ECOGRAPHY

Research

Historical demography and climate driven distributional changes in a widespread Neotropical freshwater species with high economic importance

E. A. Oliveira, M. F. Perez, L. A. C. Bertollo, C. C. Gestich, P. Ráb, T. Ezaz, F. H. S. Souza, P. Viana, E. Feldberg, E. H. C. Oliveira and M. B. Cioffi

E. A. Oliveira, M. F. Perez (https://orcid.org/0000-0002-4642-7793) ✉ (manolofperez@gmail.com), L. A. C. Bertollo, C. C. Gestich, F. H. S. Souza and M. B. Cioffi, Depto de Genética e Evolução, Univ. Federal de São Carlos (UFSCar), São Carlos, SP, Brazil. EAO also at: Secretaria de Estado de Educação de Mato Grosso – SEDUC-MT, Cuiabá, MT, Brazil. – P. Ráb, Laboratory of Fish Genetics, Inst. of Animal Physiology and Genetics, Czech Academy of Sciences, Liběchov, Czech Republic. – T. Ezaz, Inst. for Applied Ecology, Univ. of Canberra, Canberra, ACT, Australia. – P. Viana and E. Feldberg, Inst. Nacional de Pesquisas da Amazônia, Coordenação de Biodiversidade, Laboratório de Genética Animal, Petrópolis, CEP, Brazil. – E. H. C. Oliveira, Laboratório de Cultura de Tecidos e Citogenética, SAMAM, Inst. Evandro Chagas, Ananindeua, PA, Brazil.

Ecography
43: 1–14, 2020
doi: 10.1111/ecog.04874

Subject Editor: Gretta Pecl
Editor-in-Chief:
Jens-Christian Svenning
Accepted 10 May 2020

The Neotropical region exhibits the greatest worldwide diversity and the diversification history of several clades is related to the puzzling geomorphologic and climatic history of this region. The freshwater Amazon ecoregion contains the main hydrographic basins of the Neotropical region that are highly dendritic and ecologically diverse. It contains a rich and endemic fish fauna, including one of its most iconic and economically important representatives, the bony-tongue *Arapaima gigas* (Teleostei, Osteoglossiformes). Here, we evaluated the projected distribution of the genus in different historical periods (Present, Last Glacial Maximum, Last Interglacial Maximum and Near Future) and interpreted these results in light of the genomic diversity and modeled historical demography. For that, we combined species distribution models, population genetic analysis using SNPs and deep learning model selection. We analyzed a representative sample of the genus from the two basins where it naturally occurs, four localities in the Amazon (Am) and three in the Tocantins-Araguaia (To-Ar) basin, as well as individuals from three fish farms. We inferred a potentially smaller distribution in the glacial period, with a possible refuge in central Am. Our genetic data agrees with this result, suggesting a higher level of genetic diversity in the Am basin, compared to that observed in To-Ar. Our deep learning model comparison indicated that the To-Ar basin was colonized by the population from the Am basin. Considering a global warming scenario in the near future, *A. gigas* could reach an even larger range, especially if anthropogenic related dispersal occurs, potentially invading new areas and impacting their communities.

Keywords: climate change, DArTseq, deep learning, fish, historical demography, neotropical diversity

NORDIC SOCIETY OIKOS

www.ecography.org

Introduction

The Neotropical region exhibits one of the greatest biodiversity levels worldwide (Antonelli et al. 2018a, Rull 2018). Within this region, Amazonia is identified as the major source of this outstanding species richness (Antonelli et al. 2018b, Fine and Lohmann 2018). The Amazon River (Am) basin contains the most speciose fish fauna in the world (Reis et al. 2016), a likely result of its complex geomorphologic and climatic history (Figueiredo et al. 2009, 2010, Hoorn et al. 2017, van Soelen et al. 2017). This intricate history is also reflected by puzzling biogeographic patterns (Ribas et al. 2012, Dagosta and Pinna 2017). The Tocantins-Araguaia basin (To-Ar), although not a real tributary of the Amazon basin, as it flows directly to the Atlantic (Carvalho and Albert 2011), is considered part of the freshwater Amazon ecoregion (Albert and Reis 2011, Dagosta and Pinna 2017). A high level of endemic species is observed in To-Ar basin (higher than other basins in the Brazilian shield and comparable to the levels observed in lowland Amazonia), probably due to historical connections to the Am basin (Rossetti and Valeriano 2007, Albert and Reis 2011, Dagosta and Pinna 2017). Assessing genetic diversity in taxa occurring in regions with such puzzling geomorphologic history may provide insights on the evolutive processes acting on the biodiversity of these basins.

Among the fish orders with representatives in both Am and To-Ar basins, the Osteoglossiformes is one of the first three sister lineages to all other modern teleosts (Greenwood et al. 1966, Arratia 1999, Near et al. 2012, Hilton and Lavoué 2018) and living forms are restricted to freshwater (Myers 1949). Osteoglossiformes species are naturally distributed or introduced to all continents, with the exception of Antarctica (Adite et al. 2005, Nelson et al. 2016, Hilton and Lavoué 2018). Currently, this order comprises six families, namely Pantodontidae, Notopteridae, Gymnarchidae, Mormyridae, Osteoglossidae and Arapaimidae. However, authors recently have considered Arapaimidae as subfamily Arapaiminae, within the Osteoglossidae family (Nelson et al. 2016, Cavin 2017). The Arapaiminae subfamily is represented by only two extant genera, the monotypic African *Heterotis* Rüppell, 1829, and the South American *Arapaima* Müller, 1843 (Nelson et al. 2016). The genus *Arapaima* is considered monotypic by many authors, with *A. gigas* as the only valid species (but see Castello et al. 2013, Stewart 2013a, b). Stewart (2013a, b) carefully analyzed existing types of this species and advocated that they might represent easily diagnosable species. As a result, the author formally described the new species *A. leptosoma* (Stewart 2013b) and validated a previously described species (Stewart 2013a). Therefore, some authors accept as much as five species for *Arapaima* (Castello et al. 2013). However, recent publications that analyzed the genetic diversity using molecular markers, considered *Arapaima* as a monotypic genus (Hrbek et al. 2005, 2007, Farias et al. 2019, Torati et al. 2019). Here, taking into account the morphological classification performed in the museum after the voucher deposit, we decided to consider the specimens in all our analyses as monotypic (*A. gigas*), but discussing our results considering the potential for cryptic species to occur.

Arapaima gigas is widely distributed across a large portion of the To-Ar and Am basins in Brazil, Peru, Colombia, Ecuador and Guyana (Reis et al. 2003, Castello 2008), and represents one of the largest freshwater fish species, with some individuals reaching up to 200 kg of body mass and up to three meters in length (Stone 2007, Bezerra et al. 2013, Nelson et al. 2016). The species presents rapid growth, typically reaching ~60–80 cm in the first year of life (Arantes et al. 2010) and is considered one of the fish species with highest aquaculture potential (Ono 2007), because of its high nutrition value and low fat (< 5%) content (dos Santos Fogaça et al. 2011).

Previous studies on the population genetic diversity of *A. gigas* with molecular markers generally pointed to the absence of genetic structuring in the Am and lower To-Ar basins, and a more pronounced structuring in the upper To-Ar. These previous analyses, however, focused on few mtDNA sequences (Hrbek et al. 2005), nuclear markers associated with repetitive regions (Hrbek et al. 2007, Vitorino et al. 2015, 2017, Farias et al. 2019), and randomly distributed genotypes along the genome (Torati et al. 2019). Here, we sampled specimens in the field and deposited vouchers for both wild populations and fish farms, obtaining SNP markers with DArTSeq (Kilian et al. 2012). Differently from most technologies that obtain random genomic sequences, such as CRoPS (Complexity Reduction of Polymorphic Sequences), GBS (Genotype by Sequencing) and RAD (Restriction Site Associated DNA) (van Orsouw et al. 2007, Baird et al. 2008, Elshire et al. 2011), DArT (Diversity Arrays Technology) is a genome-complexity reduction method that enriches for hypomethylated regions of the genome allowing active genomic regions to be recovered (Jaccoud et al. 2001, Kilian et al. 2012). Therefore, it is possible to detect markers that may be under the effect of selective pressures. Besides, when coupled with next-generation sequencing (NGS) technologies (DArTSeq), thousands of SNPs can be generated in a relatively short amount of time, making them powerful tools for genomic investigation (Kilian et al. 2012).

By using such datasets containing hundreds to thousands of loci (Garrick et al. 2015), coupled with demographic model selection strategies (Carstens et al. 2013a), it is possible to compare complex scenarios and make more accurate estimates of demographic parameters from more realistic models (Knowles 2009, Thomé and Carstens 2016). Recently, deep learning methods were incorporated in population genetics (Sheehan and Song 2016, Schrider and Kern 2018 and references therein) and applied for demographic model comparison (Flagel et al. 2019, Villanea and Schraiber 2019). These procedures have the advantage of making use of the information present in the large and multivariate datasets obtained with NGS, without the need of reducing this information with summary statistics (Flagel et al. 2019).

Here, we used species distribution models (SDMs) to assess the potential range of *A. gigas* during past and future

climate change. We also obtained genotypic data for thousands of loci with DArTSeq procedure, that were used to estimate the genetic diversity in different sampling sites and to estimate population structure along *A. gigas* distribution. The SDMs and genetic diversity results were used to generate demographic models for *A. gigas*, that were compared against each other with a deep learning approach and explicitly test whether: 1) specimens from the two basins can be considered a single genetic population; 2) populations from both basins expanded their range to reach the current distribution; 3) the population located in Am was colonized from the To-Ar basin with a founding event followed by expansion; or 4) the opposite colonization pathway occurred, with To-Ar being colonized from Am.

Material and methods

Individuals examined and DNA extraction

We collected *A. gigas* individuals from seven localities in Am (four sampling sites) and To-Ar (three sampling sites) river basins. Besides, we analyzed samples from three different fish farms (Fig. 1 and Table 1). We sampled individuals using traps, and after capture, the animals were transported to the research station. The Brazilian environmental agencies ICMBIO/SISBIO (license no. 48290-1) and SISGEN (A96FF09) authorized the collections and voucher individuals were identified and deposited (Table 1) in the fish collections of the Museu de Zoologia da Univ. de São Paulo (MZUSP). We collected liver fragments of all individuals and stored them in 100% ethanol for DNA extraction, following Sambrook and Russell (2001). The procedures followed ethical and anesthesia procedures, in accordance with the Ethics Committee on Animal Experimentation of the Univ. Federal de São Carlos (process number CEUA 9506260315).

Paleogeographic modeling

The climatic niche for *A. gigas* was estimated from 85 wild occurrences (Fig. 2A) based on our field collections (7 points), in previous published works (Hrbek et al. 2007, 11 points; Torati et al. 2019, 4 points; Vitorino et al. 2015, 4 points; Vitorino et al. 2017, 1 point) and 57 available points in the Global Biodiversity Information Facility (GBIF) that were manually checked to avoid inconsistencies. We performed a combination of nine distribution algorithms with biomod2 (Thuiller et al. 2009), including generalized linear models (GLM; McCullagh and Nelder 1989), multivariate adaptive regression splines (MARS; Friedman 1991), classification tree analysis (CTA; Breiman et al. 1984), mixture discriminant analysis (MDA; Hastie et al. 1994), artificial neural networks (ANN; Ripley 2014), generalised boosted models (GBM; Ridgeway 1999), random forests (Breiman 2001), surface range envelop (SRE; Busby 1991) and Maximum Entropy (Maxent; Phillips et al. 2006). We used 15 bioclimatic variables from the 19 available in WorldClim (Hijmans et al. 2005) for the CCSM4 circulation model. Variables 8, 9, 18 and 19 were omitted for having artificial breaks (Bonatelli et al. 2014). To avoid high correlations between variables, for all pairwise comparisons with Pearson index > 0.85, only the variable with higher explanatory capacity was kept, after a preliminary run.

Model calibration was carried out with present climatic conditions and a 30 arc-seconds resolution, while projections for the present, last glacial maximum (LGM, 21 kya), last interglacial maximum (LIG, 120 kya) and future (2070) were performed with a 2.5′ arc-minutes resolution. A total of 5000 pseudoabsence points was simulated, using the SRE strategy with a quantile threshold of 0.005. We adopted a proportional weighted mean ensemble method with five simulations for each algorithm, keeping only simulations with TSS higher than 0.7.

DNA extraction and DArTseq genotyping

The gDNAs of all sampled individuals were analyzed under the DArTseq technology (Kilian et al. 2012) by the Diversity Arrays Technology Company (Canberra, Australia). A combination of PstI and SphI enzymes was used to construct the libraries using methods described by Kilian et al. (2012) and sequenced on the Illumina Hiseq2500 next-generation sequencer. Raw data generated by sequencing were filtered, processed and converted to high-quality genotypes by the facility, using their proprietary DArTsoft14 v1.0 software. The following filters were used to obtain the SNP markers: 1) overall call rate over 95%; 2) polymorphic information content (PIC) between 0.3 and 0.5; 3) Q-value (that measures the false discovery rate) above 2.5 and 4) minimum allele frequency of 0.5%. Linkage Disequilibrium and deviations from Hardy–Weinberg Equilibrium were estimated with dartR (Gruber et al. 2018). Genotypes were coded as an SNP matrix with loci in the rows and individuals in the columns. For each genotype, data were stored as 0 for state homozygotes, 1 for heterozygotes and 2 for alternate state homozygotes (Dryad doi: 10.5061/dryad.4qrfj6q7j).

Outlier loci detection

Signatures of directional selection were assessed for all sampling sites with the Bayesian method implemented in BayeScan (Foll and Gaggiotti 2008) with a Prior Odds of 100 and a False Discovery Rate of 0.01. Gene ontology (GO) and annotation of detected candidate loci were then evaluated in Blast2Go (Conesa et al. 2005). To achieve this, flanking regions of outlier SNPs were blasted against the National Center for Biotechnology Information (NCBI) nonredundant nucleotide database and annotated with an e-value threshold of 10^{-3} and 10^{-6}, respectively.

Genetic diversity and isolation by distance

Summary statistics for genetic diversity were calculated using GENODIVE (Meirmans and van Tienderen 2004) for each

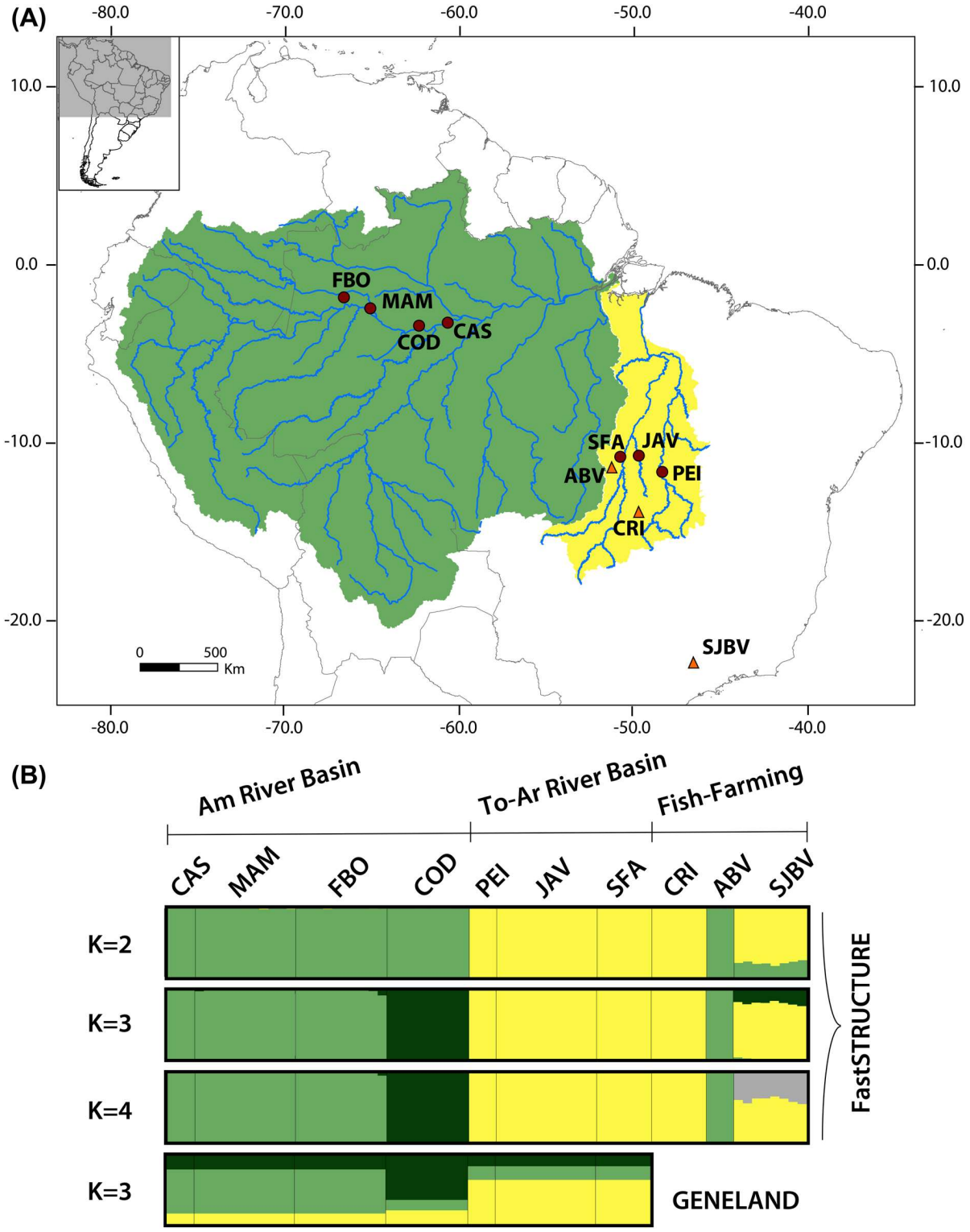

Figure 1. (A) Map of northern South America indicating the sampling sites of *Arapaima gigas* analyzed in this study from Tocantins-Araguaia (yellow) and Amazon (green) river basins, coded according to Table 1. Fish farming sampling sites are represented as orange triangles, while natural samplings are shown as red circles. (B) fastSTRUCTURE results for K from 2 to 4 and GENELAND analysis for K = 3. Each sampling site is represented as a vertical bar showing the proportion of their genome belonging to each of the K groups. Black lines separate individuals of different sampled localities.

Table 1. Sample information from *Arapaima* specimens used to develop SNPs with DArTseq and to perform genetic analyses.

Locality description	DArTseq samples	Code	Basin	Geographical coordinates	Voucher number
Castanho Lake, Manaus, AM	03	CAS	Am	3°42′48.4″S, 60°31′11.6″W	123955
Panta Leão Lake, Mamirauá Reserve, Alvarães, AM	11	MAM	Am	2°54′58.1″S, 64°49′29.1″W	123953
Lakes of the Juruá River, Mariana Sector, Fonte Boa, AM	10	FBO	Am	2°19′05.1″S, 66°16′39.5″W	–
Onças Lake, Codajás, AM	09	COD	Am	3°53′17.1″S, 62°07′36.2″W	123954
Marginal lake to the Santa Tereza River, tributary of the Tocantins River, Peixes, TO	03	PEI	To-Ar	11°54′31.7″S, 48°38′02.7″W	121642
Marginal lake to the Javaé River, tributary of the Araguaia River, Lagoa da Confusão, TO	11	JAV	To-Ar	11°00′27.1″S, 49°56′01.8″W	121639
Xavantinho River, tributary of the Araguaia River, São Félix do Araguaia, MT	06	SFA	To-Ar	11°42′07.2″S, 50°50′15.4″W	121643
Liberdade fish facility, Uirapuru and Crixás, GO	06	CRI	Farm	14°05′45.4″S, 49°55′18.3″W	121644
Mr. Roberto fish facility, São Félix do Araguaia, MT	03	ABV	Farm	11°39′33.1″S, 51°26′23.3″W	121641
Rio Doce fish facility, São João da Boa Vista, SP	08	SJBV	Farm	22°01′14.9″S, 46°54′08.1″W	121645

sampled locality, as estimates of expected heterozygosity (H_E), observed heterozygosity (H_O) and inbreeding coefficient (G_{IS}). Allelic richness (A_R) was calculated to correct for heterogeneous sample sizes, by using the rarefaction method in diveRsity (Keenan et al. 2013). Pairwise F_{ST} (Weir and Cockerham 1984) was also calculated among sampling sites, with significance evaluated using 10 000 permutations after a Bonferroni correction with $\alpha = 0.05$.

Isolation by distance (IBD) was tested only for natural-occurrence populations and for each basin separately.

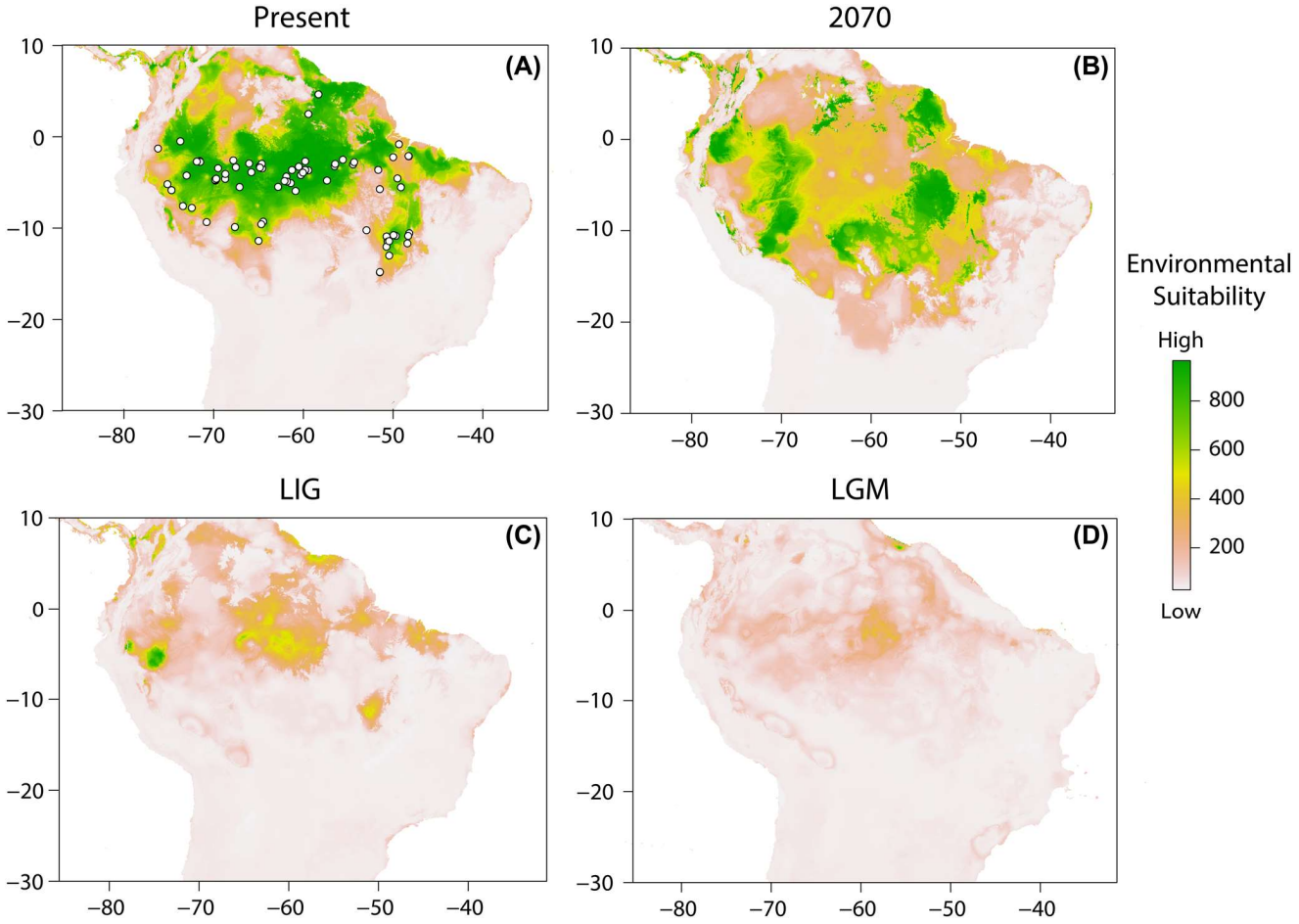

Figure 2. Climatic distribution modeling in *A. gigas*, constructed based on current distribution points. Suitable climatic areas are shown according to a gradient for present (A), future (2070 – B), last interglacial maximum (LIG, 120 kya – C) and last glacial maximum (LGM, 21 kya – D).

The straight-line distance was used between populations SFA and JAV in To-Ar, as they get connected in the wet season. For the Am basin and comparisons considering PEI in To-Ar, stream distances were used. We performed a Mantel test (Mantel 1967) and canonical redundancy analysis (RDA), a method combining PCA and multiple regressions, which decomposes the genetic variance based on allele frequencies (Orsini et al. 2012). For RDA, stream and straight-line distances were transformed into coordinates with R-command cmdscale. The obtained spatial coordinates were then converted to third-degree orthogonal polynomials, with a modified version of the scripts from Meirmans (2015). We also obtained the spatial component of the total genetic variation by multiplying the percentage of constrained variation by the overall value of F_{ST}, as suggested by Meirmans (2015).

Population structure

Population genetic structure for all collected sampling sites was investigated with the non-spatial method fastSTRUCTURE ver. 1.0 (Raj et al. 2014). This method is a variation of the popular Bayesian clustering method STRUCTURE (Pritchard et al. 2000), optimized for large genotype datasets. Data preparation and analysis were performed with the aid of the 'lizards-are-awesome' pipeline (Melville et al. 2017). Population genetic structure of the wild occurring populations was also assessed with the spatially explicit strategy implemented in GENELAND (Guillot et al. 2011), under the correlated frequencies prior with 500 000 iterations and a thinning of 200. Runs in fastSTRUCTURE were repeated for a range of K (number of populations) from 1 to 11 and in GENELAND from 1 to 8. Results from both analyses were processed with the online tool CLUMPAK (Kopelman et al. 2015), which simplifies the use of DISTRUCT (Rosenberg 2004) and CLUMPP (Jakobsson and Rosenberg 2007) to summarize and plot the results, respectively.

Demographic model selection and parameter estimation

We simulated genetic data similar to our wild-occurrence dataset in ms (Hudson 2002), considering four possible scenarios for the demographic history of *A. gigas*: 1) one panmictic population harboring all samples collected from the two basins studied here; 2) a reduction in the ancestral population, followed by expansion on both basins; 3) colonization of the Am basin from the To-Ar, simulated as a founding event followed by exponential population expansion; and 4) To-Ar basin being colonized from Am, also simulated as a founder event followed by expansion (Fig. 3). Such scenarios were conceived based on the results of the population structure estimates (Fig. 1B) and the SDM projections (Fig. 2).

To perform our coalescent simulations, we adopted a uniform prior for generation time, spanning from 4 to 5 yr (Hrbek et al. 2005). A mutation rate (μ) of 1.25×10^{-9} mutations per site per year was used, calculated from the splitting times and amount of genome differences from *A. gigas* to *Scleropages formosus* (Vialle et al. 2018), the closest species with a sequenced genome. We performed data simulations (20 000 for each model) with scripts modified from Perez et al. (2016), using empirical sample sizes. Values of θ were calculated for each simulation using μ and the effective population size (N_e) sampled from a uniform distribution from 100 to 500 000 individuals (Hrbek et al. 2005 suggest ca 150 000 females killed by year in the transition of the 19th to 20th centuries, based on harvest estimates). Divergence time for Am and To-Ar basin (τ_2) was sampled from a uniform distribution from 200 thousand yr ago (kya) to 2 million yr ago (Mya). This period of time includes the age estimate of Tocantins river achieving its modern course (Plio-Pleistocene boundary, Rossetti and Valeriano 2007; 1.8 million yr ago (Mya), Silva-Santos et al. 2018). For divergence time of population COD from other Am localities, a uniform distribution between 0 and 200 kya was used. Founder effect magnitude during colonization (θrF-A; used in the models 2 and 3), was computed as the ratio between the θ values during the colonization and the value for the ancient population (drawn from a uniform distribution ranging from 0.001 to 0.1). The intensity of population expansion after colonization (θrC-A) was estimated as the ratio between the θ value in the current and in the ancient population (sampled from 0.1 to 1).

Competing demographic scenarios were compared using a recent approach described in Flagel et al. (2019). This strategy is based on converting the SNP matrices into images and extracting information via convolutional neural networks (CNN; for a review on the main architectures and applications of CNN see Christin et al. 2019). We loaded our simulated data from ms into python as NumPy arrays containing individuals in the lines and loci as columns. The genotypes were coded as 0 (black) for the reference state and 1 (white) as the alternate state. Then, for each simulation, we clustered individuals by genetic distance and transposed the matrix to keep each individual in a column and markers as lines (Supplementary material Appendix 1 Fig. A1). The resulting NumPy arrays containing all simulations were then shuffled and 10 000 random simulations were separated to be used as a validation set, while the remaining 50 000 were used as training data. The training data was submitted to a CNN based on the architecture suggested by Flagel et al. (2019) with slight modifications (Supplementary material Appendix 1 Fig. A1). Briefly, it consists in three (the first layer containing 250 and the other two 125 neurons) one-dimensional convolutional layers with a kernel size of 2. These convolutional layers apply filters (kernels) by sliding them along (convolving) the input image and generate new layers by calculating the scalar product of the kernel and the image being convolved. By doing such operations, they extract features of the input image, such as edges and formats. Our convolutional layers were interleaved with average-pooling layers, that recover the average value of the area covered by the kernel, reducing the dimensionality by extracting dominant features of the data. Then, two fully connected layers with 125 neurons connect all the previous neurons (flattened in a single dimensional

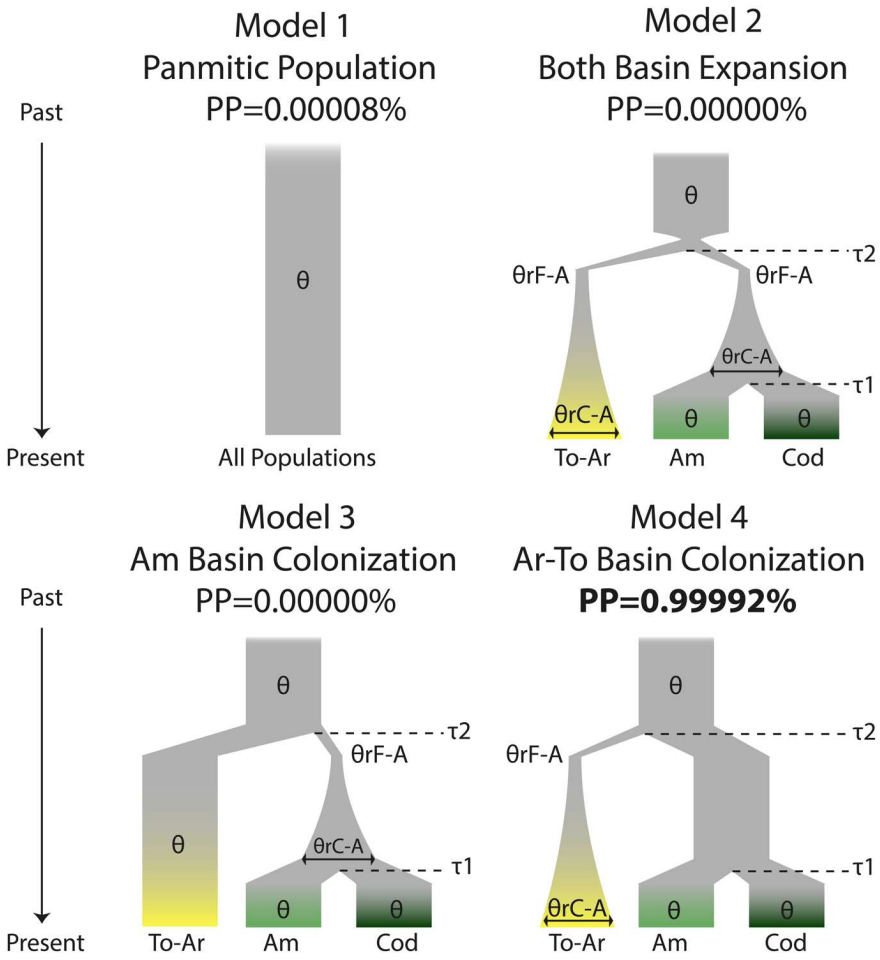

Figure 3. Graphic representation of the simulated scenarios tested in *A. gigas*. Model 1 considers the sampled distribution as a simple panmictic population at all times; model 2 simulates a reduction followed by expansion on both basins; model 3 simulates a colonization event in the Amazon basin; and model 4 a colonization event in the Tocantins-Araguaia basin. Symbols represent the estimated parameters of theta in the current populations (θ) in all models, as well as divergence times between basins (τ_2) and of the COD population (τ_1), population sizes during founder event (θrF-A) and ratio of population sizes during and after the founder event (θrC-A). PP – posterior probability.

layer). Finally, these layers compute the probability for each model using a sigmoid function with a final output layer with four neurons, corresponding to the four scenarios used to simulate the data. The CNN was run with a mini-batch size of 250. Rectified linear unit activation functions, that usually learn image patterns faster, were used with the convolutional layers together with a dropout (to avoid overfitting) of 25% and 50% of neurons after pooling and densely connected layers, respectively. We evaluated the learning performance with a loss function of categorical cross-entropy and updated the network weights during training with Adam optimization (Kingma and Ba 2015).

After training the neural network, we performed a cross-validation power analysis of our CNN approach using 2000 simulations per model to evaluate the performance of our method to identify correctly the simulated scenario (Supplementary material Appendix 1 Fig. A2). After that, our empirical dataset was submitted to the trained CNN to select the most likely demographic scenario for *A. gigas* (Fig. 3). After selecting the preferred scenario, we conducted 100 000 simulations for parameter estimation, using the same CNN architecture described above. Accuracy for estimates of each parameter was evaluated with root mean square error (RMSE), a measure of the standard deviation of the residuals, and Spearman's ρ. All scripts used in model comparison and parameter estimation are available in github (<https://github.com/manolofperez/CNN_DemographyArapaima>).

Results

Paleogeographic modeling

After evaluating variable correlations and explanatory capacity, only seven bioclimatic variables (bio 1 – annual mean temperature, 2 – mean diurnal temperature range,

3 – isothermality, 4 – temperature seasonality, 14 – precipitation of driest month, 15 – precipitation seasonality and 16 – precipitation of wettest quarter) were maintained. Projections for present showed the most continuous range. Although largely congruent with *A. gigas* occurrence points, the potential current distribution was more widespread and predicted areas that are outside floodplains and within rapids systems, where the species is not expected to occur (Fig. 2A). The future projection was the most widespread of the modelled periods (Fig. 2B), with suitable areas outside the contemporary species distribution, with a highly suitable area in central Brazil and another in western Amazonia. A similar result for future conditions was observed by Oberdorff et al. (2015), that projected increasing distributional ranges for *A. gigas* in future periods. The LGM projection exhibited the most restricted range, with a stable area in the central Amazon and another in the northern coastline of South America. The projection for the LIG (Fig. 2C) presented more stable areas than the LGM (Fig. 2D), especially in central and western Amazonia, in the To-Ar basin, and in a few areas along the northern South American coast.

DArTseq genotyping and genetic diversity

Sequencing of DArTSeq markers resulted in an average of 2 559 000 reads per sample. A total of 2364 high-quality filtered SNPs, with an average read depth after filtering of 43.1, were obtained in the 70 samples genotyped, with 3.14% missing data. A minor allele frequency of 16% in average was observed, 4% of loci comparisons showed significant LD and 7% of the loci showed significant HWE deviation after Bonferroni correction (alpha = 0.01). We detected 18 loci as candidate outliers in BayeScan. Only four of them returned blast hits and presented GO terms related to DNA-binding transcription factor activity (GO: 0000113), heparan sulfate sulfotransferase activity (GO: 0034483), TIMP family protein binding (GO: 0098769) and LEM domain binding (GO: 0097726). We maintained candidate SNPs for all further analyses, as their removal rendered similar results (data not shown).

Diversity levels were higher for Am Basin populations when compared with To-Ar for all diversity indexes calculated, with H_E and H_O values at least one order of magnitude higher (Table 2, Fig. 4). Samples from ABV and SJBV fish farms showed diversity levels similar to Am basin localities, while the diversity levels of the individuals from CRI fish farm were similar to the To-Ar basin (Table 2, Fig. 4). The inbreeding estimator (G_{IS}) presented negative values, that indicates outbreeding, in most localities, except for MAM and FBO (Am basin), with 0.052 and 0.130, respectively, JAV in the To-Ar basin (0.004), and ABV fish farm (0.113). Pairwise F_{ST} ranged from 0.035 (between MAM and FBO) to 0.771 (between CAS and JAV). In general, higher values were observed in pairwise comparisons of Am basin and To-Ar localities (Supplementary material Appendix 1 Fig. A3). After applying a Bonferroni correction, all F_{ST} values were significant (Supplementary material Appendix 1 Fig. A3). Fish farming samples showed diverse patterns, with CRI showing higher F_{ST} values when compared with Am basin localities, ABV more dissimilar to To-Ar localities, and SJBV showing moderate values of F_{ST} with all other localities (Supplementary material Appendix 1 Fig. A3).

Isolation by distance

Comparisons of genetic and geographic distance with Mantel test suggested a small non-significant correlation between these two variables in Am basin (r = 0.3057; p = 0.2917) and high non-significant relationship in To-Ar basin (r = 0.8771; p = 0.1250). Redundancy analysis pointed to a significant correlation in Am basin (RDA = 0.5532; p = 0.0417), that resulted in a value that indicates absence of IBD when multiplied by F_{ST} (RDA*F_{ST} = 0.0457), according to Meirmans (2015). RDA for To-Ar basin localities was not able to select any variables, suggesting an absence of correlation between geographic and genetic distances.

Population structure

Results from the chooseK command in fastSTRUCTURE suggested a maximum marginal likelihood with K = 2, and K = 4 as the model complexity required to explain the data. Therefore, we decided to show clustering results for 2–4 groups. All results grouped all To-Ar localities, along with CRI fish farm. Sampling sites from Am basin were also grouped when K = 2 was used, alongside with samples from

Table 2. Genetic diversity levels in *Arapaima* sample sites. *A* – average number of alleles; A_R – allelic richness; H_O – observed heterozygosity; H_E – expected heterozygosity; G_{IS} – inbreeding coefficient.

Locality	Basin	*A*	A_R	H_O	H_E	G_{IS}
CAS	Am	1.226	1.532	0.190	0.142	−0.336
MAM	Am	1.361	1.700	0.218	0.230	0.052
FBO	Am	1.355	1.652	0.200	0.229	0.130
COD	Am	1.306	1.589	0.229	0.183	−0.255
PEI	To-Ar	1.023	1.217	0.019	0.015	−0.265
JAV	To-Ar	1.064	1.240	0.040	0.040	0.004
SFA	To-Ar	1.058	1.239	0.044	0.037	−0.200
CRI	Farm	1.048	1.225	0.037	0.030	−0.233
ABV	Farm	1.235	1.564	0.144	0.167	0.133
SJBV	Farm	1.186	1.417	0.144	0.112	−0.292

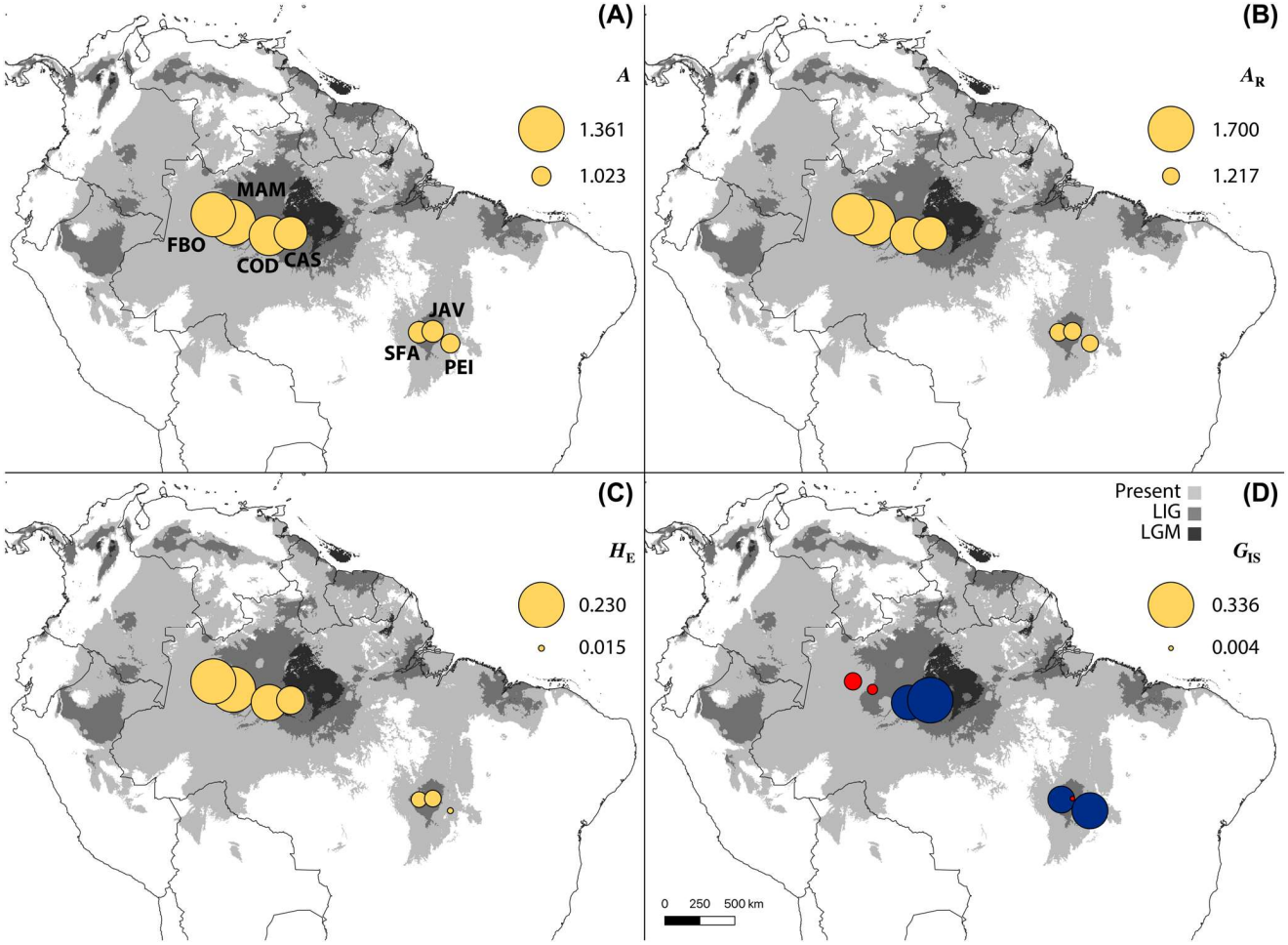

Figure 4. Map of northern South America showing the predicted suitable areas (environmental suitability higher than 200) for Present (light grey), LIG (grey) and LGM (dark grey). Circles represent the value of genetic diversity indexes (A) number of alleles – A; (B) allelic richness – A_R; (C) expected heterozygosity – H_E and (D) inbreeding coefficient – G_{IS} (with circle colors representing positive values in blue and negative in red). Circle sizes are proportional to the value of the genetic parameter, and the highest and lowest values are shown.

ABV fish farm. When K = 3 and K = 4 were used, COD was allocated alone in a new group. All samples from SJBV fish farm showed admixed ancestry, with part of their genomes assigned with samples from both basins, but with most of their genome belonging to To-Ar basin (Fig. 1). Geneland results suggested three as the optimum number of genetic clusters. The obtained result was largely congruent with K = 3 in fastSTRUCTURE, grouping all To-Ar samples in one group, COD alone in a second group and the remaining localities from Am basin in a third cluster (Fig. 1).

Demographic model selection

Based on the congruence of the results from the two clustering analyses performed (Fig. 1), we decided to use three groups in the simulations of the demographic scenarios (Fig. 3). After 20 epochs, our CNN showed an accuracy of 0.9007 and 0.8815 in the training and in the validation set, respectively. Our cross-validation procedure showed a high proportion of simulations correctly predicted to their generating model, and the scenarios for both basin expansion (model 2) and Am colonization (model 3) were the most difficult to differentiate with our approach, as they were confounded with each other (80.9 and 83.3% of correct predictions for model 2 and model 3, respectively). The panmictic scenario was the most easily diagnosable, as it showed very high proportions of correct predictions (99.9%) (Supplementary material Appendix 1 Fig. A2). When the empirical data was submitted to the trained CNN, the most likely scenario was To-Ar basin colonization (Fig. 3), with a posterior probability (PP) of 0.99992, while Am basin colonization and both basin expansion showed the lowest probabilities (PP = 0.00000 for both).

Parameter estimation (Table 3) based on the selected scenario suggested that our dataset contains information to estimate more accurately the N_e (RMSE = 0.186; Spearman's $\rho = 0.765$) and splitting time of the Am and To-Ar basins (τ_2; RMSE = 0.294; Spearman's $\rho = 0.757$). The divergence time for COD (τ_1; RMSE = 0.612; Spearman's $\rho = 0.621$), magnitude

of the founder event (θrF-A; RMSE = 0.738; Spearman's ρ = 0.498) and magnitude of growth since the founder event (θrC-A; RMSE = 0.665; Spearman's ρ = 0.516) showed lower estimation capacity. The recovered N_e values presented a similar magnitude of the estimates from Hrbek et al. (2005), with a median value of 144 089 individuals (interval = 135 560–152 782 individuals). The separation of the Am and To-Ar basins was estimated in the Pleistocene (median = 922.06 kya; interval = 864.81–977.48 kya).

Discussion

Our SDM projection for the present conditions recovered a potential distribution larger than the current natural occurrence of *A. gigas*. This result can be related to the use of only bioclimatic variables in our SDM approach, without incorporating the presence of floodplains or waterfalls in the models. Though useful, incorporating such information would preclude projections of *A. gigas* distribution in past and future periods. Similar strategies are being used by other authors when analyzing freshwater fishes (Bagley et al. 2013, Oberdorff et al. 2015, McMahan et al. 2017). The paleogeographic reconstructions obtained here indicated that during the last glacial period, *A. gigas* distribution was constrained under severe climatic conditions, with suitable climatic habitats scattered and restricted to refugial areas (Fig. 2C–D, 4). These population size fluctuations may have resulted in an accentuated genetic drift caused by bottlenecks (Wright 1931, Lande 1988), especially in To-Ar according to our demographic model results. However, care should be taken when correlating these two results, as the time period recovered for colonization of To-Ar is older than our SDM projections. Therefore, these approaches can be viewed as complementary, giving insights about different evolutionary periods in *A. gigas* history. Altogether, these features possibly played a major role in shaping the modern genetic diversity observed in *A. gigas*, implying that the pattern of differentiation observed between distinct populations is most probably affected by hydrological and historical climate features.

Another important result for the species conservation is that the fish farm ABV is located within the area of the To-Ar basin, but presented genotypes associated to the Am population (Fig. 1B). This is of special concern, as the distribution models for the future (Fig. 2B) indicate an invasive potential for areas outside the current species distribution. Many of those predicted areas outside the current distribution are in central Brazil, where climate change would increase temperatures and precipitation levels, promoting a reduction of savanna (Moncrieff et al. 2016). Though presenting a suitable climate, these areas present higher elevations, an absence of suitable lentic habitats for spawning, fast water currents and rapids, that would likely inhibit *A. gigas* invasion. However, some areas would still be potentially invaded, especially as the release of individuals from fish farms in nearby rivers could facilitate that effect. In fact, according to information from local farmers and fisherman during our sample expeditions, the locality PEI is probably the result of introductions from other To-Ar natural localities, as the species do not use to occur in the region before the 50s. Because the collected individuals were occurring in the wild and our pairwise F_{ST} results pointed to a unique genetic variation for this locality (Supplementary material Appendix 1 Fig. A3), we decided to analyze it together with the other wild occurrences. Such invasions can greatly impact the invaded communities, by spreading diseases and affecting the ecosystems, as *A. gigas* is one of the largest freshwater species and a generalist carnivore. Moreover, admixture of local populations with alien alleles can cause outbreeding depression (Weeks et al. 2011), potentially yielding infertile offspring. As the flood dynamics, associated with the reproductive biology of *Arapaima*, in the two basins are markedly different (Albert and Reis 2011), contamination of natural populations with introduced individuals from a different basin can have negative results in their fitness. Likewise, the reduced environmental suitability recovered in central Amazon (Fig. 2B) is also of conservation concern. This pattern can be related to a potential change in the Amazon phytophysiognomy towards a savannah-like habitat, as result of a drier climate (Nobre et al. 2016, Lovejoy and Nobre 2019).

Genetic diversity among populations

In agreement with our distribution models for the past, we detected lower genetic diversity levels for localities from the To-Ar basin compared to the Am basin ones, even when correcting for small sample sizes with A_R (Table 2, Fig. 4). Lower genetic diversity in To-Ar localities, especially in the upper portion of the basin was also observed in other *A. gigas* studies (Vitorino et al. 2015, 2017, Farias et al. 2019, Torati et al. 2019). In addition to the differences in the genetic diversity levels, we recovered structure separating populations from both basins and high significant F_{ST} values between them (higher than 0.7 for 9 comparisons; Supplementary material

Table 3. Parameter estimates obtained from the preferred demographic model. N_e – effective population size; τ_1 – divergence time (in years) between COD and the other Am localities; τ_2 – divergence time (in years) between the two basins; θrF-A – population size reduction ratio during colonization; θrC-A – ratio between the current and the ancient population.

Parameter	RMSE	Spearman's ρ	Median	Interval
N_e	0.186	0.765	144 090	135 560–152 782
τ_1	0.612	0.621	89 489	84 205–96 683
τ_2	0.294	0.757	922 059	864 815–977 476
θrF-A	0.738	0.498	0.0571	0.054–0.059
θrC-A	0.665	0.516	0.5266	0.496–0.567

Appendix 1 Fig. A3), a similar pattern to that recovered with microsatellite markers (Farias et al. 2019), also suggesting substantial differences in the genetic distribution. Previous structure analyses of *A. gigas* populations from Am and lower To-Ar basins pointed to the absence of genetic structure (Hrbek et al. 2005, 2007) or resulted in highly separate genetic clusters in the two basins (Araripe et al. 2013, Farias et al. 2019). However, a recent study analyzing the genomic polymorphism through ddRAD sequencing of *A. gigas* including samples from Am, upper and lower To-Ar basins found a high genetic structure between the two basins, with the lower To-Ar showing mixed ancestry (Torati et al. 2019). We also found some genetic substructure within the Am basin, with COD being assigned to a separate group from the remaining localities. These results are in consonance with a hypothesis of more than one species be present in the genus *Arapaima* (Castello et al. 2013). However, we decided not to address formal taxonomic suggestions here, as we believe that an integrative species delimitation approach (Carstens et al. 2013b) would be necessary, coupling genomic data with other sources of information (e.g. morphological and cytogenetic), besides including specimens from all previously described morphotypes (Stewart 2013a, b).

The detected genetic diversity and population structure patterns can be related to several aspects of *Arapaima* biology. *Arapaima gigas* is considered a sedentary species (i.e. low migratory activity) with a preference for low-oxygenated lentic environments and having specialized parental care (Hrbek et al. 2005). The hydrological dynamics of the regions where *A. gigas* is currently found shows long flooding periods and flow cycles allowing the migration of these fishes to neighbor lakes within the same basin, a process known as lateral migration (Castello 2008, Farias et al. 2019). This migratory pattern can be responsible for the observed pattern of genetic groups including most or all samples within each analyzed basin, coupled with high genetic differentiation among populations both in intra and inter-basin comparisons (Supplementary material Appendix 1 Fig. A3). Moreover, along with several other fish species, this species has shown a decline in genetic diversity due to the loss of natural habitats and commercial over-exploitation (Allan et al. 2005, Castello et al. 2011). In fact, its obligate air-breathing behavior and the lentic environments where these fishes inhabit make it an easy target for fishing.

Demographic history

Our results indicate a scenario in which the ancient Am basin population colonized the To-Ar basin (Fig. 3). The parameter estimation step suggested that the colonization of To-Ar took place during the Pleistocene (τ_2; median = 922.06 kya). This estimate is older than the time periods used in our SDM projections, and caution is necessary to associate these results. These are complementary results, and the obtained estimate for To-Ar colonization is in agreement to the suggested Plio-Pleistocene age for the definitive splitting of the Am and To-Ar basins based on river sediments (Rossetti and Valeriano 2007). Although the ages estimated for the separation of these two basins from dated phylogenetic trees based on mtDNA in *Inia* (Hrbek et al. 2014) and in mtDNA and two nuclear markers in *Salminus* (Machado et al. 2018) were older than our estimates, the confidence intervals were also placed on the Plio-Pleistocene boundary. The estimated effective population size was also highly concordant with a previous estimation of this parameter using cpDNA markers (Hrbek et al. 2005). The remaining estimated parameter showed a lower accuracy when simulated data were evaluated by RMSE and Spearman's ρ, and they should be considered with care. The estimated values suggested a very recent separation of the COD population (τ_1; median = 89.49 kya; interval = 84.20–96.68 kya), a strong bottleneck during the foundation of the To-Ar basin (θrF-A; median = 0.0571; interval = 0.054–0.059), and a current population size for To-Ar that is approximately half of the size estimated for the Am basin (θrC-A; median = 0.5266; interval = 0.496–0.567). Among those estimates, the result for current population size of To-Ar was unexpectedly high. This is probably related to limitations of our method to estimate this parameter, as there is much more suitable habitat for the species in the Am basin (Fig. 2A). Also, the time estimate for the separation of the COD population was placed in the Pleistocene glaciations, which could be related to these events.

The model comparison and parameter estimation approach adopted here, based on CNN, allows taking information directly from the SNP matrices, without the use of summary statistics. Besides, contrary to other model testing approaches based on a rejection step that discards most of the simulations and retain only a small part that is more similar to the empirical data (e.g. ABC; Csillery et al. 2012), CNN use information from the whole set of simulations to learn how to distinguish among concurrent scenarios. These features resulted in a high capacity to distinguish among the simulated colonization models in our dataset (Supplementary material Appendix 1 Fig. A2).

The most likely demographic scenario recovered was the colonization of To-Ar basin from an ancient Am basin population, in accordance to the genetic diversity observed. Our analysis showed a more restricted distribution for *A. gigas* in the past, especially during glacial periods. Present distribution is more widespread and continuous, while future predictions indicate a distribution range shift towards south and a more fragmented distribution (Fig. 2B), potentially involving extinction of populations in central Amazon, as a result of global warming. Such scenario can result in invasive potential of new areas, increased by the presence of fish farms containing specimens located outside the natural occurrence area. This result is of special concern as *A. gigas* is one of the largest freshwater species and a generalist predator, characteristics that might cause a high impact in the invaded communities.

Data availability statement

Data available from the Dryad Digital Repository: <https://doi.org/10.5061/dryad.4qrfj6q7j> (Oliveira et al. 2020).

Funding – MFP was supported by Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) (proc. no. 2017/10240-0). MBC was supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (proc. no. 302449/2018-3), FAPESP (proc. no. 2018/22033-1) and CAPES/Alexander von Humboldt (proc. no. 88881.136128/2017-01). PR was supported by the project EXCELLENCE CZ.02.1.01/0.0/0.0/15_003/0000460 OP RDE. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

Author contributions – The two first authors contributed equally to this work; EAO, MFP, LACB and MBC designed research; EAO, MFP and MBC analyzed the data; EAO, MFP and MBC leaded the writing. LACB, CCG, PR, TE, FHSS, PV, EF and EHCO revised the manuscript.

Conflicts of interest – The authors declare no conflict of interests.

Permits – Collections were authorized by Brazilian environmental agencies ICMBIO/SISBIO (license no. 48290-1) and SISGEN (A96FF09). Procedures followed ethical and anesthesia conducts, in accordance with the Ethics Committee on Animal Experimentation of the Univ. Federal de São Carlos (process number CEUA 9506260315).

References

Adite, A. et al. 2005. Ontogenetic, seasonal and spatial variation in the diet of *Heterotis niloticus* (Osteoglossiformes: Osteoglossidae) in the Sô River and Lake Hlan, Benin, west Africa. – Environ. Biol Fish. 73: 367–378.

Albert, J. S. and Reis, R. E. 2011. Historical biogeography of neotropical freshwather fishes. – Univ. California Press.

Allan, J. D. et al. 2005. Overfishing of inland waters. – BioScience 55: 1041–1051.

Antonelli, A. et al. 2018a. Conceptual and empirical advances in Neotropical biodiversity research. – PeerJ 6: e5644.

Antonelli, A. et al. 2018b. Amazonia is the primary source of Neotropical biodiversity. – Proc. Natl Acad. Sci. USA 115: 6034–6039.

Arantes, C. C. et al. 2010. Population density, growth and reproduction of arapaima in an Amazonian river-floodplain. – Ecol. Freshwater Fish. 19: 455–465.

Araripe, J. et al. 2013. Dispersal capacity and genetic structure of *Arapaima gigas* on different geographic scales using microsatellite markers. – PLoS One 8: e54470.

Arratia, G. 1999. The monophyly of Teleostei and stem-group teleosts. – In: Arratia, G. and Schultze, H.-P. (eds), Mesozoic fishes 2 – systematics and fossil record. Verlag Dr. Friedrich Pfeil, pp. 265–334.

Bagley, J. C. et al. 2013. Paleoclimatic modeling and phylogeography of least killifish, *Heterandria formosa*: insights into Pleistocene expansion–contraction dynamics and evolutionary history of North American coastal plain freshwater biota. – BMC Evol. Biol. 13: 223.

Baird, N. A. et al. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. – PLoS One 3: e3376.

Bezerra, R. F. et al. 2013. Pirarucu, *Arapaima gigas*, the amazonian giant fish is briefly reviewed. – Nova Science Publishers.

Bonatelli, I. A. S. et al. 2014. Interglacial microrefugia and diversification of a cactus species complex: phylogeography and palaeodistributional reconstructions for *Pilosocereus aurisetus* and allies. – Mol. Ecol. 23: 3044–3063.

Breiman, L. 2001. Random forest. – Mach. Learn. 45: 5–32.

Breiman, L. et al. 1984. Classification and regression trees. – Chapman and Hall.

Busby, J. 1991. BIOCLIM – a bioclimate analysis and prediction system. – Plant Prot. Q. 6: 8–9.

Carstens, B. C. et al. 2013a. Model selection as a tool for phylogeographic inference: an example from the willow *Salix melanopsis*. – Mol. Ecol. 22: 4014–4028.

Carstens, B. C. et al. 2013b. How to fail at species delimitation. – Mol. Ecol. 22: 4369–4383.

Carvalho, T. P. and Albert, J. S. 2011. The Amazon–Paraguay divide. – In: Albert, J. S. and Reis, R. E. (eds), Historical biogeography of neotropical freshwater fishes. Univ. of California Press, pp. 193–202.

Castello, L. 2008. Nesting habitat of *Arapaima gigas* (Schinz) in Amazonian floodplains. – J. Fish Biol. 72: 1520–1528.

Castello, L. et al. 2011. Modeling population dynamics and conservation of arapaima in the Amazon. – Rev. Fish Biol. Fish. 21: 623–640.

Castello, L. et al. 2013. O que sabemos e precisamos fazer a respeito da conservação do pirarucu (*Arapaima* spp.) na Amazônia. – In: Figueiredo, E. S. A. (ed.), Biologia, conservação e manejo participativo de pirarucus na Pan-Amazônia. IDSM, pp. 17–32.

Cavin, L. 2017. Freshwater fishes: 250 million years of evolutionary history. – ISTE Press Elsevier.

Christin, S. et al. 2019. Applications for deep learning in ecology. – Methods Ecol. Evol. 10: 1632–1644.

Conesa, A. et al. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. – Bioinformatics 21: 3674–3676.

Csillery, K. et al. 2012. abc: an R package for approximate Bayesian computation (ABC). – Methods Ecol. Evol. 3: 475–79.

Dagosta, F. C. P. and Pinna, M. de 2017. Biogeography of Amazonian fishes: deconstructing river basins as biogeographic units. – Neotrop. Ichthyol. 15: 1–24.

dos Santos Fogaça, F. H. et al. 2011. Yield and composition of pirarucu fillet in different weight classes. – Acta Sci. Anim. Sci. 33: 95–99.

Elshire, R. J. et al. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. – PLoS One 6: e19379.

Farias, I. P. et al. 2019. The largest fish in the world's biggest river: genetic connectivity and conservation of *Arapaima gigas* in the Amazon and Araguaia-Tocantins drainages. – PLoS One 14: e0220882.

Figueiredo, J. et al. 2009. Late Miocene onset of the Amazon River and the Amazon deep-sea fan: evidence from the Foz do Amazonas Basin. – Geology 37: 619–622.

Figueiredo, J. et al. 2010. Late Miocene onset of the Amazon River and the Amazon deep-sea fan: evidence from the Foz do Amazonas Basin: reply. – Geology 37: 619–622.

Fine, P. V. A. and Lohmann, L. G. 2018. Importance of dispersal in the assembly of the Neotropical biota. – Proc. Natl Acad. Sci. USA 115: 5829–5831.

Flagel, L. et al. 2019. The unreasonable effectiveness of convolutional neural networks in population genetic inference. – Mol. Biol. Evol. 36: 220–238.

Foll, M. and Gaggiotti, O. 2008. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a bayesian perspective. – Genetics 180: 977–993.

Friedman, J. 1991. Multivariate adaptive regression splines (with discussion). – Ann. Stat. 19: 1–67.

Garrick, R. C. et al. 2015. The evolution of phylogeographic data sets. – Mol. Ecol. 24: 1164–1171.

Greenwood, P. H. et al. 1966. Phyletic studies of teleostean fishes, with a provisional classification of living forms. – Bull. Am. Mus. Nat. Hist. 131: 339–456.
Gruber, B. et al. 2018. dartr: an r package to facilitate analysis of SNP data generated from reduced representation genome sequencing. – Mol. Ecol. Resour. 18: 691–699.
Guillot, G. et al. 2011. Population genetics analysis using R and Geneland. – DTU Library.
Hastie, T. et al. 1994. Flexible discriminant analysis by optimal scoring. – J. Am. Stat. Assoc. 89: 1255–1270.
Hijmans, R. J. et al. 2005. Very high resolution interpolated climate surfaces for global land areas. – Int. J. Climatol. 25: 1965–1978.
Hilton, E. J. and Lavoué, S. 2018. A review of the systematic biology of fossil and living bony-tongue fishes, Osteoglossomorpha (Actinopterygii: Teleostei). – Neotrop. Ichthyol. 16: 1–35.
Hoorn, C. et al. 2017. The Amazon at sea: onset and stages of the Amazon River from a marine record, with special reference to Neogene plant turnover in the drainage basin. – Global Planet Change 153: 51–65.
Hrbek, T. et al. 2005. Population genetic analysis of *Arapaima gigas*, one of the largest freshwater fishes of the Amazon basin: implications for its conservation. – Anim. Conserv. 8: 297–308.
Hrbek, T. et al. 2007. Conservation strategies for *Arapaima gigas* (Schinz, 1822) and the Amazonian várzea ecosystem. – Braz. J. Biol. 67: 909–917.
Hrbek, T. et al. 2014. A new species of river dolphin from Brazil or: how little do we know our biodiversity. – PLoS One 9: e83623.
Hudson, R. R. 2002. Generating samples under a Wright–Fisher neutral model of genetic variation. – Bioinformatics 18: 337–338.
Jaccoud, D. et al. 2001. Diversity arrays: a solid state technology for sequence information independent genotyping. – Nucleic Acids Res. 29: e25.
Jakobsson, M. and Rosenberg, N. A. 2007. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. – Bioinformatics 23: 1801–1806.
Keenan, K. et al. 2013. DiveRsity: an R package for the estimation and exploration of population genetics parameters and their associated errors. – Methods Ecol. Evol. 4: 782–788.
Kilian, A. et al. 2012. Diversity arrays technology: a generic genome profiling technology on open platforms. – In: Pompanon, F. and Bonin, A. (eds), Data production and analysis in population genomics. Humana Press, pp. 67–89.
Kingma, D. P. and Ba, J. L. 2015. Adam: a method for stochastic optimization. – ICLR: International Conference on Learning Representations.
Knowles, L. L. 2009. Statistical phylogeography. – Annu. Rev. Ecol. Evol. Syst. 40: 593–612.
Kopelman, N. M. et al. 2015. Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. – Mol. Ecol. Resour. 15: 1179–1191.
Lande, R. 1988. Genetics and demography in biological conservation. – Science 241: 1455–1460.
Lovejoy, T. E. and Nobre, C. 2019. Amazon tipping point: last chance for action. – Sci. Adv. 5: eaba2949.
Machado, C. B. et al. 2018. Bayesian analyses detect a history of both vicariance and geodispersal in Neotropical freshwater fishes. – J. Biogeogr. 45: 1313–1325.
Mantel, N. 1967. The detection of disease clustering and a generalized regression approach. – Cancer Res. 27: 209–220.
McCullagh, P. and Nelder, J. A. 1989. Generalized linear models. – Chapman and Hall/CRC Press.
McMahan, C. D. et al. 2017. Pleistocene to holocene expansion of the black-belt cichlid in Central America, *Vieja maculicauda* (Teleostei: Cichlidae). – PLoS One 12: e0178439.
Meirmans, P. G. 2015. Seven common mistakes in population genetics and how to avoid them. – Mol. Ecol. 24: 3223–3231.
Meirmans, P. G. and van Tienderen, P. H. 2004. Genotype and genodive: two programs for the analysis of genetic diversity of asexual organisms. – Mol. Ecol. Notes 4: 792–794.
Melville, J. et al. 2017. Identifying hybridization and admixture using SNPs: application of the DArTseq platform in phylogeographic research on vertebrates. – R. Soc. Open Sci. 4: 161061.
Moncrieff, G. R. et al. 2016. The future distribution of the savannah biome: model-based and biogeographic contingency. – Phil. Trans. R. Soc. B 371: 1–10.
Myers, G. S. 1949. Salt-tolerance of fresh-water fish groups in relation to zoogeographical problems. – Contrib. Zool. 28: 315–322.
Near, T. J. et al. 2012. Resolution of ray-finned fish phylogeny and timing of diversification. – Proc. Natl Acad. Sci. USA 109: 13698–13703.
Nelson, J. S. et al. 2016. Fishes of the world. – Wiley.
Nobre, C. A. et al. 2016. Land-use and climate change risks in the Amazon and the need of a novel sustainable development paradigm. – Proc. Natl Acad. Sci. USA 113: 10759–10768.
Oberdorff, T. et al. 2015. Opinion paper: how vulnerable are Amazonian freshwater fishes to ongoing climate change? – J. Appl. Ichthyol. 31: 4–9.
Oliveira, E. A. et al. 2020. Data from: Historical demography and climate driven distributional changes in a widespread Neotropical freshwater species with high economic importance. – Dryad Digital Repository, <https://doi.org/10.5061/dryad.4qrfj6q7j>.
Ono, E. A. 2007. Perspectivas para o aumento da oferta de juvenis de pirarucu. – Panorama Aquic. 17: 45–47.
Orsini, L. et al. 2012. Genomic signature of natural and anthropogenic stress in wild populations of the waterflea *Daphnia magna*: validation in space, time and experimental evolution. – Mol. Ecol. 21: 2160–2175.
Perez, M. F. et al. 2016. Model-based analysis supports interglacial refugia over long-dispersal events in the diversification of two South American cactus species. – Heredity 116: 550–557.
Phillips, S. J. et al. 2006. Maximum entropy modeling of species geographic distributions. – Ecol. Model. 190: 231–259.
Pritchard, J. K. et al. 2000. Inference of population structure using multilocus genotype data. – Genetics 155: 945–959.
Raj, A. et al. 2014. fastSTRUCTURE: variational inference of population structure in large SNP data sets. – Genetics 197: 573–589.
Reis, R. E. et al. 2003. Check list of the freshwater fishes of South and Central America. – EdiPUCRS.
Reis, R. E. et al. 2016. Fish biodiversity and conservation in South America. – J. Fish Biol. 89: 12–47.
Ribas, C. C. et al. 2012. A palaeobiogeographic model for biotic diversification within Amazonia over the past three million years. – Proc. R. Soc. B 279: 681–689.
Ridgeway, G. 1999. The state of boosting. – Comput. Sci. Stat. 31: 172–181.
Ripley, B. D. 2014. Pattern recognition and neural networks. – Cambridge Univ. Press.
Rosenberg, N. A. 2004. DISTRUCT: a program for the graphical display of population structure. – Mol. Ecol. Notes 4: 137–138.

Rossetti, D. F. and Valeriano, M. M. 2007. Evolution of the lowest amazon basin modeled from the integration of geological and SRTM topographic data. – CATENA 70: 253–265.
Rull, V. 2018. Neotropical diversification: historical overview and conceptual insights. – PeerJ Preprints.
Sambrook, J. and Russell, D. W. 2001. Molecular cloning, a laboratory manual. – Cold Spring Harbor Laboratory Press.
Schrider, D. R. and Kern, A. D. 2018. Supervised machine learning for population genetics: a new paradigm. – Trends Genet. 34: 301–312.
Sheehan, S. and Song, Y. S. 2016. Deep learning for population genetic inference. – PLoS Comput. Biol. 12: e1004845.
Silva-Santos, R. et al. 2018. Molecular evidences of a hidden complex scenario in *Leporinus* cf. *friderici*. – Front. Genet. 9: 47.
Stewart, D. J. 2013a. Re-description of *Arapaima agassizii* (Valenciennes), a rare fish from Brazil (Osteoglossomorpha: Osteoglossidae). – Copeia 2013: 38–51.
Stewart, D. J. 2013b. A new species of *Arapaima* (Osteoglossomorpha: Osteoglossidae) from the Solimões River, Amazonas state, Brazil. – Copeia 2013: 470–476.
Stone, R. 2007. The last of the leviathans. – Science 316: 1684–1688.
Thomé, M. T. C. and Carstens, B. C. 2016. Phylogeographic model selection leads to insight into the evolutionary history of four-eyed frogs. – Proc. Natl Acad. Sci. USA 113: 8010–8017.
Thuiller, W. et al. 2009. BIOMOD – a platform for ensemble forecasting of species distributions. – Ecography 32: 369–373.
Torati, L. S. et al. 2019. Genetic diversity and structure in *Arapaima gigas* populations from Amazon and Araguaia-Tocantins river basins. – BMC Genet. 20: 13.
van Orsouw, N. J. et al. 2007. Complexity reduction of polymorphic sequences (CRoPS): a novel approach for large-scale polymorphism discovery in complex genomes. – PLoS One 2: e1172.
van Soelen, E. E. et al. 2017. A 30 Ma history of the Amazon River inferred from terrigenous sediments and organic matter on the Ceará Rise. – Earth Planet Sci. Lett. 474: 40–48.
Vialle, R. A. et al. 2018. Whole genome sequencing of the Pirarucu (*Arapaima gigas*) supports independent emergence of major teleost clades. – Genome Biol. Evol. 10: 2366–2379.
Villanea, F. A. and Schraiber, J. G. 2019. Neanderthal and modern humans. – Nat. Ecol. Evol. 3: 39–44.
Vitorino, C. A. et al. 2015. Genetic diversity of *Arapaima gigas* (Schinz, 1822) (Osteoglossiformes: Arapaimidae) in the Araguaia-Tocantins basin estimated by ISSR marker. – Neotrop. Ichthyol. 13: 557–568.
Vitorino, C. A. et al. 2017. Low genetic diversity and structuring of the arapaima (Osteoglossiformes, Arapaimidae) population of the Araguaia-Tocantins basin. – Front. Genet. 8: 1–10.
Weeks, A. R. et al. 2011. Assessing the benefits and risks of translocations in changing environments: a genetic perspective. – Evol. Appl. 4: 709–725.
Weir, B. S. and Cockerham, C. C. 1984. Estimating F-statistics for the analysis of population structure. – Evolution 38: 1358–1370.
Wright, S. 1931. Evolution in Mendelian populations. – Genetics 16: 97–159.

Supplementary material (available online as Appendix ecog-04874 at <www.ecography.org/appendix/ecog-04874>). Appendix 1.