Statistics assignment

Richard Burns

Question 1; Marine microbial diversity

Initial descriptive statistics suggest that there was no significant change in microbial diversity based on the latitude at which a marine sample is collected, although diversity had the potential to be lower in samples collected in a tropical setting (see fig 1). Samples collected In January had on average a higher diversity than those collected in August, although data collected in January also had a much higher spread of diversity scores (See fig 2).

Figure 1

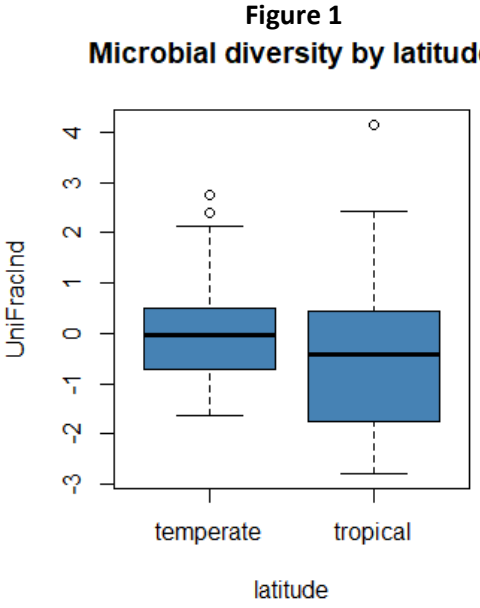
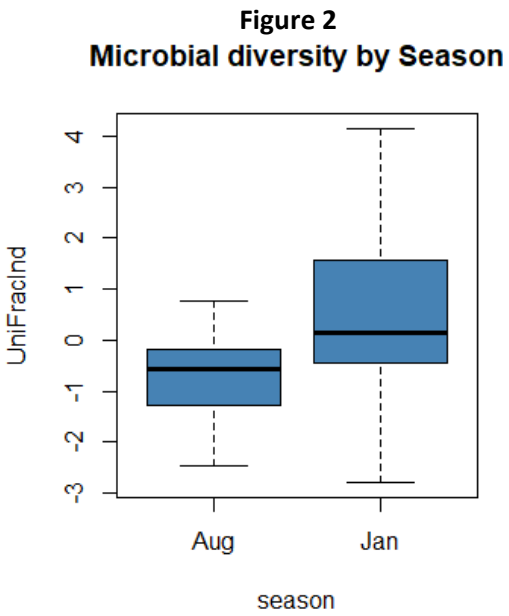


Figure 2



Paired *t*-tests were used to compare the differences in the microbial diversity of the marine samples based on season of collection and the latitude at which the sample was collected (n=20 for each group). Results suggested that season had a significant effect on sample diversity (p=0.01604,df=19), however latitude had none (p=0.31130,df=19). A two-factor ANOVA was also used to test for interaction between both latitude and season on sample diversity, however again there was no apparent interaction (p=0.846, 1d.f.).

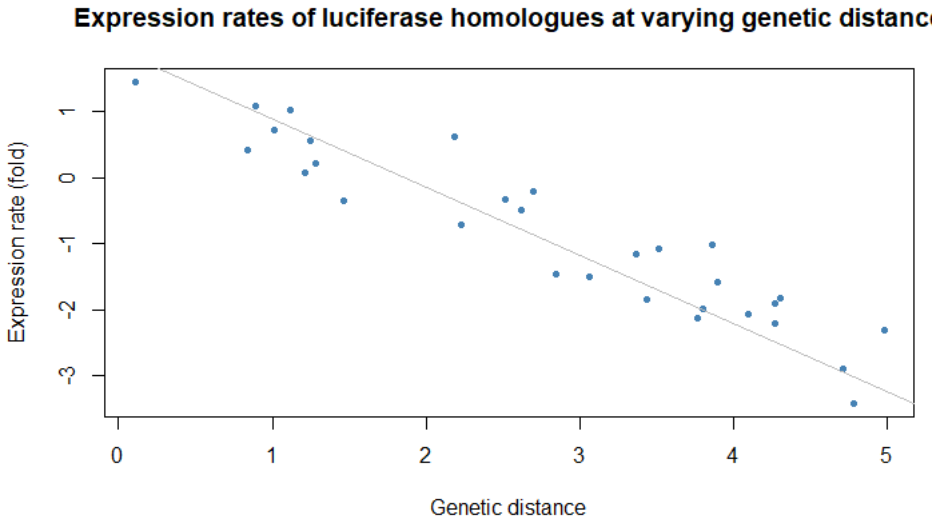
Question 2; Luciferase expression with genetic distance

Dataset two suggests that there is a strong negative correlation between the level of luciferase gene homologue expression in *Brassicaceae sp.* and the pairwise genetic distance between the homologue in question and a putative luciferase gene in *Arabidopsis thaliana* (Figure 3). As the number of amino acid substitutions between a homologue the reference luciferase gene increases, the expression level of that homologue decreases. A linear regression model can be fitted to the data with valid model assumptions (normally distributed errors, independent, linear and homoscedastic data).

Assuming the statistical validity of the model; the observed relationship could be due to genetic divergence of the luciferase gene amongst *Brassicaceae sp.*, in which populations of a species become separated by reproductive barriers and consequently accrue individual genetic mutations, in conjunction with the speciation of the *Brassicaceae* family (Palumbi, 1994) (Wu et al., 2012). We would expect the expression rate of a given homologue to decrease as the genetic distance

increases, as it becomes decreasingly likely to observe an increasingly diverged gene if selecting at random.

Figure 3: A linear regression model of the relationship between genetic distance and the expression rate of a homologue



References

Palumbi, S. (1994). Genetic Divergence, Reproductive Isolation, and Marine Speciation. *Annual Review of Ecology and Systematics*, 25(1), pp.547-572.

Wu, X., Northcott, P., Dubuc, A., Dupuy, A., Shih, D., Witt, H., Croul, S., Bouffet, E., Fults, D., Eberhart, C., Garzia, L., Van Meter, T., Zagzag, D., Jabado, N., Schwartzentruber, J., Majewski, J., Scheetz, T., Pfister, S., Korshunov, A., Li, X., Scherer, S., Cho, Y., Akagi, K., MacDonald, T., Koster, J., McCabe, M., Sarver, A., Collins, V., Weiss, W., Largaespada, D., Collier, L. and Taylor, M. (2012). Clonal selection drives genetic divergence of metastatic medulloblastoma. *Nature*, 482(7386), pp.529-533.

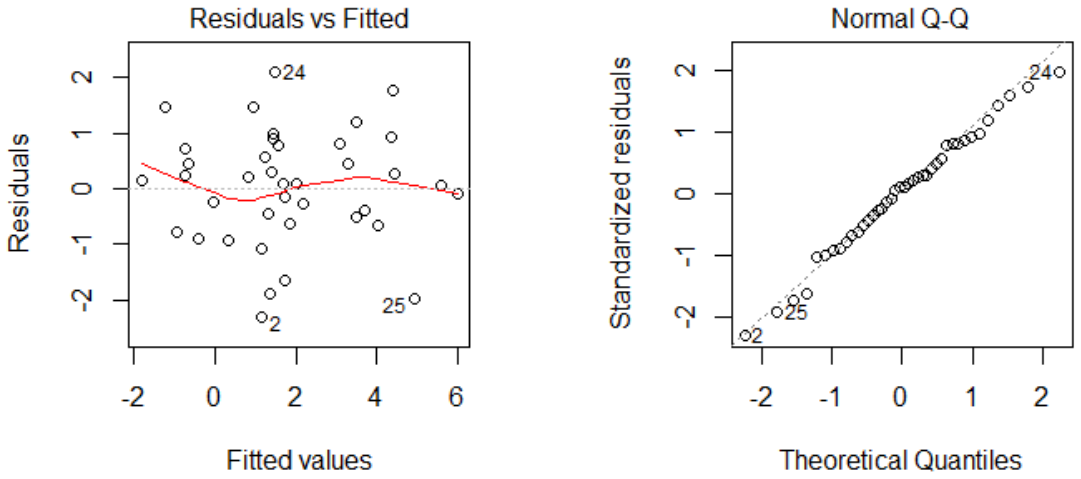
Question 3; HIV load & clinical data

The model selected to best fit the data provided treats the HIV viral load (log10 ppml) of the patient as a linear combination of CD4+ count, sampling site, Shannon population diversity and pairwise genetic distance of individual viruses present, also including interactions between regressors. It can be defined as follows:

Viral Load ~ Distance score + Shannon score + tissue site + CD4 count + Shannon score: CD4 count + Shannon score: tissue site + Distance score:CD4 count

(Multiple R-squared =0.7803, p=6.573e-09 32d.f.). We used a backwards stepwise regression method to optimize a starting linear model containing all of the independent variables. Subsequent steps were evaluated by their AIC scores, any subsequent models with a higher AIC than the previous were rejected, and any with a lower score were accepted as the new 'best' model from which a new stepwise optimization occurred. The best model displayed above had an AIC score of 16.43, the lowest score produced from the automatic model selection method, and summary plots show a strong normal distribution and linearity in the model, as shown in figures 4 and 5 below:

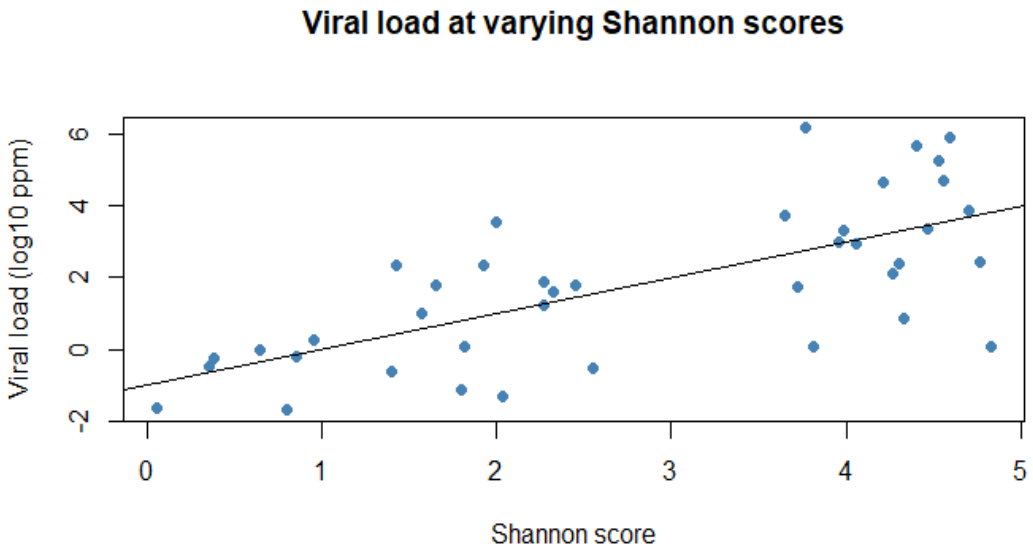
Figure 4 & 5: Summary plots of the 'best' fitted linear model.



In the data collected there was a clear positive correlation between the Shannon scores observed in the sample collected and the viral load of the sample (p=6.888e-07,1d.f.), shown in figure 6, suggesting a strong positive relationship between the diversity present in HIV virus within a patients CNS and their viral load. There was seemingly no direct relationship between a patient's viral load and the genetic distance of a virus from the reference HIV viral genome (p=0.2872,1d.f.). Paired t-tests for viral load (normally distributed as proved by Shapiro Wilk p=0.2576) and both CD4+ count and site of sample showed no apparent direct relationship between the two in either case (p=0.1787, 19d.f.) and (p=0.7748, 19d.f.) respectively, accepting the null hypothesis in both cases that there was no real difference between their means.

An anova of the model best fitted to the dataset shows that all but tissue site and the interaction between tissue site and Shannon score seem to have a significant effect on the viral load of the patient. Particularly significant relationships in the model are the Shannon score (p=3.088e-10, 1d.f.), as reflected in our observations earlier and in figure 6, and the CD4 count – Shannon score interaction (p=0.001742, 1d.f.).

Figure 6: A plot showing the model of the relationship between viral load and Shannon score in the patient's dataset.



Appendix

R script for dataset 1

```
attach(Marine_diversity)

#Initial viewing of the data to observe obvious trends and data distribution

boxplot(UniFracInd~season,xlab="season",ylab="UniFracInd",main="Microbial diversity by Season",col="steel blue")

boxplot(UniFracInd~latitude,xlab="latitude",ylab="UniFracInd",main="Microbial diversity by latitude",col="steel blue")

#Test for normal distribution of the UnifracID over the samples

shapiro.test(UniFracInd)

#We know that the data is normally distributed,so can perform a T-test on the data

#First the data is separated by season

UnifracJan<- subset(UniFracInd, season=="Jan")

UnifracAug<-subset(UniFracInd,season=="Aug")

#Performed a T-test on the two subsets to see if there is a significant difference between diversity by season

t.test(UnifracJan,UnifracAug,paired=TRUE)

#p=0.01604, reject null hypothesis, season has significant effect on diversity

#Repeated for latitude

UnifracTrop<-subset(UniFracInd,latitude=="tropical")

UnifracTemp<-subset(UniFracInd,latitude=="temperate")

t.test(UnifracTemp,UnifracTrop,paired=TRUE)

#p=0.3113, accept null hypothesis, latitude has no significant effect on diversity

#Performed Anova on the dataset to test for interactions between season and latitude on diversity

MarineAnova1<-lm(UniFracInd~season*latitude)

anova(MarineAnova1)

#Anova agrees with t-test results that there is no significant interaction between diversity and latitude

#But that there is an interaction between season and diversity.

#Anova also suggests there is no joint interaction between season and latitude on diversity (p=0.846)
```

R script for dataset 2

```
attach(luciferase_phylo)

plot(distance,expression_fold)

luciferanova1<-lm(distance~expression_fold)

abline(luciferanova1,col="red")

#GLM assumptions

#Data is independent.

#Normality of error:
```

```
hist(luciferanova1$residuals)

qqPlot(luciferanova1$residuals)

shapiro.test(luciferanova1$residuals)

#p=0.4197, residuals are normally distributed

#homoscedasticity:line seems even across Residuals vs fitted plot

par(mfrow=c(2,2))

plot(luciferanova1)

#linearity

#data points seem fairly linear in res-fitted despite spread, intial plot also linear
```

R script for dataset 3

```
full.model<-lm(VLoad~.,data = HIV_load)

step.model<-stepAIC(full.model,direction = "both",trace = FALSE)

summary(step.model)

#gives VLoad ~ CD4 + score_shannon + score_distance, tissue dropped

backwards_final=step(lm(VLoad ~ CD4 * score_shannon * score_distance),direction = "backward")

#formula VLoad ~ CD4 + score_shannon + score_distance + CD4:score_shannon + CD4:score_distance

#AIC=20.02

#Tried reintroducing tissue variable

final_final=step(lm(VLoad ~ score_distance + CD4 + score_shannon + CD4:score_shannon +
score_distance:CD4),scope=c(lower=~score_shannon,upper=~
CD4*score_shannon*score_distance*tissue),direction="both")

#formula VLoad ~ score_distance + CD4 + score_shannon + tissue + CD4:score_shannon +
score_distance:CD4 + score_shannon:tissue

#AIC=16.43

final_HIV<-lm(VLoad ~ score_distance + CD4 + score_shannon + tissue + CD4:score_shannon +
score_distance:CD4 + score_shannon:tissue)

#anova of final_HIV shows high variance in tissue variable, however removing it increases AIC

#VLoad ~ score_distance + CD4 + score_shannon + CD4:score_shannon + score_distance:CD4

#AIC=20.02

#Test viral load data for normality

hist(VLoad)

shapiro.test(VLoad)

#Paired t test prep

VLbrain<-subset(VLoad,tissue=="brain")

VLcord<-subset(VLoad,tissue=="spinalCord")

t.test(VLbrain,VLcord,paired = TRUE)

VLHi<-subset(VLoad,CD4=="hi")
```

```
VLLo<-subset(VLoad,CD4=="lo")
t.test(VLHi,VLLo,paired = TRUE)
summary(final_HIV)
anova(final_HIV)
```