**Case Study 4: Hybridization of isolated populations of corkwing wrasse**

In today’s activity, we will reproduce the results of Faust, Ellika, et al. "Cleaner fish escape salmon farms and hybridize with local wrasse populations." *Royal Society Open Science* 5.3 (2018): 171752.

Prior to 1990s there were no records of corkwing wrasse (*Symphodus melops)* in Northern Scandinavian region. Since then, they are occasionally collected in field surveys in mid to northern Norway (Figure A). This opened the question of 1) whether the range is expanding due to increased water temperature further north or 2) whether the observations were the result of accidental release from salmon aquaculture facilities. These small predatory fish are collected and used a cleaner fish to reduce sea lice infestations of salmon starting in the 1980s.



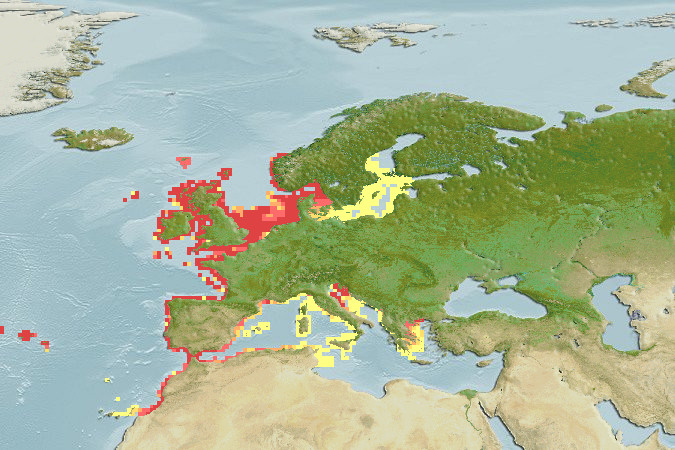
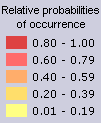
. 

Figure A. Native range with probabilities of occurrence.

They used 2bRAD, a reduced representation method for generating genotypes, on 240 fish collected from 6 regions that spanned the Northern, Western and Southern part of the species range.

Their sample map from the paper is below.



Figure B. Sample locations. Kristiansand, Stromstad and Kungbacka are the southern population, Stavanger and Austevoll make up the Western population and Flatanger is the northern expansion range.

Let’s see if we can replicate their genetic clustering analysis from Figures 2 and 3 in their paper. Go complete the first section of the R analysis script on Data filtering and PCA/DAPC.

1. **Explain** what the different filtering functions are doing to the genotype data. How many loci remained for the PCA and DAPC analyses?

* Missingno “loci” vs. “geno”=
* Informloci =

1. Copy the PCA scatterplot in the space below and circle the putative hybrids. What differences, if any, do you see between your plot and Figure 3 in the paper?
2. Copy the discriminant 1 plot in the space below and circle the putative hybrids.

Now, let’s test the population structure and possible hybridization of the samples using sNMF, a tool that estimates ancestry coefficients similar to STRUCTURE that was used in the original study. This tool uses sparse non-negative matrix factorization approach KH48whereas STRUCTURE uses a Bayesian iterative algorithm.

In the study, they ran STRUCTURE with all SNPs but then tested hybridization with a smaller subset of SNPs. In our example, we will compare the full data set (n= 4372 SNPs) to the reduced data set (n=200 SNPs). *Note- the color/population assignments may change between data sets BUT the order of samples on the x-axis will remain the same. Don’t get fixated on color but more on general patterns.*

1. What observations can you make about ancestry assignment between the full and reduced SNP sets? What population(s) were F, FKH50 and FKH67 assigned to?
2. How do your results compare to Figure 2 of the paper? Do your results align with their result of K=2 being the optimal K?

Let’s try another tool, SNAPCLUST, a maximum-likelihood approach that uses Expectation-Maximization (EM) algorithm.

1. What K is best from this tool? How does that match up with the original study (like Q5 above)?
2. How do the results change when you sequentially add on more complex hybrid types? Which model had the best fit?
3. Does this match Figure 4 in the original study?

Last test of the day, Patterson’s D statistics.

D statistics, sometimes called ABBA-BABA tests are a simple way to test for hybridization. Let P1, P2, P3, O be four taxa (populations, species, individuals) with a phylogeny ((P1,P2),P3),O), with P1 and P2 being the focal species, P3 the tested species potentially introgressing with P1 or P2, and the outgroup (O) used to polarize variants as ancestral (A) or derived (B) (Figure C). Under incomplete lineage sorting only (no hybridization), the number of discordant SNPs grouping P1 and P3 together (BABA patterns) should be roughly equal to the number of SNPs grouping P1 and P2 together (ABBA). The formula to compute the D statistic is based on allele frequencies in the program DSUITE (https://onlinelibrary.wiley.com/doi/10.1111/1755-0998.13265) and allows one or multiple individuals per taxon:

For each SNP:

num = (P1 – P2)(P3 − o )

den = (P1 + P2 – 2P1P2)(P3 + o − 2yo )

Summed across all SNPs:

D = sum(num) / sum(den)

If a single sequence (haploid individual) is used per population, this formula equals to:

D = (nBABA - nABBA) / (nBABA+nABBA)

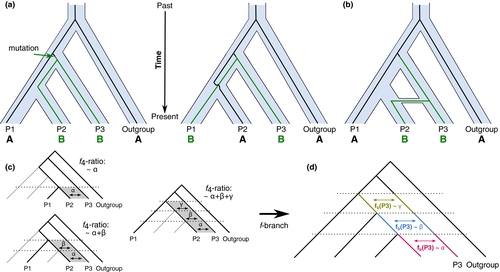


Figure C. Basic principles behind the D statistics. (a) Example genealogies showing the sharing of derived alleles, denoted as 'B' between populations P2 and P3 (the ABBA pattern) and between P1 and P3 (the BABA pattern) as a result of incomplete lineage sorting. In a scenario without gene flow, both patterns are assumed to be equally likely (but see Eriksson & Manica, 2012 for exceptions). (b) Gene flow between P2 and P3 introduces additional loci with ABBA patterns, which would lead to a positive D statistic.

DSUITE also outputs z-scores which are the number of standard errors that D deviates from 0. An absolute z-score of 3 is generally accepted as significant. If P1 and P2 share equal amounts of alleles with P3, D will be 0, or at least the absolute z-score will be below 3. In the output file, P1 and P2 are ordered so that nABBA>= nBABA. As a result, the D statistic is always positive and all the results, including the f4-ratio and other statistics, reflect evidence of excess allele sharing between P3 and P2 for each trio.

Using DSUITE to get a whole-genome point estimate of the D statistic is particularly useful to test multiple different populations for hybridization and to get an overview who hybridized with whom. Whole-genome estimates of the D statistic do not require whole-genome sequencing and work well with RAD data and UCEs. However, we would also recommend you treat results with less than 50 ABBA or 50 BABA patterns with caution.

Sometimes we may want to run an f4 test instead of the D statistic. This is a test of whether four taxa are two pairs of sister taxa ((P1,P2),(P3,O)). F4 tests can also be used to estimate the admixture proportion if combined in a smart way. DSUITE computes the f4 ratio tests to estimate admixture proportions.

1. How many trios have significant evidence of gene flow between P2 and P3?
2. Based on the highest f4 ratio, which of the southern regions (Kristiansand or Stromstad) shares the most alleles with Flatanger?

***Redundancy Analysis (RDA), a genotype by environment association test***

RDA is a multivariate ordination technique that can analyze many loci and environmental predictors simultaneously. For this reason, we can input all of the SNPs and environmental predictors at once, with no need to correct for multiple tests. RDA determines how groups of loci co-vary in response to the multivariate environment, and can better detect processes that result in weak, multilocus molecular signatures relative to univariate tests.

RDA can be used on both individual and population-based sampling designs. The distinction between the two may not be straightforward in all cases. A simple guideline would be to use an individual-based framework when you have individual coordinates for most of your samples, and the resolution of your environmental data would allow for a sampling of environmental conditions across the site/study area. For population-level data, you would input the genetic data as allele frequencies within demes.

The code to run the RDA is simple. RDA runs relatively quickly on most data sets, though you may need a high memory node on a cluster for very large data sets (i.e., millions of SNPs).

We will also import climate data from the World Climate database (covers 1970-2000). This is not complete for all our sampling locations, so we will reduce our individuals (down to 225) to those with climate data. We also have highly collinear predictors and only test three of the four we downloaded.

1. Which environmental predictor was thrown out?

Once you’ve run the RDA, note that we will have as many constrained (“RDA”) axes as we have predictors in the model. All residual variance is then modeled by PCA (the unconstrained “PC” axes). 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.

1. How much variation is explained by the environmental predictors? (hint- adjusted.r.squared)
2. Explain how the wrasse populations (colored based on collection sites) are distributed in relation to the environmental predictors (red vectors). For example, what drives the split between the Southern (Kungsbacka, Stomstad and Kristiansand) and Western populations (Stavanger and Austevoll)? How about The Southern population from Flatanger?