Differential Expression Analysis

First we will load the necessary libraries.

```
library(tidyverse)
library(DESeq2)
library(magrittr)
```

The data has been extracted from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE226134 from the GSE226134_CK_10___norm.xlsx file.

```
data <- read.csv("data/input/JoinedWide.csv")

property <- as.data.frame(cbind(data$SegmentDisplayName, data$METASTATIC))
names(property)=c("SegmentDisplayName","METASTATIC")
normCountData <- data[,59:ncol(data)]
row.names(normCountData) <- data$SegmentDisplayName
```

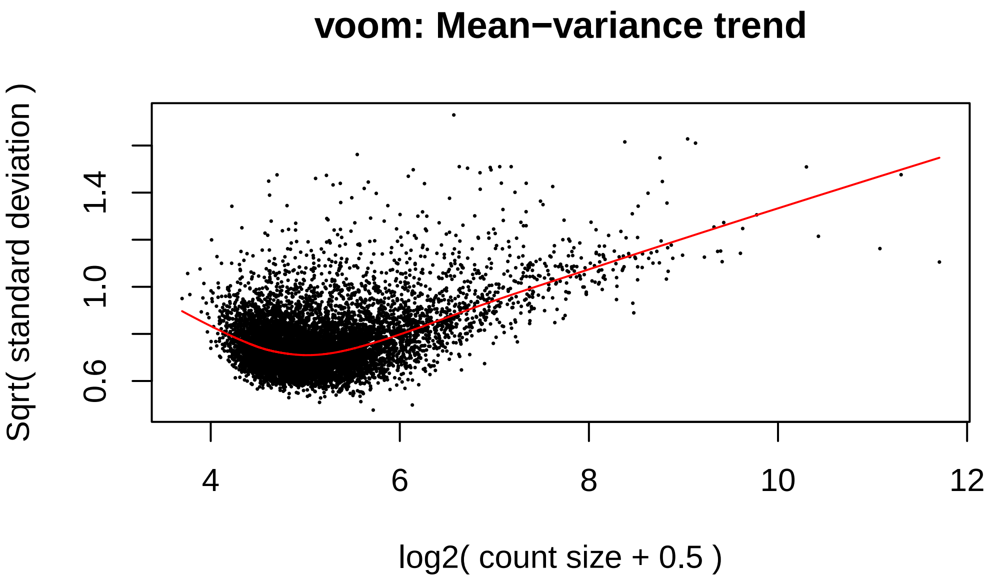
Typical log-2 fold change scores are around 1.5-2. Here we need to drastically increase in order to be more specific for our results.

```
# Load the limma library
library(limma)
```

```
Warning: package 'limma' was built under R version 4.1.3


Attaching package: 'limma'


The following object is masked from 'package:DESeq2':

    plotMA
```

```
The following object is masked from 'package:BiocGenerics':

    plotMA
```

```
library(edgeR)
design <- model.matrix(~property$METASTATIC)

y <- DGEList(counts=t(normCountData),group=property$METASTATIC)
keep <- filterByExpr(y)
y <- y[keep,,keep.lib.sizes=FALSE]
y <- calcNormFactors(y)
v <- voom(y,design,plot=TRUE)
```



```
fit <- lmFit(v,design)

contrast <- c(-1,1)
fit2 <- contrasts.fit(fit,contrast)
fit2 <- eBayes(fit2)


de_genes <- topTable(fit2,coef=1,number=Inf,adjust.method="fdr",sort.by="B",lfc=log2(250))

threshold <- 0.05/9223
sig_genes <- de_genes %>% filter(P.Value < threshold)
```

```
nrow(sig_genes) #this is saying that all the genes are differentially expressed
```

```
[1] 487
```

To check our findings, we have use a t-test

```
#reference from lab 7
boolM = property$METASTATIC=="True"
ttests = apply(normCountData,2,function(x){t.test(x[boolM],x[!boolM])$p.value})

ttests <- ttests %>%
  as.data.frame() %>%
  set_colnames("pval") %>%
  rownames_to_column("gene") %>%
  arrange(pval)

pval_threshold <- 0.05/nrow(ttests) #bonferroni corrections

DEGenes_ttest <- ttests %>% filter(pval < pval_threshold)
nrow(DEGenes_ttest) #so there are 11 differentially expressed genes
```

```
[1] 11
```

Similar to a t-test, we can use the non-parametric Wilcoxon rank sum test to see differentially expressed genes.

```
# Define a custom function to perform wilcoxon rank sum test and return p-value
#reference from lab 7
boolM = property$METASTATIC=="True"
wilcoxon_tests = apply(normCountData,2,function(x){wilcox.test(x[boolM],x[!boolM], alterna

wilcoxon_tests <- wilcoxon_tests %>%
  as.data.frame() %>%
  set_colnames("pval") %>%
  rownames_to_column("gene") %>%
  arrange(pval)

pval_threshold <- 0.05/nrow(wilcoxon_tests) #bonferroni corrections

DEGenes_wilcoxon_tests <- wilcoxon_tests %>% filter(pval < pval_threshold)
```

```r
nrow(DEGenes_wilcoxon_tests) #so there are 2 differentially expressed genes
```

```
[1] 2
```

```r
DEGenes_wilcoxon_tests$gene %in% DEGenes_ttest$gene
```

```
[1] FALSE  TRUE
```

```r
DEGenes_ttest$gene %in% row.names(sig_genes)
```

```
 [1]  TRUE  TRUE  TRUE  TRUE FALSE FALSE  TRUE FALSE FALSE  TRUE  TRUE
```

```r
DEGenes_wilcoxon_tests$gene %in% row.names(sig_genes)
```

```
[1] FALSE  TRUE
```