Transcript length determines mRNA stability: the implications for mRNA and protein expression levels in *Schizosaccharomyces pombe*

ABSTRACT

S. pombe's prominence as a eukaryotic model organism is evidence by its popularity in the primary literature, with its unicellular nature making it the perfect candidate for use in a multitude of high throughput studies. These studies have contributed to the identification and characterisation of numerous proteins and their encoding genes, with others revealing the transcriptional mechanisms mediating their expression. However, research has negated the potential for the regulation of gene expression to be influenced by the stability of mRNA transcripts. Hence, using data from several high throughput studies, we investigated the influence of mRNA stability on mRNA and protein levels in cells of *S. pombe*. Using structured hypothesis testing and statistical analysis, we reveal that mRNA stability significantly correlates with mRNA and protein levels, with the most stable transcripts encoding nucleolar proteins. We suppose that prior evidenced mRNA length's determination of stability imposes these seemingly regulatory relationships.

INTRODUCTION

Schizosaccharomyces pombe's role as a prominent eukaryotic model organism shouldn't be underestimated. This species' demonstration of representative cellular effects has prompted the identification of numerous proteins' function and subsequent characterisation of their encoding genes (Yanagida, 2002). Genes presenting a conserved mechanism for regulating protein-mediated cellular responses, through adaption of expression. Classical signalling pathways typically result in altered accessibility of DNA to transcriptional machinery, often following receptor-mediated intracellular signal amplification (Quivy et al., 2004). Resulting in transcription factors and chromatin remodelling facilitating synergistic expression of multiple proteins for transient cellular responses (Tuteja, 2009, Qiao et al., 2013, Alvaro and Thorner, 2016). However, transcriptional level mechanisms of regulating gene expression don't only modulate protein levels. Instead, modulation is dictated by a complex amalgamation of influences on transcription, mRNA stability, translation, protein stability and turnover (Shalem et al., 2008). With cells coordinating an integral 'balancing act' of influences to accomplish optimal protein levels required for delivery of specific cellular responses (Rothman, 2010). Protein stability and turnover is well-studied, with little research focussing on the potential regulatory influence of mRNA stability on gene expression. Our research used genome-wide data from multiple high throughput studies on *S.* pombe (Marguerat et al., 2012, Hasan et al., 2014, Matsuyama et al., 2006) to investigate the influence of mRNA stability on mRNA and protein levels in the cell, identifying the cellular location of proteins encoded by the most stable transcripts.

RESULTS

Hypothesis

There will be a significant positive correlation between mRNA stability (half-life in minutes) and mRNA copies per cell in genes of *S. pombe*.

Statistical testing

The Spearman rank correlation quantified the level of correlation between mRNA stability and mRNA copies per cell and identified its statistical significance*.

Result

There is a highly significant positive correlation between mRNA stability and mRNA levels (Spearman rank correlation, r=0.48, $P<2.2\times10^{-16}$). Figure 1 shows a scatter graph demonstrating this correlation, with genes revealing a logarithmic average mRNA stability and mRNA copy number of 1.52 and 0.44 respectively. The areas featuring the highest density of genes appear consistent with these averages, with most deviation from them seen in genes following the trend with the highest stability mRNA transcripts.

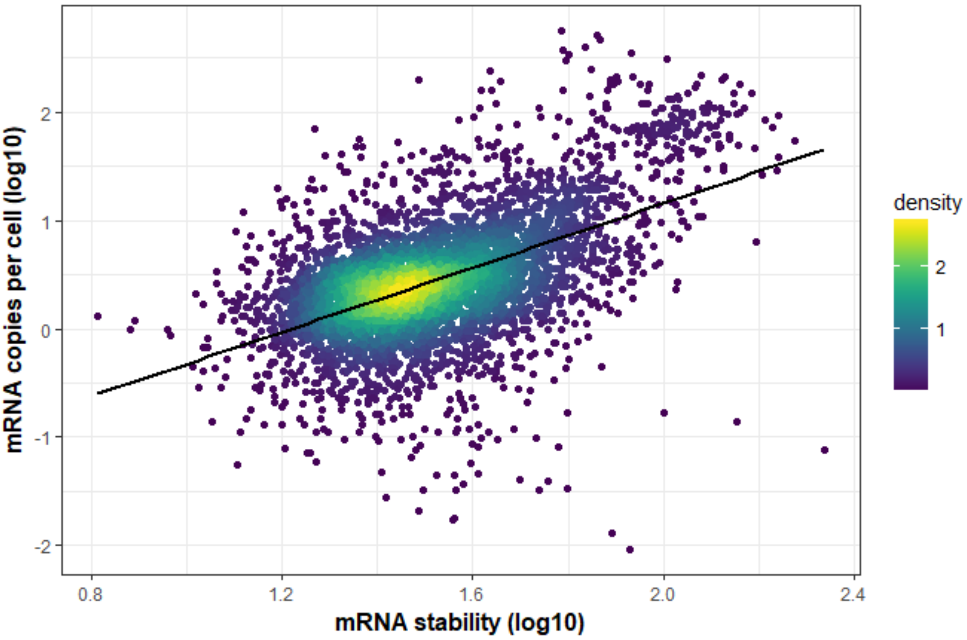

Figure 1 – Correlation of mRNA levels and mRNA stability in S. pombe. *Scatter plot representing the logarithmic relationship between mRNA stability (half-life in minutes) and mRNA levels (copies per cell) of S. pombe genes, with linear regression line. Colour intensity is indicative of gene density.*

Interpretation

Our hypothesis appears to have been proven correct with results revealing a highly significant correlation between mRNA stability and mRNA levels (figure 1). This genome representative relationship appears to imply that due to the resistance of high stability transcripts to cellular degradation, they achieve longer intracellular persistence than low stability transcripts prone to higher rates of degradation. Resulting in higher levels of more stable mRNAs than less stable ones at any time (Trcek et al., 2011).

The significant correlation between mRNA stability and mRNA levels does not prove that stability is the sole causative factor instigating the changes in mRNA levels seen between genes. However, it is possible that stability may be contributing to a regulatory effect. Given longer length transcripts generally take longer to synthesise, and the negative correlation between transcript length and stability in human studies (Feng and Niu, 2007). Further investigation could seek to clarify whether the variation in mRNA levels seen between genes is a product of mRNA length's effect on stability.

Hypothesis

There will be a significant positive correlation between mRNA stability (half-life) and protein levels (copies per cell) in genes of *S. pombe*.

Statistical testing

The Spearman rank correlation quantified the level of correlation between mRNA stability and protein levels. To control for the effect of the number of mRNA copies per cell on the association between these two variables, a partial correlation was also utilised. Both tests identified statistical significance of their associated correlation*.

Result

There is a highly significant positive correlation between mRNA stability and protein copies per cell (Spearman rank correlation: r=0.59, $P<2.2\times10^{-16}$). When excluding the effect of mRNA copies in the cell at any one time, there was also a highly significant partial correlation between these two variables (Spearman partial correlation: r=0.39, $P<2.2\times10^{-16}$). Figure 2 shows a scatter graph demonstrating the correlation between mRNA stability and protein levels on a logarithmic scale, with the average of protein copies having a logarithmic value of 3.63. The most deviation from this average in genes following this trend is those with the highest mRNA stabilities.

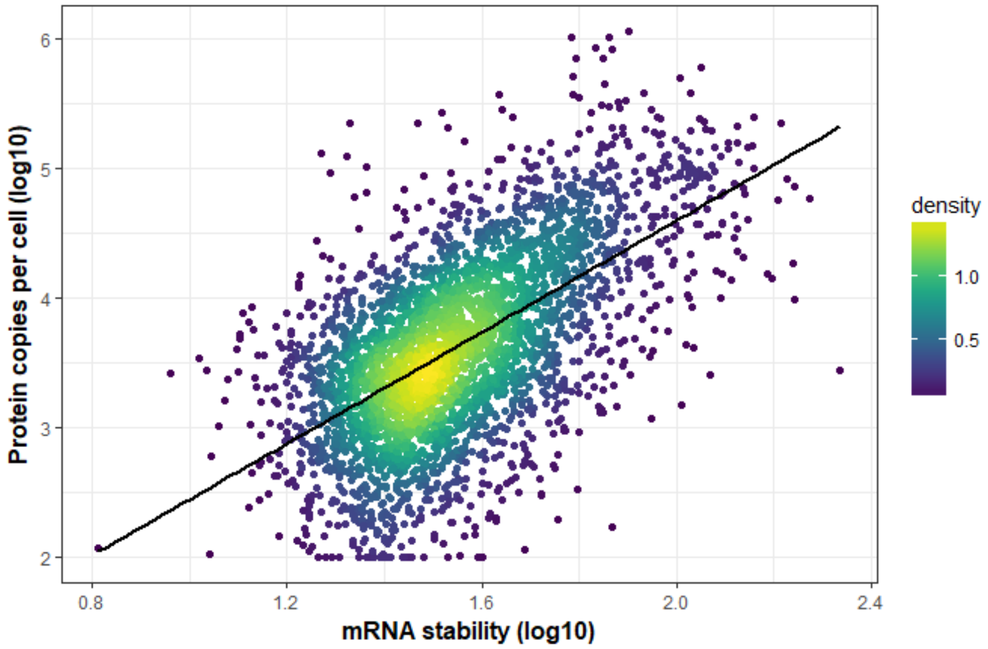

Figure 2 – Correlation of protein levels and mRNA stability in S. pombe. *Scatter plot representing the logarithmic relationship between mRNA stability (half-life in minutes) and protein copies per cell of S. pombe genes, with linear regression line. Colour is indicative of gene density.*

Interpretation

This hypothesis appears to have been proven correct, with mRNA stability being positively correlated with protein levels (figure 2). This correlation seems consistent with the relationship seen between mRNA stability and mRNA levels in the cell through the intracellular persistence of different stability transcripts. High stability transcripts persist for longer leaving greater opportunity for ribosomal attachment and translation initiation. But unstable transcripts' liability to degradation leaves less opportunity for translational events (Misquitta et al., 2006). This concept, with the more higher stability transcripts than unstable transcripts in the cell, and the decreased correlation between mRNA stability and protein levels (when excluding the influence of mRNA copies), reveal a combinative effect. That both the large number of transcripts and the increased opportunity for ribosomal attachment, both contribute to high protein levels derived from genes encoding very stable mRNAs.

Given the prior inference that mRNA length rather than stability regulates transcript levels, and the negative correlation between transcript length and stability (Feng and Niu, 2007). Potentially, longer length mRNAs' increased time undergoing transcription, provides enough opportunity for prior transcribed unstable mRNA to be degraded. Thus, ensuring that consistently low levels of mRNA are maintained to satisfy the low protein requirements of this gene, with the opposite being the case for more stable, shorter length transcripts.

Hypothesis

Genes that encode proteins expressed in the nucleolus are derived from mRNA transcripts with significantly greater stability than genes encoding proteins expressed in any other organelle.

Statistical testing

The mRNA half-lives of protein coding genes were transformed through a base 10 logarithm, with subsequent exclusion of genes with a mRNA stability of less than 1.7 on the logarithmic scale. The two-sample Wilcoxon (Mann-Whitney) rank sum test was used to investigate the relationship between the stability of mRNA transcripts encoding proteins of the nucleolus and transcripts encoding proteins of other complexes and organelles (Mitochondrion, Nucleus, Nuclear envelope, Nuclear dots, Golgi). *All statistical analyses were performed using R studio with significance determined at the 0.05 level (Team, 2017, Wickham, 2009, Kassambara, 2018, Venables and Ripley, 2002, Garnier, 2018)*.

Results

Genes encoding nucleolar proteins (median=1.96) had significantly greater mRNA stability than those encoding proteins of other complexes and organelles (Mann-Whitney: Mitochondrion; median=1.80, W=317, n_1=14, n_2=82, p=7.75×10^{-3}. Nucleus; median=1.79, W=200.5, n_1=14, n_2=19, p=0.0147. Nuclear dots; median=1.74, W=179, n_1=14, n_2= 15, p=1.33×10^{-3}. Golgi; median=1.75, W=168, n_1=14, n_2=14, p=7.944×10^{-4}). Except the nuclear envelope whose transcripts had no significant difference in stability to those of the nucleolus (Mann-Whitney: median=1.72, W=53, n_1=14, n_2=5, p=0.107).

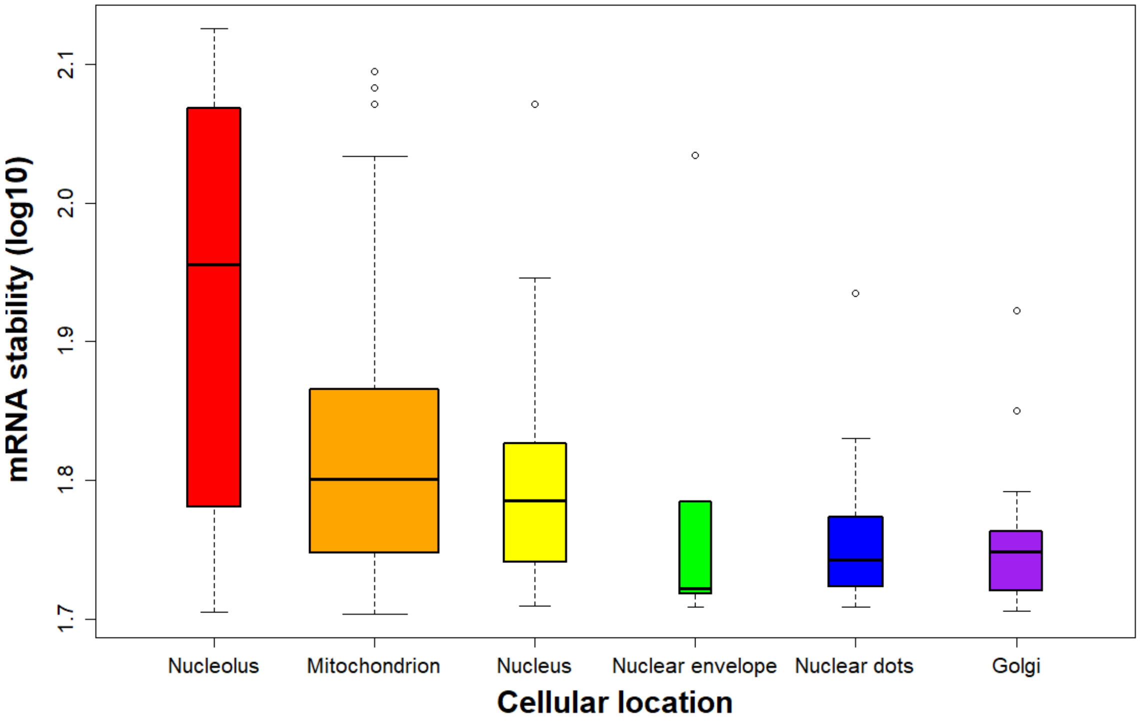

Figure 3 – mRNA stability versus the cellular location of proteins in S. pombe. *Boxplot representing the stabilities of the most stable mRNAs (logarithmic mRNA stability >1.7) against the cellular location of their encoding proteins. Box width is indicative of the number of genes encoding the most stable mRNAs.*

Interpretation

Our hypothesis appears to have been proven correct, with the nucleolus having significantly more stable mRNA transcripts than the: mitochondrion, nucleus, nuclear dots and Golgi (figure 3). Given that the organelles with the next highest median mRNA stabilities to the nucleolus were the mitochondrion and the nucleus, and the positive correlation observed between mRNA stability and protein levels. It is possible that proteins with the highest expression levels are located at the nucleolus, which seems consistent with its function. Since the nucleolus synthesises ribosomes, required for the translation of mRNA to protein, a vital constituent of organelles (McLeod et al., 2014). It would seem a cellular priority to dedicate the largest numbers of proteins to the nucleolus to safeguard and regulate a function depended on by other cellular components. These complex relationships simultaneously highlight the co-dependency of organelles, especially the mitochondrion exchanging ATP for proteins mediating its functionality. This such relationship may evidence the endosymbiont hypothesis of mitochondrial evolution, with such an exchange seemingly derivative of a mutually beneficial symbiosis (Gray, 2012).

The few stable mRNAs encoding proteins of the nuclear envelope and a large outlier skewing its mean may explain the lack of a significant relationship with the stability of mRNAs encoding nucleolar proteins (figure 3).

CONCLUSION

Our investigation has successfully identified that both mRNA and protein levels significantly correlate with mRNA stability in genes of *S. pombe*, with transcripts encoding proteins in the nucleolus being significantly more stable than those of proteins in almost all other organelles. A major implication of these conclusions was that transcript length's determination of stability (Feng and Niu, 2007), means that it itself may be an indirect regulator of protein levels. However, a noteworthy limitation is in the suggested combinative effect of mRNA levels and stability, a linear correlation was assumed between mRNA and protein levels when this is likely a much more complex relationship (Liu et al., 2016). The selection of genes with the most stable mRNAs was determined using a logarithmic stability of 1.7. This was not an implicit valued defined by the literature but was instead selected upon inspection of the correlation depicted in figure 1, such arbitrariness limits the comparability of this aspect of the investigation. Despite caveats, we have identified a vital regulator of protein levels in the cell, mRNA length, revealing potential implications for its use as a specificity factor of mRNA targeting drugs.

REFERENCES

ALVARO, C. G. & THORNER, J. 2016. Heterotrimeric G Protein-coupled Receptor Signaling in Yeast Mating Pheromone Response. *J Biol Chem,* 291**,** 7788-95.

FENG, L. & NIU, D. K. 2007. Relationship between mRNA stability and length: an old question with a new twist. *Biochem Genet,* 45**,** 131-7.

GARNIER, S. 2018. viridis: Default Color Maps from 'matplotlib'.

GRAY, M. W. 2012. Mitochondrial evolution. *Cold Spring Harb Perspect Biol,* 4**,** a011403.

HASAN, A., COTOBAL, C., DUNCAN, C. D. & MATA, J. 2014. Systematic analysis of the role of RNA-binding proteins in the regulation of RNA stability. *PLoS Genet,* 10**,** e1004684.

KASSAMBARA, A. 2018. ggpubr: 'ggplot2' Based Publication Ready Plots.

LIU, Y., BEYER, A. & AEBERSOLD, R. 2016. On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell,* 165**,** 535-50.

MARGUERAT, S., SCHMIDT, A., CODLIN, S., CHEN, W., AEBERSOLD, R. & BÄHLER, J. 2012. Quantitative analysis of fission yeast transcriptomes and proteomes in proliferating and quiescent cells. *Cell,* 151**,** 671-83.

MATSUYAMA, A., ARAI, R., YASHIRODA, Y., SHIRAI, A., KAMATA, A., SEKIDO, S., KOBAYASHI, Y., HASHIMOTO, A., HAMAMOTO, M., HIRAOKA, Y., HORINOUCHI, S. & YOSHIDA, M. 2006. ORFeome cloning and global analysis of protein localization in the fission yeast Schizosaccharomyces pombe. *Nat Biotechnol,* 24**,** 841-7.

MCLEOD, T., ABDULLAHI, A., LI, M. & BROGNA, S. 2014. Recent studies implicate the nucleolus as the major site of nuclear translation. *Biochem Soc Trans,* 42**,** 1224-8.

MISQUITTA, C. M., CHEN, T. & GROVER, A. K. 2006. Control of protein expression through mRNA stability in calcium signalling. *Cell Calcium,* 40**,** 329-46.

QIAO, Y., GIANNOPOULOU, E. G., CHAN, C. H., PARK, S. H., GONG, S., CHEN, J., HU, X., ELEMENTO, O. & IVASHKIV, L. B. 2013. Synergistic activation of inflammatory cytokine genes by interferon-γ-induced chromatin remodeling and toll-like receptor signaling. *Immunity,* 39**,** 454-69.

QUIVY, V., CALOMME, C., DEKONINCK, A., DEMONTE, D., BEX, F., LAMSOUL, I., VANHULLE, C., BURNY, A. & VAN LINT, C. 2004. Gene activation and gene silencing: a subtle equilibrium. *Cloning Stem Cells,* 6**,** 140-9.

ROTHMAN, S. 2010. How is the balance between protein synthesis and degradation achieved? *Theor Biol Med Model,* 7**,** 25.

SHALEM, O., DAHAN, O., LEVO, M., MARTINEZ, M. R., FURMAN, I., SEGAL, E. & PILPEL, Y. 2008. Transient transcriptional responses to stress are generated by opposing effects of mRNA production and degradation. *Mol Syst Biol,* 4**,** 223.

TEAM, R. C. 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria

TRCEK, T., LARSON, D. R., MOLDÓN, A., QUERY, C. C. & SINGER, R. H. 2011. Single-molecule mRNA decay measurements reveal promoter- regulated mRNA stability in yeast. *Cell,* 147**,** 1484-97.

TUTEJA, N. 2009. Signaling through G protein coupled receptors. *Plant Signal Behav,* 4**,** 942-7.

VENABLES, W. N. & RIPLEY, B. D. 2002. Modern Applied Statistics with S. Fourth ed. New York: Springer.

WICKHAM, H. 2009. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

YANAGIDA, M. 2002. The model unicellular eukaryote, Schizosaccharomyces pombe. *Genome Biol,* 3**,** COMMENT2003.

SUPPLEMENTARY METHODS

```
##############################################################################
# Importing, tidying and exploratory analysis on mRNA stability  and mRNA copies per cell #
##############################################################################

#Setting working directory

setwd("~/Big Data Biology")


#Loaded data in R

load("data/fission_yeast_data.2018-11-21.Rda")

#Creating subset of data with only numeric values for mRNA copies per cell by removing, missing values (NA's)

mRNA_copies_omit_na <- gene[!is.na(gene$mRNA_copies_per_cell),]


#Drawing histogram to visualise if mRNA copies per cell resembles a normal distribution

hist(mRNA_copies_omit_na$mRNA_copies_per_cell)


#Histogram demonstrates a large majority of values with a low number of mRNA copies per cell, so will transform data to see a better range of values. Transforming data through a base 10 logarithm.

logmRNA_copies <- log10(mRNA_copies_omit_na$mRNA_copies_per_cell)


#Adding transformed variable to the new subset

mRNA_copies_omit_na$logmRNA_copies <- logmRNA_copies


#Checking number of rows has decreased in comparison to original data

nrow(gene)

nrow(mRNA_copies_omit_na)


#Row number decreased because of removing of missing values.

#Checking data to see if NA's of mRNA copies column have been removed and that transformed data column has been added

head(gene)

head(mRNA_copies_omit_na)
```

```
#Missing values from first 6 rows appear to be removed

#Drawing histogram of transformed mRNA copies per cell

hist(mRNA_copies_omit_na$logmRNA_copies)


#Hard to tell from histogram the exact normality of the data but it hardly resembles the typical graphical structure of a normal distribution. Sample size is too large to use shapiro.test function so using Q-Q plot to better visualise if the data is normal.

#Drawing q-q plot to see if mRNA copies seem normally distributed

#Installing ggpubr package (required package)

install.packages("ggpubr")


#Loading ggpubr package

library("ggpubr")


#Drawing Q-Q plot with 95% confidence intervals

ggpubr::ggqqplot((mRNA_copies_omit_na$logmRNA_copies), title = "Q-Q Plot",

xlab = "Theorectical Quantiles", ylab="Sample Quantiles")


#Many of the points of the Q-Q pot largely deviate from the 'qqline' especially at the upper tail of the data. This demonstrates that the mRNA copies per cell variable does not much resemble a normal distribution.

#Testing to see if any of the data on mRNA stability column is non-numeric (NA)

which(is.na(mRNA_copies_omit_na$mRNA.stabilities))


#Many genes appear to have no recorded data on mRNA stability.

#Creating a new gene subset with all missing values from the mRNA stability variable removed

mRNA_stability_omit_na <- gene[!is.na(gene$mRNA.stabilities),]


#Drawing histogram to visualise mRNA stability data

hist(mRNA_stability_omit_na$mRNA.stabilities)


#Histogram barely resembles structure of a normal distribution. Transforming mRNA stability data through a base 10 logarithm to better visualise range of values
```

```
logmRNA_stability <- log10(mRNA_stability_omit_na$mRNA.stabilities)


#Adding transformed data to the new subset
mRNA_stability_omit_na$logmRNA_stability <- logmRNA_stability


#Checking to see if the number of rows has changed by comparison to original data
nrow(gene)
nrow(mRNA_stability_omit_na)


#Row number decreased due to removal of NA's
#Checking to see if NA's of mRNA stability column have been removed and transformed mRNA stability data has been added
head(gene)
head(mRNA_stability_omit_na)


#NA's appear to be removed and data added
#Drawing histogram to see if transformed mRNA stability data appears normally distributed
hist(mRNA_stability_omit_na$logmRNA_stability)


#Histogram appears structurally similar to that of a normal distribution so will run shapiro.test function on data to confirm, as it meets the requirements of the test.
#Running shapiro.test
shapiro.test(mRNA_stability_omit_na$logmRNA_stability)


#Despite appearing normally distributed, shapiro.test shows highly significant deviation from a normal distribution. However, this may be due to the sensitivity of the shapiro.test to large numbers of small deviations in large samples producing a significant.
#Drawing Q-Q plot with 95% confidence intervals to investigate further
ggpubr::ggqqplot((log10(mRNA_stability_omit_na$logmRNA_stability)), title = "Q-Q Plot",
              xlab = "Theorectical Quantiles", ylab="Sample Quantiles")


#It appears that according to the Q-Q plot, the data has very little deviation from the 'qqline' indicating its general consistency with a normal distribution
```

```
################################################################################
#                              Statistical analysis                              #
################################################################################

# However, with one of our continuous variables (mRNA copies per cell) being non-normal and the other being normal (mRNA stability), a non-parametric spearman rank test will be utilised.

#Adding transformed variables to the original data set as variable sizes must be equal with test omitting missing values.

logmRNA_copies <- log10(gene$mRNA_copies_per_cell)

logmRNA_stability <- log10(gene$mRNA.stabilities)

gene$logmRNA_copies <- logmRNA_copies

gene$logmRNA_stability <- logmRNA_stability


#Running Spearman rank test

cor.test(gene$logmRNA_copies, gene$logmRNA_stability, method="spearman",

         use="complete.cases")

#It is apparent that there is a highly significant moderate positive correlation between mRNA stability and mRNA copies per cell.

################################################################################
#                                     Figure                                     #
################################################################################

#Making new subset and removing missing values within either column to make row numbers equal for use in figure

plot1 <- subset(gene, select=logmRNA_copies:logmRNA_stability)

plot2 <- na.omit(plot1)


#Loading required packages

library(MASS)

library(ggplot2)

library(viridis)


#Drawing scatterplot of mRNA stability versus mRNA copies per cell with colour intensity indicative of the density of genes in a given area

get_density <- function(x, y, ...)

{dens <- MASS::kde2d(x, y, ...)

ix <- findInterval(x, dens$x)
iy <- findInterval(y, dens$y)

ii <- cbind(ix, iy)

return(dens$z[ii])}

set.seed(1)

dat <- data.frame( x = plot2$logmRNA_stability, y = plot2$logmRNA_copies)

dat$density <- get_density(dat$x, dat$y, n = 100)

ggplot(dat, aes(x,y)) + geom_point(aes(x, y, color = density)) +

scale_color_viridis()+ ylab("mRNA copies per cell (log10)")+ xlab("mRNA stability (log10)")+

theme_bw() + theme(axis.title.x = element_text(color = "black", size = 12, face = "bold"),

 axis.title.y = element_text(color = "black", size = 12, face = "bold")) +

 geom_smooth(method=lm,se=FALSE,colour="black", linetype="solid")

 ################################################################################
#               Explanatory analysis on mRNA stability and protein copies per cell                #
 ################################################################################

#Drawing histogram of protein copies per cell variable

hist(gene$protein_copies_per_cell)


#Large majority of genes have smaller levels of protein per cell so transforming variable through a base 10 logarithm to see better range of values and add it to the data set

logprotein_copies <- log10(gene$protein_copies_per_cell)

gene$logprotein_copies <- logprotein_copies


#Drawing histogram to see if data is normally distributed

hist(gene$logprotein_copies)


#Checking that new transformed variable has been added to the dataset

head(gene)


#Appears that the variable has been added but all of first 6 rows are missing values.

#Creating a new subset containing the transformed protein copies variable with missing values removed.

protein_copies_omit_na <- gene[!is.na(gene$logprotein_copies),]


#Checking structure of transformed variable
```

```
str(protein_copies_omit_na$logprotein_copies)


#Variable satisfies the requirements of the shapiro.test (<5000 values) so will use function to assess normality

shapiro.test(protein_copies_omit_na$logprotein_copies)


#Shapiro.test test reveals highly significant deviation from a normal distribution but again, this is likely due to the large sample size.

#Running Q-Q plot to visualise if the logged protein copies variable is normal

ggpubr::ggqqplot((protein_copies_omit_na$logprotein_copies), title = "Q-Q Plot",

            xlab = "Theoretical Quantiles", ylab = "Sample Quantiles")


#The protein copies per cell appear very normal with only a few points towards the lower tail of the data demonstrating much deviation beyond the 95% confidence interval. The data on mRNA stability reveals greater deviation beyond these confidence intervals hence a non-parametric will be used.

################################################################################
#                              Statistical analysis                            #
################################################################################

#Given the data's lack of normality, running the non-parametric Spearman's correlation test

cor.test(gene$logprotein_copies, gene$logmRNA_stability, method="spearman",

       use="complete.cases")

#The spearman's correlation test reveals a highly significant moderate positive correlation between mRNA stability and protein copies in the cell. However, given the biological context it would be likely that the number of mRNA copies in the cell would likely to be influencing the correlation seen between the mRNA stability and protein copies. Hence a partial correlation test will be used to see correlation between mRNA stability and protein copies whilst controlling for the number of mRNA copies.

#Installing required package

install.packages("ppcor")


#Loading required package

library(ppcor)


#Creating new subset

no_na <- subset(protein_copies_omit_na, !is.na(logmRNA_stability) | !is.na(logmRNA_copies))
```

```
#Selecting which genes have missing values as partial correlation does not allow missing values

miss_val <- which(is.na(no_na[26:28]), arr.ind=TRUE)


#Removing missing values

no_na<- no_na[-c(miss_val[,1]),]


#Running the partial correlation test

pcor.test(no_na$logmRNA_stability, no_na$logprotein_copies, no_na$logmRNA_copies,

          method = "spearman")


#The number of mRNA copies in the cell has been controlled for and the partial correlation test has calculated a smaller but still moderate correlation. Given this, it is apparent that the persistence of an mRNA to remain the in the cell due to its stability has a large and potentially greater influence on protein levels in the cell than the number of mRNA transcripts.
################################################################################
#                                   Figure                                     #
################################################################################
#Drawing scatterplot of mRNA stability versus protein copies per cell with colour intensity indicative of the density of genes in a given area

theme_set(theme_bw(base_size = 16))

get_density <- function(x, y, ...) {

  dens <- MASS::kde2d(x, y, ...)

  ix <- findInterval(x, dens$x)

  iy <- findInterval(y, dens$y)

  ii <- cbind(ix, iy)

  return(dens$z[ii]) }

set.seed(1)

dat <- data.frame( x = no_na$logmRNA_stability, y = no_na$logprotein_copies)

dat$density <- get_density(dat$x, dat$y, n = 100)

ggplot(dat, aes(x,y)) + geom_point(aes(x, y, color = density)) +

  scale_color_viridis()+ ylab("Protein copies per cell (log10)")+

  xlab("mRNA stability (log10)") + theme_bw() +

  theme(axis.title.x = element_text(color = "black", size = 12, face = "bold"),
     axis.title.y = element_text(color = "black", size = 12, face = "bold"))+
  geom_smooth(method=lm,se=FALSE,colour="black", linetype="solid")
##############################################################################
#     Explanatory analysis on the cellular locations of the proteins from genes from genes     #
#                                  encoding the most stable mRNA transcripts                                   #
##############################################################################

#Making subsets separating genes above a 1.7 log10 mRNA stability into the cellular
location of their encoded proteins

Golgi <- subset(no_na, Golgi==1 & logmRNA_stability>=1.7)

Mitochondrion <- subset(no_na, Mitochondrion==1 & logmRNA_stability>=1.7)

Nuclear_dots <- subset(no_na, Nuclear_dots==1 & logmRNA_stability>=1.7)

Nuclear_envelope <- subset(no_na, Nuclear_envelope==1 & logmRNA_stability>=1.7)

Nucleolus <- subset(no_na, Nucleolus==1 & logmRNA_stability>=1.7)

Nucleus <- subset(no_na, Nucleus==1 & logmRNA_stability>=1.7)


#Checking the number of genes in each row

nrow(Golgi)

nrow(Mitochondrion)

nrow(Nuclear_dots)

nrow(Nuclear_envelope)

nrow(Nucleolus)

nrow(Nucleus)


#Drawing boxplot to visually compare each organelle's median mRNA stability

boxplot(Nucleolus$logmRNA_stability, Mitochondrion$logmRNA_stability,
Nucleus$logmRNA_stability,

     Nuclear_envelope$logmRNA_stability, Nuclear_dots$logmRNA_stability,

     Golgi$logmRNA_stabilit,

     ylab=expression(bold("mRNA stability (log10)")),

     xlab=expression(bold("Cellular location")),

     names=c("Nucleolus","Mitochondrion", "Nucleus", "Nuclear envelope",

          "Nuclear dots", "Golgi"),

     varwidth = T,cex.lab=1.9, cex.axis= 1.3, boxlwd=2,

     col=c('red', 'orange', 'yellow','green', 'blue','purple'))
```

```
#From the boxplot it is apparent that the Nucleolus has the highest median for mRNA stability.

#Drawing histogram to check if the most stable mRNAs of the Nucleolus appear to be structurally consistent with that of a normal distribution

hist(Nucleolus$logmRNA_stability)


#Structure of the logged mRNA stability does not appear to be consistent with a normal distribution hence running shapiro.test to check normality

shapiro.test(Nucleolus$logmRNA_stability)


#Q-Q confirms that variable is in fact normally distributed

ggpubr::ggqqplot((Nucleolus$logmRNA_stability), title = "Q-Q Plot",

              xlab = "Theorectical Quantiles", ylab="Sample Quantiles")


#Running shapiro.test to confirm that all other organelle subsets having mRNA stabilities that deviate from normality

shapiro.test(Mitochondrion$logmRNA_stability)

shapiro.test(Nucleus$logmRNA_stability)

shapiro.test(Nuclear_envelope$logmRNA_stability)

shapiro.test(Nuclear_dots$logmRNA_stability)

shapiro.test(Golgi$logmRNA_stability)


#Genes encoding proteins of all other organelles appear to have mRNA stabilities that are not normally distributed.
###############################################################################
#                              Statistical analysis                           #
###############################################################################

#Running non-parametric two-sample Wilcoxon (Mann-Whitney) test of the mRNA stabilities encoding proteins of the Nucleolus against those encoding proteins of all other organelles.

wilcox.test(Mitochondrion$logmRNA_stability, Nucleolus$logmRNA_stability, paired=F)

wilcox.test(Nucleolus$logmRNA_stability, Nucleus$logmRNA_stability, paired =F)

wilcox.test(Nucleolus$logmRNA_stability, Nuclear_envelope$logmRNA_stability, paired=F)

wilcox.test(Nucleolus$logmRNA_stability, Nuclear_dots$logmRNA_stability, paired=F)

wilcox.test(Nucleolus$logmRNA_stability, Golgi$logmRNA_stability, paired=F)
```

```
#The genes that encode protein located in the Nucleolus have highly significantly greater
mRNA stabilities than those located in any other organelle except the Nuclear envelope.
However, given that the number of genes located in that subset is so small (n=4) and the
presence of a large outlier this is likely to skew the average of the data to remove any
significant difference.

##############################################################################
#                                    Figure                                  #
##############################################################################

#Drawing boxplot of the mRNA stabilities against their protein's cellular location using
subsets with most stable mRNA transcripts (>=1.7log mRNA stability)

boxplot(Nucleolus$logmRNA_stability, Mitochondrion$logmRNA_stability,
Nucleus$logmRNA_stability,

    Nuclear_envelope$logmRNA_stability, Nuclear_dots$logmRNA_stability,

    Golgi$logmRNA_stabilit,

    ylab=expression(bold("mRNA stability (log10)")),

    xlab=expression(bold("Cellular location")),

    names=c("Nucleolus","Mitochondrion", "Nucleus", "Nuclear envelope",

        "Nuclear dots", "Golgi"),

    varwidth = T,cex.lab=1.9, cex.axis= 1.3, boxlwd=2,

    col=c('red', 'orange', 'yellow','green', 'blue','purple'))
```