Analysing Neutral Population Structure

James O'Reilly
Student Number: r0773125

1 Hierarchical genetic structure

1.1 What is your outcome of the AMOVA?

Table 1 shows how much variance is detected at each stratification. We expect variations within samples to give the greatest amount of variation for populations that are not significantly differentiated. Sigma represents the variance, σ, for each hierarchical level and to the right is the percent of the total variance.

Level of Variance	σ	%
variations between groups	0.207	8.21
variations between samples within groups	0.386	15.28
variations within samples	1.932	76.51
total variations	2.525	100

Table 1: Components of variance given by the AMOVA.

The ϕ-statistics output by the AMOVA provide the population differentiation statistics (see Figure 2). These are used to test hypotheses about population differentiation. We would expect a higher ϕ-statistic to represent a higher amount of differentiation.

Hypothesis	ϕ
phi-samples-total	0.207
phi-samples-pop	0.386
phi-group-total	1.932

Table 2: Phi-statistics given by the AMOVA.

1.2 What is an AMOVA exactly? Explain with your own words.

Analysis of Molecular Variance (AMOVA) is a method of estimating population differentiation directly from molecular data and testing hypotheses about this differentiation. In our case, the molecular data is microsatellite data. An AMOVA treats any kind of raw molecular data as a Boolean vector d_i, with 1 indicating the presence of a marker (eg. microsatellite) and 0 indicating its absence. The pairwise squared Euclidean distances between vectors d_i and d_j are then calculated using the equation

$$\delta_{ij} = (\mathbf{d_i} - \mathbf{d_j})'\mathbf{W}(\mathbf{d_i} - \mathbf{d_j}) \tag{1}$$

where W is a weighting matrix. Often it is the identity matrix but this can be changed based on how you want to weight molecular change at different locations.

These distances are then arranged into a distance matrix, and partitioned into submatrices which correspond to the subpopulations present. The submatrices on the diagonal of the distance matrix are pairs of individuals in the same population while those on the off-diagonal represent individuals from different populations. The sums of the diagonals in the matrix and submatrices yield sums of squares for the various hierarchical levels of the population which are then analysed using a nested analysis of variance. The nested ANOVA allows for hypothesis tests of between-group and within-group differences at several hierarchical levels. The AMOVA ultimately outputs

the variance explained at each level of the structure, as well as the phi-statistics discussed previously.

Phi-statistics are effectively a hypothesis about differentiation between the population and its component subpopulations, which can be tested using the null distribution of the variance components. **If the variance of the subpopulations does not significantly differ from the null distribution of the variance of the population, the hypothesis that those subpopulations are differentiated from the larger population would be rejected.** The null distributions are calculated by resampling with permutation. Hypothesis testing is carried out relative to these resampling distributions.

1.3 How do you interpret the hierarchical pattern?

The results of the AMOVA given in Tables 1 and 2 show that most of the variance (76.51%) arises from within samples. This is strong evidence for panmixia. However, there is clearly some hierarchical structure present, both at the sample and group level. 15% of the total variance is observed at the level of subpopulations, while 8% is at the level of grouping (coastal vs inland). To test if the differences observed between populations at different levels is statistically significant, a null distribution was constructed and the observed data was compared to this distribution. The results in Figure 1 indicate that the results are statistically significant. Overall, there is strong evidence that there is some hierarchical pattern, with statistically significant variance at the group and sub-population levels. However, most of the variance arises from within samples, which indicates random mating.

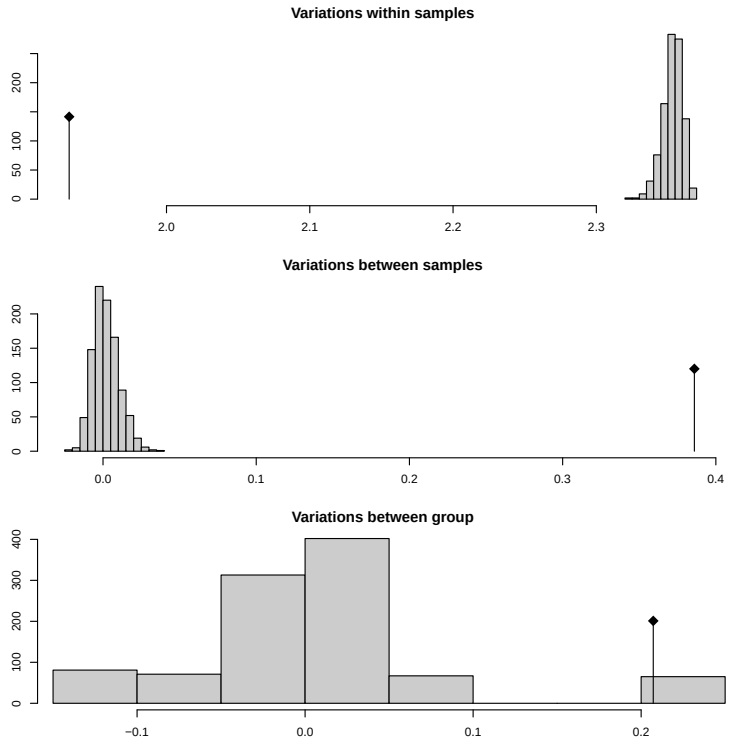

Figure 1: Significance tests for variation at different levels. The black line represents the observed data and the null distribution is in grey.

2 Outlier detection

We can use OutFLANK uses likelihood on a trimmed distribution of F_{ST} values to infer the distribution of F_{ST} for neutral markers. We can then use this distribution to determine if the F_{ST} for given loci are excessively high or low and therefore determine if they are outliers and

potentially subject to divergent or balancing selection. Importantly, trimming the distribution of F_{ST} values at the tails ensures that our estimated null (neutral) distribution is not affected by outliers, and therefore will perform better when testing for outliers. The extent to which we trim the original F_{ST} distribution therefore significantly impacts the sensitivity to outliers and the false-positive ratio.

In this assignment, we prune low-quality SNPs with missing data above a given percentage threshold. Pruning low-quality SNPs has two main effects:

- **Removes low heterozygosity loci:** Missing values are replaced by some placeholder value (in this case 9). If a locus has a large number of missing data, then this placeholder value will be present for many individuals and therefore reduce the heterozygosity of the locus. The distribution of F_{ST} is quite different for loci with low heterozygosity. These loci carry little information about F_{ST} because their sampling variance is so high. Removing these loci allows for a better estimation of the null distribution, and therefore improved outlier detection.
- **Reduces the amount of F_{ST} values contributing to the distribution:** By removing the loci, we also reduce the number of F_{ST} values which we can use to estimate the null distribution. The primary affect of this is that the that the distribution of observed neutral F_{ST} values will have larger variance and, therefore, overlap with selected loci even more, reducing the sensitivity to outliers.

Pruning low-quality SNPs therefore both increases and decreases sensitivity to outliers by affecting the estimation of the null distribution in different ways. Here we will prune at using thresholds of 40%, 25%, and 10%. Pruning at each of these thresholds and then running OutFLANK returned no outliers. This is likely because the SNPs have a large amount of missing data, and removing these SNPs significantly decreases the sensitivity to outliers by increasing the variance of the null distribution. In particular, pruning at a threshold of 10% left only 19 SNPs from which to estimate the null distribution (and this is before trimming). The null distributions along with histograms are given in Figure 5.

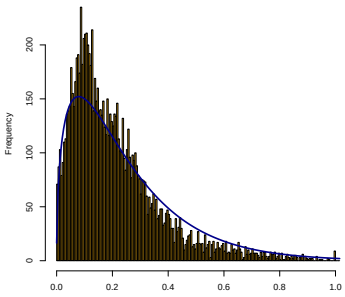

Figure 2

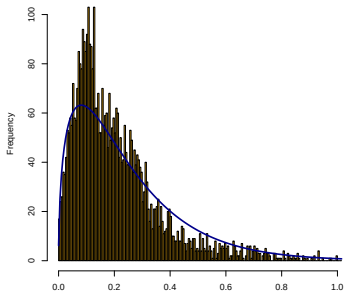

Figure 3

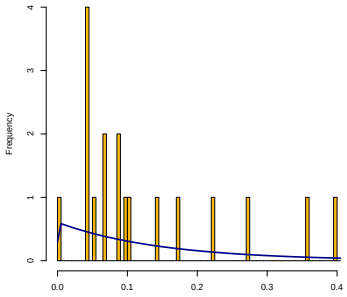

Figure 4

Figure 5: F_{ST} null distributions and histograms given by outflank with pruning at 40%, 25%, and 10% (from left to right)

One potential solution to increase sensitivity to outliers is to increase the trimming at the tails of the F_{ST} distribution. This ensures the core of the F_{ST} distribution on which the null model is fitted will have lower variance, and should therefore increase sensitivity to outliers.[1] Note that this again reduces the number of F_{ST} values used to fit the distribution and so also decreases the sensitivity, as previously discussed.

Why else might one not detect a signal of selection in this dataset? The neutral loci might be strongly differentiated. In these cases, it can be extremely difficult to statistically detect the signal of selection. Outlier approaches can only detect loci that are substantially more differentiated than expected of neutral loci, because the distribution of possible F_{ST} values these neutral loci are so broad. As a result, loci affected by only weak selection are unlikely to be detected.

References

[1] Whitlock, M. C., and Lotterhos, K. E. Reliable detection of loci responsible for local adaptation: inference of a null model through trimming the distribution of f st. *The American Naturalist 186*, S1 (2015), S24–S36.