TRANSBORAN Project

Transboundary population structure of sardine, European hake and blackspot seabream in the Alboran Sea and adjacent waters: a multidisciplinary approach

Final report on European hake genetics

Kenza Mokhtar-Jamaï (INRH), Alessia Cariani (UniBo), Rachele Corti (UniBo), Elisabetta Piazza (UniBo), Massimiliano Babbucci (University of Padova), Naiara Rodriguez-Ezpeleta (AZTI), Mohammed Malouli Idrissi (INRH)

Coordinators of Genetics: Alessia Cariani (UniBo), Carolina Johnstone (IEO), Kenza Mokhtar-Jamaï (INRH)

Introduction and Background

Defining the genetic spatial boundaries of commercial marine species is a key issue for the sustainable management of fisheries (Waples et al., 2008). Indeed, structured populations may display variations in their life history traits such as fecundity, growth and mortality rates, and their abundance. Therefore, stock assessment should be based on known population boundaries. However, historically stock boundaries have been mainly based on economic, social and political factors instead of biological ones, which can lead to mismatch between biological and fisheries management units (Reiss et al., 2009 and references therein). In the worst scenario it can result in loss of genetic diversity (Hauser et al., 2002) and the ability of the species to adapt to environmental changes and to local population extinction.

The European hake, *Merluccius merluccius*, is widely distributed in the eastern Atlantic Ocean (from Mauritania to Norway) and in the Mediterranean Sea and is one of the most important demersal species in these areas from a commercial point of view (FAO, 2018 ; ICES, 2018). In the Northeast Atlantic Ocean, hake is managed by the International Council for the Exploration of the Sea (ICES) assuming two stocks: the northern stock and the southern stock separated without biological evidence by the Capbreton Canyon. In the Southeast Atlantic Ocean, hake is managed by the Fishery Committee for the Eastern Central Atlantic (CECAF) and Morocco assuming one stock. Within the Mediterranean Sea, hake is managed by countries through the 27 geographical subareas (GSA) defined by the General Fisheries Commission for the Mediterranean (GFCM) assuming one stock per GSA.

Several genetic studies highlighted the need for defining new assessment and management units for the European hake within the Northeast Atlantic Ocean and evidenced that the actual separation between the northern and the southern stock is not supported by genetic data (see for example Lundy et al., 1999; Castillo et al., 2005; Pita et al., 2011, 2014; Milano et al., 2014; Westgaard et al., 2017; Leone et al., 2019). However within the Mediterranean Sea, only few studies aimed at disentangling the population genetic structure of hake with an exhaustive sampling (Cimmaruta et al., 2005; Pita et al., 2010, 2014) but see (Milano et al., 2014) where they found genetic differentiation between Western, Central and Eastern Mediterranean.

In this context, the FAO Copemed II Transboran project aimed at investigating the spatial population structure of sardine, European hake and blackspot seabream in the Alboran Sea and adjacent waters following a multidisciplinary approach. The goal of this project is therefore to identify the stock units and to determine if the current GSA boundaries are appropriate spatial scale of assessment and management for these species.

The aim of the present work, is to determine the population genetic structure of the European hake within the Alboran Sea and adjacent waters using microsatellite and SNP genetic markers.

Materials and methods

Sampling design

Fifteen locations were sampled within the Alboran Sea, and neighbouring Mediterranean waters and from adjacent Atlantic Ocean. A 0.5 cm^3 piece of white skeletal muscle was taken from around 40 hakes per location and stored in non-denaturated ethanol 96% (see Table 1 and Figure 1).

Table 1. Location of Hake samples.

Area	Sampling location	Population ID	GFCM GSA	Sample size
Atlantic Ocean	Agadir (Morocco)	AGA	/	40
Atlantic Ocean	Mehdia (Morocco)	MHD	/	40
Atlantic Ocean	Huelva (Spain)	HUE	/	40
Atlantic Ocean	Cadiz (Spain)	CDZ	/	40
North Alboran Sea	Estepona (Spain)	ETP	1	40
North Alboran Sea	Malaga (Spain)	MLG	1	33
North Alboran Sea	Roquetas (Spain)	RQT	1	38
South Alboran Sea	M'Diq (Morocco)	MDQ	3	40*
South Alboran Sea	Nador (Morocco)	NDR	3	40
Mediterranean Sea	Ghazaouet (Algeria)	GHZ	4	41
Mediterranean Sea	Annaba (Algeria)	ANB	4	41
Mediterranean Sea	Tabarka (Tunisia)	TBK	12	40
Mediterranean Sea	Gulf of Tunis (Tunisia)	GTU	12	40
Mediterranean Sea	Torrevieja (Spain)	TOR	6	40
Mediterranean Sea	Castellon (Spain)	CAS	6	40

* For the SNP dataset only 37 individuals were considered

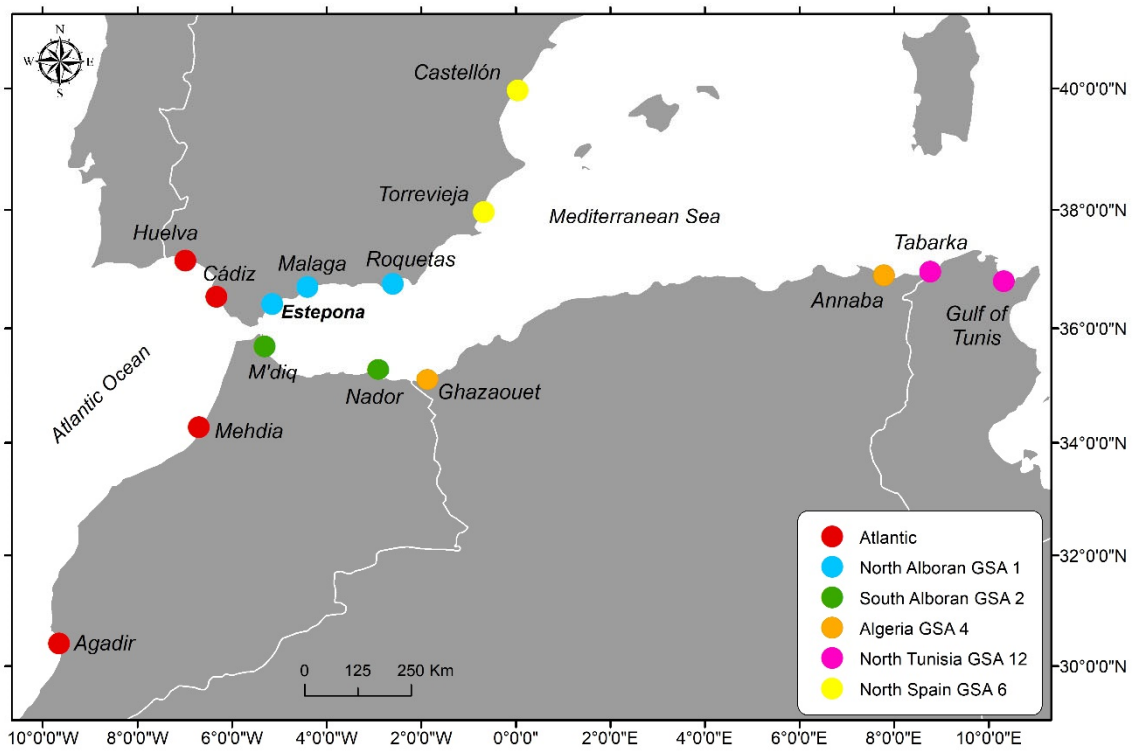


Figure 1. Map of sampling location.

DNA extraction

Total genomic DNA was extracted from 593 individuals using the pureLink DNA Invitrogen Kit, following the manufacturer's protocol.

Microsatellite genotyping

All individuals were genotyped at six microsatellite loci: Mmer-hk20, Mmer-hk9b, Mmer-hk3b, Mmer-hk29, Mmer-hk34b (Morán et al., 1999) and Mmer-UEAW01 (Rico et al., 1997). PCR was carried out in a total volume of 10 µl containing: 10 ng DNA, 0.2 µM of each primer, 0.2 mM of each deoxinucleotide, 1.0 mM $MgCl_2$ (for Mmer-hk29, Mmer-hk34b and Mmer-UEAW01) or 1.5 mM $MgCl_2$ (for Mmer-hk20, Mmer-hk9b and Mmer-hk3b), 1 X PCR buffer and 2 U Platinum Taq DNA Polymerase (Invitrogen). Amplification was performed in a 96-well Applied Biosystems Proflex thermocycler as follows: 94°C for 2 min; 35 cycles of 94°C for 30 s, locus annealing temperature (50°C for Mmer-hk3b and 55°C for Mmer-hk20, Mmer-hk9b, Mmer-hk29 and Mmer-UEAW01) for 30 s, 72°C for 1 min. For the locus Mmer-hk34b the PCR program was: 95°C for 5 min; 40 cycles of 95°C for 1 min, 51°C for 1 min, 72°C for 55 s and a final elongation step at 72°C for 30 min. Amplified fragments were analyzed on an ABI 3500 Genetic Analyzer with GeneScan 600 LIZ internal size standard (Applied Biosystems). GeneMapper v.3.5 software (Applied Biosystems) was used to score alleles.

Scoring errors due to stutter fragments, large allele dropout and the presence of null alleles was checked with MICRO-CHECKER 2.2.3 (Van Oosterhout et al., 2004). Null allele frequencies were estimated for each loci and sample following the algorithm of (Dempster et al., 1977) with FREENA (Chapuis and Estoup, 2007) for microsatellite loci.

SNP genotyping

A custom Genotyping By Sequencing (GBS) assay was developed by Thermo Fisher Scientific, from a set of 917 SNP of hake which were selected from Milano et al., 2014 and Leone et al., 2019, using the following criteria and ranking: (1) outlier loci identified by the two studies, (2) significant P value of computed single locus F_{ST} values and common to Atlantic and Mediterranean populations, (3) significant P value of computed single locus F_{ST} values for Atlantic populations and (4) higher F_{ST} values. This set of markers was passed through a Thermo Fisher Scientific's quality control process. The quality check was performed using European Hake, *Merluccius merluccius*, reference genome (GenBank assembly accession: GCA_900312545.1). 591 SNP markers which passed the quality step were then submitted to AgriSeq™ primer design phase. The primer designs were in-silico checked for specificity and sensitivity of the intended target/marker regions using hake reference genome. Total of 574 SNPs were designed and were contained within 554 amplicons/target regions. Next, targeted sequencing was performed using the developed custom GBS hake panel.

The DNA samples were prepared for sequencing using the AgriSeq™ HTS Library Kit (Applied Biosystems). DNA samples were quantified using the Quant-iT DNA Assay kit (Thermo Fisher Scientific) on the Fluoroskan™ Microplate Fluorometer. DNA concentrations were normalized to 3.3 ng/µL for a total of 10 ng DNA per 10 µL reaction. Normalized DNA was combined with the Ion AgriSeq primer panel and AgriSeq amplification master mix. For amplification of genomic targets, the following thermocycling programs were used: 99°C for 2 min, then 16 cycles of 99°C for 15s and 60°C for 4 min. Amplicons were then prepared for ligation with pre-ligation enzyme digestion at 50°C for 10 min, 55°C for 10 min, and 60°C for 20 min. IonCode ™ Barcode Adapters 1-1248 kit were ligated to the digested products with barcoding enzyme and buffer. Labelled amplicons were then pooled, cleaned up, amplified and normalized. Following library preparation, libraries were loaded onto an Ion 540 ™ sequencing Chip Kit via the Ion 540™ Kit-Chef and Ion Chef. Sequencing was then performed on the Ion S5 system (Thermo Fisher, Inc. Waltham, MA). A total of 73M of reads per sequencing run were generated. The reads were then de-multiplexed to individual samples using barcode sequences.

For each sample, the sequenced reads from the targeted regions were then mapped to the hake reference genome using TMAP- Torrent Mapping Alignment Program followed by genotyping using TVC-Torrent Variant Caller. The genotypes were reported in different formats TOP, TOP/BOT and actual alleles using AgriSum Toolkit.

Loci and individuals with a call rate below 80% were removed. Monomorphic loci and loci with a minor allele frequency (MAF) lower than 1% were identified using adegenet R-package (Jombart, 2008) and were removed from the final dataset.

Linkage disequilibrium, Hardy-Weinberg equilibrium and genetic diversity

Linkage disequilibrium was tested for each pairs of loci in each sample using the probability test implemented in GENEPOP 4.0 (Rousset, 2008).

Tests for Hardy-Weinberg equilibrium within sample for each locus and over all loci were conducted with GENEPOP 4.0 using the exact test. Single and multilocus (Weir and Cockerham, 1984) *f* estimator of F_{IS} were computed with the same software.

Allele frequencies, observed (Ho) and Nei's (Nei, 1973) unbiased expected (He) heterozygosities were estimated using GENETIX 4.0.5 (Belkhir et al., 2004). Allelic richness [Ar(g)] and private allelic richness [Ap(g)] were computed with a rarefaction procedure using HP-RARE (Kalinowski, 2005) with the minimum number of genes set to 66 genes for microsatellite loci and 44 genes for SNP loci.

SNP Outlier detection

Outlier loci, potentially under selection, were identified using the independent approaches in Bayescan 2.1 (Foll and Gaggiotti, 2008) and R package *pcadapt* (Luu et al., 2017) with the whole dataset.

Bayescan relies on differences in allele frequencies between subpopulations to identify candidate loci under selection. Subpopulation specific FST coefficients are divided into two components by logistic regression. The population-specific component β is shared by all loci, and the locus-specific component α is shared by all populations. When the locus-specific component α is significantly different from zero means that the component is necessary to explain the diversity pattern, and points out a departure from neutrality at the given locus. Specifically, positive values of alpha indicate diversifying selection whereas negative values suggest balancing/purifying selection. For each locus a posterior probability is computed and SNPs with prior odds (PO) greater than a threshold are considered outliers since PO express how likely is the model with selection compared to the neutral. For each dataset we ran Bayescan with the following parameters: 50,000 burn-in period, 5,000 number of iterations, 20 pilot runs and thinning interval size set to 10. We set prior odd to 10 since the data set comprised less than 1000 SNPs and we set the false discovery rate (FDR) to 5% (Benjamini and Hochberg, 1995).

Furthermore, the method implemented in the R package pcadapt was run to detect local adaptation. This method relies on PCA to detect population structure, and it allows the identification of genetic markers putatively involved in biological adaptation considering their relationship with the population structure. The statistical and computational approach implemented in the package can be summarized in different steps. First, a PCA is used to ascertain population structure and identify the number of principal components K to consider. The number of PCs can be inferred by using the graphical approach based on the scree plot of the eigenvalues and applying the Cattell's rule. The components corresponding to eigenvalues to the left of the straight line should be kept. Then, each SNP is regressed by the K principal components. The results are reported as a vector of z-scores between each SNP, and the first K components obtained with the linear regression. Based on this vector, the outliers are then

identified using a multi-dimensional approach. The Mahalanobis distance is used as a statistic to measure the distance between each point and its mean. It calculates the distance between the covariance matrix of the z-scores and the vector of the z-score means. Finally, Mahalanobis distances are transformed in p-values based on the correlations between SNPs and the K principal components to performed multiple hypothesis testing. The q-value procedure is recommended to choose a threshold for p-values FDR approach (Benjamini and Hochberg, 1995).

The function pcadapt() with a large number of PCs (K= 20) was applied to the dataset to identify the optimal number of principal components K by using scree plot (Cattell's rule) or scoreplot (population structure). As a result, the pcadapt() function was run with the optimal K (K=4) by using Mahalanobis method and a threshold for MAF of 0.05 choosing as cut-off α = 0.1.

Population genetic structure analyses

The population genetic structure analyses were performed on microsatellite dataset and on the whole, neutral and outlier datasets for SNP.

(Weir and Cockerham, 1984) ϑ estimator of F_{ST} was computed between all pairs of samples with GENEPOP 4.0. For microsatellite dataset, pairwise F_{ST} estimates were also computed following the excluding null allele method in FREENA (Chapuis and Estoup, 2007), as null alleles can induce overestimation of genetic distance.

The measure of genetic differentiation D_{EST} (Jost, 2008) was computed with GENALEX 6.5 (Peakall and Smouse, 2012) using 999 permutations for microsatellite dataset.

The significance of pairwise genotypic differentiation between samples was tested using the exact test implemented in GENEPOP 4.0.

To evaluate the number of clusters (K) from individual's genotypes without prior information on their geographical locations, the Bayesian method implemented in STRUCTURE 2.3.4 (Pritchard et al., 2000; Falush et al., 2003, 2007) was used under the admixture model with correlated allele frequencies among clusters. Ten independent runs were performed for each K using 500 000 iterations and a burn-in period of 50 000. The evaluation of the best K was performed according to Pritchard's criterion (Pritchard et al., 2000) using STUCTURE HARVESTER online (Earl and vonHoldt, 2012). We used CLUMPP 1.1 (Jakobsson and Rosenberg, 2007) to merge the results across the 10 runs and DISTRUCT 1.1 (Rosenberg, 2004) to visualize the results.

TESS3 algorithm, combining genetic and geographic data to compute spatial ancestry coefficients, was also used to estimate and visualize population structure (Caye et al., 2016, 2018). The method uses individual spatial coordinates to better discriminate among putative populations using a graph based non-negative matrix factorization (Caye et al., 2016). The data were converted to tess3 matrix format using tess2tess3() function implemented in tess3r R package (Caye et al., 2018). The algorithm was run by using tess3() function to estimate spatial population structure for K values from 1 to 10 with 20 repetitions for each K. To identify the K that better described the data cross validation score was inspected, and ancestry coefficients

were analyzed only for values of K from 2 to 4. The corresponding Qmatrix of each K was visualized by using plotQ function implemented in pophelper R package. Values of the Qmatrix were interpolated on the geographic map of the study area using ggtess3Q() R function, showing the distribution of gene pools over the seascape.

A locus-by-locus analysis of molecular variance (AMOVA, n= 1000 permutations) was carried out in ARLEQUIN 3.5. (Excoffier et al., 2005) using the groups defined by STRUCTURE (see results).

Principal coordinate analysis (PCoA) was performed with GENALEX 6.5 (Peakall and Smouse, 2012) to investigate relative genetic distances among samples. PCoA is a multivariate technique that allows one to find and plot the major patterns within a multivariate data set (e.g. multiple loci and multiple samples). Here it was computed based on the matrix of pairwise codominant genotypic genetic distance among all pairs of geographical population samples, with the default options of Triangular Distance Matrix and Covariance-Standardized.

The pattern of isolation by distance (IBD) was tested through the correlation between pairwise $F_{ST}/(1- F_{ST})$ values and the logarithm of the geographical distances between samples by a Mantel test (n= 10 000 permutations) with GENETIX 4.05.

The level of significance was adjusted using a FDR, whenever multiple tests were conducted (Benjamini and Hochberg, 1995).

Gene-environment association analysis

Based on the geographical position of the 15 sampling sites, 7 macro-areas have been identified following the depth range of the species in the Mediterranean Sea and Atlantic Ocean (0-500 m) (Casey and Pereiro, 1995; Recasens et al., 1998) and the FAO GSA. For each macro-area, the centroid and its geographical coordinates (latitude and longitude) were identified and were used for individual coordinates. Three environmental variables (salinity, temperature and chlorophyll a) were selected, based on their influence on the distribution of European hake and their variation between sampling locations. Monthly mean values of physical and biochemical environmental variables were downloaded from E.U. Copernicus Marine Service Information (https://marine.copernicus.eu/). Salinity (sal, psu) and temperature (temp, °C) data were extracted from the Global Ocean Physics Reanalysis (product identifier GLOBAL_REANALYSIS_PHY_001_030) from 2014 to 2018 on a grid with 1/12° x 1/12° horizontal resolution (approximatively 8 km) and 50 vertical levels of thickness increasing with deph (0-5500 m). Chlorophyll a concentrations (chl, milligram m^{-3}) were extracted from the Global Ocean Biogeochemistry Hindcast (product identifier GLOBAL_REANALYSIS_BIO_001_029) from 2014 to 2018 on a grid with 1/4° x 1/4° horizontal resolution and 75 vertical levels of thickness increasing with depth (0-5500 m). Average values were extracted over the temporal windows of 5 years (2014-2018) based on macro-areas extension keeping only surface values.

In order to detect multicollinearity, we applied the function vifstep in the usdm R package (Naimi et al., 2014; Naimi, 2017) which calculates the Variance Inflation Factor (VIF) for the set of variables using a stepwise procedure. Highly correlated variables with a VIF greater than

10 indicating collinearity problem (Dormann et al., 2013) are excluded. Then, only salinity and chlorophyll a were kept. Spatial distance matrix between locations were calculated from geographical coordinates using gcd.hf function to account for earth curvature. Spatial variables were obtained by computing the distance-based Moran Eigenvector's Maps (dbMEMs) using pcnm function. Four dbMEMs were identified and one was removed after multicollinearity analysis.

To assess how much of the genetic variation could be explained by the variation in the set of environmental and spatial variables, Redundacy Analysis (RDA) was performed. Principal components of PCA on genetic distance matrices were used as response variables, and environmental and spatial variables as explanatory variables. We performed distinct RDAs using the whole, neutral and outlier datasets as response variable. To perform the PCA missing data were replaced with mean allele frequency. The matrix of allele frequencies was then transformed with the Hellinger approach using the decostand function, and PCA on this standardized matrix was performed by using the prcomp function. Following Selmoni et al., (2020) we extracted and used first principal components that reach the 80% of cumulative variance as response variable to compute RDA by using rda function in vegan R package, setting the option scale = TRUE. For the explanatory variables we combined the spatial dbMEMs factors and the environmental factors. We applied an ANOVA with 1000 permutations to assess the significance of the global model, and calculated the adjusted coefficient of determination (R^2adj) using the RSquareAdj function in vegan R package to determine the variation explained by the model. Then, we applied the ordistep function with 1000 permutations to perform both forward and backward selection of explanatory variables that best explained the variability of the response variable (optimal model). Also in this case, we evaluated the significance of the model and each variable by applying marginal ANOVAs with 1000 permutations, and we calculated he adjusted coefficient of determination, as previously explained. Distance among individuals and the relationship with the environmental variables were visualised by using biplots with option display = ("lc", "wa") and default scaling = 2 (Bernatchez et al., 2019; Selmoni et al., 2020).

To visualize the gradient of selected environmental variables, we applied the ordisurf function of vegan R package which fits a smooth surface for each variable providing the diagrams wih isolines. We then used the envfit function in vegan R package to identify vectors pointing in direction of the largest increase in variable value (Oksanen, 2015). Finally, we used the varpart function to partition the variation in genomic data with respect to environmental and spatial variables. We assessed the proportion of variability explained by each set of variables through the adjusted coefficient of determination (R^2adj) and Venn's diagrams.

Results

Microsatellite genotypic data, Linkage disequilibrium, Hardy-Weinberg equilibrium and genetic diversity

All the microsatellite loci were polymorphic with a number of alleles ranging from 16 for Mmer-hk3b to 62 for Mmer-hk9b with a mean value of 38 alleles per locus. No failure of PCR

amplification was observed, except one individual of ETP at locus Mmer-hk29 that was coded as missing data.

Over all sample, significant linkage disequilibrium among loci Mmer-hk20 and Mmer-hk9b was found ($P < 0.05$ after FDR correction) but was not generalized in all samples. Therefore, we did not consider these 2 loci physically linked. Besides, the linkage disequilibrium was also assessed with 1000 permutations using GENETIX 4.0.5 and no global linkage disequilibrium was found ($P > 0.05$ after FDR correction).

No evidence of large allele dropout or scoring errors due to stutters was found. However, evidence of null allele was observed at locus Mmer-hk9b in MHD, at locus Mmer-hk29 in all samples and locus Mmer-hk34b in all samples except CDZ, MDQ, NDR and ANB (Table S1, Supporting Information). The estimates of null allele frequencies varied between 0.01 (for Mmer-hk9b in GHZ) to 0.353 (for Mmer-hk29 in MLG) (Table S1, Supporting Information).

Multilocus F_{IS} values ranged between 0.13 for CDZ and ANB and 0.2 for GHZ (Table 2). For each locus F_{IS} values ranged from -0.13 (for Mmer-hk3b in NDR) to 0.74 (for Mmer-hk29 in MLG) (Table S1, Supporting Information). Over all loci, significant heterozygote deficits were found in all samples (after FDR correction). However, heterozygote deficit was not generalized for all loci in all samples (except for Mmer-hk29 and Mmer-hk34b) and in 77% of the cases those heterozygote deficits matched the evidence of null allele using MICRO-CHECKER and high null allele frequencies for Mmer-hk29 and Mmer-hk34b (Table S1, Supporting Information).

Observed and unbiased expected heterozygosities varied between 0.72 for GHZ and 0.82 for CDZ and between 0.88 for ANB, TBK, TOR, CAS and 0.93 for MHD, CDZ, MDQ, respectively (with a mean value of 0.76 and 0.90 respectively) (Table 2). The allelic richness Ar(66) ranged from 20.1 for HUE to 22.33 for MLG and the private allelic richness Ap(66) from 0.06 for ETP and MDQ to 0.76 for AGA (Table 2).

Table 2. Estimators of genetic diversity at microsatellite loci.

Sample	Ho	He	Ar(66)	Ap(66)	*f*
AGA	0.75	0.92	21.10	0.76	**0.19**
MHD	0.75	0.93	22.05	0.65	**0.19**
HUE	0.79	0.92	20.10	0.34	**0.15**
CDZ	0.82	0.93	21.66	0.34	**0.13**
ETP	0.76	0.9	20.45	0.06	**0.16**
MLG	0.78	0.92	22.33	0.38	**0.16**
RQT	0.75	0.90	22.18	0.33	**0.18**
MDQ	0.78	0.93	21.69	0.06	**0.17**
NDR	0.77	0.90	20.85	0.28	**0.15**
GHZ	0.72	0.89	21.25	0.17	**0.20**
ANB	0.77	0.88	21.18	0.19	**0.13**
TBK	0.76	0.88	21.81	0.46	**0.15**
GTU	0.73	0.89	20.17	0.44	**0.19**
TOR	0.74	0.88	21.40	0.31	**0.17**
CAS	0.73	0.88	20.85	0.45	**0.17**

Ho, observed heterozygosity; He, unbiased expected heterozygosity; Ar(66) rarefied allelic richness and Ap(66) rarefied private allelic richness (with rarefaction size of 66) ; *f*, Weir & Cockerham's (1984) *f* estimator of F_{IS} with significant values in bold (0.05 threshold after FDR correction).

SNP genotypic data, Linkage disequilibrium, Hardy-Weinberg equilibrium and genetic diversity

After excluding loci with a call rate lower than 80% (71 loci), monomorphic loci (36 loci) and loci with a MAF lower than 1% (14 loci), the linkage disequilibrium and Hardy Weinberg tests were performed on 453 loci. Three individuals with a call rate lower than 80% were removed from MDQ.

Over all sample, significant linkage disequilibrium among 15 pairs of loci were found ($P < 0.05$ after FDR correction). Over all loci, all samples presented significant heterozygote deficits (after FDR correction). Therefore, all loci involved in linkage disequilibrium and presenting departure from Hardy-Weinberg expectations in at least one sample were removed from subsequent analyses (192 loci, keeping 261 loci).

Multilocus F_{IS} values ranged between -0.02 for MDQ and 0.04 for NDR and GHZ (Table 3). Over all loci, significant heterozygote deficits were found in 3 samples (NDR, GHZ and TOR) out of 15 (after FDR correction) (Table 3).

Observed and unbiased expected heterozygosities varied between 0.30 for ANB, GTU, TOR and CAS and 0.33 for AGA, ETP, MLG, MDQ and GHZ and between 0.30 for GTU and 0.34 for ETP and GHZ, respectively (with a mean value of 0.32 and 0.32 respectively) (Table 3). The allelic richness Ar(44) ranged from 1.99 for AGA, MHD, HUE, CDZ, RQT, NDR, ANB, TBK, GTU, TOR and CAS to 2 for GHZ, ETP, MLG and MDQ (Table 3).

Table 3. Estimators of genetic diversity at (261) SNP loci.

Sample	Ho	He	Ar(44)	*f*
AGA	0.33	0.33	1.99	0.01
MHD	0.32	0.32	1.99	0.01
HUE	0.32	0.33	1.99	0.02
CDZ	0.32	0.32	1.99	0.02
ETP	0.33	0.34	2	0.02
MLG	0.33	0.33	2	0.02
RQT	0.32	0.32	1.99	-0.01
MDQ	0.33	0.33	2	-0.02
NDR	0.32	0.33	1.99	**0.04**
GHZ	0.33	0.34	2	**0.04**
ANB	0.3	0.31	1.99	0.01
TBK	0.32	0.32	1.99	-0.01
GTU	0.3	0.3	1.99	0.02
TOR	0.3	0.31	1.99	**0.03**
CAS	0.3	0.31	1.99	0.02

Ho, observed heterozygosity ; He, unbiased expected heterozygosity ; Ar(44) rarefied allelic richness (with rarefaction size of 44) ; *f*, Weir & Cockerham's (1984) *f* estimator of F_{IS} with significant values in bold (0.05 treshold after FDR correction).

SNP outlier detection

Bayescan method identified 59 SNPs putatively under selection, while *pcadapt* identified 55 SNPs. In order to minimize the detection of false positives, only 31 loci that were identified in common with both methods were considered as outlier loci. All loci not detected as outlier by any of the 2 software pcadapt and Bayescan, considering both balancing and diversifying outliers, were considered as neutral (178 loci).

Population genetic structure

Overall F_{ST} values were 0.0053, 0.0377, 0.0047 and 0.1360 using microsatellite, whole, neutral and outlier datasets respectively and the exact test indicated a significant differentiation ($P < 0.001$). Pairwise F_{ST} values (Table 4) ranged from -0.0037 (TOR vs. CAS) to 0.0184 (HUE vs. ANB), from -0.0036 (MLG vs. GHZ) to 0.1196 (CDZ vs. GTU), from -0.0034 (MLG vs. GHZ) to 0.0188 (CDZ vs. GTU) and from -0.0061 (MLG vs. GHZ) to 0.3729 (CDZ vs. GTU) using microsatellite, whole, neutral and outlier datasets respectively.

No significant differences were observed between pairwise F_{ST} and pairwise corrected for null alleles (t-test, $P = 0.783$) for microsatellite dataset.

Pairwise D_{EST} values (Table 4) ranged from -0.0396 (HUE vs. AGA) to 0.1613 (CAS vs. CDZ) for microsatellite dataset.

Pairwise comparisons of genetic differentiation, displayed a high degree of structure with 80 and 91 significant tests out of 105 for the whole and outlier datasets respectively (Table 4).

However, for the neutral and microsatellite datasets the level of structure was moderate with respectively 21 and 30 significant tests out of 105 (Table 4).

For all datasets, the Atlantic formed 1 genetic unit, except for the outlier dataset that unveiled genetic differentiation between AGA and HUE; AGA and CDZ and MHD and CDZ.

For neutral and microsatellite datasets, North Alboran (GSA 1) formed 1 genetic unit, whereas for the whole and outlier datasets RQT was genetically differentiated from ETP and MLG.

For all datasets, South Alboran (GSA 3) formed 1 genetic unit, except for the outlier dataset that revealed genetic differentiation between MDQ and NDR.

For neutral and microsatellite datasets, Algeria (GSA 4) formed 1 genetic unit, whereas for the whole and outlier datasets GHZ was genetically differentiated from ANB.

For all datasets, North Tunisia (GSA 12) formed 1 genetic unit such as North Spain (GSA 6).

Atlantic was genetically differentiated from the Mediterranean for all datasets. However, only for the whole and outlier datasets, all Atlantic samples were genetically differentiated from all the Mediterranean samples.

The genetic structure between each GSA, based on pairwise comparisons of genetic differentiation, depended on the considered dataset (see Table 4 for details).

Table 4. Pairwise F_{ST} (lower left) and D_{EST} (upper right) values for: a) microsatellite, b) whole, c) neutral, and d) outlier datasets respectively. Significant values are in bold (0.05 threshold after FDR correction).

a) MICROSATELLITE LOCI		ATLANTIC				NORTH ALBORAN			SOUTH ALBORAN		ALGERIA		NORTH TUNISIA		NORTH SPAIN	
		AGA	MHD	HUE	CDZ	ETP	MLG	RQT	MDQ	NDR	GHZ	ANB	TBK	GTU	TOR	CAS
ATLANTIC	AGA	-	0.0149	-0.0396	-0.0198	0.0126	0.0183	0.0334	-0.0369	0.0578	0.0147	0.091	0.0741	0.1221	0.0864	0.0555
	MHD	0.0013	-	0.0038	-0.0062	0.0849	-0.0201	0.0487	-0.0086	0.107	0.0813	0.1586	0.103	0.1079	0.1113	0.1072
	HUE	-0.0036	0.0003	-	-0.0175	0.0376	0.048	0.0629	-0.02	0.0684	0.0621	0.1587	0.0819	0.1214	0.1457	0.105
	CDZ	-0.0017	-0.0005	-0.0015	-	0.0268	0.0523	0.0819	-0.0352	0.0544	0.0699	0.1595	0.1275	0.1255	0.1507	0.1613
NORTH ALBORAN	ETP	0.0011	0.0081	0.0040	0.0024	-	0.027	0.0136	0.0262	0.0317	-0.0104	0.0535	0.0339	0.0747	0.0504	0.0505
	MLG	0.0017	-0.0018	0.0044	0.0045	0.0029	-	-0.0055	-0.0163	0.0763	0.0381	0.0791	0.0258	0.0616	0.0692	0.0518
	RQT	0.0034	0.0046	**0.0063**	**0.0078**	0.0014	-0.0006	-	0.0497	0.0451	0.02	-0.0006	-0.0221	0.0098	-0.0011	-0.0042
SOUTH ALBORAN	MDQ	-0.0032	-0.0007	-0.0017	-0.0028	0.0027	-0.0014	0.0048	-	0.0483	0.0205	0.1029	0.0716	0.0947	0.0864	0.0672
	NDR	0.0059	0.0103	0.0069	0.0052	0.0035	0.0079	0.0051	0.0047	-	0.044	0.0512	0.045	0.0265	0.028	0.0343
ALGERIA	GHZ	0.0016	0.0083	**0.0066**	0.0071	-0.0014	0.0041	0.0024	0.0021	0.0052	-	0.0389	0.0353	0.07	0.0419	0.03
	ANB	**0.0108**	**0.0176**	**0.0184**	0.0176	**0.0070**	**0.0095**	-0.0000	**0.0115**	0.0066	0.0053	-	0.0244	0.0192	0.002	-0.0072
NORTH TUNISIA	TBK	**0.0083**	**0.0109**	**0.0091**	**0.0134**	0.0042	0.0030	-0.0027	0.0076	0.0056	0.0046	0.0034	-	0.0221	0.0058	-0.0075
	GTU	**0.0131**	**0.0109**	**0.0129**	0.0126	**0.0089**	0.0067	0.0012	**0.0096**	**0.0032**	**0.0087**	0.0026	0.0029	-	0.0096	-0.0052
NORTH SPAIN	TOR	**0.0100**	**0.0121**	**0.0166**	0.0163	**0.0064**	0.0081	-0.0001	0.0095	0.0036	**0.0056**	0.0003	0.0008	0.0013	-	-0.0264
	CAS	**0.0064**	**0.0115**	**0.0118**	0.0171	**0.0064**	0.0060	-0.0005	0.0073	0.0043	**0.0039**	-0.0010	-0.0010	-0.0007	-0.0037	-

b) WHOLE SNP LOCI		ATLANTIC				NORTH ALBORAN			SOUTH ALBORAN		ALGERIA		NORTH TUNISIA		NORTH SPAIN	
		AGA	MHD	HUE	CDZ	ETP	MLG	RQT	MDQ	NDR	GHZ	ANB	TBK	GTU	TOR	CAS
ATLANTIC	AGA	-														
	MHD	0.0018	-													
	HUE	0.0036	0.0033	-												
	CDZ	0.0078	0.0045	0.0007	-											
NORTH ALBORAN	ETP	**0.0150**	**0.0153**	**0.0147**	**0.0185**	-										
	MLG	**0.0166**	**0.0163**	**0.0167**	**0.0198**	-0.0019	-									
	RQT	**0.0497**	**0.0545**	**0.0602**	**0.0660**	**0.0193**	**0.0143**	-								
SOUTH ALBORAN	MDQ	**0.0112**	**0.0110**	**0.0173**	**0.0172**	0.0051	0.0045	**0.0244**	-							
	NDR	**0.0169**	**0.0188**	**0.0224**	**0.0260**	0.0017	0.0002	**0.0106**	0.0036	-						
ALGERIA	GHZ	**0.0179**	**0.0170**	**0.0180**	**0.0227**	-0.0005	-0.0036	**0.0146**	0.0051	-0.0017	-					
	ANB	**0.0801**	**0.0855**	**0.0950**	**0.1038**	**0.0457**	**0.0410**	**0.0112**	**0.0499**	**0.0314**	**0.0387**	-				
NORTH TUNISIA	TBK	**0.0778**	**0.0840**	**0.0911**	**0.0969**	**0.0419**	**0.0391**	**0.0116**	**0.0451**	**0.0283**	**0.0354**	0.0004	-			
	GTU	**0.0965**	**0.1034**	**0.1118**	**0.1196**	**0.0583**	**0.0534**	**0.0208**	**0.0613**	**0.0422**	**0.0494**	0.0053	0.0026	-		
NORTH SPAIN	TOR	**0.0903**	**0.0933**	**0.0943**	**0.1030**	**0.0387**	**0.0338**	**0.0103**	**0.0567**	**0.0330**	**0.0340**	**0.0207**	0.0214	**0.0291**	-	
	CAS	**0.0856**	**0.0874**	**0.0919**	**0.0992**	**0.0356**	**0.0332**	0.0064	**0.0505**	**0.0294**	**0.0309**	0.0114	0.0124	0.0198	0.0002	-

c) NEUTRAL SNP LOCI

		ATLANTIC				NORTH ALBORAN			SOUTH ALBORAN		ALGERIA		NORTH TUNISIA		NORTH SPAIN	
		AGA	MHD	HUE	CDZ	ETP	MLG	RQT	MDQ	NDR	GHZ	ANB	TBK	GTU	TOR	CAS
ATLANTIC	AGA	-														
	MHD	0.0016	-													
	HUE	-0.0033	0.0010	-												
	CDZ	0.0003	-0.0003	0.0002	-											
NORTH ALBORAN	ETP	0.0015	0.0012	0.0011	0.0020	-										
	MLG	0.0043	0.0021	0.0031	0.0034	-0.0005	-									
	RQT	**0.0099**	0.0089	**0.0095**	**0.0092**	0.0015	0.0014	-								
SOUTH ALBORAN	MDQ	0.0065	0.0024	0.0062	0.0039	0.0009	0.0003	0.0075	-							
	NDR	0.0022	0.0009	0.0011	0.0010	-0.0020	-0.0007	0.0031	0.0027	-						
ALGERIA	GHZ	0.0026	0.0014	0.0024	0.0049	-0.0015	-0.0034	0.0030	0.0010	-0.0017	-					
	ANB	**0.0104**	0.0077	**0.0080**	**0.0122**	0.0003	0.0020	0.0004	0.0098	0.0031	0.0033	-				
NORTH TUNISIA	TBK	**0.0122**	0.0088	0.0091	0.0096	-0.0006	0.0008	0.0002	0.0085	0.0025	0.0036	-0.0013	-			
	GTU	**0.0185**	**0.0180**	**0.0164**	**0.0188**	0.0056	0.0063	0.0023	0.0127	0.0075	0.0067	0.0026	-0.0011	-		
NORTH SPAIN	TOR	**0.0152**	**0.0119**	**0.0126**	**0.0139**	0.0027	0.0035	-0.0003	**0.0103**	0.0048	0.0062	-0.0026	-0.0018	-0.0019	-	
	CAS	**0.0158**	**0.0115**	**0.0135**	**0.0143**	0.0017	0.0057	0.0012	**0.0110**	0.0061	0.0060	-0.0024	-0.0008	-0.0007	-0.0019	-

d) OUTLIER SNP LOCI

		ATLANTIC				NORTH ALBORAN			SOUTH ALBORAN		ALGERIA		NORTH TUNISIA		NORTH SPAIN	
		AGA	MHD	HUE	CDZ	ETP	MLG	RQT	MDQ	NDR	GHZ	ANB	TBK	GTU	TOR	CAS
ATLANTIC	AGA	-														
	MHD	0.0070	-													
	HUE	**0.0207**	0.0122	-												
	CDZ	**0.0359**	**0.0166**	-0.0007	-											
NORTH ALBORAN	ETP	**0.0525**	**0.0602**	**0.0626**	**0.0646**	-										
	MLG	**0.0525**	**0.0640**	**0.0731**	**0.0750**	-0.0047	-									
	RQT	**0.1592**	**0.1917**	**0.2152**	**0.2251**	**0.0707**	**0.0559**	-								
SOUTH ALBORAN	MDQ	**0.0237**	**0.0336**	**0.0473**	**0.0493**	**0.0208**	**0.0147**	**0.0845**	-							
	NDR	**0.0590**	**0.0806**	**0.0924**	**0.0965**	0.0073	0.0008	**0.0372**	**0.0128**	-						
ALGERIA	GHZ	**0.0548**	**0.0667**	**0.0747**	**0.0753**	-0.0008	-0.0061	**0.0503**	0.0100	-0.0047	-					
	ANB	**0.2533**	**0.2957**	**0.3279**	**0.3359**	**0.1755**	**0.1592**	**0.0566**	**0.1746**	**0.1176**	**0.1444**	-				
NORTH TUNISIA	TBK	**0.2415**	**0.2852**	**0.3130**	**0.3210**	**0.1663**	**0.1561**	**0.0587**	**0.1651**	**0.1119**	**0.1351**	0.0055	-			
	GTU	**0.2887**	**0.3342**	**0.3634**	**0.3729**	**0.2091**	**0.1955**	**0.0809**	**0.2088**	**0.1526**	**0.1773**	0.0003	0.0017	-		
NORTH SPAIN	TOR	**0.2701**	**0.2992**	**0.3142**	**0.3212**	**0.1282**	**0.1131**	**0.0441**	**0.1861**	**0.1091**	**0.1127**	**0.1067**	**0.1155**	**0.1296**	-	
	CAS	**0.2494**	**0.2798**	**0.2994**	**0.3064**	**0.1221**	**0.1053**	**0.0229**	**0.1657**	**0.0908**	**0.0996**	**0.0720**	**0.0739**	**0.0912**	0.0028	-

According to Pritchard's criterion, the best K value for the whole and neutral datasets was K = 2 (Figure S1, Supporting Information) and K = 3 for the outlier dataset (Figure S1, Supporting Information), whereas STRUCTURE failed to detect any structure with the microsatellite dataset (Figure S1, Supporting Information). For all the SNP datasets the first cluster grouped the Atlantic, North Alboran (except RQT), South Alboran samples and GHZ, while the second cluster gouped RQT, ANB, North Tunisia and North Spain samples for the whole and neutral datasets (Figure 2). The outlier dataset separated the second previous cluster into 2 different clusters: one grouping RQT and North Spain samples and another gouping ANB and North Tunisia samples.

Following the cross-validation criterion for TESS3, the best K values corresponds either to a plateau or an increase in the curve. For whole dataset, the cross-validation curve started to increase from K = 3 (Figure S2, Supporting Information) indicating three main ancestral groups (Figure 3 and Figure S3, Supporting Information), corresponding to the following geographical clines: (1) Atlantic, North Alboran (except RQT) and South Alboran samples and GHZ; (2) ANB and North Tunisia samples; (3) RQT and North Spain samples. For neutral dataset, the cross-validation curve started to increase from K = 2 (Figure S2, Supporting Information) indicating two main ancestral groups (Figure 3 and Figure S3, Supporting Information): (1) Atlantic, South Alboran samples and GHZ; (2) ANB, North Alboran, North Tunisia and North Spain samples. For outlier dataset, the cross-validation curve decreased rapidly from K = 1 to K = 2, confirming the main division into two ancestral groups. From K = 3 the values of cross-validation started to decrease slowly indicating a subtle substructure (Figure S2, Supporting Information). Four ancestral groups were detected (Figure 3 and Figure S3, Supporting Information), corresponding to the following geographical area: (1) Atlantic samples (except AGA); (2) AGA and North and South Alboran samples; (3) ANB and North Tunisia samples; (4) North Spain samples. For microsatellite dataset, the cross-validation curve started to increase from K = 2 (Figure S2, Supporting Information) indicating two main ancestral groups (Figure 3 and Figure S3, Supporting Information): (1) Atlantic, North Alboran (except RQT) samples and MDQ; (2) NDR, RQT, Algeria, North Tunisia and North Spain samples.

The AMOVA confirmed the significant genetic structure among the STRUCTURE groups, among samples within group and within samples for all datasets ($P < 0.001$, except among samples within group for the neutral dataset $P = 0.09$; Table 5). The percentage of genetic variation explained by differences among groups was 5, 0.8 and 16.5 % for the whole, neutral and outlier datasets respectively (Table 5). However, the largest variation was due to within populations differences with 93.9, 99.1 and 80.8% for whole, neutral and outlier datasets respectively (Table 5).

The PCoA also confirmed STRUCTURE results (Figure 4) as 83.4, 72.3 and 83% of variance from the first principal coordinate distinguished Atlantic, North Alboran (except RQT), South Alboran samples and GHZ from RQT, ANB, North Tunisia and North Spain samples for whole, neutral and outlier datasets respectively. The second principal coordinate, accounting for 9.9, and 12.8% of the variance allowed to differentiate RQT and North Spain samples from ANB and North Tunisia samples for the whole and outlier datasets respectively. For the microsatellite dataset, the first principal coordinate (63.4 %) distinguished 3 groups: 1) Atlantic

samples, MDQ and GHZ, 2) NDR, GHZ and ETP and 3) RQT, ANB, North Tunisia and North Spain samples (Figure 4).

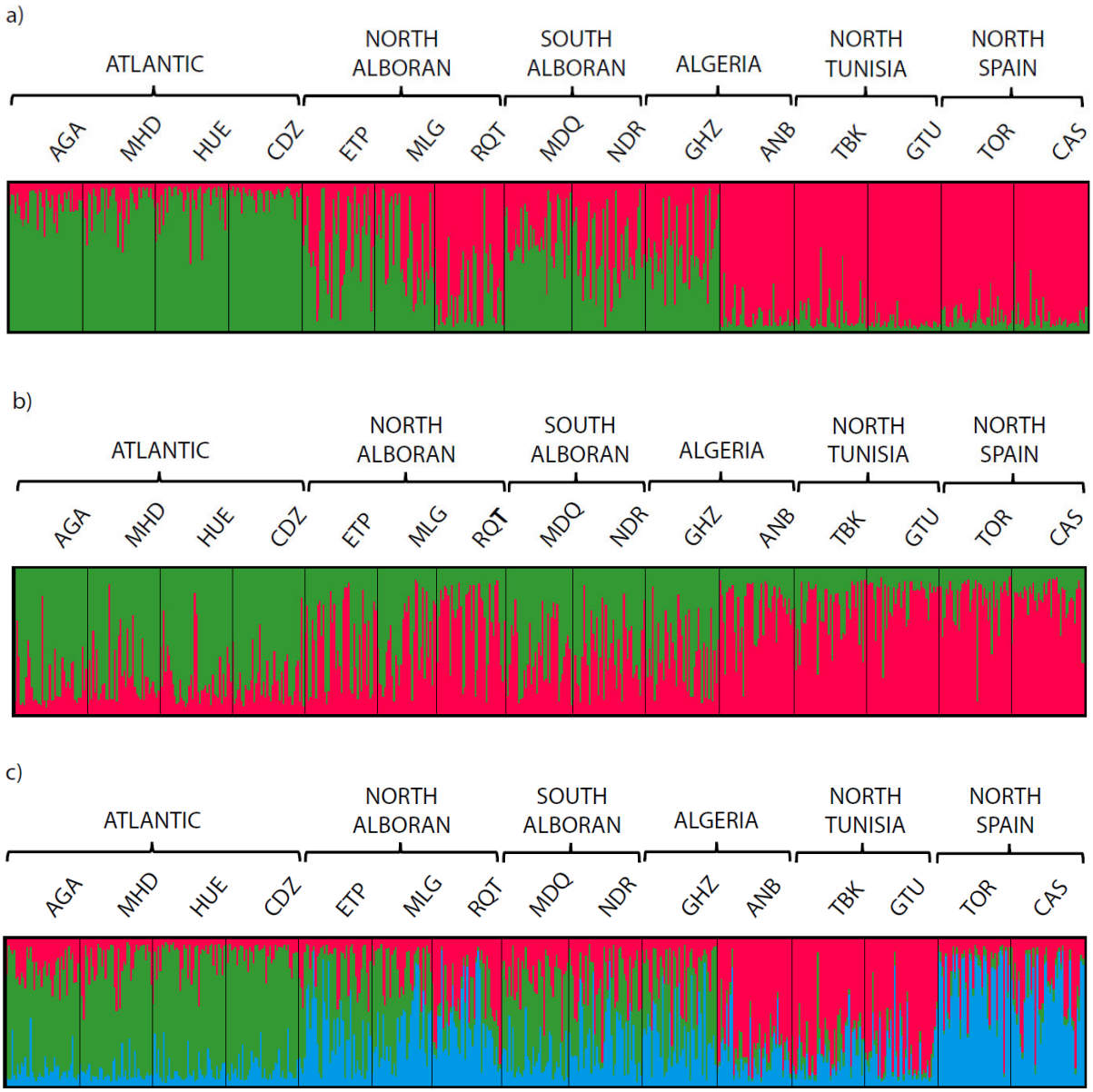
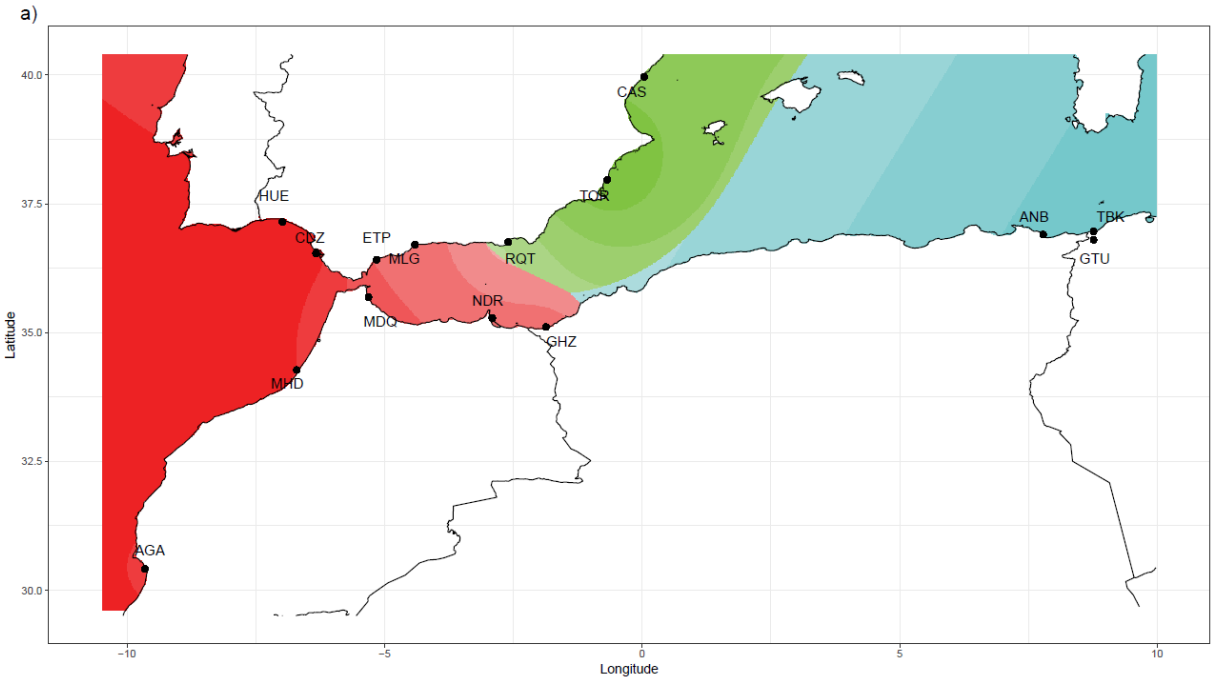
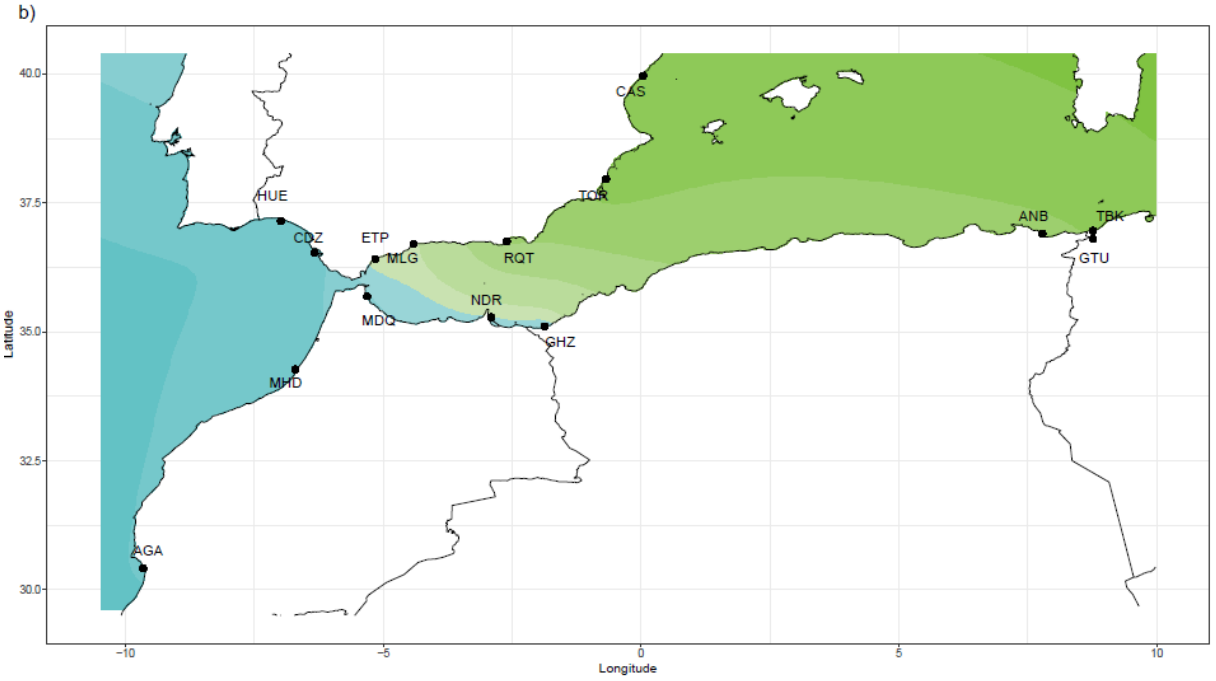


Figure 2. Graphical representation of the Bayesian clustering analysis with STRUCTURE for: a) whole, b) neutral and c) outlier datasets of SNP.


a)
CAS
TOR
HUE
CDZ
ETP
MLG
RQT
ANB
TBK
GTU
NDR
MDQ
GHZ
MHD
AGA
Latitude
Longitude
40.0
37.5
35.0
32.5
30.0
−10
−5
0
5
10


b)
CAS
TOR
HUE
CDZ
ETP
MLG
RQT
ANB
TBK
GTU
NDR
MDQ
GHZ
MHD
AGA
Latitude
Longitude
40.0
37.5
35.0
32.5
30.0
−10
−5
0
5
10

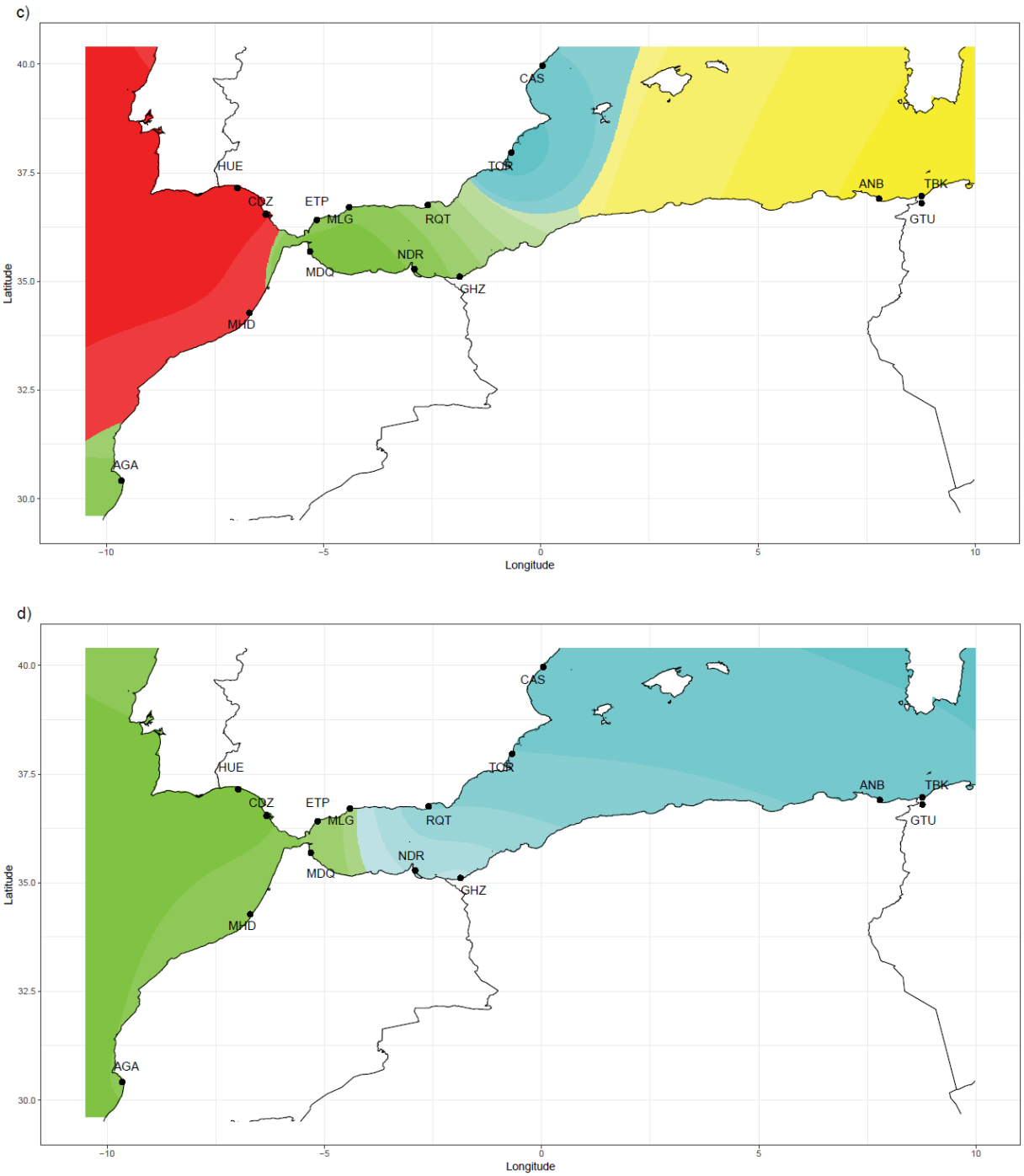
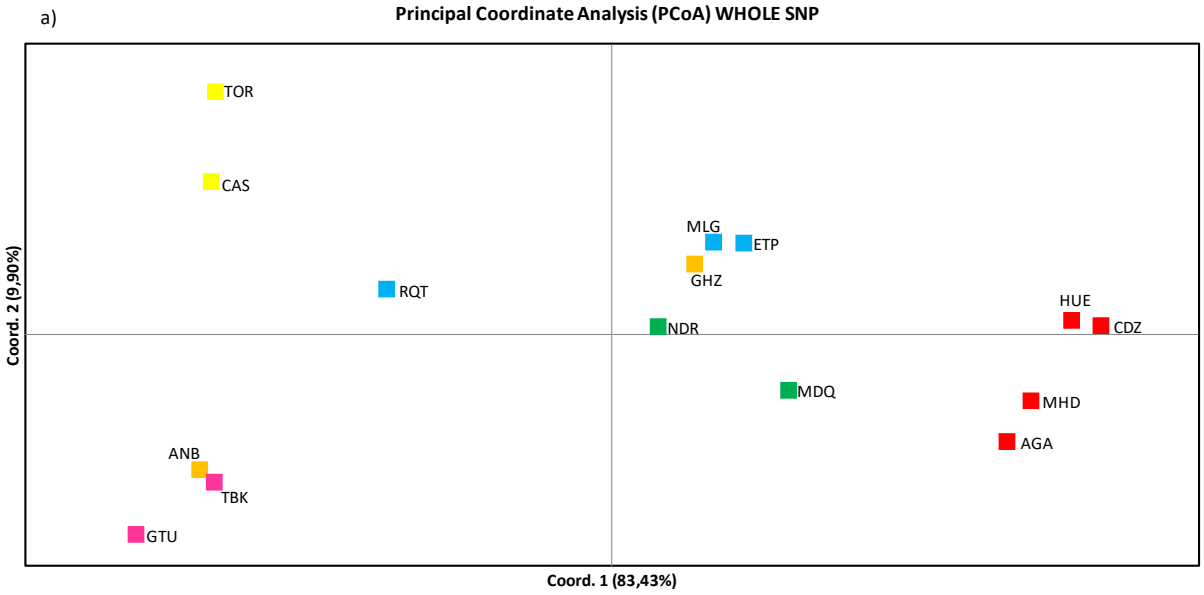


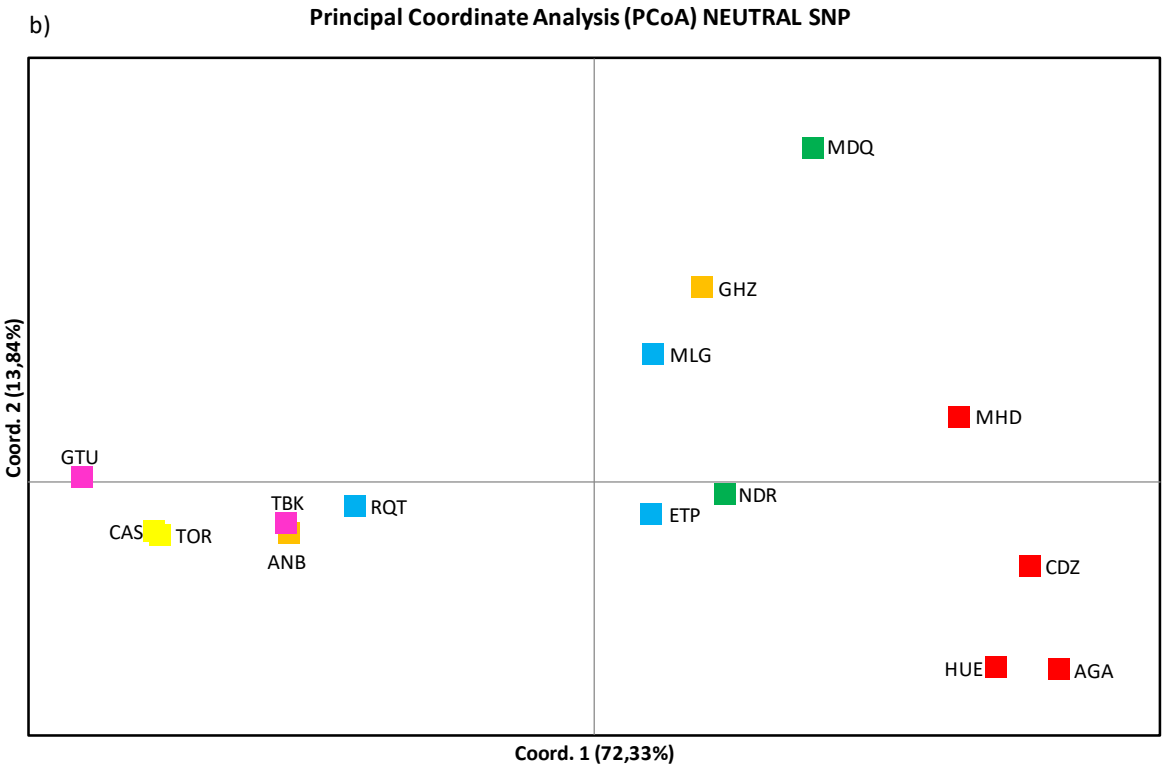
Figure 3. Interpolation of the values of the ancestry coefficient on the geographical map of the study area for: a) whole, b) neutral, c) outlier datasets of SNP and d) microsatellite dataset.

Table 5. Analysis of molecular variance (AMOVA) following STRUCTURE groups for: a) whole, b) neutral and c) outlier datasets of SNP.

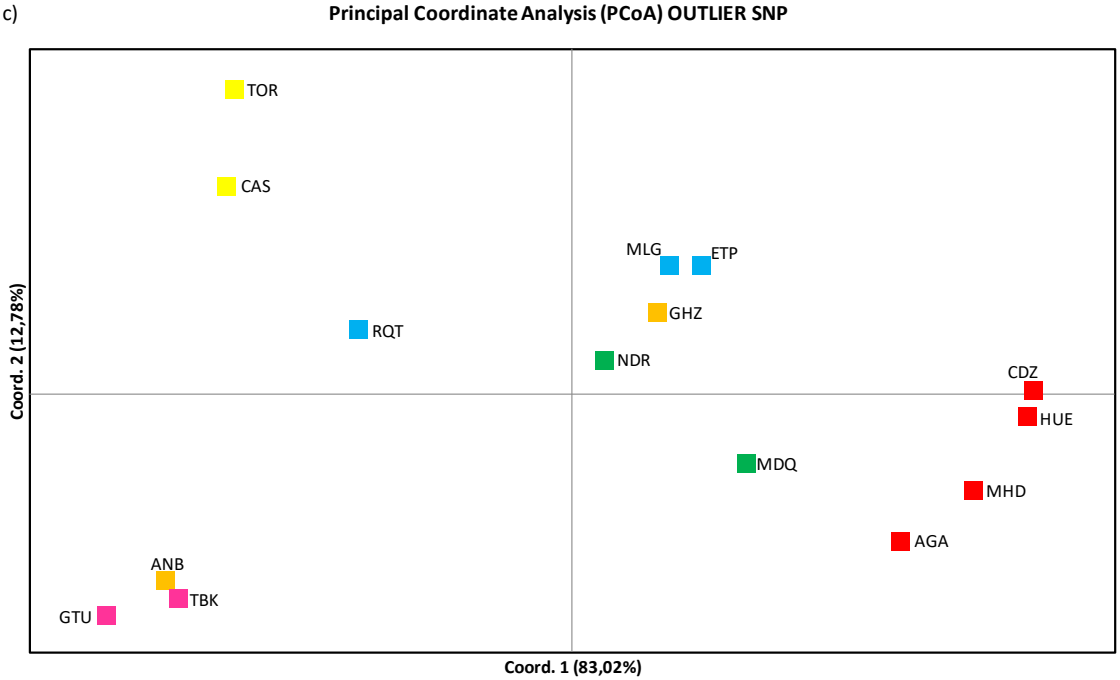
	Source of Variation	Percentage of variance (%)	*P*-value
a)	**STRUCTURE clusters K = 2**		
	Among groups	5	<0.001
	Among samples within group	1.1	<0.001
	Within samples	93.9	<0.001
b)	**STRUCTURE clusters K = 2**		
	Among groups	0.8	<0.001
	Among samples within group	0.1	0.09
	Within samples	99.1	<0.001
c)	**STRUCTURE clusters K = 3**		
	Among groups	16.5	<0.001
	Among samples within group	2.7	<0.001
	Within samples	80.8	<0.001



b)


Principal Coordinate Analysis (PCoA) NEUTRAL SNP
MDQ
GHZ
MLG
MHD
GTU
TBK
RQT
NDR
CAS
TOR
ETP
ANB
CDZ
HUE
AGA
Coord. 2 (13,84%)
Coord. 1 (72,33%)

c)


Principal Coordinate Analysis (PCoA) OUTLIER SNP
TOR
CAS
MLG
ETP
RQT
GHZ
NDR
CDZ
HUE
MDQ
MHD
AGA
ANB
TBK
GTU
Coord. 2 (12,78%)
Coord. 1 (83,02%)

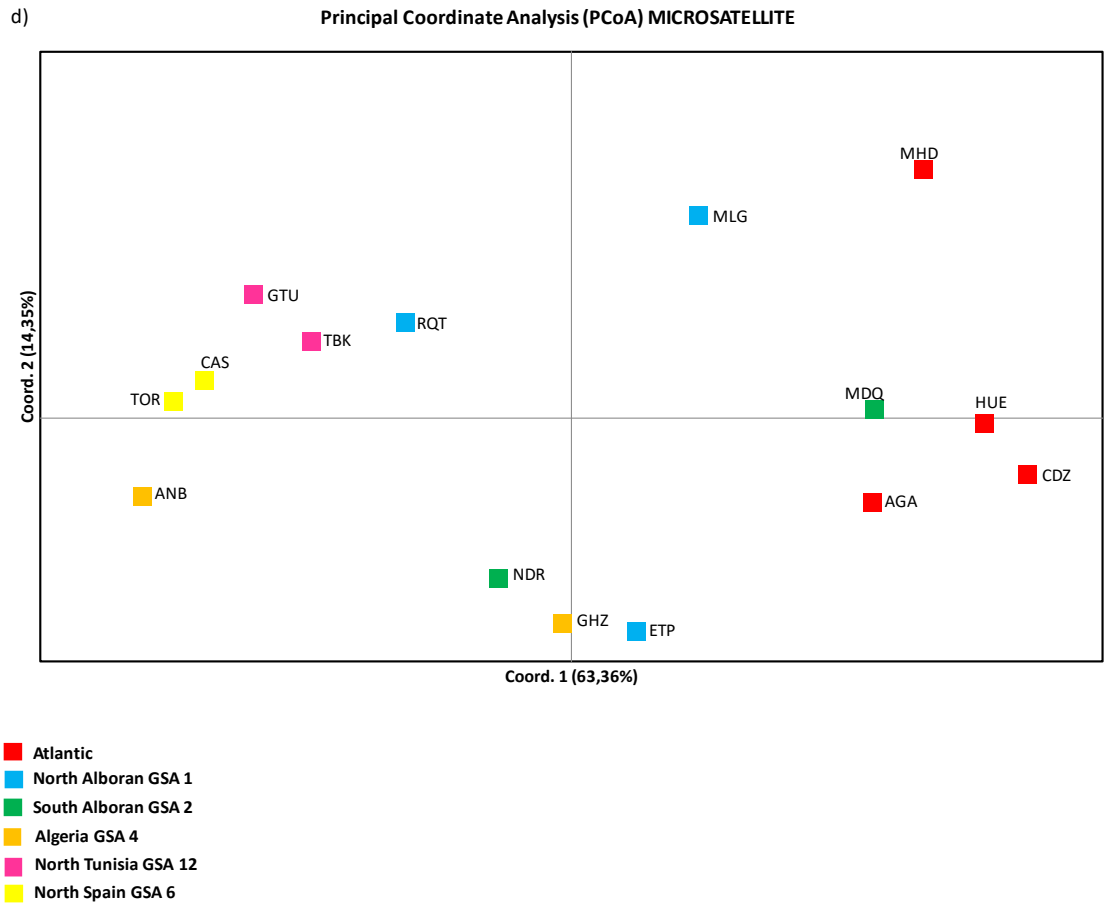


Figure 4. Principal coordinate analysis (PCoA) for: a) whole, b) neutral, c) outlier and d) microsatellite datasets.

No pattern of IBD was found with neither of the datasets (R = -0.312, P = 0.978; R = -0.278, P = 0.985; R = -0.383, P = 0.997; R = -0.240, P = 0.979 for whole, neutral, outlier and microsatellite datasets respectively).

Gene-environment association analysis

The global RDA model was significant ($p < 0.05$) and explained 0.73% of genetic variation for the whole dataset. The first two axes of the RDA accounted for 0.46% of the genetic variation. Spatial factors (dbMEM2 and dbMEM3), salinity and chlorophyll a were identified for the RDA using whole dataset and following the ordistep procedure (Figure S4, Supporting Information). The marginal ANOVAs, considering each independent explanatory variable selected by ordistep function, showed that spatial factors (dbMEM2 and dbMEM3), salinity and chlorophyll a were all significant predictors of the genetic variation ($p < 0.01$). The results of variation partitioning indicated that environmental variables (5.2%) had a higher fraction of explained variance than spatial variables (0.7%). Shared variance resulted in 0.8% of explained variance.

The global RDA model was significant ($p < 0.05$) and explained 0.40% of genetic variation for the neutral dataset. The first two axes of the RDA accounted for 0.24% of the genetic variation.

Only environmental factors (sal and chl) were identified for the RDA using neutral dataset and following the ordistep procedure (Figure S4, Supporting Information). The marginal ANOVAs, considering each independent explanatory variable selected by ordistep function, showed that salinity and chlorophyll a were all significant predictors of the genetic variation ($p < 0.01$). The results of variation partitioning indicated that chlorophyll a (0.3%) had approximately the same individual fraction of explained variance than salinity (0.1%). Shared variance (0.6%) resulted in the highest amount of explained variance.

The global RDA model was significant ($p < 0.05$) and explained 5.40% of genetic variation for the whole dataset. The first two axes of the RDA accounted for 4.59% of the genetic variation. Spatial factors (dbMEM2 and dbMEM3), salinity and chlorophyll a were identified for the RDA using outlier dataset and following the ordistep procedure (Figure S4, Supporting Information). The marginal ANOVAs, considering each independent explanatory variable selected by ordistep function, showed that spatial factors (dbMEM2 and dbMEM3), salinity and chlorophyll a were all significant predictors of the genetic variation ($p < 0.01$). The results of variation partitioning indicated that environmental variables (17.1%) had a higher fraction of explained variance than spatial variables (3%). Shared variance (3%) resulted in the same amount of explained variance as spatial variables.

Discussion and conclusion

Genetic markers

The present study of population genetic structure of European hake used two kind of genetic markers, microsatellite and SNP loci. The level of genetic variability, based on observed and expected heterozygosities, found in this study was in the same order of magnitude as previous studies for microsatellite (Lundy et al., 1999; Castillo et al., 2005; Pita et al., 2011, 2016) and SNP (Milano et al., 2014; Westgaard et al., 2017; Leone et al., 2019) markers. However, the 6 microsatellite loci used in this study were less powerful than the 261 SNP loci to unveil population structure for hake at a small spatial scale. This could be due to the low number of amplified microsatellite markers and 2 loci out of the 6 were out of Hardy Weinberg equilibrium, mainly explained by high level of null allele frequencies at these loci.

Population genetic structure of the European hake within the Alboran Sea and adjacent waters

Our study unveiled an overall population structure in the European hake in the Alboran Sea and among adjacent waters. Despite some genetic differentiation between Atlantic and Mediterranean samples, our results evidenced that Alboran Sea samples were genetically closer to Atlantic samples than to eastern Mediterranean samples (Annaba from Algeria or North Tunisia and North Spain samples). This result agreed with previous studies which found that Atlantic hakes from Morocco (Roldán et al., 1998), Portugal (Lundy et al., 1999; Castillo et al., 2004) or Spain (Tanner et al., 2014) clustered closer to Mediterranean hakes than to Northeast Atlantic hakes (from Galician shelf to up north). Several studies argued that the Gibraltar Strait acts as a barrier to gene flow for hake, separating the Atlantic and Mediterranean populations, with limited exchange (Lundy et al., 1999; Castillo et al., 2004, 2005; Pita et al., 2010, 2014). Nevertheless, none of these studies analyzed simultaneously samples from both side of the Gibraltar Strait (north and south, east and west) and neither in

the close vicinity of the Gibraltar Strait. Our results, with a more exhaustive sampling design around the Gibraltar Strait, support the existence of ongoing gene flow from the Atlantic to the Mediterranean Sea with a west-to-east introgression of the Atlantic genetic component into the Mediterranean Sea and the Alboran Sea being a transition zone between both water bodies as previously suggested for hake (Milano et al., 2014). However, whether this gene flow is the result of passive dispersal of eggs and larvae or the active migration of juvenile or adult has to been determined.

Within the Mediterranean Sea, the genetic divergence observed between Alboran Sea versus North Tunisia and North of Spain samples indicated limited gene flow between Alboran Sea and eastern adjacent Mediterranean waters. The Almeria Oran front (Tintore et al., 1988) has been widely reported as a genetic break in several marine species (Patarnello et al., 2007) and European hake (Cimmaruta et al., 2005; Pita et al., 2014). However, Roquetas sample from North Alboran was genetically closer to North Spain samples than North Alboran ones. Nevertheless, our sampling came from landings and the exact fishing location of Roquetas was uncertain to argue wether the genetic split was located at the Almeria-Oran front, but this point should be further investigated with an appropriate sampling design.

Our results, underlined that environmental factors (salinity and chlorophll a) better explained the genetic variation at SNP loci compared to spatial factors suggesting that environmental factors could be also shaping the genetic structure of hake. Milano et al. (2014) found significant correlations between allele frequencies of several SNP outlier loci and seawater surface temperature and salinity and argued that hake populations might be adapted to local conditions. Based on two allozyme loci (*Gapdh and Gpi-2*), a previous study showed a strong correlation between genetic variation and salinity and temperature values, suggesting that these two environmental factors may play a role for selective processes and maintaining the genetic structure of hake.

Management implications

In the Mediterranean Sea, the European hake is assessed and managed following the 27 GSA and assuming one stock per GSA. This study highlighted the need for taking into account the genetic connectivity between the different GSA especially between GSA 4 (Algeria) and GSA 3 (South Alboran) with Ghazaouet sample being genetically closer to South Alboran samples than to Annaba sample and between GSA 4 (Algeria) and GSA 12 (North Tunisia) with Annaba sample being genetically closer to North Tunisia samples than to Ghazaouet one. However, this study could not provide information on where the exact boundary should be located among GSA 3, 4 and 12 as more samples from GSA 4 are needed. Moreover, the limit between GSA 1 (North Alboran) and GSA6 (North Spain) should be further investigated as Roquetas sample seemed to be genetically closer to North Spain samples than to North Alboran samples. Therefore, the assessment and management of European hake within the western Mediterranean Sea should be occurred jointly among countries as already suggested (Benchoucha et al., 2012). Finally, the ongoing gene flow between the Atlantic and Mediterranean populations of the European hake suggests that a cohesive stock management of this species among the ICES, CECAF and GFCM should be considered.

References

Belkhir, K., Borsa, P., Chikhi, L., Raufaste, N., and Bonhomme, F. (2004). GENETIX 4.05, logiciel sous Windows TM pour la génétique des populations. Laboratoire, Génome, Populations, Interactions, CNRS UMR 5000, Université de Montpellier II, Montpellier, France.

Benchoucha, Pérez Gil, J. L., Ainouche, N., Jarbui, O., Baro, J., Elouamari, N., et al. (2012). Advances in preparing a joint assessment of European hake, Merluccius merluccius, stock for GSAs 01, 02, 03 and 04 of the GFCM (Algeria, Morocco and Spain). Paper presented at the Working Group on Stock Assessment of Demersal Species (SCSA-SAC, GFCM), (Split, Croatia, 5-9 November 2012). GCP/IN /028/SPA-GCP/INT/006/EC. CopeMed II Occasional Papers n^{o} 14: 19 pp.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 57, 289–300.

Bernatchez, S., Xuereb, A., Laporte, M., Benestan, L., Steeves, R., Laflamme, M., et al. (2019). Seascape genomics of eastern oyster (Crassostrea virginica) along the Atlantic coast of Canada. *Evolutionary Applications* 12, 587–609. doi:10.1111/eva.12741.

Casey, J., and Pereiro, J. (1995). “European hake (M. merluccius) in the North-east Atlantic,” in Hake (Dordrecht: Springer), 125–147.

Castillo, A. G. F., Alvarez, P., and Garcia-Vazquez, E. (2005). Population structure of Merluccius merluccius along the Iberian Peninsula coast. *ICES Journal of Marine Science* 62, 1699–1704. doi:10.1016/j.icesjms.2005.06.001.

Castillo, A. G. F., Martinez, J. L., and Garcia-Vazquez, E. (2004). Fine Spatial Structure of Atlantic Hake (Merluccius merluccius) Stocks Revealed by Variation at Microsatellite Loci. *Mar. Biotechnol.* 6, 299–306. doi:10.1007/s10126-004-3027-z.

Caye, K., Deist, T. M., Martins, H., Michel, O., and François, O. (2016). TESS3: fast inference of spatial population structure and genome scans for selection. *Molecular Ecology Resources* 16, 540–548. doi:10.1111/1755-0998.12471.

Caye, K., Jay, F., Michel, O., and François, O. (2018). Fast inference of individual admixture coefficients using geographic data. *The Annals of Applied Statistics* 12, 586–608. doi:10.1214/17-AOAS1106.

Chapuis, M.-P., and Estoup, A. (2007). Microsatellite null alleles and estimation of population differentiation. *Mol Biol Evol* 24, 621–631. doi:10.1093/molbev/msl191.

Cimmaruta, R., Bondanelli, P., and Nascetti, G. (2005). Genetic structure and environmental heterogeneity in the European hake (Merluccius merluccius). *Molecular Ecology* 14, 2577–2591. doi:10.1111/j.1365-294X.2005.02595.x.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* 39, 1–38.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., et al. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* 36, 27–46. doi:10.1111/j.1600-0587.2012.07348.x.

Earl, D. A., and vonHoldt, B. M. (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genet Resour* 4, 359–361. doi:10.1007/s12686-011-9548-7.

Excoffier, L., Laval, G., and Schneider, S. (2005). Arlequin (version 3.0): An integrated software package for population genetics data analysis. *Evol Bioinform Online* 1, 47–50.

Falush, D., Stephens, M., and Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 164, 1567–1587.

Falush, D., Stephens, M., and Pritchard, J. K. (2007). Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Molecular Ecology Notes* 7, 574–578. doi:10.1111/j.1471-8286.2007.01758.x.

FAO (2018). The State of Mediterranean and Black Sea Fisheries. General Fisheries Commission for the Mediterranean. Rome. 172 pp. Licence: CC BY-NC-SA 3.0 IGO.

Foll, M., and Gaggiotti, O. (2008). A Genome-Scan Method to Identify Selected Loci Appropriate for Both Dominant and Codominant Markers: A Bayesian Perspective. *Genetics* 180, 977–993. doi:10.1534/genetics.108.092221.

Hauser, L., Adcock, G. J., Smith, P. J., Ramírez, J. H. B., and Carvalho, G. R. (2002). Loss of microsatellite diversity and low effective population size in an overexploited population of New Zealand snapper (Pagrus auratus). *PNAS* 99, 11742–11747. doi:10.1073/pnas.172242899.

ICES (2018). Report of the Working Group for the Bay of Biscay and the Iberian Waters Ecoregion (WGBIE). ICES CM 2018/ACOM:12. 585 pp.

Jakobsson, M., and Rosenberg, N. A. (2007). CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* 23, 1801–1806. doi:10.1093/bioinformatics/btm233.

Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* 24, 1403–1405. doi:10.1093/bioinformatics/btn129.

Jost, L. (2008). GST and its relatives do not measure differentiation. *Molecular Ecology* 17, 4015–4026. doi:10.1111/j.1365-294X.2008.03887.x.

Kalinowski, S. T. (2005). HP-RARE 1.0: a computer program for performing rarefaction on measures of allelic richness. *Molecular Ecology Notes* 5, 187–189. doi:10.1111/j.1471-8286.2004.00845.x.

Leone, A., Álvarez, P., García, D., Saborido-Rey, F., and Rodriguez-Ezpeleta, N. (2019). Genome-wide SNP based population structure in European hake reveals the need for harmonizing biological and management units. *ICES Journal of Marine Science* 76, 2260–2266. doi:10.1093/icesjms/fsz161.

Lundy, C. J., Moran, P., Rico, C., Milner, R. S., and Hewitt, G. M. (1999). Macrogeographical population differentiation in oceanic environments: a case study of European hake (Merluccius merluccius), a commercially important fish. *Molecular Ecology* 8, 1889–1898. doi:10.1046/j.1365-294x.1999.00789.x.

Luu, K., Bazin, E., and Blum, M. G. B. (2017). pcadapt: an R package to perform genome scans for selection based on principal component analysis. *Molecular Ecology Resources* 17, 67–77. doi:10.1111/1755-0998.12592.

Milano, I., Babbucci, M., Cariani, A., Atanassova, M., Bekkevold, D., Carvalho, G. R., et al. (2014). Outlier SNP markers reveal fine-scale genetic structuring across European hake populations (Merluccius merluccius). *Molecular Ecology* 23, 118–135. doi:10.1111/mec.12568.

Morán, P., Lundy, C., Rico, C., and Hewitt, G. M. (1999). Isolation and characterization of microsatellite loci in European hake, Merlucius merlucius (Merlucidae, Teleostei). *Molecular Ecology* 8, 1357–1358. doi:10.1046/j.1365-294X.1999.00701_4.x.

Naimi, B. (2017). usdm: uncertainty analysis for species distribution models. r package version 1.1–18. RDocumentation https://cran.r-project.org/web/packages/usdm/usdm.pdf.

Naimi, B., Hamm, N. A. S., Groen, T. A., Skidmore, A. K., and Toxopeus, A. G. (2014). Where is positional uncertainty a problem for species distribution modelling? *Ecography* 37, 191–203. doi:10.1111/j.1600-0587.2013.00205.x.

Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences of the United States of America* 70, 3321–3323. doi:VL - 70.

Oksanen, J. (2015). Vegan: an introduction to ordination. URL http://cran. r-project.org/web/packages/vegan/vignettes/introvegan. pdf 8, 19.

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

Peakall, R., and Smouse, P. E. (2012). GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research—an update. *Bioinformatics* 28, 2537–2539. doi:10.1093/bioinformatics/bts460.

Pita, A., Leal, A., Santafé-Muñoz, A., Piñeiro, C., and Presa, P. (2016). Genetic inference of demographic connectivity in the Atlantic European hake metapopulation (Merluccius merluccius) over a spatio-temporal framework. *Fisheries Research* 179, 291–301. doi:10.1016/j.fishres.2016.03.017.

Pita, A., Pérez, M., Balado, M., and 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. doi:10.1016/j.seares.2013.11.003.

Pita, A., Pérez, M., Cerviño, S., and Presa, P. (2011). What can gene flow and recruitment dynamics tell us about connectivity between European hake stocks in the Eastern North Atlantic? *Continental Shelf Research* 31, 376–387. doi:10.1016/j.csr.2010.09.010.

Pita, A., Presa, P., and Pérez, M. (2010). Gene flow, multilocus assignment and genetic structuring of the European hake (Merluccius merluccius). *Thalassas* 26, 129–133.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959.

Recasens, L., Lombarte, A., Morales-Nin, B., and Tores, G. J. (1998). Spatiotemporal variation in the population structure of the European hake in the NW Mediterranean. *Journal of Fish Biology* 53, 387–401. doi:10.1111/j.1095-8649.1998.tb00988.x.

Reiss, H., Hoarau, G., Dickey-Collas, M., and Wolff, W. J. (2009). Genetic population structure of marine fish: mismatch between biological and fisheries management units. *Fish and Fisheries* 10, 361–395.

Rico, C., Ibrahim, K. M., Rico, I., and Hewitt, G. M. (1997). Stock composition in North Atlantic populations of whiting using microsatellite markers. *Journal of Fish Biology* 51, 462–475. doi:10.1111/j.1095-8649.1997.tb01504.x.

Roldán, M. I., García-Marín, J. L., Utter, F. M., and Pla, C. (1998). Population genetic structure of European hake, Merluccius merluccius. *Heredity* 81, 327–334. doi:10.1046/j.1365-2540.1998.00383.x.

Rosenberg, N. A. (2004). DISTRUCT: a program for the graphical display of population structure. *Molecular Ecology Notes* 4, 137–138. doi:10.1046/j.1471-8286.2003.00566.x.

Rousset, F. (2008). GENEPOP'007: a complete re-implementation of the GENEPOP software for Windows and Linux. *Molecular Ecology Resources* 8, 103–106. doi:10.1111/j.1471-8286.2007.01931.x.

Selmoni, O., Lecellier, G., Vigliola, L., Berteaux-Lecellier, V., and Joost, S. (2020). Coral cover surveys corroborate predictions on reef adaptive potential to thermal stress. *Sci Rep* 10, 19680. doi:10.1038/s41598-020-76604-2.

Tanner, S. E., Pérez, M., Presa, P., Thorrold, S. R., and Cabral, H. N. (2014). Integrating microsatellite DNA markers and otolith geochemistry to assess population structure of European hake (Merluccius merluccius). *Estuarine, Coastal and Shelf Science* 142, 68–75. doi:10.1016/j.ecss.2014.03.010.

Tintore, J., La Violette, P. E., Blade, I., and Cruzado, A. (1988). A study of an intense density front in the Eastern Alboran Sea: the Almeria-Oran Front. *Journal of Physical Oceanography* 18, 1384–1397.

Van Oosterhout, C. V., Hutchinson, W. F., Wills, D. P. M., and Shipley, P. (2004). MICRO-CHECKER: software for identifying and correcting genotyping errors in microsatellite data. *Molecular Ecology Notes* 4, 535–538. doi:10.1111/j.1471-8286.2004.00684.x.

Waples, R. S., Punt, A. E., and Cope, J. M. (2008). Integrating genetic data into management of marine resources: how can we do it better? *Fish and Fisheries* 9, 423–449.

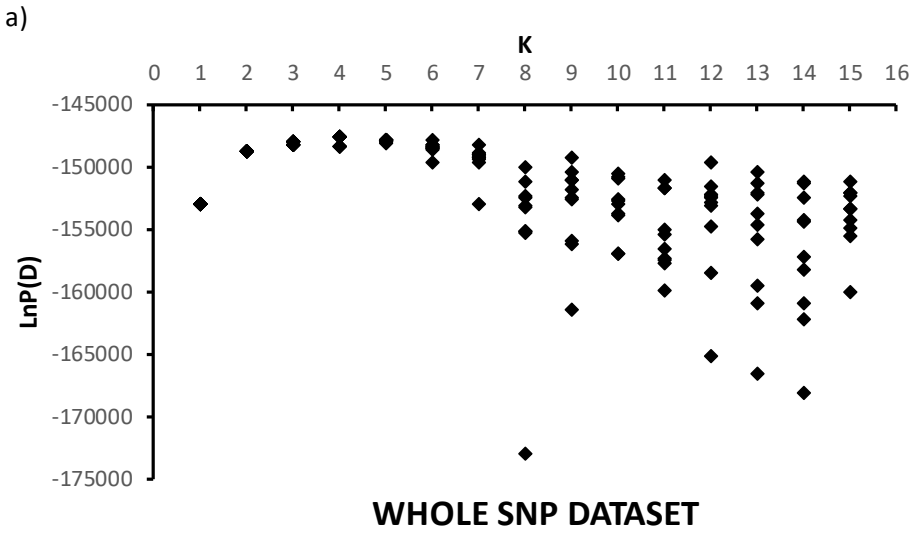
Weir, B. S., and Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution* 38, 1358–1370.

Westgaard, J.-I., Staby, A., Aanestad Godiksen, J., Geffen, A. J., Svensson, A., Charrier, G., et al. (2017). Large and fine scale population structure in European hake (Merluccius merluccius) in the Northeast Atlantic. *ICES Journal of Marine Science* 74, 1300–1310. doi:10.1093/icesjms/fsw249.

SUPPORTING INFORMATION

Table S1. *f* estimator of F_{IS} per sample and locus (significant values in bold; 0.05 threshold after FDR correction) and null allele frequencies given in brackets when MICRO-CHECKER detected the occurrence of null alleles. Nnull: no null allele detected.

Sample \ Locus	Mmer-hk20	Mmer-hk9b	Mmer-hk3b	Mmer-hk29	Mmer-hk34b	Mmer-UEAW01
AGA	0.109	0.07	-0.025	**0.577** (0.277)	**0.273** (0.126)	0.038
MHD	0.014	**0.08** (0.036)	0.043	**0.6** (0.285)	**0.325** (0.15)	0.06
HUE	-0.047	-0.014	-0.013	**0.683** (0.326)	**0.21** (0.093)	0.028
CDZ	0.024	-0.006	0.029	**0.571** (0.271)	**0.142** (Nnull)	-0.023
ETP	-0.036	0.017	0.092	**0.689** (0.322)	**0.17** (0.073)	0.023
MLG	-0.008	-0.022	-0.027	**0.743** (0.353)	**0.228** (0.102)	-0.018
RQT	0.01	0.033	0.069	**0.637** (0.302)	**0.264** (0.13)	-0.019
MDQ	0.061	0.029	-0.072	**0.681** (0.325)	**0.193** (Nnull)	0.056
NDR	-0.088	-0.002	-0.129	**0.676** (0.318)	**0.222** (Nnull)	**0.139** (Nnull)
GHZ	0.06	**0.049** (Nnull)	-0.035	**0.738** (0.35)	**0.253** (0.115)	0.033
ANB	-0.006	-0.004	-0.031	**0.481** (0.227)	**0.21** (Nnull)	0.019
TBK	0.016	-0.021	0.072	**0.642** (0.299)	**0.118** (0.055)	0.034
GTU	0.028	-0.031	0.111	**0.735** (0.352)	**0.286** (0.127)	-0.047
TOR	-0.077	**0.031** (Nnull)	0.044	**0.622** (0.294)	**0.321** (0.152)	-0.009
CAS	**0.064** (Nnull)	-0.001	0.047	**0.682** (0.324)	**0.192** (0.085)	-0.013



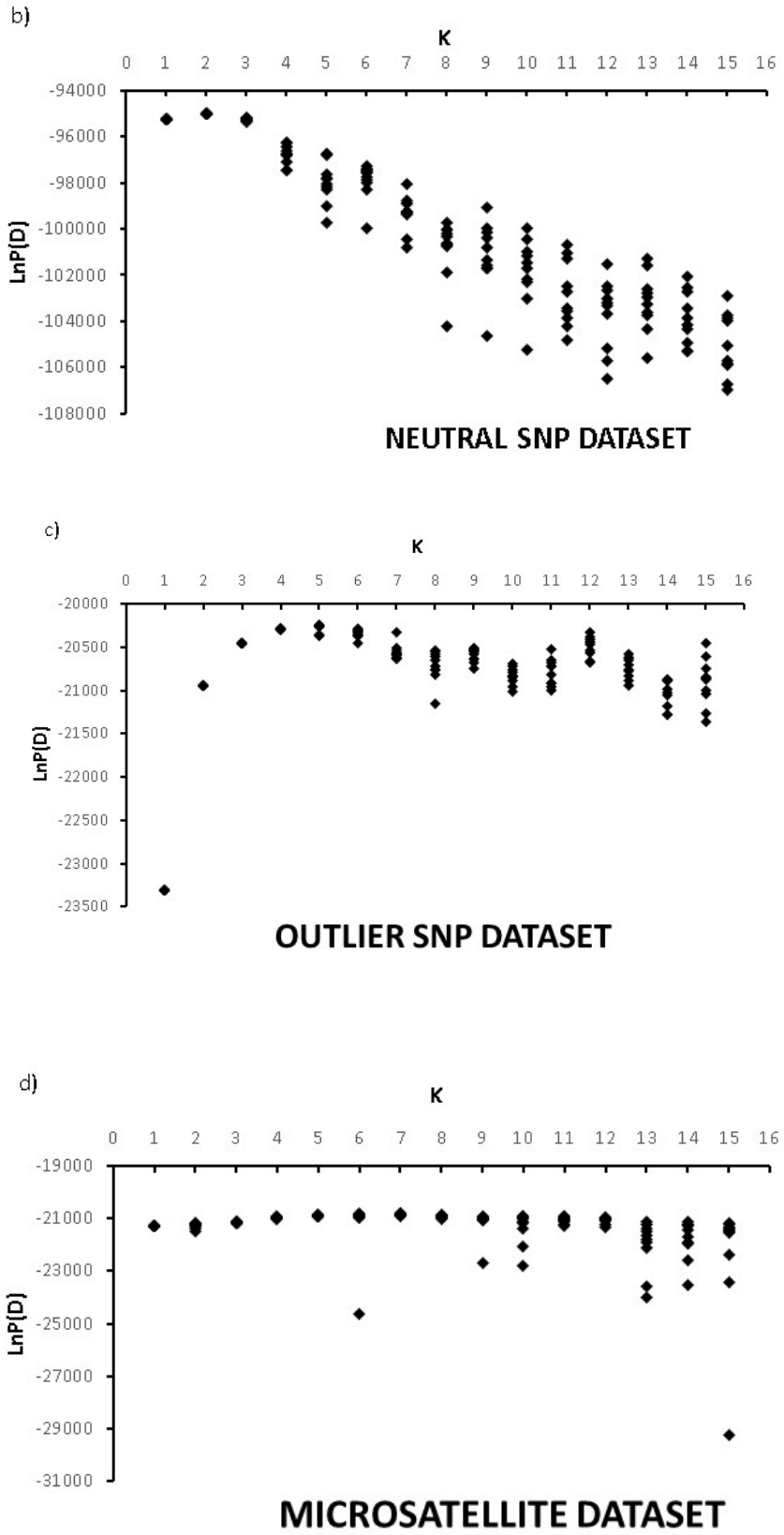
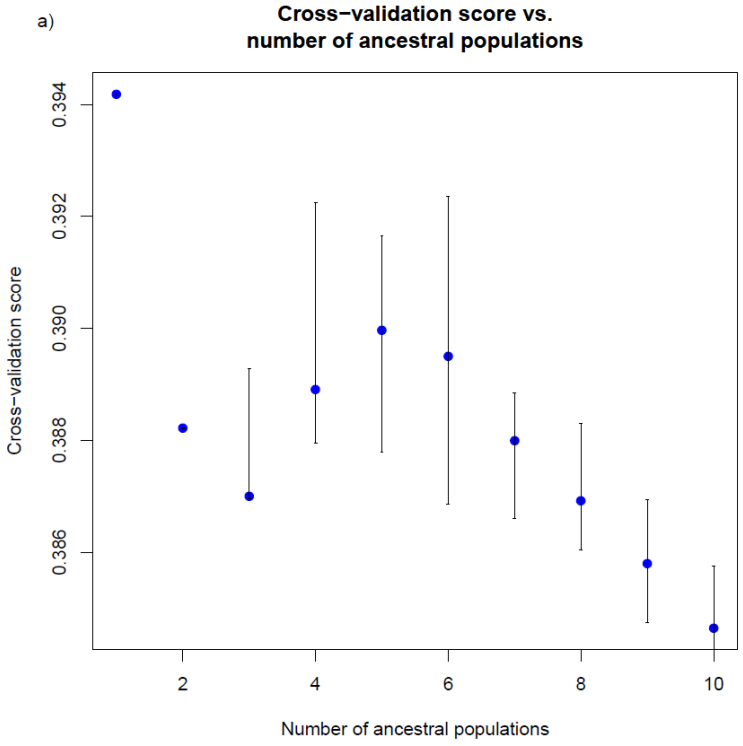
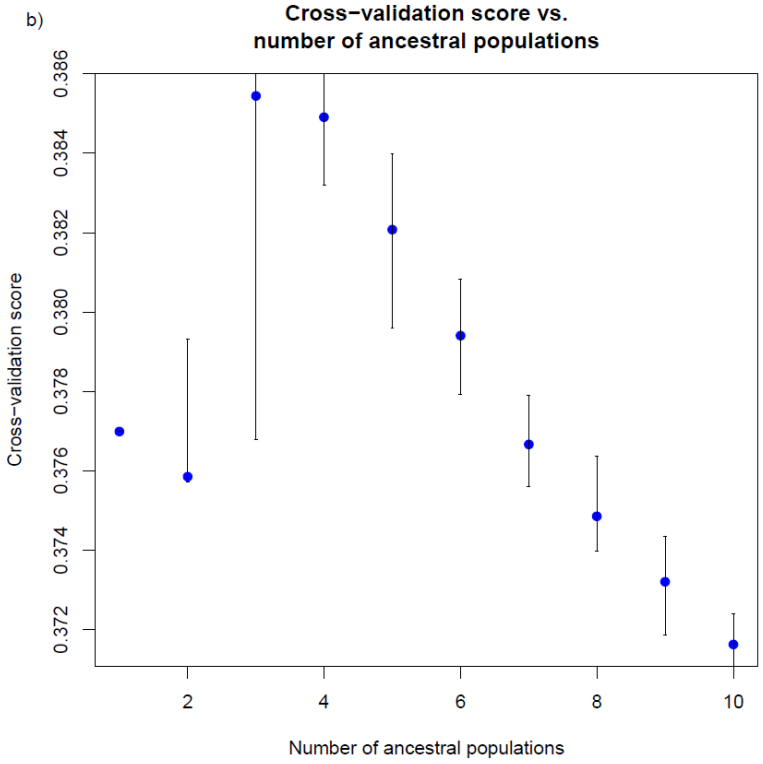


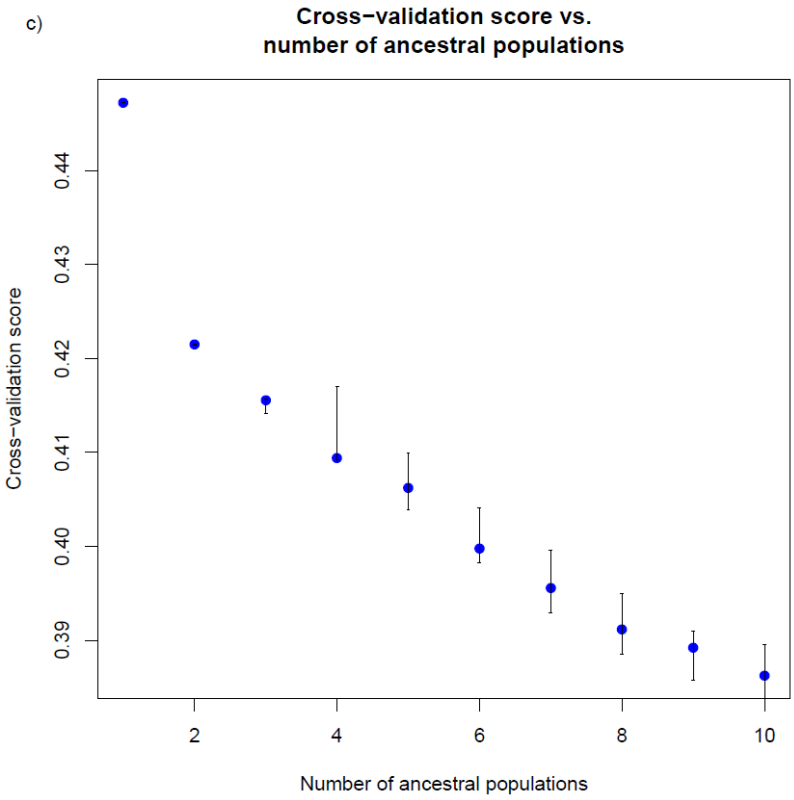
Figure S1. Plot of LnP(D) as a function of the number of clusters (K) across the 10 runs for: a) whole, b) neutral, c) outlier and d) microsatellite datasets respectively.

a)


Cross−validation score vs.
number of ancestral populations
Cross−validation score
0.386
0.388
0.390
0.392
0.394
2
4
6
8
10
Number of ancestral populations

b)


Cross−validation score vs.
number of ancestral populations
Cross−validation score
0.372
0.374
0.376
0.378
0.380
0.382
0.384
0.386
2
4
6
8
10
Number of ancestral populations



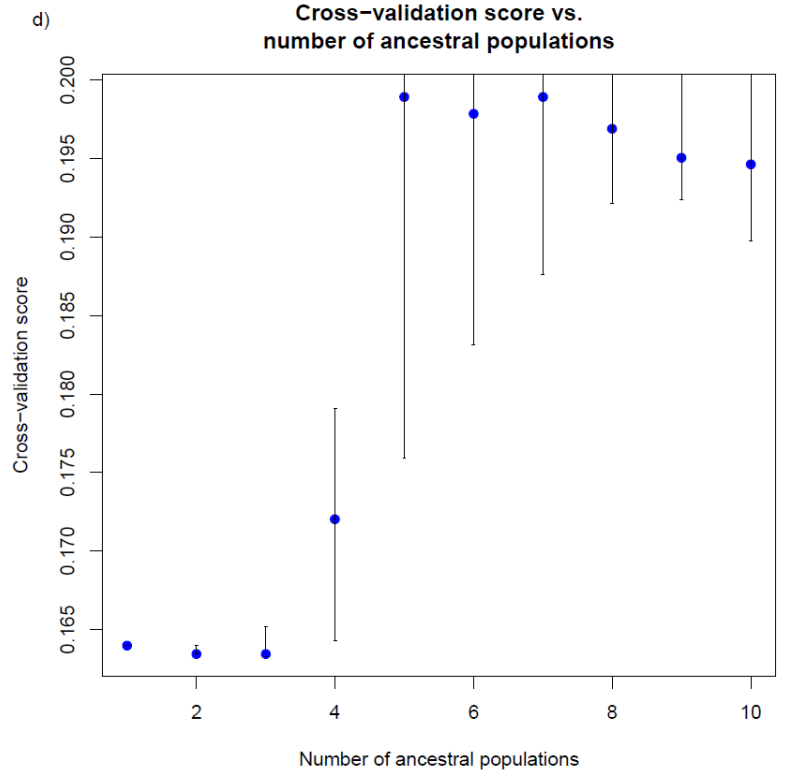


Figure S2. Cross-validation sore profile showing putative ancestral population for: a) whole, b) neutral, c) outlier and d) microsatellite datasets respectively.

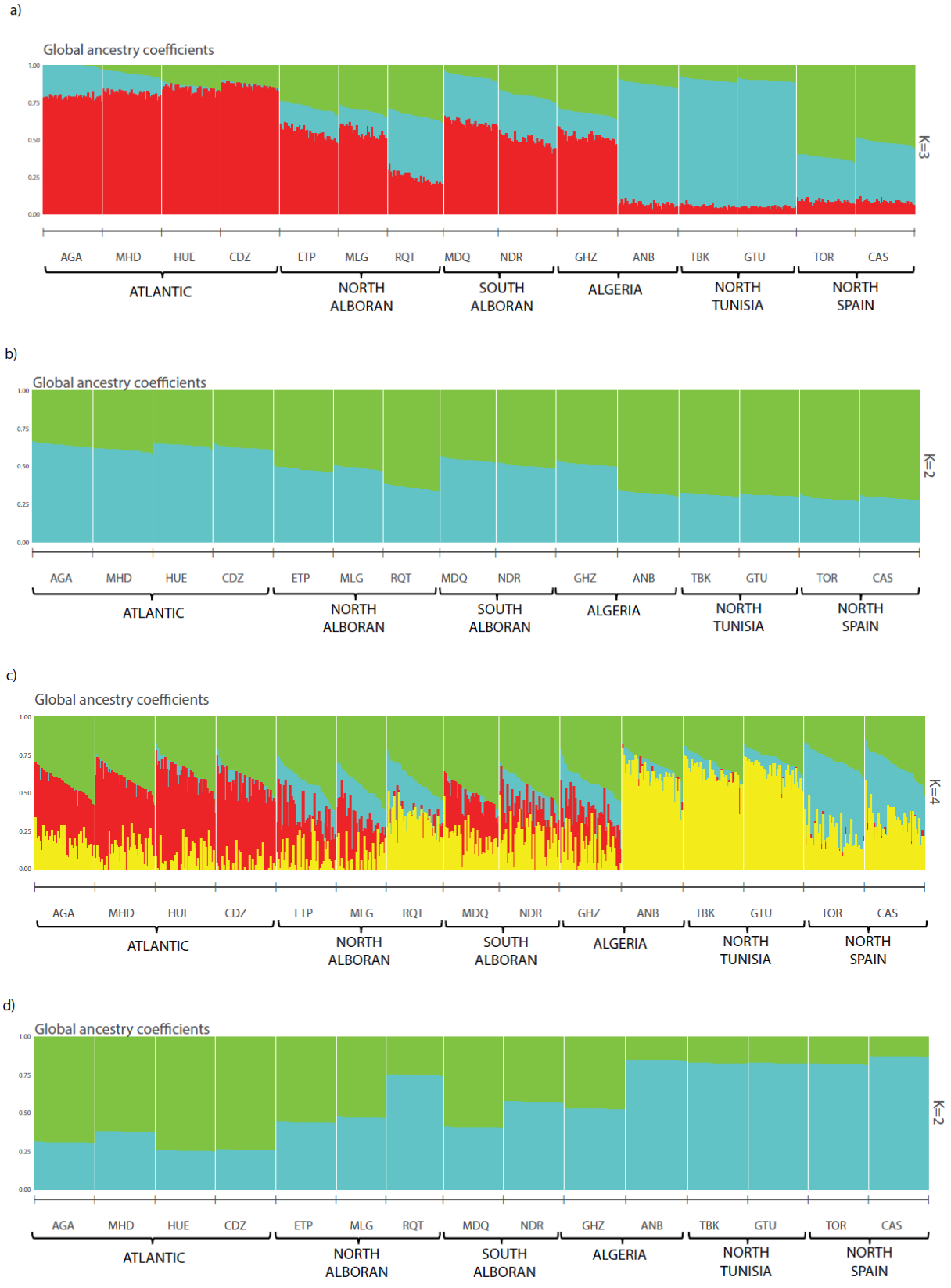
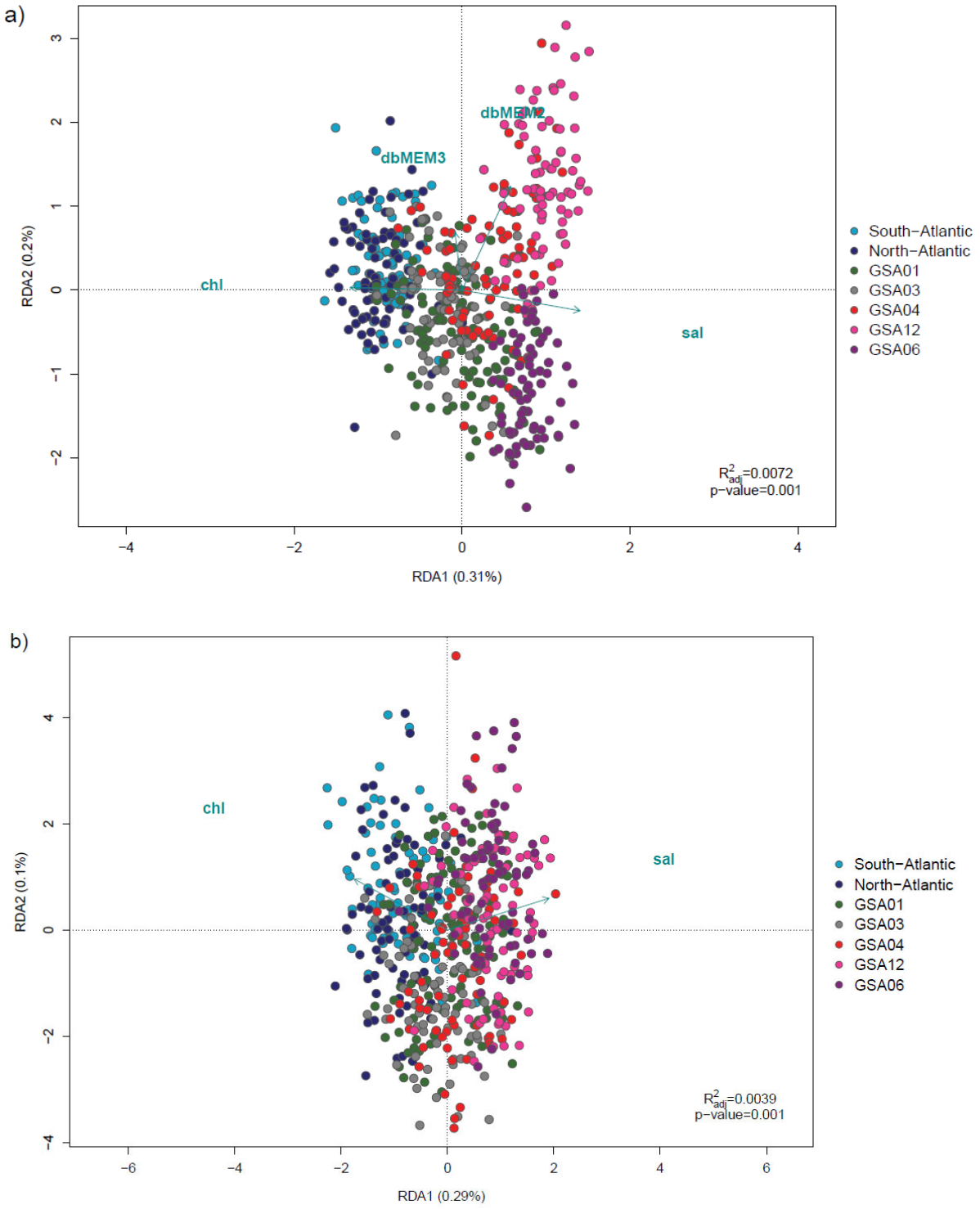


Figure S3. Barplot representation of the Qmatrix for: a) whole, b) neutral, c) outlier an d) microsatellite datasets respectively.


a)
dbMEM2
dbMEM3
chl
sal
RDA2 (0.2%)
RDA1 (0.31%)
R^2_{adj}=0.0072
p−value=0.001
South−Atlantic
North−Atlantic
GSA01
GSA03
GSA04
GSA12
GSA06
b)
chl
sal
RDA2 (0.1%)
RDA1 (0.29%)
R^2_{adj}=0.0039
p−value=0.001
South−Atlantic
North−Atlantic
GSA01
GSA03
GSA04
GSA12
GSA06

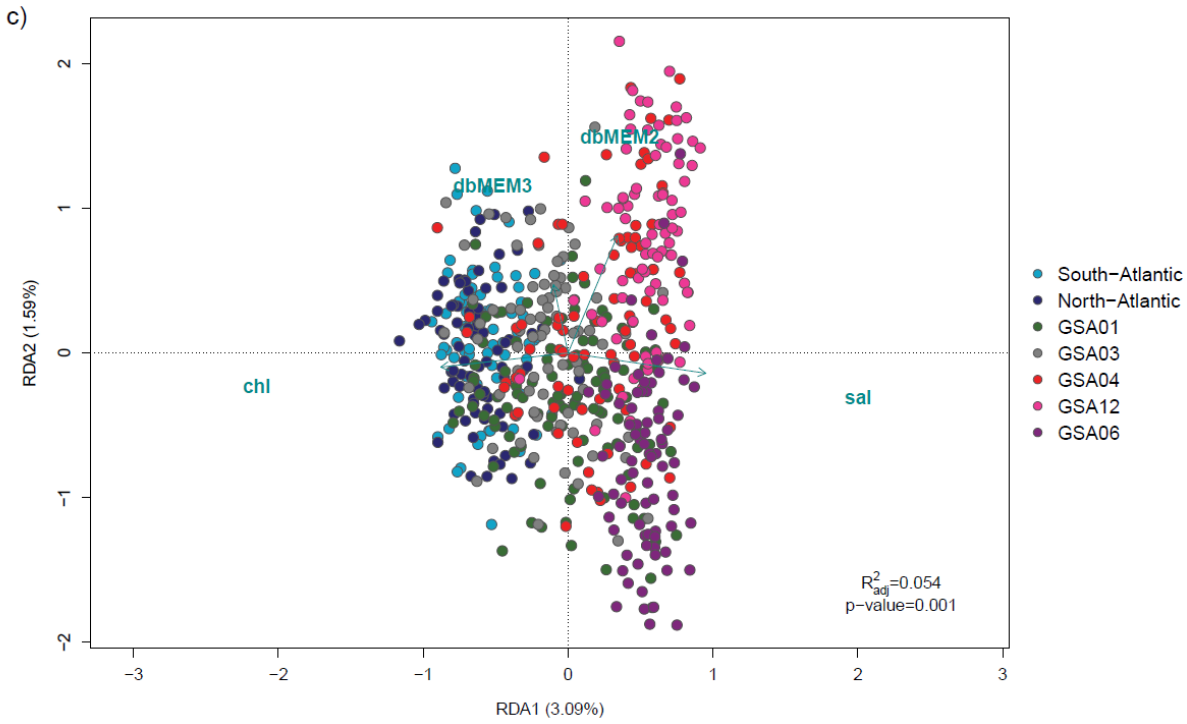


Figure S4. Redundancy analysis (RDA) performed on: a) whole, b) neutral and c) outlier datasets following the ordistep procedure. Significant variables are shown with arrows.