Sperm Whale population genetics summary for discussion

2024-09-06

Contents

sperm whale summary **1**
summary of results: . . . 1
Data summary . . . 1
relatedness . . . 1
Population structure . . . 2
Population structure conclusion . . . 7
genetic diversity . . . 7

Mitochondrial data **9**

sperm whale summary

The goal of this analysis was to determine if there is population structure between sperm whale populations across the gulf southern Atlantic coast

summary of results:

1. Related individuals
   - There were closely related individuals in the data ranging from full sibs, half-sibs, and cousins (or similar).
   - The siblings tended to be in the same geographic area, but not exclusively so.
2. Lack of population structure:
   - There is no evidence for population structure in the nuclear genome.
   - However, the mitochondial data suggest structure across populations.
   - This makes sense given the lack of dispersal of females.
3. Low genetic diversity
   - There is relatively low genetic diversity for the nuclear data.
   - Mitochondial diversity is very low. This is likely due to a bottleneck ~125k years ago.
   - Despite this, inbreeding is not high.

Data summary

The data consist of 73 sperm whale samples from across the Gulf and southern East Coast. We had 4,441 single nucleotide polymorphisms that were of high quality across all of the samples (data generated using NextRad). See the figure below for sample locations.

relatedness

I estimated relatedness between individuals and found related individuals ranging from full siblings to cousins.

- Full siblings: 4 pairs
- Half siblings: 8 pairs
- Cousins: 4 pairs

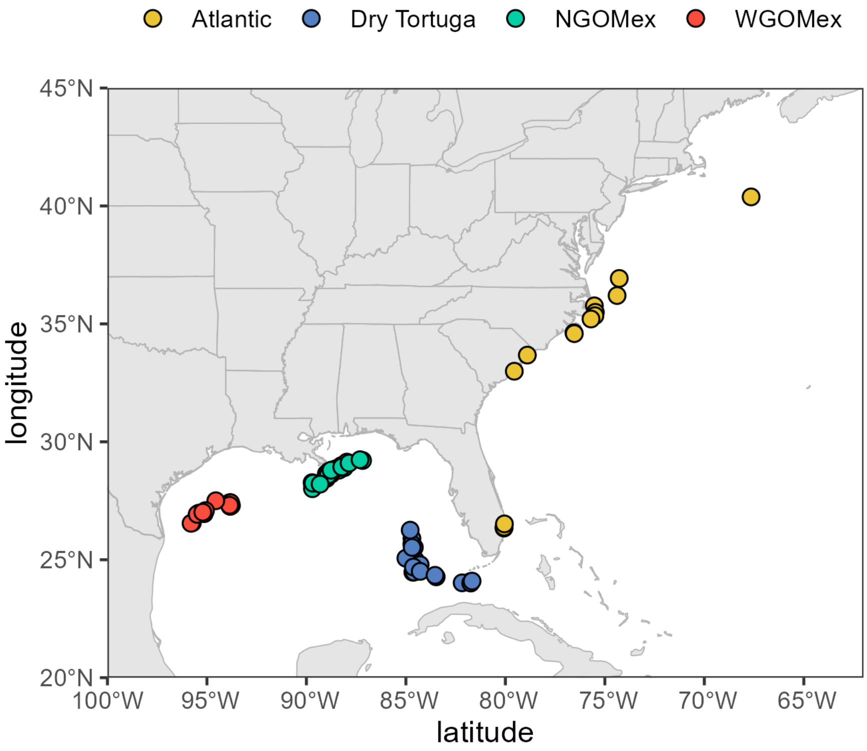


Figure 1: Sample locations colored by putative population

Related individuals were generally found within a region or between adjacent regions. See the figure below showing pairwise estimates of relatedness (theta). The expectations for theta are:

- full sib: 0.25
- half sib: 0.125
- 3rd degree (cousins, and other more complex relationships): 0.065.
- Beyond 3rd degree is harder to detect and usually our detection limit.

All the full siblings were collected at the same time and location, so were probably the same family group.

Population structure

We can assess population structure in 4 main ways:

1. Principal component analysis
   - Here, we take a hypothesis free approach. The basic idea is to summarize the variation across all SNPs where the first axis explains the largest amount of variation, then second axis the second largest amount of variation, and so on.
   - When population structure is present individuals will cluster separately in the plot.
2. Admixture analysis
   - Here we're trying to estimate ancestry of populations given our data. In short we're trying to cluster individuals into groups that share genetic variation.
   - We can then identify the most likely number of populations from this analysis.
3. Discriminant analysis of principal components (DAPC)
   - DAPC take a set of predictor variables and tries to identify combinations that maximize differences between individuals from some predifined groups.
   - Basically, we're asking if we can identify consistent differences between populations.

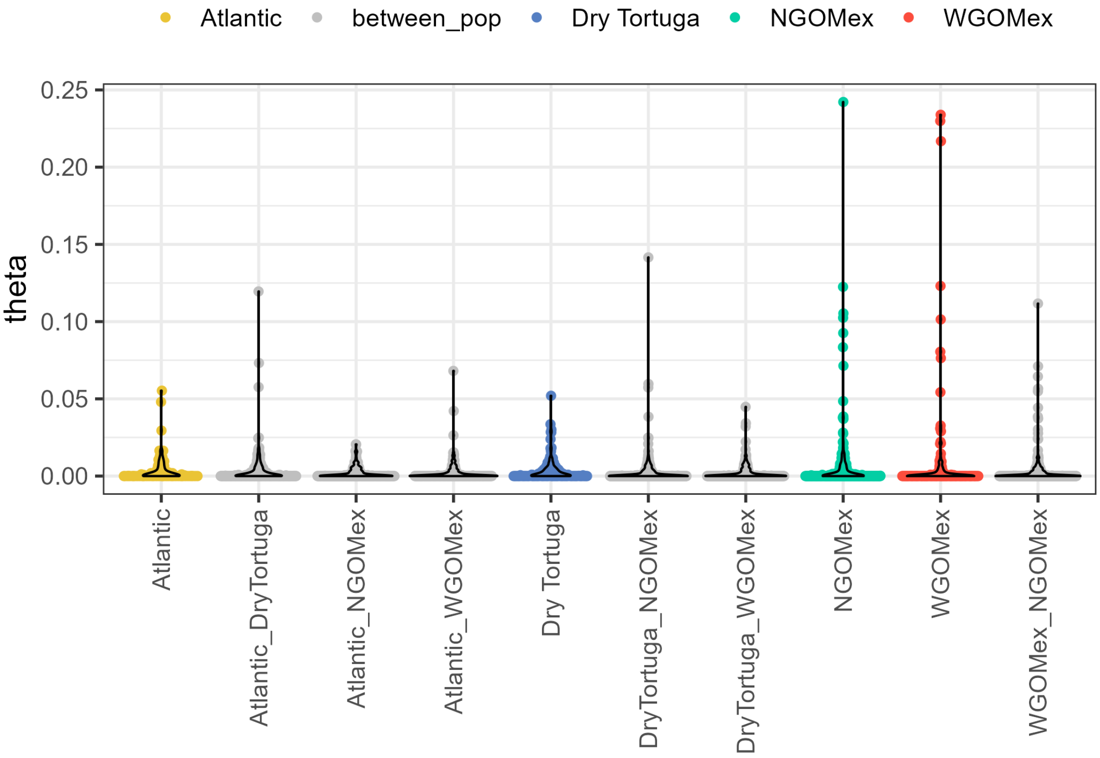


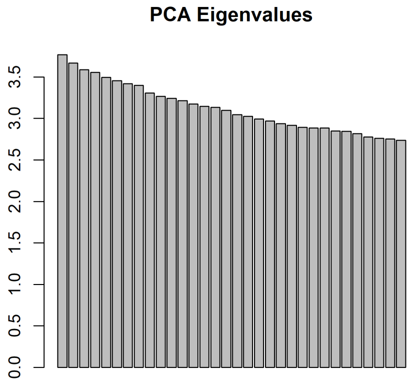
Figure 2: Pairwise relatedness estimates. Notice that most highly related individuals are within a population. Grey points are between population comparisons, colored points within population.

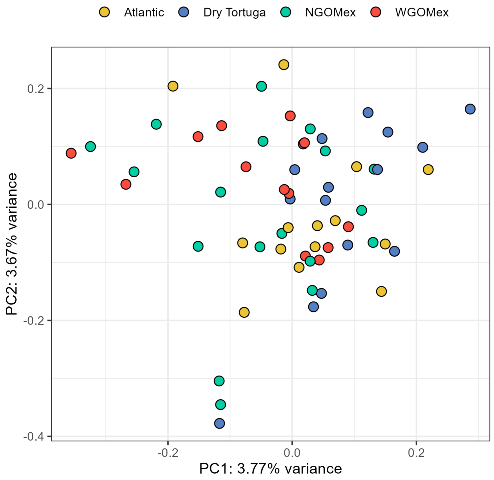
   - This method relies on using PCA to reduce the dimensionality of the data and then uses these PCAs to discriminate between groups.
   - Therefore, we need the PCA to explain the variation in the data well for it to be able to discriminate between populations.
4. Fst
   - This is a measure of allelic differentiation between groups.
   - A value of 1 are populations that share no variation at all and 0 are populations that are the same.
   - In practice values of 1 are not seen genome-wide (even between species)
   - For general reference (but rules are challenging and presence of population structure depends on many factors)
     - 0-0.05: no/weak differentiation
     - 0.05-0.15: moderate differentiation
     - 0.15-0.25: strong differentiation
     - 0.25 and >: very very, strong differentiation. Possibly different species.

PCA

When population structure is present in data we expect we expect to see a "broken hockey stick" when structure is present in the PCA eigen values (left figure, below). That is, the first few PCs explain large amounts of variation and the proportion of variation explained drops off abruptly. We also expect that populations would form distinct clusters in the PCA plot.

We see no evidence for population strucutre. There is no broken hockey stick (below, left) and no clustering in the PCA (below, right).





Admixture

First we can look at the cross validation (CV) error estimates, to determine our most likely number of populations. Note that this is blind to the actual population of samples. The lowest CV error is the most likely number of populations (that is, it gives accurate and stable results). Here, most the most likely K is 1. and there is no evidence of population structure.

Here are the actual results as well, just to see. In part because sometimes it is hard to determine the best number of populations, but it is still clear there is structure in the data. Each bar is an individual and the color of the bar indicates the assigned ancestry ("K". the color of the ancestries don't mean anything).

Individuals within a population are basically randomly being assigned to various ancestries. This suggests that the CV scores are accurate and there is no structure. If structure was present, we would expect individuals within a region to have similar ancestry.

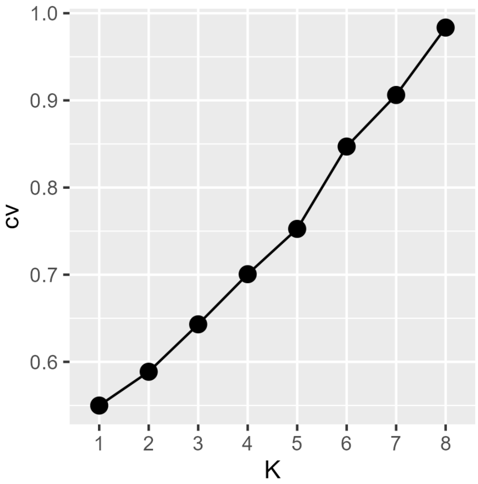


Figure 3: CV scores to identify the number of populations. Lowest CV score indicates the most likely number of populations

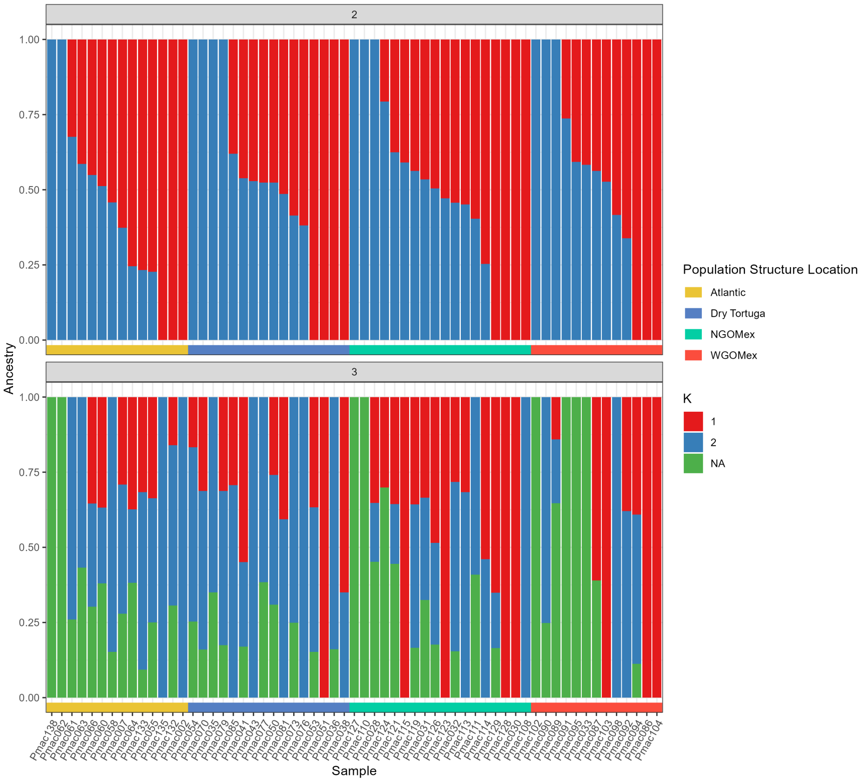
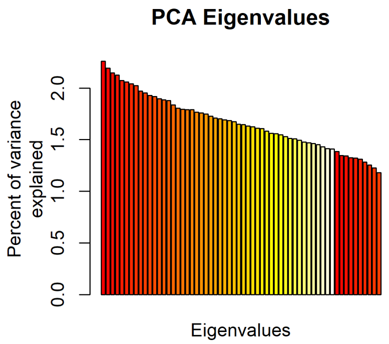


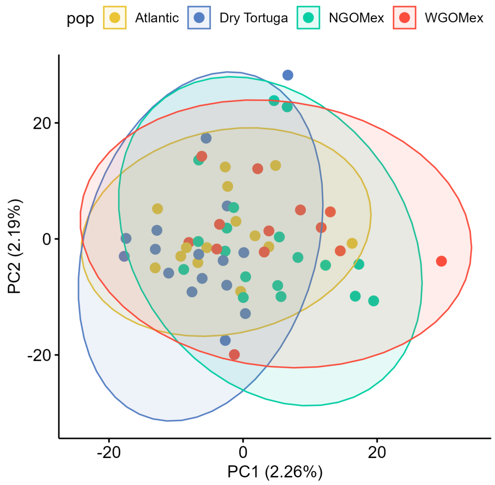
Figure 4: Admixture plot for K=2 and K=3. Individuals are ordered by their ancestry in K=2 within populations.

dapc

Next we will run DAPC. This is nice to run because we are able to quantify if we can accurately assign individuals to populations.

Remember that for this to work well we need a principal components that explain vairation in the data. In our case, our PC's do not. Below are the same plots from the PCA section above, but calculated using a different approach (just to make sure our method isn't causing problems).





Nevertheless, let's run the DAPC. There are two main considerations:

1. How many populations to consider?
2. How many PCs to include. Too many risks overfitting the model.

I'm going to skip over the details of these issues and go right to the results.

I ran the DAPC in two ways:

1. Assuming two populations
2. Assuming four populations.

Within each of these I used two different methods to determine how many principal components to included (i.e., point 2 above).

The accuracy of our DAPC in assigning individuals to populations was determined using training–testing partitioning.

70% training- 30% test. Repeat 100x for each DAPC model.

- A good model would be somewhere around 90% accuracy.
- Random would be 25% for the 4 population set. 50% for the 2 population set.

k	num_PCs	mean_accuracy	sdev
2	1	0.453	0.103
2	5	0.547	0.127
4	3	0.231	0.0818
4	17	0.278	0.104

These results suggest that the DAPC assignments are basically random and have nearly no predictive power for population origin.

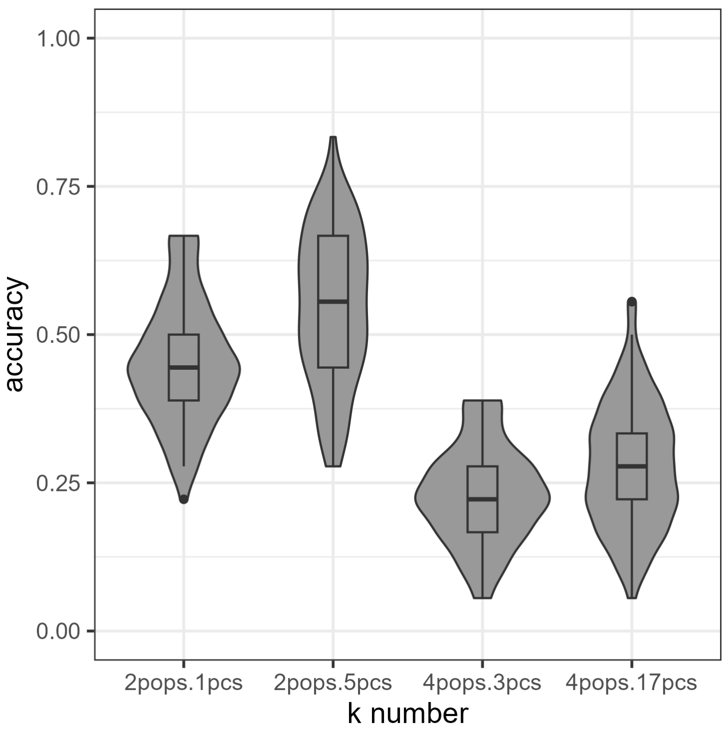


Figure 5: DAPC accuracy with 2 and 4 populations and the K-1 method vs. cross validation method for choosing PCs

Fst

Remeber the categories for Fst:

- 0-0.05: no/weak differentiation
- 0.05-0.15: moderate differentiation
- 0.15-0.25: strong differentiation
- 0.25 and >: very very, strong differentiation. Possibly different species.

See figure 6, below: find weak/no differentiation between populations.

Population structure conclusion

From the RADseq data, there is no/little evidence of population structure. PCA, Admixture, DAPC, and Fst all support this with none of them having any predictive power or giving typical signals of structure in the data.

genetic diversity

Estimates of genetic diversity were estimated a few ways:

1. Tajima's pi
2. Observed heterozygosity
3. Expected heterozygosity

Tajima's pi is the pairwise genetic diversity tells us about the overall level of genetic diversity within the population. Figure 2 here shows some typical levels for various species. Mammals are typically 0.001-0.01.

Overall, inconsistent and minor differences.

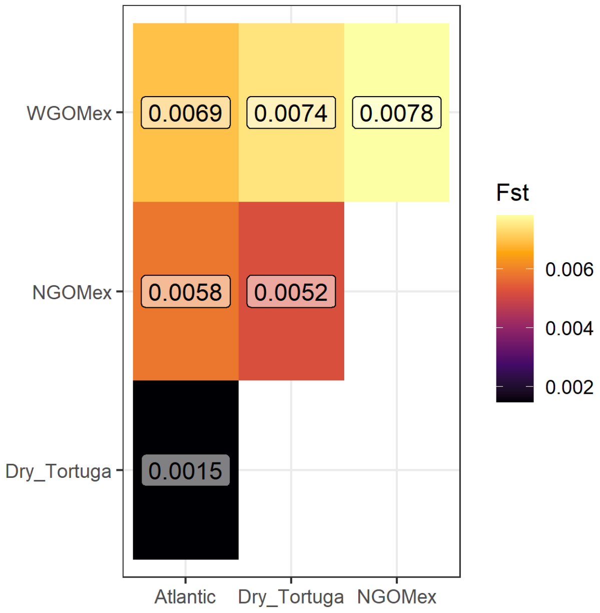
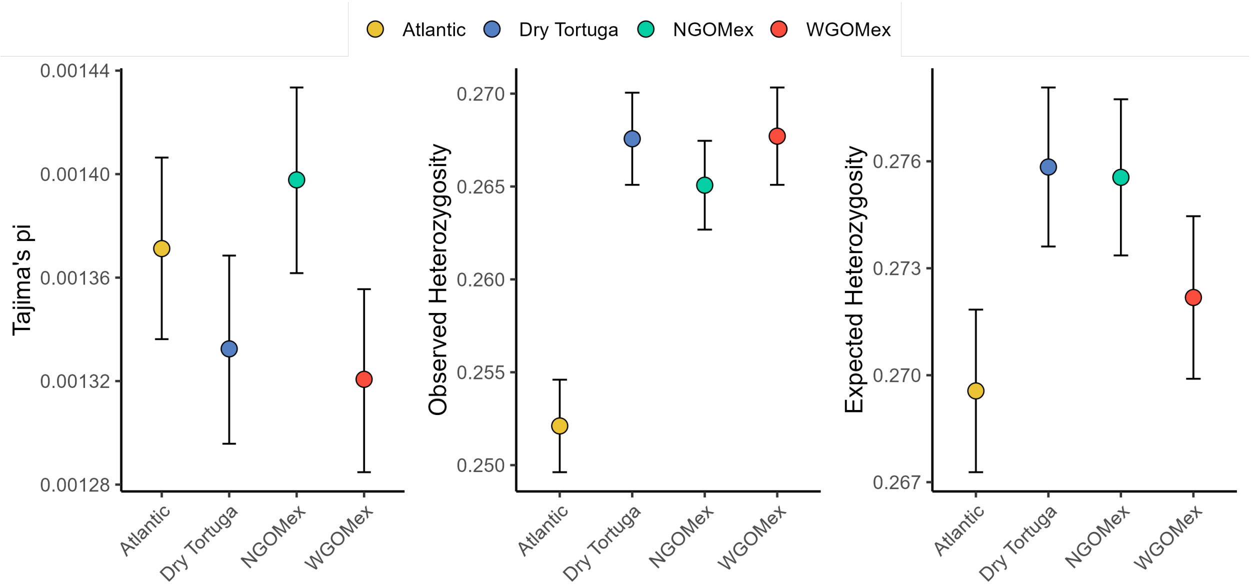


Figure 6: Pairwise Fst

- Pi: WGOMex has lower than NGOMex.
- Observed heterozygosity: Atlantic lower than all other populations.
- Expected heterozygosity: Atlantic lower than Dry Tortuga and NGOMex

But it is likely not justified to make these population splits as there isn't population structure.



However, inbreeding is not high in any population:

- Atlantic: 0.1017
- Dry_Tortuga: 0.0629
- NGOMex: 0.0653
- WGOMex: 0.0572

Effective population size

Estimated via CurrentNe:

- Ne point estimate: 123.75
- Lower bound of the 90% CI: 100.51
- Upper bound of the 90% CI: 152.37

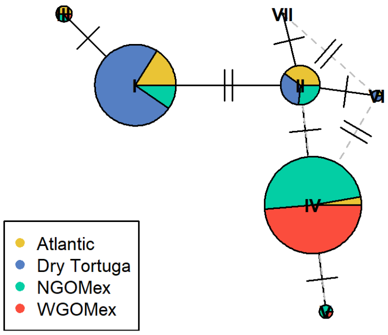
Accounting for sib relationships:

- Ne point estimate: 136.53
- Lower bound of the 90% CI: 109.98
- Upper bound of the 90% CI: 169.49

Mitochondrial data

Despite lack of structure in the RAD data, there does appear to be structure in the mitochondrial haplotypes.

Overall, mitochondrial diversity is very low: `0.00157`, haplotype diversity is high: `0.7388`. These values are very similar to other populations of sperm whales in the Gulf of California. This signal of low nucleotide diversity with higher haplotype diversity is indicative of a bottleneck and expansion. This paper from Morin et al found evidence for a bottleneck 125,000 years ago and expanded out from a Pacific refuge to re-colonize the Atlantic. From the same study, global Mitogenome nucleotide diversity is 0.093% and haplotype diversity is 0.975.


III
VII
I
II
VI
IV
V
● Atlantic
● Dry Tortuga
● NGOMex
● WGOMex