# Principal components analysis separates RNA-Seq samples by experimental variables

Molly Davis

2023-09-06

---

## Introduction

This Markdown analyzes RNA-Seq gene expression data from data file `results/muday-144-SL5_counts-salmon.txt` documented in `Documentation/muday-144_sample_sheet.xlsx` using functions from the DESeq library.

Note that the version of `results/muday-144-SL5_counts-salmon.txt` is: `f49ae22ec21c40fafa41d4325fc0752e8a016843`

Question:

- Are the samples from different experimental conditions and genotypes well separated by PC1 and PC2?

---

## Background

The data set analyzed here comes from heat-treated and control samples collected at 15, 30, 45, and 75 minutes of heat treatment. The analysis shown here will investigate if the ARE genotype shows greater variance compared to other genotypes VF36 and OE3.

---

## Results

Load required library from the Bioconductor project:

Load required data:

```
counts = read.csv('results/muday-144-SL5_counts-salmon.txt',sep = "\t",stringsAsFactors = F,
                  header=T, check.names = FALSE, row.names = "gene_name")
counts = counts[,-ncol(counts)]
```

Define a function that builds a Experiment Meta data frame:

```
library(stringr)
makeExperimentMetaDataFrame = function(fname="results/muday-144-SL5_counts-salmon.txt") {
  experiment_data = read.csv(fname,
                             sep = "\t",
                             stringsAsFactors = F,
                             header=T,
                             check.names = FALSE,
                             row.names = "gene_name")
  sample_names = colnames(experiment_data)
  boolean_vector = str_detect(sample_names,'[VFA]\\.\\d\\d\\.\\d\\d\\.\\d')
  sample_names = sample_names[boolean_vector]
  genotype = sapply(strsplit(sample_names,"\\."),function(x){x[[1]]})
  time = sapply(strsplit(sample_names,"\\."),function(x){x[[3]]})
  temperature = sapply(strsplit(sample_names,"\\."),function(x){x[[2]]})
  to_return = data.frame(genotype,time,temperature)
  rownames(to_return) = sample_names
  return(to_return)
}
```

Define a function that tests treatment versus control in samples that have the same variety and treatment duration as well as time duration:

```
getDEgenes <- function(counts, sampleData) {
  cts <- as.matrix(counts)
  coldata <- sampleData
  coldata <- coldata[,c("genotype", "time", "temperature")]
  coldata$genotype <- factor(coldata$genotype, levels = c("A", "F", "V"))
  coldata$time <- factor(coldata$time, levels = c("15", "30", "45", "75"))
  coldata$temperature <- factor(coldata$temperature, levels = c("28", "34"))

  cts <- cts[, rownames(coldata)]
  dds <- DESeqDataSetFromMatrix(countData = round(cts),
                                colData = coldata,
                                design = ~ time + temperature)
  featureData <- data.frame(gene=rownames(cts))
  mcols(dds) <- DataFrame(mcols(dds), featureData)
  dds <- DESeq(dds, minReplicatesForReplace=Inf)
  keep <- rowSums(counts(dds)) >= 10
  dds <- dds[keep,]
  resultsNames(dds)
  res05 <- results(dds, alpha=0.05)
  numSignGenes<- sum(res05$pvalue < 0.05, na.rm=TRUE) # Not adjusted
  numSignGenes<- sum(res05$padj < 0.05, na.rm=TRUE) # Adjusted
  res05 <- res05[order(res05$pvalue),]
  vsd <- vst(dds, blind=FALSE)
}
```

Define a Function that create PCA plots that show control and treatment groups clusters over the time durations:

```
PCA_plots <- function(vsd, sampleData, title) {
  options(ggrepel.max.overlaps = Inf)
  pcaData <- plotPCA(vsd, intgroup=c("temperature", "genotype", "time"), returnData=TRUE)
  percentVar <- round(100 * attr(pcaData, "percentVar"))
  ggplot(pcaData, aes(PC1, PC2, color=time, shape=genotype)) +
    #geom_label_repel(aes(label = rownames(sampleData))) +
    scale_color_manual(values =  c("15" ="#56B4E9","30" = "#009E73",
                                   "45" = "#E69F00", "75" = "#CC79A7")) +
    scale_fill_manual(values =  c("15" ="#56B4E9","30" = "#009E73",
                                  "45" = "#E69F00", "75" = "#CC79A7")) +
    geom_point(size =4, aes(fill =time, alpha=temperature)) +
    geom_point(size=4) +
    scale_shape_manual(values= c("A"= 23, "F"= 22, "V"= 25)) +
    scale_alpha_manual(values=c("28"=0.1, "34"=1)) +
    #geom_mark_circle(aes(color = as.factor(time)), expand = unit(0.5,"mm"))+
    xlab(paste0("PC1: ",percentVar[1],"% variance")) +
    ylab(paste0("PC2: ",percentVar[2],"% variance")) +
    ggtitle(title) +
    theme_bw() +
    theme(aspect.ratio = 1)+
    theme(plot.title = element_text(hjust = 0.5)) +
    theme(text = element_text(size = 15))
}
```

Run meta dataframe function:

```
sampleData = makeExperimentMetaDataFrame()
```

Sanity Check:

```
counts <- counts[,unique(rownames(sampleData))]
check <- all(colnames(counts) == rownames(sampleData))
if(check == "TRUE"){
print("Columns and rows match. Please Continue the analysis!")
}else {
print("Columns and rows do not match. Please fix this before continuing!")
}
```

```
## [1] "Columns and rows match. Please Continue the analysis!"
```

Run all data and make a DESeq analysis with the design time + temperature.

Run just OE3 genotype data and make a DESeq analysis with the design time + temperature.

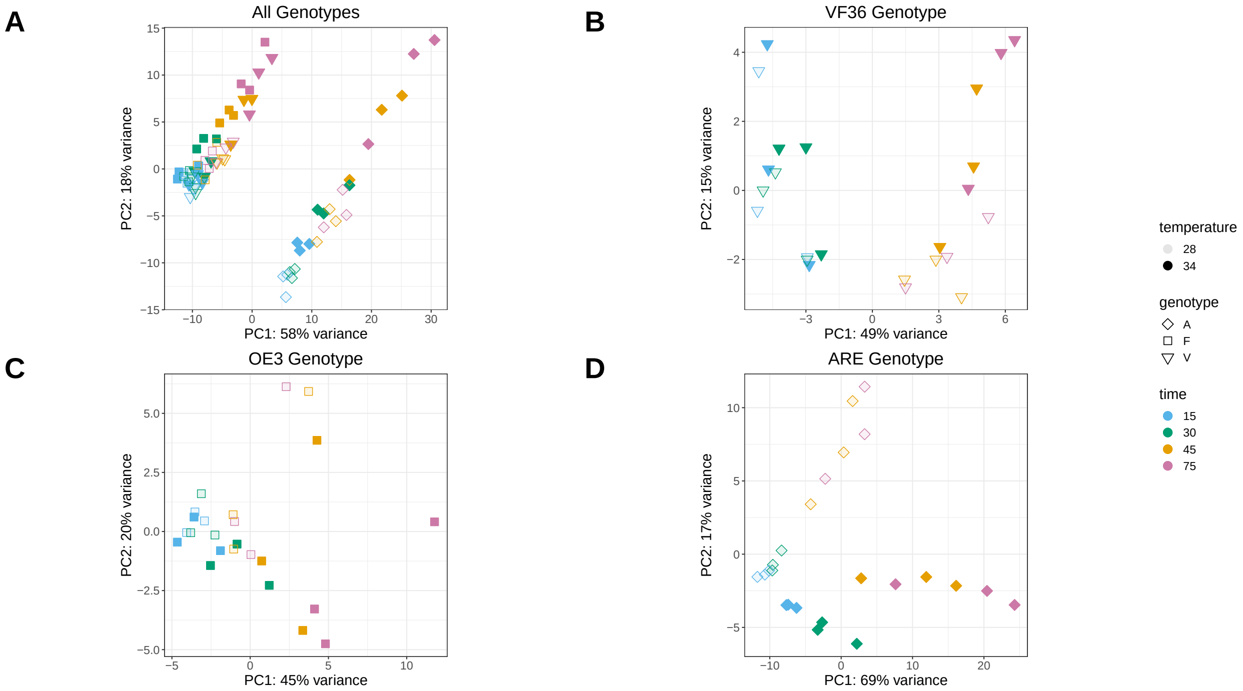
Run just VF36 genotype data and make a DESeq analysis with the design time + temperature.

Run just ARE genotype data and make a DESeq analysis with the design time + temperature.

## PCA Plots

```
combined<- ggpubr::ggarrange(All_PCA, VF36_PCA, OE3_PCA, ARE_PCA, # list of plots
                  labels = "AUTO",
                  font.label = list(size = 30),
                  common.legend = T, # COMMON LEGEND
```

```
                    legend = "right", # legend position
                    align = "hv", # Align them both, horizontal and vertical
                    nrow = 2,
                    ncol = 2)
combined
```



A) All of the genotypes are plotted together on the first figure and were also all analyzed together. There is an obvious separation between ARE and the other two genotypes OE3 and VF36.

B) The VF36 genotype PC1 cleanly separates points into two visually obvious groups, one containing only earlier-stage sample points and the second containing only-later-stage points

C) This plot and the code used to produce it included OE3 genotype samples only and had no clear indication of clustering or separation of points except for maybe a small cluster for 75 minutes.

D) The ARE genotype has an obvious visual separation between the treatment (heat stress) samples and the control (no-heat stress sample) in the PC2 dimension. There is also a clear grouping based on early-versus-later sample points in the PC1 dimension.

```
# Save pca plots as a PDF
setwd("../72_F3H_PollenTube/")
pdf(file= "Muday-144-Combined-PCA-Plots.pdf", width=18, height=10)
combined
dev.off()
```

```
## pdf
##   2
```

---

## Discussion

The samples used in this analysis have significant results regarding separation by PC1 and PC2. The ARE genotype contains the most significant samples due to having the highest variance in PC1. ARE samples at 34 degrees Celsius around 45 to 75 minutes had the largest variance of all other samples. This makes the samples the most significant in the dataset.

---

## Conclusion

The ARE genotype had the most significant samples with high variation and clear separation for warmer temperatures over long periods of time.