Population genomics data supports introgression between western Iberian *Squalius* freshwater fish species

Sofia L. Mendes[1], Maria M. Coelho[1*], Vitor C. Sousa[1*&]

[1] cE3c – Centre for Ecology, Evolution and Environmental Changes, Faculdade de Ciências da Universidade de Lisboa, Lisbon, Portugal

[*]These authors contributed equally to this work

[&]corresponding author: vmsousa@fc.ul.pt

Abstract

In freshwater fish, processes of population divergence and speciation are often linked to the geomorphology of rivers and lakes that isolate populations. However, current geographical isolation does not necessarily imply total absence of gene flow during the divergence process. Here, we focused on four species of the genus *Squalius* in Portuguese rivers: *S. carolitertii*, *S. pyrenaicus*, *S. aradensis* and *S. torgalensis*. Previous studies based on eight nuclear and mitochondrial markers revealed incongruent patterns, with nuclear loci suggesting that *S. pyrenaicus* was a paraphyletic group since its northern populations were genetically closer to *S. carolitertii* than to its southern populations. For the first time, we successfully applied a genomic approach to the study of the relationship between these species, using a Genotyping by Sequencing (GBS) approach. Our results revealed a species tree with two main lineages: (i) *S. carolitertii* and *S. pyrenaicus*; (ii) *S. torgalensis* and *S. aradensis*. Regarding *S. carolitertii* and *S. pyrenaicus*, we found evidence for past introgression between the two in the northern part of *S. pyrenaicus* distribution. This introgression reconciles previous mitochondrial and nuclear incongruent results and explains the apparent paraphyly of *S. pyrenaicus*. Our demographic modelling estimates are consistent across models, suggesting that the northern populations of *S. pyrenaicus* received a contribution of approximately 90% from *S. carolitertii* and 10% from southern *S. pyrenaicus*. This illustrates that even in freshwater species currently allopatric, we are able to detect past gene flow

events in present-day genomes, suggesting speciation is more complex than simply allopatric.

Key-words: Iberian freshwater fish; introgression; speciation; demographic modelling

Introduction

Answering questions regarding how populations diverge and ultimately originate new species is a major goal of evolutionary biology. Speciation is assumed to occur due to a systematic reduction in gene flow through time until reproductive isolation is achieved and populations maintain phenotypic and genetic distinctiveness (Seehausen *et al.*, 2014). The most accepted hypothesis is that divergence happens in a strictly allopatric scenario in the absence of gene flow, due to barriers (geological, hydrological, etc.). Without gene flow, genetic incompatibilities are expected to accumulate through time which can lead to reproductive isolation (Sousa and Hey, 2013). However, there are now several studies based on phenotypic and genomic data suggesting that past gene flow is common in several species, including in humans (e.g. Green *et al.*, 2010; Dasmahapatra *et al.*, 2012; Lamichhaney *et al.*, 2015; vonHoldt *et al.*, 2016; de Manuel *et al.*, 2016). Nevertheless, despite the growing number of examples of gene flow between species, it is still unclear whether gene flow accompanies the divergence process or if populations first get isolated and then come into contact after a period of time, i.e. a secondary contact (Sousa and Hey, 2013). Thus, to understand the process of speciation it is important to characterize the timing and mode of gene flow. The study of these processes has been revolutionized by the possibility of generating genome-wide data from multiple individuals of closely related species to obtain large numbers of polymorphic genetic markers scattered across the genome, either by reduced representation (e.g. genotyping by sequencing)

or whole genome sequencing (Davey *et al.*, 2011; Andrews *et al.*, 2016). These types of data have been used in the study of speciation and the relationship between species in several taxa, from insects (e.g. Dasmahapatra et al. 2012; Bagley et al. 2017) to mammals (e.g. McManus et al. 2015; Figueiró et al. 2017), including freshwater fish (e.g. Hohenlohe et al. 2010; Meier et al. 2017).

Due to their outstanding diversity and remarkable adaptive radiations, freshwater fish species have been widely used as model systems to study speciation (Seehausen and Wagner, 2014). A variety of scenarios have been described to explain the differentiation of different freshwater fish populations, including transitions from marine to freshwater habitats (e.g. Jones et al. 2012; Terekhanova et al. 2014), adaptation to extreme environments (e.g. Pfenninger et al. 2015), and differentiation along water depth clines (e.g. Barluenga et al. 2006; Gagnaire et al. 2013). Another important factor for speciation in freshwater fish is the geomorphology of rivers and lakes, given that such events are often associated with geological barriers isolating populations (Seehausen and Wagner, 2014). However, this does not mean that currently geographically separated populations have always been isolated since the configuration of river and lake systems can change over geological time. In fact, several studies document both past and ongoing introgression in freshwater fish, both in species that have evolved with and without geographical isolation (Redenbach and Taylor, 2002; Jones *et al.*, 2013; Hohenlohe *et al.*, 2013; Gante *et al.*, 2016). Nonetheless, vicariant events resulting from

geographical barriers imposed by the geomorphology of lakes and rivers remains the most accepted explanation for the abundance of freshwater fish species (Seehausen and Wagner, 2014). One geographical area where isolation and the configuration of the drainage systems are assumed to have fuelled the origin of a multitude of endemic fish species is the Iberian Peninsula (Sousa-Santos *et al.*, 2019).

The freshwater fish fauna of the Iberian Peninsula includes several endemic species (Mesquita *et al.*, 2007). Among these, a diverse group are the “chubs” from the genus *Squalius* Bonaparte, 1837, in which there are currently eight species and a hybrid complex described in the peninsula (Perea *et al.*, 2016). In Portuguese rivers, apart from the hybrid complex, four species can be found: *Squalius carolitertii*, *Squalius pyrenaicus*, *Squalius torgalensis* and *Squalius aradensis* (Figure 1), distributed along a temperature cline, with increasing temperatures from north to south (Jesus *et al.*, 2017). Two of the species have rather wide distribution ranges: *Squalius carolitertii* (Doadrio, 1988) is endemic to the northern region of the peninsula and can be found in the northern rivers up to the Mondego basin, while *Squalius pyrenaicus* (Gunther, 1868) has a more southern distribution range and is considered to be present in the Tagus, Sado, Guadiana and south-eastern basins (Coelho *et al.*, 1995, 1998). On the other hand, the two other species are confined only to small river systems in the southwestern area of the country: *Squalius torgalensis* (Coelho *et al.*, 1998) is endemic to the Mira river basin and *Squalius aradensis* (Coelho *et al.*,

1998) is restricted to basins (e.g. Arade) in the extreme southwestern area (Coelho *et al.*, 1998).

The relationship between these species has been investigated (e.g. Brito et al. 1997; Sanjur et al. 2003; Mesquita et al. 2007; Waap et al. 2011; Sousa-Santos et al. 2019) and estimates based on fossil calibrations, nuclear and mitochondrial markers date their most recent common ancestor to ≈14 Mya (Perea *et al.*, 2010; Sousa-Santos *et al.*, 2019). *S. torgalensis* and *S. aradensis* were found to be sister species, forming one clade distinct from the clade of sister species *S. carolitertii* and *S. pyrenaicus*, based on both mitochondrial (mtDNA) and nuclear markers (Brito *et al.*, 1997; Mesquita *et al.*, 2007; Almada and Sousa-Santos, 2010; Waap *et al.*, 2011; Sousa-Santos *et al.*, 2019). However, while the mtDNA trees cluster different populations of *S. pyrenaicus* from different river basins together (Brito *et al.*, 1997; Mesquita *et al.*, 2007), phylogenetic trees produced concatenating 7 nuclear genes suggest that *S. pyrenaicus* individuals from the Tagus river basin cluster with *S. carolitertii*, instead of clustering with *S. pyrenaicus* from other river basins further south (e.g. Guadiana, Sado, Almargem), which form a separate clade (Waap *et al.*, 2011; Sousa-Santos *et al.*, 2019).

While the previous work provided valuable information to understand the diversity and taxonomy of these species (Coelho *et al.*, 1995; Brito *et al.*, 1997; Mesquita *et al.*, 2005; Henriques *et al.*, 2010), their evolutionary history was mostly investigated based on single mtDNA gene trees and recently complemented with seven nuclear markers (Waap *et al.*,

2011; Sousa-Santos *et al.*, 2019). Investigating the history of species based on single genes can be problematic due to highly stochastic events of genetic drift and mutational processes (Hey and Machado, 2003). Moreover, when species diverged relatively recently, gene trees might not reflect the underlying species tree due to incomplete lineage sorting and/or gene flow (Hey and Machado, 2003). Thus, although seven nuclear genes constitute an improvement over phylogenies based only on mitochondrial DNA, it still provides a limited picture of the genome. Therefore, this work had two major goals: (i) first, to characterize the genome-wide patterns of genetic differentiation and reconstruct the species tree for these four *Squalius* species in Portuguese river basins; (ii) second, to investigate the possibility of introgression between *S. carolitertii* and *S. pyrenaicus*, given the previously reported incongruent results between mtDNA and nuclear markers. To achieve these goals, we successfully obtained genome-wide single nucleotide polymorphisms (SNPs) through a Genotyping by Sequencing (GBS) protocol.

Methods

Sampling and sequencing

A total of 65 individuals were sampled from 8 different locations, as displayed in Figure 1. For each species, at least one sampling location from a representative drainage system was sampled. For *S. carolitertii*, individuals were collected from the Mondego basin (n=10). For *S. pyrenaicus*, in the northern part of its distribution individuals were collected from the Ocreza river (n=10) and Canha stream (n=10), both tributaries of the Tagus basin. Specimens were also collected in the Lizandro basin (n=10). From here on, we use “northern *S. pyrenaicus*” to refer to *S. pyrenaicus* from Ocreza, Canha and Lizandro. In the southern part of the distribution, *S. pyrenaicus* was sampled in the Guadiana (n=2) and Almargem (n=8) basins, which we refer to as “southern *S. pyrenaicus*”. For *S. aradensis*, individuals were collected from the Arade (n=5) basin. For *S. torgalensis* individuals were collected in the Mira basin (n=10). Detailed locations with GPS coordinates and fishing licenses from the Portuguese authority for conservation of endangered species [ICNF (Instituto de Conservação da Natureza e das Florestas)] can be found in Table S1.

Fish were collected by electrofishing (300V, 4A), and total genomic DNA was extracted from fin clips using a phenol-chloroform protocol adapted from Taggart *et al.*, 1992. DNA was quantified using Qubit® 2.0 Fluorometer (Live Technologies). The samples were subjected to a paired-

end Genotyping by Sequencing (GBS) protocol (adapted from Elshire *et al.*, 2011), performed in outsourcing at Beijing Genomics Institute (BGI, www.bgi.com). The DNA samples were sent to the facility mixed with DNAstable Plus (Biomatrica) to preserve DNA at room temperature during shipment. Briefly, upon arrival, DNA was fragmented using the restriction enzyme ApeKI and the fragments were amplified after adaptor ligation (Elshire *et al.*, 2011). The resulting library was sequenced using Illumina Hiseq2000.

Obtention of a high-quality SNP dataset

First, the quality of the sequences of each individual was assessed using FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). To compile the information from all individual reports, we used MultiQC (Ewels *et al.*, 2016) to merge and summarize the individual FastQC reports. Second, we used the program *process_radtags* from Stacks version 2.2 (Catchen *et al.*, 2013; Rochette *et al.*, 2019) to trim all reads to 82 base pairs and discard reads with low quality scores, using the default settings for the window size (0.15x the length of the read) and the base quality threshold (10 in Phred score). Given the absence of a reference genome for any of the species in this study, we built a reference catalog of all loci using a *denovo* assembly approach on Stacks version 2.2 (Catchen *et al.*, 2013; Rochette *et al.*, 2019). To determine the best parameters for the construction of the catalog, we followed the approach recommended by

Paris et al. 2017 (Figure S1). We decided to allow a maximum of 2 differences between sequences within the same individual (M=2) and a maximum of 4 differences between sequences from different individuals (n=4) for them to be considered the same locus on the catalog. We also required a minimum depth of coverage of 4x for every locus on the catalog (m=4). After building the catalog, given the possibility that forward and reverse sequences of the same fragment were treated as different loci, similar reads within the catalog were clustered using CD-HIT version 4.7 (Li and Godzik, 2006; Fu *et al.*, 2012). We used CD-HIT-EST from the CD-HIT package with a word length of 6 and a sequence identity threshold of 0.85.

Once we clustered similar reads within the catalog, this was treated as a reference and the reads from each individual were aligned against it using BWA-MEM from BWA version 0.7.17-r1188 (Li, 2013) with default parameters. The output alignments of BWA were sorted and unmapped reads were removed using Samtools version 1.8 (Li and Durbin, 2009). To call genotypes for each individual at each site and identify SNPs we used the method implemented on Freebayes v1.2.0 (Garrison and Marth, 2012). We applied further filters to keep only SNPs present in all sampling sites in at least 50% of the individuals using VCFtools version 0.1.15 (Danecek *et al.*, 2011).

To discard sites and genotypes that are more likely to be the result of sequencing or mapping errors, we applied filters on the minor allele frequency (MAF ≥ 0.01) and on the depth of coverage, keeping only genotypes with a depth of coverage (DP) between ¼ to 4 times the

individual median DP, after assessing the effect of different filtering options (Tables S2 and S3). The different filters were applied using a combination of options from VCFtools version 0.1.15 (Danecek *et al.*, 2011) and BCFtools version 1.6 (Li *et al.*, 2009). Finally, individuals with more than 50% missing data were removed from the dataset.

Characterization of the global patterns of genetic differentiation

To quantify the levels of differentiation between sampling locations, we calculated the pairwise F_{ST} using the Hudson estimator (Hudson *et al.*, 1992). Given that the sampling locations may not correspond to populations, we investigated fine population structure with individual-based methods. To understand how individuals cluster, we conducted a principal component analysis (PCA). The number of significant principal components was determined with the Tracy-Widom test (Patterson *et al.*, 2006) on all eigenvalues. Furthermore, individual ancestry proportions were estimated with the sparse Non-negative Matrix Factorization method (sNMF) (Frichot *et al.*, 2014). We tested values of K between 1 and 8, performing 100 repetitions for each K value. All calculations were performed in RStudio version 1.1.383 and R version 3.4.4 and the PCA and sNMF were performed using the package LEA (Frichot and François, 2015).

Inference of a population and species tree

Given that our sampling included different species and populations within species, we used the SNP data to reconstruct a species and population tree describing the relationships between the populations using TreeMix (Pickrell and Pritchard, 2012). We explored a scenario with no migration, as well as models allowing for up to two migration events. Since we do not have an outgroup, the position of the root was not specified, and thus the resulting trees are unrooted.

Effect of linked SNPs

It is noteworthy that PCA, sNMF and TreeMix methods assume that SNPs are independent, and thus results can be affected by linked SNPs in our dataset. Given the absence of a reference genome, we lack information on the location of the SNP markers. To verify if the results were influenced by potential linkage of SNP markers, we produced a dataset by dividing each scaffold of the catalog into blocks of 200 base pairs, which is larger than the mean size of GBS loci. We then selected the SNP with less missing data per block to generate a dataset with a single SNP per block. Using this “single SNP” dataset, we repeated the three aforementioned analysis.

Detection of introgression between *S. carolitertii* and *S. pyrenaicus*

To test for possible past introgression between *S. carolitertii* and *S. pyrenaicus* in the northern area of *S. pyrenaicus* distribution, we used the D-statistic (Durand *et al.*, 2011), which was used to distinguish between ancestral polymorphism and introgression by looking at four different populations related through a fixed species tree: two sister populations (P1 and P2), a third population that could be the source of introgressed genes (P3) and has a common ancestor to P1 and P2, and one outgroup (Pout). We explored four different possible species trees to perform different tests. In scenario A, we tested for introgression between *S. carolitertii* (P3) and two sister populations (P1 and P2) from *S. pyrenaicus*, one from the northern and another from the southern part of its distribution. In B, we tested if *S. pyrenaicus* populations from the south (P3) are more closely related to *S. carolitertii* (P1) or populations from the northern part of *S. pyrenaicus* distribution (P2). Considering the possibility of a geographical cline in admixture proportions between *S. carolitertii* and *S. pyrenaicus* in the northern part of *S. pyrenaicus* distribution, we also tested if the northernmost sampling site of *S. pyrenaicus* (P3, Ocreza – see Figure 1) showed more signs of introgression with *S. carolitertii* (P1) than the other northern *S. pyrenaicus* (P2, scenario C). The opposite (all northern *S. pyrenaicus* as sister populations P1 and P2 and *S. carolitertii* as the potential source of introgressed genes P3) corresponds to scenario D. In all cases, the outgroup (Pout) was either *S. torgalensis* or *S. aradensis*. All possible combinations of the populations shown in the figure were tested. We used *S. pyrenaicus* Almargem as the southern *S. pyrenaicus*

population as *S. pyrenaicus* Guadiana was represented by only one individual after removing individuals with more than 50% missing data (see results). Significance of D-statistic values was assessed using a jackknife approach, dividing the dataset into 25 blocks and converting z-scores into p-values assuming a standard normal distribution ($p<0.01$). These computations were done in RStudio version 1.1.383 and R version 3.4.4 using custom scripts, available at Dryad.

If introgression between populations occurred in the relatively recent past, we would expect individuals within the same population to show different degrees of introgression. To test this hypothesis, we calculated the D-statistic for each individual of P2 for the same scenarios as above.

Demographic modelling of the divergence of *S. carolitertii* and *S. pyrenaicus*

We compared alternative divergence scenarios of the northern *S. pyrenaicus* from *S. carolitertii* and the southern *S. pyrenaicus* to test and quantify past introgression events. We used the composite likelihood method based on the joint site frequency spectrum (SFS) implemented in *fastsimcoal2* version 2.6 (Excoffier *et al.*, 2013). First, we compared the fit of three models to the observed SFS: “Admixture”, “No Admixture C-PN” and “No Admixture PN-PS”. The Admixture model assumes that the northern *S. pyrenaicus* received a contribution alpha (α) from the southern *S. pyrenaicus* and 1-alpha (1-α) from *S. carolitertii* at the time of the split. Note that the estimates of alpha not only indicate the most likely species

tree but also quantify the level of introgression. If alpha=0 then the northern *S. pyrenaicus* is more closely related to *S. carolitertii*, whereas if alpha=1 then the northern and southern *S. pyrenaicus* are closer to each other. Values of alpha in between 0 and 1 indicate that the northern *S. pyrenaicus* received a contribution from both species, and hence indicate introgression. We compared the likelihood of this admixture model to two models without admixture. In the “No Admixture C-PN” model, *S. carolitertii* (C) and the northern *S. pyrenaicus* (PN) share a more recent common ancestor. On the other hand, in the “No Admixture PN-PS”, the northern (PN) and southern *S. pyrenaicus* (PS) have a more recent common ancestor. To be able to compare the likelihood values directly, models need to have the same number of parameters. Thus, to ensure the same number of parameters, in the models without admixture we allowed for a bottleneck associated with the split of the northern *S. pyrenaicus* from *S. carolitertii* and the southern *S. pyrenaicus*, respectively, mimicking a founder effect. All parameters were scaled in relation to a reference effective size, which was arbitrarily set to be the effective size (*Ne*) of *S. carolitertii*. Considering the results of this first three models, we then compared the fit of three more complex models to distinguish between a hybrid origin of the northern *S. pyrenaicus* and a secondary contact: “Hybrid Origin”, “C-PN + Sec Contact PS-PN” and “PN-PS + Sec Contact C-PN”. The “Hybrid Origin” model is identical to the previous “Admixture” model. However, to ensure the same number of parameters as the two other models, we allowed for a bottleneck after the split and hybridization, mimicking a founder event. The “C-PN + Sec

Contact PS-PN" model assumes that *S. carolitertii* (C) and the northern *S. pyrenaicus* (PN) share a more recent common ancestor followed by a secondary contact between southern and the northern *S. pyrenaicus* (PS-PN). Finally, the "PN-PS + Sec Contact C-PN" model assumes that the northern (PN) and the southern (PS) *S. pyrenaicus* share a more recent common ancestor followed by a secondary contact between *S. carolitertii* and the northern *S. pyrenaicus* (C-PN).

To obtain an observed SFS without missing data, we built the joint 3D-SFS by sampling 3 individuals from *S. carolitertii* and the southern *S. pyrenaicus*, and 4 individuals from the northern *S. pyrenaicus*. Given the lack of an outgroup, we could not identify the ancestral state of alleles and hence used the minor allele frequency spectrum. To sample individuals without missing data, we used the initial dataset but without the MAF filter, and each scaffold was divided into blocks of 200bp (which is larger than the average length of the GBS loci). For each block we sampled the individuals from each population with less missing data keeping only the sites with data across all individuals. Given that the SFS is affected by the depth of coverage, only genotypes with a depth of coverage >10x were used (Nielsen *et al.*, 2011). This resulted in an observed SFS with 7,047 SNPs. For each model we performed 100 independent runs with 100 cycles, approximating the SFS with 100,000 coalescent simulations. To convert the relative divergence times estimated into absolute time in million years (Mya), we assumed a generation time of 3 years for these species (Magalhães *et al.*, 2003; Almada and Sousa-Santos, 2010).

Results

Obtention of a high-quality SNP dataset

After the initial processing removing low quality reads and trimming all reads to 82 base pairs, we obtained a mean of 5,891,239 high-quality reads per individual. After mapping all the reads from each individual to the catalog, the median depth of coverage per sample was 47x. Filtering based on MAF ≥ 0.01 and depth of coverage between ¼ to 4x of the individual median resulted in 19 individuals with more than 50% of missing data, which were removed. The final dataset had a total of 25,353 SNPs, with 40.32% missing data, and was comprised of 46 individuals, as follows: *S. carolitertii* (n=10), *S. pyrenaicus* Ocreza (n=6), *S. pyrenaicus* Lizandro (n=4), *S. pyrenaicus* Canha (n=6), *S. pyrenaicus* Almargem (n=5), *S. pyrenaicus* Guadiana (n=1), *S. torgalensis* (n=9), *S. aradensis* (n=5).

Characterization of the global patterns of genetic differentiation

The pairwise F_{ST} estimates of genetic differentiation between sampling locations are shown in Table 1. Overall, the higher levels of genetic differentiation are between the two southwestern species (*S. torgalensis* and *S. aradensis*) and the two more widely distributed species (*S. carolitertii* and *S. pyrenaicus*) (F_{ST}>0.352). On the other hand, we find the lower levels of genetic differentiation within northern *S. pyrenaicus* and

between them and *S. carolitertii* (F_{ST}<0.165). Indeed, we find lower levels of genetic differentiation between the northern *S. pyrenaicus* and *S. carolitertii* (F_{ST}<0.165) than between the northern and the southern *S. pyrenaicus* (F_{ST}>0.201). Interestingly, the levels of differentiation found between both *S. carolitertii* and the northern *S. pyrenaicus* and the southern *S. pyrenaicus* are comparable to those found between *S. torgalensis* and *S. aradensis*.

The PCA results show that the first three principal components explain approximately 26% of the variation (Figure S2A), although the Tracy-Widom tests (Patterson *et al.*, 2006) indicate that the first five components have a significant effect (p<0.01) (Figure S2B). We only show the first three PCs because these have a clear biological interpretation. The first principal component (Figure 2A and 2B) explains the higher percentage of the variance (≈16%) and clearly separates two groups: one formed by *S. carolitertii* and *S. pyrenaicus* and another formed by *S. aradensis* and *S. torgalensis*. This is consistent with the higher pairwise F_{ST} values obtained between these two groups. The second principal component (PC2) explains a much lower percentage of the variance (≈6%) and separates *S. aradensis* from *S. torgalensis* (Figure 2A and 2C). Finally, PC3 affects *S. carolitertii* and *S. pyrenaicus* and separates the southern *S. pyrenaicus* from a cluster formed by *S. carolitertii* and the northern *S. pyrenaicus* (Figure 2B and 2C). It is not possible to distinguish between individuals from *S. carolitertii* and the different sampling locations of northern *S. pyrenaicus*.

The estimation of ancestry proportions and the most likely number of clusters with sNMF (Frichot et al. 2014) suggests that the data are consistent with four populations (Figure 3), with K=4 having the smallest cross-entropy value (Figure S3). Interestingly, while individuals from the two southwestern species (*S. aradensis* and *S. torgalensis*) are clustered according to their species, individuals from *S. carolitertii* and the northern *S. pyrenaicus* are clustered together, whereas the southern *S. pyrenaicus* form a fourth cluster (Figure 3). Three individuals from *S. pyrenaicus* Almargem appear to share a high ancestry proportion with *S. carolitertii* and the northern *S. pyrenaicus*. However, these particular individuals have the higher percentages of missing data in that location. Moreover, virtually all individuals in the dataset exhibit some small proportion from groups other than the one they are assigned to, which can be due to statistical noise or shared ancestral polymorphism.

Inference of a population and species tree

We inferred a species tree based on the covariance of allele frequencies across all SNPs, modelling changes in allele frequencies through time due to genetic drift using TreeMix (Pickrell and Pritchard, 2012). This unrooted tree (Figure 4) shows a clear separation between two groups: one comprising *S. aradensis* and *S. torgalensis* and the other comprising *S. carolitertii* and *S. pyrenaicus*. Moreover, *S. aradensis* and *S. torgalensis* appear as sister species, in accordance with the F_{ST}, PCA and

sNMF results. Within the group of *S. carolitertii* and *S. pyrenaicus*, we found two main lineages: the southern *S. pyrenaicus* (here represented by *S. pyrenaicus* Almargem) and the one of *S. carolitertii* and the northern *S. pyrenaicus*. This is in agreement with the PCA and sNMF, where these two clusters were also detected, as well as with the F_{ST} results that indicated a lower level of differentiation between northern *S. pyrenaicus* populations and *S. carolitertii* than between northern and southern *S. pyrenaicus* populations. Attempts to produce a species tree with one or two migration events were unsuccessful as different runs of the TreeMix program did not produce consistent results.

Effect of linked SNPs

To verify if the results were influenced by the fact that some SNPs could be linked, we produced a dataset with only one SNP per block of 200 base pairs. This dataset comprised 3,901 SNPs and the overall percentage of missing data was ≈42.48%. The results of PCA, sNMF and TreeMix analysis were consistent with those from the initial dataset of 25,353 SNPs (Figures S4- S7). This indicates that our results are not influenced by the possibility that some SNPs are linked. Hence, further analyses were done using the initial dataset.

Detection of introgression between *S. carolitertii* and *S. pyrenaicus*

The results for the D-statistic (ABBA/BABA test) calculated per population are displayed in Figure 5. The exact number of SNPs that showed the ABBA or BABA pattern and p-values can be found in Table S4.

To test for introgression between *S. carolitertii* and *S. pyrenaicus*, we used the first topology, where the two *S. pyrenaicus* groups are sister species, with *S. carolitertii* as the source of potential introgression (Figure 5A). We obtained significantly positive values of D for all population combinations, independently of the outgroup used, reflecting an excess of sites where northern *S. pyrenaicus* populations (P2) share the same allele with *S. carolitertii* (P3), which can be interpreted as a sign of introgression or a more recent shared ancestry. On the other hand, when we tested the hypothesis that *S. carolitertii* and the northern *S. pyrenaicus* share a more recent ancestry, most of the combinations of sampling locations resulted in positive D-statistic values; however, these were not significantly different from zero for most comparisons (Figure 5 B). The exception was the significant positive values of D when the northern *S. pyrenaicus* population is Ocreza. The overall pattern is in agreement with those from the PCA and sNMF and with the species tree inferred with TreeMix, suggesting that *S. carolitertii* and northern *S. pyrenaicus* share a more recent common ancestor, even though the trend for positive (non-significant) D values is consistent with gene flow between northern and southern *S. pyrenaicus* and/or between *S. carolitertii* and northern *S. pyrenaicus*.

If *S. carolitertii* diverged at different times from the northern *S. pyrenaicus* populations, or if introgression occurred after divergence of the

different northern *S. pyrenaicus* populations, we would expect differences in D-statistics among the northern *S. pyrenaicus*. To investigate the possibility of such a geographical cline, we tested whether the northernmost sampling location of *S. pyrenaicus* (Ocreza) is closer to *S. carolitertii* than the other northern *S. pyrenaicus* locations, by computing D-statistics according to a topology where *S. carolitertii* and *S. pyrenaicus* Ocreza are sister populations. The estimated D-values were always significantly positive (Figure 5C), indicating that *S. pyrenaicus* Ocreza shares more alleles with the other northern *S. pyrenaicus* than with *S. carolitertii*. Contrarily, when the sister populations are both from the northern area of *S. pyrenaicus* distribution and P3 is *S. carolitertii*, D is never significantly different from zero (Figure 5D). This indicates that *S. pyrenaicus* Ocreza is not closer to *S. carolitertii*, suggesting that all northern *S. pyrenaicus* populations share similar numbers of derived alleles with *S. carolitertii*. This is consistent with the species tree inferred with TreeMix, showing that all northern *S. pyrenaicus* have a common ancestor that diverged from *S. carolitertii* after the divergence of the southern *S. pyrenaicus* (Figure 4). However, a scenario of introgression between *S. carolitertii* and the ancestor of the northern *S. pyrenaicus* (i.e. prior to the divergence of the different northern *S. pyrenaicus* populations) can also explain these results.

In the case of recent introgression events, we would expect to find differences in the D-statistic values among individuals from a given population. To detect evidence of such relatively recent introgression

between species, we computed the D-statistic by individual (Figure S8 and Table S5). Overall, we found no significant variation among different individuals from the same population, suggesting that introgression events are likely pre-dating the divergence of populations.

Demographic modelling of divergence of *S. carolitertii* and *S. pyrenaicus*

For the first three models tested (Figure 6 A-C), which were intended at investigating whether an introgression scenario was a better fit for the data than a simply bifurcating tree, the “Admixture” models achieved a higher likelihood than the models without admixture (“No admixture C-PN” and “No admixture PN-PS”) (Figure 6 and Table S6), suggesting that the northern *S. pyrenaicus* received a contribution from both *S. carolitertii* and the southern *S. pyrenaicus*. Estimates under this model indicate that, at the time of the split, the northern *S. pyrenaicus* received a contribution of 11.35% from the southern *S. pyrenaicus* and the remaining 88.65% from *S. carolitertii* (Figure 6A). This model suggests that the three populations have similar population sizes, although slightly lower for the southern *S. pyrenaicus*, large ancestral sizes for both species and a relatively recent split of the northern *S. pyrenaicus*, approximately 10% of the split time of *S. carolitertii* and southern *S. pyrenaicus* (Table S7-A).

Based on this result, we compared three models to distinguish between a scenario of hybrid origin of the northern *S. pyrenaicus* and

secondary contact (Figure 6 D-F). We obtained very similar likelihoods between models, although the model where the northern *S. pyrenaicus* originates via a simultaneous contribution of *S. carolitertii* and the northern *S. pyrenaicus* ("Hybrid Origin") achieving a slightly higher likelihood (Figure 6 and Table S6). Under this model, we estimated that, at the time of its origin, the northern *S. pyrenaicus* received a contribution of 90.78% from *S. carolitertii* and 9.22% from the southern *S. pyrenaicus* (Figure 6D). Despite the fact that this model has a slightly higher likelihood, we note that the difference in likelihood between these three models is small, and hence, with current data, we have no power to clearly distinguish between a hybrid origin and a secondary contact. The three models indicate overall similar relative times (Table S7B), although the second-best model ("PN-PS + Sec Contact PN-C") infers an older divergence time of the northern *S. pyrenaicus* than the other models.

Discussion

In this work, our goal was to investigate the evolutionary relationship between populations of *S. carolitertii*, *S. pyrenaicus*, *S. torgalensis* and *S. aradensis* using genome-wide data (SNPs) obtained through Genotyping by Sequencing, as well as test for the possibility of past introgression between *S. carolitertii* and *S. pyrenaicus* in the northern part of *S. pyrenaicus* distribution. We successfully obtained a high-quality set of SNP markers for these four species from GBS data without a reference genome.

Inferring a species tree from population genomic data

Taken together, our results indicate a species tree composed of two main lineages: (i) *S. torgalensis* and *S. aradensis* and (ii) *S. carolitertii* and *S. pyrenaicus*. This is evidenced by the pairwise F_{ST} (Table 1), as well as by the PCA results (Figure 2), and the species tree inferred with TreeMix (Figure 4). This finding is in agreement with phylogenies previously obtained for cytochrome b (Brito *et al.*, 1997; Sanjur *et al.*, 2003; Mesquita *et al.*, 2007; Perea *et al.*, 2010, 2016; Sousa-Santos *et al.*, 2019) and nuclear genes (Almada and Sousa-Santos, 2010; Waap *et al.*, 2011; Sousa-Santos *et al.*, 2019). Recent estimates based on one mitochondrial, seven nuclear genes and fossil calibration place the divergence between the two main lineages at approximately 14 Million years ago (Mya) (Sousa-Santos *et al.*, 2019). At that time, the configuration of the river systems in the Iberian Peninsula was very different from today, characterized by many

endorheic basins (river basins that did not flow to the ocean, but instead to lakes). Various Iberian rivers (e.g. Tagus, Guadiana) were composed of several endorheic lakes, instead of having a continuous flow to the ocean. It has been proposed that the isolation of the ancestral of the species in one of those endorheic basins (the Lower Tagus paleobasin, approximately in the current location of the Tagus river mouth – see Sousa-Santos *et al.*, 2019 for a map of the Iberian endorheic basins) led to the differentiation of the *S. aradensis* and *S. torgalensis* lineage, whose ancestor posteriorly migrated south, to the current distributions of the two species (Sousa-Santos *et al.*, 2007, 2019). The isolation and differentiation of the ancestral into *S. torgalensis* and *S. aradensis* in distinct river basins might have then been facilitated by the uplift of a mountain range in the area (the Caldeirão mountains in the south of Portugal) (Mesquita *et al.*, 2005), with the most recent estimates of their divergence pointing to 4 Mya (Sousa-Santos *et al.*, 2019).

For the second lineage, comprising *S. carolitertii* and *S. pyrenaicus*, we find overall relatively lower genetic differentiation between the northern *S. pyrenaicus* and *S. carolitertii* than between northern and southern *S. pyrenaicus* (Table 1 and Figures 2 and 3) and the species tree inferred with TreeMix (Figure 4) shows a more recent common ancestor between *S. carolitertii* and the northern *S. pyrenaicus*. These results could, in principle, be explained by two different scenarios: (i) *S. carolitertii* and the northern *S. pyrenaicus* share a more recent common ancestor but evolved

independently in the absence of gene flow; (ii) the northern *S. pyrenaicus* appear closer to *S. carolitertii* due to extensive introgression between them.

Introgression between *S. carolitertii* and *S. pyrenaicus*

Previous studies suggested the possibility of introgression to explain incongruent topologies obtained with nuclear and mitochondrial markers (Waap *et al.*, 2011; Sousa-Santos *et al.*, 2019) and described *S. pyrenaicus* as paraphyletic in relation to *S. carolitertii* (Sousa-Santos et al. 2019). Our results indicate that introgression very likely occurred between *S. carolitertii* and *S. pyrenaicus*, which reconciles discrepancies in previous results of mitochondrial and nuclear markers.

Our estimates from the demographic modelling based on the joint three population site frequency spectrum showed that a scenario of introgression is more likely than one without any gene flow (Figure 6A-C), indicating that the divergence of *S. pyrenaicus* and *S. carolitertii* involved events of introgression, and thus the species tree cannot be simply explained by a bifurcating tree. These are simple models but, nonetheless, indicate that northern *S. pyrenaicus* seems to be a mixture of *S. carolitertii* and the southern *S. pyrenaicus* lineage, with a higher proportion from *S. carolitertii.* This could explain why *S. pyrenaicus* from the Tagus, Guadiana and adjacent small basins cluster together in previously inferred mtDNA phylogenies (Brito *et al.*, 1997; Mesquita *et al.*, 2007) but seem to group in different clusters on nuclear (Waap *et al.*, 2011; Sousa-Santos *et al.*, 2019)

and genome-wide data (this study). The fact that we infer a relatively small admixture contribution from the southern *S. pyrenaicus* is probably the reason why this introgression was not detected with the D-statistics for all the northern *S. pyrenaicus* populations used (Figure 5B). However, D-values tend to be positive and are in fact significant when the northern *S. pyrenaicus* population is Ocreza, suggesting some shared alleles between northern and southern *S. pyrenaicus*, which would not be expected in the case of a simple bifurcating tree where *S. carolitertii* and the northern *S. pyrenaicus* share a more recent common ancestor. Moreover, the consistency of the results obtained for the D-statistic independently of the northern *S. pyrenaicus* used indicates that the introgression had to be older than the isolation of different populations in tributaries of the Tagus basin (Ocreza and Canha, on opposite margins of the main river). In fact, the introgression had to be older than the isolation of *S. pyrenaicus* in Lizandro, which is a small independent basin and is not connected to the Tagus, although it might have been colonized from there, at a time when connections were still present, as it has been hypothesised for other small basin nearby (Colares) (Sousa-Santos *et al.*, 2007).

The “Hybrid Origin” model (Figure 6D) assumes that the time of the admixture with the southern *S. pyrenaicus* is the same as with *S. carolitertii*, which corresponds to a scenario of hybrid speciation. Indeed, this result raises the possibility that *S. pyrenaicus* from Tagus drainage is the result of hybridization between the southern *S. pyrenaicus* lineage and *S. carolitertii*, which could have happened during the changes of endorheic paleo-

drainage systems. Hybrid speciation has been invoked in several instances to explain incongruences between nuclear and mtDNA markers and has been proposed in various organisms (e.g. Hermansen *et al.*, 2011; Leducq *et al.*, 2016; Lamichhaney *et al.*, 2018), including freshwater fish species (e.g. *Gila,* DeMarais *et al.*, 1992; *Cottus,* Nolte *et al.*, 2005).

However, our estimates suggest that a secondary contact scenario could not be discarded. Despite the fact the “Hybrid Origin” model achieves a higher likelihood, we note that there is a small difference between the likelihood of the models of hybrid speciation and secondary contact (Figure 6 D-F). Therefore, we are not able to distinguish between the two scenarios with certainty. A scenario of secondary contact would have been possible if, for example, after *S. pyrenaicus* was already disseminated throughout the Tagus and southern drainages (e.g. Guadiana), connections were re-established between the Tagus and the northern basins, allowing gene flow between *S. carolitertii* and *S. pyrenaicus.* The geological record seems to indicate that river connections were maintained until later in time between the Douro paleobasin and the paleobasin in the current location of the Tagus spring (Upper Tagus paleobasin) than between that and the Guadiana (Sousa-Santos *et al.*, 2019), which could have allowed the secondary contact between the northern *S. pyrenaicus* and *S. carolitertii*. This would be in agreement with the “PN-PS + Sec Contact C-PN” model, which achieves the second highest likelihood amongst the three more complex models (Table S6).

Whether the genomic patterns observed in the northern *S. pyrenaicus* are the result of hybrid speciation or a secondary contact, a genome with contributions from different species provides genetic diversity upon which natural selection can act (Martin and Jiggins, 2017). While introgression can generate negative fitness effects resulting of incompatibilities between loci from the different species, it also possible that introgression brings beneficial alleles which might help the population to better cope with the environment (Martin and Jiggins, 2017). Although signs of adaptation to the different climatic types of the Iberian peninsula (Atlantic in the north and Mediterranean in the south) have been investigated in Iberian *Squalius* (Jesus *et al.*, 2017), the northern *S. pyrenaicus* populations have not been examined so far. Therefore, it remains to be investigated if having a genome with contributions from different species provides an advantage to the northern *S. pyrenaicus*, as has been found in other species (e.g., in swordtail fish, hybrids between two species living at different altitudes seem to be better adapted to intermediate altitudes (Culumber *et al.*, 2012)). Nonetheless, the system provides an exciting opportunity for future work to answer fundamental questions on the consequences of hybridization at the genome level.

Conclusions

In face of the incongruent results between mitochondrial and nuclear markers, previous studies have suggested that populations from the Tagus river basin could correspond to a new taxon (Waap *et al.*, 2011; Sousa-Santos *et al.*, 2019). Overall, our results indicate that the patterns observed in the Tagus are most likely the result of introgression. Although the “Hybrid Origin” model has the best likelihood, we note that the fit to the SFS is not perfect (Figure S9). This suggests that the mode of speciation can be even more complex, e.g. involving further changes in the past effective sizes. Whole-genome data would be required to obtain more SNPs to distinguish between a hybrid origin for the northern *S. pyrenaicus* and secondary contact.

Our work shows evidence for past gene flow between currently allopatric freshwater fish species, estimating that the northern populations of *S. pyrenaicus* received approximately 90% from *S. carolitertii*. Furthermore, our results illustrate that even in freshwater species currently found in isolated river drainages, divergence can be more complex than a simply allopatric model, involving periods of past gene flow. This illustrates how the use of population genomics data based on SNP polymorphisms could detect past gene flow events and resolve species evolutionary history. Moreover, this work adds to the growing list of examples where hybridization has been reported and opens the door to future studies to elucidate how such “hybrid”/introgressed genomes cope with incompatibilities, but also can have a higher potential to adapt to new environments due to their increased genetic diversity.

Acknowledgements

We would like to thank Tiago Jesus and Miguel Machado for the preparation of the samples. This work was funded by the strategic project UID/BIA/00329/2013 (2015-2018) and UID/BIA/00329/2019 granted to cE3c from the Portuguese National Science Foundation, Fundação para a Ciência e a Tecnologia. VCS is funded by EU H2020 programme (Marie Sklodowska-Curie grant 799729).

Conflict of interest

The authors declare that they have no conflict of interest.

References

Almada V, Sousa-Santos C (2010). Comparisons of the genetic structure of Squalius populations (Teleostei, Cyprinidae) from rivers with contrasting histories, drainage areas and climatic conditions based on two molecular markers. *Mol Phylogenet Evol* **57**: 924–931.

Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nat Rev Genet* **17**: 81–92.

Bagley RK, Sousa VC, Niemiller ML, Linnen CR (2017). History, geography and host use shape genomewide patterns of genetic variation in the redheaded pine sawfly (Neodiprion lecontei). *Mol Ecol* **26**: 1022–1044.

Barluenga M, Stölting KN, Salzburger W, Muschick M, Meyer A (2006). Sympatric speciation in Nicaraguan crater lake cichlid fish. *Nature* **439**: 719–723.

Brito RM, Briolay J, Galtier N, Bouvet Y, Coelho MM (1997). Phylogenetic Relationships within Genus Leuciscus (Pisces , Cyprinidae) in Portuguese Fresh Waters , Based on Mitochondrial DNA Cytochrome b Sequences. *Mol Phylogenet Evol* **8**: 435–442.

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013). Stacks: An analysis tool set for population genomics. *Mol Ecol* **22**: 3124–3140.

Coelho MM, Bogutskaya NG, Rodrigues JA, Collares-Pereira MJ (1998). Leuciscus torgalensis, and L. aradensis, two new cyprinids for Portuguese fresh waters. *J Fish Biol* **52**: 937–950.

Coelho MM, Brito RM, Pacheco TR, Figueiredo D, Pires AM (1995). Genetic variation and divergence of Leuciscus pyrenaicus and L. carolitertii (Pisces, Cyprinidae). *J Fish Biol* **47**: 243–258.

Culumber ZW, Shepard DB, Coleman SW, Rosenthal GG, Tobler M (2012). Physiological adaptation along environmental gradients and replicated hybrid zone structure in swordtails (Teleostei: Xiphophorus). *J Evol Biol* **25**: 1800–1814.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, *et al.* (2011). The variant call format and VCFtools. *Bioinformatics* **27**: 2156–2158.

Dasmahapatra KK, Walters JR, Briscoe AD, Davey JW, Whibley A, Nadeau NJ, *et al.* (2012). Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* **487**: 94–98.

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat Rev Genet* **12**: 499–510.

DeMarais BD, Dowling TE, Marsh PC, Douglas ME, Minckley WL (1992). Origin of Gila seminuda (Teleostei□: Cyprinidae) through introgressive hybridization□: Implications for evolution and

conservation. *Evolution (N Y)* **89**: 2747–2751.

Durand EY, Patterson N, Reich D, Slatkin M (2011). Testing for ancient admixture between closely related populations. *Mol Biol Evol* **28**: 2239–2252.

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, *et al.* (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* **6**: 1–10.

Ewels P, Magnusson M, Lundin S, Käller M (2016). MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**: 3047–3048.

Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M (2013). Robust Demographic Inference from Genomic and SNP Data. *PLoS Genet* **9**.

Figueiró H V., Li G, Trindade FJ, Assis J, Pais F, Fernandes G, *et al.* (2017). Genome-wide signatures of complex introgression and adaptive evolution in the big cats. *Sci Adv* **3**: e1700299.

Frichot E, François O (2015). LEA: An R package for landscape and ecological association studies. *Methods Ecol Evol* **6**: 925–929.

Frichot E, Mathieu F, Trouillon T, Bouchard G, François O (2014). Fast and efficient estimation of individual ancestry coefficients. *Genetics* **196**: 973–983.

Fu L, Niu B, Zhu Z, Wu S, Li W (2012). CD-HIT: Accelerated for clustering

the next-generation sequencing data. *Bioinformatics* **28**: 3150–3152.

Gagnaire PA, Pavey SA, Normandeau E, Bernatchez L (2013). The genetic architecture of reproductive isolation during speciation-with-gene-flow in lake whitefish species pairs assessed by rad sequencing. *Evolution (N Y)* **67**: 2483–2497.

Gante HF, Matschiner M, Malmstrøm M, Jakobsen KS, Jentoft S, Salzburger W (2016). Genomics of speciation and introgression in Princess cichlid fishes from Lake Tanganyika. *Mol Ecol* **25**: 6143–6161.

Garrison E, Marth G (2012). Haplotype-based variant detection from short-read sequencing. *http://arxiv.org/abs/12073907*.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, *et al.* (2010). A Draft Sequence of the Neandertal Genome. *Science (80-)* **328**: 710–722.

Henriques R, Sousa V, Coelho MM (2010). Migration patterns counteract seasonal isolation of Squalius torgalensis, a critically endangered freshwater fish inhabiting a typical Circum-Mediterranean small drainage. *Conserv Genet* **11**: 1859–1870.

Hermansen JS, Sæther SA, Elgvin TO, Borge T, Hjelle E, Sætre GP (2011). Hybrid speciation in sparrows I: Phenotypic intermediacy, genetic admixture and barriers to gene flow. *Mol Ecol* **20**: 3812–3822.

Hey J, Machado CA (2003). The study of structured populations - New

hope for a difficult and divided science. *Nat Rev Genet* **4**: 535–543.

Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA (2010). Population Genomics of Parallel Adaptation in Threespine Stickleback using Sequenced RAD Tags (DJ Begun, Ed.). *PLoS Genet* **6**: e1000862.

Hohenlohe PA, Day MD, Amish SJ, Miller MR, Kamps-Hughes N, Boyer MC, *et al.* (2013). Genomic patterns of introgression in rainbow and westslope cutthroat trout illuminated by overlapping paired-end RAD sequencing. *Mol Ecol* **22**: 3002–3013.

Hudson RR, Slatkint M, Maddison WP (1992). Estimation of Levels of Gene Flow From DNA Sequence Data. *Genetics* **589**: 583–589.

Jesus TF, Moreno JM, Repolho T, Athanasiadis A, Rosa R, Almeida-Val VMF, *et al.* (2017). Protein analysis and gene expression indicate differential vulnerability of Iberian fish species under a climate change scenario (S Rutherford, Ed.). *PLoS One* **12**: e0181325.

Jones JC, Fan S, Franchini P, Schartl M, Meyer A (2013). The evolutionary history of Xiphophorus fish and their sexually selected sword: A genome-wide approach using restriction site-associated DNA sequencing. *Mol Ecol* **22**: 2986–3001.

Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, *et al.* (2012). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* **484**: 55–61.

Lamichhaney S, Berglund J, Almén MS, Maqbool K, Grabherr M, Martinez-Barrio A, *et al.* (2015). Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature* **518**: 371–375.

Lamichhaney S, Han F, Webster MT, Andersson L, Grant BR, Grant PR (2018). Rapid hybrid speciation in Darwin's finches. *Science (80-)* **359**: 224–228.

Leducq JB, Nielly-Thibault L, Charron G, Eberlein C, Verta JP, Samani P, *et al.* (2016). Speciation driven by hybridization and chromosomal plasticity in a wild yeast. *Nat Microbiol* **1**: 1–10.

Li H (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *http://arxiv.org/abs/13033997*.

Li H, Durbin R (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754–1760.

Li W, Godzik A (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**: 1658–1659.

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

Magalhães MF, Schlosser IJ, Collares-Pereira MJ (2003). The role of life history in the relationship between population dynamics and environmental variability in two Mediterranean stream fishes. *J Fish*

Biol **63**: 300–317.

de Manuel M, Kuhlwilm M, Frandsen P, Sousa VC, Desai T, Prado-Martinez J, *et al.* (2016). Chimpanzee genomic diversity reveals ancient admixture with bonobos. *Science (80-)* **354**: 477–481.

Martin SH, Jiggins CD (2017). Interpreting the genomic landscape of introgression. *Curr Opin Genet Dev* **47**: 69–74.

McManus KF, Kelley JL, Song S, Veeramah KR, Woerner AE, Stevison LS, *et al.* (2015). Inference of gorilla demographic and selective history from whole-genome sequence data. *Mol Biol Evol* **32**: 600–612.

Meier JI, Sousa VC, Marques DA, Selz OM, Wagner CE, Excoffier L, *et al.* (2017). Demographic modelling with whole-genome data reveals parallel origin of similar Pundamilia cichlid species after hybridization. *Mol Ecol* **26**: 123–141.

Mesquita N, Cunha C, Carvalho GR, Coelho MM (2007). Comparative phylogeography of endemic cyprinids in the south-west Iberian Peninsula: Evidence for a new ichthyogeographic area. *J Fish Biol* **71**: 45–75.

Mesquita N, Hänfling B, Carvalho GR, Coelho MM (2005). Phylogeography of the cyprinid Squalius aradensis and implications for conservation of the endemic freshwater fauna of southern Portugal. *Mol Ecol* **14**: 1939–1954.

Nielsen R, Paul JS, Albrechtsen A, Song YS (2011). Genotype and SNP

calling from next-generation sequencing data. *Nat Rev Genet* **12**: 443–451.

Nolte AW, Freyhof J, Stemshorn KC, Tautz D (2005). An invasive lineage of sculpins, Cottus sp. (Pisces, Teleostei) in the Rhine with new habitat adaptations has originated from hybridization between old phylogeographic groups. *Proc R Soc B Biol Sci* **272**: 2379–2387.

Paris JR, Stevens JR, Catchen JM (2017). Lost in parameter space: a road map for stacks. *Methods Ecol Evol* **8**: 1360–1373.

Patterson N, Price AL, Reich D (2006). Population structure and eigenanalysis. *PLoS Genet* **2**: 2074–2093.

Perea S, Böhme M, Zupancic P, Freyhof J, Sanda R, Ozuluğ M, *et al.* (2010). Phylogenetic relationships and biogeographical patterns in Circum-Mediterranean subfamily Leuciscinae (Teleostei, Cyprinidae) inferred from both mitochondrial and nuclear data. *BMC Evol Biol* **10**: 265.

Perea S, Cobo-Simon M, Doadrio I (2016). Cenozoic tectonic and climatic events in southern Iberian Peninsula: Implications for the evolutionary history of freshwater fish of the genus Squalius (Actinopterygii, Cyprinidae). *Mol Phylogenet Evol* **97**: 155–169.

Pfenninger M, Patel S, Arias-Rodriguez L, Feldmeyer B, Riesch R, Plath M (2015). Unique evolutionary trajectories in repeated adaptation to hydrogen sulphide-toxic habitats of a neotropical fish (Poecilia

mexicana). *Mol Ecol* **24**: 5446–5459.

Pickrell JK, Pritchard JK (2012). Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. *PLoS Genet* **8**.

Redenbach Z, Taylor EB (2002). Evidence for historical introgression along a contact zone between two species of char (Pisces: Salmonidae) in northwestern North America. *Evolution (N Y)* **56**: 1021–1035.

Rochette NC, Rivera-Colón AG, Catchen JM (2019). Stacks 2: Analytical Methods for Paired-end Sequencing Improve RADseq-based Population Genomics. *bioRxiv.* www.biorxiv.org/content/10.1101/615385v1.

Sanjur OI, Carmona JA, Doadrio I (2003). Evolutionary and biogeographical patterns within Iberian populations of the genus Squalius inferred from molecular data. *Mol Phylogenet Evol* **29**: 20–30.

Seehausen O, Butlin RK, Keller I, Wagner CE, Boughman JW, Hohenlohe PA, *et al.* (2014). Genomics and the origin of species. *Nat Rev Genet* **15**: 176–192.

Seehausen O, Wagner CE (2014). Speciation in Freshwater Fishes. *Annu Rev Ecol Evol Syst* **45**: 621–651.

Sousa-Santos C, Collares-Pereira MJ, Almada V (2007). Reading the history of a hybrid fish complex from its molecular record. *Mol Phylogenet Evol* **45**: 981–996.

Sousa-Santos C, Jesus TF, Fernandes C, Robalo JI, Coelho MM (2019).

Fish diversification at the pace of geomorphological changes: evolutionary history of western Iberian Leuciscinae (Teleostei: Leuciscidae) inferred from multilocus sequence data. *Mol Phylogenet Evol* **133**: 263–285.

Sousa V, Hey J (2013). Understanding the origin of species with genome-scale data: modelling gene flow. *Nat Rev Genet* **14**: 404–414.

Taggart JB, Hynes RA, Prodöuhl PA, Ferguson A (1992). A simplified protocol for routine total DNA isolation from salmonid fishes. *J Fish Biol* **40**: 963–965.

Terekhanova N V., Logacheva MD, Penin AA, Neretina T V., Barmintseva AE, Bazykin GA, *et al.* (2014). Fast Evolution from Precast Bricks: Genomics of Young Freshwater Populations of Threespine Stickleback Gasterosteus aculeatus. *PLoS Genet* **10**.

vonHoldt BM, Cahill JA, Fan Z, Gronau I, Robinson J, Pollinger JP, *et al.* (2016). Whole-genome sequence analysis shows that two endemic species of North American wolf are admixtures of the coyote and gray wolf. *Sci Adv* **2**: e1501714.

Waap S, Amaral AR, Gomes B, Coelho MM (2011). Multi-locus species tree of the chub genus Squalius (Leuciscinae: Cyprinidae) from western Iberia: New insights into its evolutionary history. *Genetica* **139**: 1009–1018.

929

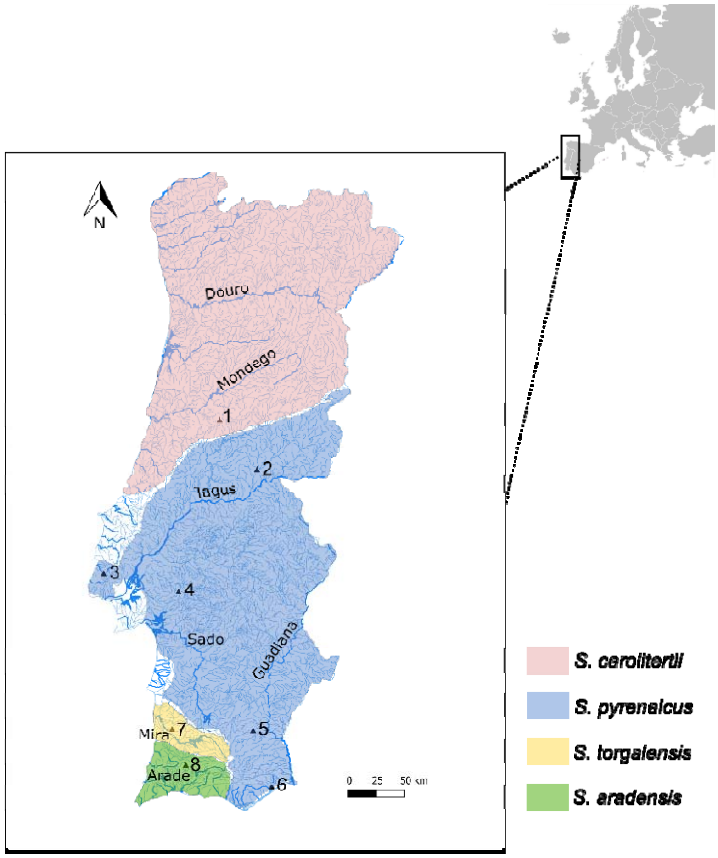

Figure 1 – Distribution range of the four *Squalius* species in Portuguese rivers and sampling locations: (1) Mondego; (2) Ocreza; (3) Lizandro; (4) Canha; (5) Guadiana; (6) Almargem; (7) Mira; (8) Arade.

936

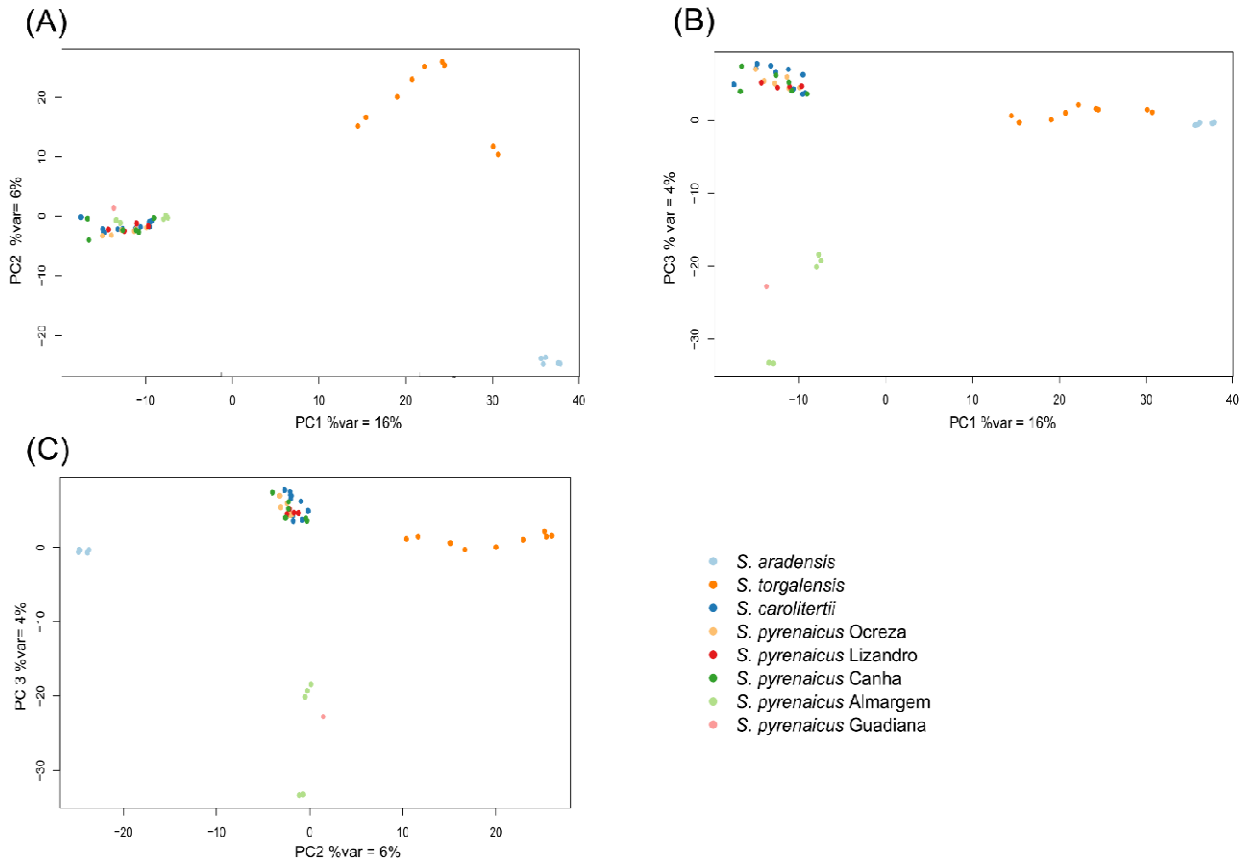

Figure 2 - Results for the first three components of the Principal Components Analysis: (A) PC1 and PC2; (B) PC1 and PC3; (C) PC2 and PC3. Each point corresponds to one individual. The PCA was calculated based on the dataset with 25,353 SNPs, filtered with MAF ≥0.01 and keeping only SNPs with a depth of coverage between ¼ and 4 times the individual median depth of coverage.

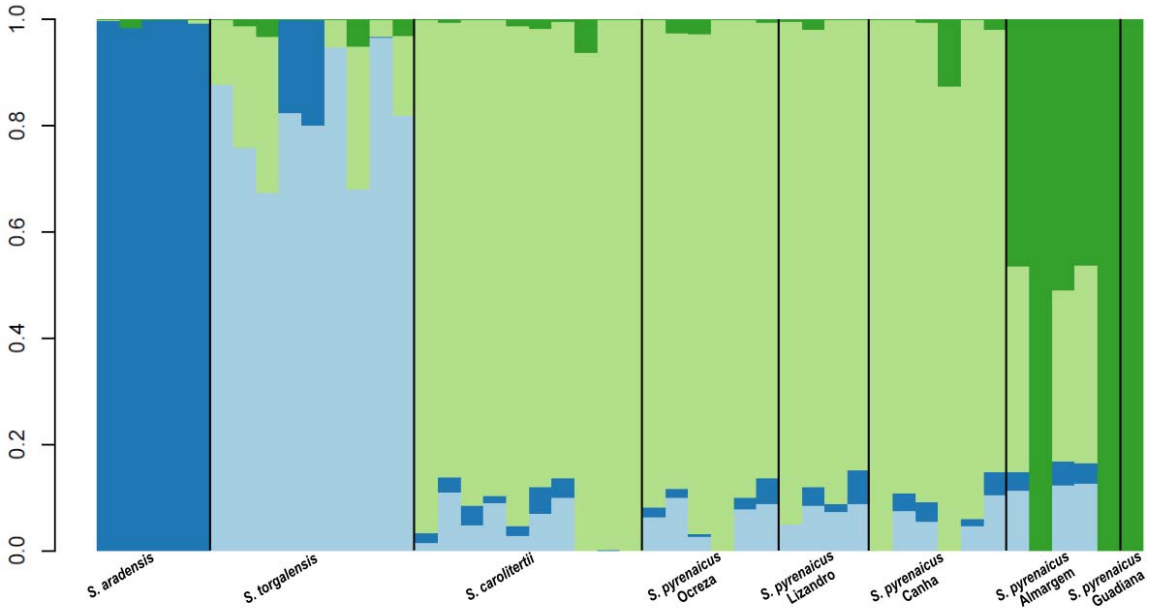

Figure 3 - Ancestry proportions inferred with sNMF for four ancestral populations (K=4). Each vertical bar corresponds to one individual and the proportion of each colour corresponds to the estimated ancestry proportion from a given cluster. The individuals are grouped per sampling locations separated by black lines. Ancestry proportions were inferred based on the dataset with 25,353 SNPs, filtered with MAF ≥0.01 and keeping only SNPs with a depth of coverage between ¼ and 4 times the individual median depth of coverage.

937

938

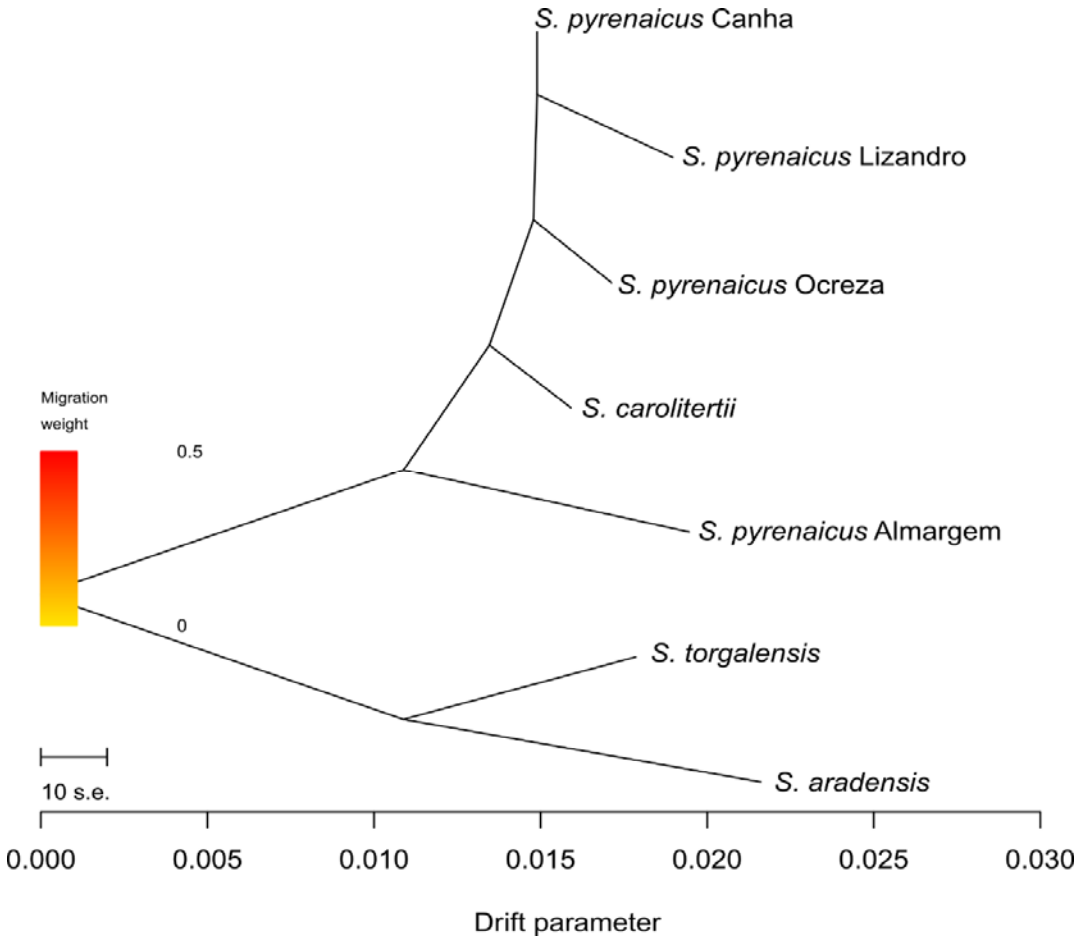

Figure 4 - Species tree graph obtained with TreeMix. This is an unrooted tree and branch lengths are represented in units of genetic drift, i.e. the longer a given branch the stronger the genetic drift experienced in that lineage, which could be due to longer divergence times and/or smaller effective sizes. The species tree was inferred based on the dataset with 25,353 SNPs, filtered with MAF ≥0.01 and keeping only SNPs with a depth of coverage between ¼ and 4 times the individual median depth of coverage.

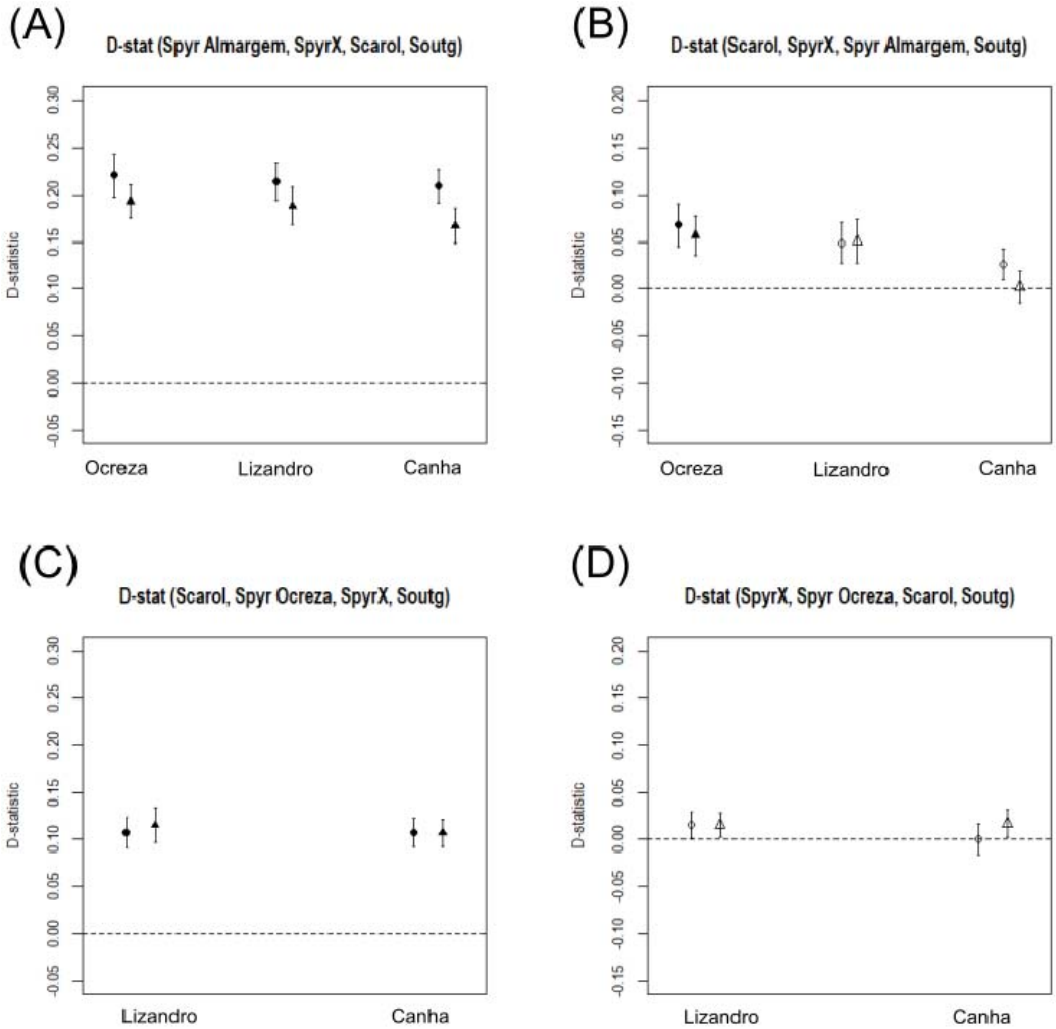

Figure 5 - Results of the D-statistic calculated for different topologies. For each topology (A to D), the results are presented according to the northern *S. pyrenaicus* sampling location (S. pyrX) used. "S.carol" stands for *S. carolitertii*, "S.pyr Almargem" stands for *S. pyrenaicus* Almargem, "*S.pyr* Ocreza" stands for *S. pyrenaicus* Ocreza and "Outg" for outgroup. Results obtained with each outgroup are represented by a different symbol (circles for *S. aradensis* and triangles for *S. torgalensis*). Full symbols represent significant D values ($p<0.01$). The D-statistic was calculated based on the dataset with 25,353 SNPs, filtered with MAF ≥0.01 and keeping only SNPs with a depth of coverage between ¼ and 4 times the individual median depth of coverage.

949

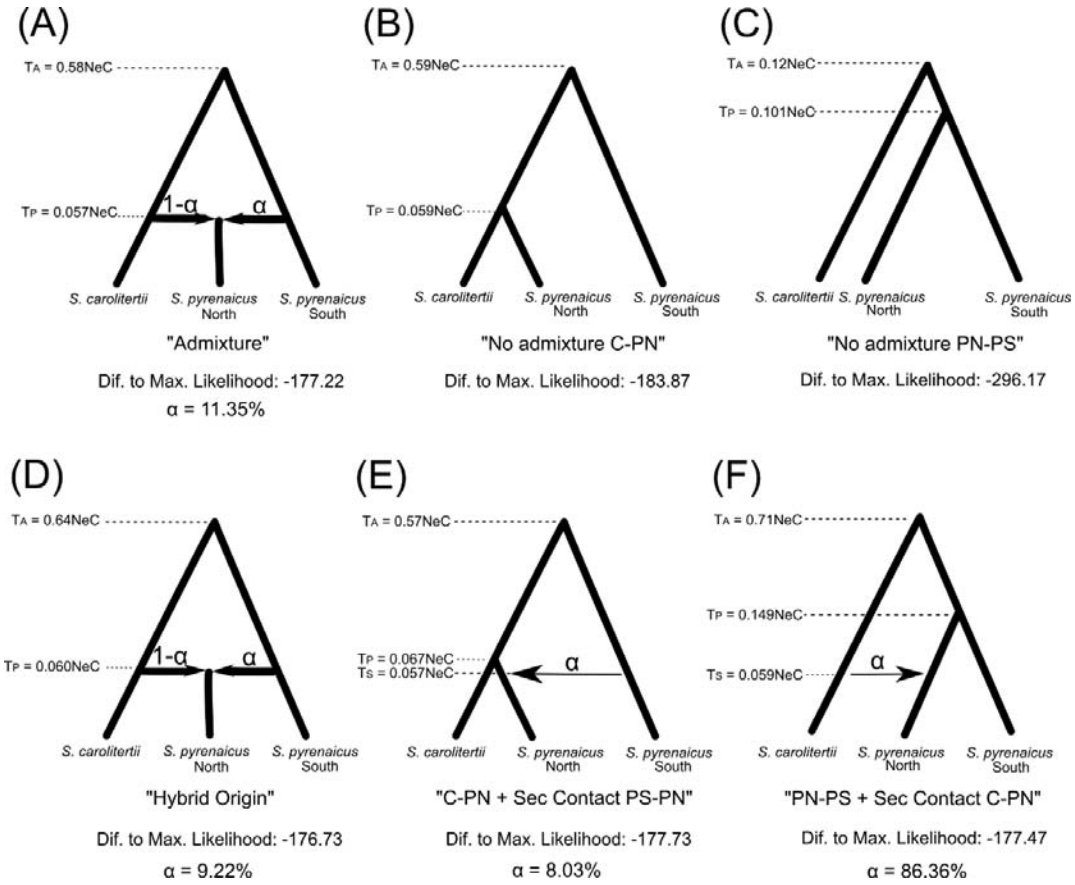

Figure 6 – Schematic representation of the likelihood of the models tested with *fastsimcoal2* and percentages of admixture inferred. The name given to each model is indicated below the schematic representation, as well the difference to maximum likelihood (Dif. To Max. Likelihood) which is the difference in log10 units between the estimated likelihood and the maximum likelihood if there was a perfect fit to the observed site frequency spectrum. The closer to zero (less negative values), the better the fit. α indicates the percentage of admixture estimated. Models (A) to (C) have 8 parameters and therefore are directly comparable. Models (D) to (F) have 9 parameters and are also directly comparable. The inferred divergence times (calculated in relation to the Ne of *S. carolitertii* (NeC)) are indicated on each model: T_A – divergence from the ancestral; T_P – divergence of the northern *S. pyrenaicus*; T_S – secondary contact.

Table 1 – Pairwise F_{ST} calculated between the different sampling locations. *S. pyrenaicus* Guadiana was deliberately left out as there is only one individual from this sampling location.

	S. carolitertii	*S. pyrenaicus* Ocreza	*S. pyrenaicus* Lizandro	*S. pyrenaicus* Canha	*S. pyrenaicus* Almargem	*S. torgalensis*	*S. aradensis*
S. carolitertii	-	0.126	0.165	0.081	0.217	0.377	0.368
S. pyrenaicus Ocreza	-	-	0.161	0.070	0.234	0.401	0.391
S. pyrenaicus Lizandro	-	-	-	0.092	0.271	0.427	0.414
S. pyrenaicus Canha	-	-	-	-	0.201	0.364	0.352
S. pyrenaicus Almargem	-	-	-	-	-	0.400	0.390
S. torgalensis	-	-	-	-	-	-	0.225
S. aradensis	-	-	-	-	-	-	-