The influence of native populations' genetic history on the reconstruction of invasion routes: the case of a highly invasive aquatic species

Thomas Brazier, Emira Cherif, Jean-François Martin, André Gilles, Simon Blanchet, Yahui Zhao, Marine Combe, R. J. S. Mccairns, Rodolphe E. Gozlan

The influence of native populations' genetic history on the reconstruction of invasion routes: The case of a highly invasive aquatic species

Brazier Thomas[1,2], Cherif Emira[3*], Martin Jean-François[4], Gilles André[5], Blanchet Simon[6], Zhao Yahui[7], Combe Marine[3], McCairns R.J.Scott[1] & Gozlan Rodolphe.E[3]

[1] DECOD (Ecosystem Dynamics and Sustainability), INRAE, Institut Agro, IFREMER, Rennes, France

[2] University of Rennes, CNRS, ECOBIO (Ecosystems, Biodiversity, Evolution) - UMR 6553, Rennes, France

[3] ISEM, Univ Montpellier, CNRS, IRD, Montpellier, France

[4] CBGP, Montpellier SupAgro, INRA, CIRAD, IRD, Univ. Montpellier, Montpellier, France

[5] UMR 1467 RECOVER, Aix Marseille Univ., INRAE, Centre St Charles, 3 place Victor Hugo, 13331, Marseille, France

[6] CNRS, Station d'Ecologie Théorique et Expérimentale (SETE), Moulis, France

[7] Institute of Zoology, Chinese Academy of Sciences, Chaoyang District, Beijing 100101, China

Brazier Thomas and Cherif Emira should be considered joint first authors.

[*]Correspondence

Emira Cherif, ISEM, Univ Montpellier, CNRS, IRD, Montpellier, France

Email: emira.cherif@ird.fr ORCID: 0000-0001-9365-7500

Thomas Brazier ORCID: 0000-0001-5990-7545

Martin Jean-François ORCID: 0000-0001-9176-4476

Blanchet Simon ORCID: 0000-0002-3843-589X

Zhao Yahui ORCID: 0000-0002-4615-596X

Combe Marine ORCID: 0000-0002-0708-9234

McCairns R.J.Scott ORCID: 0000-0002-0392-7413

Gozlan Rodolphe.E ORCID: 0000-0003-1773-3545

Abstract

Insufficient data on the origins of the first introduced propagule and the initial stages of invasion complicate the reconstruction of a species' invasion history. Phylogeography of the native area profoundly shapes the genomic patterns of the propagules on which subsequent demographic processes of the invasion are based. Thus, a better understanding of this aspect helps to disentangle native and invasive histories. Here, we used genomic data together with clustering methods, explicit admixture tests combined with ABC models and Machine Learning algorithms, to compare patterns of genetic structure and gene flow of native and introduced populations, and infer the most likely invasion pathways of the highly invasive freshwater fish *Pseudorasbora parva*. This species is the vector of a novel lethal fungal-like pathogen (*Sphaerothecum destruens*) that is responsible for the decline of several fish species in Europe. We found that the current genetic structuring in the native range of *P. parva* has been shaped by waves of gene flow from populations in southern and northern China. Furthermore, our results strongly suggest that the genetic diversity of invasive populations results from recurrent global invasion pathways of admixed native populations. Our study also illustrates how the combination of admixture tests, ABC and Machine Learning can be used to detect high-resolution demographic signatures and reconstruct an integrative biological invasion history.

KEYWORDS

Invasion pathways, aquatic biological invasion, demographic inference, population modeling, population genomics, Approximate Bayesian Computation.

Declarations

Funding

This study is part of the project GENESIS ANR-AF 13-ADAP-0005-001 supported by the Agence Nationale de la Recherche. S.B. was co-funded by the project PROBIS (Biodiversa) and E.C. was funded by an IRD postdoctoral fellowship. R.E.G., A.G., J.F.M. and S.B. received funding for this research from ANR Bioadapt.

Conflicts of interest

The authors declare that they have no conflict of interest.

Availability of data and material

All sequencing data (GBS reads) are publicly-available under the BioProject ID PRJNA799666. The original genetic data yielded by genotyping (in vcf format), additional metadata (e.g. geographical coordinates) and all scripts necessary to reproduce the results presented in the paper are available in a publicly-available Github repository: https://github.com/ThomasBrazier/PseudorasboraGENESIS. Data are available from the authors upon reasonable request**.**

Authors' contributions

T.B., E.C. performed research, analysed the data. T.B., E.C. and R.E.G. wrote the manuscript; R.E.G., R.J.S.McC., J.F.M., A.G. and S.B. designed research and contributed with data; Y.Z., R.E.G., M.C. contributed with data; R.J.S.McC. assisted with analyses. All authors commented on the manuscript.

1 INTRODUCTION

Invasive species are significant contributors to global change and often lead to biotic homogenization and biodiversity losses (Chapin Iii et al. 2000; Clavero & García-Berthou 2005; Didham et al. 2005; Villéger et al. 2011; Simberloff 2013). Globalization of trade and growth of worldwide transportation are the main drivers of non-native species introductions (Hulme 2009). Major ecological impacts from biological invasions include destabilization of trophic networks (Stiers et al. 2011; Gallardo et al. 2016), competition for resources and habitats (Graebner et al. 2012; Perdereau et al. 2011), predation on native species (Salo et al. 2007) and the transmission of novel infectious pathogens (Crowl et al. 2008). Invasions can also have evolutionary consequences, such as genome introgression or altering selection pressures on native species (Crispo et al. 2011; Mooney & Cleland 2001; Sinama et al. 2013; Philips et al. 2006). Yet invasions offer an incredible framework to study adaptation to new environmental conditions and to understand how small introduced populations with a supposedly low level of genetic diversity manage to colonize large areas (Bossdorf et al. 2005; Peischl & Excoffier 2015; Sax et al. 2007; Roman & Darling 2007; Facon et al. 2006), as well as to understand the spread and emergence of pathogens co-introduced together with their non-native hosts. Reconstructing the invasion history of a species is challenging. Typically, few individuals may constitute the initial colonizing group, but data on the amount of propagule pressure as well as its origin and time of first introduction are often lacking. Historical records of introduction events, census data or environmental monitoring projects often miss the first stages of invasion (Holsbeek et al. 2008; Mergeay et al. 2005). Molecular ecology allows the reconstruction of events that have not been directly observed but have left a genomic signature (Cristescu 2015). Through the use of new demographic inference methods that permit deviations from the assumptions of the mutation-drift equilibrium model (which is often the case for recent colonization events), it is now possible to identify source populations or even to model the short-term demo-genetic processes underpinning invasion history (Beichman et al. 2018; Cabrera & Palsbøll 2017; Shafer et al. 2015; Estoup & Guillemaud 2010). Recent population genetic studies on

97 invasion pathways have shown that invasion patterns may be more complex than previously
98 thought (Lombaert et al. 2014). A single introduction event is rare, and multiple introductions
99 and/or admixture between source populations are often the cornerstone of successful
100 invasions (Estoup et al. 2016; Roman & Darling 2007). Since introduced populations inherit
101 genetic variation primarily shaped by long-term evolutionary processes in the native area,
102 incorporating this information in the interpretation of the results is important to avoid
103 erroneous conclusions (Rius & Turon 2020).
104 The topmouth gudgeon, *Pseudorasbora parva* (Temminck & Schlegel), is a small freshwater
105 cyprinid fish. More importantly, it is an interesting study system for biological invasion for
106 several reasons: i) it is a highly successful invader that has spread and established very
107 quickly in diverse environments throughout the world, ii) it has a well-documented history
108 and iii) its invasion is associated with a pathogen co-introduction. It also has a very large
109 native distribution in Eastern Asia (Eastern China, Taiwan, Korea and Japan) with broad
110 environmental tolerance from continental climates to tropical ones (Gozlan 2012; Zhang &
111 Zhao 2016). It was initially accidentally released outside its native range via aquaculture
112 partnership exchanges of Chinese carp between China and former countries of the Soviet
113 Bloc (Gozlan et al. 2010). During the 1960s, multiple introductions of *P. parva* took place all
114 around the Black Sea area, followed by further introductions in the 80s in Eurasia and North
115 Africa. However, introduction records lack precise information on the source populations.
116 After these initial phases of human-made introductions, natural local colonization of entire
117 river networks occurred across major European rivers to the Middle East. Its life-history
118 traits, which include early maturity (1 year) coupled with nest guarding behavior that ensures
119 an increased likelihood of survival for juveniles and short longevity, have been identified as
120 key traits to explain its rapid establishment and spread (Gozlan et al. 2010; Gozlan 2012,
121 Gozlan et al. 2020), although its long-term invasion success depends on high genetic
122 diversity to drive adaptation to the new environment (Roman & Darling 2007). A major
123 biodiversity problem associated with *P. parva* is the fungus-like pathogen, the rosette agent
124 *Sphaerothecum destruens*, for which it acts as a vector, and that has subsequently spread

to invaded areas causing mortality in a large number of native freshwater fish species in Europe (Andreou & Gozlan 2016; Combe & Gozlan 2018).

Genetic structure and phylogeography, based upon classical mitochondrial and nuclear markers (microsatellites), have been extensively described in both native and invasive *P. parva* populations, suggesting the existence of two genetic lineages within non-native populations in Europe (Hardouin et al. 2018; Simon et al. 2011, 2015). This general pattern has been recently confirmed by a study using 13,785 single nucleotide polymorphisms in Slovakian and Turkish introduced populations (Baltazar-Soares et al. 2020). Yet, the fine-scale genetic structure of the native populations and the gene flow that has shaped it, as well as the demographic dynamics underpinning the invasion pathways, remain unresolved. Assessing the source populations within the native range by extensive sampling, along with high-throughput genotyping, is crucial to detect high-resolution signatures of demographic history (Muirhead et al. 2008; Beichman et al. 2018; Shafer et al. 2015). Although such large datasets require advanced statistical methods that are computationally demanding, Approximate Bayesian Computation (ABC) and associated Machine Learning algorithms take advantage of large sets of summary statistics, exploring huge parameter spaces with reduced computational effort (Beaumont et al. 2002; Cabrera & Palsbøll 2017; Pudlo et al. 2016; Raynal et al. 2017; Rey et al. 2015).

Here we took advantage of genomic signatures, ABC models and Machine Learning algorithms to i) characterize the genetic structure of native populations and test whether they aggregate on homogeneous and coherent demes and ii) retrace *P. parva*'s introduction history from Asia to Europe. To do so, we first applied genetic clustering methods and explicit tests of admixture to describe the phylogeography of both native and invasive ranges. We then identified putative source populations of introduced demes with population assignment tests. We finally used ABC model-based procedures to infer the most probable demographic scenarios of *P. parva* invasion and reconstruct an integrative biological invasion history.

2 MATERIALS AND METHODS

153 **2.1 Sampling material**

154 We sampled *P. parva* from 21 discrete sites across its overall distribution in Asia, comprising
155 sixteen different river catchments in the historical Chinese native range (Figure S1, Table 1,
156 sites numbered from S1 to S18). Additional samples were obtained from Japan and
157 neighboring Asian invasive populations in Tibet and South-East China (sites: S19, S20). We
158 also sampled *P. parva* across thirteen locations within the invasive European and Middle-
159 Eastern range (Figure S1, Table 1). The sampling methods included fish traps, electric
160 fishing and micro-mesh seine netting where appropriate. In the field, the fish were
161 euthanized with an overdose of anesthetic, initially preserved in ethanol and subsequently
162 stored at -70 °C. Each fish was measured and a fin-clip taken and stored in ethanol.

163 **2.2 GBS sequencing**

164 In total, 858 DNA samples (746 individuals in total, with 75 individuals replicated across at
165 least one sequencing lane) were genotyped for single nucleotide polymorphism (SNP)
166 markers by first digesting genomic DNA with PstI, followed by genotyping-by-sequencing
167 (GBS), yielding an average of 2,702,000 raw sequencing reads per sample. SNP calling was
168 performed using programs comprising the Stacks (v1.46) bioinformatics *de novo* pipeline
169 (Catchen et al. 2013). Replicated individuals were used to estimating genotyping error and
170 only markers with less than 1% errors were retained. A minimum read depth of 20 was
171 required for each marker, ultimately yielding 3999 validated SNP markers. To reduce linkage
172 disequilibrium amongst markers, only a single SNP (first position) was called for each locus
173 stack. To prevent biases due to large proportions of missing data, populations with more
174 than 70% missing data, individuals with more than 60% missing data and loci with more than
175 45% of missing data were removed from the dataset. After trimming, the final dataset
176 contained 300 individuals from eighteen sites in the historic Asian range (including invasive
177 populations in Tibet and South-West China) and 168 individuals from eleven invasive sites
178 across Europe, Turkey and Iran, with on average 16 individuals per sampled site (Figure S1,
179 Table 1).

180 **2.3 Genetic diversity**

181 Statistics of genetic diversity were estimated with the ‘adegenet’ and ‘hierfstat’ R packages
182 (Goudet 2005; Goudet & Jombart 2015; Jombart 2008; Jombart & Ahmed 2011) for R
183 version 3.5.3 (R Core Team, 2019) (Table S1). Sensitivity analyses on summary statistics
184 showed that biases were minimized without loss of power when estimates were inferred from
185 2,000 to 3,000 loci with the lowest proportion of missing data (Figure S2). Consequently, the
186 dataset for inferring population assignment consisted of 2,112 SNPs to reduce
187 computational load (maximum 45% missing data per locus); however, up to 3000 SNPs
188 (maximum ~50% missing data) were retained for other analyses to keep most of the genetic
189 information. Simulations have shown that even loci with high levels of missing data (> 50%)
190 can retain meaningful information (Huang & Knowles 2016), as it has been empirically
191 assessed for population structure (Chan et al. 2017).

192 **2.4 Genetic clustering**

193 Genetic clustering of sampled sites was assessed independently in the Asian range (native
194 and non-native sites) and within the invasive European range. Putative demes were defined
195 as groups of individuals sharing a gene pool. Results from an iterative K-means method and
196 a model-based Bayesian clustering method were compared for cross-validation.
197 Discriminant Analysis of Principal Components (DAPC), implemented in the R package
198 ‘adegenet’ (Jombart 2008; Jombart & Ahmed 2011) was first used to determine genetic
199 clusters within native and invasive regions and then to predict membership of individuals
200 within invasive sites to native clusters. The most probable number of genetic clusters, K, was
201 searched within the distribution of K among 1,000 independent clustering iterations, based
202 on the ‘goodfit’ and ‘min’ criteria of the Bayesian Information Criterion (BIC). Overfitting was
203 prevented by a cross-validation procedure to define the optimal number of Principal
204 Components required to discriminate amongst these K clusters.

205 Additionally, we used STRUCTURE 2.3 to infer the number of genetic clusters and to
206 estimate admixture between them (Pritchard et al. 2000). The admixture model with
207 correlated allele frequencies was parameterized with a fixed Lambda value (parameter of the

allele frequencies distribution) directly estimated from the data. Sampling location was set as a prior to improve inferences on weak genetic structure. STRUCTURE was performed for K=1-21 (number of sampled sites + 3) for Asian sites and K=1-14 for European and Middle-Eastern sites. Twenty replicates were computed for each K value, with 100,000 sampling iterations after a burn-in of 100,000. The most probable K was assessed after considering the smallest value of K minimizing differences of likelihood (i.e. the plateau method of Pritchard et al. 2000), the highest value of ΔK given using Evanno's method, computed with STRUCTURE HARVESTER (Earl & vonHoldt 2012; Evanno et al. 2005) and the Puechmaille statistics (Puechmaille 2016), implemented in the STRUCTURE SELECTOR web interface (Li & Liu 2018). We then used CLUMPP to aggregate STRUCTURE replicates to produce mean individual admixture proportions with a 'greedy' search algorithm over 1,000 repetitions (Jakobsson & Rosenberg 2007). Convergence among parameter sampling chains was assessed via CLUMPP's H' statistic of similarity amongst several replicates. Additionally, within-chain convergence for parameter estimates (i.e. ancestry coefficient Alpha and Ln Likelihood) was assessed with diagnostics implemented in the R package 'coda' (Plummer et al. 2006).

Finally, based on the observed partitioning of genetic variance, sampled sites were pooled into demes (also referred to as populations) that made sense biologically (i.e. continuous gene flow among sites) and geographically. Admixture proportions inferred with STRUCTURE helped to delineate putative demes and were computed as the mean of the major ancestry coefficients (Q) in a given site. Sampled sites with a mean major ancestry coefficient (Q) under 70% were considered as admixed sites. Sampled sites that did not cluster well based on genetic markers were clustered into putative demes based on geographical and historical data. The consistency of our final clustering of sampled sites in genetic populations was assessed by a hierarchical AMOVA implemented in the 'poppr' R package (Kamvar et al. 2014, 2015) and the significance of variance proportions was tested with 1,000 random permutations (Table S3). Maps were drawn from the R package 'maps' with the 'world' database (Becker et al. 2018). The main river network was drawn from

'RNaturalEarth' (South 2017).

2.5 Assignment to source populations with supervised machine learning

The R package 'AssignPOP' was used to assign invasive individuals to candidate source populations (i.e. native demes) with a machine learning classification algorithm (Chen et al. 2018). The Support Vector Machine (SVM) classification algorithm was trained on a set of individuals from known populations (i.e. candidate source populations in the native range), and then individuals from unknown populations (i.e. invasive populations) were assigned to one of the candidate source populations in the training dataset. Using a Monte Carlo procedure with K-fold cross-validation, a set of known loci and individuals were randomly sampled in multiple iterations to train the classification algorithm, while the remaining known loci and individuals were used to iteratively test the accuracy of the predictive model. One hundred training iterations were performed for different proportions of training individuals (i.e. 0.5, 0.7 and 0.9) and training loci (i.e. 0.25, 0.5 and 1) to select the best training sample size. The training error rate was assessed as the reassignment success of known individuals to their source population. Source populations of invasive individuals were predicted from posterior assignment probabilities. Only individuals for whom the highest probability was at least twice that of the second were retained as confident assignments.

2.6 Inference of invasion history with ABC

Approximate Bayesian Computation (ABC) was used to infer past demographic events shaping contemporary genetic diversity (Supporting information). We simulated large datasets under various invasion and admixture demographic scenarios and estimated the probability that data were observed under a given demographic scenario (Estoup et al. 2012). Scenarios were designed in a two-step hierarchical procedure of increasing complexity, with the second step derived from findings at step 1 (Figure S3, S4; see supplementary method for details). In the first step, we tested 3 independent sets of scenarios about the origins in the native range of 3 independent invasive demes (Western Europe, Eastern Europe and Iran; see Results). For each independent invasive deme, we tested if the source population was one of the three candidates or an admixture between the

three candidates. To avoid the trap of infinite combinations of exhaustive scenarios, source populations predicted with population assignment served as initial candidate source populations. Choice of candidate source populations was also cross-validated with a Maximum-likelihood phylogenetic tree (Figure S8) inferred by the program TreeMix version 1.13 (Pickrell & Pritchard 2012). As the Italian sample's origin was ambiguous in previous genetic clustering results, we performed the ABC analysis with and without Italian samples. The comparison of both replicated scenarios allowed us to assess the sensitivity of the selected scenario to Italian individuals (Table 2). The second step was dedicated to resolving a more complex invasion pattern, comparing competing worldwide invasion pathways encompassing all invasive demes. Source populations of invasive demes were found to be admixed (results of step 1, described below). Therefore, plausible hypotheses to test at step 2 were: (1) three independent introductions from three independent admixed source populations; (2) three independent introductions from a single admixed native population; or (3) a single continental introduction from an admixed native population.

Demo-genetic scenarios were simulated and summary statistics estimated, with DIYABC version 2.1 (Cornuet et al. 2014). Prior probabilities of scenarios were set to uniform. Parameter prior distributions were first set to a biologically reasonable range of values, then confidence in priors checked (*via* DIYABC test of goodness of fit), with distributions refined iteratively until the simulated scenarios fit the data (final parameter spaces given in supplementary methods). For each scenario, 10,000 simulations were conducted, estimating all available summary statistics in DIYABC for each (Cornuet et al. 2014). The best model was selected using two machine-learning algorithms trained on the simulated datasets: a Neural Network algorithm implemented in the R package 'abc' (Csilléry et al. 2012) and a Random Forest algorithm in the 'abcrf' package (Pudlo et al. 2016). One thousand Neural Networks were trained with 5 to 12 units of hidden neural layers. Parameter sets were weighted by an Epanechnikov kernel. The tolerance rate was set to 0.2. Other configuration values were set to the default value. Power of ABC model selection with Neural Network was evaluated by leave-one-out cross-validation repeated 100 times. Independently, 1,000 trees

were grown in the Random Forest training set, with linear discriminant analysis scores added to summary statistics when it reduced the prior error rate. Power of Random Forest was evaluated by out-of-bag prior misclassification error rate. Lastly, the quality of the selected scenario was checked by comparing the marginal posterior predictive distributions to the observed values of the summary statistics. The marginal posterior predictive distribution was computed from 10,000 simulations under the selected scenario, parameterized with estimated posterior parameter distributions as priors.

The complexity of the selected scenario at step 2, combined with a restricted number of markers (3,000 SNPs), reduced the power to infer demographic parameters jointly with a reasonable confidence interval. Hence, demographic parameters of the selected scenarios were estimated at step 1 with a regression-based method adjusted by local linear regression (Blum & François 2010; Csilléry et al. 2012). Parameters were weighted by an Epanechnikov kernel. One million simulations were produced under the selected scenario to explore parameter space, but only simulations closest to the observed dataset were retained for parameter estimation (i.e. tolerance rate). Confidence in parameter estimates was checked with leave-one-out cross-validation repeated 1,000 times, thus estimating prediction errors.

2.7 Genotype phasing and imputation

The haplotype phase and missing data of the 468 sequenced individuals were inferred using Beagle 5.1 (Browning & Browning 2007; Browning et al. 2018) to perform population migration modeling and admixture tests. Beagle uses the localized haplotype-cluster model and applies an iterative approach to infer the most likely haplotype pair for each individual. At each iteration, phased input data are used to build a localized haplotype-cluster model. Once the model is built, phased haplotypes for each individual are sampled from the induced diploid HMM conditional on the individual's genotypes. The sampled haplotypes are the input for the next iteration, and so forth. In the final iteration, the Viterbi algorithm selects the most likely haplotypes for each individual, conditional on the diploid HMM and the individual's genotype data. For each copy of each individual, missing alleles are randomly imputed according to allele frequencies, and the data for each individual are phased by randomly

ordering the genotypes (Browning & Browning 2007).

2.8 Native population migration modeling

The modeling of the native population splits and mixtures was performed using TreeMix, a statistical model inferring the patterns of population splits and mixtures in multiple populations (Pickrell & Pritchard 2012). A Maximum-Likelihood tree was constructed using genome-wide allele frequency and genetic drift approximation. One thousand bootstraps were performed to assess the robustness of the inferred Maximum-Likelihood (ML) tree. Migration edges were then added sequentially to connect pairs of populations when allele frequency covariance excess was detected. For this analysis, we tested mainland China populations. To strengthen the migration model, we added the genetically close-related Tibet sample to the north-central China deme and considered Japan's population as an outgroup. We ran TreeMix with migration events ranging from zero to four and the TreeMix composite model incorporating known admixture with the '-cor_mig' option (Pickrell et al. 2012) based on STRUCTURE admixture results. The robustness of the tree and the migration edges were confirmed by 1,000 bootstraps using GNU Parallel (Tange 2011) and 'treemix.bootstrap' function implemented in the R package BITE (Milanesi et al. 2017).

2.9 Native population admixture test

We selected one deme from each mainland China region (North, North East, Central, South East and South) to perform admixture tests. We used the four-population test of the D-statistic (Green et al. 2010; Patterson et al. 2012) implemented in Popstats (Skoglund et al. 2015) to test for admixture and gene flow directionality within the native populations. The notation used by Popstats for the D-statistic is D(O, P3; P1, P2), where O is the outgroup, P3 the test population and P1, P2 the sister populations. A significant negative D indicates that P3 exchanged genes with P1; conversely, a positive D indicates that P3 exchanged genes with P2 (Durand et al. 2011). The D-statistic was estimated for each combination of demes. D-statistic significance was assessed by block jackknife of 5kb and the standard error (SE) was used to estimate the Z-score (Skoglund et al. 2015).

3 RESULTS

3.1 Complex genetic structure of the native range

We searched for the genetic structure of the native range in order to aggregate sampled sites into fewer homogeneous and consistent demes, hence simplifying further demographic inferences based on ABC modeling. The highest level of structure estimated by STRUCTURE was K=5-6 (Figure 1a, Table S3, Supplementary information). To select the most plausible number of genetic clusters K, multiple criteria were assessed and two clustering methods cross-validated (Table S3, Supplementary information). The modal value of the distribution of ΔK suggested K=3 in the native area, whereas the shape of the plateau of ln(Pr(X|K)) suggested a value of K between 5 and 7. At K=3, Northern China was well separated from Central China/Japan and Southern China. However, at K=6-7, a better resolution on Central China (S4, S6, S10, S11) and Tibet was obtained, with signals of a north-south admixture. The overall patterns of clustering and admixture were similar between K=6 and K=7 with a spurious cluster at K=7. Convergence tests assessed that most sampling chains were convergent, yet some of the sampling chains remained non-convergent even after a burn-in of 100,000 and 100,000 sampling iterations. The highest values of H' was for K=3 (0.99), though H' reached 0.81 for K=6. Moreover, the DAPC

approach yielded very similar results with a most plausible K=5. STRUCTURE and DAPC clustering were congruent in revealing the same separation between Northern and Southern China. The K that we finally selected was that meeting statistical, geographical and historical compromises. Hence, K=6 was chosen as the better compromise between statistical parsimony and the highest level of genetic structure.

3.2 Definition of native putative demes

The definition of native putative demes based on genetic clustering was crucial to building invasion route scenarios (Supporting information). Most putative demes in the native region were consistent (South China, North China, Japan), despite high uncertainty in some sites (Figure 2a, Figure S6). S3 in particular had a small sample size and was strongly admixed, causing high uncertainty for its assignment to a putative deme. As a consequence, S3 was removed from the dataset for any further analysis. Two admixed demes, composed of sites with a major ancestry coefficient lower than 70%, were created in Central China. These groupings were also justified on the basis of previous studies reporting this region as a zone of secondary contact between Northern and Southern populations (Hardouin et al. 2018; Simon et al. 2011). S11 was assigned at 65% to the North East China deme, but its geographic proximity and connectivity with S10 supported the constitution of an admixed Central China deme encompassing S10 and S11. Lastly, the admixed S13 site was placed into the North China deme because of its location within the same river basin (Figure 2a). While the Tibetan population is genetically representative of the Central China deme (ancestry coefficient > 75%), for historical reasons (i.e. recent Tibetan introduction of *P. parva* and few commercial exchanges with Europe), this population was not considered part of the deme. On the other hand, S19 and S20 were genetically well clustered with S9 and S18 (ancestry coefficient >99%) and were considered part of the South China deme. Six native demes were finally defined: North China, North-East China, Central China, South-East China, South China and Japan (Figure 2a).

3.3 Two main gene flow directions shaped the native populations

In the obtained consensus tree (Figure 1b), almost all the nodes were well supported (75%-100%). The Central-North East China node was supported by 47% of the bootstraps. The model with four migration events showed stable migration edges after multiple runs of TreeMix and TreeMix composite models (Figure 1b). The migration edges showed two main southern and northern origins of gene flow: gene flow from the South mainly to the Central and South East populations with 45% and 33% migration weights and gene flow from the North to the North East China population with 9% migration weight (Figure 1b).

Admixture tests confirmed gene flow origins followed a directional pattern. The four-population test assumed the population configuration D(O, P3; P1, P2), with Northern and Southern demes as our test populations (P3), the remaining demes as P1 and P2 and Japan as the outgroup O. D-statistics showed an excess of shared derived polymorphism, highlighting genes exchanged between i) the South population and the South East and Central ones and ii) between the North population and the Central and North East ones (Figure 1c, Table S4).

3.4 Genetic structure of the invasive range

Genetic clustering in the invasive range was used to narrow the putative number of non-native populations' origins to test (i.e. the number of scenarios). STRUCTURE results indicated a clear genetic structure in the European invasive range with K=3, supported by DAPC results, despite some uncertainty around the actual highest level of genetic clustering (Table S3). Indeed, for all values of K between 2 and 9, Turkey and Bulgaria clustered together without admixture; Iran formed a similarly distinct cluster (Figures S5a-S5b). Conversely, Western Europe showed a pattern of admixture for all values of K and admixture persisted unless K was equal to the number of sites (Figures S5b-S5c). This indicated that Western Europe is most likely a single deme with a strong sub-structure. Furthermore, hierarchical STRUCTURE analysis performed on the subset of West European sites confirmed a substructure at this scale. Within West European sites (Aus, Bel, Hun, Ita, Pol, Spa, UK), K=3-4 was the most probable grouping according to Evanno's method and

K=5 according to the plateau method (Figure S5c). K=5 was displayed for the Western European sub-structure (Figure S5c) because spurious admixture indicated overfitting for higher values (Figure S5b). H' was high (0.98) for K=3-4-5 in the STRUCTURE analysis of the Western European deme. Under Geweke's diagnostic and trace plots, sampling chains seemed convergent or close to convergence, yet some seemed to show departure from stationarity. Heidelberg and Welsh's diagnostic supported that most of the Markov Chains were a stationary distribution. Additionally, Gelman and Rubin's diagnostic assessed that replicated chains converged on similar values. Finally, despite the strong sub-structuring observed, the Western European deme was considered as a single consistent genetic population in ABC scenarios. Hence, we retained three independent demes in the invasive range formed by Western Europe, Eastern Europe and Iran.

3.5 Population assignment to source populations

'AssignPOP' training was efficient, with an assignment accuracy greater than 90% (Figure S7) and all putative demes confidently discriminated by the training algorithm. Predictions with the SVM algorithm confidently assigned 100 individuals (59%) to a source population, with a relative posterior probability >2. Assignment to source populations showed multiple origins in sampled sites, especially in Eastern Europe and Iran (Figure 2b), congruent with DAPC. Posterior membership probabilities of invasive individuals assigned to native Asian clusters with DAPC showed a putative origin in North East China for the Western Europe deme (cluster) and South/South East China with admixture for Bulgaria and Turkey. The Italian site showed a different origin than the rest of Western Europe, closer to Eastern Europe. The Iranian population was separated from all others, linked to South East China or Japan.

3.6 ABC inference of source populations and invasion pathways

ABC simulations were first used to infer the most probable source population in Asia of each one of the three main demes identified in the non-native area (step 1), followed by discrimination between different competing invasion pathways at a global scale (step 2).

Step 1: Source populations of non-native populations

Following prior calibration, all scenarios could be fitted to the data, with 112 summary statistics estimated to compare scenarios. Prior error rates were low for both Neural Network and Random Forest (Table 2), indicating a good predictive power. There was no confusion between scenarios, and the marginal posterior predictive distribution improved the goodness-of-fit of the selected model. Demographic inferences in Western Europe, Eastern Europe and Iran (the three main demes identified from the clustering approaches) all predicted introductions issued from admixture events (scenario 4, Figure S4), supported by high posterior probabilities (Table 2). The choice of candidate source populations was congruent with phylogenies inferred with Maximum Likelihood in Treemix (Figure S8). Each invasive population formed a group with an Asian native population, with relatively short distances between them for European demes. The Iranian population was particular, exhibiting long drift from the most recent common ancestor shared with Japan.

Step 2: Invasion pathways

As in step one, prior fitting to the data was achieved, although model selection was based on 256 estimated summary statistics. Prior error rates were also low (Table 2), giving confidence in subsequent inferences. Scenario 1, modelling three independent introductions from independent admixed populations leading to the three observed invasive demes, was selected with strong support from both Neural Network (99%) and Random Forest (86% of votes) algorithms (Table 2, Figure 2c).

The three different invasive populations clearly formed three distinct groups with admixed origins. Western Europe origins were in a Northern part of China (North-East China admixed with Central China and North China). Removal of the ambiguous Italian sample did not change the selected scenario. Eastern Europe origins were in a Southern part of China (admixture between Central China, South East China and South China). Lastly, Iran origins were an admixture between China (South East China, North East China) and Japan (Figure 2c).

3.7 Invasion process and founding populations

Parameters linked to invasion history (i.e. time of invasion and bottleneck severity of the founding population) were estimated with reasonable confidence intervals by local linear regression (tolerance rate of 0.05 for Eastern and Western Europe; 0.005 for Iran; Figure S9). The Iranian invasion was estimated to have occurred 66 generations ago ($CI_{95\%}$ = 44; 90), with an effective population size of 217 ($CI_{95\%}$ = 185; 253) for the founding group. The Eastern European invasion was estimated to be more recent at 39 generations ago ($CI_{95\%}$ = 18; 62) and with a smaller effective population size 90 ($CI_{95\%}$ = 32; 148). The effective population size of the Western European founding group required two million simulations to be estimated with a higher degree of confidence at 1,436 individuals ($CI_{95\%}$ = 1,001; 1,825); furthermore, the time of divergence was estimated at 40 generations ($CI_{95\%}$ = 20; 63).

4 DISCUSSION

Population genomics approaches used to inform biological invasions can be challenged by the discrepancy between the large timescale of genetic mutations and the smaller timescale of human-mediated invasions, leading to potentially flawed analyses (Fitzpatrick et al. 2012). Nevertheless, accurate assignments and estimates of geographic origin and number of introductions can be obtained through extensive sampling of native and invasive ranges, the use of high-resolution of molecular markers and a highly genetically structured source population (Excoffier & Heckel 2006; Dlugosch & Parker 2008). Our extensive sampling across both native and invasive areas, the genome-wide markers used in this analysis, as well as previous knowledge on native genetic structuring (Hardouin et al. 2018), allowed us to highlight long-term phylogeography and recurrent gene flow. In turn, this knowledge of complex genetic structure in the native range was an essential element allowing us to successfully model probable invasion pathways. Our results suggest that the genomic diversity of invaders was shaped long before introduction by the presence of geographical barriers and by human-mediated gene flow in the native range. The combination of appropriate and complementary methods for populations departing from H-W equilibrium and appropriate clustering within demes allowed us to infer

source populations and an invasion scenario despite high admixture within native demes, thus reconciling the genetic history of the native range with recent genomic patterns of invasion.

4.1 The Asian history of *P. parva* shaped by paleogeography and anthropization

Paleogeographic influence

Large-scale genetic structure was consistent with the known phylogeography of the species (Hardouin et al. 2018; Simon et al. 2011). Likewise, the retained fine-scale structure in the native range was consistent with non-genetic data (historical, geographical and morphological), suggesting that the estimated value of K is representative of the actual genetic structure. Asian phylogeography revealed a complex genetic structure involving locally high gene flow and variable admixture. Results of STRUCTURE, TreeMix and D-statistics taken together suggest that the current genetic structuring of the native *P. parva* range has been shaped by waves of gene flow originating from southern and northern populations (edges of the natural native distribution). Fish phylogeography in China has been greatly influenced by geological events (Chiang et al. 2013; Li 1981) that can represent major barriers to gene flow (Brandley et al. 2010). The complex geological history of the South China landmass acted as a barrier to gene flows and induced vicariance in common cyprinid species (Yang et al. 2016). In China, notable genetic divergence occurred between the North and the South, with the Qinling Mountains acting as a strong biogeographic barrier between the temperate climate of the North and the subtropical climate of the South (Yuan et al. 2012; Dong et al. 2011) resulting in a wide range of local adaptations to various ecological conditions that might facilitate establishment in the invaded area.

Yet within both biogeographical regions, gene flow was restricted and populations were mainly differentiated per river basins. In China north to the Qinling Mountains, the two main demes (i.e. North China and North East China; Figure 2a) were influenced by two different major river networks, the Amur River and the Yellow River (S1, S2 and S3 are located on the Haihe and Huaihe Rivers but both rivers were influenced by the Yellow River in history). South China (sites S9, S18; Figure 2a) was structured by the Yangtze River. Indeed, most

native demes were congruent with the architecture of the main river networks. Historically, five *Pseudorasbora* species, based on morphological characteristics and their association with major river basins, were described as endemic to China (Nichols 1928). These five species actually corresponded to the same *P. parva* species (Gozlan 2012), though genetic clustering was largely in agreement with the distribution of historical morphotypes (Nichols 1928). Yet admixture was also observed in locations at the boundaries between river basins (e.g. site S16 at the frontier between the Yellow and Amur river basins), indicating that recurrent migration has happened between demes.

Wild populations in Japan split from Chinese continental populations 12.1 Mya (Hardouin et al. 2018), consistent with Japan's separation from the Eurasian continent 15 Mya (Barnes 2003). Yet our results demonstrated significant gene flow with the South East deme. Two mtDNA lineages co-exist in Japan, including one closely related to continental populations (Watanabe et al. 2000), and recent hybridizations with Chinese populations have been described (Hardouin et al. 2018). Across Honshū Island, *P. parva* rapidly expanded its distribution with translocations of commercial cyprinids into ponds, which may have facilitated the introduction and the spread of the Chinese lineage (Konishi et al. 2003, 2009). However, resolving the origins and contemporary phylogeography of the Japanese population within the native distribution of *P. parva* with sufficient power would require more extensive sampling in Japan.

Anthropogenic influence

Genetic structure with long-term admixture has been reported in South China for other fishes, explained by coastal land and tributaries between river basins, especially for the Yangtze and Pearl Rivers (Yang et al. 2016). Admixed populations may be the result of anthropogenic modifications to the hydrological landscape, especially new dispersal pathways such as canals. Recent secondary contacts between the North and South may have been facilitated by the increased structural connectivity between river basins. The construction of the Lingqu Canal 2,200 years ago connected the Yangtze River (sites S9, S18 in South China and to a lesser extent S6 and S4 in South East China; Figure 2a) to the

Pearl River (Fengshu 1990). Moreover, the Beijing-Hangzhou Grand Canal connected the Yangtze River to the Huai He, Hai He and Yellow rivers in the east of China (AD 581-618, Sui Dynasty), a region at the core of the two admixed regions (Central China, North East China and South East China: Figure 2a). Secondary contact due to human-induced corridors can increase local admixture between populations (Crispo et al. 2011), and canals serving as ecological corridors between two interbreeding cyprinid species have ultimately become hybrid zones (Guivier et al. 2019). However, the most important vector for long-distance dispersal in the native range is aquaculture expansion throughout central China. The two most admixed demes correspond to areas of intensive aquaculture in China. In Central China, many ponds and reservoirs were intensively stocked in the 1950s-60s, with *P. parva* eventually becoming the dominant species (Zhao et al. 2015; Gong & Tu 1991). Massive human translocations for aquaculture have often led to biotic homogenization (Olden et al. 2004) and increased pathogen dispersal into the pool of future invaders (Price et al. 2016). Moreover, consecutive introductions represent a type of punctuated gene flow that produces admixed genotypes with high genetic diversity but unpredictable evolutionary effects (Crispo et al. 2011; Hasselman et al. 2014). In native locations, admixture is known for negative effects such as outbreeding depression and loss of local adaptation (Côte et al. 2014; Huff et al. 2011; Hufford et al. 2012), but in novel environments, admixture may increase adaptive potential for translocated individuals (Verhoeven et al. 2011). For example, admixture in the wild between divergent sculpin populations increased their genetic diversity, lineage differentiation, and facilitated colonization of new habitats (Nolte et al. 2005, 2009; Stemshorn et al. 2011).

4.2 The invasion

Invasion starts in Asia

Invasive populations have been successfully established at the edges of the native distribution. In South-West China, sites S19 and S20 are recently introduced populations (1980s) from South-East China, and perfectly clustered with S9, S18 (Figure 2a; see also Hardouin et al. 2018). The Tibetan population was also introduced, commercial exchange

being most probably the pathway for this invasion given that many living fishes are sold in Tibetan markets and are traditionally released into rivers (Gozlan 2012). Although Tibet shows similarities with Central China, a continuum of populations between them were not sampled, making it difficult to determine if Tibet should be clustered with this deme or another.

Out of Asia

Three invasive demes were consistently discriminated as (1) Iran, (2) a small Balkan-Anatolian deme (Eastern Europe) and (3) a large pan-European deme (Western Europe). Population differentiation (F_{ST}) was overall higher between the three invasive demes, than between the invasive demes and their source populations, suggesting independent introductions with different genomic backgrounds (Figure 3, Table S5). In addition, clear genetic sub-structuring was observed in the Western European deme. More populations sampled in this region and more markers would then be essential to more accurately depict fine-scale genetic structure of the West European deme. The Iranian population was particular, forming a distinct deme completely isolated from all European populations, consistent with observations of Hardouin et al. (2018), who revealed that Iran has a different mtDNA lineage. Differentiation between Eastern and Western Europe may be imputed to the former USSR era when on one side, the eastern introduction first opens and on the other commercial exchanges with Western Europe were limited (Britton & Gozlan 2013). The Danube river ensures structural connectivity which has favored colonization among a large central European region (Weiss et al. 2002), from Germany to Bulgaria through Austria and Hungary. However, regional biogeography of the Balkan region could also partially explain the observed population structure (Oikonomou et al. 2014). In addition, rapid range expansion of *P. parva* over long distances in the invasive range likely relied on human-mediated dispersal (e.g. unintentional releases, aquaculture exchanges) rather than natural dispersal, as *P. parva* favours lentic habitats. The observed genetic structure of invasive locations may suggest multiple independent introductions without subsequent gene flow. This assumption is strengthened by an invasive population genetic diversity similar to the

native populations (Table S1). A recent study showed a clear reduction in genetic diversity in the French invasive deme of *P. parva* compared to native demes, highlighting that the introduction was accompanied by a substantial loss of genetic diversity (Combe et al. 2022). However, in a previous study, high genetic diversity among invasive *P. parva* has been reported, first by Simon et al. (2015), and later attributed by Hardouin et al. (2018) to unbalanced sampling. Baltazar-Soares et al. (2020) have since highlighted a high genomic diversity attributed to a lack of admixture due to a recovery of the genetic bottleneck associated with the introduction. Here, with more balanced sampling between invasive and native populations, coupled with observed admixture in the invasive population (Figure 2b, Figure S5), our results suggest an invasive genetic diversity shaped by past invasions of admixed native populations.

It is common to observe strong genetic structure within demes for continental invasive fishes when using nuclear markers. This is partly explained by human driven long-distance dispersal and geographical barriers to gene flow (Sanz et al. 2013; Simon et al. 2011). Although we detected signals of differentiation with nuclear SNP data, mtDNA has revealed that Western European locations were characterized by the same mitochondrial lineage and, thus, constitute a single genetic population sharing a common ancestor (Simon et al. 2011). Low migration rates in a stepping-stone model induced by human-mediated secondary introductions might explain the relative differentiation between Western European locations, and might be a sign of an invasive bridgehead (van Boheemen et al. 2017). Invasive populations could also act as source populations for a secondary invasion, even over long-distances (Karsten et al. 2015; Lombaert et al. 2014).

4.3 Origins of invasion: admixture and multiple introductions

Our models described major demographic events, and gave congruent results between the two independent Machine Learning algorithms, supporting three independent introductions at the origins of the three invasive demes. Multiple introductions would, thus, multiply the chances of successful establishment outside

the native range. Results point towards a southern Chinese origin of the Eastern European deme, with alleles coming from and/or shared between the Central, South East and South China demes (Figure 2c). Similar patterns have been demonstrated via population assignment tests, revealing multiple origins in the same location indicative of multiple introductions or admixture within invasive demes (Guillemaud et al. 2011). Studies have also highlighted that admixture is likely a common characteristic for invasive species (Genton et al. 2005; Kolbe et al. 2004; Rius & Darling 2014); however, it is often difficult to disentangle multiple introductions from admixture before introduction. The reconstruction of admixture history may depend on the genetic structure observed in the native area. Admixture after introduction may be inferred when source populations are clearly differentiated and admixture is demonstrated in the invasive area, as in the case of the mussel, *Mytella charruana* (Gillis et al. 2009). However, this is not the case for *P. parva,* which shows high rates of admixture within the inferred source populations, suggesting that human-mediated gene flow within the native range shaped genetic variation in the pool of future invaders. The Central deme corresponded to the Wuhan region, at the core of Chinese aquaculture, with more than two million fishes produced annually, and frequent translocations from surrounding regions to restock ponds (Zhao et al. 2015). Consequently, admixture in the native range before the translocation into Europe may be at the origin of the Eastern European deme. The original introduction probably took place in Romania, at the epicenter of carp aquaculture in the 1960s, in the Nucet Fisheries Research Centre in 1961 (Gozlan et al. 2010), although several other introductions took place at the same time throughout the Black Sea region. Introduction was most likely accidental, as *P. parva* is a common contaminant of carp stockings (Wolter & Röhr 2010). Assuming between one and two generations per year among eastern European populations (Gozlan 2008), it complies with our estimates of the time of invasion (ca. 18 to 62 generations), and the time of Asian fish export increases. Likewise, development of aquaculture in this part of Europe may have subsequently facilitated the spread of *P. parva* in Turkey as early as 1982 (Ekmekçi & Kırankaya 2006; Britton & Gozlan 2013).

667 The Western Europe deme was probably colonized from a first introduction in Hungary,
668 noticed in the Paks Fisheries Farm in 1963 (Gozlan et al. 2010), which then spread
669 throughout Western Europe. The source populations mainly differed from those of the
670 Eastern Europe deme, although both share a connection with Central China. Western
671 Europe populations also include an admixed north Chinese origin to both the North East
672 China and North China demes. The invasion time was approximately the same as for
673 Eastern Europe, coincident with a trading cooperation period between countries of the
674 Eastern Bloc and China (Britton & Gozlan 2013). On average, genetic divergence amongst
675 source demes was relatively high (mean F_{ST} = 0.27; Figure 3, Table S5), suggesting limited
676 gene flow and a lower likelihood of native admixture, compared to that for Eastern Europe
677 source demes (mean F_{ST} = 0.13; Figure 3, Table S5). Moreover, individuals of the Hungarian
678 location were systematically assigned to two genetic clusters (Figure S5c), making it the
679 most admixed location of Western Europe, and suggested further that admixture most likely
680 happened from multiple introductions in the same invasive area. Temporal dynamics of
681 invasive mosquitoes in the US have shown the same admixture-like clustering pattern
682 (Fonseca et al. 2010). Range expansion across European countries most probably began in
683 the Hungarian admixed region, and its likely chronology has been reconstructed from census
684 data (Gozlan et al. 2010). *P. parva* colonized Austria as early as 1982 (Weber 1984), then
685 Germany from 1984 to 1987 (Arnold 1985; Lelek & Köhler 1989), and Belgium by 1992
686 (Vandelannoote & Yseboodt 1998). Austrian, Belgian and Polish sites clustered together,
687 suggesting a common ancestry. Their connectivity along the Danube and Rhine river
688 systems may have facilitated natural dispersal (Hegediš et al. 2007; Leuven et al. 2009).
689 However, natural dispersal cannot explain the observed high genetic structure, nor the
690 expansion over geographical barriers (e.g. the English Channel, the Alps), and so human-
691 mediated dispersal is more likely (Gozlan et al. 2002, Aparicio et al. 2012, Caiola & Sostoa
692 2002). Thus, Hungary could be the invasive bridgehead that initiated the colonization of
693 Western Europe, followed by secondary bottlenecks associated with a series of founding
694 events during westward range expansion. This invasive bridgehead effect scenario has also

been suggested by Combe et al. (2022) for a French deme, which probably originated from a successful invasive population. Surprisingly, Italy showed different origins than other Western European locations in population assignment tests (Central China instead of North-East China; Figure 2b), but was nonetheless considered in the Western European deme based on genetic clustering results. The origin of this small population was difficult to explain, as it may be part of an unsampled fourth invasive population linked with Slovakia that successfully spread across the Italian peninsula (Carosi et al. 2016; Záhorská & Kováč 2009). This uncertainty in the Italian origin underlies the necessity for the same extensive sampling in Europe as in Asia, especially in the Southeastern Europe that was believed to be the core of past introductions (e.g. Slovenia, Hungary, Romania, Bulgaria). This punctual sampling represents an inherent limitation to the definition of putative demes, and subsequently, to the number of invasion pathways inferred.

The Iranian population origin was complex to infer because of an ancient admixed origin between South East China and Japan. Iran showed a pattern of assignment similar to that of the Japanese admixture in nuclear markers between endemic and Chinese populations, potentially explaining why it was not possible to disentangle China and Japan as putative sources. In line with our results, a recent study (Ganjali et al. 2020), based on two mtDNA markers, identified three Iranian matrilineal haplotypes belonging to two distinct lineages, an older Japanese lineage and a Chinese lineage, corresponding to a recent natural dispersal from Azerbaijan. They also highlighted an admixture of highly divergent Japanese and Chinese lineages (Hardouin et al. 2018). Many questions remain, mainly because the history of Japanese populations has not been fully resolved. The high number of private alleles in Iran (Table S1) and the long drift separating the Iranian population from the most recent common ancestor with Japan (Figure S8) might suggest that the true source population has not been sampled. Thus, additional sampling would be required to effectively determine the origins of Iranian *P. parva*.

4.4 A successful invasion

The surprisingly large effective population sizes in Western Europe likely enhanced the

27

probability of introduced populations survival, and may explain its invasiveness. Furthermore, the co-introduction of non-native host-pathogen systems such as *P. parva*-*S. destruens* may create new host-pathogen interactions that can be detrimental to naïve native host species, hence providing a competitive advantage for the invaders (Price et al. 1986; Andreou & Gozlan 2016; Combe & Gozlan 2018; Vilcinskas & Knoll 2015). The gene flow resulting from the multiple introductions highlighted in Western Europe largely increased the effective population size as well as genetic diversity. However, effective population sizes were smaller in Iran and Eastern Europe, although the bottleneck they experienced was not severe, with 100 to 200 individuals in the founder populations. *P. parva* probably overcame the loss of diversity induced by bottlenecks due to admixture between native populations within its new environment and prior to range expansion. Our results demonstrated multiple source populations coming from a wide range of climatic conditions, the most flagrant being the admixture in the Wuhan region between subtropical and temperate populations, which represent different morphotypes (Nichols 1928; Gozlan et al. 2020). Multiple introductions associated with admixture can either increase or decrease the adaptive potential of invasive populations (Barker et al. 2019; Verhoeven et al. 2011). However, if from genetically distinct sources, multiple introductions can mitigate the negative effects of bottlenecks associated with invasion, as genetically diverse populations are less affected by the deleterious effects of inbreeding depression (Verhoeven et al. 2011). Moreover, they can increase individual fitness through heterosis, contributing to the invasion's success (Rius & Darling 2014; Vallejo-Marín et al. 2021), which is most likely the case in *P. parva*. Hence, native populations with broad and diverse biogeographic distribution may act as genetic and ecological diversity reservoirs.

5 CONCLUSION

Our results shed light on the importance of grasping the genetic history of the native range populations to better understand the effects of introductions and admixture on the success and adaptive potential of invasive populations. Our study also draws a picture of the complex demo-genetic history of an invasive species' source populations and its spread across

recurrent global invasion pathways. Reconstruction of such invasion pathways is crucial for setting up conservation biology approaches and management to prevent further non-native species introductions and potential associated pathogens.

Acknowledgements

We would like to thank all co-authors of Gozlan et al. 2020 for their help in collecting *P. parva* samples across the native and invasive range as well as B. Haenfling. We would also like to thank Eric Lombaert for kindly agreeing to discuss the best practices of Approximate Bayesian Computation methods. This study has been conducted in accordance with the relevant animal or human ethics approvals.

References

Andreou D, Gozlan, RE (2016) Associated disease risk from the introduced generalist pathogen *Sphaerothecum destruens*: management and policy implications. Parasitology 143:1204-1210. https://doi.org/10.1017/S003118201600072X

Aparicio E, Peris B, Torrijos L, Prenda J, Nieva A, Perea S (2012) Expansion of the invasive *Pseudorasbora parva* (*Cyprinidae*) in the Iberian Peninsula: first record in the Guadiana River basin. Cybium 36:585-586

Arnold A (1985) *Pseudorasbora parva* (Schlegel 1842) nun auch in der DDR. *Z. Binnenfisch DDR* 32:182-183

Baltazar Soares M, Blanchet S, Cote J, Tarkan AS, Záhorská E, Gozlan RE, Eizaguirre C (2020) Genomic footprints of a biological invasion: Introduction from Asia and dispersal in Europe of the topmouth gudgeon (*Pseudorasbora parva*). Mol Ecol 29:71-85

Barker BS, Cocio JE, Anderson SR, Braasch JE, Cang FA, Gillette HD et al (2019) Potential limits to the benefits of admixture during biological invasion. Mol Ecol 28:100-113. https://doi.org/10.1111/mec.14958

Barnes GL (2003) Origins of the Japanese Islands: The New" Big Picture". *Nichibunken Japan Review* 3-50

Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian computation in population genetics. Genetics 162:2025-2035

Becker RA, Wilks AR, Brownrigg R, Minka TP, Deckmyn A (2018) maps: Draw Geographical Maps. R package version 3.3. 0. https://CRAN.R-project.org/package=maps

Beichman AC, Huerta-Sanchez E, Lohmueller KE (2018) Using genomic data to infer historic population dynamics of nonmodel organisms. Annu Rev Ecol Evol Syst 49:433-456. https://doi.org/10.1146/annurev-ecolsys-110617-062431

Blum MG, François O (2010) Non-linear regression models for Approximate Bayesian Computation. Stat Comput 20:63-73. https://doi.org/10.1007/s11222-009-9116-0

Bossdorf O, Auge H, Lafuma L, Rogers WE, Siemann E, Prati D (2005) Phenotypic and genetic differentiation between native and introduced plant populations. Oecologia 144:1-11. https://doi.org/10.1007/s00442-005-0070-z

Brandley MC, Guiher TJ, Pyron RA, Winne CT, Burbrink FT (2010) Does dispersal across an aquatic geographic barrier obscure phylogeographic structure in the diamond-backed watersnake (*Nerodia rhombife*r)? Mol Phylogenetics Evol 57:552-560.

https://doi.org/10.1016/j.ympev.2010.07.015

Britton JR, Gozlan RE (2013) Geo-politics and freshwater fish introductions: How the Cold War shaped Europe's fish allodiversity. Glob Environ Change 23:1566-1574. https://doi.org/10.1016/j.gloenvcha.2013.09.017

Browning BL, Zhou Y, Browning SR (2018) A one-penny imputed genome from next-generation reference panels. Am J Hum Genet 103:338-348. https://doi.org/10.1016/j.ajhg.2018.07.015

Browning SR, Browning BL (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet 81:1084-1097. https://doi.org/10.1086/521987

Cabrera AA, Palsbøll PJ (2017) Inferring past demographic changes from contemporary genetic data: A simulation based evaluation of the ABC methods implemented in DIYABC. Mol Ecol Resour 17:e94-e110. https://doi.org/10.1111/1755-0998.12696

Caiola N, De Sostoa A (2002) First record of the Asiatic cyprinid *Pseudorasbora parva* in the Iberian Peninsula. J Fish Biol 4:1058-1060. https://doi.org/10.1006/jfbi.2002.2103

Carosi A, Ghetti L, Lorenzoni M (2016) Status of *Pseudorasbora parva* in the Tiber river basin (Umbria, central Italy) 20 years after its introduction. Knowl Manag Aquat Ecosyst 417:22. https://doi.org/10.1051/kmae/2016009

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. Mol Ecol 22:3124-3140. https://doi.org/10.1111/mec.12354

Chan KO, Alexander AM, Grismer LL, Su Y-C, Grismer JL, Quah ESH, Brown RM (2017) Species delimitation with gene flow: A methodological comparison and population genomics approach to elucidate cryptic species boundaries in Malaysian Torrent Frogs. Mol Ecol 26:5435–5450

Chapin Iii FS, Zavaleta ES, Eviner VT, Naylor RL, Vitousek PM, Reynolds HL et al (2000) Consequences of changing biodiversity. Nature 405:234-242. https://doi.org/10.1038/35012241

Chen KY, Marschall EA, Sovic MG, Fries AC, Gibbs HL, Ludsin SA (2018) Assign POP: An r package for population assignment using genetic, non□genetic, or integrated data in a machine□learning framework. Methods Ecol Evol 9:439-446. https://doi.org/10.1111/2041-210X.12897

Chiang TY, Lin HD, Zhao J, Kuo PH, Lee TW, Hsu KC (2013) Diverse processes shape deep phylogeographical divergence in *Cobitis sinensis* (*Teleostei*: *Cobitidae*) in East Asia. J Zool Syst Evol Res 51:316-326. https://doi.org/10.1111/jzs.12030

Clavero M, García-Berthou E (2005) Invasive species are a leading cause of animal extinctions. Trends Ecol Evol 20:110. https://doi.org/10.1016/j.tree.2005.01.003

Combe M, Gozlan RE (2018) The rise of the rosette agent in Europe: An epidemiological enigma. Transbound Emerg Dis 65:1474-1481. https://doi.org/10.1111/tbed.13001

Combe M, Cherif E, Charrier A, Barbey B, Chague M, Carrel G et al (2022) Towards unravelling the Rosette agent enigma: Spread and emergence of the co-invasive host-pathogen complex, *Pseudorasbora parva-Sphaerothecum destruens*. Sci. Total Environ. 806:150427. https://doi.org/10.1016/j.scitotenv.2021.150427

Cornuet JM, Pudlo P, Veyssier J, Dehne-Garcia A, Gautier M, Leblois R et al (2014) DIYABC v2.0: a software to make approximate Bayesian computation inferences about population history using single nucleotide polymorphism, DNA sequence and microsatellite data. Bioinformatics 30:1187-1189. https://doi.org/10.1093/bioinformatics/btt763

Côte J, Roussel JM, Le Cam S, Evanno G (2014) Outbreeding depression in Atlantic salmon revealed by hypoxic stress during embryonic development. Evol Biol 41:561-571. https://doi.org/10.1007/s11692-014-9289-0

Crispo E, Moore JS, Lee-Yaw JA, Gray SM, Haller BC (2011) Broken barriers: Human-induced changes to gene flow and introgression in animals: An examination of the ways in which humans increase genetic exchange among populations and species and the consequences for biodiversity. BioEssays 33:508-518. https://doi.org/10.1002/bies.201000154

Cristescu ME (2015) Genetic reconstructions of invasion history. Mol Ecol 24:2212-2225. https://doi.org/10.1111/mec.13117

Crowl TA, Crist TO, Parmenter RR, Belovsky G, Lugo AE (2008) The spread of invasive species and infectious disease as drivers of ecosystem change. Front Ecol Environ 6:238-246. https://doi.org/10.1890/070151

Csilléry K, François O, Blum MG (2012) abc: an R package for approximate Bayesian computation (ABC). Methods Ecol Evol 3:475-479. https://doi.org/10.1111/j.2041-210X.2011.00179.x

Didham RK, Tylianakis JM, Hutchison MA, Ewers RM, Gemmell NJ (2005) Are invasive species the drivers of ecological change? Trends Ecol Evol 20:470-474. https://doi.org/10.1016/j.tree.2005.07.006

Dlugosch KM, Parker IM (2008) Founding events in species invasions: Genetic variation, adaptive evolution, and the role of multiple introductions. Mol Ecol 17:431–449

Dong Y, Zhang G, Neubauer F, Liu X, Genser J, Hauzenberger C (2011) Tectonic evolution of the Qinling orogen, China: review and synthesis. J Asian Earth Sci 41:213-237. https://doi.org/10.1016/j.jseaes.2011.03.002

Durand EY, Patterson N, Reich D, Slatkin M (2011) Testing for ancient admixture between closely related populations. Mol Biol Evol 28:2239-2252. https://doi.org/10.1093/molbev/msr048

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

Ekmekçi FG, Kirankaya ŞG (2006) Distribution of an invasive fish species, *Pseudorasbora parva* (Temminck & Schlegel, 1846) in Turkey. Turk J Zool 30:329-334

Estoup A, Guillemaud T (2010) Reconstructing routes of invasion using genetic data : Why, how and so what? Mol Ecol 19:4113-4130. https://doi.org/10.1111/j.1365-294X.2010.04773.x

Estoup A, Lombaert E, Marin JM, Guillemaud T, Pudlo P, Robert CP, Cornuet JM (2012) Estimation of demo-genetic model probabilities with Approximate Bayesian Computation using linear discriminant analysis on summary statistics. Mol Ecol Resour 12:846-855. https://doi.org/10.1111/j.1755-0998.2012.03153.x

Estoup A, Ravigné V, Hufbauer R, Vitalis R, Gautier M, Facon B (2016) Is there a genetic paradox of biological invasion? Annu Rev Ecol Evol Syst 47:51-72. https://doi.org/10.1146/annurev-ecolsys-121415-032116

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol 14:2611-2620. https://doi.org/10.1111/j.1365-294X.2005.02553.x

Excoffier L, Heckel G (2006) Computer programs for population genetics data analysis: a survival guide. Nat Rev Genet 7:745–758

Facon B, Genton BJ, Shykoff J, Jarne P, Estoup A, David P (2006) A general eco-evolutionary framework for understanding bioinvasions. Trends Ecol Evol 21:130-135. https://doi.org/10.1016/j.tree.2005.10.012

Fengshu Z (1990) Preliminary study of geological geomorphological conditions of ling canal. Carsologica Sinica 1

Fitzpatrick BM, Fordyce JA, Niemiller ML, Reynolds RG (2012) What can DNA tell us about biological invasions?. Biol. Invasions 14:245-253.

Fonseca DM, Widdel AK, Hutchinson M, Spichiger SE, Kramer LD (2010) Fine□scale spatial and temporal population genetics of *Aedes japonicus*, a new US mosquito, reveal multiple introductions. Mol Ecol 19:1559-1572. https://doi.org/10.1111/j.1365-294X.2010.04576.x

Gallardo B, Clavero M, Sánchez MI, Vilà M (2016) Global ecological impacts of invasive species in aquatic ecosystems. Glob Chang Biol 22:151-163. https://doi.org/10.1111/gcb.13004

Ganjali Z, Esmaeili HR, Zarei F, Sayyadzadeh G, Eagderi S, Gozlan RE (2020) West Asian colonisation of topmouth gudgeon, *Pseudorasbora parva* (*Teleostei*: *Gobionidae*): Genetic admixture at the crossroad of Europe and east Asia. Freshw Biol 66:699–715. https://doi.org/10.1111/fwb.13671

Genton BJ, Shykoff JA, Giraud T (2005) High genetic diversity in French invasive populations of common ragweed, *Ambrosia artemisiifolia*, as a result of multiple sources of introduction. Mol Ecol 14:4275-4285. https://doi.org/10.1111/j.1365-294X.2005.02750.x

Gillis NK, Walters LJ, Fernandes FC, Hoffman EA (2009) Higher genetic diversity in introduced than in native populations of the mussel *Mytella charruana*: evidence of population admixture at introduction sites. Divers Distrib 15:784-795. https://doi.org/10.1111/j.1472-4642.2009.00591.x

Gong M, Tu F (1991) Fishery in contemporary China. Contemporary China Press, Beijing

Goudet J (2005) Hierfstat, a package for R to compute and test hierarchical F□statistics. Mol Ecol Notes 5:184-186. https://doi.org/10.1111/j.1471-8278 .2004.00828.x

Goudet J, Jombart T (2015) hierfstat: estimation and tests of hierarchical F-statistics. R package version 0.04-22. 10

Gozlan RE, Andreou D, Asaeda T, Beyer K, Bouhadad R, Burnard D et al (2010) Pan□ continental invasion of *Pseudorasbora parva*: towards a better understanding of freshwater fish invasions. Fish Fish 11:315-340. https://doi.org/10.1111/j.1467-2979.2010.00361.x

Gozlan RE, Pinder AC, Shelley J (2002) Occurrence of the Asiatic cyprinid *Pseudorasbora parva* in England. J Fish Biol 61:298-300. https://doi.org/10.1006/jfbi.2002.2042

Gozlan RE, Záhorská E, Cherif E, Asaeda T, Britton JR, Chang CH et al (2020) Native drivers of fish life history traits are lost during the invasion process. Ecol Evol 10:8623-8633. https://doi.org/10.1002/ece3.6521

Gozlan RE (2008) Introduction of non□native freshwater fish: is it all bad? Fish Fish 9:106-115. https://doi.org/10.1111/j.1467-2979.2007.00267.x

Gozlan RE (2012) *Pseudorasbora parva* temminck and schlegel (topmouth gudgeon). Handb Glob Freshw Invasive Species, Abingdon: Earthscan

Graebner RC, Callaway RM, Montesinos D (2012) Invasive species grows faster, competes better, and shows greater evolution toward increased seed size and growth than exotic non-invasive congeners. Plant Ecol 213:545-553. https://doi.org/10.1007/s11258-012-0020-x

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M et al (2010) A draft sequence of the Neandertal genome. Science 328:710-722. https://doi.org/10.1126/science.1188021

Guillemaud T, Ciosi M, Lombaert E, Estoup A (2011) Biological invasions in agricultural settings: insights from evolutionary biology and population genetics. C R Biol 334:237-246. https://doi.org/10.1016/j.crvi.2010.12.008

Guivier E, Gilles A, Pech N, Duflot N, Tissot L, Chappaz R (2019) Canals as ecological corridors and hybridization zones for two cyprinid species. Hydrobiologia 830(1):1-16. https://doi.org/10.1007/s10750-018-3843-1

939 Hardouin EA, Andreou D, Zhao Y, Chevret P, Fletcher DH, Britton JR, Gozlan RE (2018)
940 Reconciling the biogeography of an invader through recent and historic genetic patterns: the
941 case of topmouth gudgeon *Pseudorasbora parva*. Biol Invasions 20(8):2157-2171.
942 https://doi.org/10.1007/s10530-018-1693-4

943 Hasselman DJ, Argo EE, McBride MC, Bentzen P, Schultz TF, Perez□Umphrey AA,
944 Palkovacs EP (2014) Human disturbance causes the formation of a hybrid swarm between
945 two naturally sympatric fish species. Mol Ecol 23:1137-1152.
946 https://doi.org/10.1111/mec.12674

947 Price PW, Westoby M, Rice B, Atsatt PR, Fritz RS, Thompson JN, Mobley K (1986) Parasite
948 mediation in ecological interactions. Annu. Rev. Ecol. Syst. 17:487–505.

949 Hegediš A, Lenhardt M, Mićković B, Cvijanović G, Jarić I, Gačić Z (2007) Amur sleeper
950 (*Perccottus glenii* Dubowski, 1877) spreading in the Danube River basin. J Appl Ichthyol
951 23:705-706. https://doi.org/10.1111/j.1439-0426.2007.00867.x

952 Holsbeek G, Mergeay J, Hotz H, Plötner J, Volckaert FAM, De Meester L (2008) A cryptic
953 invasion within an invasion and widespread introgression in the European water frog
954 complex: consequences of uncontrolled commercial trade and weak international legislation.
955 Mol Ecol 17:5023-5035. https://doi.org/110.1111/j.1365-294X.2008.03984.x

956 Huang H, Knowles LL (2016) Unforeseen Consequences of Excluding Missing Data from
957 Next-Generation Sequences: Simulation Study of RAD Sequences. Syst Biol 65:357–365.
958 https://doi.org/10.1093/sysbio/syu046

959 Huff DD, Miller LM, Chizinski CJ, Vondracek B (2011) Mixed□source reintroductions lead to
960 outbreeding depression in second□generation descendents of a native North American fish.
961 Mol Ecol 20:4246-4258. https://doi.org/10.1111/j.1365-294X.2011.05271.x

962 Hufford KM, Krauss SL, Veneklaas EJ (2012) Inbreeding and outbreeding depression in
963 *Stylidium hispidum*: implications for mixing seed sources for ecological restoration. Ecol Evol
964 2:2262-2273. https://doi.org/10.1002/ece3.302

965 Hulme PE (2009) Trade, transport and trouble: managing invasive species pathways in an
966 era of globalization. J Appl Ecol 46:10-18. https://doi.org/10.1111/j.1365-2664.2008.01600.x

967 Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation
968 program for dealing with label switching and multimodality in analysis of population structure.
969 Bioinformatics 23:1801-1806. https://doi.org/10.1093/bioinformatics/btm233

970 Jombart T, Ahmed I (2011) adegenet 1.3-1: new tools for the analysis of genome-wide SNP
971 data. Bioinformatics 27:3070-3071. https://doi.org/10.1093/bioinformatics/btr521

972 Jombart T (2008) adegenet: a R package for the multivariate analysis of genetic markers.
973 Bioinformatics 24:1403-1405. https://doi.org/10.1093/bioinformatics/btn129

974 Kamvar ZN, Brooks JC, Grünwald NJ (2015) Novel R tools for analysis of genome-wide
975 population genetic data with emphasis on clonality. Front Genet 6:208.
976 https://doi.org/10.3389/fgene.2015.00208

977 Kamvar ZN, Tabima JF, Grünwald NJ (2014) Poppr: an R package for genetic analysis of
978 populations with clonal, partially clonal, and/or sexual reproduction. PeerJ 2:e281.
979 https://doi.org/10.7717/peerj.281

980 Karsten M, Jansen van Vuuren B, Addison P, Terblanche JS (2015) Deconstructing
981 intercontinental invasion pathway hypotheses of the Mediterranean fruit fly (*Ceratitis*
982 *capitata*) using a Bayesian inference approach: are port interceptions and quarantine
983 protocols successfully preventing new invasions? Divers Distrib 21:813-825. https://doi.org/
984 10.1111/ddi.12333

985 Kolbe JJ, Glor RE, Schettino LR, Lara AC, Larson A, Losos JB (2004) Genetic variation
986 increases during biological invasion by a Cuban lizard. Nature 431:177-181.
987 https://doi.org/10.1038/nature02807

988 Konishi M, Hosoya K, Takata K (2003) Natural hybridization between endangered and

989 introduced species of *Pseudorasbora*, with their genetic relationships and characteristics
990 inferred from allozyme analyses. J Fish Biol 63:213-231. https://doi.org/10.1046/j.1095-
991 8649.2003.00146.x

992 Konishi M, Sakano H, Iguchi KI (2009) Identifying conservation priority ponds of an
993 endangered minnow, *Pseudorasbora pumila*, in the area invaded by *Pseudorasbora parva*.
994 Ichthyol Res 56:346. https://doi.org/10.1007/s10228-009-0106-1

995 Lelek A, Köhler C (1989) Zustandsanalyse der Fischartengemeinschaften im Rhein (1987–
996 1988) *Fischökologie* 1:47-64

997 Leuven RS, van der Velde G, Baijens I, Snijders J, van der Zwart C, Lenders HR, bij de
998 Vaate A (2009) The river Rhine: a global highway for dispersal of aquatic invasive species.
999 Biol Invasions 11:1989. https://doi.org/10.1007/s10530-009-9491-7

1000 Li S (1981) Studies on zoogeographical divisions for fresh water fishes of China. Science
1001 Press

1002 Li YL, Liu JX (2018) StructureSelector: A web□based software to select and visualize the
1003 optimal number of clusters using multiple methods. Mol Ecol Resour 18:176-177.
1004 https://doi.org/10.1111/1755-0998.12719

1005 Lombaert E, Guillemaud T, Lundgren J, Koch R, Facon B, Grez A et al (2014)
1006 Complementarity of statistical treatments to reconstruct worldwide routes of invasion: the
1007 case of the Asian ladybird *Harmonia axyridis*. Mol Ecol 23:5979-5997.
1008 https://doi.org/10.1111/mec.12989

1009 Mergeay J, Verschuren D, De Meester L (2005) Cryptic invasion and dispersal of an
1010 American Daphnia in East Africa. Limnol Oceanogr 50:1278-1283.
1011 https://doi.org/10.4319/lo.2005.50.4.1278

1012 Milanesi M, Capomaccio S, Vajana E, Bomba L, Garcia JF, Ajmone-Marsan P, Colli L
1013 (2017) BITE: an R package for biodiversity analyses. BioRxiv 181610.
1014 https://doi.org/10.1101/181610

1015 Mooney HA, Cleland EE (2001) The evolutionary impact of invasive species. Proc Natl Acad
1016 Sci U S A 98:5446-5451. https://doi.org/10.1073/pnas.091093398

1017 Muirhead JR, Gray DK, Kelly DW, Ellis SM, Heath DD, Macisaac HJ (2008) Identifying the
1018 source of species invasions: sampling intensity vs. genetic diversity. Mol Ecol 17:1020-1035.
1019 https://doi.org/10.1111/j.1365-294X.2008.03669.x

1020 Nichols JT (1928) Chinese fresh-water fishes in the American Museum of Natural History's
1021 collections: a provisional check-list of the fresh-water fishes of China. Bulletin of the AMNH;
1022 v. 58, article 1

1023 Nolte AW, Freyhof J, Stemshorn KC, Tautz D (2005) An invasive lineage of sculpins, *Cottus*
1024 sp.(*Pisces*, *Teleostei*) in the Rhine with new habitat adaptations has originated from
1025 hybridization between old phylogeographic groups. Proc R Soc Lond B Biol Sci 272:2379-
1026 2387. https://doi.org/10.1098/rspb.2005.3231

1027 Nolte AW, Gompert Z, Buerkle CA (2009) Variable patterns of introgression in two sculpin
1028 hybrid zones suggest that genomic isolation differs among populations. Mol Ecol 18:2615-
1029 2627. https://doi.org/10.1111/j.1365-294X.2009.04208.x

1030 Oikonomou A, Leprieur F, Leonardos ID (2014) Biogeography of freshwater fishes of the
1031 Balkan Peninsula. Hydrobiologia 738:205-220. https://doi.org/10.1007/s10750-014-1930-5

1032 Olden JD, Poff NL, Douglas MR, Douglas ME, Fausch KD (2004) Ecological and
1033 evolutionary consequences of biotic homogenization. Trends Ecol Evol 19:18-24.
1034 https://doi.org/10.1016/j.tree.2003.09.010

1035 Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y et al (2012) Ancient admixture
1036 in human history. Genetics 192:1065-1093. https://doi.org/10.1534/genetics.112.145037

1037 Peischl S, Excoffier L (2015) Expansion load: recessive mutations and the role of standing

genetic variation. Mol Ecol 24:2084-2094. https://doi.org/10.1111/mec.13154

Perdereau E, Dedeine F, Christidès JP, Dupont S, Bagnères AG (2011) Competition between invasive and indigenous species: an insular case study of subterranean termites. Biol Invasions 13:1457-1470. https://doi.org/10.1007/s10530-010-9906-5

Phillips BL, Brown GP, Webb JK, Shine R (2006) Invasion and the evolution of speed in toads. Nature 439:803-803. https://doi.org/10.1038/439803a

Pickrell JK, Pritchard JK (2012) Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. PLoS Genet 8:e1002967. https://doi.org/10.1371/journal.pgen.1002967

Pickrell JK, Patterson N, Barbieri C, Berthold F, Gerlach L, Güldemann T et al (2012) The genetic prehistory of southern Africa. Nat Commun 3(1):1-6. https://doi.org/10.1038/ncomms2140

Plummer M, Best N, Cowles K, Vines K (2006) CODA: convergence diagnosis and output analysis for MCMC. R news 6:7-11

Price SJ, Garner TWJ, Cunningham AA, Langton TES, Nichols RA (2016) Reconstructing the emergence of a lethal infectious disease of wildlife supports a key role for spread through translocations by humans. Proc R Soc B 283:20160952. https://doi.org/10.1098/rspb.2016.0952

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945-959. https://doi.org/10.1093/sysbio/sys038

Pudlo P, Marin JM, Estoup A, Cornuet JM, Gautier M, Robert CP (2016) Reliable ABC model choice via random forests. Bioinformatics 32:859-866. https://doi.org/10.1093/bioinformatics/btv684

Puechmaille SJ (2016) The program structure does not reliably recover the correct population structure when sampling is uneven: subsampling and new estimators alleviate the problem. Mol Ecol Resour 16:608-627. https://doi.org/10.1111/1755-0998.12512

Raynal L, Marin JM, Pudlo P, Ribatet M, Robert CP, Estoup A (2019) ABC random forests for Bayesian parameter inference. Bioinformatics 35:1720-1728. https://doi.org/10.1093/bioinformatics/bty867

Rius M, Darling JA (2014) How important is intraspecific genetic admixture to the success of colonising populations? Trends Ecol Evol 29:233-242. https://doi.org/10.1016/j.tree.2014.02.003

Rius M, Turon X (2020) Phylogeography and the Description of Geographic Patterns in Invasion Genomics. Front Ecol Evol 8:595711. https://doi.org/10.3389/fevo.2020.595711

Roman J, Darling JA (2007) Paradox lost: genetic diversity and the success of aquatic invasions. Trends Ecol Evol 22:454-464. https://doi.org/10.1016/j.tree.2007.07.002

Salo P, Korpimäki E, Banks PB, Nordström M, Dickman CR (2007) Alien predators are more dangerous than native predators to prey populations. Proc R Soc Lond B Biol Sci 274:1237-1243. https://doi.org/10.1098/rspb.2006.0444

Sanz N, Araguas RM, Vidal O, Diez-del-Molino D, Fernández-Cebrián R, García-Marín JL (2013) Genetic characterization of the invasive mosquitofish (*Gambusia* spp.) introduced to Europe: population structure and colonization routes. Biol Invasions 15:2333-2346. https://doi.org/10.1007/s10530-013-0456-5

Sax DF, Stachowicz JJ, Brown JH, Bruno JF, Dawson MN, Gaines SD et al (2007) Ecological and evolutionary insights from species invasions. Trends Ecol Evol 22:465-471. https://doi.org/10.1016/j.tree.2007.06.009

Shafer AB, Gattepaille LM, Stewart RE, Wolf JB (2015) Demographic inferences using short□read genomic data in an approximate Bayesian computation framework: In silico evaluation of power, biases and proof of concept in Atlantic walrus. Mol Ecol 24:328-345.

https://doi.org/10.1111/mec.13034

Simberloff D (2013) Biological invasions: much progress plus several controversies. Contributions to Science 7-16. https://doi.org/10.2436/20.7010.01.158

Simon A, Britton R, Gozlan R, Van Oosterhout C, Volckaert FA, Hänfling B (2011) Invasive cyprinid fish in Europe originate from the single introduction of an admixed source population followed by a complex pattern of spread. PLOS one 6:e18560. https://doi.org/10.1371/journal.pone.0018560

Simon A, Gozlan RE, Britton JR, Van Oosterhout C, Hänfling B (2015) Human induced stepping-stone colonisation of an admixed founder population: the spread of topmouth gudgeon (*Pseudorasbora parva*) in Europe. Aquat Sci 77:17-25. https://doi.org/10.1007/s00027-014-0374-3

Sinama M, Gilles A, Costedoat C, Corse E, Olivier JM, Chappaz R, Pech N (2013) Non-homogeneous combination of two porous genomes induces complex body shape trajectories in cyprinid hybrids. Front Zool 10:1-16. https://doi.org/10.1186/1742-9994-10-22

Skoglund P, Mallick S, Bortolini MC, Chennagiri N, Hünemeier T, Petzl-Erler ML et al (2015) Genetic evidence for two founding populations of the Americas. Nature 525:104-108. https://doi.org/10.1038/nature14895

South A (2017) Rnaturalearth: world map data from natural earth. R package version 0.1.0. https://CRAN.R-project.org/package=rnaturalearth.

Stemshorn KC, Reed FA, Nolte AW, Tautz D (2011) Rapid formation of distinct hybrid lineages after secondary contact of two fish species (*Cottus* sp.). Mol Ecol 20:1475-1491. https://doi.org/10.1111/j.1365-294X.2010.04997.x

Stiers I, Crohain N, Josens G, Triest L (2011) Impact of three aquatic invasive species on native plants and macroinvertebrates in temperate ponds. Biol Invasions 13:2715-2726. https://doi.org/10.1007/s10530-011-9942-9

Tange O (2011) Gnu parallel-the command-line power tool. The USENIX Magazine 36:42-47

Team RC (2019) R: A language and environment for statistical computing (version 3.1.2) Vienna, Austria. R Foundation for Statistical Computing; 2014

Vallejo-Marín M, Friedman J, Twyford AD, Lepais O, Ickert-Bond SM, Streisfeld M et al (2021) Population genomic and historical analysis suggests a global invasion by bridgehead processes in *Mimulus guttatus*. Commun Biol 4:1-12. https://doi.org/10.1038/s42003-021-01795-x

van Boheemen LA, Lombaert E, Nurkowski KA, Gauffre B, Rieseberg LH, Hodgins KA (2017) Multiple introductions, admixture and bridgehead invasion characterize the introduction history of *Ambrosia artemisiifolia* in Europe and Australia. Mol Ecol 26:5421-5434. https://doi.org/10.1111/mec.14293

Vandelannoote A, Yseboodt R (1998) Atlas van de Vlaamse beek-en riviervissen. Water-Energik-vLario

Verhoeven KJ, Macel M, Wolfe LM, Biere A (2011) Population admixture, biological invasions and the balance between local adaptation and inbreeding depression. Proc R Soc Lond B Biol Sci 278:2-8. https://doi.org/10.1098/rspb.2010.1272

Vilcinskas A (2015) Pathogens as Biological Weapons of Invasive Species. PLoS Pathog 11:e1004714. https://doi.org/10.1371/journal.ppat.1004714

Villéger S, Blanchet S, Beauchard O, Oberdorff T, Brosse S (2011) Homogenization patterns of the world's freshwater fish faunas. Proc Natl Acad Sci U S A 108:18003-18008. https://doi.org/10.1073/pnas.1107614108

Watanabe K, Iguchi KI, Hosoya K, Nishida M (2000) Phylogenetic relationships of the Japanese minnows, *Pseudorasbora* (*Cyprinidae*), as inferred from mitochondrial 16S rRNA gene sequences. Ichthyol Res 47:43-50. https://doi.org/10.1007/BF02674312

Weber E (1984) Die ausbreitung der pseudokeilfleckbarben im donauraum. Österreichs Fischerei 37:63-65

Weiss S, Persat H, Eppe R, Schlötterer C, Uiblein F (2002) Complex patterns of colonization and refugia revealed for European grayling *Thymallus thymallus*, based on complete sequencing of the mitochondrial DNA control region. Mol Ecol 11:1393-1407. https://doi.org/10.1046/j.1365-294x.2002.01544.x

Wolter C, Röhr F (2010) Distribution history of non□native freshwater fish species in Germany: how invasive are they? J Appl Ichthyol 26:19-27. https://doi.org/10.1111/j.1439-0426.2010.01505.x

Yang JQ, Hsu KC, Liu ZZ, Su LW, Kuo PH, Tang WQ et al (2016) The population history of *Garra orientalis* (*Teleostei*: *Cyprinidae*) using mitochondrial DNA and microsatellite data with approximate Bayesian computation. BMC Evol Biol 16:73. https://doi.org/10.1186/s12862-016-0645-9

Yuan JH, Cheng FY, Zhou SL (2012) Genetic structure of the tree peony (*Paeonia rockii*) and the Qinling Mountains as a geographic barrier driving the fragmentation of a large population. PLoS One 7:e34955. https://doi.org/10.1371/journal.pone.0034955

Záhorská E, Kováč V (2009) Reproductive parameters of invasive topmouth gudgeon *Pseudorasbora parva* (Temminck and Schlegel, 1846) from Slovakia. J Appl Ichthyol 25:466-469. https://doi.org/10.1111/j.1439-0426.2009.01190.x

Zhang C, Zhao Y (2016) Species diversity and distribution of inland fishes in China. Science Press, Beijing

Zhao Y, Gozlan RE, Zhang C (2015) Current state of freshwater fisheries in China. Wiley-Blackwell, Oxford

List of Tables and Figures

Table 1. Population properties with site specific sample, geographic range, geographical coordinates, and (N) number of individuals.

Table 2. Inference of the demographic model based on Neural Network and Random Forest. Selected models at each step are given for each algorithm (bold values are those of the selected model). 10,000 simulated datasets per scenario for training. Prior error rate estimated with leave-one-out cross-validation for Neural network and out-of-bag error rate for Random Forest. Random Forest does not provide posterior probability of each scenario, but rather the number of votes in favor to each one. Subsequently, posterior probability of the Random Forest corresponds to the posterior probability of the selected model.

Fig. 1 Genetic structure and population modeling in the native range. **(a)** Posterior admixture proportions of individuals estimated with STRUCTURE for 18 sampled sites (n=300) in the native range (Asia) and K=6 genetic clusters (proportions aggregated with CLUMPP from 20 independent replicates). **(b)** ML tree based on TreeMix with block size of 500 SNPs. The migration edges pointing from South China to Central China, South Central China, and North East China have migration weights of 45%, 33%, and 2,7%. The migration edge pointing from North China to North East China has a migration weight of 9%. **(c)** D-statistics for testing admixture in native populations. Populations are noted D(O, P3, P1, P2). Negative values indicate gene flow between P3 and P1 while positive values indicate gene flow between P3 and P2. Confidence interval at 95% of the D-statistic estimated by block jackknife of 5kb. Null expectation is a D-statistic of 0. Demes abbreviations are North=North China, South=South China, Central=Central China, NorthE=North East China, SouthE=South East China

Fig. 2 Definition of demes in the native range, and source populations predicted for the invasive range. **(a) (Native range, Asia)** Mean admixture proportions estimated with STRUCTURE in the native range (Asia), for the chosen K=6. Native sampled sites are pooled into putative demes for subsequent analyses. Pie chart colors correspond to the genetic clusters inferred by STRUCTURE, as in Figure 1. **(Invasive range, Europe & Middle-East)** Assignment predictions of invasive individuals to native demes with AssignPOP's SVM algorithm with a relative posterior probability >2 (n=292 for training, n=100 for predictions). Pie chart colors correspond to the putative demes defined in the native range to which invasive individuals were assigned. **(b)** Posterior assignment probabilities to putative native demes estimated with AssignPop. **(c)** The demo-genetic scenario that was inferred with Approximate Bayesian Computation demonstrating three independent introductions from three independent admixed source populations. Branch lengths are not scaled. Bottleneck events are represented in thin red lines in branches. Colored branches correspond to invasive demes history.

Fig. 3 Population differentiation. Pairwise F_{ST} (Weir & Cockerham's F_{ST}) between native and invasive demes. F_{ST} values were estimated from 3,000 loci.

1235

39

Table 1. Population properties with site specific sample, geographic range, geographical coordinates, and (N) number of individuals.

			Coordinates		
Country	Pop	Range	Longitude	Latitude	N
Japan	Jap	Asia	139.43	35.67	18
China	S1	Asia	115.56	37.55	22
	S2	Asia	117.12	34.81	17
	S3	Asia	118.59	33.19	8
	S4	Asia	118.57	31.40	19
	S6	Asia	118.97	30.63	15
	S9	Asia	119.57	28.12	16
	S10	Asia	110.32	25.27	21
	S11	Asia	113.11	29.15	15
	S13	Asia	111.55	32.56	14
	S14	Asia	110.99	34.62	10
	S15	Asia	122.52	40.10	18
	S16	Asia	124.99	45.03	13
	S17	Asia	122.93	42.64	22
	S18	Asia	118.27	40.9	10
	S19	Asia	116.89	43.3	16
	S20	Asia	115.17	24.74	20
Tibet	Tib	Asia	99.94	23.56	26
Austria	Aus	Europe	99.53	23.35	17
Belgium	Bel	Europe	94.37	29.63	17
Bulgaria	Bul1	Europe	14.72	48.19	10
	Bul2	Europe	4.8	50.94	10
Hungary	Hun	Europe	26.85	44.06	22
Iran	Ira	Europe	26.7	43.99	10
Italy	Ita	Europe	18.87	46.63	18
Poland	Pol	Europe	54.78	37.05	12
Spain	Spa	Europe	10.52	44.77	19
Turkey	Tur	Europe	17.19	51.19	19
UK	UK	Europe	2.53	41.57	14

Table 2. Inference of the demographic model based on Neural Network and Random Forest. Selected models at each step are given for each algorithm (bold values are those of the selected model). 10,000 simulated datasets per scenario for training. Prior error rate estimated with leave-one-out cross-validation for Neural network and out-of-bag error rate for Random Forest. Random Forest does not provide posterior probability of each scenario, but rather the number of votes in favor to each one. Subsequently, posterior probability of the Random Forest corresponds to the posterior probability of the selected model.

Scenario	Neural Network						Random Forest						
	Prior error rate	Posterior probabilities				Selected	Prior error rate	Number of votes				Posterior probability	Selected
		1	2	3	4			1	2	3	4		
Step 1. Source populations of invasive demes													
Western E.	0.03	0.03	0.02	0.1	**0.85**	4	0.05	175	175	192	**458**	0.87	4
Western E. (w/o Italy)	0.025	0.03	0.04	0.02	**0.92**	4	0.04	172	188	164	**476**	0.85	4
Eastern E.	0.04	0.02	0.03	0.03	**0.92**	4	0.06	104	121	136	**639**	0.79	4
Iran	0.04	0.04	0.12	0.04	**0.80**	4	0.07	60	210	36	**694**	0.8	4
Step 2. Global invasion pathways													
Global	0.09	**0.99**	0.006	0.004	-	1	0.03	**859**	124	17	-	0.47	1

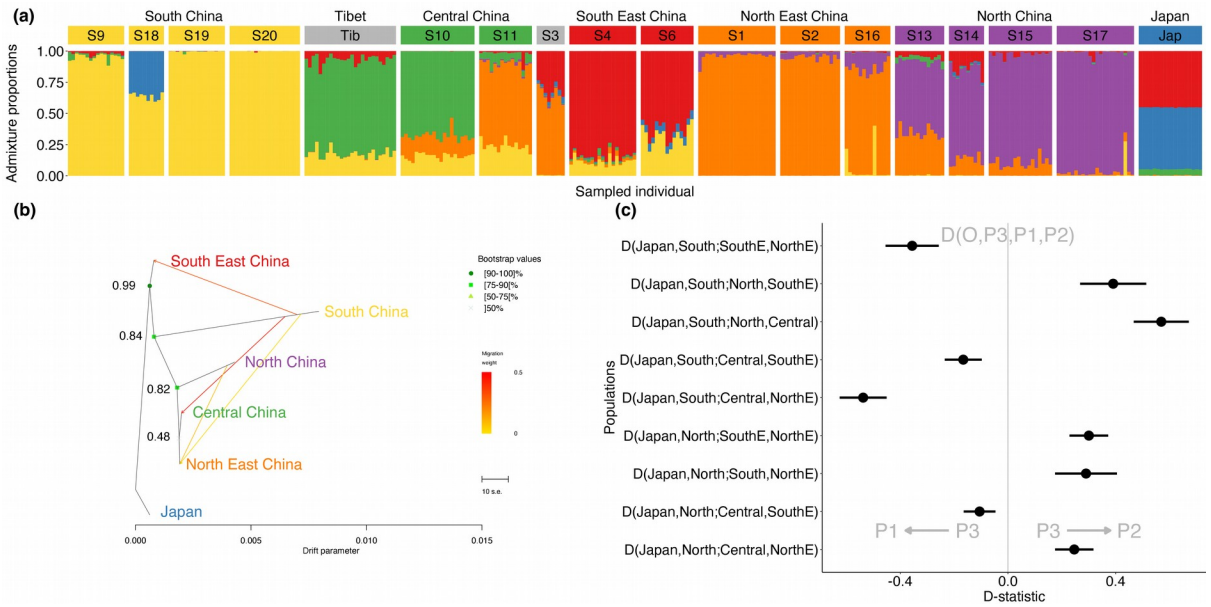

Fig. 1 Genetic structure and population modeling in the native range. **(a)** Posterior admixture proportions of individuals estimated with STRUCTURE for 18 sampled sites (n=300) in the native range (Asia) and K=6 genetic clusters (proportions aggregated with CLUMPP from 20 independent replicates). **(b)** ML tree based on TreeMix with block size of 500 SNPs. The migration edges pointing from South China to Central China, South Central China, and North East China have migration weights of 45%, 33%, and 2,7%. The migration edge pointing from North China to North East China has a migration weight of 9%. **(c)** D-statistics for testing admixture in native populations. Populations are noted D(O, P3, P1, P2). Negative values indicate gene flow between P3 and P1 while positive values indicate gene flow between P3 and P2. Confidence interval at 95% of the D-statistic estimated by block jackknife of 5kb. Null expectation is a D-statistic of 0. Demes abbreviations are North=North China, South=South China, Central=Central China, NorthE=North East China, SouthE=South East China

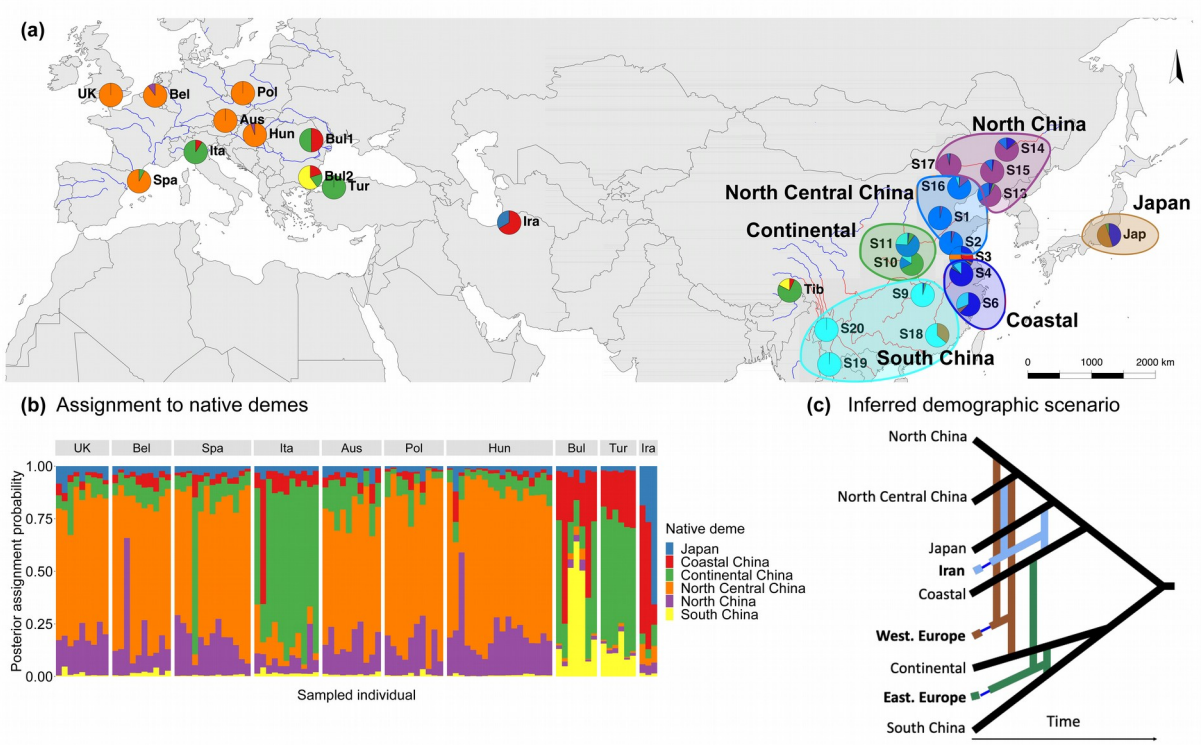

Fig. 2 Definition of demes in the native range, and source populations predicted for the invasive range. **(a) (Native range, Asia)** Mean admixture proportions estimated with STRUCTURE in the native range (Asia), for the chosen K=6. Native sampled sites are pooled into putative demes for subsequent analyses. Pie chart colors correspond to the genetic clusters inferred by STRUCTURE, as in Figure 1. **(Invasive range, Europe & Middle-East)** Assignment predictions of invasive individuals to native demes with AssignPOP's SVM algorithm with a relative posterior probability >2 (n=292 for training, n=100 for predictions). Pie chart colors correspond to the putative demes defined in the native range to which invasive individuals were assigned. **(b)** Posterior assignment probabilities to putative native demes estimated with AssignPop. **(c)** The demo-genetic scenario that was inferred with Approximate Bayesian Computation demonstrating three independent introductions from three independent admixed source populations. Branch lengths are not scaled. Bottleneck events are represented in thin red lines in branches. Colored branches correspond to invasive demes history.

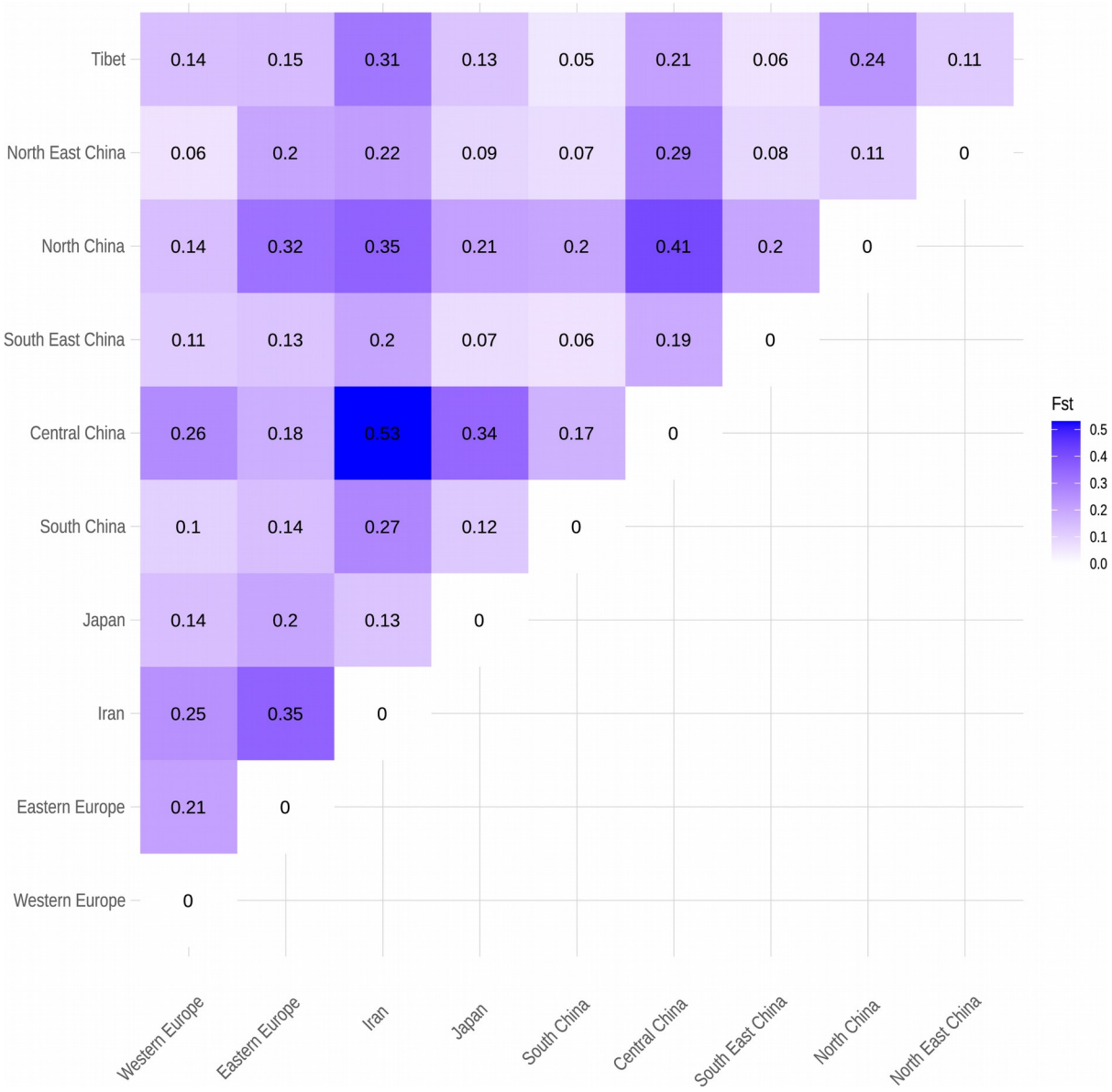

Fig. 3 Population differentiation. Pairwise F_{ST} (Weir & Cockerham's F_{ST}) between native and invasive demes. F_{ST} values were estimated from 3,000 loci.