Essential genes are more highly conserved than nonessential genes in the fission yeast, Schizosaccharomyces pombe.

Exam Number: Y3858139

Abstract:

A high gene conservation rate indicates that the gene has remained unchanged throughout evolution. The evolutionary conservation rate can inform one about the function of the gene. Previous studies have shown genes with an essential function are more highly conserved than nonessential genes. However, one study analysing the *Saccharomyces cerevisiae* genome did not find any significant difference in conservation rate between the two categories. Therefore, further investigation was needed. In this study, gene conservation was compared to several factors using analysis of the fission yeast genome. The results show a significantly higher rate of gene conservation in essential genes compared to nonessential genes. There was a negative correlation observed between knockout fitness and gene conservation. There was positive correlation with gene conservation and two factors that are high in essential genes - gene expression and mRNA stability. Overall, one can conclude that the evidence does support the hypothesis.

Introduction:

Gene conservation is the process where a gene does not alter majorly throughout its evolution. The conversation rate can differ enormously between protein coding genes, with a higher conversion rate signifying a slower gene evolution (Choi et al. 2007). The gene conversion rate can inform one about the properties of the gene. One such example is that it has been predicted using the 'knockout hypothesis' that essential genes are more evolutionarily conserved than non-essential genes (Hurst & Smith 1999). Essential genes are indispensable for the survival of the organism. Examples include genes coding for proteins involved in key cellular processes such as transcription and translation. Essential genes also have certain properties that distinguish them such as a higher expression (Wang et al. 2015) and high protein stability (Chen et al. n.d.). In the early stages of the Millennium, a study (Jordan et al. 2002) confirmed the knockout hypothesis to be true in *Escherichia coli* through the use of genomic analysis. The study shown that essential genes have smaller synonymous (Ks) and nonsynonymous (Ka) substitution rates. The KS/KA ratio is used as a measurement of evolutionary conservation as it determine the selective pressure on a gene (Wang et al. 2009). Only recently had there been enough data available to expand on this prediction. The knockout hypothesis was tested that in 23 different bacteria genomes, and agreed that essential genes in had smaller KS/KA rates (Luo et al. 2015), and thus greater evolutionary conservation than non-essential genes. The hypothesis is not without debate

however. One study (Hirsh & Fraser 2001) found no significant difference in the evolutionarily conserved rate of essential and nonessential genes in *Saccharomyces cerevisiae*. Therefore, the topic warrants further exploration to progress onto a clear answer. To further investigate this thesis, genomic analysis of the model organism, fission yeast *Schizosaccharomyces pombe* was carried out to see whether gene conversation is higher in essential genes. This was done by comparing gene conservation to gene essentiality and properties associated with it.

Gene Conversation in Fission Yeast:

The top 100 genes with the highest phyloP scores were visualized in Pombase (Anon n.d.). PhyloP is the measurement of evolutionary conservation at individual alignment sites. Therefore, a higher phyloP score signifies that the gene is more evolutionarily conserved (Grech et al. 2018). The visualization shown that 8/10 of the genes were viable for deletion (**fig.1**), showing that the majority of the highest conserved genes in fission yeast are non-essential. This rejects the hypothesis.

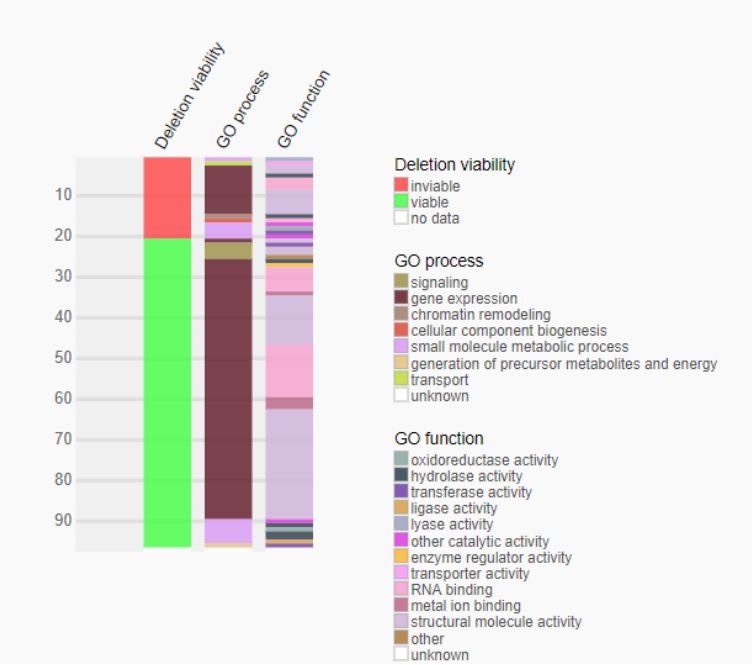

Figure 1. The visualisation shows the deletion viability, GO process and GO functions of the 100 most conserved genes is the *Schizosaccharomyces pombe* genome. 77/100 genes had deletion viability.

However, the properties of the gene list is similar to the characterization of essential genes. The visualization shown that the majority of the genes (77%) were involved in the process of gene expression, while next two highest categories were cellular signalling and metabolic processes, all of which are essential processes (**fig.1**) (Radhakrishnan et al. 2010;

Gutteridge et al. 2007),. The functions of the 100 most highly conserved genes included structural molecule activity and RNA binding. (**fig.1)**

Overall, the visualization did not give any definite answer to the hypothesis, so more analysis needed to be carried out.

Essential genes are more conserved than non-essential genes:

Firstly, it was predicted that essential genes have a higher rate of evolutionary conservation than nonessential genes. The phyloP score was compared in both categories of genes to answer the question. It was shown that essential genes had larger phyloP scores than nonessential than predicted by chance (Wilcoxon Rank Sum test, P < 2.2e-16, **fig. 2**). All analysis was performed using R.studio.

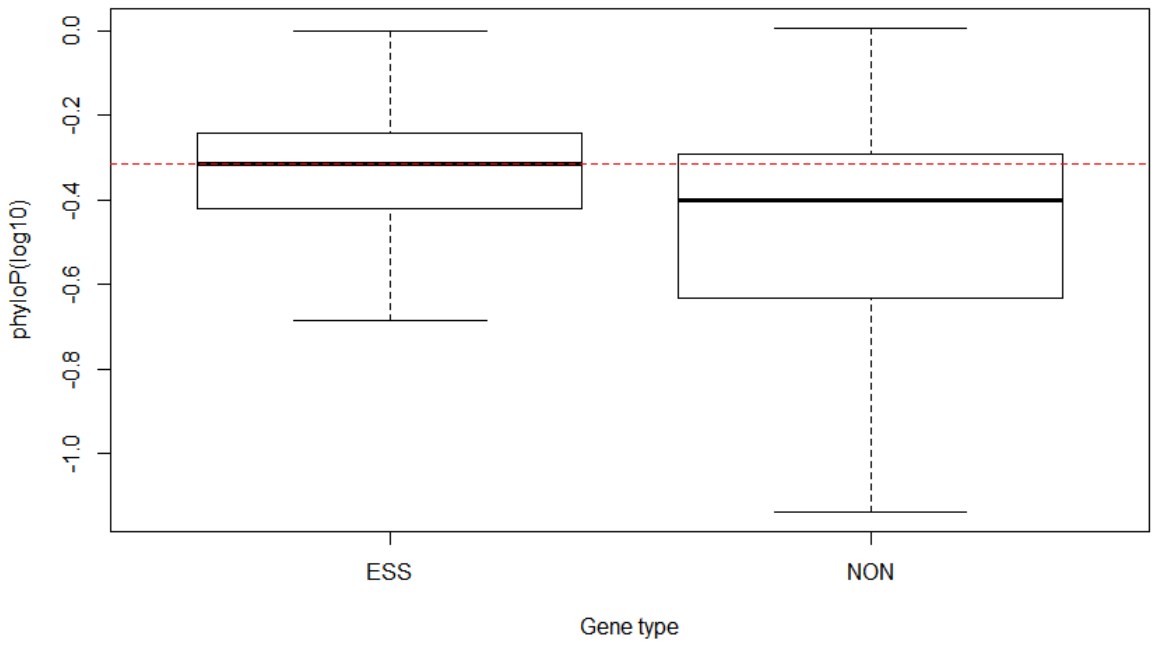

Figure 2. PhyloP scores in two categories of gene - Essential and nonessential in *Schizosaccharomyces pombe*. Genomic analysis shown essential genes have a higher mean phyloP than nonessential (Wilcoxon Rank Sum test, P < 2.2e-16). The graph shows the median and the upper and lower quartile of each category. The red dotted line signifies the median PhyloP score of the essential gene category.

This supports the hypothesis that essential genes have a higher evolutionary conservation rate than non-essential. This was expected as significant mutations to essential genes would result becoming non-functional (Kemphues 2005).

The result agrees with the previous studies of Jordan and Lou, that show essential genes are more conserved. The result disagree with the visualization in Pombase which shown that

the majority of high conserved genes were viably deleted. There might be conflicting evidence as the gene list used in Pombase is approximately 50x smaller than the data set used for the Wilcoxon test. Additionally, one limitation of this result is that the PhyloP was the only conservation score used, therefore not giving the most accurate result. In future analysis, alternative conservation scores such as PhastCons (Siepel et al. 2005) would be used in conjunction with PhyloP.

Overall, the result agreed with previous studies that the hypothesis is true. However, because only one type of conservation score is used and because it conflicts with the previous analysis in Pombase, further analysis using a range of conservation scores will needed to be carried out.

Gene conservation and knockout fitness:

It was been predicted that gene conservation and knockout fitness would be negatively correlated. The colony size with the knockout gene was used as a proxy for knockout fitness and compared the PhyloP score (Malecki et al. 2016). Knockout fitness can be used as an alternative measurement to show how essential genes are (Kim et al. 2019). As a gene knockout with a larger colony size signifies that the organism can grow and proliferate without it, showing the gene that was knocked out is nonessential to the survival of the cell. Gene conservation was slightly significantly negatively correlated to knockout fitness (Spearman rank correlation, R = -0.143, P=6.109e-16, **fig.3**).

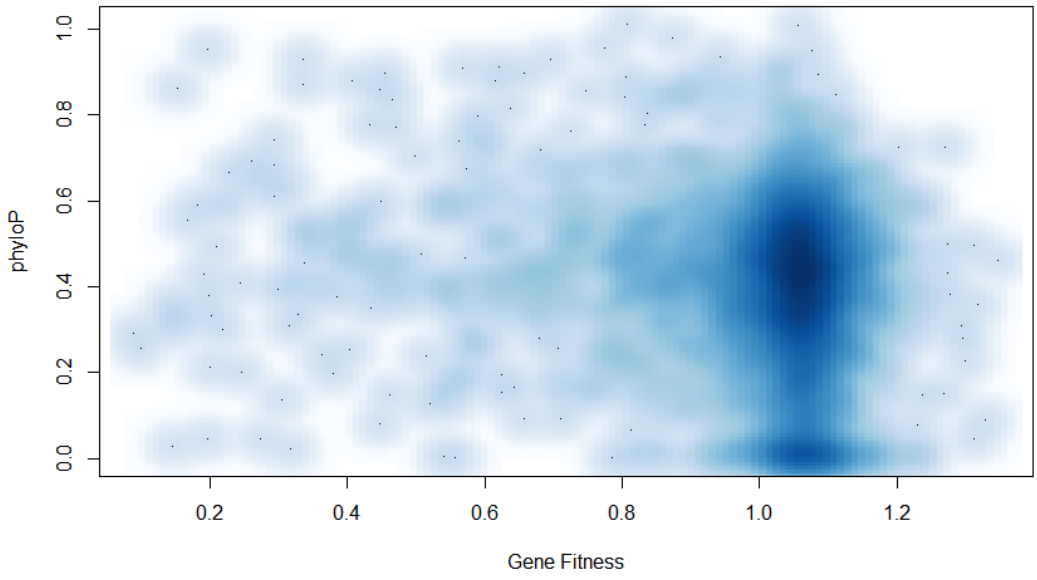

Figure 3. PhyloP scores were compared the gene knockout fitness to compared evolutionary conservation rates and knockout fitness in the *Schizosaccharomyces pombe* genome. Colony size was used as a proxy for gene

knock fitness. There was a significant slight negative correlation. (Spearman rank correlation, R = -0.143, P = 6.109e-16,)

This agrees with the previous result of essential genes being more evolutionary conserved than nonessential genes. It also provides evidence to argue Hurst and Smith's Knockout Hypothesis is correct. However, seeing as the correlation was only considered slightly negative (0.-143), it was not be considered a strong enough result to confirm the hypothesis. Therefore, more exploration into this would be needed.

Overall, gene conservation and knockout fitness show a highly significant weak positive correlation. Knockout fitness is an indicator to how essential the gene is the colony survival. Therefore, the correlation links to the previous hypothesis.

Gene conservation and gene expression:

To test if higher conserved genes had a higher expression rate, phyloP scores were compared to RPKM scores (The RNA expression level from RNA-seq, from proliferating cells) (Atkinson et al. 2018). It was shown that higher phyloP scores positively correlated with RPKM significantly (Spearman rank correlation R = 0.283, P < 2.2e-16, **fig.4**) which agrees with the hypothesis. The correlation was weak.

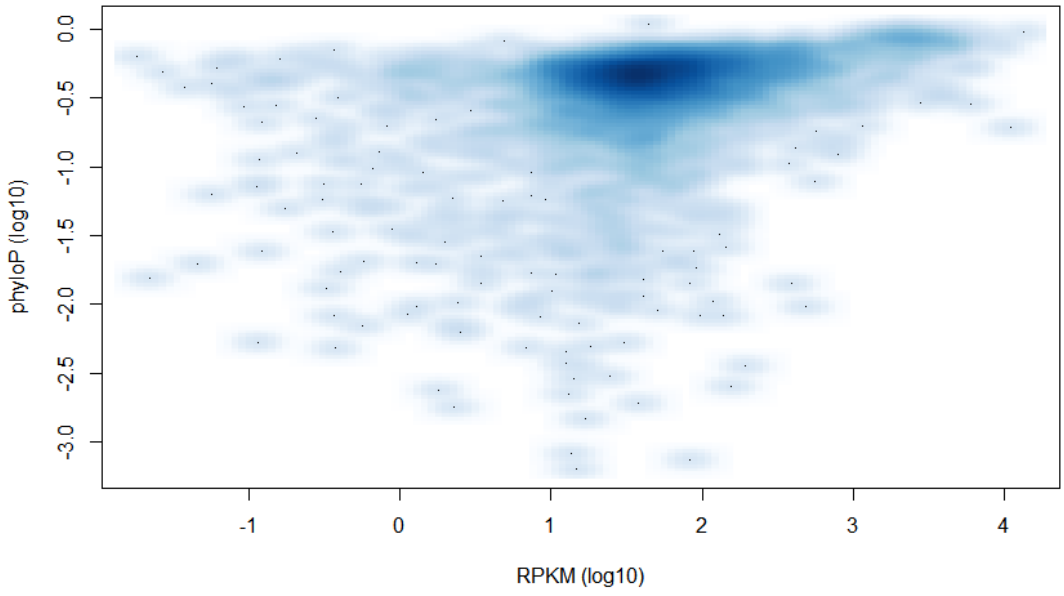

Figure 4. Log10 PhyloP scores was used to measure conservation rate and compared to the gene expression rate (log10 (RPKM)) in *Schizosaccharomyces pombe* genome. There was a significant weak positive correlation, (Spearman rank correlation R = 0.283, P < 2.2e-16, fig.4)

Previous studies have shown that higher expressed genes have a slower evolutionary rate (Pál et al. 2001) however the reason why this occurs is debated. Gene expression was compared to gene conservation as it has already been established that essential genes are higher expressed than non-essential genes (Wang et al. 2015) (Pál et al. 2001). Thus, it could be considered that higher gene expression is a marker of essential genes. However, it is debated that higher levels of gene expression is not determine by function it codes, but rather a mechanism of protein misfolding. (Drummond et al. 2005).

Overall, it is shown that higher gene conservation is weakly correlated to gene expression despite the reason why being debated. Higher gene expression is also considered a characteristic of essential gene thus linking the three categories. However, there is debate within the previous studies, so this result does not hold significant weigh in the hypothesis of Essential gene being highly conserved.

Gene Conservation and mRNA stability:

It was predicted that gene conservation and mRNA stability would be positively correlated. The measurement used was the half-life of the mRNA in minutes, with a larger half-life showing a more stable mRNA (Hasan et al. 2014). The result show gene conservation is weakly positively correlated to mRNA stability than predicted by chance (Spearman rank correlation, R = 0.295, P < < 2.2e-16, fig.5).

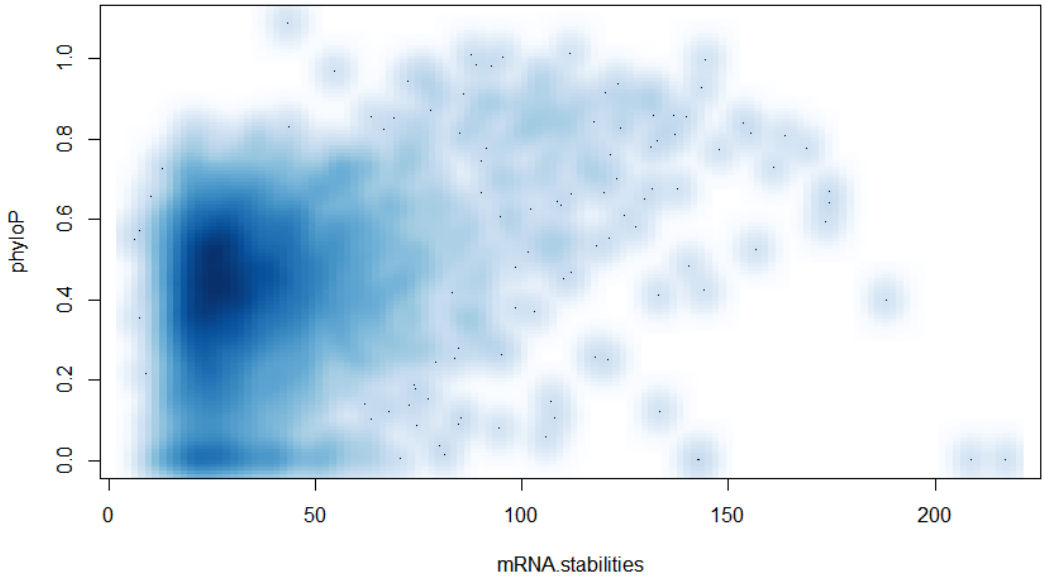

Figure 5. Log10 PhyloP score was used to measure conservation rate and compared to mRNA stabilities in *Schizosaccharomyces pombe* genome. There was a significant positive correlation (Spearman rank correlation, R = 0.295, P < < 2.2e-16)

There is little related studies into the two factors, although it has been proved that mutations at conserved sites transcribe more stable proteins (Sullivan et al. 2012)

It is logical that genes with higher conservation rate are more stable; if it hypothesised that more conserved genes are essential to the cell. It has already been established that essential genes from more stable mRNA complexes than nonessential genes (Zhang et al. 2015) suggesting that high conservation rate mRNA stability are linked features of essential genes.

Overall, the result shows that gene conservation and mRNA stability is significantly weakly correlated. High mRNA stability has also been linked with essential genes.

Conclusion:

In conclusion, the findings agree with the majority of previous studies suggesting that essential genes are more conserved than nonessential genes. Comparing phyloP in essential and nonessential genes, gave a significantly higher result in essential genes. However, to create a more accurate result, future analysis would need to use a range of conservation scores. Knockout fitness can be used as another indicator to how essential a gene is, by comparing gene conversation and knockout fitness there was shown to be significantly but weak correlation. Plus, gene conservation was compared to factor that has been established to signify essential protein coding genes - high expression and high mRNA stability. Both factors shown a weak positive correlation to gene conversion, which was expected due to the previous studies in the topics. However, there is some debate to whether the function of the gene relating to expression, meaning that it cannot be used as measurement for Gene essentiality. Overall, despite the weak correlations and limitations of the data, the evidence just suggest that essential genes are more highly conserved than non-essential genes in the fission yeast Schizosaccharomyces pombe.

References:

Anon, Pombase. Available at: https://www.pombase.org/ [Accessed May 6, 2019].

Atkinson, S.R. et al., 2018. Long noncoding RNA repertoire and targeting by nuclear exosome, cytoplasmic exonuclease, and RNAi in fission yeast. RNA , 24(9), pp.1195–1213.

Chen, H. et al., 2018. New insights on human essential genes based on integrated analysis. bioRxiv. Available at: http://dx.doi.org/10.1101/260224.

Choi, J.K. et al., 2007. Impact of transcriptional properties on essentiality and evolutionary rate. Genetics, 175(1), pp.199–206.

Drummond, D.A. et al., 2005. Why highly expressed proteins evolve slowly. Proceedings of the National Academy of Sciences of the United States of America, 102(40), pp.14338–14343.

Grech, L. et al., 2018. Fitness Landscape of the Fission Yeast Genome. bioRxiv, p.398024. Available at: https://www.biorxiv.org/content/10.1101/398024v1 [Accessed May 3, 2019].

Gutteridge, A., Kanehisa, M. & Goto, S., 2007. Regulation of metabolic networks by small molecule metabolites. BMC bioinformatics, 8, p.88.

Hasan, A. et al., 2014. Systematic analysis of the role of RNA-binding proteins in the regulation of RNA stability. PLoS genetics, 10(11), p.e1004684.

Hirsh, A.E. & Fraser, H.B., 2001. Protein dispensability and rate of evolution. Nature, 411(6841), pp.1046–1049.

Hurst, L.D. & Smith, N.G., 1999. Do essential genes evolve slowly? Current biology: CB, 9(14), pp.747–750.

Jordan, I.K. et al., 2002. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. Genome research, 12(6), pp.962–968.

Kemphues, K., 2005. Essential genes, WormBook.

Kim, E. et al., 2019. A network of human functional gene interactions from knockout fitness screens in cancer cells. Life science alliance, 2(2). Available at: http://dx.doi.org/10.26508/lsa.201800278.

Luo, H., Gao, F. & Lin, Y., 2015. Evolutionary conservation analysis between the essential and nonessential genes in bacterial genomes. Scientific reports, 5, p.13210.

Malecki, M. et al., 2016. Functional and regulatory profiling of energy metabolism in fission yeast. Genome biology, 17(1), p.240.

Pál, C., Papp, B. & Hurst, L.D., 2001. Highly expressed genes in yeast evolve slowly. Genetics, 158(2), pp.927–931.

Radhakrishnan, K. et al., 2010. Quantitative understanding of cell signaling: the importance of membrane organization. Current opinion in biotechnology, 21(5), pp.677–682.

Siepel, A. et al., 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome research, 15(8), pp.1034–1050.

Sullivan, B.J. et al., 2012. Stabilizing proteins from sequence statistics: the interplay of conservation and correlation in triosephosphate isomerase stability. Journal of molecular biology, 420(4-5), pp.384–399.

Wang, D. et al., 2009. How do variable substitution rates influence Ka and Ks calculations? Genomics, proteomics & bioinformatics, 7(3), pp.116–127.

Wang, T. et al., 2015. Identification and characterization of essential genes in the human genome. Science, 350(6264), pp.1096–1101.

Zhang, R. et al., 2015. Genes with stable DNA methylation levels show higher evolutionary conservation than genes with fluctuant DNA methylation levels. Oncotarget, 6(37), pp.40235–40246.

Supplementary methods:

```
############################ Big Data Assessment ############################

############################### Introduction ################################

# The data set I will be exploring will be the fission yeast Schizosaccharomyces pombe.
# The data shows many properties of Fission yeast - both Quantitative and Categorical.
# I will be exploring the measurement of gene conservation and whether this has an impact
on other  factors.
#Gene conservation shows how slowly the gene has evolved.

################################# Set up ################################
# Set working directory
setwd("M:/Big data/Data")

# Packages
# For figures
library(KernSmooth)
# Wand, M. P. and Jones, M. C. (1995). Kernel Smoothing. Chapman and Hall, London.


############################ Load the data ################################

# Load the Fission yeast data
data <-load(url("http://www-
users.york.ac.uk/~dj757/BIO00047I/data/fission_yeast_data.2018-11-21.Rda"))

################################Explore the data################################

# The object I have chosen is gene conservation. Let's answer the question of what type of
genes
# are more conserved.
# I can do this by make a gene list and looking on Pombase

# Look at class
```

```
class(gene$conservation.phyloP)
# "numeric"
# Gene conservation is quantitative data.
# The higher the value of the phyloP score, the more conserved the gene is

# Firstly I need to make a subset of the gene table that contain only protein-coding genes
# As these are the only ones I'm interested in.
prot <- subset(gene, protein_coding ==1)

# To visualize my data, I need to make a histogram of this subset
hist(prot$conservation.phyloP)
# The histogram shows the frequency of the phyloP score in protein coding gene.
# Shows not normally distributed

# I'm interested in the top 100 conserved genes
# I need to make a subset of the prot table that contain only genes with high conservation
# Using the histogram, I can estimate for around 100 gene
conserved <- subset(prot, conservation.phyloP > 0.80)
# Count how many you get
nrow(conserved)
# [1] 97

# I can now extract a gene list from this table and output this to a file,
# so we can use it in PomBase
# Name gene list
top_conversationgenelist <- conserved$gene
# Output a table of data to a file called top_conversationgenelist.txt
write.table(top_conversationgenelist,file="top_conversationgenelist.txt",col.names =
F,quote=F, row.names = F)

# Now I can put my gene list into Pombase
# By looking at the visualization of my data in Pombase, the data shows that majority of the
gene GO process is
# gene expression.
# It also shows that the genes function are essential process such as RNA binding and
enzyme binding activity.
# This suggests that the greater conserved genes, have an important role to play in the cell.
# Now I can ask questions.

# My first question I want to answer is are essential genes more conserved than non
essential?
# I can show this by using a box and whiskers plot

# Firstly, I need to make two subset of the data:
# 1) the essential protein-coding genes 2) non-essential protein-coding genes
ess.prot <- subset(gene, protein_coding == 1 & essential==1)
non.ess.prot <- subset(gene, protein_coding == 1 & essential==0)

# Now, I can make a box plot comparing the two categories to gene conservation
boxplot(
  log10(ess.prot$conservation.phyloP),
  log10(non.ess.prot$conservation.phyloP),
  ylab=" phyloP(log10)",
  xlab="Gene type",
```

```
  names=c("ESS","NON")
)
ess.med <- median(ess.prot$conservation.phyloP,na.rm=T)
abline(h=log10(ess.med),col="red",lty=2)
# The red dotted line shows the median conservation of the essential and shows you that the
non essential
# median is below this. It makes it easier to visualize.
# However, the plot is hard to read due to the outliers.
# I used log as numbers were small

# Remove the outliers
boxplot(
  log10(ess.prot$conservation.phyloP),
  log10(non.ess.prot$conservation.phyloP),
  outline =  FALSE,
  ylab="Conservation (mean phyloP(log10))",
  xlab="Gene type",
  names=c("ESS","NON")
)
ess.med <- median(ess.prot$conservation.phyloP,na.rm=T)
abline(h=log10(ess.med),col="red",lty=2)
# This makes the data much easier to read.
# This plot shows that essential genes are more conserved
# Let's test if this difference is significant by using a wilcoxon test.
wilcox.test(ess.prot$conservation.phyloP,non.ess.prot$conservation.phyloP)
# < 2.2e-16
# This confirms that the difference is significant.

# Now that I've answered that question, we can ask another.
# Are conservation of genes and gene expression correlated ?
# To find this out, we can make a scatter graph.

# Firstly, lets look at gene expression distribution to show its not normally distributed
hist(prot$gene.expression.RPKM)
# Using the Log10 function will make the graph easier to visualize.
hist(log10(prot$gene.expression.RPKM))

# Now, I can make a scatter graph to see if gene conservation and gene
# expression are correlated
plot(
  prot$gene.expression.RPKM,
  prot$conservation.phylo
)

# Use log10 to make the graph easier to visualize
plot(
  log10(prot$gene.expression.RPKM),
  log10(prot$conservation.phylo)
)
# This looks better, but it is still hard to read due to number of points.

# I can use the Kern Function to produce a Smooth Scatter Graph
smoothScatter(
  log10(prot$gene.expression.RPKM),
```

```
  log10(prot$conservation.phylo),
  xlab = ‘RPKM (log10)',
  ylab = 'phyloP (log10)' )
# This is a lot more easy on the eye. The darker the blue, the higher the
# frequency of genes
# The factors look positively correlated

# It also looks like the two factors are correlated.
# I will carry out a cor.test to confirm if it is significant.
cor.test(prot$gene.expression.RPKM,prot$conservation.phylo)
# Pearson's product-moment correlation
# p-value < 2.2e-16
# cor = 0.2833064

# The P-value shows there is significance, while cor values show there is a small correlation.
# Therefore, I can interrupt that there is a significant small correlation.

# solid.media.KO.fitness measures colony size as a proxy for knockout 'fitness'
# Therefore, bigger the colony the less essential the gene is for survival
# Make a histogram of Fitness
hist(prot$solid.media.KO.fitness)
# Doesn't look normally distributed

# Make a scatter to show correlation
smoothScatter(
  prot$solid.media.KO.fitness,
  (prot$conservation.phylo),
  xlab = ' Knockout Fitness',
  ylab = 'phyloP')
# The factors look negatively correlated

# Do a core test to show significance
cor.test(prot$solid.media.KO.fitness,prot$conservation.phylo)
# p-value = 6.109e-16, -0.1430888
# The results show a slight negative correlation

# Now, i will see are highly conserved genes more stable?
# Make a histogram of mRNA stability
hist(prot$mRNA.stabilities)
# Doesn't look normally distributed

# Make a scatter graph to show correlation
smoothScatter(
  prot$mRNA.stabilities,
prot$conservation.phylo,
  xlab = 'mRNA.stabilities (mins)',
  ylab = 'Gene conservation  (phyloP)')
# Shows a positive correlation, lets see if it looks better with log10 function
smoothScatter(
  log10(prot$mRNA.stabilities),
log10(prot$conservation.phylo),
  xlab = 'mRNA.stabilities (mins(log10))',
  ylab = 'Gene conservation  (phyloP(log))')
# I prefer the graph without the log10
```

```
# Turn off pdf
dev.off()
# Revert to original graph
smoothScatter(
  prot$mRNA.stabilities,
prot$conservation.phylo,
  xlab = 'mRNA.stabilities',
  ylab = 'phyloP')
# See if it is significant
cor.test(prot$mRNA.stabilities,prot$conservation.phylo)
# p-value < 2.2e-16, cor = 0.294545
# Shows that there is a significant positive correlation
```