**Dataset 1**

**Results**

**Marine microbial diversity is influenced by season.**

A two-way analysis of variance was conducted on the influence of independent variables (latitude and season) on the log relative diversity of marine microbial diversity of seawater, each variable had two levels (equatorial/temperate and Aug/Jan respectively) (Figure 1). Season was the only effect to be statistically significant using a critical value of P<0.05. The main effect for latitude gave an F ratio of F(1, 16) = 1.004, p=0.331, indicating a non-significant difference between diversity of equatorial (M = 1.828, SD = 2.037) and temperate samples (M = 1.172, SD = 2.017). The main effect for season gave an F ratio of F(1, 16) = 18.25, p=0.001, indicating a significant difference between samples from January (M = 0.101, SD = 1.351) and August (M = 2.900, SD = 1.510). The interaction effect was not significant, F(1, 16) = 0.211, p=0.652.

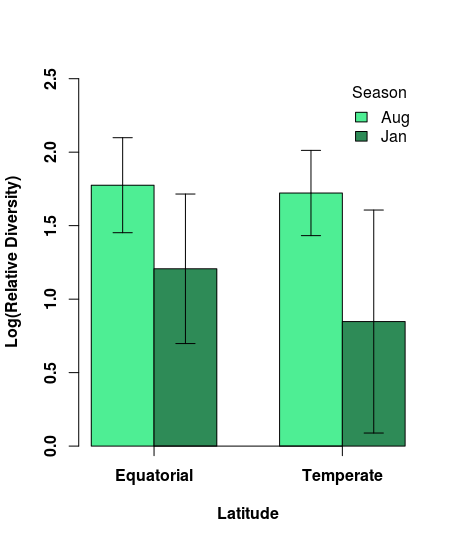


Figure 1 | Marine microbial diversity levels are sensitive to only seasonal changes. Samples were collected from two different latitudes during two different seasons; sample sizes were N = 5 for each of the four seawater sample types. Microbial diversity, latitude and season were modelled by linear model. Significance between groups was tested by ANOVA and assessed using P<0.05. Barplot showing means and 95% confidence intervals of log relative fitness for each of the four seawater sample types.

**Dataset 2**

**Results**

**Luciferase homologue expression increases with genetic distance.**

Luciferase expression levels were plotted against the average amino acid substitutions (genetic distance) and were found to be highly positively correlated (Pearson product-moment correlation, r(28) = 0.986, p=2.2e-16), Figure 2.

Two models were proposed for expressing the relationship between expression levels and genetic distance: a linear model and a polynomial model. The model that best represented the relationship between the variables was found to be polynomial (ANOVA – F(1, 27)=12.542, p=0.001).

Luciferase expression levels could be predicted from average amino acid substitutions by the following formula (when x = average amino acid substitutions): Expression Level = 4.42100 + 0.911552x + 0.07229x2, R2 = 0.979. This model shows that the luciferase homolog expression exponentially increases with genetic distance, as seen in Figure 2.

Linear regression makes the assumptions that errors are normally distributed, homogeneity of variance and linearity. When looking at Residuals vs Fitted and NormalQ-Q plots for the model (Figure 3) the Normal Q-Q plot shows that there is a mostly normal distribution, but the Residuals vs Fitted plot indicates that there may not be homogeneity as the variance appears to decrease with genetic distance. As this model is exponential it doesn’t fit the assumption of linearity.

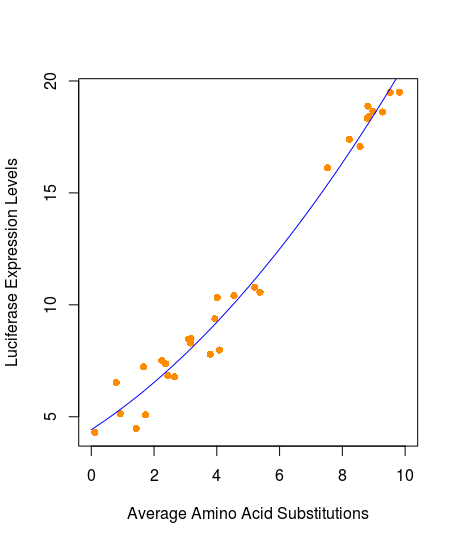
An evolutionary driver for maintaining an amino acid substitution is for the change to have had no effect or to give an evolutionary advantage. The evolutionary benefit of having increased luciferase expression may be the driving force for maintaining amino acid changes and therefore increasing genetic distance with increased luciferase expression. 

Figure 2 Luciferase expression is positively correlated with amino acid substitutions. A scatter plot of Luciferasse expression levels against the average amino acid substitutions, showing luciferase expression exponentially increasing with the average amino acid substituions from the reference, with fitted line for the quadratic model (R2 = 0.979).

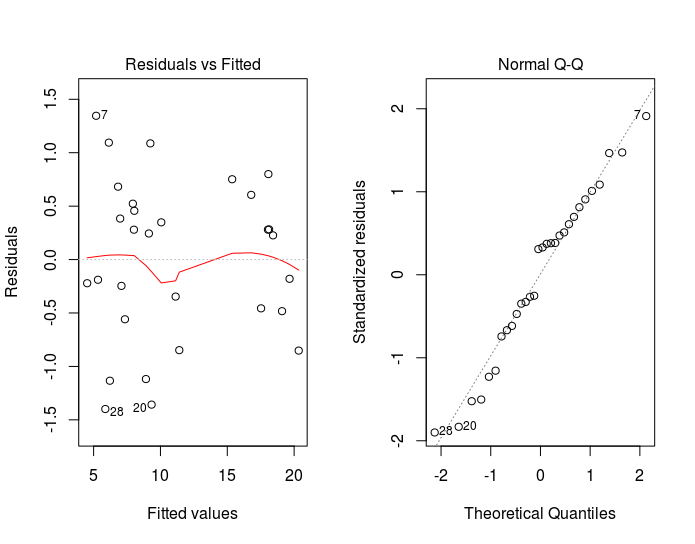


Figure 3 | Disgnostic plots for the quadratic model of luciferase expression given by the average amino acid substitutions. The residuals vs Fitted plots showing the level of homogeneity of variance by the spread of the data points (left), and the Normal Q-Q plot suggesting that the data conforms to the assumption of normal errors with all the data points fitted closely to the central dotted line (right).

**Dataset 3**

**Method**

A total of 40 samples were taken from a patient for a period of 40 weeks, with data on viral load, Average Shannon population diversity (diversity), mean pairwise genetic distance (distance), sample location (tissue type) and CD4+ count being recorded. The two tissue types were spinal cord (n=20) and brain (n=20).

The following assumptions were used and tested prior to modelling: linearity, independent variables and homoscedasticity. A critical value of p<0.05 was used to measure significance.

Models were curated both manually and automatically; models were then compared to find the best model to show the relationship of viral load and other independent factors. During manual model curation the starting model was the maximal model. Each term was considered individually and the least significant term of the highest order was removed until removal of a term would cause a significant difference. Models were compared to the minimal model and maximal model, log versions of these models were also used.

**Results**

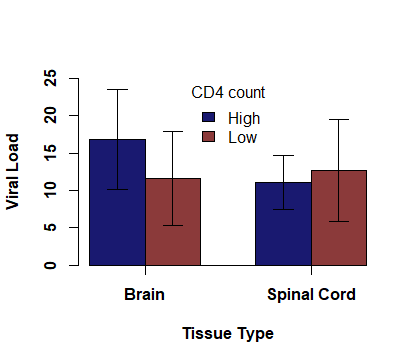
The ability of the immune system to penetrate the CNS and the effect this had on the viral load was assessed, overall the CD4 count and tissue type were found to be non-significant factors in viral load(ANOVA – F(1,36) = 0.736, p=0.397 and F(1,36) = 1.145, p=0.292 respectively). In the brain a difference in viral load can be seen that corresponds to the CD4 count, with higher CD4 counts generally seen with higher viral loads, possibly indicating that the immune system is able to penetrate this area of the CNS (CD4 high: M = 16.865, SD = 7.670; CD4 low: M = 11.566, SD = 7.161). This can’t be seen in the spinal cord, with viral load being similar levels irrespective of the immune response (CD4 high: M = 11.106, SD = 4.084; CD4 low: M = 12.690, SD = 7.789). There is some interaction between the tissue type and the CD4 count - although this was not considered statistically significant (ANOVA – F(1,36) = 2.526, p = 0.121). Due to the high error rates seen, these results aren’t fully conclusive.

Figure 4 | HIV viral load is not influenced by tissue type of immune reaction. Samples were collected from two different sites in the body and the general CD4 count measured; sample sizes were N = 20 for each tissue type and N = 20 for each CD4 level. Viral load, tissue type and CD4 count were modelled by linear model. Significance between groups was tested by ANOVA and assessed using p<0.05. Barplot showing means and 95% confidence intervals of viral load.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table 1 | Main experimental factors predicting HIV viral load**. ANOVA analysis of variance. Statistics describe linear models and their significance levels. Bold values indicate statistically significant factors that were maintained in the manually curated model. | | | | |
| Experimental Factor | Distance | Shannon Diversity | CD4 level | Tissue type |
| Significance | F(1,35) = 13.121  **P=0.001** | F(1,37) = 0.2436  P=0.625 | F(1,31)=0.926  P=0.343 | F(1,36)=0.007  P=0.936 |

Initial exploration of the data showed that the only variables showing a significant relationship were viral load and distance, a strong positive correlation can be seen with viral load increasing with genetic distance (Pearson’s coefficient, r(38)=0.495, p=0.002) (Figure 4). This is seen in the models curated as other factors play a non-significant role in predicting viral load; the main factors and their significance are listed in Table 1.

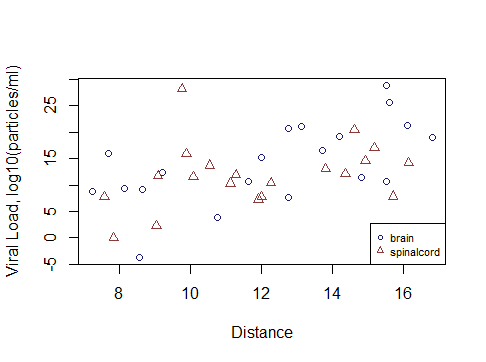


Figure 5 | Scatter plot showing the strong correlation between viral load (viral population expressed as log10(particles per ml) and distance (mean pairwise genetic distance to a reference sequence). Tissue did not play a significant role in predicting viral load.

The concluding automatically curated backwards stepwise model was equivalent to the maximal model and the forward stepwise model gave viral load as a measure of Distance + Diversity, meaning that patient viral load is higher for more genetically distant and diverse viruses. Comparison of these models showed that maximal model to be a better model than the simplified one, but not significantly (F(13, 24)=1.609, p=0.1514). Due to this, and the principle of parsimony, I would suggest that the forward stepwise model is best due to simplistic models with fewer parameters are preferred.

The manually curated model gave a minimal adequate model for viral load as viral load increasing with genetic distance (Figure 5). Comparative to the log maximal model the manually curated minimal adequate model was not a significantly better fit (F(14, 38)=0.9929, p=0.4892); therefore I would also conclude that the manually curated model is more suitable for describing the model due to the principle of parsimony. Further analysis of the model may be required without inclusion of the outlying data point that can be seen, as it has a Cook’s distance greater than 0.5 meaning it may be influential.

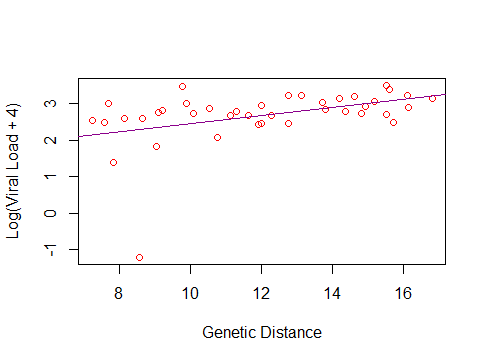


Figure 6 | Scatter plot showing the manually curated model for viral load. Log(viral load) plotted against genetic distance with the model plotted as a line.

Total word count = 999.

**References**

R Core Team (2017). “R: A language and environment for statistical computing”. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.