Dataset 1

1) How does microbial diversity change with latitude?

Figure 1 presents the boxplot of microbial diversity samples collected in tropical and temperate latitudes. It can be observed that tropical latitude samples have greater diversity than temperate samples, including median value of -0.081 for tropical and -1.522 for temperate compared to the reference sample. Shapiro-Wilk test was performed to determine normality of the data and both datasets are normally distributed (tropical p-value = 0.2311, temperate p-value = 0.3583). Welch two sample t-test was performed (t = -2.2743, df= 34.928, p-value = 0.0292) determining that there is a significant difference in means between samples. Hence, there is a greater microbial diversity in the tropical latitude than in temperate.

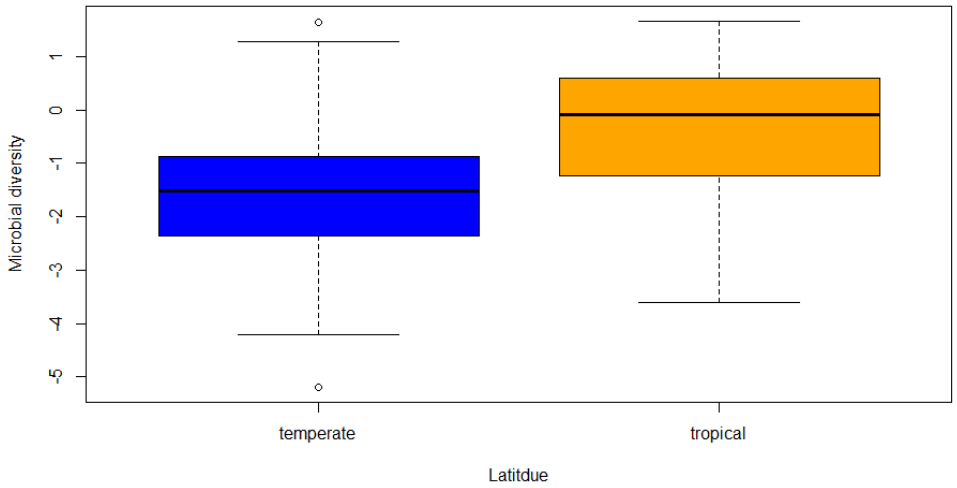

Fig 1. Diagram presenting the effect of latitude on the microbial diversity in the samples.

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

Figure 2 presents the boxplot of microbial diversity in samples collected in August and January. It can be observed that samples collected in August have greater range of microbial diversity compared to ones collected in January. January samples have greater median (-1.1989 vs -0.6160). Shapiro-Wilk test was performed to assess the normality of data (January p-value = 0.3489, August p-value = 0.6797) and August data is slightly insignificant, thus does not follow the normal distribution. However, Welch two sample t-test is applicable (t= -1.4289, df= 37.273, p-value = 0.1614) determining that means in two samples are not equal, thus significantly different. Hence, microbial diversity is greater in January than in August compared to the reference sample.

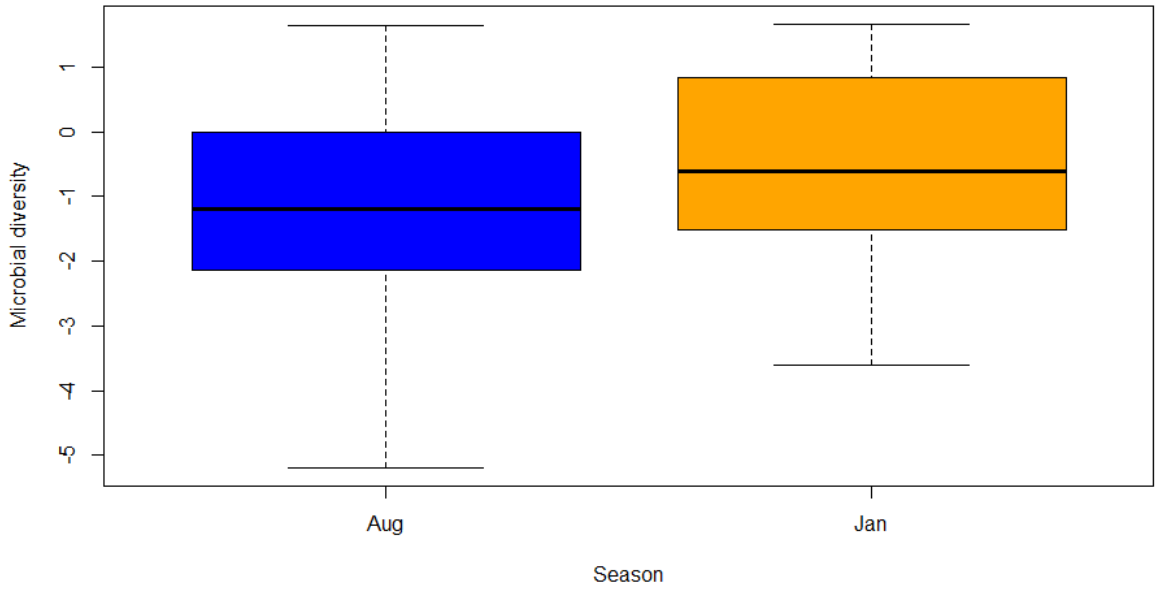

Fig 2. Figure presenting the effect of the time of sample collection (August and January) on microbial diversity in the samples.

3) Is there an interaction between the season and location?

Figure 3 presents an interaction plot between a season and location of the collected samples, which shows that in tropical latitude mean microbial diversity content is higher for both seasons compared to temperate latitude. Hence, tropical location of sample collection provides greater microbial diversity for both seasons. Figure 4 presents interaction between location and season, and it can be observed that samples collected in August in temperate location have significantly lower mean diversity compared to tropical. Moreover, approximately doubled diversity increase can be observed for August samples when collected in temperate latitude and tropical. This leads to a conclusion, that time of the sample collection does not affect the microbial diversity as significantly in tropical latitude, compared to the temperate latitude where the mean difference between the month more than double.

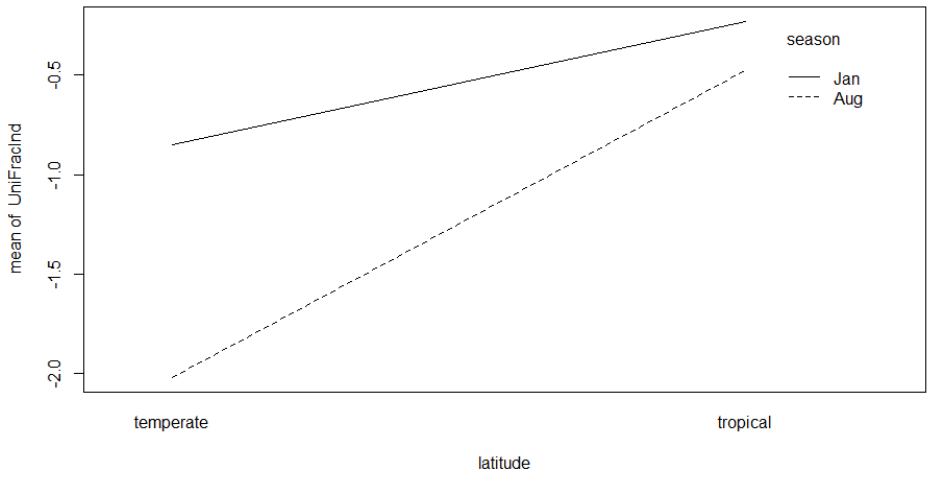

Fig. 3. Figure presenting an interaction plot between season and location (latitude) compared against reference microbial diversity content.

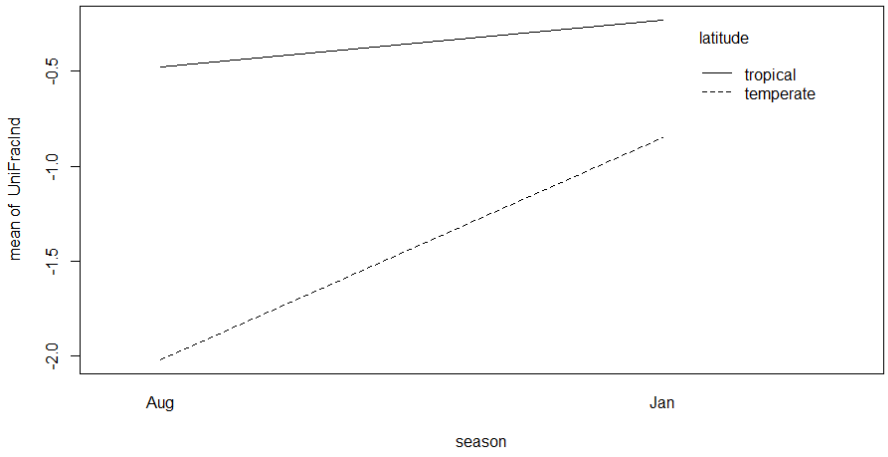

Fig. 4 Figure presenting an interaction plot between location (latitude) and season compared against reference microbial diversity content.

DATASET 2:

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

Figure 5 presents a relationship between homologue expression and genetic distance and it can be observed that it is a negative correlation – the greater the genetic distance is, the lower homologue expression is measured.

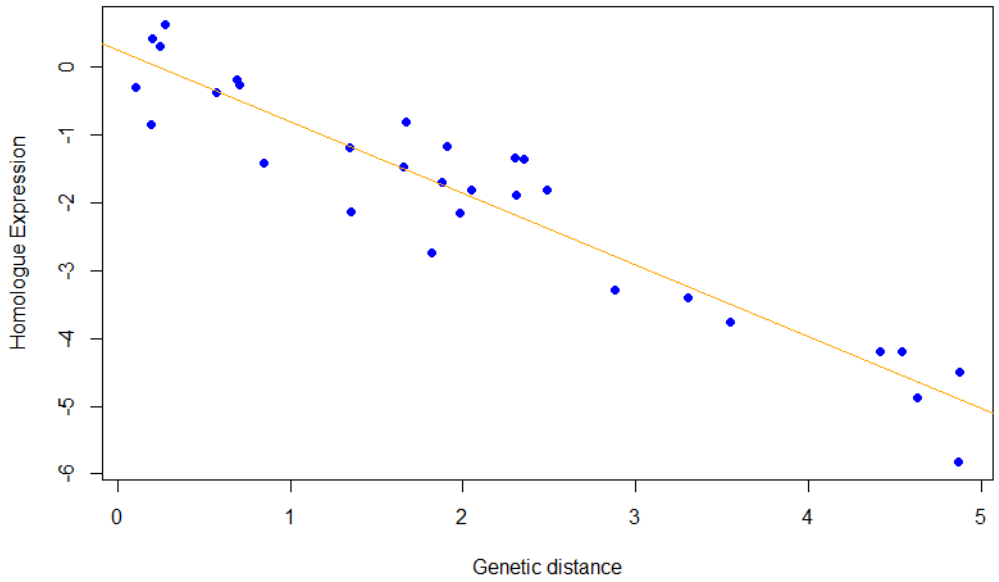

Fig 5. Scatter plot presenting a relationship between putative luciferase homologue expression change and genetic distance.

2). Comment on whether the model assumptions are valid.
Chosen model presented in Fig 5 is based on the relationship of homologue expression and genetic distance (Expression ~ distance). Collected data prove that assumptions are valid as the negative correlation can be observed and the ANOVA analysis determined that data is significantly different (df=1, F-value = 228.84, p-value = 5.281e-15).

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

Discovered negative relationship between homologue expression and evolutionary genetic distance is based on the accumulation of amino acid substitutions and their negative effect on expression. It has been proven that "the intensity of gene expression relates inversely to the rate of protein sequence evolution on a genomic scale. Additionally, the most highly expressed genes show the lost total number substitutions per polypeptide." (Kumar S., Subramanian S., 2004)

DATASET 3

Prior to developing a model which would explain viral load in terms of the other variables, a relation between viral load and CD4+ cell count, tissues (brain and spinal cord), Shannon population diversity and genomic distance was investigated and presented in the Fig.6. Additionally, ANOVA analysis was performed on a model aggregating all of the variables which indicated the (VLoad ~ score_shannon) variation between VLoad and Shannon diversity is significant (df=1, F-value = 13.8301, p-value 0.001), showing a potential strong relation between each other.

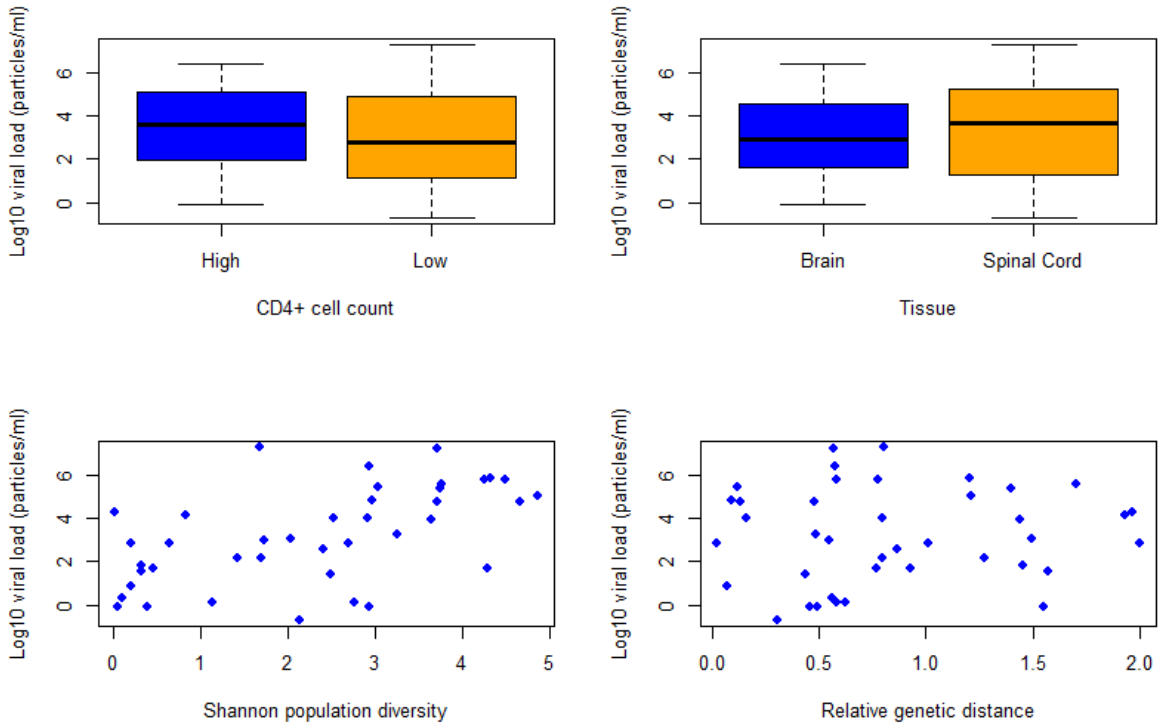

Fig 6. Figure presenting relation between viral load count (log10) and each variable, such as CD4+ cell count, tissue (brain, spinal cord), Shannon population diversity and relative genetic distance.

Automated model choice was applied utilising forward and backward stepwise regression with lowest AIC score determining the best descriptor of the data.

Three models provided similar result and are presented in the Table 1.

Table 1. Models with the lowest AIC score.

Model	AIC score
VLoad~ score_shannon + score_distance	**49.481**
VLoad~ score_shannon + score_distance + tissue	51.217
VLoad ~ score_shannon	51.302

VLoad= viral load, score_shannon = Shannon population diversity, score_distance = Relative genetic distance.

Model with lowest AIC score (VLoad~score_shannon +score_distance) was chosen since it is the simplest and allows to combine two factors such as Shannon population diversity and relative genetic distance to explain behaviour of the data. ANOVA analysis of the model determined significant difference in data variance between viral load and Shannon score (df=1, F-value = 18.6926, p-value = 0.00011) and variance in score distance as slightly insignificant (df=1, F-value = 3.7086, p-value =

0.061). Even though, there is a slight insignificance of the relative genetic distance, the AIC difference values between final model and a model utilising only Shannon diversity are minor (<2) and a model combining two factors may be more robust in data explanation.

References:

1. Kumar, S. and Subramanian, S. (2004). Gene Expression Intensity Shapes Evolutionary Rates of the Proteins Encoded by the Vertebrate Genome. Genetics, 168(1), pp.373-381.

CODE

DATASET 1

```
#Dataset 1
marine = read.delim('part_1_student_1063.tdf', header = T)
str(marine)
attach(marine)
summary(marine)
#boxplot diversity vs latitude
div_lat = boxplot(UniFracInd ~ latitude, xlab='Latitdue',
      ylab='Microbial diversity', col=c('blue','orange'))
div_lat$stats        # distribution stats
#boxplot diversity vs season
div_seas = boxplot(UniFracInd ~ season, xlab='Season',
      ylab='Microbial diversity', col=c('blue','orange'))
div_seas$stats       #distribution stats
#shapiro-wilk test to check if data is normally distributed
norm_tropical = shapiro.test(UniFracInd[latitude=="tropical"])
norm_temp = shapiro.test(UniFracInd[latitude=="temperate"])
norm_jan = shapiro.test(UniFracInd[season=="Jan"])
norm_aug = shapiro.test(UniFracInd[season=="Aug"])
#T-test to assess the data
t_latitude = t.test(UniFracInd ~ latitude)
t_season = t.test (UniFracInd ~ season)
# Q3
#Interaction model between season and latitdue
model_lat_seas = lm(UniFracInd ~ season * latitude)
par(mfrow=c(2,2))
plot(model_lat_seas)

#Assesing quality
drop1(model_lat_seas, test = 'F')
```

```
#Interaction plots
par(mfrow=c(2,1))
interaction.plot(season,latitude,UniFracInd)
interaction.plot(latitude,season,UniFracInd)
################################################################################
```

DATASET 2

```
#DATASET 2

RNA = read.delim('part_2_student_1063.tdf', header = TRUE)
str(RNA)
attach(RNA)
# Relationship between expression and distance
plot(expression_fold~distance, xlab='Genetic distance', ylab='RNA expression', pch=19, col='blue')
#Creating a model
rna_model = lm(expression_fold~distance)
#ANOVA to determine signifficant difference in variation
anova(rna_model)
par(mfrow=c(2,2))
#diagnostic plot
plot(rna_model)
par(mfrow=(c(1,1)))
# how does expression change with distance
plot(distance, expression_fold, xlab= "Genetic distance", ylab="Homologue Expression", pch=19
, col='blue')
abline(rna_model, col='orange')
```

DATASET 3

```
#Reading the data + checking the data structre

HIV = read.delim('part_3_student_1063.tdf', header=TRUE)

str(HIV)

summary(HIV)

attach(HIV)

par(mfrow=c(2,2))

#Plots to represent the relationship between viral load and cell count,tissue,diversity, and evolutionary distance

plot(VLoad ~ CD4, xlab = 'CD4+ cell count', ylab='Log10 viral load (particles/ml)', names=c('High','Low'),

    col=c('blue','orange'))

plot(VLoad ~ tissue, xlab = 'Tissue', ylab='Log10 viral load (particles/ml)', names=c('Brain','Spinal Cord'),

    col=c('blue','orange'))

plot(VLoad ~ score_shannon, xlab='Shannon population diversity',ylab='Log10 viral load (particles/ml)',

   pch=19, col='blue')

plot(VLoad ~ score_distance, xlab='Relative genetic distance', ylab='Log10 viral load (particles/ml)',

    pch=19, col='blue')

par(mfrow=c(1,1))

#Automated model selection

hiv_model= lm(VLoad ~ CD4 * tissue * score_shannon * score_distance)

anova(hiv_model)    # data quality assesment

# Automated stepwise regression forward/backward, both

backwards = step(hiv_model,direction ='backward')

forwards = step(lm(VLoad ~ 1), scope=(~ score_shannon* CD4 * tissue * score_distance),

          direction = 'forward')

both = step(lm(VLoad ~ score_shannon + score_distance), scope = c(lower=~score_shannon, upper=~score_shannon*score_distance*tissue))

#Final model determined by the lowest AIC score

final_model = lm(VLoad ~ score_shannon + score_distance)

anova(final_model)
```