Queen Mary University of London

School of Biological and Chemical Sciences

Coursework Coversheet

Student Details Place barcode in the dashed box below.	MARK
1 7 0 4 8 4 3 6 7	
GRIGORIADIS, Dionysios	

BIO782P		PARKER, Joe
STATISTICS FOR BIOINFORMATICIANS		
WEEK 1 ASSESSMENT		
Friday, 1 December 2017, 5:00 PM		

Feedback

If generic feedback is provided on the module webpage then please indicate by ticking here: ☐

General comments

Suggestions for improvement

Declaration [to be completed by student]

I certify that this coursework that I am submitting is my own work, that it has not been copied in part or in whole from any other person, and that any ideas or quotations from the work of other people, published or otherwise, are properly referenced. I have read and understood the School guidelines on plagiarism and I am aware that penalties will be applied for any plagiarism or other poor academic practice.

	30/11/2017

Dataset 1: Marine microbial diversity

The first goal analysing the experimental data was to examine if and how the location of the sampling (equatorial or temperate waters of the North Atlantic) affects the observed microbial diversity (MD) (relatively expressed to a reference sample). For this reason, MD data (n=20) were divided into two distinct groups, depending on their sampling location, “equatorial” group (n=10) for equatorial waters and “temperate” group (n=10) for temperate waters. Interestingly, the MD of the “temperate” group was significantly higher compared to the “equatorial” group (Welch Two Sample t-test, P=0.0000002), suggesting that the temperate waters of the North Atlantic show increased MD compared to the equatorial waters (Figure 1).

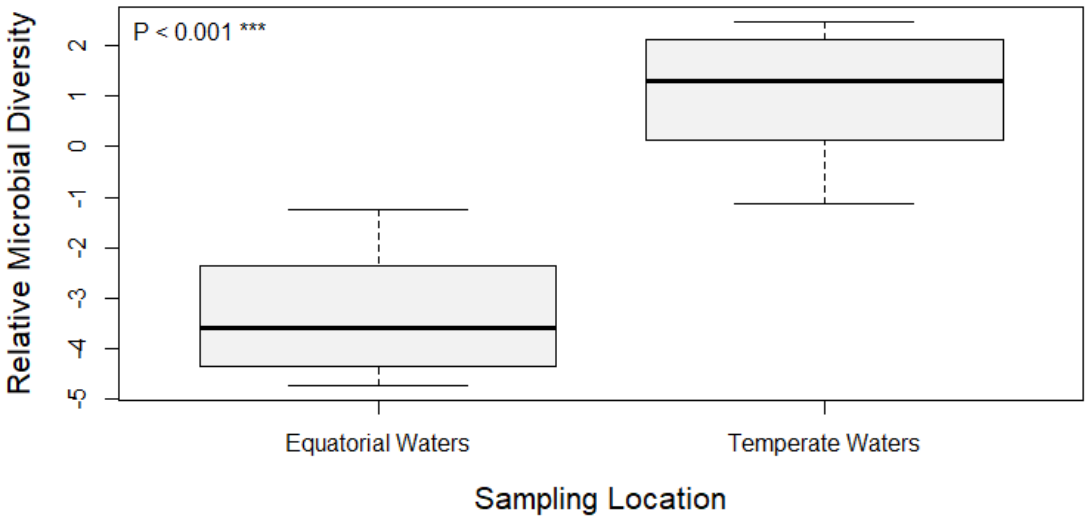

Figure 1. Relative microbial diversity levels in different sampling locations. The plot shows the relative MD in different sampling location groups (equatorial and temperate waters of North Atlantic). Microbial diversity appears to be significantly higher in temperate waters compared to equatorial waters (Welch Two Sample t-test, equatorial mean=-3.56 vs. temperate mean=1.04, P=0.0000002377). *P ≤ 0.05, **P ≤ 0.01, ***P≤0.001.

The potential correlation between the time of the year that sampling conducted with the MD was also investigated. MD data (n=20) were divided into two groups depending on the time of the year that cruise took place, “January” group (n=10) for the January cruise and “August” group (n=10) for the August cruise. The analysis revealed that regardless the time of the year that sampling conducted the MD remained unaltered suggesting that probably the microbial diversity of the North Atlantic waters remains relatively stable while seasons change (Figure 2).

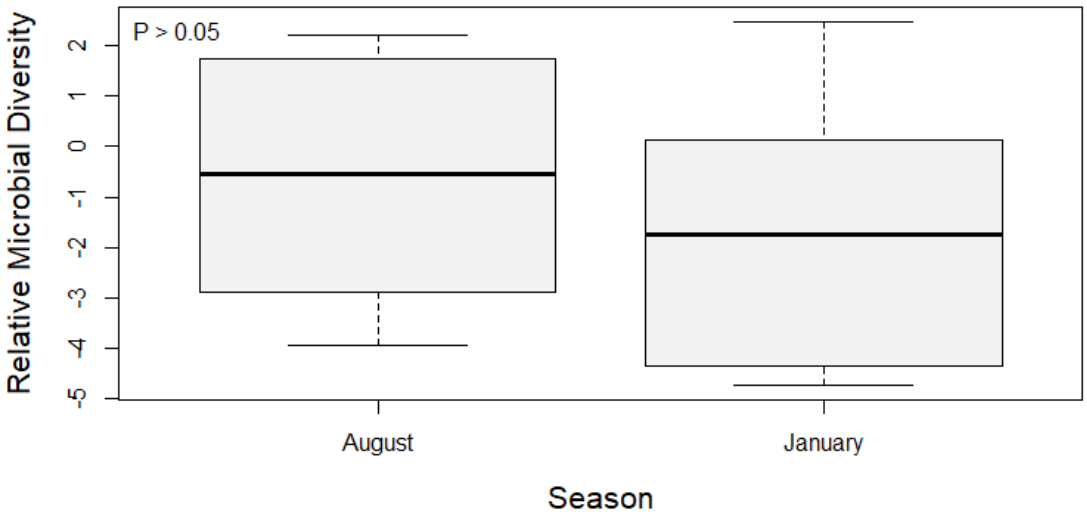

Figure 2. Relative microbial diversity levels in different sampling seasons. The plot shows the relative MD in different sampling season groups (August and January). Microbial diversity appears to remain unaffected by the different time of the year that sampling occurred. (Welch Two Sample t-test, January mean=-1.687 vs. August mean=-0.628, P=0.3674).

In terms of experimental design, there could not be an interaction between season and location of sampling as equal times (5) samples had been collected from equatorial waters in both January and August, while the same sampling methodology was applied to temperate waters as well. A Pearson's chi-square test of independence was performed to examine the relation between the location and season of sampling. As expected, these variables were found to be totally independent (X-squared = 0, df = 1, p-value = 1).

Dataset 2: Pairwise nucleotide substitutions and RNA expression levels

As the the plot of the putative 'luciferase' homologue expression levels (Expression) against the genetic distance from the report gene measured in amino acid substitutions (Distance) suggests, there might be a positive linear interaction between these two continuous variables. To test this hypothesis, a simple linear regression was calculated to predict the putative Expression based on the measured Distance. A significant regression equation was found ($F(1,14) = 284.2$, $P = 3.402e\text{-}16$ – ANOVA, $R^2 = 0.9103$) . Based on this regression (Figure 3):

Expression = 1.67 x Distance + 2.49

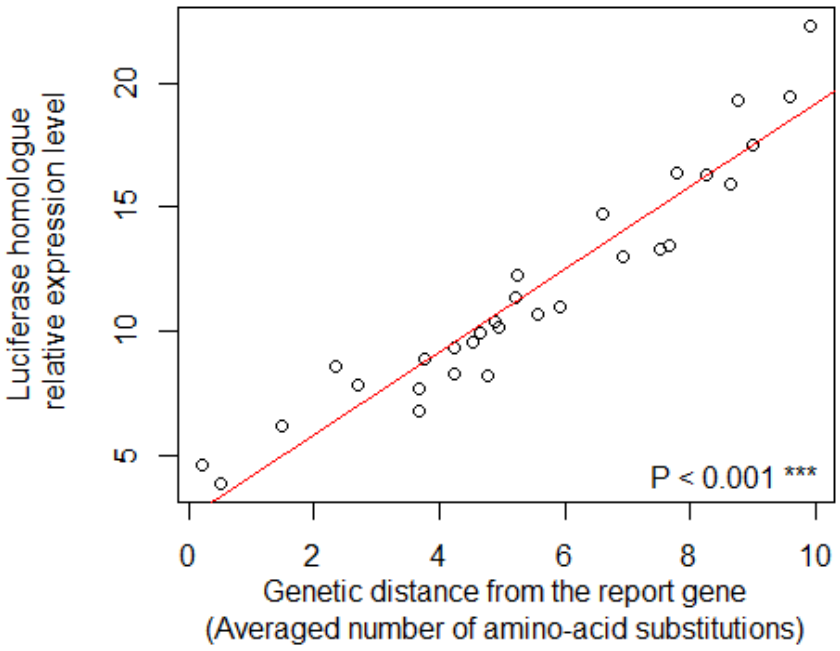

Figure 3. The scatter diagram and the regression line of Luciferase homologue expression level of plants within the Brassicaceae family against the genetic distance from the report gene of Arabidopsis Thaliana measured in averaged amino-acid substitutions. The regression equation is: Expression = 1.67 x Distance + 2.49, (F(1,14) = 284.2, P = 3.402e-16, $R^2 = 0.9103$).

The assumptions of this proposed model were also checked for validity. The independence of the data points is guaranteed from the experimental design knowing that the luciferase expression for a given genetic distance tells us nothing about the error of any other data point. The normality of the residuals is also validated, as the residuals fall on a line on the diagnostic Q-Q plot. Furthermore, the relationship between the explanatory and the response variable seems to be linear, as there is no pattern in the diagnostic residuals vs. fitted values plots. However, homoscedasticity assumption is not valid. More specifically, the residuals (or standardized residuals) vs. fitted plot of the diagnostic plots indicate that the variance is not constant over all predicted values of the response variable. For this problem to be overcame, several transformations were applied to this model, with the square-root transformation of the luciferase homologue expression leading to a linear model with better assumptions.

The homology found between these genes of plants within the Brassicaceae family and the luciferase gene is not that surprising, as it has been evidenced in the past that Arabidopsis genes encode enzymes of the firefly luciferase (Staswick, Tiryaki and Rowe, 2002). It is normal

that several mutations are being found between the species of this family which alter the amino acid sequence of the encoding protein and affect gene expression. Species that benefit more from the expression of this gene are trying to find ways to enhance its expression. Thus, mutations are accumulated in specific regions enhancing the gene expression by altering the epigenetic profile and/or by facilitating gene transcription.

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

To have an insight into which of the measured variables (Average Shannon population diversity, mean pairwise genetic distance from individual viruses present in each weekly sample to a reference sequence, tissue and CD4+ cell counts in the patient's general circulation) have an explanatory power on the response variable of viral population size, the relationship of each of these variables with the viral load is tested. A Pearson product-moment correlation coefficient was computed and demonstrated that there was no correlation between Shannon diversity and viral load ($r = 0.048$, $p = 0.7672$), while there was a significant positive correlation between genetic distance and viral load in contrast to the genetic distance ($r = 0.367$, $p = 0.01983$). Furthermore, CD4+ cell counts in the patient's general circulation was not affecting the viral load values (viral load values of "high" CD4+ patients mean = 12.02 vs. viral load values of "low" CD4+ patients mean = 10.72, Welch Two Sample t-test, P=0.601). Similarly, the tissue from which the sample was collected were found unrelated to the viral load values (viral load values from spinal cord mean = 10.77 vs. viral load values from brain mean = 11.98, Welch Two Sample t-test, P= P=0.5742). Given all these observations, only the variable representing mean pairwise genetic distance from samples' individual viruses to the reference sequence seems to explain the observed changes of the viral load.

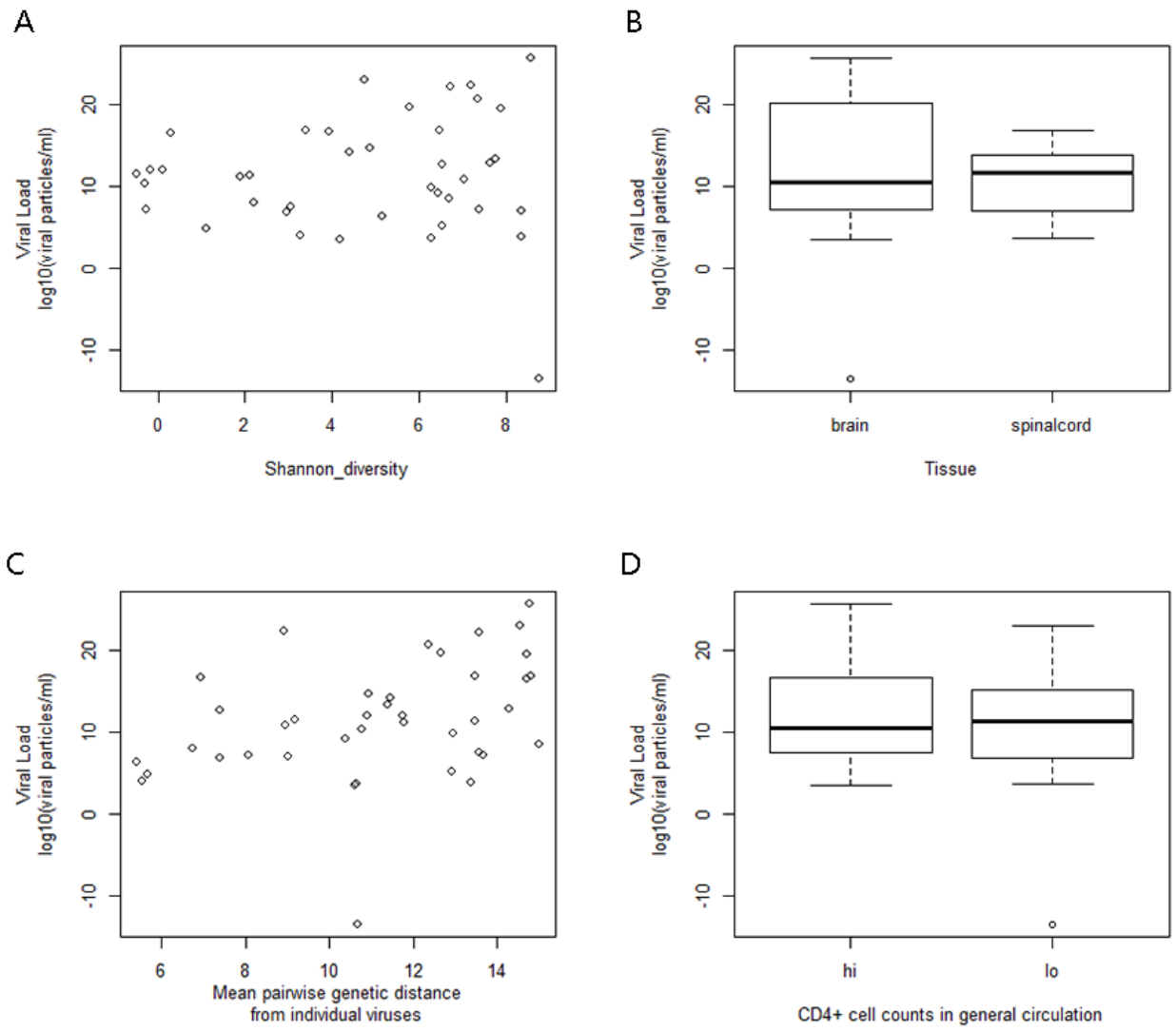

Figure 4. Viral population sizes values are plotted against the other explanatory variables: A) Average Shannon population diversity. B) Sampling tissue C) mean pairwise genetic distance from individual viruses present in each weekly sample to a reference sequence. D) CD4+ cell counts in the patient's general circulation.

To further test this hypothesis, stepwise variable selection procedures (Forward selection, Backward selection, Stepwise Selection) were followed to fit the best model that explains viral load in terms of the other variables. The resulted models were compared together based on the Akaike information criterion (AIC). The best resulted model (AIC = 154.91) was a simple linear that predicts the viral load based on the genetic distance. The assumptions of these model were valid. Based on this regression:

Viral load = 0.92 x Genetic Distance + 1.18

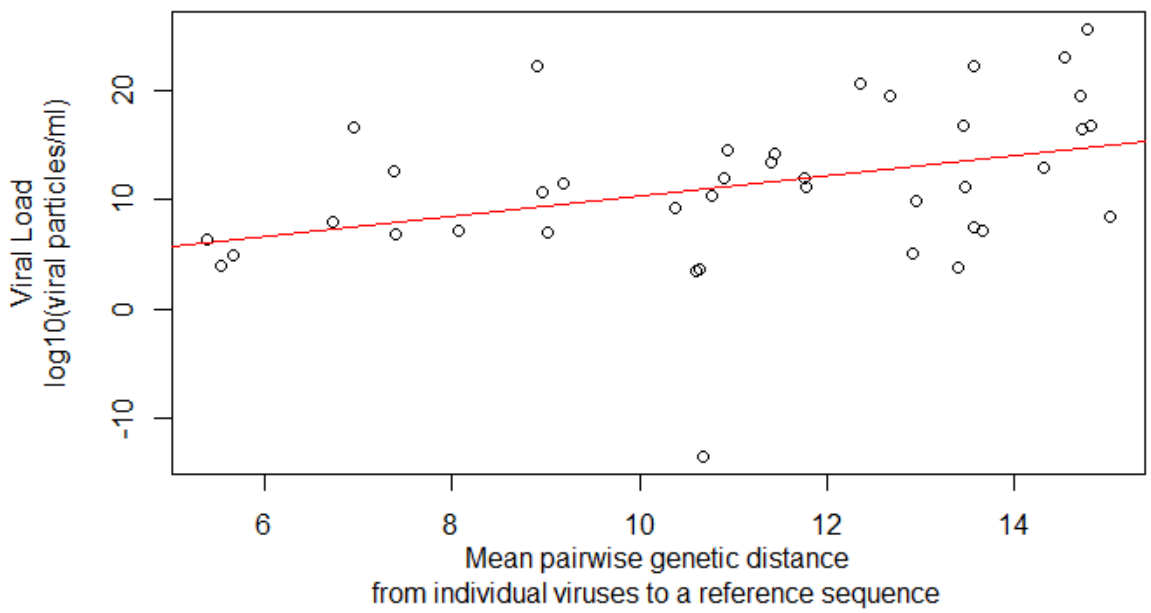

Figure 5. The scatter diagram and the regression line of Viral population size -here expressed as log10(number of viral particles per ml)- against the mean pairwise genetic distance from individual viruses present in each weekly sample to a reference sequence. The regression equation is: Viral load = 0.92 x Genetic Distance + 1.18. AIC = 154.91.