WEEK 1 ASSIGNMENT REPORT.

STUDENT NAME/ID: Gabriel Wee Kiat Lim/ 180703634

Dataset 1: Marine microbial diversity

1. How does microbial diversity change with latitude?

The distribution of the Microbial diversity index, UniFracInd, was first examined. A Normal Q-Q plot was also generated to examine the distribution visually (Figure: 1-1)

A Shapiro Test was done to test whether the distribution of the Microbial diversity index, UniFracInd is normal. The p-value of 0.866 demonstrates that the distribution is not significantly different from a normal distribution. [1]

The mean UniFracInd in temperate waters is 0.97 and in tropical waters is -0.26. (Figure 1-2). There is more microbial diversity change in temperate waters than in tropical waters. A two-sample t-test has shown this difference to be significant (t(38,1)= 2.839, p-value=0.007249) [1]

In the diagnostic plots, in the normal Q-Q plot, the residuals follow a straight line. (Figure 1-3).

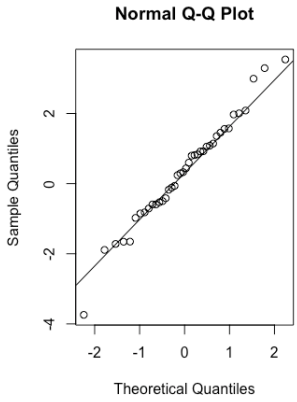

Figure 1-1: Normal Q-Q plot of UniFracInd

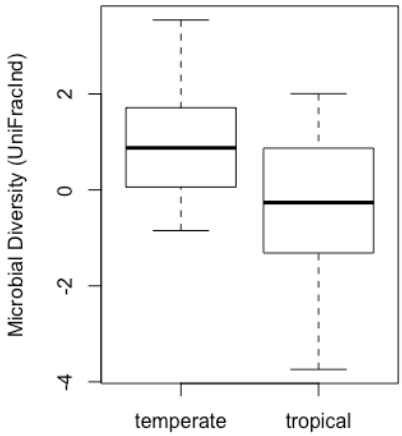

Figure 1-2: Boxplot comparing microbial diversity in temperate and tropical regions.

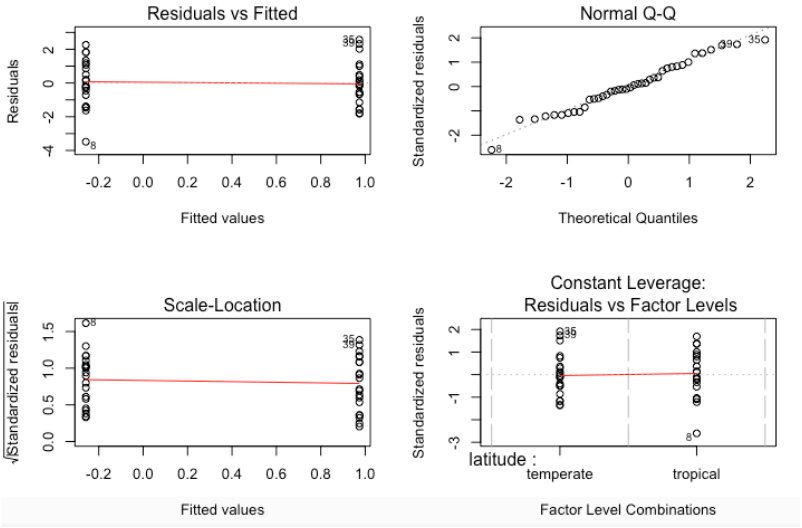

Figure 1-3: Diagnostic plot of UniFracInd modelled as a function of latitude

2. How does microbial diversity change with time of year?

The normal Q-Q plot in the diagnostic plot, shows a slight S-curve to the plot. A Kruskal-Wallis test was carried out as it does not assume a normal distribution (Figure 1-4) [2]

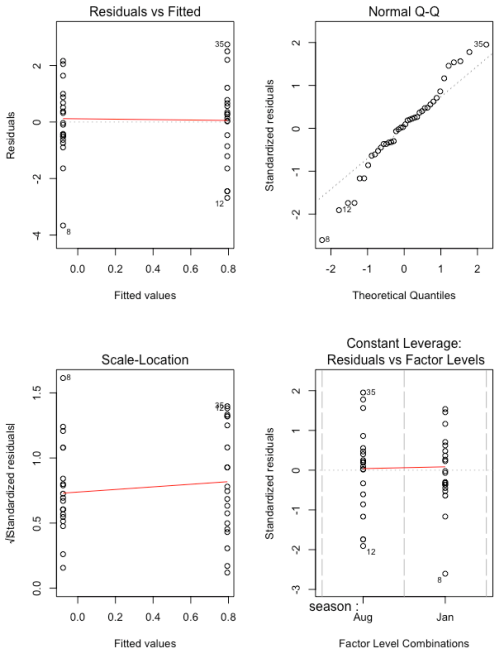

Figure 1-4 : Diagnostic plot of UniFracInd modelled as a function of season

The median UniFracInd in August is 0.99 and in January is -0.14. A Kruskal-Wallis test (p-value = 0.05146) has shown that there was no significant difference between the median microbial diversity UniFracInd in August and January. (Figure 1-5)

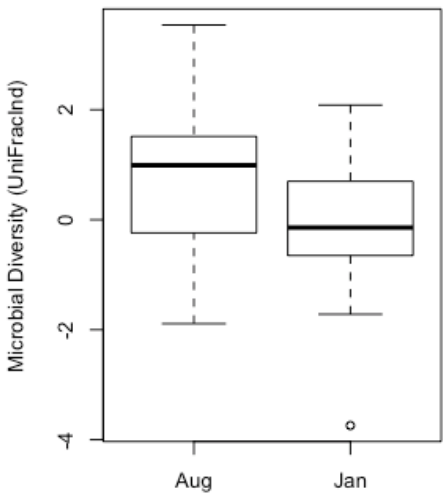

Figure 1-6: Boxplot comparing microbial Diversity in August and January.

3. Is there an interaction between the season, and location?

To investigate to see if there any interactions between the factors season and latitude, a two-way ANOVA was carried out. An interaction plot showed no crossing over of lines, therefore appear to have no interaction between the factors season and latitude. (Figure 1-7)

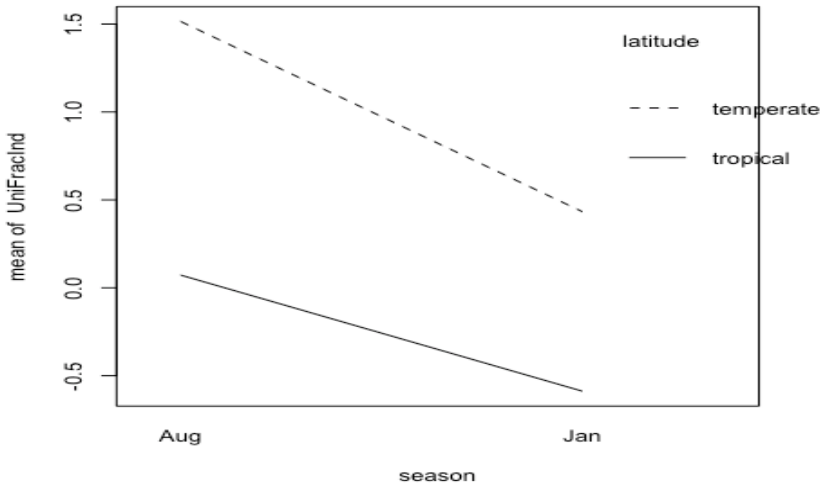

Figure 1-7: An interaction plot with UniFracInd and factors season and latitude

A two-way ANOVA also showed that there was no significant interaction between season and latitude on microbial diversity [Pr(>F)= 0.61902] (Figure 1-8)

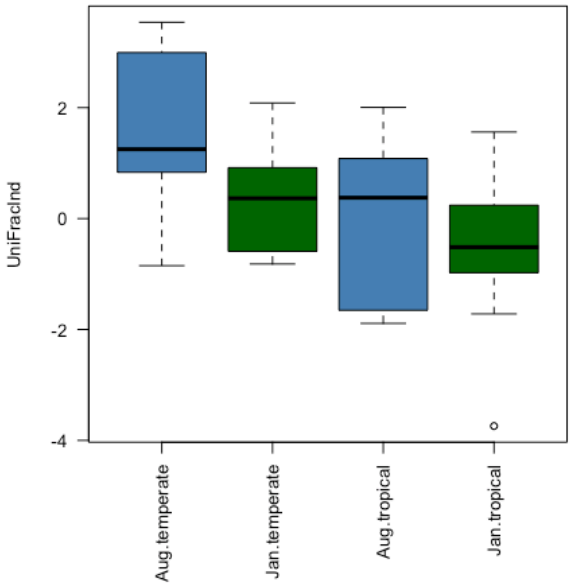

Figure 1-8: Boxplot showing median, range of microbial diversity given the different seasons and latitudes.

Dataset 2: Pairwise nucleotide substitutions and RNA expression levels

1. **How does the putative 'luciferase' homologue expression change with genetic distance (amino acid substitutions)?**

The data expression fold is a measure for the change in the level of expression of a gene in RNA sequencing and microarray [4]

The data shows that expression fold, and therefore luciferase homologue expression, increases significantly with increase in genetic distance (amino acid substituions). [Figure 2-3]

Comment on whether the model assumptions are valid.

Regression assumptions are independence, normally distributed errors, homoscedasticity of and linearity in parameters. [3]

In the model where expression fold is a function of distance, the Residual vs Fitted plot showed homoscedasticity, with no trend in the scatter of the residuals, however the Normal Q-Q plot displayed a slightly S-shape, indicating non-normality. Hence this assumption is likely not to be valid. (Figure 2-1)

Log transformation of the model (*lm(log(expression_fold)~ log(distance)*), produced a relatively straight normal Q-Q plot, but resulted in heteroscedasticity shown in the Residual vs Fitted plot (Figure 2-2) [3]

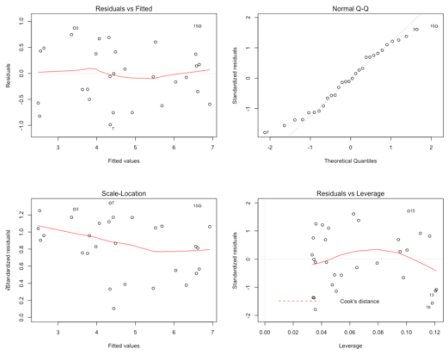

Figure 2-1 : Diagnostics plot of model where expression fold is a function of distance

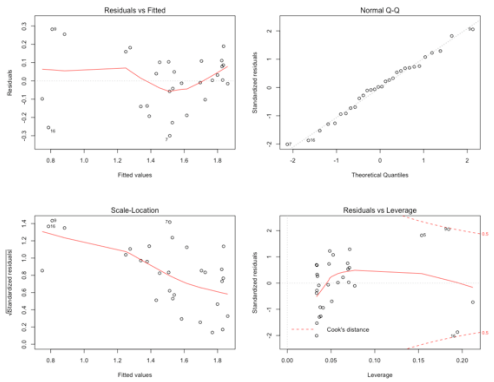

Figure 2-2: Diagnostics plot of model where log(expression fold) is a function of log(distance)

2. **Assuming your model is *statistically* valid, can you guess what effect is responsible for the relationship you've found?**

Regression analysis has shown that luciferase homologue expression increases significantly with genetic distance (DF = 28, p-value= 1.08e-13, in model where expression fold is a function of distance). In model where log(expression fold) is a function of log(distance)(DF=28, p-value= 3.48e-12). (Figure 2-3)

The line of best fit is given by the equation which demonstrates the relationship:

Expression_fold = 2.043+ (0.977 x distance)

Evolutionary difference increases down the phylogenetic tree. Along with it genetic distance and amino-acid substitutions increases, which will have affected the coding loci and translation in a way that resulted in increased expression of luciferase homologue.

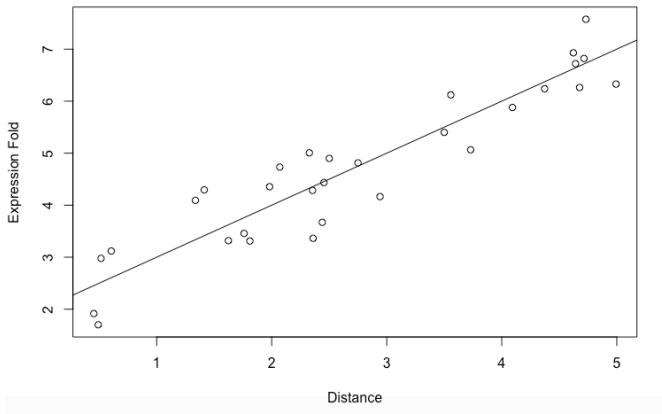

Figure 2-3: Scatterplot of Expression_fold versus genetic distance

Dataset 3: HIV viral load and within-patient population dynamics

Exploratory plots were made for HIV viral load against CD4, tissue parts, score_shannon and score_distance. There was overlap for the boxplots for the tissue part of brain and spinal cord, as there was overlap for boxplots of high and low CD cell counts. (Figure 3-1 and 3-2) Therefore the difference in the median is likely not significant. There does not appear to be a clear pattern for the scatterplot with score_distance (Figure 3-3) There appears to be a relationship for the scatterplot with score_shannon(Figure 3-4)

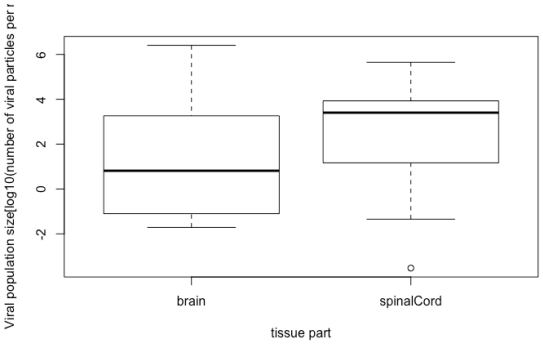

Figure 3-1 Boxplots for tissue parts brain and spinal cord

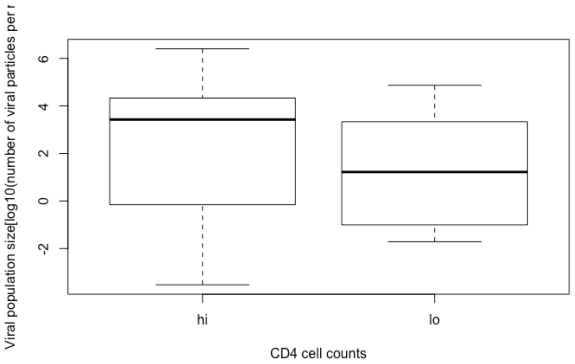

Figure 3-2 Boxplots for hi and lo CD cell counts

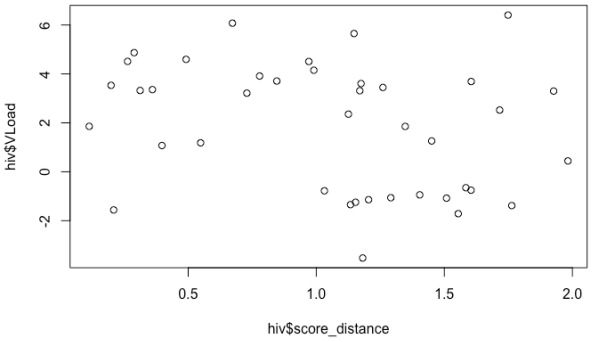

Figure 3-3 Scatterplot for HIV viral load and score_distance

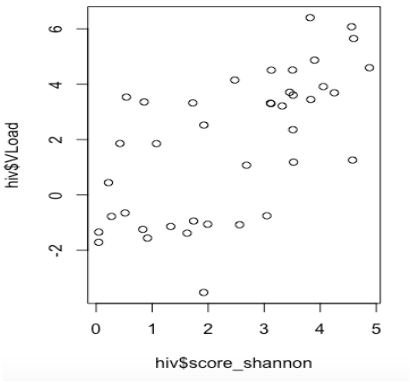

Figure:3.4 : Scatterplot for Viral load and score_shannon

The model (mymod_3) with predictor *score_shannon* was found to the model with the best fit. The adjusted R2 is 0.4049 F(1,38)= 27.54, p-value: 6.114e-06. (Figure 3-5).

As sampling is high-risk and painful , the second best-fit model (forward_model), which factors in score_shannon and score_distance might also be considered.

The relationship between viral load and average Shannon population diversity is shown in Figure 3-4. A Spearman correlation test showed that there was a significant relationship between the two variables (p-value = 1.285e-06). [2]

This shows that drug therapy that extends over long periods of time, like the 40 weeks stated in this case, produces clear difference in HIV sequence that is quantified by Shannon population diversity (score_shannon).[5]

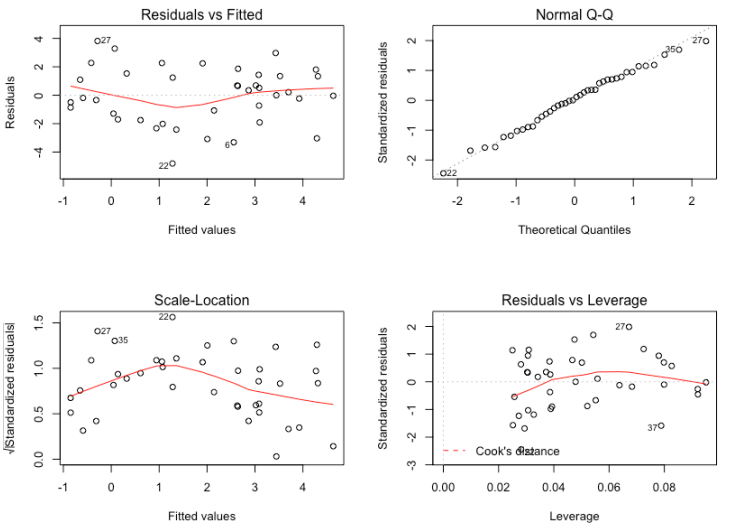

Figure 3-5: Diagnostic plots where viral load is a function of score_shannon

The relationship between viral load and average Shannon population diversity is shown in Figure 3-4. A Spearman correlation test showed that there was a significant relationship between the two variables (p-value = 1.285e-06). [2]

References:

1. Ennos, R & Johnson, M (2018). *Statistical and data handling skills in biology*. 4th ed. London: Pearson.

2. Teetor, P (2011). *R Cookbook*. London: O'Reilly Media.

3. Grafen, A; Hails, R (2002). *Modern statistics for the life sciences*. London: Oxford University Press.

4. Tusher, Virginia Goss; Tibshirani, Robert; Chu, Gilbert *(2001).* Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America. 98 (18): 5116–5121.*

5. Gall A, et al (2013) Restriction of V3 region sequence divergence in the HIV-1 envelope gene during antiretroviral treatment in a cohort of recent seroconverters. Retrovirology. 2013;10:8. doi: 10.1186/1742-4690-10-8.