Unraveling Gene Expression Patterns in Retinitis Pigmentosa: Insights from Two-Way Clustering and Correlation Analysis Across Multiple Tissue Types

Avery Holloman

2024-08-12

```
# I committed out the installed packages so I could Knit into pdf
#install.packages("readxl")
#install.packages("ggplot2")
#install.packages("ggpubr")

# Libraries
library(readxl)
library(ggplot2)
library(ggpubr)
```

```
# Loading excel file from computer
file_path <- "C:/Users/jacob/OneDrive/Desktop/R Studio Projects 2024/MalaCards - Genes associate
d with Retinitis Pigmentosa.xlsx"
data <- read_excel(file_path)
```

```
## Warning: Expecting numeric in A1334 / R1334C1: got 'Copyright LifeMap Sciences
## Inc. and the Weizmann Institute of Sciences. May not be used for any
## non-academic research purpose without explicit written permission from LifeMap
## Sciences.'
```

```
# Showing the first few rows of the data to get a better grasp of data
head(data)
```

```
## # A tibble: 6 × 7
##      ID Symbol Description                      Category Score Molecular Variation
##   <dbl> <chr>  <chr>                            <chr>    <dbl> <chr>     <chr>    
## 1     1 CRX    Cone-Rod Homeobox                Protein… 1210. Genetic … Pathogen…
## 2     2 RPGR   Retinitis Pigmentosa GTPase R… Protein…  956. Genetic … Pathogen…
## 3     3 PRPH2  Peripherin 2                     Protein…  941. Genetic … Pathogen…
## 4     4 EYS    Eyes Shut Homolog                Protein…  941. Genetic … Pathogen…
## 5     5 PRPF8  Pre-MRNA Processing Factor 8     Protein…  938. Genetic … Pathogen…
## 6     6 CNGB1  Cyclic Nucleotide Gated Chann… Protein…  936. Genetic … Pathogen…
```

```
# Making sure 'Molecular' and 'Variation' columns are numerical value
data$Molecular_numeric <- as.numeric(gsub("[^0-9]", "", data$Molecular))
data$Variation_numeric <- as.numeric(gsub("[^0-9]", "", data$Variation))

# log transformation to y-axis as outliers brought errors
data$Molecular_log <- log10(data$Molecular_numeric + 1)
data$Variation_log <- log10(data$Variation_numeric + 1)

# To make Scatter plot better with correlations for Panel A, I used the log transformation
panel_A <- ggplot(data, aes(x = Score, y = Molecular_log)) +
  geom_point(color = "blue", size = 3, alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE, color = "darkred", linewidth = 1.2) +
  stat_cor(method = "pearson", label.x = 200, label.y = max(data$Molecular_log, na.rm = TRUE) -
0.5, color = "darkgreen", size = 6) +
  theme_minimal() +
  labs(x = "60S ribosomal protein L22 (Score)",
       y = "Log10 of ribosomal protein L5 (Molecular)",
       title = "a)") +
  theme(plot.title = element_text(color = "darkblue", size = 14),
        axis.text.x = element_text(angle = 45, hjust = 1))

# To make Scatter plot better with correlations for Panel B, I used the log transformation
panel_B <- ggplot(data, aes(x = Score, y = Variation_log)) +
  geom_point(color = "purple", size = 3, alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE, color = "darkorange", linewidth = 1.2) +
  stat_cor(method = "pearson", label.x = 200, label.y = max(data$Variation_log, na.rm = TRUE) -
0.5, color = "darkgreen", size = 6) +
  theme_minimal() +
  labs(x = "60S ribosomal protein L22 (Score)",
       y = "Log10 of Hnf2 (Variation)",
       title = "b)") +
  theme(plot.title = element_text(color = "darkblue", size = 14),
        axis.text.x = element_text(angle = 45, hjust = 1))

# I adjusted the layout as the outliers could not be identified to be removed
combined_plot <- ggarrange(panel_A, panel_B, ncol = 2, nrow = 1, labels = c("A", "B"))
```

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

```
## Warning: Removed 1114 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 1114 rows containing non-finite outside the scale range
## (`stat_cor()`).
```

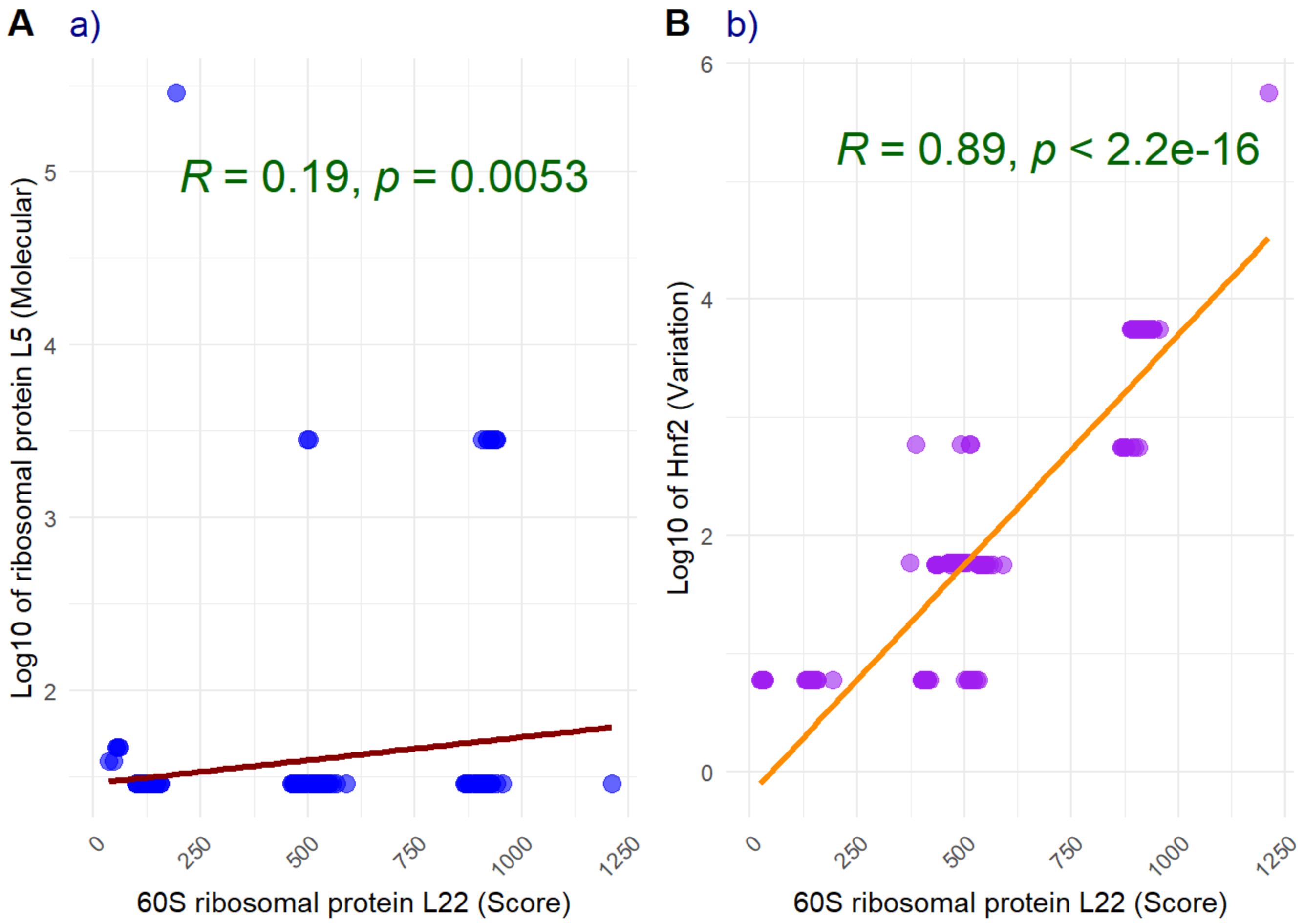
```
## Warning: Removed 1114 rows containing missing values or values outside the scale range
## (`geom_point()`).
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 1130 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 1130 rows containing non-finite outside the scale range
## (`stat_cor()`).
```

```
## Warning: Removed 1130 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

```
# Now lets take a look with the combined plot
print(combined_plot)
```


A
a)
R = 0.19, p = 0.0053
Log10 of ribosomal protein L5 (Molecular)
60S ribosomal protein L22 (Score)
B
b)
R = 0.89, p < 2.2e-16
Log10 of Hnf2 (Variation)
60S ribosomal protein L22 (Score)