**Case Study 7: Local adaptation of** **corkwing wrasse**

We are returning to the corkwing wrasse data set that we analyzed during the hybridization week. Our goal today is to identify loci responding to environmental selection. Here we can use two different types of tests, 1) those that model neutral genetic structure among individuals or populations and then identify SNPs which are outside of that distribution (FST outliers), or 2) those that directly correlate environmental variables with SNP allele frequencies (genotype-by-environment associations, GEA).

In Rstudio, set your working directory using the command (change kitchens to your netID):

setwd("/scratch/group/kitchen-group/MARB\_689\_Molecular\_Ecology/class\_working\_directories/kitchens/")

Make a new subdirectory as follows:

dir.create(file.path("11\_localAdaptation"))

Now, update your working directory:

setwd("/scratch/group/kitchen-group/MARB\_689\_Molecular\_Ecology/class\_working\_directories/kitchens/11\_localAdaptation")

Then, copy over today’s activity:

file.copy("/scratch/group/kitchen-group/MARB\_689\_Molecular\_Ecology/11\_local\_adaptation/LA\_wrasse.R", ".")

Open up the R script using the File > Open File on the top bar. Load the libraries listed on the top of the script.

1. We will read in the same VCF file as prior lab, located in the 07\_hybrids directory. As a reminder,

* How many SNPs are present in this data set?
* How many individuals?
* Where were these fish collected from?

2. What are the environmental variables that we downloaded to test later? What is the resolution of this data?

**Method 1: Differentiation-based outlier detection**

This method is most useful for detecting loci of large effect, or those with large differences in allele frequencies between locations. The approach does not examine environmental selection, and detects outliers using genetic data only (FST). We will run ***pcadapt*** (Luu et al. 2017), a tool based on principal component analysis where the FST of each SNP is regressed against each PC, with outliers extracted based on z-scores. This approach is not impacted by admixture (population mixing) and does not require us to define populations in advance. It calculates a ***genomic inflation factor*** (GIF) to correct for population structure or other factors not explicitly accounted for.

Under PCAdapt section, load in the genetic data.

path\_to\_file<-"/scratch/group/kitchen-group/MARB\_689\_Molecular\_Ecology/07\_hybrids/west.filt.maf0.01.recode.vcf"

filename <- read.pcadapt(path\_to\_file, type = "vcf")

In the first step, a PCA is performed on the centered and scaled genotype matrix. Specify number of PCs to retain. Ideally this should be the minimum number of PCs needed to describe “neutral” population genetic structure, but we don’t know what this is yet. For now, we’ll pick a number (20) that should be overkill, and then refine with a scree plot and score plot.

x <- pcadapt(input = filename, K = 20)

Look at the scree plot, plot(x, option = "screeplot", K=5).

3. What number of PCs should be retained based on the “elbow” method?

Now, look at the score plot (PCA), to see the individual scores of each sample along PC 1 and PC2. plot(x, option = "scores", pop = poptab\_gen$Sample\_ID)

4. How are the groups separating along PC1? What about PC2? And PC3?

Next, each SNP is regressed on K=3 principal components. The multi-dimensional (Mahalanobis) distance of each SNP from the mean covariance is measured using the D statistic. x <- pcadapt(filename, K = 3)

Let’s look at a summary of the output.

**maf:** minor allele frequencies

**loadings:** correlations between each genetic marker and PC

**sing.values:** K ordered squared root of the proportion of variance explained by each PC

**scores:** projections of the individuals onto each PC

**zscores:** z-scores from p-values

**stat:** Mahalanobis distances (by default)

**gif:** genomic inflation factor estimated from stat

**chi2.stat:** rescaled statistics stat/gif that follow a chi-squared distribution

**pvalues:** z-scores

We can apply a false discovery rate (FDR) to the p-values to account for the multiple tests performed on all loci. Those with q-values less than a chosen alpha are the outliers.

5. At an alpha level of 0.1, how many outliers are found?

**Method 2: GEA-based outlier detection**

The second class of outlier methods is those explicitly using correlation between the allele frequencies of SNPs and environmental variables putatively selecting them. These tests have their own pros (e.g. explicitly testing the reason for underlying positive selection), and cons (e.g. some programs have lots of false positives due to not controlling for population structure, or a poor choice of test statistic not accounting for multiple tests).

**Redundancy Analysis**

We will repeat the RDA multivariate ordination from the prior exercise to now identify the loci that are correlated with the axes (environmental variables) driving the spread in the data.

* It determines how groups of loci covary in response to the multivariate environment, as opposed to univariate approaches that apply multiple tests per locus. This makes it more useful for detecting weaker, polygenic signatures of adaptation.
* RDA performs multivariate linear regression on genetic and environmental data, producing a matrix of fitted values. Then PCA of the fitted values is used to produce canonical axes, which are linear combinations of the environmental predictors.
* RDA doesn’t require corrections for multiple tests because it analyzes all genomic and environmental data simultaneously. However, post processing includes checking and modification of the GIF and p-value distribution, with application of the FDR threshold.

The proportion of the variance explained by the environmental predictors is given under the “Proportion” column for “Constrained”; this is equivalent to the R2 of a multiple regression. Just like in multiple regression, this R2 will be biased and should be adjusted based on the number of predictors.

6. What is the adjusted R2? Does this value make sense in terms of the number of SNPs that are correlated with the environmental predictors?

You can run a formal test of statistical significance of each constrained axis using: anova.cca(wrasse.rda, by="axis"). We can assess both the full model and each constrained axis using F-statistics (Legendre et al, 2010). The null hypothesis is that no linear relationship exists between the SNP data and the environmental predictors. This can take a long time to run, thus, the screeplot provides an informal (and quick) way to determine how many constrained axes to include when we search for candidate SNPs.

7. Based on the scree plot, how many constrained axes may be informative?

Variance Inflation Factors can help identify redundant predictors. If values are below 10, then multicollinearity among these predictors is not be a problem. vif.cca(wrasse.rda)

Now, plot the RDA results as we have before. Here, the SNPs are in grey (in the center of each plot), and the individuals are the colored circles based on their sampling site. The red vectors are the environmental predictors. The relative arrangement of these items in the ordination space reflects their relationship with the ordination axes, which are linear combinations of the predictor variables.

**Identify RDA candidates:** We’ll use the loadings of the SNPs (their location) in the ordination space to determine which SNPs are candidates for local adaptation. The SNP loadings are stored as “species” in the RDA object. We’ll extract the SNP loadings from the three constrained axes, based on our assessment of the scree plot above.

load.rda <- summary(wrasse.rda)$species[,1:3]

If we look at histograms of the loadings on each RDA axis, we can see their (relatively normal) distribution. SNPs loading at the center of the distribution are not showing a relationship with the environmental predictors; those loading in the tails are, and are more likely to be under selection as a function of those predictors (or some other predictor correlated with them).

I’ve written a simple function to identify SNPs that load in the tails of these distributions. We’ll start with a 3 standard deviation cutoff (two-tailed p-value = 0.0027). As with all cutoffs, this can be modified to reflect the goals of the analysis and our tolerance for true positives vs. false positives. For example, if you needed to be very conservative and only identify those loci under very strong selection (i.e., minimize false positive rates), you could increase the number of standard deviations to 3.5 (two-tailed p-value = 0.0005). This would also increase the false negative rate. If you were less concerned with false positives, and more concerned with identifying as many potential candidate loci as possible (including those that may be under weaker selection), you might choose a 2.5 standard deviation cutoff (two-tailed p-value = 0.012).

8. How many RDA candidate loci were found on each axis?

9. Which environmental predictor(s) seem to be most strongly related to candidate SNPs in the RDA1/RDA2 plot? How about the RDA2/RDA3 plot?

10. Why do we see candidate SNPs loading in the center of the ordination space in both plots?

11. Which environmental predictors are most strongly correlated with the first three RDA axes? Going back to Question 8, can you make sense of the loci associated with the environmental predictors?

**Latent Factor Mixed Models**

LFMM is a regression model that includes unobserved variables (latent factors) that correct the model for confounding effects. The latent factors are estimated simultaneously with the environmental and response variables, which can help improve power when environment and demography are correlated.

For the univariate LFMM test, we could run a separate set of tests for each predictor, but this would be 4372 SNPs x 3 predictors, or 13,116 tests! Instead, for LFMM, we’ll perform a PCA on the environmental predictors and use the first principal component (PC) as a synthetic predictor. This will reduce our ability to interpret the output, since the PC predictor will be a linear combination of the original three variables, but it will reduce the number of corrections needed for multiple tests. Your decision of how to handle multiple predictors for a univariate GEA test will depend on the study goals and characteristics of the data set.

There are many ways to run PCA in R; we’ll use the rda function in vegan (Oksanen et al., 2016). We’ll center and scale the predictors (scale=T), since they’re in different units. We’ll then determine the proportion of the environmental variance explained by each PC axis & investigate how the original predictors correlate with the first PC axis.

pred.pca <- rda(pred,scale=T)

summary(pred.pca)$cont

12. How much variation in the data is explained by the first PC?

13. What are the strongest correlations of the predictors along PC1? What about PC2?

LFMM requires an estimate of the number of populations in the data (K). To determine the most likely value of K, we’ll use PCA, noting that there are many different approaches for determining “K” from genetic data. It is also reasonable to run LFMM with different values of K, if there is uncertainty.

We’ll use a broken stick criterion to determine K. The broken stick stopping rule states that principal components should be retained as long as observed eigenvalues are higher than corresponding random broken stick components.

14. What is the first PC to fail the broken stick rule?

Next, run LFMM with your chosen K value.

Next, we post-process the model output. Decisions made here can dramatically impact the candidate markers you identify.

The steps for post-processing are:

1. Look at the genomic inflation factor (GIF), which gives us a sense for how well the model has accounted for confounding factors in the data. An appropriately calibrated set of tests will have a GIF of ***around 1***. An elevated GIF may suggest the data is overly liberal in identifying candidate SNPs. If the GIF is less than one, the test may be too conservative.
2. Plot the p-values to see how application of the GIF influences the p-value distribution. We want to see a relatively flat histogram (most loci not under selection) with a peak near zero, indicating candidate adaptive markers. We see a very large peak with the unadjusted p-values, and a much smaller peak with the GIF-adjusted p-values (note differences in the scale of the y-axis).
3. Modify the GIF (if needed) and re-plot the p-values to identify the best possible fit to the “ideal” p-value distribution.
4. Apply a False Discovery Rate control method to the p-values by converting to q-values.
5. Identify candidates as those below a given FDR threshold. The False Discovery Rate is the expected proportion of false positives among the list of positive tests (see Storey and Tibshirani, 2003).

15. What is our GIF? Is this liberal or conservative?

16. How many loci were detected as outliers along the PC1 synthetic variable?

17. How could you change this analysis to be more or less conservative?

A common approach in studies that use GEAs and populations differentiation methods is to use many tests and then look for the overlap across the detections. This can be helpful for some questions (e.g. where minimizing false positive detections is of greatest importance), but can be overly conservative in many cases and bias detections against weaker signatures of selection.

18. Is there any overlap in loci from PCAdapt, RDA and LFMM?

19. Now that you have identified loci under selection, what would you do next?