

Complex population structure revealed using genomic data from Tamanend's bottlenose dolphins (*Tursiops erebennus*)

Marine Biology

Nicole L. Vollmer*, Lynsey A. Wilcox Talbot, Antoinette M. Gorgone, Aleta Hohn, Andrew J. Read, Eric S. Zolman, Patricia E. Rosel

*Corresponding Author: nicole.vollmer@noaa.gov

Cooperative Institute for Marine and Atmospheric Studies, Rosenstiel School of Marine, Atmospheric, and Earth Science, University of Miami, Miami, FL, USA

ONLINE RESOURCE 1

Description of Current Stock Delineations in North Carolina

According to NOAA's most current stock assessment reports, the NNCES stock occurs primarily within the Pamlico Sound estuarine system during the warmer months of July-August, and the SNCES stock primarily within estuarine and nearshore waters (≤ 3 km from shore) between the Little River Inlet (near the North/South Carolina border) and the New River in North Carolina during the colder months of January-February (Hayes et al. 2021). These two estuarine stocks overlap in central North Carolina around Cape Lookout during July-August. The NM stock occupies coastal waters from Long Island, New York to Assateague, Virginia during July-August, and reaching its southern limit of Cape Lookout, North Carolina only during winter. The SM stock ranges from northern Florida to Cape Lookout, North Carolina during fall and winter months and as far north as Assateague, Virginia during the warm months of July-August. The delimitation of both the NM and SM stocks, for management purposes, is between the shoreline and 20 m isobath north of Cape Hatteras and between the shoreline and 200 m isobath south of Cape Hatteras (Hayes et al. 2021). These two coastal migratory stocks may overlap during spring and fall in northern North Carolina to southern Virginia, however a spatial shift in the distribution of the SM stock in summer 2004 suggests the northern boundary may differ yearly (Hayes et al. 2021). Adding another level of complexity, overlap between these coastal and estuarine stocks also occurs during certain seasons in waters ≤ 3 or ≤ 1 km from shore (depending on season and stock). See Online Resource 3 Figure A for maps of the seasonal distributions of the stocks discussed above.

Choosing Samples to Represent Current Stocks

For samples collected in North Carolina and South Carolina, a fishnet grid was created in ArcGIS v10.7.1 setting each grid to 3.7800 km wide x 3.3085 km tall. Per grid cell: one sample was chosen for those cells in which there were < 5 samples, two samples were chosen if ≥ 5 samples were inside the cell, and to the extent possible no two samples collected on the same day were used. For all samples, those with the best DNA quality/quantity and/or extracted using a DNeasy Blood and Tissue Kit (Qiagen) were prioritized based on SNPsaurus recommendations. When available, those collected during summer months (June, July, and August) were chosen to ensure the highest likelihood of including representatives of the SM, NM, NNCES, and SNCES stocks. Samples from South Carolina coastal waters (presumed to be part of the SC/GA stock) were included to examine if any of the estuarine or coastal populations in North Carolina had extended ranges further south, or vice versa, as little is understood regarding the northern

boundary of the SC/GA stock (Hayes et al. 2018). Only samples of the coastal ecotype were used to avoid including individuals from the Western North Atlantic Offshore Stock.

Genomic DNA Library Preparation, Sequencing, Reference Assembly, and Mapping

Genotyping-by-sequencing libraries were created by fragmenting genomic DNA and ligating short adapter sequences to the ends of fragments with Nextera DNA Flex reagent (Illumina, Inc). The Nextera reaction was scaled for fragmenting 50 ng of genomic DNA, although 75 ng of genomic DNA was used to compensate for DNA degradation in the samples and to increase fragment sizes. Next, fragmented DNA was amplified for 27 cycles at 74°C, with one primer matching the adapter and extending 10 nucleotides into the genomic DNA with the selective sequence GTGTAGAGCC. Therefore, only fragments starting with a sequence that can be hybridized by the selective sequence of the primer will be amplified efficiently. The nextRAD libraries were then sequenced on a HiSeq 4000 creating 150 bp single-end reads. Adaptor trimming was conducted in in BBDuk v38.87 (BBMap tools, <http://sourceforge.net/projects/bbmap/>) using the following options: ktrim=r k=17 hdist=1 mink=8 ref=nextera.fa.gz minlen=100 ow=t qtrim=r trimq=20. Trimmed reads were mapped to the *T. truncatus* reference genome (GenBank assembly GCA_011762595.1) which was first indexed before aligning reads using BWA v0.7.17-r1188 (Li 2013; BWA-MEM with default parameters was used for the alignment). The output files were converted into BAM format using SAMtools v1.11 (Li et al. 2009). After merging the three bam files for each sample, SAMtools was used to assess mapping quality for each merged bam file (i.e., sample) to the reference genome. To calculate the average read depth at each position for each sample, a custom script using “samtools depth” was used. Read depths were then averaged across all samples to get the overall average read depth for the dataset. To calculate alignment information for each sample, a custom script using “samtools flagstat” was used. From each flagstat output file produced, the percentage of reads mapped for each sample was collated and averaged together to get the mean mapping rate for the dataset. Custom scripts are available on Zenodo: <https://doi.org/10.5281/zenodo.14919600>

SNP Calling and Filtering

In Stacks v2.5.4 (Catchen et al. 2013), the “gstacks” module was run with default parameters, and all samples were assigned to the same group in the population map because it was unclear how many genetically distinct populations were present in the dataset. Next, in the “populations” module, to minimize the effect of linkage disequilibrium between loci, only the first SNP per locus was analyzed using --write-single-snp, and --ordered-export was used to make sure that if loci overlap in the genome only one was called.

A total of 871,890,600 raw sequence reads (pre-trimming) were obtained (754,498,874 post adapter trimming), resulting in a raw SNP catalog of 753,326 SNPs (Online Resource 2 Table C). The average mapping rate of each of the merged bam files (i.e., samples) to the reference genome was 99.88% (ranging 99.86 - 99.91%) and the overall read depth at each position for the dataset was 9.82 (ranging 3.53 - 14.47). Custom scripts are available on Zenodo: <https://doi.org/10.5281/zenodo.14919600>

HWE Analyses in dartR

To perform HWE and LD analyses in dartR, the four populations identified with STRUCTURE were analyzed separately. For HWE, using the gl.report.hwe function, the Exact

method (Wigginton et al. 2005) for determining statistical significance was used with $\alpha = 0.05$, along with the False Discovery Rate (FDR) method based on Benjamini and Yekutieli (2001) to correct for multiple tests. All other HWE options used were the default.

Analysis of Outlier SNP Loci

To examine outlier loci using Arlequin, because there was no obvious evidence of hierarchical structure present from the STRUCTURE analyses, a non-hierarchical infinite island model was used. The number of coalescent simulations = 100,000, simulating 100 demes per group with a minimum and maximum expected heterozygosity of 0.0 and 1.0, respectively. For the Distance Method, “Number of different alleles” was chosen. For BAYESCAN, all default parameters were applied except the thinning interval was set to 500, 1,000 pilot runs were completed, and prior odds = 100. All default parameters were selected for OutFLANK and pcadapt. For the latter, initially $K = 100$ was tested and after examining the scree plot it was determined that $K = 4$ was best for the final PCA analysis.

STRUCTURE Reanalysis with Equal Sample Sizes

The estimation of the best number of populations using STRUCTURE can be biased when there is unequal representation (in terms of sample size) in the pre-defined populations being analyzed (Meirmans 2019). Although our study conducted STRUCTURE analyses without using any *a priori* population information, we attempted to address bias due to unequal sample sizes resulting from the Coastal Population ($n = 84$) being as much as 6 times the size of the smallest population identified (Southern Estuarine $n = 14$) by rerunning the Bayesian analysis using a subset of individuals from the Coastal Population. Specifically, 18 samples that were assigned to the Coastal Population based on the initial STRUCTURE run (using 141 individuals and 6,479 SNP loci) were randomly chosen. These 18 samples (Online Resource 2 Table E) were combined with all remaining samples that were either assigned to the three estuarine populations ($n = 52$) or were unassigned ($n = 5$) and STRUCTURE was rerun using all the same parameters as described in the main text except using 500,000 iterations. This resulted in the best $K = 2$ (Online Resource 2 Table E), and after running CLUMPP and using a Q -cutoff of 0.50 as described in the main text, individuals were assigned to each K . One group contained all of the Northern Estuarine ($n = 21$) and Coastal ($n = 18$) individuals in addition to 12 from the North-central Population and four originally unassigned individuals, and the second group contained all 14 of the Southern Estuarine Population, four from the North-central Population, and one originally unassigned individual. Each of these two groups was run again through STRUCTURE to identify hierarchical population structure as described in the main text, using the same parameters as above except testing $K = 1 - 5$ (Online Resource 2 Table E). For the larger group ($n = 55$), the best $K = 3$, and after running CLUMPP and assigning individuals using a Q -cutoff of 0.50 the remaining Coastal and North-central Estuarine individuals were each assigned to their own population and all but one of the Northern Estuarine individuals was assigned to the third population (one Northern Estuarine individual grouped with the Coastal Population). For the smaller group containing all 14 of the Southern Estuarine Population ($n = 19$), the best $K = 1$. Overall, results were exceptionally congruent with the original population assignments from analysis of the full data set.

REFERENCES

- Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. *Ann Stat* 29:1165-1188. <https://doi.org/10.1214/aos/1013699998>
- Catchen J, Hohenlohe P, Bassham S, Amores A, Cresko W (2013) Stacks: an analysis tool set for population genomics. *Mol Ecol* 22:3124-3140. <https://doi.org/10.1111/mec.12354>
- Hayes SA, Josephson E, Maze-Foley K, Rosel PE et al (2018) US Atlantic and Gulf of Mexico Marine Mammal Stock Assessments - 2017. NOAA Technical Memorandum, NMFS-NE-245:378
- Hayes SA, Josephson E, Maze-Foley K, Rosel PE, Byrd B, Turek J et al (2021) US Atlantic and Gulf of Mexico Marine Mammal Stock Assessments 2020. NOAA Technical Memorandum, NMFS-NE-271:403
- Li H (2013) Aligning sequence reads, clone sequences an assembly contigs with BWA-MEM. *arXiv:1303.3997v2 [q-bio.GN]* <https://doi.org/10.48550/arXiv.1303.3997>
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R and the 1000 Genome Project Data Processing Subgroup (2009) The sequence alignment/map format and SAMtools. *Bioinformatics* 25:2078-2079. <https://doi.org/10.1093/bioinformatics/btp352>
- Meirmans PG (2019) Subsampling reveals that unbalanced sampling affects STRUCTURE results in a multi-species dataset. *Heredity* 122:276-287. <https://doi.org/10.1038/s41437-018-0124-8>
- Wigginton JE, Cutler DJ, Abecasis GR (2005) A note on exact tests of Hardy-Weinberg equilibrium. *Am J of Hum Genet* 76:887-893. <https://doi.org/10.1086/429864>