MOLECULAR ECOLOGY

Genomic survey provides insights into the evolutionary changes that occurred during European expansion of the invasive mosquitofish (Gambusia holbrooki)

Journal:	*Molecular Ecology*
Manuscript ID	MEC-15-0854.R1
Manuscript Type:	Original Article
Date Submitted by the Author:	10-Dec-2015
Complete List of Authors:	Vera, Manuel; University of Girona, Biology, Genetics Area; University of Santiago de Compostela, Genetics Diez-del-Molino, David; University of Girona, Biology Garcia-Marin, Jose Luis; University of Girona, Biology
Keywords:	RADtag sequencing, SNPs, population genomics, Invasive Species, genetic drift, natural selection

Manuscript, December 2015

Do not distribute without authors' permission

Genomic survey provides insights into the evolutionary changes that occurred during European expansion of the invasive mosquitofish (*Gambusia holbrooki)*

Vera, M. [1*]

Díez-del-Molino, D [1,$]

García-Marín, J.L.[1]

[1]LIG-UdG, Department of Biology, Faculty of Sciences, Universitat de Girona. Campus of Montilivi, 17071 Girona, Spain

* to whom correspondence should be addressed

TLF +34 630 75 69 62

Fax +34 972 41 82 77

e-mail: manuel.verar@udg.edu

Running title: Population Genomics for the invasive *Gambusia holbrooki*

$ Present address: Research Department of Genetics, Evolution, and Environment, University College London, WC1E 6BT, London, United Kingdom

Abstract

Biological invasions rank among the main global threats for biodiversity. The eastern mosquitofish (*Gambusia holbrooki*) is considered one of the 100 world worst invasive species due to its high adaptation capability to new environments. Using the restriction-site associated DNA tags (RADtags), introduced European locations were compared against native USA mosquitofish populations to analyze genomic changes that occurred during invasive process of European locations. After filtering, 7,724 RADtags containing only one SNP were retained for population studies. Comparative genomics indicated that 186 of these RADtags matched sequences in the transcriptome of *Xyphophorus maculatus*, the most closely related genome available. Genomic analyses showed that invasive populations show high reductions in diversity. Further, analyses of population structuring based on these data are concordant with previous analyses based on microsatellites. It is concluded that during the invasion process genetic drift was the main evolutionary force affecting patterns of diversity and population structure. While recognizing that positive selection could be masked by the strong drift during founder events, adaptive processes were evidenced in a reduced number of RADtags (less than 2%), with only one of these in a putative coding region. Surprisingly, balancing selection was detected in several coding RADtags, suggesting that the preservation of polymorphism in specific genes could be more important than the average population diversity for the population maintenance at any location, particularly for the survival of introduced populations.

Keywords: RADtag sequencing, SNPs, population genomics, invasive species, genetic drift, natural selection.

Introduction

The introduction and spread of nonindigenous species in new regions is one of the main global threats to biodiversity (Clavero & García-Berthou 2005). Freshwater ecosystems are among the most affected by invasions (WWF 2014). Genetic diversity is expected to decline during biological invasion episodes due to founder effects associated with a reduced number of introduced individuals. If the low effective population sizes (Ne) are maintained over time, or at least during several generations following introduction, the effects of subsequent genetic drift may limit the success of the introduced population by increasing inbreeding depression and reducing adaptive potential (Allendorf & Lundquist 2003; Blackburn *et al*. 2015). However, multiple introductions from different source populations (Kolbe *et al.* 2004) or migration (i.e. gene flow) between newly-established populations of alternative origin in the invaded range may maintain or even recover the population genetic diversity of the species in these territories (Facon *et al.* 2008; Keller & Taylor 2010). While several evolutionary processes related to successful invasions, such as the relative importance of propagule pressure to retain enough genetic diversity for establishment, the best traits to predict invasion success or the differences in local adaptation among invaders are still unknown (Bock *et al*. 2015). The Eastern mosquitofish, *Gambusia holbrooki*, is a small, ovoviviparous poeciliid species native to the southeastern USA that was introduced worldwide as biological control agent for mosquito populations (reviewed in Garcia-Berthou *et al.* 2005). Originally supported by governmental health agencies, these introductions resulted in an important threat to native species of fish and amphibians (see Stockwell & Henkanaththegedara 2011). For this reason, and together with *G. affinis*, *G. holbrooki* is listed as one of the 100 world worst invasive species (Lowe *et al.* 2000). Several ecological factors contribute to the worldwide invasive success of poeciliid fish,

including a short period for population recovery (Chapman & Warburton 2006; Deacon *et al.* 2011), multiple paternity (Neff *et al*. 2008), and high dispersal capabilities (Rehage & Sih 2004). Recently translocated populations of *G. affinis* from shared ancestors in Hawaii and in North America displayed divergences among introduced populations reflecting environmental adaptation (reviewed in Reznick & Ghalambor 2001). In fact, mosquitofish quickly adapt in response to environmental factors such as the presence of predators (Langerhans *et al*. 2007), salinity and thermal gradients (Congdon 1994; Meffe *et al.* 1995; Purcell *et al.* 2012), and pollutants (e.g. Tatara *et al*. 2002). Some of the above adaptations were maintained over time even in the absence of barriers to gene flow (Purcell *et al*. 2012).

Mosquitofish arrived to Europe with the release of only 12 individuals in a Spanish pond in 1921 (Vidal *et al.* 2010). The following year, 200 descendants of these introduced individuals were transferred to Italy, and subsequent human translocations spread *G. holbrooki* throughout the Mediterranean countries (Garcia-Berthou *et al.* 2005). The wide ecological tolerance and aggressive behavior has been claimed to facilitate the integration of *G. holbrooki* in these invaded fish communities (Ribeiro *et al.* 2008). A mtDNA survey throughout European *G. holbrooki* populations confirmed a common source for the majority of these populations and indicated a notorious uniformity among Iberian ones (Vidal *et al.* 2010). These analyses linked the European populations to the native Type I group found in northern USA basins that were invaded after the glacial ice-retreat (Wooten *et al.* 1988; Scribner & Avise 1993). Genetic diversity in these American populations is lower than in southern non-glaciated areas along the Atlantic coast of the USA (Wooten *et al*. 1988; Scribner & Avise 1993; Hernández-Martich *et al*. 1995). Surprisingly, substantial genetic diversity was observed at nuclear markers among Iberian populations from distinct river basins

(Vidal *et al*. 2012; Sanz *et al*. 2013), with considerable variation also in life-history traits such as size-at-maturity, reproductive effort, and gonadal size (Benejam *et al*. 2009; Carmona-Catot *et al*. 2011). Reduced mitochondrial variation and significant nuclear divergence were also observed among introduced Australian populations of the guppy (*Poecilia reticulata*), another worldwide distributed poeciliid. However, in spite of founder effects related with the introductions, substantial adaptive variation was apparently retained in these guppy populations (Lindholm *et al*. 2005). Deacon *et al*. (2011) showed that guppy populations founded from a single gravid female can preserve enough adaptive potential to colonize new territories. Compared to Australian guppies, European mosquitofish populations apparently retained most of the genetic variation of their American sources (Sanz *et al*. 2013), and apparently also the invasive potential acquired during the post-glacial colonization of the northernmost American basins (Díez-del-Molino *et al*. 2013).

Improvements in speed, cost and accuracy of next-generation sequencing (NGS) and bioinformatics tools are increasing the availability of genetic resources for non-model organisms (Helyar *et al.* 2011). These advances are fostering genomic studies in ecology and evolution through the fast development of new molecular markers such as single nucleotide polymorphisms loci (SNPs) (Allendorf *et al.* 2010; Davey *et al.* 2011). High throughput sequencing likely will be a key tool to analyze genomic basis of adaptation of non-native species (Blanchet 2012), and genomic studies may help disentangle the genetic processes that occurred during invasions, including the discovery of the sources of genetic variation in invasive populations, the role of expansion load, the relative importance of propagule pressure vs. genetic diversity for successful establishment, or the role of chromosomal rearrangements (Bock *et al*. 2015). Genomic approaches have been applied to study introgression events between

native and introduced populations of freshwater fish species (e.g Hohenlohe *et al*. 2011; Hand *et al*. 2014), although the number of studies of invasive processes in freshwater ecosystems is still limited (see Critescu 2015). SNPs occur in non-coding and coding regions thus the study of neutral variation and of genomic regions under selection (e.g. Bruneaux *et al.* 2013). Additionally, the potential for their high-throughput genotyping offers the possibility to obtain extremely large collections of data at relatively modest and decreasing costs (Davey *et al.* 2011). Genotyping-by-sequencing (GBS) techniques are based on the massive sequencing of a reduced part of the genome followed by SNP discovery (Elshire *et al.* 2011). High-throughput sequencing of restriction-site associated DNA tags (RADtags) (Miller *et al.* 2007; Baird *et al.* 2008) allows efficient identification of SNPs because the coverage of each RADtag is extremely high (e.g. Andrew *et al*. 2013; Fraser *et al*. 2015), and a large proportion of the identified RADtags are well represented in all analyzed individuals (Davey *et al.* 2011; Wang *et al.* 2013). Recently, this methodology has been successfully applied to SNP discovery in non-models species allowing researchers to perform genomic studies related with ecological, evolutionary and population genetic issues (see Narum *et al*. 2013 for review).

In this work, we used the RAD sequencing approach to explore genomic changes between native American and introduced European mosquitofish populations. Comparisons of diversity levels and population structure between the two geographical territories provided insights of the relative relevance of the evolutionary forces (drift, migration, selection) during the invasive process. We also noted limitations of current statistical approaches to detect directional positive selection involved in local adaption after recent strong drift events. Furthermore, due to the scarcity of molecular markers in *G. holbrooki*, the genomic resources obtained in this study are also introduced.

Material and methods

Sampling sites

A total of 96 individuals from eight different locations, including three North American native locations and five European invaded locations, were analyzed (Table 1, Fig.1). According with previous genetic analyses using microsatellite loci and mtDNA sequences, American Everglades (EV) and Florence (FLO) locations represented the Type-II populations of *G.holbrooki* as defined by Wooten et al (1988), while Potomac (PO) which has been suggested as genetically close to the source populations involved in the European introduction (Vidal *et al.* 2010; Sanz *et al.* 2013), represented the Type-I. The four Spanish collections represented the upper and lower reaches of two geographically close rivers (see Díez-del-Molino *et al*. 2013). Individuals were stored in 95% ethanol and frozen at -20ºC until assayed.

RADtag sequencing

Total genomic DNA was extracted from muscle tissue using the DNeasy Blood and Tissue Kit (Qiagen) following the manufacturer's instructions. Subsequently, 3 µg of high molecular weight (> 50Kb) DNA samples fluorometrically quantified by Qubit (Invitrogen, Carlsbad, CA) (concentration ≈ 30 ng/ µl) were sent to FLORAGENEX (Eugene, Oregon) to generate and sequence RADtags following the methods outlined by Baird *et al*. (2008), Hohenlohe *et al*. (2010) and Emerson *et al*. (2010). Briefly, sequencing adaptors and individual barcodes were ligated to fragments of total genomic DNA obtained by digestion using SbfI restriction enzyme (enzyme target: 5'…CCTGCA/GG… 3', giving cohesive extremes). The resulting fragments were sequenced from the restriction sites. Individually barcoded RAD samples were jointly sequenced on the Illumina HiSeq2000 platform with single-end 100-bp chemistry.

Reads were separated by individual, and sequencing barcodes were removed after the sequencing run, resulting in RADtags of 90 bp of length.

RADtag database description and SNP identification

Because the genome of *G. holbrooki* is not available, the individual VM43 collected in Vilanova de Muga (Spain) was used as a model skeleton framework for the construction of RADtags due to the high quality of its sequences. The Illumina sequence reads from this sample were trimmed from the 3' end to a total length of 90 bp. FLORAGENEX custom PERL scripts were used to cluster 100% identical sequences represented between 5 and 500 times in VM43 sequence dataset. Based on the observed coverage distribution from VM43, these sequences were posited to be sequence reads from low-copy or single dose RADtags in the *G. holbrooki* genome. The VM43 assembly was condensed to FASTA format and then aligned against itself using BWA software (Li & Durbin 2009). BWA alignment parameters were *aln* function, edit distance (-n) of 3 and the iterative search disabled. This self-alignment was performed to identify sequences within the assembly that carried substantial homology to one another. BWA alignment criteria also permitted a maximum 3bp mismatch between reference and query (96.6% sequence identity). After alignment, any cluster with more than two observed haplotypes during alignment was discarded as a potential paralog or duplicated sequence in the assembly. The final VM43 assembly contained 44,398 RADtag clusters (see results). Using the program BLAST+ (NCBI), this final assembly was locally aligned against the *Xyphophorus maculatus* transcriptome (23,236 sequences), representing the most closely related fish species transcriptome available in GenBank. Hrbek *et al*. (2007) included *Gambusia* and *Xyphophorus* genus in a monophyletic cluster of Central and North American poeciliid species that likely diverged during the Oligocene (33-22 Mya). BLASTN searches were

performed using a word size (length of the initial exact match) of 90 for accuracy, and with an e-value threshold of $< 10^{-5}$ to consider the match significant. In case of multiple hits, only the best match was kept. Only identity percentages ≥ 90% were considered. These stringent searching conditions were used to ensure that the RADtags matches against *X. maculatus* transcriptome were reliable. Statistics on output files were performed using R custom scripts.

The VM43 assembly was used for mapping raw Illumina sequence reads for the rest of analyzed samples into the constructed RADtags using BOWTIE software version 0.11.3 (Langmead *et al.* 2009). For each individual, the ref_map.pl pipeline included in STACKS program version 1.19 (Catchen *et al*. 2011, 2013) was used to analyze the BAM files provided directly by FLORAGENEX. This ref_map.pl pipeline executes sequentially the STACKS programs for building the different loci and calling their SNPs (pstacks), creating a catalog (cstacks), matching each sample against the catalog (sstacks) and finally indexing the data into the database (index_radtags.pl). STACKS also loaded outputs into a MySQL database to facilitate their visualization. This database was used for mining and filtering the data. The number of mismatches allowed between loci when building the catalog (-n) was 0 and the minimum coverage depth to report a stack in an individual was 10X. To reduce genotyping errors and spurious SNPs, we selected only RADtags with a single SNP for population genomic analyses. After this data filtering, all RADtags containing repetitive motifs were identified and removed for subsequent analyses using SciROKO 3.4 (Kofler *et al*. 2007) with default parameters. RADtags containing a single SNP included in the above list among the ones matching against the *X. maculatus* transcriptome were selected as a subset of RADtags (hereafter named MXT) to describe population diversity patterns from coding regions, and further annotation analyses.

Genetic diversity and population structure

The computer program POPULATIONS, included in STACKS, was used to process the SNP dataset across all analyzed locations. Diversity indices were estimated without allele frequency restriction and with a minimum percentage of 70 % genotyped individuals to process a locus for that population (- r option). The minimum stack depth (- m option) required for individuals at a RADtag was 10X. POPULATIONS provided the corresponding GENEPOP files (--genepop option) which were further transformed for the subsequent analyses using PGDSpider 2.0.7.4 (Lischer & Excoffier 2012). ARLEQUIN 3.5 (Excoffier & Lischer 2010) was used to estimate genomic measures of genetic diversity [number and percentage of polymorphic loci, mean number of alleles per locus and expected heterozygosity (HE)] at each of the studied locations and the overall and pairwise population differentiation (F_{ST}) values (significance for F_{ST} values estimated by 10,000 permutations). Analyses of the molecular variance (AMOVA; Excoffier *et al.* 1992) carried out with ARLEQUIN were used to study the distribution of genetic variation within and among population groups according to geographical hierarchical models of population clustering. The number of population units (K) was estimated using the Bayesian Markov Chain Monte-Carlo clustering approach implemented in the STRUCTURE 2.3.3 program (Pritchard *et al.* 2000) and running the output files in the STRUCTURE HARVESTER program (Earl & VonHoldt 2012) which implements the methodology proposed by Evanno et al. (2005) for K estimation. Due to computational limitations, a total of 10,000 burn-in steps and 20,000 replicates were performed for each run, with ten replicates for each K value tested (K= 1-9). The contribution of genetic drift to differentiation among populations was explored using the search algorithm implemented in TREEMIX program (Pickrell & Pritchard 2012). TREEMIX fits a population graph (i.e. a phylogenetic tree that incorporates

admixture) on the basis of the allele frequencies and a Gaussian approximation to genetic drift, allowing patterns of splits and mixtures in multiple populations to be inferred. Files were translated into TREEMIX format using the eight locations as clusters. We searched the best model to fit the data using a range of migrational events from 0 to 8. Because the EV location was situated in a suggested glacial refugium for the species, we run TREEMIX using this location as root for all the trees. Independence of all SNPs was assumed using the *–K 1* label.

Finally, a F_{ST} outlier approximation was used to determine the SNPs showing higher F_{ST} (i.e. positive selection) or lower F_{ST} values (i.e. balancing or purifying selection) than expected under a neutral model of selection using the approach of Beaumont & Balding (2004) implemented in the software BAYESCAN 2.1 (Foll & Gaggiotti 2008). This approximation was selected over the method proposed by Beaumont & Nichols (1996) implemented in LOSITAN software (Antao *et al.* 2008) because BAYESCAN considers scenarios where the effective size and the immigration rate may differ among analyzed locations (Foll & Gaggiotti 2008) as expected for our locations (Diez-del-Molino *et al*. 2013). RADtags were analyzed in five models: (1) including all locations, (2) pooling the samples in the two best population groups proposed by STRUCTURE analyses (EV + FLO group vs PO + European samples group, see results section), (3) only USA samples, (4) only European samples and (5) only Spanish samples (Table 1). The runs were performed using prior odds value of 10, with 100,000 iterations and a burn-in of 50,000 iterations. Only loci with a posterior probability (P) ≥ 0.99 and a FDR < 0.01 were considered as outliers. SNPs in the outlier RADtags from the MXT subset were located inside the *X. maculatus* genes and the two resulting sequences (one for each SNP variant) were translated to determine whether SNP variants lead synonymous or non-synonymous changes using the program BIOEDIT 7.1 (Hall 1999).

Finally, these genes were functionally categorized according the three Gene Ontology (GO) terminology categories (biological process, molecular function and cellular component) using the DAVID web-service (Jiao *et al.* 2012).

Results

RADtag development and genomic resources

The raw VM43 assembly consisted of 3,215,755 sequence reads. After quality filtering procedures, a total of 54,939 RADtag clusters were coalesced from 2,886,102 sequence reads, representing the 89.7% of the total sequence data from this individual. After homology filtering, the VM43 assembly retained 44,398 RADtags, representing 3.995 Mb of mosquitofish genomic sequence. The number of RADtags of VM43 that matched against *X. maculatus* mRNAs was 562 with the word size of 90. The stringent conditions used in our search ensured that these RADtags represented coding regions also in *G. holbrooki* (although many other *G. holbrooki* coding regions likely remain undetected and present in our 44,398 RADtag dataset).

The total number of reads for the 96 mosquitofish specimens analyzed was 394,574,734. Due to low quality results, one individual for each American location (EV, FLO and PO) were not used to construct the SNP database and therefore only 93 mosquitofish were used for subsequent analyses. The number of reads per individual ranged from 612,917 for the individual BL42 to 6,350,615 for the individual FLO06 with a mean number of sequence reads per sample 4,178,503.5 (± 824828.2). Per individual, the number of RADtags ranged from 36,418 for BL42 to 72,011 for EV16 (mean 60,683.6 ± 5,837.1) for a coverage higher than 5X.

SNP calling and datasets

The number of RADtags present in the database constructed from the BAM files of the 93 individuals studied was 44,271 with 36,052 of these containing SNPs. The number

296 of RADtags with a single SNP was 7,724 (Supplementary Table 1), and the SNPs
297 presented a homogeneous distribution along the RADtag sites/nucleotides
298 (Supplementary Figure 1). Of all these single SNPs, 5,259 involved transitions (A/G
299 and C/T) and 2,465 transversions (A/C, A/T, C/G and G/T), with A/G being the most
300 common (2,667) and A/T the least common (558). Of the 7,724 filtered RADtags, 46
301 showed repeat motifs and were excluded for subsequent analyses (marked in red into
302 Supplementary Table 1).

303 *Genetic variability and population structure analyses*

304 After filtering for population genotyping, the number of RADtags retained was 7,621,
305 185 of them creating the MXT subset of undoubtedly coding regions (marked in green
306 into Supplementary Table 1). Considering the 7,621 SNPs, the most variable location
307 according with the number of polymorphic loci, the percentage of polymorphic loci and
308 gene diversity (H_E), was the American EV while the least variable was the Italian CO
309 (Table 2). Significant reductions in diversity levels were found in comparisons between
310 USA and European locations (Mann-Whitney U tests; $P \leq 0.025$ for the different
311 genetic estimators at all comparisons). PO was the least diverse of USA locations
312 studied, although this reduction was not significant (Mann-Whitney U tests; $P > 0.600$
313 for the different genetic estimators at all comparisons). Reduced diversity was also
314 observed in the Italian CO location when compared with Spanish ones (Table 2), but
315 again non-significant (Mann-Whitney U tests with $P > 0.150$ for all analyzed local
316 diversity indexes).

317 Global estimate of population differentiation (F_{ST}) was 0.416 ($P < 0.001$). Average
318 pairwise F_{ST} values ranged from 0.045 between EP and PP locations, to 0.547 between
319 EV and CO. The number of genetic homogenous units (K) inferred following the
320 Evanno's method with the STRUCTURE results was 2 (Delta value $_{(K=2)}$ = 1264.2,

Delta value $_{(K=4)}$ = 76.3, Delta value $_{(K=3,5,6,7,8,9)}$ < 1.2). One of the clusters grouped together EV and FLO mosquitofish, while the other grouped American PO mosquitofish with all the European fish (Fig.4). Considering K= 3, mosquitofish from EV and FLO locations separated in two different clusters (Fig.4). Any hierarchical AMOVA model suggested significant genomic divergence among locations within group. AMOVA arranged by studied countries resulted in worst grouping of the locations compared to the one based on the two groups identified by STRUCTURE. The proportion of variation due to differences among groups (F_{CT}) of this later AMOVA model (EV + FLO vs PO + European samples, see Table 4), was six fold larger than the proportion within groups (F_{SC}). The AMOVA models grouping the American EV or FLO locations with European mosquitofishes resulted in larger divergences among locations within group and non-significant divergence among groups (P>0.05, Table 4), corroborating historical records and previous genetic evidences regarding the origin of European mosquitofish populations from some location in the native northern range of the Type-I *G. holbrooki*. Diversity patterns showed by the MXT RADtags subset agreed fully with observations from the complete data set (Table 2 and 4).

TREEMIX results reported the best model fit for the tree containing three migration events (Fig. 5). The tree supported the colonization pathway suggested in previous studies (i.e. Southern USA (EV+PLO) > Northern USA (PO) > European samples). It also suggested a prominent role of genetic drift in the steps of these mosquitofish expansions, with a substantial increase (three to four fold) in drift in American PO and European mosquitofish locations as compared with the southernmost American locations EV and FLO. Results indicated admixture within European locations.

346 *Outlier detection and gene annotation*

347 From the 7,621 markers analyzed, selection scans indicated that a total of 185 SNPs
348 had a posterior probability larger than 0.99 of being outlier (Supplementary Table 2;
349 Fig. 4). Eighty-five markers were associated to positive alpha values (positive
350 directional selection,) while 100 were associated to negative alpha values (balancing or
351 purifying selection). Only four out of these 185 outliers were listed among the MXT
352 subset representing coding regions in *G.holbrooki.* One of them (RADtag 11446)
353 showed positive selection, while the other three showed balancing selection (Table 5).
354 Selection scans did not detect outliers RADtags when comparisons were restricted to
355 the USA locations, or in comparisons of pooled individuals according to the two
356 population groups suggested by STRUCTURE. Analysis including only the Spanish
357 locations detected 308 outlier RADtags, all of them candidates for balancing or
358 purifying selection, with only ten of these from the MXT subset (Table 5). The number
359 of outliers increased to 346 when analyzing all European locations (all of them again
360 related with balancing or purifying selection, and seven of them from MXT subset).
361 The three outliers RADtags from the MXT subset putatively under balancing selection
362 when comparing all samples, remained under balancing selection in the analysis
363 restricted to the Spanish or European locations. Overall, 11 different RADtags from the
364 MXT subset resulted outliers showing negative alpha value in more than one of the
365 different geographical analyses performed. The allele frequencies of the two variants of
366 the SNP contained in these RADtag were intermediate, suggesting balancing rather
367 than purifying selection. The outlier candidate for positive selection (RADtag 11446)
368 showed a fixed variant for EV and FLO (G nucleotide) whereas the American PO and
369 the European locations for another (A nucleotide). All the outlier SNPs of the MXT
370 subset, except RADtag 37,928 that corresponds to an uncharacterized/unknown protein

(Table 5), were annotated using the available information from *X.maculatus*. All the annotated SNPs were located inside the coding region of the annotated genes. Six of them SNPs leaded to silent synonymous change while the other four leaded to non-synonymous change.

The ten outlier mosquitofish associated to genes were involved in signal transduction, apoptosis, cell morphogenesis, locomotion, organ development and catabolic processes (Supplementary Table 3). When compared to human genome, they were enriched in three different GO terms (FDR < 0.05): GO: 0005737 cytoplasm (cellular component category) found in eight genes (ARFGEF2, ABTB1, CYFIP2, FEZ1, RARB, SLC9A3R2, UBR4 and ZRANB1), GO: 0005515 protein binding (molecular function category) also found in eight genes (ARFGEF2, ABTB1, CYFIP2, FEZ1, MPP2, SLC9A3R2, UBR4 and ZRANB1), and GO: 0016881 amino-acid ligase activity (molecular function) found in two genes (UBR4 and ZRANB1).

Discussion

Genomic resources for mosquitofish

The application of NGS techniques to non-model organisms is fuelling the availability of genomic resources, providing a huge number of valuable molecular information to analyze genetic diversity, population structure and adaption in these organisms (Helyar *et al.* 2011). In this study a trustable genomic database containing information on thousands of SNPs that can further the study *G. holbrooki,* or other close-related poeciliid species such as *G. affinis* or *Poecilia reticulata*, is presented

Despite of its rank as one of the 100 world's worst invasive species (Lowe *et al.* 2000), the number of molecular markers available to study mosquitofish was extremely low [30 microsatellite loci (Parker *et al.* 1998; Zane *et al.* 1999; Spencer *et al.* 1999; Purcell *et al.* 2011) and five associated-gene SNPs designed for *G.holbrooki* (Vidal *et al.*

2012)]. This study increases substantially the number of makers available to study mosquitofish, by generating a 36,052 polymorphic RADtags. This total is similar to those described in databases for other species created using a similar number of individuals and techniques (Narum *et al.* 2013; Richards *et al.* 2013; Wang *et al.* 2013). As suggested by the homogeneous distribution of filtered SNP positions in the RADtags, sequencing errors were not responsible for the observed SNPs (Fig. 2). A transition/transversion (ts/tv) ratio close to 0.50 would be obtained when a large portion of detected SNPs resulted from sequencing errors because a random nucleotide error has a probability of 1/3 to be a transition (A↔G, C↔T) and 2/3 to be a transversion (A↔C, A↔T, C↔G, G↔T). However, we obtained a ts/tv ratio of 2.13, and similar to those reported in other species analysed with RADtag sequencing (*Solanum melongena* 1.65, Barchi *et al*. 2011; *Anguilla anguilla* 1.60, Pujolar *et al*. 2013), and fish species analysed using other sequencing approaches (*Sparus aurata* 1.38, Cenadelli *et al*. 2007; *Salmo salar* 1.37, Hayes *et al*. 2007; *Scophthalmus maximus* 1.35-1.88, Vera *et al*. 2011, 2013).

Mosquitofish population genomics: local diversity and population structure

A cline of decreasing genetic diversity towards northern mosquitofish populations was observed in American locations (Table 2). In fact, the genetic diversity levels of mosquitofish collected in PO location were even lower than those observed in post-glacially founded freshwater Alaskan populations of three-spine stickleback, *Gasterosteus aculeatus,* using the same methodology (Hohenlohe *et al.* 2010). In the Northern hemisphere, the genetic variability of many species of animals and plants has been influenced by the Quaternary ice ages (Hewitt 2000). The effects of these climatic fluctuations resulted in range expansions and contractions, commonly depicted in the population structure of species with a contrasting differentiation between northern and

421 southern populations. In North America, comparative phylogeographical congruence
422 across many taxa indicates historical biogeographical signal and a major role for
423 climate cycles in intraspecific diversification along the Gulf-Atlantic Coastal Plain
424 (Avise 1992). The Florida peninsula has been proposed as glacial refuge for many fish
425 species (Bermingham & Avise 1986; Bernatchez & Wilson 1998; Bagley *et al.* 2013).
426 Values of SNPs diversity among mosquitofish *G.holbrooki* populations agreed with the
427 patterns of divergence observed in other species, confirming at a genomic scale the
428 vicariance between northern and southern populations. In fact, increased diversity in
429 Florida and other southern mosquitofish populations yields support for this region as a
430 glacial refuge for this species (Scribner & Avise 1993).

431 Vidal *et al*. (2010) conducted mtDNA sequence analyses on the same samples
432 employed in present study and found that the EV mosquitofish were related with the
433 southern and more diverse Type II group, whereas FLO and PO collections contained
434 haplotypes characteristic of the northern Type I group. The most common mtDNA
435 haplotype observed at PO (Hol1) was also the most common among introduced
436 European populations, suggesting that the source of the invasion came from a
437 mosquitofish population close to PO. Population differentiation patterns reported in
438 here also agreed with recent observations using microsatellite loci (Sanz *et al*. 2013,
439 Díez-del-Molino *et al*. 2013). Those studies assigned the highest levels of diversity to
440 the EV location from Florida (H_S= 0.867), and confirmed PO as the genetically closest
441 population to introduced European mosquitofish populations (Fig. 3). Significant
442 reductions in diversity levels (measured as H_E and percentage of polymorphic loci)
443 have been detected during the invasion process as consequence of sequential founder
444 effects (Lockwood *et al*. 2005). In our results, American mosquitofish locations

harbored larger levels of genetic diversity compared to the introduced European populations (Table 2).

At a global level, the results presented here agree with other studies that identify (EV) as the more diverse location and the likely source for the USA northern natural post-glacial colonization, whereas the least diverse (CO) probably represents the last invasive event among the European locations studied (Table 2; Graputto *et al*. 2006). Genetic losses of variation have been described for many invasive species, both in animals and plants, but these not necessarily imply also the loss of adaptive potential (Duglosch & Parker 2008). For example, in *Hypericum canariense*, a native endemic plant of the Canary Islands introduced in USA Pacific coast and Hawaii Islands, approximately 45 % of the diversity was lost in each invasive event but evidences for adaptive genetic changes were also shown for increased growth in terms of both survival and reproduction (Duglosch & Parker 2008).

Northern USA locations genetically close to PO have been suggested to act as "invasive bridgeheads" to the European invasion (Diez-del-Molino *et al.* 2013). Invasive bridgeheads are described as particularly successful invasive populations that accumulate the evolutionary changes necessary to become invaders, and as such the source of colonists for remote new habitats (Lombaert *et al.* 2010). Following the hypothesis of Díez-del-Molino *et al*. (2013), northern USA locations would have acquired their invasive potential due to the stress of post-glacial expansion process. Therefore, the invasive potential of mosquitofish would predate recent anthropic introductions. In fact, populations from peripheral ranges of distribution often show low genetic diversity levels but display greater stress-adaptation (Hardie & Hutchings 2010). Analyses with STRUCTURE and AMOVA depict a close genomic relation

between PO and the European locations, confirming the status of the former as the source for European locations.

Evolutionary forces driving invasive process: genetic drift versus selection

Understanding genetic changes involved in invasive potential as well as developing genetic control measures are key topics to prevent and control invasions (Lee 2002; Handley *et al*. 2011). Mosquitofish diversity loses at genomic scales during the European colonization affected equally coding and non-coding regions (Table 2). In agreement with theoretical expectations at early stages of invasions (Allendorf & Lundquist 2003), such losses reflect the effects of genetic drift rather than other evolutionary forces. Genetic drift related to the founder events during the post-glacial colonization has been invoked as responsible for the genetic depauperation of northern USA *G. holbrooki* populations (Scribner & Avise 1993). Strong genetic drift would differentiate introduced mosquitofish populations by randomly fixing allelic variants, resulting in significantly high F_{ST} values detected either in the global analysis (F_{ST} = 0.416) and in pairwise comparisons among studied locations (Table 3). Unshared genetic drift also explained a major part of the differentiation seen between PO and the other American collections in the TREEMIX analysis.

Mosquitofish have colonized all Mediterranean coastal basins in the Iberian Peninsula competing and displacing such native fishes as *Valencia hispanica* and *Aphanius iberus* (Rincon *et al*. 2002; Carmona-Catot *et al*. 2013), and are currently spreading towards the headstreams of all major Iberian river basins (e.g. Oscoz *et al.* 2008). Substantial divergence has been observed among Iberian mosquitofish populations from distinct river basins at life-history traits such as size-at-maturity, reproductive effort, or gonadal-size (Benejam *et al.* 2008; Carmona-Catot *et al.* 2011). Certainly, phenotypic plasticity enables organisms to respond to varying environments (reviewed in Murren *et*

494 *al.* 2015), but more often, the success of invasive species after introduction has been
495 associated to rapid adaptive response to novel selective processes during spread and
496 colonization of newer habitats (Barrett 2015). Despite the prominent role of founder
497 events during invasions, it appears that genetic drift does not eliminate adaptive
498 potential because many fitness-related traits may be polygenic and do not lose variation
499 as quickly as individual loci do (Duglosh & Parker 2008). Accordingly, substantial
500 additive variation could be retained during bottlenecks (e.g. Lindholm *et al.* 2005). Our
501 outlier analyses indicated 85 out of 7,724 studied RADtags to be under positive
502 selection pressure when comparing all locations. The lack of positive selective response
503 in comparisons involving only European collections indicated that allele changes in
504 RADtags related to positive selection occurred prior to the European mosquitofish
505 invasion. For instance, the RADtag 11446 showed a fixed G nucleotide at the studied
506 SNP at EV and FLO locations, while PO and all European locations had an A at that
507 site, resulting in a non-synonymous amino acid substitution. Positive selection may be
508 confused with demographic fluctuations because they leave similar signals in genomic
509 variation (Currat *et al.* 2006). Genetic drift tends to increase differentiation among
510 populations (i.e. high F_{ST} values), therefore outliers methods based on F_{ST} values are
511 limited to detect positive selection under strong genetic drift. The inclusion of
512 populations which suffered severe bottlenecks increases the false-positive outlier rate
513 for directional selection in BAYESCAN (Foll & Gaggiotti 2008). Microsatellite
514 analyses suggested that most of the mosquitofish diversity present in Europe can be
515 attributed to a single American source (Sanz *et al.* 2013), and since the introduction of
516 mosquitofish in Spain in 1921 consisted of only 12 individuals, it can be inferred that
517 several additional bottlenecks occurred during subsequent spread along the Iberian
518 Peninsula (Diez-del-Molino *et al*. 2013). It is very likely that not enough time has

519 elapsed to dismiss the effect of genetic drift on the outlier analyses performed. Hence
520 the relevance of adaptive pressures in European mosquitofish populations could remain
521 underestimated.
522 Interestingly, a larger number of outliers detected in this study were related to
523 balancing selection either in analyses involving all locations or in those restricted to the
524 European locations. Among Spanish location, nine RADtags undoubtedly associated to
525 coding regions in mosquitofish (MXT RADtag subset) showed balancing selection and
526 three of these consist of non-synonymous substitutions. The low FDR values (< 0.010,
527 Table 5) suggest that these results are reliable. Genes associated to these outliers were
528 related with cell morphogenesis, locomotion, organ development and some catabolic
529 processes which are important for individual development/integrity (Supplementary
530 Table 3). Accordingly, these results suggest that retention of genetic variability in some
531 specific genes is essential for the survival of both native and invasive mosquitofish
532 populations.
533 Putative loci under balancing selection have been often detected in genomic surveys of
534 population structure within species (e.g. Makinen *et al*. 2008). Nevertheless these loci
535 have received less attention than those showing positive directional selection because of
536 the more evident relationship of positive selection with local adaptation (Belanger-
537 Deschenes *et al*. 2013; Fisher *et al*. 2014) . Balancing selection in isolated small
538 populations retained diversity levels of MHC immunity in the guppies (Van Oosterhout
539 *et al*. 2006) and panthers (Castro-Prieto *et al*. 2011). High genetic variability in MHC
540 genes is necessary for the pathogen detection that triggers the immune response
541 essential for survival (Luo *et al*. 2012). Loci retaining polymorphism under putative
542 balancing selection has been also detected in yellow perch (*Perca flavescens*)
543 inhabiting polluted environments (Belanger-Deschenes *et al*. 2013).

Genetic drift associated with founder events at early stages of invasion tends to maximize genetic differentiation among populations by randomly fixing allele variants. Our results indicate that genetic drift shaped diversity and divergence among studied mosquitofish populations by increasing allele frequency variance among populations and consequently the estimates of F_{ST}. In such scenario polymorphic loci retaining similar allele frequencies in all populations would be rare and such diversity could be putatively maintained by balancing selection. There appears to be no studies documenting the effect of strong drift on the detection of outlier loci under balancing selection using genomic population data. By combining information from different approaches, the detection of outliers may improve by reducing rates of false positives (Excoffier *et al*. 2009; De Mita *et al*. 2013; De Villemereuil *et al*. 2014). Although BAYESCAN may be useful to detect outliers under some evolutionary scenarios (Perez-Figueroa *et al*. 2010), it is prone to yield false positive results when hierarchical substructure is present among locations. To minimize this potential error, Excoffier *et al*. (2009) proposed a method to detect outliers under a hierarchical island model where locations within groups exchange more migrants than locations from different groups. This methodology is implemented in ARLEQUIN 3.5 software, and such analysis was performed as follows: [number of simulations= 50,000; demes simulated (d) = 100; number of groups (k)= 10] including all locations grouped as suggested by STRUCTURE and AMOVA results (EV+FLO – PO+Europe). The results obtained confirmed the same four RADtag outliers suggested by BAYESCAN among the MXT RADtags subset. Thus, RADtag 11446 (CYFIP2 gene) was identified to be experiencing positive selection and RADtags 13210 (ZRANB1 gene), 36169 (ARFGEF2 gene), and 37928 (uncharacterized protein) were identified again to be experiencing balancing selection.

As far as we know, this is the first genomic evidence of the role of genetic drift and adaptation in the invasive process of a highly successful freshwater invasive fish, *G. holbrooki*. We added genomic evidence to explain how pre-invasive evolutionary changes, such as postglacial expansion in America, are relevant when exploring changes on levels of genetic diversity as a consequence of invasions. Finally, our results support a prominent role of genetic drift as one of the main evolutionary forces driving the invasive process, but preserving polymorphisms by balancing selection in specific genes could be more important than the average population diversity for the maintenance of local populations and for their invasive success.

Acknowledgements

Authors wish to thank FLORAGENEX team for their technical assistance, Dr. J. Boone for his comments about RAD technology, Dr. J. Catchen for his help with data analyses and Dr. J. Alvarado-Bremer (Texas A&M University) for grammatical improvements. The authors are also grateful to Dr. J. Gow, Dr. P. Hohenlohe and two anonymous reviewers for their constructive comments on the earlier version of this manuscript. DDM was supported by a predoctoral fellowship from the University of Girona (BR10/15). This study has been supported by the Spanish Ministry of Science and Innovation project CGL2009-12877-C02-02, and the AGAUR 2009 SGR/144 financial budget of the Autonomous Government of Catalonia.

References

Allendorf FW, Lundquist LL (2003) Introduction: population biology, evolution, and control of invasive species. *Conservation Biology* **17**, 24–30.

Allendorf FW, Hohenlohe PA, Luikart G (2010) Genomics and the future of conservation genetics. *Nature Reviews Genetics* **11**, 697-709.

Andrew RL, Kane NC, Baute GJ, Grassa CJ, Rieseberg LH (2013) Recent nonhybrid origin of sunflower ecotypes in a novel habitat. *Molecular Ecology* **22**, 799-813.

Antao T, Lopes A, Lopes RJ, Beja-Pereira A, Luikart G. LOSITAN: A workbench to detect molecular adaptation based on a F_{st}-outlier method. *BMC Bioinformatics* **9**, 323.

Avise JC (1992) Molecular population-structure and the biogeographic history of a regional fauna - a case-history with lessons for conservation biology. *Oikos* **63**, 62-76.

Bagley JC, Sandel M, Travis J, de Lourdes Lozano-Vilano M, Johnson JB (2013) Paleoclimatic modeling and phylogeography of least killifish, Heterandria formosa: insights into Pleistocene expansion-contraction dynamics and evolutionary history of North American Coastal Plain freshwater biota. *BMC Evolutionary Biology* **13**, 223.

Baird NA, Etter PD, Atwood TS*, et al.* (2008) Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *Plos One* **3**, e3376.

Barchi L, Lanteri S, Portis E*, et al.* (2011) Identification of SNP and SSR markers in eggplant using RAD tag sequencing. *BMC Genomics* **12**, 304.

Barrett SCH (2015) Foundations of invasion genetics: the Baker and Stebbins legacy. *Molecular Ecology* **9**, 1927-1941.

Beaumont MA, Nichols RA (1996) Evaluating loci for the use in the genetic analysis of population structure. *Proceedings of the Royal Society, Series B: Biological Sciences* **263**, 1619-1626.

Beaumont MA, Balding DJ (2004) Idenntfying adaptative genetic divergence among populations from genome scans. *Molecular Ecology* **13**, 969-980.

Belanger-Deschenes S, Couture P, Campbell PGC, Bernatchez L (2013) Evolutionary change driven by metal exposure as revealed by coding SNP genome scan in wild yellow perch (*Perca flavescens*). *Ecotoxicology* **22**, 938–957.

Benejam L, Alcaraz C, Sasal P, Simon-Levert G, García-Berthou E (2008) Life history and parasites of the invasive mosquitofish (*Gambusia holbrooki*) along a latitudinal gradient. *Biological Invasions* **11**, 2265–2277.

Benejam L, Benito J, García-Berthou E (2009) Decreases in Condition and Fecundity of Freshwater Fishes in a Highly Polluted Reservoir. *Water, Air, & Soil Pollution* **210**, 231–242.

Bermingham E, Avise JC (1986) molecular zoogeography of freshwater fishes in the southeastern United States. *Genetics* **113**, 939-965.

Bernatchez L, Wilson CC (1998) Comparative phylogeography of nearctic and palearctic fishes. *Molecular Ecology* **7**, 431-452.

Blackburn TM, Lockwood JL, Cassey P (2015) The influence of numbers on invasion success. *Molecular Ecology* **24**, 1942-1953.

Blanchet S (2012) The use of molecular tools in invasion biology: an emphasis on freshwater ecosystems. *Fisheries Management and Ecology* **19**, 120–132.

Bock DG, Caseys C, Cousens RD, *et al*. (2015) What we still don't know about invasion genetics. *Molecular Ecology* **24**, 2277-2297.

Bruneaux M, Johnston SE, Herczeg G*, et al.* (2013) Molecular evolutionary and population genomic analysis of the nine-spined stickleback using a modified restriction-site-associated DNA tag approach. *Molecular Ecology* **22**, 565-582.

Carmona-Catot G, Benito J, García-Berthou E (2011) Comparing latitudinal and upstream-downstream gradients: life history traits of invasive mosquitofish. *Diversity and Distributions* **17**, 214–224.

Carmona-Catot G, Magellan K, García-Berthou E (2013) Temperature-specific competition between invasive mosquitofish and an endangered cyprinodontid fish. *PloS One* **8**, e54734.

Castro-Prieto A, Wachter B, Melzheimer J, Thalwitzer S, Sommer S (2011) Diversity and evolutionary paterns of immune genes in free-ranging Namibian leopards (*Panthera pardus pardus*). *Journal of Heredity* **102**, 653-665.

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) Stacks: Building and Genotyping Loci De Novo From Short-Read Sequences. *G3-Genes Genomes Genetics* **1**, 171-182.

Catchen JM, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. *Molecular Ecology* **22**, 3124-3140.

Cenadelli S, Maran V, Bongioni G*, et al.* (2007) Identification of nuclear SNPs in gilthead seabream. *Journal of Fish Biology* **70**, 399-405.

Chapman P, Warburton K (2006) Postflood movements and population connectivity in gambusia (*Gambusia holbrooki*). *Ecology of Freshwater Fish* **15**, 357–365.

Clavero M, Garcia-Berthou E (2005) Invasive species are a leading cause of animal extinctions. *Trends in Ecology and Evolution* **20**, 110.

Collins JE, White S, Searle SMJ, Stemple DL (2012) Incorporating RNA-seq data into the zebrafish Ensembl genebuild. *Genome Research* **22**, 2067-2078.

Congdon B (1994) Salinity-related fitness differences amongst GPI genotypes in the mosquitofish *Gambusia holbrooki* (Poeciliidae: Teleostei). *Biological Journal of the Linnean Society* **53**, 343–352.

Critescu ME (2015) Genetic reconstructions of invasion history. *Molecular Ecology* **24**, 2212-2225.

Currat M, Excoffier L, Maddison W, *et al.* (2006) Comment on "Ongoing adaptive evolution of ASPM, a brain size determinant in Homo sapiens" and "Microcechaplin, a gene regulating brain size, continues to evolve adaptively in humans". *Science* **313**, 172.

Davey JW, Hohenlohe PA, Etter PD*, et al.* (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics* **12**, 499-510.

Deacon AE, Ramnarine IW, Magurran AE (2011) How reproductive ecology contributes to the spread of a globally invasive fish. *PLoS One* **6**, e24416.

De Mita S, Thuillet AC, Gay L, *et al*. (2013) Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. *Molecular Ecology* **22**, 1383–1399.

De Villemereuil P, Frichot É, Bazin É, François O, Gaggiotti OE (2014) Genome scan methods against more complex models: when and how much should we trust them? *Molecular Ecology* **23**, 2006–2019.

Diez-del-Molino D, Carmona-Catot G, Araguas R-M*, et al.* (2013) Gene Flow and Maintenance of Genetic Diversity in Invasive Mosquitofish (*Gambusia holbrooki*). *Plos One* **8**, e82501.

Dlugosh KM, Parker IM (2008) Founding events in species invasions: genetic variation, adaptive evolution, and the role of multiple introductions. *Molecular Ecology* **17**, 431-449.

Earl DA, vonHoldt BM (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetic Resources* **4**, 359-361.

Elshire RJ, Glaubitz JC, Sun Q *et al*. (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species.*PLoS ONE* **6**, e19379.

Emerson KJ, Merz CR, Catchen JM*, et al.* (2010) Resolving postglacial phylogeography using high-throughput sequencing. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 16196-16200.

Evanno G, Regnaut S, Goudet J (2005). Detecting the number of clusters of individuals using the software STRUCTURE : a simulation study. *Molecular Ecology* **14**, 2611-2620.

Excoffier L, Smouse PE, Quattro JM (1992) Analysis of moleculrar variance inferred from metric distances among DNA hapltypes: Application to human mitochondrial DNA restriction data. *Genetics* **131**, 479-491.
Excoffier L, Hoffer T, Foll M (2009) Detecting loci under selection in a hierarchically structured population. *Heredity* **103**, 285-298.
Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources* **10**, 564-567.
Facon B, Pointier JP, Jarne P, Sarda V, David P (2008) High genetic variance in life-history strategies within invasive populations by way of multiple introductions. *Current Biology* **18**, 363-367.
Fraser BA, Kunstner A, Reznick DN, Dreyer C, Weigel D (2015) Population genomics of natural and experimental populations of guppies (*Poecilia reticulata*). *Molecular Ecology* **24**, 389-408.
Fischer MC, Foll M, Heckel G, Excoffier L (2014) Continental-scale footprint of balancing and positive selection in a small rodent (*Microtus arvalis*). *PLoS ONE* **9**, e112332.
Foll M, Gaggiotti O (2008) A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics* **180**, 977-993.
Garcia-Berthou E, Alcaraz C, Pou-Rovira Q*, et al.* (2005) Introduction pathways and establishment rates of invasive aquatic species in Europe. *Canadian Journal of Fisheries and Aquatic Sciences* **62**, 453-463.
Grapputo A, Bisazza A, Pilastro A (2006) Invasion success despite reduction of genetic diversity in the European populations of eastern mosquitofish (*Gambusia holbrooki*). *Italian Journal of Zoology* **73**, 67–73.
Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program fro Windows 95/98/NT. *Nucleid Acids Symposium Series* **41**, 95-98.
Hand BK, Hether TD, Kovach RP, *et al.* (2014) Genomics and introgression: Discovery and mapping of thousands of species-diagnostic SNPs using RAD sequencing. *Current Zoology* **61**, 146-154.
Handley LJL, Estoup A, Evans DM*, et al.* (2011) Ecological genetics of invasive alien species. *Biocontrol* **56**, 409-428.
Hardie DC, Hutchings JA (2010) Evolutionary ecology at the extremes of species' ranges. *Environmental Reviews* **18**, 1-20.
Hayes B, Laerdahl JK, Lien S*, et al.* (2007) An extensive resource of single nucleotide polymorphism markers associated with Atlantic salmon (*Salmo salar*) expressed sequences. *Aquaculture* **265**, 82-90.
Helyar SJ, Hemmer-Hansen J, Bekkevold D*, et al.* (2011) Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. *Molecular Ecology Resources* **11**, 123-136.
Hernandez-Martich JD, Novak JM, Smith MH, Johns PE (1995) Genetic structure of mosquitofish populations in the Altamaha and Ogeechee drainages of Georgia: Reporting an undescribed form in the Ocmulgee River. *Biochemical Systematics and Ecology* **23**, 617-625.
Hewitt G (2000) The genetic legacy of the Quaternary ice ages. *Nature* **405**, 907-913.
Hohenlohe PA, Bassham S, Etter PD*, et al.* (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD Tags. *Plos Genetics* **6**, e1000862.

748 Hohenlohe PA, Amish SJ, Catchen JM, Allendorf FW, Luikart G (2011) Next-
749 generation RAD sequencing identifies thousands of SNPs for assessing
750 hybridization between rainbow and westslope cutthroat trout. *Molecular*
751 *Ecology Resources* **11**, 117-122.
752 Hrbek T, Seckinger J, Meyer A (2007) A phylogenetic and biogeographic perspective
753 on the evolution of poeciliid fishes. *Molecular Phylogenetics and Evolution* **43**,
754 986-998.
755 Jiao X, Sherman BT, Huang DW, *et al*. (2012) DAVID-WS: a stateful web service to
756 facilitate gene/protein list analysis. *Bioinformatics* **28**, 1805-1806.
757 Keller SR, Taylor DR (2010) Genomic admixture increases fitness during a biological
758 invasion. *Journal of Evolutionary Biology* **23**:1720-1731.
759 Kofler R, Schlötterer C, Lelley T (2007) SciRoKo: A new tool for whole genome
760 microsatellite search and investigation. *Bioinformatics* **23**, 1683–1685.
761 Kolbe JJ, Glor RE, Rodriguez Schettino L, *et al.* (2004) Genetic variation increases
762 during biological invasion by a Cuban lizard. *Nature* **431**, 177–181.
763 Langerhans RB, Gifford ME, Joseph EO (2007) Ecological speciation in Gambusia
764 fishes. *Evolution* 61, 2056–2074.
765 Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient
766 alignment of short DNA sequences to the human genome. *Genome Biology* **10**.
767 Lee CE (2002) Evolutionary genetics of invasive species. *Trends in Ecology &*
768 *Evolution* **17**, 386-391.
769 Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler
770 transform. *Bioinformatics* **25**, 1754-1760.
771 Lindholm AK, Breden F, Alexander HJ, *et al*. (2005) Invasion success and genetic
772 diversity of introduced populations of guppies *Poecilia reticulata* in Australia.
773 *Molecular Ecology* **14**, 3671–82.
774 Lischer HEL, Excoffier L (2012) PGDSpider: An automated data conversion tool for
775 connecting population genetics and genomics programs. *Bioinformatics* **28**, 298-
776 299.
777 Lockwood JL, Cassey P, Blackburn T (2005) The role of propagule pressure in
778 explaining species invasions. *Trends in Ecology and Evolution* **20**, 223-228.
779 Lombaert E, Guillemaud T, Cornuet J-M*, et al.* (2010) Bridgehead Effect in the
780 Worldwide Invasion of the Biocontrol Harlequin Ladybird. *Plos One* **5**, e9743.
781 Lowe S, Browne M, Boudjelas S, De Poorter M (2000) 100 of the World's worst
782 invasive alien species: a selection from the global invasive species database (ed.
783 ISSG), 12 p. World Conservation Union (IUCN).
784 Luo MF, Pan HJ, Liu ZJ, Li M (2012) Balancing selection and genetic drift at major
785 histocompatibility complex class II genes in isolated populations of golden
786 snub-nosed monkey (*Rhinophitecus roxellana*). *BMC Evolutionary Biology* **12**,
787 207.
788 Makinen HS, Cano JM, Merila J (2008) Identifying footprints of directional and
789 balancing selection in marine and freshwater three-spined stickleback
790 (*Gasterosteus aculeatus*) populations. *Molecular Ecology* **17**, 3565–82.
791 Meffe GK, Weeks SC, Mulvey M, Kandl KL (1995) Genetic differences in thermal
792 tolerance of eastern mosquitofish (*Gambusia holbrooki* Poeciliidae) from
793 ambient and thermal ponds. *Canadian Journal of Fisheries and Aquatic*
794 *Sciences* **52**, 2704–2711.
795 Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA (2007) Rapid and cost-
796 effective polymorphism identification and genotyping using restriction site
797 associated DNA (RAD) markers. *Genome Research* **17**, 240-248.

Murren CJ, Auld JR, Callahan H, *et al.* (2015) Constraints on the evolution of phenotypic plasticity: limits and costs of phenotype and plasticity. *Heredity* in press, doi:10.1038/hdy.2015.8

Narum SR, Buerkle CA, Davey JW, Miller MR, Hohenlohe PA (2013) Genotyping-by-sequencing in ecological and conservation genomics. *Molecular Ecology* **22**, 2841-2847.

Neff BD, Pitcher TE, Ramnarine IW (2008) Inter-population variation in multiple paternity and reproductive skew in the guppy. *Molecular Ecology* **17**,2975–2984.

Oscoz J, Miranda R, Leunda P (2008) Additional records of eastern mosquitofish *Gambusia holbrooki* (Girard, 1859) for the River Ebro basin (Spain). *Aquatic Invasions* **3**, 108–112.

Parker KM, Hughes K, Kim TJ, Hedrick PW (1998) Isolation and characterization of microsatellite loci from the Gila topminnow (*Poeciliopsis occidentalis*) and their utility in guppies (*Poecilia reticulata*). *Molecular Ecology* **7**, 361-363.

Perez-Figueroa A, Garcia-Pereira MJ, Saura M, Rolan-Alvarez E, Caballero A (2010) Comparing three different methods to detect selective loci using dominant markers. *Journal of Evolutionary Biology* **23**, 2267–2276.

Pickrell JK, Pritchard JK (2012) Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genetics* **8**, e1002967

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959.

Pujolar JM, Jacobsen MW, Frydenberg J*, et al.* (2013) A resource of genome-wide single-nucleotide polymorphisms generated by RAD tag sequencing in the critically endangered European eel. *Molecular Ecology Resources* **13**, 706-714.

Purcell KM, Ling N, Stockwell CA (2012) Evaluation of the introduction history and genetic diversity of a serially introduced fish population in New Zealand. *Biological Invasions* 14, 2057–2065.

Purcell KM, Lance SL, Jones KL, Stockwell CA (2011) Ten novel microsatellite markers for the western mosquitofish *Gambusia affinis*. *Conservation Genetics Resources* **3**, 361-363.

Rehage JS, Sih A (2004) Dispersal behavior, boldness, and the link to invasiveness: a comparison of four Gambusia species. *Biological Invasions* **6**, 379–391.

Reznick D, Ghalambor C (2001) The population ecology of contemporary adaptations: What empirical studies reveal about the conditions that promote adaptive evolution. *Genetica* **112-113**, 183– 198.

Ribeiro F, Elvira B, Collares-Pereira M (2008) Life-history traits of non-native fishes in Iberian watersheds across several invasion stages: a first approach. *Biological Invasions* **10**, 89-102.

Richards PM, Liu MM, Lowe N*, et al.* (2013) RAD-Seq derived markers flank the shell colour and banding loci of the *Cepaea nemoralis* supergene. *Molecular Ecology* **22**, 3077-3089.

Rincón PA, Correas AM, Morcillo F, Risueño P, Lobón-Cerviá J (2002) Interaction between the introduced eastern mosquitofish and two autochthonous Spanish toothcarps. *Journal of Fish Biology* **61**, 1560–1585.

Sanz N, Araguas RM, Vidal O*, et al.* (2013) Genetic characterization of the invasive mosquitofish (*Gambusia* spp.) introduced to Europe: population structure and colonization routes. *Biological Invasions* **15**, 2333-2346.

Scribner KT, Avise JC (1993) Cytonuclear genetic architecture in mosquitofish populations and the possible roles of introgressive hybridization. *Molecular Ecology* **2**, 139-149.

Spencer CC, Chlan CA, Neigel JE*, et al.* (1999) Polymorphic microsatellite markers in the western mosquitofish, *Gambusia affinis*. *Molecular Ecology* **8**, 157-158.

Stockwell CA, Henkanaththegedara S (2011) Evolutionary conservation biology. In Ecology Evolutionary Poeciliid Fishes, chapter 12, page 424. University of Chicago Press, Chicago.

Tatara CP, Mulvey M, Newman MC (2002) Genetic and demographic responses of mercury-exposed mosquitofish (*Gambusia holbrooki*) populations: temporal stability and reproductive components of fitness. *Environmental Toxicology and Chemistry* **21**, 2191–2197.

Van Oosterhout C, Joyce DA, Cummings SM, *et al*. (2006) Balancing selection, random genetic drift, and genetic variation at the Major Histocompatibility Complex in two wild populations of guppies (*Poecilia reticulata*). *Evolution* **60**, 2562-2574.

Vera M, Alvarez-Dios J-A, Fernandez C*, et al.* (2013) Development and validation of Single Nucleotide Polymorphisms (SNPs) markers from two transcriptome 454-runs of turbot (*Scophthalmus maximus*) using high-throughput genotyping. *International journal of molecular sciences* **14**, 5694-5711.

Vera M, Alvarez-Dios JA, Millan A*, et al.* (2011) Validation of single nucleotide polymorphism (SNP) markers from an immune Expressed Sequence Tag (EST) turbot, *Scophthalmus maximus*, database. *Aquaculture* **313**, 31-41.

Vidal O, Garcia-Berthou E, Tedesco PA, Garcia-Marin J-L (2010) Origin and genetic diversity of mosquitofish (*Gambusia holbrooki*) introduced to Europe. *Biological Invasions* **12**, 841-851.

Vidal O, Sanz N, Araguas RM*, et al.* (2012) SNP diversity in introduced populations of the invasive *Gambusia holbrooki*. *Ecology of Freshwater Fish* **21**, 100-108.

Wang XQ, Zhao L, Eaton DAR, Li DZ, Guo ZH (2013) Identification of SNP markers for inferring phylogeny in temperate bamboos (Poaceae: Bambusoideae) using RAD sequencing. *Molecular Ecology Resources* **13**, 938-945.

Wooten MC, Scribner KT, Smith MH (1988) Genetic variability and systematics of *Gambusia* in the southeastern United States. *Copeia*, 283-289.

WWF (2014) Living Planet Report: species and spaces, people and places. Ed. McLellan R, Iyengar L, Jeffries B, Oerlemans N. WWF, Gland, Switzerland, 178 pp.

Zane L, Nelson WS, Jones AG, Avise JC (1999) Microsatellite assessment of multiple paternity in natural populations of a live bearing fish, *Gambusia holbrooki*. *Journal of Evolutionary Biology* **12**, 61-69.

Data Accessibility

GENEPOP input files, BayeScan results and BLASTN output file have been deposited in Dryad (Number identification: 10.5061/dryad.j059v). Filtered consensus RADtag sequences containing only one SNP are supplied on the Supporting Information (Supplementary Table 1). BAM files for all the individuals included in the database have been deposited in NCBI-SRA repository (number: SRP066107; Bioproject: PRJNA301984; Biosamples: SAMN04262351 – SAMN04262443).

Author Contributions Box

JLGM and MV designed the study. MV, DDM and JLGM analysed data. MV wrote the paper with the contributions of JLGM and DDM. JLGM funded the reagents.

Figure Legends

Figure 1. Geographical distribution of the sampling sites analysed. Location codes are indicated on Table 1.

Figure 2. Bayesian analyses of population structure carried out with STRUCTURE. Each vertical bar represents one individual, and the colour proportion for each bar represents the posterior probability of assignment of each individual to the different clusters (K) inferred by the program.

Figure 3. Maximum-likelihood tree generated by TREEMIX. The three migration events are depicted in heatmap colored arrows. EV location was used as root (see text)

Figure 4. Posterior odds of the selection model (Log10(PO)) and locus-specific FST for each SNP in the whole dataset according to the BAYESCAN results. Dotted line indicates the choosen threshold of significance corresponding to a probability $P \geq 0.99$ of being under selection. Loci above that threshold are indicated in red if belong to the anonymous database, and in green if associated to genes.

Supporting information

Supplementary Table 1.Information for the RADtags containing one SNP included on the *Gambusia holbrooki* genomic database. Identification number on the database (Catalog ID), corresponding RADtag for the individual VM43 used as framework for the construction of RADtags (VM43 RADid), RADtag consensus sequence, number of individuals presenting the RADtag (N), SNP position inside RADtag and its variants are showed. RADtags in green and red represent markers matching against genes in *X. maculatus* and containing repeat motifs, respectively.

Supplementary Table 2. Outliers loci detected with the program BAYESCAN. Only RADtags with a posterior probability (P) > 0.99000 and a False Discovery Rate (FDR) < 0.01 are showed. RADtags associated to X. maculatus genes are showed in green.

Supplementary Table 3. Gene Ontology (GO) terms of level 4 for the three GO categories provided by DAVID program (a maximum of three for each category).

Supplementary Figure 1. Number of SNPs per nucleotide position inside the RADtags. The first five nucleotide positions correspond with the restriction enzyme target.

925 Table 1. Sampling sites of the present study including the number of individuals analysed (N)

Sampling Site	Code	River	Country	Geographical Coordinates	N
Everglades	EV	Everglades	USA	25º 26’ N 80º 46’ W	12
Florence	FLO	Florence	USA	36º 01’ N 79º 57’ W	12
Potomac	PO	Potomac	USA	38º 39’ N 77º 11’ W	12
Besalú	BL	Fluvià	Spain (Europe)	42º 11' N 02º 44' E	12
Sant Pere Pescador	PP	Fluvià	Spain (Europe)	42º 11' N 03º 04' E	12
Vilanova de Muga	VM	Muga	Spain (Europe)	42º 17' N 03º 02' E	12
Empúria-Brava	EP	Muga	Spain (Europe)	42º 14' N 03º 07' E	12
Coltano	CO	Coltano	Italy (Europe)	43º 38' N 10º 24' E	12

926

927

928 Table 2. Diversity levels of the studied sampling sites. Number of individuals analysed (N), Gene diversity (H_E), mean number of alleles per
929 locus (N alleles) and their standard deviations are included. Location codes are showed on Table 1. Results for all RADtags (Full dataset) and for
930 the subset of RADtags matching against the *X. maculatus* transcriptome (XMT subset) are included.

Variable	EV	FLO	PO	BL	PP	VM	EP	CO	America	Europe
N	11	11	11	12	12	12	12	12	33	60
Full dataset										
Usable Loci	7614	7609	7621	7617	7619	7619	7619	7620	7604	7621
Polymorphic Loci	4781	3232	1760	716	975	884	947	650	7395	1191
% Polymorphic Loci	62.79	42.48	23.09	9.40	12.80	11.60	12.43	8.53	97.25	15.63
H_E	0.135	0.116	0.060	0.034	0.041	0.039	0.042	0.033	0.149	0.047
SD	0.154	0.168	0.134	0.116	0.124	0.119	0.124	0.125	0.155	0.127
N alleles	1.629	1.424	1.231	1.094	1.128	1.116	1.125	1.085	1.976	1.156
SD	0.490	0.499	0.422	0.293	0.334	0.321	0.331	0.280	0.174	0.364
MXT subset										
Usable Loci	184	184	185	185	185	185	185	184	184	185
Polymorphic Loci	107	85	44	14	24	21	22	16	181	27
% Polymorphic Loci	58.15	46.20	23.78	7.57	12.97	11.35	11.89	8.70	98.37	14.59
H_E	0.124	0.118	0.063	0.033	0.043	0.044	0.042	0.034	0.146	0.048
SD	0.154	0.168	0.137	0.117	0.125	0.130	0.125	0.115	0.153	0.133
N alleles	1.573	1.465	1.238	1.076	1.130	1.114	1.119	1.092	1.984	1.146
SD	0.507	0.500	0.427	0.265	0.337	0.318	0.325	0.290	0.127	0.354

931

932

933

934 Table 3. Pairwise FST values among the studied locations.

	EV	FLO	PO	BL	PP	VM	EP	CO
EV	0.000							
FLO	0.259	0.000						
PO	0.475	0.436	0.000					
BL	0.537	0.508	0.258	0.000				
PP	0.528	0.499	0.220	0.205	0.000			
VM	0.536	0.507	0.234	0.255	0.145	0.000		
EP	0.527	0.495	0.206	0.225	0.045	0.065	0.000	
CO	0.547	0.520	0.264	0.337	0.307	0.315	0.292	0.000

935

936 All P-values < 0.05

937

938 Table 4. AMOVA results applying different hierarchical population structure models. Differences among populations (F_{ST}), among populations
939 within group (F_{SC}) and among groups (F_{CT}) and their percentage (%) of explained variance (% Var) are shown.

Model	F_{ST}	F_{SC}	% Var	F_{CT}	% Var
America vs Europe					
Full dataset	0.484	0.318	24.05	0.243	24.30
	P < 0.001	P < 0.001		P = 0.019	
MXT subset	0.509	0.314	22.50	0.284	28.38
	P < 0.001	P < 0.001		P = 0.019	
America-Spain-Italy					
Full dataset	0.450	0.330	27.06	0.179	17.93
	P < 0.001	P < 0.001		P = 0.095	
MXT subset	0.474	0.320	24.81	0.226	22.55
	P < 0.001	P < 0.001		P = 0.010	
EV+FLO - PO+Europe					
Full dataset	0.571	0.217	11.90	0.452	45.18
	P < 0.001	P < 0.001		P = 0.028	
MXT subset	0.606	0.177	8.47	0.521	52.11
	P < 0.001	P < 0.001		P = 0.037	
EV+PO - FLO+Europe					
Full dataset	0.439	0.398	37.11	0.068	6.82
	P < 0.001	P < 0.001		P = 0.059	
MXT subset	0.442	0.424	41.14	0.031	3.09
	P < 0.001	P < 0.001		P = 0.074	
PO+FLO - EV+Europe					
Full dataset	0.423	0.411	40.19	0.021	2.14
	P < 0.001	P < 0.001		P = 0.159	
MXT subset	0.436	0.429	42.39	0.012	1.23
	P < 0.001	P < 0.001		P = 0.143	

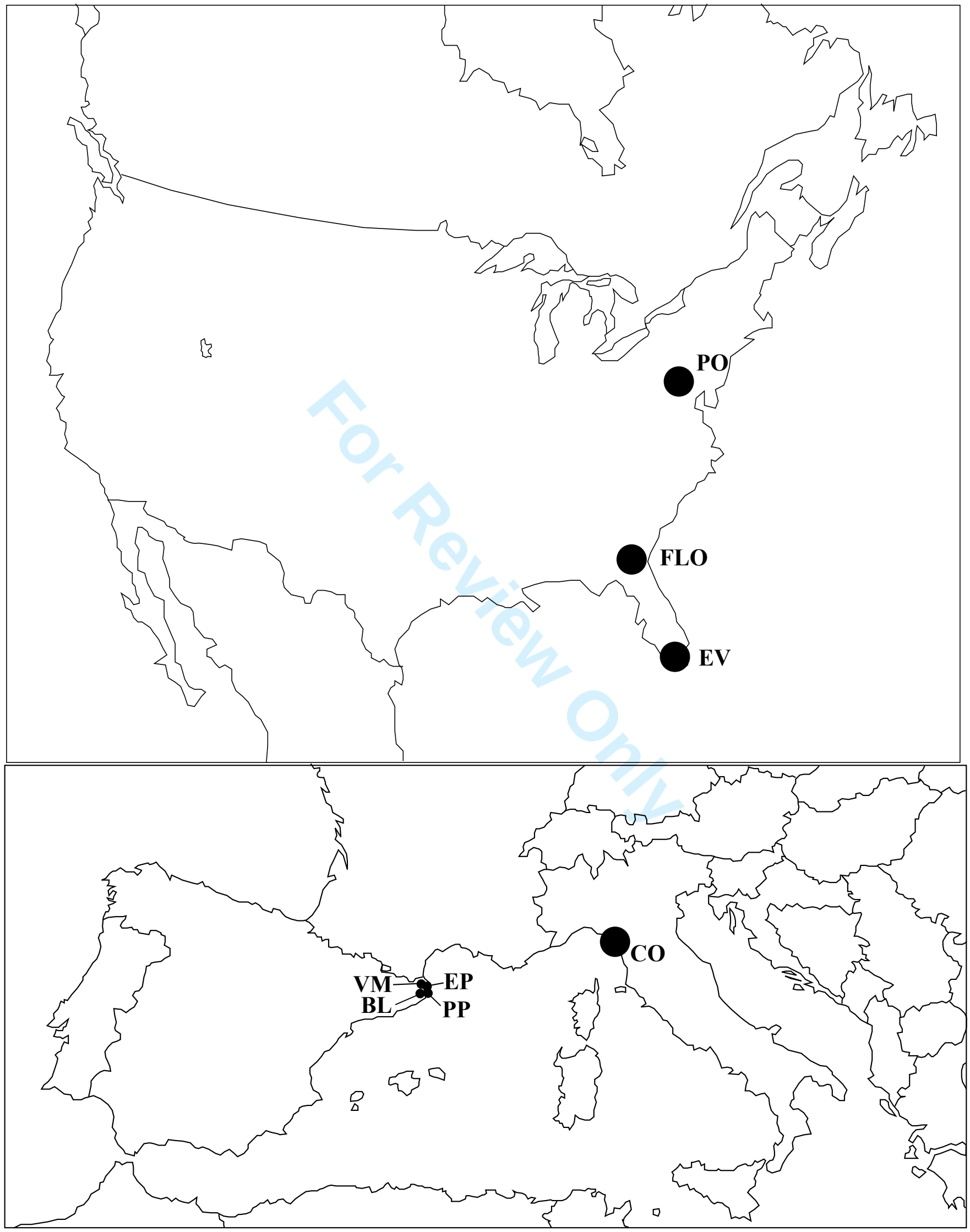
940

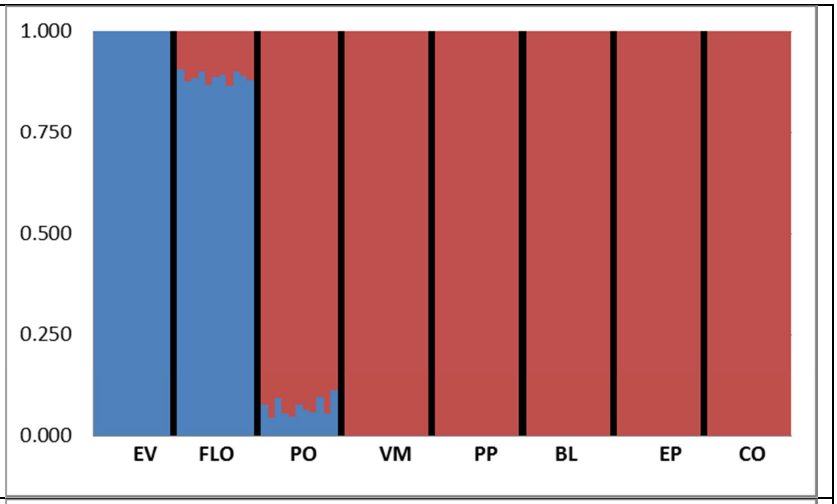
941

Table 5. Annotated outlier loci detected in the studied locations. The posterior probability (P), False Discovery Rate (FDR), the locus specific component shared by all the populations (alpha value; a positive value of alpha suggests diversifying selection, whereas negative values suggest balancing or purifying selection), matching sequence over *X. maculatus* transcriptome, annotation, Gene Approved Symbol, the two SNP variants, change effect and the interval of the allele frequency for one of the variants (Allele Freq Interval) are shown.

Comparison	Catalog_ID	P	qval (FDR)	alpha	F_{ST}	*X. maculatus* GenBank	Annotation	Approved Symbol	SNP Variants	Change effect	Allele Freq Interval
All Locations	11446	0.999	0.0005	19.994	0.897	XM_005816281.1	cytoplasmic FMR1-interacting protein 2	CYFIP2	G>A	NonSyn (V/M)	G= 0.000/1.000*
	13210	1.000	0.0000	-26.589	0.184	XM_005816270.1	ubiquitin thioesterase ZRANB1	ZRANB1	C>A	NonSyn (A/S)	C= 0.500-0.682
	36169	0.999	0.0001	-22.211	0.244	XM_005806231.1	brefeldin A-inhibited guanine nucleotide-exchange protein 2	ARFGEF2	T>C	Syn (L/L)	T= 0.500-0.875
	37928	1.000	0.0000	-28.351	0.162	XM_005812435.1	uncharacterized/unknown		G>A		G= 0.500-0.625
European Locations	1078	1.000	0.0000	-2.2170	0.405	XM_005812632.1	retinoic acid receptor RXR-beta-A	RARB	C>T	Syn (N/N)	C= 0.292-0.750
	2804	0.999	0.0002	-2.3534	0.378	XM_005800147.1	E3 ubiquitin-protein ligase UBR4	UBR4	C>T	Syn (N/N)	C= 0.182-0.542
	8684	0.999	0.0003	-2.4497	0.360	XM_005800006.1	ankyrin repeat and BTB/POZ domain-containing protein BTBD11-A	ABTB1	T>C	Syn (D/D)	T= 0.708-0.833
	13210	1.000	0.0000	-2.5361	0.342	XM_005816270.1	ubiquitin thioesterase ZRANB1	ZRANB1	C>A	NonSyn (A/S)	C= 0.625-0.682
	17945	1.000	0.0000	-2.3183	0.384	XM_005811728.1	MAGUK p55 subfamily member 2	MPP2	T>C	NonSyn (E/G)	T= 0.227-0.727
	36169	0.999	0.0003	-2.3787	0.375	XM_005806231.1	brefeldin A-inhibited guanine nucleotide-exchange protein 2	ARFGEF2	T>C	Syn (L/L)	T= 0.750-0.875
	37928	1.000	0.0000	-2.5656	0.336	XM_005812435.1	uncharacterized/unknown		G>A		G= 0.500-0.625
Spanish Locations	1078	0.996	0.0029	-2.0452	0.447	XM_005812632.1	retinoic acid receptor RXR-beta-A	RARB	C>T	Syn (N/N)	C= 0.417-0.750
	2804	0.994	0.0036	-2.1375	0.428	XM_005800147.1	E3 ubiquitin-protein ligase UBR4	UBR4	C>T	Syn (N/N)	C= 0.182-0.375
	3379	0.996	0.0027	-2.1121	0.433	XM_005803525.1	fasciculation and elongation protein zeta-1	FEZ1	C>T	Syn (K/K)	C= 0.208-0.591
	8684	0.992	0.0042	-2.1101	0.434	XM_005800006.1	ankyrin repeat and BTB/POZ domain-containing protein BTBD11-A	ABTB1	T>C	Syn (D/D)	T= 0.750-0.833
	13210	0.996	0.0028	-2.2151	0.412	XM_005816270.1	ubiquitin thioesterase ZRANB1	ZRANB1	C>A	NonSyn (A/S)	C= 0.625-0.682
	17945	0.997	0.0026	-2.0296	0.450	XM_005811728.1	MAGUK p55 subfamily member 2	MPP2	T>C	NonSyn (E/G)	T= 0.227-0.727
	24342	0.995	0.0033	-1.9499	0.468	XM_005798108.1	sodium/hydrogen exchanger 2	SLC9A3R2	T>C	NonSyn (I/T)	T= 0.458-0.875
	29072	0.993	0.0040	-1.9583	0.466	XM_005811982.1	WD repeat-containing protein 20	WDR20	C>T	Syn (S/S)	C=0.167-0.708
	36169	0.991	0.0045	-2.0405	0.450	XM_005806231.1	brefeldin A-inhibited guanine nucleotide-exchange protein 2	ARFGEF2	T>C	Syn (L/L)	T= 0.750-0.875
	37928	0.998	0.0017	-2.2478	0.404	XM_005812435.1	uncharacterized/unknown		G>A		G= 0.500-0.625

* G variant fixed for EV and FLO and A variant fixed for PO and all European sample



K=2	
K=3