**Case Study 3: Population Structure and Haplotype Networks**

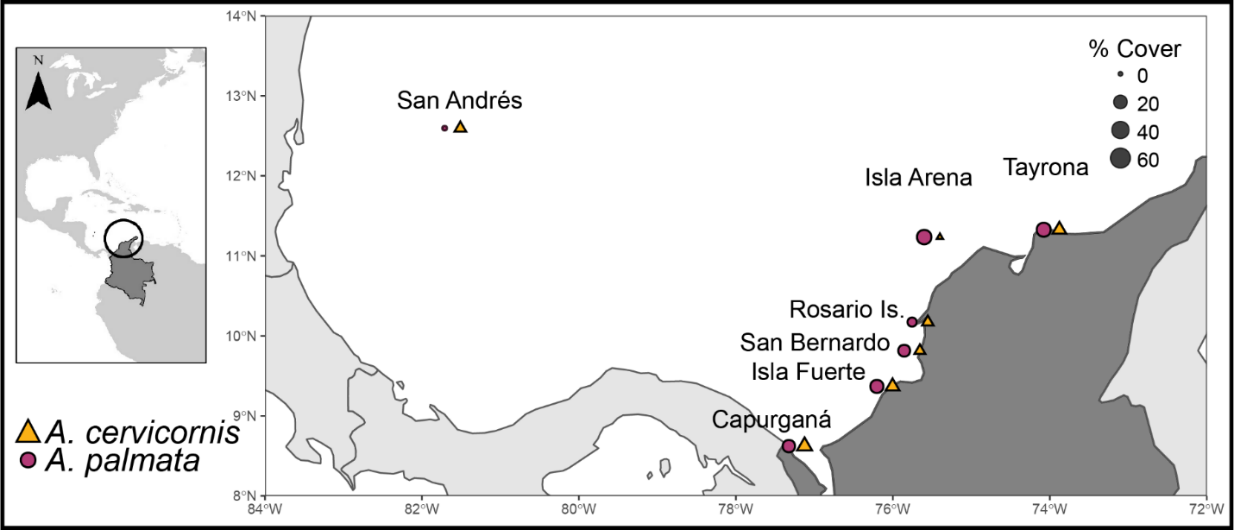
**Part I: *A. cervicornis* population structure follow-up questions**

Start an interactive Rstudio session on the HPRC (Grace Portal) as you did last week. When you log in, your prior files (Rmarkdown and R script) will still be loaded in the top left pane. Re-load the R libraries required for last week’s Rmarkdown file and pick up with the *A. cervicornis* population structure section. Recall, you may need to you’re your working directory:

> setwd("/scratch/group/kitchen-group/MARB\_689\_Molecular\_Ecology/class\_working\_directories/kitchens/03\_geneFlow/")

Answer the following questions **after** completing those sections of R code.

1. From the population structure analysis, what population does the *A. cervicornis* samples from San Andres belong to? What other countries/regions form a genetic cluster with San Andres? And, what side of the Caribbean break (East or West) does it fall out with?
2. In the space below, draw on the Panama-Colombia Gyre and Caribbean Current on the sample map. **Explain** how these ocean currents contribute to the gene flow and population structure observed in the *A. cervicornis* samples in the space under the map.



Note: Capurgana = Urba Gulf

1. García-Urueña & Garzón-Machado 2020 reported an overall decrease in *A. cervicornis* coverage at all Colombia sites with limited to no evidence of recruitment (i.e., babies). When encountered, *A. cervicornis* reefs were often isolated patches of relict colonies. Given our population structure results, what sort of restoration strategy might you suggest to ***enhance A. cervicornis coral coverage***?
2. What strategy would you suggest to ***enhance A. cervicornis genetic diversity***? (hint- this strategy is probably different than the one above).
3. From the AMOVA test, how much allelic variation is found among populations? Was this result significant?

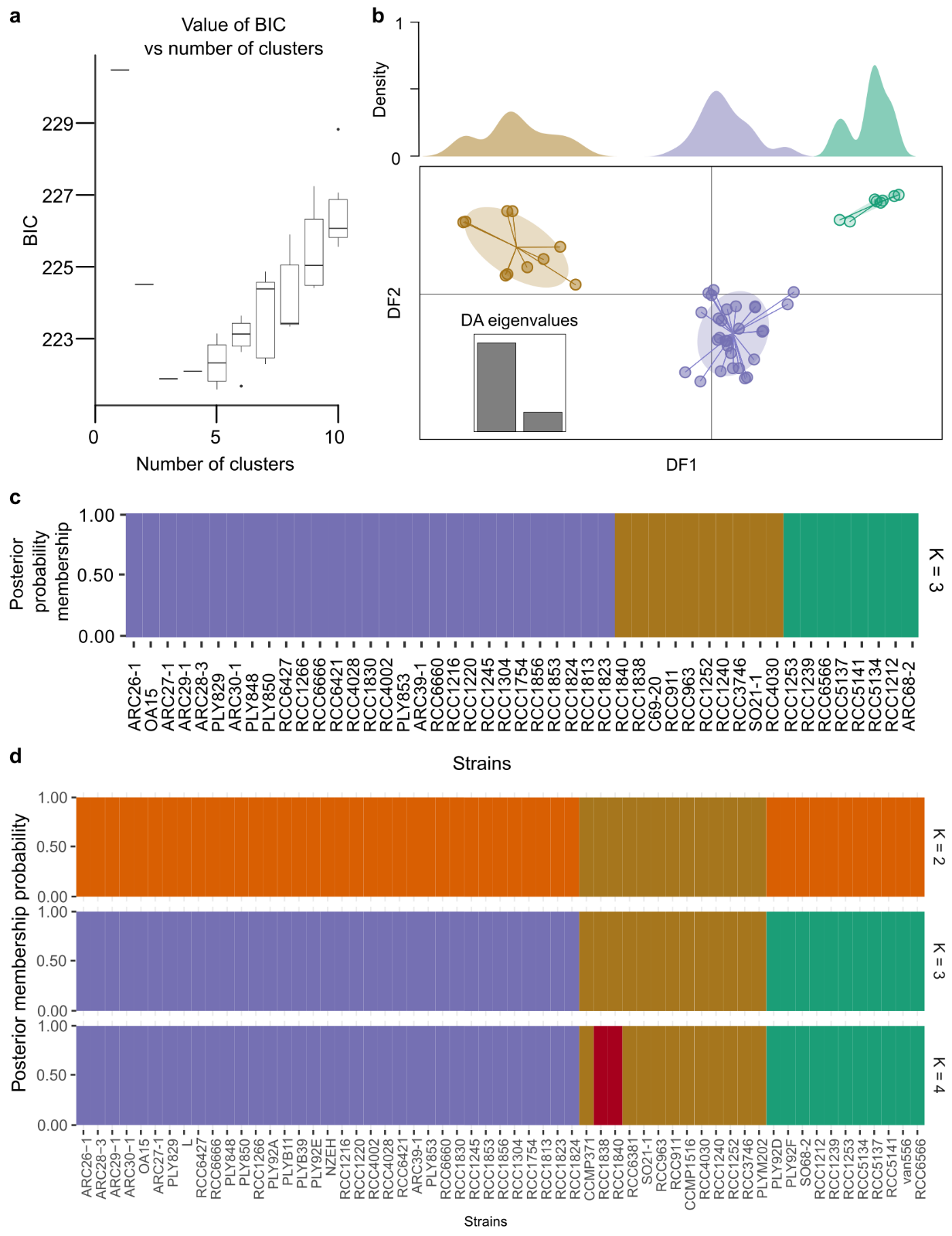
**Part II: Haplotype Networks**

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. Set your working directory and make a new subdirectory as follows:

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

> dir.create(file.path("06\_phyloGeo"))

> setwd("/scratch/group/kitchen-group/MARB\_689\_Molecular\_Ecology/class\_working\_directories/kitchens/06\_phyloGeo")

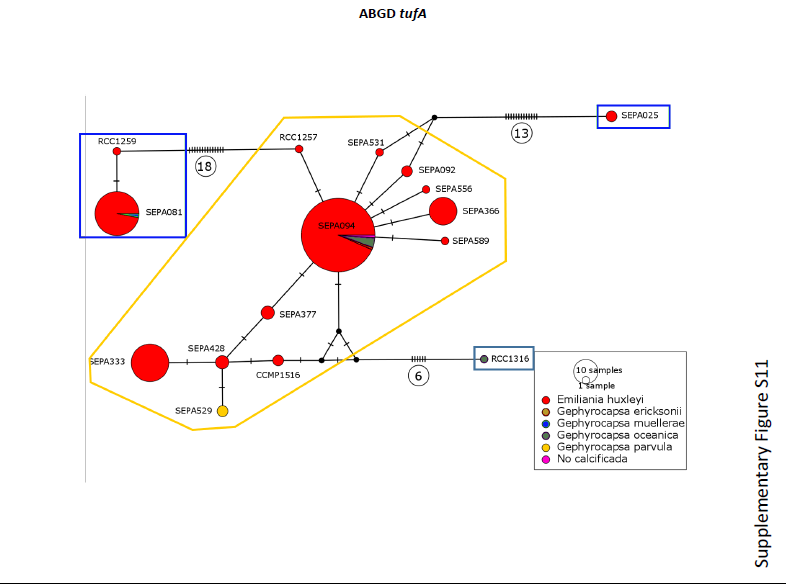
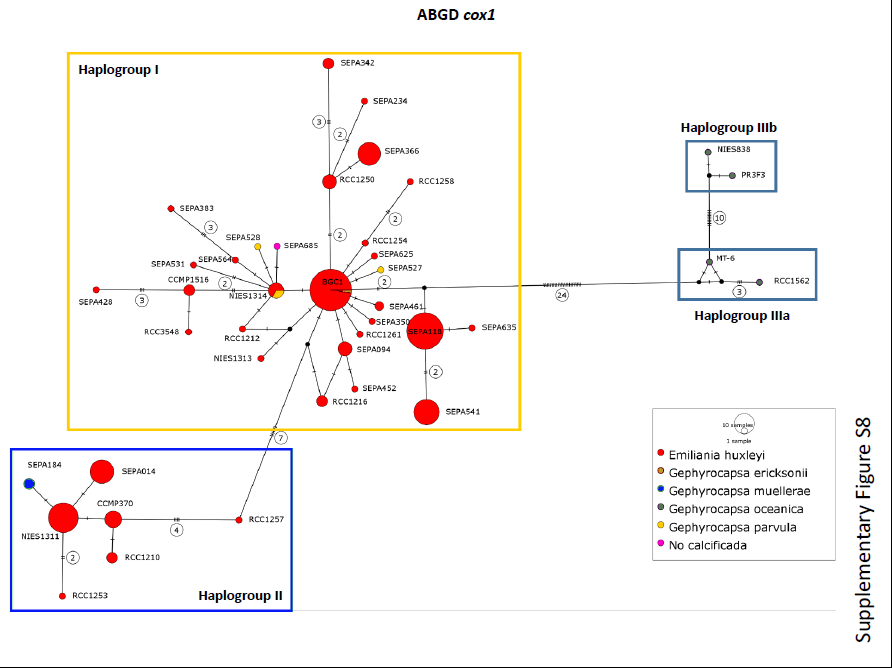
Then, copy over today’s activity:

> file.copy("/scratch/group/kitchen-group/MARB\_689\_Molecular\_Ecology/06\_phylogeo/week6\_netTrees.R", ".")

Open up the R script using the File > Open File on the top bar.

(Re)Load the libraries listed on the top.

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.

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.*