Low coverage whole genome sequencing reveals the underlying structure of European sardine populations

Population genomics of European sardines

Rute R. da Fonseca[1,2*], Paula F. Campos[2,3], Alba Rey-Iglesia[4], Gustavo V. Barroso[5], Lucie A. Bergeron[6], Manuel Nande[3], Fernando Tuya[7], Sami Abidli[8,9], Montse Pérez[11], Isabel Riveiro[11], Pablo Carrera[11], Alba Jurado-Ruzafa[12], M. Teresa G. Santamaría[12], Rui Faria[3+], André M. Machado[3], Miguel M. Fonseca[3], Elsa Froufe[3], L. Filipe C. Castro[3,13*]

[1]Center for Macroecology, Evolution and Climate, The Globe Institute, University of Copenhagen, Denmark

[2]The Bioinformatics Centre, Department of Biology, University of Copenhagen, Copenhagen, Denmark

[3]CIIMAR – Interdisciplinary Centre of Marine and Environmental Research – University of Porto, Porto, Portugal

[4]Centre for GeoGenetics, Natural History Museum Denmark, University of Copenhagen, Østervoldgade 5-7, 1350 Copenhagen, Denmark

[5]Department of Ecology and Evolutionary Biology, University of California, Los Angeles, USA

[6]Section for Ecology and Evolution, University of Copenhagen, Denmark

[7]Grupo en Biodiversidad y Conservación, IU-ECOAQUA, Universidad de Las Palmas de Gran Canaria, Las Palmas, 35017, Canary Islands, Spain

[8]Institut Supérieur des Sciences Biologiques Appliquées de Tunis (ISSBAT), Université de Tunis El Manar, Tunisia

[9]Faculté des Sciences de Bizerte (FSB), Zarzouna7021, Université de Carthage, Tunisia

[11]Instituto Español de Oceanografía, Centro Oceanográfico de Vigo, 36390 Vigo, Spain

[12]Centro Oceanográfico de Canarias, Instituto Español de Oceanografía, Dársena Pesquera, Santa Cruz de Tenerife 38180, Spain

[13]Department of Biology, Faculty of Sciences, U. Porto – University of Porto, Portugal

[+]Currently at CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, InBIO, Laboratório Associado, Universidade do Porto, Vairão, Portugal

*Corresponding authors

Rute R da Fonseca: rfonseca@sund.ku.dk

L. Filipe C. Castro: filipe.castro@ciimar.up.pt

Abstract

Whole genome sequence data is an ideal tool for characterizing processes in ecology and evolution. Despite the lowering in sequencing costs, it can be challenging to produce a genome and high-coverage resequencing data for a non-model species. New population genomics data analysis pipelines based on genotype likelihoods allow for a significant reduction in cost by efficiently extracting information from low coverage sequence data. We demonstrate the robustness of such approaches with a genomic data set consisting of two draft genomes of the European sardine (*Sardina pilchardus*, Walbaum 1792), and resequencing data (~1.5 X depth) for 78 individuals from 12 sampling locations across the 5,000 Km of the species' distribution range (from the Eastern Mediterranean to the archipelagos of Madeira and Azores). Our results clearly show at least three genetic clusters. One includes individuals from Azores and Madeira (two archipelagos in the Atlantic), the second corresponds to Iberia (the center of the sampling distribution), and the third gathers the Mediterranean samples and those from the Canary Islands. This suggests at least two important barriers to gene flow, even though these do not seem complete, with individuals from Iberia showing some degree of admixture. These results together with the genetic resources generated for this commercially important taxon provide a baseline for further studies aiming at identifying the nature of these barriers between Sardine populations, and information for transnational stock management of this highly exploited species towards sustainable fisheries.

Keywords

European sardine, low coverage sequencing, population structure, marine fish

Background

Population structuring in the absence of obvious physical barriers have puzzled biologists for centuries. In oceanic environment, strong genetic structure is mainly expected in species with limited ability to disperse and/or philopatric. Most marine animals are capable of long-distance dispersal facilitated by the combination of a long larval pelagic phase, high fecundity, large population sizes and adult migratory behavior, all of which contribute to low genetic structure (J. Faria, Froufe, Tuya, Alexandrino, & Pérez-Losada, 2013). Yet, many studies have shown that several species have higher spatial genetic differentiation than expected considering their high dispersal potential (Palero, Abelló, Macpherson, Gristina, & Pascual, 2008; Pérez-Ruzafa, González-Wangüemert, Lenfant, Marcos, & García-Charton, 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 in the absence of complete physical barriers (Rui Faria, Johannesson, & Stankowski, 2021). This is especially relevant for conservation efforts, including stock management of commercially important species (J. 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. The Almeria-Oran Front, a well-defined oceanographic break situated east of the Strait of Gibraltar, has been pointed 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, Volckaert, & Castilho, 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. 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, Caldeira, Estrada-Allis, Froufe, & Couvelard, 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 (J. 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). Despite the importance of the sardine fishery, stock delineation for management purposes is still a matter of debate throughout the species distribution range (ICES 2006, FAO 2008), and is especially relevant in the recent scenario low biomass level [ICES, 2020]. The European sardine occurs from the southern Celtic Sea and the North Sea to Mauritania and Senegal, including the Azores, Madeira and the Canary Archipelagos, being also abundant in the Mediterranean (Parrish, Serra, & Grant, 1989). As other marine pelagic fish, *S. pilchardus* shows schooling and migratory behavior, large effective population sizes and great dispersal capabilities, both at the

larval and adult stages. In line with expectations from its life-history traits, some population genetics studies focusing on various sites across its global distribution have detected very low levels of genetic differentiation using allozymes (M. Chlaida et al., 2009; Malika Chlaida, Kifani, Lenfant, & Ouragh, 2006; Laurent, Caneco, Magoulas, & Planes, 2007; Spanakis, Tsimenides, & Zouros, 1989), mitochondrial DNA (mtDNA) (Atarhouch et al., 2006; Tinti et al., 2002), and microsatellites (Gonzalez & Zardoya, 2007; Kasapidis, Silva, Zampicinini, & Magoulas, 2011). However, the observed phenotypic variation in gill raker counts and head length led to the division in two subspecies of sardine, S. *pilchardus pilchardus*, present in the eastern Atlantic, from the North Sea to southern Portugal, and *S. pilchardus sardina* in the Mediterranean Sea and northwest African coast (Andreu & B., 1969; Parrish et al., 1989). This division was also supported by mitochondrial haplotype frequency differences (Atarhouch et al., 2006), with a suggestion for a contact zone around the Strait of Gibraltar (Atarhouch et al., 2006).

Additionally, a study covering most of the European sardine Atlantic range using allozymes and microsatellites detected a significant differentiation of the peripheral populations from Madeira and the Azores (Kasapidis et al., 2011). These apparent contradictions probably have multiple causes, including differences in terms of samples sizes, geographic coverage of populations, number and resolution power of the markers used, and tools used to characterize population differentiation and substructure among the different studies. Thus, studies that incorporate more even sample sizes of populations spanning the species distribution range, using a large number of markers across the genome, will be important to provide a more consistent and complete picture of the genetic structure of European sardine populations. The lowering in sequencing costs and the development of bioinformatics tools to analyze large volumes of data

now allows for the use of a wide range of sequence data in non-model organisms research (R.R. da Fonseca et al., 2016), increasing the power and the resolution of population-level comparisons. However, whole-genome resequencing of several individuals across many populations is still not accessible for most research groups on population genomics, especially for species with large genome size. One cost-effective approach that has been successfully applied in population genomics is the use of low-coverage whole-genome sequencing data (Clucas, Lou, Therkildsen, & Kovach, 2019; Rute R. da Fonseca et al., 2019).

In this study, a genomic data set consisting of low coverage sequencing data was produced for the European sardine to investigate the population structure across its entire range, i.e., North-West Atlantic, including the Macaronesian Archipelagos of Madeira, Azores and Canaries, and the Mediterranean Sea. Consistent results were observed using different subsets of the data after conservative filtering of the available draft genome assemblies. The access to patterns of genetic variation across the genome allowed the identification of the few genomic regions that seem resistant to gene flow, where the species' genetic structure is more evident. An initial analysis of admixture patterns was used to select unadmixed individuals for high coverage sequencing to assess fluctuations in population size through time and recombination rates in a cost-effective approach. Whole mitochondrial genomes were also recovered from the low coverage data, enabling a comparison between markers with different modes of inheritance.

Materials and Methods

Sample collection and DNA extraction

Samples were collected from 12 different geographical locations encompassing the species' current distribution range (Figure 1). Samples from three locations (SpainN, SpainSE and SpainSW; n=15) were collected during oceanographic surveys. The remaining 68 specimens, from nine distinct geographic locations, were sampled at local markets (Table S1).

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 ºC for 10minutes, to increase DNA yield. DNA concentration and purity were verified using a Nanodrop Spectrophotometer and a Qubit Fluorometer. Sequencing data was commercially obtained at Novogene (China).

Sequencing data pre-processing

To assess the patterns of genetic differentiation of the European sardine, 78 samples were sequenced to around 2 X depth of coverage (Additional file 1: Table S1) and six of these samples were further re-sequenced to 20 X depth. We first produced a dataset of whole genome sequences for 53 individuals (low depth, ~1.5 X) from 12 sampling locations across the 5000 Km distribution range of the species. After an initial assessment of the pattern of population structure, we selected the six seemingly unadmixed individuals (A1, M1, T1, T6, G5 and G8) for sequencing to high depth to allow for analysis based on genotype calls, such as the PSMC approach to estimate variation of effective population size through time (Li & Durbin, 2011), and further re-sequenced 25 individuals from the populations at the edges of the distribution (Azores, Madeira, Canary Islands, Trieste and Greece) as these were shown to be part of two

separate clusters relative to the individuals in the center of the distribution, overrepresented in the first batch. We also processed the sequencing data from a sample used to assemble the genome of *S. pilchardus* using an all-Illumina approach [1] (depth of 45X; sampling location: Porto). Raw Illumina reads were first processed with Trimmomatic (version 0.36) (Bolger, Lohse, & Usadel, 2014) for removal of adapter sequences and trimming bases with quality <20 and discarded reads with length <80. Clean reads were mapped to two *S. pilchardus* reference genomes (Louro et al., 2018; Machado et al., 2018). A fully assembled mitochondria from Machado et al. (2018) was added to the assembly from Louro et al. (2018) after removal of individual contigs matching mitochondrial DNA. Reads showing a mapping hit were further filtered for mapping quality >25. PCR duplicates were removed with Picard MarkDuplicates (version 1.95; http://picard.sourceforge.net) and local realignment around indels was done with GATK (DePristo et al., 2011).

The subsequent methods description and results presented in the main paper refer to analyses performed using the assembly from Louro et al. (2018). Details regarding the assembly from Machado et al. (2018) are in the Supplementary Information.

Assembly filtering

After calculating the depth of coverage per scaffold using the individual with sequence data of 45X, only scaffolds with depth between 35 X and 65 X were considered to avoid misassembled regions. Furthermore, to avoid including contigs that can be assigned to sex chromosomes, we further removed scaffolds that did not have at least 90% of the depth of scaffold 1 (the largest and most representative of the assembly) in all individuals, resulting in 405 scaffolds. Within

these, the 50 scaffolds that contained more than 100,000 informative sites (sites that do not overlap repeat annotations, with maximum 20% of missing data in each population, only considering bases with mapping quality > 30 and base quality >20) were used in all subsequent analyses. Methods appropriate for low coverage NGS data (Korneliussen, Albrechtsen, & Nielsen, 2014; Meisner & Albrechtsen, 2018; Skotte, Korneliussen, & Albrechtsen, 2013; Soraggi, Wiuf, & Albrechtsen, 2018) were used throughout the analyses and applied to all samples. The regions corresponding to repeats were masked in the reference genomes and excluded from subsequent analyses.

Population structure

NGSadmix version 32 (Skotte et al., 2013) was used to detect population structure using autosomal data. NGSadmix infers population structure using genotype likelihoods (that contain all relevant information on the uncertainty of the underlying genotype (R.R. da Fonseca et al., 2016). As it does not require the exact genotypes, it is adequate for low-depth sequencing data (Skotte et al., 2013). NGSadmix was run for K equal 2 and 3 for sites present in a minimum of 75% of the individuals: a total of 16,683 SNP sites for the 78 samples. The program was run with different seed values until convergence was reached.

A principal component analysis (PCA) using the same SNP set was done with PCAngsd (Meisner & Albrechtsen, 2018) which estimates the covariance matrix for low depth NGS data in an iterative procedure based on genotype likelihoods. Genotype likelihoods for all individuals were generated with ANGSD (Korneliussen et al., 2014) (options -GL 1 -doGlf 2 -minQ 20 -minMapQ 30 -minInd 63 -minMaf 0.05 -C 50 -baq 2 -remove_bads 1 -uniqueOnly 1 -SNP_pval 1e-6).

Complete mitochondrial consensus sequences were obtained from the shotgun resequencing data by choosing the most common base per position (-doFasta 2 in ANGSD (Korneliussen et al., 2014)).The software RAxML (Stamatakis, 2014) with 100 rapid bootstrap replicates was used to estimate a maximum likelihood mitochondrial phylogenetic tree under the GTR+GAMMA model of sequence evolution. The tree figure was produced in iTol (Letunic & Bork, 2011).

Assessment of genetic diversity and population differentiation

We used methods based on the site frequency spectrum (SFS) (Korneliussen, Moltke, Albrechtsen, & Nielsen, 2013; Nielsen, Korneliussen, Albrechtsen, Li, & Wang, 2012) to estimate nucleotide diversity, the neutrality test statistic Tajima's D and genome-wide fixation index (F_{ST}) values. Briefly, after estimating the SFS, posterior sample allele frequencies are calculated using the global SFS as prior. SFSs estimated separately were used to obtain joint SFSs for population pairs, which are then used to estimate F_{ST} using the option -whichFst 1. Sites considered for analyses were allowed to have a maximum of 1.5 X of the median depth across all samples, and 20% missing data.

Variant calling and historical effective population size estimation

Variants were called for the seven individuals for which high coverage data was available using GATK version 4.0.7.0 (Van der Auwera et al., 2013). Briefly, first variants were 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. We used the default filter of GATK (--phred-scaled-global-read-mismapping-rate 45;--base-quality-score-threshold 18; --min-base-quality-score 10).

We estimated the historical effective population size (N_e) with PSMC (Li & Durbin, 2011) using: i) the “mpileup” command in samtools (Li et al., 2009) with options -C50, -Q 30 -q 30; bcftools “view -c” command ; the vcfutils.pl “vcf2fq” option, masking regions of low coverage (less than half of the mean depth per sample) and excessive coverage (more than twice the mean). The PSMC inference was done using the recommended number of time intervals of 64 (-p "4+25*2+4+6") (Li & Durbin, 2011), an upper limit of time to the most recent common ancestor (TMRCA) of 15 and an initial value of $r = \theta/\rho$ of 5. Further 100 bootstraps replicates were done per sample to assess the variance of the *Ne* estimates. The results were scaled using an estimate of the generation time of 2 years and a mutation rate of 2E-09 substitutions per base per generation (estimate for *Clupea harengus* (Feng et al., 2017)).

Results

Population structure across the distribution range

The results show that the European sardines seem to be part of at least three structured populations (Figure 2A). When setting the number of expected clusters to 2 (k=2; Figure 2A, 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 2A, 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, Tunisia and Trieste. Azores shows two individuals with ancestry from the Central region; Canaries’ individuals show some admixed ancestry with the Western cluster; and one

individual from Greece showed some ancestry from the Western cluster. The same clustering was observed in a preliminary analysis done with a subset of 53 individuals mapped to a different genome assembly (see Methods for details). We obtained similar results regarding the population structure analysis with two drafts assemblies of scaffold N50s below 100 kb (Figure 2A and Figure S1), after using stringent filtering to remove low quality scaffolds.

The organization in three separate clusters can also be observed in the principal component analysis (PCA; Figure 2B). The first two PCs explained 9.8 % and 7.2 % 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.

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.

Assessment of genetic diversity and population differentiation

There were high levels of genetic diversity in all sampling locations (Table 1), in agreement with

previous results showing that *S. pilchardus* has the highest genome-wide heterozygosity compared to other fish species with similar geographic distributions (Barry, Broquet, & Gagnaire, 2020). The populations in the Center were the most diverse consistent with the observed patterns of admixture (Figure 2). In general, we observed lower values of genetic differentiation as measured by F_{ST} for comparisons within regions (distances ranging from 144 to 2,033 Km show F_{ST} values between 0 and 0.03), the exception being the values observed for the pairs including the Canary Islands and the locations on the East (Figure 4 and Table S2). The highest values of F_{ST} included comparisons with Madeira (West) and sampling locations in the Center (top value of 0.96). A correlation between F_{ST} and distance can only be observed for comparisons between the Center and the East, with the highest values pertaining to comparisons with Greece, and the lowest values associated with Tunisia (Figure S3).

We also assessed the variation in F_{ST} along the genome (Figure S4) and found that regions with elevated F_{ST} variance reflect the pattern of population structure observed in Figure 2, whereas regions of low F_{ST} variance agree with the observed clustering in the mitochondrial data, where the Center and the East are indistinguishable in terms of individual haplotypes (Figure S5). The F_{ST} along the genome is generally positively correlated with the recombination rate estimated using the iSMC approach from (Barroso, Puzović, & Dutheil, 2019) (Figure S6), however areas of the genome that have a high F_{ST} show lower recombination rates, and areas with very low F_{ST} are associated with the highest recombination rates (Figure S7).

Temporal dynamics of effective population size

All individuals (West, Center and East) share a similar demography until 400,000 years ago (ya);

291 at that point, the effective population size (N_e) starts to increase. N_e peaks at around 150,000 ya
292 for individuals from the Atlantic (West and Center), which then see N_e decreasing to its
293 minimum, remaining stable during the glacial period. The N_e of the Mediterranean populations
294 (East) increased much more (2-to >10x) than that of the Atlantic ones and remained high until
295 later. The decline in N_e started at around ~60,000 years ago, but the bootstrap analysis revealed
296 high uncertainty in the estimate of N_e for recent times (Figure S8), unlike what was observed for
297 the Atlantic individuals. Nevertheless, in most cases, the N_e of the populations from the East
298 remained higher than those in the Atlantic.

299 ## Discussion

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

307 Altogether, the assessment of nuclear genome sequences by means of individual ancestry
308 information, principal component analysis (Figure 2) and differentiation (F_{ST}) among populations
309 from different geographic regions (Figure 4), supports that the European sardine comprises
310 three main stocks: “West” that includes individuals from Azores and Madeira (part of the
311 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 (Figure 2). 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 & B., 1969; Atarhouch et al., 2006; Parrish et al., 1989), suggesting the existence of a phylogeographic break between the South of Portugal and Mediterranean populations. However, our work shows that the Spanish Mediterranean populations belong to the Central (Iberian) and not to the Mediterranean genetic cluster, while the population from the Canary Islands has a Mediterranean ancestry. Thus, although the regions around the Strait of Gibraltar and the Almeria-Oran Front have been suggested to form a phylogeographic break for this species, our work shows that this is not a complete barrier.

The differentiation between Azores/Madeira and the other populations shown here is in agreement with the results from Kasapidis et al (2011). 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. Whether this genetic division is caused by currents, isolation by distance and lack of suitable habitat between these regions, local adaptation to different environmental conditions or other reasons, needs to be further investigated.

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, McCracken, & Peters, 2014). Patterns suggesting admixture between the three genetic clusters were also observed with the nuclear data in all populations except Madeira. However, we cannot exclude that this could instead be the result of incomplete lineage sorting. 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. However, a similar pattern of strong admixture is observed in nuclear contigs showing low F_{ST} across all comparisons (Figure S5). Although the contrast of differentiation between genomic regions with high vs low F_{ST} could suggest that that the former genomic regions potentially retain the ancestral pattern of structure, because of the low recombination rates that hinder introgression, which would contrast with the latter that could be eroded by gene flow between populations. However, similar signatures can be obtained if there is low diversity in low recombination regions, where the effect of background selection is stronger resulting in inflated F_{ST} values (Cruickshank & Hahn, 2014; Ravinet et al., 2017). The fact that F_{ST} values across the genome seem to follow the same trend across different population comparisons (Figure S4), points towards the influence of the same processes and genomic features (e.g., recombination rate) across populations.

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 from the East and Center clusters (Figure S3), as expected under a process of isolation by distance, suggesting that indeed at least some of the patterns observed with nuclear and mtDNA genomes can indeed be created by gene flow between populations from different genetic clusters except Madeira. Although this needs to be further confirmed using model-based approaches, if true, it provides additional support that the genetic barriers involved in the differentiation between these three genetic clusters are only partial. Under this scenario, the admixture observed in the Canary Islands could be a result of gene flow from Madeira, while the admixture in the Azores could have resulted from gene flow with the Central populations, which in this case seems to be bi-directional. Furthermore, the admixture observed between populations from the Central and Eastern clusters could suggest bidirectional gene flow between populations from Iberia and Mediterranean populations outside Iberia.

Patterns of admixture suggesting 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, Castilho, Caetano, Marcato, & Patarnello, 2006), a species that shares several traits with sardines (Checkley, Asch, & Rykaczewski, 2017), could potentially explain the admixed ancestry of the Canary Islands and the Eastern cluster (Figure 2). A recolonization event during the Eemian interglacial period (130-

115 kya) could have led to admixture with populations from Mediterranean refugia, resulting in the spike in N_e observed in Mediterranean populations (Figure 5) over that period, as variation in the temporal values of N_e can be caused by gene flow between populations (Mazet, Rodríguez, Grusea, Boitard, & Chikhi, 2016). Indeed, population structure is known to influence genetic variation in a way that mimics population size change (Rodríguez et al., 2018). However, we cannot exclude that the higher N_e observed in the Mediterranean individuals, which was remained higher for a longer period than in Atlantic populations (Figure 5), could have resulted from favorable and stable environmental conditions for sardines in the Mediterranean during the last glacial period, as suggested for other Clupeid species (R. Faria, Weiss, & Alexandrino, 2012).

Conclusions

Unlike reduced-representation approaches (Andrews, Good, Miller, Luikart, & Hohenlohe, 2016), low depth whole-genome resequencing allows for sequencing in smaller batches, which can be very cost-effective, besides having a much lower turnover time, allowing for a flexible project design. In this study, we took advantage of state-of-the-art bioinformatic tools to extract information from low depth sequencing data, which enabled us to assess patterns of genetic variation across the genome of the European sardine. We were able to show how different parts of the nuclear genome yield population structure and diversity patterns congruent with those observed in previous studies that were seemingly contradictory. This approach also recovers full mitochondrial genomes for comparison with genetic sequences with

different modes of inheritance, also previously highlighted as being essential for a complementary insight on sardine population history (Gonzalez & Zardoya, 2007).

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 are 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 taxon, offers precious information 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, 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 Danish National Research Foundation for its funding of the Center for Macroecology, Evolution, and Climate (grant DNRF96). 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 the project VALORMAR (reference nr. 24517), supported by COMPETE2020, LISBOA2020, ALGARVE2020, PORTUGAL2020, through ERDF. It was also supported by the 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

Alvarado Bremer, J. R., 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

Andreu, B., & 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

Atarhouch, T., Rüber, L., Gonzalez, E. G., Albert, E. M., Rami, M., Dakkak, A., & Zardoya, R. (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/

455 j.ympev.2005.08.003

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

459 Barry, P., Broquet, T., & Gagnaire, P.-A. (2020). Life tables shape genetic diversity in marine
460 fishes. *BioRxiv*, 2020.12.18.423459. doi: 10.1101/2020.12.18.423459

461 Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina
462 sequence data. *Bioinformatics (Oxford, England)*, *30*(15), 2114–2120. doi:
463 10.1093/bioinformatics/btu170

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

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

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

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

475 Cruickshank, T. E., & Hahn, M. W. (2014). Reanalysis suggests that genomic islands of speciation
476 are due to reduced diversity, not reduced gene flow. *Molecular Ecology*, *23*(13), 3133–
477 3157. doi: 10.1111/mec.12796

478 da Fonseca, R.R., Albrechtsen, A., Themudo, G. E., Ramos-Madrigal, J., Sibbesen, J. A., Maretty,

L., … Pereira, R. J. (2016). Next-generation biology: Sequencing and data analysis approaches for non-model organisms. *Marine Genomics*, *30*, 3–13. doi: 10.1016/j.margen.2016.04.012

da Fonseca, Rute R., Ureña, I., Afonso, S., Pires, A. E., Jørsboe, E., Chikhi, L., & Ginja, C. (2019). Consequences of breed formation on patterns of genomic diversity and differentiation: the case of highly diverse peripheral Iberian cattle. *BMC Genomics*, *20*(1), 334. doi: 10.1186/s12864-019-5685-2

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V, Maguire, J. R., Hartl, C., … Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491–498. doi: 10.1038/ng.806

Díaz-Jaimes, P., Uribe-Alcocer, M., Rocha-Olivares, A., García-de-León, F. J., Nortmoon, P., & Durand, J. D. (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

Faria, J., Froufe, E., Tuya, F., Alexandrino, P., & Pérez-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, Rui, 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

Feng, C., Pettersson, M., Lamichhaney, S., Rubin, C.-J., Rafati, N., Casini, M., … Andersson, L.

(2017). Moderate nucleotide diversity in the Atlantic herring is associated with a low mutation rate. *ELife*, *6*. doi: 10.7554/eLife.23907

Francisco, S. M., Faria, C., Lengkeek, W., Vieira, M. N., Velasco, E. M., & Almada, V. C. (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)*.

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

Letunic, I., & Bork, P. (2011). Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Research*, *39*(Web Server issue), W475-8. doi: 10.1093/nar/gkr201

Li, H., & Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, *475*(7357), 493–496. doi: 10.1038/nature10231

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (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., Veríssimo, A., Sabatino, S. J., … Canario, A. V. M. (2018). A haplotype-resolved draft genome of the European sardine (Sardina pilchardus). *BioRxiv*, 441774. doi: 10.1101/441774

Machado, A., Tørresen, O., Kabeya, N., Couto, A., Petersen, B., Felício, M., … C. Castro, L. F. (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., Rodríguez, 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. Retrieved from http://dx.doi.org/10.1371%252Fjournal.pone.0037558

Palero, F., Abelló, 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

Pérez-Ruzafa, Á., González-Wangüemert, M., Lenfant, P., Marcos, C., & García-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

Ravinet, M., Faria, R., Butlin, R. K., Galindo, J., Bierne, N., Rafajlović, M., … Westram, A. M. (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

Rodríguez, W., Mazet, O., Grusea, S., Arredondo, A., Corujo, J. M., Boitard, S., & Chikhi, L. (2018). The IICR and the non-stationary structured coalescent: towards demographic inference with arbitrary changes in population structure. *Heredity*, *121*(6), 663–678. doi: 10.1038/s41437-018-0148-0

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

Soraggi, S., Wiuf, C., & Albrechtsen, A. (2018). Powerful Inference with the D-Statistic on Low-Coverage Whole-Genome Data. *G3 (Bethesda, Md.)*, *8*(2), 551–566. doi: 10.1534/g3.117.300192

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

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. doi: 10.1093/bioinformatics/btu033

Stefanni, S., Castilho, R., Sala-Bozano, M., Robalo, J. I., Francisco, S. M., Santos, R. S., … Mariani, S. (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

Tinti, F., Di Nunno, C., Guarniero, I., Talenti, M., Tommasini, S., Fabbri, E., & Piccinetti, C. (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

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., … DePristo, M. A. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Current Protocols in Bioinformatics*, *43*(1110), 11.10.1-11.10.33. doi: 10.1002/0471250953.bi1110s43

Data Accessibility

The Illumina fastq files raw reads are deposited in *NCBI Sequence Read Archive* (Bioproject accession no.: PRJNA688514, samples accession no.: SRR13325046, SRR13324980, SRR13324981, SRR13324982, SRR13324985, SRR13324986, SRR13324988, SRR13324989, SRR13324991, SRR13324992, SRR13324999, SRR13325000, SRR13325002, SRR13325003, SRR13325007, SRR13325008, SRR13325010, SRR13325011, SRR13325013, SRR13325014, SRR13325015, SRR13325018, SRR13325019, SRR13325021, SRR13325022, SRR13325023, SRR13325024, SRR13325025, SRR13325026, SRR13325027, SRR13325028, SRR13325029, SRR13325030, SRR13325031, SRR13325032, SRR13325033, SRR13325034, SRR13325035, SRR13325036, SRR13325037, SRR13325038, SRR13325039, SRR13325040, SRR13325041, SRR13325042, SRR13325043, SRR13325044, SRR13325045, SRR13325047, SRR13325048, SRR13325049, SRR13325050, SRR13325051, SRR13325052, SRR13325053, SRR13325054, SRR13325055, SRR13325056, SRR13325057, SRR13325058, SRR13325059, SRR13324977, SRR13324978, SRR13324979, SRR13324983, SRR13324984, SRR13324987, SRR13324990, SRR13324993, SRR13324994, SRR13324995, SRR13324996, SRR13324997, SRR13324998, SRR13325001, SRR13325005, SRR13325006, SRR13325004, SRR13325009, SRR13325012, SRR13325016, SRR13325017, SRR13325020, SRR13325060). The data will be public upon acceptance of the manuscript for publication.

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.

Tables

Table 1. Average number of pairwise differences (nucleotide diversity, tP), segregating sites (S) and Tajima's D across all scaffolds with more than 100K informative sites for all individuals.

Population	Putative cluster	tP	S	Tajima's D	Number of sites
Azores	WEST	0.0051	0.0064	-0.99	8,074,633
Madeira	WEST	0.0047	0.0056	-0.73	20,868,807
Spain (N)	CENTER	0.0059	0.0067	-0.93	40,847,936
Vigo	CENTER	0.0058	0.0066	-0.87	24,061,285
Porto	CENTER	0.0057	0.0065	-0.87	17,105,311
Spain (SW)	CENTER	0.0060	0.0068	-0.91	31,995,207
Spain (SE)	CENTER	0.0060	0.0068	-0.88	40,784,255
Canaries	EAST	0.0052	0.0056	-0.44	73,353,851
Tunisia	EAST	0.0055	0.0063	-0.97	40,954,036
Trieste	EAST	0.0053	0.0071	-1.26	23,914,974
Greece	EAST	0.0054	0.0066	-1.02	26,678,378

Figure captions

Figure 1. 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.

Figure 2. A) Population structure plot showing the ancestry of each individual (vertical bar) to two (above) and three (below) genetic clusters. **B)** Distribution of individuals based on the first two components of the principal component analysis. Variance explained by each component is shown in parenthesis.

Figure 3. Maximum likelihood tree (unrooted) obtained using full mitochondrial DNA sequences. Red circles indicate branches with 100 % bootstrap support. Colors represent the main ancestry of each individual (for K=3 as in Figure 2A).

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

Figure 5. Pairwise sequentially Markovian coalescent (PSMC) estimates for seven individuals across the sampling range. The last glacial period before the Holocene (115,000 – c. 11,700 years ago) is highlighted.