Reanalysis of sardine genomic data based on 2023 panel recommendations

P.R. Teske[1], A. Emami-Khoyi[1,2] and S. Ockhuis[1]

Marine biogeography is strongly shaped by water temperature (Murawski 1993) and is likely the direct result of species' thermal tolerance ranges (Pörtner 2007). This implies that populations of ectotherms are primarily subdivided into evolutionary units that are adapted to different temperature-defined water masses (although other and often temperature-linked factors such as nutrient concentrations, pH, salinity etc. also play a role). Adaptation to local environmental conditions is thus an important force that drives ecological divergence, and strong regional differences in selection can theoretically result in population divergence in the absence of absolute dispersal barriers.

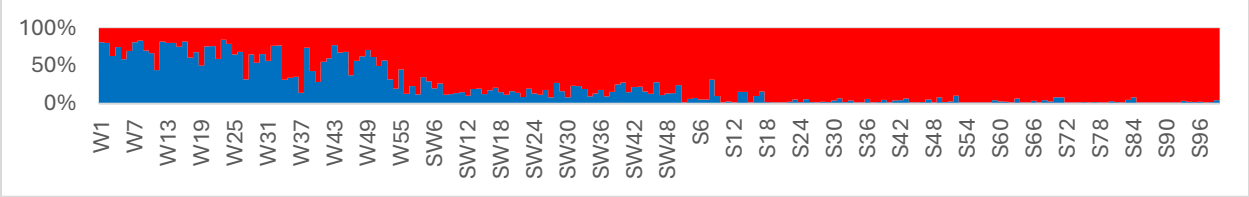
With this in mind, Teske et al. (2021) used the following common approach to identify the most informative genetic markers (outlier SNPs) in the genomic (ddRADseq) and exome (RNA-seq) datasets as a means of detecting genetic population structure that was neither evident in the complete dataset, nor in a dataset of selectively neutral variables. First, we used outlier detection by means of genotype-environment association (GEA) using gINLAnd (Guillot et al. 2014). We had environmental data from the following environmental variables: sea-surface temperature (SST), salinity, nitrogen (N), phosphate (P), silicone (Si) and dissolved O_2. The p-values of correlation analyses showed that all variables were significantly correlated with temperature and, given the importance of this variable as the primary driver of population divergence, we selected two datasets (the mean of the warmest and the mean of the coolest 5% of temperature measurements near the site of interest, see Teske et al. 2021 for details) for GEA. In addition, we used Bayescan (Foll & Gaggiotti 2008), which identifies outliers on the basis of genetic structure. Our final genomic outlier dataset included only SNPs that were identified by both methods. This is a commonly used approach in this field. The resulting dataset based on RNA-seq data was considerably more informative than the ddRADseq dataset, a major difference being that the former was suitable to assign several individuals to one of the two stock components identified (cool-temperate stock, or CTS, and warm-temperate stock, or WTS) with almost 100% probability.

[1] Centre for Ecological Genomics and Wildlife Conservation, University of Johannesburg, Auckland Park 2006, South Africa
[2] Institute of Wildlife Management and Nature Conservation, Hungarian University of Agriculture and Life Sciences, Gödöllő, Hungary

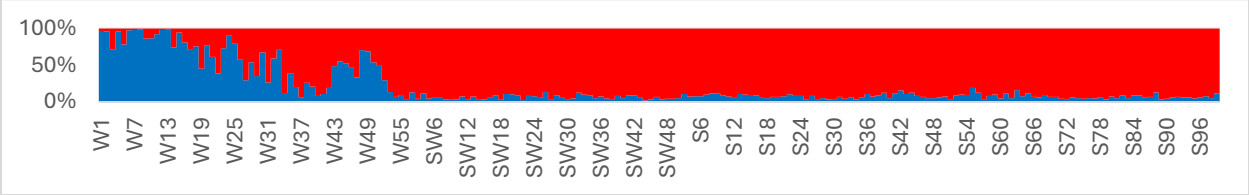
Following the recommendation by the 2023 panel to include additional environmental variables in the GEA approach (MARAM/IWS/2024/Sardine/BG2), we have now explored the ddRADseq dataset as follows:

First, to assess the relative information content of all six variables (in this case, with highest and lowest 5% of SST in the study region combined) detected using GEA (using the univariate program gINLAnd) were analysed separately in STRUCTURE (Pritchard et al. 2000), with 'locprior' enabled (Falush et al. 2003). We also used the genome scan methods PCadapt (Luu et al. 2017) and outFLANK (Whitlock & Lotterhos 2015) to identify outliers based on genetic structure. Patterns of spatial genetic structure for each is shown in Fig. 1. Not all individuals are shown on the x-axis, but the affiliation with three regions should be clear (total number of individuals per region: W = 55, SW = 48, S = 99)

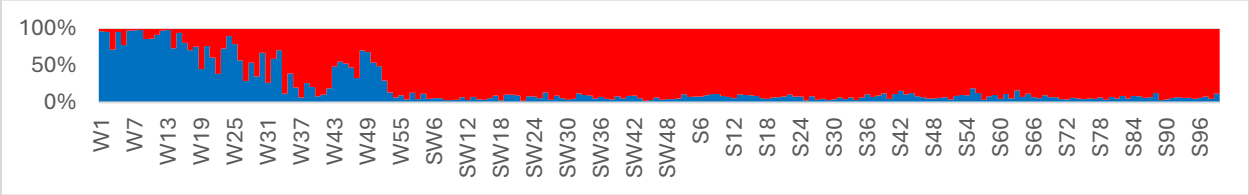
a) Temperature



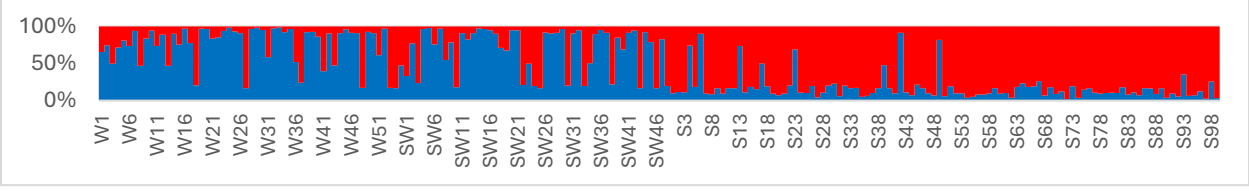
b) Salinity



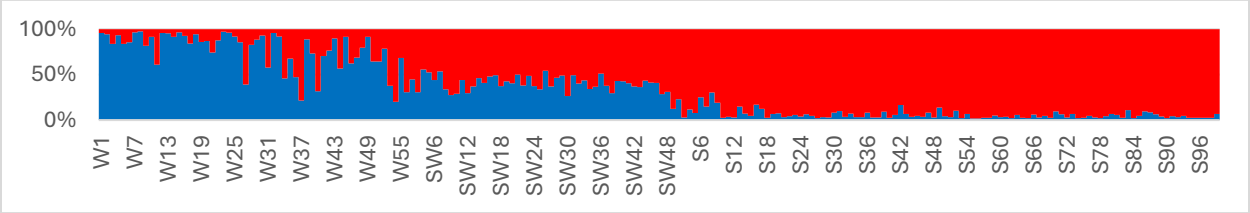
c) Dissolved O_2



d) N



e) P



f) Si

g) F_{ST} -based outliers (PCadapt and outFLANK)

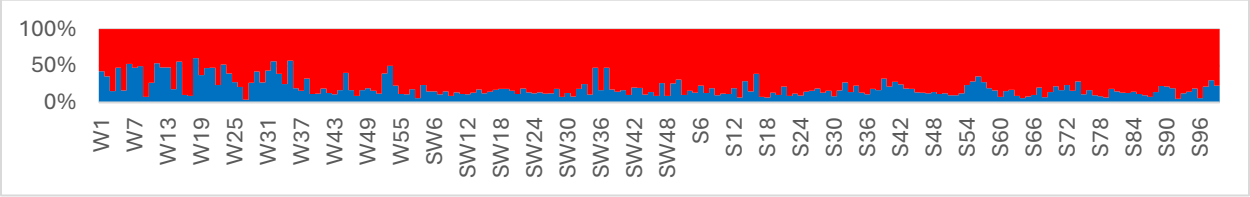


Fig. 1. Genetic structure based on outlier loci detected for individual environmental variable in GEA (a-f) and genome scans (g) using STRUCTURE; blue = CTS, red = WTS, W = west coast, SW = southwest coast, S = south coast.

The previously found intermediate status of sardines caught on the SW coast was not universally recovered here. For example, individuals from the SW coast were more similar to those from the W coast particularly for N, but more similar to those from the S coast for salinity and dissolved O_2. No clear trend was found for F_{ST} -based outliers, and this method was not explored further.

Next, we explored the magnitude of the correlation coefficient between pairs of environmental variables (as an alternative to the previous approach based on P-values) as a justification for excluding variables that are strongly correlated with others (Fig. 2).

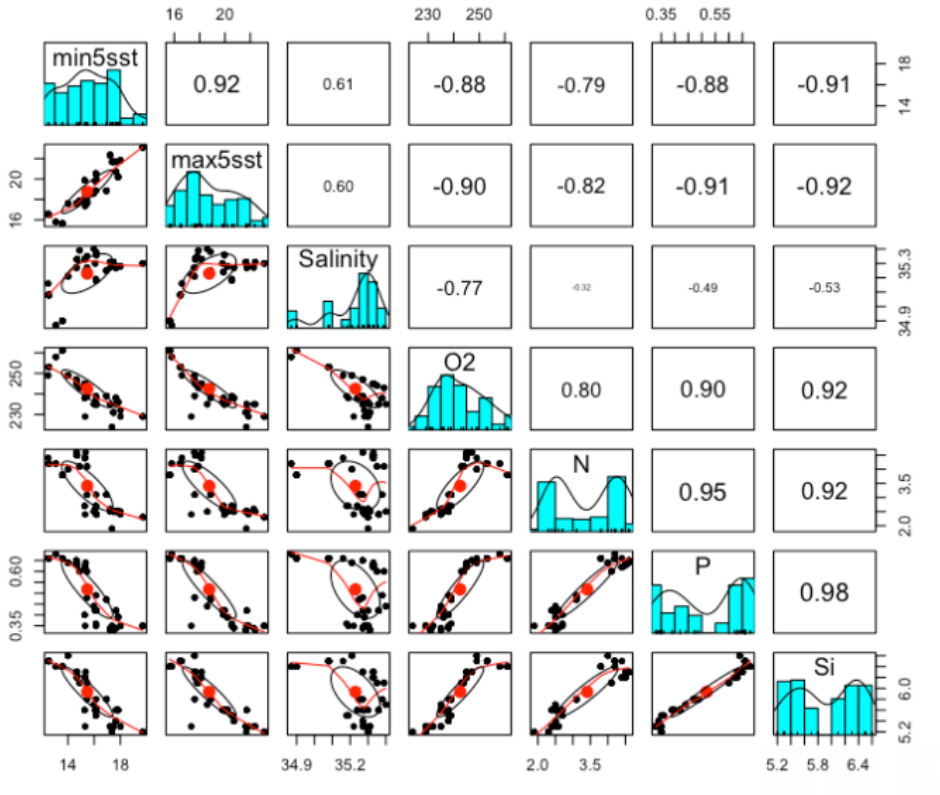
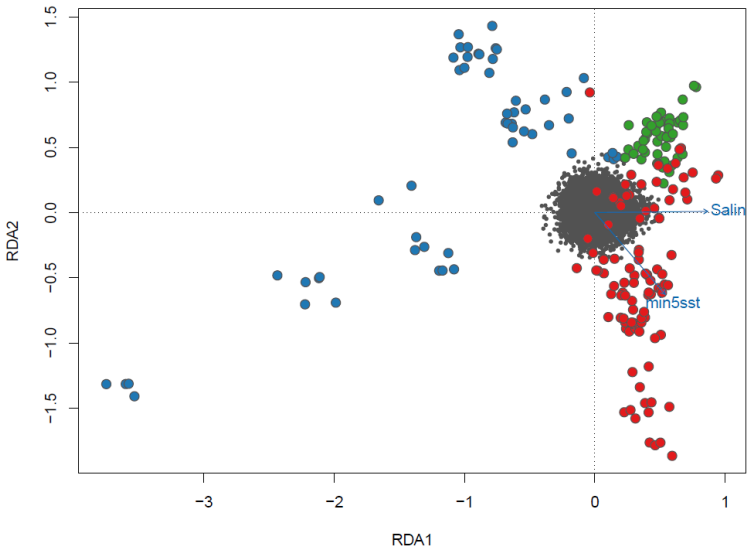


Fig. 2. Pairwise correlation coefficients calculated using the R package 'vegan' (Oksanen et al. 2016).

Applying the rule that $|r| > 0.7$ indicates strong correlations, only salinity is not strongly correlated with temperature (r = 0.6 for max5sst and 0.61 for min5sst). Comparatively low negative correlations among other pairs of variables were also found for salinity and N (r = -0.32), for salinity and P (r = -0.42), and for salinity and Si (r = -0.53).

Based on this finding, we explored outlier loci obtained from two different combinations of environmental variables: salinity + min5sst (of particular interest because it is assumed that west coast upwelling excludes south coast sardines), as well as salinity + N. A second GEA method, redundancy analysis (RDA in 'vegan', a multivariate ordination technique) was used to jointly estimate outliers for these combinations (Fig. 3). Note how the SW coast individuals in a) form a distinct cluster, which in b) is distinct from both the W coast and the S coast. The latter is a result of individuals found on the SW coast having salinity-correlated SNPs like those on the S coast, but N-correlated SNPs like those on the W coast (see Fig. 1).

a)



b)

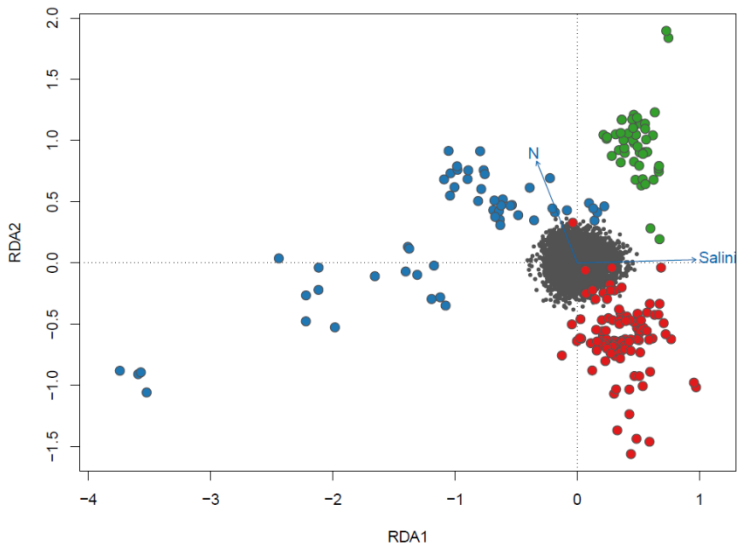
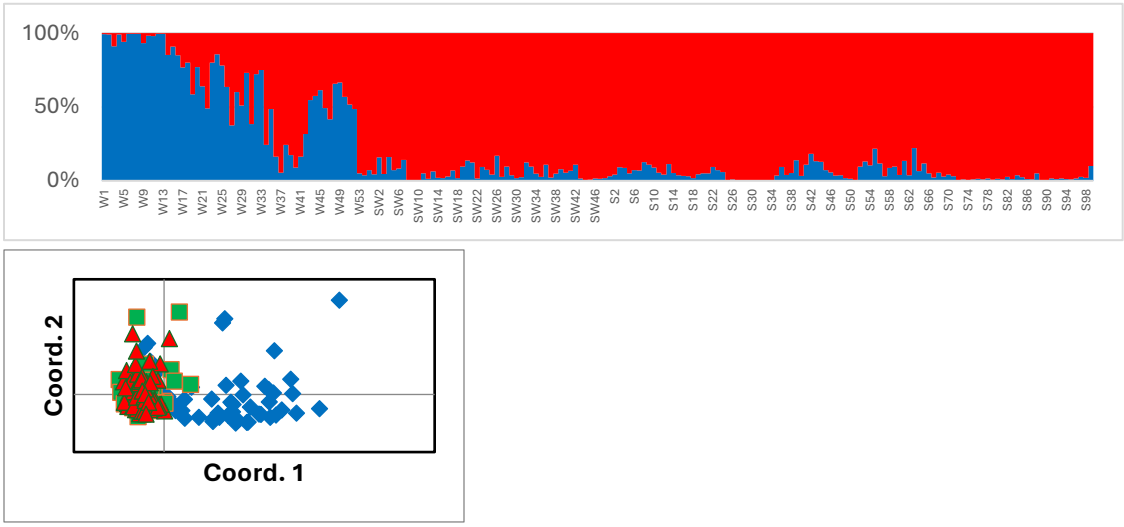


Fig. 3. Redundancy analysis (RDA) ordination plots identifying GEA outliers (grey dots) based on (a) salinity + min5sst and (b) salinity + N (environmental variables are represented by blue arrows). Individual sardines are assigned to three regions: W = blue, SW = green, S = red.

STRUCTURE barplots and ordination plots from Principle Coordinates Analysis (PCoA) using GenAlEx (Peakall & Smouse 2012) that combine the outliers for min5sst+salinity and salinity+N obtained from RDA with those obtained from gINLAnd are shown in Fig. 4 (unlike previously, these are not consensus outlier SNPs identified by both programs, but the total number of SNPs identified to increase the overall number of outliers). Results for the two datasets are very similar. As for the RNA-seq data from Teske et al. (2021), numerous individuals were assigned 100% to either the CTS (blue) or the WTS (red). A strong effect of salinity-correlated SNPs is evident in that individuals from the SW are more similar to those from the S coast, which differs from the previous results obtained for temperature on its own (see also Fig. 1). On the W coast, a gradient in relative ancestry proportions is evident from north to south, with the northernmost sites having the highest proportion of CTS ancestry.

a) min5sst + salinity



b) salinity + nitrogen

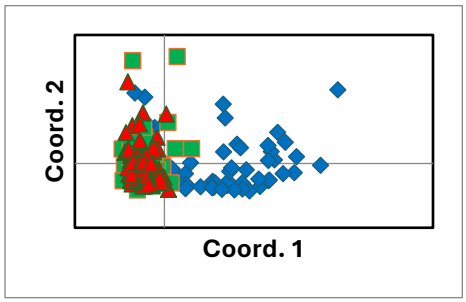


Fig. 4. STRUCTURE barlotsplots and Principal Coordinates Analyses (PCoA) constructed from (a) combined min5sst and salinity outlier loci and (b) combined salinity and nitrogen from gINLAnd and RDA; blue = CTS, red = WTS, W = west coast, SW = southwest coast, S = south coast.

A comparison of different values of *K* (number of populations) from 1 to 4 using ΔK in CLUMPAK (Kopelman et al. 2015) confirms that 2 populations (as shown in Fig. 4) is the optimal number for this dataset (example only shown for min5sst+salinity, Fig. 5).

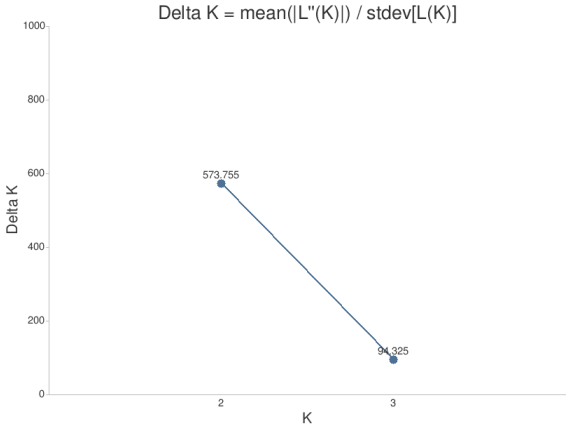


Fig. 5. Optimal number of pouplations (*K*) based on ΔK.

An assessment of migration rates between the three regions using BA3-SNPs software (Mussmann et al. 2019) produces the following results (e.g. [W][W] means self-recruitment into W, [W][SW] means migration into W from SW etc.):

m[W][W]: 0.980	m[W][SW]: 0.003	m[W][S]: 0.008
m[SW][W]: 0.319	m[SW][SW]: 0.674	m[SW][S]: 0.007
m[S][W]: 0.117	m[S][SW1]: 0.006	m[S][S]: 0.877

In each case, migration rates into a particular region sum up to 1. These values can potentially be converted into individual sardines using a mutation rate and an estimate of effective population size (N_e).

Conclusion:

Exploration of the ddRADseq data from Teske et al. (2021) using additional environmental variables, as recommended in MARAM/IWS/2024/Sardine/BG2, indicates that the dataset contains much additional information, and assignment of individuals to either CTS and WTS with 100% probability is possible (as was the case for the more informative RNA-seq dataset). The latter only included a very small number of sardines because of financial constraints.

However, the present dataset is not yet suitable to assess mixing proportions between regions and replace the approach described in MARAM/IWS/2024/Sardine/P5 that is based on the original dataset. The reason for this is that despite being more informative, F_{ST} between regions is presently too low to reliably distinguish between migrants and residents. Faubet et al. (2007) recommended that estimating migration rates requires an $F_{ST} > 0.05$ between the regions of interest, but the highest F_{ST} found here was 0.029 (between W and SW using the min5sst+salinity dataset).

It is possible that applying more stringent outlier selection criteria for the gINLAnd analyses, or using only the RDA outliers (given that this was the only method that supported some distinctness of SW from both W and S, Fig. 3b) may be a suitable way to increase F_{ST} between the major regions (at least between W vs. SW+S). Incorporating multiple correlated environmental variables (e.g. min5sst and N in the same dataset, given that their outlier datasets are not identical) may also be an option. This will be explored further during the next week.

References:

Falush D et al. (2003) Genetics 164: 1567–1587.

Faubet P et al. (2007) Mol Ecol 16: 1149-1166.

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

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

Kopelman MN (2015) Mol Ecol Res 15: 1179-1191.

Luu K et al. (2017). Mol Ecol Res17: 67-77.

Murawski SA (1993) Trans Amer Fish Soc 122: 647-658.

Mussmann SM (2019) Methods Ecol Evol 10: 1808-1813.

Oksanen J et al. (2016) vegan: Community Ecology Package. R package version 2.3-5.

Peakall R, Smouse PE (2012) Bioinformatics 28: 2537-2539.

Pörtner HO (2007) Phil Trans Roy Soc B 362: 2233-2258.

Pritchard JK et al (2000) Genetics155:945-959.

Whitlock MC, Lotterhos KE (2015) Am Nat 186: S24-36.