# Project 1

Lily Hoefner 3/14/2021

## Introduction

Datasets: The four datasets in this project are very old datasets produced by my FRI lab. They look at the effects of deleting the Htz1 gene in yeast. The nonessential Htz1 gene codes for a histone variant, H2Az, which DNA wraps around to maintain its structure. The datasets are for yeast that: Have been heatshocked and have the Htz1 gene, have been heatshocked and do not have the Htz1 gene, have been treated with rapamycin and have the Htz1 gene, have been treated with rapamycin and do not have the Htz1 gene. These datasets interested me because my lab studies yeast histone genes, and I am curious about the effects of Rapamycin on histones. Rapamycin is drug currently being studied as a treatment for human histone-related neurological diseases like Huntington's Disease.

Variables: The variables include the name of the gene (Gene_Name), gene starting position(Gene_Start), the gene stopping position(Gene_Stop), the chromosome the gene is located on (Chromosome), transcription levels (Count), and whether the gene is on the positive or negative strand (Strand).

Predictions: I predict that the deletion of the Htz1 gene has an impact on the transcription levels of other genes in the yeast genome. More specifically, I predict that other genes in the genome will be downregulated by the deletion of Htz1. I think the yeast treated with Rapamycin will be more affected by the deletion of Htz1 than the yeast treated with heatshock.

## Loading the packages

```
library(tidyverse)
```

```
## -- Attaching packages ------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.1     v dplyr   1.0.0
## v tidyr   1.1.2     v stringr 1.4.0
```

```
## v readr   1.3.1     v forcats 0.5.0

## -- Conflicts ---------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(purrr)
library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'

## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
library(psych)
```

```
##
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve
```

# Importing the datasets

```
#Import the datasets saved as .csv files on my computer
WT<-read.csv(file='WT.csv')
Htz1<- read.csv(file = 'Htz1.csv')
RapWT<-read.csv(file='RapWT.csv')
RapHtz1<-read.csv(file='RapHtz1.csv')
```

# Tidying the data

```
#The data did not need tidying, but I did remove the extra columns before joining
Htz1<-Htz1 %>%
  select(Gene_Name,Count)
RapHtz1<-RapHtz1 %>%
  select(Gene_Name, Count)
RapWT<-RapWT%>%
  select(Gene_Name, Count)

#Join the four datasets by the common variable "Gene_Name"
Data<-WT%>%
  left_join(Htz1, by="Gene_Name")%>%
  rename(Count_WT=Count.x)%>%
  rename(Count_Htz1=Count.y)%>%
  left_join(RapHtz1, by="Gene_Name")%>%
  rename(Count_RapHtz1=Count)%>%
  left_join(RapWT, by="Gene_Name")%>%
  rename(Count_RapWT=Count)
head(Data)
```

```
##   Chromosome Transcript_Start Transcript_Stop Gene_Name Ignore_Color_Score
## 1       chrI             1664            2229   YAL068C                251
## 2       chrI             2419            2867 YAL067W-A                111
## 3       chrI             7057            9191   YAL067C                255
## 4       chrI            11404           12011   YAL065C                111
## 5       chrI            11985           12586 YAL064W-B                111
## 6       chrI            13202           13803 YAL064C-A                111
##   Strand Gene_Start Gene_Stop Count_WT Count_Htz1 Count_RapHtz1 Count_RapWT
## 1      -       1806      2169        0          0             0           0
## 2      +       2479      2707        0          0             0           0
## 3      -       7234      9016        1          0             0           2
## 4      -      11564     11951        7          0             6           1
## 5      +      12045     12426        2          1             0           3
## 6      -      13362     13743       12          7             5          11
```

# Exploring the data with dplyr functions

```
# Use mutate to create a column for gene length, use mutate to create a column for th
Data<-Data%>%
  mutate(Gene_Length = abs(Gene_Start-Gene_Stop), Normalized_WT=Count_WT/Gene_Length,
  select(-Ignore_Color_Score, -Count_Htz1, -Count_WT, -Count_RapWT, -Count_RapHtz1)
```

```
#Create a column for the difference between Normalized_WT and Normalized Htz1, create
Data<-Data%>%
  mutate(Heat_Difference=Normalized_WT-Normalized_Htz1, Rap_Difference=Normalized_Rap
  mutate(Heat_Regulation=case_when(Heat_Difference>0~"down", Heat_Difference<0~"up"),
head(Data)
```

```
##   Chromosome Transcript_Start Transcript_Stop Gene_Name Strand Gene_Start
## 1       chrI             1664            2229   YAL068C      -       1806
## 2       chrI             2419            2867 YAL067W-A      +       2479
## 3       chrI             7057            9191   YAL067C      -       7234
## 4       chrI            11404           12011   YAL065C      -      11564
## 5       chrI            11985           12586 YAL064W-B      +      12045
## 6       chrI            13202           13803 YAL064C-A      -      13362
##   Gene_Stop Gene_Length Normalized_WT Normalized_Htz1 Normalized_RapWT
## 1      2169         363  0.0000000000     0.000000000      0.000000000
## 2      2707         228  0.0000000000     0.000000000      0.000000000
## 3      9016        1782  0.0005611672     0.000000000      0.001122334
## 4     11951         387  0.0180878553     0.000000000      0.002583979
## 5     12426         381  0.0052493438     0.002624672      0.007874016
## 6     13743         381  0.0314960630     0.018372703      0.028871391
##   Normalized_RapHtz1 Heat_Difference Rap_Difference Heat_Regulation
## 1         0.00000000    0.0000000000    0.000000000            <NA>
## 2         0.00000000    0.0000000000    0.000000000            <NA>
## 3         0.00000000    0.0005611672    0.001122334            down
## 4         0.01550388    0.0180878553   -0.012919897            down
## 5         0.00000000    0.0026246719    0.007874016            down
## 6         0.01312336    0.0131233596    0.015748031            down
##   Rap_Regulation
## 1           <NA>
## 2           <NA>
## 3           down
## 4             up
## 5           down
## 6           down
```

```
#Filter for genes that were upregulated by the deletion of Htz1.
#More genes than I expected were upregulated by the deletion of Htz1.  However, some
Upregualted<-Data%>%
  filter(Heat_Regulation=="up"|Rap_Regulation=="up")%>%
  select(Gene_Name, Chromosome, Heat_Regulation, Rap_Regulation)
head(Upregualted)
```

```
##   Gene_Name Chromosome Heat_Regulation Rap_Regulation
## 1   YAL065C       chrI            down            up
```

```
## 2   YAL063C       chrI            down             up
## 3   YAL060W       chrI              up           down
## 4   YAL059W       chrI            down             up
## 5   YAL054C       chrI            down             up
## 6   YAL046C       chrI              up           <NA>
```

```
#Filter for genes on chromosome M and select the regulation columns.
#INTERSTING DISCOVERY: MITOCHONDRIAL DNA IS ONLY UPREGULATED (NEVER DOWNREGULATED) BY
ChromosomeM<-Data%>%
  filter(Chromosome=="chrM")%>%
  select(Gene_Name, Heat_Regulation, Rap_Regulation)
head(ChromosomeM)
```

```
##   Gene_Name Heat_Regulation Rap_Regulation
## 1     Q0055              up             up
## 2     Q0065              up             up
## 3     Q0045              up             up
## 4     Q0070              up             up
## 5     Q0050              up             up
## 6     Q0060              up             up
```

```
#Arrange the Data to see which heatshocked genes were most up or down regulated by th
#Looking into the functions of these genes could reveal more about the role of Htz1 i
Data1<-Data%>%
  select(Gene_Name, Heat_Difference, Rap_Difference)
ArrangedbyHeat<-Data1%>%
  arrange(desc(Heat_Difference))
head(ArrangedbyHeat)
```

```
##   Gene_Name Heat_Difference Rap_Difference
## 1   YPL144W       17.391499      12.892617
## 2 YJL047C-A       11.762963       9.474074
## 3   YAL003W       11.602837      10.405268
## 4   YGL076C        9.036703       9.575812
## 5   YHR174W        9.029680       4.907915
## 6 YMR194C-B        7.806122       8.020408
```

```
#Arrange the Data to see which Rapamycin treated genes were most up or down regulated
ArrangedbyRap<-Data1%>%
  arrange(desc(Rap_Difference))
head(ArrangedbyRap)
```

```
##   Gene_Name Heat_Difference Rap_Difference
## 1   YPL144W       17.391499      12.892617
## 2   YAL003W       11.602837      10.405268
## 3   YGL076C        9.036703       9.575812
## 4 YJL047C-A       11.762963       9.474074
## 5 YMR194C-B        7.806122       8.020408
## 6   YHR174W        9.029680       4.907915
```

```
#Use group_by and summarize to see the mean normalized heatshocked WT, heatschocked d
#As seen below, genes on chromosome 1 experienced the most downregulation with the de
MeansbyChromosome<-Data%>%
  group_by(Chromosome)%>%
  summarise(Mean_WT_Count= mean(Normalized_WT), Mean_Htz1_Count=mean(Normalized_Htz1)
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
head(MeansbyChromosome)
```

```
## # A tibble: 6 x 5
##   Chromosome Mean_WT_Count Mean_Htz1_Count Mean_RapWT_Count Mean_RapHtz1_Count
##   <chr>              <dbl>           <dbl>            <dbl>              <dbl>
## 1 chrI              0.399           0.168            0.291              0.115
## 2 chrII             0.112           0.0882           0.0958             0.0742
## 3 chrIII            0.144           0.105            0.150              0.120
## 4 chrIV             0.0934          0.0709           0.105              0.0858
## 5 chrIX             0.0489          0.0349           0.0653             0.0456
## 6 chrM              0.202           0.345            0.198              0.278
```

```
Differences<-MeansbyChromosome%>%
  mutate(Heat_Difference=Mean_WT_Count-Mean_Htz1_Count, Rap_Difference=Mean_RapWT_Cou
  select(Chromosome, Heat_Difference, Rap_Difference)
Differences
```

```
## # A tibble: 17 x 3
##    Chromosome Heat_Difference Rap_Difference
##    <chr>                <dbl>          <dbl>
```

```
##  1 chrI               0.231          0.176
##  2 chrII              0.0241         0.0216
##  3 chrIII             0.0396         0.0300
##  4 chrIV              0.0225         0.0192
##  5 chrIX              0.0139         0.0197
##  6 chrM              -0.143         -0.0806
##  7 chrV               0.0382         0.0263
##  8 chrVI              0.0134         0.0104
##  9 chrVII             0.0519         0.0489
## 10 chrVIII            0.0568         0.0421
## 11 chrX               0.0639         0.0613
## 12 chrXI              0.0476         0.0469
## 13 chrXII             0.0469         0.0295
## 14 chrXIII            0.0542         0.0502
## 15 chrXIV             0.0169         0.0194
## 16 chrXV              0.0356         0.0280
## 17 chrXVI             0.0805         0.0553
```

# Summary statistics for each numeric variable

```
#After removing the non-numeric variables, I used the describe() function to get the
#The mean column of the table says that mean heat difference is greater than mean Rap

SummaryStats<-Data%>%
  select(-Chromosome, -Gene_Name, -Strand, -Heat_Regulation, -Rap_Regulation)%>%
  describe()%>%
  select(-vars, -trimmed, -mad, -kurtosis, -se)

SummaryStats %>%
  kbl() %>%
  kable_styling()
```

| | n | mean | sd | median |
|---|---|---|---|---|
| Transcript_Start | 5816 | 4.528086e+05 | 3.219520e+05 | 3.978335e+05 |
| Transcript_Stop | 5816 | 4.545215e+05 | 3.219491e+05 | 3.992275e+05 |
| Gene_Start | 5816 | 4.529275e+05 | 3.219549e+05 | 3.979535e+05 |
| Gene_Stop | 5816 | 4.543988e+05 | 3.219527e+05 | 3.990920e+05 |
| Gene_Length | 5816 | 1.471338e+03 | 1.130428e+03 | 1.200000e+03 |
| Normalized_WT | 5816 | 1.301832e-01 | 7.472044e-01 | 2.636970e-02 |

| | n | mean | sd | median | |
|---|---|---|---|---|---|
| Normalized_Htz1 | 5816 | 8.606390e-02 | 4.445147e-01 | 1.783260e-02 | |
| Normalized_RapWT | 5816 | 1.209870e-01 | 5.707076e-01 | 3.231850e-02 | |
| Normalized_RapHtz1 | 5816 | 8.426690e-02 | 3.520602e-01 | 2.304810e-02 | |
| Heat_Difference | 5816 | 4.411930e-02 | 4.279233e-01 | 7.130100e-03 | |
| Rap_Difference | 5816 | 3.672010e-02 | 3.379507e-01 | 7.560100e-03 | |

# Summary statistics for variables grouped by Chromosome

```
#Here I grouped by the categorical variable "Chromosome".  I chose to look at summary
#We can see that Chromosome I and Chromosome M had the highest mean Heat_Difference a
SummaryStatsGrouped<-Data %>%
  select(-Transcript_Start, -Transcript_Stop, -Strand, -Gene_Name, -Gene_Start, -Gene
  filter(Chromosome==c("chrI", "chrM", "chrX", "chrXVI"))

Stats<-describeBy(SummaryStatsGrouped,SummaryStatsGrouped$Chromosome)
Stats
```

```
##
##  Descriptive statistics by group
## group: chrI
##                    vars  n    mean      sd  median trimmed     mad min     max
## Chromosome*           1 23    1.00    0.00    1.00    1.00    0.00   1    1.00
## Gene_Length           2 23 1452.22 1071.07 1242.00 1300.05 1018.55 300 4068.00
## Normalized_WT         3 23    0.42    1.91    0.02    0.02    0.02   0    9.19
## Normalized_Htz1       4 23    0.26    1.15    0.01    0.01    0.01   0    5.53
## Normalized_RapWT      5 23    0.11    0.36    0.02    0.03    0.02   0    1.74
## Normalized_RapHtz1    6 23    0.07    0.26    0.01    0.02    0.02   0    1.24
## Heat_Difference       7 23    0.17    0.76    0.01    0.01    0.01   0    3.66
## Rap_Difference        8 23    0.03    0.10    0.01    0.01    0.01   0    0.50
##                     range skew kurtosis     se
## Chromosome*          0.00  NaN      NaN   0.00
## Gene_Length        3768.00 1.06     0.30 223.33
## Normalized_WT         9.19 4.19    16.25   0.40
## Normalized_Htz1       5.53 4.19    16.24   0.24
## Normalized_RapWT      1.74 4.10    15.77   0.07
## Normalized_RapHtz1    1.24 4.12    15.87   0.05
## Heat_Difference       3.67 4.19    16.25   0.16
## Rap_Difference        0.50 3.99    15.09   0.02
```

```
## ------------------------------------------------------------
## group: chrM
##                     vars n    mean      sd  median trimmed     mad    min
## Chromosome*            1 5    1.00    0.00    1.00    1.00    0.00   1.00
## Gene_Length            2 5 2556.00 1857.59 2505.00 2556.00 2557.48 756.00
## Normalized_WT          3 5    0.29    0.19    0.28    0.29    0.13   0.05
## Normalized_Htz1        4 5    0.54    0.52    0.43    0.54    0.26   0.07
## Normalized_RapWT       5 5    0.27    0.18    0.28    0.27    0.18   0.05
## Normalized_RapHtz1     6 5    0.41    0.29    0.39    0.41    0.21   0.07
## Heat_Difference        7 5   -0.25    0.34   -0.15   -0.25    0.13  -0.85
## Rap_Difference         8 5   -0.14    0.11   -0.11   -0.14    0.07  -0.31
##                       max   range  skew kurtosis     se
## Chromosome*          1.00    0.00   NaN      NaN   0.00
## Gene_Length        5013.00 4257.00  0.16    -2.01 830.74
## Normalized_WT         0.56    0.51  0.20    -1.63   0.09
## Normalized_Htz1       1.41    1.34  0.78    -1.23   0.23
## Normalized_RapWT      0.52    0.47  0.08    -1.78   0.08
## Normalized_RapHtz1    0.83    0.76  0.29    -1.65   0.13
## Heat_Difference      -0.03    0.83 -0.99    -1.02   0.15
## Rap_Difference       -0.03    0.28 -0.63    -1.39   0.05
## ------------------------------------------------------------
## group: chrX
##                     vars  n    mean      sd  median trimmed    mad   min     max
## Chromosome*            1 88    1.00    0.00    1.00    1.00   0.00  1.00    1.00
## Gene_Length            2 88 1529.15 1083.53 1301.00 1399.67 746.49 51.00 7413.00
## Normalized_WT          3 88    0.14    0.47    0.02    0.04   0.02  0.00    2.96
## Normalized_Htz1        4 88    0.10    0.31    0.02    0.03   0.02  0.00    2.11
## Normalized_RapWT       5 88    0.16    0.44    0.03    0.05   0.03  0.00    2.50
## Normalized_RapHtz1     6 88    0.10    0.22    0.02    0.04   0.02  0.00    1.12
## Heat_Difference        7 88    0.05    0.18    0.01    0.01   0.01 -0.07    1.51
## Rap_Difference         8 88    0.07    0.24    0.01    0.01   0.01 -0.10    1.72
##                      range skew kurtosis     se
## Chromosome*           0.00  NaN      NaN   0.00
## Gene_Length        7362.00 2.29     8.86 115.51
## Normalized_WT         2.96 4.86    23.66   0.05
## Normalized_Htz1       2.11 4.96    25.48   0.03
## Normalized_RapWT      2.50 3.83    14.46   0.05
## Normalized_RapHtz1    1.12 3.37    10.96   0.02
## Heat_Difference       1.58 6.32    44.86   0.02
## Rap_Difference        1.81 4.87    26.20   0.03
## ------------------------------------------------------------
## group: chrXVI
##                     vars   n    mean      sd  median trimmed    mad    min
## Chromosome*            1 114    1.00    0.00    1.00    1.00   0.00   1.00
## Gene_Length            2 114 1408.68 1060.51 1115.00 1246.00 804.31 189.00
## Normalized_WT          3 114    0.25    1.73    0.03    0.03   0.02   0.00
## Normalized_Htz1        4 114    0.09    0.37    0.01    0.02   0.02   0.00
## Normalized_RapWT       5 114    0.20    1.30    0.03    0.04   0.03   0.00
## Normalized_RapHtz1     6 114    0.08    0.30    0.02    0.03   0.02   0.00
## Heat_Difference        7 114    0.17    1.63    0.01    0.01   0.01  -0.72
## Rap_Difference         8 114    0.12    1.21    0.01    0.01   0.01  -0.27
```

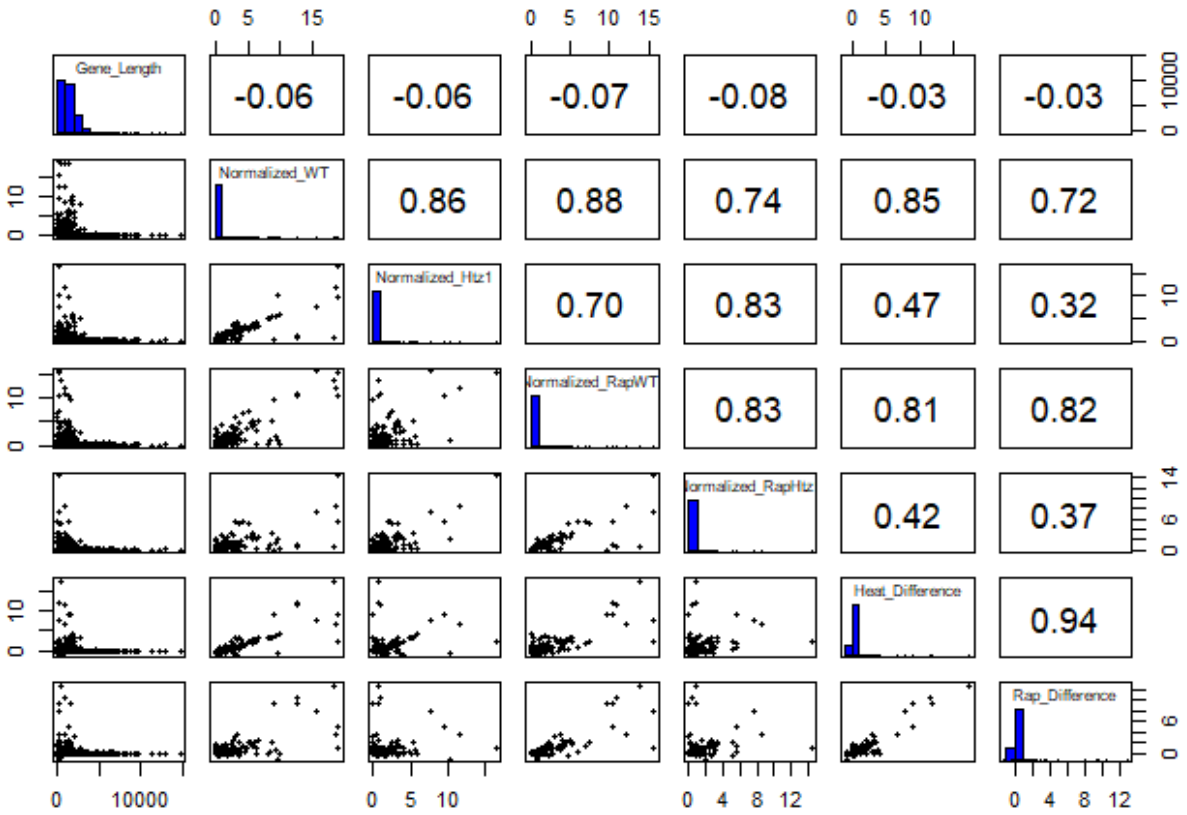
```
##                       max   range  skew kurtosis    se
## Chromosome*          1.00    0.00   NaN      NaN  0.00
## Gene_Length       5604.00 5415.00  1.50     2.29 99.33
## Normalized_WT       18.18   18.18  9.76    97.48  0.16
## Normalized_Htz1      3.50    3.50  7.52    61.72  0.03
## Normalized_RapWT    13.67   13.67  9.81    98.04  0.12
## Normalized_RapHtz1   3.01    3.01  8.65    81.07  0.03
## Heat_Difference     17.39   18.11 10.31   105.85  0.15
## Rap_Difference      12.89   13.16 10.38   106.76  0.11
```

# Correlation matrix

```
#Build a correlation matrix of the numeric variables
#I chose to remove some of the less relevant variables for simplification.  The corre
data_num <- Data %>%
  select_if(is.numeric)%>%
  select(-Gene_Start, -Gene_Stop, -Transcript_Start, -Transcript_Stop)
cor(data_num, use = "pairwise.complete.obs")
```

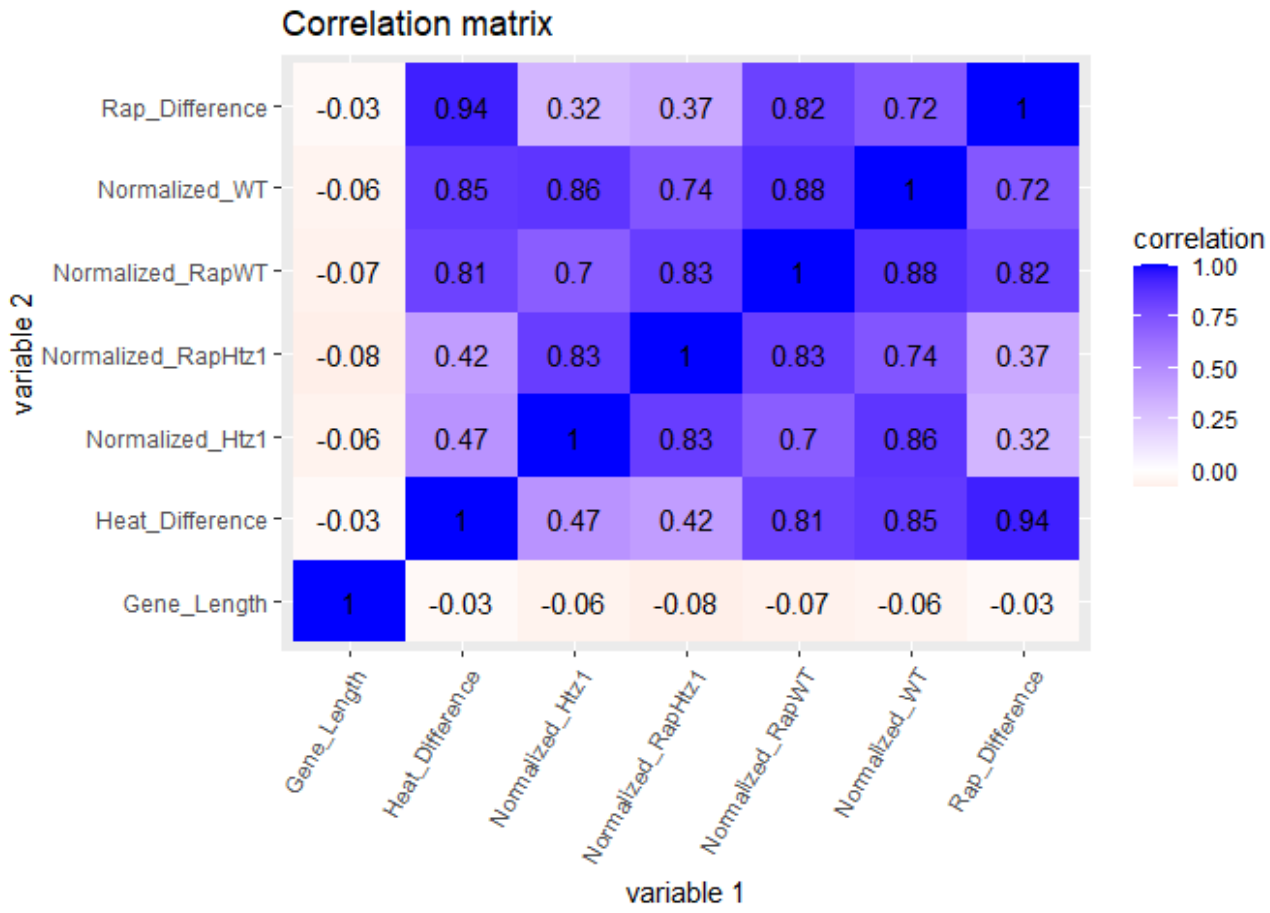
```
##                   Gene_Length Normalized_WT Normalized_Htz1 Normalized_RapWT
## Gene_Length        1.00000000   -0.05588144     -0.06490737      -0.06917137
## Normalized_WT     -0.05588144    1.00000000      0.86226238       0.88285547
## Normalized_Htz1   -0.06490737    0.86226238      1.00000000       0.69989840
## Normalized_RapWT  -0.06917137    0.88285547      0.69989840       1.00000000
## Normalized_RapHtz1 -0.08352314    0.73547416      0.83158620       0.83475306
## Heat_Difference   -0.03015162    0.85042373      0.46683966       0.81453474
## Rap_Difference    -0.02980155    0.72472457      0.31563463       0.81912629
##                   Normalized_RapHtz1 Heat_Difference Rap_Difference
## Gene_Length               -0.08352314     -0.03015162    -0.02980155
## Normalized_WT              0.73547416      0.85042373     0.72472457
## Normalized_Htz1            0.83158620      0.46683966     0.31563463
## Normalized_RapWT           0.83475306      0.81453474     0.81912629
## Normalized_RapHtz1         1.00000000      0.42039603     0.36792270
## Heat_Difference            0.42039603      1.00000000     0.93758201
## Rap_Difference             0.36792270      0.93758201     1.00000000
```

```
pairs.panels(data_num,
             method = "pearson", # correlation coefficient method
             hist.col = "blue", # color of histogram
             smooth = FALSE, density = FALSE, ellipses = FALSE)
```




# Correlation heatmap

```
#Build a correlation heatmap
cor(data_num, use = "pairwise.complete.obs") %>%
  # Save as a data frame
  as.data.frame %>%
  # Convert row names to an explicit variable
  rownames_to_column %>%
  # Pivot so that all correlations appear in the same column
  pivot_longer(-1, names_to = "other_var", values_to = "correlation") %>%
  ggplot(aes(rowname, other_var, fill=correlation)) +
  # Heatmap with geom_tile
  geom_tile() +
  # Change the scale to make the middle appear neutral
  scale_fill_gradient2(low="red",mid="white",high="blue") +
  # Overlay values
  geom_text(aes(label = round(correlation,2)), color = "black", size = 4) +
  # Give title and labels
  labs(title = "Correlation matrix", x = "variable 1", y = "variable 2")+theme(axis.t
```

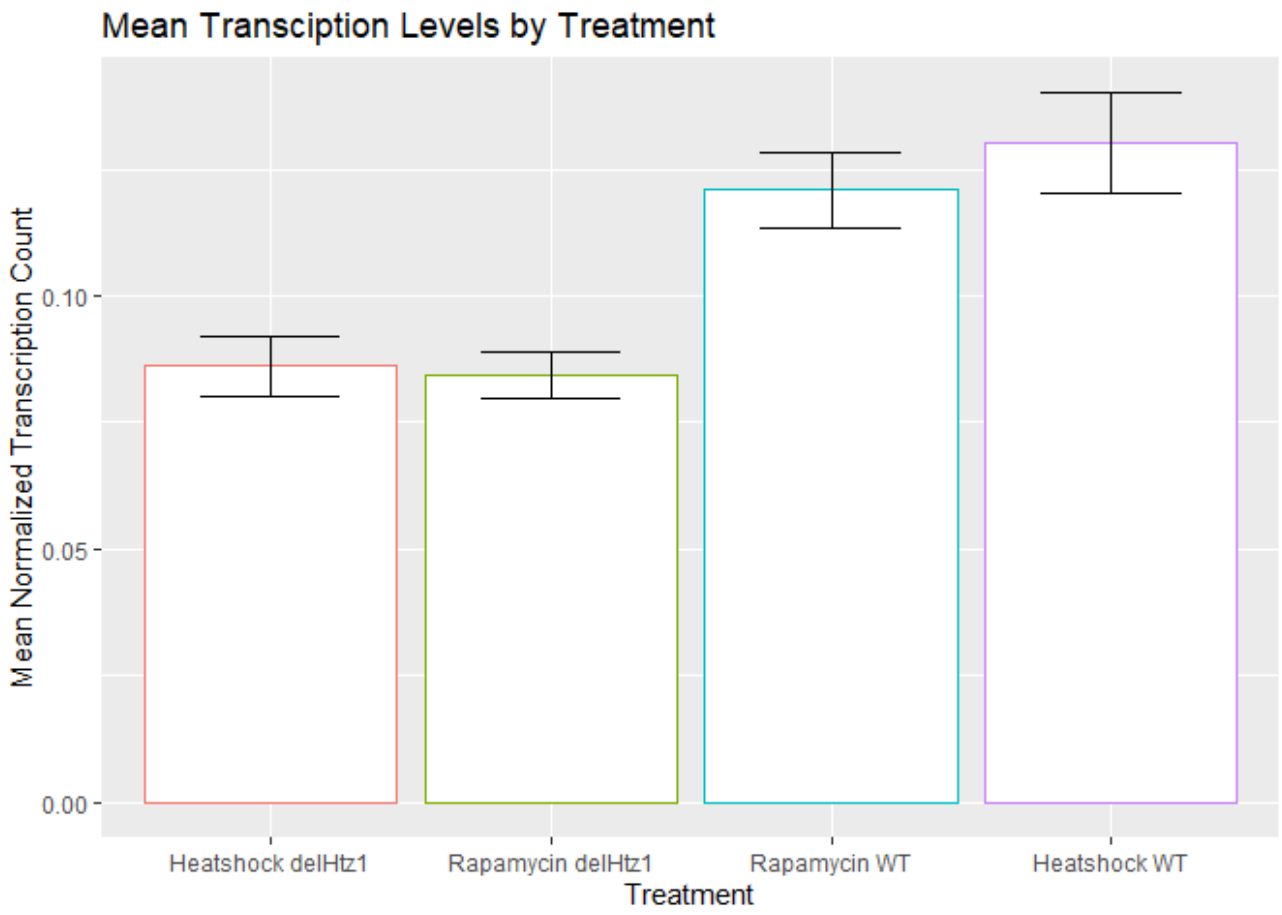



```
#As seen in the correlation heatmap, several variables had no correlation while other
```

# Visualizing the data

```
#This graph looks at each of the four categories of genes (Heatshock WT, Heatshock de
#As seen below, the Heatshock WT yeast had the highest mean transcription levels, the
Data%>%
  select(Chromosome, Normalized_WT, Normalized_Htz1, Normalized_RapWT, Normalized_Rap
  pivot_longer(cols = c(`Normalized_WT`, `Normalized_Htz1`, `Normalized_RapWT`, `Norm
  ggplot(aes(x = Treatment, color=Treatment)) +
  geom_bar(aes(y = Transcripts), stat="summary", fun=mean, fill="white", show.legend
```

```
## No summary function supplied, defaulting to `mean_se()`
```



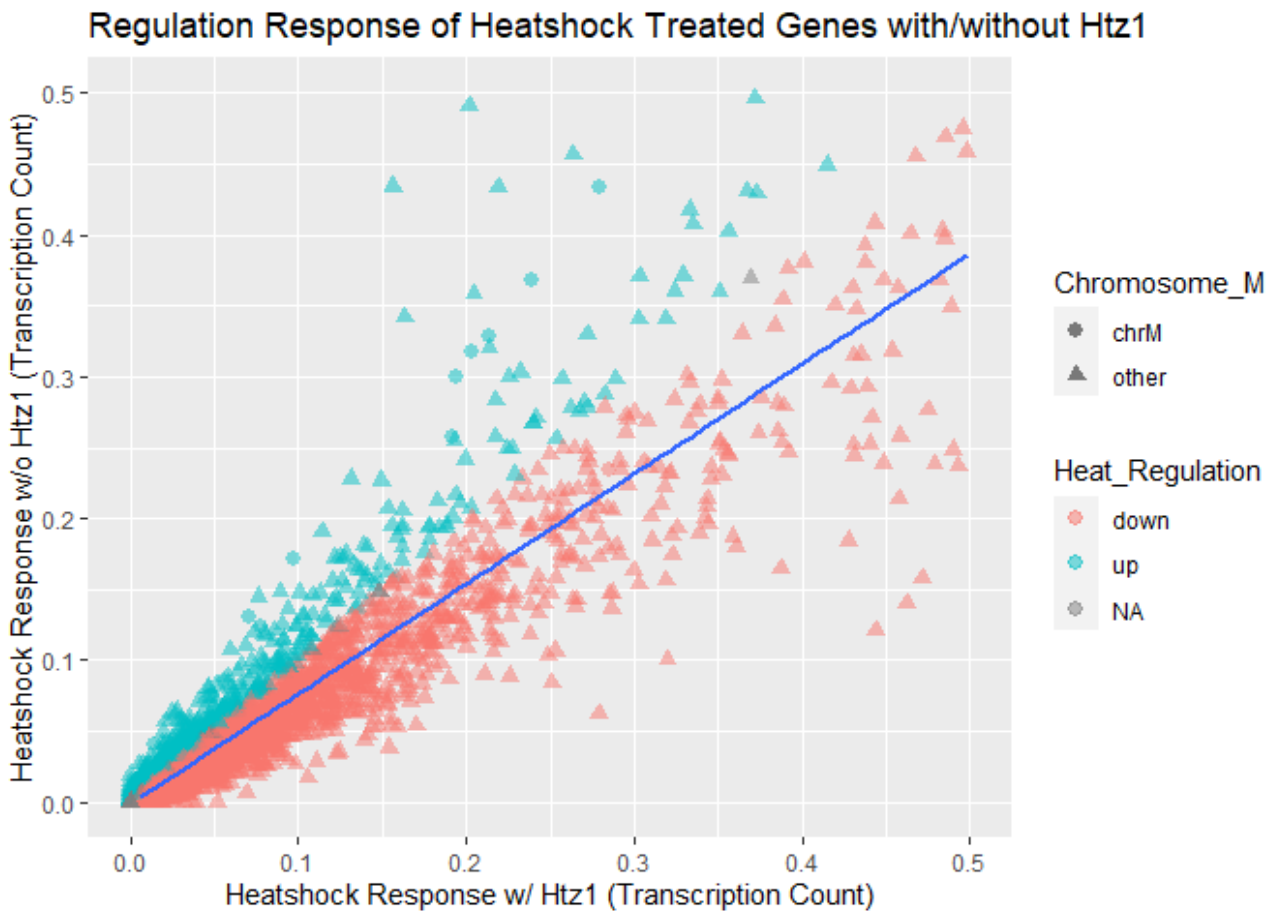

```
#This graph looks at heatshocked yeast, and plots the transcription levels of those w
#One can see that the slope of the line is less than 1, implying downregulation of ge
Data%>%
  mutate(Chromosome_M=case_when(Chromosome=="chrM"~"chrM", Chromosome!="chrM"~"other"
  ggplot(aes(x=Normalized_WT, y=Normalized_Htz1))+geom_point(alpha = 1/2, size = 2.5,
```

```
## `geom_smooth()` using formula 'y ~ x'

## Warning: Removed 207 rows containing non-finite values (stat_smooth).

## Warning: Removed 207 rows containing missing values (geom_point).

## Warning: Removed 1 rows containing missing values (geom_smooth).
```




```
#I ran a correlation test on the x and y variables to see the relationship between th
cor.test(Data$Normalized_WT, Data$Normalized_Htz1)
```

```
##
##  Pearson's product-moment correlation
##
## data:  Data$Normalized_WT and Data$Normalized_Htz1
## t = 129.82, df = 5814, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8555205 0.8687119
## sample estimates:
##       cor
## 0.8622624
```

# PCA

```
#Keep only numeric variables and scale each numeric variable.  Perform pca with the f
Datanumeric<-Data%>%
  select_if(is.numeric)%>%
  scale()
pca<-Datanumeric%>%
  prcomp()

#Visualize results of the PCA
#First we see the standard deviations for each principal component. If we squared the
#Next, we can see how each variable impacts each principal component.  For example, w
pca
```

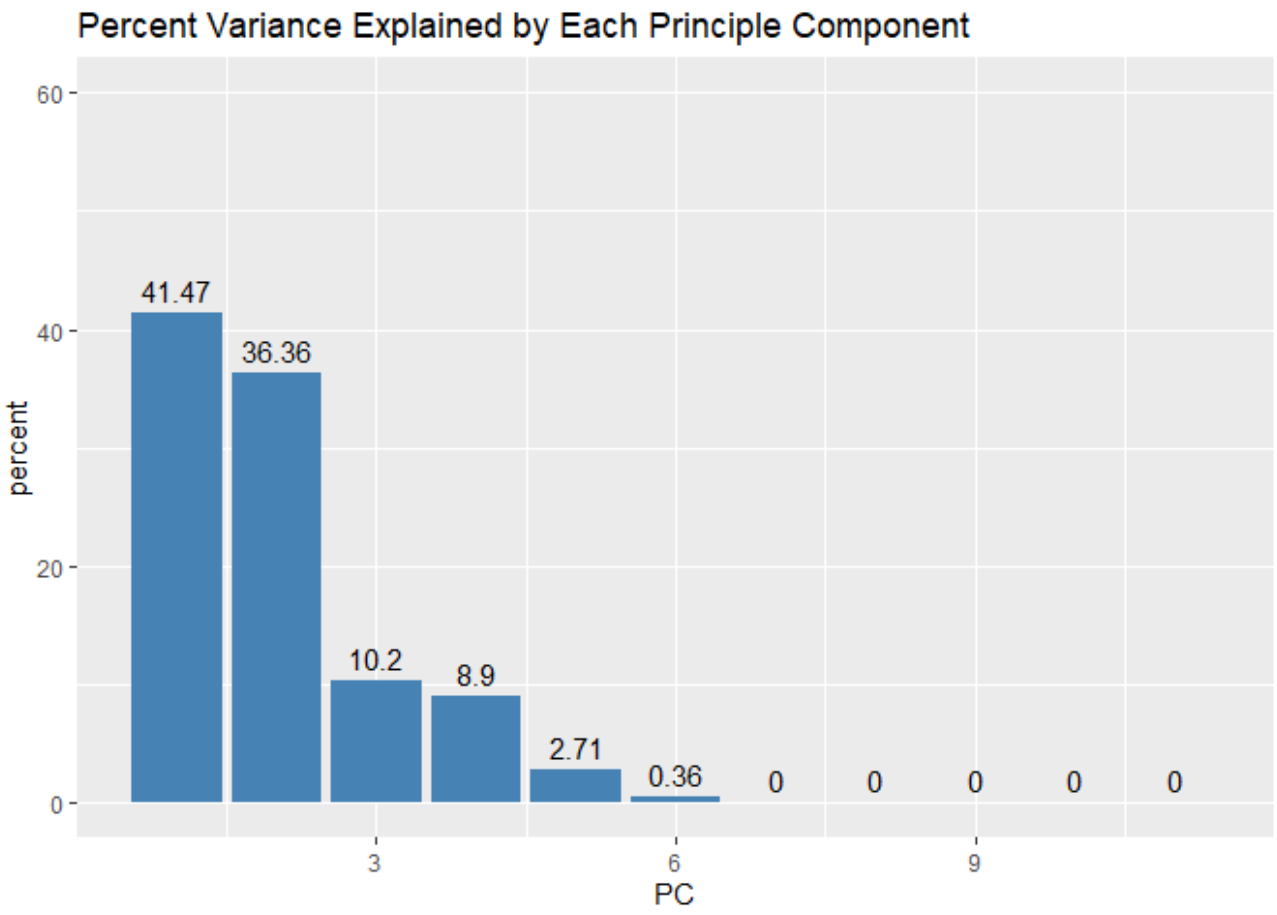
```
## Standard deviations (1, .., p=11):
##  [1] 2.135802e+00 1.999906e+00 1.059242e+00 9.894923e-01 5.462736e-01
##  [6] 1.980472e-01 4.319488e-04 2.872379e-04 3.829691e-15 2.418809e-15
## [11] 1.186485e-15
##
## Rotation (n x k) = (11 x 11):
##                            PC1          PC2          PC3          PC4
## Transcript_Start   -0.02009070 -0.499555277 -0.004129116  0.002290376
## Transcript_Stop    -0.02022426 -0.499547051 -0.005324834 -0.001047737
## Gene_Start         -0.02009261 -0.499555227 -0.004135779  0.002271298
## Gene_Stop          -0.02022636 -0.499546998 -0.005318099 -0.001030675
## Gene_Length        -0.03805427  0.003316631 -0.336723884 -0.940426388
## Normalized_WT       0.45418616 -0.015225375  0.036355243 -0.043839107
## Normalized_Htz1     0.37372239 -0.014920233  0.470493273 -0.197481260
## Normalized_RapWT    0.45350225 -0.021688135 -0.023564582  0.002898737
## Normalized_RapHtz1  0.37454443 -0.024962533  0.460740653 -0.164488894
## Heat_Difference     0.40485016 -0.011086578 -0.425254689  0.128589755
## Rap_Difference      0.37566133 -0.010620690 -0.519770933  0.176251537
##                            PC5           PC6           PC7           PC8
## Transcript_Start    0.004403572  0.0007775031  7.206104e-01 -4.803324e-01
## Transcript_Stop     0.004281360  0.0008881885 -6.930620e-01 -5.192881e-01
## Gene_Start          0.004389308  0.0008558799 -1.377612e-02  4.998084e-01
## Gene_Stop           0.004293596  0.0008569389 -1.376740e-02  4.998121e-01
## Gene_Length        -0.027268106  0.0002999400  2.509562e-03  6.810089e-05
## Normalized_WT       0.425983245 -0.1192068704 -1.649159e-05  3.655662e-06
## Normalized_Htz1     0.481808940  0.3964626926  2.256561e-05 -7.699076e-06
## Normalized_RapWT   -0.442973621  0.1394344035  1.939036e-05 -2.285821e-06
## Normalized_RapHtz1 -0.545850799 -0.3193519719 -1.483990e-05  7.136855e-06
## Heat_Difference     0.243327275 -0.6199835612 -5.223678e-05  1.438080e-05
## Rap_Difference     -0.179422856  0.5681521400  4.820457e-05 -1.129496e-05
##                             PC9          PC10          PC11
## Transcript_Start    5.056503e-14  1.956098e-13  3.932042e-14
## Transcript_Stop    -7.224494e-16 -1.899750e-13 -1.805760e-13
## Gene_Start         -7.069625e-01 -1.346175e-02  4.808581e-03
## Gene_Stop           7.069577e-01  1.346165e-02 -4.808548e-03
```

```
## Gene_Length         -2.482242e-03 -4.726603e-05  1.688359e-05
## Normalized_WT        1.277067e-02 -4.460431e-01  6.288482e-01
## Normalized_Htz1     -7.597320e-03  2.653527e-01 -3.741041e-01
## Normalized_RapWT    -8.810412e-03  6.197875e-01  4.397953e-01
## Normalized_RapHtz1   5.434999e-03 -3.823368e-01 -2.713025e-01
## Heat_Difference     -7.313752e-03  2.554485e-01 -3.601408e-01
## Rap_Difference       5.217181e-03 -3.670139e-01 -2.604295e-01
```

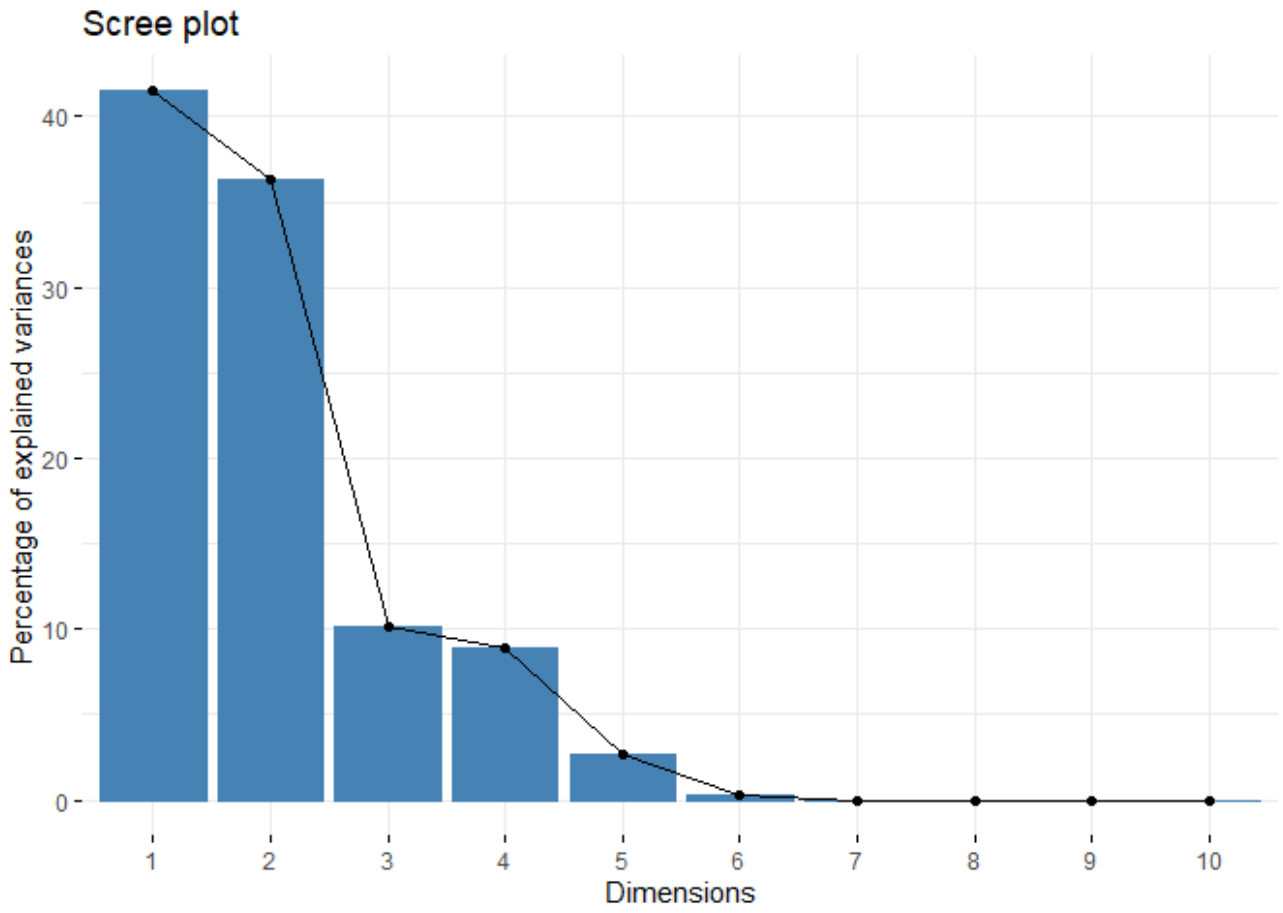
```
# Determine the percentage of variance explained by each component with sdev
percent <- 100* (pca$sdev^2 / sum(pca$sdev^2))
percent
```

```
##  [1] 4.146953e+01 3.636024e+01 1.019993e+01 8.900864e+00 2.712863e+00
##  [6] 3.565699e-01 1.696179e-06 7.500509e-07 1.333321e-28 5.318759e-29
## [11] 1.279770e-29
```

```
# Visualize the percentage of variance explained by each component
#As seen in the graph, 41.5% is explained by PC1 and 36.4% is explained by PC2.
perc_data <- data.frame(percent = percent, PC = 1:length(percent))
ggplot(perc_data, aes(x = PC, y = percent)) +
  geom_col(fill="steelblue") +
  geom_text(aes(label = round(percent, 2)), size = 4, vjust = -0.5) +
  ylim(0, 60) + ggtitle("Percent Variance Explained by Each Principle Component")
```




```
#Construct a scree plot using the package and determine how many principal components
#Based on this, I chose to use the first 2 principle components. Collectively, these
fviz_screeplot(pca)
```




```
# Visualize the rotated data
head(pca$x)
```

```
##             PC1      PC2        PC3        PC4        PC5          PC6
## [1,] -0.2693958 2.818629  0.2542834  0.9707977 0.05383369 -0.011958731
## [2,] -0.2650142 2.814192  0.2944585  1.0831134 0.05712531 -0.011987747
## [3,] -0.3156958 2.784472 -0.1710683 -0.2089390 0.01910834 -0.010258855
## [4,] -0.2403566 2.756461  0.2695199  0.9413121 0.05519445 -0.076094302
## [5,] -0.2496896 2.753985  0.2363068  0.9593635 0.05099266  0.001016184
## [6,] -0.1714810 2.742810  0.2479419  0.9520736 0.04823809  0.002140593
##                PC7           PC8          PC9         PC10          PC11
## [1,]  5.592615e-05  1.497335e-04 5.278438e-15 3.677614e-16 -3.885781e-16
## [2,]  2.087660e-05 -1.338860e-04 5.236371e-15 3.400058e-16 -4.718448e-16
## [3,] -2.350538e-04  1.588504e-05 5.263476e-15 2.359224e-16 -2.185752e-16
## [4,]  1.117273e-05  1.776983e-04 5.253610e-15 2.151057e-16 -4.232725e-16
## [5,]  2.568177e-05 -1.344630e-04 5.168121e-15 3.330669e-16 -4.510281e-16
## [6,]  1.734385e-05  1.761105e-04 5.186010e-15 2.810252e-16 -3.781697e-16
```

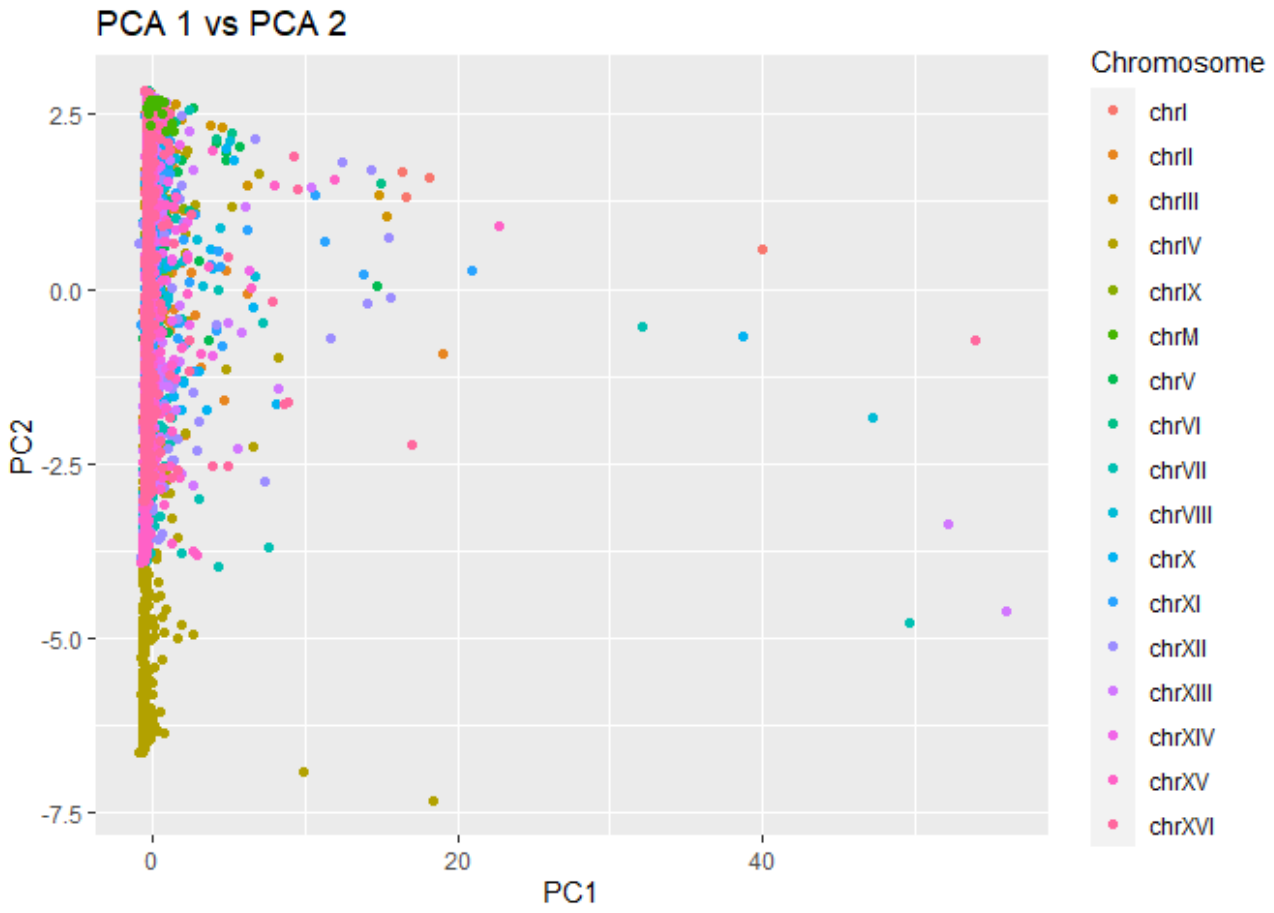
```
pca_data <- data.frame(pca$x)

#Add the variable chromosome back into the data
```

```
pca_data <- data.frame(pca$x, Chromosome = Data$Chromosome)
head(pca_data)
```

```
##          PC1      PC2        PC3        PC4        PC5          PC6
## 1 -0.2693958 2.818629  0.2542834  0.9707977 0.05383369 -0.011958731
## 2 -0.2650142 2.814192  0.2944585  1.0831134 0.05712531 -0.011987747
## 3 -0.3156958 2.784472 -0.1710683 -0.2089390 0.01910834 -0.010258855
## 4 -0.2403566 2.756461  0.2695199  0.9413121 0.05519445 -0.076094302
## 5 -0.2496896 2.753985  0.2363068  0.9593635 0.05099266  0.001016184
## 6 -0.1714810 2.742810  0.2479419  0.9520736 0.04823809  0.002140593
##             PC7           PC8          PC9         PC10          PC11
## 1  5.592615e-05  1.497335e-04 5.278438e-15 3.677614e-16 -3.885781e-16
## 2  2.087660e-05 -1.338860e-04 5.236371e-15 3.400058e-16 -4.718448e-16
## 3 -2.350538e-04  1.588504e-05 5.263476e-15 2.359224e-16 -2.185752e-16
## 4  1.117273e-05  1.776983e-04 5.253610e-15 2.151057e-16 -4.232725e-16
## 5  2.568177e-05 -1.344630e-04 5.168121e-15 3.330669e-16 -4.510281e-16
## 6  1.734385e-05  1.761105e-04 5.186010e-15 2.810252e-16 -3.781697e-16
##   Chromosome
## 1       chrI
## 2       chrI
## 3       chrI
## 4       chrI
## 5       chrI
## 6       chrI
```

```
#Plot the first and second principle components coloring by chromosome
#The clusters: Two clusters that stand out are chromosome M (pictured in green near t
ggplot(pca_data, aes(x = PC1, y = PC2, color= Chromosome)) +
  geom_point()+ggtitle("PCA 1 vs PCA 2")
```




# References

Graziotto J.J., Cao K., Collins F.S., Krainc D. 2012. Rapamycin activates autophagy in Hutchinson-Gilford progeria syndrome: implications for normal aging and age-dependent neurodegenerative disorders. Autophagy. 8(1), 147–151. doi: 10.4161/auto.8.1.18331

Morano K, Grant C, Moye-Rowley, WS. 2012. The response to heat shock and oxidative stress in Saccharomyces cerevisiae. Genetics, 190(4), 1157-1195. doi: 10.1534/genetics.111.128033

Mühlhofer M, Berchtold E, Stratil CG, Csaba G, Kunold E, Bach NC, Sieber SA, Haslbeck M, Zimmer R, Buchner J. 2019. The Heat Shock Response in Yeast Maintains Protein Homeostasis by Chaperoning and Replenishing Proteins. Cell Rep. 24;29(13):4593-4607.e8. doi: 10.1016/j.celrep.2019.11.109.

National Human Genome Research Institute. 09 May 2013. 1996: Yeast Genome Sequenced. www.genome.gov/25520379/online-education-kit-1996-yeast-genome-sequenced.