**Analysis of the Relationship Between Essential and Non-Essential *Schizosaccharomyces pombe* Genes Reveals a Correlation Between mRNA Function and localisation on mRNA Stability**

**Abstract**

mRNA stability has been shown to influence gene expression and is regulated by post-transcriptional mechanisms. Improper post-transcriptional regulation has been shown to contribute to several diseases including cancer, inflammatory disease, and Alzheimer’s. Understanding the influences of mRNA stability could lead to the development of new therapies. In this report, I test: whether essentiality affects gene expression; the relationship between mRNA stability and expression to essentiality of mRNAs, as well as the relationship between mRNA stability and localisation. It was found that essential genes were expressed more highly than non-essential genes in general (W = 3039370, p-value < 2.2e-16); essential mRNAs were significantly more stable than non-essential mRNAs overall (W = 2454908, p-value = 3.743e-08); more stable mRNAs tend to be more highly expressed (W = 9156457, p-value < 2.2e-16); mRNA stability differs significantly based on the localisation of the proteins they encode (ANOVA: F = 43.510; d.f. = 5; p< 2e-16) and that differences in stability between essential and non-essential mRNAs were dependant on their localisation (ANOVA: F = 3.709; d.f. = 5; p= 0.00241). It is suggested that differences in post-transcriptional regulation are a contributor to the pattern of expression observed, and that essentiality is not the factor affecting mRNA stability, but it is instead based on the function of the proteins they encode.

**Introduction**

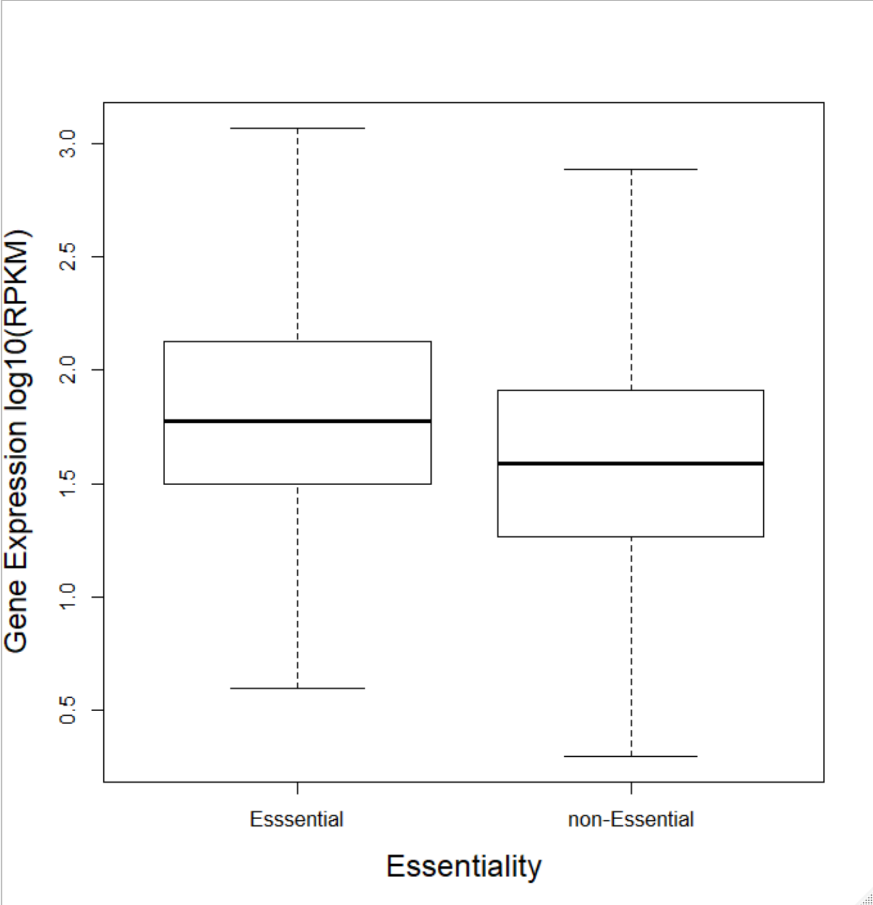
It is generally agreed that post-transcriptional regulation of gene expression is essential for the proper functioning of cells. It has been shown that malfunctioning post-transcriptional regulation can be responsible for various diseases including cancer, inflammatory disease, and Alzheimer’s. The fate of mRNAs are influenced in a many ways. RNA capping and addition of a poly(A) tail increase the stability of mRNAs resulting in increased expression. The RNA interference (RNAi) pathway is another mechanism of regulation, which silences and/or degrades mRNAs that it interacts with (D. Cheneval et al., 2010). RNAi is thought to modulate the expression ~30% of all mammalian protein-coding genes (W. Filipowicz et al., 2008) and ~60% of human protein-coding genes (R. C. Friedman et al., 2009). The stability of mRNAs appears to be incredibly influential on the overall expression of a gene with just small changes in stability resulting in large changes in steady-state levels (D. Cheneval et al., 2010). The stability of mRNAs is regulated by a subset of RNA binding proteins (RBPs) with functions ranging from control of polyadenylation to dampening of miRNA mediated decay, part of the RNAi pathway (M. D. Díaz-Muñoz et al., 2018).

This study aims to assess whether essential and non-essential protein-coding mRNAs interact with the post transcription machinery differently and whether this pattern persists when these mRNAs are categorised based on the localisation of the proteins they encode to determine whether essentiality affects post-transcriptional regulation.

**Results and Discussion**

**Section 1: Are Essential Genes Expressed Differently to Non-Essential Genes?**

It was hypothesised that Essential protein-coding genes would be more highly expressed than non-essential genes. This is due to their role in core cellular processes and as a result, necessitate more frequent expression as well as at a greater level. I tested this by comparing the expression of 1369 essential Schizosaccharomyces pombe genes and 3771 non-essential genes using a Wilcoxon signed-rank test. The expression data were taken from S. R Atkinson et al., 2018 and the essentiality data from D-U. Kim et al., 2010, https://www.pombase.org/. provided the protein-coding gene data. Data analysis revealed that essential genes were more highly expressed than non-essential genes (figure 1) (W = 3039370, p-value < 2.2e-16). Similar studies have found that essential genes in *Escherichia coli* and *Bacillus subtilis* are also expressed more highly than non-essential genes (E. P. C. Rocha et al., 2004). The results seen in these data may be due to increased stability of essential gene mRNAs, as a consequence of differing post-transcriptional regulation, resulting in increased steady-state levels of expression.



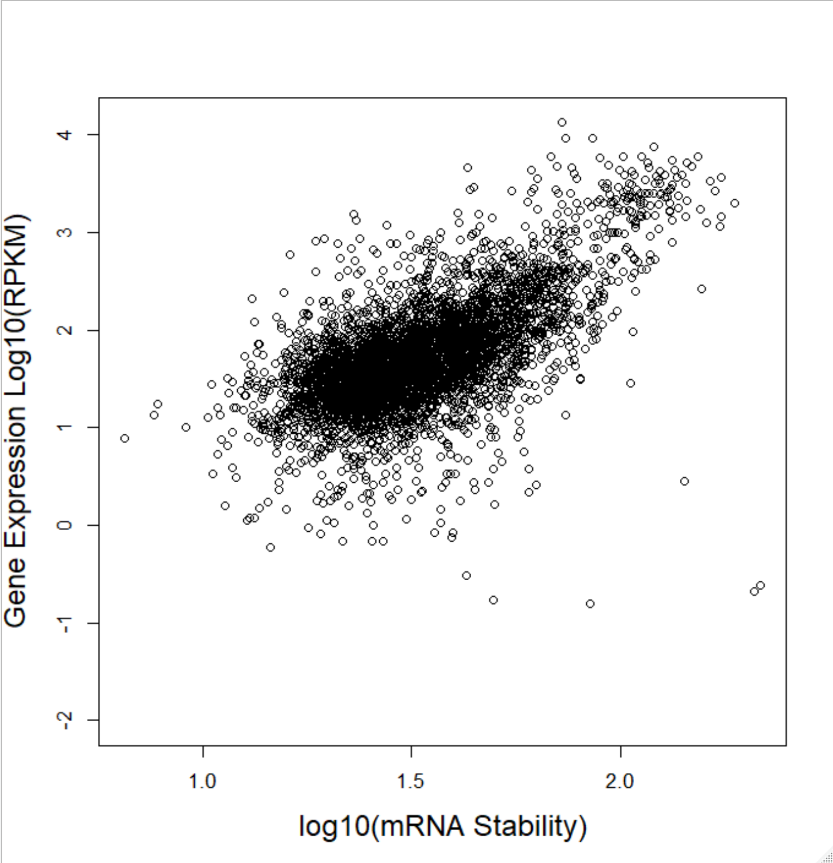
**Figure 1: Essential genes are more highly expressed than non-essential genes on average.**

A comparison between the expression of 1369 essential Schizosaccharomyces pombe genes and 3771 non-essential genes was made using a Wilcoxon signed-rank test. Data analysis revealed that essential genes were more highly expressed than non-essential genes (W = 3039370, p-value < 2.2e-16).

**Section 2: The relationship between mRNA Stability and RPKM in Essential and Non-Essential mRNAs**

It was hypothesised that higher mRNA stability is correlated with higher expression, as the mechanisms of post-transcriptional regulation would suggest. I used a Wilcoxon signed-rank test to determine the relationship between the stability of mRNAs and their level of expression (RPKM). mRNA stability data were taken from (A. Hasan et al., 2014). It was found that overall more stable mRNAs were expressed more highly (figure 2) (W = 9156457, p-value < 2.2e-16). This confirmed that the difference in mRNA stability observed in section 1 could have been the result of varying mRNA stabilities suggesting that essential mRNAs are more stable. Increased RNA stability is generally considered to increase expression levels (D. Cheneval et al., 2010; W. Filipowicz et al., 2008; R. C. Friedman et al., 2009; R. de Sousa Abreu et al., 2009; M. D. Díaz-Muñoz et al., 2018).

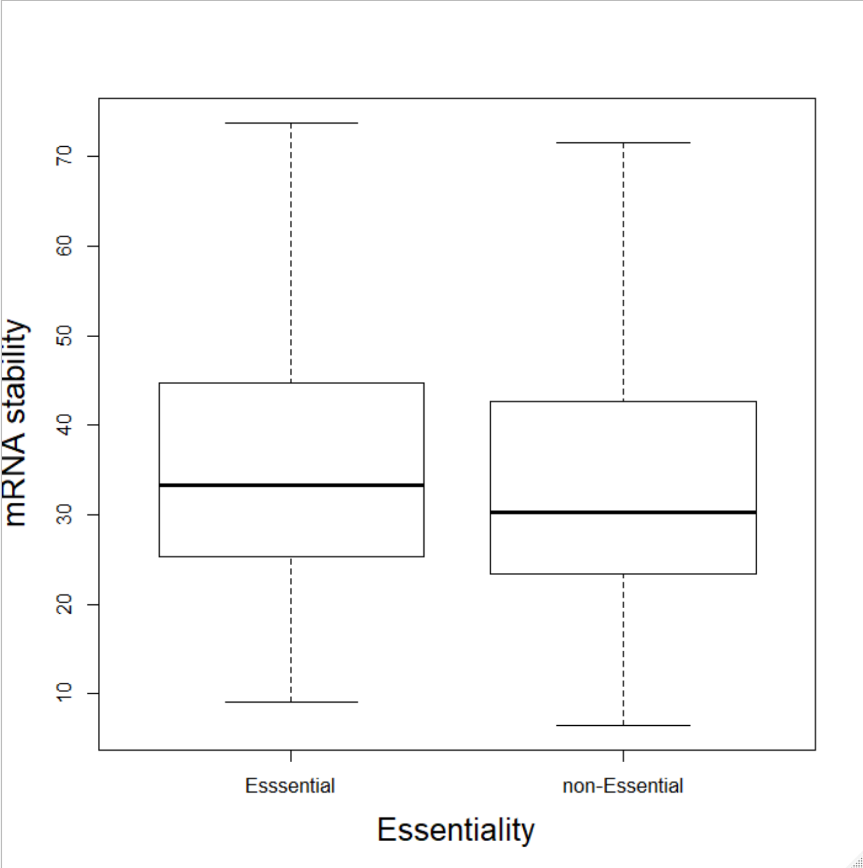
After the influence of stability on expression had been verified I used a Wilcoxon signed-rank test to analyse the difference between the stability of essential and non-essential mRNAs. It was hypothesised that essential mRNAs would have greater stability due to their increased expression. The results (figure 3) showed that essential mRNAs were more stable than non-essential mRNAs on average (W = 2454908, p-value = 3.743e-08). These results suggested that the differences in expression observed in section 1 are due in part to differences in post-transcriptional regulation. Post-transcriptional regulation is known to involve modulation of mRNA stability (D. Cheneval et al., 2010; W. Filipowicz et al., 2008; R. C. Friedman et al., 2009; R. de Sousa Abreu et al., 2009; M. D. Díaz-Muñoz et al., 2018). It is unclear to what extent post-transcriptional regulation is affecting the differences in expression or if differences in expression are correlated with essentiality directly. There is evidence to suggest that mRNAs involved in a particular process can be co-regulated in groups called regulons (A. Hasan et al., 2014).



**Figure 2: mRNA stability increases with expression.**

A Wilcoxon signed rank test was used to determine the relationship between the stability of mRNAs and their level of expression (RPKM). It was found that overall more stable mRNAs were expressed more highly (figure 2) (W = 9156457, p-value < 2.2e-16).

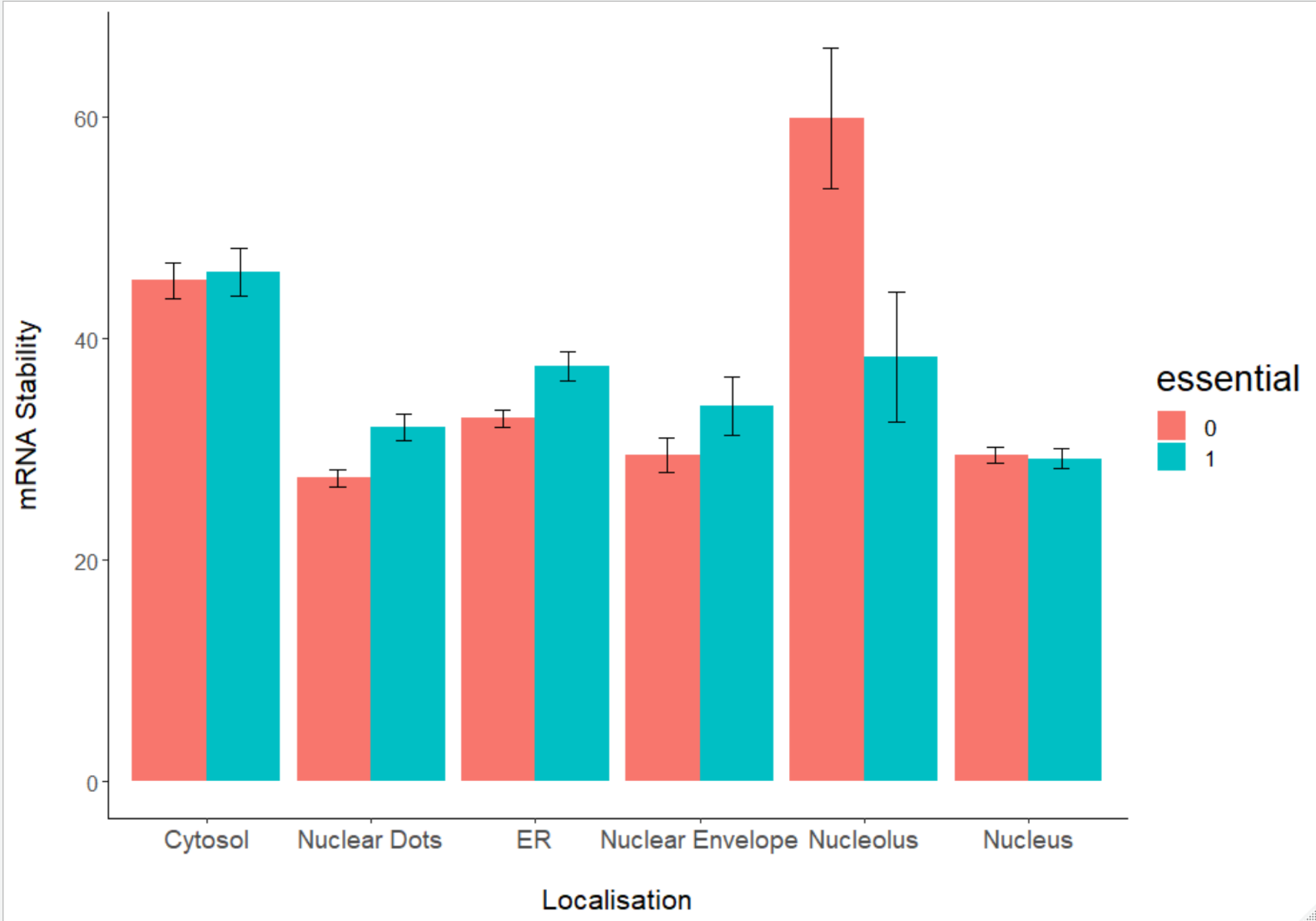
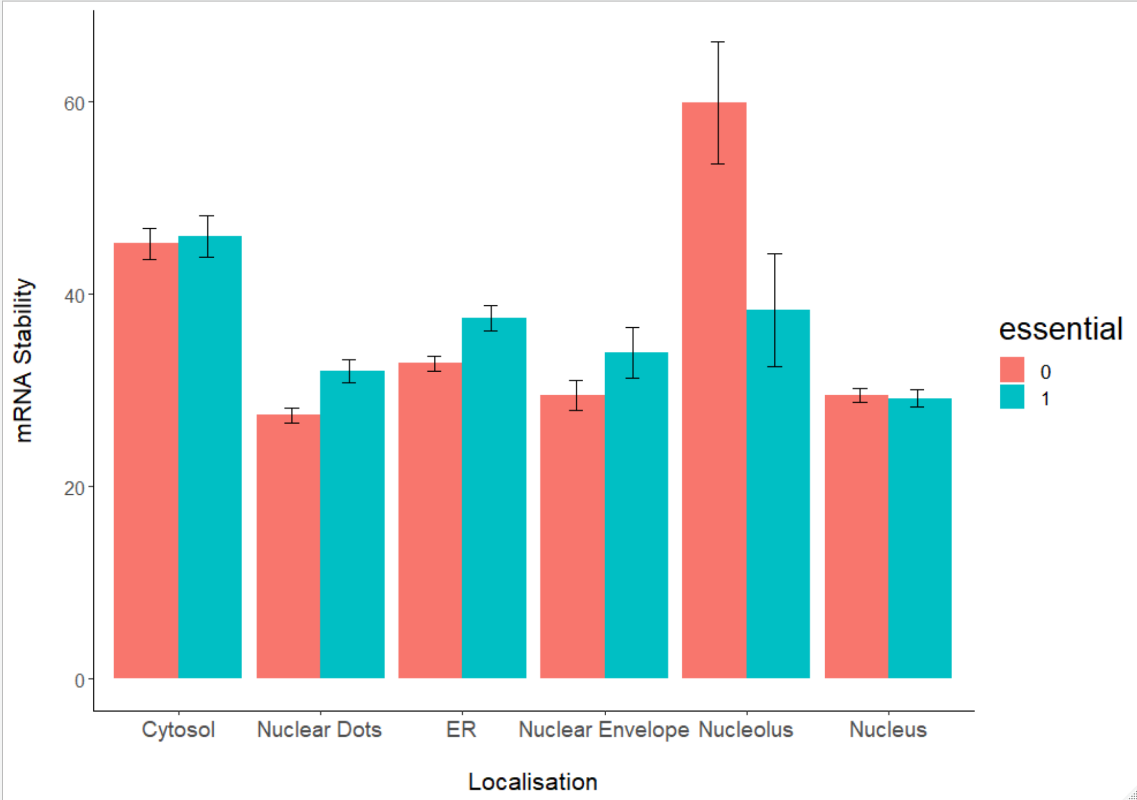
**Figure 3: Essential mRNAs are more stable than non-essential mRNAs on average.** *The box indicates the 1st and 3rd quartiles, and the dark line indicates the median. Maximum and minimum data points are indicate by the bars.* A Wilcoxon signed rank test was used to analyse the difference between the stability of essential and non-essential mRNAs. The results showed that essential mRNAs were more stable than non-essential mRNAs on average (W = 2454908, p-value = 3.743e-08)



**Section 3: The Relationship Between mRNA Stability and Localisation in Essential and Non-Essential mRNAs**

Based on the results of section 2, a difference between the stability of mRNAs encoding proteins of different functions was hypothesised, as essential and non-essential genes generally carry out different functions and are regulated differently as a result. Due to this, it was predicted that there would be differences in the stabilities of mRNAs with respect to their localisations, due to different compartments carrying out different processes. It was also hypothesised that in different protein localisations you would see differences in the relationship between mRNA stability and essentiality if we assume that essentiality does not influence mRNA stability separately from the effect of differing functions.

A two-way ANOVA compared the relationship between localisation and essentiality on the stability of their respective mRNAs. Localisation data was taken from A. Matsuyama et al., 2006. Results (Figure 4) showed a significant main effect of localisation on mRNA stability (ANOVA: F = 43.510; d.f. = 5; p< 2e-16), no significant main difference between essential and non-essential mRNAs (ANOVA: F = 1.357; d.f. = 1; p= 0.24422). An interaction was observed (ANOVA: F = 3.709; d.f. = 5; p= 0.00241), which explains the difference in the effect of essentiality compared to section 2, suggesting that differences seen in previous sections were due to differing proportions of mRNAs of different functions, created by grouping by essentiality. This may be skewed by the unique pattern observed in the nucleolus data. The data may not be reliable as none of the assumptions for ANOVA were met. Sample sizes for nucleolus and nuclear envelope localisations were comparatively small. Function influencing post-transcriptional regulation is supported by the existence of regulons, which coregulate mRNAs post-transcriptionally. Regulons each have unique binding motifs for RBPs (A. Hasan et al., 2014). The existence of regulons may not fully explain the differences observed, as it was found that only 16% of mRNAs were affected by RBP deletion mutants, suggesting that other factors affected RNA stability. Testing with a larger range of localisations and the incorporation of other potential influences of mRNA stability is required to find out more.



Essentiality

Non-Essential

Essential

**Figure 4: mRNA stability varies based on localisation and the difference between the stability of essential and non-Essential mRNAs is dependent on their localisation.** *Bars indicate mean Mean mRNA stability with error bars indicating ± 1 standard error of the mean (SEM).* A two-way ANOVA was used to compare the effect and relationship between localisation and essentiality on the stability of their respective mRNAs. Results showed a significant main effect of localisation on mRNA stability (ANOVA: F = 43.510; d.f. = 5; p< 2e-16), but no significant main difference between essential and non-essential mRNAs (ANOVA: F = 1.357; d.f. = 1; p= 0.24422). An interaction was observed (ANOVA: F = 3.709; d.f. = 5; p= 0.00241).

**Conclusions:**

The results showed that essential genes were expressed more highly than non-essential genes overall. Further testing revealed that essential mRNAs were significantly more stable overall, suggesting that differences in post-transcriptional regulation may be contributing to the pattern of expression. It was also found that mRNA stability differs significantly based on their localisation and that differences in stability between essential and non-essential mRNAs were dependant on their localisation. It was inferred that essentiality may not affect mRNA stability and that the differences due to the function of the proteins they encode instead. All localisations will need to be compared to test this fully. My explanation for these differences was that mRNAs coding for proteins involved in different processes are regulated differently based on the process/processes they are involved in. This idea is supported by the findings of A. Hasan et al., 2014 who found that mRNAs coding for proteins involved in different processes were organised into regulons with unique binding motifs for RBPs. The data from section 3 may not be representative, as the assumptions of ANOVA were not met and the sample sizes for certain localisations were comparatively small. It was also suggested that not all mRNAs are organised into regulons and that there are likely other factors affecting mRNA stability (A. Hasan et al., 2014), for example, there is evidence that increased concentrations of mRNA can increase mRNA stability (S. Nouaille et al., 2017). Comparing the relationship between mRNA function and mRNA stability directly and comparing the impact of various factors on these values could shed more light on the relationship between post-transcriptional regulation mRNA function.

**References**

A. Hasan, C. Cotobal, C. D. S. Duncan, J. Mata, (2014), “Systematic Analysis of the Role of RNA-Binding Proteins in the Regulation of RNA Stability.” *PLoS Genetics* 10(11), e1004684

A. Matsuyama, , R. Arai, Y. Yashiroda, A. Shirai, A. Kamata, S. Sekido, Y. Kobayashi, A. Hashimoto, M. Hamamoto, Y. Hiraoka, S. Horinouchi, M. Yoshida, (2006), “ORFeome Cloning and Global Analysis of Protein Localization in the Fission Yeast Schizosaccharomyces Pombe.” *Nature Biotechnology* 24(7), pp.841–847.

D. Cheneval, T. Kastelic, P. Fuerst, C. N. Parker, (2010), “A Review of Methods to Monitor the Modulation of mRNA Stability: A Novel Approach to Drug Discovery and Therapeutic Intervention.” *Journal of Biomolecular Screening* 15(6), pp.609–622.

D-U. Kim, J. Hayles, D. Kim, V. Wood, H-O. Park, M. Won, H-S. Yoo, et al., (2010), “Analysis of a Genome-Wide Set of Gene Deletions in the Fission Yeast Schizosaccharomyces Pombe.” *Nature Biotechnology* 28(6), pp.617–723.

E. P. C. Rocha, A. Danchin, (2004), “An Analysis of Determinants of Amino Acids Substitution Rates in Bacterial Proteins.” *Molecular Biology and Evolution* 21(1), pp.108–116.

M. D. Díaz-Muñoz, M. Turner, (2018), “Uncovering the Role of RNA-Binding Proteins in Gene Expression in the Immune System.” *Frontiers in Immunology* 9(May), pp.1094.

“Pombase.” n.d. Accessed April 14, 2020. <https://www.pombase.org/>.

R. C. Friedman, K. Kai-How Farh, C. B. Burge, D. P. Bartel, (2009), “Most Mammalian mRNAs Are Conserved Targets of microRNAs.” *Genome Research* 19(1), pp.92–105.

R. de Sousa Abreu, L. O. Penalva, E. M. Marcotte, C. Vogel, (2009), “Global Signatures of Protein and mRNA Expression Levels.” *Molecular bioSystems* 5(12), pp.1512–1526.

S. R Atkinson, S. Marguerat, D. A. Bitton, M. Rodríguez-López, C. Rallis, J-F. Lemay, C. Cotobal, M. Malecki, P. Smilaowski, J. Mata, P.Korber, F. Bachland, J. Bäler, (2018). “Long Noncoding RNA Repertoire and Targeting by Nuclear Exosome, Cytoplasmic Exonuclease, and RNAi in Fission Yeast.” *RNA* 24(9), pp.1195–1213.

S. Nouaille, S. Mondeil, A-L. Finoux, C. Moulis, L. Girbal, M. Cocaign-Bousquet, (2017). “The Stability of an mRNA Is Influenced by Its Concentration: A Potential Physical Mechanism to Regulate Gene Expression.” *Nucleic Acids Research* 45(20), pp.11711–11724.

W. Filipowicz, S. N. Bhattacharyya, N. Sonenberg, (2008), “Mechanisms of Post-Transcriptional Regulation by microRNAs: Are the Answers in Sight?” *Nature Reviews. Genetics* 9(2), pp.102–114.

Word Count: 1491

**Supplementary Methods**

#######################################################################################

# #

# IMPORTING THE DATA #

# #

#######################################################################################

setwd("~/Biology/R/Biology/Big Data Biology/Data")

#First I load in the two data sets required for my analysis

load(url("https://www-users.york.ac.uk/~dj757/BIO00047I/data/yeast\_data.28-02-2020.Rda"))

angeli <- read.delim("AnGeLiDatabase.txt",h=T)

#I then check to see if the data has been imported correctly

ls()

class(gene)

nrow(gene)

ncol(gene)

head(gene)

class(angeli)

nrow(angeli)

ncol(angeli)

head(angeli)

#Here I insert cytosol an ER localisation data from the angeli database into the gene dataframe

#First I remove gene ontology columns

#I get a the vector of the column names

nam <- names(angeli)

#Using grep I can locate gene ontology or "GO." columns

#I invert the grep, so it does not locate these columns

not.go.columns <- grep("GO.",nam,invert=T)

#using this I create a subset of angeli without GO columns

angeli2 <- angeli[,not.go.columns]

#I also need to remove 'FYPO', (fission yeast phenotype ontology) or PFAM groups

#so I remove them with grep also

not.fypo.or.pfam.columns <- grep("FYPO|PF",names(angeli2),invert=T)

angeli3 <- angeli2[,not.fypo.or.pfam.columns]

#I create a subset with just the information rows so I can locate the data I need

info <- angeli3[1:7,]

View(info)

#I also create a subset with only the rows with data

gene.data <- angeli3[8:7012,]

#renaming first column

names(gene.data)[1]="gene"

#I create subset of the data containing just the two columns I need

my.data.col <- which(names(gene.data) == "Cytosol")

my.data.col2 <- which(names(gene.data) == "ER")

angeli4 <- gene.data[,c(1,my.data.col, my.data.col2)]

View(angeli4)

#I combine angeli4 with the gene dataframes, and save it as an Rda file to save

#time when reopening the script

gene\_final<-merge(gene,angeli4,by="gene",all=T)

save(gene\_final,file="gene\_finaldata.Rda")

#Finally I check that the dataframe has been generated correctly

class(gene\_final) #data.frame

nrow(gene\_final)#how many rows?

#7009

ncol(gene\_final)#columns?

#25

head(gene\_final)#first few rows

#I can use this file for future use without having to import the whole database again

load(file="gene\_finaldata.Rda")

#######################################################################################

# #

# DATA ANALYSIS #

# #

#######################################################################################

######

######Section 1: Are Essential Genes Expressed Differently to Non-Essential Genes?

######

#first I make a new data frame with just protein coding data

prot\_pre <- subset(gene\_final, protein\_coding ==1)

#Checking rows have been removed

nrow(prot\_pre)

#I make the essential column into a factor (it’s currently numeric) This will change the

#data from continuous to discrete categorical which is what it should be.

prot\_pre$essential <- as.factor(prot\_pre$essential)

#I use this to make sure any rows that are missing data are removed

prot <- subset(prot\_pre, !is.na(essential))

nrow(prot) #fortunately none were removed at this stage

#I now split the protein coding data into essential and non essential proteins

ess <- subset(prot, essential==1)

non.ess <- subset(prot, essential==0)

#Plotting the difference in expression of essential and non-essential genes

boxplot(

log10(ess$gene.expression.RPKM),

log10(non.ess$gene.expression.RPKM),

cex.axis=1.05,

cex.lab=1.5,

ylab="Gene Expression log10(RPKM)",

xlab="Essentiality",

names=c("Esssential","non-Essential"),

outline=F)

#log 10 in order to visualise data more clearly

#Outliers removed for a clearer plot

#I use a wilcoxon signed rank test to determine significance

wilcox.test(ess$gene.expression.RPKM, non.ess$gene.expression.RPKM)

#the difference observed is very unlikely to occur by chance (Null hypothesis rejected) (W = 3039370, p-value < 2.2e-16)

#essential genes are more highly expressed than non-essential genes in general

######

######Section 2: The relationship between mRNA Stability and RPKM in Essential and Non-Essential mRNAs

######

#Plotting the relationship between mRNA stability and RPKM

plot(log10(prot$mRNA.stabilities), log10(prot$gene.expression.RPKM),

xlab = "log10(mRNA Stability)",

ylab= "Gene Expression Log10(RPKM)",

cex.axis=1.05,

cex.lab=1.5)

#logs of both datasets were taken to visualise the trend easier

#I use another wilcoxon signed rank test to determine significance

wilcox.test(prot$mRNA.stabilities, prot$gene.expression.RPKM)

#Null hypothesis rejected (W = 9156457, p-value < 2.2e-16)

#More stable mRNAs tend to be more highly expressed

#Plotting the relationship between esentiality and mRNA stability

boxplot(ess$mRNA.stabilities, non.ess$mRNA.stabilities,

outline=F,

ylab="mRNA stability",

cex.axis=1.05,

xlab="Essentiality",

cex.lab=1.6,

names=c("Esssential","non-Essential"))

#Outliers removed for a clearer plot

#Third wilcoxon signed rank test to determine significance

wilcox.test(ess$mRNA.stabilities, non.ess$mRNA.stabilities)

#Null hypothesis rejected (W = 2454908, p-value = 3.743e-08)

#mRNA of essential genes is more stable on average

######

######Section 3: The Relationship Between mRNA Stability and Localisation in Essential and Non-Essential mRNAs

######

#First I need to make localisation a group variable

#I create subsets of each localisation

Prot\_dotssubset <- subset(prot, Nuclear\_dots ==1)

Prot\_envsubset <- subset(prot, Nuclear\_envelope ==1)

Prot\_nolussubset <- subset(prot, Nucleolus ==1)

Prot\_nucsubset <- subset(prot, Nucleus ==1)

Prot\_cysubset <- subset(prot, Cytosol ==1)

Prot\_ersubset <- subset(prot, ER ==1)

#And change the data from binary (0 or 1) to a category e.g nucleolus ("nolus")

Prot\_dotssubset$localisation="dots" #nuclear dots

Prot\_envsubset$localisation="n.en" #nucler envelope

Prot\_nolussubset$localisation="nolus" #nucleolus

Prot\_nucsubset$localisation="nuc" #nucleus

Prot\_cysubset$localisation="cyt" #cytoplasm

Prot\_ersubset$localisation="ER" #ER

#I re-merge the subsets, but now all localisations are in one column

library(dplyr)#for the bind\_rows function

locale <- bind\_rows(Prot\_dotssubset,Prot\_envsubset,

Prot\_nolussubset,Prot\_nucsubset,

Prot\_cysubset,Prot\_ersubset,)

##################################################################################

#I remove N/a values so that the summary function works

locale\_mrna<- subset(locale, !is.na(mRNA.stabilities))

library(Rmisc)#for the summarySE function

#I create summary data for the plot

summary <- summarySE(locale\_mrna, measurevar = "mRNA.stabilities", groupvars = c("localisation", "essential"))

summary

#Two-way ANOVA is carried out with mRNA stability being an explanatory variable and

#localisation and essentiality as response variables

mod <- aov(data = locale\_mrna, mRNA.stabilities ~ localisation \* essential)

summary(mod)

#Significant main effect of localisation on mRNA stability (ANOVA: F = 43.510; d.f. = 5; p< 2e-16)

#No significant main effect of essentiality on mRNA stability (ANOVA: F = 1.357; d.f. = 1; p= 0.24422)

#There was an interaction (ANOVA: F = 3.709; d.f. = 5; p= 0.00241)

#Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

#

#TukeyHSD provides detail on the differences of each localisations

TukeyHSD(mod)

#I chose not to refer to them in the text as the intention was just to determine

#whether there was a main effect and an interaction

#Interesting that the Nucleoulus pattern is completely different from the rest however

#Testing the normal distribution assumption

shapiro.test(mod$residuals)

#Null hypothesis rejected (W = 0.81428, p-value < 2.2e-16)

#this assumption is not met as the data is not normally distributed

#Testing equal variances assumption

library(car)#for the leveneTest function which can test for heterogeneity of variance

leveneTest(mRNA.stabilities ~ localisation \* essential, data = locale\_mrna)

#Null hypothesis rejected (p-value < 2.2e-16)

#equal variances assumption not met

library(ggplot2)#For ggplot function

#barplot of mRNA stability with respect to localisation an essentiality

ggplot(data = summary, aes(x=localisation, y=mRNA.stabilities, fill=essential,)) +

geom\_bar(stat= "identity",position=position\_dodge())+

geom\_errorbar(aes( ymin = mRNA.stabilities - se,

ymax = mRNA.stabilities + se),

width = 0.2,

position = position\_dodge(0.8),

colour = c("Black"), size = 0.5) +

ylab("mRNA Stability") +

scale\_x\_discrete(breaks=c("dots", "n.en", "nolus", "nuc", "cyt", "ER"),

labels=c("Nuclear Dots", "Nuclear Envelope", "Nucleolus", "Nucleus", "Cytosol", "ER"))+

theme(axis.title.x = element\_text(size = 16, margin = margin(t = 20, r = 40, b = 0, l = 0)),

axis.title.y = element\_text(size = 16, margin = margin(t = 1, r = 20, b = 0, l = 0)),

legend.title = element\_text(size= 20),

legend.text = element\_text(size = 13),

axis.text.x = element\_text(size = 14),

axis.text.y = element\_text(size = 12),

panel.background = element\_rect(fill = "white"),

axis.line.x = element\_line(color = "black"),

axis.line.y = element\_line(color = "black"),

legend.key = element\_blank())+

xlab("Localisation")