**Case Study 5: Haplotype Networks and Redundancy Analysis**

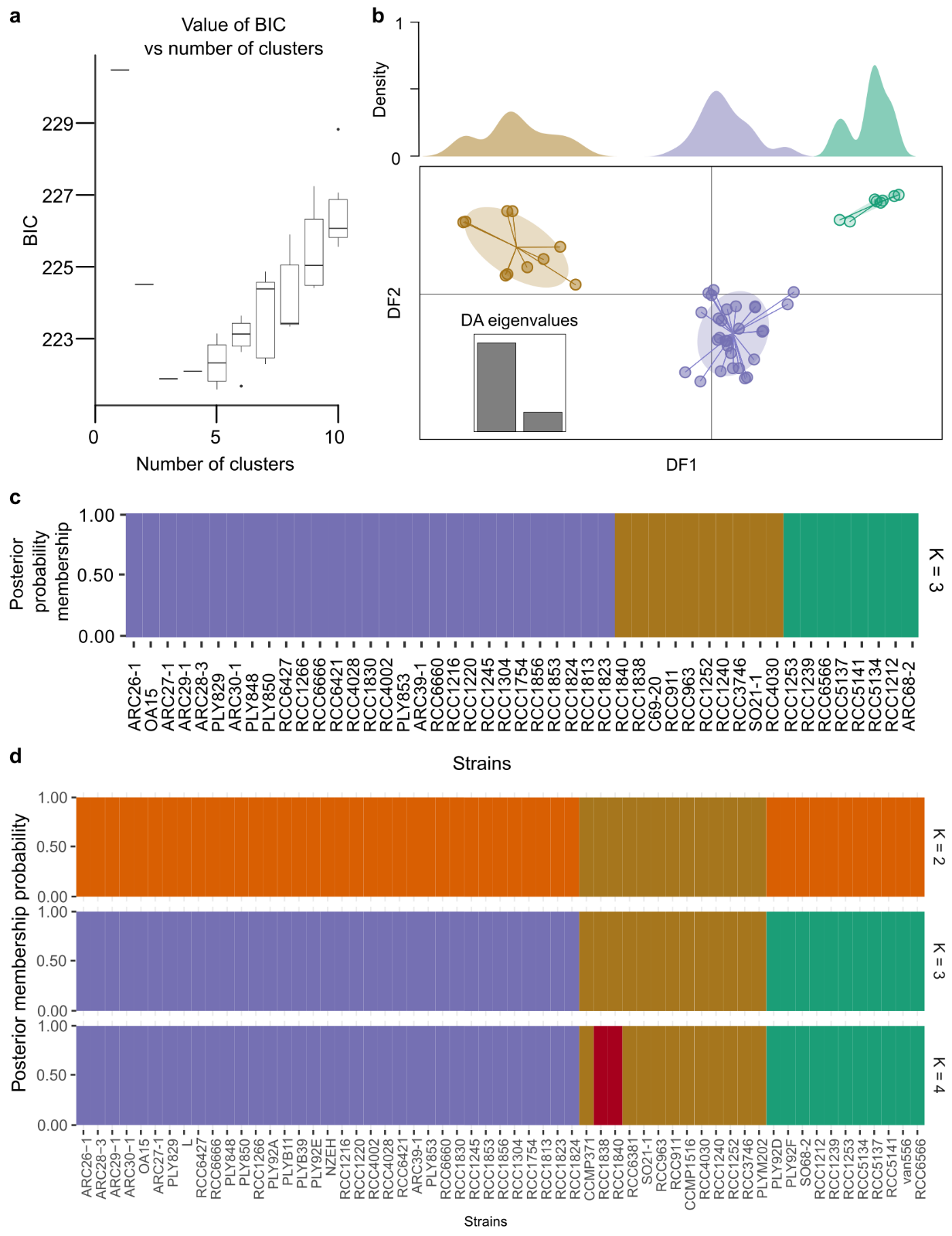
**Part 1: Haplotype Networks**

What is a ***haplotype***? It is a unique sequence of nucleotide bases over a region of the genome. Individuals that share the exact same sequence for this entire region share a haplotype, and are therefore inferred to be closely related to one another. The longer the piece of the genome, the lower the chances of individuals being identical for the full sequence (= more haplotypes).

Haplotypes can be used to compare individuals within and between populations to infer relatedness and recent evolutionary history. A ***haplotype network*** is a visualization of the genotypic relationships among observed haplotypes within a population or species. It is a common tool in phylogeography. In the networks, the circles represent the haplotypes and the size of those circles is proportional to the number of individuals with that observed haplotype in the sampled population, and mutational steps are symbolized by dashes.

Today we are going to use both mtDNA and cpDNA to see if we can recover the genetic groups (A1(b-d), A2 and B) identified using nuclear markers (SNPs) in this week’s paper:

Bendif, El Mahdi, et al. "Rapid diversification underlying the global dominance of a cosmopolitan phytoplankton." The ISME Journal 17.4 (2023): 630-640.



**Figure 1.** **Discriminant analysis in principal component (DAPC) based on 2,086,643 SNPs from 59 genomes.** Copied from Supplement of Bendif et al. 2023, ISME Journal.

To do this, I extracted mitochondrial cytochrome oxidase subunit 1 (*cox1* or *COI*) and plastid elongation factor tu (*tufA*) from 17 strains on NCBI. These strains are listed below with their geographic source and morphotype, if known.

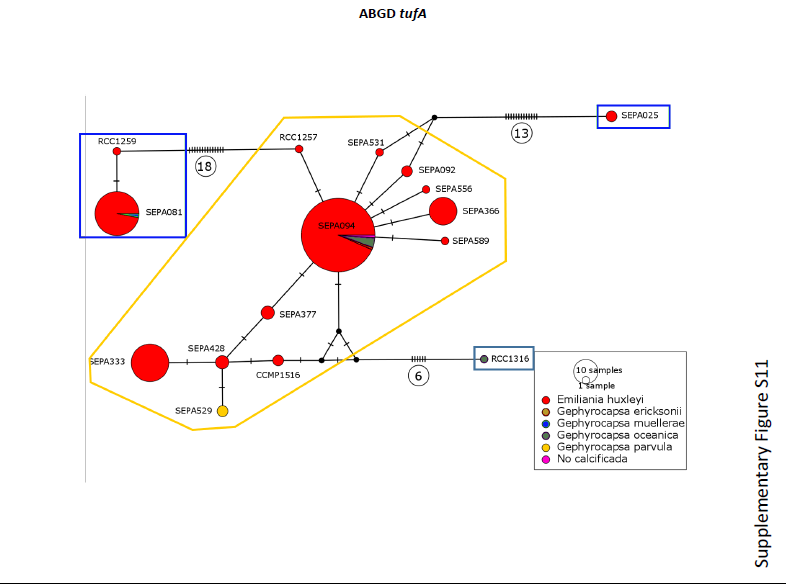
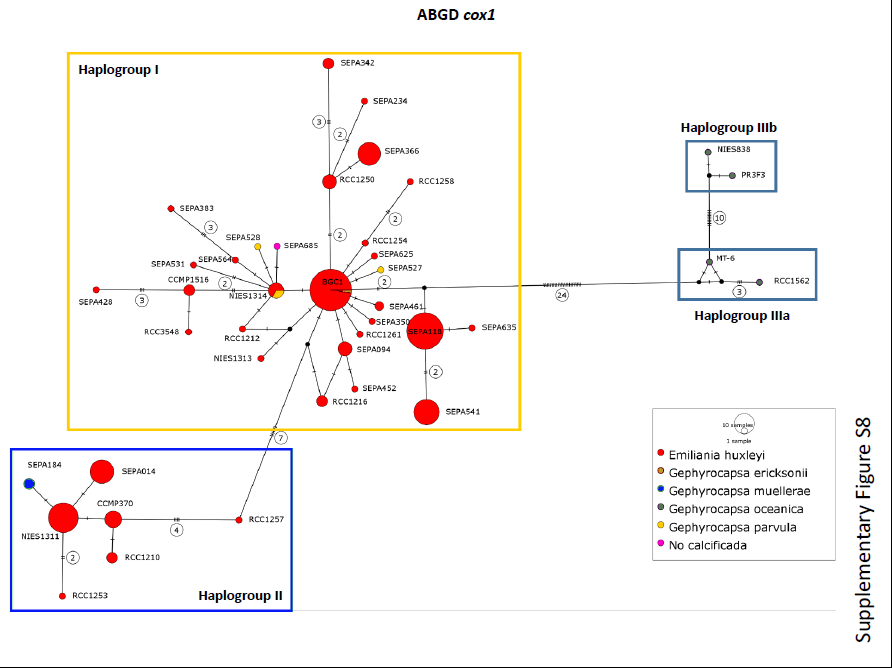
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Strain ID** | **Region/Ocean** | **Morphogroup** | **Morphotype** | **Nuclear Cluster Assignment** |
| Van556 | Pacific Ocean |  |  | B |
| RCC1253 | Japanese coast | B | O | B |
| RCC4030 | South East Pacific | A | A | A2 |
| PLYM217/CCMP1516 | Pacific Ocean | A | A | A2 |
| PLYM202 | South Pacific | A | A | A2 |
| CCMP371 | Sargasso Sea | A | A | A2 |
| RCC4028 | South East Pacific | A | A overcalcified | A1b |
| RCC4002 | South East Pacific | A | A overcalcified | A1b |
| RCC1216 | Tasman Sea | A | R | A1b |
| PLYM219/NZEH | South Pacific | A | R | A1b |
| PLY92A | English Channel | A | A | A1c |
| L | Oslo Fjord | A | A | A1d |
| CS-369 | Tasman Sea |  |  | Not tested |
| CCMP3266 | Tasman Sea | A | R | Not tested |
| BOF92 | North Atlantic |  |  | Not tested |
| RCC174 | English Channel | B | B | Not tested |
| RCC1217 | Tasman Sea |  |  | Not tested |

We will first make a haplotype network using the cox1 gene in R. Login into you VPN and then go to the RStudio interactive app on the HPRC grace portal. Set your working directory. Then, copy over today’s activity

> file.copy(“/scratch/group/kitchen-group/09\_phylogeo/week9\_netTrees.R”, “.”)

Open the R script on your interactive session. And begin.

1. How many segregating sites are there between the 17 strains? What does that tell you about *cox1* as a marker gene?
2. How many cox1 haplotypes were found?
3. Compare your results to the following figures (<https://www.frontiersin.org/articles/10.3389/fmars.2021.785763/full>) :



**Figure 2. Median joining haplotype network based on alignment of partial *cox1* sequences (n=205 strains) and *tufA* chloroplast sequences (n=82 strains).** Copied from von Dassow et al. 2021, Frontiers in Marine Science.

In the figures above, ***haplogroups*** are shown in boxes after running Automatic Barcode Gap Discovery (ABGD) that sorts the sequences into hypothetical species based on the barcode gap, which can be observed whenever the divergence among organisms belonging to the same species is smaller than divergence among organisms from different species. Unfortunately, I could not find this tool. Let’s just believe their results, I guess.

1. In the table below, fill in each sample’s mitochondrial haplotype (***hint***- look at merged table):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Strain ID** | **Morphogroup** | **Nuclear Cluster Assignment** | **Mito haplotype** | **Plastid haplotype** |
| Van556 |  | B |  |  |
| RCC1253 | B | B |  |  |
| RCC4030 | A | A2 |  |  |
| PLYM217/CCMP1516 | A | A2 |  |  |
| PLYM202 | A | A2 |  |  |
| CCMP371 | A | A2 |  |  |
| RCC4028 | A | A1b |  |  |
| RCC4002 | A | A1b |  |  |
| RCC1216 | A | A1b |  |  |
| PLYM219/NZEH | A | A1b |  |  |
| PLY92A | A | A1c |  |  |
| L | A | A1d |  |  |
| CS-369 |  | Not tested |  |  |
| CCMP3266 | A | Not tested |  |  |
| BOF92 |  | Not tested |  |  |
| RCC174 | B | Not tested |  |  |
| RCC1217 |  | Not tested |  |  |

1. Now repeat the process with the *tufA* data and fill in the corresponding column above.
2. Copy over the two NJ trees in the space below and annotate them according to the genetic cluster assignments. Can you start to see ***haplogroups***?
3. In the space below summarize the results of the nuclear, mtDNA and cpDNA results on species delimitation in this activity and the two published papers. ***Reminder*** *that we only looked at 1 mtDNA and cpDNA gene here and a subset of strains.*

***Part 2: 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 covary 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 return to the wrasse hybridization dataset.

Return to R to load in the VCF from Case Study 4 (07\_hybrids). 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?

Note this activity was heavily copied from: <https://popgen.nescent.org/2018-03-27_RDA_GEA.html>

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?