KU LEUVEN

Evolutionary and Quantitative Genetics – I0D53A
Evaluation sheet - Assignment 2
Landscape-genomics

Due date of the assignment: 29.11.2023 at noon

Your family and first name: *Plewa Daria*
Your student number: r0976669

You are asked to prepare landscape genetic models for allelic richness and pairwise F_{ST}.

1. Which statistical analyses do you propose to investigate barrier impact? Do you think there is a straightforward way to measure barrier impact? Why or why not?

To investigate the barrier impact we should take into account the descriptive analysis of populations, the correlation of explanatory variables, and later AMOVA, Non-parametric / Parametric Regression / Multiple Regressions Models models.

With the use of descriptive analyses, we would be able to determine the data distribution and consequently what kind of analysis we can undertake. The correlation of explanatory variables can show us the intercorrelated variables that can have an impact on the final result. Finally, we would be able to carry out the model analysis. AMOVA (Analysis of Molecular Variance) will show us the differences between the AR (Allele Richness) and FST (Pairwise Genetic Differentiation). Non-Parametric Regression models could be interpreted as AR while the Parametric ones as FST. And Multiple Regressions Models models we would be able the get the values corresponding to AR and FST.

With the use of R2 we can compare the results of our model to the real ones and evaluate its accuracy, and thanks to AIK (Akaike Information Criterion) we can choose the best model.

Unfortunately, the exact level of barrier impact could be hard to achieve with basic statistical approaches. With more advanced analysis like Machine Learning and Deep Learning probably we would be able to obtain more realistic results, without biases as in more primary methods. We need to remember that the barrier impact can be caused by different kinds of events such as ecological, geographical, environmental, and historical processes - and a model taking all those effects into account with certain weights without other biases could be the most suitable one. Consequently, with more advanced models, we should obtain the result that measures the barrier impact the best.

2. **Please provide me a table with the output of the statistical analysis, choose the best model, and interpret the results in a few lines. What do the results actually mean?**

Comparison of AIC for AR and FST

Model	AIC - AR	AIC - FST
1	1.092017	-57.79286
2	1.600424	-61.27569
3	**-1.399841**	-59.66505
4	11.40008	-61.1325
5	0.4261399	-56.53184
6	19.29477	-62.68645
7	11.12163	**-64.09174**
8		-62.07646
9		-58.14681
10		-57.36302
11		-56.8902
12		**-64.81343**
13		-55.9681
14		-53.90724
15		-59.02749

Best model for AR is the model with the lowest AIC - consequently, the 3rd model with the AIC value -1,399841, and for FST model 12 is the best with the AIC value equal to -64,81343, a bit worse is model 7 with AIC -64,09174.

Summary of best AR model

AR ~ all_barriers + log10.upstream.distance.				
Residuals				
Min	1Q	Median	3Q	Max
-1.78440	-0.36128	-0.01934	0.37699	1.79081
Coefficients				
	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	7.6421	0.9128	8.372	1.28e-07
all_barriers	-0.1592	0.0363	-4.386	0.000356
log10.upstream.distance.	1.0760	0.5143	2.092	0.050864
Adjusted R-squared (R^2)		0.7684		
$AR = 7.6421 - 0.1592 * all.barriers + 1.0760 * up.stream.distance + x_E$				

The best AR model takes into account the barriers and distance to the sea (upstream.distance) with model: $AR = 7.6421 - 0.1592 * all.barriers + 1.0760 * up.stream.distance + x_E$. In this model the barriers have a degrading effect on AR - with each barrier, the AR is degrading by -0.1592. Simultaneously the distance to the sea is increasing the AR value - with each kilometer, the AR would be growing by 1.0760. The model obtained adjusted R^2 equal 0,7684 which is quite high - consequently, the model has a high power and is able to explain 77% of the variation.

Summary of the best FST model

fst ~ all_barriers				
Residuals				
Min	1Q	Median	3Q	Max
-0.142227	-0.043015	-0.005367	0.037996	0.230181
Coefficients				
	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	0.0451032	0.0085997	5.245	3.84e-07
all_barriers	0.0064181	0.0004581	14.011	< 2e-16
Adjusted R-squared (R^2)		0.4831		
$FST = 0.0451032 + 0.0064181 * all.barriers + x_E$				

The best FST model consists only of barriers with an equation $FST = 0.0451032 + 0.0064181 * all.barriers + x_E$. With each barrier, the FST and the inbreeding would be growing by 0.0064181. It has not as great power as the previous model regarding the AR because only 0,4831 - but still we would be able the explain 48% of the variation.

3**. Do you have suggestions for improving the study?**

The study was conducted all right in view of the obtained data. However, we did obtain a small amount of samples. Next time similar analysis could be conducted with the use of a bigger research sample. Or with the use of bootstrap sampling - that however could also cause the growth of biases. Additionally using more advanced models, especially Deep Learning could significantly improve the R^2 score of models - however, that would be even more useful if we took into account more ecological, historical, environmental, and geographical data.