == metadata.subset$title)
```

```
[1] TRUE
```

```r
# Remove title column then use it as row names
colData <- metadata.subset[,-1]
rownames(colData) <- metadata.subset[,1]
head(colData)
```

```
                 disease_state ifn_status patient_id   state
001_L0038_HC_T healthy control         HC      L0038 control
003_L0140_HC_T healthy control         HC      L0140 control
004_T4631_HC_T healthy control         HC      T4631 control
005_T5210_HC_T healthy control         HC      T5210 control
006_T5466_HC_T healthy control         HC      T5466 control
007_T5502_HC_T healthy control         HC      T5502 control
```

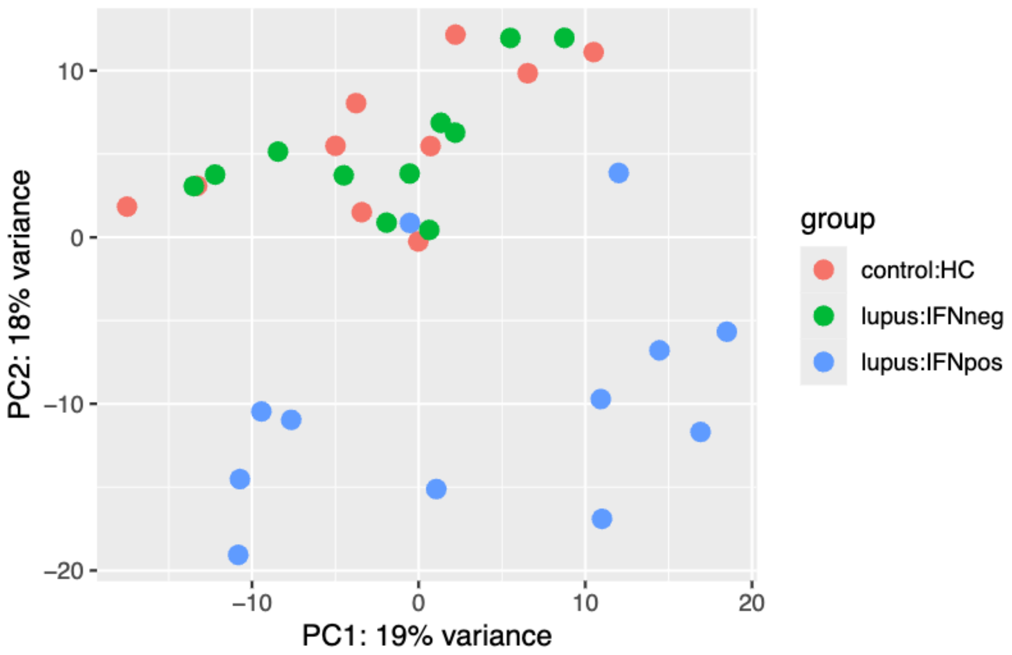
Set up DESeq object

```r
dds <- DESeqDataSetFromMatrix(countData = counts.subset, #you already have the matrix
                              colData = colData,
                              design = ~state)
```

```r
keep.inds <- rowSums(counts(dds)) >= 10
dds <- dds[keep.inds,]
```

```r
#PCA analysis
vsd <- vst(dds, blind = FALSE)
plotPCA(vsd, intgroup = c("state", "ifn_status"))
```

```
using ntop=500 top features by variance
```



Running DESeq2

```
dds <- DESeq(dds)
```

Extract Results

```
res <- results(dds)
head(res)
```

```
log2 fold change (MLE): state lupus vs control
Wald test p-value: state lupus vs control
DataFrame with 6 rows and 6 columns
           baseMean log2FoldChange     lfcSE      stat    pvalue      padj
          <numeric>      <numeric> <numeric> <numeric> <numeric> <numeric>
5S_rRNA     2.78039      0.4759046  0.457248  1.040802  0.297967        NA
7SK         2.33480      0.3570971  0.489993  0.728781  0.466136        NA
A1BG       81.11733     -0.0226099  0.176306 -0.128243  0.897957  0.974910
A1BG-AS1   21.03977      0.2993010  0.264304  1.132410  0.257462  0.698095
A1CF        3.11018      0.3174058  0.566481  0.560311  0.575267        NA
```

```
A2M      251.71157     -0.0294669  0.287036 -0.102659  0.918233  0.980429
```

```r
summary(res)
```

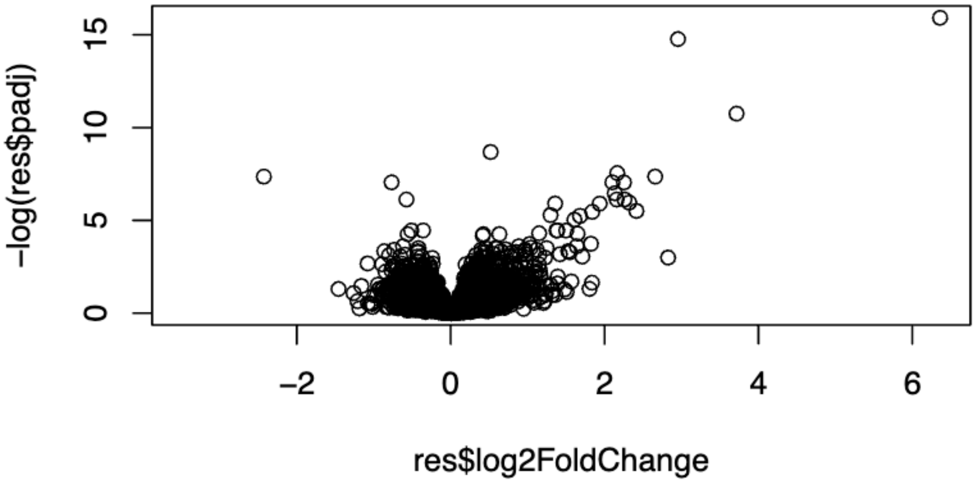
```
out of 31491 with nonzero total read count
adjusted p-value < 0.1
LFC > 0 (up)       : 110, 0.35%
LFC < 0 (down)     : 44, 0.14%
outliers [1]       : 0, 0%
low counts [2]     : 16500, 52%
(mean count < 16)
[1] see 'cooksCutoff' argument of ?results
[2] see 'independentFiltering' argument of ?results
```

After adjusting p-value threshold

```r
res_p05 <- results(dds, alpha=0.05)
summary(res_p05)
```

```
out of 31491 with nonzero total read count
adjusted p-value < 0.05
LFC > 0 (up)       : 61, 0.19%
LFC < 0 (down)     : 17, 0.054%
outliers [1]       : 0, 0%
low counts [2]     : 15890, 50%
(mean count < 14)
[1] see 'cooksCutoff' argument of ?results
[2] see 'independentFiltering' argument of ?results
```

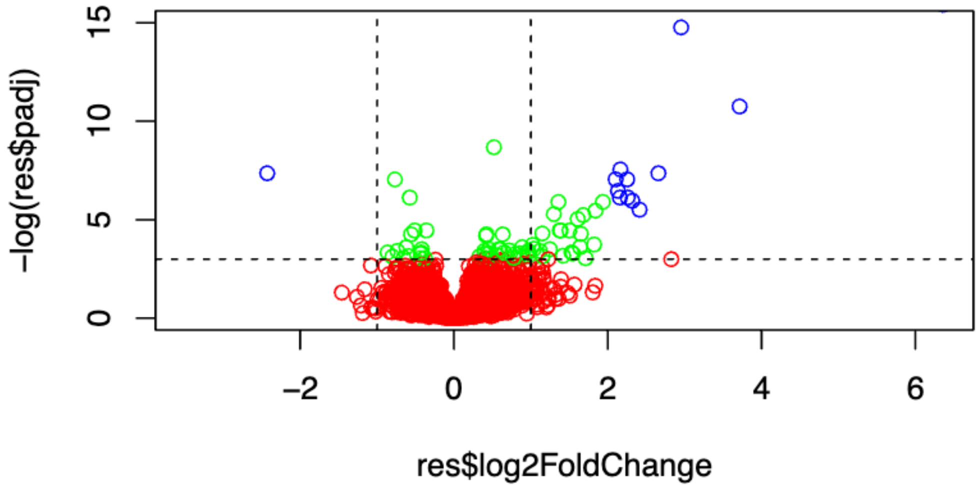
```r
write.csv(res_p05, "deseq_results_tc_SLE.csv")
plot(res$log2FoldChange, -log(res$padj))
```



Vizualizations

```
mycols <- rep("green", nrow(res))
mycols[ abs(res$log2FoldChange) > 2] = "blue"
mycols[ res$padj > 0.05 ] = "red"

plot(res$log2FoldChange, -log(res$padj), ylim=c(0,15), col=mycols)
abline(v=c(-1,1), lty=2)
abline(h=-log(0.05), lty=2)
```

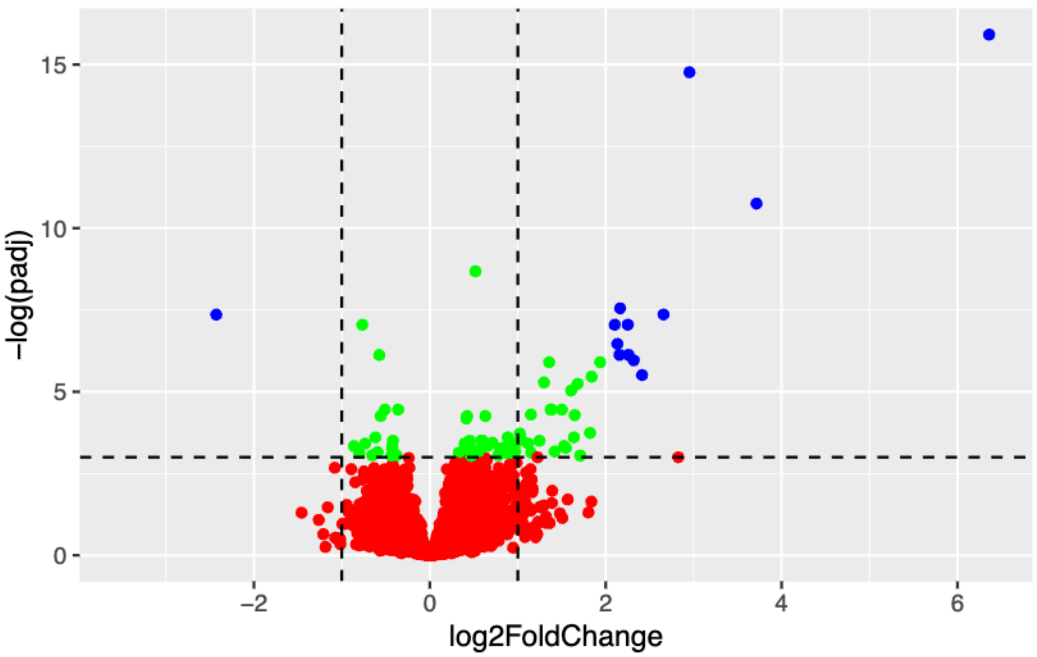


```
results <- as.data.frame(res)

library(ggplot2)

ggplot(results) +
  aes(log2FoldChange, -log(padj)) +
  geom_point(col=mycols) +
  geom_vline(xintercept = c(-1,+1), linetype=2) +
  geom_hline(yintercept = -log(0.05), linetype=2)
```

```
Warning: Removed 16500 rows containing missing values or values outside the scale range
(`geom_point()`).
```



Extract top genes

```
top.genes <- results %>% filter(padj <= 0.05 & abs(log2FoldChange) >= 2)
head(top.genes)
```

```
        baseMean log2FoldChange     lfcSE     stat       pvalue         padj
CMPK2   218.7990       2.251249 0.4504244 4.998062 5.790927e-07 8.690444e-04
GSTM1   124.9589       2.953416 0.4497644 6.566585 5.148243e-11 3.862984e-07
IFI27   885.4340       6.360160 0.9304322 6.835705 8.160257e-12 1.224610e-07
IFI44  1207.3932       2.102190 0.4196298 5.009631 5.453450e-07 8.690444e-04
IFI44L 2167.9605       3.714660 0.6325324 5.872679 4.288088e-09 2.145045e-05
IFIT1   285.4165       2.132057 0.4382361 4.865088 1.144060e-06 1.560810e-03
```

#Save Results

```
save(top.genes, file = "top_genes.RData")
write.csv(res, file="results.csv")
```