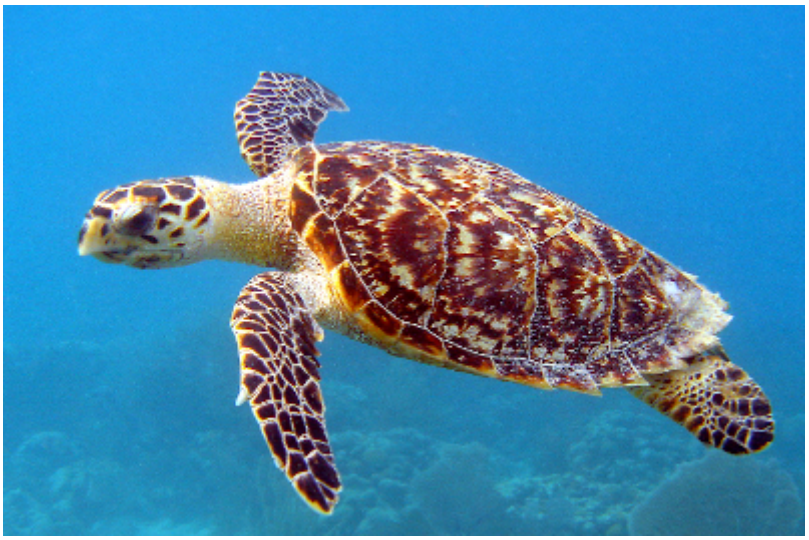
Practical on population subdivision and demographic inferences

You are interested in inferring the history of a population of sea turtles.



You collected some DNA data from 40 (diploid) individuals from 4 different locations (A, B, C, D). You have 10 individuals for each location. For each sample, you have access to 2000 SNPs (not base pairs, only the polymorphic sites). The data on alleles/haplotypes is stored in `turtle.csv` (80 rows) while the data on genotypes (40 rows) is stored in `turtle.genotypes.csv`. Genotypes are encoded as 0/1/2 as homozygous for the ancestral, heterozygous, homozygous for the derived allele, respectively. Individuals for each location are ordered in all files (i.e. first 10 entries on the genotype data are for location A, the second 10 entries are for B, etc etc).

1. Test whether there is population subdivision in this sample and, if so, at which extent.
2. Assess whether there has been isolation by distance in this species, knowing that the geographical distance of each population from a putative origin is: A (5 km), B (10 km), C (12 km), D (50 km).

Some considerations for each task.

1. There are several ways of doing it. (a) You may want to infer a tree from genetic distances. In R you can use `dist` to calculate eucledian distances between individuals based on their genotypes. From the resulting distance matrix you can infer a tree using `tree <- hclust(distance_matrix)` and then plot it with `plot(tree)`. (b) You can do a Principal Component Analysis. In R you can use `pca <- prcomp(data ...)` to calculate principal components which will be stored in `pca$x`. You can plot the first two/three components afterwards. For this analysis it is better to consider only SNPs with an allele frequency of at least 0.05. (c) You can calculate average FST across all SNPs for each pair of subpopulation and see whether it is larger than zero, and whether it changes for each comparison. My suggestion would be to start with point (c) as discussed during the lecture.
2. You can test whether genetic distance correlated with geographical distance.

Extra bonus question: assuming that we have access to reference information from 3 known subpopulations in the area, how would you perform an admixture analysis in this sample? Allele frequency for a set of markers for 3 subpopulations are stored in `turtle_markers.csv`, with the first column indicating the genomic position of the marker and the other columns showing the derived allele frequency at each subpopulation. How would you do that?