Population genomics reveals the underlying structure of the small pelagic European sardine and suggests low connectivity within Macaronesia

Rute da Fonseca[1], Paula Campos[2], Alba Rey de la Iglesia[1], Gustavo Barroso[3], Lucie Bergeron[4], Manuel Nande[2], Fernando Tuya[5], Sami Abidli[6], Montse Pérez[7], Isabel Riveiro[7], Pablo Carrera[7], Alba Jurado-Ruzafa[8], M Teresa G Santamaria[8], Rui Faria[9], Andre Machado[2], Miguel Fonseca[2], Elsa Froufe[2], and L Filipe C Castro[2]

[1]University of Copenhagen Faculty of Health and Medical Sciences
[2]University of Porto
[3]University of California Los Angeles
[4]University of Copenhagen Department of Biology
[5]Universidad de Las Palmas de Gran Canaria
[6]Université de Tunis El Manar
[7]Instituto Español de Oceanografía Centro Oceanográfico de Vigo
[8]Instituto Español de Oceanografía Centro Oceanográfico de Canarias
[9]University of Sheffield

July 30, 2022

Abstract

The European sardine (*Sardina pilchardus*, Walbaum 1792) is indisputably a commercially important species. Previous studies using uneven sampling or a limited number of makers have presented sometimes conflicting evidence for the genetic structure of *S. pilchardus* populations. Here we show that whole genome data from 108 individuals from 16 sampling areas across 5,000 Km of the species' distribution range (from the Eastern Mediterranean to the archipelago of Azores) supports at least three genetic clusters. One includes individuals from Azores and Madeira, with evidence of substructure separating these two archipelagos in the Atlantic. Another cluster broadly corresponds to the center of the distribution including the sampling sites around Iberia, separated by the Almeria-Oran front from the third cluster that includes all of the Mediterranean samples, except those from the Alboran Sea. Individuals from the Canary Islands appear as belonging to the same ancestral group as those from the Mediterranean. This suggests at least two important geographical barriers to gene flow, even though these do not seem complete, with many individuals from around Iberia and the Mediterranean showing some patterns compatible with admixture with other genetic clusters. Genomic regions corresponding to the top outliers of genetic differentiation are located in areas of low recombination indicative that genetic architecture also has a role in shaping population structure. These regions include genes related to otolith formation, a calcium carbonate structure in the inner ear previously used to distinguish *S. pilchardus* populations. Our results provide a baseline for further characterization of physical and genetic barriers that divide European sardine populations, and information for transnational stock management of this highly exploited species towards sustainable fisheries.

Background

Population structuring in the absence of obvious physical barriers have puzzled biologists for centuries. In oceanic environment, strong genetic structure is expected. Most marine animals are capable of exchanging migrants across large distances, and their genetic structure results from a combination of a long larval pelagic phase, high fecundity, large population sizes and adult migratory behavior (Faria et al., 2013). Yet, many studies have shown that several species have higher spatial genetic differentiation than expected considering their high dispersal potential (Palero et al., 2008; Pérez-Ruzafa et al., 2006). In the case of marine fish, structure can range from a lack of differentiation between oceans to significant structure within an ocean basin, challenging the simple concept of "open seas" and the assumption of high connectivity in marine environments (Graves, 1998). Assessing the existence of population structure in marine species capable of long-distance dispersal is essential to identify the various factors involved in population differentiation and diversification in the absence of complete physical barriers (Faria et al., 2021). This is especially relevant for conservation efforts, including stock management of commercially important species (Faria et al., 2013).

The Mediterranean Sea and the contiguous Northeastern Atlantic Ocean have been the focus of several phylogeographic and population genetic studies on marine fish marine fish (e.g. Patarnello et al., 2007; Tine et al., 2014). The Almeria-Oran Front, a well-defined oceanographic break situated east of the Strait of Gibraltar, has been suggested to be responsible as responsible for hindering gene flow between Mediterranean and Atlantic fish populations of many fish species but it is far from being an universal barrier (Patarnello et al., 2007). The less studied Macaronesia, a group of archipelagos (Azores, Madeira and Canaries) separated from the Euro-African mainland by c. 100–1,900 km, has also been the target of several phylogeographic studies (e.g. Kasapidis et al. 2011 or Sá-Pinto et al.,2008). This area is characterized by the presence of several oceanographic currents, e.g., the North Atlantic Current, the Azores Current and the Canary Current (Sala et al., 2013), that together with the apparent lack of physical barriers can strengthen the potential for gene flow. Therefore, it is not surprising that several studies have reported low population genetic differentiation within the Macaronesian region for different taxa (Faria et al., 2013), including fishes (Francisco et al., 2011; Stefanni et al., 2015). Species distributed across these regions can thus inform us about the existence of cryptic substructure and possible barriers to gene flow between populations.

One of the most important pelagic fish resources in Atlantic waters is the European sardine, *Sardina pilchardus* Walbaum, 1792. This species has an enormous economic value, especially in Southern Europe and Morocco, where it is the main target of the purse-seine fleets in Portugal and Spain, representing a major source of income for local economies (ICES, 2013). Recently reported low biomass levels (ICES, 2020) led to recommendation to reduced fishing in Southern Europe, with great economic impact. It also prompted us to reevaluate the current population structure of *S. pilchardus* aiming at the ongoing discussion on how genetic information can contribute to stock delineation for management purposes (Caballero-Huertas et al., 2022).

The European sardine has a broad distribution from the Eastern Mediterranean to the North-East Atlantic, including the Azores, Madeira and the Canary Archipelagos, and is found along the African coast down to Senegal (Parrish et al., 1989). As other marine pelagic fish, *S. pilchardus* shows schooling and migratory behavior and high dispersal capabilities, both at the larval and adult stages. In agreement, low levels of genetic differentiation were detected across the species distribution using allozymes (Chlaida et al., 2009; Chlaida et al., 2006; Laurent et al., 2007; Spanakis et al., 1989), mitochondrial DNA (mtDNA) (Atarhouch et al., 2006; Tinti et al., 2002), and microsatellites (Gonzalez & Zardoya, 2007; Kasapidis et al., 2011). Nevertheless, phenotypic variation in gill raker counts and head length (Andreu, 1969; Parrish et al., 1989) and mitochondrial haplotype frequency differences (Atarhouch et al., 2006) led to the proposal of two subspecies: *S. pilchardus pilchardus* (North Sea to southern Portugal), and *S. pilchardus sardina* (Mediterranean Sea and northwest African coast). Accordingly, otolith shapes differ between Atlantic and Mediterranean sardines (Jeema et al., 2015), and further suggest a subdivision between the Northern Mediterranean and the Alboran-Algero-Provençal basin (Jeema et al., 2015; Alemany & Alvarez, 1993). A study using 15 allozymes supports the latter (Ramon & Castro, 1997), but, unlike the otolith shapes, these markers suggest discontinuity caused by the Almeria-Oran front. When considering a large fraction of the European sardine Atlantic range, allozymes

and microsatellites suggest that Madeira and Azores form a significantly differentiated group (Kasapidis et al., 2011). This mosaic of regional population structure built by several independent studies has been mostly justified by geographical barriers that potentially hinder gene flow, expected to be high for the abundant and mobile *S. pilchardus*. The phenotypic differences between groups might also have arisen from retention of adaptive phenotypes, and population structure in the Mediterranean was found to be associated with environmental variables (Antoniou et al., 2022). This prompted us to raise questions about the contributions of genomic architecture to the basis for the observed present-day population structure.

In this study, we produced an European sardine genomic data set consisting of whole genome nuclear data and complete mitochondrial genomes for 88 individuals that were analyzed together with data from 20 sardine individuals from a previous study (Barry et al., 2022), in a total of 108 samples from a total of 16 locations across 5000 km of the species distribution range. This enabled investigating previously suggested barriers to gene flow, mapping the major genetic clusters that characterize *S. pilchardus* in a large part of its distribution, the comparison between markers with different modes of inheritance but also to get a first insight into the genomics barriers contributing to the observed population structure.

Materials and Methods

Sample collection and DNA extraction

Samples were collected from 17 different geographical locations encompassing a large part of the species' current distribution range (Figure 1A, Table 1). A total of 15 samples from three were collected during oceanographic surveys and the remaining 73 specimens, from ten distinct geographic locations, were sampled at local markets (Table S1). Sequence data for the samples from Bay of Biscay, Gulf of Cadiz, Mar Menor and the Gulf of Lion (n=20) were obtained from Barry et al. (2022), adding further sampling locations to our dataset.

Total genomic DNA was extracted using Qiagen's DNeasy Blood & Tissue Kit (Hilden, Germany) according to the manufacturer's instructions, with the following modifications, prior to elution in 100ul AE buffer, samples were incubated at 37 ^{o}C for 10minutes, to increase DNA yield. DNA concentration and purity were verified using a Nanodrop Spectrophotometer and a Qubit Fluorometer. A commercial service (Novogene, China) produced Truseq Nano DNA libraries and sequenced paired-end reads (150 base pairs (bp)) in a Novaseq6000. To assess the patterns of genetic differentiation of the European sardine, 81 samples were sequenced to $<$ 3 X sequencing depth (i.e. each position of the genome is covered by 3 reads) and seven to 20 X sequencing depth (details in Additional file 1: Table S1). Raw data for 20 sardine individuals from (Barry et al., 2022) was further processed using the same procedure as described in the next section (sequencing depth between 15 and 22 X). Table S1 indicates the assignment of samples to the different subsets considered for further analysis.

Assembly filtering and sequencing data pre-processing

Individual contigs in the reference genome (GenBank assembly accession: GCA_900499035.1) of *S. pilchardus* (Louro et al., 2018) matching mitochondrial DNA (unique assignment to mitochondrial DNA without nuclear genome flanking regions) were identified via BLAST (version 2.6.0+) (Camacho et al., 2008) using the mitochondrial genome (mtDNA) assembled by Machado et al. (2018) as a query. Matching contigs were removed from the assembly file and replaced by the mtDNA of Machado et al. (2018) to enable the extraction of individual mtDNA sequences from all individuals after mapping of resequencing data.

Regions of low complexity in the reference genome of *S. pilchardus* (Louro et al., 2018) were detected with GenMap (version v1.3.0) (Pockrandt et al., 2020) using a k-mer of 100bp. We calculated the normalized depth per scaffold by using the sequencing depth of scaffold 1 as a reference to identified potentially misassemblies

(e.g. unmerged haploid scaffolds or collapsed repeats regions). Regions of mappability below 1 (meaning that more than one 100bp kmer matched the region, indicating duplications) or identified as repeats in all the other scaffolds, and sites with data missing in more than 25% of the individuals were excluded from all subsequent analyses.

Raw Illumina reads for all 108 samples were first processed with Trimmomatic (version 0.36) (Bolger et al., 2014) for removal of adapter sequences and trimming bases with quality <20 and discarded reads with length <80. Clean reads were mapped to the genome assembly using bwa-mem version: 0.7.17-r1188 (Li, 2013) and samtools version: 1.7 (Li et al. 2009) was used to retain reads with mapping quality >25. PCR duplicates were removed with Picard MarkDuplicates (version 1.95; http://picard.sourceforge.net) and only reads were both pairs were retained were considered for the local realignment around indels with GATK version 3.6-0-g89b7209 (DePristo et al., 2011) and further analyses. The mapping and base quality options -minQ 20 -minMapQ 30 were used in all subsequent analyses with ANGSD (Korneliussen et al., 2014).

Population structure

Beagle files with the nuclear genome positions of single nucleotide polymorphisms (SNPs) were produced by ANGSD (Korneliussen et al., 2014) using the following options: -GL 1 -doGlf 2 -minMaf 0.05 -C 50 -baq 2 -remove_bads 1 -uniqueOnly 1 -SNP_pval 1e-6. Linkage disequilibrium (LD) was estimated as r^2 values for all SNP pairs minimum 500 kbp apart with ngsLD v1.1.1 (Fox et al., 2019) and a LD decay curve was plotted using the script provided by (Fox et al., 2019) using 0.05% of all estimated r^2 values. This indicated that a distance threshold of 2,000 bp was adequate for linkage pruning. A total of 560,735 SNPs were obtained using all samples (n=108, Table S1), and 319,236 SNPs were found to be in putatively neutral regions of the genome (see below). Admixture proportions were estimated by running NGSadmix version 32 (Skotte eKt al., 2013) for K equal 2 and 3 with 300 seed values, ensuring convergence (convergence was not reached for K = 4 and above). A principal component analysis (PCA) using the same SNP set was obtained with PCAngsd version 0.1 (Meisner & Albrechtsen, 2018).

The mitochondrial genome (mtDNA) for each individual was obtained as a consensus sequence of the reads mapped to the mitochondrial DNA sequence included in the reference genome. The individual sequences were generated in ANGSD (Korneliussen et al., 2014) using the option -doFasta 2 for sites with a minimum sequencing depth of 10X (-setMinDepth 10). An haplotype network was designed using mitochondrial SNPs with minor allele frequency >30% (total of 26 SNPs) in POPART (Leigh & Bryant, 2015) with the Median Joining Network algorithm (Bandelt at al., 1999).

Population differentiation

We used methods based on the site frequency spectrum (SFS) (Korneliussen et al., 2013; Nielsen et al., 2012) to obtain the genome-wide fixation index (F_{ST}) values in ANGSD (Korneliussen et al., 2014). We calculated F_{ST} for the populations (six) containing at least 10 individuals with each geographical region represented by two locations (Figure 2). First, we generated unfolded SAF files (angsd -bam bamList -doSaf 1 -anc ANC -GL 1), then we estimated the folded SFS for each pair of populations (realSFS safidx1 safidx2 -fold 1). Each joint folded SFS was then used to estimate F_{ST} (-whichFst 1 -fold 1). To detect genomic windows of high differentiation in each region, we estimated the population branch statistic (Yi et al., 2010) for non-overlapping windows of 50 kb in ANGSD using this same approach with three populations with 10 individuals each (Table S1). The individuals were chosen from those that had 100% assignment to one of the three ancestral populations defined by NGSadmix in preliminary analysis using all genomic positions that passed the filters described above. Genomic positions within windows with PBS values below the 90^{th} percentile (putatively neutral) were used in all analyses presented.

Maritime geographical pairwise distances (https://sea-distances.org/) were calculated using the seaport nearest to the sampling location. Average distances were considered for merged populations. A Mantel test

implemented in ade4 (Dray & Dufour, 2007) was used to test the statistical significance of the correlation between the geographic and the genetic distance matrices.

We assessed the gene content of the top outlier PBS windows for each region by running tblastn (Camacho et al., 2008) (BLAST version 2.6.0+) of the zebra fish proteome (ENSEMBL version GRCz11_pep) against the reference genome using the option "-evalue 0.000001". Phenotypes associated with each gene were extracted from ENSEMBL using Biomart (Smedley et al., 2009).

Recombination rate

Variants were called using GATK version 4.0.7.0 (DePristo et al., 2011) for one representative individual per region (Table S1). Briefly, variants were first called for each individual with HaplotypeCaller in BP-RESOLUTION mode, then combining those GVCF files for each sample into a single one using CombineGVCFs per scaffold of interest, and finally joint genotyping with GenotypeGVCF. The default filter of GATK (–phred-scaled-global-read-mismapping-rate 45; –base-quality-score-threshold 18; –min-base-quality-score 10) was used. Recombination rates for 100 kb non-overlapping windows along the genome were estimated using the iSMC approach from (Barroso et al., 2019). We fitted an iSMC model with 40-time intervals and five categories of recombination rates to the samples from each population and optimized parameters in composite likelihood fashion (Barroso & Dutheil, 2021). We then obtained recombination landscapes of single-nucleotide resolution by performing posterior decoding in each diploid using the estimated parameters, and computed a consensus map for each sample by averaging over (for each site) the posterior estimates of rho = 4*Ne*r from all diploids. The final map of 100 kb resolution was obtained by further averaging the single-nucleotide estimates over 100 kb in non-overlapping windows.

Results

Population structure

The admixture analysis conducted in NGSadmix showed that European sardines seem to be part of at least three structured populations (Figure 1B). When setting the number of expected clusters to 2 (k=2; Figure 1B, top), one of the clusters is prevalent in the Center region, while the other is more frequent in both Western and Eastern regions, as well as the Canary Islands. Individuals with admixed ancestry from these two clusters were observed at all sampling sites, except at Madeira. For k=3 (Figure 1B, bottom), one of these clusters (West-East-Canaries) splits into two: one frequent in the Mediterranean and Canaries and the other in the West (Madeira and Azores). Admixed ancestry between the three main clusters was observed in individuals from the Central region, Eastern Iberia, South of France, Tunisia and from the Aegean Sea. Canaries' individuals show some admixed ancestry with the Western cluster; and two individuals from The Adriatic share some ancestry from the Western cluster. The proportion of individuals with admixed ancestry is lower in populations located in the extremes of the distribution: Madeira, Azores in the West, Aegean and Adriatic Sea in the East as well as Bretagne and Bay of Biscay S in the central group.

The organization in three separate clusters can also be observed in the principal component analysis (PCA; Figure 1C). The first two PCs explained 5.8 % and 4.6 % of the total variation. Both PC1 and PC2 separate the samples into three clusters. Sampling locations do not form individual groups within the three main clusters, reflecting the high amount of admixture observed in Figure 1B, except for Madeira and Azores, which appear separated in the PC1.

The main clusters observed in the mitochondrial haplotype network combine haplotypes that are not geographically confined to a region, suggestive of gene flow mainly between Center and East (Figure 1D), which is in agreement with their proximity in PC2 (Figure 1C). The West group dominates a centrally branching haplogroup.

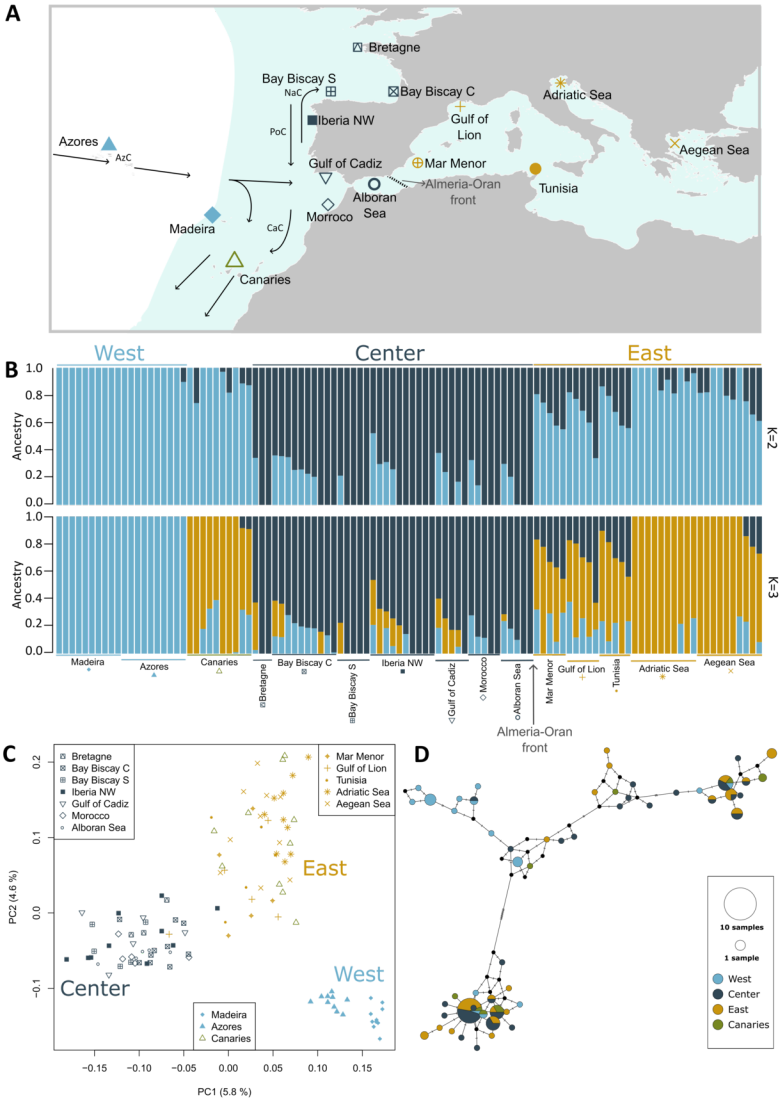

Figure 1

Figure 1. ***A)*** *Sampling sites across the species distribution (blue, adapted from FAO). The color of each circle represents the most frequent genetic cluster for K=3 (Figure 2A). Surface currents are represented by arrows: Azores current (AzC); Canary Current (CaC); Portugal Current (PoC); Navidad Current (NaC). The Almeria-Oran Front (AO) is shown as a dashed line.* ***B)*** *Population structure plot showing the ancestry of each individual (vertical bar) to two (above) and three (below) genetic clusters.* ***C)*** *Distribution of individuals based on the first three components of the principal component analysis. Variance explained by each component is shown in parenthesis.* ***D)*** *Population net (PopART, median-joining network) obtained using mitochondrial variants with minor allele frequency above 30%; mutations shown as Hatch marks. Colors represent the main ancestry of each individual (for K=3 as in Figure 1B) except for samples from the Canary Islands which are depicted in green.*

The organization in three separate clusters can also be observed in the principal component analysis (PCA;

Figure 2B). The first two PCs (Figure 2A, left) explained 7.9 % and 5.0 % of the total variation. PC1 separates the West-East clusters from the Center, and PC2 partitions the Western cluster from the Eastern populations. Sampling locations do not form individual groups within the three main clusters, except for Madeira, Azores and partially Greece, reflecting the high amount of admixture observed in Figure 2A. PC3 further separated the individuals from Azores from those of Madeira (Figure 2A right).

The phylogenetic tree of complete mitogenomes shows two well supported clades at the extremes, with some haplotypes branching from the middle part of the tree (Figure 3 and Figure S2). This agrees with the genetic clusters observed for the nuclear data belonging to regions with low F_{ST} variance (Figure S3 and Figure S4). While the central haplotypes are more common in the West group, the groups of haplotypes at each extreme of the tree are not geographically confined to a region, suggestive of high gene flow between the Center and the East.

Population differentiation

The levels of nucleotide diversity were comparable across regions, although slightly lower for populations in the West (Table S2). In general, we observed lower values of genetic differentiation as measured by F_{ST} for comparisons within regions (distances ranging from 982 to 1,943 Km show F_{ST} values between 0.004 and 0.03; Figure 2 and Table S3). The highest values of F_{ST} included comparisons with Madeira (West) and sampling locations in the Center region and the Aegean Sea ($0.044 < F_{ST} < 0.05$). We do not find evidence of isolation by distance across the sampled range (*Mantel* test; P-value = 0.32). Areas of the genome that correspond to the top PBS outliers per region (the window where one population is the most dissimilar to the other two) show lower recombination rates, and areas of very low differentiation are associated with the highest recombination rates (Figure 3). The gene content of these three outlier windows includes protein involved in otholith formation (East), vasculature and organ formation (Center) and blood coagulation (West) (Table S4).

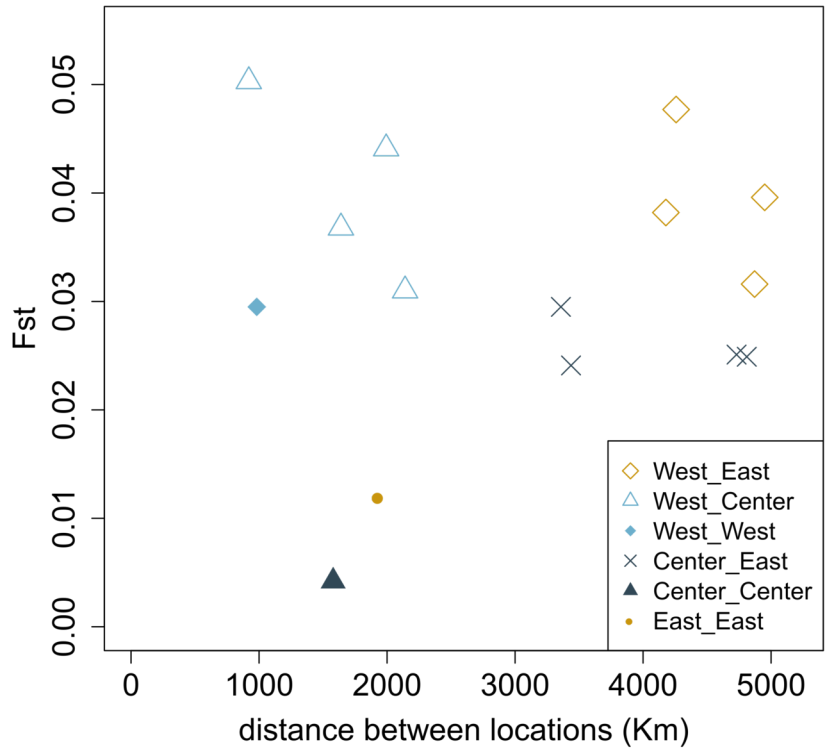

Figure 2

Figure 1: **Figure 4.** Pairwise F_{ST} between populations based on nuclear data (see Table S2 for details on the individual populations).

Figure 2. Average pairwise F_{ST} between populations with minimum of 10 individuals (Table S2) calculated across 50 kb windows located in putatively neutral regions of the genome.

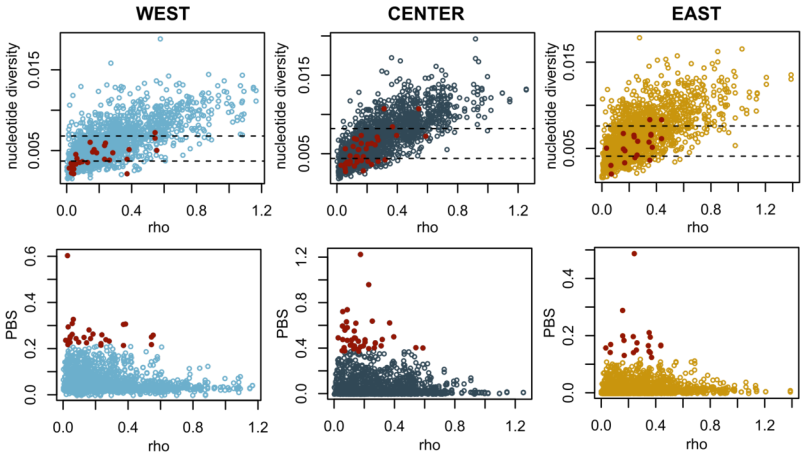

Figure 3

Figure 3. *Summary statistics calculated across 50 kb windows located in putatively neutral regions of the genome located in comparison with the top outliers (99th percentile) of the PBS analysis for each region (red dots).* ***A)*** *Nucleotide diversity* vs *recombination rate.* ***B)*** *PBS* vs *recombination rate.*

Discussion

In this study, we present the first analysis of population structure in European sardine across a large part of its distribution range using whole-genome sequencing data. A number of mechanisms have been suggested to explain how population structure can evolve in an environment without any complete physical barrier to gene flow, including local adaptation, habitat discontinuity, different habitat preferences and behavior, sexual selection, oceanographic currents , isolation by distance and limited dispersal capabilities (Bremer et al., 2005; Díaz-Jaimes et al., 2010; Faria et al., 2021; Kumar & Kumar, 2018; Patarnello et al., 2007).

Altogether, the assessment of nuclear genome sequences by means of individual ancestry information, principal component analysis (Figure 1B and Figure 1C) and differentiation (F_{ST}) among populations from different geographic regions (Figure 2), supports that the European sardine comprises three main stocks: "West" that includes individuals from Azores and Madeira (part of the Macaronesian region in the Atlantic), "Central" that corresponds to Iberia (the center of the sampling distribution), and "East" that gathers the Mediterranean samples and those from the Canary Islands. The observed genetic differentiation between Mediterranean and Atlantic populations (except the Canary Islands) is in agreement with previous phenotypic and genetic studies based on mtDNA (Andreu, 1969; Atarhouch et al., 2006; Parrish et al., 1989), suggesting the existence of a phylogeographic break between the South of Portugal and Mediterranean populations. The Almeria-Oran Front is likely to be responsible for reduction in gene flow between populations in each side, as previously observed in sardine (Ramon & Castro, 1997) and other species (Pita et al., 2014; Ouagajjou & Presa, 2015), as the Spanish Mediterranean populations have only a small proportion of ancestry associated with the Central cluster (Figure 1B). Instead, the shared pattern of ancestry of the individuals from the Alboran Sea and the Gulf of Cadiz indicates that the Strait of Gibraltar is not such a strong barrier as previously suggested (Jeema et al., 2015; Alemany & Alvarez, 1993; Ramon & Castro, 1997).

The population from the Canary Islands has a Mediterranean ancestry, and its divergence from the Western group has also been suggested by Kasapidis et al (2011) using microsatellite data, which further revealed a

high differentiation between Azores/Madeira and the other Atlantic populations. Notably, populations from these two archipelagos cluster together genetically, despite Madeira being geographically closer to Canary Islands and almost at the same distance to Iberia as it is to Azores. This strongly suggests a barrier to gene flow between the region formed by these two archipelagos and the other populations analyzed in this study, including Canary Islands and Iberia. This genetic division can be caused by currents, isolation by distance and lack of suitable habitat between these regions, local adaptation to different environmental conditions or other reasons. The fact that we did not observe a pattern consistent with isolation by distance and that we excluded markers putatively under selection argue against the latter. Nonetheless, this needs further investigation.

The higher differentiation of sardine populations from Azores and Madeira is also clear in the mitogenome tree (Figure 3). Although two other main clades are observed, they are formed by haplotypes from individuals with a very different nuclear-based ancestry. Thus, it is not easy to objectively pinpoint the geographic origin of these mtDNA clades.

Discordance between differentially inherited markers can simply result from stochastic patterns of lineage sorting, but it can also be indicative of introgression (Lavretsky et al., 2014). Patterns suggesting admixture between the three genetic clusters were also observed with the nuclear data in all populations except Madeira. Given the lower effective population size of mtDNA when compared to nuclear DNA, we would expect to see it more sorted within each region. The fact that haplotypes from the main clades in the mitochondrial tree are present across almost the entire distribution could eventually favor introgression over incomplete lineage sorting.

An important piece of information that can help us to disentangle the role of gene flow versus shared ancestral polymorphism is the geographic pattern of differentiation. Genetic differentiation is lower between closer geographic populations within the East and Center clusters (Figure 2). Furthermore, we observed that the proportion of individuals with pure nuclear ancestry is higher in populations that are geographically more distant from populations with a different ancestry, suggesting that at least some of the patterns observed with nuclear and mtDNA genomes can indeed be created by gene flow between populations from these genetic clusters. Although this needs to be further confirmed using model-based approaches, if true, it provides additional support that the barriers involved in the differentiation between these three genetic clusters are only partial. Furthermore, the ancestry patterns observed between populations from the Central and Eastern clusters could suggest bidirectional gene flow between populations from Iberia and Mediterranean populations outside Iberia, which is also supported by the more even distribution of the two main mitochondrial haplotypes between these regions.

Western cluster ancestry is also observed in populations from the Center, Canaries' Island and mainly in the Western Mediterranean populations. Although these patterns are compatible with admixture, gene flow between populations from the Eastern and Western clades are more difficult to explain. This discordance between molecular markers can also reflect the fact that regional populations of sardines seem to undergo periodic extinctions and recolonizations (Grant & Bowen, 1998). A recolonization of the Mediterranean from a refugium in the West African coast, as it has been suggested for anchovies (Magoulas et al., 2006), a species that shares several traits with sardines (Checkley et al., 2017), could potentially explain the admixed ancestry of the Canary Islands and the Eastern cluster (Figure 2).

Finally, we found that genomic regions corresponding to the top outliers of genetic differentiation are located in areas of low recombination (Figure 3), suggesting that genetic architecture can be contributing in some extent to the observed pattern of population structure. Interestingly, one of these regions include genes related to otolith formation, and otolith shapes have been found to divide the Atlantic and Mediterranean sardines (Jeema et al., 2015).

Conclusions

Our main results provide evidence for three main genetic clusters of sardine populations across the analyzed specimens, suggesting at least two important barriers to gene flow. Although these do not seem complete, with gene flow possibly occurring between the three main phylogeographic regions identified, they seem to be strong enough to maintain populations genetically differentiated following their own evolutionary trajectory. Our results thus offer an important baseline for further studies trying to identify the nature of these and other possible barriers between sardine populations, which can be compared with the phylogeographic patterns of other organisms with a similar distribution. Finally, the differentiation patterns reported here together with the genetic resources generated for this commercially important species, offers information of strategic importance for transnational stock management of this highly exploited species towards sustainable fisheries.

Acknowledgments

All figures were edited in Inkscape (http://www.inkscape.org/). Thanks to Alessandro Laio, Amélia Fonseca, Ludovic Dickel, Patrícia Campos, Sara Rocha, Yorgos Athanasidis and Sabour Brahim, for supplying tissue samples. We would also like to thank Anders Albrechtsen, Katherine Richardson, Lounes Chikhi, Jonas Meisner, Jørgen Bendtsen, Rasmus Heller, Ricardo Pereira, and Stephen Sabatino for advice. The authors gratefully acknowledge the following for funding their research: Villum Fonden Young Investigator Grant VKR023446 (R.D.F.); Fundação para a Ciência e a Tecnologia (FCT), Portugal, Scientific Employment Stimulus Initiative, grants CEECIND/00627/2017 to E.F and CEECIND/01799/2017 to P.F.C.. R.D.F. thanks the VILLUM FONDEN for the Center for Global Mountain Biodiversity (grant no 25925). M.P. and I.R. thank the Axencia Galega de Innovación (GAIN), Xunta de Galicia, Spain, for its funding of the AQUACOV and MERVEX Research Groups (grants IN607B 2018/14 and IN607-A 2018/4) and IMPRESS project supported by Spanish MICINN through grant RTI2018-099868-B-I00. R. F. is currently funded by FEDER through the Operational Competitiveness Factors Program (COMPETE) and by FCT (project "Hybrabbid", grants PTDC/BIA-EVL/30628/2017 and POCI-01-0145-FEDER-030628). E.F. research was funded by the project The Sea and the Shore, Architecture and Marine Biology: The Impact of Sea Life on the Built Environment Project No. POCI-01-0145-FEDER-029537, co-financed by COMPETE 2020, Portugal 2020 and the European Union through the European Regional Development Fund (ERDF). L.F.C.C research was funded by: project VALORMAR (reference nr. 24517), supported by COMPETE2020, LISBOA2020, ALGARVE2020, PORTUGAL2020, through ERDF; strategic funding UIDB/04423/2020 through FCT and ERDF, in the framework of the programme PT2020. We thank the scientific and technical staff and the crew of the PELACUS0315 and SARLINK oceanographic surveys conducted by the Instituto Español de Oceanografía. Alboran Sea samples were collected during the SARLINK oceanographic survey. Samples from Galicia, Cantabrian Sea and Bay of Biscay were collected during the PELACUS 0315 Oceanographic survey, funded by the EU through the European Maritime and Fisheries Fund (EMFF) within the National Program of collection, management and use of data in the fisheries sector and support for scientific advice regarding the Common Fisheries Policy.

References

Alemany, F. & Alvarez, F. (1993). Growth differences among sardine (*Sardina pilchardus* WaIb.) populations in Western Mediterranean. *Sci. Mar.*, 57, 229—23

Andreu, B. (1969). Las branquispinas en la caracterización de las poblaciones de Sardina pilchardus (Walb). *Las Branquispinhas En La Caracterizacion de Las Poblaciones de Sardina Pilchardus (Walb.)*, *33*(1), 425–607. Retrieved from http://hdl.handle.net/10261/166805

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, *17*(2), 81–92. doi: 10.1038/nrg.2015.28

Antoniou, A., Manousaki, T., Ramírez, F., Cariani, A., Cannas, R., et al. (2021) *Sardines at a junction: seascape genomics reveals ecological and oceanographic drivers of variation in the NW Mediterranean Sea.* Authorea. doi: 10.22541/au.162854295.58977249/v1

Atarhouch, T., Rüber, L., Gonzalez, E. G., Albert, E. M., Rami, M. et al. (2006). Signature of an early genetic bottleneck in a population of Moroccan sardines (*Sardina pilchardus*). *Molecular Phylogenetics and Evolution*, *39*(2), 373–383. doi: 10.1016/j.ympev.2005.08.003

Bandelt H, Forster P, Röhl A (1999). Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol*. 16(1), 37–48.

Barroso, G. V., Puzović, N., & Dutheil, J. Y. (2019). Inference of recombination maps from a single pair of genomes and its application to ancient samples. *PLOS Genetics*, *15*(11), e1008449. doi: 10.1371/journal.pgen.1008449

Barroso, G. V. & Dutheil, J. Y. (2019). Mutation rate variation shapes genome-wide diversity in Drosophila melanogaster. bioRxiv 2021.09.16.460667; doi: https://doi.org/10.1101/2021.09.16.460667

Barry, P., Broquet, T., & Gagnaire, P.-A. (2022). Age-specific survivorship and fecundity shape genetic diversity in marine fishes. *Evolution Letters, 6*(1), 46-62. doi:10.1002/evl3.265

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)*, *30*(15), 2114–2120. doi: 10.1093/bioinformatics/btu170

Bremer, J. R. A., Viñas, J., Mejuto, J., Ely, B., & Pla, C. (2005). Comparative phylogeography of Atlantic bluefin tuna and swordfish: the combined effects of vicariance, secondary contact, introgression, and population expansion on the regional phylogenies of two highly migratory pelagic fishes. *Molecular Phylogenetics and Evolution*, *36*(1), 169–187. doi: 10.1016/J.YMPEV.2004.12.011

Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K., & Madden T.L. (2008). BLAST+: architecture and applications. *BMC Bioinformatics*, 10,421.

Caballero-Huertas, M., Frigola-Tepe, X., Coll, M., Munoz, M., Vinas, J. (2022). The current knowledge status of the genetic population structure of the European sardine (Sardina pilchardus): uncertainties to be solved for an appropriate fishery management. *Reviews in Fish Biology and Fisheries*, 32**,** 745–763. doi: 10.1007/S11160-022-09704-Z

Checkley, D. M., Asch, R. G., & Rykaczewski, R. R. (2017). Climate, Anchovy, and Sardine. *Annual Review of Marine Science*, *9*(1), 469–493. doi: 10.1146/annurev-marine-122414-033819

Chlaida, M., Laurent, V., Kifani, S., Benazzou, T., Jaziri, H. et al. (2009). Evidence of a genetic cline for *Sardina pilchardus* along the Northwest African coast. *ICES Journal of Marine Science*, *66*(2), 264–271. doi: 10.1093/icesjms/fsn206

Chlaida, Malika, Kifani, S., Lenfant, P., & Ouragh, L. (2006). First approach for the identification of sardine populations *Sardina pilchardus* (Walbaum 1792) in the Moroccan Atlantic by allozymes. *Marine Biology*, *149*(2), 169–175. doi: 10.1007/s00227-005-0185-0

Clucas, G. V., Lou, R. N., Therkildsen, N. O., & Kovach, A. I. (2019). Novel signals of adaptive genetic variation in northwestern Atlantic cod revealed by whole-genome sequencing. *Evolutionary Applications*, *12*(10), 1971–1987. doi: 10.1111/eva.12861

Cruickshank, T. E., & Hahn, M. W. (2014). Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology*, *23*(13), 3133–3157. doi: 10.1111/mec.12796

Diaz-Jaimes, P., Uribe-Alcocer, M., Rocha-Olivares, A., Garcia-de-Leon, F. J., Nortmoon, P. et al. (2010). Global phylogeography of the dolphinfish (*Coryphaena hippurus*): The influence of large effective population size and recent dispersal on the divergence of a marine pelagic cosmopolitan species. *Molecular Phylogenetics and Evolution*, *57*(3), 1209–1218. doi: 10.1016/J.YMPEV.2010.10.005

Dray S., Dufour A. (2007). The ade4 Package: Implementing the Duality Diagram for Ecologists. *Journal of Statistical Software*, 22(4), 1–20. doi: 10.18637/jss.v022.i04.

Faria, J., Froufe, E., Tuya, F., Alexandrino, P., & Perez-Losada, M. (2013). Panmixia in the endangered slipper lobster scyllarides latus from the Northeastern Atlantic and Western Mediterranean. *Journal of Crustacean Biology*, *33*(4), 557–566. doi: 10.1163/1937240X-00002158

Faria, R., Weiss, S., & Alexandrino, P. (2012). Comparative phylogeography and demographic history of European shads (*Alosa alosa* and *A. fallax*) inferred from mitochondrial DNA. *BMC Evolutionary Biology*, *12*(1), 194. doi: 10.1186/1471-2148-12-194

Faria, R., Johannesson, K., & Stankowski, S. (2021). Speciation in marine environments: Diving under the surface. *Journal of Evolutionary Biology*, *34*(1), 4–15. doi: 10.1111/jeb.13756

Fox, E.A., Wright, A.E., Fumagalli, M., & Vieira, F.G. (2019). ngsLD: evaluating linkage disequilibrium using genotype likelihoods. *Bioinformatics*, 35(19):3855 – 3856. doi.org/10.1093/bioinformatics/btz200

Francisco, S. M., Faria, C., Lengkeek, W., Vieira, M. N., Velasco, E. M. et al. (2011). Phylogeography of the shanny *Lipophrys pholis* (Pisces: Blenniidae) in the NE Atlantic records signs of major expansion event older than the last glaciation. *Journal of Experimental Marine Biology and Ecology*, *403*(1–2), 14–20. doi: 10.1016/J.JEMBE.2011.03.020

Gonzalez, E. G., & Zardoya, R. (2007). Relative role of life-history traits and historical factors in shaping genetic population structure of sardines (*Sardina pilchardus*). *BMC Evolutionary Biology*, *7*(1), 197. doi: 10.1186/1471-2148-7-197

Grant, W., & Bowen, B. (1998). Shallow population histories in deep evolutionary lineages of marine fishes: insights from sardines and anchovies and lessons for conservation. *Journal of Heredity*, *89*(5), 415–426. doi: 10.1093/jhered/89.5.415

Graves, J. (1998). Molecular insights into the population structures of cosmopolitan marine fishes. *Journal of Heredity*, *89*(5), 427–437. doi: 10.1093/jhered/89.5.427

ICES. (2013). *Report of the Working Group on Southern Horse Mackerel, Anchovy and Sardine (WGHANSA).*

ICES. (2020). *Report of the Working Group on Southern Horse Mackerel, Anchovy and Sardine (WGHANSA).*

Jemaa, S., Bacha, M., Khalaf, G., Dessailly, D., Rabhi, K. and Amara, R. (2015) What can otolith shape analysis tell us about population structure of the European sardine, *Sardina pilchardus*, from Atlantic and Mediterranean waters? *Journal of Sea Research*, 96, 11–17. Doi: 10.1016/j.seares.2014.11.002.

Kasapidis, P., Silva, A., Zampicinini, G., & Magoulas, A. (2011). Evidence for microsatellite hitchhiking selection in European sardine (*Sardina pilchardus*) and implications in inferring stock structure. *Scientia Marina*, *76*(1), 123–132.

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics*, *15*(1), 356. doi: 10.1186/s12859-014-0356-4

Korneliussen, T. S., Moltke, I., Albrechtsen, A., & Nielsen, R. (2013). Calculation of Tajima's D and other neutrality test statistics from low depth next-generation sequencing data. *BMC Bioinformatics*, *14*(1), 289. doi: 10.1186/1471-2105-14-289

Kumar, R., & Kumar, V. (2018). A review of phylogeography: biotic and abiotic factors. *Geology, Ecology, and Landscapes*, *2*(4), 268–274. doi: 10.1080/24749508.2018.1452486

Laurent, V., Caneco, B., Magoulas, A., & Planes, S. (2007). Isolation by distance and selection effects on genetic structure of sardines *Sardina pilchardus* Walbaum. *Journal of Fish Biology*, *71*(sa), 1–17. doi: 10.1111/j.1095-8649.2007.01450.x

Lavretsky, P., McCracken, K. G., & Peters, J. L. (2014). Phylogenetics of a recent radiation in the mallards and allies (Aves: Anas): inferences from a genomic transect and the multispecies coalescent. *Molecular Phylogenetics and Evolution*, *70*, 402–411. doi: 10.1016/j.ympev.2013.08.008

Leigh, JW, Bryant D (2015). PopART: Full-feature software for haplotype network construction. *Methods Ecol Evol*, 6(9),1110–1116.

Li H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v1 [q-bio.GN]

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J. et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. doi: 10.1093/bioinformatics/btp352

Louro, B., Moro, G. De, Garcia, C. M. E. V. R., Cox, C., Verissimo, A. et al. (2018). A haplotype-resolved draft genome of the European sardine (*Sardina pilchardus*). *BioRxiv*, 441774. doi: 10.1101/441774

Machado, A., Torresen, O., Kabeya, N., Couto, A., Petersen, B. et al. (2018). "Out of the Can": A Draft Genome Assembly, Liver Transcriptome, and Nutrigenomics of the European Sardine, *Sardina pilchardus*. *Genes*, *9*(10), 485. doi: 10.3390/genes9100485

Magoulas, A., Castilho, R., Caetano, S., Marcato, S., & Patarnello, T. (2006). Mitochondrial DNA reveals a mosaic pattern of phylogeographical structure in Atlantic and Mediterranean populations of anchovy (*Engraulis encrasicolus*). *Molecular Phylogenetics and Evolution*, *39*(3), 734–746. doi: 10.1016/J.YMPEV.2006.01.016

Mazet, O., Rodriguez, W., Grusea, S., Boitard, S., & Chikhi, L. (2016). On the importance of being structured: instantaneous coalescence rates and human evolution—lessons for ancestral population size inference? *Heredity*, *116*(4), 362–371. doi: 10.1038/hdy.2015.104

Meisner, J., & Albrechtsen, A. (2018). Inferring Population Structure and Admixture Proportions in Low Depth NGS Data. *BioRxiv*, 302463. doi: 10.1101/302463

Nielsen, R., Korneliussen, T., Albrechtsen, A., Li, Y., & Wang, J. (2012). SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. *PLoS ONE*, *7*(7), e37558. doi: 10.1371%252Fjournal.pone.0037558

Yassine Ouagajjou, & Pablo Presa (2015). The connectivity of Mytilus galloprovincialis in northern Morocco: A gene flow crossroads between continents. Estuarine, Coastal and *Shelf Science*, 152, 1-10. doi: 10.1016/j.ecss.2014.10.016.

Palero, F., Abello, P., Macpherson, E., Gristina, M., & Pascual, M. (2008). Phylogeography of the European spiny lobster (*Palinurus elephas*): Influence of current oceanographical features and historical processes. *Molecular Phylogenetics and Evolution*, *48*(2), 708–717. doi: 10.1016/J.YMPEV.2008.04.022

Parrish, R. H., Serra, R., & Grant, W. S. (1989). The Monotypic Sardines, *Sardina* and *Sardinops*: Their Taxonomy, Distribution, Stock Structure, and Zoogeography. *Canadian Journal of Fisheries and Aquatic Sciences*, *46*(11), 2019–2036. doi: 10.1139/f89-251

Patarnello, T., Volckaert, F. A. M. J., & Castilho, R. (2007). Pillars of Hercules: is the Atlantic-Mediterranean transition a phylogeographical break? *Molecular Ecology*, *16*(21), 4426–4444. doi: 10.1111/j.1365-294X.2007.03477.x

Perez-Ruzafa, A., Gonzalez-Wanguemert, M., Lenfant, P., Marcos, C., & Garcia-Charton, J. A. (2006). Effects of fishing protection on the genetic structure of fish populations. *Biological Conservation*, *129*(2), 244–255. doi: 10.1016/J.BIOCON.2005.10.040

Pita, A., Perez, M., Balado, M. & Presa, P. (2014). Out of the Celtic cradle: The genetic signature of European hake connectivity in South-western Europe. *Journal of Sea Research*, 93, 90-100. https://doi.org/10.1016/j.seares.2013.11.003.

Pockrandt, C., Alzamel, M., Iliopoulos, C.S., Reinert, K. (2020) GenMap: Ultra-fast Computation of Genome Mappability. *Bioinformatics*, 2020. doi.org/10.1093/bioinformatics/btaa222

Ravinet, M., Faria, R., Butlin, R. K., Galindo, J., Bierne, N. et al. (2017). Interpreting the genomic landscape of speciation: a road map for finding barriers to gene flow. *Journal of Evolutionary Biology*, *30*(8), 1450–1477. doi: 10.1111/jeb.13047

Ramon, M.M. & Castro J.A. (1997). Genetic variation in natural stocks of *Sardina pilchardus* (sardines) from the western Mediterranean Sea. Heredity, 78, 520—528

Sa-Pinto, A., Branco, M., Sayanda, D. and Alexandrino, P. (2008), Patterns of colonization, evolution and gene flow in species of the genus *Patella* in the Macaronesian Islands. *Molecular Ecology*, 17: 519-532. doi.org/10.1111/j.1365-294X.2007.03563.x

Sala, I., Caldeira, R. M. A., Estrada-Allis, S. N., Froufe, E., & Couvelard, X. (2013). Lagrangian transport pathways in the northeast Atlantic and their environmental impact. *Limnology and Oceanography: Fluids and Environments*, *3*(1), 40–60. doi: 10.1215/21573689-2152611

Skotte, L., Korneliussen, T. S., & Albrechtsen, A. (2013). Estimating individual admixture proportions from next generation sequencing data. *Genetics*, genetics.113.154138-. doi: 10.1534/genetics.113.154138

Smedley, D., Haider, S., Ballester, B. *et al.* (2009). BioMart – biological queries made easy. *BMC Genomics*. 10, 22. https://doi.org/10.1186/1471-2164-10-22

Spanakis, E., Tsimenides, N., & Zouros, E. (1989). Genetic differences between populations of sardine, *Sardina pilchardus*, and anchovy, *Engraulis encrasicolus*, in the Aegean and Ionian seas. *Journal of Fish Biology*, *35*(3), 417–437. doi: 10.1111/j.1095-8649.1989.tb02993.x

Stefanni, S., Castilho, R., Sala-Bozano, M., Robalo, J. I., Francisco, S. M. et al. (2015). Establishment of a coastal fish in the Azores: recent colonisation or sudden expansion of an ancient relict population? *Heredity*, *115*(6), 527–537. doi: 10.1038/hdy.2015.55

Tine, M., Kuhl, H., Gagnaire, PA. *et al.* European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation. *Nat Commun* **5,** 5770 (2014). https://doi.org/10.1038/ncomms6770

Tinti, F., Di Nunno, C., Guarniero, I., Talenti, M., Tommasini, S. et al. (2002). Mitochondrial DNA sequence variation suggests the lack of genetic heterogeneity in the Adriatic and Ionian stocks of *Sardina pilchardus*. *Marine Biotechnology (New York, N.Y.)*, *4*(2), 163–172. doi: 10.1007/s10126-002-0003-3

Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z.X. et al (2010) Sequencing of 50 human exomes reveals adaptation to high altitude. *Science*, 329(5987), 75-8. doi: 10.1126/science.1190371.

Author contributions

R.D.F. and L.F.C. designed the study; F.T., M.N., S.A., M.P., I.R., P.C., A.J-R., M.T.G.S. organized and executed the sample collection; P.F.C., A.R-I. and E.F. performed the laboratory work; R.D.F. analyzed

the data with contributions from G.B., L.B., R.F., A.M.M.; R.D.F., P.F.C., E.F. and L.F.C. wrote the manuscript with contributions from all authors. All authors have read and approved the manuscript.