BIO782P:
Statistics and Bioinformatics
2018/19

Assignment 1

Friha Zafar

Student Number: 180832277

Dataset 1: Marine microbial diversity

The marine microbial diversity in the seawater was investigated at two different latitudes (temperate and tropical) and in two different seasons, summer (August) and winter (January). The microbial diversity was measured using the UniFrac Index at the two latitudes **(Figure 1)**. It can be inferred that the different latitudes have similar microbial diversity as shown in **Figure 1** where temperate and tropical have similar median Unifrac Index values. Furthermore, in **Figure 1** there are large error bars which overlap. The unpaired Student *t*-test was conducted to find any significant difference between two means of the UniFrac Index values for the temperate and tropical latitudes, which gave a *p*-value of 0.7579. This was greater than the *alpha(α)*-value of 0.05 therefore the null hypothesis (that there was no change in microbial diversity with latitude) was accepted and the alternative hypothesis (there was a change in microbial diversity with latitude) was rejected. Thus, latitude does not have a significant effect on the diversity of microbiomes in the seawater.

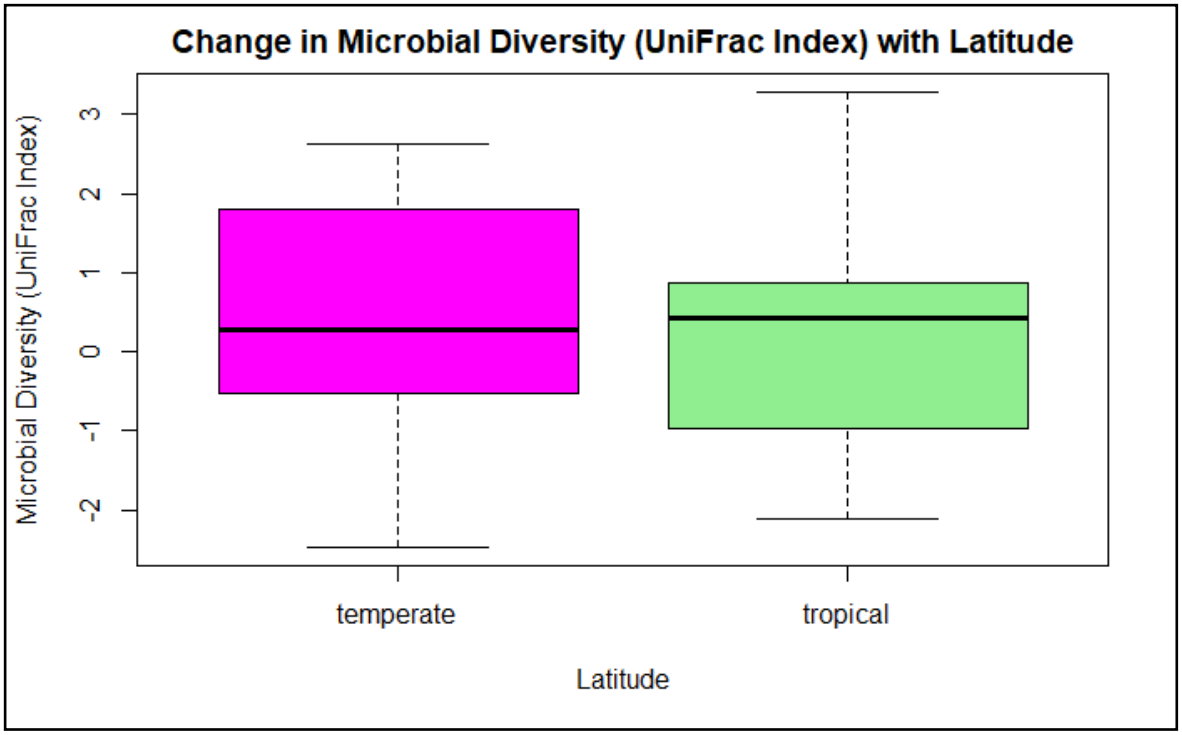

Figure 1. A boxplot illustrating the difference in marine microbial diversity (UniFrac Index) between the two latitudes, temperate(pink) and tropical(green) using the dataset: 'part_1_student_1068'.

Nevertheless, a significant difference was seen in the median UniFrac Index values in **Figure 2** which shows the change in microbial diversity in August and January. The boxplot **(Figure 2)** suggests a greater range in the seawater microbial diversity in August than in January. The unpaired Student *t*-test was conducted on the microbial diversities for both seasons which indicated a significant difference (*p*-value= 0.03043) which may suggest that August has significantly greater diversity than January.

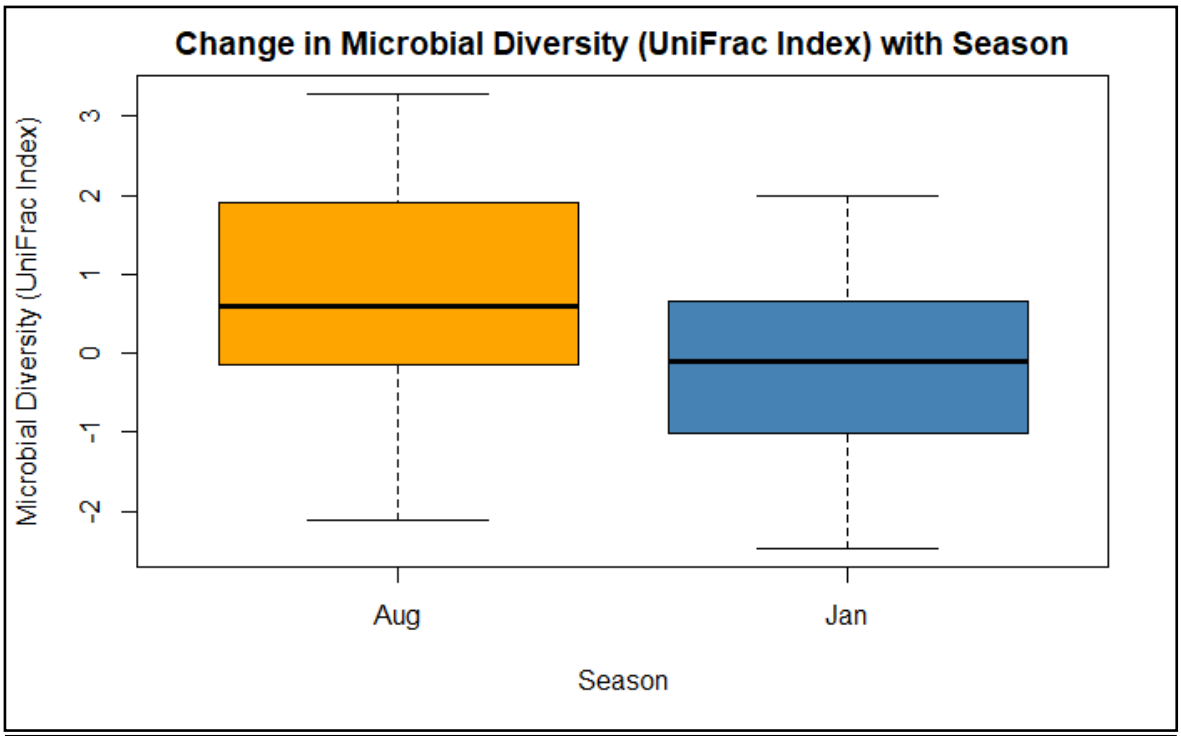

Figure 2. A boxplot illustrating the change in marine microbial diversity (UniFrac Index) in two different seasons, 'August' (orange) and 'January' (blue) using the dataset: 'part_1_student_1068'.

The interaction between the microbial diversity during the different seasons and their locations was investigated. A linear model was made for which the diagnostic plots showed that the residuals were homoscedastic with a normal distribution. An ANOVA was conducted on this model where the variable 'Season' was independently statistically significant (at α=0.05). According to Ladau et al. (2013) microbial diversity is low in the summer in temperate regions, while in tropical regions they are high. This is seen in **Figure 3** where the samples taken in August-tropical had a greater median and range than in August-temperate. However, the model showed no significant relationship between season and latitude. Hence, no significant interaction was detected between the microbial diversity in different latitudes and seasons.

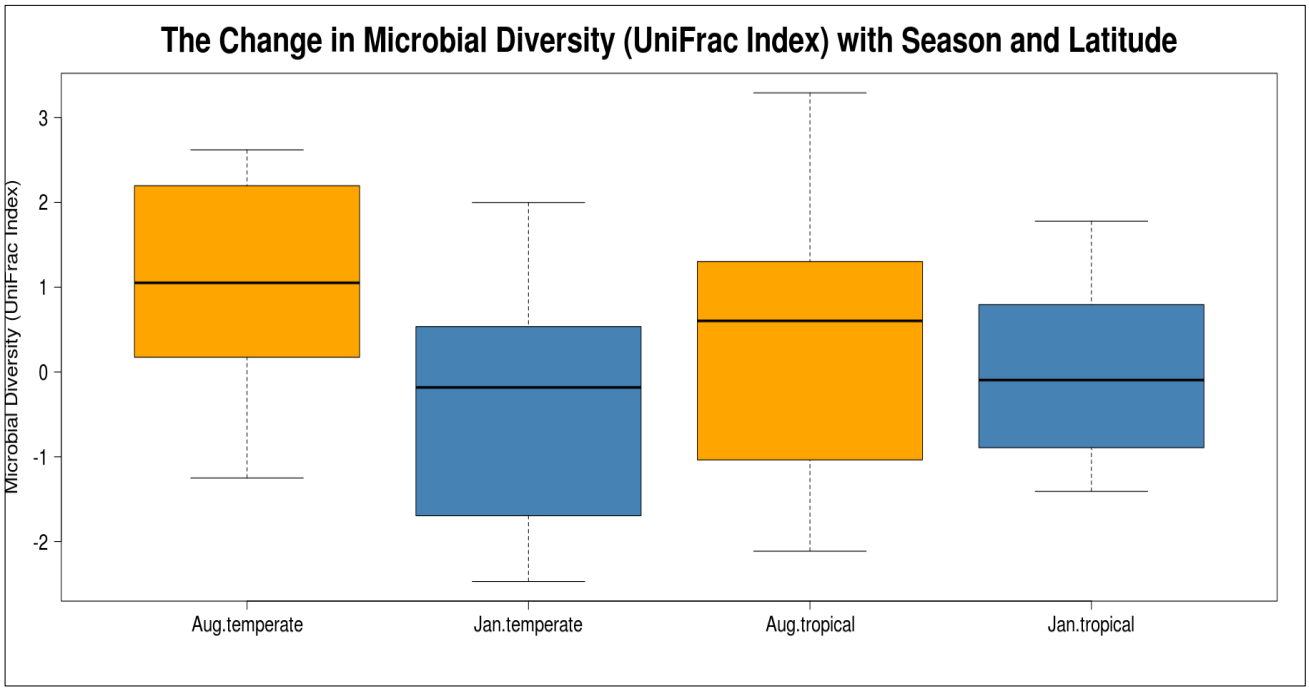

Figure 3. The change in microbial diversity (Unifrac Index) in summer (August in orange) and winter (January in blue) at different latitudes (temperate and tropical) using the dataset: 'part_1_student_1068'.

Dataset 2: Pairwise nucleotide substitutions and RNA expression levels

Putative luciferase genes from the *Brassicaceae* family were investigated for their gene expressions and their pairwise genetic distances with the genome of *Arabidopsis thaliana* which is a closely-related species. A very weak negative correlation was found (Pearson's *r*=-0.1077147) **(Figure 4)** between the putative 'luciferase' homologue expression change and the genetic distance (amino acid substitutions) due to the large scatter. Hence no significant relationship could be established **(Figure 4)**. Thus, the putative 'luciferase' homologue expression does not change with genetic distance.

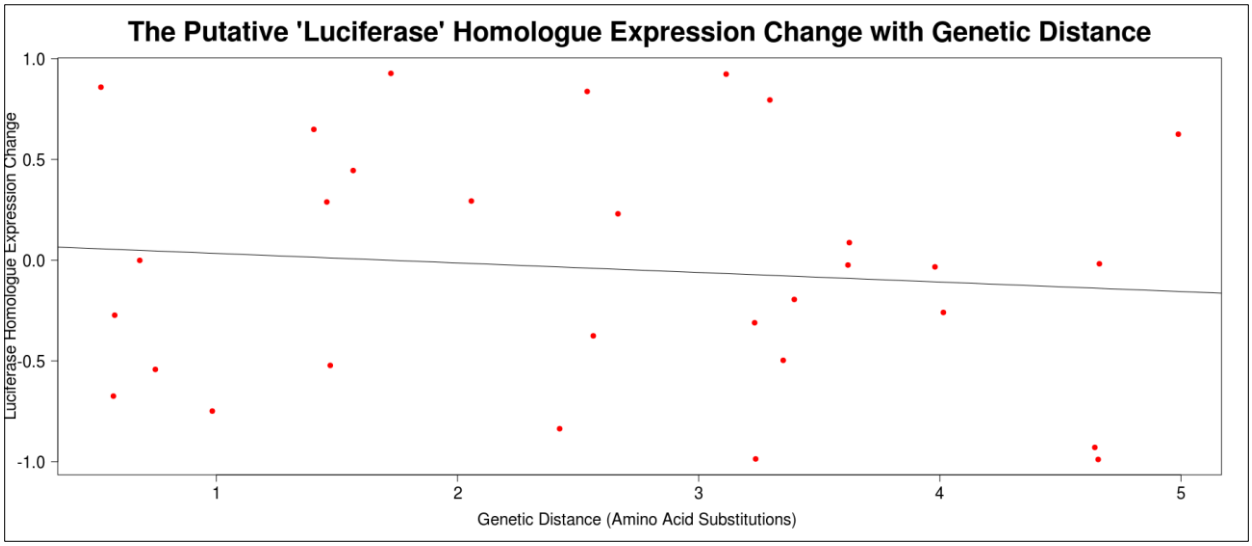

Figure 4. A linear regression plot showing the effect of the 'putative' luciferase homologue expression change on their genetic distance measured by amino acid substitutions. The dataset used: 'part_2_student_1068'.

The linear model's assumptions were found to be valid as the diagnostic plots indicated that the model assumptions were met since a linear pattern was seen, where the residuals were normally distributed and showed homoscedasticity (Kim, 2015). Furthermore, no extreme residuals were found therefore no outliers influenced the regression line (Kim, 2015). Thus, the model assumptions were met, and the model is valid.

The weak negative correlation **(Figure 4)** may suggest that as the genetic distance increases between the putative 'luciferase' homologues, the expression slightly decreases however the effect was found to be insignificant. This suggests that the homologues located in similar positions in the genome, may have similar expression changes as they may share the same transcription factors, thus expressed at similar times (Chen, de Meaux & Lercher, 2010). Yet there are some that have negative expression changes **(Figure 4)** meaning those 'putative' homologues with greater

genetic distances do not have similar expression changes and may not be in the same genomic locations as the others (Chen, de Meaux & Lercher, 2010).

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

A Human Immunodeficiency Virus (HIV)-positive patient was investigated for HIV viral load in brain and spinal cord for 40 weeks. The average Shannon population diversity and mean pairwise genetic distance was measured by comparing the *env* gene of the samples against the reference gene.

Automated and manual linear models were fit against the data, the best model was significant at an α-value of 0.05 after running ANOVA comparing all models. This model showed that CD4+ cell count and the tissue types had significant differences in HIV viral load. The Student *t*-test conducted for the CD4+ cell count gave a *t*-score of -3.4244 and a *p*-value of 0.001821, therefore a significant difference in the viral load between low and high CD4+ cell counts was established. Low viral load was seen when CD4+ cell count was high, possibly due to drug therapies such as the highly active antiretroviral therapy (HAART) (Autran, 1997) **(Figure 5)**. However, there was an anomaly which could suggest the opposite, since the value was outside the upper quartile **(Figure 5)**. High HIV viral load was seen when the cell counts were low possibly due to reduced effects of the therapy causing CD4+ depletion (Autran, 1997).

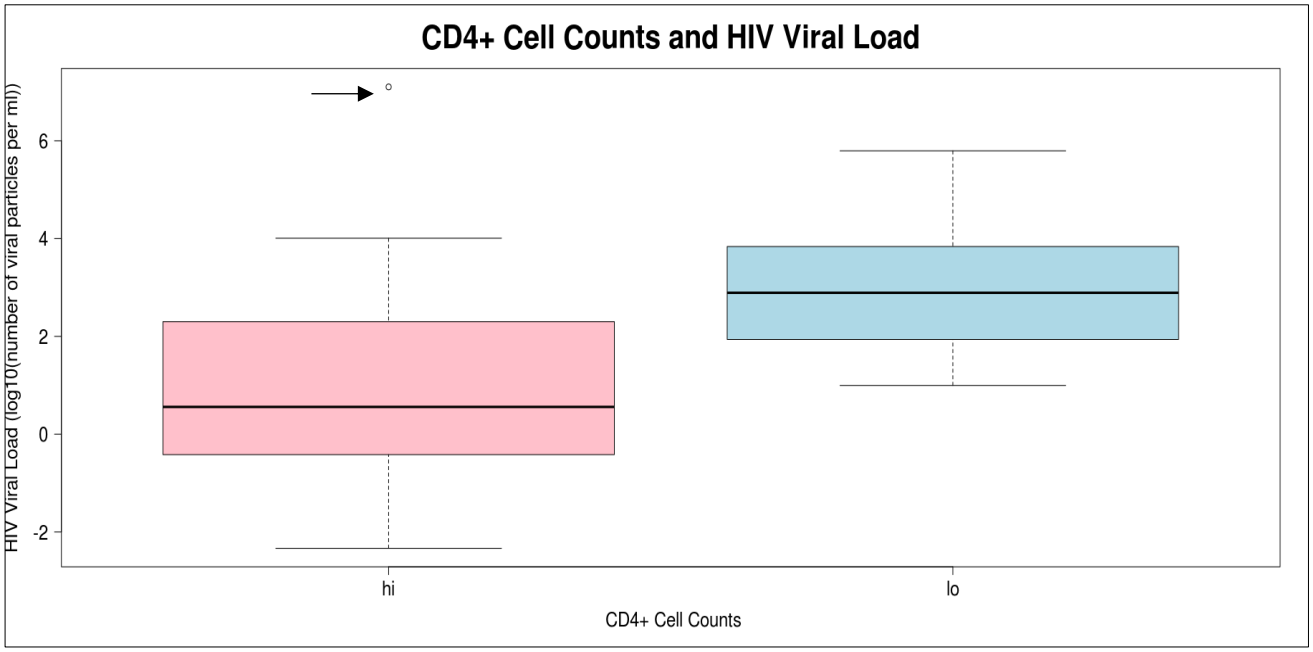

Figure 5. A boxplot showing the patient's HIV viral load at log10(number of viral particles per ml) and the CD4+ cell counts from the flow cytometry which is shown as high(hi) and low(lo). The anomaly is indicated by the arrow (→). The dataset used: 'part_3_student_1068'.

A significant difference was seen in HIV viral load when comparing the tissue types, brain and spinal cord, where spinal cord had greater HIV viral load than the brain **(Figure 6)**. The Student *t*-test gave a *t*-score of -2.7823 and a *p*-value of 0.008363 thus there was a significant difference in the viral load between the tissue types. HIV relies on monocytes that differentiate into macrophages upon entering the central nervous system (CNS) in order to infect the brain (Ivey, MacLean & Lackner, 2009), this delay may have caused the reduced viral load in the brain.

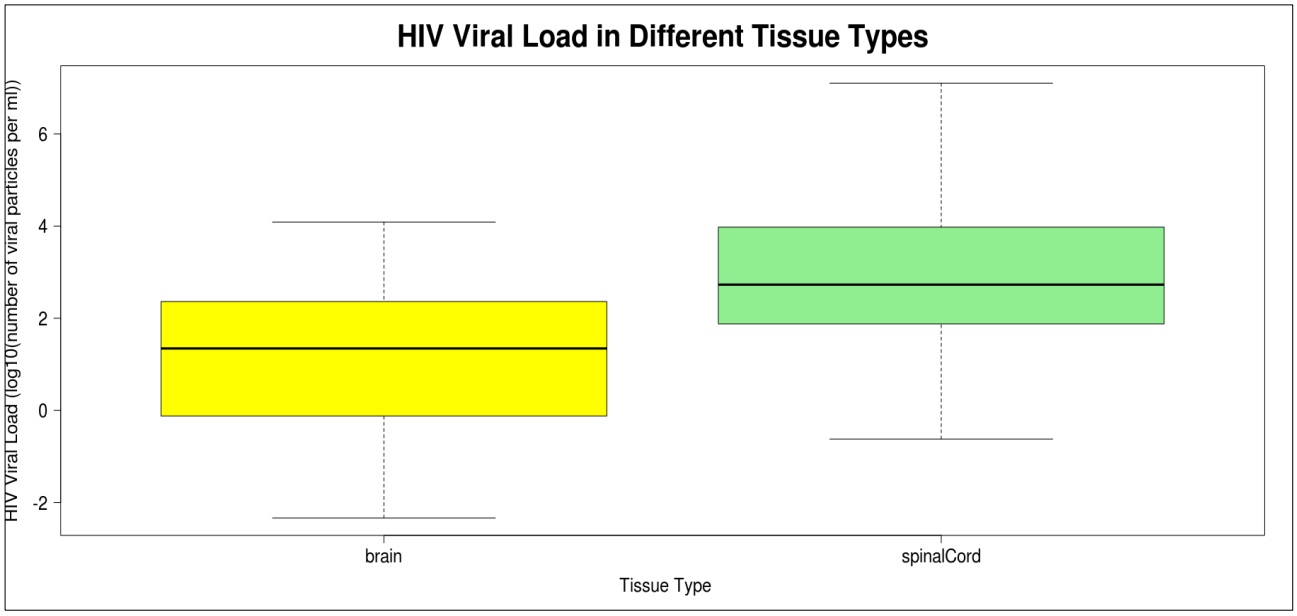

Figure 6. A boxplot showing the patient's HIV viral load at log10(number of viral particles per ml) and the different tissue types (brain (yellow) and spinal cord(green)) from which the samples were taken. The dataset used: 'part_3_student_1068'.

Yet, no significant relationship was observed between viral load and the average Shannon-Weiner diversity score and the mean pairwise genetic distance. Possibly suggesting that the diversity in the *env* gene did not affect the clearance of the virus and therefore did not affect the viral load. Furthermore, a positive correlation (Pearson's *r*= 0.5435939) was seen in the effect of time on viral load **(Figure 7)**, possibly because HIV causes the patient's immune system to be reduced, thus reducing the clearance of the virus. However, the relationship was not significant since the *r*-value was too low. High diversity in the *env* gene may not aid in understanding drug resistance, there may be a bias in the amplification the *env* gene therefore not all variations may be detected (Maldarelli et al., 2013). The diversity of the HIV genes *pro-pol* may aid in future studies to understand the effect of diversity on the increase in viral load (Maldarelli et al., 2013).

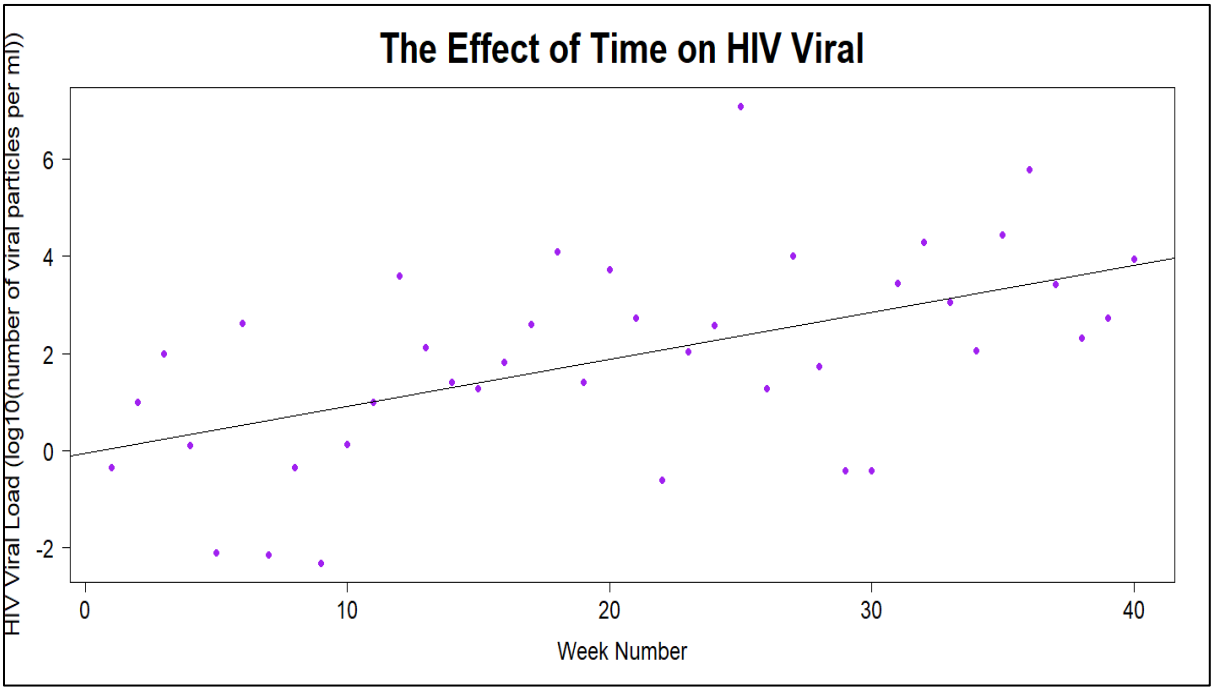

Figure 7. A linear regression plot showing the patient's HIV viral load at log10(number of viral particles per ml) and the number of week when the samples were taken. The dataset used: 'part_3_student_1068edited'.

References

Autran, B., Carcelain, G., Li, T., Blanc, C., Mathez, D., & Tubiana, R. et al. (1997). *Positive Effects of Combined Antiretroviral Therapy on CD4+T Cell Homeostasis and Function in Advanced HIV Disease*. *Science, 277*(5322), 112-116. doi: 10.1126/science.277.5322.112

Chen, W., de Meaux, J., & Lercher, M. (2010). *Co-expression of neighbouring genes in Arabidopsis: separating chromatin effects from direct interactions*. BMC Genomics, 11(1), 178. doi: 10.1186/1471-2164-11-178

Ivey, N., MacLean, A., & Lackner, A. (2009*). Acquired immunodeficiency syndrome and the blood-brain barrier. Journal Of Neurovirology, 15*(2), 111-122. doi: 10.1080/13550280902769764

Ladau, J., Sharpton, T., Finucane, M., Jospin, G., Kembel, S., & O'Dwyer, J. et al. (2013). Erratum: Global marine bacterial diversity peaks at high latitudes in winter. *The ISME Journal*, *7*(9), 1876-1876. doi: 10.1038/ismej.2013.76

Kim, B. (2015). *Understanding Diagnostic Plots for Linear Regression Analysis | University of Virginia Library Research Data Services + Sciences.* Retrieved from https://data.library.virginia.edu/diagnostic-plots/ Accessed, November, 24, 2018.

Maldarelli, F., Kearney, M., Palmer, S., Stephens, R., Mican, J., & Polis, M. et al. (2013). HIV Populations Are Large and Accumulate High Genetic Diversity in a Nonlinear Fashion. *Journal Of Virology*, *87*(18), 10313-10323. doi: 10.1128/jvi.01225-12