Khan Inan
BI-GY 7653 NGS
Week 9 hw assignment

Q1.1. Report the output indicating what type of object the dds variable is, how many genes, and how many samples are stored in the object? [ 1 point ]

```
class: DESeqDataSet
dim: 28595 8
#Dimension is 28595 genes by 8 samples
metadata(1): version
assays(1): counts
rownames(28595): Pdac_HC_000007FG0000100 Pdac_HC_000007FG0000200 ... Pdac_HC_chr9G0155000
  Pdac_HC_chr9G0155100
rowData names(0):
colnames(8): PDAC253.htseq_count.txt PDAC282.htseq_count.txt ... PDAC306.htseq_count.txt
  PDAC318.htseq_count.txt
colData names(1): condition
```

Q2.1 Enter `dds` at the console to summarize your object. Report the output for your answer and how many genes were retained after removing the low count genes [ 1 point ].

```
class: DESeqDataSet
dim: 20029 8
#20029 genes were retained after removing the low count genes
metadata(1): version
assays(4): counts mu H cooks
rownames(20029): Pdac_HC_000007FG0000100 Pdac_HC_000007FG0000200 ... Pdac_HC_chr9G0154800
  Pdac_HC_chr9G0155000
rowData names(22): baseMean baseVar ... deviance maxCooks
colnames(8): PDAC253.htseq_count.txt PDAC282.htseq_count.txt ... PDAC306.htseq_count.txt
  PDAC318.htseq_count.txt
colData names(2): condition sizeFactor
```

Q3.1 Include the hierarchical clustering result, your plotPCA command, and the PCA plot for you answer [ 1 point ].

```
rld <- rlog(dds)
dists <- dist(t(assay(rld)))
plot(hclust(dists))
plotPCA(rld)
```

Q3.2a Do the samples cluster by sugar composition phenotype in the hierarchical clustering? Explain.

**I would say no because hierarchical clustering is not really used for any kind of stats analysis. However, it does provide us distances between profiles of gene expression, so we can come to a better understanding of the similarity and difference levels between different profiles.**

Q3.2b Does the PCA separate samples by sugar composition? If so, on which axis?

a. No. The PCA does not separate the treatments on either access.
b. Yes. The PCA separates all points on the PC1 axis
c. Yes. The PCA mostly separates points on PC1 except one outlier.
d. Yes. The PCA separates all points on the PC2 axis.
e. Yes. The PCA mostly separates points on PC2 except one outlier.
**f. <u>Yes. The PCA separates points on both PC1 and PC2 axes.</u>**
g. None of the above.

Q3.2c In many contexts, the treatment will induce large effects on the transcriptome that often cause biological replicates from the same treatment to cluster together. Another factor that could cause samples to cluster together is batch effects. Which of the following are potential sources of batch effects in RNA-seq analysis?

a. sequencer effects (libraries sequenced on different sequencing machines)
b. lane effects (libraries sequenced on different lanes)
c. day effects (libraries are prepared on different days)
d. reagent effects (libraries are prepared from different batches of reagents)
e. technician effect (libraries are prepared by different technicians)
**f. <u>all of the above</u>**

Q3.2d Without any additional information about how the experiment was conducted, do you see any evidence of a batch effect? Why or why not?

**In terms of evidence of the batch effect, it is present on one of the groups that has an outlier sample. The low sucrose group has several samples that seem to be clustered together, while one of the samples is farther away from the rest which suggests that this group has been affected by the batch effect in some way**

Q3.3 Read the section “More information on results columns” of the DESeq2 vignette. Which of the following are “independent filtering” steps are taken by DESeq2 to automatically to drop from consideration genes with suspect, or problematic, p-values? Select all that apply [ 1 point ].

**a. <u>Set to NA (missing data) the log2 fold change estimates, p values and adjusted p values for genes with zero counts across all samples and baseMean = 0</u>**
b. Set to NA the p values and adjusted p values of genes with high mean expression as defined by a Z-score greater than 3
**c. <u>Set to NA the adjusted p value of genes with a low mean normalized count.</u>**
**d. <u>Set to NA the p values and adjusted p values of genes with extreme outlier counts (as detected by Cook’s distance)</u>**
e. None of the above.

Q3.4a. Report the table of results for the three candidate genes. For each of the candidate genes, which sugar-type (high sucrose or low sucrose) has higher expression? What is fold-change expressed on the decimal scale (convert from log2 to linear scale)? [ 1 point ]

```
                     baseMean log2FoldChange     lfcSE      stat      pvalue        padj
                    <numeric>      <numeric> <numeric> <numeric>   <numeric>   <numeric>
Pdac_HC_chr14G0023100  2915.858       5.653391  0.630509   8.96639 3.06399e-19 2.62293e-15
Pdac_HC_chr14G0022900  2839.320       9.978814  1.205430   8.27822 1.25031e-16 4.28130e-13
Pdac_HC_chr14G0028200   590.073       0.429695  0.249691   1.72091 8.52678e-02 2.95845e-01

#For each gene: Pdac_HC_chr14G0022900, Pdac_HC_chr14G0023100 and Pdac_HC_chr14G0028200, low sucrose has higher expression co
mpared to high sucrose since the log2fold change values are greater than 1 and the numerator is lowsucrose while the denomin
ator is highsucrose. The fold-change expressed on the decimal scale:

Pdac_HC_chr14G0023100 2^5.653391= 50.33155
Pdac_HC_chr14G0022900 2^9.978814= 1009.072
Pdac_HC_chr14G0028200 2^0.429695= 1.346949

# Fold-change expressed on the decimal scale for Pdac_HC_chr14G0023100 is 50.33155, for Pdac_HC_chr14G0022900 is 1009.072 an
d for Pdac_HC_chr14G0028200 is 1.346949.
```

Q3.4b Sometimes its useful to report the normalized counts for each gene as a figure or table. For your answer, report the normalized counts for each candidate gene as a table [ 1 point ]

```
#For Pdac_HC_chr14G002820 gene:
                          count   condition
PDAC253.htseq_count.txt 365.4494 highSucrose
PDAC282.htseq_count.txt 544.6490 highSucrose
PDAC286.htseq_count.txt 624.9751 highSucrose
PDAC316.htseq_count.txt 475.6036 highSucrose
PDAC266.htseq_count.txt 727.8490  lowSucrose
PDAC273.htseq_count.txt 524.2659  lowSucrose
PDAC306.htseq_count.txt 701.2766  lowSucrose
PDAC318.htseq_count.txt 760.5131  lowSucrose

#For Pdac_HC_chr14G0023100 gene:
                            count   condition
PDAC253.htseq_count.txt   22.43685 highSucrose
PDAC282.htseq_count.txt  106.79149 highSucrose
PDAC286.htseq_count.txt  269.12122 highSucrose
PDAC316.htseq_count.txt   57.73196 highSucrose
PDAC266.htseq_count.txt 9356.85264  lowSucrose
PDAC273.htseq_count.txt 6044.40094  lowSucrose
PDAC306.htseq_count.txt 2713.47034  lowSucrose
PDAC318.htseq_count.txt 4760.05711  lowSucrose

#For Pdac_HC_chr14G0022900 gene:
                              count   condition
PDAC253.htseq_count.txt    2.494259 highSucrose
PDAC282.htseq_count.txt    2.086440 highSucrose
PDAC286.htseq_count.txt   19.885037 highSucrose
PDAC316.htseq_count.txt    0.500000 highSucrose
PDAC266.htseq_count.txt 9441.867455  lowSucrose
PDAC273.htseq_count.txt 6666.762989  lowSucrose
PDAC306.htseq_count.txt 2941.984454  lowSucrose
PDAC318.htseq_count.txt 3642.977490  lowSucrose
```

Q3.5 Use the plotMA function to generate MA plots for the `res` and `res.shrunk` objects. Report the plots in your assignment document and explain your observations. Why it is appropriate to report the shrunken estimates. [ 1 point ].

**It is important because the log2 fold values tend to be over exaggerated and may provide inaccurate values for genes with low expressions. So shrinking this values results in a more accurate estimate where conclusions can be drawn from.**

Q4.1a. Report your results from the lfcShrink output table for the candidate genes.

```
res.shrunkOrdered[ row.names(resOrdered) %in% c('Pdac_HC_chr14G0022900','Pdac_HC_chr14G0023100','Pdac_HC_chr14G0028200'), ]

log2 fold change (MMSE): condition lowSucrose vs highSucrose
Wald test p-value: condition lowSucrose vs highSucrose
DataFrame with 3 rows and 5 columns
                      baseMean log2FoldChange     lfcSE      pvalue        padj
                     <numeric>      <numeric> <numeric>   <numeric>   <numeric>
Pdac_HC_chr14G0023100  2915.858       5.372701  0.629431 3.06399e-19 2.62293e-15
Pdac_HC_chr14G0022900  2839.320       8.824298  1.168327 1.25031e-16 4.28130e-13
Pdac_HC_chr14G0028200   590.073       0.324961  0.217307 8.52678e-02 2.95845e-01
```

Q4.1b. Statisticians make the distinction between statistically signficant and biologically significant. Using a criterion that a statistically differentially expressed gene **must also show at least a two-fold change in expression (on linear scale) to be biologically meaningful**, which genes do you consider to be differentially expressed? Please specify the gene(s), the FDR threshold you applied.

**The genes I selected are Pdac_HC_chr14G0022900 and Pdac_HC_chrg0023100. These two genes are expressed differentially and they have a value for the high log2 FC that is greater than 1. This means that in the statistical sense these genes are differentially expressed and there is a two-fold change in expression. These genes have a low adjusted p-value. Since I wanted to be strict in terms of which genes I select, the FDR threshold that I selected is 0.01. With this selected threshold the chance of a gene being a false positive is around 1%. Both my selected genes have a adjusted p value that is larger than 0.01 and they also have a log2FC that is larger than 1, so it is safe to say that these genes are differentially expressed compared to other genes that do not meet these prerequisites.**

Q4.1c Which of the candidate genes do you think could be responsible for the sugar composition trait on the chromosome 14 sugar QTL?

**I would say that the two genes I previously selected (Pdac_HC_chr14G0022900 and Pdac_HC_chrg0023100) can be considered responsible for the sugar composition trait. This is mainly because I was able to show that these genes are indeed differentially expressed. Gene Pdac_HC_chr14G0022900 has a log value of 8.824 which is fairly higher compared to the other gene but since both genes do have a log value of greater than one, they can be considered to be responsible for that specific trait**