Population genomics data supports introgression between Western Iberian *Squalius* freshwater fish species from different drainages

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

[1] cE3c – Centre for Ecology, Evolution and Environmental Changes, Departamento de Biologia Animal, Faculdade de Ciências da Universidade de Lisboa, Campo Grande, 1749-016 Lisbon, Portugal

[†] equal contribution

*corresponding authors: vmsousa@fc.ul.pt and mmcoelho@fc.ul.pt

Abstract

In freshwater fish, processes of population divergence and speciation are often linked to the geomorphology of rivers and lakes that create barriers isolating 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 other southern populations. Here, for the first time, we successfully applied a genomic approach to the study of the relationship between these species, using a Genotyping by Sequencing approach to obtain single nucleotide polymorphisms (SNPs). Our results revealed a species tree with two main lineages: (i) *S. carolitertii* and *S. pyrenaicus*; (ii) *S. torgalensis* and *S. aradensis*. Moreover, regarding *S. carolitertii* and *S. pyrenaicus*, we found evidence for past introgression between these two species 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*. Although we cannot distinguish a scenario of hybrid speciation from secondary contact, our estimates are consistent across models, suggesting that the northern populations of *S. pyrenaicus* received approximately 80% from *S. carolitertii* and 20% from southern *S. pyrenaicus*. This illustrates that even in freshwater species currently found in isolated river drainages, we are able to detect past gene flow events in present-day genomes, suggesting that speciation is more complex than simply allopatric.

Key-words: Iberian freshwater fish; *Squalius*; 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 systematically reduction in gene flow through time until reproductive isolation is achieved and populations maintain phenotypic and genetic distinctiveness (Seehausen et al. 2014). The most acceptable 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; 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 freshwater fish speciation is the geomorphology of the rivers and lakes, since the formation of geological barriers isolates 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; Hohenlohe et al. 2013; Jones et al. 2013; Gante et al. 2016). Nonetheless, 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 is 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 an 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 and Guadiana basins (Coelho et al. 1995; Coelho et al. 1998). On the other hand, the two other species are confined to much smaller 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 small drainages (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), the trees produced using nuclear genes (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), 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

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

132

133 **Methods**

134

135 **Sampling and sequencing**

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

150 All fish were collected by electrofishing (300V, 4A), and total genomic DNA was
151 extracted from fin clips using an adapted phenol-chloroform protocol (Sambrook et al.
152 1989). DNA was quantified using Qubit® 2.0 Fluorometer (Live Technologies). The
153 samples were subjected to a paired-end Genotyping by Sequencing (GBS) protocol
154 (adapted from Elshire et al. 2011), performed in outsourcing at Beijing Genomics
155 Institute (BGI, www.bgi.com). The DNA samples were sent to the facility mixed with
156 DNAstable Plus (Biomatrica) to preserve DNA at room temperature during shipment.
157 Briefly, upon arrival, DNA was fragmented using the restriction enzyme ApeKI and the

158 fragments were amplified after adaptor ligation (Elshire et al. 2011). The resulting
159 library was sequenced using Illumina Hiseq2000.

160

161 **Obtention of a high-quality SNP dataset**

162 First, the quality of the sequences of each individual was assessed using
163 FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). To compile the
164 information from all individual reports, we used MultiQC (Ewels et al. 2016) to merge
165 and summarize the individual FastQC reports. Second, we used the program
166 *process_radtags* from Stacks version 2.2 (Catchen et al. 2013) to trim all reads to 82
167 base pairs and discard reads with low quality scores, using the default settings for the
168 window size (0.15x the length of the read) and the base quality threshold (10 in phred
169 score). Given the absence of a reference genome for any of the species in study, we
170 built a reference catalog of all loci using a *denovo* assembly approach on Stacks
171 version 2.2 (Catchen et al. 2013). To determine the best parameters for the
172 construction of the catalog, we followed the approach recommended by Paris et al.
173 2017 (Figures S1 and S2). We decided to allow a maximum of 2 differences between
174 sequences within the same individual (M=2) and a maximum of 4 differences between
175 sequences from different individuals (n=4) for them to be considered the same locus on
176 the catalog. We also required a minimum depth of coverage of 4x for every locus on
177 the catalog (m=4). After building the catalog, given the possibility that forward and
178 reverse sequences of the same fragment were treated as different loci, similar reads
179 within the catalog were clustered using CD-HIT version 4.7 (Li and Godzik 2006; Fu et
180 al. 2012). We used CD-HIT-EST from the CD-HIT package with a word length of 6 and
181 a sequence identity threshold of 0.85.

182 Once we clustered similar reads within the catalog, this was treated as a
183 reference and the reads from each individual were aligned against it using BWA-MEM

184 from BWA version 0.7.17-r1188 (Li 2013) with default parameters. The output
185 alignments of BWA were sorted and unmapped reads were removed using Samtools
186 version 1.8 (Li and Durbin 2009). To call genotypes for each individual at each site and
187 identify SNPs we used the method implemented on Freebayes v1.2.0 (Garrison and
188 Marth 2012). We applied further filters to keep only SNPs present in all sampling sites
189 in at least 50% of the individuals using VCFtools version 0.1.15 (Danecek et al. 2011).

190 To discard sites and genotypes that are more likely to be the result of
191 sequencing or mapping errors, we applied filters on the minor allele frequency (MAF ≥
192 0.01) and on the depth of coverage, keeping only genotypes with a depth of coverage
193 (DP) between ¼ to 4 times the individual median DP, after assessing the effect of
194 different filtering options (Tables S2 and S3). The different filters were applied using a
195 combination of options from VCFtools version 0.1.15 (Danecek et al. 2011) and
196 BCFtools version 1.6 (Li et al. 2009). Finally, individuals with more than 50% missing
197 data were removed from the dataset.

198

199 **Characterization of the global patterns of genetic differentiation**

200 To quantify the levels of differentiation between sampling locations, we
201 calculated the pairwise F_{ST} using the Hudson estimator (Hudson et al. 1992). Given
202 that the sampling locations may not correspond to populations, we investigated fine
203 population structure with individual-based methods. To understand how individuals
204 cluster, we conducted a principal component analysis (PCA). The number of significant
205 principal components was determined with the Tracy-Widom test (Patterson et al.
206 2006) on all eigenvalues. Furthermore, individual ancestry proportions were estimated
207 with the sparse Non-negative Matrix Factorization method (sNMF) (Frichot et al. 2014).
208 We tested values of K between 1 and 8, performing 100 repetitions for each K value.

209 All calculations were performed in RStudio version 1.1.383 and R version 3.4.4 and the
210 PCA and sNMF were performed using the package LEA (Frichot and François 2015).

211

212 **Inference of a population and species tree**

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

219

220 **Effect of linked SNPs**

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

229

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

231 To test for possible past introgression between *S. carolitertii* and *S. pyrenaicus*
232 in the northern area of *S. pyrenaicus* distribution, we used the D-statistic (Durand et al.
233 2011), which was used to distinguish between ancestral polymorphism and

234 introgression by looking at four different populations related through a fixed species
235 tree: two sister populations (P1 and P2), a third population that could be the source of
236 introgressed genes (P3) and has a common ancestor to P1 and P2, and one outgroup
237 (Pout). We explored four different possible species trees to perform different tests. In
238 scenario A, we tested for introgression between *S. carolitertii* (P3) and two sister
239 populations (P1 and P2) from *S. pyrenaicus*, one from the northern and another from
240 the southern part of its distribution. In B, we tested if *S. pyrenaicus* populations from
241 the south (P3) are more closely related to *S. carolitertii* (P1) or populations from the
242 northern part of *S. pyrenaicus* distribution (P2). Considering the possibility of a
243 geographical cline in admixture proportions between *S. carolitertii* and *S. pyrenaicus* in
244 the northern part of *S. pyrenaicus* distribution, we also tested if the northern most
245 sampling site of *S. pyrenaicus* (Ocreza – see Figure 1) showed more signs of
246 introgression with *S. carolitertii* than the other northern *S. pyrenaicus*, which
247 corresponds to scenario C. The opposite (all northern *S. pyrenaicus* as sister
248 populations and *S. carolitertii* as the potential source of introgressed genes)
249 corresponds to scenario D. In all cases, the outgroup (Pout) was either *S. torgalensis*
250 or *S. aradensis*. All possible combinations of the populations shown in the figure were
251 tested. We used *S. pyrenaicus* Almargem as the southern *S. pyrenaicus* population as
252 *S. pyrenaicus* Guadiana was represented by only one individual after removing
253 individuals with more than 50% missing data (see results). Significance of D-statistic
254 values was assessed using a jackknife approach, dividing the dataset into 25 blocks
255 and converting z-scores into p-values assuming a standard normal distribution
256 ($p<0.01$). These computations were done in RStudio version 1.1.383 and R version
257 3.4.4 using custom scripts, available at Dryad.

258 If introgression between populations occurred in the relatively recent past, we
259 would expect individuals within the same population to show different degrees of

260 introgression. To test this hypothesis, we calculated the D-statistic for each individual of
261 P2 for the same scenarios as above.
262
263 **Demographic modelling of the divergence of *S. carolitertii* and *S. pyrenaicus***
264 We compared alternative divergence scenarios of the northern *S. pyrenaicus*
265 from *S. carolitertii* and the southern *S. pyrenaicus* to test and quantify past
266 introgression events. We used the composite likelihood method based on the joint site
267 frequency spectrum (SFS) implemented in *fastsimcoal2* (Excoffier et al. 2013). First,
268 we compared the fit of three models to the observed SFS: "Admixture", "No Admixture
269 C-PN" and "No Admixture PN-PS". The Admixture model assumes that the northern *S.*
270 *pyrenaicus* received a contribution alpha (α) from the southern *S. pyrenaicus* and 1-
271 alpha (1-α) from *S. carolitertii* at the time of the split. Note that the estimates of alpha
272 not only indicate the most likely species tree but also quantify the level of introgression.
273 If alpha=0 then the northern *S. pyrenaicus* is more closely related to *S. carolitertii*,
274 whereas if alpha=1 then the northern and southern *S. pyrenaicus* are closer to each
275 other. Values of alpha in between 0 and 1 indicate that the northern *S. pyrenaicus*
276 received a contribution from both species, and hence indicate introgression. We
277 compared the likelihood of this admixture model to two models without admixture, i.e.
278 with alpha=0 or alpha=1. In the "No Admixture C-PN" model, *S. carolitertii* and the
279 northern *S. pyrenaicus* share a more recent common ancestor (i.e. alpha=0). On the
280 other hand, in the "No Admixture PN-PS", the northern and southern *S. pyrenaicus*
281 have a more recent common ancestor (i.e. alpha=1). To be able to compare the
282 likelihood values directly, models need to have the same number of parameters. Thus,
283 to ensure the same number of parameters, in the models without admixture we allowed
284 for a bottleneck associated with the split of the northern *S. pyrenaicus* from *S.*
285 *carolitertii* and the southern *S. pyrenaicus*, respectively, mimicking a founder effect. All
286 parameters were scaled in relation to a reference effective size, which was arbitrarily

287 set to be the effective size (*Ne*) of *S. carolitertii*. Considering the results of this first
288 three models (see results), we then compared the fit of three more complex models to
289 distinguish between a hybrid origin of the northern *S. pyrenaicus* and a secondary
290 contact: “Hybrid Origin”, “C-PN + Sec Contact PN-PS” and “PN-PS + Sec Contact PN-
291 C”. The “Hybrid Origin” model is identical to the previous “Admixture” model. However,
292 to ensure the same number of parameters as the two other models, we allowed for a
293 bottleneck after the split and hybridization, mimicking a founder event. The “C-PN +
294 Sec Contact PN-PS” model assumes that *S. carolitertii* (C) and the northern *S.*
295 *pyrenaicus* (PN) share a more recent common ancestor followed by a secondary
296 contact between the northern and the southern *S. pyrenaicus* (PN-PS). Finally, the
297 “PN-PS + Sec Contact PN-C” model assumes that the northern (PN) and the southern
298 (PS) *S. pyrenaicus* share a more recent common ancestor followed by a secondary
299 contact between the northern *S. pyrenaicus* and *S. carolitertii* (PN-C).

300 To obtain an observed SFS without missing data, we built the joint 3D-SFS by
301 sampling 2 individuals from *S. carolitertii* and the southern *S. pyrenaicus*, and 3
302 individuals from the northern *S. pyrenaicus*. Given the lack of an outgroup, we could
303 not identify the ancestral state of alleles, and hence used the minor allele frequency
304 spectrum. To sample individuals without missing data, we used the initial dataset but
305 without the MAF filter, and each scaffold was divided into blocks of 200bp (which is
306 larger than the average length of the GBS loci), and for each block we sampled the
307 individuals from each population with less missing data keeping only the sites with data
308 across all individuals. Given that the SFS is affected by the depth of coverage, only
309 genotypes with a depth of coverage >10x were used (Nielsen et al. 2011). This
310 resulted in an observed SFS with 6,753 SNPs. For each model we performed 50
311 independent runs with 50 cycles, approximating the SFS with 100,000 coalescent
312 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).

316 **Results**

317

318 **Obtention of a high-quality SNP dataset**

319 After the initial processing removing low quality reads and trimming all reads to
320 82 base pairs, we obtained a mean of 5,891,239 high quality reads per individual. After
321 mapping all the reads from each individual to the catalog, the median depth of
322 coverage per sample was 47x. Filtering based on MAF ≥ 0.01 and depth of coverage
323 between ¼ to 4x of the individual median resulted in 19 individuals with more than 50%
324 of missing data, which were removed. The final dataset had a total of 25,353 SNPs,
325 with 40.32% missing data, and was comprised of 46 individuals, as follows: *S.*
326 *carolitertii* (n=10), *S. pyrenaicus* Ocreza (n=6), *S. pyrenaicus* Lizandro (n=4), *S.*
327 *pyrenaicus* Canha (n=6), *S. pyrenaicus* Almargem (n=5), *S. pyrenaicus* Guadiana
328 (n=1), *S. torgalensis* (n=9), *S. aradensis* (n=5).

329

330 **Characterization of the global patterns of genetic differentiation**

331 The pairwise F_{ST} estimates of genetic differentiation between sampling
332 locations are shown in Table 1. Overall, the higher levels of genetic differentiation are
333 between the two southwestern species (*S. torgalensis* and *S. aradensis*) and the two
334 more widely distributed species (*S. carolitertii* and *S. pyrenaicus*) (F_{ST}>0.352). On the
335 other hand, we find the lower levels of genetic differentiation within northern *S.*
336 *pyrenaicus* and between them and *S. carolitertii* (F_{ST}<0.165). Indeed, we find lower
337 levels of genetic differentiation between the northern *S. pyrenaicus* and *S. carolitertii*
338 (F_{ST}<0.165) than between the northern and the southern *S. pyrenaicus* (F_{ST}>0.201).
339 Interestingly, the levels of differentiation found between both *S. carolitertii* and the
340 northern *S. pyrenaicus* and the southern *S. pyrenaicus* are comparable to those found
341 between *S. torgalensis* and *S. aradensis*.

342 The PCA results show that the first three principal components explain
343 approximately 26% of the variation (Figure S3), although the Tracy-Widom tests
344 (Patterson et al. 2006) indicate that the first five components have a significant effect
345 ($p<0.01$) (Figure S4). We only show the first three PCs because these have a clear
346 biological interpretation. The first principal component (Figure 2A and 2B) explains the
347 higher percentage of the variance (≈16%) and clearly separates two groups: one
348 formed by *S. carolitertii* and *S. pyrenaicus* and another formed by *S. aradensis* and *S.*
349 *torgalensis*. This is consistent with the higher pairwise F_{ST} values obtained between
350 these two groups. The second principal component (PC2) explains a much lower
351 percentage of the variance (≈6%) and separates *S. aradensis* from *S. torgalensis*
352 (Figure 2A and 2C). Finally, PC3 affects *S. carolitertii* and *S. pyrenaicus* and separates
353 the southern *S. pyrenaicus* from a cluster formed by *S. carolitertii* and the northern *S.*
354 *pyrenaicus* (Figure 2B and 2C). It is not possible to distinguish between individuals
355 from *S. carolitertii* and the different sampling locations of northern *S. pyrenaicus*.

356 The estimation of ancestry proportions and the mostly likely number of clusters
357 with sNMF (Frichot et al. 2014) suggests that the data are consistent with four
358 populations (Figure 3), with K=4 having the smallest cross-entropy value (≈0.364)
359 (Figure S5). Interestingly, while individuals from the two southwestern species (*S.*
360 *aradensis* and *S. torgalensis*) are clustered according to their species, individuals from
361 *S. carolitertii* and the northern *S. pyrenaicus* are clustered together, leaving the
362 southern *S. pyrenaicus* in a fourth cluster (Figure 3). Three individuals from *S.*
363 *pyrenaicus* Almargem appear to share a high ancestry proportion with *S. carolitertii* and
364 the northern *S. pyrenaicus*. However, these particular individuals have the higher
365 percentages of missing data in that location. Moreover, virtually all individuals in the
366 dataset exhibit some small proportion from groups other than the one they are
367 assigned to, which can be due to statistical noise or shared ancestral polymorphism.

368

369 **Inference of a population and species tree**

370 We inferred a species tree based on the covariance of allele frequencies across
371 all SNPs, modelling changes in allele frequencies through time due to genetic drift
372 using TreeMix (Pickrell and Pritchard 2012). This unrooted tree (Figure 4) shows a
373 clear separation between two groups: one comprising *S. aradensis* and *S. torgalensis*
374 and the other comprising *S. carolitertii* and *S. pyrenaicus*. *S. aradensis* and *S.*
375 *torgalensis* appear as sister species, in accordance with the F_{ST}, PCA and sNMF
376 results. Within the group of *S. carolitertii* and *S. pyrenaicus*, we found two main
377 lineages: the southern *S. pyrenaicus* (here represented by *S. pyrenaicus* Almargem)
378 and the one of *S. carolitertii* and the northern *S. pyrenaicus*. This is in agreement with
379 the PCA and sNMF, where these two clusters were also detected, as well as with the
380 F_{ST} results that indicated a lower level of differentiation between northern *S. pyrenaicus*
381 populations and *S. carolitertii* than between northern and southern *S. pyrenaicus*
382 populations. Attempts to produce a species tree with one or two migration events were
383 unsuccessful as different runs of the TreeMix program did not produce consistent
384 results.

385

386 **Effect of linked SNPs**

387 To verify if the results were influenced by the fact that some SNPs could be
388 linked, we produced a dataset with only one SNP per block of 200 base pairs. This
389 dataset comprised 3,901 SNPs and the overall percentage of missing data was
390 ≈42.48%. The results of PCA, sNMF and TreeMix analysis were consistent with those
391 from the initial dataset of 25,353 SNPs (Figures S6- S11). This indicates that our
392 results are not influenced by the possibility that some SNPs are linked. Hence, further
393 analyses were done using the initial dataset.

394

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

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

399 To test for introgression between *S. carolitertii* and *S. pyrenaicus*, we used the
400 first topology, where the two *S. pyrenaicus* groups are sister species, with *S. carolitertii*
401 as the source of potential introgression (Figure 5A). We obtained significantly positive
402 values of D for all population combinations, independently of the outgroup used,
403 reflecting an excess of sites where the northern *S. pyrenaicus* populations (P2) shares
404 the same allele with *S. carolitertii* (P3), which can be interpreted as a sign of
405 introgression or a more recent shared ancestry. On the other hand, when we tested the
406 hypothesis that *S. carolitertii* and the northern *S. pyrenaicus* share a more recent
407 ancestry, most of combinations of sampling locations resulted in positive D-statistic
408 values, however these were not significantly different from zero for most values (Figure
409 5 B). The exception were the significant positive values of D when the northern *S.*
410 *pyrenaicus* population is Ocreza. The overall pattern is in agreement with those from
411 the PCA and sNMF and with the species tree inferred, suggesting that *S. carolitertii*
412 and northern *S. pyrenaicus* share a more recent common ancestor, even though the
413 trend for positive (non-significant) D values is consistent with some gene flow between
414 northern and southern *S. pyrenaicus* and/or between *S. carolitertii* and northern *S.*
415 *pyrenaicus*.

416 If *S. carolitertii* diverged at different times from the northern *S. pyrenaicus*
417 populations, or if introgression occurred after divergence, we would expect differences
418 in D-statistics among the northern *S. pyrenaicus*. To investigate the possibility of such
419 a geographical cline, we tested whether the northern most sampling location of *S.*
420 *pyrenaicus* (Ocreza) is closer to *S. carolitertii* than the other northern *S. pyrenaicus*
421 locations, by computing D-statistics according to a topology where *S. carolitertii* and *S.*

422 *pyrenaicus* Ocreza are sister populations. The estimated D-values were always
423 significantly positive (Figure 5C), indicating that *S. pyrenaicus* Ocreza shares more
424 alleles with the other northern *S. pyrenaicus* than with *S. carolitertii*. Contrarily, when
425 the sister populations are both from the northern area of *S. pyrenaicus* distribution and
426 P3 is *S. carolitertii*, D is never significantly different from zero (Figure 5D). This
427 indicates that *S. pyrenaicus* Ocreza is not closer to *S. carolitertii*, suggesting that all
428 northern *S. pyrenaicus* populations share similar numbers of derived alleles with *S.*
429 *carolitertii*. This is consistent with the species tree inferred with TreeMix, showing that
430 all northern *S. pyrenaicus* have a common ancestor that diverged from *S. carolitertii*
431 after the divergence of the southern *S. pyrenaicus* (Figure 4). However, a scenario of
432 introgression between *S. carolitertii* and the ancestor of the northern *S. pyrenaicus* (i.e.
433 prior to the divergence of the different northern *S. pyrenaicus* populations) could also
434 lead to the same results.

435 In the case of recent introgression events, we would expect to find differences
436 in the D-statistic values among individuals from a given population. To detect evidence
437 of such relatively recent introgression between species, we computed the D-statistic by
438 individual (Figure S12 and Table S5). Overall, we found no significant variation among
439 different individuals from the same population, suggesting that introgression events are
440 likely pre-dating the divergence of populations.

441

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

443 For the first three models tested (Figure 6 A-C), which were intended at investigating
444 whether an introgression scenario was a better fit for the data than a simply bifurcating
445 tree, the “Admixture” models achieved a higher likelihood than the models without
446 admixture (“No admixture C-PN” and “No admixture PN-PS”) (Figure 6 and Table S6),
447 suggesting that the northern *S. pyrenaicus* received a contribution from both *S.*

448 *carolitertii* and the southern *S. pyrenaicus*. Estimates under this model indicate that, at
449 the time of split, the northern *S. pyrenaicus* received a contribution of 14.3% from the
450 southern *S. pyrenaicus* and the remaining 85.7% from *S. carolitertii* (Figure 6A). This
451 model suggests that the three populations have similar population sizes, although
452 slightly higher for the northern *S. pyrenaicus*, large ancestral sizes for both species and
453 a relative recent split of the northern *S. pyrenaicus* in comparison with the split of *S.*
454 *carolitertii* and southern *S. pyrenaicus*. (Table S7-A).

455 Based on this result, we compared three models to distinguish between a
456 scenario of hybrid origin of the northern *S. pyrenaicus* and secondary contact (Figure 6
457 D-F). We obtained very similar likelihoods between models, with the model of a
458 common origin for *S. pyrenaicus* followed by secondary contact ("PN-PS + Sec Contact
459 PN-C") achieving a slightly higher likelihood (Figure 6 and Table S6). Under this model
460 we estimated that, at the time of the secondary contact, the northern *S. pyrenaicus*
461 received a contribution of 80.29% from *S. carolitertii* and that the effective sizes of the
462 three populations are similar (Figure 6F). Despite the fact that this model has a slightly
463 higher likelihood, we note that the difference in likelihood between these three models
464 is small, and hence with current data we have no power to distinguish this from the
465 hybrid origin model. All three models indicate similar relative times, with a recent
466 divergence of the northern *S. pyrenaicus* (Table S7B). For the best model ("PN-PS +
467 Sec Contact PN-C"), the relative time of the secondary contact is approximately half of
468 the divergence time of the northern *S. pyrenaicus*. Finally, all six models suggest that
469 the ancestral population of the three lineages had a small effective size (Table S7A and
470 B).

471

472 **Discussion**

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

479

480 **Inferring a species tree from population genomic data**

481 Taken together, our results indicate a species tree composed of two main
482 lineages: (i) *S. torgalensis* and *S. aradensis* and (ii) *S. carolitertii* and *S. pyrenaicus*.
483 This is evidenced by the pairwise F_{ST} results indicating lower levels of differentiation
484 within each lineage than between the two lineages, as well as by the PCA results
485 (Figure 2) and the species tree inferred with TreeMix (Figure 4). This is in agreement
486 with phylogenies previously obtained for cytochrome b (Brito et al. 1997; Sanjur et al.
487 2003; Mesquita et al. 2007; Perea et al. 2010; Sousa-Santos et al. 2019) and nuclear
488 genes (Almada and Sousa-Santos 2010; Waap et al. 2011; Sousa-Santos et al. 2019).
489 The divergence between the two main lineages has recently been estimated, based on
490 one mitochondrial and seven nuclear genes, to be approximately 14 Million years ago
491 (Mya) (Sousa-Santos et al. 2019). At that point, the configuration of the river systems in
492 the Iberian Peninsula was very different from today, characterized by many endorheic
493 basins (basins that did not flow to the ocean). The Tagus was composed of several
494 endorheic lakes and it has been suggested that the isolation of one of them, the Lower
495 Tagus (approximately in the current location of the Tagus and Sado river mouths) was
496 related to the isolation of the ancestor of *S. torgalensis* and *S. aradensis*. This
497 ancestral could have become isolated in this paleobasin when connections to other

498 freshwater masses ceased and then migrated south once connections were re-
499 established, reaching the current distributions of *S. torgalensis* and *S. aradensis*
500 (Sousa-Santos et al. 2007; Sousa-Santos et al. 2019). The uplift of a mountain range in
501 this area of the south of Portugal (the Caldeirão mountains) has been proposed to have
502 facilitated the isolation and divergence of the ancestral of *S. torgalensis* and *S.*
503 *aradensis* in the Mira and Arade river basins respectively (Mesquita et al. 2005), with
504 the most recent estimates of their divergence pointing to 4 Mya (Sousa-Santos et al.
505 2019).

506

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

508 For the second lineage, comprising *S. carolitertii* and *S. pyrenaicus*, we find
509 overall relatively lower genetic differentiation between the northern *S. pyrenaicus* and
510 *S. carolitertii* than between northern and southern *S. pyrenaicus* (Table 1 and Figures 2
511 and 3) and the species tree inferred with TreeMix shows a more recent common
512 ancestor between *S. carolitertii* and the northern *S. pyrenaicus*. These results could, in
513 principle, be explained by two different scenarios: (i) *S. carolitertii* and the northern *S.*
514 *pyrenaicus* share a more recent common ancestor but evolved independently in the
515 absence of gene flow; (ii) the northern *S. pyrenaicus* appear closer to *S. carolitertii* due
516 to extensive introgression between them.

517 Previous studies suggested the possibility of introgression to explain
518 incongruent topologies obtained with nuclear and mitochondrial markers (Waap et al.
519 2011; Sousa-Santos et al. 2019) and described *S. pyrenaicus* as paraphyletic in
520 relation to *S. carolitertii* (Sousa-Santos et al. 2019). Our results indicate that
521 introgression very likely occurred between *S. carolitertii* and *S. pyrenaicus*, which
522 reconciliates previous incongruencies between mitochondrial and nuclear marker
523 results.

524 Our estimates from the demographic modelling based on the joint population
525 site frequency spectrum showed that a scenario of introgression ("Admixture" model) is
526 more likely than one without any gene flow (Figure 6 A-C), indicating that the
527 divergence of *S. pyrenaicus* and *S. carolitertii* involved events of gene flow, and thus
528 the species tree cannot be simply explained by a bifurcating tree. These are simple
529 models but, nonetheless, indicate that northern *S. pyrenaicus* seems to be a mixture of
530 *S. carolitertii* and the southern *S. pyrenaicus* lineage, with a higher proportion from *S.*
531 *carolitertii* (Figure 6A-C). This could explain why *S. pyrenaicus* from the Tagus and
532 Guadiana cluster together in previously inferred mtDNA phylogenies but seem to group
533 in different clusters on nuclear and genome-wide data. The fact that we infer a
534 relatively small admixture contribution from the southern *S. pyrenaicus* (≈14%) is
535 probably the reason why this introgression was not detected with the D-statistics for all
536 the northern *S. pyrenaicus* populations used (Figure 5B). However, D-values tend to be
537 positive and are in fact significant when the northern *S. pyrenaicus* populations is
538 Ocreza, suggesting some shared alleles between northern and southern *S. pyrenaicus*,
539 which would not be expected in the case of a simple bifurcating tree where *S.*
540 *carolitertii* and the northern *S. pyrenaicus* share a more recent common ancestor.
541 Moreover, the consistency of the results obtained for the D-statistic independently of
542 the northern *S. pyrenaicus* used indicate that the introgression had to be older than the
543 isolation of different populations in tributaries of the Tagus basin (Ocreza and Canha,
544 on opposite margins of the main river). In fact, the introgression had to be older than
545 the isolation of *S. pyrenaicus* in Lizandro, which is not connected to the Tagus basin,
546 although it might have been colonized from there, at a time when connections were still
547 present, as it has been hypothesised for other small basin nearby (Colares) (Sousa-
548 Santos et al. 2007).

549 The "Admixture" model assumes that the time of the admixture with the
550 southern *S. pyrenaicus* is the same as with *S. carolitertii*, which corresponds to a

551 scenario of hybrid speciation. Indeed, this result raises the possibility that *S. pyrenaicus*
552 from Tagus drainage is a different species resulting from hybridization between the
553 southern Guadiana drainage lineages and *S. carolitertii* lineages, which could have
554 happened during the changes of endorheic paleo-drainage systems. In fact, hybrid
555 speciation has been invoked to explain incongruences between nuclear and mtDNA
556 markers and has been proposed in several instances in freshwater fish (DeMarais et al.
557 1992; Nolte et al. 2005; Meier, Marques, et al. 2017).

558 However, our estimates suggest that a secondary contact scenario could not be
559 discarded. Interestingly, despite the very high contribution from *S. carolitertii* to
560 northern *S. pyrenaicus,* the best model supports that both *S. pyrenaicus* populations
561 share a common ancestor followed by secondary contact with significant introgression
562 of approximately 80% from *S. carolitertii* ("PN-PS + Sec Contact PN-C" model – Figure
563 6F). However, we note that there is a small difference between the likelihood of the
564 models of hybrid speciation and secondary contact (Figure 6 D-F). Therefore, we are
565 not able to distinguish between the two scenarios with certainty. The history of the
566 hydrological basins seems to suggest that connections between the Lower Tagus
567 paleobasin and the Guadiana paleobasin ceased before those between the Upper
568 Tagus and the Douro paleobasins (the last two located in present day Spain, near
569 present day Tagus and Douro river springs, respectively) (Sousa-Santos et al. 2019).
570 Thus, a secondary contact between the northern *S. pyrenaicus* and *S. carolitertii* would
571 have been possible due to the maintenance of that connection between the Upper
572 Tagus and Douro paleobasins for a longer period. A secondary contact between the
573 northern and southern *S. pyrenaicus* would also be possible through re-establishment
574 of connections between the Tagus and Guadiana basins. The possibility that the Tagus
575 and Guadiana basins were connected more recently has been proposed to explain the
576 presence of a common lineage in these two basins for another Iberian endemic
577 cyprinid (*Iberochondrostoma lemmingii*) (Lopes-Cunha et al. 2012). Another possibility

578 is that the introgression of southern *S. pyrenaicus* lineages into the northern *S.*
579 *pyrenaicus* was not caused by the re-establishment of connections between the Tagus
580 and the Guadiana, but between the Tagus and the Sado (see Figure 1). This would
581 have been possible if the Lower Tagus and the paleobasin that originated the Sado
582 (Alvalade paleobasin) were connected at a time where *S. pyrenaicus* was already
583 present in the Alvalade paleobasin (Sousa-Santos et al. 2019).

584

585 **Final remarks**

586 In face of the incongruent results between mitochondrial and nuclear markers,
587 previous studies have suggested that populations from the Tagus river basin could
588 correspond to a new taxa (Waap et al. 2011; Sousa-Santos et al. 2019). Overall, our
589 results indicate that the patterns observed in the Tagus are most likely the result of
590 introgression, even though we are not able to reject the hypothesis that the northern *S.*
591 *pyrenaicus* is a new taxon resulting from hybrid speciation. Indeed, estimates suggest
592 that a secondary contact is as good to explain our data. We note that the models we
593 considered are still a major simplification and that our models do not fit exactly the
594 observed SFS (Figure S13). This suggests that the mode of speciation can be even
595 more complex, e.g. involving further changes in the past effective sizes. Future studies
596 should focus on whole-genome data, which would be required to obtain more SNPs to
597 distinguish between a hybrid origin for the northern *S. pyrenaicus* and secondary
598 contact. Furthermore, such studies should include sampling of two key locations that
599 are missing from our dataset: the Zêzere river (a tributary of the Tagus) and the Sado
600 basin. The Zêzere river has consistently been a source of incongruences in mtDNA
601 phylogenies, with authors suggesting both *S. pyrenaicus* and *S. carolitertii* can be
602 found in this river (Brito et al. 1997; Almada and Sousa-Santos 2010; Sousa-Santos et
603 al. 2016). On the other hand, *S. pyrenaicus* from the Sado, although clustering with the
604 Guadiana individuals in both mitochondrial and nuclear markers on phylogenetic

605 analysis (Brito et al. 1997; Waap et al. 2011), have been described as very
606 differentiated from other southern *S. pyrenaicus* (Sousa-Santos et al. 2007; Sousa-
607 Santos et al. 2019) and could also be important to understand the origin of the northern
608 *S. pyrenaicus*.

609 Our work shows evidence for past gene flow between currently allopatric
610 freshwater fish species, estimating that the northern populations of *S. pyrenaicus*
611 received approximately 80% from *S. carolitertii*. Furthermore, our results illustrate that
612 even in freshwater species currently found in isolated river drainages, divergence can
613 be more complex than a simply allopatric model, involving periods of past gene flow.
614 This work adds to the growing list of examples where hybridization has been reported
615 and opens the door to future studies to elucidate how such “hybrid”/introgressed
616 genomes cope with incompatibilities, but also can have a higher potential to adapt to
617 new environments due to their increased genetic diversity.

618

619 **References**

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

624 Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA. 2016. Harnessing the
625 power of RADseq for ecological and evolutionary genomics. Nat. Rev. Genet.
626 17:81–92.

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

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

632 Brito RM, Briolay J, Galtier N, Bouvet Y, Coelho MM. 1997. Phylogenetic Relationships
633 within Genus Leuciscus (Pisces , Cyprinidae) in Portuguese Fresh Waters ,
634 Based on Mitochondrial DNA Cytochrome b Sequences. Mol. Phylogenet. Evol.
635 8:435–442.

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

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

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

644 Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE,
645 Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools.
646 Bioinformatics 27:2156–2158.

647 Dasmahapatra KK, Walters JR, Briscoe AD, Davey JW, Whibley A, Nadeau NJ, Zimin
648 A V., Hughes DST, Ferguson LC, Martin SH, et al. 2012. Butterfly genome reveals
649 promiscuous exchange of mimicry adaptations among species. Nature 487:94–98.

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

653 DeMarais BD, Dowling TE, Marsh PC, Douglas ME, Minckley WL. 1992. Origin of Gila
654 seminuda (Teleostei : Cyprinidae) through introgressive hybridization :
655 Implications for evolution and conservation. Evolution (N. Y). 89:2747–2751.

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

658 Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE.
659 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high
660 diversity species. PLoS One 6:1–10.

661 Ewels P, Magnusson M, Lundin S, Käller M. 2016. MultiQC: Summarize analysis
662 results for multiple tools and samples in a single report. Bioinformatics 32:3047–
663 3048.

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

666 Figueiró H V., Li G, Trindade FJ, Assis J, Pais F, Fernandes G, Santos SHD, Hughes
667 GM, Komissarov A, Antunes A, et al. 2017. Genome-wide signatures of complex
668 introgression and adaptive evolution in the big cats. Sci. Adv. 3:e1700299.

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

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

673 Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: Accelerated for clustering the next-
674 generation sequencing data. Bioinformatics 28:3150–3152.

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

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

681 Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read
682 sequencing. arXiv:1207.3907v2.

683 Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H,
684 Zhai W, Fritz MHY, et al. 2010. A Draft Sequence of the Neandertal Genome.
685 Science (80-.). 328:710–722.

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

689 Hey J, Machado CA. 2003. The study of structured populations - New hope for a
690 difficult and divided science. Nat. Rev. Genet. 4:535–543.

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

694 Hohenlohe PA, Day MD, Amish SJ, Miller MR, Kamps-Hughes N, Boyer MC, Muhlfeld
695 CC, Allendorf FW, Johnson EA, Luikart G. 2013. Genomic patterns of
696 introgression in rainbow and westslope cutthroat trout illuminated by overlapping
697 paired-end RAD sequencing. Mol. Ecol. 22:3002–3013.

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

700 Jesus TF, Moreno JM, Repolho T, Athanasiadis A, Rosa R, Almeida-Val VMF, Coelho
701 MM. 2017. Protein analysis and gene expression indicate differential vulnerability
702 of Iberian fish species under a climate change scenario.Rutherford S, editor. PLoS
703 One 12:e0181325.

704 Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, Swofford R, Pirun
705 M, Zody MC, White S, et al. 2012. The genomic basis of adaptive evolution in
706 threespine sticklebacks. Nature 484:55–61.

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

710 Lamichhaney S, Berglund J, Almén MS, Maqbool K, Grabherr M, Martinez-Barrio A,
711 Promerová M, Rubin C-J, Wang C, Zamani N, et al. 2015. Evolution of Darwin's
712 finches and their beaks revealed by genome sequencing. Nature 518:371–375.

713 Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with
714 BWA-MEM. http://arxiv.org/abs/1303.3997.

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

717 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G,
718 Durbin R. 2009. The Sequence Alignment/Map format and SAMtools.

719 Bioinformatics 25:2078–2079.

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

722 Lopes-Cunha M, Aboim MA, Mesquita N, Alves MJ, Doadrio I, Coelho MM. 2012.
723 Population genetic structure in the Iberian cyprinid fish Iberochondrostoma
724 lemmingii (Steindachner, 1866): Disentangling species fragmentation and
725 colonization processes. Biol. J. Linn. Soc. 105:559–572.

726 Magalhães MF, Schlosser IJ, Collares-Pereira MJ. 2003. The role of life history in the
727 relationship between population dynamics and environmental variability in two
728 Mediterranean stream fishes. J. Fish Biol. 63:300–317.

729 de Manuel M, Kuhlwilm M, Frandsen P, Sousa VC, Desai T, Prado-Martinez J,
730 Hernandez-Rodriguez J, Dupanloup I, Lao O, Hallast P, et al. 2016. Chimpanzee
731 genomic diversity reveals ancient admixture with bonobos. Science (80-.).
732 354:477–481.

733 McManus KF, Kelley JL, Song S, Veeramah KR, Woerner AE, Stevison LS, Ryder OA,
734 Project GAG, Kidd JM, Wall JD, et al. 2015. Inference of gorilla demographic and
735 selective history from whole-genome sequence data. Mol. Biol. Evol. 32:600–612.

736 Meier JI, Marques DA, Mwaiko S, Wagner CE, Excoffier L, Seehausen O. 2017.
737 Ancient hybridization fuels rapid cichlid fish adaptive radiations. Nat. Commun.
738 8:1–11.

739 Meier JI, Sousa VC, Marques DA, Selz OM, Wagner CE, Excoffier L, Seehausen O.
740 2017. Demographic modelling with whole-genome data reveals parallel origin of
741 similar Pundamilia cichlid species after hybridization. Mol. Ecol. 26:123–141.

742 Mesquita N, Cunha C, Carvalho GR, Coelho MM. 2007. Comparative phylogeography
743 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, Abdoli A, Doadrio I. 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.

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

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

774 Sambrook J, Fritsch EF, Maniatis T. 1989. Molecular Cloning: A Laboratory Manual.

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

778 Seehausen O, Butlin RK, Keller I, Wagner CE, Boughman JW, Hohenlohe PA, Peichel
779 CL, Saetre G-P, Bank C, Brännström Å, et al. 2014. Genomics and the origin of
780 species. Nat. Rev. Genet. 15:176–192.

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

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

785 Sousa-Santos C, Jesus TF, Fernandes C, Robalo JI, Coelho MM. 2019. Fish
786 diversification at the pace of geomorphological changes: evolutionary history of
787 western Iberian Leuciscinae (Teleostei: Leuciscidae) inferred from multilocus
788 sequence data. Mol. Phylogenet. Evol. 133:263–285.

789 Sousa-Santos C, Robalo JI, Pereira AM, Branco P, Santos JM, Ferreira MT, Sousa M,
790 Doadrio I. 2016. Broad-scale sampling of primary freshwater fish populations
791 reveals the role of intrinsic traits, inter-basin connectivity, drainage area and
792 latitude on shaping contemporary patterns of genetic diversity. PeerJ 4:e1694.

793 Sousa V, Hey J. 2013. Understanding the origin of species with genome-scale data:

794 modelling gene flow. Nat. Rev. Genet. 14:404–414.

795 Terekhanova N V., Logacheva MD, Penin AA, Neretina T V., Barmintseva AE, Bazykin
796 GA, Kondrashov AS, Mugue NS. 2014. Fast Evolution from Precast Bricks:
797 Genomics of Young Freshwater Populations of Threespine Stickleback
798 Gasterosteus aculeatus. PLoS Genet. 10.

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

802

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) granted to cE3c from the Portuguese National Science Foundation, Fundaçao para a Ciência e a Tecnologia. VS is funded by EU H2020 programme (Marie Skłodowska-Curie grant 799729).

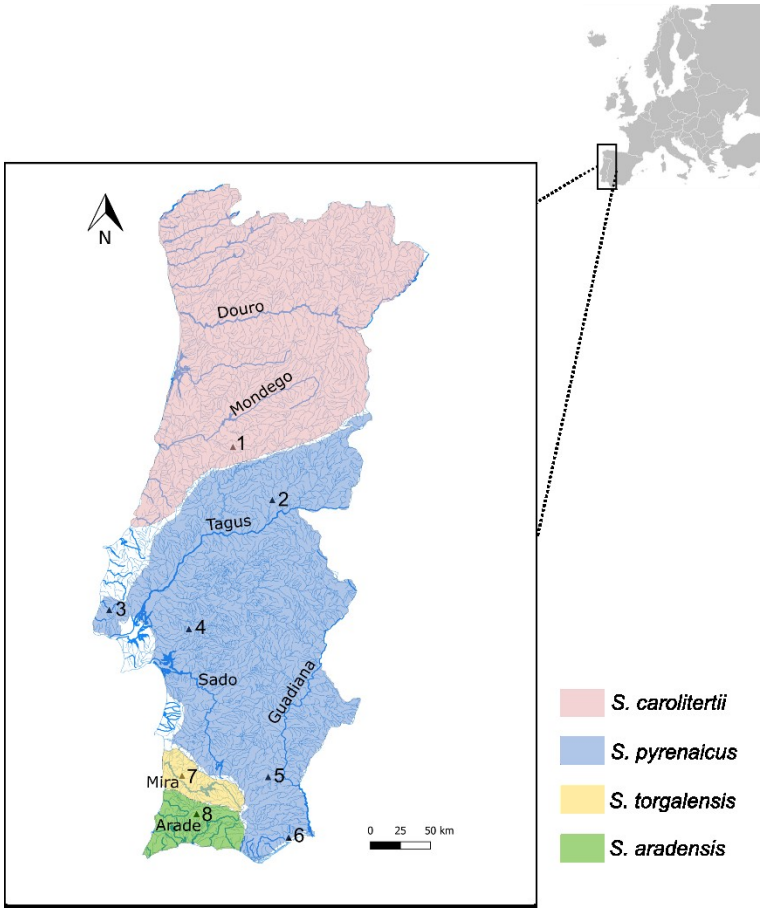

Figure 1 – Distribution range of the four *Squalius* species in Portuguese rivers and sampling locations: (1) Mondego; (2) Ocreza; (3) Lizandro; (4) Canha; (5) Guadiana; (6) Almargem; (7) Mira; (8) Arade.

816

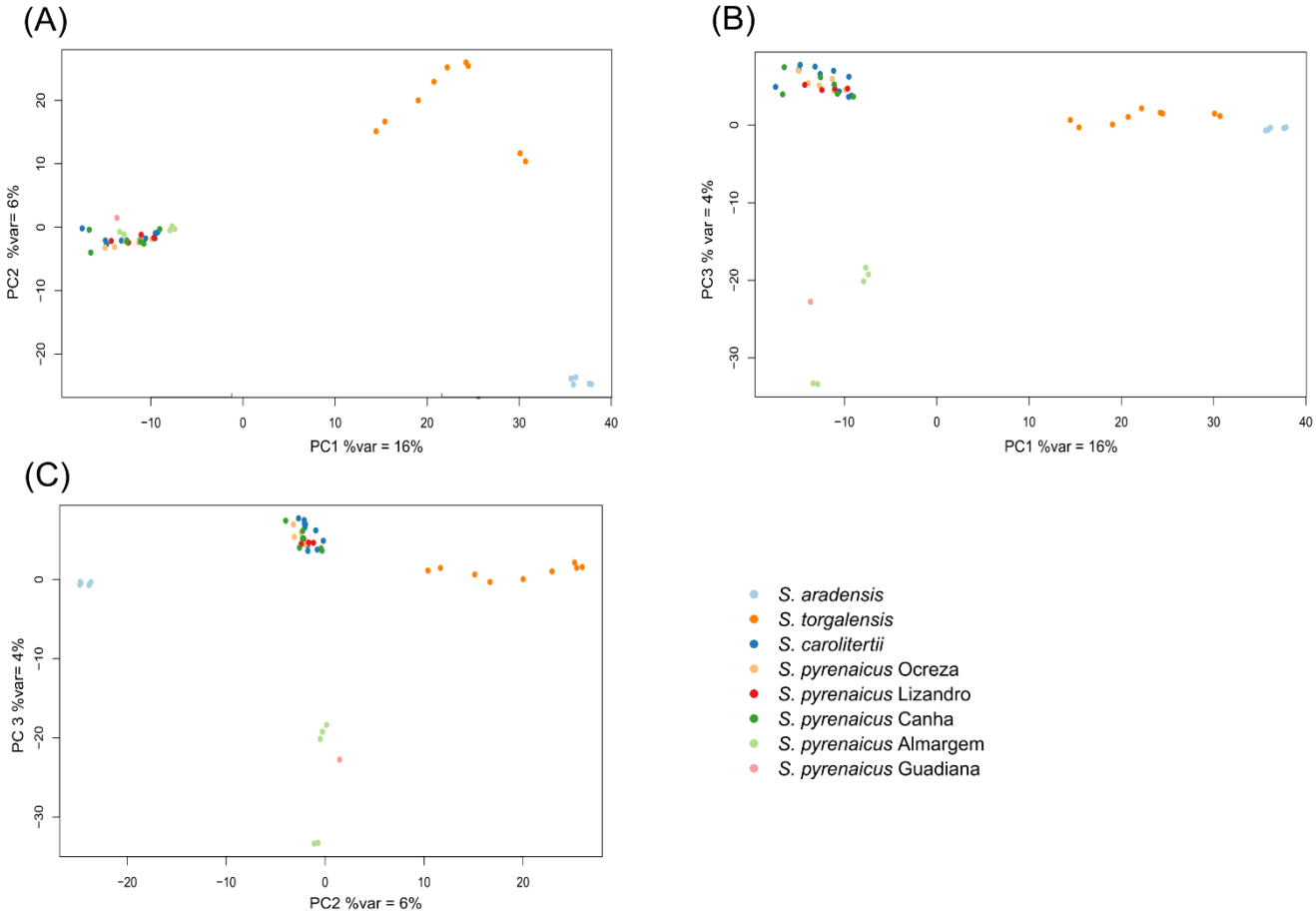

Figure 2 - Results for the first three components of the Principal Components Analysis: (A) PC1 and PC2; (B) PC1 and PC3; (C) PC2 and PC3. Each point corresponds to one individual. The PCA was calculated based on the dataset with 25,353 SNPs, filtered with MAF ≥0.01 and keeping only SNPs with a depth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

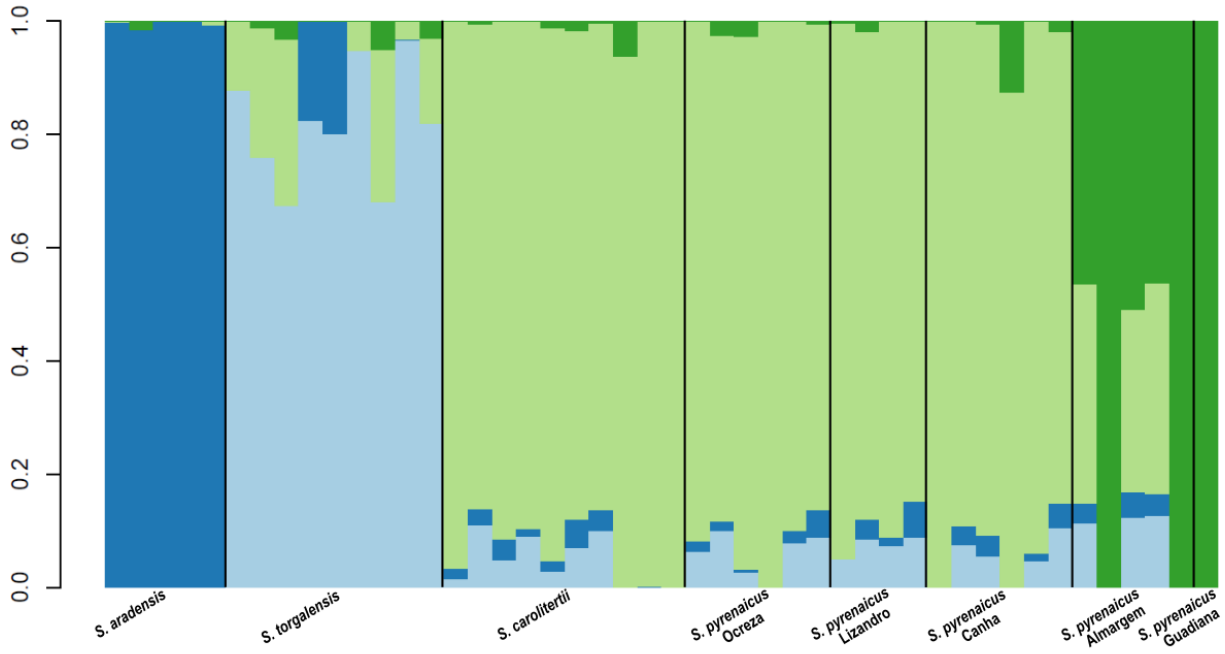

Figure 3 - Ancestry proportions inferred with sNMF for four ancestral populations (K=4). Each vertical bar corresponds to one individual and the proportion of each colour corresponds to the estimated ancestry proportion from a given cluster. The individuals are grouped per sampling locations separated 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.

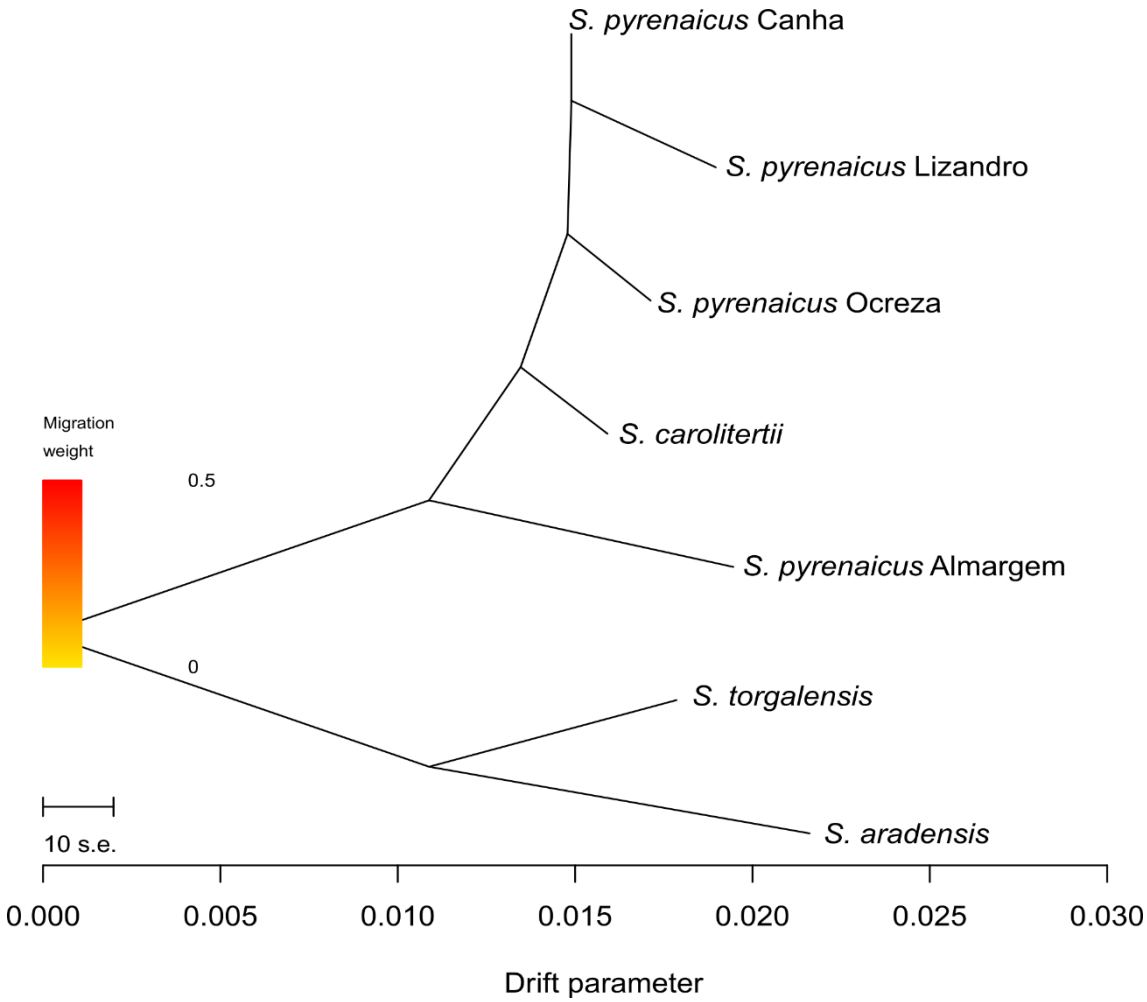

Figure 4 - Species tree graph obtained with TreeMix. This is an unrooted tree and branch lengths are represented in units of genetic drift, i.e. the longer a given branch the stronger the genetic drift experienced during that branch, which could be due to longer divergence times and/or smaller effective 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.

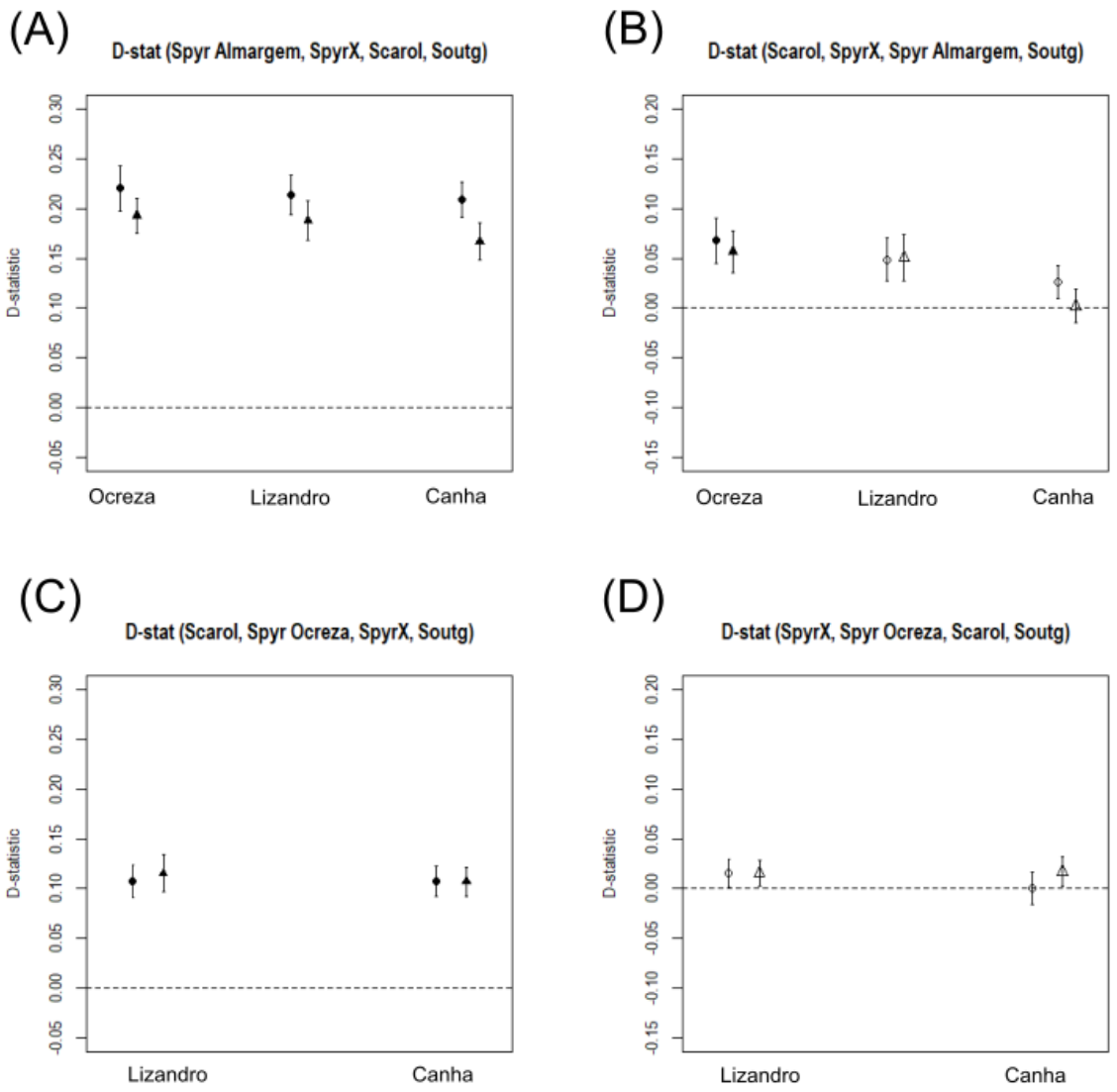

Figure 5 - Results of the D-statistic calculated for different topologies. For each topology (A to D), the results are presented according to the northern *S. pyrenaicus* sampling location (S. pyrX) used. "S.carol" stands for *S. carolitertii*, "S.pyr Almargem" stands for *S. pyrenaicus* Almargem, "*S.pyr* 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.

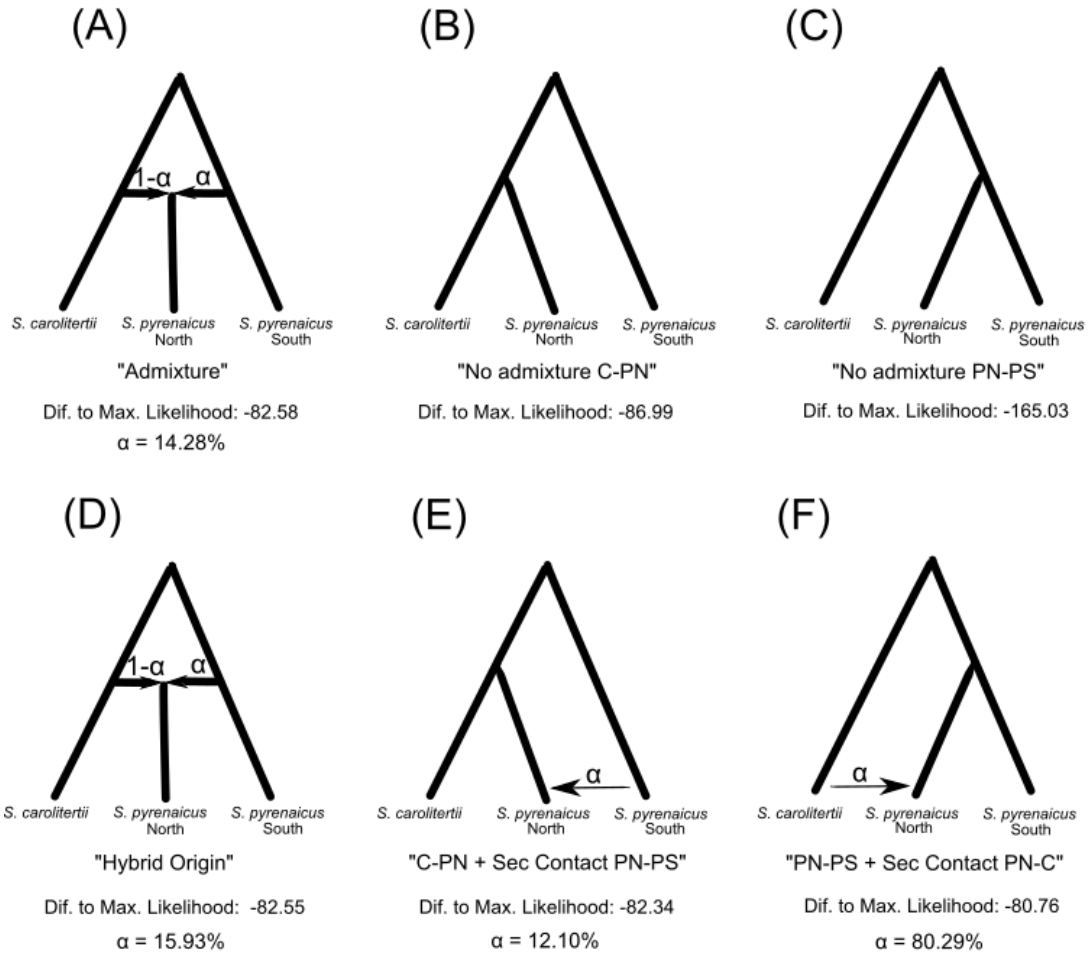

Figure 6 – Schematic representation of the likelihood of the models tested with *fastsimcoal2* and percentages of admixture inferred. The name given to each model is indicated below the schematic representation, as well the difference to maximum likelihood (Dif. To Max. Likelihood) which is the difference 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.

845 **Table 1 – Pairwise FST calculated between the different sampling locations.** *S. pyrenaicus* Guadiana
846 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	-	-	-	-	-	-	-

847

848