Description of planned whole-genomic sequencing

P. Teske[1]

The need for population genomic data to inform sardine fisheries management is well illustrated by the fact the genomic work by Teske et al. (2021) is now being used in the development of a two-stock fishing model (e.g. de Moor et al. 2017, 2022, 2023a,b). However, the original genomic dataset on which the genomic study was based was generated almost a decade ago, at a time when reduced representation approaches (e.g. RADseq) were considered the ideal way to generate genomic data from a large number of individuals.

While the study proved useful in achieving its primary aim, which was to determine the origin of the sardines that participate in the KwaZulu-Natal sardine run, its limitations to the development of fishing models quickly became apparent. These include the fact that only a small portion of the genome was sequenced, that data from some regions and seasons was lacking (e.g. only one site on the west coast was sampled in summer), that no DNA from sardine eggs passed quality screening (so there was uncertainty whether sardines found in a particular area were actually spawning there or just travelling through), and that the gene regions for distinguishing between temperature-adapted regional stocks are likely not under selection themselves, but merely linked to such regions, making the genomic findings difficult to interpret. Perhaps the sample size was also too small to assess migration rates.

Following the recommendation by the panel in 2022 that more informative molecular data should be generated (see MARAM/IWS/2024/Sardine/BG2 for details), we submitted a proposal titled “SARDINE-SEQ: Stock structure of southern African sardines investigated by whole-genome sequencing” to the NRF in early 2024. The funds provided by the NRF are usually insufficient for work of this nature and previously only covered 1/3 of the work, so we have established collaborations with the University of Oslo and Southwest Fisheries Science Center (US). The details are as follows:

[1] Centre for Ecological Genomics and Wildlife Conservation, University of Johannesburg, Auckland Park 2006.

Aim:

The aim of this project is to generate and apply whole genome datasets for improving the long-term sustainable fisheries management of sardines in South Africa, Namibia and Mozambique.

Objectives:

The above aim will be achieved by identifying gene regions under selection that can be used to determine stock structure in southern Africa's sardine resource with unprecedented detail. Two approaches will be used: a) low cover whole genome sequencing (lcWGS) of 96 sardines from different marine regions and countries to identify genes regions that are most informative to distinguish between regional stocks; assembly will use a fully annotated South African S. sagax genome that will be generated by colleagues in the US; and b) development of a panel of these regions using GTseq to subsequently process a very large number of samples to assess stock ranges and mixing; as this method is suitable for DNA of lower quality than lcWGS and the previous methods, we anticipate that it may also be used for sardine eggs, which could not be analysed previously.

Value:

Although the approach outlined in the proposal uses cutting-edge technology, we expect that the research findings will be straightforward and easy to understand, and particularly relevant to fisheries management in South Africa. In fact, if the actual gene regions that are under thermal selection are found, the results will be even simpler compared to the previous ddRADseq and exome datasets, which mostly showed gradients in stock structure rather than examples of sardines that are completely affiliated with one of the two stocks.

A whole-genome sequencing approach is more likely to be successful in finding highly diagnostic gene regions that can be used to distinguish between stocks, which is necessary for sustainable management of the fishery for this two-stock resource.

Sampling:

To reduce costs and minimise the amount of time required to acquire a comprehensive dataset, we will use a combination of previously sequenced sardine samples that showed particularly strong affiliation with either of the two stock components, in addition to new

samples (approximately 1/3 of the total). Most of the new samples have already been collected, including sardines from Namibia that were captured offshore in 2021 between 21°21.1'S and 13°02.3'E and 21°21.4'S and 13°03.0'E, and Mozambican sardines captured near Inhassoro in 2023 (Fig.2). Additional samples will be collected opportunistically and during DFFE-funded pelagic biomass spawner surveys in 2024 and 2025. The latter will also provide an opportunity to acquire eggs; although the DNA quality of these tends to be comparatively low and they are not suitable for whole-genome sequencing, we will use these for a subsequent low-cost approach (GTseq). The bulk of the GTseq samples will be acquired during 2025 and early 2026.

DNA extraction and library preparation:

DNA will be extracted from 96 specimens from all representative regions, using the Qiagen Blood and Tissue Kit. DNA will be shipped on dry ice to the University of Oslo, where libraries will be created using an in-house protocol (Cobb et al. 2024: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0300865). This plate-based, tagmentase-based protocol is optimized for low levels of DNA, and requires as little as 6 ng of DNA. DNA input is normalized, and after library preparation using Illumina-compatible dual-index adapters, individual libraries will be pooled and sequenced on an Illumina NovaSeq Plus using 150 bp paired reads to approximately 6-14 fold coverage per individual.

The use of facilities in Oslo is crucial to make this project financially feasible as previous projects could not be funded by the NRF because of financial constraints, despite receiving reviewer approval.

Assembly and analysis of whole-genome data:

As for the whole-genome data generation, the collaboration with highly experienced international partners inworking with the South African participants on analysing these data using cutting-edge software is crucial to the success of this project. Please see the "Participating members" section for more detail on the expertise of Luciano Beheregaray (Flinders University, Australia) and Bastiaan Star (University of Oslo, Norway).

Reads will be aligned to a fully annotated reference genome of Sardinops sagax that will be generated by colleagues from the US, who are working on a phylogenomic study of sardines.

It is anticipated the several million SNPs (single nucleotide polymorphisms) will be identified, i.e. several orders of magnitude more than the tens of thousands generated previously (Teske et al. 2021). These will be used to clarify adaptive and neutral patterns of genetic structure, the strength and distribution of selective gradients across environments, and to identify the ranges of regional populations and assess their levels of gene flow and demographic histories. In addition, we will identify structural variants (SVs) that differ between geographical regions, i.e. larger chromosomal segments that may be several thousand nucleotides in length (Matschiner et al.2022), and that could not be identified with the short fragments generated for the previous RADseq and RNAseq approaches. SVs that differ between genomes may include deletions, duplications, insertions, inversions and translocations (Sudmant et al. 2015), and these have been shown to be of considerable importance in the evolution of new phenotypes that may not differ morphologically, but that are adapted to different environments, or display different behaviours. For example, in Atlantic cod, chromosomal inversions define migratory and stationary populations (Matschiner et al. 2022).

Datasets will be analysed using high-performance computing platforms, including the CSIR's Lengau cluster in Cape Town.

Population structure will be assessed with the genome data using both SVs (following the protocol outlined in Matschiner et al. 2022) and SNPs. Both selectively neutral and environmentally selected SNPs will be identified. The latter are particularly important to identify diagnostic SNPs useful to differentiate between populations (or stock components), and will be identified using consensus between Fst-based genome scans using BayeScan (Foll & Gaggiotti 2008) and gene-environment association analyses, using the univariate gINLAnd (Guillot et al. 2014) and the multivariate RDA (Forester et al. 2018). The program ADMIXTURE (Alexander et al., 2009) will be used as a tool for maximum likelihood estimation of individual ancestry proportions. Relationships between populations identified in this way will be analysed for long-term gene flow trends using Treemix (Pickrell and Pritchard, 2012) and trends in effective population size over the past 100 generations, using the program GONE v1 (Santiago et al. 2020), in both cases using the selectively neutral SNPs.

GTseq of diagnostic markers:

Following the identification of SNPs that are particularly suitable to distinguish between the Namibian, western South African and southern South African stocks (we anticipate that the Mozambican samples represent sardine run individuals, and are thus part of the western

South African stock), we will use GTseq (genotyping-in-thousands by sequencing; Campbell et al. 2015) to process a large number of additional samples, including eggs. This is a long-term endeavour that will exceed the available funding for this project, so we will raise funds for it separately while gradually increasing sample sizes. It is envisaged that GTseq analyses can eventually be conducted following annual spawner biomass surveys to assess shifts in stock ranges and levels of mixing.

References:

Alexander D.H. et al. (2009). Genome Res 19, 1655–1664.

Campbell NR et al. (2015) Mol Ecol Resour 15:855-867.

Cobb L et al. (2024) PloS One 19: e0300865.

de Moor CL et al (2017). Can J Fish Aquat Sci 74: 1895-1903

de Moor CL (2022) DFFE: Branch Fisheries Document FISHERIES/2022/OCT/SWG-PEL/34.

de Moor CL (2023a) DFFE: Branch Fisheries Document FISHERIES/2023/JUL/SWG-PEL/10.

de Moor CL (2023b) DFFE: Branch Fisheries Document FISHERIES/2023/JUL/SWG-PEL/13

Foll M, Gaggiotti O (2008) Genetics 180:977–993.

Forester BR et al. (2018) Molecular Ecology 27: 2215-2233.

Guillot G et al. (2014) Spat Stat 8:145–155.

Lavoue S et al. (2007) Mol Phylogenet Evol 43:1096–1105.

Matschiner M et al. (2022) Nature Ecol Evol 6:469–481.

Phillips SJ et al. (2006) Ecol Model 190:231-259.

Santiago E et al. (2020) Mol Biol Evol 37, 3642–3653.

Sudmant PH et al (2015) Nature 526:75–81.

Teske PR et al. (2021) Sci Adv 7:abf4514